qwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmrtyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmrtyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmrtyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmrtyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmrtyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmrtyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmrtyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnm

|  |
| --- |
| Προηγμένη Εξόρυξη Δεδομένων  Τελική Εργασία  Ακ. Έτος: 2008-2009  Απόστολος Κρητικός (249) |

Περιεχόμενα

[Εισαγωγή 3](#_Toc224055926)

[Περιγραφή του προβλήματος 3](#_Toc224055927)

[Δομή του εγγράφου 3](#_Toc224055928)

[Εφαρμογή τεχνικών εξόρυξης δεδομένων στο δοσμένο σύνολο δεδομένων 4](#_Toc224055929)

[Ζητούμενο 1ο: 4](#_Toc224055930)

[Απάντηση: 4](#_Toc224055931)

[Ζητούμενο 2ο: 6](#_Toc224055932)

[Απάντηση: 7](#_Toc224055933)

[Ζητούμενο 3ο: 10](#_Toc224055934)

[Απάντηση: 10](#_Toc224055935)

[Ζητούμενο 4ο: 13](#_Toc224055936)

[Απάντηση: 13](#_Toc224055937)

# Εισαγωγή

Καλούμαστε να απαντήσουμε στο πρόβλημα που περιγράφεται αμέσως μετά κάνοντας χρήση της θεωρίας που διδάχθηκε στα πλαίσια του μαθήματος «Προηγμένη Εξόρυξη Δεδομένων» στα πλαίσια του Μεταπτυχιακού Προγράμματος Σπουδών του τμήματος Πληροφορικής του Α.Π.Θ. με εισηγητή των κ. Συμεωνίδη Παναγιώτη.

## Περιγραφή του προβλήματος

Ας υποθέσουμε ότι είμαστε οι διαχειριστές της  βάσης δεδομένων της εταιρίας FoodMart. ΆΗ FoodMart είναι μια μεγάλη αλυσίδα παντοπωλείων με πωλήσεις στις Ηνωμένες Πολιτείες, το Μεξικό, και τον Καναδά. Το εμπορικό τμήμα της εταιρείας FoodMart θέλει να αναλύσει όλες τις πωλήσεις των προϊόντων της και την αγοραστική συμπεριφορά των πελατών της που έγιναν κατά τη διάρκεια του ημερολογιακού έτους του 1997.

Ά

## Δομή του εγγράφου

Παρακάτω απαντούμε στα ερωτήματα με τη σειρά που ζητούνται. Σε κάθε τμήμα παρατίθεται το ζητούμενο και αμέσως μετά η απάντηση, εμπλουτισμένη με γραφήματα, screenshots ή ότι άλλο κρίνετε απαραίτητο.

# Εφαρμογή τεχνικών εξόρυξης δεδομένων στο δοσμένο σύνολο δεδομένων

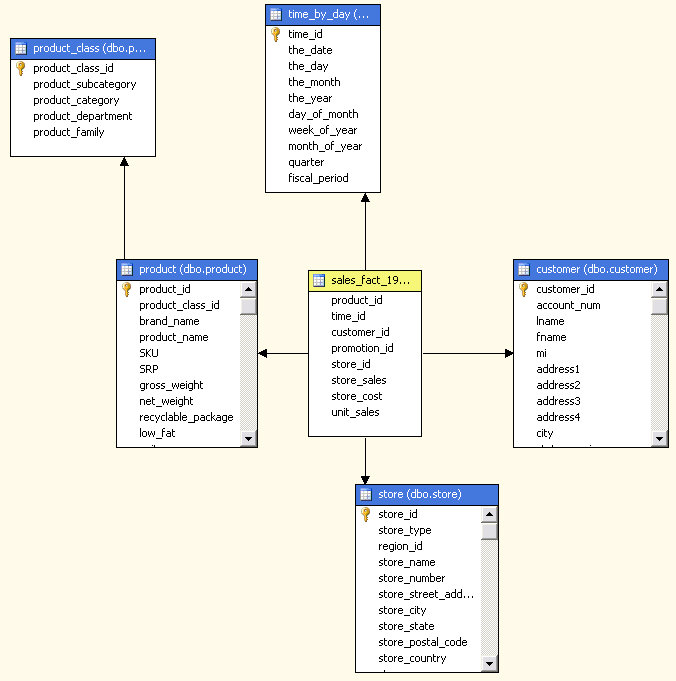
## Ζητούμενο 1ο:

Χρησιμοποιώντας τα στοιχεία που αποθηκεύονται στην βάση δεδομένων της επιχείρησης,  να χτίσετε μια πολυδιάστατη δομή δεδομένων (ένα κύβο) για να επιτρέψουμε τους γρήγορους χρόνους απόκρισης της βάσης, όταν ρωτούν σε αυτήν οι εμπορικοί αναλυτές της εταιρείας.

