Πανεπιστήμιο Πειραιώς – Τμήμα Πληροφορικής
Πρόγραμμα Μεταπτυχιακών Σπουδών
«Προηγμένα Συστήματα Πληροφορικής»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**Επιχειρηματική Ευφυΐα και Τεχνικές Εξόρυξης Γνώσης. Οι λύσεις που παρέχονται από τα SSAS.**
Ονοματεπώνυμο Φοιτητή	**Παπαοικονόμου Εμμανουήλ του Χρήστου**
Αριθμός Μητρώου	**ΜΠΣΠ/07028**
Κατεύθυνση	**Συστήματα Υποστήριξης Αποφάσεων**
Επιβλέπων	**Δημήτρης Δεσπότης, Καθηγητής**

Ημερομηνία Παράδοσης **Νοέμβριος 2010**

Τριμελής Εξεταστική Επιτροπή

(υπογραφή)	(υπογραφή)	(υπογραφή)
Δ. Δεσπότης Καθηγητής	Ν. Αλεξανδρής Καθηγητής	Θ. Παναγιωτόπουλος Καθηγητής

Περίληψη

Στην παρούσα εργασία θα παρουσιαστεί η έννοια της Επιχειρηματικής Ευφυΐας (Business Intelligence – BI), οι πληροφορίες που χρειάζονται για την χρήση των εφαρμογών BI σε ένα οργανισμό, οι λύσεις που προσφέρονται από την Microsoft και η σχέση του BI με την Εξόρυξη Γνώσης. Θα παρουσιαστούν επίσης, οι κυριότερες τεχνικές Εξόρυξης Γνώσης και τα βασικότερα επιχειρησιακά προβλήματα που μπορούν να επιλυθούν με την χρήση τεχνικών αυτών. Θα δείξουμε τις λύσεις BI που προσφέρει η Microsoft μέσα από τα Analysis Services (SSAS) του SQL Server (2005/2008) και πιο συγκεκριμένα θα δείξουμε τα κυριότερα χαρακτηριστικά των βασικών αλγορίθμων Εξόρυξης Γνώσης που μπορούν να υλοποιηθούν από τα SSAS (Decision Trees, Clustering, Association Rules), τις παραμέτρους που ρυθμίζουν τον κάθε αλγόριθμο και τις εντολές DMX που μπορούν να χρησιμοποιηθούν για την υλοποίηση τους. Τέλος, θα παρουσιαστεί η ανάπτυξη μιας εφαρμογής χρήσης των μοντέλων Εξόρυξης Γνώσης μέσα από το Business Intelligence Development Studio (BIDS) και τα SSAS του SQL Server χρησιμοποιώντας την μεθοδολογία ανάπτυξης έργων Εξόρυξης Γνώσης CRISP-DM.

Abstract

In this work we will present the concept of Business Intelligence (Business Intelligence - BI), the information needed for the usage of BI in an organization, the solutions offered by Microsoft and the relationship between BI and Data Mining. The main data mining techniques will be presented and the basic operational problems that can be solved using these techniques. We will show the BI solutions offered by Microsoft through the Analysis Services (SSAS) of SQL Server (2005/2008); more specifically we will show the main characteristics of the basic data mining algorithms that can be implemented from SSAS platform (Decision Trees, Clustering, Association Rules), the parameters utilizing each algorithm and the DMX commands that can be used to implement them. Finally, we will present the development process of an application using Data Mining models through the Business Intelligence Development Studio (BIDS) and SQL Server's SSAS using the CRISP-DM Data Mining project development methodology.

Περιεχόμενα

Περιεχόμενα Πινάκων

Περιεχόμενα Εικόνων

Περιεχόμενα Διαγραμμάτων

Ευχαριστίες

Ολοκληρώνοντας την συγγραφή της παρούσας εργασίας θεωρώ υποχρέωση μου να ευχαριστήσω ορισμένα άτομα, που με υποστήριξαν σε όλη τη φάση του σχεδιασμού και της υλοποίησης της διατριβής για την απόκτηση του Μεταπτυχιακού τίτλου σπουδών «Προηγμένα Συστήματα Πληροφορικής» του τμήματος Πληροφορικής του Πανεπιστημίου Πειραιώς. Θέλω να ευχαριστήσω τον καθηγητή κ. Δεσπότη Δημήτρη για την υποστήριξη και την επίβλεψη του σε όλη την διάρκεια της υλοποίησης της διατριβής, τον κ. Σταμάτη Μπίλα, διευθυντή πληροφορικής της εταιρείας ΒΙΟΣΕΡ για την παροχή των πολύτιμων συμβουλών του κατά τον σχεδιασμό της διατριβής, τους συναδέλφους μου στην εταιρεία Datamine για την συμπαράσταση τους και την παροχή πολύτιμων ιδεών για την υλοποίηση της εφαρμογής μέσα από τα SSAS και προπάντων τους δικούς μου ανθρώπους για την αμέριστη συμπαράσταση τους σε όλη την διάρκεια συγγραφής και ανάπτυξης της εργασίας.

Τέλος, θέλω να αφιερώσω την παρούσα εργασία στον φίλο μου Αντρέα, που έφυγε πολύ νωρίς.

Παπαοικονόμου Εμμανουήλ
Αθήνα, 8/10/2010

Εισαγωγή

Ο όρος της Επιχειρηματικής Ευφυΐας (Business Intelligence – BI) είναι επίκαιρος στις μέρες μας, λόγω του ολοένα και αυξανόμενου αριθμού των δεδομένων που υπάρχουν σε ένα οργανισμό και της ανάγκης για ακριβή και έγκαιρη παροχή πληροφόρησης για την λήψη αποφάσεων σχετικά με την υλοποίηση της στρατηγικής του οργανισμού. Σε πολλές περιπτώσεις τα δεδομένα που διαχειρίζονται οι εφαρμογές BI μπορεί να κρύβουν πολύτιμη γνώση, η οποία είναι δύσκολο να την αντιληφθούν οι χρήστες. Για το λόγο αυτό, μπορούν να χρησιμοποιηθούν τεχνικές Ανακάλυψης Γνώσης από Βάσεις Δεδομένων, οι οποίες μπορούν να ανακαλύψουν τάσεις, συσχετίσεις και εξαρτήσεις μεταξύ των δεδομένων και γενικά να «σκάψουν» βαθιά στα δεδομένα για να ανακαλύψουν τον «χρυσό» που μπορεί να περιέχουν. Τέτοιες τεχνικές ανακάλυψης γνώσης είναι οι τεχνικές Εξόρυξης Γνώσης (Data Mining). Οι τεχνικές Εξόρυξης Γνώσης έχουν παρουσιαστεί εδώ και αρκετά χρόνια στον χώρο της Στατιστικής, των Μαθηματικών, της Πληροφορικής και γενικά στον χώρο της επεξεργασίας δεδομένων και ένα από τα μεγάλη προβλήματα που (θεωρούμε ότι) υπήρχαν ήταν η δυσκολία στην χρήση των εφαρμογών που υποστήριζαν τις τεχνικές Εξόρυξης Γνώσης, η δυσκολία στην επεξήγηση των αποτελεσμάτων (συχνά απαιτούνταν η παρουσία κάπου αναλυτή από τον χώρο της Εξόρυξης Γνώσης για να αναλύσει τα αποτελέσματα) και η δυσκολία στην οπτικοποίηση των αποτελεσμάτων των αλγορίθμων Εξόρυξης Γνώσης. Μια λύση στις δυσκολίες αυτές προσφέρουν τα σύγχρονα συστήματα εφαρμογών BI που προσπαθούν να απλοποιήσουν την παρουσίαση των αποτελεσμάτων του BI και γενικώς να κάνουν τον χρήστη των εφαρμογών αναλυτή των δικών του δεδομένων. Μια τέτοια εφαρμογή BI παρέχεται και από την Microsoft μέσω των Analysis Services του SQL Server (εκδόσεις 2005/2008).

Σκοπός της παρούσας εργασίας είναι να παρουσιαστεί η έννοια της Επιχειρηματικής Ευφυΐας, οι πληροφορίες που χρειάζονται για την χρήση των εφαρμογών BI σε ένα οργανισμό, οι λύσεις που προσφέρονται από την Microsoft και η σχέση του BI με την Εξόρυξη Γνώσης. Σχετικά με την Εξόρυξη Γνώσης θα παρουσιαστούν οι βασικοί λόγοι που ωθούν έναν οργανισμό στην χρήση των τεχνικών της, θα παρουσιαστούν οι κυριότερες τεχνικές Εξόρυξης Γνώσης και τα βασικότερα επιχειρησιακά προβλήματα που μπορούν να επιλυθούν με την χρήση τεχνικών Εξόρυξης Γνώσης. Επίσης θα παρουσιαστεί η λύση BI που προσφέρει η Microsoft μέσα από τα Analysis Services (SSAS) του SQL Server (2005/2008). Πιο συγκεκριμένα θα παρουσιαστούν τα κυριότερα χαρακτηριστικά των βασικών αλγορίθμων Εξόρυξης Γνώσης που μπορούν να υλοποιηθούν από τα SSAS (Decision Trees, Clustering, Association Rules), οι παράμετροι που ρυθμίζουν τον κάθε αλγόριθμο και οι εντολές DMX που μπορούν να χρησιμοποιηθούν για την υλοποίηση τους. Τέλος, θα παρουσιαστεί μια εφαρμογή χρήσης των μοντέλων Εξόρυξης Γνώσης μέσα από το Business Intelligence Development Studio (BIDS) και τα SSAS της Microsoft χρησιμοποιώντας την μεθοδολογία ανάπτυξης έργων Εξόρυξης Γνώσης CRISP-DM.

Στο **Κεφάλαιο 1** θα παρουσιαστεί η έννοια της Επιχειρηματικής Ευφυΐας (BI), η τυπική αρχιτεκτονική ενός συστήματος BI, η χρήση του BI (οι ανάγκες πληροφόρησης) σε όλα τα επίπεδα ενός οργανισμού, οι λύσεις που παρέχονται από την Microsoft και τα βήματα που πρέπει να γίνουν από την πλευρά του Αναλυτή για την ανάπτυξη ενός έργου BI.

Στο **Κεφάλαιο 2** θα παρουσιαστεί η σχέση της Εξόρυξης Γνώσης (Data Mining) με την Επιχειρηματική Ευφυΐα, η αρχιτεκτονική ενός συστήματος Εξόρυξης Γνώσης και οι βασικότερες τεχνικές Εξόρυξης Γνώσης. Πιο συγκεκριμένα, θα παρουσιαστεί η τεχνική της Ομαδοποίησης (Classification) και η υλοποίηση της με Δέντρα Απόφασης, η τεχνική της Ομαδοποίησης (Clustering) και η τεχνική των Κανόνων Συσχέτισης (Association Rules).

Στη συνέχεια, στο **Κεφάλαιο 3** θα παρουσιαστούν τα κυριότερα Επιχειρησιακή Προβλήματα τα οποία μπορούν να επιλυθούν με τεχνικές Εξόρυξης Γνώσης. Συνοπτικά αναφέρουμε την Εύρεση Κερδοφόρων Πελατών, την Κατανόηση Αναγκών Πελατών, την Διατήρηση Πελατών – Churn Ανάλυση, την Πρόβλεψη Πωλήσεων και Αποθεμάτων, την δημιουργία Αποτελεσματικών Καμπανιών Marketing και τον Εντοπισμό και Πρόληψη Απάτης.

Επίσης θα αναφέρουμε την χρήση που μπορεί να γίνει των τεχνικών Εξόρυξης Γνώσης σε μια εφαρμογή CRM (Customer Relationship Management).

Στο **Κεφάλαιο 4** θα γίνει παρουσίαση του περιβάλλοντος SSAS (SQL Server Analysis Services) για την υλοποίηση έργων Επιχειρηματικής Ευφυΐας μέσα από το Business Intelligence Development Studio (BIDS) της Microsoft αλλά και με την χρήση των κατάλληλων εντολών DMX (Data Mining Extensions). Πιο συγκεκριμένα θα παρουσιαστούν οι τεχνικές Εξόρυξης Γνώσης, που παρέχονται μέσα από τα SSAS. Θα παρουσιαστεί η αρχιτεκτονική του συστήματος Εξόρυξης Γνώσης στα SSAS, η εφαρμογή καθενός από τους αλγορίθμους Εξόρυξης Γνώσης που παρέχονται και οι κυριότερες εντολές DMX που παρέχονται για το σκοπό αυτό. Στη συνέχεια, θα γίνει μια συνοπτική παρουσίαση των κυριότερων τεχνικών Εξόρυξης Γνώσης που παρέχονται από τα SSAS. Του αλγορίθμου Microsoft Decision Trees (Κατασκευή του Δέντρου, Παράμετροι Αλγορίθμου, Ανάλυση συσχετίσεων, Παλινδρόμηση και Εντολές DMX), του αλγορίθμου Microsoft Clustering (Μέθοδοι αλγορίθμου (EM Ομαδοποίηση, K-Means Ομαδοποίηση), δεδομένα που απαιτούνται για τον Αλγόριθμο, Παράμετροι Αλγορίθμου και Εντολές DMX) και του αλγορίθμου Microsoft Association Rules (τι είναι Υποστήριξη και Εμπιστοσύνη, δεδομένα που απαιτούνται για τον Αλγόριθμο, Παράμετροι Αλγορίθμου και Εντολές DMX).

Τέλος, στο **Κεφάλαιο 5** θα γίνει μια πλήρης παρουσίαση χρήσης μιας εφαρμογής Εξόρυξης Γνώσης μέσα από το BIDS. Για την ανάπτυξη της εφαρμογής θα χρησιμοποιηθεί η μεθοδολογία ανάπτυξης έργων Εξόρυξης Γνώσης CRISP-DM. Η εφαρμογή που θα παρουσιαστεί αφορά το παράδειγμα δημιουργίας μιας Στοχευμένης Διαφημιστικής Εκστρατείας και περιλαμβάνει ένα πλήρη κύκλο σχεδιασμού, ανάπτυξης, ελέγχου και χρήσης διαφόρων μοντέλων Εξόρυξης Γνώσης μέσα από το BI Development Studio (Visual Studio 2008) της Microsoft.

1. Επιχειρηματική Ευφυΐα (Business Intelligence)

Το πρώτο βήμα σε ένα οργανισμό για την αποτελεσματική λήψη αποφάσεων, είναι να τεθούν συγκεκριμένοι και μετρήσιμοι στόχοι σχετικά με την υλοποίηση της στρατηγικής του οργανισμού. Αφού καθοριστούν αυτοί οι στόχοι, το επόμενο βήμα είναι να υπάρξουν ακριβείς και χρήσιμες πληροφορίες προς τα στελέχη, που θα λάβουν τις αποφάσεις, σχετικά με την εγκυρότητα των στόχων και για το αν μπορούν να επιτευχθούν. Οι πληροφορίες αυτές θα χρησιμεύσουν σαν βάση για την λήψη αποφάσεων και σαν ανατροφοδότηση σχετικά με τον έλεγχο της αποτελεσματικότητας των αποφάσεων αυτών. Για τον οργανισμό λοιπόν, το βασικό στοιχείο είναι οι πληροφορίες να είναι διαθέσιμες την κατάλληλη στιγμή προς τα στελέχη, που θα λάβουν τις αποφάσεις. Επομένως, το ερώτημα για τον οργανισμό είναι πώς ο οργανισμός θα λάβει και θα διανέμει τις πληροφορίες αυτές.

Σύμφωνα με το παραπάνω πρόβλημα που αντιμετωπίζουν οι οργανισμοί, της έγκυρης παροχής πληροφορίας σε όλο τον οργανισμό, η Επιχειρηματική Ευφυΐα είναι:

«η διαδικασία παροχής ακριβούς και χρήσιμης πληροφορίας στα κατάλληλα στελέχη λήψης αποφάσεων εντός του αναγκαίου χρονικού πλαισίου για να υποστηρίξει την αποτελεσματική λήψη απόφασης».

Η επιχειρηματική ευφυΐα δεν είναι μόνο απλοί αριθμοί και διαγράμματα τυπωμένα σε μια αναφορά ή στην οθόνη ενός υπολογιστή. Τα αριθμητικά δεδομένα σε κάποιες γραμμές μιας αναφοράς, που δείχνουν για παράδειγμα στοιχεία πωλήσεων ή δεδομένα παραγωγής, μπορεί να είναι ακριβείς πληροφορίες αλλά δεν είναι επιχειρηματική ευφυΐα, έως ότου αυτά τεθούν σε μια τέτοια μορφή ώστε να γίνουν εύκολα κατανοητά από το στέλεχος που λαμβάνει αποφάσεις και χρειάζεται αυτές τις πληροφορίες εγκαίρως. Για παράδειγμα, μια συνοπτική περίληψη της ικανοποίησης των πελατών, σε μια επιχείρηση, η οποία μπορεί να γίνει εύκολα κατανοητή από τα στελέχη, δεν είναι επιχειρηματική ευφυΐα, έως ότου οι πληροφορίες αυτές παραδίδονται την σωστή, έτσι ώστε να επηρεάζουν ουσιαστικά την καθημερινή λήψη αποφάσεων, σχετικά με την εξυπηρέτηση των πελατών.

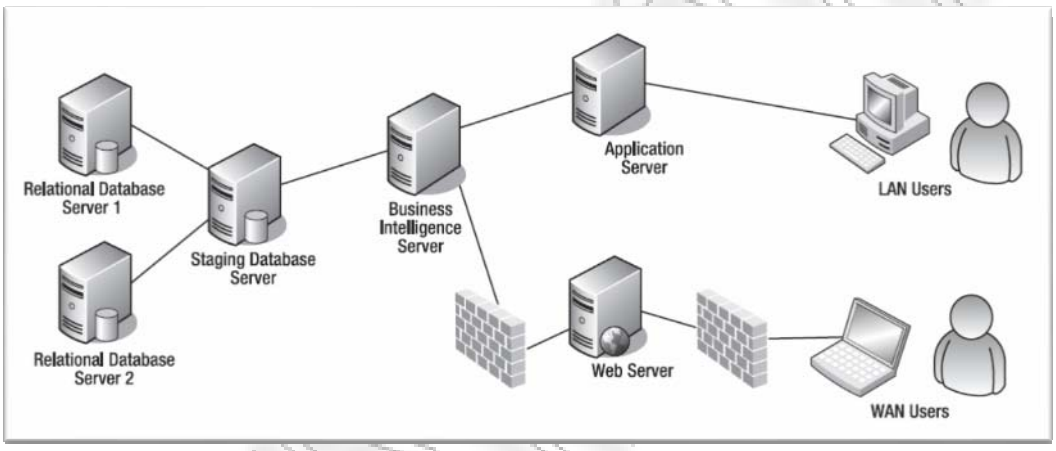
Ένας άλλος ορισμός της Επιχειρηματικής Ευφυΐας» είναι:

«*ένα μεγάλο σύνολο εφαρμογών και τεχνολογιών για την συγκέντρωση, αποθήκευση, ανάλυση, κοινοποίηση και παροχή πρόσβασης σε δεδομένα που βοηθούν τις επιχειρήσεις να λαμβάνουν καλύτερες επιχειρηματικές αποφάσεις*». (Gartner).

Σύμφωνα, λοιπόν με τον παραπάνω ορισμό η επιχειρηματική ευφυΐα (Business Intelligence – BI) είναι ένα σύνολο μεθόδων για αποθήκευση και παρουσίαση σημαντικών πληροφοριών έτσι ώστε ο κάθε ενδιαφερόμενος σε ένα οργανισμό να μπορεί εύκολα και γρήγορα να παίρνει απαντήσεις στα ερωτήματα του και να λαμβάνει τις πιο σωστές αποφάσεις. Στον ορισμό που παρουσιάστηκε παραπάνω αλλά και στην ίδια την ονομασία του όρου «επιχειρηματική ευφυΐα» γίνεται λόγος για μια επιχείρηση, δηλαδή για ένα οργανισμό που προσβλέπει σε πωλήσεις και σε αντίστοιχο κέρδος (σε όρους χρημάτων). Βέβαια, ο παραπάνω ορισμός της επιχειρηματικής ευφυΐας μπορεί να επεκταθεί και σε άλλους οργανισμούς, εκτός από επιχειρηματικούς, όπως οργανισμούς σχετικά με την υγεία (νοσοκομεία, κλινικές κλπ), κυβερνητικούς οργανισμούς, φιλανθρωπικούς οργανισμούς κλπ. Παρατηρούμε, με λίγα λόγια ότι η επιχειρηματική ευφυΐα δεν αναφέρεται τόσο σε αποφάσεις μιας επιχείρησης που ενδιαφέρεται μόνο για το εμπορικό κέρδος αλλά τόσο στην ποιότητα των αποφάσεων, που λαμβάνει ο οργανισμός με τα εργαλεία BI που του παρέχονται.

Επομένως, ο σωστός ορισμός θα ήταν η βελτίωση της απόδοσης των αποφάσεων που λαμβάνει ο οργανισμός. Ωστόσο, για εμπορικούς σκοπούς έχει επικρατήσει ο όρος Business Intelligence, που είναι πιο εμπορικός όρος από κάποιον άλλο όρος, όπως για παράδειγμα Performance Intelligence. Επαναλαμβάνουμε λοιπόν ότι παρόλο που ο όρος BI περιλαμβάνει τον όρο Business (Επιχείρηση) δεν αναφέρεται μόνο σε επιχειρηματικούς οργανισμούς αλλά και σε άλλους οργανισμούς, που στόχο έχουν την βελτίωση της ποιότητας και της απόδοσης των αποφάσεων, που λαμβάνουν με την χρήση συγκεκριμένων εφαρμογών και τεχνολογιών.

Παρακάτω παρουσιάζεται μια τυπική αρχιτεκτονική ενός συστήματος Business Intelligence. Χρησιμοποιείται μια staging βάση, για την σταδιακή συλλογή δεδομένων από διάφορες σχεσιακές Βάσεις Δεδομένων και ένας server για την υποστήριξη των διαδικασιών Business Intelligence (όλες οι λειτουργίες Business Intelligence που θα τρέχουν, για παράδειγμα εφαρμογές OLAP συστήματα για διαχείριση κύβων και διαστάσεων, εφαρμογές Εξόρυξης Γνώσης, αναφορές κα). Στην συνέχεια, τα αποτελέσματα της διαδικασίας Business Intelligence μοιράζονται μέσω ενός Application Server στους χρήστες του τοπικού δικτύου του οργανισμού για την προβολή πληροφοριών που θα υποστηρίξει την εργασία τους είτε μέσω ενός web server στους χρήστες ενός ευρύτερου δικτύου (είτε μεταξύ διαφορετικών τμημάτων του οργανισμού σε διαφορετικές τοποθεσίες είτε μεταξύ διαφορετικών επιχειρήσεων).



Διάγραμμα 1: Τυπική φυσική οργάνωση μιας εφαρμογής BI

Business Intelligence σε Πολλαπλά Επίπεδα Λειτουργίας

Μια λύση BI πρέπει να χρησιμοποιείται σε όλα τα επίπεδα ενός οργανισμού έτσι ώστε να υποστηρίζει την λήψη αποτελεσματικών αποφάσεων. Ωστόσο, η χρήση μια λύσης BI σε όλα τα επίπεδα του οργανισμού δεν απαιτεί και την ίδια ποιότητα και ποσότητα πληροφοριών σε κάθε επίπεδο. Τα διαφορετικά επίπεδα διοίκησης σε ένα οργανισμό χρειάζονται και μια διαφορετική μορφή της λύσης BI για την αποτελεσματική υποστήριξη στην λήψη αποφάσεων. Για την καλύτερη παρουσίαση των διαφόρων επιπέδων ενός οργανισμού και του είδους των δεδομένων που απαιτούνται, θα χρησιμοποιήσουμε στις παρακάτω παραγράφους το σχήμα μιας πυραμίδας, η οποία δείχνει τα αναφερόμενα επίπεδα του οργανισμού.

Η κορυφή της Πυραμίδας

Τα στελέχη που λαμβάνουν αποφάσεις στα υψηλά επίπεδα ενός οργανισμού χρειάζονται να βλέπουν την μεγάλη εικόνα (big picture) της επιχείρησης. Είναι υπεύθυνοι για τον καθορισμό μακροπρόθεσμων στόχων για τον οργανισμό και πρέπει να έχουν μια ευρεία άποψη του χώρου ευθύνης τους και όχι να ασχολούνται τόσο πολύ με λεπτομέρειες.

Διάγραμμα 2: Στόχοι σε κάθε επίπεδο του Οργανισμού

Η λύση Business Intelligence που παρέχεται στα υψηλά επίπεδα πρέπει να αντιμετωπίζει τα παραπάνω χαρακτηριστικά. Οι μετρικές που παρέχονται στα στελέχη λήψης αποφάσεων πρέπει να αθροίζονται σε υψηλό επίπεδο (highly summarized measures) και να είναι πολύ συνοπτικά. Η κάθε μετρική δεν αρκεί μόνο να παρουσιάζεται με ένα συγκεκριμένο αριθμό αλλά να παρουσιάζεται ένας δείκτης κατάστασης που παρουσιάζει εάν η τιμή της μετρικής αυτή η βρίσκεται μέσα σε αποδεκτά όρια ή ξεφεύγει από τους στόχους. Αυτές οι μετρικές που αθροίζονται σε υψηλό επίπεδο αναφέρονται σαν *Βασικοί Δείκτες Απόδοσης* (*Key Performance Indicators – KPIs*).

Τα KPIs χρησιμοποιούνται για να προσφέρουν στα υψηλά στελέχη που λαμβάνουν αποφάσεις, ένα γρήγορο τρόπο να καθορίσουν την «υγεία», την εύρυθμη λειτουργία των διαφόρων τμημάτων του οργανισμού. Τα KPIs συνήθως αναπαριστώνται με κάποιο γράφημα (σαν το κοντέρ μέτρησης ταχύτητας των αυτοκινήτων ή σαν ένα φανάρι ρύθμισης της κυκλοφορίας) για να δείξουν άμεσα την κατάσταση των τμημάτων του οργανισμού.

Διάγραμμα 3: Απαιτούμενες μετρικές σε κάθε επίπεδο του Οργανισμού

Επειδή τα στελέχη στα υψηλά τμήματα του οργανισμού ασχολούνται με τον καθορισμό μακροπρόθεσμων στόχων και με την κατεύθυνση του οργανισμού προς την επίτευξη αυτών των στόχων, δεν απαιτούν ανατροφοδότηση πληροφοριών από το σύστημα Business Intelligence την ίδια στιγμή. Δηλαδή ο χρόνος που μεσολαβεί ανάμεσα στην εμφάνιση μιας συναλλαγής στο σύστημα και της εισαγωγής των πληροφοριών στο σύστημα BI μπορεί να είναι σχετικά μεγάλος (**Χρόνος Απόκρισης**). Με λίγα λόγια, τα υψηλά στελέχη χρειάζονται να βλέπουν τις τάσεις της απόδοσης του οργανισμού και δεν είναι ανάγκη να βλέπουν τα αποτελέσματα των καθημερινών συναλλαγών του οργανισμού.

Διάγραμμα 4: Χρόνος απόκρισης και ανατροφοδότηση πληροφορίας σε κάθε επίπεδο του Οργανισμού

Μεσαίο Επίπεδο

Τα στελέχη που λαμβάνουν αποφάσεις στα μεσαία επίπεδα του οργανισμού και χρησιμοποιούν την εφαρμογή Business Intelligence είναι στελέχη που ελέγχουν την λειτουργία των διαφόρων τμημάτων του οργανισμού, των οποίων η λειτουργία έχει ανατεθεί σε αυτούς. Τα στελέχη αυτά θέτουν βραχυπρόθεσμους στόχους και καθορίζουν το πλάνο λειτουργίας των τμημάτων τους. Τα μεσαία στελέχη είναι σε ένα επίπεδο, που δεν ενδιαφέρονται για παροχή πληροφοριών σχετικά με τις καθημερινές λειτουργίες του οργανισμού. Έτσι τα στελέχη αυτά απαιτούν από την εφαρμογή Business Intelligence δεδομένα, που είναι μεν αθροιστικά σε ένα υψηλότερο επίπεδο αλλά πολλές φορές επιτρέπουν την πλοήγηση προς τα κάτω (drill down) σε πληροφορίες για να έχουν μια πιο λεπτομερής πληροφόρηση. Τα στελέχη αυτά χρησιμοποιούν κυρίως εκτυπωμένες αναφορές και αλληλεπιδραστικά συστήματα που επιτρέπουν την ανακάλυψη πληροφοριών. Τα μεσαία στελέχη συνήθως χρησιμοποιούν πληροφορίες που εξάγονται μέσα διάφορες εφαρμογές Εξόρυξης Γνώσης (Data Mining).

Επειδή τα μεσαία στελέχη βρίσκονται κοντά στις καθημερινές λειτουργίες του οργανισμού, ο χρόνος απόκρισης, που αναφέραμε παραπάνω (χρόνος μεταξύ μιας συναλλαγής του οργανισμού και της ανατροφοδότησης την πληροφορίας στο BI σύστημα) μπορεί να είναι σχετικά μικρός. Συνήθως τα στελέχη χρειάζονται να βλέπουν τα αποτελέσματα και να ενημερώνονται καθημερινά. Σε άλλες όμως περιπτώσεις, τα στελέχη αυτά ενδιαφέρονται για τάσεις σε εβδομαδιαία ή μηνιαία βάση.

Η Ευρεία Βάση

Στην ευρεία βάση της πυραμίδας της εφαρμογής Business Intelligence, βρίσκονται τα στελέχη που φροντίζουν για τις καθημερινές λειτουργίες του οργανισμού. Τα στελέχη αυτά θέτουν τους καθημερινούς λειτουργικούς στόχους και λαμβάνουν αποφάσεις για τον καθορισμό πόρων για την επομένη εβδομάδα, την επόμενη μέρα ή ίσως και για την επόμενη βάρδια στην επιχείρηση. Συνήθως τα στελέχη αυτά απαιτούν από τα συστήματα Business Intelligence να έχουν μεγάλη διαθεσιμότητα και υπευθυνότητα.

Όπως τονίσαμε, τα στελέχη στο τελευταίο επίπεδο της πυραμίδας χρησιμοποιούν δεδομένα σε μεγάλη λεπτομέρεια. Οι μετρικές μπορεί να αθροίζονται σε ένα υψηλότερο επίπεδο αλλά η πλοήγηση σε κατώτερο επίπεδο (drill down) είναι απαιτούμενο από την εφαρμογή. Συνήθως τα στελέχη μπορούν να χρησιμοποιούν κάποιες εφαρμογές Εξόρυξης Γνώσης για να ανακαλύψουν κάποιες τάσεις στα δεδομένα και κάποιες συσχετίσεις ανάμεσα στις πληροφορίες που παρέχονται σε καθημερινό επίπεδο.

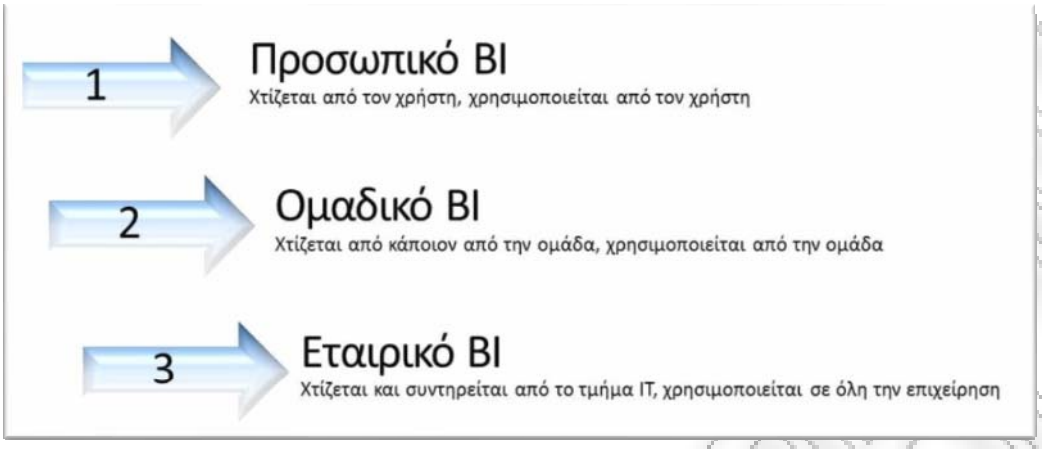
Επειδή στο επίπεδο αυτό ελέγχονται οι καθημερινές λειτουργίες, ο χρόνος απόκρισης πρέπει να είναι μικρός και τα συστήματα Business Intelligence πρέπει να αντιδρούν άμεσα στις διάφορες αλλαγές στην ανατροφοδότηση των πληροφοριών. Με λίγα λόγια, κάθε συναλλαγή στον οργανισμό πρέπει να αντανακλά σαν αλλαγή των δεδομένων στο σύστημα Business Intelligence.

Από το Εταιρικό BI στο Προσωπικό BI

Οι σημερινές λύσεις Business Intelligence που παρέχονται στους χρήστες συνήθως έχουν χαμηλή ανταπόκριση από τους απλούς χρήστες ενώ στηρίζονται ως επί το πλείστον σε κάποιον υπεύθυνο του τμήματος IT της εταιρείας. Οι χρήστες δηλαδή των εφαρμογών BI σε ένα οργανισμό δεν έχουν εύκολα πρόσβαση σε επιχειρηματικά δεδομένα, στηρίζονται σε κάποιον από το τμήμα IT για να δημιουργήσει για αυτούς και να εξαγάγει μια αναφορά και είναι αρκετά δύσκολο να διαμοιράσουν την πληροφορία αυτή στους υπόλοιπους συναδέλφους και συνεργάτες τους (η δυσκολία αυτή έγκειται συνήθως στην έλλειψη εμπειρίας και στην κατανόηση της χρήσης πολύπλοκων εφαρμογών πληροφορικής).

Στόχος των εταιρειών που προσφέρουν υπηρεσίες Business Intelligence, όπως για παράδειγμα της Microsoft, στις εφαρμογές της οποίας θα αναφερθούμε στην συνέχεια, είναι η δυνατότητα στους χρήστες των εφαρμογών αυτών, πέρα από την δυνατότητα δημιουργίας μοντέλων Business Intelligence για όλη την επιχείρηση, να γίνουν οι ίδιοι αναλυτές των δεδομένων και να δημιουργούν τα διάφορα μοντέλα που θέλουν, να παίρνουν τα αποτελέσματα που θέλουν και να κάνουν διορθώσεις όπου χρειάζεται. Ο όρος που έχει χρησιμοποιήσει η Microsoft για τον σκοπό αυτό είναι «*BI for the Masses*», όλοι δηλαδή οι χρήστες των εφαρμογών της Microsoft σε ένα οργανισμό να έχουν την δυνατότητα να πραγματοποιήσουν λειτουργίες Επιχειρηματικής Ευφυΐας.

Με λίγα λόγια, στόχος είναι να παρέχονται οι δυνατότητες στους χρήστες να αναλύουν τα δεδομένα που χρειάζονται, σύμφωνα με τις προσωπικές τους ανάγκες και να λαμβάνουν τις αποφάσεις που θέλουν – και όλα αυτά χωρίς την υποχρεωτική παρουσία κάπου αναλυτή ή συμβούλου πληροφορικής, ο οποίος θα πρέπει να δημιουργήσει αυτός κάποιο πολύπλοκο μοντέλο (πχ κάποιο μοντέλο Εξόρυξης Γνώσης). Αυτό έχει σαν αποτέλεσμα τα μέλη μιας επιχείρησης να έχουν πρόσβαση άμεσα σε πληροφορίες που χρειάζονται, και να παίρνουν τις αποφάσεις που επιθυμούν.



Διάγραμμα 5: Τρία πλαίσια χρήσης BI

Η δυνατότητα που παρέχεται στα μέλη μιας επιχείρησης να έχουν πρόσβαση στις πληροφορίες που χρειάζονται (ή καλύτερα να έχουν πρόσβαση στην Γνώση που χρειάζονται) έχει σαν αποτέλεσμα την λήψη γρηγορότερων και πιο σχετικών αποφάσεων (σχετικών με το εκάστοτε πρόβλημα που αντιμετωπίζει ο κάθε χρήστης της επιχείρησης).

Για πετύχει τον σκοπό αυτό, η Microsoft προσφέρει τις νέες υπηρεσίες Business Intelligence μέσα από γνωστά εργαλεία, που επί χρόνια τα χρησιμοποιούν οι περισσότεροι χρήστες και το περιβάλλον τους είναι φιλικό – η βάση δεδομένων SQL Server (2005 ή 2008) και η ολοκλήρωση της με περιβάλλον Office (με την χρήση του Excel, του Visio και άλλα).

Η λύση της Microsoft: Business Intelligence και SQL Server

Ο Microsoft SQL Server (2005/2008) παρέχει εργαλεία για να υποστηρίξει όλες τις πλευρές του Business Intelligence. Τα Integration Services (SSIS) παρέχουν στους χρήστες την δυνατότητα να δημιουργήσουν αυτοματοποιημένες διαδικασίες για να καθαρίσουν τα δεδομένα και να τα μεταφέρουν προς την αποθήκη δεδομένων, που θα χρησιμοποιηθεί στην διαδικασία BI και για να εξασφαλίσουν ότι υπάρχουν ακριβείς πληροφορίες την χρονική στιγμή που απαιτούνται. Στα Analysis Services (SSAS) είναι διαθέσιμες πολλές δυνατότητες OLAP, όπως οι Δείκτες Απόδοσης (KPIs), ερωτήσεις πάνω στις διαστάσεις των κύβων που δημιουργούνται (MDX – Multidimensional Expressions) και το μοντέλο δημιουργίας κύβων, όπου μπορούμε να δούμε τα δεδομένα από διάφορες πλευρές (διαστάσεις) και να εκτελούμε λειτουργίες όπως αθροίσεις, προβολή συγκεντρωτικών αποτελεσμάτων και άλλες λειτουργίες από την χρήση των κύβων. Επίσης, οι λειτουργίες Εξόρυξης Γνώσης που παρέχονται από τα SSAS βοηθούν τους χρήστες να ανακαλύψουν πρότυπα στα δεδομένα και να προβλέψουν αποτελέσματα, λειτουργίες που δεν θα μπορούσαν να εκτελεστούν διαφορετικά στα δεδομένα. Τέλος, τα Reporting Services (SSRS) με τις δυνατότητες δημιουργίας αναφορών βοηθάνε στον διαμοιρασμό της πληροφορίας σε όλον τον οργανισμό και ιδίως στους υπευθύνους λήψης αποφάσεων σχετικά με την στρατηγική του οργανισμού.

Πως λειτουργεί η πλατφόρμα BI της Microsoft;

Η πλατφόρμα Business Intelligence της Microsoft προορίζεται για επιχειρήσεις οι οποίες επιθυμούν να πραγματοποιείται η έξυπνη λήψη αποφάσεων σε ολόκληρο τον οργανισμό και να καθίσταται εύκολη η συνεργασία, ανάλυση, κοινή χρήση των επιχειρηματικών πληροφοριών για όλους όσους εργάζονται στον οργανισμό, από μια κεντρικά διαχειριζόμενη και ασφαλής προέλευση. Μια λύση BI για να ανταποκριθεί στους σκοπούς της, θα πρέπει να σχεδιάζεται με βάση τις ανάγκες κάθε χρήστη και να έχει την ευελιξία να φιλοξενήσει όλη την δομημένη και αδόμητη πληροφορία που χρησιμοποιεί καθημερινά ο οργανισμός.

Η δημιουργία μιας σταθερής λύσης BI πρέπει να βασίζεται στα ίδια τα δεδομένα τα οποία χρησιμοποιούν οι χρήστες για να λάβουν αποφάσεις. Ο **SQL Server** έχει την δυνατότητα να συγκεντρώσει όλα τα στοιχεία, οπουδήποτε και αν είναι αυτά. Ο SQL Server είναι γνωστός για την δύναμη και την επεκτασιμότητα του και αποτελεί μια λύση που μπορεί να λειτουργήσει σε ένα οργανισμό, για την αποθήκευση τεράστιων ποσοτήτων δεδομένων και την υποστήριξη μεγάλων αριθμών ερωτημάτων από τους χρήστες.

Το επόμενο βήμα μιας λύσης BI είναι να δοθεί στους ανθρώπους του οργανισμού ένας τρόπος να συγκεντρώσουν και να οργανώσουν όλη την πληροφορία που χρησιμοποιούν απ' όπου και αν προέρχεται – από το ηλεκτρονικό ταχυδρομείο, το Internet, από δεδομένα του οργανισμού ή από οπουδήποτε αλλού. Οι λύσεις BI που παρέχονται από τον Microsoft SQL Server είναι μια πλήρως ενοποιημένη οικογένεια προϊόντων (client server και εργαλεία προγραμματισμού) που είναι ενσωματωμένη στην σουίτα εφαρμογών **Microsoft Office** και παρέχει τις κατάλληλες πληροφορίες, την κατάλληλη στιγμή και στην κατάλληλη μορφή. Οι λύσεις BI παρέχουν εύχρηστες πληροφορίες απευθείας στα σημεία που οι χρήστες των εφαρμογών δουλεύουν, συνεργάζονται και λαμβάνουν αποφάσεις. Συνδέοντας τη στρατηγική με τις μετρήσεις, οι οργανισμοί μπορούν να αποκτήσουν το ανταγωνιστικό πλεονέκτημα της ταχύτερης λήψης καλύτερων αποφάσεων, σε όλα τα επίπεδα του οργανισμού.

Οι λύσεις Business Intelligence της Microsoft παρέχουν επιχειρηματικές πληροφορίες σε οποιονδήποτε εντός του οργανισμού, ενσωματώνοντας δύο κύρια στοιχεία:

- **Πλατφόρμα Business Intelligence**, η οποία λειτουργεί με τον Microsoft SQL Server και συμπεριλαμβάνει το σύστημα διαχείρισης σχεσιακών βάσεων δεδομένων, τις υπηρεσίες SQL Server Integration Services, SQL Server Analysis Services, SQL Server Reporting Services και τις δυνατότητες SQL Server Data Mining.

- **Εφαρμογές Microsoft Office**, που παρέχουν πληροφορίες μέσω των εργαλείων τα οποία οι χρήστες ήδη γνωρίζουν και εμπιστεύονται. Οι χρήστες έχουν τη δυνατότητα χρήσης ισχυρών, αλληλεπιδραστικών υπολογιστικών φύλλων μέσα από το Excel, χρησιμοποιώντας βελτιωμένες δυνατότητες δημιουργίας γραφημάτων, σύνθεσης τύπων και ενισχυμένη δυνατότητα. Με υπολογιστικά φύλλα, τα οποία βρίσκονται στον server του οργανισμού, οι χρήστες μπορούν να κάνουν ευρεία χρήση πληροφοριών με αξιοπιστία, γνωρίζοντας ότι οι πληροφορίες είναι πιο ασφαλείς και κεντρικά διαχειριζόμενες, αλλά και προσβάσιμες σε διαφόρους συναδέλφους τους, πελάτες και συνεργάτες τους μέσω του Web.
  Οι δυναμικοί πίνακες αποτελεσμάτων συνδυάζουν την ισχύ της ανάλυσης και της Εξόρυξης Γνώσης με τις αναφορές σε πραγματικό χρόνο. Οι στρατηγικοί χάρτες που παρέχονται από το σύστημα διευκολύνουν την οπτικοποίηση σημαντικών περιοχών, έτσι οι χρήστες μπορούν να δουν τάσεις, να αναγνωρίσουν έγκαιρα προβληματικές περιοχές, να μεγιστοποιήσουν τις επιτυχημένες περιοχές και να παρακολουθήσουν την απόδοση ως προς βασικούς στόχους του οργανισμού σε πραγματικό χρόνο.

Για την υποστήριξη μιας περισσότερο εμπεριστατωμένης λήψης αποφάσεων σε ένα ευρύτερο επίπεδο στον οργανισμό, θα πρέπει να υπάρχει μια σταθερή στρατηγική παρακολούθησης των επιδόσεων. Η εφαρμογή Office Performance Point Server χρησιμοποιείται για να υποστηρίξει επιχειρηματικά σενάρια, παρέχοντας άμεση προβολή των βασικών κινητήριων δυνάμεων της επιχείρησής. Μέσω της εφαρμογής αυτής οι χρήστες μπορούν «με μια ματιά» να καταλάβουν στην επιχείρηση τι λειτουργεί και τι όχι, ποιες είναι οι ευκαιρίες που πρέπει να επωφεληθούν από αυτές και που πρέπει να ληφθεί άμεση δράση.

Βελτίωση επιχειρηματικής απόδοσης

Οι λειτουργίες Business Intelligence του SQL Server βοηθούν κυρίως τους οργανισμούς να βελτιώσουν την απόδοση τους, παρέχοντας τις παρακάτω δυνατότητες:

- **Παρακολούθηση** και **ανάλυση** οικονομικών, λειτουργικών πληροφοριών, στοιχείων πελατών και πληροφοριών ανθρωπίνων πόρων σε ολόκληρο τον οργανισμό, παρέχοντας στους χρήστες τις σωστές πληροφορίες, τη σωστή στιγμή και σε εύχρηστη μορφή.
- **Πίνακες αποτελεσμάτων** (**scorecards)**. Επεκτείνει την προσέγγιση των επιχειρηματικών πληροφοριών και εισάγει κάθε εργαζόμενο στη διαδικασία διαχείρισης απόδοσης μέσω των πινάκων αποτελεσμάτων. Οι πίνακες Αποτελεσμάτων μπορούν να εγκατασταθούν σε ένα Web server έτσι ώστε να είναι προσβάσιμοι από διάφορα σημεία και από όσους θέλουν να λάβουν την πληροφορία που χρειάζονται.
- **Αναλυτικά στοιχεία**. Παρέχει τη δυνατότητα στους χρήστες να λαμβάνουν καλύτερες και ταχύτερες αποφάσεις μέσω των σύνθετων δυνατοτήτων αναλύσεων και οπτικοποίησης δεδομένων.
- **Σχεδιασμός**. Παρέχει τη δυνατότητα εκτέλεσης ενός αποτελεσματικού σχεδιασμού προϋπολογισμού και προβλέψεων, βοηθώντας τους χρήστες να δημιουργούν λεπτομερή μοντέλα και ευέλικτα σχέδια, συγχρονισμένα σε όλα τα τμήματα και σε ολόκληρη την ιεραρχία του οργανισμού.

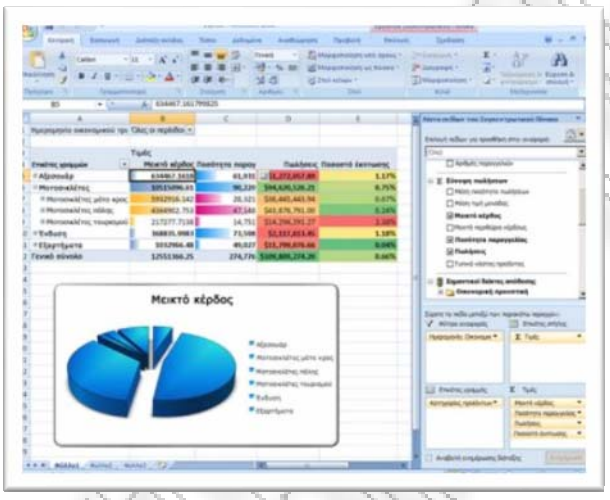
Διάγραμμα 6: Περιβάλλον εργασίας του Microsoft BI για την δημιουργία scorecard

Διανομή πληροφοριών μέσω του Microsoft Office

Όπως τονίσαμε παραπάνω, τα αποτελέσματα της ανάλυσης των εφαρμογών Business Intelligence της Microsoft, μπορούν να διανεμηθούν μέσα από την πλατφόρμα Microsoft Office, η οποία αποτελεί ένα γνώριμο περιβάλλον εργασίας των χρηστών και ουσιαστικά υπάρχει σε όλους τους υπολογιστές στον οργανισμό. Οι λειτουργίες που παρέχονται για την διανομή των πληροφοριών, μέσω του Microsoft Office παρουσιάζονται παρακάτω:

- **Ανάπτυξη σύνθετης ανάλυσης** με τη χρήση υπολογιστικών φύλλων του Excel.
- **Χρήση οικείων εργαλείων ανάλυσης**. Το Microsoft BI παρέχει πληροφορίες στην επιφάνεια εργασίας με το οικείο και εύχρηστο περιβάλλον Microsoft Office. Μπορούν να ενσωματωθούν εύκολα πληροφορίες από κάθε προέλευση δεδομένων που είναι διαθέσιμη στην εταιρεία, συμπεριλαμβανομένων των αποθηκών δεδομένων και των εταιρικών εφαρμογών.
- **Πραγματοποίηση σύνθετης ανάλυσης**. Χρησιμοποιώντας το Excel (που συμπεριλαμβάνει τη αυξημένη χωρητικότητα υπολογιστικών φύλλων, την σύνθεση τύπων με τη χρήση επιχειρηματικών όρων και την προηγμένη δυνατότητα ταξινόμησης και φιλτραρίσματος) το Microsoft BI υποστηρίζει τη σύνθετη ανάλυση που μπορεί να οδηγήσει σε καλύτερη πληροφόρηση και λήψη αποφάσεων.
- **Βελτίωση της ανάλυσης υπολογιστικών φύλλων**. Με τις υπηρεσίες Excel Services και SQL Server Analysis Services, η άμεση σύνδεση με την προέλευση δεδομένων παρουσιάζει μετά-δεδομένα, διαστάσεις και μετρήσεις σε κατανοητούς επιχειρηματικούς όρους. Οι χρήστες μπορούν εύκολα να τροποποιήσουν αναφορές και να ανανεώσουν δεδομένα έτσι ώστε να διασφαλίζεται η ενημερωμένη ακρίβεια ανά πάσα στιγμή.
- **Κοινή χρήση, διαχείριση και έλεγχος υπολογιστικών φύλλων**. Οι χρήστες μπορούν να κάνουν ευρεία και αξιόπιστη κοινή χρήση επιχειρηματικών δεδομένων. Οι υπηρεσίες Excel Services εκτελούν υπολογισμούς στην πλευρά του διακομιστή (server) και παρέχουν πρόσβαση σε δεδομένα και αναλυτικά στοιχεία σε πραγματικό χρόνο, μέσω δυναμικών υπολογιστικών φύλλων που αποδίδονται ως HTML. Μπορεί να γίνει αποτελεσματική κοινή χρήση μιας κεντρικής έκδοσης των δεδομένων, βοηθώντας παράλληλα στην προστασία των ευαίσθητων ή ιδιωτικών πληροφοριών. Οι πληροφορίες που περιέχονται στα έγγραφα (πχ οικονομικά μοντέλα) θωρακίζονται, καθώς περιορίζεται η πρόσβαση σε τμήματα του υπολογιστικού φύλλου.

- **Επαναχρησιμοποίηση των μοντέλων υπολογιστικών φύλλων σε ανάπτυξη εφαρμογών**. Ο διαχωρισμός της εργασίας ανάπτυξης από την εργασία της επιχειρηματικής ανάλυσης, όχι μόνο διευκολύνει τη συντήρηση, αλλά επίσης μειώνει τα κόστη ανάπτυξης. Μπορεί να συγχρονιστεί η ανάπτυξη με τη χρήση των υπηρεσιών Excel Services σε σελίδες του Office SharePoint Server και να γίνει κλήση επιχειρηματικών μοντέλων σε υπολογιστικά φύλλα απευθείας από εφαρμογές μέσω Web services που παρέχονται. Οι υπηρεσίες Excel Services παρέχουν τη δυνατότητα στους χρήστες να δημιουργούν ισχυρές λύσεις χωρίς την ανάμιξη προγραμματιστή και παρέχουν πολλούς τρόπους αξιοποίησης και επαναχρησιμοποίησης της επιχειρηματικής λογικής και των αναφορών που δημιουργούνται από τους ειδικούς, δημιουργώντας έτσι ισχυρές, επεκτάσιμες και κλιμακούμενες εφαρμογές με βελτιωμένη ασφάλεια και ταυτόχρονη εξοικονόμηση χρόνου και χρήματος.



Διάγραμμα 7: Υπηρεσίες Excel Services με λίστα δεικτών βασικής απόδοσης (KPIs).

Πλεονεκτήματα του Microsoft Business Intelligence

Μερικά από τα πλεονεκτήματα, που παρέχει η λύση BI της Microsoft, παρουσιάζονται παρακάτω.

- **Αποτελεσματική και αποδοτική σύνδεση ατόμων με πληροφορίες.** Το BI διευκολύνει τους υπεύθυνους λήψης αποφάσεων στην πρόσβαση και στην ανάλυση των πληροφοριών ανά πάσα στιγμή και οπουδήποτε. Οι ενημερωμένες πληροφορίες είναι διαθέσιμες στη θέση εργασίας, συνεργασιών και λήψης αποφάσεων, είτε στην επιφάνεια εργασίας των χρηστών είτε στο Web.
- **Ενίσχυση των εργαζομένων.** Όταν τα αναλυτικά δεδομένα είναι άμεσα διαθέσιμα και κατανοητά, οι εργαζόμενοι μπορούν να αναλαμβάνουν δράση και να υποστηρίζουν τη συνολική στρατηγική της εταιρείας πιο εύκολα. Το BI συμπεριλαμβάνει ισχυρά εργαλεία δυναμικών πινάκων με τα εταιρικά αποτελέσματα, με αναλύσεις και αναφορές, έτσι ώστε ο καθένας μέσα στην εταιρεία να είναι σε θέση να λάβει καλύτερες αποφάσεις πιο γρήγορα.
- **Απλοποίηση της συνεργασίας και της κοινής χρήσης.** Παρέχεται η βελτίωση της αποτελεσματικότητας της επιχείρησής χρησιμοποιώντας τις τεχνολογίες επιχειρηματικής ευφυΐας από τα προϊόντα BI της Microsoft. Μέσω της ενοποίησης των εφαρμογών BI με την πλατφόρμα του Microsoft Office (η οποία χρησιμοποιείται από την πλειονότητα των χρηστών προσωπικών υπολογιστών και είναι ένα γνώριμο περιβάλλον εργασίας για τους χρήστες) επιτρέπει την κοινή χρήση των πληροφοριών εύκολα, σε

ένα διαχειριζόμενο web περιβάλλον με ενισχυμένη ασφάλεια μεταξύ των εργαζομένων του οργανισμού και τους συναδέλφους, τους πελάτες και τους συνεργάτες τους. Το κυριότερο είναι ότι υπάρχει μια κεντρική θέση για παρακολούθηση των βασικών δεικτών απόδοσης του οργανισμού, την πρόσβαση σε αναφορές, την ανάλυση των δεδομένων καθώς και την κοινή χρήση των εγγράφων μέσα στον οργανισμό και την πρόσβαση σε σχετική θεματική ύλη.