1. Δημιουργήστε Ά ένα κύβο (Cube) με τα εξής στοιχεία:
   1. Πίνακας γεγονότων: Sales\_fact\_1997
   2. Πίνακες διαστάσεων: Product, Time By Day, Store, Customer
   3. Μετρικά: store\_sales, store\_cost και unit\_sales.
2. Απεικονίστε στο Excel γραφήματα για τα αποτελέσματα για τα εξής :
   1. Πωλήσεις και κόστος πωλήσεων της εταιρείας για το πρώτο τρίμηνο του 1997 ανά πολιτεία της USA.
   2. Ποια 5 καταστήματα σημείωσαν τις μεγαλύτερες πωλήσεις για το πρώτο τρίμηνο του 1997;
   3. Ποια 5 προϊόντα σημείωσαν τις μεγαλύτερες πωλήσεις το μήνα Ιανουάριο του 1997;

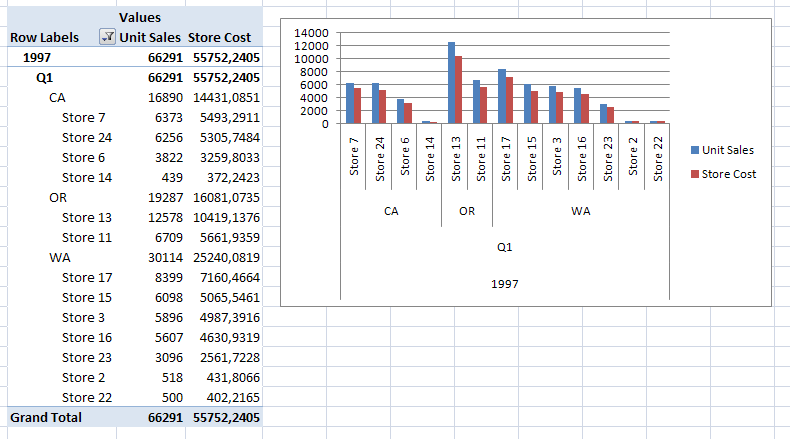
## Απάντηση:

1. Αρχικά δημιουργώ τον κύβο με βάση τα στοιχεία που αναφέρονται στο (1). Το σχήμα του κύβου που παράχθηκε φαίνεται παρακάτω γραφικά.

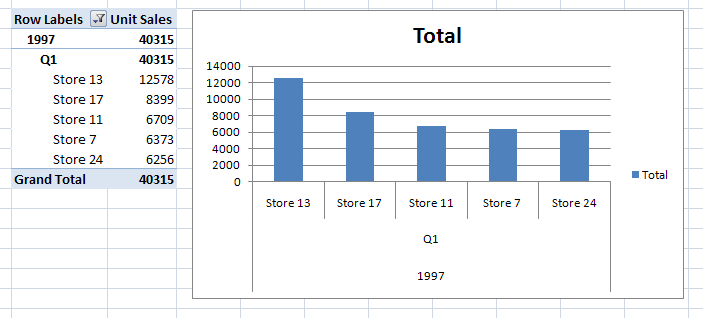


Στη συνέχεια πραγματοποιούμε σύνδεση με τον OLAP SQL ‐Server μέσω του Excel 2007 κάνοντας εισαγωγή εξωτερικών δεδομένων και πλέον έχουμε πρόσβαση στον κύβο που μόλις δημιουργήσαμε.