- **Βελτίωση της ευθυγράμμισης στόχων.** Το BI βελτιώνει την ευθυγράμμιση των στόχων σε ολόκληρο τον οργανισμό. Τα στελέχη μπορούν να διατυπώνουν σαφώς τη στρατηγική, να ορίζουν στόχους, να παρακολουθούν την απόδοση, να πραγματοποιούν ανάλυση και στη συνέχεια να λαμβάνουν ενημερωμένες αποφάσεις που στηρίζουν τη συνολική στρατηγική της εταιρείας. Οι διευθυντές μπορούν να καθιερώνουν εύκολα γραμμές αρμοδιοτήτων σε ένα χάρτη στρατηγικής και οι εργαζόμενοι μπορούν να προσαρμόζουν τους στόχους τους με τους εταιρικούς στόχους.
- **Παροχή επιχειρηματικών πληροφοριών σε ολόκληρο τον οργανισμό.** Το BI υποστηρίζει ολόκληρο το εύρος των αναγκών επιχειρηματικών πληροφοριών του οργανισμού. Ο στρατηγικός σχεδιασμός είναι απλούστερος όταν χρησιμοποιούνται οικεία εργαλεία, η διαχείριση πληροφοριών είναι ευκολότερη σε ένα κεντρικό και πλήρως ενοποιημένο περιβάλλον BI, ενώ η ανάπτυξη είναι πιο οικονομική όταν χρησιμοποιείται ένα πρότυπο περιβάλλον ανάπτυξης. Με το οικείο περιβάλλον εργασίας του Microsoft Office, την πλατφόρμα επιχειρηματικών πληροφοριών του SQL Server στο παρασκήνιο και την προσαρμοσμένη ανάπτυξη μέσω του Microsoft Visual Studio, το Business Intelligence υποστηρίζει τον κάθε εργαζόμενο — υπεύθυνους πληροφοριών, επαγγελματίες IT και προγραμματιστές — στον οργανισμό.
- **Μείωση των αναγκών εκπαίδευσης.** Με το BI, τα άτομα στον οργανισμό μπορούν να αλληλεπιδρούν με δεδομένα, όπου χρειάζεται. Χρησιμοποιώντας εργαλεία τα οποία είναι οικεία, προσβάσιμα και ευρέως υποστηριζόμενα, τα κόστη εκπαίδευσης μειώνονται ενώ περιορίζεται σημαντικά και η μεταβατική περίοδος εκπαίδευσης.
- **Δυνατότητα σύνθετης ανάλυσης και αναφορών.** Εμπλουτισμένες λειτουργίες πινάκων αποτελεσμάτων, που υποστηρίζονται από αναφορές, διαγράμματα, γραφήματα και αναλύσεις, παρέχουν τη δυνατότητα στους εργαζόμενους να παρακολουθούν τους βασικούς δείκτες απόδοσης ως προς τους βασικούς επιχειρηματικούς στόχους. Η κατανόηση και η ανάλυση των σχέσεων μεταξύ των KPIs και των εταιρικών στόχων σημαίνει ότι μπορούν οι εργαζόμενοι να κατανοήσουν καλύτερα τον τρόπο λειτουργίας της επιχείρησής σήμερα, όχι για παράδειγμα στο τέλος του μήνα, του τριμήνου ή του έτους (ανάλογα με την συχνότητα έκδοσης σχετικών αναφορών απόδοσης), οπότε θα ήταν πολύ αργά για να ληφθούν κάποια μέτρα για την απόδοση του οργανισμού.

Απλά βήματα χρήσης μιας λύσης BI από την πλευρά του Αναλυτή

Το πρόβλημα σε ένα οργανισμό είναι πώς θα εκμεταλλευτούν σε έπακρο τις δυνατότητες που προσφέρει μια λύση Business Intelligence με την χρήση του SQL Server και των εφαρμογών του Microsoft Office. Όλα πρέπει να ξεκινήσουν από την κατανόηση της επιχειρηματικής λειτουργίας του οργανισμού και από τον καθορισμό των στόχων που θέλει να επιτύχει. Δεν πρέπει μόνο να σκέφτονται τα δεδομένα που συγκεντρώνουν κατά τη διάρκεια της λειτουργίας του οργανισμού αλλά πρέπει να σκέφτονται τι θέλουν να μάθουν από τα δεδομένα. Τι πληροφορίες χρειάζονται για τις λειτουργίες του οργανισμού. Πώς συγκεντρώνονται τώρα τα δεδομένα αυτά ή πως θα πρέπει να συγκεντρώνονται στο μέλλον.

Αφού αποφασιστεί το τι χρειάζεται ο οργανισμός από τα δεδομένα, πρέπει να καθοριστούν οι δείκτες απόδοσης (KPIs). Οι δείκτες απόδοσης είναι οι μετρικές που καθορίζουν την επιτυχία και ενημερώνουν πως λειτουργεί ο οργανισμός. Τα θέματα που πρέπει να αντιμετωπιστούν είναι που θα βρίσκονται τα δεδομένα που χρησιμοποιούνται, πώς θα αναλυθούν τα δεδομένα για να ανακαλυφθούν τάσεις σε αυτά, να εντοπιστούν ευκαιρίες και να μειωθούν οι αναποτελεσματικότητες.

Η διαδικασία δημιουργίας μιας λύσης Business Intelligence, από την πλευρά του αναλυτή παρουσιάζεται στα παρακάτω βήματα:

1. Πρέπει να καθοριστεί από τον αναλυτή μια **στρατηγική σκέψη** για τον οργανισμό, καθορίζοντας τι πρέπει να μάθουν οι υπεύθυνοι από τα δεδομένα που χρησιμοποιούνται. Πρέπει να χρησιμοποιηθούν από τα πρώτα βήματα οι υπεύθυνοι του οργανισμού στην διαδικασίας, για να ληφθεί υπόψη η άποψη τους με σκοπό να δημιουργηθεί μια αποτελεσματική BI στρατηγική.
2. Πρέπει να δημιουργηθεί μια **λίστα με τα δεδομένα** που παράγονται και χρησιμοποιούνται σήμερα, συμπεριλαμβάνοντας στοιχεία επικοινωνίας με τους πελάτες, τιμολόγια, λογιστικές κινήσεις, δραστηριότητα στο διαδίκτυο, στοιχεία πωλήσεων και μισθοδοσίας κ.α.
3. Πρέπει να γίνουν **συζητήσεις με το προσωπικό** για να εντοπιστούν αν υπάρχουν άλλα δεδομένα που χρειάζονται οι εργαζόμενοι να χρησιμοποιούν έτσι ώστε να ελέγχουν την καλή λειτουργία του οργανισμού. Πρέπει να καθοριστούν τυχόν κενά, δεδομένα δηλαδή που απαιτούνται αλλά δεν συγκεντρώνονται. Επίσης, ποια είναι τα υπάρχοντα δεδομένα τα οποία θα ήταν περισσότερο χρήσιμα αν χρησιμοποιούνταν διαφορετικά σε άλλες αναφορές. Όλες οι απαιτήσεις σχετικά με τα δεδομένα πρέπει να καταγραφούν σε μια λίστα.
4. Πρέπει να γίνει **κατηγοριοποίηση των δεδομένων** σύμφωνα με τον τύπο τους, για παράδειγμα πληροφορίες πελατών, στοιχεία μάρκετινγκ, οικονομικές πληροφορίες, στοιχεία πωλήσεων μέσω internet, κλπ. Πρέπει οι κατηγορίες που θα καταγραφούν να αντικατοπτρίζουν το πώς αντιλαμβάνεται ο αναλυτής την λειτουργία του οργανισμού και πώς θα προσφέρουν νόημα σε όλους τους εργαζόμενους στον οργανισμό.
5. Πρέπει να γίνει **έλεγχος από τον αναλυτή** ότι έχουν καταγραφεί όλες οι πληροφορίες. Είναι αναγκαίο να γίνουν έλεγχοι και αναθεωρήσεις μαζί με τους διοικούντες τον οργανισμό έτσι ώστε να επιτευχθεί η πληρότητα των δεδομένων.
6. Αφού ο αναλυτής έχει μάθει για τα δεδομένα που πρέπει να παρακολουθεί, πρέπει να εργαστεί μαζί με τους υπευθύνους των τμημάτων για να καταγράψει τα δεδομένα στις κατάλληλες **μετρικές**. Με λίγα λόγια, πρέπει να δημιουργηθούν οι κατάλληλοι **δείκτες απόδοσης** (Key Performance Indicators – KPIs).
7. Σχεδιασμός μιας **αρχικής κάρτας καταγραφής των στόχων (scorecard)**, που να περιλαμβάνει όλους του δείκτες απόδοσης, η οποία πρέπει να αναθεωρηθεί από τον αναλυτή μαζί με τους υπευθύνους του οργανισμού. Η scorecard πρέπει να τροποποιηθεί μαζί με τα δεδομένα εισόδου έτσι ώστε να επιβεβαιωθεί ότι οι υπεύθυνοι των τμημάτων καθορίζουν την λειτουργία του τομέα τους βάσει των μετρικών που υπάρχουν στην scorecard. Με λίγα λόγια, **αλλαγή νοοτροπίας** των υπευθύνων των τμημάτων έτσι ώστε να καθορίζουν την λειτουργία τους βάσει καθορισμένων δεικτών απόδοσης.
8. Αφού σχεδιαστεί ένα πρότυπο της λειτουργικής κάρτας επιδόσεων, τα δεδομένα της οποίας χρησιμοποιούν οι υπεύθυνοι του οργανισμού για να παρακολουθούν την λειτουργία του οργανισμού κάνοντας τις απαραίτητες διορθωτικές κινήσεις, πρέπει να καθοριστούν τα **πρότυπα αναφορών** που θα χρησιμοποιούνται για να μοιράζουν την αξία της πληροφόρησης σε όλο τον οργανισμό.
9. **Σχεδιασμός του συστήματος BI** που θα χρησιμοποιηθεί, έτσι ώστε να λαμβάνει αυτόματα τα δεδομένα που χρειάζεται για να λειτουργήσει.
10. **Δημιουργία των καρτών απόδοσης** που θα χρησιμοποιηθούν στο σύστημα και **αυτοματοποίηση της διαδικασίας εξαγωγής των αναφορών** με τα αποτελέσματα των δεικτών απόδοσης. Το τελευταίο βήμα δημιουργεί το **Business Intelligence σύστημα**, που θα «τρέξει» στον οργανισμό για να υποστηρίξει την αποτελεσματικότητα της λειτουργίας τους και για να μεγιστοποιήσει τις ευκαιρίες ανάπτυξης.

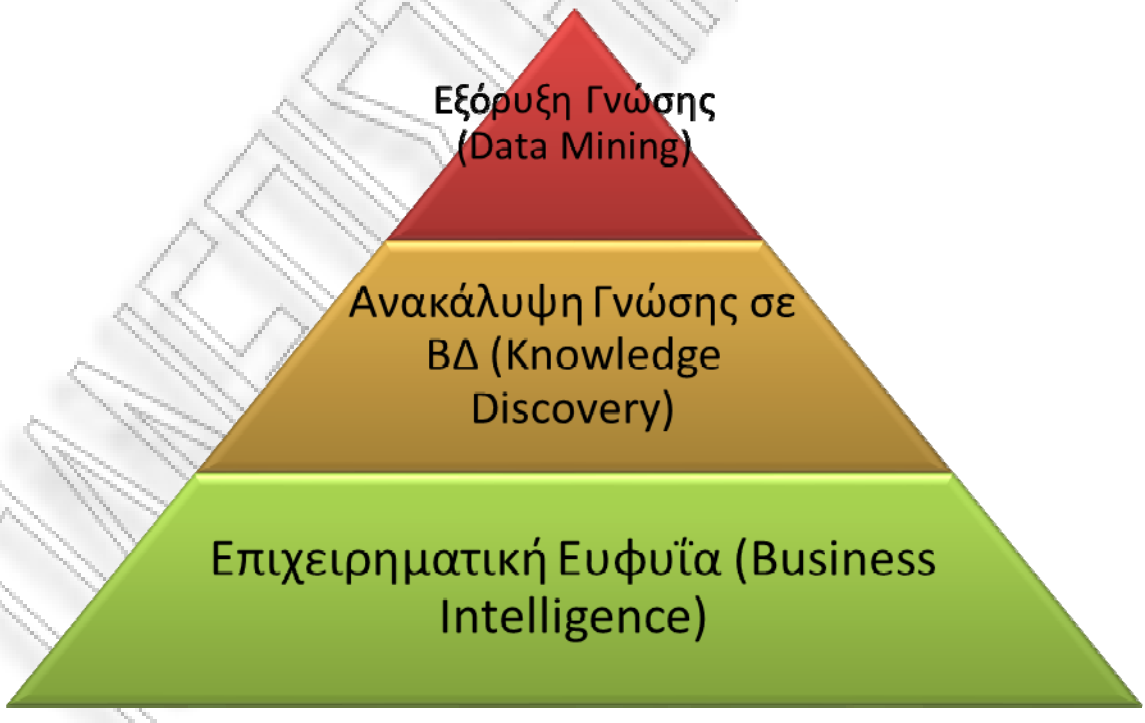
2. Εξόρυξη Γνώσης

Η λειτουργία της Εξόρυξης Γνώσης σαν μέρος του BI

Η επιχειρηματική ευφυΐα (Business Intelligence), όπως την παρουσιάσαμε στο προηγούμενο κεφάλαιο, είναι ένα σύνολο μεθόδων που χρησιμοποιούνται σε έναν οργανισμό για αποθήκευση και παρουσίαση σημαντικών πληροφοριών, έτσι ώστε ο κάθε ενδιαφερόμενος (εργαζόμενοι, διοικητικοί υπάλληλοι, διευθυντές που λαμβάνουν αποφάσεις κλπ) στον οργανισμό να μπορεί εύκολα και γρήγορα να παίρνει απαντήσεις στα ερωτήματα του και να λαμβάνει τις πιο σωστές αποφάσεις. Σε πολλές περιπτώσεις όμως, τα δεδομένα που διαχειρίζονται οι εφαρμογές BI μπορούν να «κρύβουν» απαντήσεις σε πολλά ερωτήματα, τα οποία όμως ακόμα δεν έχουν δημιουργηθεί. Τα δεδομένα μπορεί να κρύβουν τάσεις (trends), συσχετίσεις και εξαρτήσεις σε ένα τέτοιο επίπεδο λεπτομέρειας, έτσι ώστε ένας άνθρωπος από μόνος του (βλέποντας δηλαδή τα αποτελέσματα της ανάλυσης χωρίς κάποια περαιτέρω επεξεργασία, ιδίως με την χρήση κάποιας εξειδικευμένης εφαρμογής ή κάποιου εργαλείου) να μην μπορεί να τα αντιληφθεί. Οι συσχετίσεις αυτές στα δεδομένα, με λίγα λόγια η ευφυΐα που υπάρχει μέσα στο τεράστιο πλήθος των δεδομένων, μπορεί να καταγραφεί χρησιμοποιώντας τεχνικές Ανακάλυψης Γνώσης από μεγάλες Βάσεις Δεδομένων (Knowledge Discovery in Databases – KDD).

Οι τεχνικές Ανακάλυψης Γνώσης από Βάσεις Δεδομένων περιλαμβάνουν τις τεχνικές του προσδιορισμού των επιχειρηματικώς προβλημάτων προς Επίλυση (Business Cases Definition), την προετοιμασία των Δεδομένων (Data Preparation), τις τεχνικές Εξόρυξης Δεδομένων (Data Mining) και την Αποτίμηση των Αποτελεσμάτων (Evaluation).

Η θέση των τεχνικών Εξόρυξης Γνώσης στον οργανισμό, σε σχέση με τις λειτουργίες Business Intelligence σε ένα οργανισμό παρουσιάζεται στο παρακάτω διάγραμμα.



Διάγραμμα 8: Σχέση Εξόρυξης Γνώσης και BI

Τι είναι η Εξόρυξη Γνώσης (Data mining)

Ο όρος Εξόρυξη Γνώσης μπορεί απλά να περιγραφεί σαν την διαδικασία εξαγωγής (εξόρυξης) γνώσης και ανακάλυψη προτύπων από μεγάλες ποσότητες δεδομένων. Η βασική έννοια του όρων που χρησιμοποιείται σχετίζεται με την ανακάλυψη σημαντικών στοιχείων μέσα από ένα μεγάλο πλήθος δεδομένων.

Αυτή η δυνατότητα που προσφέρει η Εξόρυξη Γνώσης, η ανακάλυψη δηλαδή δεδομένων και προτύπων (patterns) ανάμεσα στα δεδομένα μέσα από ένα τεράστιο πλήθος πληροφοριών αλλά και η άμεση ανάγκη να μετατραπούν αυτά τα δεδομένα σε χρήσιμες πληροφορίες και γνώση έχουν επιτελέσει στο να θεωρείται η τεχνική της Εξόρυξης Γνώσης μια πολύ αναγκαία τεχνική για τους οργανισμούς. Η πληροφορία και η γνώση που ανακτώνται από τις τεχνικές Εξόρυξης Γνώσης μπορούν να χρησιμοποιηθούν σε πολλές εφαρμογές – από ανάλυση αγορών, ανακάλυψη απάτης, διατήρηση πελατών μέχρι τον έλεγχο παραγωγής και σε διάφορα άλλα επιστημονικά πεδία.

Η Εξόρυξη Γνώσης μπορεί να θεωρηθεί σαν αποτέλεσμα της φυσικής εξέλιξης της τεχνολογίας πληροφοριών. Η τεχνολογία των βάσεων δεδομένων μέσα από την εξέλιξη της παρέχει τις παρακάτω λειτουργίες:

- Συλλογή δεδομένων και δημιουργία Βάσεων Δεδομένων
- Διαχείριση δεδομένων (αποθήκευση και ανάκτηση δεδομένων)
- Προηγμένη Ανάλυση δεδομένων (συμπεριλαμβανομένων τις Αποθήκες δεδομένων και την Εξόρυξη Γνώσης)

Πληθώρα πληροφοριών – ανάγκες που δημιουργούνται

Η σταθερή και τεράστια ανάπτυξη της τεχνολογίας πληροφορικής και των τεχνολογιών συλλογής δεδομένων και των μηχανισμών δημιουργίας και διαχείρισης Βάσεων Δεδομένων έχουν σαν αποτέλεσμα την παροχή μεγάλων ποσοτήτων δεδομένων και πληροφοριών.

Τα δεδομένα πλέον αποθηκεύονται σε διάφορα είδη Βάσεων Δεδομένων και σε διάφορες άλλες θέσεις αποθήκευσης. Αυτό έχει σαν αποτέλεσμα την ανάγκη δημιουργίας μιας νέας αρχιτεκτονικής αποθήκευσης δεδομένων, την Αποθήκη Δεδομένων (Data Warehouse), η οποία συγκεντρώνει δεδομένα από ετερογενείς πηγές δεδομένων κάτω από ένα κοινό σχήμα σε ένα μοναδικό σημείο, με σκοπό να εξυπηρετήσει την λήψη αποφάσεων.

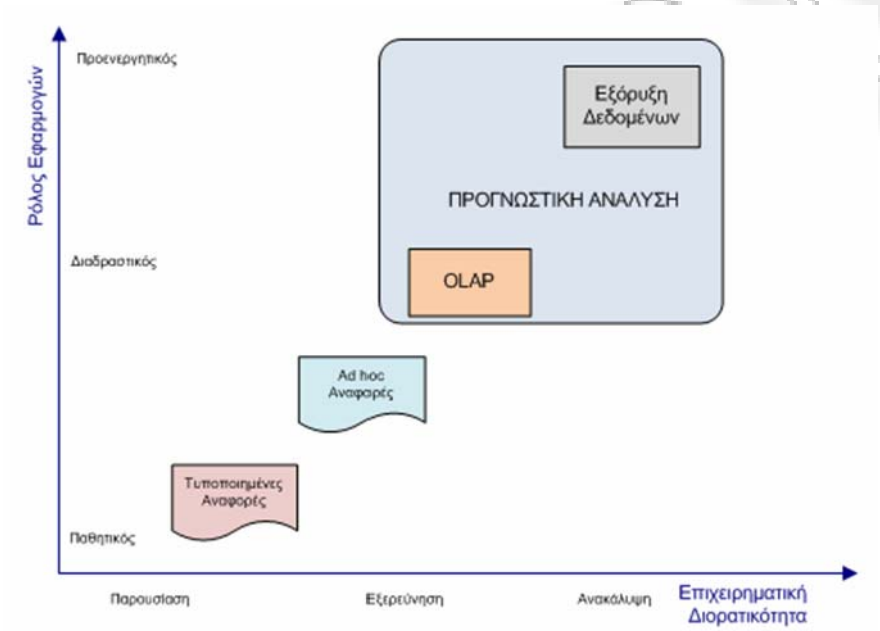
Οι τεχνολογίες της Αποθήκης Δεδομένων περιλαμβάνουν τον καθαρισμό δεδομένων, την ολοκλήρωση δεδομένων και την αναλυτική επεξεργασία δεδομένων (OLAP – On-Line Analytical Processing), η οποία περιλαμβάνει τεχνικές ανάλυσης με λειτουργικότητες όπως άθροιση, συγκέντρωση δεδομένων και σύνοψη αποτελεσμάτων καθώς και την δυνατότητα προβολής της πληροφορίας από διάφορες πηγές.

Παρόλο που τα εργαλεία OLAP υποστηρίζουν την ανάλυση πολυδιάστατων δεδομένων και την λήψη αποφάσεων, **είναι αναγκαία η χρήση επιπλέον εργαλείων** για ανάλυση σε μεγαλύτερο βάθος, όπως η κατηγοριοποίηση δεδομένων (data classification), η ομαδοποίηση (clustering) των διαφόρων οντοτήτων των δεδομένων και η κατανόηση των αλλαγών των δεδομένων στην διάρκεια του χρόνου. Η πληθώρα των δεδομένων αυτών καθώς και η ανάγκη ύπαρξης δεδομένων για μεγαλύτερη ανάλυση έχει χαρακτηριστεί σαν μια κατάσταση με πλούσια δεδομένα αλλά με φτωχές πληροφορίες. Η τεράστια αύξηση των ποσοτήτων που συγκεντρώνονται από πολλές πηγές (διάφοροι τύποι Βάσεων Δεδομένων, αρχεία χρηστών, παγκόσμιος ιστός κ.α.) και η αποθήκευση τους σε διάφορες αποθήκες δεδομένων έχει υπερβεί την ανθρώπινη ικανότητα κατανόησης των πληροφοριών, χωρίς την χρήση εξειδικευμένων εργαλείων.

Η συλλογή δεδομένων σε τεράστιους αποθηκευτικούς χώρους έχει σαν αποτέλεσμα την δημιουργία αρχείων ιστορικών δεδομένων (data archives) που σπάνια χρησιμοποιούνται από

τους χρήστες. Έτσι, σημαντικές αποφάσεις λαμβάνονται όχι βάσει των πλούσιων πληροφοριών που υπάρχουν αλλά περισσότερο στην διαίσθηση των ατόμων που λαμβάνουν αποφάσεις, απλώς επειδή οι λήπτες αποφάσεων δεν έχουν τα κατάλληλα εργαλεία για να εξάγουν την πολύτιμη γνώση που υπάρχει μέσα στην πληθώρα των δεδομένων.

Όλα αυτά έχουν δημιουργήσει όλες τις απαιτήσεις για την δημιουργία εξελιγμένων μηχανισμών ανάλυσης δεδομένων. Τα εργαλεία Εξόρυξης Γνώσης μπορούν να ανακαλύψουν πρότυπα στα δεδομένα και να συμβάλλουν αποτελεσματικά στην διαμόρφωση της στρατηγικής του οργανισμού μέσα από την γνώση που προσφέρουν. Ο ρόλος της λειτουργίας της Εξόρυξης Γνώσης, όσον αφορά στον γενικό ρόλο των διάφορων εφαρμογών σε ένα οργανισμό και στην επιχειρηματική διορατικότητα (δηλαδή κατά πόσο μπορούμε να κατανοήσουμε τις λειτουργίες σε ένα οργανισμό και τις ανάγκες του) παρουσιάζεται στο παρακάτω διάγραμμα. Εκτός από τις εφαρμογές που χρησιμοποιούν αναφορές (τυποποιημένες αναφορές ή αναφορές που έχουν σχεδιαστεί για συγκεκριμένους σκοπούς), οι οποίες παρουσιάζουν στοιχεία σχετικά με τον οργανισμό (στοιχεία με Πίνακες Αποτελεσμάτων, Διαγράμματα, στατιστικά στοιχείο κλπ.), υπάρχουν και οι ομάδες εφαρμογών, που υποστηρίζουν την **Προγνωστική Ανάλυση** (**Predictive Analysis**).



Διάγραμμα 9: Τοποθέτηση Εφαρμογών Εξόρυξης Γνώσης στον οργανισμό

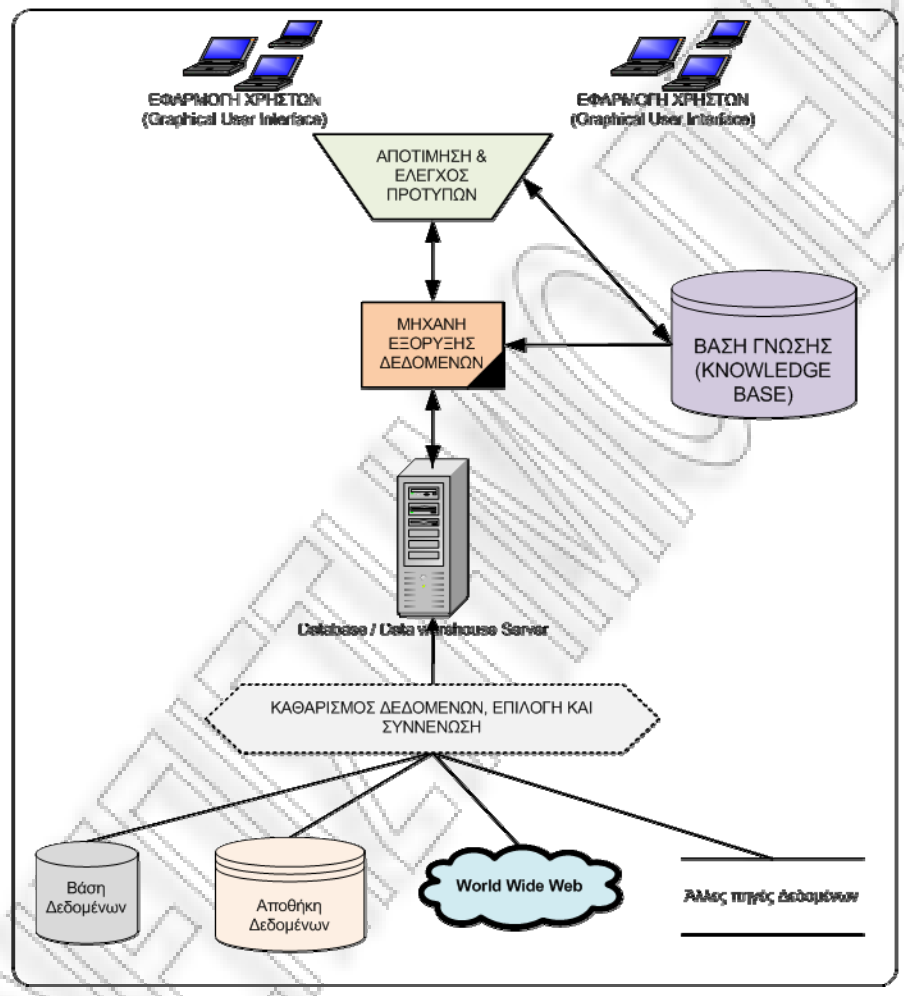
Με τον όρο **Προγνωστική Ανάλυση** εννοούμε την χρήση τεχνικών από την Στατιστική, την Εξόρυξη Γνώσης και την Θεωρία Παιγνίων, οι οποίες αναλύουν τρέχοντα και ιστορικά δεδομένα για να κάνουν προβλέψεις για μελλοντικά γεγονότα. Στους οργανισμούς, τα μοντέλα προγνωστικής ανάλυσης εκμεταλλεύονται τα πρότυπα τα οποία έχουν βρεθεί σε ιστορικά δεδομένα και σε δεδομένα συναλλαγών. Τα δεδομένα αυτά συγκεντρώνονται, δημιουργούνται τα απαραίτητα στατιστικά μοντέλα και μοντέλα Εξόρυξης Γνώσης και γίνονται οι κατάλληλες προβλέψεις για να ανακαλύψουν πιθανά ρίσκα και ευκαιρίες. Στην συνέχεια, όταν εισάγονται καινούργια δεδομένα στα προγνωστικά μοντέλα, τα αποτελέσματα αποτιμώνται καλύτερα και επανεξετάζονται και όπου απαιτείται διορθώνονται ή βελτιώνονται για το καλύτερο κάθε φορά αποτέλεσμα.

Η λειτουργία της Εξόρυξης Γνώσης περιλαμβάνει την ενσωμάτωση τεχνικών από πολλούς τομείς, όπως η τεχνολογία Βάσεων Δεδομένων και Αποθηκών Δεδομένων, Στατιστική, Μηχανική Μάθηση, Αναγνώριση Προτύπων, Νευρωνικά Δίκτυα, Οπτικοποίηση Δεδομένων,

Ανάκτηση Πληροφοριών, Επεξεργασία Εικόνας και Σήματος, Χωρική και Χρονική Ανάλυση δεδομένων κα. Δίνεται έμφαση κυρίως στις αποτελεσματικές και επεκτάσιμες τεχνικές Εξόρυξης Γνώσης.

Αρχιτεκτονική ενός συστήματος Εξόρυξης Γνώσης

Μία τυπική αρχιτεκτονική ενός συστήματος Εξόρυξης Γνώσης παρουσιάζεται στο παρακάτω διάγραμμα.



Διάγραμμα 10: Αρχιτεκτονική ενός τυπικού συστήματος Εξόρυξης Γνώσης

Τα βασικά επίπεδα που αποτελούν ένα σύστημα Εξόρυξης Γνώσης, παρουσιάζονται παρακάτω:

- **Βάση Δεδομένων, Αποθήκη Δεδομένων, World Wide Web** ή **Άλλες Πηγές Δεδομένων**: Οι πηγές που προέρχονται τα δεδομένα, τα οποία εισάγονται στο σύστημα Εξόρυξης Γνώσης. Οι πηγές μπορεί να περιέχουν δεδομένα με διαφορετική μορφή και οι πηγές να είναι ετερογενείς και διαφορετικής τεχνολογίας. Για το λόγο αυτό, απαιτείται η παρεμβολή διαφόρων τεχνικών Συνένωσης, Επιλογής και Καθαρισμού (ETL – Extract, Transform, Load) των δεδομένων έτσι ώστε να έχουν την κατάλληλη δομή και περιεχόμενο για να εισαχθούν στην Βάση ή Αποθήκη Δεδομένων του επόμενου επιπέδου.

- **Database / Data Warehouse server**: Το επίπεδο που περιέχει τον server, που διαχειρίζεται την Βάση ή την Αποθήκη Δεδομένων του συστήματος Εξόρυξης Γνώσης. Το επίπεδο αυτό διαχειρίζεται τα απαραίτητα δεδομένα, που απαιτούνται σύμφωνα με τις απαιτήσεις των χρηστών της εφαρμογής Εξόρυξης Γνώσης.
- **Βάση Γνώσης (Knowledge Base)**: Το επίπεδο αυτό περιέχει την κύρια γνώση, που χρησιμοποιείται για να καθοδηγήσει την έρευνα ή να αποτιμήσει την χρησιμότητα των προτύπων, που προκύπτουν από την διαδικασία. Η γνώση αυτή μπορεί να περιέχει διάφορες ιεραρχίες εννοιών, που χρησιμοποιούνται για να οργανώσουν τα χαρακτηριστικά σε διάφορα επίπεδα αφαίρεσης. Επίσης, στο επίπεδο αυτό μπορεί να περιέχονται οι πεποιθήσεις των χρηστών, οι οποίες μπορούν να χρησιμοποιηθούν για την αξιολόγηση του ενδιαφέροντος του κάθε προτύπου. Τέλος, άλλα παραδείγματα της γνώσης, που περιέχει το επίπεδο είναι κάποια όρια και περιορισμοί που εφαρμόζονται στις τεχνικές Εξόρυξης Γνώσης και κάποια άλλα μετά-δεδομένα (όπως οι περιγραφές δεδομένων από ετερογενείς πηγές δεδομένων).
- **Μηχανή Εξόρυξης Γνώσης**: Είναι το πιο σημαντικό επίπεδο του συστήματος Εξόρυξης Γνώσης και αποτελείται από ένα σύνολο λειτουργικών μονάδων για εκτέλεση λειτουργιών, όπως η κατηγοριοποίηση, η ανάλυση συσχετίσεων, η πρόβλεψη, η ανάλυση ομάδων, η ανάλυση ακραίων τιμών και άλλες λειτουργίες Εξόρυξης Γνώσης.
- **Αποτίμηση και Έλεγχος Προτύπων**: Το επίπεδο αυτό κυρίως περιέχει χρήσιμες μετρικές και αλληλεπιδρά με τις διάφορες μεθόδους Εξόρυξης Γνώσης, έτσι ώστε να επικεντρωθεί η έρευνα σε ενδιαφέροντα πρότυπα. Επίσης, μπορεί να χρησιμοποιήσει διάφορα όρια στην εξαγωγή των αποτελεσμάτων, έτσι ώστε να παραχθούν πρότυπα, που όντως προσφέρουν κάποιο ενδιαφέρον. Το επίπεδο της αποτίμησης των προτύπων μπορεί να ενσωματωθεί στις λειτουργίες Εξόρυξης Γνώσης, ανάλογα βέβαια με την τεχνική που χρησιμοποιείται κάθε φορά (για παράδειγμα, μπορούν να εφαρμοστούν συγκεκριμένα όρια στα αποτελέσματα της Τεχνικής Ομαδοποίησης (clustering) έτσι ώστε να παραχθούν μόνο συγκεκριμένες ομάδες, με ενδιαφέροντα αποτελέσματα ή να χρησιμοποιηθούν κάποια συγκεκριμένα όρια σε μια τεχνική εύρεσης Κανόνων Συσχέτισης, έτσι ώστε να ανακαλυφθούν χρήσιμοι κανόνες για κάποια χαρακτηριστικά των δεδομένων). Για το λόγο αυτό, είναι σημαντικό να χρησιμοποιηθεί η αξιολόγηση των προτύπων σε όλη την διαδικασία Εξόρυξης Γνώσης, σε μεγάλο βάθος έτσι ώστε να περιοριστεί η έρευνα των προτύπων σε πραγματικά ενδιαφέροντα αποτελέσματα.
- **Εφαρμογή (Graphical User Interface)**: Το επίπεδο αυτό αναλαμβάνει την επικοινωνία μεταξύ των χρηστών και του συστήματος Εξόρυξης Γνώσης, επιτρέπει στους χρήστες να αλληλεπιδρούν με το σύστημα χρησιμοποιώντας ερωτήματα ή λειτουργίες Εξόρυξης Γνώσης και παρέχει πληροφορίες για τα αποτελέσματα της διαδικασίας Εξόρυξης Γνώσης. Επιπλέον, το επίπεδο αυτό επιτρέχει στους χρήστες να περιηγηθούν στα σχήματα και τα πρότυπα δεδομένων στη Βάση Δεδομένων του συστήματος, να αξιολογήσουν τα παραγόμενα πρότυπα και να τα οπτικοποιήσουν σε διάφορες μορφές (διάφορα διαγράμματα που οπτικοποιούν τα αποτελέσματα της Εξόρυξης Γνώσης για να γίνουν καλύτερα αντιληπτά από τους χρήστες).

Τεχνικές – Αλγόριθμοι Εξόρυξης Γνώσης

Στην συνέχεια του Κεφαλαίου θα παρουσιάσουμε τις βασικότερες τεχνικές Εξόρυξης Γνώσης, που χρησιμοποιούνται από τις μεγάλες εφαρμογές Εξόρυξης Γνώσης. Ειδικότερα, θα παρουσιάσουμε την τεχνική της Κατηγοριοποίησης (Classification) (με πιο γνωστή εφαρμογή της τα Δέντρα Απόφασης) για την τοποθέτηση των δεδομένων σε προκαθορισμένες ομάδες, την τεχνική της Ομαδοποίησης (Clustering) για την ανακάλυψη ομάδων ανάμεσα στα δεδομένα και την τεχνική των Κανόνων Συσχέτισης (Association Rules), με την κύρια χρήση τους την Ανάλυση του Καλαθιού Αγοράς των πελατών (Market Basket Analysis) για την ανακάλυψη συσχετίσεων μεταξύ προϊόντων μέσα από ένα σύνολο συναλλαγών. Η παρουσίαση των τεχνικών αυτών Εξόρυξης Γνώσης δεν θα γίνει σε μεγάλο βάθος, αλλά θα παρουσιαστούν σε γενικές γραμμές τα βασικότερα χαρακτηριστικά τους. Οι αλγόριθμοι Εξόρυξης Γνώσης, που θα αναφερθούμε χρησιμοποιούνται και από την πλατφόρμα διαχείρισης Επιχειρησιακής Ευφυΐας της Microsoft, τα SSAS που θα παρουσιαστούν στο Κεφάλαιο 4.

Κατηγοριοποίηση – Classification

Η κατηγοριοποίηση (classification) είναι η πιο γνωστή και πιο δημοφιλής τεχνική της Εξόρυξης Γνώσης. Η κατηγοριοποίηση δεδομένων είναι μια διαδικασία που αποτελείται από δύο βήματα. Στο πρώτο βήμα, ένα μοντέλο ταξινόμησης (ένας ταξινομητής – classifier) χτίζεται και περιγράφει ένα προκαθορισμένο σύνολο κατηγοριών δεδομένων ή εννοιών. Το βήμα αυτό είναι το βήμα της εκμάθησης (ή εκπαίδευσης), όπου ένας αλγόριθμος ταξινόμησης χτίζει το μοντέλο με την ανάλυση (ή την «εκμάθηση από») ενός συνόλου εκπαίδευσης που απαρτίζεται από εγγραφές (πλειάδες) από την βάση δεδομένων και τις σχετικές ετικέτες που περιγράφουν τις κατηγορίες. Κάθε πλειάδα, Χ, θεωρείται ότι ανήκει σε μια προκαθορισμένη κατηγορία όταν καθορίζεται από ένα άλλο χαρακτηριστικό των δεδομένων, που ονομάζεται χαρακτηριστικό της κατηγορίας (class label attribute). Το χαρακτηριστικό της κατηγορίας έχει διακριτές τιμές και είναι μη-ταξινομημένο. Κάθε τιμή του χαρακτηριστικού της ομάδας χρησιμοποιείται για να προσδιορίζει την ίδια την κατηγορία ή τάξη. Η τεχνική, λοιπόν, της κατηγοριοποίησης απεικονίζει τα δεδομένα σε προκαθορισμένες ομάδες ή κατηγορίες (classes). Ουσιαστικά, για κάθε τιμή γνωρίσματος από μια βάση δεδομένων προσπαθούμε να την εκχωρήσουμε σε μια ομάδα, μέσα από ένα προκαθορισμένο σύνολο από ομάδες.

Επειδή το χαρακτηριστικό της κάθε παρέχεται στον αλγόριθμο κατηγοριοποίησης και οι κατηγορίες (κλάσεις) καθορίζονται πριν εξεταστούν τα δεδομένα, η τεχνική ονομάζεται και ***εποπτευόμενη μάθηση***. Οι κλάσεις καθορίζονται συνήθως κοιτάζοντας τα χαρακτηριστικά δεδομένων που είναι γνωστό ότι ανήκουν στις κλάσεις αυτές. Η πρόβλεψη, στη συνέχεια μπορεί να θεωρηθεί σαν η τοποθέτηση της τιμής ενός γνωρίσματος σε μια κλάση, από το σύνολο των πιθανών κλάσεων που έχουν οριστεί κατά την διάρκεια εκτέλεσης του αλγορίθμου.

Ένα είδος κατηγοριοποίησης είναι η αναγνώριση προτύπου (pattern recognition), όπου ένα πρότυπο εισόδου τοποθετείται σε μια από τις διάφορες κατηγορίες, με βάση την εγγύτητα του (κατά ποσό πλησιάζει η τιμή της μεταβλητής εισόδου δηλαδή) ως προς τις προκαθορισμένες κατηγορίες.

Όλες οι προσεγγίσεις στην εκτέλεση των τεχνικών κατηγοριοποίησης προϋποθέτουν την γνώση των δεδομένων. Για το λόγο αυτό, χρησιμοποιείται ένα ποσοστό του συνόλου δεδομένων για την εκπαίδευση της μεθόδου (δεδομένα εκπαίδευσης – training data) και τον καθορισμό των παραμέτρων που απαιτούνται από την μέθοδο κατηγοριοποίησης.

Το πρόβλημα της Κατηγοριοποίησης συνήθως υλοποιείται σε δύο φάσεις.

1. Δημιουργούμε ένα μοντέλο με την βοήθεια των δεδομένων εκπαίδευσης. Πιο συγκεκριμένα, το βήμα αυτό έχει σαν **είσοδο** τα δεδομένα εκπαίδευσης καθώς και τον ορισμό των κατηγοριών για τα δεδομένα και σαν **έξοδο** έχει τον ορισμό του μοντέλου που αναπτύχθηκε. Το μοντέλο αυτό κατηγοριοποιεί τα δεδομένα, τα εντάσσει δηλαδή σε συγκεκριμένες κατηγορίες (κλάσεις).

2. Εφαρμόζουμε το μοντέλο που έχουμε δημιουργήσει κατηγοριοποιώντας τις πλειάδες των δεδομένων που θέλουμε να εξετάσουμε. Συνήθως τα δεδομένα αυτά προέρχονται μέσα από μια Βάση δεδομένων.

Το βασικό πρόβλημα που πρέπει να αντιμετωπιστεί στο βήμα αυτό είναι η **ακρίβεια της κατηγοριοποίησης**. Πρέπει να υπολογιστεί η ακρίβεια και η αξιοπιστία του μοντέλου που παράγεται, βρίσκοντας το κατάλληλο σύνολο δεδομένων για να γίνει η μέτρηση αυτή. Στην περίπτωση που χρησιμοποιηθεί το σύνολο εκπαίδευσης, που χρησιμοποιήθηκε για να δημιουργηθεί το μοντέλο, τότε το αποτέλεσμα μπορεί να είναι αισιόδοξο, γιατί το μοντέλο θα τείνει να ταιριάζει με το σύνολο εκπαίδευσης (γιατί το σύνολο δεδομένων αυτό χρησιμοποιήθηκε για να κατασκευαστεί το μοντέλο και είναι πολύ πιθανό κάποιες ανωμαλίες στα δεδομένα αυτά να έχουν διορθωθεί). Έτσι πρέπει να χρησιμοποιηθεί ένα σύνολο δεδομένων με διαφορετικές πληροφορίες, τα οποία έχουν επιλεχθεί τυχαία από το γενικό σύνολο.

Η ακρίβεια του μοντέλου κατηγοριοποίησης για ένα δοσμένο σύνολο δεδομένων είναι το ποσοστό των δοκιμαστικών δεδομένων, τα οποία κατηγοριοποιούνται σωστά από το μοντέλο. Είναι σημαντικό να συμπεριλαμβάνεται στα δεδομένα εκπαίδευσης και το αντίστοιχο χαρακτηριστικό (μια επιπλέον στήλη) που προσδιορίζει την κάθε ομάδα. Η τιμή του χαρακτηριστικού αυτού στα δεδομένα ελέγχου (που ουσιαστικά προσδιορίζει την ομάδα) συγκρίνεται με την προβλεπόμενη κατηγοριοποίηση (με την τοποθέτηση δηλαδή σε ομάδα) που προτείνει το μοντέλο κατηγοριοποίησης. Έτσι για όλο το σύνολο δεδομένων ελέγχου παράγεται ένα ποσοστό με τις πλειάδες δεδομένων που οποίες έχουν κατηγοριοποιηθεί σωστά. Στην περίπτωση που το ποσοστό αυτό δεν είναι χαμηλότερο από ένα καθορισμένο όριο, τότε το μοντέλο κατηγοριοποίησης θεωρείται αποδεκτό και μπορεί να χρησιμοποιηθεί για να κατηγοριοποιήσει μελλοντικά σύνολα δεδομένων.

Μέθοδοι Κατηγοριοποίησης

Για την επίλυση του προβλήματος της κατηγοριοποίησης, υπάρχουν τρείς βασικές μέθοδοι:

- **Καθορισμός των ορίων**, όπου η κατηγοριοποίηση εκτελείται με την διαίρεση του χώρου των δεδομένων σε περιοχές, όπου η κάθε περιοχή συνδέεται με μια κατηγορία.
- **Χρήση κατανομών πιθανότητας**, όπου για κάθε κατηγορία δίνεται μια συνάρτηση κατανομής πιθανότητας, υπολογισμένη για ένα σημείο. Έτσι υπάρχει μια εκτίμηση της πιθανότητας ένα σημείο να ανήκει σε μια κατηγορία.
- **Χρήση εκ των υστέρων πιθανοτήτων**, όπου με δεδομένη μια τιμή καθορίζουμε μια πιθανότητα ότι το σημείο που επιλέγουμε ανήκει σε μια κατηγορία. Χρησιμοποιείται μια εκ των υστέρων πιθανότητα (posterior probability). Έτσι για την υλοποίηση της κατηγοριοποίησης, καθορίζονται εκ των υστέρων οι πιθανότητες κάθε σημείο να ανήκει σε κάθε κατηγορία και στη συνέχεια επιλέγονται για το κάθε σημείο οι κατηγορίες που εμφανίζουν την μεγαλύτερη πιθανότητα.

Κατηγοριοποίηση με Δέντρα Απόφασης

Με την τεχνική της δημιουργίας δέντρων απόφασης, κατασκευάζεται ένα δέντρο για να υποστηρίξει την διαδικασία της κατηγοριοποίησης. Ένα δέντρο απόφασης (Decision Tree) είναι μια τεχνική μοντελοποίησης πρόβλεψης, στο οποίο η ρίζα και κάθε εσωτερικός κόμβος έχει χαρακτηριστεί με μία ερώτηση. Τα τόξα που προέρχονται από κάθε κόμβο αντιπροσωπεύουν κάθε πιθανή απάντηση στην σχετική ερώτηση. Κάθε φύλο του δέντρου αντιπροσωπεύει μια πρόβλεψη της λύσης στο πρόβλημα που εξετάζει το δέντρο.

Ένα μοντέλο δέντρου απόφασης είναι ένα υπολογιστικό μοντέλο που υλοποιείται από τρία μέρη:

- Από το δέντρο απόφασης.
- Από τον αλγόριθμο, που θα δημιουργήσει το δέντρο.
- Από ένα αλγόριθμο, που εφαρμόζει το δέντρο στα δεδομένα και λύνει το υπό εξέταση πρόβλημα.

Η προσέγγιση του δέντρου απόφασης είναι από τις πιο διαδεδομένες μεθόδους επίλυσης προβλημάτων κατηγοριοποίησης. Με το δέντρο απόφασης μοντελοποιείται η ίδια η διαδικασία της κατηγοριοποίησης. Όταν χτιστεί το μοντέλο του δέντρου απόφασης, εφαρμόζεται σε κάθε πλειάδα των δεδομένων ο αλγόριθμος του μοντέλου για να αποφασίσει σε ποια κατηγορία ανήκει η κάθε πλειάδα. Η τεχνική αυτή προσπαθεί να αντιμετωπίσει δύο βήματα: την κατασκευή του δέντρου και την εφαρμογή του δέντρου στην Βάση Δεδομένων.

Το δέντρο απόφασης που κατασκευάζεται για την επίλυση του προβλήματος της κατηγοριοποίησης έχει τα παρακάτω χαρακτηριστικά:

Έστω ότι έχουμε ένα σύνολο δεδομένων, που έχουν συγκεκριμένα γνωρίσματα και ένα σύνολο κατηγοριών. Το δέντρο απόφασης είναι το δέντρο που σχετίζεται με τα δεδομένα και έχει τις παρακάτω ιδιότητες.

- Κάθε εσωτερικός κόμβος παίρνει το όνομα του από ένα γνώρισμα των δεδομένων.
- Κάθε τόξο παίρνει το όνομα του από το κατηγόρημα, το οποίο μπορεί να εφαρμοστεί στο γνώρισμα (εσωτερικός κόμβος) που συνδέεται με τον πατέρα κόμβο. Λέγοντας κατηγόρημα, εννοούμε τις λέξεις κλειδιά οι οποίες χαρακτηρίζουν τον κόμβο πατέρα σε σχέση με το γνώρισμα στο οποίο καταλήγει το τόξο.
- Κάθε φύλλο παίρνει το όνομα του από το όνομα της κάθε κατηγορίας.

Έτσι για την λύση του προβλήματος της κατηγοριοποίησης, η διαδικασία της χρήσης δέντρων απόφασης περιλαμβάνει τα παρακάτω βήματα:

1. Δημιουργία του δέντρου απόφασης, χρησιμοποιώντας τα δεδομένα εκπαίδευσης
2. Για κάθε πλειάδα της βάσης δεδομένων, εφαρμογή του δέντρου απόφασης για τον προσδιορισμό της κατηγορίας στην οποία ανήκει η πλειάδα

Υπάρχουν πολλά **πλεονεκτήματα** από την χρήση των Δέντρων Απόφασης για την εκτέλεση της λειτουργίας της Κατηγοριοποίησης. Τα Δέντρα Απόφασης είναι εύκολα στην χρήση και αποτελεσματικά. Με την χρήση τους, μπορούν να δημιουργηθούν κανόνες, οι οποίοι είναι εύκολοι στο να κατανοηθούν και να τους ερμηνεύσουν οι χρήστες. Επίσης, τα Δέντρα Απόφασης μπορούν να αποδώσουν καλύτερα σε μεγάλες Βάσεις Δεδομένων, επειδή το μέγεθος του δέντρου, που δημιουργείται δεν εξαρτάται από το μέγεθος της Βάσης Δεδομένων. Μπορούμε, τέλος, να κατασκευάσουμε δέντρα για δεδομένα που έχουν πολλά γνωρίσματα.

Υπάρχουν όμως και **μειονεκτήματα** από την χρήση των Δέντρων Απόφασης. Πρώτο πρόβλημα που δημιουργείται είναι ότι τα Δέντρα Απόφασης δεν μπορούν εύκολα να διαχειριστούν συνεχή δεδομένα. Έτσι, οι τιμές των γνωρισμάτων θα πρέπει να χωριστούν σε κατηγορίες για να μπορέσουν να τα διαχειριστούν τα Δέντρα Απόφασης. Επίσης, είναι δύσκολος ο χειρισμός ελλιπών δεδομένων, γιατί εάν λείπουν αρκετές τιμές σε κάποια γνωρίσματα δεν θα μπορούν να βρεθούν οι σωστές διακλαδώσεις του δέντρου. Ένα άλλο πρόβλημα δημιουργείται από το γεγονός ότι για την δημιουργία του δέντρου χρησιμοποιούνται δεδομένα εκπαίδευσης, με αποτέλεσμα να παρουσιαστεί το πρόβλημα της υπέρ-προσαρμογής, όπου τα αποτελέσματα του μοντέλου που παράγεται προσαρμόζονται σε μεγάλο βαθμό στα δεδομένα εκπαίδευσης (στο συγκεκριμένο πρόβλημα δηλαδή που πάνε να επιλύσουν) και ως εκ τούτου τυχόν νέα δεδομένα στο μοντέλο να παράγουν διαφορετικά ή λανθασμένα αποτελέσματα.