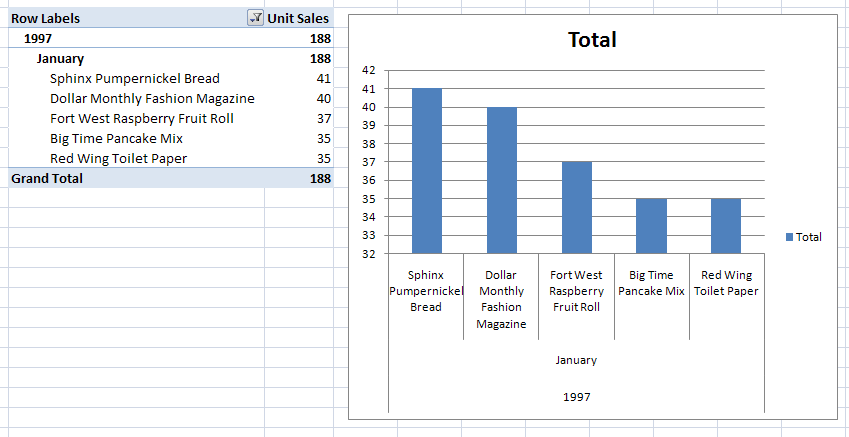
1. Έτσι μπορούμε να τρέξουμε τα ερωτήματα:
   1. Πωλήσεις και κόστος πωλήσεων της εταιρείας για το πρώτο τρίμηνο του 1997 ανά πολιτεία της USA.



* 1. Ποια 5 καταστήματα σημείωσαν τις μεγαλύτερες πωλήσεις για το πρώτο τρίμηνο του 1997;



* 1. Ποια 5 προϊόντα σημείωσαν τις μεγαλύτερες πωλήσεις το μήνα Ιανουάριο του 1997;



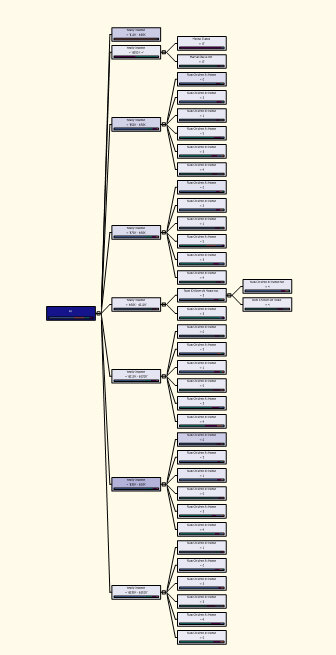
## Ζητούμενο 2ο:

Το εμπορικό τμήμα θέλει να αυξήσει την ικανοποίηση των πελατών και το ποσοστό διατήρησης αυτών. Το πρόγραμμα καρτών μέλους θα επαναπροσδιοριστεί για να εξυπηρετήσει τους πελάτες καλύτερα και για να παρέχει υπηρεσίες που εκπληρώνουν περισσότερο τις προσδοκίες των πελατών. Προκειμένου να επαναπροσδιοριστεί το πρόγραμμα Κάρτας Μέλους, το εμπορικό τμήμα θέλει να αναλύσει τις συναλλαγές πωλήσεων και να ανακαλύψει τα πρότυπα μεταξύ των δημογραφικών πληροφοριών των πελατών (φύλλο, συζυγική κατάσταση, ετήσιο εισόδημα, αριθμός παιδιών στο σπίτι, αριθμός αυτοκινήτων, εκπαίδευση) και της κάρτας που αυτοί χρησιμοποίησαν.  Με αυτήν την γνώση, οι κάρτες θα επαναπροσδιοριστούν βασισμένες στα χαρακτηριστικά των πελατών που τις χρησιμοποίησαν. Να δημιουργηθεί ένα μοντέλο για να εκπαιδευτούν τα στοιχεία των πωλήσεων και να χρησιμοποιηθεί ο αλγόριθμος Decision Tree της Microsoft για να βρεθούν τα κύρια στοιχεία που συντελούν στην επιλογή μιας  κάρτας μέλους(π.χ. χρυσή κάρτα μέλους, ασημένια κάρτα μέλους κτλ.) . Η κύρια διάσταση που θα χρησιμοποιηθεί θα είναι ο πίνακας των Πελατών, ενώ η κύρια ιδιότητα που θα διερευνηθεί θα είναι αυτή της κάρτας μέλους (Member\_Card). Κατόπιν να επιλεχθεί  ένας κατάλογος δημογραφικών χαρακτηριστικών από τα οποία ο αλγόριθμος (Decision Tree της Microsoft) θα καθορίσει τα πρότυπα: Ενδεικτικά αναφέρονται τα παρακάτω χαρακτηριστικά : φύλλο, συζυγική κατάσταση, ετήσιο εισόδημα, αριθμός παιδιών στο σπίτι, αριθμός αυτοκινήτων και εκπαίδευση.