Ζητήματα σχετικά με τους αλγορίθμους Δέντρων Απόφασης

Οι κυριότεροι παράγοντες στην απόδοση ενός αλγορίθμου δημιουργίας Δέντρων Απόφασης είναι το μέγεθος του συνόλου εκπαίδευσης και η επιλογή των καλύτερων γνωρισμάτων διάσπασης, τα γνωρίσματα δηλαδή που καθορίζουν την διάσπαση των κόμβων του δέντρου και κατά συνέπεια το σχήμα του. Υπάρχουν επιπλέον και άλλα ζητήματα, τα οποία αντιμετωπίζονται από τους αλγορίθμους δημιουργίας Δέντρων Απόφασης. Τα ζητήματα αυτά παρουσιάζονται παρακάτω:

- **Επιλογή Γνωρισμάτων για την διάσπαση και Διάταξή τους**. Το ποια γνωρίσματα χρησιμοποιούνται ως γνωρίσματα επηρεάζουν σε μεγάλο βαθμό την απόδοση του Δέντρου Απόφασης, που δημιουργείται. Έτσι πολλές φορές, η επιλογή του γνωρίσματος πραγματοποιείται είτε από την εξέταση των συνόλου των δεδομένων εκπαίδευσης αλλά επίσης απαιτείται και η εμπεριστατωμένη γνώμη ειδικών του τομέα, για τον καθορισμό των γνωρισμάτων διάσπασης. Επίσης, η σειρά με την οποία επιλέγονται τα γνωρίσματα διάσπασης είναι σημαντική και επηρεάζει την μορφή του δέντρου αλλά και τον χρόνο, που απαιτείται για την δημιουργία του. Η διάταξη των γνωρισμάτων επηρεάζει επίσης και τις διασπάσεις, οι οποίες πραγματοποιούνται στο δέντρο. Σε μερικά γνωρίσματα, εάν το εύρος των τιμών τους είναι μικρό, ο αριθμός των διασπάσεων είναι προφανής βάσει των τιμών αυτών. Σε μερικές περιπτώσεις όμως, εάν το πεδίο είναι συνεχές ή έχει ένα μεγάλο εύρος τιμών, ο αριθμός των διασπάσεων που θα γίνουν δεν είναι εύκολο να καθοριστεί.
- **Δομή του Δέντρου**. Η απόδοση της λειτουργίας ενός Δέντρου Απόφασης για κατηγοριοποίηση εξαρτάται σε μεγάλο βαθμό από την δομή του Δέντρου. Σε γενικές γραμμές θέλουμε να υπάρχει ένα ισοζυγισμένο δέντρο με τα λιγότερα δυνατά επίπεδα (για να φτάσουμε σύντομα στο αποτέλεσμα της κατηγοριοποίησης). Όμως, στην περίπτωση αυτή μπορεί να χρειάζονται πιο πολύπλοκες συγκρίσεις και πολλαπλές διακλαδώσεις, για να παραχθεί ένα τέτοιο δέντρο. Επίσης, πολλοί αλγόριθμοι δημιουργούν μόνο δυαδικά δέντρα, με αποτέλεσμα τα δέντρα αυτά να είναι αρκετά βαθιά. Έτσι μπορεί να απαιτείται να γίνουν πολλές συγκρίσεις για να βρεθεί το τελικό αποτέλεσμα. Ωστόσο, επειδή αυτές οι συγκρίσεις μπορεί να είναι πιο απλές από την περίπτωση που έχουν ένα πιο χαμηλό δέντρο αλλά με πολύπλοκες διασταυρώσεις, το τελικό αποτέλεσμα της απόδοσης του δέντρου να είναι καλύτερη στα δυαδικά δέντρα. Σε κάθε περίπτωση όμως, η απόδοση εξαρτάται και από τον τύπο των δεδομένων, στα οποία θα γίνουν οι έλεγχοι.
- **Κριτήρια Τερματισμού**. Η δημιουργία του Δέντρου Απόφασης σταματά όταν τα δεδομένα εκπαίδευσης κατηγοριοποιούνται τέλεια, ανάλογα με το πρόβλημα. Υπάρχουν όμως περιπτώσεις που πρέπει να σταματήσει νωρίτερα η δημιουργία του δέντρου, έτσι ώστε να αποτραπεί η επέκταση του σε μεγάλο βάθος. Τα κριτήρια τερματισμού της χτισίματος του δέντρου αποτελεί ένα συμβιβασμό ανάμεσα στην ακρίβεια της κατηγοριοποίησης και την απόδοσης της λειτουργίας κατηγοριοποίησης του δέντρου. Επίσης, η διαδικασία χτισίματος του δέντρου πρέπει να σταματήσει για να αποφευχθεί η υπέρ-προσαρμογή του δέντρου στα δεδομένα εκπαίδευσης.
- **Δεδομένα Εκπαίδευσης και Πολυπλοκότητα Δέντρου**. Η δομή του Δέντρου Απόφασης εξαρτάται, όπως έχουμε τονίσει, από τα δεδομένα εκπαίδευσης. Εάν το σύνολο των δεδομένων εκπαίδευσης είναι πολύ μικρό, το δέντρο που θα παραχθεί ίσως να μην μπορεί να δουλέψει καλά με πιο γενικά δεδομένα. Στην αντίθετη περίπτωση, αν τα δεδομένα εκπαίδευσης είναι πάρα πολλά, τότε υπάρχει περίπτωση το δέντρο που θα δημιουργηθεί να παρουσιάσει υπέρ-προσαρμογή. Σε γενικές γραμμές, η πολυπλοκότητα σε χρόνο και σε χώρο των Δέντρων Απόφασης εξαρτάται από το μέγεθος των δεδομένων εκπαίδευσης (q), τον αριθμό των γνωρισμάτων (h) και το σχήμα του δέντρου που προκύπτει. Στην χειρότερη περίπτωση, το δέντρο απόφασης που θα χτιστεί μπορεί να είναι αρκετά βαθύ και όχι πολύ πυκνό. Έτσι καθώς χτίζεται το δέντρο για κάθε ένα από τους κόμβους, κάθε γνώρισμα θα εξετάζεται για να καθοριστεί αν είναι το καλύτερο, για να χρησιμοποιηθεί για την διάσπαση.
- **Κλάδεμα Δέντρου**. Όταν χτιστεί ένα Δέντρο Απόφασης, ίσως να είναι απαραίτητες μερικές τροποποιήσεις στο δέντρο για να βελτιώσουν την απόδοση του κατά την διάρκεια της κατηγοριοποίησης. Η φάση του κλαδέματος (pruning) μπορεί αφαιρέσει τμήματα του δέντρου έτσι ώστε να μειωθούν περιττές συγκρίσεις στο δέντρο (στην περίπτωση που υπάρχουν τμήματα του δέντρου που σχετίζονται με όχι και τόσο σημαντικά γνωρίσματα), να διαγράψει περιττά υπό-δέντρα ή να συνδυάσει τμήματα του δέντρου έτσι ώστε να μειωθεί το συνολικό μέγεθος του δέντρου για να επιτευχθεί καλύτερη απόδοση.

Ομαδοποίηση – Clustering/Segmentation

Η διαδικασία της ομαδοποίησης ενός συνόλου φυσικών ή αφηρημένων αντικειμένων σε κατηγορίες που περιέχουν παρόμοια αντικείμενα ονομάζεται **ομαδοποίηση (clustering)**. Μια ομάδα (cluster) είναι μια συλλογή από αντικείμενα που είναι παρόμοια μεταξύ τους (έχουν κοινά χαρακτηριστικά) και διαφέρουν από τα αντικείμενα που ανήκουν σε άλλες ομάδες. Μια ομάδα αντικειμένων μπορεί να αντιμετωπιστεί συλλογικά ως ένα σύνολο και έτσι μπορεί να θεωρηθεί ως μια μορφή συμπίεσης δεδομένων.

Η τεχνική της ομαδοποίησης (clustering) είναι παρόμοια με την τεχνική της κατηγοριοποίησης, όπως την παρουσιάσαμε παραπάνω, με την διαφορά όμως ότι οι ομάδες δεν είναι καθορισμένες από την αρχή αλλά ορίζονται μέσα από τα ίδια τα δεδομένα. Για το λόγο αυτό η τεχνική αναφέρεται και σαν ***μη εποπτευόμενη μάθηση***, γιατί οι ομάδες δημιουργούνται κατά την διάρκεια της λειτουργίας του αλγορίθμου. Η ομαδοποίηση επιτυγχάνεται με τον καθορισμό της ομοιότητας ανάμεσα στα δεδομένα, ως προς τα χαρακτηριστικά τους που εισάγονται στον αλγόριθμο. Παράδειγμα χρήσης της ομαδοποίησης σε ένα οργανισμό είναι η κατανόηση της πελατειακής βάσης του οργανισμού, με την τοποθέτηση των πελατών σε ομάδες, βάσει κάποιων συγκεκριμένων χαρακτηριστικών τους. Με την βοήθεια της ομαδοποίησης, ο οργανισμός μπορεί να αναγνωρίσει εν δυνάμει πελάτες βάσει των χαρακτηριστικών τους και να τους τοποθετήσει στις ομάδες, που έχει ανακαλύψει ο αλγόριθμος της ομαδοποίησης. Το αποτέλεσμα της ομαδοποίησης μπορεί επίσης να χρησιμοποιηθεί για να εντοπίσει τα δεδομένα, τα οποία δεν ανήκουν σε κάποια ομάδα (ή είναι αρκετά «μακριά» από τις ομάδες που έχουν βρεθεί). Τα δεδομένα αυτά ανήκουν συνήθως στα όρια των δεδομένων και πολλές φορές, ανάλογα με τον σκοπό της ανάλυσης, τα δεδομένα αυτά μπορεί να είναι περισσότερο σημαντικά από τις απλές περιπτώσεις. Παραδείγματα χρήσης του εντοπισμού των ακραίων τιμών είναι ο έλεγχος απάτης από ασυνήθιστες συναλλαγές πιστωτικών καρτών (για παράδειγμα μια μη συνηθισμένη αγορά πολύ μεγάλης αξίας) ή η αντιμετώπιση παράνομων πράξεων σε εταιρείας ηλεκτρονικού εμπορίου (για παράδειγμα η καταχώρηση ενός πολύ μεγάλου αριθμού παραγγελιών από τον ίδιο λογαριασμό).

Για την καλύτερη κατανόηση των αποτελεσμάτων της ομαδοποίησης, οι ομάδες που παράγονται λαμβάνουν κάποιο χαρακτηρισμό, για να μπορεί αυτός που διαβάζει τα αποτελέσματα να καταλάβει τι περιλαμβάνει η κάθε ομάδα. Ο χαρακτηρισμός αυτός που μπαίνει σε κάθε ομάδα είναι σημαντικός, αφού δεν γνωρίζουμε εκ των προτέρων τις ομάδες αλλά αυτές δημιουργούνται στην διάρκεια εκτέλεσης του αλγορίθμου ομαδοποίησης. Την υποχρέωση χαρακτηρισμού των ομάδων την έχει κάποιος, ο οποίος θα είναι καλός γνώστης του προβλήματος – συνήθως το άτομο το οποίο θα «τρέξει» τον αλγόριθμο – και μέσα από την ανάλυση των ομάδων και τα χαρακτηριστικά που έχουν τα μέλη κάθε ομάδας να μπορεί να δώσει το όνομα της κάθε ομάδας.

Σαν μια διαδικασία Εξόρυξης Γνώσης, η ανάλυση ομάδων (cluster analysis) μπορεί να χρησιμοποιηθεί σαν μια αυτόνομη εφαρμογή για να προσφέρει γνώση σχετικά με το σύνολο των δεδομένων και πως αυτά σχετίζονται μεταξύ τους, για να παρουσιάσει τα χαρακτηριστικά της κάθε ομάδας ή να βοηθήσει τους αναλυτές να επικεντρωθούν σε μια συγκεκριμένη ομάδα δεδομένων για περαιτέρω ανάλυση. Εναλλακτικά, τα αποτελέσματα της ομαδοποίησης μπορούν να χρησιμοποιηθούν σαν ένα πρωταρχικό βήμα επεξεργασίας των δεδομένων, για χρήση τους από άλλες τεχνικές Εξόρυξης Γνώσης, όπως για παράδειγμα η επιλογή χαρακτηριστικών για την χρήση τους στον αλγόριθμο κατηγοριοποίησης.

Τα εργαλεία που παρέχουν τις τεχνικές Εξόρυξης Γνώσης, όπως για παράδειγμα το περιβάλλον Business Intelligence Development Studio του SQL Server, δίνει στους χρήστες την δυνατότητα οπτικοποίησης των ομάδων (clusters) και κάποια στατιστικά στοιχεία σχετικά με τα μέλη της κάθε ομάδας, έτσι ώστε ο χρήστης να κατανοήσει τα αποτελέσματα και να χαρακτηρίσει σωστά την κάθε ομάδα.

Προβλήματα που προκύπτουν

Η εφαρμογή της τεχνικής της ομαδοποίησης σε πραγματικά δεδομένα σε μία Βάση Δεδομένων, πρέπει να αντιμετωπίσει τα παρακάτω προβλήματα:

- Ο χειρισμός των ακραίων σημείων είναι αρκετά δύσκολος. Τα ακραία σημεία είναι ορισμένες τιμές, που πολλές φορές δεν μπορούν να τοποθετηθούν σε κάποια συγκεκριμένη ομάδα. Οι ακραίες τιμές προκύπτουν, τις περισσότερες φορές από λανθασμένη καταχώρηση δεδομένων στην βάση, με αποτέλεσμα να υπάρχουν κάποιες τιμές οι οποίες «οπτικά» δεν ανήκουν σε κάποια από τις ομάδες που ανακαλύπτονται. Κάποιοι αλγόριθμοι ομαδοποίησης θεωρούν τις ακραίες τιμές σαν ξεχωριστές ομάδες ενώ άλλοι αλγόριθμοι μεγαλώνουν τις ομάδες που ανακαλύπτουν, ώστε να περιλαμβάνουν κάποιες από αυτές τις ακραίες τιμές που υπάρχουν.
- Από την φύση των αλγορίθμων ομαδοποίησης και από τα δυναμικά δεδομένα που υπάρχουν σε μία Βάση δεδομένων, η σύσταση των ομάδων που δημιουργούνται, μπορεί να αλλάξει στην πορεία του χρόνου.
- Όπως τονίσαμε παραπάνω, ο χαρακτηρισμός που ορίζεται σε μία ομάδα εξαρτάται από τον εκάστοτε αναλυτή των ομάδων. Σε αντίθεση με την κατηγοριοποίηση, που οι ομάδες είναι προκαθορισμένες, στην ομαδοποίηση οι ομάδες είναι δυναμικές και για το λόγο αυτό πρέπει να οριστούν κάποιες ετικέτες σε κάθε ομάδα, έτσι ώστε ο αναγνώστης των ομάδων να μπορεί να αντιληφθεί την σημασία της κάθε ομάδας.
- Σε συνέχεια με το παραπάνω ζήτημα, σε ένα πρόβλημα ομαδοποίησης δεν υπάρχει μια μόνο σωστή λύση αλλά μπορούν να βρεθούν πολλές απαντήσεις. Ο ακριβής αριθμός των ομάδων δεν είναι προκαθορισμένος και δεν είναι και τόσο εύκολο να βρεθεί. Για το λόγο αυτό, την ανάλυση των αποτελεσμάτων του αλγορίθμου ενδεχομένως να πρέπει να την αναλάβει κάποιος ειδικός, ενώ πολλές εφαρμογές που προσφέρουν τεχνικές Εξόρυξης Γνώσης επιτρέπουν στους χρήστες να αλλάζουν διάφορες παραμέτρους των αλγορίθμων και να αναλύουν τα αποτελέσματα των ομάδων. Στην περίπτωση που οι ομάδες σε κάθε εκτέλεση του αλγορίθμου δεν διαφέρουν πολύ μεταξύ τους, συμπεραίνουμε ότι οι ομάδες που προκύπτουν είναι αυτές που αναλύουν πιο σωστά τα δεδομένα.
- Ένα άλλο πρόβλημα που πρέπει να αντιμετωπιστεί σε μια τεχνική ομαδοποίησης είναι τι δεδομένα θα χρησιμοποιηθούν. Σε αντίθεση με την κατηγοριοποίηση που οι ομάδες είναι γνωστές και ξέρουμε σε γενικές γραμμές τα γνωρίσματα που θα εισαχθούν στον αλγόριθμο για την τοποθέτηση των δεδομένων στις ομάδες, στην ομαδοποίηση δεν γνωρίζουμε ποια γνωρίσματα χαρακτηρίζουν την κάθε ομάδα. Για το λόγο αυτό, η τεχνική της ομαδοποίησης θεωρείται μη επιβλεπόμενη μάθηση. Σε πολλές εφαρμογές Εξόρυξης Γνώσης εισάγονται διάφορα γνωρίσματα των δεδομένων και αναλύονται τα τελικά αποτελέσματα. Έτσι, είναι και στην περίπτωση αυτή αναγκαία η συμβολή κάποιου ειδικού που θα αναλύσει τα αποτελέσματα των ομάδων και θα καθορίσει την ορθότητα των ομάδων που θα ανακαλυφθούν.

Απαιτήσεις από τις τεχνικές Ομαδοποίησης

Οι τυπικές απαιτήσεις που θα πρέπει να διαθέτει μια τεχνική ομαδοποίησης για την χρήση της σε μια διαδικασία Εξόρυξης Γνώσης παρουσιάζονται παρακάτω:

- **Επεκτασιμότητα**: Πολλοί αλγόριθμοι ομαδοποίησης δουλεύουν καλά με ένα μικρό σύνολο δεδομένων (μερικές εκατοντάδες), ωστόσο μια βάση δεδομένων μπορεί να περιέχει πολλαπλάσια δεδομένα. Η ομαδοποίηση σε ένα δείγμα των δεδομένων μπορεί να οδηγήσει σε «προκατειλημμένα» αποτελέσματα (που εξαρτώνται πολύ από το ίδιο το σύνολο δεδομένων που χρησιμοποιείται στον αλγόριθμο ομαδοποίησης). Για το λόγο αυτό απαιτείται η χρήση αρκετά επεκτάσιμων αλγορίθμων, που να μπορούν να διαχειρίζονται ένα μεγάλο σε μέγεθος σύνολο δεδομένων.
- **Δυνατότητα αντιμετώπισης διαφορετικών τύπων χαρακτηριστικών**: Πολλοί αλγόριθμοι ομαδοποίησης έχουν σχεδιαστεί για να δουλεύουν με συγκεκριμένους τύπους δεδομένων, όπως για παράδειγμα αριθμητικά δεδομένα. Ωστόσο, υπάρχει η

ανάγκη χρήσης αλγορίθμων ομαδοποίησης που να λειτουργούν με διάφορους τύπους δεδομένων, όπως δυαδικές τιμές, κατηγορικές τιμές, συνεχείς τιμές ή τον συνδυασμό κάποιων από αυτούς τους τύπους δεδομένων.

- **Ανακάλυψη ομάδων με αυθαίρετο σχήμα**: Πολλοί αλγόριθμοι ομαδοποίησης κατασκευάζουν τις ομάδες βάσει συγκεκριμένων μετρικών, όπως η Ευκλείδεια απόσταση των σημείων από το κέντρων της κάθε ομάδας. Οι αλγόριθμοι που βασίζονται σε τέτοιου είδους μετρικές τείνουν να κατασκευάζουν ομάδες με παρόμοιο σχήμα (πχ σφαιρικό σχήμα) και πυκνότητα. Ωστόσο, μια ομάδα πρέπει να είναι οποιουδήποτε σχήματος. Είναι σημαντικό να υπάρχουν αλγόριθμοι που να μπορούν να εντοπίζουν ομάδες με αυθαίρετο σχήμα.
- **Δυνατότητα αντιμετώπισης «δεδομένων με θόρυβο»**: πολλές βάσεις δεδομένων περιέχουν δεδομένα με ελλείπεις, λανθασμένες ή άγνωστες τιμές. Αυτά τα δεδομένα θεωρούνται ότι περιέχουν θόρυβο, επηρεάζοντας το τελικό αποτέλεσμα της ομαδοποίησης. Μερικοί αλγόριθμοι είναι αρκετά ευαίσθητοι σε τέτοιου είδους δεδομένων με αποτέλεσμα να παράγουν ομάδες με φτωχή ποιότητα.
- **Δυνατότητα αυξητικής διαμόρφωσης των ομάδων και ευαισθησία στην σειρά εισόδου των δεδομένων**: Μερικοί αλγόριθμοι ομαδοποίησης δεν μπορούν να διαχειριστούν ανανεωμένα δεδομένα (για παράδειγμα ενημερώσεις των δεδομένων της Βάσης) σε υπάρχοντες ομάδες και, αντ' αυτού πρέπει να προσδιορίσουν μια καινούργια ομαδοποίηση από την αρχή. Επίσης, κάποιοι αλγόριθμοι είναι ευαίσθητοι στην σειρά, με την οποία εισάγονται τα δεδομένα στο μοντέλο ομαδοποίησης. Αυτό έχει σαν αποτέλεσμα να παράγονται εντελώς διαφορετικές ομάδες, ανάλογα με την σειρά εισόδου των χαρακτηριστικών εισόδου στους αλγορίθμους. Είναι σημαντικό να αναπτύσσονται αυξητικοί αλγόριθμοι (που να διαμορφώνουν αυξητικά τις ομάδες καθώς ανανεώνονται τα δεδομένα στην Βάση) καθώς και αλγόριθμοι που δεν είναι ευαίσθητοι στην σειρά χρήσης των χαρακτηριστικών εισόδου.
- **Δυνατότητα ερμηνείας και χρησιμότητα**: Οι χρήστες των αποτελεσμάτων της ομαδοποίησης περιμένουν οι ομάδες να είναι αναγνώσιμες, κατανοητές και χρήσιμες. Αυτό απαιτείται, γιατί η ομαδοποίηση μπορεί να πρέπει να συνδεθεί με μια συγκεκριμένη ερμηνεία των αποτελεσμάτων της καθώς και με άλλες εφαρμογές.

Κατηγοριοποίηση των βασικότερων τεχνικών Ομαδοποίησης

Υπάρχουν πολλοί αλγόριθμοι ομαδοποίησης στην βιβλιογραφία. Είναι δύσκολο να παρουσιαστεί μια προκαθορισμένη κατηγοριοποίηση των τεχνικών ομαδοποίηση γιατί οι κατηγορίες αυτές μπορεί να επικαλύπτουν η μία την άλλη, έτσι ώστε ο κάθε αλγόριθμος να περιέχει χαρακτηριστικά από διάφορες κατηγορίες. Παρακάτω παρουσιάζουμε μια προσπάθεια κατηγοριοποίησης των αλγορίθμων ομαδοποίηση προσπαθώντας να δώσουμε μια γενική εικόνα των αλγορίθμων. Σε γενικές γραμμές, οι τεχνικές ομαδοποίησης μπορούν τοποθετηθούν στις παρακάτω κατηγορίες:

- **Μέθοδοι Διαμερισμού του χώρου (Partitioning methods)**: Δεδομένης μιας βάσης δεδομένων με n στοιχεία (πλειάδες), μια μέθοδος διαμερισμού του χώρου κατασκευάζει k διαμερίσματα δεδομένων, όπου κάθε διαμέρισμα αποτελεί μία ομάδα (cluster) και ισχύει k <= n. Δηλαδή, κατατάσσει τα δεδομένα σε k ομάδες, κάθε μια από τις οποίες ικανοποιεί τις ακόλουθες απαιτήσεις: (1) κάθε ομάδα πρέπει να περιέχει τουλάχιστον ένα αντικείμενο, και (2) κάθε αντικείμενο πρέπει να ανήκει σε ακριβώς μια ομάδα.

  Δεδομένου του αριθμού των διαμερισμάτων για την κατασκευή k, μια μέθοδος διαμερισμού του χώρου δημιουργεί μια αρχική ομάδα. Στη συνέχεια, χρησιμοποιεί μια επαναληπτική τεχνική, η οποία προσπαθεί να βελτιώσει τον διαμερισμό με την μετακίνηση των αντικειμένων από τη μία ομάδα στην άλλη. Το γενικό κριτήριο ενός καλού διαχωρισμού είναι ότι τα αντικείμενα στην ίδια ομάδα είναι «κοντινά» ή σχετίζονται μεταξύ τους, ενώ τα αντικείμενα των διαφορετικών ομάδων είναι «πολύ μακριά» ή πολύ διαφορετικά. Υπάρχουν διάφορα είδη κριτηρίων για την αξιολόγηση της ποιότητας των διαχωρισμών στις ομάδες.

Για την επίτευξη μιας γενικής βελτιστοποίησης των τεχνικών διαμερισμού, θα έπρεπε να απαριθμήσουμε όλες τις πιθανές ομάδες που μπορούν να παραχθούν. Αντί όμως για αυτή την επίπονη διαδικασία, οι περισσότερες τεχνικές χρησιμοποιούν μερικές δημοφιλείς ευρεστικές μεθόδους, όπως για παράδειγμα, (1) ο αλγόριθμος k-means, όπου η κάθε ομάδα αντιπροσωπεύεται από τη μέση τιμή των αντικειμένων στην ομάδα αυτή, και (2) ο αλγόριθμος k-medoids, όπου κάθε ομάδα αντιπροσωπεύεται από ένα από τα αντικείμενα που βρίσκονται κοντά στο κέντρο της ομάδας. Αυτές οι ευρεστικές μέθοδοι ομαδοποίησης λειτουργεί καλά για την εύρεση ομάδων με σφαιρικό σχήμα, σε μικρές και μεσαίες βάσεις δεδομένων.

- **Ιεραρχικές Μέθοδοι (Hierarchical methods)**: Μια ιεραρχική μέθοδος δημιουργεί μια ιεραρχική αποσύνθεση του διαθέσιμου συνόλου των δεδομένων. Η μέθοδος μπορεί να χαρακτηριστεί είτε ως μέθοδος συσσώρευσης δεδομένων (από κάτω προς τα πάνω) είτε ως μέθοδος διάσπασης (από πάνω προς τα κάτω) με βάση τον τρόπο με τον οποίο σχηματίζεται η ιεραρχική διάσπαση των δεδομένων.

  Η προσέγγιση της συσσώρευσης των δεδομένων, ξεκινά με το κάθε σημείο δεδομένων που αποτελεί μια ξεχωριστή ομάδα. Στην συνέχεια συγχωνεύει το κάθε σημείο ή τις ομάδες, που είναι κοντά το ένα στο άλλο, έως ότου όλες οι ομάδες συγχωνευθούν σε μια ομάδα (το ανώτερο επίπεδο της ιεραρχίας) ή μέχρι να ισχύσει ένα κριτήριο τερματισμού. Η προσέγγιση της διάσπασης, ξεκινά με όλα τα σημεία δεδομένων στην ίδια ομάδα. Σε κάθε διαδοχική επανάληψη, μια ομάδα χωρίζεται σε μικρότερες ομάδες, μέχρι που τελικά το κάθε σημείο δεδομένων ανήκει σε μια συστάδα, ή έως ότου ισχύσει ένα κριτήριο τερματισμού.

  Οι ιεραρχικές μέθοδοι υποφέρουν από το γεγονός ότι μόλις ένα βήμα (συγχώνευση ή διάσπαση) πραγματοποιηθεί, δεν μπορεί ποτέ να αναιρεθεί. Αυτή η έλλειψη ευελιξίας είναι χρήσιμη στο βαθμό που οδηγεί σε μικρότερα έξοδα υπολογισμού, γιατί ο αλγόριθμος δεν χρειάζεται να ανησυχεί για τον συνδυαστικό αριθμό διαφορετικών επιλογών, που μπορεί να έχει. Ωστόσο, τέτοιες τεχνικές δεν μπορεί να διορθώσουν λανθασμένες αποφάσεις. Υπάρχουν δύο προσεγγίσεις για τη βελτίωση της ποιότητας της ιεραρχικής ομαδοποίησης: (1) εκτελείται προσεκτική ανάλυση της συσχέτισης των δεδομένων της κάθε ομάδας σε κάθε ιεραρχική κατάτμηση, ή (2) ενσωματώνεται μια τεχνική συσσώρευσης, με την δημιουργία αρχικά πολύ μικρών ομάδων (micro clusters) και στην συνέχεια χρησιμοποιείται μια επαναληπτική τεχνική ομαδοποίησης η οποία εφαρμόζει την ομαδοποίηση στις μικρές αυτές ομάδες για να δημιουργήσει μεγαλύτερες αλλάζοντας θέση, όπου απαιτείται στις μικρές αυτές ομάδες.

- **Μέθοδοι βασισμένοι στην Πυκνότητα (Density-based methods)**: Οι περισσότερες μέθοδοι διαμερισμού χωρίζουν τα δεδομένα σε ομάδες βάσει της απόστασης μεταξύ των δεδομένων. Τέτοιες μέθοδοι μπορούν να βρουν μόνο ομάδες με σφαιρικό σχήμα και συναντούν δυσκολία στην ανακάλυψη ομάδων με αυθαίρετα σχήματα. Άλλες μέθοδοι ομαδοποίησης έχουν αναπτυχθεί με βάση την έννοια της πυκνότητας. Η γενική ιδέα είναι να συνεχίσει να αυξάνεται η συγκεκριμένη ομάδα δεδομένων, εφόσον η πυκνότητα (αριθμός των αντικειμένων ή σημείων δεδομένων στην ομάδα) στη «γειτονιά» ξεπεράσει κάποιο όριο. Το όριο αυτό είναι, για κάθε σημείο δεδομένων μέσα σε μια συγκεκριμένη ομάδα, η γειτονιά μιας δεδομένης ακτίνας πρέπει να περιέχει έναν ελάχιστο αριθμό σημείων. Μια τέτοια μέθοδος μπορεί να χρησιμοποιηθεί για το φιλτράρισμα του θορύβου (ακραίες τιμές), καθώς και να ανακαλύψει ομάδες αυθαίρετου σχήματος.
- **Μέθοδοι βασισμένοι σε πλέγμα (Grid-based methods)**: Οι μέθοδοι που είναι βασισμένοι στο πλέγμα διαχωρίζουν το σύνολο των αντικειμένων σε έναν πεπερασμένο αριθμό κελιών που σχηματίζουν μια δομή πλέγματος. Όλες οι εργασίες ομαδοποίησης εκτελούνται πάνω στην δομή του πλέγματος (δηλαδή, στον διαχωρισμό του συνόλου δεδομένων). Το βασικό πλεονέκτημα αυτής της προσέγγισης είναι η γρήγορη επεξεργασία των δεδομένων, η οποία είναι συνήθως ανεξάρτητη από τον αριθμό των δεδομένων και εξαρτάται μόνο από τον αριθμό των κελιών σε κάθε διάσταση στον διαχωρισμένο χώρο των δεδομένων.

- **Μέθοδοι βασισμένοι σε Μοντέλα (Model-based methods)**: Οι μέθοδοι που βασίζονται σε μοντέλα υποθέτουν ένα μοντέλο για κάθε μία από τις ομάδες και βρίσκουν τη καλύτερη εφαρμογή των δεδομένων στο συγκεκριμένο μοντέλο. Ένα μοντέλο, στο οποίο που βασίζεται ο αλγόριθμος μπορεί να εντοπίσει ομάδες με την κατασκευή μιας συνάρτηση πυκνότητας, η οποία αντικατοπτρίζει την χωρική κατανομή των σημείων δεδομένων. Το μοντέλο οδηγεί επίσης σε έναν τρόπο για τον αυτόματο προσδιορισμό του αριθμού των ομάδων που βασίζονται σε στατιστικές μεθόδους, λαμβάνοντας υπόψη τον «θόρυβο» ή τις ακραίες τιμές, παράγοντας έτσι μια ισχυρή ομαδοποίηση. Ο ΕΜ είναι ένας αλγόριθμος που εκτελεί ανάλυση των ομάδων με βάση στατιστικά μοντέλα, προσδοκώντας την βέλτιστη λύση (χρησιμοποιεί πιθανότητες για να καθορίσει το πόσο «καλά» ανήκει ένα σημείο δεδομένων σε μια ομάδα).

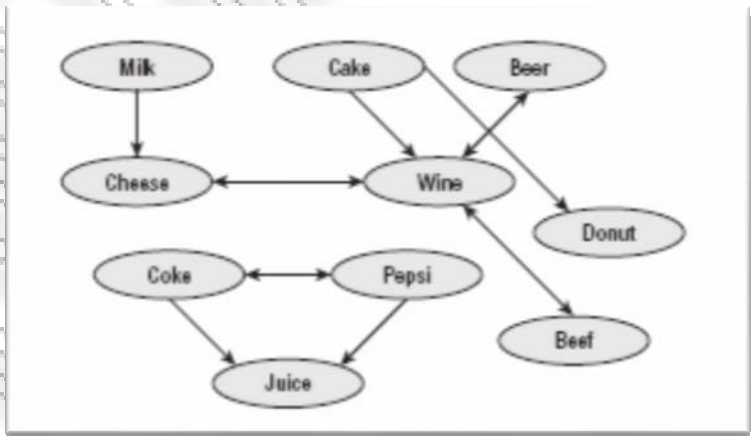
Η επιλογή της μεθόδου ομαδοποίησης βασίζεται στον τύπο των δεδομένων που είναι διαθέσιμα και στον συγκεκριμένη σκοπό της εφαρμογής, που θα χρησιμοποιήσει τον αλγόριθμο. Αν η ανάλυση της ομαδοποίησης χρησιμοποιηθεί σαν ένα εργαλείο για την εκμάθηση των δεδομένων, είναι καλύτερο να χρησιμοποιηθούν διάφοροι αλγόριθμοι στα ίδια δεδομένα για να δει ο χρήστης τι αποτελέσματα τα αποκαλύψει.

Συσχετίσεις – Associations

Οι κανόνες συσχέτισης είναι μια από τις πιο σημαντικές τεχνικές Εξόρυξης Γνώσης και είναι ίσως η πιο κοινή μορφή ανακάλυψης προτύπων σε συστήματα μη εποπτευόμενης μάθησης. Η μορφή Εξόρυξης Γνώσης των κανόνων συσχέτισης είναι αυτή που μοιάζει περισσότερο στην διαδικασία που έχουν πολλοί άνθρωποι στο νου τους όταν προσπαθούν να κατανοήσουν την λογική της ίδια της διαδικασίας της Εξόρυξης Γνώσης, η οποία είναι η εξόρυξη «χρυσού» (γνώσης) μέσα από μια τεράστια Βάση Δεδομένων. Ο χρυσός στην προκειμένη περίπτωση των κανόνων συσχέτισης είναι οι κανόνες που εξάγονται και παρουσιάζουν κάποιο ενδιαφέρον και προσφέρουν μια πληροφορία για τα δεδομένα τα οποία πιθανόν να μην τα γνωρίζουν οι χρήστες και να μην μπορούσαν να τα καθορίσουν ρητά. Οι τεχνικές των κανόνων συσχέτισης ανακτούν όλα τα ενδιαφέρονται συχνά πρότυπα που υπάρχουν στο σύνολο δεδομένων.

Τα συχνά πρότυπα είναι πρότυπα (patterns) που εμφανίζονται πολύ συχνά στα δεδομένα. Συνήθως τα πρότυπα, που ονομάζονται και σύνολα αντικειμένων (itemsets) είναι ομάδες αντικειμένων (προϊόντων) που εμφανίζονται μαζί σε μια συναλλαγή, όπως για παράδειγμα σε μια αγορά προϊόντων (πχ γάλα και ψωμί). Η εξόρυξη γνώσης μέσα από τα συχνά πρότυπα μπορεί να ανακαλύψει ενδιαφέροντες στατιστικές συσχετίσεις μεταξύ των αντικειμένων, που αναλύονται. Για παράδειγμα, η αγορά ενός προϊόντος όταν αγοράζεται μαζί με ένα άλλο προϊόν, αντιπροσωπεύει ένα κανόνα συσχέτισης. Η ανακάλυψη των προτύπων και των συσχετίσεων ανάμεσα σε ένα πολύ μεγάλο όγκο δεδομένων είναι χρήσιμη στις λειτουργίες μάρκετινγκ, στην λήψη αποφάσεων, στην διαφήμιση και στην διοίκηση επιχειρήσεων κυρίως σε καταστήματα λιανικής πώλησης. Μια δημοφιλής περιοχή εφαρμογής των κανόνων συσχέτισης είναι η **ανάλυση του καλαθιού αγοράς** (**market basket analysis**) η οποία μελετά τις αγοραστικές συνήθειες των πελατών μέσα από το σύνολο των συναλλαγών πωλήσεων.

Ο συνεχώς αυξανόμενος όγκος δεδομένων που συγκεντρώνεται από τις συναλλαγές καθημερινά και οι πληροφορίες που αποθηκεύονται καθιστούν αναγκαία την ύπαρξη ενός μηχανισμού ανακάλυψης προτύπων και συσχετίσεων μεταξύ των προϊόντων που πωλούνται. Η λειτουργία της ανάλυσης είναι να ψάχνει για ομάδες προϊόντων τα οποία αγοράζονται από κοινού ή στην σειρά από τους πελάτες. Η τυπική χρήση των κανόνων συσχέτισης που ανακαλύπτονται είναι η ανακάλυψη ομάδων προϊόντων που αγοράζονται μαζί και η χρήση κανόνων για μετέπειτα πωλήσεις πολλαπλών προϊόντων (cross-selling), για επανασχεδιασμό των καταλόγων προϊόντων (product catalogues) καθώς και την ανάλυση της αγοραστικής συμπεριφοράς των καταναλωτών για σχεδιασμό των καμπανιών μάρκετινγκ.



Διάγραμμα 11: Συσχετίσεις μεταξύ προϊόντων

Συχνά οι κανόνες συσχέτισης για να έχουν νόημα πρέπει να ικανοποιούν ένα ελάχιστο όριο ***υποστήριξης*** (support) και ένα ελάχιστο όριο ***εμπιστοσύνης*** (confidence).

Η **υποστήριξη** ενός στοιχείου (ή μιας ομάδας στοιχείων) είναι το ποσοστό των συναλλαγών στις οποίες εμφανίζεται το στοιχείο αυτό (ή η ομάδα) στο σύνολο των συναλλαγών.

Για παράδειγμα για τον κανόνα

A => B (αν κάποιος αγοράσει το προϊόν Α μπορεί να αγοράσει και το προϊόν Β)

η υποστήριξη είναι το ποσοστό των περιπτώσεων που εμφανίζονται σε μια συναλλαγή το προϊόν Α και το προϊόν Β ως προς το σύνολο των συναλλαγών.

Η **εμπιστοσύνη** αναφέρεται στον κανόνα και είναι το ποσοστό των συναλλαγών που εμφανίζεται το προϊόν Β και το προϊόν Α ως προς τον αριθμό των συναλλαγών που εμφανίζεται το προϊόν Α. Η εμπιστοσύνη δηλαδή αναφέρεται στην ίδια την φύση του κανόνα και δείχνει την ισχύ του (πόσες φορές δηλαδή εμφανίζεται ο κανόνας που αναφέρεται) ενώ η υποστήριξη μετρά πόσες φορές εμφανίζεται ο κανόνας στο σύνολο των δεδομένων. Θέλουμε η υποστήριξη να είναι σχετικά μεγάλος αριθμός αλλά όταν ο αριθμός αυτός πλησιάζει το 1 (το 100% δηλαδή του συνόλου των συναλλαγών) τότε ο κανόνας δεν είναι χρήσιμος αφού μας δείχνει κάτι που ισχύει πάντα! Έτσι για να δεχτούμε κάποιον κανόνα θέτουμε αρχικά κάποια ελάχιστα όρια εμπιστοσύνης και υποστήριξης που πρέπει να πληροί και στην συνέχεια μελετάμε περισσότερο τους κανόνες με υψηλές τιμές στα ποσοστά αυτά.

3. Επιχειρησιακά Προβλήματα – Επίλυση με τεχνικές Εξόρυξης Γνώσης

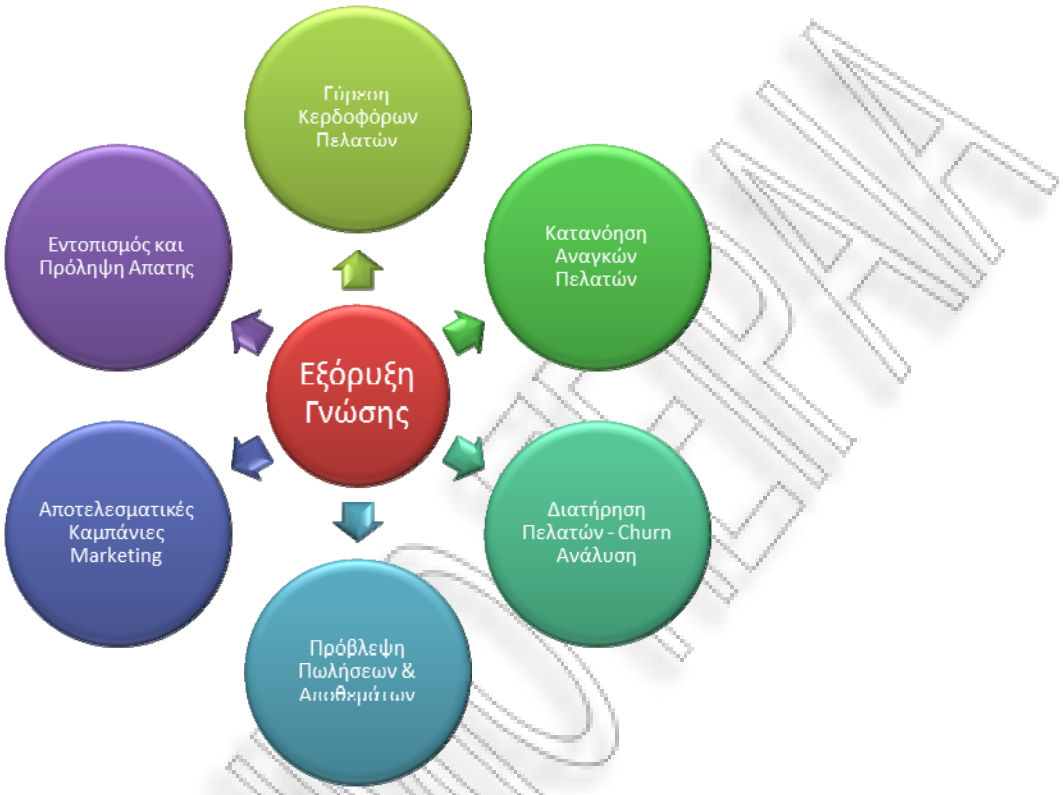
Οι τεχνικές Εξόρυξης Γνώσης μπορούν να χρησιμοποιηθούν σχεδόν σε όλες τις επιχειρηματικές εφαρμογές, απαντώντας σε διάφορους τύπους επιχειρηματικών ερωτήσεων. Δεδομένης της πληθώρας των εφαρμογών που είναι διαθέσιμα σήμερα, το μόνο που απαιτείται είναι το κίνητρο και η τεχνογνωσία για να απαντηθούν πολλά προβλήματα που απασχολούν τις επιχειρήσεις. Παρακάτω αναφέρουμε μερικά παραδείγματα επιχειρηματικών προβλημάτων που αντιμετωπίζονται με τις τεχνικές Εξόρυξης Γνώσης.

Εύρεση Κερδοφόρων Πελατών

Για την εύρεση των κερδοφόρων πελατών σε ένα οργανισμό προσφέρονται αρκετές τεχνικές Εξόρυξης Γνώσης για βοηθήσουν τον σκοπό αυτό. Το πρώτο βήμα που πρέπει να γίνει είναι η ομαδοποίηση των πελατών σε συγκεκριμένες ομάδες, που έχουν κάποια καθορισμένα χαρακτηριστικά. Η τεχνική που το προσφέρει αυτό είναι η Ομαδοποίηση (Clustering), όπου βάσει των χαρακτηριστικών των πελατών δημιουργούνται διάφορες ομάδες πελατών, που δεν είναι καθορισμένες από την αρχή. Η ομαδοποίηση των πελατών επιτρέπει των διαχωρισμό τους, μέσα από το μεγάλο σύνολο που έχει ο οργανισμός και την τοποθέτηση τους σε ομάδες. Οι τεχνικές της ομαδοποίησης είναι σημαντικές γιατί βοηθούν στην ανακάλυψη ομάδων, που με «γυμνό μάτι», βλέποντας δηλαδή μόνο τα δεδομένα, δεν μπορεί κάποιος να τις αντιληφθεί επιτρέποντας έτσι την ομαδοποίηση πελατών βάσει των χαρακτηριστικών τους, ιδίως σε περιπτώσεις οργανισμών με πολλές χιλιάδες πελάτες.

Στην συνέχεια με την χρήση των τεχνικών Κατηγοριοποίησης, ο οργανισμός είναι σε θέση να κατανοήσει την σχέση μεταξύ του κέρδους και των χαρακτηριστικών των πελατών. Μια τεχνική κατηγοριοποίησης είναι η χρήση Δέντρων Απόφασης (Decision Trees), όπου μπορούν να τοποθετηθούν οι πελάτες σε προκαθορισμένες ομάδες, όπως για παράδειγμα σε Επισφαλείς Πελάτες, σε πελάτες Υψηλού Ρίσκου ή σε πελάτες Υψηλής Κερδοφορία. Η διαχείριση κινδύνου (risk management) είναι μια τεχνική που χρησιμοποιείται κυρίως από τραπεζικούς οργανισμούς, όπου με την χρήση τεχνικών κατηγοριοποίησης είναι σε θέση να εντάξουν τους πελάτες τους, βάσει συγκεκριμένων χαρακτηριστικών τους σε προκαθορισμένες ομάδες, ξεχωρίζοντας έτσι τους πελάτες που έχουν υψηλό κίνδυνο, ιδίως στη περίπτωση εξόφλησης οφειλών, πληρωμών δόσεων κλπ. Τέλος με την εύρεση και την χρήση Κανόνων Συσχέτισης (Association Rules) μπορεί ο οργανισμός να κατανοήσει τις προτιμήσεις των πελατών και να συνδυάσει τις διάφορες αγοραστικές συνήθειες τους.

Έχοντας ο οργανισμός τα διάφορα μοντέλα εξόρυξης Γνώσης που έχουν δημιουργηθεί, είναι σε θέση να ξεχωρίσει τους νέους πελάτες και να προβλέψει τους εν δυνάμει κερδοφόρους πελάτες, που έρχονται στον οργανισμό βάσει των χαρακτηριστικών τους.



Διάγραμμα 12: Εφαρμογές Εξόρυξης Γνώσης

Κατανόηση Αναγκών Πελατών

Η κατανόηση των αναγκών των πελατών είναι ένα από τα πιο σημαντικά προβλήματα που προσπαθούν να επιλύσουν οι επιχειρήσεις. Η κατανόηση των αναγκών του πελάτη σε μια επιχείρηση μπορεί να οδηγήσει σε δημιουργία συστάσεων αγορών (διαφημιστικές εκστρατείες, προσφορές κλπ) προς τον κάθε πελάτη. Οι πελάτες που τους παρέχονται σωστές και έγκαιρες προτάσεις αγοράς προϊόντων είναι περισσότερο επικερδείς (γιατί θα αγοράσουν περισσότερα προϊόντα) και θα είναι περισσότερο «πιστοί» πελάτες (γιατί νοιώθουν μια πιο δυνατή σχέση με τον προμηθευτή τους). Για παράδειγμα, σε πολλά ηλεκτρονικά καταστήματα, όταν ένας πελάτης αγοράζει ένα προϊόν, λαμβάνει και κάποιες προτάσεις αγοράς για άλλα προϊόντα, που ενδεχομένως να τον ενδιαφέρουν. Οι προσφορές αγοράς συνήθως συνοδεύονται μαζί με κάποιες εκπτώσεις στην τιμή των προϊόντων, κάνοντας έτσι τον πελάτη να νοιώθει ευχαριστημένος από την εξυπηρέτηση του πωλητή και ικανοποιημένος από την ανταπόκριση της εταιρείας απέναντί του. Οι συστάσεις αγορών γίνονται με την ανάλυση της αγοραστικής συμπεριφοράς όλων των πελατών του οργανισμού (γίνεται εξαγωγή κανόνων συσχέτισης στις πωλήσεις των προϊόντων) και με την εφαρμογή των κανόνων αυτών στα προσωπικά δεδομένα ενός συγκεκριμένου πελάτη για την ανακάλυψη της σχέσης μεταξύ των προϊόντων που αγοράζει ο πελάτης.

Μια άλλη τεχνική για την κατανόηση των αναγκών των πελατών είναι η ομαδοποίηση των πελατών βάσει των χαρακτηριστικών τους σε κάποιες ομάδες. Οι ομάδες αυτές, αρχικά δεν είναι γνωστές και απαιτείται συνήθως η συμμετοχή κάποιου ειδικού αναλυτή, προκειμένου να χαρακτηρίσει τις ιδιαιτερότητες της κάθε ομάδας και να της αποδώσει κάποιο συγκεκριμένο χαρακτηριστικό, κυρίως βάσει κάποιων χαρακτηριστικών των πελατών που ανήκουν στην κάθε ομάδα. Έτσι, θέτοντας κάποιο συγκεκριμένο χαρακτηρισμό σε κάθε ομάδα (κάποιο όνομα με λίγα λόγια σε κάθε ομάδα που να περιγράφει τα μέλη της), κάθε πελάτης (καινούργιος η παλιός

πελάτης) μπορεί να τοποθετηθεί σε κάποια από τις ομάδες και να αντιμετωπιστεί διαφορετικά από την επιχείρηση.

Διατήρηση Πελατών – Churn Ανάλυση

Με τον ολοένα και αυξανόμενο αριθμό των ανταγωνιστικών υπηρεσιών, που είναι διαθέσιμες σήμερα, οι επιχειρήσεις απαιτείται να επικεντρώνονται στις προσπάθειες διατήρησης των υπαρχόντων πελατών τους προσφέροντας τους υψηλό επίπεδο εξυπηρέτησης και ικανοποίησης. Σε ένα τέτοιο ανταγωνιστικό περιβάλλον, που δραστηριοποιούνται συγκεκριμένου είδους επιχειρήσεις, όπως οι τηλεπικοινωνίες, οι τράπεζες και οι ασφαλιστικοί οργανισμοί, απαιτείται να μειωθούν οι φθορές των πελατών τους. Οι εταιρείες που δραστηριοποιούνται στους παραπάνω τομείς τις περισσότερες φορές (ιδίως οι εταιρείες κινητής τηλεφωνίας) αντιμετωπίζουν σοβαρά τον ανταγωνισμό τους και αντιμετωπίζουν το πρόβλημα φυγής των πελατών, καθώς αυτοί έχουν πολλές πιθανότητες να πάνε στους ανταγωνιστές. Τις περισσότερες φορές, οι επιχειρήσεις τείνουν να ανταποκρίνονται στην «φθορά» ενός πελάτη αντιδραστικά, όταν αυτός έχει αρχίσει την διαδικασία να τερματίσει την υπηρεσία του στην εταιρεία, σε περιπτώσεις κυρίως που δεν είναι ικανοποιημένος από τις υπηρεσίες που δέχεται ή όταν έχει βρει κάποια καλύτερη υπηρεσία στον ανταγωνιστή. Στην περίπτωση αυτή, η πιθανότητα να αλλάξει γνώμη ο πελάτης είναι σχεδόν μηδενική. Η churn analysis είναι η ανάλυση που γίνεται από μια εταιρεία προς τους πελάτες της και μελετά το κατά πόσο αυτοί έχουν πιθανότητα να πάνε στους ανταγωνιστές καθώς και τους λόγους φυγής τους.

Η πιο σωστή αντίδραση, λοιπόν που πρέπει να υπάρξει από μια επιχείρηση είναι να δράσει δυναμικά υιοθετώντας μια προγνωστική στρατηγική. Απαιτείται λοιπόν να γίνει μια προσεκτική ανάλυση την παλαιότερης χρήσης των υπηρεσιών από τον πελάτη, της απόδοσης των υπηρεσιών από πλευράς της επιχείρησης και να δημιουργηθούν συγκεκριμένα πρότυπα συμπεριφοράς των πελατών και μοντέλων προβλέψεων, έτσι ώστε να μπορεί να προβλεφθεί μια αποχώρηση του πελάτη στο κοντινό μέλλον. Αναλύοντας την συμπεριφορά των πελατών και τους λόγους, που αυτοί είναι πιθανόν να αποχωρήσουν, αυξάνεται η πιθανότητα η εταιρεία να πράξει αποτρεπτικά, με σωστό τρόπο έτσι ώστε να διατηρήσει τους πελάτες της και να προσελκύσει νέους. Μια καλή αντίδραση, από πλευράς της επιχείρησης είναι η δημιουργία συγκεκριμένων προσφορών και εκπτώσεων στον πελάτη με σκοπό να αυξήσουν την πιθανότητα παραμονής του στην εταιρεία αλλά και η προσφορά υπηρεσιών, που είναι καλύτερες από τον ανταγωνισμό. Οι τεχνικές Κατηγοριοποίησης πελατών σε ομάδες με συγκεκριμένα χαρακτηριστικά, ορισμένα μοντέλα πρόβλεψης της συμπεριφοράς των πελατών καθώς και η εξαγωγή κάποιων κανόνων συσχέτισης, βάσει της αγοραστικής συμπεριφοράς των πελατών είναι ορισμένα μοντέλα Εξόρυξης Γνώσης που μπορούν να βοηθήσουν τις επιχειρήσεις στο πρόβλημα της διατήρησης των πελατών τους.

Πρόβλεψη Πωλήσεων και Αποθεμάτων

Οι τεχνικές Εξόρυξης Γνώσης που μπορούν να υποστηρίξουν την πρόβλεψη πωλήσεων και αποθεμάτων σε ένα οργανισμό είναι η ανάλυση χρονοσειρών (time series). Τα πρώτα βήματα που πρέπει να γίνουν είναι η δόμηση των πωλήσεων και των αποθεμάτων σαν χρονολογικές σειρές. Αυτό συνήθως γίνεται μέσα από την αποθήκη δεδομένων (data warehouse) που είναι αποθηκευμένα τα δεδομένα και τα οποία μπορούν να δομηθούν ανάλογα με τις ανάγκες. Στη συνέχεια μπορεί να γίνει μια πρόβλεψη, βάσει των ιστορικών στοιχείων, χρησιμοποιώντας Χρονολογικές Σειρές (Time Series) ή δέντρα απόφασης με υποστήριξη παλινδρόμησης.

Για τις επιχειρήσεις οι οποίες πωλούν προϊόντα διαφορετικών ειδών, μια ανάλυση της συμπεριφοράς των πελατών της μπορεί να οδηγήσει σε μια επιτυχής διασταυρωμένη πώληση (cross sell) των προϊόντων της. Αυτό μπορεί άμεσα να οδηγήσει σε αυξημένη κερδοφορία ανά πελάτη και να δυναμώσει την σχέση της εταιρείας με αυτόν. Η ανάλυση προβλέψεων μπορεί να

βοηθήσει στην ανάλυση των δαπανών, της χρήσης και άλλες συμπεριφορές των πελατών και να βοηθήσει την επιχείρηση να πωλήσει πολλαπλά προϊόντα στην κατάλληλη στιγμή.