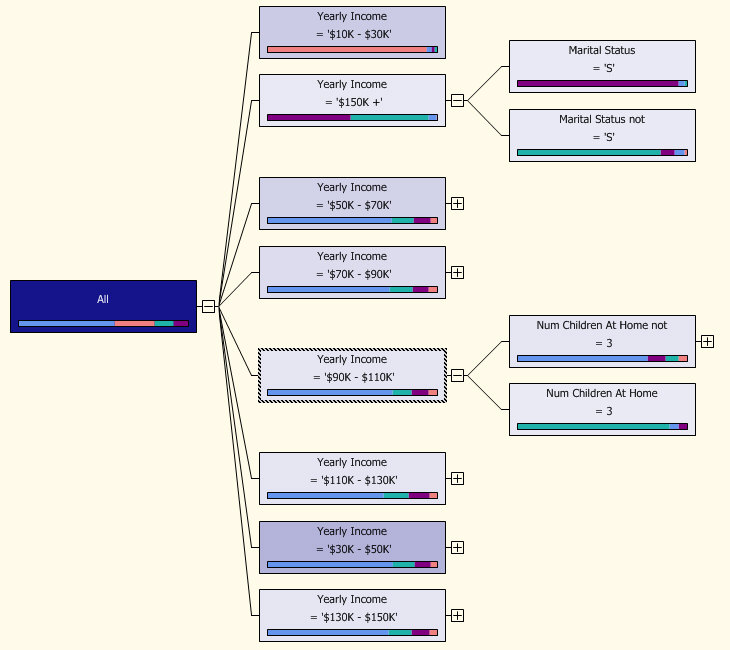
1. Δημιουργήστε 2 δέντρα απόφασης για πρόβλεψη της ιδιότητας Member Card από τις ιδιότητες: Gender, Marital Status, Num Cars Owned, Num Children At Home, Yearly Income, Education.
2. Προτείνεται εκείνο το μοντέλο δέντρου απόφασης που πετυχαίνει ακριβέστερο προσδιορισμό των κατόχων χρυσής κάρτας.

## Απάντηση:

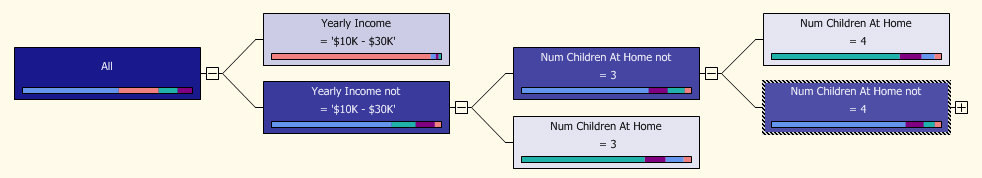
1. Αρχικά δημιουργώ το δένδρο χρησιμοποιώντας τις default τιμές, ακολουθώντας πιστά τις οδηγίες του εγχειριδίου του εργαστηριακού μαθήματος. Καταλήγω στην παρακάτω κατάσταση:



Για εποπτικούς λόγους παραθέτουμε και μια απεικόνιση με λιγότερα επίπεδα:

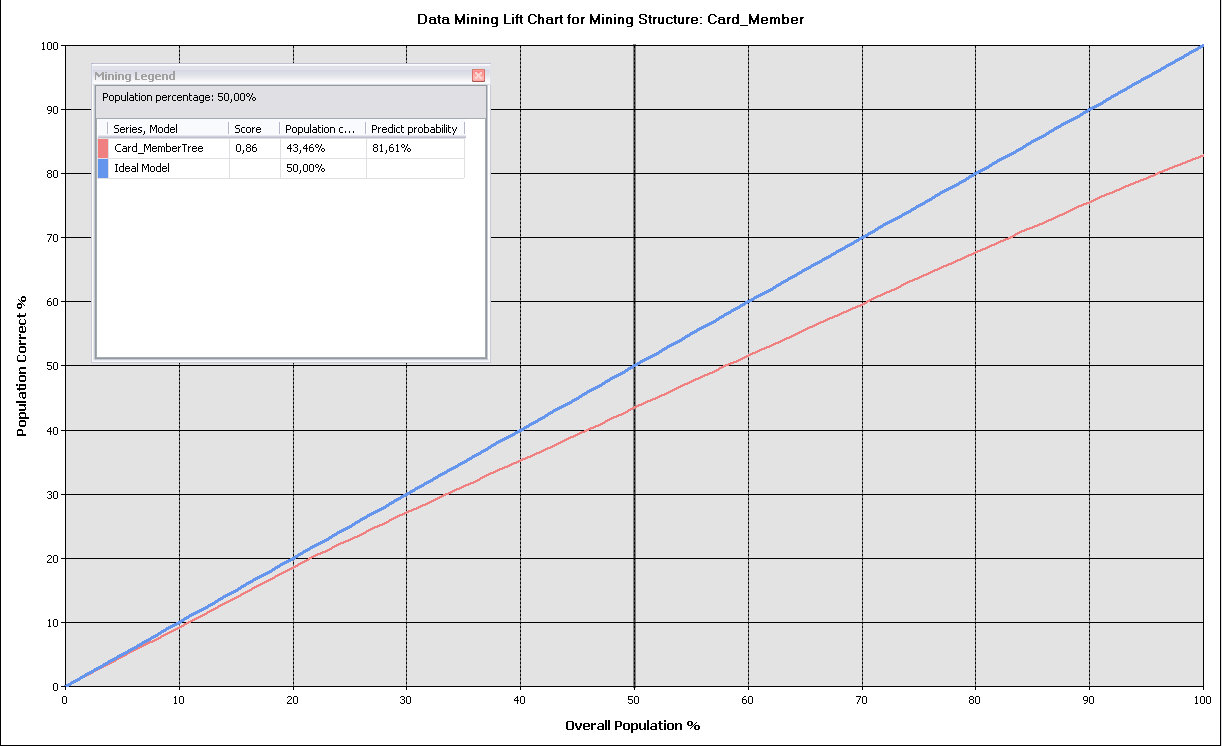


Τώρα κατασκευάζουμε ένα δεύτερο δένδρο επεμβαίνοντας στην μεταβλητή Minimum Support και αναθέτοντας της την τιμή 300.

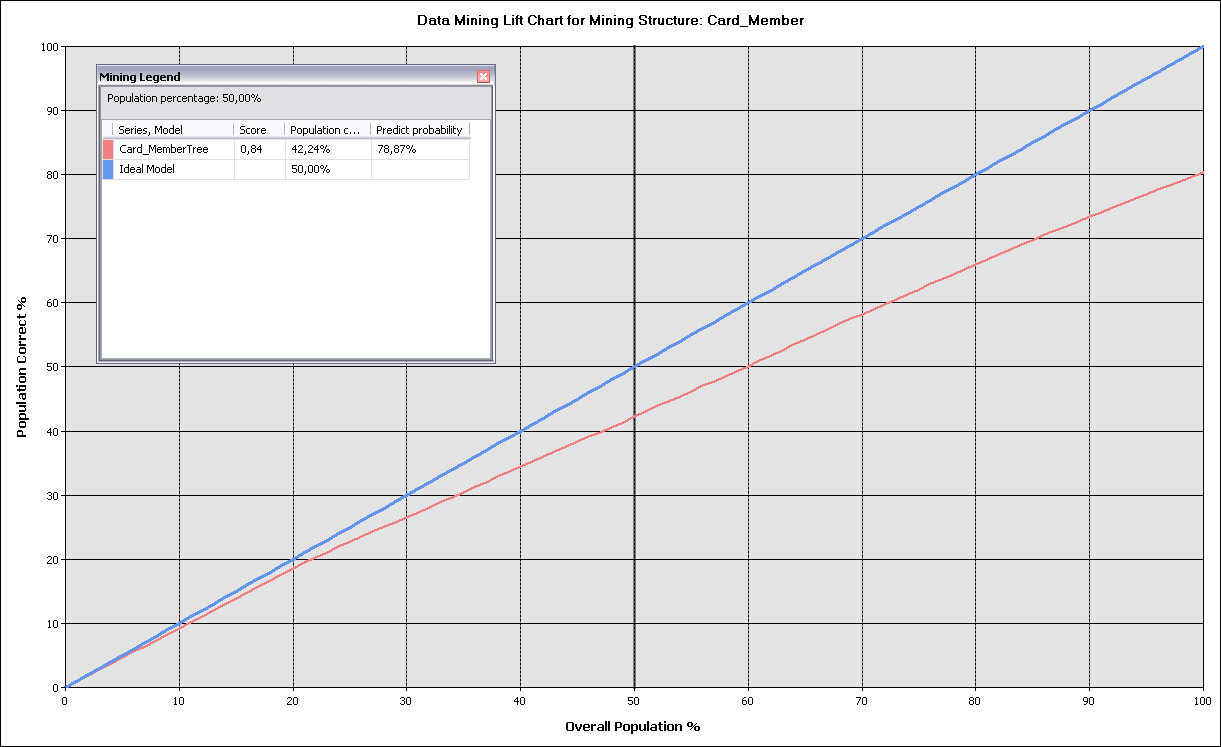


Η παράμετρος αυτή προσδιορίζει το ελάχιστο όριο της υποστήριξης – support, για να γίνει ένα itemset δεκτό, στην πρώτη φάση του αλγορίθμου Microsoft Association Algorithm.

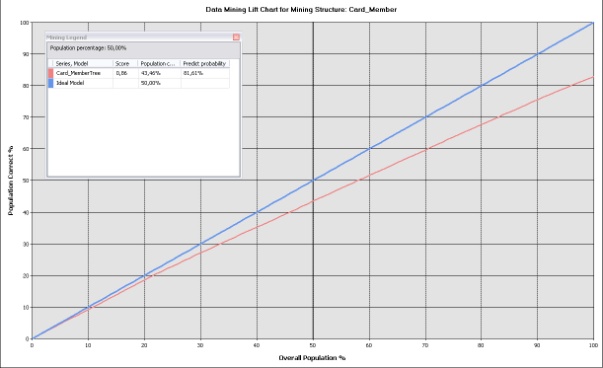
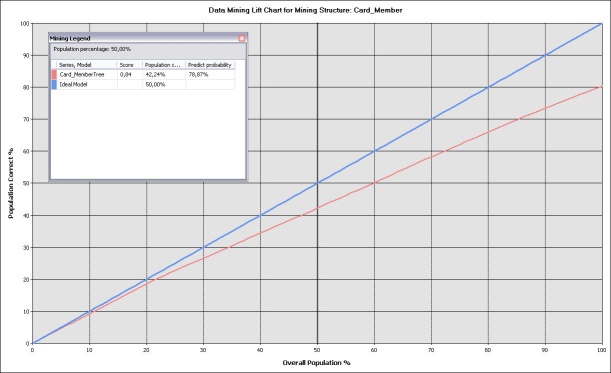
1. Παραθέτουμε εδώ τα LiftCharts των δύο δένδρων για να κάνουμε τη σύγκριση.



***LiftChart πρώτης προσέγγισης***



***LiftChart δεύτερης προσέγγισης***

***LiftChart Πρώτης vs LiftChart Δεύτερης προσέγγισης***

Από την αντιπαράθεση των δύο LiftCharts γίνεται φανερό στον αναγνώστη ότι ή δεύτερη προσέγγιση είναι χειρότερη της πρώτης. Μετά από εξαντλητική επιβολή αλλαγών αλλά και συνδυαστικών αλλαγών στις παραμέτρους για παραγωγή δένδρων δεν μπορέσαμε να εντοπίσουμε κάποια φόρμουλα που να οδηγεί σε καλύτερο αποτέλεσμα. Έτσι παραθέσαμε το αποτέλεσμα μιας από τις πολλές μας προσπάθειες και καταγράφουμε την παρατήρησή μας αυτή.

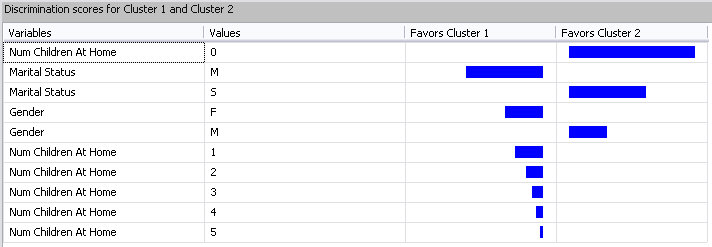
## Ζητούμενο 3ο:

Το εμπορικό τμήμα της εταιρείας FoodMart έχει καθορίσει ένα χρηματικό προϋπολογισμό για να δημιουργήσει τρεις εκδόσεις του εβδομαδιαίου ενημερωτικού περιοδικού που εκδίδει.  ΘΆέλει να τρέξει μερικές διαδικασίες εξόρυξης δεδομένων, μέσω των στοιχείων πωλήσεων που διαθέτει, για να προσδιορίσει τρεις ομάδες πελατών. ΆΜε βάση τα χαρακτηριστικά αυτών των ομάδων, θα είναι σε θέση να επιλέξει  τον τύπο των διαφημίσεων και προσφορών  που θα παρεμβάλλει σε κάθε έκδοση του εβδομαδιαίου του περιοδικού. ΆΘα είναι σε θέση, επίσης, να ξέρει  σε ποια κατηγορία πελατών θα αποσταλεί ή κάθε μία από τις τρεις εκδόσεις του περιοδικού.

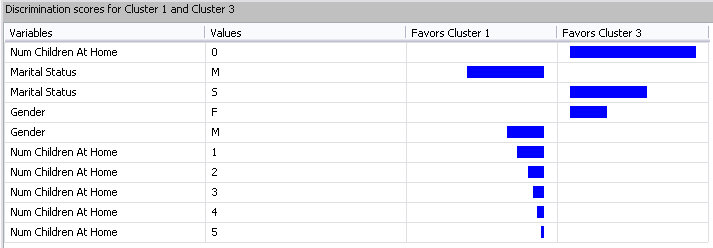
1. Δημιουργείστε με τον k-Means 2 μοντέλα ομαδοποίησης με 3 ομάδες πελατών βάσει των δημογραφικών τους χαρακτηριστικών. Ενδεικτικά αναφέρονται τα παρακάτω δημογραφικά χαρακτηριστικά: φύλλο, συζυγική κατάσταση, ετήσιο εισόδημα, αριθμός παιδιών στο σπίτι, αριθμός αυτοκινήτων και εκπαίδευση.
2. Δώστε τα αποτελέσματα σύγκρισης μεταξύ Cluster1 και Cluster2, Cluster1 και Cluster3, και Cluster2 και Cluster3.
3. Προσδιορίστε ποιο από τα δύο μοντέλα τελικά προτείνετε καθώς επίσης και τους λόγους της απόφασης σας.

## Απάντηση:

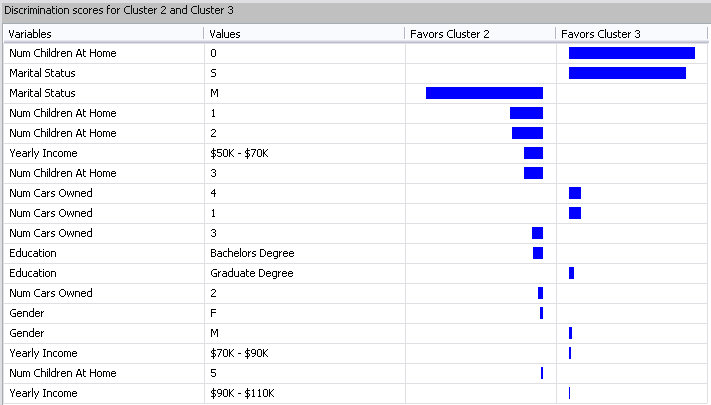
Αντιμετωπίζω τα υποερωτήματα (1) + (2) μαζί. Τα μοντέλα γίνονται εμφανή με βάσει τους παρακάτω discrimination πίνακες.