Επίσης, τεχνικές Εξόρυξης Γνώσης που προσφέρονται, όπως η ανάλυση του καλαθιού του πελάτη (market basket analysis) με την χρήση κανόνων συσχέτισης μπορούν να βοηθήσουν στην πρόβλεψη των αγορών που θα πραγματοποιήσει ο πελάτης. Πρακτικά αναλύεται η αγοραστική συμπεριφορά των αγοραστών, και με την εξαγωγή των κατάλληλων κανόνων η επιχείρηση μπορεί να βελτιώσει τις πωλήσεις συγκεκριμένων προϊόντων προς τον πελάτη, καθώς και να πετύχει την βέλτιστη τοποθέτηση συγκεκριμένων προϊόντων σε κοντινή απόσταση στα ράφια των καταστημάτων (για παράδειγμα τοποθέτηση δίπλα – δίπλα προϊόντων που πωλούνται πολύ συχνά μαζί) έτσι ώστε να αυξηθούν οι πωλήσεις των προϊόντων αυτών αλλά και να προγραμματιστεί καλύτερα το ύψος των αποθεμάτων από τα προϊόντα που αναλύονται.

Αποτελεσματικές Καμπάνιες Marketing

Τα τμήματα των επιχειρήσεων που ασχολούνται με το μάρκετινγκ προϊόντων έρχονται συνεχώς αντιμέτωπα με την πρόκληση της διαχείρισης του αυξανόμενου αριθμού ανταγωνιστικών προϊόντων, τις διαφορετικές προτιμήσεις των πελατών και την διαφορετικότητα των μεθόδων (κανάλια) που είναι διαθέσιμα για την επικοινωνία με τον κάθε καταναλωτή. Το αποτελεσματικό μάρκετινγκ είναι μια διαδικασία κατανόησης της διαφορετικότητας που υπάρχει σε ένα οργανισμό (σχετικά με τα θέματα που αναφέραμε παραπάνω) και στην προσαρμογή του μάρκετινγκ για μεγαλύτερη κερδοφορία. Το μάρκετινγκ εξαρτάται σε μεγάλο βαθμό από τις ακριβείς πληροφορίες που είναι διαθέσιμες σχετικά με την εκτέλεση των καμπανιών μάρκετινγκ, τις στοχευμένες προωθητικές ενέργειες κα. Μόνο με ένα πλήρες προφίλ του πελάτη μπορούν οι προωθητικές ενέργειες να είναι στοχευμένες προς τον σωστό πελάτη έτσι ώστε να αυξήσουν τα ποσοστά ανταπόκρισης και να μειώσουν το κόστος της καμπάνιας.

Αρχικά, μια τεχνική Εξόρυξης Γνώσης, που μπορεί να χρησιμοποιηθεί είναι η ομαδοποίηση πελατών για την εύρεση ομάδων πελατών, που έχουν κάποια κοινά χαρακτηριστικά. Η ομαδοποίηση των πελατών γίνεται συνήθως βάσει των δημογραφικών, των γεωγραφικών και άλλων χαρακτηριστικών τους. Η κατανόηση των ομάδων που ανακαλύπτονται, βοηθάει το τμήμα μάρκετινγκ να εντοπίσει πελάτες με συγκεκριμένα χαρακτηριστικά και να εφαρμόσει σε αυτούς συγκεκριμένες καμπάνιες. Επίσης η τεχνική της κατηγοριοποίησης με την χρήση δέντρων απόφασης βοηθάει στην τοποθέτηση πελατών σε προκαθορισμένες κατηγορίες, που έχουν κάποια συγκεκριμένα χαρακτηριστικά, που ενδιαφέρουν το τμήμα μάρκετινγκ. Ομοίως και σε αυτή την περίπτωση, σε κάθε πελάτη που ανήκει σε κάποιες συγκεκριμένες κατηγορίες εφαρμόζονται προκαθορισμένες καμπάνιες.

Εκτός από τις στοχευμένες καμπάνιες, που αναφέρονται σε συγκεκριμένους πελάτες, υπάρχουν και οι στοχευμένες διαφημίσεις, κυρίως σε διάφορα sites λιανικής πώλησης. Για την πραγματοποίηση των στοχευμένων διαφημίσεων, χρησιμοποιούνται τεχνικές Εξόρυξης Γνώσης με ανάλυση των προτύπων πλοήγησης στις σελίδες (sequence analysis) καθώς και εύρεση προτύπων σχετικά με την online αγορά προϊόντων βάσει των ιστορικών πωλήσεων και των προτιμήσεων των πελατών, έτσι ώστε να εμφανιστούν στοχευμένες διαφημίσεις στους επισκέπτες του site.

Εντοπισμός και Πρόληψη Απάτης

Η απάτη σε μια εταιρεία είναι ένα πολύ σοβαρό πρόβλημα που πρέπει να αντιμετωπίσει. Η απάτη περιλαμβάνει διάφορα προβλήματα, όπως ανακριβείς αιτήσεις για χορήγηση πιστωτικών καρτών, παράνομες συναλλαγές, κλοπή ταυτότητας και ψευδείς απαιτήσεις ασφάλειας και αποτελούν μάστιγα για πολλές επιχειρήσεις και κυρίως για εταιρείες που ασχολούνται με εκδόσεις πιστωτικών καρτών, ασφαλιστικές εταιρείες, εμπόρους λιανικής πώλησης, προμηθευτές επιχειρήσεων, εταιρείες παροχής υπηρεσιών, κα. Μια λύση

αντιμετώπισης των προβλημάτων αυτών είναι ο εντοπισμός ανωμαλιών, τα δεδομένα δηλαδή που δεν ταιριάζουν με το γενικότερο σύνολο. Για τον εντοπισμό των ανωμαλιών, μπορούν να χρησιμοποιηθούν τεχνικές ομαδοποίησης (clustering) για τον εντοπισμό ομάδων βάσει των χαρακτηριστικών των δεδομένων. Τα δεδομένα τα οποία βρίσκονται έξω από τις ομάδες και έχουν ακραίες τιμές (outliers) ξεχωρίζουν εύκολο από το σύνολο. Οι τιμές αυτές αναλύονται ξεχωριστά και μπορούν εύκολα να εντοπιστούν σε αυτές τυχών σφάλματα.

Ένα άλλο πρόβλημα είναι ο εντοπισμός παράνομων ή εσφαλμένων κινήσεων, σε μια σειρά συναλλαγών στην διάρκεια του χρόνου. Για παράδειγμα, μια τράπεζα μπορεί να εντοπίσει αν μια συναλλαγή μια πιστωτικής κάρτας δεν είναι ορθή, αν τα χαρακτηριστικά της συναλλαγής ξεπερνούν κάποιο όριο πέρα από το συνηθισμένο (για παράδειγμα μια πολύ μεγάλη συναλλαγή που ξεχωρίζει ή μια συναλλαγή μέσω internet ενώ ο χρήστης δεν πραγματοποιεί τέτοιες συναλλαγές). Στην περίπτωση αυτή ένα μήνυμα μπορεί να ενημερώσει κάποιον υπεύθυνο για την ανωμαλία της συναλλαγής και αυτός με την σειρά του επικοινωνήσει με τον πελάτη για να μάθει αν όντως αυτός πραγματοποίησε την συναλλαγή αυτή. Οι τεχνικές Εξόρυξης Γνώσης είναι σημαντικές σε αυτές τις περιπτώσεις καθώς οι συναλλαγές ημερησίως μπορεί να είναι αρκετές χιλιάδες και δεν μπορεί κάποιος να παρακολουθεί όλες τις συναλλαγές που γίνονται. Έτσι με την βοήθεια της ομαδοποίησης, ξεχωρίζουν αμέσως οι συναλλαγές που διαφέρουν από το σύνολο και γίνονται οι απαραίτητες ενέργειες από τους υπευθύνους.

Επίσης, οι τεχνικές εντοπισμού ανωμαλιών μπορούν να χρησιμοποιηθούν για τον έλεγχο της καταχώρησης δεδομένων σε ένα σύστημα (data entry validation). Αναλύονται δηλαδή όλες οι καταχωρήσεις δεδομένων σε ένα σύστημα και εντοπίζονται οι περιπτώσεις εκείνες όπου η τιμή που δίνεται δεν ταιριάζει με το σύνολο, για παράδειγμα η καταχώρηση μιας μεγάλης ηλικίας σε ένα σύστημα που καταγράφει στοιχεία εφήβων. Ο έλεγχος γίνεται την στιγμή της καταχώρησης και το σύστημα ενημερώνει ανάλογα τον χρήστη στην περίπτωση που εντοπιστεί κάποιο πιθανό σφάλμα.

Η πρόληψη απάτης μπορεί να αποφευχθεί με την δημιουργία μοντέλων κατηγοριοποίησης, βάσει συγκεκριμένων χαρακτηριστικών των πελατών. Οι προκαθορισμένες ομάδες μπορούν να περιλαμβάνουν άτομα, που είναι πιθανό να είναι κακοπληρωτές πελάτες σε μια εταιρεία, να έχουν υψηλό πιστωτικό ρίσκο σε μια τράπεζα ή να είναι πελάτες που μπορεί σύντομα να αποχωρήσουν από την εταιρεία. Στην περίπτωση αυτή τα δέντρα απόφασης μπορούν να χρησιμοποιηθούν για την κατηγοριοποίηση νέων πελατών και έτσι οι επιχειρήσεις να μπορούν να κατατάξουν τους πελάτες σε κάποια από τις κατηγορίες προλαμβάνοντας έτσι μελλοντικά προβλήματα.

Εξόρυξη Γνώσης στο CRM

Προκειμένου να επηρεάσει η Εξόρυξη Γνώσης τον οργανισμό, πρέπει να έχει σχέση με την διαδικασία του οργανισμού την οποία υποστηρίζει. Η Εξόρυξη Γνώσης αποτελεί τμήμα μιας πολύ μεγαλύτερης σειράς βημάτων που πραγματοποιούνται μεταξύ του οργανισμού και των πελατών της. Η τρόπος με τον οποίο η Εξόρυξη Γνώσης επηρεάζει τον οργανισμό εξαρτάται από την ίδια την επιχειρηματική διαδικασία και όχι την διαδικασία της Εξόρυξης Γνώσης αυτή καθεαυτή.

Το θέμα το οποίο πρέπει να σημειωθεί στο σημείο αυτό είναι ότι τα αποτελέσματα της Εξόρυξης Γνώσης είναι διαφορετικά από άλλες διαδικασίες που επηρεάζουν τα δεδομένα του οργανισμού. Στις περισσότερες τυποποιημένες διαδικασίες, που αλληλεπιδρούν με τα δεδομένα, σχεδόν όλα τα αποτελέσματα που παρουσιάζονται στους χρήστες είναι πράγματα τα οποία γνωρίζουν ότι υπάρχουν ήδη στην Βάση Δεδομένων. Για παράδειγμα, μια αναφορά που δείχνει την κατανομή των πωλήσεων ανά κατηγορία προϊόντων ή ανά περιοχή, μπορεί εύκολα να γίνει κατανοητή από τον χρήστη γιατί μπορεί να γνωρίζει διαισθητικά ότι τέτοιου είδους πληροφορία υπάρχει στην βάση. Εάν η εταιρεία αρχίσει να πωλεί σε διαφορετικές περιοχές μια χώρας, δεν υπάρχει πρόβλημα στο να παρουσιάσουν μια τέτοια καινούργια πληροφορία για να γίνει κατανοητή η επιχειρηματική διαδικασία.

Η Εξόρυξη Γνώσης, από την άλλη πλευρά, εξάγει πληροφορία από την Βάση Δεδομένων την οποία ο χρήστης δεν γνωρίζει ότι υπάρχει. Οι σχέσεις μεταξύ των μεταβλητών και της συμπεριφοράς των πελατών (που δεν είναι άμεσα ορατή) είναι η πολύτιμη πληροφορία που η Εξόρυξη Γνώσης ευελπιστεί να ανακαλύψει. Και επειδή ο χρήστης δεν μπορεί να ξέρει εκ των προτέρων τι έχει ανακαλύψει η Εξόρυξη Γνώσης, είναι ένα πολύ μεγάλο άλμα το να λάβει ο χρήστης την έξοδο του συστήματος Εξόρυξης Γνώσης και να το μεταφράσει σε λύση ενός επιχειρησιακού προβλήματος.

Οι χρήστες των marketing εφαρμογών πρέπει να κατανοήσουν τα αποτελέσματα της Εξόρυξης Γνώσης, πριν τα θέσουν σε δράση. Επειδή η Εξόρυξη Γνώσης συνήθως περιλαμβάνει εξαγωγή «κρυφών» προτύπων της συμπεριφοράς των πελατών, η διαδικασία κατανόησης των αποτελεσμάτων Εξόρυξης Γνώσης μπορεί να είναι λίγο περίπλοκη. Το κλειδί της κατανόησης των αποτελεσμάτων της Εξόρυξης Γνώσης είναι να τοποθετήσουν τους χρήστες σε ένα πλαίσιο όπου νοιώθουν άνετα με τις πληροφορίες που δέχονται και να τις μελετήσουν μέχρι να καταλάβουν ότι δεν μπορούσαν να δουν προηγουμένως.

Οι χρήστες πρέπει να βλέπουν τα αποτελέσματα της Εξόρυξης Γνώσης σε ένα τέτοιο πλαίσιο, έτσι ώστε να μπορούν να τα κατανοήσουν. Εάν μπορέσουν να καταλάβουν αυτά που έχουν ανακαλυφθεί από την Εξόρυξη Γνώσης, θα μπορέσουν να τα εμπιστευθούν και να τα χρησιμοποιήσουν. Υπάρχουν, επομένως δύο θέματα σε αυτό το πρόβλημα:

1. Παρουσίαση των αποτελεσμάτων της διαδικασίας Εξόρυξης Γνώσης με ένα κατανοητό τρόπο, και
2. Δυνατότητα των χρηστών να αλληλεπιδρούν με τα αποτελέσματα έτσι ώστε να μπορούν να λάβουν απαντήσεις σε απλές ερωτήσεις τους.

Χρήση τεχνικών Εξόρυξης Γνώσης στο CRM

Μια εφαρμογή CRM (Customer Relationship Management) είναι μια διαδικασία η οποία διαχειρίζεται τις αλληλεπιδράσεις μεταξύ του οργανισμού και των πελατών της. Οι κύριοι χρήστες του CRM είναι χρήστες του τμήματος μάρκετινγκ, οι οποίοι ψάχνουν να αυτοματοποιήσουν την διαδικασία της αλληλεπίδρασης με τους πελάτες. Για να υπάρχει επιτυχία, οι χρήστες μάρκετινγκ πρέπει πρώτα να ανακαλύψουν ομάδες, οι οποίες περιλαμβάνουν πελάτες ή προσδοκώμενους πελάτες με δυνατότητα υψηλής απόδοσης κέρδους. Στην συνέχεια κατασκευάζουν και εκτελούν διαφημιστικές καμπάνιες, οι οποίες θα επηρεάσουν θετικά την συμπεριφορά αυτών των πελατών.

Πρώτη χρήση – Ομάδες Πελατών

Το πρώτο βήμα, ο εντοπισμός των τμημάτων της αγοράς, απαιτεί σημαντικές πληροφορίες σχετικά με μελλοντικούς πελάτες και την αγοραστική συμπεριφορά τους. Στην θεωρία, όσο πιο πολλά δεδομένα υπάρχουν, τόσο το καλύτερο. Στην πράξη όμως, οι τεράστιες ποσότητες δεδομένων συχνά εμποδίζουν τους χρήστες του μάρκετινγκ, οι οποίοι προσπαθούν να ανακαλύψουν από τον τεράστιο όγκο πληροφοριών την παραμικρή πληροφορία η οποία μπορεί να φανεί χρήσιμη. Οι εφαρμογές Εξόρυξης Γνώσης αυτοματοποιούν την διαδικασία αναζήτησης στο πλήθος των δεδομένων για να βρουν πρότυπα, τα οποία είναι χρήσιμοι δείκτες της αγοραστικής συμπεριφοράς των πελατών. Τα πρότυπα αυτά αναπαριστούν ομάδες πελατών, με κοινά χαρακτηριστικά.

Δεύτερη χρήση –Campaign Management Systems

Μετά την εξόρυξη των δεδομένων, οι χρήστες μάρκετινγκ πρέπει να τροφοδοτήσουν τα αποτελέσματα στα συστήματα διαχείρισης των διαφημιστικών καμπανιών (campaign management systems – CMS) τα οποία, όπως υποδηλώνει και το όνομα τους, διαχειρίζονται τις διαφημιστικές εκστρατείες του οργανισμού σε προκαθορισμένα τμήματα της αγοράς. Στο παρελθόν, η σχέση μεταξύ Εξόρυξης Γνώσης και CMS γινόταν κυρίως με την μεσολάβηση κάποιου χρήστη. Στις χειρότερες περιπτώσεις, απαιτούνταν η δημιουργία ενός φυσικού αρχείου (σε μια δισκέτα ή ένα CD), το οποίο μετέφερε κάποιος σε έναν άλλο υπολογιστή και τα δεδομένα φορτώνονταν στην Βάση Δεδομένων του μάρκετινγκ. Σε πιο σύγχρονες εφαρμογές, η επικοινωνία των δύο συστημάτων γίνεται αυτόματα. Η ολοκλήρωση των συστημάτων αυτών παρουσιάζει μια ευκαιρία για τους οργανισμούς, έτσι ώστε να αποκτήσουν ανταγωνιστικό πλεονέκτημα.

Όσο πιο κοντά δουλεύουν τα συστήματα Εξόρυξης Γνώσης και CMS, τόσο καλύτερα είναι τα αποτελέσματα για τον οργανισμό.

Τρίτη χρήση – Marketing

Η Εξόρυξη Γνώσης βοηθά τους χρήστες μάρκετινγκ να στοχεύσουν τις εκστρατείες διαφήμισης με μεγαλύτερη ακρίβεια, καθώς επίσης και να ευθυγραμμίσουν τις εκστρατείες πιο στενά με τις ανάγκες, τα θέλω και την στάση των πελατών τους. Εάν οι απαιτούμενες πληροφορίες υπάρχουν στην βάση δεδομένων, η διαδικασία Εξόρυξης Γνώσης μπορεί να μοντελοποιήσει σχεδόν κάθε δραστηριότητα του πελάτη. Το κλειδί στην περίπτωση αυτή είναι η εύρεση προτύπων σχετικά με τα τρέχοντα προβλήματα του οργανισμού.

Οι τυπικές ερωτήσεις που αντιμετωπίζει η διαδικασία Εξόρυξης Γνώσης περιλαμβάνουν τα παρακάτω:

- Ποιοι πελάτες είναι οι περισσότερο πιθανοί να εγκαταλείψουν μια υπηρεσία του οργανισμού (πχ συμβόλαιο κινητής τηλεφωνίας);
- Ποια είναι η πιθανότητα ένας πελάτης να αγοράσει τουλάχιστον ένα δεδομένο ποσό (για παράδειγμα 100 ευρώ) από ένα συγκεκριμένο κατάλογο προϊόντων, που μοιράζεται δια αλληλογραφίας;
- Ποια άτομα είναι πολύ πιθανό να ανταποκριθούν σε μια συγκεκριμένη αγορά και να γίνουν πελάτες του οργανισμού;

Οι απαντήσεις στα ερωτήματα αυτά μπορούν να βοηθήσουν στην διατήρηση των πελατών και να αυξήσουν τα ποσοστά ανταπόκρισης στην διαφημιστική εκστρατεία, η οποία, με τη σειρά της, αυξάνει την αγορά, τις σύνθετες πωλήσεις προϊόντων (cross-selling) και την απόδοση της επένδυσης (ROI) στα συστήματα διαχείρισης διαφημιστικών εκστρατειών.

Τέταρτη χρήση – Scoring

Η διαδικασία της Εξόρυξης Γνώσης κατασκευάζει μοντέλα με την χρήση δεδομένων εισόδου από την Βάση Δεδομένων για να προβλέψει την συμπεριφορά των πελατών. Για παράδειγμα, μια τέτοια συμπεριφορά μπορεί να είναι η αποχώρηση ενός πελάτη από μια ετήσια συνδρομή σε ένα περιοδικό, η αγορά πολλαπλών προϊόντων, η προθυμία του πελάτη να χρησιμοποιήσει μια πιστωτική κάρτα για αγορές ή μια χρεωστική κάρτα και άλλα παραδείγματα. Η πρόβλεψη που παράγεται από τα μοντέλα Εξόρυξης Γνώσης ονομάζονται συνήθως βαθμολογίες (scores). Μια βαθμολογία (συνήθως μια αριθμητική τιμή) αποδίδεται σε κάθε εγγραφή στην Βάση Δεδομένων και υποδεικνύει την πιθανότητα, το οποίου η εγγραφή έχει σημειωθεί με αυτή την βαθμολογία, θα παρουσιάσει μια συγκεκριμένη συμπεριφορά.

Για παράδειγμα, αν το μοντέλο προσπαθεί να προβλέψει την αποχώρηση του πελάτη, μια υψηλή βαθμολογία μπορεί να σημαίνει ότι ο πελάτης είναι πολύ πιθανό να αποχωρήσει ενώ μια μικρή βαθμολογία μπορεί να δείχνει το αντίθετο. Μετά την βαθμολόγηση του συνόλου των πελατών, οι εγγραφές με μια συγκεκριμένη ομάδα μπορούν να επιλεχθούν για να τοποθετηθούν σε μια συγκεκριμένη και στοχοθετημένη διαφημιστική εκστρατεία για την προσπάθεια διατήρησης των πελατών αυτών στην εταιρεία.

Τα συστήματα διαχείρισης διαφημιστικών εκστρατειών (CMS) χρησιμοποιούν τα αποτελέσματα των μοντέλων Εξόρυξης Γνώσης για να ενισχύσουν την εστίαση σε συγκεκριμένους πελάτες, με αποτέλεσμα την αύξηση του ρυθμού ανταπόκρισης και της αποτελεσματικότητας των εκστρατειών. Στην ιδανική περίπτωση, οι χρήστες μάρκετινγκ που κατασκευάζουν τις διαφημιστικές εκστρατείες να πρέπει να είναι σε θέση να εφαρμόσουν κάθε μοντέλο που βρίσκεται στα δικά τους συστήματα (CMS) σε ένα καθορισμένο σύνολο πελατών (για παράδειγμα ένα σύνολο πελατών που έχει προκύψει από ένα αλγόριθμο Εξόρυξης Γνώσης που πραγματοποιεί Ομαδοποίηση Πελατών).

4. Microsoft SQL Server Analysis Services

Οι τεχνικές Εξόρυξης Γνώσης, όπως τις παρουσιάσαμε στο Κεφάλαιο 3, είναι ένα σύνολο από εξελιγμένα εργαλεία και αλγορίθμους, που επιτρέπουν στους αναλυτές και στους τελικούς χρήστες να λύσουν προβλήματα, που αναφέρονται σε τεράστιους όγκους δεδομένων και που σε κανονικές συνθήκες, η ανάλυση των δεδομένων αυτών θα απαιτούσε χρόνο – που πρακτικά τα προβλήματα αυτά δεν θα μπορούσαν ποτέ να επιλυθούν. Για την εφαρμογή των τεχνικών Εξόρυξης Γνώσης υπάρχουν αρκετές εμπορικές εφαρμογές, που μπορούν να χρησιμοποιηθούν από τους προγραμματιστές, αναλυτές και γενικώς από τους χρήστες που ενδιαφέρονται για τα αποτελέσματα μιας διαδικασίας της Εξόρυξης Γνώσης.

Μια σπουδαία και πολύ γνωστή εμπορική εφαρμογή για υλοποίηση Εξόρυξης Γνώσης είναι ο SQL Server της Microsoft (εκδόσεις 2005 και 2008), που εκτός από την διαχείριση Βάσεων Δεδομένων, παρέχει την πλατφόρμα διαχείρισης της Επιχειρησιακής Ευφυΐας (Business Intelligence – BI) με τις παρακάτω λειτουργίες:

- SQL Server Analysis Services (SSAS), που παρέχει λειτουργίες OLAP ανάλυσης (δημιουργία κύβων, διαστάσεων κλπ) καθώς και βασικές τεχνικές Εξόρυξης Γνώσης (για παράδειγμα Microsoft Decision Trees, Microsoft Clustering, Microsoft Association Rules, Microsoft Time Series κλπ).
- SQL Server Integration Services (SSIS), που παρέχει λειτουργίες για την μεταφορά δεδομένων από διάφορες πηγές σε ένα κεντρικό σημείο διαχείρισης, ελέγχου και διόρθωσης δεδομένων, προγραμματισμού λειτουργιών ανάγνωσης και εισαγωγής δεδομένων κλπ.
- SQL Server Reporting Services (SSRS), για την δημιουργία αναφορών.

Σκοπός του κεφαλαίου αυτού είναι να παρουσιαστούν οι βασικές τεχνικές Εξόρυξης Γνώσης, που παρέχονται από τον SQL Server και ομαδοποιούνται κάτω από τον γενικό τίτλο SQL Server Analysis Services (SSAS). Η ονομασία SSAS θα χρησιμοποιηθεί στην συνέχεια της παρούσας εργασίας. Στο επόμενο κεφάλαιο θα παρουσιαστούν διάφορα παραδείγματα χρήσης διαφόρων τεχνικών Εξόρυξης Γνώσης που παρέχουν τα SSAS, με την παρουσίαση των βασικών βημάτων που πρέπει να εκτελέσει ο χρήστης μέσα από το περιβάλλον Business Intelligence Development Studio (BIDS) το οποίο είναι κάτω από το Microsoft Visual Studio 2008.

Εξόρυξη Γνώσης με τα SSAS

Όπως έχουμε τονίσει, η Εξόρυξη Γνώσης είναι ένα σύνολο εργαλείων που επιτρέπει στους χρήστες να ανακαλύψουν πρότυπα και τάσεις ανάμεσα σε καθορισμένα σύνολα δεδομένων σε μία επιχείρηση. Τέτοια ανάλυση είναι χρήσιμη σε συγκεκριμένες επιχειρηματικές λειτουργίες, όπως η πρόβλεψη μελλοντικών τιμών και η καλύτερη κατανόηση των αποτελεσμάτων που έχει επιτύχει η επιχείρηση στο παρελθόν, έτσι ώστε να βελτιώσει το μέλλον της. Με την χρήση των εργαλείων των SSAS, ο χρήστης μπορεί να δημιουργήσει **Δομές Εξόρυξης Γνώσης** (**data mining structures**), οι οποίες περιέχουν **Μοντέλα Εξόρυξης Γνώσης** (**data mining models**). Πρακτικά, ένα μοντέλο Εξόρυξης Γνώσης αναπαριστά την εφαρμογή ενός συγκεκριμένου αλγορίθμου Εξόρυξης Γνώσης. Ο Αλγόριθμος είναι μια μαθηματική συνάρτηση που πραγματοποιεί μια συγκεκριμένη μορφή ανάλυσης σε ένα σχετιζόμενο σύνολο δεδομένων.

Μοντέλα Εξόρυξης Γνώσης (Data Mining Models)

Η έννοια του μοντέλου Εξόρυξης Γνώσης είναι παρόμοια με την έννοια ενός πίνακα σε μια σχεσιακή Βάση δεδομένων. Το μοντέλο Εξόρυξης Γνώσης καθορίζει πως θα αναλυθούν τα

δεδομένα και πως θα γίνει η πρόβλεψη σε αυτά. Με λίγα λόγια, το μοντέλο Εξόρυξης Γνώσης περιέχει τον ορισμό των στηλών δεδομένων που θα χρησιμοποιηθούν από τον Αλγόριθμο Εξόρυξης Γνώσης, που αναπτύσσεται κατά την διάρκεια της διαδικασίας Εξόρυξης Γνώσης.

Στον παρακάτω πίνακα παρουσιάζουμε ένα παράδειγμα ορισμού χρήσης των στηλών δεδομένων για ένα πίνακα δεδομένων[1] που εισάγονται σε ένα αλγόριθμο κατηγοριοποίησης με χρήση Δέντρων Απόφασης.

Customer ID	Name	Age	Gender	Education	Byer
1	User1	25	M	Graduate	Y
2	User2	35	F	College	N
3	User3	30	M	High School	Y

Ο αλγόριθμος Κατηγοριοποίησης στο παράδειγμα που αναφέρουμε χρησιμοποιείται για να προβλέψει αν κάποιος πελάτης θα αγοράσει ένα προϊόν, βάσει των διαφόρων χαρακτηριστικών του, όπως η ηλικία του (Age), το φύλο του (Gender) και το επίπεδο εκπαίδευσης του (Education). Για τον σκοπό αυτό, θα χρησιμοποιηθεί ο αλγόριθμος Δέντρων Απόφασης για να προβλέψει την τιμή της στήλης Byer.

Για το παραπάνω παράδειγμα, ο ορισμός χρήσης των στηλών δεδομένων (Customer ID, Name, Age, Gender, Education, Byer) ορίζεται όπως παρακάτω:

Customer ID	Name	Age	Gender	Education	Byer
1	User1	25	M	Graduate	Y
2	User2	35	F	College	N
3	User3	30	M	High School	Y

Αν μια στήλη στο σύνολο δεδομένων αναγνωρίζει μια γραμμή δεδομένων (case – περίπτωση) στο μοντέλο, τότε η χρήση της στήλης πρέπει να οριστεί ως *Κλειδί* (*Key*) (η στήλη *customer id* αναγνωρίζει μοναδικά την κάθε εγγραφή, οπότε η χρήση της ορίζεται σαν κλειδί). Μια στήλη που ορίζεται σαν *Δεδομένο Εισόδου* (*Input*) τότε χρησιμοποιείται σαν παράμετρος εισόδου στον αλγόριθμο (οι στήλες *age*, *gender*, *education* του παραπάνω πίνακα ορίζονται ως στήλες εισόδου δεδομένων στην αλγόριθμο δέντρων Απόφασης). Ένας αλγόριθμος Εξόρυξης Γνώσης μπορεί να έχει μια ή περισσότερες στήλες για πρόβλεψη. Η στήλη που ορίζεται ως στήλη *Πρόβλεψης* (*Predict*) τότε η στήλη χρησιμοποιείται και για στήλη Εισόδου, ενώ αν η στήλη οριστεί ως *Πρόβλεψη Μόνο* (*PredictOnly*) τότε η στήλη των δεδομένων θα χρησιμοποιηθεί μόνο για πρόβλεψη και όχι και σαν στήλη δεδομένων εισόδου (η χρήση της στήλη δεδομένων *byer* ορίζεται ως στήλη πρόβλεψης). Τέλος αν μια στήλη δεν προσφέρει κάποιο γνώση (όπως για παράδειγμα η στήλη name, όπου το όνομα του πελάτη δεν προσφέρει κάποια πληροφορία σχετικά με το αν ο πελάτης θα αγοράσει από το κατάστημα ή όχι) τότε η χρήση της στήλης αυτής *δεν λαμβάνεται υπόψη* (*Ignore*) από τον αλγόριθμο.

Ένα μοντέλο Εξόρυξης Γνώσης πρέπει να εκπαιδευθεί πριν χρησιμοποιηθεί. Η εκπαίδευση του μοντέλου πραγματοποιείται με την εισαγωγή σε αυτό ένα σύνολο δεδομένων

[1] Τα δεδομένα του παραδείγματος προέρχονται από την Βάση Δεδομένων AdventureWorksDW2008 που παρέχεται από την Microsoft και χρησιμοποιείται για εκπαίδευση των χρηστών στις τεχνολογίες Business Intelligence (και όχι μόνο) της Microsoft. Τα δεδομένα της ίδιας Βάσης θα χρησιμοποιηθούν και στο Κεφάλαιο 5, για την παρουσίαση των βασικών βημάτων χρήσης μιας εφαρμογής Data Mining μέσα από το περιβάλλον SSAS του MS SQL Server.

εκπαίδευσης. Στη συνέχεια, το μοντέλο Εξόρυξης Γνώσης μπορεί να χρησιμοποιηθεί για πρόβλεψη, εισάγοντας σε αυτό καινούργια δεδομένα.

Δομές Εξόρυξης Γνώσης (Data Mining Structures)

Για την επίλυση ενός προβλήματος του οργανισμού με την χρήση τεχνικών Εξόρυξης Γνώσης κατασκευάζεται ένα μοντέλο Εξόρυξης Γνώσης, επιλέγεται δηλαδή ένας αλγόριθμος Εξόρυξης Γνώσης και ορίζονται οι στήλες των δεδομένων που θα χρησιμοποιηθούν στο μοντέλο αυτό. Ένα επιχειρησιακό πρόβλημα όμως, δύναται να επιλυθεί με την χρήση περισσοτέρων από ένα διαφορετικών αλγορίθμων Εξόρυξης Γνώσης. Για το λόγο αυτό, χρησιμοποιείται η έννοια της δομής δεδομένων, η οποία περιέχει ένα ή περισσότερα μοντέλα Εξόρυξης Γνώσης.

Σε μια δομή Εξόρυξης Γνώσης μπορούν να χρησιμοποιηθούν πολλοί αλγόριθμοι Εξόρυξης Γνώσης, για να επιλύσουν το ίδιο πρόβλημα του οργανισμού. Αυτό παρατηρείται πολλές φορές κατά την διάρκεια επίλυσης ενός προβλήματος BI με την χρήση τεχνικών Εξόρυξης Γνώσης, να χρησιμοποιούνται δηλαδή διαφορετικοί αλγόριθμοι. Αυτό δίνει το πλεονέκτημα στον χρήστη να κάνει τον ορισμό των δεδομένων μια φορά και να εφαρμόσει σε αυτά διάφορες τεχνικές Εξόρυξης Γνώσης και στο τέλος να επιλέξει τον αλγόριθμο ο οποίος επιστρέφει (ή καλύτερα φαίνεται να επιστρέφει) τα καλύτερα αποτελέσματα.

Εννοιολογικά, ένα μοντέλο Εξόρυξης Γνώσης είναι σαν έναν πίνακα δεδομένων, που περιέχει τον ορισμό και την χρήση των στηλών δεδομένων ενώ μια δομή Εξόρυξης Γνώσης είναι ένα σχέδιο της Βάσης Δεδομένων το οποίο μπορεί να περιέχει ένα ή περισσότερα μοντέλα. Σε όρους αντικειμενοστραφούς προγραμματισμού, μια δομή Εξόρυξης Γνώσης είναι συνώνυμη με μια Κλάση, ενώ τα μοντέλα Εξόρυξης Γνώσης τα οποία υπάρχουν στην ίδια δομή μπορούν να παρουσιαστούν σαν τα στιγμιότυπα της Κλάσης αυτής.

Τεχνικές Εξόρυξης Γνώσης

Ένα από τα πιο βασικά ζητήματα της Εξόρυξης Γνώσης στα SSAS είναι η *κατανόηση* του τι πραγματικά κάνει το κάθε μοντέλο (αλγόριθμος) και στη συνέχεια η *δημιουργία* μιας Δομής Εξόρυξης Γνώσης, η οποία να περιέχει τον κατάλληλο αλγόριθμο που υποστηρίζει κατάλληλα τις επιχειρησιακές ανάγκες. Επίσης, μέσα από την κατανόηση των λειτουργιών της κάθε τεχνικής, ο χρήστης μπορεί να διαμορφώσει κατάλληλα τις *παραμέτρους* του χρησιμοποιούμενου αλγορίθμου, έτσι ώστε να έχει το επιθυμητό αποτέλεσμα – η διαμόρφωση των παραμέτρων των αλγορίθμων που θα χρησιμοποιηθούν αποτελεί επίσης ένα από τα βασικά ζητήματα της Εξόρυξης Γνώσης στα SSAS. Τέλος, ένα σημαντικό θέμα είναι η *παρουσίαση* των *αποτελεσμάτων* Εξόρυξης Γνώσης του κάθε χρησιμοποιούμενο Μοντέλου στους τελικούς χρήστες.

Οι **κυριότερες τεχνικές Εξόρυξης Γνώσης**, που παρέχονται μέσα από το περιβάλλον του SSAS είναι οι παρακάτω.

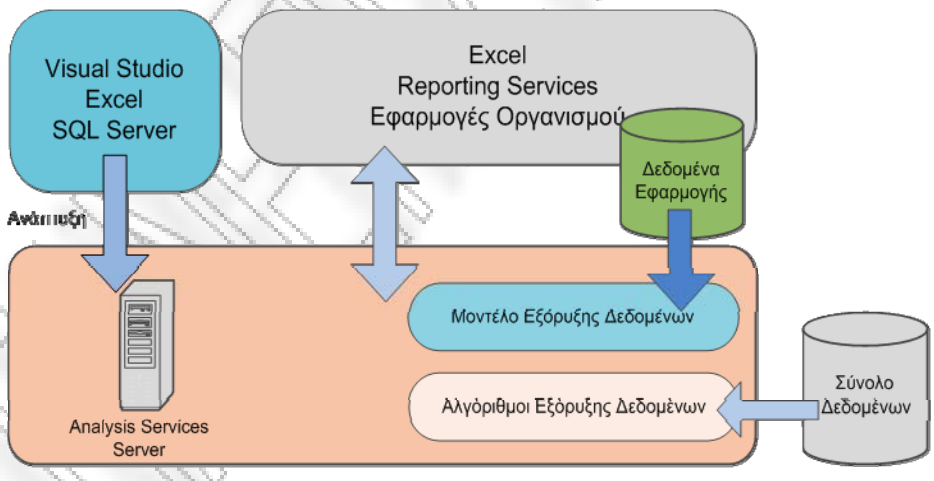
- **Classification (Κατηγοριοποίηση)**: Η κατηγοριοποίηση αναφέρεται στην πρόβλεψη μιας ή περισσότερων σταθερών μεταβλητών βάσει των τιμών πολλαπλών μεταβλητών εισόδου. Οι τεχνικές αυτές χρησιμοποιούνται συνήθως σε περιπτώσεις εταιρειών που έχουν ένα μεγάλο όγκο ιστορικών δεδομένων, υψηλής ποιότητας (ακρίβειας). Ο αλγόριθμος που χρησιμοποιείται για κυρίως κατηγοριοποίηση και περιλαμβάνεται στα SSAS είναι ο Microsoft Decision Trees, ενώ μπορούν επίσης να χρησιμοποιηθούν οι αλγόριθμοι Microsoft Naïve Bayes και Microsoft Neural Network.
- **Clustering (Ομαδοποίηση)**: Η τεχνική της ομαδοποίησης αναφέρεται στην συγκέντρωση δεδομένων σε κατηγορίες (ομάδες – segment/buckets) βάσει ενός συνόλου δεδομένων εισόδου ή γνωρισμάτων. Όλα τα γνωρίσματα και οι τιμές εισόδου έχουν την ίδια βαρύτητα στον προσδιορισμό των ομάδων. Συνήθως οι τεχνικές ομαδοποίησης βοηθούν του χρήστες να καταλάβουν καλύτερα τις σχέσεις μεταξύ των γνωρισμάτων μέσα από ένα πολύ μεγάλο σύνολο δεδομένων. Επίσης, υπάρχουν

εταιρείες που χρησιμοποιούν τις τεχνικές αυτές για να κάνουν «έξυπνες» προβλέψεις, όπως για παράδειγμα μια οντότητα που ανήκει σε μια ομάδα, θα συμπεριφέρεται συνήθως σαν και τις άλλες οντότητες της ίδιας ομάδας. Ο αλγόριθμος που χρησιμοποιείται στα SSAS είναι ο Microsoft Clustering.

- **Association (Συσχέτιση)**: Η τεχνική της συσχέτισης προσπαθεί να βρει συσχετίσεις μεταξύ των μεταβλητών σε ένα σύνολο δεδομένων. Ο στόχος του αλγορίθμου είναι να βρει σύνολα αντικειμένων, που έχουν κάποιες σχέσεις μεταξύ τους (συνήθως βάσει στοιχείων πωλήσεων προϊόντων). Η πιο γνωστή εφαρμογή τέτοιου είδους αλγορίθμων είναι η ανάλυση του Καλαθιού Πωλήσεων, όπου δημιουργούνται κανόνες συσχέτισης για τα αντικείμενα που αγοράζονται από τους καταναλωτές για να βρεθούν οι συσχετίσεις ανάμεσα τους. Ο αλγόριθμος που χρησιμοποιείται στα SSAS είναι ο Microsoft Association.
- **Forecasting (Πρόβλεψη) / Regression (Παλινδρόμηση)**: Η πρόβλεψη περιλαμβάνει μια διαδικασία, που είναι παρόμοια με την Κατηγοριοποίηση, με την έννοια ότι επιχειρείται μια πρόβλεψη μιας τιμής, βάσει πολλαπλών τιμών εισόδου. Η διαφορά εδώ είναι ότι η πρόβλεψη γίνεται σε μια συνεχή τιμή. Στην περίπτωση της πρόβλεψης, συνήθως τα δεδομένα εισόδου σχετίζονται με μια χρονολογική σειρά (Time Series). Συνήθως οι εταιρείες χρησιμοποιούν την τεχνική αυτή για να προβλέψουν το ύψος των πωλήσεων ενός προϊόντος, βάσει της λιανικής τιμής του, της θέσης του στα ράφια του καταστήματος κλπ. Ο αλγόριθμος που χρησιμοποιείται στα SSAS είναι ο Microsoft Time Series.
- **Ανάλυση Ακολουθιών (Sequence Analysis)**: Ομαδοποίηση και παρουσίαση συγκεντρωτικών πληροφοριών για διάφορες ακολουθίες δεδομένων, όπως για παράδειγμα οι ακολουθίες των επισκέψεων στις σελίδες ενός web site. Ο αλγόριθμος ανάλυσης ακολουθιών που χρησιμοποιείται είναι ο Microsoft Sequence Clustering.

Αρχιτεκτονική Συστήματος Εξόρυξης Γνώσης στα SSAS

Στο παρακάτω διάγραμμα παρουσιάζουμε την αρχιτεκτονική ενός συστήματος Εξόρυξης Γνώσης, με την χρήση των αλγορίθμων Εξόρυξης Γνώσης που παρέχονται μέσα από τα SSAS του Microsoft SQL Server.



Διάγραμμα 13: Αρχιτεκτονική Συστήματος Εξόρυξης Γνώσης

Μέσα από το περιβάλλον ανάπτυξης του Visual Studio, το Excel ή με την χρήση των κατάλληλων εντολών του SQL Server, πραγματοποιείται η ανάπτυξη των Μοντέλων Εξόρυξης

Γνώσης, με την χρήση των αντίστοιχων αλγορίθμων, ανάλογα με τις ανάγκες του οργανισμού. Η ανάπτυξη του έργου Εξόρυξης Γνώσης πραγματοποιείται σε ένα server, ο οποίος υποστηρίζει τα Analysis Services του SQL Server. Στην συνέχεια, οι αλγόριθμοι Εξόρυξης Γνώσης φορτώνονται με το κατάλληλο σύνολο δεδομένων, για την εκπαίδευση και την δημιουργία των μοντέλων και των δομών Εξόρυξης Γνώσης. Τα μοντέλα Εξόρυξης Γνώσης που παράγονται θα τροφοδοτηθούν, στη συνέχεια, με πραγματικά δεδομένα, ανάλογα πάντα με το επιχειρησιακό πρόβλημα που επιχειρούν να επιλύσουν. Για την παρουσίαση των αποτελεσμάτων της Εξόρυξης Γνώσης χρησιμοποιούνται διάφορες αναφορές (κυρίως μέσα από τα Reporting Services του SQL Server), χρησιμοποιείται το Excel, το οποίο με την χρήση συγκεκριμένων add-ins (data mining add-ins) χρησιμοποιεί τις δυνατότητες Εξόρυξης Γνώσης του SQL Server για να εφαρμόσει τα μοντέλα εξόρυξης στα δικά του δεδομένα (σε ένα φύλλο εργασίας του excel). Τέλος, μπορούν να αναπτυχθούν συγκεκριμένες εφαρμογές στον οργανισμό (κυρίως βασισμένες στο .NET Framework της Microsoft και το ADOMD.NET[2] provider για την επικοινωνία μιας .NET εφαρμογής με τα Analysis Services του SQL Server) για να χρησιμοποιήσουν τα μοντέλα Εξόρυξης Γνώσης, που έχουν δημιουργηθεί.

Μια γρήγορη ματιά στους Αλγόριθμους

Decision Trees	Βρίσκει τα αποτελέσματα κατηγοριοποίησης που βασίζεται στις τιμές ενός συνόλου δεδομένων εκπαίδευσης
Association Rules	Αναγνωρίζει συσχετίσεις μεταξύ ομάδων αντικειμένων
Clustering	Κατηγοριοποιεί δεδομένα σε διακριτές ομάδες βάσει οποιονδήποτε χαρακτηριστικών τους
Sequence Clustering	Τοποθετεί δεδομένα σε ομάδες βάσει μιας σειράς από γεγονότα του παρελθόντος
Time Series	Αναλύει και προβλέπει γεγονότα που βασίζονται σε χρονολογική σειρά
Neural Nets	Ψάχνει για να ανακαλύψει συσχετίσεις στα δεδομένα που δεν βασίζονται στην διαίσθηση των μελετητών
Linear Regression	Προσδιορίζει την συσχέτιση μεταξύ στηλών δεδομένων σε ένα σύνολο δεδομένων για να προβλέψει ένα αποτέλεσμα
Logistic Regression	Προσδιορίζει την συσχέτιση μεταξύ στηλών με σκοπό να αποτιμήσει την πιθανότητα μια στήλη να περιέχει (να είναι σε) μια συγκεκριμένη κατάσταση

Πίνακας 1: Γρήγορη παρουσίαση αλγορίθμων Εξόρυξης Γνώσης των SSAS

Εφαρμογή των Αλγορίθμων

[2] Για περισσότερες πληροφορίες σχετικά με το ADOMD.NET, μπορείτε να επισκεφτείτε την επίσημη σελίδα της Microsoft http://msdn.microsoft.com/en-us/library/ms123483.aspx

Η επιλογή του βέλτιστου αλγόριθμου για να χρησιμοποιηθεί σε ένα επιχειρησιακό πρόβλημα μπορεί να είναι μια πρόκληση για τον αναλυτή. Ενώ ο χρήστης μπορεί να χρησιμοποιήσει διαφορετικούς αλγορίθμους για να εκτελέσει την ίδια επιχειρησιακή λειτουργία, κάθε αλγόριθμος μπορεί να παράγει ένα διαφορετικό αποτέλεσμα και κάποιοι αλγόριθμοι μπορεί να παράγουν περισσότερους από ένα διαφορετικούς τύπους αποτελεσμάτων. Για παράδειγμα, οι χρήστες μπορούν να χρησιμοποιήσουν τον αλγόριθμο Microsoft Decision Trees όχι μόνο για πρόβλεψη, αλλά και για να μειώσουν τον αριθμό των στηλών που θα χρησιμοποιηθούν σε ένα άλλο αλγόριθμο Εξόρυξης Γνώσης, γιατί ένα δέντρο απόφασης μπορεί να αναγνωρίσει τις στήλες οι οποίες δεν επηρεάζουν το τελικό μοντέλο Εξόρυξης Γνώσης.

Επίσης, οι χρήστες δεν είναι απαραίτητο να χρησιμοποιήσουν τους αλγορίθμους ανεξάρτητα. Σε μια λύση Εξόρυξης Γνώσης επιτρέπεται η χρήση διαφορετικών αλγορίθμων για την μελέτη και κατανόηση των δεδομένων και στην συνέχεια να χρησιμοποιηθούν άλλοι αλγόριθμοι για να προβλέψουν ένα συγκεκριμένο αποτέλεσμα βάσει των δεδομένων του μοντέλου. Για παράδειγμα, μπορεί να γίνει χρήση ενός μοντέλου ομαδοποίησης (clustering) για να γίνει αναγνώριση προτύπων και να ομαδοποιηθούν τα δεδομένα σε ομάδες που περιέχουν ομογενή δεδομένα. Στην συνέχεια, τα ομαδοποιημένα δεδομένα μπορούν να χρησιμοποιηθούν για να παραχθεί ένα καλύτερο μοντέλο δέντρου απόφασης.

Τα διάφορα μοντέλα Εξόρυξης Γνώσης στα SSAS μπορούν να χρησιμοποιηθούν για πρόβλεψη τιμών, για παραγωγή συγκεντρωτικών πληροφοριών από ένα σύνολο δεδομένων, για να βρουν κρυφές συσχετίσεις ανάμεσα στα δεδομένα κα. Για να βοηθηθεί ο χρήστης στην απόφαση, που πρέπει πάρει για το ποια λύση Εξόρυξης Γνώσης να χρησιμοποιήσει, παρουσιάζουμε ένα συνοπτικό πίνακα με κάποιες προτάσεις αλγορίθμων Εξόρυξης Γνώσης των SSAS που μπορούν να χρησιμοποιηθούν για διάφορα προβλήματα.

Πρόβλημα	Αλγόριθμος που χρησιμοποιείται
Πρόβλεψη διακριτής τιμής ενός χαρακτηριστικού. Πρόβλεψη εάν ο παραλήπτης μιας στοχευόμενης διαφημιστικής επιστολής θα αγοράσει ένα προϊόν.	Microsoft Decision Trees Microsoft Clustering Microsoft Neural Network
Πρόβλεψη συνεχούς τιμής ενός χαρακτηριστικού. Πρόβλεψη των πωλήσεων του επόμενου έτους.	Microsoft Decision Trees Microsoft Time Series
Πρόβλεψη μια αλληλουχίας γεγονότων. Πρόβλεψη των επιλογών ιστοσελίδων (πατήματα του ποντικιού σε URLs) σε web site μιας εταιρείας.	Microsoft Sequence Clustering
Εύρεση ομάδων κοινών προϊόντων σε μια συναλλαγή. Ανάλυση του καλαθιού αγοράς σε μια επιχείρηση βάσει των ιστορικών στοιχείων των συναλλαγών στην επιχείρηση, έτσι ώστε να γίνει πρόταση αγοράς επιπλέον προϊόντων σε κάποιον πελάτη.	Microsoft Association Microsoft Decision Trees
Εύρεση ομάδων με όμοια αντικείμενα. Ομαδοποίηση πελατών βάσει των δημογραφικών τους πληροφοριών για να γίνει καλύτερη κατανόηση της συσχέτισης μεταξύ των χαρακτηριστικών των πελατών.	Microsoft Clustering Microsoft Sequence Clustering

Πίνακας 2: Χρήση αλγορίθμων Εξόρυξης Γνώσης για διάφορα ζητήματα

	Κατηγοριοποίηση	Ομαδοποίηση	Εκτίμηση	Συσχέτιση	Πρόβλεψη
Decision Trees	✓		✓	✓	
Association Rules				✓	
Clustering		✓			
Sequence Clustering		✓			
Time Series					✓
Neural Nets	✓		✓		
Linear Regression			✓		
Logistic Regression	✓		✓		

Πίνακας 3: Αλγόριθμοι Εξόρυξης Γνώσης και Λειτουργίες που εκτελούν

Θέματα σχετικά με την αρχιτεκτονική της Εφαρμογής

Για την υλοποίηση μιας BI εφαρμογής, είναι σημαντικό να ληφθούν υπόψη ορισμένες απαιτήσεις σχετικά με την αρχιτεκτονική της εφαρμογής:

- Καθορισμός των τύπων των επιχειρησιακών προβλημάτων που πρέπει να αντιμετωπιστούν.
- Αναθεώρηση της ποιότητας των δεδομένων, που υπάρχουν στην διάθεσή μας – θα χρησιμοποιηθεί κάποια διαδικασία ETL για καθαρισμό και επικύρωση των δεδομένων, θα χρησιμοποιηθούν αθροιστικές συναρτήσεις (aggregates), θα αφαιρεθούν οι κενές (nulls) τιμές, πως θα γίνει διαχείριση των μη-φυσικών και ακραίων τιμών (outliers);.
- Ποια δεδομένα πρέπει να χρησιμοποιηθούν στο μοντέλο – επιλογή πινάκων της Βάσης Δεδομένων και στηλών από την σχεσιακή Βάση ή ορισμένων διαστάσεων από τον Κύβο, άλλοι βοηθητικοί πίνακες (nested tables), ποιες στήλες θα οριστούν ως κλειδιά, ως inputs και ως στήλες πρόβλεψης (predictable).
- Ποιοι αλγόριθμοι θα χρησιμοποιηθούν στο μοντέλο επίλυσης. Στο σημείο αυτό είναι δυνατόν να προστεθούν επιπλέον αλγόριθμοι, καθώς αναπτύσσεται το μοντέλο.
- Πώς να γίνει ο έλεγχος και η επικύρωση (validation) των αποτελεσμάτων του μοντέλου. Ποιοι αλγόριθμοι έχουν βγάλει τα πιο χρήσιμα και τα πιο ακριβή αποτελέσματα.
- Ποιοι χρήστες θα δουν τα αποτελέσματα. Επίσης, ποια εργαλεία θα χρησιμοποιηθούν για την οπτικοποίηση των αποτελεσμάτων.
- Πως θα γίνεται η συντήρηση και η επέκταση του μοντέλου. Επίσης κάθε πότε θα γίνεται εισαγωγή καινούργιων δεδομένων στο μοντέλο.
- Τέλος, κάθε πότε θα γίνεται επαναξιολόγηση των αποτελεσμάτων του μοντέλου. Ποιες μετρικές πρέπει να χρησιμοποιηθούν έτσι ώστε να δείχνουν την ακρίβεια και την χρησιμότητα του μοντέλου.