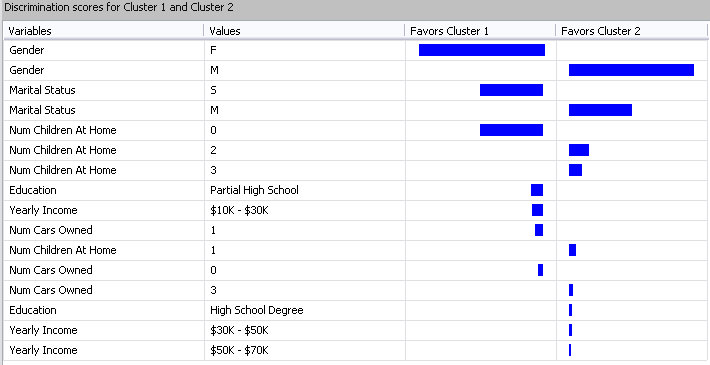
***Μοντέλο 1ο – Clusters 1 & 2***



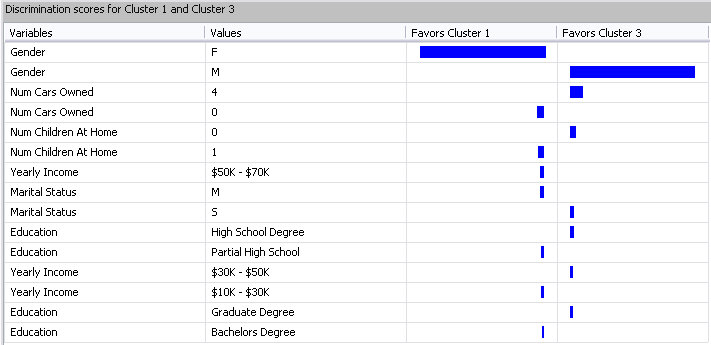
***Μοντέλο 1ο – Clusters 1 & 3***



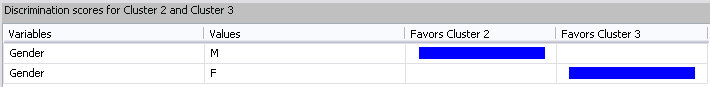
***Μοντέλο 1ο – Clusters 2 & 3***



***Μοντέλο 2ο – Clusters 1 & 2***



***Μοντέλο 2ο – Clusters 1 & 3***

******

***Μοντέλο 2ο – Clusters 2 & 3***

Αξίζει να σημειωθεί ότι στο δεύτερο μοντέλο επιβάλαμε την τιμή 50 στην μεταβλητή MODELING\_CARDINALITY. Η παράμετρος αυτή επηρεάζει τον αριθμό των υποψήφιων μοντέλων που θα δημιουργηθούν από τον αλγόριθμο που κάνει clustering. Το 50 είναι ο μέγιστος αριθμός.

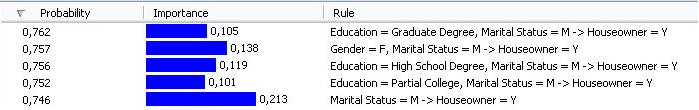
3. Αν παρατηρήσει κανείς τους discrimination matrices που παραθέσαμε για το πρώτο και δεύτερο μοντέλο αντίστοιχα (σε αντιπαραβολή) θα διαπιστώσει ότι στο δεύτερο μοντέλο η διαφοροποίηση μεταξύ των clusters είναι εμφανέστερη ενώ το χαρακτηριστικό στο οποίο διαφέρουν πρόδηλο. Στο πρώτο μοντέλο, αν και η διαφοροποίηση είναι επίσης εμφανής, δεν έχουμε τόσο καλή διαφάνεια ενώ περισσότερα του ενός χαρακτηριστικά εμπλέκονται.

## Ζητούμενο 4ο:

Θέλουμε να βρούμε συσχετίσεις μεταξύ των ιδιοτήτων των πελατών. Από την διάσταση Customer του προηγούμενου κύβου, δημιουργήστε Association Rules από τις εξής ιδιότητες (θα είναι και input και predictable):

1. City, Education, Gender, Houseowner, Marital Status (support, confidence αφήστε τις default τιμές)
2. Αναφέρετε τους 5 κανόνες με την μεγαλύτερη τιμή Probability
3. Αποτυπώστε το Dependency Network

## Απάντηση:

1. 

2. 5 top-probability rules:

Rule 1: Αν κάποιος έχει τελειώσει ανώτατη σχολή και είναι παντρεμένος, έχει δική του κατοικία. (0,762 - probability)

Rule 2: Αν μία γυναίκα είναι παντρεμένη, έχει δική της κατοικία (0,757 - probability)

Rule 3: Αν κάποιος έχει τελειώσει λύκειο και είναι παντρεμένος, έχει δική του οικεία (0,756 - probability)

Rule 4: Αν κάποιος έχει φοιτήσει μερικώς σε κολλέγιο και είναι παντρεμένος έχει δική του κατοικία (0,752 - probability)

Rule 5: Αν κάποιος είναι παντρεμένος, έχει δική του κατοικία (0,746 - probability)

3.