Data Mining Extension (DMX)

Η δημιουργία των μοντέλων και των δομών Εξόρυξης Γνώσης γίνεται κυρίως μέσα από το περιβάλλον Microsoft Visual Studio και συγκεκριμένα από το Business Intelligence Development Studio (BIDS) το οποίο εγκαθίσταται κάτω από το περιβάλλον του VS. Το BIDS εγκαθίσταται στον υπολογιστή του χρήστη με την εγκατάσταση του Microsoft SQL Server (εκδόσεις 2005 και 2008) και δίνει δυνατότητα στους χρήστες να δημιουργήσουν λύσεις Analysis Services, Integration Services και Reporting Services. Το περιβάλλον του BIDS θα το δούμε καλύτερα στο κεφάλαιο 5, όπου θα παρουσιάσουμε τα βήματα, που πρέπει να ακολουθήσει ένας χρήστης για να επιλύσει ένα επιχειρησιακό πρόβλημα με την χρήση των τεχνικών Εξόρυξης Γνώσης.

Εκτός από το BIDS, οι χρήστες έχουν την δυνατότητα να δημιουργήσουν μοντέλα Εξόρυξης Γνώσης και να εκτελέσουν ερωτήματα σε αυτά με την χρήση ερωτημάτων DMX (Data Mining EXtensions). Οι εντολές DMX είναι παρόμοιες με τις εντολές SQL που είναι γνωστές σχεδόν σε όλους όσους ασχολούνται με τον τομέα της διαχείρισης δεδομένων και όχι μόνο. Η εκτέλεση των ερωτημάτων DMX γίνεται κατευθείαν στον SQL Server (μέσα από ένα Analysis Services project που δημιουργείται από το Management Studio του SQL Server) και εμφανίζει τα αποτελέσματα των ερωτημάτων μέσω του Management Studio.

Στην συνέχεια θα παρουσιάσουμε κάποιες βασικές εντολές DMX που χρησιμοποιούνται για την κατασκευή μοντέλων Εξόρυξης Γνώσης και την υποβολή ερωτημάτων σε αυτά. Επίσης, στην συνέχεια του κεφαλαίου, μετά από την προβολή των βασικότερων τεχνικών Εξόρυξης Γνώσης (Κατηγοριοποίηση, Ομαδοποίηση και Κανόνες Συσχέτισης) θα παρουσιάσουμε τις κυριότερες εντολές DMX για την κάθε μία τεχνική.

Δημιουργία Μοντέλου Εξόρυξης Γνώσης (CREATE MINING MODEL)

Για την δημιουργία ενός μοντέλου Εξόρυξης Γνώσης, χρησιμοποιείται η εντολή CREATE MINING MODEL όπως φαίνεται παρακάτω:

```
CREATE MINING MODEL <name>
(
<ορισμός στηλών>
) USING <algorithm>[(<parameters>)]
[WITH DRILLTHROUGH]
```

Το όνομα του μοντέλου (<name>) δίνεται από τον χρήστη, ενώ ο αλγόριθμος <algorithm> που θα χρησιμοποιηθεί επιλέγεται με τον ορισμό της αντίστοιχης παραμέτρου. Ενδεικτικά, αναφέρουμε κάποιες τιμές για την επιλογή του αλγορίθμου Εξόρυξης Γνώσης, όπως:

- Microsoft_Decision_Trees
- Microsoft_Clustering
- Microsoft_Neural_Networks
- Microsoft_Naive_Bayes
- Microsoft_Association_Rules
- Microsoft_Logistic_Regression

Στην συνέχεια ο χρήστης πρέπει να ορίσει τις στήλες δεδομένων, οι οποίες θα χρησιμοποιηθούν από τον αλγόριθμο. Για τον ορισμό των στηλών δεδομένων, ο χρήστης

πρέπει να δώσει το *όνομα* της στήλης, τον *τύπο* των δεδομένων που περιέχονται στην στήλη δεδομένων καθώς και την χρήση της στήλης αυτής (για παράδειγμα αν θα χρησιμοποιηθεί για πρόβλεψη ή να την αγνοήσει σαν στήλη ένας συγκεκριμένος αλγόριθμος).

Κάποιες από τις διαθέσιμες επιλογές για τον τύπο των δεδομένων είναι οι:

- Text
- Long
- Double
- Boolean
- Date

Ενώ οι διαθέσιμες επιλογές για τον τύπο των περιεχομένων είναι οι:

- Key
- Key Time
- Discrete
- Continuous
- Discretized

Οι διαθέσιμες επιλογές για την χρήση των στηλών δεδομένων είναι οι:

- Predict
- Predict Only

Ένα παράδειγμα δημιουργίας ενός μοντέλου Εξόρυξης Γνώσης με την χρήση του αλγορίθμου Δέντρων Απόφασης είναι το παρακάτω:

```
CREATE MINING MODEL MyModel
(
[CustID] LONG KEY,
[Gender] TEXT  DISCRETE,
[Marital Status] TEXT DISCRETE,
[Education] TEXT DISCRETE,
[Home Ownership] TEXT DISCRETE PREDICT,
[Age] LONG CONTINUOUS,
[Income] DOUBLE CONTINUOUS
) USING Microsoft_Decision_Trees
```

Δημιουργία Μοντέλου με Εσωτερικούς Πίνακες (nested tables)

Οι εσωτερικοί πίνακες χρησιμοποιούνται για να συνδυάσουν περιεχόμενα δυο ή περισσοτέρων πινάκων δεδομένων. Για παράδειγμα, στην περίπτωση της ανάλυσης του καλαθιού αγοράς των πελατών, χρειάζεται να γίνει ανάλυση της αγοραστικής συμπεριφοράς των πελατών βάσει των συναλλαγών (των αγορών τους) που πραγματοποιούν αυτοί στην εταιρεία. Ως εκ τούτου, χρειάζεται να συνδυαστεί ο πίνακας που περιέχει τα δεδομένα των πελατών με τον πίνακα που περιέχει όλες τις συναλλαγές των πελατών για το κατάστημα. Πράγματι, γίνεται η ένωση των πινάκων πάνω στον κωδικό πελάτη (που υπάρχει και στους δύο πίνακες) για να συνδυαστούν τα αποτελέσματα. Ένα άλλο παράδειγμα είναι να συνδυαστούν τα στοιχεία των πελατών με τα προϊόντα που αγοράζουν και να εφαρμοστεί σε αυτά ένας αλγόριθμος Δέντρων Απόφασης για να πραγματοποιηθεί μια τοποθέτηση των πελατών σε συγκεκριμένες κατηγορίες.

Για τον ορισμό του εσωτερικού πίνακα, χρησιμοποιείται εντολή TABLE και δίνεται το όνομα του πίνακα που περιέχει τα δεδομένα. Στην συνέχεια από τον εσωτερικό πίνακα που έχουμε δηλώσει, καθορίζουμε την στήλη του δεδομένων, από την οποία θέλουμε να τροφοδοτήσουμε πληροφορίες το μοντέλο μας.

Ένα παράδειγμα δήλωσης ενός μοντέλου Εξόρυξης Γνώσης που θα χρησιμοποιήσει εσωτερικό πίνακα είναι το παρακάτω:

```
CREATE MINING MODEL MyModel
(
[CustID] LONG KEY,
[Gender] TEXT  DISCRETE,
[Marital Status] TEXT DISCRETE,
[Education] TEXT DISCRETE,
[Home Ownership] TEXT DISCRETE PREDICT,
[Age] LONG CONTINUOUS,
[Income] DOUBLE CONTINUOUS,
[Products] TABLE
([Product Name] TEXT KEY)
) USING Microsoft_Decision_Trees
```

Εκπαίδευση Μοντέλου (INSERT INTO)

Πριν χρησιμοποιηθεί ένα μοντέλο Εξόρυξης Γνώσης, πρέπει να εκπαιδευτεί με την εισαγωγή σε αυτό δεδομένα από την Βάση. Η εκπαίδευση του μοντέλου πραγματοποιείται είτε μέσα από το BIDS είτε χρησιμοποιώντας την εντολή DMX INSERT INTO:

```
INSERT INTO
  [MINING MODEL | MINING STRUCTURE]
<όνομα μοντέλου ή δομής>
[( <column list> )]
<προέλευση δεδομένων >
```

Τα δεδομένα μπορεί να προέρχονται από διαφορετικές πηγές (ο ορισμός <προέλευση δεδομένων>). Οι διαθέσιμες επιλογές που έχει ο χρήστης είναι οι:

- Data Query. Ένα ερώτημα σε ένα πίνακα της Βάσης Δεδομένων.
- DMX Query. Ένα άλλο ερώτημα DMX, δηλαδή δεδομένα από κάποιο άλλο μοντέλο Εξόρυξης Γνώσης.
- MDX Query. Ερώτημα σε ένα κύβο (ή σε μια διάσταση του) που έχει οριστεί στη λύση Εξόρυξης Γνώσης που δουλεύουμε.
- Stored Procedure.

Πιο συγκεκριμένα, για την εισαγωγή δεδομένων στο μοντέλο προκειμένου να εκπαιδευτεί έχουμε τις παρακάτω εντολές:

OPENQUERY

Η εντολή αυτή επιστρέφει δεδομένα από υπάρχουσα προέλευση δεδομένων, για παράδειγμα ένα πίνακα από την Βάση Δεδομένων. Η εντολή OPENQUERY χρησιμοποιεί την υπάρχουσα σύνδεση με την Βάση Δεδομένων που έχει δημιουργηθεί στην λύση Εξόρυξης

Γνώσης που δουλεύουμε. Ο χρήστης μετά από την δήλωση OPENQUERY μπορεί να δηλώσει την προέλευση δεδομένων και στην συνέχεια να γράψει ένα ερώτημα SQL για να φέρει τα δεδομένα από ένα συγκεκριμένο πίνακα της Βάσης.

Ένα παράδειγμα εκπαίδευσης του μοντέλου με την χρήση της εντολής OPENQUERY είναι το παρακάτω:

```
INSERT INTO MyModel
([CustID], [Gender],[Marital Status],
 [Education], [Home Ownership], [Age], [Income])
OPENQUERY(MyDataSource,
  'SELECT ID, Age, Gender, MaritalStatus,
   Education, HomeOwnership, Age, Income FROM MyTable')
```

Ερωτήματα σε πολλαπλές πηγές (SHAPE)

Η εντολή SHAPE χρησιμοποιείται για να δημιουργήσει ερωτήματα σε πολλαπλές πηγές δεδομένων και να δημιουργήσει ένα εσωτερικό πίνακα (nested table). Με την εντολή SHAPE ουσιαστικά μπορούμε να συνδυάσουμε δεδομένα από πολλαπλές προελεύσεις σε ένα ιεραρχικό πίνακα.

```
INSERT INTO MyModel
([CustID], [Gender],[Marital Status],
 [Education], [Home Ownership], [Age], [Income])    Products(SKIP, [Product Name])
SHAPE {OPENQUERY('MyDataSource',
  'SELECT ID, Age, Gender, MaritalStatus,
   Education, HomeOwnership, Age, Income FROM MyTable')}
APPEND
( {OPENQUERY('MyDataSource',
  'SELECT CustId, Product FROM Purchases')}
        RELATE ID TO CustID)
        AS Products
```

Πρόβλεψη (PREDICTION JOIN)

Μια από τις βασικές χρήσεις των μοντέλων Εξόρυξης Γνώσης είναι η πρόβλεψη της τιμής ενός χαρακτηριστικού. Πιο συγκεκριμένα, μπορεί να χρησιμοποιηθεί το μοντέλο Εξόρυξης Γνώσης για να προβλεφθεί η κατάσταση της στήλης δεδομένων σε μια εξωτερική προέλευση δεδομένων, για παράδειγμα η τιμή σε μια στήλη ενός πίνακα σε μια εξωτερική Βάση Δεδομένων. Η βασική εντολή για την πρόβλεψη είναι η παρακάτω:

```
SELECT [TOP <count> ]
<expression-list> FROM <model>
[
[NATURAL] PREDICTION JOIN
<source data> AS <alias>
[ ON <column-mapping> ]
[ WHERE <filter expression> ]
```

```
[ ORDER BY <expression> ]
]
```

Η εντολή SELECT επιλέγει τις στήλες δεδομένων, για τις οποίες ο αλγόριθμος θα προβλέψει τις τιμές τους. Η έννοια του SELECT στην DMX είναι ίδια με την έννοια του SELECT στην SQL, όπου καθορίζονται οι οποίες θα επιστραφούν από την προέλευση δεδομένων.

Ίδια έννοια επίσης, με την εντολή JOIN στην SQL για την ένωση δυο ή περισσότερων πινάκων έχει η εντολή της DMX PREDICTION JOIN ON η οποία συνδέει το μοντέλο Εξόρυξης Γνώσης με την προέλευση δεδομένων, η οποία περιέχει τα δεδομένα που θα χρησιμοποιηθούν στο μοντέλο. Τέλος, στην περίπτωση που το όνομα της στήλη δεδομένων στο μοντέλο Εξόρυξης Γνώσης είναι ίδιο με το όνομα της στήλης στον πίνακα δεδομένων, τότε μπορεί να χρησιμοποιηθεί η εντολή NATURAL PREDICTION JOIN και να μην χρησιμοποιηθεί η εντολή ON.

Συνοπτική Περιγραφή των αλγορίθμων του SSAS

Στην συνέχεια θα παρουσιάσουμε συνοπτικά τους αλγορίθμους Εξόρυξης Γνώσης, που παρέχονται μέσα από τα Analysis Services του SQL Server. Η παρουσίαση θα είναι συνοπτική και δείξουμε επιφανειακά κάποια βασικά χαρακτηριστικά των αλγορίθμων. Στην συνέχεια του κεφαλαίου αυτού θα περιγράψουμε με μεγαλύτερη λεπτομέρεια τους συγκεκριμένους αλγορίθμους, που αναφέραμε παραπάνω (Decision Trees, Clustering, Association Rules).

- ***Microsoft Decision Trees***: Είναι ο πιο συχνά χρησιμοποιούμενος αλγόριθμος, λόγω της ευελιξίας του, ο οποίος λειτουργεί με διακριτά και συνεχή γνωρίσματα. Ο αλγόριθμος χρησιμοποιείται για προβολή και για πρόβλεψη. Υπάρχουν διάφοροι παράμετροι, για διαμόρφωση του αλγορίθμου, με την πιο σημαντική η να είναι η COMPLEXITY_PENALTY. Με τον καθορισμό του αριθμού (συνήθως προς τα κάτω – μειώνοντάς τον) μπορεί να μειωθεί η πολυπλοκότητα του μοντέλου, μειώνοντας τον αριθμό των μεταβλητών εισόδου που πρέπει να εξεταστούν. Πρακτικά αυτό σημαίνει ότι μειώνεται το μέγεθος του δέντρου απόφασης (ο αριθμός των κόμβων στο σύνολο των αποτελεσμάτων).

  Το πιο σύνηθες εργαλείο οπτικοποίησης του αλγορίθμου είναι το Decision Tree Viewer. Επιτρέπει στον χρήστη να δει τους κόμβους που αντιπροσωπεύουν περισσότερο τις μεταβλητές, που προβλέπουν την επιλεγμένη τιμή και περιλαμβάνουν υποστηρικτική πληροφορία σχετικά με τις τιμές και τα δεδομένα που υπάρχουν σε κάθε κόμβο του δέντρου απόφασης.

- ***Microsoft Clustering***: Ο αλγόριθμος διαχωρίζει τα δεδομένα σε «έξυπνες» ομάδες, χρησιμοποιώντας μια επαναληπτική διαδικασία για να τοποθετήσει τα δεδομένα σε ομάδες (clusters) με κοινά χαρακτηριστικά. Μια σημαντική παράμετρος του αλγορίθμου είναι η CLUSTERING_METHOD, η οποία καθορίζει την μέθοδο που θα χρησιμοποιηθεί για την δημιουργία των ομάδων. Οι διαθέσιμες επιλογές είναι η κλιμακούμενη EM (Expectation Maximization), η μη-κλιμακούμενη (vanilla) EM, η κλιμακούμενη K-means και η μη-κλιμακούμενη K-means. Οι τεχνικές K-means θεωρούνται «αυστηρές» τεχνικές, με την άποψη ότι δημιουργούν τις ομάδες και στη συνέχεια κάθε περίπτωση δεδομένων (case) ανήκει ακριβώς σε μία ομάδα. Οι τεχνικές EM ακολουθούν την αντίθετη προσέγγιση και επιτρέπουν κάθε περίπτωση να ανήκει σε παραπάνω από μια ομάδα, χρησιμοποιώντας πιθανότητες. Περισσότερα για τις τεχνικές ομαδοποίησης θα δούμε στη συνέχεια.

  Αφού δημιουργηθεί το μοντέλο, είναι σημαντικό να χρησιμοποιηθούν τα διαθέσιμα εργαλεία οπτικοποίησης για καλύτερη κατανόηση των ομάδων που δημιουργούνται. Στο Cluster Diagram Viewer ο χρήστης μπορεί να μετονομάσει μια ομάδα, καθώς δουλεύει και να δει την πληροφορία για τα γνωρίσματα που σχετίζονται με κάποια ομάδα.

- ***Microsoft Association Rules****:* Ο αλγόριθμος παράγει ομάδες από συσχετιζόμενα αντικείμενα μέσα από τις στήλες των γνωρισμάτων που εισάγονται σε αυτόν. Το αποτέλεσμα του αλγορίθμου πολλές φορές αναφέρεται και σαν ανάλυση του καλαθιού της αγοράς «market basket analysis» γιατί αναλύει την συμπεριφορά αγορών και τα προϊόντα που αγοράζονται μαζί.

  Μια από τις διαθέσιμες παραμέτρους διαμόρφωσης των αποτελεσμάτων του αλγορίθμου είναι η παράμετρος MAXIMUM_ITEMSET_SIZE, η οποία καθορίζει τον μέγιστο αριθμό συσχετίσεων μεταξύ γνωρισμάτων, που ανακαλύπτει ο αλγόριθμος. Η προεπιλεγμένη τιμή είναι το 3, που σημαίνει ότι μέχρι 3 αντικείμενα που πωλούνται μαζί θα φανούν σαν αποτελέσματα στα εργαλεία οπτικοποίησης του αλγορίθμου.

  Το πιο σημαντικό εργαλείο οπτικοποίησης είναι το Rules Viewer, το οποίο δείχνει οποιοδήποτε υποσύνολο συσχετίσεων για κάθε γνώρισμα εισόδου στον αλγόριθμο μαζί με τις αντίστοιχες τιμές της πιθανότητας (probability) και της σπουδαιότητας (importance) για κάθε συσχέτιση.

Επιλογή Γνωρισμάτων στους Αλγορίθμους Εξόρυξης Γνώσης

Οι περισσότεροι αλγόριθμοι Εξόρυξης Γνώσης χρησιμοποιούν τεχνικές επιλογής γνωρισμάτων για να καθοδηγήσουν την επιλογή των πιο χρήσιμων γνωρισμάτων. Η επιλογή γνωρισμάτων χρησιμοποιείται γιατί στην ανάπτυξη του μοντέλου Εξόρυξης Γνώσης, συνήθως το σύνολο δεδομένων περιέχει περισσότερες πληροφορίες από αυτές που απαιτούνται για την κατασκευή του μοντέλου ή πολλές στήλες δεδομένων περιέχουν πληροφορίες που δεν είναι χρήσιμες στην ακρίβεια της πρόβλεψης. Αν υποθέσουμε ότι οι πόροι του υπολογιστή πλέον δεν είναι ένα πολύ σοβαρό ζήτημα (οι πόροι δηλαδή που απαιτούνται σχετικά με την ταχύτητα του επεξεργαστή, την διαθεσιμότητα της μνήμης και την ταχύτητα ανάγνωσης από τον δίσκο), στόχος μας είναι να μειώσουμε τις περιττές στήλες, γιατί μπορεί να υποβαθμίσουν την ποιότητα των προτύπων που ανακαλύπτονται. Οι λόγοι, που μπορεί να συμβαίνει αυτό παρουσιάζονται παρακάτω:

- Μερικές στήλες δεδομένων μπορεί να περιέχουν πλεονάζουσες πληροφορίες (θόρυβος) ή να είναι ελλείπεις. Αυτός ο θόρυβος (δεδομένα δηλαδή που δεν έχουν κάποιο νόημα ή κάποια χρήση) μπορεί να κάνουν δύσκολη την ανακάλυψη χρήσιμων προτύπων από τα δεδομένα.
- Για την ανακάλυψη προτύπων υψηλής ποιότητας, πολλοί αλγόριθμοι Εξόρυξης Γνώσης απαιτούν πολύ μεγάλα σύνολα δεδομένων εκπαίδευσης με πολλές διαστάσεις. Όμως, στους περισσότερους αλγορίθμους το σύνολο δεδομένων για εκπαίδευση είναι αρκετά μικρό.

Για τους παραπάνω λόγους, η τεχνική της επιλογής γνωρισμάτων βοηθάει στην επίλυση τέτοιων προβλημάτων, της ύπαρξης πολλών δεδομένων με λίγη όμως πληροφορία ή την ύπαρξη λίγων δεδομένων με μεγάλη αξία πληροφορίας.

Η βασική ιδέα πίσω από την επιλογή γνωρισμάτων είναι σχετικά απλή. Χρησιμοποιούνται στατιστικές συναρτήσεις για να υπολογιστεί η εν δυνάμει επιρροή του κάθε χαρακτηριστικού εισόδου σχετικά με την μεταβλητή που προσπαθεί να προβλέψει ο αλγόριθμος. Στην συνέχεια επιλέγεται το πλέον σημαντικό χαρακτηριστικό για να εισαχθεί στο μοντέλο. Οι τεχνικές της επιλογής χαρακτηριστικών δεν εφαρμόζονται μόνο στα χαρακτηριστικά εισόδου αλλά και στα χαρακτηριστικά εξόδου (τις μεταβλητές προς πρόβλεψη), για τις περιπτώσεις που ο αλγόριθμος πρέπει να προβλέψει ένα μεγάλο πλήθος μεταβλητών (για παράδειγμα σε ένα μοντέλο εύρεσης κανόνων συσχέτισης σε ένα «καλάθι αγορών» όπου τα αντικείμενα προς πρόβλεψη είναι τα προϊόντα και πιθανόν να είναι χιλιάδες από αυτά που προσπαθεί να προβλέψει ο αλγόριθμος). Έτσι επιλέγεται η τεχνική επιλογής χαρακτηριστικών για να επιλεχθούν πρώτα τα γνωρίσματα με τα πιο σημαντικά χαρακτηριστικά, με σκοπό την επιτάχυνση της λειτουργίας του αλγορίθμου.

Η τεχνική επιλογής χαρακτηριστικών λειτουργεί υπολογίζοντας μια βαθμολογία για κάθε χαρακτηριστικό και στη συνέχεια επιλέγει μόνο αυτά που έχουν τα καλύτερα αποτελέσματα. Η λειτουργία της επιλογής πάντοτε εκτελείται **πριν την εκπαίδευση του μοντέλου**, έτσι ώστε να επιλέγει αυτόματα τα χαρακτηριστικά από ένα σύνολο δεδομένων που είναι πολύ πιθανό να χρησιμοποιηθούν στο μοντέλο. Υπάρχουν **πολλοί τρόποι για την εφαρμογή της επιλογής χαρακτηριστικών**, ανάλογα με τον τύπο των δεδομένων που χρησιμοποιούνται και με τον αλγόριθμο που θα επιλεγεί για ανάλυση

- Η βαθμολογία ενδιαφέροντος (interestingness score) χρησιμοποιείται για την κατάταξη και να ταξινόμηση των χαρακτηριστικών των στηλών που περιέχουν μη δυαδικά συνεχή αριθμητικά δεδομένα. Η βαθμολογία ενδιαφέροντος με λίγα λόγια χρησιμοποιεί την έννοια της εντροπίας (που σχετίζεται με την ποιότητα της πληροφορίας που προσφέρει το κάθε χαρακτηριστικό).
- Για τις στήλες που περιέχουν διακριτά και διακριτοποιημένα δεδομένα, μπορούν να επιλεχθούν κάποιες μετρικές όπως η Εντροπία Shannon (που μετράει την αβεβαιότητα μιας τυχαίας μεταβλητής για ένα συγκεκριμένο αποτέλεσμα) ή μία Bayesian βαθμολογία (κάποια ακυκλικά Bayesian δίκτυα που περιγράφουν την μετάβαση των καταστάσεων που μπορεί να έχει ένα χαρακτηριστικό – η περιγραφή των μετρικών αυτών ωστόσο ξεφεύγει από τα όρια της παρούσας εργασίας). Αν το μοντέλο περιλαμβάνει στήλες με συνεχείς τιμές, η βαθμολογία ενδιαφέροντος θα πρέπει να χρησιμοποιηθεί για την αξιολόγηση όλων των στηλών εισόδου, για να διασφαλιστεί η συνέπεια.

Οι τεχνικές επιλογής χαρακτηριστικών χρησιμοποιούνται εσωτερικά από όλους τους αλγορίθμους Εξόρυξης Γνώσης της Microsoft και οι χρήστες δεν είναι αναγκασμένοι να καλούν αυτή την διαδικασία με κάποιον ρητό τρόπο. Οι διαφορετικοί αλγόριθμοι χρησιμοποιούν διαφορετικά κριτήρια επιλογής γνωρισμάτων. Συνήθως υπάρχουν διάφορες παράμετροι στους αλγόριθμους που ορίζουν τα όρια επιλογής γνωρισμάτων: η παράμετρος **Maximum Input Attributes**, η παράμετρος **Maximum Output Attributes** και η παράμετρος **Maximum States**.

- **MAXIMUM_INPUT_ATTRIBUTES**
  Αν το μοντέλο περιέχει περισσότερες στήλες από τον αριθμό που καθορίζει η παράμετρος **MAXIMUM_INPUT_ATTRIBUTES**, τότε ο αλγόριθμο αγνοεί κάθε στήλη που υπολογίζεται ότι δεν είναι ενδιαφέρουσα.
- **MAXIMUM_OUTPUT_ATTRIBUTE**
  Ομοίως, αν το μοντέλο περιέχει περισσότερες προβλεπόμενες στήλες από αυτές που καθορίζει η παράμετρος **MAXIMUM_OUTPUT_ATTRIBUTES**, τότε ο αλγόριθμος αγνοεί κάθε στήλη προς πρόβλεψη που καθορίζεται ότι δεν είναι ενδιαφέρουσα.
- **MAXIMUM_STATES**
  Αν το μοντέλο περιέχει περισσότερες περιπτώσεις (cases) από τον αριθμό που καθορίζει η παράμετρος **MAXIMUM_STATES**, τότε οι λιγότερο «δημοφιλείς» περιπτώσεις ομαδοποιούνται και θεωρούνται σαν κενές τιμές (missing).

Αν κάποια από τις παραπάνω παραμέτρους τεθεί ίση με το 0, τότε απενεργοποιείται η επιλογή χαρακτηριστικών, επηρεάζοντας τον χρόνο επεξεργασίας του μοντέλου και την επίδοση του. Στη συνέχεια του κεφαλαίου παρουσιάζουμε αναλυτικά τους αλγορίθμους Εξόρυξης Γνώσης, του SQL Server και συγκεκριμένα τους αλγορίθμους Microsoft Decision Trees, Microsoft Clustering και Microsoft Association Rules.

Microsoft Decision Trees

Γενικά Στοιχεία για τον Αλγόριθμο

Ο Microsoft Decision Trees (MDT) αλγόριθμος είναι ένας αλγόριθμος, ο οποίος χρησιμοποιεί διάφορες μεθόδους για να δημιουργήσει μια δενδροειδής δομή. Ο αλγόριθμος αυτός είναι από τους πιο εύκολους αλγορίθμους για να κατανοήσει κάποιος, γιατί δημιουργεί την δενδροειδή δομή, που αναφέραμε, κατά την διαδικασία εκπαίδευσής του. Η δενδροειδής δομή χρησιμοποιείται στην συνέχεια για να υποστηρίξει Κατηγοριοποίηση (Classification), να λειτουργήσει για πρόβλεψη (Regression) καθώς και σε μερικές περιπτώσεις για Συσχέτιση (Association). Το σχήμα και το βάθος του δέντρου εξαρτώνται, εκτός από το είδος και την ποιότητα των δεδομένων και από την τιμή συγκεκριμένων παραμέτρων του αλγορίθμου, όπως θα δούμε στη συνέχεια.

Κατασκευή του Δέντρου

Ο αλγόριθμος Microsoft Decision Trees δημιουργεί ένα σύνολο από πιθανές τιμές εισόδου, στην συνέχεια πραγματοποιεί μια μέθοδο επιλογής γνωρισμάτων για να εντοπίσει τα γνωρίσματα και τις τιμές αυτές που προσφέρουν την περισσότερη πληροφορία και για νααφαιρέσει τα γνωρίσματα, των οποίων οι τιμές είναι πολύ σπάνιες (δεν προσφέρουν πολύ πληροφορία).

Ένα δέντρο κατασκευάζεται με τον προσδιορισμό των συσχετίσεων μεταξύ των τιμών εισόδου και του γνωρίσματος προς πρόβλεψη, το αποτέλεσμα του οποίου μελετάμε. Το σημαντικό σημείο του αλγορίθμου, είναι το σημείο αυτό, δηλαδή ο τρόπος με τον οποίο γίνεται η επιλογή του κάθε γνωρίσματος, το πώς μετράμε την πληροφορία που παρέχει το κάθε γνώρισμα. Αυτό επιτυγχάνεται με την δημιουργία μιας ειδικής συνάρτησης, που ονομάζεται **information gain** (κέρδος πληροφορίας, θα μπορούσαμε να το μεταφράσουμε). Το γνώρισμα με το μεγαλύτερο «κέρδος πληροφορίας» επιλέγεται για να γίνει η διάσπαση του δέντρου και πραγματοποιείται μια αναδρομική ανάλυση στα άλλα γνωρίσματα, μέχρις ότου να μην μπορεί να διασπαστεί παραπάνω το δέντρο. Η συνάρτηση για τον υπολογισμό του information gain εξαρτάται κυρίως από την τιμή ειδικών παραμέτρων του αλγορίθμου, από τον τύπο της μεταβλητής που προσπαθούμε να προβλέψουμε και από το τύπο δεδομένων των γνωρισμάτων εισόδου του αλγορίθμου.

Με λίγα λόγια, κατά την διαδικασία δημιουργίας του μοντέλου, ο αλγόριθμος εξετάζει πόσο το κάθε γνώρισμα εισόδου στο σύνολο δεδομένων επηρεάζει το αποτέλεσμα του γνωρίσματος προς πρόβλεψη. Στη συνέχεια ο αλγόριθμος χρησιμοποιεί το γνώρισμα εισόδου με την πιο δυνατή σχέση με το προβλεπόμενο γνώρισμα για να δημιουργήσει μια σειρά από διασπάσεις (splits), που ονομάζονται κόμβοι (nodes).Στην περίπτωση που κάποιο γνώρισμα εισόδου βρεθεί ότι είναι περισσότερο σημαντικό (μεγαλύτερη σχέση ανάμεσα στο γνώρισμα προς πρόβλεψη και το γνώρισμα εισόδου), ένας καινούργιος κόμβος προστίθεται στο μοντέλο. Όσο οι καινούργιοι κόμβοι προσθέτονται μέσα στο μοντέλο, αρχίζει να διαμορφώνεται μια δενδροειδής δομή.

Διακριτές και Συνεχόμενες Τιμές

Με την χρήση του αλγορίθμου μπορούν να προβλεφθούν διακριτές καθώς και συνεχόμενες τιμές. Όταν η μεταβλητή πρόβλεψης είναι διακριτή και τα δεδομένα εισόδου είναι διακριτά, γίνεται καταμέτρηση των τιμών εξόδου (οι τιμές της μεταβλητής πρόβλεψης) σε σχέση με τις τιμές των μεταβλητών εισόδου και παράγεται ένας πίνακας (matrix) με διάφορες τιμές (scores) σε κάθε κελί βάσει τον οποίων γίνεται ο διαχωρισμός του δέντρου.

Όταν η μεταβλητή πρόβλεψης είναι διακριτή και οι τιμές εισόδου είναι συνεχόμενες τιμές, τότε οι συνεχόμενες τιμές εισόδου αυτόματα γίνονται διακριτές (discretized). Ο διαχωρισμός των συνεχόμενων τιμών γίνεται βάσει συγκεκριμένων παραμέτρων, που καθορίζουν τις ομάδες (bins) στις οποίες θα «μπαίνουν» οι κάθε τιμές των μεταβλητών εισόδου. Ο χωρισμός σε ομάδες μπορεί να γίνει αυτόματα στα Analysis Services (που είναι και η προεπιλεγμένη μέθοδος) ή με τον καθορισμό συγκεκριμένης τιμής, με την οποία καθορίζει τον διαχωρισμό ο χρήστης (ποιο συγκεκριμένα με τις επιλογές **DiscretizationMethod** και **DiscretizationBucketCount**).

Για συνεχόμενες τιμές στις μεταβλητές εισόδου και στο γνώρισμα πρόβλεψης, ο αλγόριθμος χρησιμοποιεί γραμμική παλινδρόμηση για τον καθορισμό του σημείου, στο οποίο χωρίζεται το δέντρο.

Παράμετροι Αλγορίθμου

Ο αλγόριθμος Microsoft Decision Trees παρέχει μια σειρά από παραμέτρους. Οι παράμετροι αυτοί χρησιμοποιούνται για να ελέγξουν την ανάπτυξη του δέντρου, το σχήμα του, τα γνωρίσματα των μεταβλητών εισόδου/εξόδου κα. Με την διαμόρφωση των παραμέτρων του αλγορίθμου μπορούμε να βελτιώσουμε την απόδοση και την ακρίβεια του μοντέλου, που παράγεται. Στη συνέχεια παρουσιάζουμε τις βασικές παραμέτρους του αλγορίθμου Microsoft Decision Trees.

COMPLEXITY PENALTY

Ελέγχει την ανάπτυξη του δέντρου απόφασης. Η παράμετρος αυτή είναι ένας δεκαδικός αριθμός μέσα το εύρος [0, 1] και ουσιαστικά ελέγχει το εύρος του δέντρου, καθορίζοντας ένα είδος ποινής (penalty) στην πολυπλοκότητα του δέντρου. Όταν η τιμή της παραμέτρου είναι κοντά στο 0, αυξάνεται ο αριθμός των διαχωρισμών και υπάρχει μια χαμηλή ποινή για την ανάπτυξη του δέντρου με αποτέλεσμα να παράγονται μεγάλα δέντρα. Όταν η τιμή της παραμέτρου είναι κοντά στο 1, μειώνεται ο αριθμός των διαχωρισμών και η στην αύξηση του δέντρου επιβάλλεται μια ποινή, με αποτέλεσμα τα δέντρα να είναι σχετικά μικρά. Σε γενικές γραμμές, τα μεγάλα δέντρα τείνουν να έχουν θέματα υπερβολικής εκπαίδευσης, ενώ τα μικρά δέντρα ενδέχεται να χάνουν και να μην ανακαλύπτουν κάποια πρότυπα στα δεδομένα τους.

Ο συνηθισμένος τρόπος ανάπτυξης των δέντρων χρησιμοποιώντας την παράμετρο αυτή είναι να παράγονται δέντρα με διαφορετικές ρυθμίσεις και στην συνέχεια να γίνεται επικύρωση μεταξύ των δέντρων (cross-validation) έτσι ώστε να εντοπισθούν οι ρυθμίσεις με την μεγαλύτερη και την πιο σταθερή ακρίβεια. Η προκαθορισμένη τιμή της παραμέτρου καθορίζεται από τον αριθμό των γνωρισμάτων, που εισάγονται στον αλγόριθμο. Έτσι έχουμε:

- Από 1 έως 9 γνωρίσματα, προκαθορισμένη τιμή είναι το 0.5
- Από 10 έως 99 γνωρίσματα, προκαθορισμένη τιμή είναι το 0.9
- Από 100 και πάνω γνωρίσματα, η προκαθορισμένη τιμή είναι το 0.99

MINIMUM SUPPORT

Είναι η παράμετρος που χρησιμοποιείται για να καθορίσει το ελάχιστο μέγεθος κάθε κόμβου σε ένα δέντρο, το οποία είναι ο αριθμός των εγγραφών σε ένα κόμβο που απαιτούνται για να δημιουργήσουν ένα διαχωρισμό (split) στον κόμβο αυτό. Για παράδειγμα, ένα η τιμή της παραμέτρου ισούται με 20, τότε ένας διαχωρισμός στο δέντρο μπορεί να παραχθεί εφόσον υπάρχουν τουλάχιστον 20 περιπτώσεις (cases) στον κόμβο αυτό.

Η προεπιλεγμένη τιμή για την παράμετρο MINIMUM SUPPORT είναι το 10. Συχνά, όταν το σύνολο δεδομένων που χρησιμοποιείται για εκπαίδευση περιέχει πολλές εγγραφές, τότε θα πρέπει να μεγαλώσει η τιμή της παραμέτρου για να αποφευχθεί η υπέρ εκπαίδευση του

δέντρου. Υψηλή υποστήριξη για ένα κόμβο σημαίνει λιγότερους διαχωρισμούς με αποτέλεσμα ένα πιο χαμηλό δέντρο.

SPLIT METHOD

Η παράμετρος αυτή χρησιμοποιείται για να καθορίσει το σχήμα του δέντρου, δηλαδή αν το δέντρο είναι δυαδικό ή αν είναι πυκνό (θαμνώδες). Οι μέθοδοι που χρησιμοποιούνται, βάσει της τιμής της παραμέτρου είναι:

- SPLIT METHOD = 1, σημαίνει ότι κάθε κόμβος χωρίζεται με δυαδικό τρόπο, θα έχει δηλαδή δυο κόμβους ανεξαρτήτως του πλήθους των τιμών των γνωρισμάτων του κόμβου. Για παράδειγμα, εάν υπάρχει σε ένα κόμβο το γνώρισμα Εκπαίδευση με τρείς τιμές, Λύκειο, Πανεπιστήμιο και Μεταπτυχιακό και ο διαχωρισμός του δέντρου έχει καθοριστεί να γίνει με δυαδικό τρόπο (τιμή της παραμέτρου 1) τότε ο αλγόριθμος θα χωρίσει τον κόμβο σε δύο φύλλα με κριτήρια πχ. Πανεπιστήμιο και Όχι Πανεπιστήμιο.
- SPLIT METHOD = 2, σημαίνει ότι οι κόμβοι του δέντρου θα διαχωριστούν βάσει κάθε γνωρίσματος, δηλαδή θα γίνουν τόσοι διαχωρισμοί όσες και οι τιμές των γνωρισμάτων. Έτσι, για το προηγούμενο παράδειγμα με το γνώρισμα Εκπαίδευση, θα παραχθούν 3 κόμβοι ένας για κάθε τιμή της Εκπαίδευσης.
- SPLIT METHOD = 3, σημαίνει ότι ο αλγόριθμος θα χρησιμοποιήσει και τους δύο παραπάνω τρόπους διαχωρισμού. Με λίγα λόγια, ο αλγόριθμος θα επιλέξει αυτόματα την καλύτερη μέθοδο για τον διαχωρισμό, βάσει των διαθέσιμων τιμών του κάθε γνωρίσματος. **Η τιμή αυτή (3) είναι η προεπιλεγμένη τιμή**.

MAXIMUM INPUT ATTRIBUTES

Η παράμετρος αυτή θέτει το όριο στην επιλογή γνωρισμάτων και καθορίζει τον αριθμό των γνωρισμάτων εισόδου που μπορεί να διαχειριστεί ο αλγόριθμος, προτού χρησιμοποιήσει την τεχνική επιλογής γνωρισμάτων (feature selection). Όταν τα γνωρίσματα εισόδου είναι περισσότερα από τον αριθμό της παραμέτρου, η τεχνική της επιλογής γνωρισμάτων χρησιμοποιείται για να επιλέξει τα πιο σημαντικά γνωρίσματα. Όταν η τιμή είναι 0, δεν επιλέγεται η τεχνική επιλογής γνωρισμάτων. **Η προεπιλεγμένη τιμή είναι το 255**.

MAXIMUM OUTPUT ATTRIBUTES

Η παράμετρος αυτή είναι ακόμα μια παράμετρος που θέτει ένα όριο και συγκεκριμένα καθορίζει τον αριθμό των γνωρισμάτων εξόδου που μπορεί ο αλγόριθμος να διαχειριστεί μέχρι να χρησιμοποιήσει την επιλογή γνωρισμάτων. Όταν ο αριθμός των γνωρισμάτων που προσπαθεί να προβλέψει ο αλγόριθμος είναι μεγαλύτερος από την τιμή της παραμέτρου, τότε η τεχνική της επιλογής γνωρισμάτων χρησιμοποιείται για να επιλέξει τα πιο σημαντικά γνωρίσματα. Όταν η τιμή είναι 0, η επιλογή γνωρισμάτων απενεργοποιείται ενώ **η προεπιλεγμένη τιμή είναι το 255**.

Αποφυγή Υπέρ-εκπαίδευσης

Ο αλγόριθμος Microsoft Decision Trees μεγαλώνει το δέντρο αναδρομικά. Ως αποτέλεσμα, μερικές φορές ο αλγόριθμος μπορεί να παράγει ένα σχετικά μεγάλο δέντρο. Τα μεγάλα δέντρα έχουν πολλά επίπεδα και διασπάσεις και επομένως περιέχουν πολλούς κανόνες. Ωστόσο, το μέγεθος του δέντρου δεν έχει άμεση σχέση με την ποιότητα της πρόβλεψης. Στην πραγματικότητα, όσο το δέντρο γίνεται αρκετά βαθύ, τείνει να αντιπροσωπεύει κυρίως δεδομένα που έχουν χρησιμοποιηθεί για την εκπαίδευση του, αντί να δημιουργεί γενικούς κανόνες. Αυτά τα δέντρα κάνουν τέλεια δουλειά στο να κατηγοριοποιούν τα δεδομένα εκπαίδευσης. Ωστόσο, τις περισσότερες φορές έχουν πολύ κακή ακρίβεια στην πρόβλεψη που επιχειρούν σε ένα καινούριο σύνολο δεδομένων. Αυτό το πρόβλημα ονομάζεται ***υπέρ εκπαίδευση*** ή ***υπέρ προσαρμογή***.

Υπάρχουν πολλοί τρόποι για να αντιμετωπιστεί το ζήτημα της υπέρ προσαρμογής. Μερικοί αλγόριθμοι δημιουργίας Δέντρων Απόφασης περιέχουν δύο βήματα επεξεργασίας: την «καλλιέργεια» και το «κλάδεμα». Κατά την διάρκεια της καλλιέργειας, αναπτύσσεται το δέντρο, ενώ κατά την διάρκεια του κλαδέματος κόβονται κόμβοι και διακλαδώσεις από αυτό, φτιάχνοντας έτσι κανόνες που ισχύουν γενικότερα.

Ο αλγόριθμος Microsoft Decision Trees χρησιμοποιεί μια τεχνική για την αντιμετώπιση της υπέρ προσαρμογής που ονομάζεται «κλάδεμα προς τα εμπρός» (*forward pruning*). Η ανάπτυξη του δέντρου ελέγχεται από την χρήση μιας Bayesian μετρικής, που αποτρέπει την διάσπαση σε ένα κόμβο όταν δεν υπάρχει αρκετή πληροφορία για να δικαιολογεί την διάσπαση. Αυτό ελέγχεται με την παράμετρο **COMPLEXITY PENALTY**, η οποία παίρνει τιμή από το 0 έως το 1 και την παρουσιάσαμε παραπάνω. Με λίγα λόγια, αν τεθεί υψηλή τιμή (κοντά στο 1) στην παράμετρο αυτή, τότε επιβάλλονται περισσότεροι περιορισμοί κατά την διάρκεια της ανάπτυξης του δέντρου, με αποτέλεσμα να δημιουργείται τελικά ένα μικρότερο δέντρο απόφασης. Η ανάπτυξη του δέντρου ελέγχεται επίσης και από την παράμετρο **MINIMUM SUPPORT**, η οποία αποτρέπει τους κόμβους να διασπώνται, εκτός και αν υπάρχει μια συγκεκριμένη «ποσότητα» πληροφορίας που να δικαιολογεί την διάσπαση.

Ανάλυση συσχετίσεων με τον αλγόριθμο Microsoft Decision Trees

Ένα από τα μοναδικά χαρακτηριστικά του αλγορίθμου Microsoft Decision Trees είναι ότι μπορεί να χρησιμοποιηθεί για ανάλυση συσχετίσεων. Ένα μοντέλο Εξόρυξης Γνώσης μπορεί να περιέχει ένα σύνολο από δέντρα απόφασης. Αν το μοντέλο περιέχει ένα εσωτερικό πίνακα (nested table) και ο εσωτερικός πίνακας είναι προς πρόβλεψη, τότε όλα τα κλειδιά του εσωτερικού πίνακα θεωρούνται ότι είναι χαρακτηριστικά προς πρόβλεψη. Έτσι ο αλγόριθμος Microsoft Decision Trees κατασκευάζει ένα δέντρο για κάθε εσωτερικό κλειδί.

Υπάρχον όμως και περιορισμοί στην χρήση του αλγορίθμου Microsoft Decision Trees για ανάλυση συσχετίσεων. Επειδή για κάθε χαρακτηριστικό κατασκευάζεται ένα δέντρο απόφασης, αυτό μπορεί να πάρει πολύ χρόνο και να απαιτεί πολλούς πόρους του συστήματος. Ο προκαθορισμένος αριθμός των δέντρων απόφασης που μπορούν να δημιουργηθούν για κάθε χαρακτηριστικό είναι 255. Στην περίπτωση που υπάρχουν πάνω από 255 αντικείμενα, τότε ο αλγόριθμος χρησιμοποιεί την τεχνική της επιλογής χαρακτηριστικών για να επιλέξει τα πιο σημαντικά χαρακτηριστικά.

Επεκτασιμότητα και Απόδοση

Η λειτουργία της Κατηγοριοποίησης είναι μια πολύ σημαντική στρατηγική Εξόρυξης Γνώσης. Σε γενικές γραμμές, η ποσότητα της πληροφορίας που απαιτείται για να κατηγοριοποιηθούν τα δεδομένα μεγαλώνει σε απόλυτη αναλογία με τον αριθμό των δεδομένων εισόδου στον αλγόριθμο. Αυτό περιορίζει το μέγεθος των δεδομένων που μπορούν να κατηγοριοποιηθούν. Ο αλγόριθμος Microsoft Decision Trees χρησιμοποιεί τις παρακάτω μεθόδους για να αντιμετωπίσει τέτοια προβλήματα, να αυξήσει την απόδοση και να μειώσει τους περιορισμούς της μνήμης που απαιτείται:

- Επιλογή Χαρακτηριστικών για να βελτιστοποιήσει την επιλογή των χαρακτηριστικών.
- Bayesian μετρικές για να ελέγξει την ανάπτυξη του δέντρου.
- Βελτιστοποίηση στην ομαδοποίηση μεταβλητών με συνεχείς τιμές..
- Δυναμική ομαδοποίηση των μεταβλητών εισόδου για να προσδιορίσει τις περισσότερο σημαντικές τιμές.

Ο αλγόριθμος Microsoft Decision Trees είναι γρήγορος και επεκτάσιμος και έχει σχεδιαστεί να μπορεί να χρησιμοποιεί όλη την ισχύ του υπολογιστή για να παράγει ένα συνεπές μοντέλο. Εάν οι περιορισμοί στην απόδοση είναι σημαντικοί, ο χρήστης του αλγορίθμου μπορεί να βελτιώσει τον χρόνο επεξεργασίας κατά την διάρκεια της εκπαίδευσης του μοντέλου

επιλέγοντας τις μεθόδους που θα παρουσιάσουμε στη συνέχεια. Πρέπει να τονίσουμε εδώ όμως, ότι οι χρήστες που θα επιλέξουν κάποια από τις τεχνικές αυτές θα πρέπει να προσέξουν το γεγονός ότι η μείωση των χαρακτηριστικών, που εισάγονται στο μοντέλο μπορεί να βελτιώσει τον υπολογιστικό χρόνο αλλά θα επηρεάσει το αποτέλεσμα του μοντέλου και πιθανό να το κάνει λιγότερο αντιπροσωπευτικό για το σύνολο του πληθυσμού.

Για την βελτίωση λοιπόν του χρόνο επεξεργασίας του μοντέλου Δέντρων Απόφασης κατά την διάρκεια της εκπαίδευσης του, οι χρήστες μπορούν να κάνουν τα παρακάτω:

- Αύξηση της τιμής της παραμέτρου **COMPLEXITY_PENALTY** για τον περιορισμό της ανάπτυξης του δέντρου.
- Περιορισμός του αριθμού των μεταβλητών που θα χρησιμοποιηθούν για ανάλυση συσχετίσεων με σκοπό να μειωθεί ο αριθμός των δέντρων που θα κατασκευαστούν.
- Αύξηση της τιμής της παραμέτρου **MINIMUM_SUPPORT** για την αποφυγή της υπέρ προσαρμογής του δέντρου.
- Περιορισμός του αριθμού των διακριτών τιμών για κάθε χαρακτηριστικό σε 10 ή λιγότερες. Οι χρήστες μπορούν να ομαδοποιήσουν τις διακριτές τιμές με διαφορετικό τρόπο για κάθε διαφορετικό μοντέλο.

Παλινδρόμηση

Η παλινδρόμηση είναι παρόμοια με την κατηγοριοποίηση, με την μόνη διαφορά ότι η παλινδρόμηση προβλέπει συνεχείς μεταβλητές. Παρόλο που ο βασικός στόχος ενός Δέντρου Απόφασης είναι η κατηγοριοποίηση, μπορεί επίσης να χρησιμοποιηθεί και για παλινδρόμηση. Ο αλγόριθμος Microsoft Decision Trees υποστηρίζει παλινδρόμηση. Τα μοντέλα παλινδρόμησης χρησιμοποιούν κάποιες συνεχείς μεταβλητές εισόδου (regressors) για να υπολογίσουν μια συνεχή μεταβλητή (το αποτέλεσμα της παλινδρόμησης), με ένα γραμμικό τρόπο. Για παράδειγμα, για την παρακάτω συνάρτηση,

$$OutpuY = a + b * InputX + e$$

η μεταβλητή OutputY είναι η συνεχής μεταβλητή που θέλουμε να προβλέψουμε, η συνεχής μεταβλητή InputX είναι η μεταβλητή regressor και οι συντελεστές a και b είναι αυτές που καθορίζονται από την συνάρτηση της παλινδρόμησης.

Ο αλγόριθμος Microsoft Decision Trees περιέχει μια γραμμική συνάρτηση παλινδρόμηση σε κάθε φύλο – κόμβο του δέντρου. Έτσι για κάθε περίπτωση μπορεί να χρησιμοποιεί και μια διαφορετική συνάρτηση παλινδρόμησης, ενώ αν υπάρχει περίπτωση που δεν μπορεί να βρεθεί μια λύση που να ταιριάζει στο πρόβλημα, το δέντρο απλώς προβλέπει μια σταθερή τιμή.

Εντολές DMX

Δημιουργία μιας δομής Εξόρυξης Γνώσης για την χρήση της σε μια διαδικασία ανάλυσης στοχευόμενης διαφήμισης.

```
CREATE MINING STRUCTURE TargetMail
(
   [Customer Key]   LONG KEY,
   [Age]   LONG CONTINUOUS,
   [Bike Buyer]   LONG DISCRETE,
   [Commute Distance]   TEXT DISCRETE,
```

```
   [Education]   TEXT DISCRETE,
   [Gender]   TEXT DISCRETE,
   [House Owner Flag]   LONG DISCRETE,
   [Marital Status]   TEXT DISCRETE,
   [Number Cars Owned]     LONG CONTINUOUS,
   [Number Children At Home] LONG CONTINUOUS,
   [Occupation]   TEXT DISCRETE,
   [Region]   TEXT DISCRETE,
   [Total Children]   LONG CONTINUOUS,
   [Yearly Income]   LONG DISCRETIZED
)
```

Για την επεξεργασία της δομής Εξόρυξης Γνώσης, χρησιμοποιούμε τις παρακάτω εντολές:

```
INSERT INTO TargetMail
(
   [Customer Key],   [Age], [Bike Buyer],
   [Commute Distance],   [Education],
   [Gender], [House Owner Flag],
   [Marital Status], [Number Cars Owned],
   [Number Children At Home],
   [Occupation], [Region],
   [Total Children], [Yearly Income]
)
OPENQUERY ([AdventureWorks DW],
   'SELECT [CustomerKey],[Age],[BikeBuyer],
   [CommuteDistance],[EnglishEducation],
   [Gender],[HouseOwnerFlag],
   [MaritalStatus],[NumberCarsOwned],
   [NumberChildrenAtHome],[EnglishOccupation],
   [Region],[TotalChildren],[YearlyIncome]
   FROM [AdventureWorksDW].[dbo].[vTargetMail]'
)
```

Για να εισάγουμε στην δομή ένα μοντέλο Εξόρυξης Γνώσης πρέπει να αλλάξουμε την δομή και να του εισάγουμε ένα νέο μοντέλο. Για παράδειγμα, ένα μοντέλο Εξόρυξης Γνώσης χρησιμοποιείται για να προβλέψει την αγορά ενός ποδηλάτου, καθώς επίσης να προβλέψει την ηλικία του πελάτη, χρησιμοποιώντας το εισόδημα του (income), τον αριθμό των παιδιών του (number of children) και τον αριθμό των αυτοκινήτων του (cars) χρησιμοποιώντας και την παράμετρο MINIMUM_SUPPORT. Τις μεταβλητές αριθμός αυτοκινήτων και αριθμός παιδιών θα τις χρησιμοποιήσουμε για παλινδρόμηση (REGRESSOR):

```
ALTER MINING STRUCTURE TargetMail
ADD MINING MODEL TargetMailDT
(
  [Customer Key],
  [Age]   PREDICT,
  [Bike Buyer]   PREDICT,
```

```
  [Commute Distance],  [Education],
  [Gender],  [House Owner Flag],
  [Marital Status],
  [Number Cars Owned]   REGRESSOR,
  [Number Children At Home] REGRESSOR,
  [Occupation],  [Region],
  [Total Children]   REGRESSOR,
  [Yearly Income]
)
USING Microsoft_Decision_Trees(MINIMUM_SUPPORT = 15)
```

Για να εκπαιδεύσουμε το καινούργιο μοντέλο, χρησιμοποιώντας τα δεδομένα που υπάρχουν στην δομή Εξόρυξης Γνώσης, χρησιμοποιούμε την εντολή:

```
INSERT INTO TargetMailDT
```

Για να πάρουμε τον κωδικό του πελάτη, το όνομα και την πιθανότητα να αγοράσει ένα πελάτης κάποιο προϊόν χρησιμοποιούμε τις παρακάτω εντολές. Η επιλογή των πελατών γίνεται μέσα από ένα σύνολο πελατών, οι οποίοι έχουν πιθανότητα 40% και παραπάνω να αγοράσουν ένα προϊόν από την εταιρεία:

```
SELECT t.CustomerKey, t.FirstName, t.LastName,
   PredictProbability([Bike Buyer],1) as ProbBuy
FROM TargetMailDT
PREDICTION JOIN
OPENQUERY([Adventure Works DW],'SELECT * FROM vTargetMail') AS t
ON
  t.Age = TargetMailDT.Age AND
  t.CommuteDistance = TargetMailDT.[Commute Distance] AND
  t.Gender = TargetMailDT.Gender AND
  t.HouseOwnerFlag = TargetMailDT.[House Owner Flag] AND
  t.MaritalStatus = TargetMailDT.[Marital Status] AND
  t.NumberCarsOwned = TargetMailDT.[Number Cars Owned] AND
  t.NumberChildrenAtHome = TargetMailDT.[Number Children At Home] AND
  t.EnglishOccupation = TargetMailDT.Occupation AND
  t.Region = TargetMailDT.Region AND
  t.TotalChildren = TargetMailDT.[Total Children] AND
  t.YearlyIncome = TargetMailDT.[Yearly Income]
WHERE PredictProbability([Bike Buyer],1) > 0.4
```

Για να πάρουμε την προβλεπόμενη τιμή, την πιθανότητα να αγοράσει ένας πελάτης ένα προϊόν, δίνοντας σαν είσοδο τιμές κάποιων πραγματικών χαρακτηριστικών του πελάτη (για παράδειγμα χαρακτηριστικά ενός νέου πελάτη για να δούμε την συμπεριφορά του), χρησιμοποιούμε τις παρακάτω εντολές:

```
SELECT [Bike Buyer],
 PredictProbability([Bike Buyer],1),
 FROM TargetMailDT
```

```
NATURAL PREDICTION JOIN
(SELECT 25 AS Age,
  '5-10 Miles' AS [Commute Distance],
  'M' AS Gender,
   1 AS [House Owner Flag],
   'S' AS [Marital Status],
   1 AS [Number Cars Owned],
   0 AS [Number Children At Home],
   'Manual' AS Occupation,
   'Pacific' AS Region,
   0 AS [Total Children],
   45000 AS [Yearly Income]
   ) AS t
```

Microsoft Clustering

Γενικά για τον Αλγόριθμο

Ο αλγόριθμος Microsoft Clustering βρίσκει «φυσικές» ομαδοποιήσεις μέσα σε ένα πλήθος δεδομένων, για τα οποία ο εντοπισμός μιας ομάδας δεν είναι προφανής. Με λίγα λόγια ο αλγόριθμος προσπαθεί να βρει την «κρυμμένη» μεταβλητή, η οποία ομαδοποιεί τα δεδομένα. Η ομαδοποίηση μπορεί να χρησιμοποιηθεί για την τοποθέτηση των πελατών του οργανισμού σε συγκεκριμένες ομάδες με στόχο την καλύτερη «στόχευση» τους, όσον αφορά καμπάνιες μάρκετινγκ ή για την καλύτερη κατανόηση συγκεκριμένων αναγκών των πελατών. Καταλαβαίνουμε λοιπόν, ότι η δυνατότητα ενός οργανισμού να μπορεί να τοποθετήσει τους πελάτες του (και όχι μόνο) σε συγκεκριμένες ομάδες, τις οποίες δεν μπορεί με την πρώτη ματιά κάποιος να τις ανακαλύψει και να αντιμετωπίσει διαφορετικά τα προβλήματα που σχετίζονται με την κάθε ομάδα, είναι ένα δυνατό εργαλείο που διευκολύνει την καθημερινή λειτουργία του οργανισμού.

Πως λειτουργεί ο Αλγόριθμος

Ο αλγόριθμος Microsoft Clustering αρχικά προσδιορίζει τις συσχετίσεις στο σύνολο δεδομένων και δημιουργεί μια σειρά ομάδων, βάσει αυτών των δεδομένων. Ένα διάγραμμα scatter (scatter plot) είναι ένας χρήσιμος τρόπος για την οπτική αναπαράσταση του τρόπου, με τον οποίο ο αλγόριθμος ομαδοποιεί τα δεδομένα. Αρχικά το scatter plot αναπαριστά όλα τα δεδομένα και κάθε μέλος του συνόλου δεδομένων είναι ένα σημείο στο διάγραμμα. Στη συνέχεια, οι ομάδες που εντοπίζονται συγκεντρώνουν τα σημεία στο γράφημα και αναπαριστούν τις συσχετίσεις που ανακαλύπτει ο αλγόριθμος.

Μετά τον αρχικό προσδιορισμό των ομάδων, ο αλγόριθμος υπολογίζει πόσο καλά οι ομάδες αναπαριστούν τα σύνολα των δεδομένων κα προσπαθεί να βελτιώσει τις ομάδες όπου χρειάζεται έτσι ώστε να σχηματιστούν ομαδοποιήσεις που αναπαριστούν καλύτερα τα δεδομένα. Η διαδικασία της βελτίωσης των ομάδων είναι μια επαναληπτική διαδικασία η οποία τερματίζει όταν δεν μπορεί να γίνει καμία επιπλέον βελτίωση των ομάδων, που αναπαριστούν τα δεδομένα.

Χρήση του Αλγορίθμου

Η πιο κοινή χρήση του αλγορίθμου Microsoft Clustering είναι απλά η ομαδοποίηση, δηλαδή η εύρεση ομάδων στα δεδομένα και στη συνέχεια να επισημανθούν τα δεδομένα που ανήκουν σε κάθε ομάδα με ένα νέο χαρακτηριστικό, που υποδηλώνει την ομάδα που ανήκει η κάθε εγγραφή. Αφού επισημανθούν τα δεδομένα, στη συνέχεια μπορούν να χρησιμοποιηθούν για την υποβολή αναφορών και για ανάλυση των ομάδων που ανήκει η κάθε εγγραφή, όπως θα γινόταν και με κάθε άλλο χαρακτηριστικό των δεδομένων. Αυτή η διαδικασία είναι τόσο απλή, ώστε πολλά προϊόντα Εξόρυξης Γνώσης την εκτελούν αυτόματα σε ένα βήμα, χωρίς περαιτέρω παρέμβαση του χρήστη. Για παράδειγμα, η επιλογή Detect Categories στα SSAS, εκτελεί μια διαδικασία ομαδοποίησης στα δεδομένα.

Σε πολλές περιπτώσεις, απλώς απαιτείται η δημιουργία ενός μοντέλου ομαδοποίησης στα δεδομένα, ο έλεγχος του μοντέλου ομαδοποίησης για να καθοριστεί η έννοια της κάθε ομάδας και στη συνέχεια ο καθορισμός ετικετών (λέξεις κλειδιά) στα δεδομένα που ανήκουν σε κάθε ομάδα δεδομένων. Υπάρχουν όμως περιπτώσεις που τα πράγματα δεν είναι τόσο απλά, γιατί μπορεί το σύνολο δεδομένων να αποτελείται αρκετά εκατομμύρια εγγραφές και η κάθε γραμμή να περιλαμβάνει πολύπλοκούς τύπους δεδομένων (αριθμητικές και κατηγορηματικές τιμές). Σε αυτή την περίπτωση, η εκπαίδευση του μοντέλου ομαδοποίησης μπορεί να είναι αρκετά «ακριβό» σε χρόνο, λόγω των υπολογισμών και των επαναλήψεων που πρέπει να γίνουν για την ανάλυση των δεδομένων. Στην περίπτωση αυτή, μπορεί απλώς να απαιτείται να

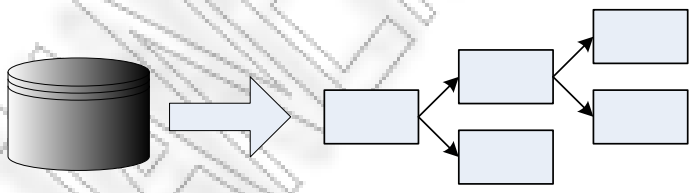
εκπαιδευτεί ένα μικρό δείγμα των δεδομένων και στην συνέχεια να εφαρμοστούν τα αποτελέσματα της ομαδοποίησης σε ένα μεγαλύτερο σύνολο δεδομένων.

Τα μοντέλα Ομαδοποίησης είναι πολύ χρήσιμα μοντέλα στο να «πετάξει» κάποιος τα δεδομένα σε αυτά και να δει τι αποτελέσματα θα προκύψουν. Ωστόσο, όπως συμβαίνει και με όλες τις τεχνικές Εξόρυξης Γνώσης, θα λάβουμε τις καλύτερες απαντήσεις όταν θέτουμε τις ερωτήσεις μας με τον σωστό τρόπο. Για παράδειγμα θέλουμε να ομαδοποιήσουμε τους πωλητές μας, βάσει των συνολικών πωλήσεων ή βάσει της αναλογίας των πωλήσεων ανά κάθε κατηγορία προϊόντων; Είναι επίσης σημαντικό να δέχεται το μοντέλο ομαδοποίησης το εισόδημα των πελατών σαν συνεχή τιμή ή πρέπει να το αναλύσουμε πρώτα σε συγκεκριμένες κατηγορίες; Ο αλγόριθμος Microsoft Clustering είναι πολύ ευέλικτος στο να υποστηρίξει όλους τις πιθανούς τύπους δεδομένων, ωστόσο θα πρέπει οι χρήστες να παρέχουν τα δεδομένα σε τέτοια μορφή, ώστε να είναι τα πλέον ενδιαφέροντα και κατάλληλα για την επίλυση του προβλήματος.

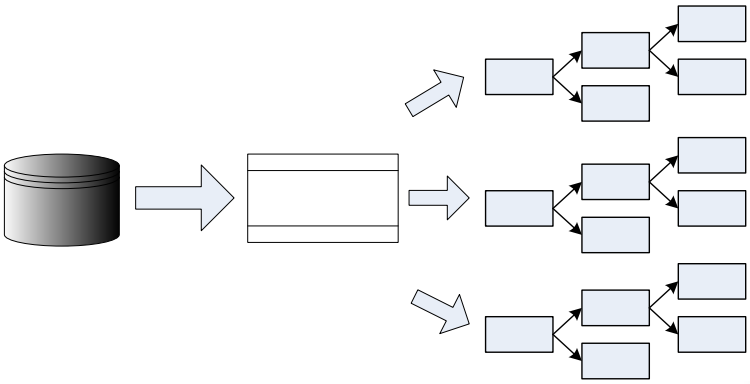
Ο αλγόριθμος Microsoft Clustering έχει μια ιδιαίτερη συμπεριφορά που συνδέεται με τον ορισμό χρήσης της κάθε στήλης δεδομένων στο μοντέλο εξόρυξης. Όταν η χρήση της στήλης δεδομένων δηλωθεί σαν **Είσοδος** (**Input**) ή σαν **Πρόβλεψη** (**Predict**), ο αλγόριθμος λειτουργεί όπως ακριβώς δηλώνεται – με την μόνη διαφορά ότι οι στήλες προς πρόβλεψη επιλέγονται από το μοντέλο αυτόματα, κατά την διάρκεια της πρόβλεψης ενώ οι στήλες εισόδου δεν επιλέγονται αυτόματα. Όταν η στήλη δηλωθεί σαν **Πρόβλεψη Μόνο** (**PredictOnly**) η στήλη αυτή έχει ειδική μεταχείριση. Οι στήλες PredictOnly στον αλγόριθμο ομαδοποίησης δεν χρησιμοποιούνται κατά την φάση της ομαδοποίησης στην εκπαίδευση του μοντέλου. Όταν το μοντέλο εκπαιδευτεί πλήρως, ο αλγόριθμος πραγματοποιεί ένα ακόμα πέρασμα στα δεδομένα εκπαίδευσης και εκχωρεί τις τιμές των ιδιοτήτων με βάση τον τρόπο με τον οποίο τα δεδομένα εκπαίδευσης ανήκουν σε κάθε ομάδα δεδομένων (clusters).

Ομαδοποίηση σαν ένα Αναλυτικό Βήμα

Η τεχνική της ομαδοποίησης συχνά χρησιμοποιείται σαν ένα ενδιάμεσο βήμα σε ένα μεγαλύτερο έργο ανάλυσης. Με την ομαδοποίηση όμοιων δεδομένων, οι χρήστες μπορούν να δημιουργήσουν καλύτερα συμπληρωματικά μοντέλα που απαντούν σε βαθύτερα ερωτήματα. Με την ανάλυση των δεδομένων βάσει των προτύπων των ομάδων, οι χρήστες μπορούν να επικεντρώσουν την προσοχή τους σε συγκεκριμένα προβλήματα. Για παράδειγμα, μπορεί να απαιτείται η δημιουργία ενός μοντέλο Δέντρων Απόφασης στο σύνολο των πελατών ενός οργανισμού για να προβλεφθεί εάν θα επαναλάβει ο πελάτης τις αγορές του τον επόμενο μήνα. Αντί να δημιουργηθεί το μοντέλο αυτό σε όλο το γενικευμένο σύνολο των πελατών, μπορεί να δημιουργηθεί αρχικά ένα μοντέλο ομαδοποίησης που διαχωρίζει τους πελάτες σε ομάδες βάσει των χαρακτηριστικών τους. Στην συνέχεια μπορεί να εφαρμοστεί το μοντέλο των Δέντρων Απόφασης σε κάποιες από τις ομάδες πελατών, που εμφανίζουν ενδιαφέροντα χαρακτηριστικά, ανάλογα με την ομάδα στην οποία ανήκουν.



Διάγραμμα 14: Κατηγοριοποίηση σε όλα τα Δεδομένα

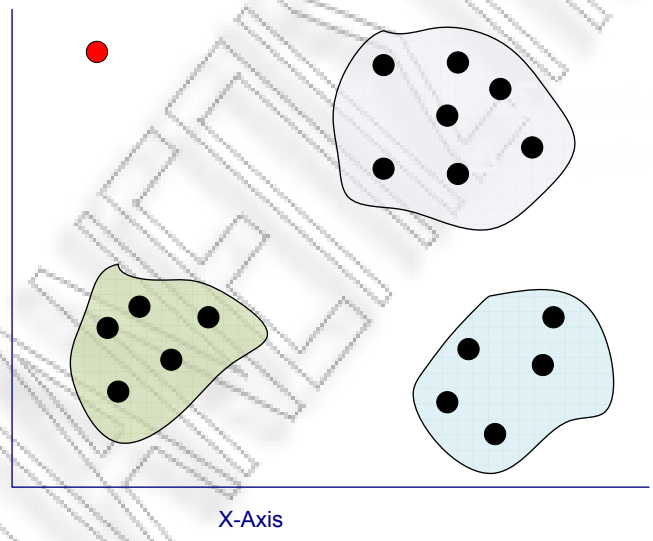


Διάγραμμα 15: Κατηγοριοποίηση στα αποτελέσματα της Ομαδοποίησης

Εύρεση Ανωμαλιών με την χρήση Ομαδοποίησης

Η ομαδοποίηση είναι ενδιαφέρουσα στο ότι δημιουργεί «ωραίες» ομάδες δεδομένων και στην συνέχεια επιτρέπει στους χρήστες να ρωτήσουν σε ποια ομάδα ανήκουν τα νέα δεδομένα που έχουν στην διάθεσή τους. Όταν τίθεται ένα ερώτημα σε ένα μοντέλο ομαδοποίησης, το μοντέλο προσδιορίζει μια πιθανότητα για την κάθε περίπτωση να ανήκει σε μια ομάδα. Τι συμβαίνει όμως όταν μια περίπτωση δεδομένων δεν ανήκει πραγματικά σε καμία από τις ομάδες που ανακαλύπτονται; Με την χρήση των χαρακτηριστικών του μοντέλο ομαδοποίησης, οι χρήστες μπορούν να προσδιορίσουν όχι μόνο που βρίσκεται η κάθε ομάδα δεδομένων αλλά και το που **δεν βρίσκονται** οι ομάδες, δηλαδή τα δεδομένα τα οποία δεν ανήκουν και δεν περικλείονται από καμία ομάδα. Οι χρήστες μπορούν να χρησιμοποιήσουν αυτή τη πληροφορία για να ανιχνεύσουν ανωμαλίες στα δεδομένα ή άλλων μορφών κακής ποιότητας δεδομένων.

Η εύρεση ανωμαλιών με την χρήση ομαδοποίησης επιτρέπει στους χρήστες να αναλύσουν ταυτόχρονα πολλαπλές διαστάσεις των δεδομένων τους για να δουν κατά πόσο ο συνδυασμός των τιμών αυτών μπορεί να ταιριάζει σε μια ομάδα δεδομένων.



Διάγραμμα 16: Εύρεση μη ομαλής Τιμής

Μέθοδοι αλγορίθμου Microsoft Clustering

Ο αλγόριθμος Microsoft Clustering παρέχει δυο μεθόδους για την δημιουργία των ομάδων και την ανάθεση σε αυτών των δεδομένων του συνόλου δεδομένων που δέχεται σαν είσοδο. Ο πρώτος αλγόριθμος είναι ο **K-means αλγόριθμος**, σύμφωνα με τον οποίο κάθε σημείο ανήκει σε ένα μόνο cluster και μια μοναδική πιθανότητα υπολογίζεται για την τοποθέτηση κάθε σημείο σε μια ομάδα. Ο δεύτερος αλγόριθμος είναι ο **EM (Expectation Maximization)**, σύμφωνα με τον οποίο ένα σημείο μπορεί να ανήκει περισσότερες από μια ομάδες και μια πιθανότητα υπολογίζεται για κάθε συνδυασμό σημείου και ομάδας (δηλαδή πόσο πιθανόν είναι να ανήκει κάθε σημείο σε κάθε ομάδα). Όπως θα δούμε στη συνέχεια, η επιλογή κάθε μεθόδου γίνεται με τον καθορισμό της παραμέτρου *CLUSTERING_METHOD* και η προεπιλεγμένη μέθοδος είναι η EM.

EM Ομαδοποίηση

Στην EM τεχνική ομαδοποίησης (EM clustering), ο αλγόριθμος πραγματοποιεί μια επαναληπτική διαδικασία για να προσδιορίσει ένα αρχικό μοντέλο ομάδων (clusters) και στη συνέχεια να τοποθετήσει τα δεδομένα στις ομάδες αυτές, καθορίζοντας την πιθανότητα κάθε σημείο να ανήκει σε μια ομάδα. Εάν έχουν βρεθεί κενές ομάδες ή εάν το μέγεθος της κάθε ομάδες (πόσα στοιχεία πρέπει να περιέχει κάθε ομάδα) είναι κάτω από ένα συγκεκριμένο όριο, τότε κ ομάδα αυτή θεωρείται σαν σημείο και ο αλγόριθμος «τρέχει» ξανά με την ομάδα αυτή σαν στοιχείο εισόδου. Ο αλγόριθμος τερματίζει όταν το μοντέλο που παράγεται ταιριάζει (fits) με τα δεδομένα που έχουν δοθεί.

Το αποτέλεσμα της μεθόδου ομαδοποίησης EM βασίζεται σε πιθανότητες (probabilistic) με την έννοια ότι κάθε σημείο ανήκει σε όλες τις ομάδες που έχουν βρεθεί αλλά ο κάθε προσδιορισμός του σημείου σε μια ομάδα σχετίζεται με μια συγκεκριμένη πιθανότητα. Η τεχνική ομαδοποίησης EM είναι η προκαθορισμένη τεχνική που χρησιμοποιείται στον αλγόριθμο Microsoft Clustering γιατί η τεχνική αυτή προσφέρει περισσότερα πλεονεκτήματα σε σχέση με την τεχνική ομαδοποίησης K-means, όπως:

- Απαιτεί ένα μόνο πέρασμα (ανάγνωση) του συνόλου δεδομένων
- Λειτουργεί με μικρότερες απαιτήσεις σε μνήμη του υπολογιστή (RAM)

Η υλοποίηση της τεχνικής ομαδοποίησης EM από την Microsoft παρέχει δυο επιλογές: την κλιμακούμενη (scalable) και την μη-κλιμακούμενη. Η προεπιλογή (κλιμακούμενη EM τεχνική) χρησιμοποιεί τις πρώτες 50.000 εγγραφές για την πρώτη ανάγνωση. Αν αυτό είναι επιτυχές για την δημιουργία των ομάδων, το μοντέλο χρησιμοποιεί αυτά τα δεδομένα. Αν το μοντέλο δεν ταιριάζει με τα δεδομένα, οι επόμενες 50.000 εγγραφές χρησιμοποιούνται. Στην μη-κλιμακούμενη τεχνική EM, ολόκληρο το σύνολο δεδομένων διαβάζεται ανεξαρτήτως μεγέθους. Αυτή η μέθοδος μπορεί να δημιουργεί περισσότερο ακριβείς ομάδες αλλά έχει μεγαλύτερες απαιτήσεις σε μνήμη ενώ η κλιμακούμενη τεχνική λειτουργεί με την χρήση ενός τοπικού buffer με αποτέλεσμα να χρησιμοποιεί καλύτερα την υπολογιστική ισχύ του επεξεργαστή του μηχανήματος. Αυτό μπορεί να έχει σαν αποτέλεσμα σε πολλές περιπτώσεις να είναι έως και τρείς φορές πιο γρήγορη η κλιμακούμενη τεχνική, ιδίως στην περίπτωση που τα δεδομένα μπορούν να χωρέσουν στην κύρια μνήμη του υπολογιστή.

K-Means Ομαδοποίηση

Η τεχνική ομαδοποίησης K-means είναι μια γνωστή τεχνική η οποία τοποθετεί ένα σημείο σε μια ομάδας ελαχιστοποιώντας τις διαφορές των σημείων της ίδιας ομάδας και μεγιστοποιώντας την απόσταση μεταξύ των ομάδων. Ο όρος means (μέσος) στο k-means αναφέρεται στο μέσο κάθε ομάδας (centroid), το οποίο είναι ένα σημείο που αρχικά έχει επιλεγεί αυθαίρετα σαν το κέντρο της ομάδας και στην συνέχεια βελτιώνεται (με μια επαναληπτική διαδικασία) έτσι ώστε να αποτελεί το πραγματικό μέσο όλων των στοιχείων της ομάδας. Ο όρος k αναφέρεται σε ένα αριθμό σημείων, που έχει χρησιμοποιηθεί στην εκκίνηση της μεθόδου. Η

τεχνική k-means υπολογίζει το τετράγωνο της Ευκλείδειας απόστασης μεταξύ κάθε σημείου στην ομάδα και την ευθεία γραμμή που αναπαριστά το μέσο της ομάδας και τοποθετεί το κάθε σημείο σε μια ομάδα όταν η απόσταση φτάσει σε ένα καθορισμένο όριο. Η διαδικασία αυτή συγκλίνει σε ένα τελικό σύνολο από k ομάδες.

Η τεχνική k-means τοποθετεί κάθε σημείο σε ακριβώς μια ομάδα και δεν επιτρέπει την αβεβαιότητα της συσχέτισης αυτής. Η τοποθέτηση ενός σημείου σε μια ομάδα εκφράζεται σαν την απόσταση τους από το μέσο (centroid) της ομάδας. Συνήθως η τεχνική k-means χρησιμοποιείται για την δημιουργία ομάδας σε συνεχείς μεταβλητές, όπου ο υπολογισμός της απόστασης από το μέσο είναι ξεκάθαρος. Ωστόσο, η υλοποίηση της τεχνικής k-means από την Microsoft επιτρέπει την χρήση της τεχνικής k-means και σε διακριτές μεταβλητές, με την χρήση πιθανοτήτων, όπου η απόσταση υπολογίζεται σαν την πιθανότητα P να ανήκει ένα σημείο στην ομάδα.

Ομοίως με την τεχνική EM, η τεχνική k-means παρέχει δυο επιλογές: την μη-κλιμακούμενη τεχνική όπου ολόκληρο το σύνολο δεδομένων διαβάζεται και εκτελεί ένα πέρασμα διαβάζοντας τα δεδομένα και την κλιμακούμενη τεχνική όπου αρχικά διαβάζονται οι πρώτες 50.000 εγγραφές και στην συνέχεια περισσότερες, εάν απαιτείται, έτσι ώστε να ταιριάζει καλύτερα η μέθοδος στα δεδομένα.

Δεδομένα που απαιτούνται για τον Αλγόριθμο

Όταν οι χρήστες προετοιμάζουν τα δεδομένα για να τα χρησιμοποιήσουν για την εκπαίδευση ενός μοντέλου ομαδοποίησης, πρέπει να έχουν καταλάβει πρώτα τις απαιτήσεις του συγκεκριμένου αλγορίθμου, συμπεριλαμβανομένου της ποσότητας των δεδομένων που απαιτούνται και πώς θα χρησιμοποιηθούν αυτά τα δεδομένα.

Οι απαιτήσεις για ένα μοντέλο ομαδοποίησης είναι οι παρακάτω:

- **Μία μοναδική στήλη που να περιέχει το κλειδί.** Κάθε μοντέλο πρέπει να περιέχει μια στήλη με αριθμητικά δεδομένα ή κείμενο τα οποία προσδιορίζουν μοναδικά κάθε εγγραφή. Τα σύνθετα κλειδιά δεν επιτρέπονται.
- **Στήλες με δεδομένα Εισόδου.** Κάθε μοντέλο πρέπει να περιέχει τουλάχιστον μια στήλη με δεδομένα εισόδου, τα οποία θα χρησιμοποιηθούν για να κατασκευάσουν τις ομάδες (clusters). Μπορούν να χρησιμοποιηθούν όσες στήλες δεδομένων εισόδου θέλουν οι χρήστες, αλλά ανάλογα με τον αριθμό των τιμών που υπάρχουν σε κάθε στήλη, οι επιπλέον στήλες εισόδου μπορούν να αυξήσουν τον χρόνο που απαιτείται για να κατασκευαστεί το μοντέλο.
- **Προαιρετικές στήλες με δεδομένα πρόβλεψης.** Ο αλγόριθμος δεν είναι απαραίτητο να περιέχει στήλες με δεδομένα που πρέπει να προβλέψει, ώστε να κατασκευάζει το μοντέλο. Ωστόσο οι χρήστες μπορούν να εισάγουν μια τέτοια στήλη με δεδομένα, σχεδόν οποιουδήποτε τύπου. Τα δεδομένα μια τέτοιας στήλης μπορούν να χρησιμοποιηθούν και σαν δεδομένα εισόδου στο μοντέλο ή μπορούν να καθορίσουν οι χρήστες ότι τα δεδομένα αυτά θα χρησιμοποιηθούν μόνο για πρόβλεψη. Για παράδειγμα, εάν οι χρήστες θέλουν να προβλέψουν το εισόδημα ενός πελάτη με την χρήση της ομαδοποίησης πάνω στα δημογραφικά χαρακτηριστικά του, όπως η περιοχή που διαμένει ή η ηλικία του, τότε στο μοντέλο μπορεί να δηλωθεί το εισόδημα σαν **PredictOnly** και να εισαχθούν οι άλλες στήλες, όπως η περιοχή και ηλικία σαν είσοδος στο μοντέλο.

Παράμετροι Αλγορίθμου

Ο χρήστης μπορεί να καθορίσει την συμπεριφορά του αλγορίθμου Microsoft Clustering με την τροποποίηση των διαφόρων παραμέτρων του. Οι προεπιλεγμένες τιμές των παραμέτρων μπορούν να διαχειριστούν την πλειονότητα των περιπτώσεων, ωστόσο υπάρχουν

περιπτώσεις που τα αποτελέσματα του αλγορίθμου μπορεί να είναι καλύτερα με την τροποποίηση των παραμέτρων του αλγορίθμου, όπως παρουσιάζονται παρακάτω:

CLUSTERING METHOD

Η παράμετρος αυτή καθορίζει την μέθοδο που θα χρησιμοποιηθεί για την δημιουργία των ομάδων (clusters) και τον ορισμό κάθε στοιχείου στην κάθε ομάδα. Η «εκδόσεις» vanilla των μεθόδων, που παρουσιάζονται στην συνέχεια αναφέρονται στις μη-κλιμακούμενες μεθόδους που εφαρμόζονται σε όλο το πλήθος δεν δεδομένων. Όπως αναφέραμε παραπάνω, οι κλιμακούμενες μέθοδοι εφαρμόζονται σε ένα συγκεκριμένο δείγμα των δεδομένων και αν δεν βρεθεί αποτέλεσμα, χρησιμοποιείται ένα επιπλέον πλήθος δοκιμαστικών δεδομένων. Οι διαθέσιμες τιμές της παραμέτρου παρουσιάζεται παρακάτω:

- Τιμή 1, κλιμακούμενη τεχνική ομαδοποίησης EM (που είναι και η **προεπιλεγμένη τιμή**).
- Τιμή 2, μη κλιμακούμενη (vanilla) τεχνική ομαδοποίησης EM
- Τιμή 3, κλιμακούμενη τεχνική ομαδοποίησης K-means.
- Τιμή 4, μη κλιμακούμενη (vanilla) τεχνική ομαδοποίησης K-means.

CLUSTER COUNT

Η παράμετρος αυτή καθορίζει στον αλγόριθμο τον αριθμό των ομάδων που πρέπει να ανακαλύψει. Συνήθως, ο αριθμός των ομάδων καθορίζεται βάσει του επιχειρηματικού προβλήματος, που προσπαθεί να αντιμετωπίσει ο αλγόριθμος και βγάζει κάποιο νόημα. Πρακτικά όσο πιο πολλά γνωρίσματα υπάρχουν τόσο περισσότερες ομάδες χρειάζονται για να περιγράψουν σωστά τα δεδομένα. Σε μερικές περιπτώσεις, αν υπάρχει ένας αρκετά μεγάλος αριθμός γνωρισμάτων, μπορούμε να οργανώσουμε έτσι τα δεδομένα έτσι ώστε ο αριθμός να μειωθεί, για παράδειγμα ομαδοποίηση βάσει κάποιου κοινού γνωρίσματος των δεδομένων.

Εάν η τιμή της παραμέτρου είναι 0, τότε θα ωθήσει τον αλγόριθμο να χρησιμοποιήσει μια ευρεστική μέθοδο για να μαντέψει τον αριθμό των ομάδων στα δεδομένα. Η ευρεστική μέθοδος δημιουργεί πολλά μικρά μοντέλα με διαφορετικό αριθμό ομάδων βάσει του συνόλου των δεδομένων και παράγει μια μετρική για κάθε ομάδα η οποία καθορίζει κατά πόσο καλά ο κάθε αριθμός των ομάδων αναπαριστά τα δεδομένα. Στην συνέχεια μια καμπύλη γραμμή εφαρμόζεται στα αποτελέσματα για να βρεθεί ένας αποτελεσματικός αριθμός ομάδων. **Η προεπιλεγμένη τιμή του αλγορίθμου είναι το 10**, δηλαδή ο αλγόριθμος πρέπει να εντοπίσει 10 ομάδες.

CLUSTER SEED

Η παράμετρος αυτή είναι ένας αριθμός που χρησιμοποιείται σαν είσοδος (seed) για την παραγωγή τυχαίου αριθμού για την αρχικοποίηση των ομάδων. Η παράμετρος αυτή επιτρέπει να ελέγξουμε την ευαισθησία των δεδομένων προς το σημείο αρχικοποίησης του αλγορίθμου. Εάν το μοντέλο του αλγορίθμου είναι σχετικά σταθερό όσο αλλάζει ο αριθμός αυτός, τότε μπορούμε να πούμε ότι η ομαδοποίηση στα δεδομένα είναι σωστή. Με την αλλαγή της τιμής του παραμέτρου αλλάζει και ο τρόπος με τον οποίο οι αρχικές ομάδες δημιουργούνται. Έτσι μπορούμε να συγκρίνουμε τα διάφορα μοντέλα που έχουν δημιουργηθεί χρησιμοποιώντας διαφορετικές τιμές εισόδου (η τιμή της παραμέτρου). Εάν η αρχική τιμή αλλάζει αλλά οι ομάδες που ανακαλύπτονται δεν διαφέρουν σημαντικά, μπορούμε να θεωρήσουμε ότι το μοντέλο είναι σχετικά σταθερό.

MINIMUM CLUSTER CASES

Η παράμετρος αυτή ελέγχει το πότε μια ομάδα θεωρείται κενή και πρέπει να διαγραφεί και να δημιουργηθεί ξανά. Καθορίζει το πλήθος δηλαδή των εγγραφών που πρέπει να περιέχει

μια ομάδα έτσι ώστε να μην θεωρείται κενή. Τυπικά, δεν είμαστε υποχρεωμένοι να αλλάζουμε την τιμή αυτής της παραμέτρου εκτός από συγκεκριμένες περιπτώσεις που το επιβάλλει η επιχειρηματική λογική. Για παράδειγμα, μπορεί να υπάρχουν κάποιες περιπτώσεις που δεν πρέπει να δημιουργήσουμε ομάδα κάτω από 10 άτομα. Ωστόσο, το να θέτουμε μια υψηλή τιμή στην παράμετρο μπορεί να οδηγήσει σε λανθασμένα αποτελέσματα. **Η προεπιλεγμένη τιμή της παραμέτρου είναι το 1**.

MINIMUM_SUPPORT

Η παράμετρος αυτή καθορίζει τον ελάχιστον αριθμό εγγραφών που απαιτούνται για να δημιουργήσουν μια ομάδα. Εάν ο αριθμός των εγγραφών στην ομάδα είναι μικρότερος από τον αριθμό της παραμέτρου, τότε η ομάδα θεωρείται κενή και ξανά δημιουργείται. Εάν καθοριστεί η τιμή της παραμέτρου πολύ υψηλή, τότε μπορεί κάποιες ομάδες που είναι σωστές να χαθούν. Εάν χρησιμοποιείται η τεχνική ομαδοποίησης EM, ενδέχεται κάποιες ομάδες να έχουν τιμή υποστήριξης (support) χαμηλότερη από την καθορισμένη τιμή της παραμέτρου. Αυτό συμβαίνει γιατί για κάθε εγγραφή αξιολογείται η συμμετοχή της σε κάθε ομάδα βάσει κάποιου συγκεκριμένου ορίου, έτσι για κάποιες ομάδες ενδέχεται να υπάρχει ένας ελάχιστος αριθμός εγγραφών κάτω από την απαιτούμενη υποστήριξη. **Η προεπιλεγμένη τιμή της παραμέτρου είναι το 1**.

STOPPING TOLERANCE

Η παράμετρος αυτή χρησιμοποιείται από τον αλγόριθμο για να καθορίσει πότε το μοντέλο έχει συγκλίνει και ο αλγόριθμος πρέπει να σταματήσει να κατασκευάζει το μοντέλο. Η σύγκλιση του αλγορίθμου επιτυγχάνεται όταν η συνολική αλλαγή στην πιθανότητα κάθε ομάδας είναι μικρότερη από την αναλογία της τιμής της παραμέτρου STOPPING_TOLERANCE διαιρούμενο με το πλήθος του μοντέλου. Αναπαριστά τον μέγιστο αριθμό των εγγραφών που μπορούν να αλλάξουν την ομάδα που ανήκουν μέχρι το μοντέλο να συγκλίνει. Η τιμή της παραμέτρου ελέγχεται σε κάθε επανάληψη του αλγορίθμου. Αν μεγαλώσουμε την τιμή της παραμέτρου, θα ωθήσουμε τον αλγόριθμο να συγκλίνει πιο γρήγορα με αποτέλεσμα να έχουμε πιο «χαλαρές» ομάδες, ενώ αν μειώσουμε τον αριθμό θα έχουμε σαν αποτέλεσμα πιο «σφιχτές» ομάδες. Αν έχουμε ένα μικρό αριθμό δεδομένων ή πολύ διακριτές ομάδες, τότε μπορούμε να θέσουμε την τιμή της παραμέτρου ίση με το 1. **Η προεπιλεγμένη τιμή είναι το 10**.

SAMPLE SIZE

Η παράμετρος αυτή σχετίζεται με τις κλιμακούμενες εκδόσεις των τεχνικών του αλγορίθμου (vanilla versions) και καθορίζει τον αριθμό των εγγραφών από το σύνολο δεδομένων που χρησιμοποιούνται σε κάθε βήμα του αλγορίθμου. Η μείωση της τιμής αυτής μπορεί να ωθήσει τον αλγόριθμο να συγκλίνει νωρίς χωρίς να χρησιμοποιήσει όλα τα απαιτούμενα δεδομένα, ιδίως αν χρησιμοποιείται μαζί με μια μεγάλη τιμή στην παράμετρο STOPPING_TOLERANCE. Αυτό μπορεί να βοηθήσει να έχουμε μια γρήγορη ομαδοποίηση σε μεγάλο πλήθος δεδομένων. Αν θέσουμε την τιμή ίση με το 0, θα ωθήσουμε τον αλγόριθμο να χρησιμοποιήσει όλη τη διαθέσιμη μνήμη που έχει για να κατασκευάσει τις ομάδες, σε ένα πέρασμα του αλγορίθμου. Αυτό όμως μπορεί να δημιουργήσει προβλήματα απόδοσης και έλλειψης μνήμης. Εάν δεν υπάρχει η απαιτούμενη μνήμη για να «περαστεί» το σύνολο δεδομένων στην μνήμη, η διαδικασία θα επιστρέψει λάθος. **Η προεπιλεγμένη τιμή είναι το 50,000**.

MAXIMUM INPUT ATTRIBUTES

Η παράμετρος αυτή ελέγχει πόσα γνωρίσματα μπορούν να χρησιμοποιηθούν στην διαδικασία της ομαδοποίησης, προτού χρησιμοποιηθεί η τεχνική της επιλογής γνωρισμάτων.

Έτσι αν υπάρχουν περισσότερα γνωρίσματα στο σύνολο δεδομένων από την τιμή της παραμέτρου, τότε χρησιμοποιείται η τεχνική επιλογής για να διαλέξει τα πιο δημοφιλή γνωρίσματα. Αυτό το όριο χρησιμοποιείται γιατί ο αριθμός των γνωρισμάτων επηρεάζει σημαντικά την απόδοση του αλγορίθμου. Εάν η παράμετρος τεθεί 0, δεν υπάρχει κάποιος μέγιστος αριθμός γνωρισμάτων. Αλλάζοντας τον αριθμό των γνωρισμάτων που χρησιμοποιούνται, μειώνεται η απόδοση του αλγορίθμου. **Η προεπιλεγμένη τιμή είναι το 255**.

MAXIMUM STATES

Η παράμετρος αυτή ελέγχει πόσες καταστάσεις (states) μπορεί να έχει ένα γνώρισμα – είναι δηλαδή ο μέγιστος αριθμός καταστάσεων των γνωρισμάτων που μπορεί να υποστηρίξει ο αλγόριθμος. Εάν ένα γνώρισμα έχει παραπάνω καταστάσεις από την τιμή της παραμέτρου, τότε οι πιο δημοφιλείς καταστάσεις επιλέγονται από τον αλγόριθμο και οι υπολειπόμενες τιμές ομαδοποιούνται κάτω από μια γενική τιμή με την ονομασία «Άλλο». Αυτό το όριο υπάρχει για το ο υπερβολικός αριθμός στοιχείων σε μια κατάσταση (cardinality) ενός γνωρίσματος μπορεί να επηρεάσει σημαντικά την απόδοση του αλγορίθμου και την μνήμη που απαιτείται για την λειτουργία του αλγορίθμου. **Η προεπιλεγμένη τιμή είναι το 100**.

Εντολές DMX

Για το παράδειγμα χρήσης των εντολών DMX για την περίπτωση της ομαδοποίησης, θα χρησιμοποιήσουμε την δομή Εξόρυξης Γνώσης TargetMail που παρουσιάσαμε παραπάνω για τα δέντρα απόφασης. Στην δομή Εξόρυξης Γνώσης που έχουμε στη διάθεσή μας, μπορούμε να την τροποποιήσουμε και να προσθέσουμε ένα μοντέλο ομαδοποίησης (clustering) με την παρακάτω εντολή:

```
ALTER MINING STRUCTURE TargetMail
ADD MINING MODEL TargetMailCL
USING Microsoft_Clustering
```

Για την εκπαίδευση του καινούργιου μοντέλου, χρησιμοποιούμε την εντολή:

```
INSERT INTO  TargetMailCL
```

Για να ανακτήσουμε τις διάφορες καταστάσεις (τιμές) μιας στήλης δεδομένων με διακριτές τιμές, δηλώνουμε:

```
SELECT DISTINCT [Region] FROM TargetMailCL
```

Για να ανακτήσουμε το εύρος τιμών μια συνεχούς μεταβλητής (ελάχιστη, μέση και μέγιστη τιμή) χρησιμοποιούμε τα παρακάτω:

```
SELECT DISTINCT RangeMin([Yearly Income]) AS [min],
 [Yearly Income] AS [mean],
 RangeMax([Yearly Income]) AS [max]
FROM TargetMailCL
```

Για να λάβουμε όλες τις περιπτώσεις δεδομένων, που υπάρχουν στην ομάδα 1, δηλώνουμε:

```
SELECT * FROM TargetMailCL.CASES WHERE IsInNode('001')
```

Για να δείξουμε για κάθε πελάτη που υπάρχει στο σύνολο δεδομένων, διάφορα στατιστικά στοιχεία, καθώς επίσης και την ομάδα που ανήκει, δηλώνουμε τα παρακάτω:

```
SELECT t.CustomerKey, t.FirstName, t.LastName, Cluster(),
   PredictHistogram(Cluster())
FROM TargetMailCL
PREDICTION JOIN
OPENQUERY([Adventure Works DW],'SELECT * FROM vTargetMail') AS t
ON
  t.Age = TargetMailCL.Age AND
  t.CommuteDistance = TargetMailCL.[Commute Distance] AND
  t.Gender = TargetMailCL.Gender AND
  t.HouseOwnerFlag = TargetMailCL.[House Owner Flag] AND
  t.MaritalStatus = TargetMailCL.[Marital Status] AND
  t.NumberCarsOwned = TargetMailCL.[Number Cars Owned] AND
  t.NumberChildrenAtHome = TargetMailCL.[Number Children At Home] AND
  t.EnglishOccupation = TargetMailCL.Occupation AND
  t.Region = TargetMailCL.Region AND
  t.TotalChildren = TargetMailCL.[Total Children] AND
  t.YearlyIncome = TargetMailCL.[Yearly Income]
```

Για να επιστρέψουμε τους πρώτους 25 πελάτες, οι οποίοι είναι πιθανόν να ανήκουν στην ομάδα 1, χρησιμοποιούμε τις παρακάτω εντολές:

```
SELECT TOP 25 t.CustomerKey, t.FirstName, t.LastName
FROM TargetMailCL
PREDICTION JOIN
OPENQUERY([Adventure Works DW],'SELECT * FROM vTargetMail') AS t
ON
  t.Age = TargetMailCL.Age AND
  t.CommuteDistance = TargetMailCL.[Commute Distance] AND
  t.Gender = TargetMailCL.Gender AND
  t.HouseOwnerFlag = TargetMailCL.[House Owner Flag] AND
  t.MaritalStatus = TargetMailCL.[Marital Status] AND
  t.NumberCarsOwned = TargetMailCL.[Number Cars Owned] AND
  t.NumberChildrenAtHome = TargetMailCL.[Number Children At Home] AND
  t.EnglishOccupation = TargetMailCL.Occupation AND
  t.Region = TargetMailCL.Region AND
  t.TotalChildren = TargetMailCL.[Total Children] AND
  t.YearlyIncome = TargetMailCL.[Yearly Income]
ORDER BY ClusterProbability('Cluster 1') DESC
```

Microsoft Association Rules

Ο αλγόριθμος Microsoft Association Rules δημιουργεί ομάδες από οντότητες και στην συνέχεια καθορίζει πόσο συχνά αυτές οι ομάδες εμφανίζονται στο σύνολο δεδομένων. Ο αλγόριθμος ξεκινά με το να ορίσει ομάδες που περιέχουν ένα στοιχείο (μια τιμή από κάθε γνώρισμα στο σύνολο δεδομένων) και ελέγχει πόσες φορές εμφανίζεται η τιμή του γνωρίσματος εισόδου που μελετάμε (για παράδειγμα η πώληση ενός προϊόντος) στις ομάδες αυτές. Στην συνέχεια ο αλγόριθμος καθορίζει ποιες ομάδες είναι αρκετά δημοφιλείς ώστε να συνεχίσουν περαιτέρω στην ανάλυση, βάσει ενός συγκεκριμένου ορίου που ονομάζεται ***ελάχιστη υποστήριξη*** **(support)**. Ο όρος *support* (ή *συχνότητα*) αναφέρεται στον αριθμό των περιπτώσεων (cases) στο σύνολο δεδομένων, που περιέχουν την ομάδα που μελετάμε (δηλαδή πόσες φορές εμφανίζεται στο σύνολο των δεδομένων η τιμή που μελετάμε). Έτσι στο αρχικό πέρασμα επιλέγονται οι ομάδες (που όπως είπαμε αρχικά περιέχουν ένα στοιχείο) τα οποία ικανοποιούν μια ελάχιστη υποστήριξη, δηλαδή πόσες φορές οι τιμές των γνωρισμάτων που ανήκουν στην ομάδα προς επιλογή εμφανίζονται στο γενικό σύνολο δεδομένων).

Ο αλγόριθμος συνεχίζει να δημιουργεί ομάδες με δυο στοιχεία, επιλέγοντας τες από τις ομάδες με ένα στοιχείο που έχουν επιλέγει από το προηγούμενο βήμα. Ο έλεγχος πραγματοποιείται στις εγγραφές, οι οποίες περιέχουν και τα δύο γνωρίσματα σε μια συναλλαγή, όπως για παράδειγμα η αγορά 2 συγκεκριμένων προϊόντων. Ομοίως και στην περίπτωση αυτή, χρησιμοποιείται ένα ελάχιστό όριο στην επιλογή των ομάδων από το γενικό σύνολο των συναλλαγών. Στην συνέχεια ο αλγόριθμος επιλέγει ομάδες με περισσότερα στοιχεία και τερματίζει στο σημείο που δεν υπάρχουν ομάδες που να ικανοποιούν το ελάχιστο όριο της υποστήριξης.

Όταν επιλεχθούν οι ομάδες δεδομένων, ο αλγόριθμος δημιουργεί κανόνες βάσει των αποτελεσμάτων. Οι κανόνες αυτοί χρησιμοποιούνται για να προβλέψουν μελλοντικές συσχετίσεις μεταξύ των γνωρισμάτων εισόδου και του γνωρίσματος εισόδου, που μελετάει ο αλγόριθμος. Μια πρόβλεψη μπορεί να ορίσει πότε ένα αντικείμενο είναι πιθανό να εμφανιστεί σε μια συναλλαγή δεδομένου την παρουσία άλλων αντικειμένων. Έτσι όταν αγοραστεί ένα συγκεκριμένο προϊόν από ένα πελάτη είναι πολύ πιθανό να αγοράζει και κάποιο άλλο προϊόν, έτσι το σύστημα μπορεί να κάνει την αντίστοιχη προσφορά προς τον πελάτη. Η πρόβλεψη αυτή μπορεί να συμπεριλάβει επίσης επιπλέον πληροφορίες όπως την πιθανότητα, την υποστήριξη και την και την εμπιστοσύνη του κάθε κανόνα.

Γενικά για τον Αλγόριθμο

Ο αλγόριθμος Microsoft Association είναι ένας αλγόριθμος εύρεσης συσχετίσεων μεταξύ γνωρισμάτων και είναι χρήσιμος κυρίως για εξαγωγή προτάσεων (recommendations) στους χρήστες. Με τον όρο προτάσεις εννοούμε τα αποτελέσματα του αλγορίθμου, σύμφωνα με τα οποία προτείνει συγκεκριμένα προϊόντα στους πελάτες βάσει προϊόντων που έχουν αγοράσει ή για κάποια προϊόντα που έχουν δείξει ενδιαφέρον για αγορά. Επίσης ο αλγόριθμος μπορεί να χρησιμοποιηθεί για ανάλυση του καλαθιού αγοράς (market basket analysis) ενός καταναλωτή.

Ο αλγόριθμος Microsoft Association Rules καθώς και ο αλγόριθμος Association Decision Trees μπορούν να χρησιμοποιηθούν για να αναλύσουν συσχετίσεις, αλλά οι κανόνες που ανακαλύπτονται από κάθε αλγόριθμο μπορεί να διαφέρουν. Σε ένα μοντέλο με δέντρα αποφάσεων, ο διαχωρισμός των κόμβων για να ορίσει κανόνες βασίζονται στην απόκτηση πληροφορίας (information gain) ενώ στον αλγόριθμο Association Rules οι κανόνες συσχέτισης, όπως θα δούμε παρακάτω, βασίζονται εξ' ολοκλήρου στην εμπιστοσύνη των κανόνων του αλγορίθμου. Αυτό μπορεί να έχει σαν αποτέλεσμα, ένας κανόνας που είναι «ισχυρός», έχει δηλαδή υψηλή εμπιστοσύνη, μπορεί να μην είναι ενδιαφέρον κανόνας γιατί δεν προσφέρει νέα πληροφορία.

Πώς λειτουργεί ο αλγόριθμος Microsoft Association

Ο αλγόριθμος Microsoft Association Rules είναι μια υλοποίηση του αλγορίθμου A-priori. Ο αλγόριθμος A-priori δεν αναλύει πρότυπα δεδομένων, αλλά δημιουργεί και στην συνέχεια υπολογίζει υποψήφιες ομάδες αντικειμένων για να εισαχθούν στο μοντέλο. Ένα αντικείμενο μπορεί να είναι ένα προϊόν ή η τιμή ενός χαρακτηριστικού, ανάλογα με τον τύπο των δεδομένων που αναλύονται. Ο αλγόριθμος αρχικά διασχίζει το σύνολο δεδομένων για να βρει αντικείμενα (ο όρος που χρησιμοποιείται είναι items, που αναφέρεται στην τιμή του κάθε γνωρίσματος (attribute) του συνόλου δεδομένων) που εμφανίζονται μαζί σε κάθε περίπτωση (σε κάθε εγγραφή του συνόλου δεδομένων).

Οι πιο απλοί τύποι δεδομένων, που εφαρμόζονται στις μεταβλητές ενός μοντέλου συσχέτισης είναι οι Boolean μεταβλητές, που αναπαριστούν ένα ΝΑΙ / ΟΧΙ ή ΕΜΦΑΝΙΣΗ / ΑΠΟΥΣΙΑ ΤΙΜΗΣ και ανατίθενται σε κάθε χαρακτηριστικό, όπως ένα προϊόν ή το όνομα κάποιου συμβάντος. Για παράδειγμα, η ανάλυση του καλαθιού αγοράς είναι ένα παράδειγμα μοντέλου συσχέτισης που χρησιμοποιεί Boolean μεταβλητές για να αναπαραστήσει την εμφάνιση ή την απουσία ενός προϊόντος από το καλάθι αγορών ενός πελάτη.

Για κάθε ομάδα αντικειμένων, ο αλγόριθμος δημιουργεί συγκεκριμένες μετρικές που αναπαριστούν την υποστήριξη και την εμπιστοσύνη. Αυτές οι μετρικές μπορούν να χρησιμοποιηθούν για να ταξινομήσουν και να εξάγουν χρήσιμους κανόνες από τις ομάδες αντικειμένων. Οι κανόνες συσχέτισης μπορούν να δημιουργηθούν και από μεταβλητές με αριθμητικές τιμές. Αν τα χαρακτηριστικά είναι συνεχείς τιμές, τότε οι αριθμοί μπορούν να γίνουν διακριτοί με το να τοποθετηθούν σε συγκεκριμένες ομάδες τιμών (buckets). Οι διακριτές αυτές τιμές, στη συνέχεια μπορούν να χρησιμοποιηθούν σαν Boolean τιμές για να εξαχθούν οι κανόνες συσχέτισης.

Υποστήριξη και Εμπιστοσύνη

Η *υποστήριξη (support)*, που μερικές φορές αναφέρεται και σαν συχνότητα, αντιπροσωπεύει τον αριθμό των περιπτώσεων που θα πρέπει να περιέχουν το αντικείμενο ή τον συνδυασμό αντικειμένων που μελετάει ο αλγόριθμος. Στον μοντέλο που παράγεται συμπεριλαμβάνονται αντικείμενα που πρέπει να έχουν την ελάχιστη υποστήριξη. Για παράδειγμα, αν για ένα σύνολο αντικειμένων {Α, Β, Γ} παράγεται ο κανόνας

Α και Β -> Γ

και καθοριστεί η ελάχιστη υποστήριξη 10, τότε πρέπει να υπάρχουν τουλάχιστον 10 περιπτώσεις που να περιέχουν και τα 3 αντικείμενα που μελετάμε για να μπορούμε να δεχθούμε τον κανόνα αυτόν.

Η *εμπιστοσύνη (confidence)* αναπαριστά το ποσοστό των περιπτώσεων που μελετάει ο αλγόριθμος σε σχέση με τα μέρη του κανόνα συσχέτισης που εφαρμόζεται στην περίπτωση αυτή και δείχνει την ισχύ του κανόνα. Για παράδειγμα για τον κανόνα, που αναφέραμε παραπάνω

Α και Β -> Γ

η εμπιστοσύνη είναι το ποσοστό των περιπτώσεων που εμφανίζεται το αντικείμενο Γ (το δεξί μέρος του κανόνα, που δείχνει και το τελικό αποτέλεσμα) διά τον αριθμό των περιπτώσεων που εμφανίζονται τα αντικείμενα {Α, Β, Γ}.

Δεδομένα που απαιτούνται για τον Αλγόριθμο

Τα δεδομένα που απαιτούνται από τον αλγόριθμο κανόνων συσχέτισης παρουσιάζονται παρακάτω:

- **Μία μοναδική στήλη που να περιέχει το κλειδί.** Κάθε μοντέλο πρέπει να περιέχει μια στήλη με αριθμητικά δεδομένα ή κείμενο τα οποία προσδιορίζουν μοναδικά κάθε εγγραφή. Τα σύνθετα κλειδιά δεν επιτρέπονται.
- **Μια μοναδική στήλη με δεδομένα προς πρόβλεψη:** Ένα μοντέλο συσχέτισης μπορεί να έχει μόνο μια στήλη με δεδομένα προς πρόβλεψη. Συνήθως η στήλη αυτή σχετίζεται με την στήλη που περιέχει το κλειδί στον εσωτερικό πίνακα (nested table), ο οποίος περιέχει την λίστα των προϊόντων που έχουν αγοραστεί από τους πελάτες. Οι τιμές αυτές πρέπει να είναι διακριτές ή να έχουν γίνει διακριτές, ύστερα από κάποια επεξεργασία.
- **Στήλες με δεδομένα Εισόδου:**. Οι στήλες με τα δεδομένα εισόδου πρέπει να έχουν διακριτές τιμές. Συνήθως τα δεδομένα για την δημιουργία του μοντέλου συσχέτισης περιέχονται σε δυο πίνακες. Για παράδειγμα, ο ένας πίνακας μπορεί να περιέχει τις πληροφορίες για τους πελάτες ενώ ο άλλος πίνακας να περιέχει τα δεδομένα με τις συναλλαγές αγορών των πελατών. Ο δεύτερος πίνακας εισάγεται στον αλγόριθμο με την δομή ενός εσωτερικού πίνακα (nested table).

Παράμετροι Αλγορίθμου

Ο αλγόριθμος εύρεσης συσχετίσεων είναι πολύ ευαίσθητος όσον αφορά στον καθορισμό των τιμών των παραμέτρων του, δηλαδή η αλλαγή στην τιμή μιας από τις παραμέτρους ενδέχεται να επηρεάσει αρκετά το μοντέλο που παράγει ο αλγόριθμος. Στην συνέχεια παρουσιάζουμε τις παραμέτρους του αλγορίθμου Microsoft Association Rules .

MINIMUM_SUPPORT

Η παράμετρος αυτή καθορίζει τον ελάχιστο αριθμό περιπτώσεων (cases) που πρέπει να περιέχει μια ομάδα αντικειμένων (itemset) έτσι ώστε ο αλγόριθμος να δημιουργήσει έναν κανόνα. Αν η τιμή που δοθεί είναι το 1, τότε ο ελάχιστος αριθμός περιπτώσεων υπολογίζεται σαν ποσοστό επί του συνολικού αριθμού των περιπτώσεων. Αν η τιμή που δοθεί είναι πολύ μικρή (για παράδειγμα 0.001) τότε ο αλγόριθμος μπορεί να χρειαστεί πολύ περισσότερο χρόνο να επεξεργαστεί τα δεδομένα και να απαιτήσει πολύ περισσότερη μνήμη.

Αν σαν τιμή δοθεί μια ακέραιη τιμή μεγαλύτερη του 1, τότε η παράμετρος καθορίζει τον ακριβή ελάχιστο αριθμό περιπτώσεων που πρέπει να περιέχει η ομάδα αντικειμένων. Ο αλγόριθμος μπορεί αυτόματα να αυξήσει την τιμή της παραμέτρου στην περίπτωση που η διαθέσιμη μνήμη είναι περιορισμένη. Με μια μικρή τιμή στην ελάχιστη υποστήριξη και στην ελάχιστη πιθανότητα, παράγονται περισσότεροι κανόνες συσχέτισης.

Η προεπιλεγμένη τιμή της παραμέτρου είναι το 0.03, που σημαίνει ότι σε μια ομάδα αντικειμένων που θα συμπεριληφθεί σε ένα κανόνα συσχέτισης, πρέπει να περιέχονται τουλάχιστον 3% των συνολικών περιπτώσεων.

MAXIMUM_SUPPORT

Η παράμετρος αυτή καθορίζει τον μέγιστο αριθμό περιπτώσεων που πρέπει να περιέχει μια ομάδα αντικειμένων σαν υποστήριξη. Η παράμετρος αυτή μπορεί να χρησιμοποιηθεί για να εξαλείψει αντικείμενα που εμφανίζονται συχνά και με αποτέλεσμα να έχουν ελάχιστη σημασία και χρησιμότητα.

Αν η τιμή είναι μικρότερη από το 1, η τιμή αναπαριστά το ποσοστό επί του συνολικού αριθμού των περιπτώσεων. Αν η τιμή είναι μεγαλύτερη του 1, αναπαριστά τον ακριβή αριθμό

περιπτώσεων που πρέπει να περιέχει η ομάδα. **Η προεπιλεγμένη τιμή είναι το 0.03** (3% επί του συνολικού αριθμού των περιπτώσεων).

MINIMUM_PROBABILITY

Η παράμετρος αυτή καθορίζει την ελάχιστη πιθανότητα ώστε ο κανόνας να είναι αληθής. Για παράδειγμα, αν δοθεί η τιμή 0.5 σημαίνει ότι κανένας κανόνας με μικρότερη από 50% πιθανότητα δεν μπορεί να παραχθεί. **Η προεπιλεγμένη τιμή είναι το 0.4**.

MINIMUM_IMPORTANCE

Η παράμετρος αυτή καθορίζει το όριο σημαντικότητας για τους κανόνες συσχέτισης. Κανόνες με σημαντικότητα ελάχιστη από την τιμή της παραμέτρου αποκλείονται από το μοντέλο.

MAXIMUM_ITEMSET_SIZE

Η παράμετρος αυτή καθορίζει τον μέγιστο αριθμό αντικειμένων που επιτρέπονται σε μια ομάδα. Η τιμή της παραμέτρου ίση με το 0, σημαίνει ότι δεν υπάρχει κανένα όριο στο μέγεθος της ομάδας αντικειμένων. Μειώνοντας το μέγιστο μέγεθος κάθε ομάδας αντικειμένων μειώνεται και ο χρόνος επεξεργασίας του αλγορίθμου, διότι ο αλγόριθμος μπορεί να αποφύγει περαιτέρω επαναλήψεις στο σύνολο δεδομένων, όταν το μέγεθος μιας ομάδας αντικειμένων φτάσει το καθορισμένο όριο. **Η προεπιλεγμένη τιμή είναι το 3**.

MINIMUM_ITEMSET_SIZE

Η παράμετρος αυτή καθορίζει τον ελάχιστο αριθμό αντικειμένων που επιτρέπονται σε μια ομάδα. Αν αυξηθεί η τιμή της παραμέτρου, το μοντέλο μπορεί να περιέχει λιγότερες ομάδες. Αυτό μπορεί να είναι χρήσιμο στην περίπτωση που θέλουμε να αγνοήσουμε τις ομάδες που περιέχουν ένα μόνο αντικείμενο. Αυξάνοντας όμως την τιμή της παραμέτρου δεν σημαίνει ότι θα μειωθεί ο χρόνος επεξεργασίας του αλγορίθμου (αφού θα παραχθούν τελικά λιγότερες ομάδες). Αυτό συμβαίνει γιατί ο αλγόριθμος πάντα ξεκινά με ομάδες του ενός αντικειμένου και αυξάνει τον αριθμό αυτό σε κάθε βήμα. Πρακτικά δηλαδή, η παράμετρος αυτή χρησιμεύει για να φιλτράρει το τελικό αποτέλεσμα των ομάδων αντικειμένων. **Η προεπιλεγμένη τιμή της παραμέτρου είναι το 1**.

MAXIMUM_ITEMSET_COUNT

Η παράμετρος αυτή καθορίζει τον μέγιστο αριθμό ομάδων αντικειμένων που παράγονται από τον αλγόριθμο. Αν δεν καθοριστεί κάποιος συγκριμένος αριθμός, χρησιμοποιείται η **προεπιλεγμένη τιμή, που είναι το 200,000**. Η παράμετρος αυτή διασφαλίζει ότι δεν θα παραχθεί ένα πολύ μεγάλος αριθμός ομάδων αντικειμένων. Στην περίπτωση που παραχθούν πολλές ομάδες, ο αλγόριθμος θα διατηρήσει συγκεκριμένο αριθμό, βάσει του βαθμού σπουδαιότητας που έχει η κάθε ομάδα.

Επιλογή Χαρακτηριστικών

Ο αλγόριθμος Microsoft Association Rules δεν χρησιμοποιεί κάποια συγκεκριμένη τεχνική αυτόματης επιλογής χαρακτηριστικών. Αντ' αυτού, ο αλγόριθμος παρέχει τις κατάλληλες παραμέτρους, που ελέγχουν τα δεδομένα που θα χρησιμοποιηθούν από τον αλγόριθμο. Αυτές οι παράμετροι, όπως είδαμε παραπάνω, καθορίζουν το μέγεθος της κάθε ομάδας αντικειμένων (itemset) ή θέτουν την μέγιστη και ελάχιστη υποστήριξη που πρέπει να πληροί μια ομάδα

αντικειμένων για να προστεθεί στο μοντέλο. Μερικά παραδείγματα χρήσης των παραμέτρων είναι τα παρακάτω:

- Για να φιλτράρουμε αντικείμενα ή γεγονότα που είναι πολύ κοινά και ουσιαστικά δεν είναι χρήσιμα και χωρίς ενδιαφέρον, μπορούμε να μειώσουμε την τιμή της παραμέτρου **MAXIMUM_SUPPORT** (που δείχνει τον μέγιστο αριθμό περιπτώσεων που πρέπει να έχει μια ομάδα αντικειμένων σαν υποστήριξη) για να αφαιρεθούν οι πολύ συχνά εμφανιζόμενες ομάδες αντικειμένων (itemsets) από το μοντέλο.
- Για να φιλτράρουμε αντικείμενα ή ομάδες αντικειμένων που είναι πολύ σπάνια, μεγαλώνουμε την τιμή της παραμέτρου **MINIMUM_SUPPORT** (που δείχνει τον ελάχιστο αριθμό περιπτώσεων που πρέπει να περιέχει μια ομάδα αντικειμένων για να παραχθεί ένας κανόνας).
- Για να φιλτράρουμε κάποιους κανόνες, μεγαλώνουμε την τιμή της παραμέτρου **MINIMUM_PROBABILITY** (που δείχνει την ελάχιστη πιθανότητα που πρέπει να πληροί ο κανόνας για να είναι αληθής).

Εντολές DMX

Για να δημιουργήσουμε ένα μοντέλο κανόνων συσχέτισης, για να προβλέψουμε τις συσχετίσεις μεταξύ των υποκατηγοριών προϊόντων, χρησιμοποιούμε τις εντολές:

```
CREATE MINING MODEL SubcategoryAssociations
([Customer ID]   LONG KEY,
  [Subcategories] TABLE PREDICT
  (  [Subcategory] TEXT KEY)
) USING Microsoft_Association_Rules
```

Για να εκπαιδεύσουμε το μοντέλο των κανόνων συσχέτισης, έχουμε:

```
INSERT INTO SubcategoryAssociations
([Customer ID],
 [Subcategories](SKIP,[Subcategory])
)
SHAPE
{
OPENQUERY([Adventure Works DW],
   'SELECT
   [OrderNumber]
   FROM
   [dbo].[vAssocSeqOrders]
   ORDER BY
   [OrderNumber]')
}
APPEND
(
 {
   OPENQUERY([Adventure Works DW],
   'SELECT
   [OrderNumber],
```

```
   [Subcategory]
   FROM
   (SELECT DISTINCT vAssocSeqLineItems.OrderNumber,
DimProductSubcategory.EnglishProductSubcategoryName
AS Subcategory
   FROM DimProduct
INNER JOIN DimProductSubcategory
ON DimProduct.ProductSubcategoryKey = DimProductSubcategory.ProductSubcategoryKey INNER JOIN
vAssocSeqLineItems ON DimProduct.ModelName = vAssocSeqLineItems.Model)
AS [CustomerSubcategories]
   ORDER BY
   [OrderNumber]')
 }
 RELATE
   [OrderNumber] TO [OrderNumber]
) AS [CustomerSubcategories]
```

Για να καθορίσουμε τις πρώτες δύο υποκατηγορίες προϊόντων, που είναι πιθανόν να αγοράσει ένα πελάτης, ο οποίος έχει αγοράσει ήδη ένα ποδήλατο δρόμου (road bike) και μια αθλητική μπλούζα (jersey), χρησιμοποιούμε τις εντολές:

```
SELECT
  Predict([Subcategories],2) as [Subcategories]
FROM
  [SubcategoryAssociations]
NATURAL PREDICTION JOIN (SELECT
  (SELECT 'Road Bikes' AS Subcategory
   UNION SELECT 'Jerseys' AS Subcategory
  ) AS Subcategories
) AS t
```

Για να καθορίσουμε την πιθανότητα κάθε πελάτης, που δεν έχει αγοράσει κράνος ποδηλάτου να αγοράσει τελικά ένα κράνος, χρησιμοποιούμε τις εντολές:

```
SELECT * FROM
(SELECT FLATTENED
  t.[CustomerKey],
  (SELECT $Probability AS ProbHelmets FROM
    Predict(Subcategories, INCLUDE_STATISTICS)
    WHERE Subcategory='Helmets') AS Sub
FROM
  [SubcategoryAssociations]
PREDICTION JOIN
  SHAPE {
  OPENQUERY([Adventure Works DW],
    'SELECT
    [CustomerKey],  [OrderNumber]
    FROM
```

```
    [dbo].[vAssocSeqOrders]
    ORDER BY
    [OrderNumber]')}
  APPEND
  ({OPENQUERY([Adventure Works DW],
    'SELECT
    [Subcategory],
    [OrderNumber]
    FROM
    (SELECT DISTINCT vAssocSeqLineItems.OrderNumber,
    DimProductSubcategory.EnglishProductSubcategoryName
    AS Subcategory
    FROM DimProduct INNER JOIN DimProductSubcategory     ON DimProduct.ProductSubcategoryKey =
    DimProductSubcategory.ProductSubcategoryKey
    INNER JOIN     vAssocSeqLineItems ON DimProduct.ModelName =  vAssocSeqLineItems.Model
    ) as [CustomerSubcategories]
    ORDER BY
    [OrderNumber]')}
    RELATE
    [OrderNumber] TO [OrderNumber])
    AS
    [CustomerSubcategories] AS t
ON
[SubcategoryAssociations].[Subcategories].[Subcategory] =
t.[CustomerSubcategories].[Subcategory]
) as s
WHERE [Sub.ProbHelmets] <> NULL
```

5. Εφαρμογή Χρήσης Μοντέλων Εξόρυξης Γνώσης

Έχουμε παρουσιάσει έως τώρα την γενική θεωρία γύρω από το Business Intelligence και πως θα μπορέσει να υποστηρίξει τις διαδικασίες σε ένα οργανισμό για την λήψη σωστών και έγκαιρων αποφάσεων σχετικά με την στρατηγική και την πορεία του οργανισμού, τις έννοιες σχετικά με την Εξόρυξη Γνώσης καθώς και τις πιο βασικές τεχνικές της και τα κύρια επιχειρησιακά προβλήματα τα οποία μπορεί να αντιμετωπίσει ένας οργανισμός και τα οποία μπορούν να επιλύσουν οι τεχνικές αυτές της Εξόρυξης Γνώσης. Επίσης, έχουμε παρουσιάσει τις λύσεις BI που προσφέρει η Microsoft, μέσα από την πλατφόρμα του SQL Server (2005 – 2008) και έχουμε επικεντρωθεί στις τεχνικές Εξόρυξης Γνώσης, που παρέχονται μέσω των SQL Server Analysis Services και πιο συγκεκριμένα στους αλγορίθμους Microsoft Decision Trees, Clustering και Association Rules.

Στο κεφάλαιο αυτό θα παρουσιάσουμε ένα παράδειγμα χρήσης των τεχνικών Εξόρυξης Γνώσης. Πιο συγκεκριμένα, θα παρουσιάσουμε τα βήματα, που πρέπει να ακολουθήσει ένας προγραμματιστής ή ένας αναλυτής, μέσα από το περιβάλλον Business Intelligence Development Studio (BIDS) του Microsoft Visual Studio για να δημιουργήσει ορισμένα μοντέλα Εξόρυξης Γνώσης προκειμένου να επιλύσει συγκεκριμένα επιχειρησιακά προβλήματα. Μέσα από την παρουσίαση βήμα προς βήμα των ρυθμίσεων του χρήστη, την δημιουργία μοντέλων Εξόρυξης Γνώσης, την αποτίμηση και έλεγχο των αποτελεσμάτων των μοντέλων και την χρήση τους για την δημιουργία προβλέψεων, θα προσπαθήσουμε να παρουσιάσουμε την χρήση του BIDS έτσι ώστε να αποτελέσει το κεφάλαιο αυτό έναν οδηγό χρήσης του περιβάλλοντος ανάπτυξης τεχνικών Εξόρυξης Γνώσης με την χρήση των εργαλείων, που παρέχονται από τη Microsoft.

Για την επίλυση των επιχειρησιακών προβλημάτων με τεχνικές Εξόρυξης Γνώσης, που θα παρουσιάσουμε στη συνέχεια θα χρησιμοποιήσουμε το μοντέλο διαδικασίας CRISP-DM το οποίο περιγράφει τα βήματα, που πρέπει να ακολουθηθούν για την επίλυση προβλημάτων σε μικρά ή μεγάλα έργα Εξόρυξης Γνώσης.

Μοντέλο Διαδικασίας CRISP-DM

Η λέξη **CRISP-DM** προέρχεται από τις λέξεις **CRoss Industry Standard Process for Data Mining** και ουσιαστικά είναι μια διαδικασία Εξόρυξης Γνώσης, που έχει καθοριστεί σαν ένα εργαλείο, το οποίο χρησιμοποιούν διάφοροι εμπειρογνώμονες στο πεδίο της Εξόρυξης Γνώσης για την αντιμετώπιση των προβλημάτων. Είναι ένα πρότυπο διαδικασίας, το οποίο έχει καθιερωθεί για να χρησιμοποιείται για μικρά ή μεγάλα έργα Εξόρυξης Γνώσης. Η CRISP-DM διαδικασία καθορίζει διάφορες φάσεις λειτουργίας ενός έργου Εξόρυξης Γνώσης και αυτό έχει σαν αποτέλεσμα το έργο (ανεξαρτήτως μεγέθους, αν είναι μικρό ή μεγάλο) να επωφελείται από την διαδικασία και να ολοκληρώνεται γρηγορότερα, φθηνότερα, πιο αξιόπιστα και να είναι πιο εύχρηστο.

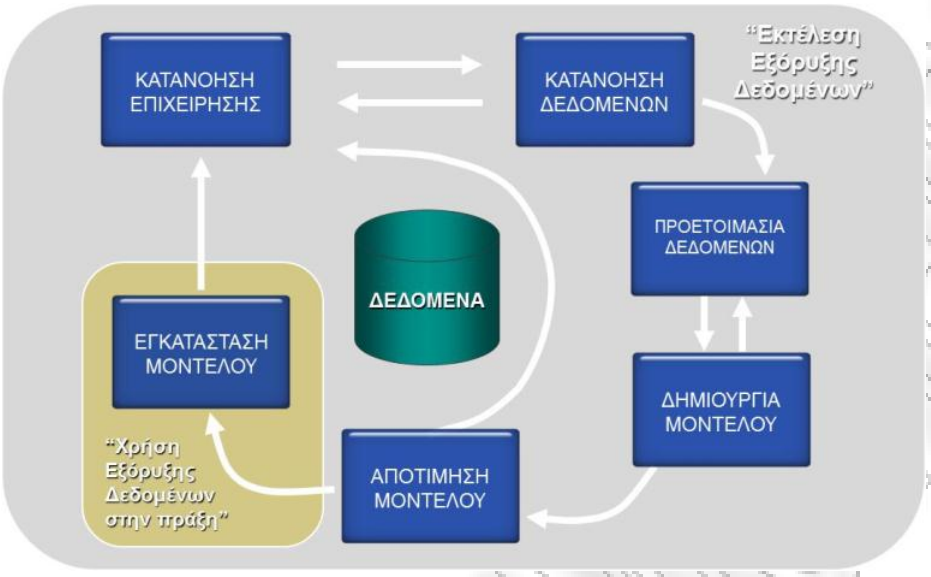
Στην συνέχεια παρουσιάζουμε το μοντέλο της διαδικασίας CRISP-DM, για τον σχεδιασμό και την εκτέλεση ενός έργου Εξόρυξης Γνώσης.

Μοντέλο Διαδικασίας

Το μοντέλο διαδικασίας Εξόρυξης Γνώσης παρέχει μια επισκόπηση του κύκλου ζωής ενός έργου Εξόρυξης Γνώσης. Περιέχει τις αντίστοιχες φάσεις του έργου, τις αντίστοιχες λειτουργίες τους, και τις σχέσεις μεταξύ αυτών των λειτουργιών.

Ο κύκλος ζωής ενός έργου Εξόρυξης Γνώσης αποτελείται από έξι στάδια. Η αλληλουχία των φάσεων δεν είναι αυστηρή, ενώ η μετακίνηση μπρός και πίσω στις διάφορες φάσεις της

διαδικασίας απαιτείται τις περισσότερες φορές. Το ποια φάση ή ποιά συγκεκριμένη λειτουργία μιας φάσης απαιτείται να εκτελεστεί στο επόμενο βήμα εξαρτάται από το αποτέλεσμα της τρέχουσας φάσης. Τα βέλη στο διάγραμμα δείχνουν τις πιο σημαντικές και συχνές εξαρτήσεις μεταξύ των φάσεων.



Διάγραμμα 17: Φάσεις του Μοντέλου διαδικασίας CRISP-DM

Ο εξωτερικός κύκλος στο σχήμα συμβολίζει την κυκλική φύση της ίδιας της Εξόρυξης Γνώσης. Η διαδικασία Εξόρυξης Γνώσης ξεκινάει να λειτουργεί μετά την διαδικασία εγκατάστασης. Στην ολοκλήρωση του κάθε κύκλου, τα διδάγματα που αντλήθηκαν κατά τη διάρκεια της διαδικασίας μπορεί να δώσουν ώθηση σε νέες και συχνά πιο συγκεκριμένες ερωτήσεις των επιχειρήσεων. Αυτό έχει σαν αποτέλεσμα οι μετέπειτα διεργασίες Εξόρυξης Γνώσης να επωφεληθούν από τις εμπειρίες των προηγούμενων κύκλων εκτέλεσης.

Παρακάτω παρουσιάζουμε μια σύντομη περιγραφή της κάθε φάσης της διαδικασίας CRISP.

Κατανόηση Επιχείρησης

Το αρχικό βήμα της διαδικασίας είναι η κατανόηση των στόχων του έργου Εξόρυξης Γνώσης και οι απαιτήσεις από την πλευρά του οργανισμού. Στόχος του βήματος αυτού είναι η μετατροπή της πληροφορίας αυτή σε ένα ορισμό προβλήματος Εξόρυξης Γνώσης και η δημιουργία ενός πρωταρχικού πλάνου επίτευξης των στόχων του προβλήματος.

Κατανόηση Δεδομένων

Το βήμα της κατανόησης των δεδομένων ξεκινά με ένα αρχικό βήμα συλλογής των δεδομένων και συνεχίζει με τις λειτουργίες σχετικά με την περαιτέρω μελέτη των δεδομένων, έτσι ώστε οι χρήστες να εξοικειωθούν μαζί τους, να εντοπίσουν πιθανά προβλήματα ποιότητας των δεδομένων και να ανακαλύψουν χρήσιμα υποσύνολα στα δεδομένα, που πιθανόν να περιέχουν κάποια κρυμμένη πληροφορία. Η κατανόηση των δεδομένων συνδέεται άμεσα με την κατανόηση της επιχείρησης, έτσι ώστε οι χρήστες που θα δημιουργήσουν το μοντέλο Εξόρυξης Γνώσης να έχουν μία πλήρης εικόνα του προβλήματος, που καλούνται να επιλύσουν και των δεδομένων που θα έχουν στην διάθεσή τους.

Προετοιμασία Δεδομένων

Η φάση της προετοιμασίας των δεδομένων καλύπτει όλες τις ενέργειες για την δημιουργία του τελικού συνόλου δεδομένων (data set) – των δεδομένων δηλαδή που θα τροφοδοτήσουν το μοντέλο Εξόρυξης Γνώσης. Οι ενέργειες της φάσης της προετοιμασίας

δεδομένων μπορεί να εκτελεστούν πολλές φορές και όχι απαραίτητα σε μια προκαθορισμένη σειρά και περιλαμβάνουν την επιλογή των χαρακτηριστικών, που θα εισαχθούν στο μοντέλο και τον απαιτούμενο μετασχηματισμό και καθαρισμό δεδομένων για τα εργαλεία μοντελοποίησης.

Δημιουργία Μοντέλου

Στην φάση της δημιουργίας του μοντέλου, επιλέγονται οι διάφορες τεχνικές Εξόρυξης Γνώσης και καθορίζονται οι τιμές των παραμέτρων τους για το καλύτερο αποτέλεσμα της δημιουργίας του μοντέλου Εξόρυξης Γνώσης. Συνήθως, υπάρχουν πολλές τεχνικές που επιλύουν τον ίδιο τύπο επιχειρησιακού προβλήματος. Ωστόσο, οι διάφορες τεχνικές που μπορεί να χρησιμοποιηθούν έχουν συγκεκριμένες απαιτήσεις από την μορφή των δεδομένων (για παράδειγμα κάποιες τεχνικές δεν λειτουργούν με συνεχείς αριθμητικές τιμές και άλλες δουλεύουν μόνο με διακριτές τιμές). Έτσι, πολλές φορές η επαναφορά στο προηγούμενο βήμα της προετοιμασίας των δεδομένων συμβαίνει πολλές φορές για την διόρθωση των δεδομένων, όπου είναι απαραίτητο.

Αποτίμηση Μοντέλου (Evaluation)

Στην φάση αυτή του έργου Εξόρυξης Γνώσης κατασκευάζεται το μοντέλο (ή τα μοντέλα) που πρέπει να έχουν μεγάλη ποιότητα, από την πλευρά της ανάλυσης δεδομένων. Πριν το τελικό βήμα της ανάπτυξης και εγκατάστασης του μοντέλου, είναι απαραίτητο να γίνει η διεξοδική αποτίμηση του μοντέλου και η επανεξέταση των βημάτων, που εκτελέστηκαν για την δημιουργία του, έτσι ώστε να είναι απόλυτα σαφές ότι το αποτέλεσμα επιτυγχάνει τους επιχειρηματικούς στόχους, για τους οποίους δημιουργήθηκε. Ένας ακόμα βασικός στόχος του βήματος αυτού είναι η εύρεση του κατά πόσο όλα τα επιχειρησιακά θέματα έχουν τεθεί προς ανάλυση και έχουν αντιμετωπιστεί όλα τα πιθανά προβλήματα ή εάν εξακολουθούν να υπάρχουν θέματα που δεν έχουν προσδιοριστεί. Τέλος, στην φάση αυτή πρέπει να παρθεί μια απόφαση για την χρήση των αποτελεσμάτων του μοντέλου Εξόρυξης Γνώσης, που θα κατασκευασθεί.

Εγκατάσταση

Η εγκατάσταση του μοντέλου Εξόρυξης Γνώσης συνήθως δεν αποτελεί και το τέλος του έργου Εξόρυξης Γνώσης. Έστω και αν σκοπός του μοντέλου είναι η αύξηση της γνώσης μέσα από τα δεδομένα, η γνώση αυτή που αποκτείται πρέπει να οργανωθεί και να παρουσιαστεί με ένα τέτοιο τρόπο, έτσι ώστε οι χρήστες του οργανισμού να μπορούν να την χρησιμοποιήσουν. Ανάλογα με τις απαιτήσεις του έργου Εξόρυξης Γνώσης, η φάση της εγκατάστασης μπορεί να περιλαμβάνει μια απλή δημιουργία μιας αναφοράς ή να είναι πιο περίπλοκη και να περιλαμβάνει μια επαναληπτική διαδικασία Εξόρυξης Γνώσης. Η επαναληπτική διαδικασία σημαίνει ότι η γνώση που λαμβάνεται, μπορεί να χρησιμοποιηθεί για βελτίωση της επιχειρηματικής λειτουργίας του οργανισμού και να προκύψουν επιπλέον ανάγκες, που πρώτα δεν ήταν ορατές και θα πρέπει να λυθούν με μια βελτιωμένη έκδοση του μοντέλου.

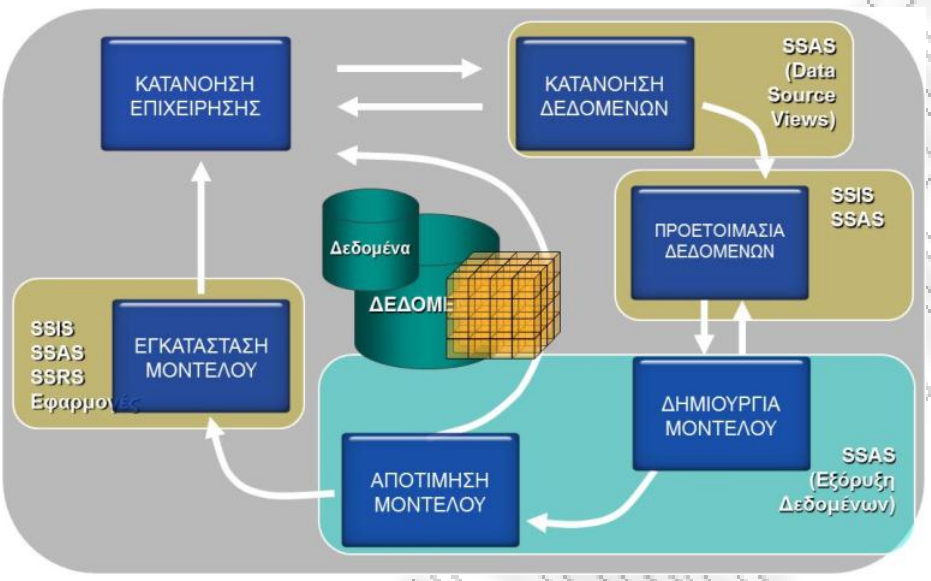
Χρήση διαδικασίας CRISP-DM μέσα από το BIDS

Στο παρακάτω διάγραμμα παρουσιάζουμε το μοντέλο διαδικασίας CRISP-DM και τα βήματα του, όπως αυτά υλοποιούνται μέσα από το περιβάλλον ανάπτυξης του Business Intelligence Development Studio (BIDS) της Microsoft.

Για το βήμα της Κατανόησης των δεδομένων μπορούμε να χρησιμοποιήσουμε τα Analysis Services και να κατασκευάσουμε την προέλευση των δεδομένων που θα χρησιμοποιηθούν από τα μοντέλα Εξόρυξης Γνώσης. Τα δεδομένα σε ένα business intelligence project μέσα από το περιβάλλον του BIDS, υπάρχουν στα Data Source Views τα οποία ουσιαστικά είναι όψεις των πραγματικών πινάκων της Βάσης Δεδομένων που θα χρησιμοποιηθεί.

Για τον βήμα της προετοιμασίας των δεδομένων συνήθως κατασκευάζονται συγκεκριμένα data integration projects μέσω του SSIS (Integration Services) τα οποία μπορεί είτε να ενώνουν δεδομένα από διαφορετικές πηγές, είτε να χρησιμοποιούν συγκεκριμένες

συνθήκες για καθαρισμό των δεδομένων και γενικά περιλαμβάνονται όλες εκείνες οι εργασίες που απαιτούνται για την προετοιμασία των δεδομένων. Επίσης, ανάλογα και με το επιχειρησιακό πρόβλημα που καλείται να επιλύσει το έργο Εξόρυξης Γνώσης, μπορεί να χρησιμοποιηθεί και μια ενδιάμεση λύση Εξόρυξης Γνώσης (μέσω των Analysis Services) για να φιλτράρει συγκεκριμένα δεδομένα. Για παράδειγμα να χρησιμοποιηθεί μια τεχνική ομαδοποίησης (clustering) για να φιλτράρει συγκεκριμένη ομάδα πελατών, που θα χρησιμοποιηθούν στην συνέχεια για τον έργο Εξόρυξης Γνώσης που αναπτύσσεται.



Διάγραμμα 18: Φάσεις του Μοντέλου διαδικασίας CRISP-DM μέσα από το BIDS

Τα βήματα της Δημιουργίας και Αποτίμησης του μοντέλου Εξόρυξης Γνώσης γίνονται καθαρά μέσα από το περιβάλλον του BIDS με την δημιουργία ενός έργου (project) Analysis Services. Μέσα στο project που δημιουργείται, όπως θα δούμε στη συνέχεια, ο χρήστης μπορεί να δημιουργήσει δομές Εξόρυξης Γνώσης (data mining structures) οι οποίες περιέχουν ένα ή περισσότερα μοντέλα Εξόρυξης Γνώσης (data mining models).

Τέλος η εγκατάσταση του μοντέλου μπορεί να πραγματοποιηθεί είτε με την δημιουργία ερωτημάτων DMX, τα οποία θα τρέχουν στον Analysis Server στην εταιρεία είτε να ενσωματωθούν σε μια .NET εφαρμογή, που θα έχει αναπτυχθεί στον οργανισμό για τον σκοπό αυτό. Τα αποτελέσματα των μοντέλων Εξόρυξης Γνώσης μπορούν επίσης να χρησιμοποιηθούν σε αναφορές, που αναπτύσσονται μέσω των reporting services (SSRS) του SQL Server. Τέλος, δίνεται η δυνατότητα στον χρήστη να εισάγει τα μοντέλα Εξόρυξης Γνώσης που έχει δημιουργήσει σε εφαρμογές του Office (κυρίως στο Excel) με την χρήση add-ins που προσφέρει η Microsoft για τον σκοπό αυτό.

Παράδειγμα Στοχευμένης Διαφημιστικής Εκστρατείας

Στη συνέχεια του κεφαλαίου θα παρουσιάσουμε μια εφαρμογή χρήσης των μοντέλων Εξόρυξης Γνώσης μέσα από το περιβάλλον του BIDS παραθέτοντας τα κυριότερα βήματα και τις επιλογές που πρέπει να κάνουμε από την αρχική δημιουργία της προέλευσης δεδομένων που θα χρησιμοποιηθούν μέχρι την πραγματοποίηση προβλέψεων με την χρήση εντολών DMX. Το παράδειγμα αναφέρεται σε μια εταιρεία πώλησης ποδηλάτων και αξεσουάρ ποδηλάτων, που επιθυμεί να δημιουργήσει μια στοχευμένη διαφημιστική εκστρατεία για την προώθηση ποδηλάτων συγκεκριμένων κατηγοριών. Το μοντέλο ανάπτυξης του έργου Εξόρυξης Γνώσης που θα ακολουθηθεί είναι το μοντέλο CRISP-DM που παρουσιάσαμε παραπάνω.

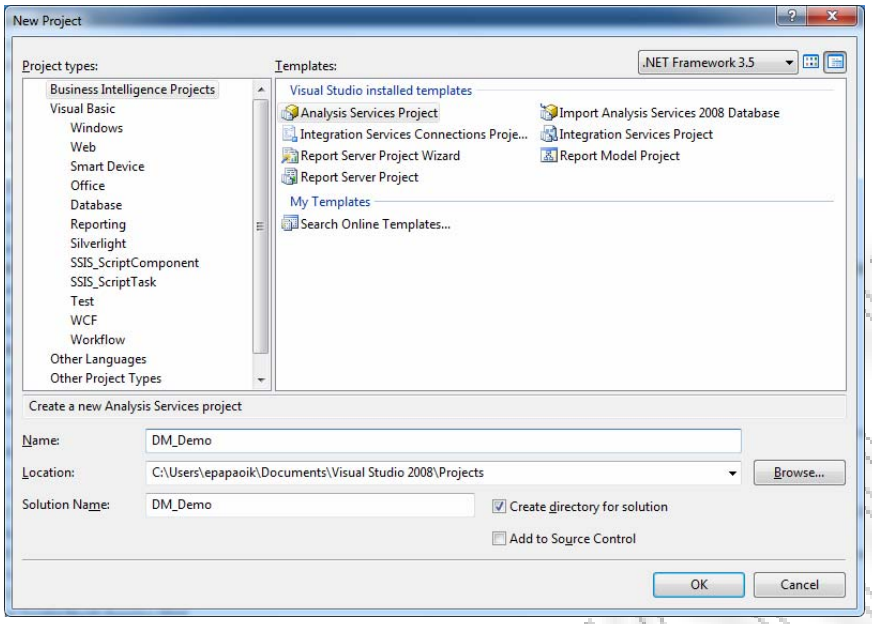
1. Προσδιορισμός του Επιχειρησιακού Προβλήματος

Το πρώτο βήμα της διαδικασίας ανάπτυξης της Εφαρμογής Εξόρυξης Γνώσης είναι ο προσδιορισμός του επιχειρησιακού προβλήματος, που θα επιλυθεί με μια εφαρμογή Εξόρυξης Γνώσης. Το τμήμα μάρκετινγκ της εταιρείας εμπορίας ποδηλάτων που αναφέραμε, ενδιαφέρεται να δημιουργήσει μια διαφημιστική εκστρατεία προώθησης ενός καινούργιου ποδηλάτου από μια συγκεκριμένη κατηγορία ποδηλάτων. Αντί να δημιουργήσει το τμήμα μάρκετινγκ μια διαφημιστική εκστρατεία για όλη την πελατειακή βάση της εταιρείας (πάνω από 18.000 πελάτες), θέλει να αναγνωρίσει ένα πιθανό σύνολο πελατών που είναι περισσότερο πρόθυμοι να αγοράσουν το καινούργιο ποδήλατο.

Το πρόβλημα που αναφέρουμε είναι ένα τυπικό πρόβλημα Κατηγοριοποίησης, όπου προσπαθούμε να κατατάξουμε τους πελάτες της εταιρείας σε μια προκαθορισμένη κατηγορία (αγοραστές ποδηλάτων συγκεκριμένης κατηγορίας). Υπάρχουν διάφοροι αλγόριθμοι επίλυσης του προβλήματος της κατηγοριοποίησης μέσα από το περιβάλλον του BIDS, όπως έχουμε δείξει στο Κεφάλαιο 4. Πιο συγκεκριμένα, οι αλγόριθμοι που θα χρησιμοποιήσουμε είναι ο αλγόριθμος Δέντρων Απόφασης και ο αλγόριθμος Naïve Bayes. Στην συνέχεια του κεφαλαίου, θα δείξουμε ότι μπορούμε να δημιουργήσουμε μια δομή που περιέχει διάφορους αλγορίθμους Εξόρυξης Γνώσης και κατόπιν θα χρησιμοποιήσουμε διάφορες τεχνικές αξιολόγησης των αποτελεσμάτων των διαφόρων αλγορίθμων για να επιλέξουμε την τεχνική με τα καλύτερα αποτελέσματα. Τέλος, θα χρησιμοποιήσουμε τον επιλεγμένο αλγόριθμο για την δημιουργία προβλέψεων σχετικά με την πρόθεση αγοράς ποδηλάτου συγκεκριμένης κατηγορίας.

Στο περιβάλλον BIDS το πρώτο που πρέπει να κάνουμε είναι να δημιουργήσουμε ένα καινούργιο Project για την επίλυση του προβλήματος της Εξόρυξης Γνώσης που αναφέρουμε.

Για να δημιουργήσουμε ένα καινούργιο Business Intelligence Project, ανοίγουμε το Visual Studio 2008 και επιλέγουμε **Create Project**, στο παράθυρο που ανοίγει **επιλέγουμε Business Intelligence Projects** και στην λίστα των Templates **επιλέγουμε Analysis Services Projects**.



Δίνουμε το όνομα DM_Demo στο καινούργιο Project που έχουμε δημιουργήσει και πατάμε **OK** για να το δημιουργήσουμε.

2. Προετοιμασία των Δεδομένων

Για την παρουσίαση των βημάτων χρήσης των τεχνικών Εξόρυξης Γνώσης μέσα από το BIDS θεωρήθηκε σκόπιμο να χρησιμοποιηθεί ένα έτοιμο σύνολο δεδομένων για την δημιουργία των μοντέλων Εξόρυξης Γνώσης. Σκοπός μας είναι να παρουσιάσουμε την χρήση των λειτουργιών του BIDS και τα βήματα που πρέπει να ακολουθήσει ο χρήστης για να ολοκληρώσει ένα πλήρη κύκλο δημιουργίας μιας εφαρμογής Εξόρυξης Γνώσης και όχι η αξιολόγηση των αποτελεσμάτων της εκτέλεσης του κάθε αλγορίθμου. Με λίγα λόγια θέλουμε να δείξουμε πώς σχεδιάζονται και εκτελούνται οι αλγόριθμοι Εξόρυξης Γνώσης και όχι να επιλύσουμε κάποιο πολύπλοκο πρόβλημα και να αναλύσουμε τα αποτελέσματα του αλγορίθμου που χρησιμοποιήθηκε.

Η βάση δεδομένων που θα χρησιμοποιηθεί είναι η Adventure Works βάση δεδομένων που παρέχεται από την Microsoft για εκπαιδευτικούς (και όχι μόνο) σκοπούς. Πιο συγκεκριμένα θα χρησιμοποιηθεί η βάση AdventureWorksDW2008[3] (η έκδοση της βάσης 2008 αναφέρεται στην έκδοση 2008 του SQL Server που έχει χρησιμοποιηθεί για την υλοποίηση της εφαρμογής που παρουσιάζουμε).

Σχετικά με την προέλευση των δεδομένων, επειδή χρειαζόμαστε πληροφορίες για τους πελάτες της εταιρείας και τα προϊόντα, θα χρησιμοποιήσουμε τους πίνακες DimCustomer (για τις πληροφορίες των πελατών) και DimProduct και DimProductSubcategory (για τις πληροφορίες των προϊόντων και κατηγοριών προϊόντων αντίστοιχα). Για την ανάλυση της αγοραστικής συμπεριφοράς των πελατών (τι έχουν αγοράσει οι πελάτες έως τώρα) θα χρειαστούμε τον πίνακα ο οποίος περιέχει τις πληροφορίες πωλήσεων της εταιρείας. Ο πίνακας αυτός είναι ο FactInternetSales, ο οποίος περιέχει πληροφορίες σχετικά με τα προϊόντα που έχουν αγοράσει οι πελάτες.

Για να δημιουργήσουμε την προέλευση δεδομένων στο BIDS, πρέπει να την ορίσουμε στο Project που έχουμε δημιουργήσει στο πρώτο βήμα. Κάτω από την λίστα Data Sources

[3] Για περισσότερες πληροφορίες σχετικά την Βάση Δεδομένων Adventure Works, μπορείτε να επισκεφτείτε την σελίδα http://msftdbprodsamples.codeplex.com/releases/view/37109

πατάμε δεξί κλικ και επιλέγουμε **New Data Source**. Στον wizard που ανοίγει, επιλέγουμε **New data source**, αν δεν υπάρχει κάποια υπάρχουσα σύνδεση. Στο πεδίο **Server Name** δίνουμε το όνομα του server που υπάρχει εγκατεστημένη η Βάση Δεδομένων AdventureWorksDW2008 (για παράδειγμα δίνουμε το **.** ή το localhost ή το όνομα του υπολογιστή μας, αν η Βάση δεδομένων έχει εγκατασταθεί τοπικά στον υπολογιστή μας), στο πεδίο Connect to a database επιλέγουμε την Βάση Δεδομένων AdventureWorksDW2008 και δίνουμε τις κατάλληλες επιλογές για την σύνδεση με την Βάση Δεδομένων (credentials) και πατάμε OK για να δημιουργηθεί η προέλευση Δεδομένων στο project που δουλεύουμε.

3. Δημιουργία του Σχήματος Δεδομένων

Το σχήμα δεδομένων που θα χρησιμοποιήσουμε είναι ουσιαστικά κάποια views των πινάκων από την Βάση Δεδομένων, που θα χρησιμοποιηθούν στο project που δουλεύουμε. Για να δημιουργήσουμε ένα καινούργιο view, επιλέγουμε με δεξί κλικ στην λίστα Data Source Views το **New Data Source View**. Στον wizard που ανοίγει, το πρώτο βήμα είναι να επιλέξουμε την προέλευση δεδομένων που δημιουργήσαμε στο προηγούμενο Βήμα. Στη συνέχεια, στην λίστα που εμφανίζεται καλούμαστε να επιλέξουμε τους πίνακες που θα χρησιμοποιήσουμε για να αντλήσουμε τα δεδομένα μας. Για παράδειγμα, μπορούμε να επιλέξουμε μόνο τον πίνακα DimCustomer και να πατήσουμε **Finish** για να κλείσει ο wizard και να μπούμε στην σελίδα σχεδιασμού του data view. Το όνομα που έχουμε δώσει στο καινούργιο data source view είναι το **Bike Buyers.dsv**.

Τα δεδομένα που θα χρησιμοποιήσουμε, όπως περιγράψαμε στο προηγούμενο βήμα, θα προέρχονται από διαφορετικούς πίνακες τους οποίους θα πρέπει να συνδέσουνε κατάλληλα για να πάρουμε τις πληροφορίες που θέλουμε. Σε γενικές γραμμές, θέλουμε να πάρουμε μια λίστα με τους πελάτες οι οποίοι έχουν αγοράσει ποδήλατα από την εταιρεία. Αυτήν την πληροφόρηση σε συνδυασμό με τα χαρακτηριστικά των πελατών θα τα χρησιμοποιήσουμε για να αναλύσουμε την αγοραστική συμπεριφορά των πελατών για να προβλέψουμε την συμπεριφορά νέων πελατών. Τις πληροφορίες που θέλουμε να χρησιμοποιήσουμε μπορούμε να τις πάρουμε από την Βάση Δεδομένων με την χρήση ερωτήματος SQL.

Μέσα στον designer του data source view ο χρήστης έχει την επιλογή να χρησιμοποιήσει πίνακες από την Βάση δεδομένων (στην πραγματικότητα δεν είναι οι πραγματικοί πίνακες της Βάσης δεδομένων αλλά όψεις (views) αυτών – έτσι οποιαδήποτε αλλαγή γίνεται στο BIDS δεν επηρεάζει την πραγματική Βάση Δεδομένων) ή να εισάγει ερωτήματα SQL τα οποία χρησιμοποιούν δεδομένα από διάφορους πίνακες της Βάσης. Για να δημιουργήσουμε ένα δικό μας SQL ερώτημα, κάνουμε δεξί κλικ στην επιφάνεια του editor και επιλέγουμε **New Named Query**.

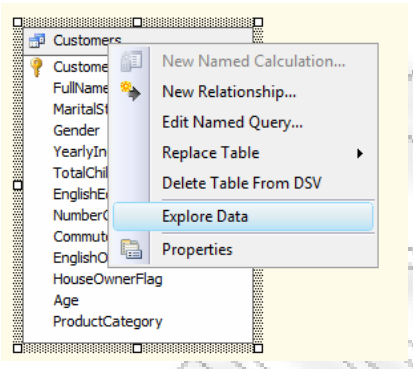
Στο παράθυρο που ανοίγει, σαν όνομα δίνουμε το **Customers** και στο πλαίσιο Query Definition δίνουμε το παρακάτω SQL ερώτημα:

```
SELECT C.CustomerKey
, C.FirstName + ' ' + ISNULL(C.MiddleName + '. ', '') + C.LastName AS FullName
, C.MaritalStatus
, C.Gender
, C.YearlyIncome
, C.TotalChildren
, C.EnglishEducation
, C.NumberCarsOwned
, C.CommuteDistance
, C.EnglishOccupation
, C.HouseOwnerFlag
, DATEDIFF(yy, C.BirthDate, GETDATE()) AS Age
```

```
, CustomerFilter.Subcategory AS ProductCategory
FROM DimCustomer AS C
INNER JOIN (SELECT C.CustomerKey, PS.EnglishProductSubcategoryName AS Subcategory
        FROM DimCustomer AS C
        INNER JOIN FactInternetSales AS S ON C.CustomerKey = S.CustomerKey
        INNER JOIN DimProduct AS P ON S.ProductKey = P.ProductKey
        INNER JOIN DimProductSubcategory AS PS ON P.ProductSubcategoryKey =
PS.ProductSubcategoryKey
        WHERE (PS.ProductCategoryKey = 1)
        GROUP BY C.CustomerKey, PS.EnglishProductSubcategoryName
        HAVING (COUNT(PS.ProductSubcategoryKey) = 1)
       ) AS CustomerFilter ON C.CustomerKey = CustomerFilter.CustomerKey
```

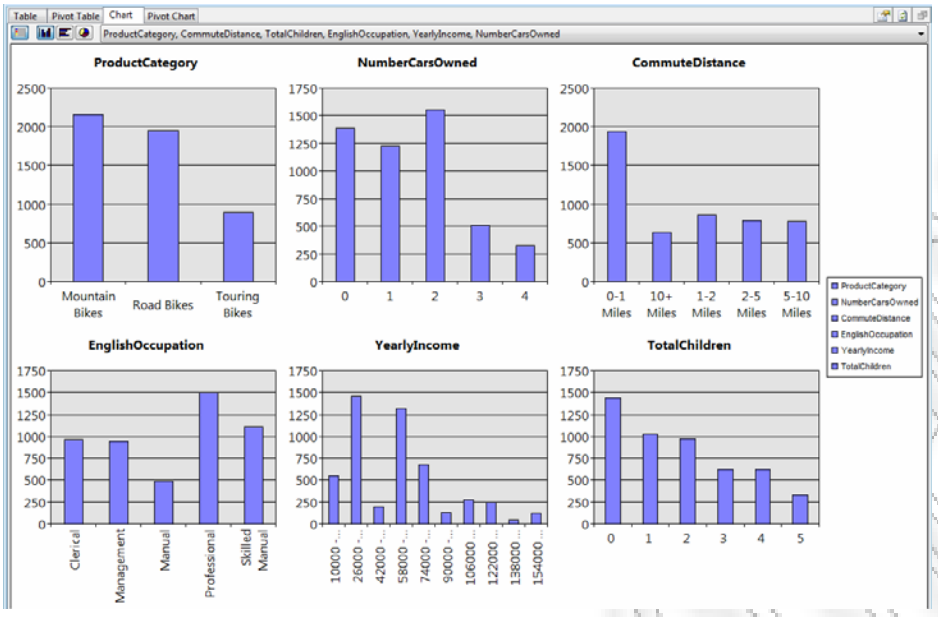
Στο παραπάνω ερώτημα, έχουμε χρησιμοποιήσει την συνθήκη να μας επιστραφούν μόνο οι πελάτες που έχουν αγοράσει ποδήλατα (PS.ProductCategoryKey = 1) και τον περιορισμό να μας επιστραφούν οι πελάτες που έχουν αγοράσει ακριβώς ένα ποδήλατο (HAVING (COUNT(PS.ProductSubcategoryKey) = 1)) από την εταιρεία και εξαιρεί τους πελάτες που έχουν αγοράσει περισσότερες από μια φορές ποδήλατα από την εταιρεία. Αυτό γίνεται καθαρά για χάριν ευκολίας και για να μειωθεί ο αριθμός των επιστρεφόμενων αποτελεσμάτων.

Στο data source view που δημιουργείται, μπορούμε πατώντας δεξί κλικ να επιλέξουμε Explore Data



Στο παράθυρο που εμφανίζεται μπορούμε να δούμε πολλές πληροφορίες σχετικά με τα δεδομένα του data view που δημιουργήσαμε.

Αρχικά μπορούμε να δούμε τα ίδια τα δεδομένα (**tab Table**) έτσι ώστε πιστοποιήσουμε ότι έχουμε δημιουργήσει το σωστό SQL ερώτημα και παίρνουμε τα σωστά αποτελέσματα. Επίσης, στο tab **Chart** μπορούμε να δούμε πληροφορίες για τα κυριότερα χαρακτηριστικά των δεδομένων που έχουμε στην διάθεσή μας. Μέσα από τα διαγράμματα αυτά μπορούμε να δούμε τις τιμές των χαρακτηριστικών που έχουμε σε διάφορες όψεις (για παράδειγμα Bar Chart, Column Chart, Pie Chart) και να εντοπίσουμε περιπτώσεις όπως δεν έχουμε επαρκή πληροφόρηση ή μονόπλευρη πληροφόρηση (για παράδειγμα χαρακτηριστικά με εύρος τιμών σε μια συγκεκριμένη περιοχή μόνο ή χαρακτηριστικά με ελάχιστες διαφορετικές τιμές). Αυτά τα χαρακτηριστικά μπορούμε εύκολα να τα εντοπίσουμε και να τα αποκλείσουμε τελικά από το data view ή να τροποποιήσουμε το SQL ερώτημα για να πάρουμε πιο σωστά αποτελέσματα.

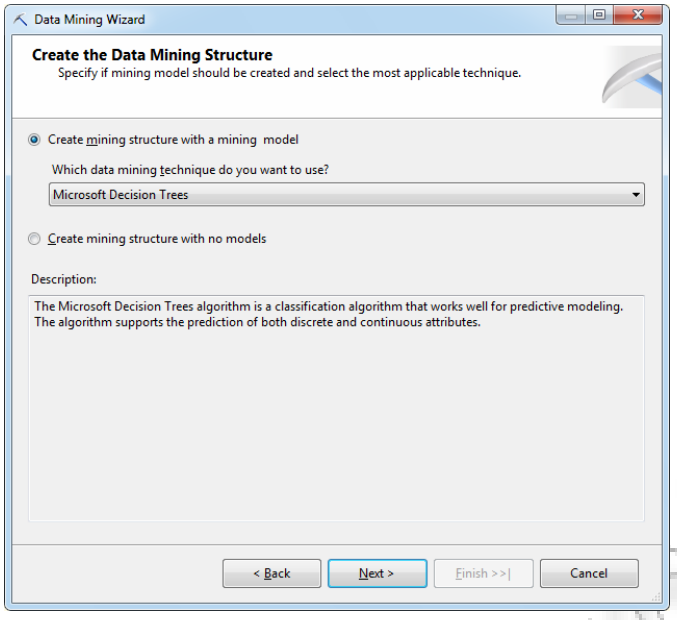


Εικόνα 1: Εξερεύνηση Δεδομένων

Οι υπόλοιπες επιλογές **Pivot Table** και **Pivot Chart** είναι επίσης πολύ χρήσιμες στην εξερεύνηση των δεδομένων που έχουμε στην διάθεσή μας, κυρίως σε περιπτώσεις που έχουμε και στήλες με αριθμητικές πληροφορίες (για παράδειγμα ποσότητες πωλήσεων προϊόντων, τιμές κλπ). Στην περίπτωση των δεδομένων που αναφέρουμε στο παράδειγμα βέβαια, δεν έχουμε στήλη με στοιχεία πωλήσεων (έχουμε μόνο την ονομασία της κατηγορίας ποδηλάτου που έχει αγοραστεί από τον πελάτη και δεν δείχνουμε για παράδειγμα την αξία αγοράς ή τις ποσότητες προϊόντων που αγόρασαν οι πελάτες) και έτσι οι επιλογές αυτές ουσιαστικά δεν θα μας δείξουν κάποια σημαντική πληροφόρηση. Ωστόσο η εξερεύνηση των δεδομένων είναι ένα πολύ σημαντικό εργαλείο που παρέχεται στον χρήστη για να κατανοήσει τα δεδομένα που έχει στην διάθεση του.

4. Κατασκευή του Μοντέλου

Ο πιο εύκολος τρόπος για να δημιουργήσουμε ένα μοντέλο Εξόρυξης Γνώσης σε μια δομή Εξόρυξης Γνώσης είναι να πατήσουμε στο μενού Mining Structure δεξί κλικ και να επιλέξουμε **New Mining Structure**. Στο παρακάτω παράθυρο που ανοίγει, δημιουργούμε μια καινούργια δομή Εξόρυξης Γνώσης επιλέγοντας για αυτή ένα καινούργιο μοντέλο δεδομένων.

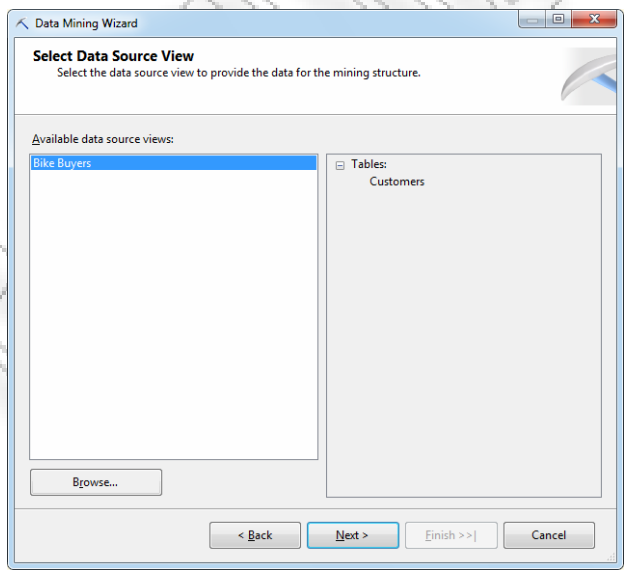


Εικόνα 2: Επιλογή Μοντέλου Εξόρυξης Γνώσης

Στο παράθυρο που εμφανίζεται, μπορούμε να επιλέξουμε από την διαθέσιμη λίστα ένα μοντέλο δεδομένων ή να δημιουργήσουμε μια κενή δομή χωρίς κάποιο συγκεκριμένο μοντέλο σε αυτή. Μπορούμε πάντα να προσθέσουμε ένα καινούριο μοντέλο δεδομένων σε μια δομή που έχουμε δημιουργήσει. Επιλέγουμε τον αλγόριθμο **Microsoft Decision Trees**.

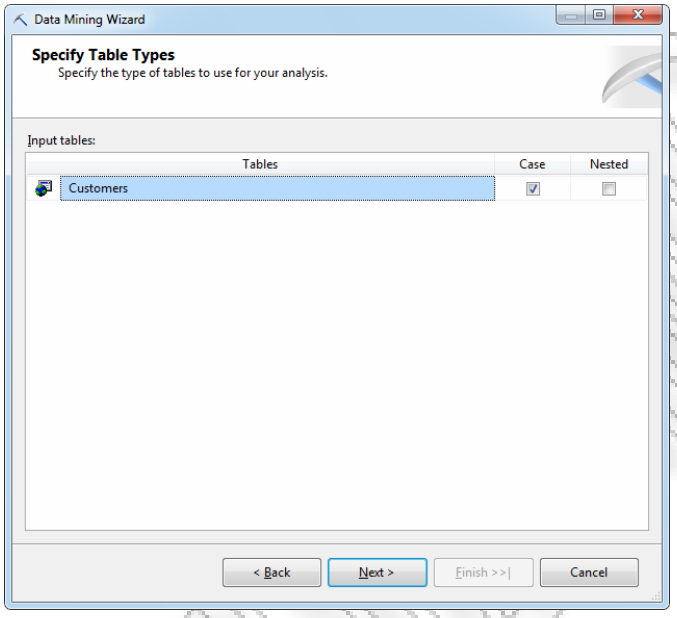
Επιλογή Προέλευσης Δεδομένων

Αφού διαλέξουμε το μοντέλο Εξόρυξης Γνώσης που θα χρησιμοποιήσουμε, το επόμενο βήμα είναι να επιλέξουμε την προέλευση δεδομένων (data source view), από την οποία θα αντλήσει δεδομένα το μοντέλο μας. Στην προκειμένη περίπτωση επιλέγουμε το view Bike Buyers, που έχουμε δημιουργήσει και βλέπουμε ότι το view αυτό περιέχει τον πίνακα **Customers**.



Επιλογή Τύπου Πίνακα Δεδομένων

Αφού επιλέξουμε τα data source view του μοντέλου μας, στο επόμενο παράθυρο πρέπει να καθορίσουμε τον τύπο του πίνακα δεδομένων, που θα εισάγουμε στο μοντέλο μας. Οι διαθέσιμες επιλογές που έχουμε είναι ο τύπος **Case** και ο τύπος **Nested**. Ο τύπος Case περιέχει τα κύρια δεδομένα, που θα χρησιμοποιήσουμε στο μοντέλο μας. Ο τύπος Nested χρησιμοποιείται στην περίπτωση που έχουμε περισσότερους από ένα πίνακες, που περιέχουν δεδομένα για να τροφοδοτήσουμε το μοντέλο μας. Στην περίπτωση που μελετάμε, για τον αλγόριθμο με Δέντρα Απόφασης δεν χρειαζόμαστε nested πίνακα, οπότε τον πίνακα Customers που θα χρησιμοποιήσουμε θα τον ορίσουμε ως case πίνακα. Ένα παράδειγμα χρήσης nested πίνακα είναι ο αλγόριθμος Κανόνων Συσχέτισης για να βρούμε κανόνες ανάμεσα σε προϊόντα τα οποία πωλούνται μαζί. Έτσι για παράδειγμα, θα μπορούσαμε να χρησιμοποιήσουμε ένα βασικό πίνακα με τα στοιχεία των προϊόντων ή των πελατών και ένα εσωτερικό πίνακα που περιέχει τα στοιχεία των πωλήσεων για τα προϊόντα αυτά.



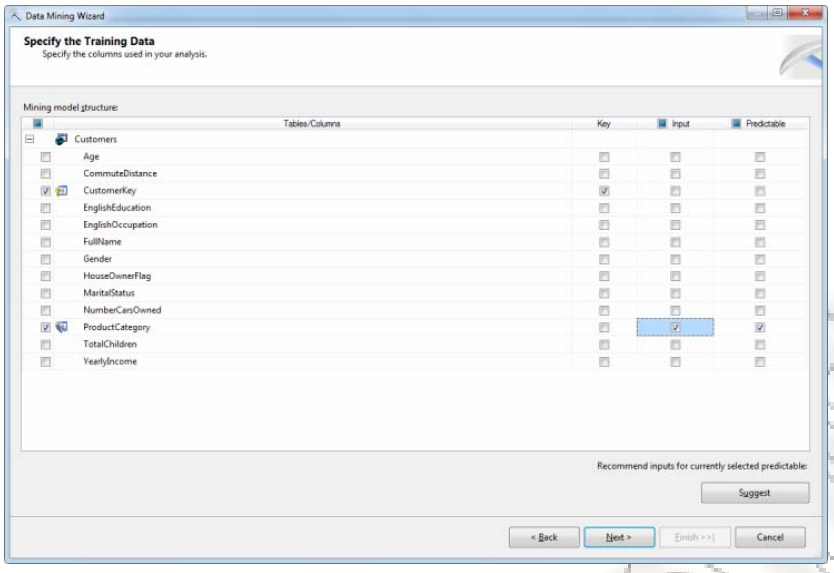
Εικόνα 3: Ορισμός Τύπου Πίνακα Δεδομένων

Σημείωση: σε ένα μοντέλο δεδομένων μπορούμε να ορίσουμε μόνο ένα Case table.

Ορισμός Στηλών Δεδομένων

Στην συνέχεια του ορισμού του μοντέλου, πρέπει να ορίσουμε τις στήλες δεδομένων, τι οποίες θα χρησιμοποιήσουμε στο μοντέλο. Οι διαθέσιμες επιλογές για τις στήλες δεδομένων είναι οι:

- **Key**: στήλη που προσδιορίζει μοναδικά την κάθε γραμμή δεδομένων.
- **Input**: στήλη που τα δεδομένα της θα χρησιμοποιηθούν σαν είσοδος στο μοντέλο.
- **Predictable**: στήλη που περιέχει δεδομένα για τα οποία ζητείται η πρόβλεψη από το μοντέλο.

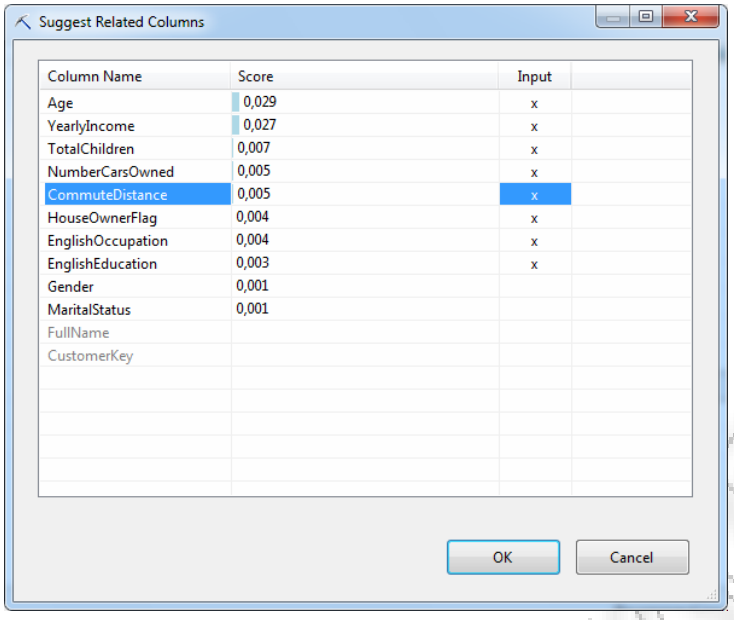


Εικόνα 4: Επιλογή στηλών Δεδομένων για εισαγωγή στο Μοντέλο

Στο παράδειγμα μας, θα ορίσουμε σαν Key στήλη την στήλη *CustomerKey*, η οποία προσδιορίζει μοναδικά την κάθε γραμμή δεδομένων (αφού περιέχει το αναγνωριστικό του κάθε πελάτη). Επειδή ο πίνακας δεδομένων περιέχει στοιχεία αγορών προϊόντων από τους πελάτες, ενδεχομένως να περιέχονται περισσότερες από μια εγγραφές για κάθε πελάτη (ο οποίος πιθανόν να έχει αγοράσει περισσότερες από μια φορές από το κατάστημα μας). Ο ορισμός της στήλης κλειδιού *CustomerKey* δεν παρουσιάζει προβλήματα ακεραιότητας (διπλές εγγραφές στον πίνακα) όπως συμβαίνει στις σχεσιακές βάσεις δεδομένων. Όπως θα δούμε όμως στη συνέχεια, μπορούμε να ορίσουμε σαν στήλη Key περισσότερες από μια στήλες δεδομένων, να ορίσουμε με λίγα λόγια ένα σύνθετο κλειδί. Αυτό θα το δούμε στην διαδικασία τροποποίησης του μοντέλου δεδομένων.

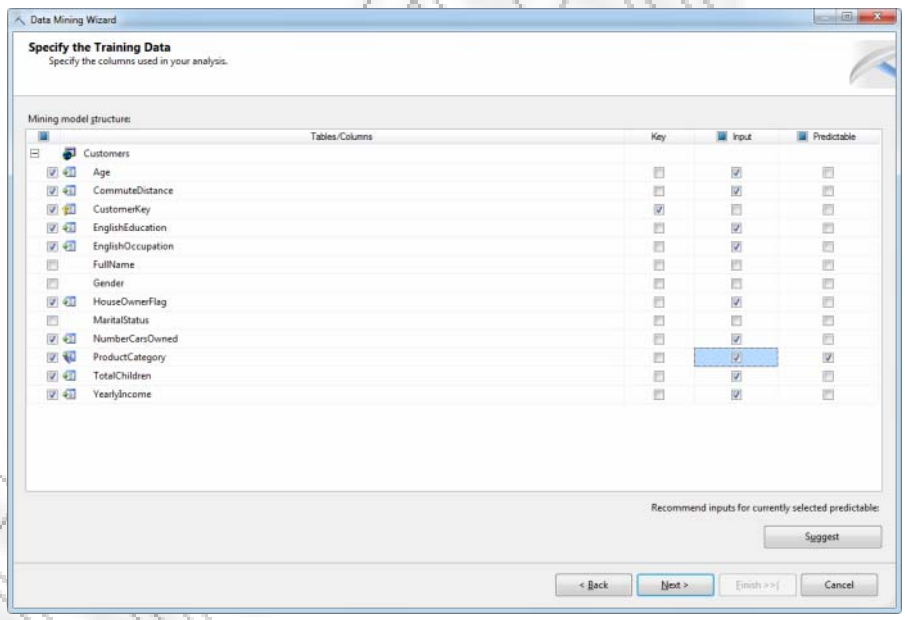
Επειδή θέλουμε ο αλγόριθμος που κατασκευάζουμε να προβλέψει τις κατηγορίες προϊόντων που μπορεί να αγοράσουν οι πελάτες του καταστήματος μας, θα επιλέξουμε την στήλη *ProductCategory* σαν **Predictable**. Επίσης θα επιλέξουμε την στήλη αυτή και σαν **Input**, γιατί θέλουμε ο αλγόριθμος να λάβει υπόψη του και τις παλαιότερες αγορές κατηγοριών προϊόντων από τους πελάτες για να καταλάβει την αγοραστική συμπεριφορά τους. Στην συνέχεια, θέλουμε να ορίσουμε τις στήλες Input που θα χρησιμοποιηθούν στο μοντέλο. Ο ένας τρόπος είναι, στην περίπτωση που γνωρίζουμε καλά τα δεδομένα που έχουμε στην διάθεσή μας και το πώς σχετίζονται μεταξύ τους, να επιλέξουμε τις στήλες που θέλουμε σαν Input. Ένας άλλος τρόπος είναι να επιλέξει ο αλγόριθμος τις στήλες εισόδου για εμάς, κάνοντας μια ανάλυση συσχετίσεων των δεδομένων μεταξύ τους.

Για να επιλέξει αυτόματα ο αλγόριθμος τις στήλες εισόδου, πατάμε στο κουμπί **Suggest** και εμφανίζεται το παρακάτω παράθυρο.



Εικόνα 5: Αυτόματη επιλογή στηλών

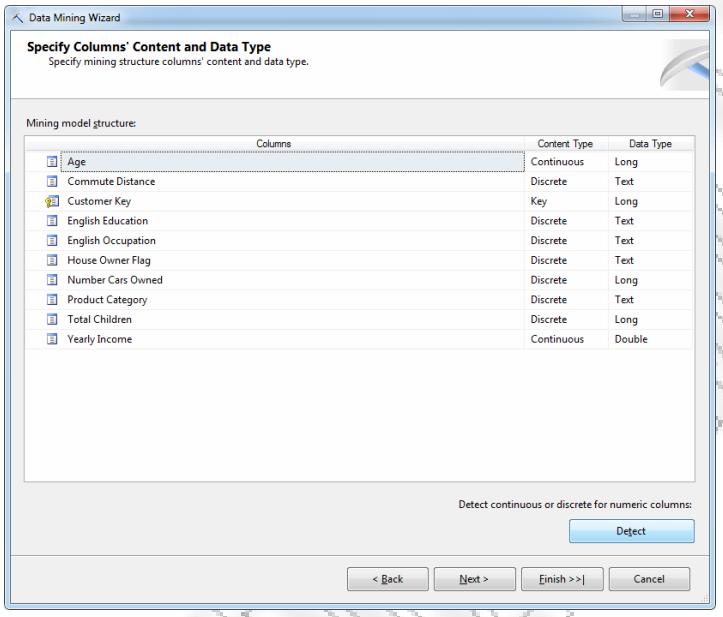
Στο παράθυρο εμφανίζονται οι στήλες δεδομένων του πίνακα και ένα αντίστοιχο score, το οποίο καθορίζει τον βαθμό συσχέτισης της κάθε στήλης σε σχέση με το σύνολο των δεδομένων. Συνήθως επιλέγουμε τις στήλες με score πάνω από το 0 πατώντας στο αντίστοιχο κελί της στήλης Input. Πατάμε OK και βλέπουμε το παράθυρο επιλογής στηλών του μοντέλου μας.



Εικόνα 6: Προβολή στηλών Δεδομένων που θα χρησιμοποιηθούν στο Μοντέλο

Ορισμός Περιεχομένου και Τύπου Στηλών Δεδομένων

Το επόμενο βήμα στην κατασκευή του μοντέλου Εξόρυξης Γνώσης είναι να ορίσουμε τον τύπο των περιεχομένων (content type) της κάθε στήλης και τον τύπο (data type) της κάθε στήλης δεδομένων. Οι διαθέσιμοι τύποι δεδομένων, όπως τους έχουμε αναφέρει και στο κεφάλαιο 4, είναι οι συνηθισμένοι τύποι δεδομένων όπως Long, Text, Double, Date κα. Ο τύπος περιεχομένων των δεδομένων ορίζει αν τα δεδομένα είναι διακριτά (Discrete), αν είναι συνεχής τιμή (Continuous), αν η στήλη δεδομένων ορίζει το κλειδί (Key) κλπ. Ένας άλλος τύπος, όπως θα δούμε στην συνέχεια, είναι να μετατρέψουμε τα συνεχή δεδομένα ή τα δεδομένα κειμένου σε διακριτές τιμές, έτσι ώστε να είναι αποδεκτές από διάφορους αλγορίθμους Εξόρυξης Γνώσης, που έχουν περιορισμούς όσον αφορά το περιεχόμενο των δεδομένων. Για παράδειγμα, ο αλγόριθμος Naïve Bayes δεν δέχεται συνεχείς τιμές. Για το λόγο μπορούμε να μετατρέψουμε κάποιες στήλες που θέλουμε σε διακριτές τιμές ορίζοντας το μέγεθος των ομάδων τιμών στις οποίες θα χωριστούν τα δεδομένα καθώς επίσης και τον τρόπο, με τον οποίο θα χωριστούν τα δεδομένα σε ομάδες έτσι ώστε να δημιουργηθούν οι διακριτές τιμές.

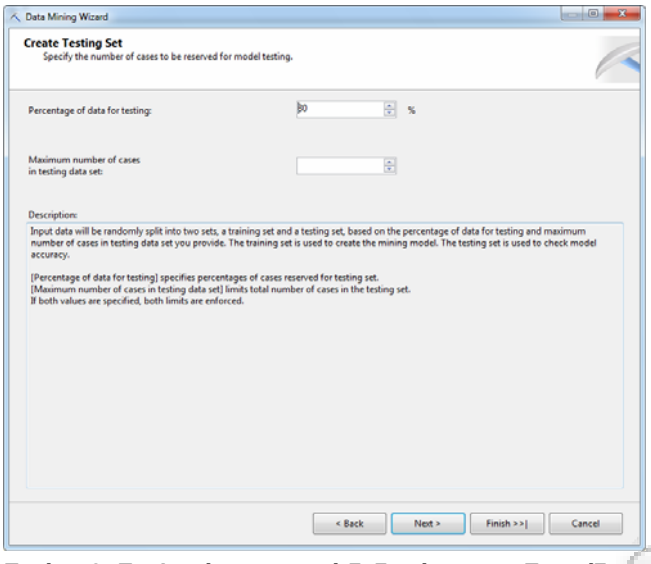


Εικόνα 7: Καθορισμός Τύπου Περιεχομένων (Content) και Τύπου Δεδομένων (Data)

Πατώντας στην επιλογή **Detect**, ο αλγόριθμος αναλύει αυτόματα τον τύπο των περιεχομένων των δεδομένων.

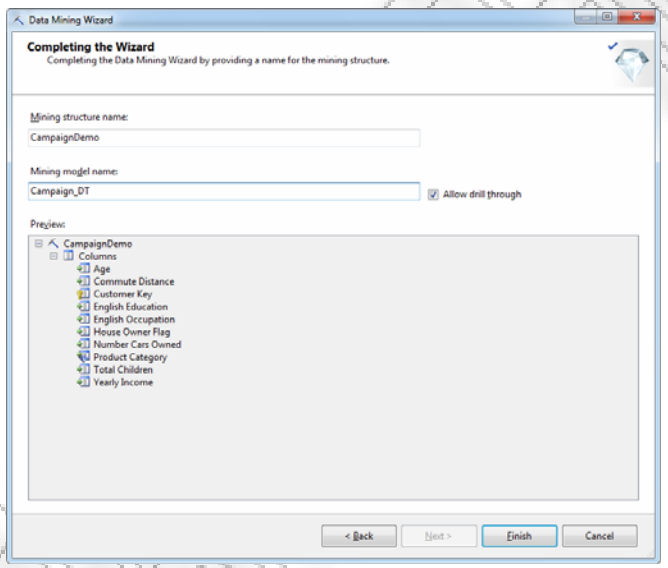
Ορισμός Συνόλου Εκπαίδευσης

Στο τελευταίο βήμα του καθορισμού τον μοντέλου Εξόρυξης Γνώσης, μπορούμε να δώσουμε τον αριθμό των δεδομένων, τα οποία θα χρησιμοποιηθούν για την εκπαίδευση του αλγορίθμου. Ο χρήστης μπορεί να δώσει ένα ποσοστό επί του συνολικού αριθμού των δεδομένων ή να δώσει ένα συγκεκριμένο αριθμό που καθορίζει το πλήθος δεδομένων εκπαίδευσης. Η συνηθισμένη τιμή που δίνουμε είναι να χρησιμοποιηθεί το 30% επί του συνολικού αριθμού των δεδομένων για την εκπαίδευση του μοντέλου Εξόρυξης Γνώσης.



Εικόνα 8: Επιλογή ποσοστού δεδομένων για Εκπαίδευση του Μοντέλου

Στο τελευταίο παράθυρο που εμφανίζεται, ο χρήστης δίνει το όνομα της Δομής Εξόρυξης Γνώσης (Mining Structure Name) και το όνομα του Μοντέλου Εξόρυξης Γνώσης (Mining Model Name) που δημιουργεί ο χρήστης. Στο παράδειγμα μας δίνουμε το όνομα **CampaignDemo** για την δομή και το όνομα **Campaign_DT** (για Decision Trees – Δέντρα Απόφασης) για το μοντέλο μας. Επίσης επιλέγουμε **Allow drill through**. Αυτή είναι μια αρκετά χρήσιμη επιλογή, που θα δούμε στη συνέχεια.

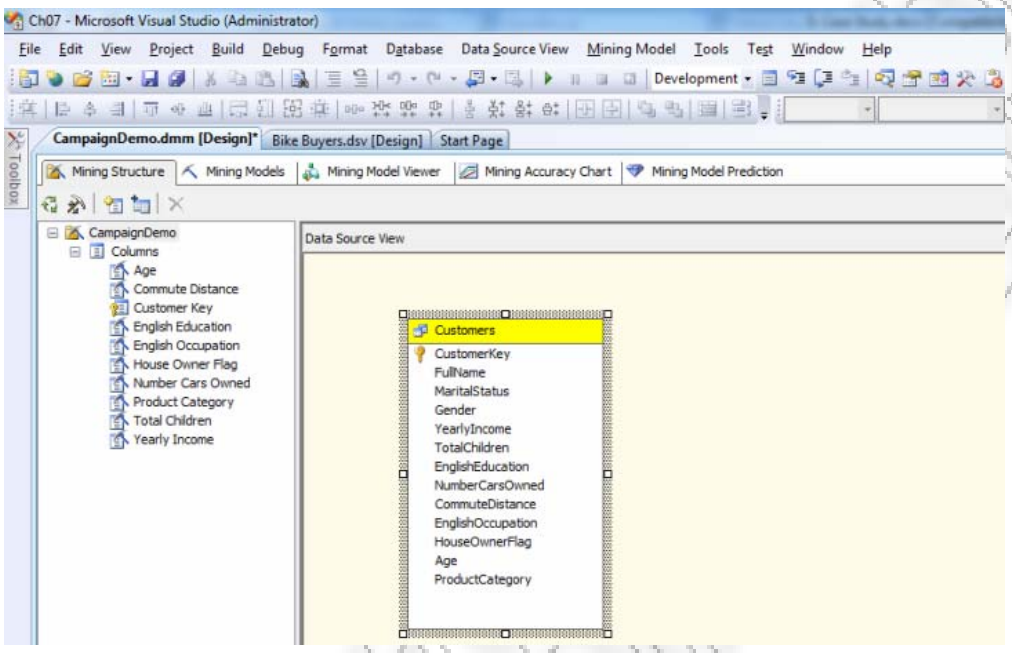


Εικόνα 9: Αποθήκευση Δομής (Structure) και Μοντέλου (Model) Εξόρυξης Γνώσης

Όπως παρατηρούμε η αρχική δημιουργία της δομής και του μοντέλου Εξόρυξης Γνώσης ορίζει μια σχέση ένα-προς-ένα για την δομή και το μοντέλο που δημιουργείται. Όπως θα δούμε στην συνέχεια, μπορούμε σε μια δομή Εξόρυξης Γνώσης να προσθέσουμε νέα μοντέλα Εξόρυξης Γνώσης, τα οποία θα χρησιμοποιούν τις ρυθμίσεις (Πίνακας Δεδομένων, Στήλες Δεδομένων που θα χρησιμοποιηθούν, Περιεχόμενο και Τύπος δεδομένων των Στηλών αυτών) που έχει ορίσει ο χρήστης στα προηγούμενα βήματα που παρουσιάσαμε. Μπορεί έτσι

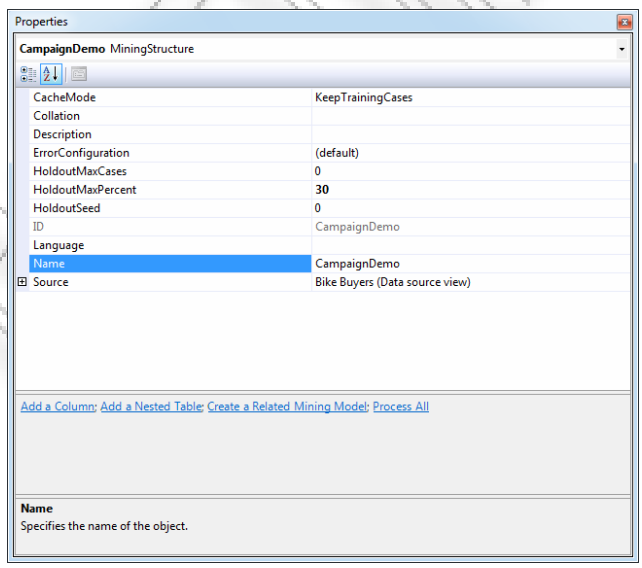
εύκολα ο χρήστης να χρησιμοποιήσει πολλά διαφορετικά μοντέλα Εξόρυξης Γνώσης για τις ίδιες ρυθμίσεις, να παράγει να αντίστοιχα αποτελέσματα του κάθε μοντέλου, να τα συγκρίνει μεταξύ τους και στην συνέχεια να επιλέξει το μοντέλο Εξόρυξης Γνώσης με τα καλύτερα αποτελέσματα για να το χρησιμοποιήσει για την πρόβλεψη που επιθυμεί.

Αφού ο χρήστης δώσει τα ονόματα που επιθυμεί στην δομή και το μοντέλο Εξόρυξης Γνώσης, ανοίγει το παράθυρο σχεδιασμού της δομής αυτής και των μοντέλων που περιέχει. Το παράθυρο σχεδιασμού (Mining Structure designer) παρουσιάζεται παρακάτω.



Εικόνα 10: Παράθυρο Data Mining Designer (Mining Structure)

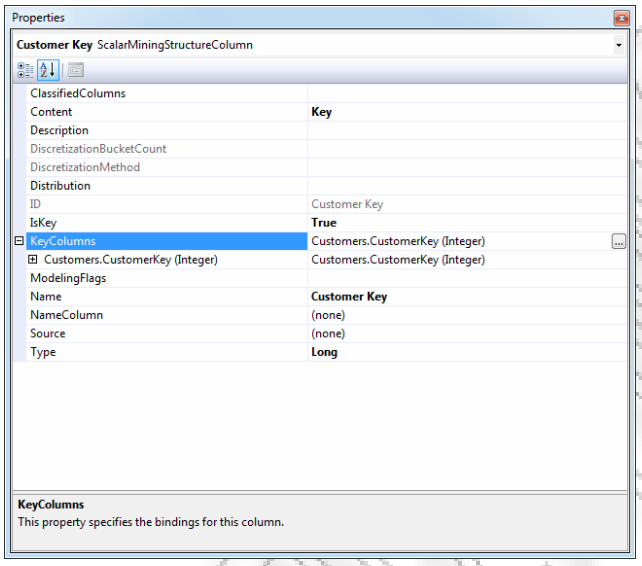
Για την δομή Εξόρυξης Γνώσης που εμφανίζεται στο παράθυρο σχεδιασμού, ο χρήστης έχει πολλές επιλογές στην διάθεση του μέσα από το παράθυρο ιδιοτήτων της δομής (το παράθυρο ιδιοτήτων της δομής εμφανίζεται πατώντας πάνω στη δομή και επιλέγοντας δεξί κλικ – Ιδιότητες ή πατώντας το F4).



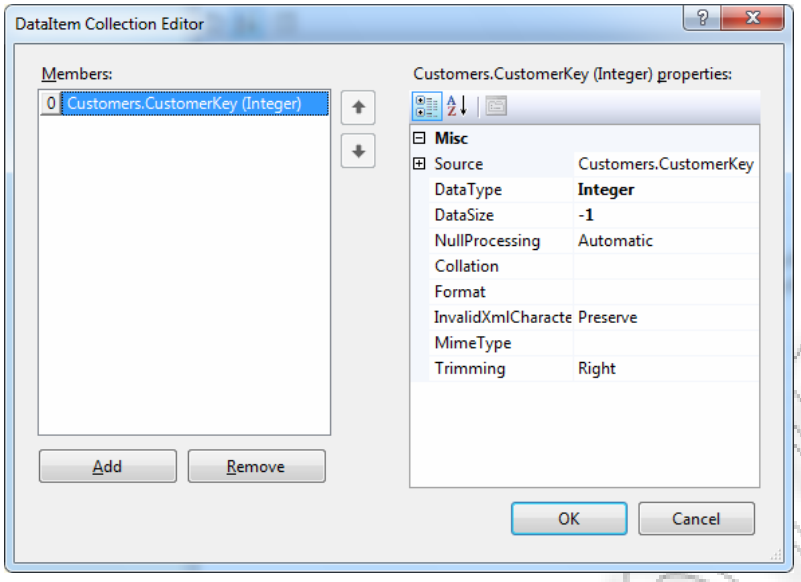
Δημιουργία Σύνθετου Κλειδιού

Μια χρήσιμη επιλογή που μας παρέχεται από την εφαρμογή είναι η τροποποίηση του κλειδιού των δεδομένων, έτσι ώστε να απεικονίζει πιο σωστά τα αποτελέσματα. Με την υπάρχουσα σχεδίαση του μοντέλου το πεδίο *CustomerKey* έχει οριστεί σαν κλειδί των δεδομένων, πράγμα που σημαίνει ότι επιλέγονται οι μοναδικοί πελάτες, οι οποίοι έχουν αγοράσει ένα προϊόν από την εταιρεία. Αυτό όμως αποκλείει και τις περιπτώσεις που κάποιοι πελάτες έχουν πραγματοποιήσει περισσότερες από μια αγορές διαφορετικών προϊόντων από την εταιρεία. Με τον ορισμό, λοιπόν του πεδίου *CustomerKey* σαν κλειδί επιστρέφονται οι μοναδικοί πελάτες, για το λόγο αυτό χρειαζόμαστε ένα σύνθετο τύπου κλειδιού, όπως για παράδειγμα την στήλη του κλειδιού *CustomerKey* και την στήλη που ορίζει την κατηγορία προϊόντων (την *ProductCategory*).

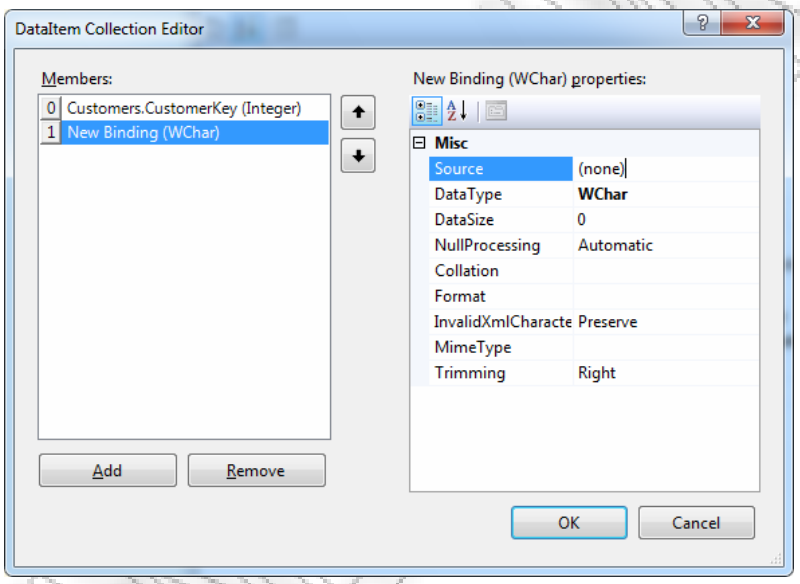
Στο tab Mining Structure, επιλέγουμε στο μοντέλο την στήλη *CustomerKey* και πατάμε F4 για να ανοίξει το παράθυρο των ιδιοτήτων της στήλης.



Στην επιλογή KeyColumns πατάμε το  για να ανοίξει το πλαίσιο επιλογή στηλών δεδομένων (**DataItem Collection Editor**).

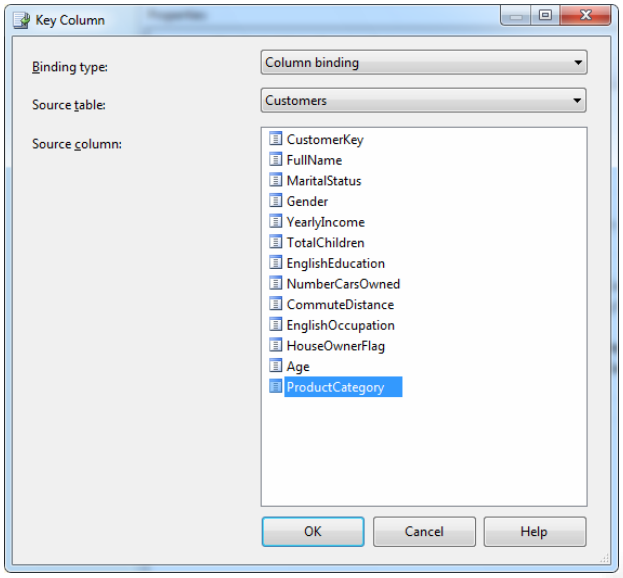


Όπως βλέπουμε, είναι επιλεγμένη η στήλη *CustomerKey* στον ορισμό του κλειδιού. Επιλέγουμε **Add** για να προσθέσουμε μια επιπλέον στήλη για να δημιουργήσουμε το σύνθετο κλειδί.



Πατάμε **OK** και στην συνέχεια πρέπει να συνδέσουμε την νέα στήλη με μια στήλη από τον πίνακα δεδομένων (στην προκειμένη περίπτωση το *ProductCategory*).

Στο παράθυρο των ιδιοτήτων, κάτω από την επιλογή KeyColumns, επιλέγουμε την καινούργια στήλη που δημιουργήσαμε (με το όνομα New Binding) και πατάμε το  Στο παράθυρο που ανοίγει, στο πεδίο Binding Type επιλέγουμε **Column Binding** και στην λίστα διαλέγουμε την στήλη *ProductCategory* και επιλέγουμε **OK**.

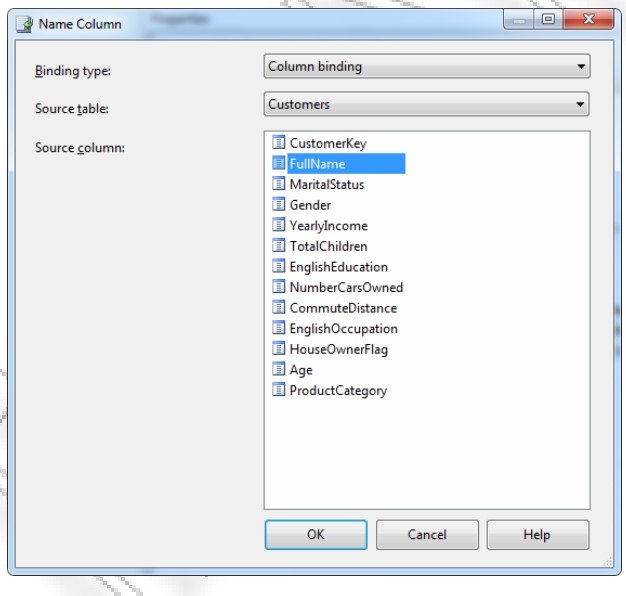


Εικόνα 11: Επιλογή στήλης για δημιουργία σύνθετου κλειδιού.

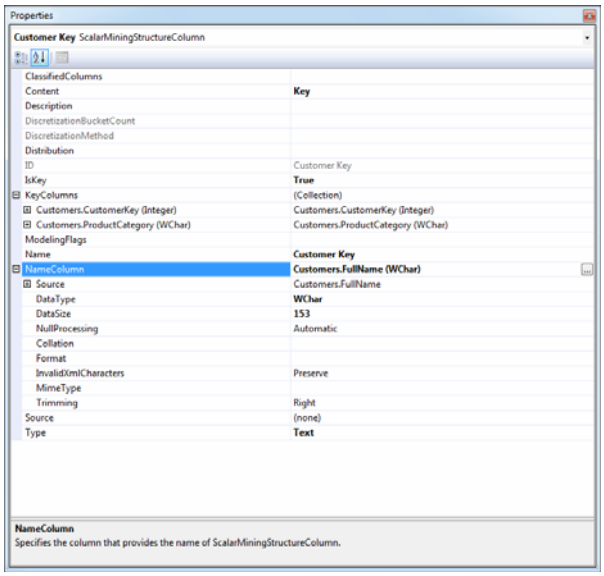
Επιλογή Στήλης για Προβολή Ονόματος

Επίσης, μια επιπλέον επιλογή που έχουμε είναι να μπορούμε να χρησιμοποιήσουμε μια διαφορετική στήλη για προβολή δεδομένων, αντί για μια στήλη με αριθμητικά δεδομένα (πχ την στήλη του κλειδιού). Για παράδειγμα, αντί για το κωδικό του πελάτη, θα θέλαμε να δείξουμε το όνομά του. Αυτή η δυνατότητα δίνεται, στο παράθυρο ιδιοτήτων της στήλης, κάτω από την επιλογή **NameColumn**.

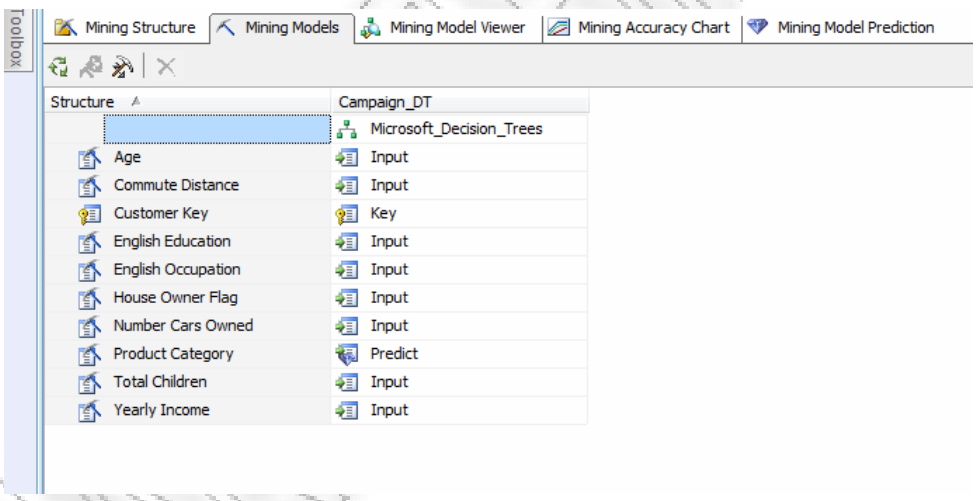
Πατώντας το  , ανοίγει το παράθυρο **Name Column**, για να διαλέξουμε την στήλη που θα παρουσιάζεται, αντί του κωδικού των πελατών.



Επιλέγουμε το *FullName* και πατάμε **OK**. Το παράθυρο των ιδιοτήτων της στήλης *CustomerKey* θα πρέπει να είναι όπως το παρακάτω:



Στο tab Mining Models μπορούμε να δούμε τα Μοντέλα Εξόρυξης Γνώσης που υπάρχουν μέσα στην δομή. Επειδή η δομή έχει δημιουργηθεί από την αρχή, όπως την παρουσιάσαμε στα προηγούμενα βήματα, έχει δημιουργηθεί μόνο ένα μοντέλο στην δομή μας. Τα περιεχόμενα του tab Mining Models εμφανίζονται στο παρακάτω παράθυρο. Στο παράθυρο αυτό βλέπουμε τις στήλες δεδομένων που έχουν οριστεί για το μοντέλο (κάτω από την στήλη με όνομα Structure) και κάτω από την στήλη με το όνομα του μοντέλου (στο παράδειγμα μας κάτω από την στήλη με όνομα **Campaign_DT**) βλέπουμε την χρήση (usage) που έχει οριστεί για την κάθε στήλη, πως θα διαχειριστεί δηλαδή την κάθε στήλη δεδομένων το μοντέλο που βλέπουμε.



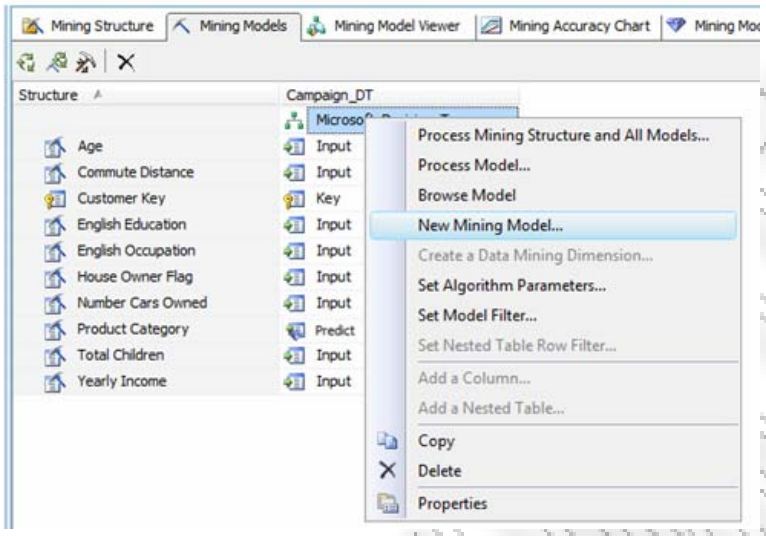
Εικόνα 12: Παράθυρο Data Mining Designer (tab Mining Models)

Αλλαγή Χρήσης (Usage) των Στηλών

Στο tab **Mining Models** βλέπουμε τα μοντέλα Εξόρυξης Γνώσης που έχουμε ορίσει στην δομή δεδομένων. Μπορούμε για κάθε στήλη δεδομένων να αλλάξουμε την χρήση του (usage) επιλέγοντας την στήλη και στην λίστα που εμφανίζεται να αλλάξουμε τιμή (για παράδειγμα για μια στήλη από **Input** να επιλέξουμε **Ignore** (αν θέλουμε να αγνοήσει το μοντέλο

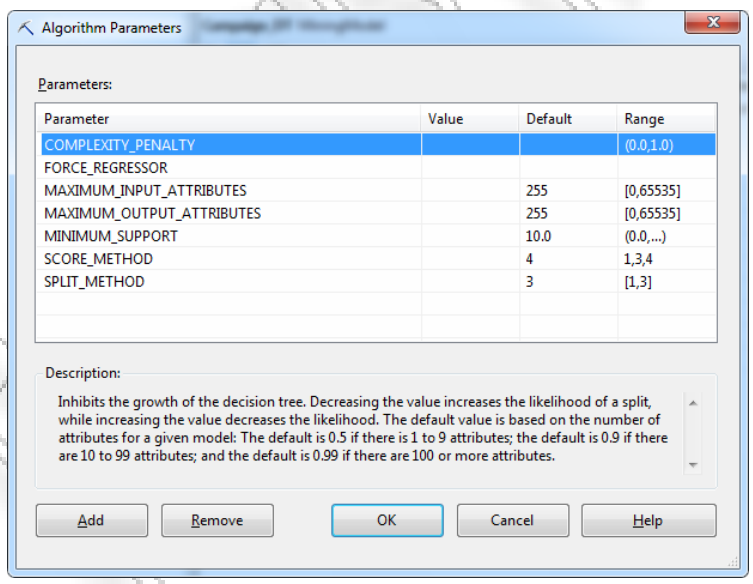
την συγκεκριμένη στήλη) ή να επιλέξουμε **Predict** (για να γίνει η πρόβλεψη για την επιλεγμένη στήλη).

Επίσης, με δεξί κλικ στο όνομα του μοντέλου, έχουμε διάφορες επιλογές στην διάθεσή μας, όπως τον καθορισμό των παραμέτρων του μοντέλου, την δημιουργία φίλτρων στα δεδομένα του μοντέλου, την επεξεργασία (process model) του μοντέλου ή και όλης της δομής δεδομένων με τα μοντέλα που περιέχει (Process mining structure and all Models). Επίσης μπορούμε να προσθέσουμε ένα καινούργιο μοντέλο Εξόρυξης Γνώσης στην δομή που δουλεύουμε.



Διαμόρφωση παραμέτρων Αλγορίθμου

Επιλέγοντας **Set Algorithm Parameters**, ανοίγει το παράθυρο ιδιοτήτων των παραμέτρων του επιλεγμένου μοντέλου.



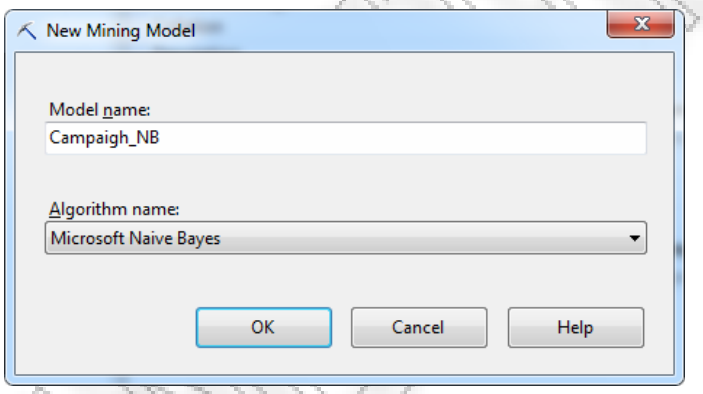
Για περισσότερες πληροφορίες σχετικά με τις παραμέτρους των μοντέλων Εξόρυξης Γνώσης και την χρήση τους, ο αναγνώστης μπορεί να ανατρέξει στο Κεφάλαιο 4, που γίνεται η αναλυτική παρουσίαση των τεχνικών Εξόρυξης Γνώσης που παρέχονται από τα SSAS.

Προσθήκη νέου Μοντέλου Εξόρυξης Γνώσης

Επιλέγοντας το **New Mining Model** στο menu που εμφανίζεται με το δεξί κλικ πάνω σε ένα υπάρχον μοντέλο, ο χρήστης μπορεί να προσθέσει ένα καινούργιο μοντέλο στην υπάρχουσα δομή Εξόρυξης Γνώσης. Η επιλογή αυτή είναι πολύ σημαντική και προσφέρει μεγάλη βοήθεια στον χρήστη. Όπως έχουμε τονίσει, ένα επιχειρησιακό πρόβλημα που επιλύεται με την χρήση τεχνικών Εξόρυξης Γνώσης δεν έχει μια και μοναδική λύση, δεν υπάρχει δηλαδή αυστηρά ένα μόνο μοντέλο Εξόρυξης Γνώσης που μπορεί να επιλύσει το πρόβλημα. Για το λόγο αυτό, είναι σύνηθες το γεγονός να δημιουργούνται πολλά μοντέλα Εξόρυξης Γνώσης και στην συνέχεια οι χρήστες να συγκρίνουν τα αποτελέσματα του κάθε μοντέλου και να επιλέγουν αυτό με τα καλύτερα αποτελέσματα. Το BIDS προσφέρει στον χρήστη την δυνατότητα αυτή καθώς επίσης και τα εργαλεία εκείνα που χρησιμοποιούνται, όπως θα δούμε στη συνέχεια, για την σύγκριση των αποτελεσμάτων των μοντέλων.

Το BIDS προσφέρει στους χρήστες την δυνατότητα να προσθέσουν επιπλέον Μοντέλα Εξόρυξης Γνώσης στην υπάρχουσα δομή. Αυτό έχει το σημαντικό πλεονέκτημα ότι το καινούργιο μοντέλο θα έχει τις ίδιες ρυθμίσεις με τα υπάρχοντα μοντέλα της δομής. Αντί λοιπόν να κάνει τις ρυθμίσεις από την αρχή για το καινούργιο μοντέλο, ο χρήστης με την επιλογή New Mining Model εισάγει ένα καινούργιο μοντέλο το οποίο παίρνει αυτόματα όλες τις ρυθμίσεις των υπαρχόντων μοντέλων. Έτσι στην συνέχεια η σύγκριση των αποτελεσμάτων του κάθε μοντέλου θα γίνει σε μια κοινή βάση.

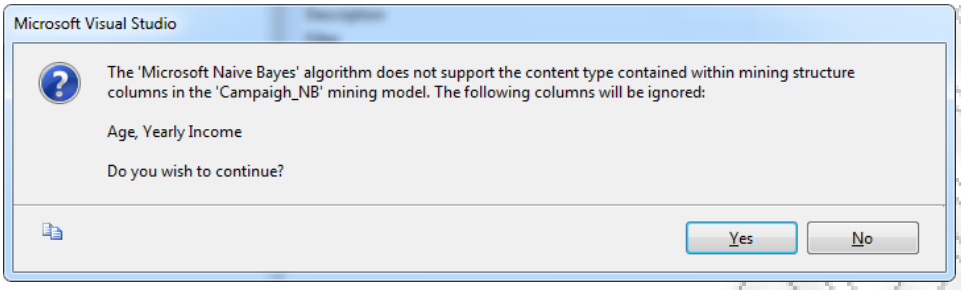
Έστω ότι για παράδειγμα θέλουμε να εισάγουμε ένα καινούργιο μοντέλο που θα χρησιμοποιήσει την τεχνική Naïve Bayes[4]. Στην επιλογή New Mining Model ανοίγει το παρακάτω παράθυρο, που δίνουμε το όνομα του καινούργιου μοντέλου και επιλέγουμε την τεχνική Εξόρυξης Γνώσης που θα χρησιμοποιήσει.



Υπάρχουν όμως κάποιοι περιορισμοί στην εισαγωγή του καινούργιου μοντέλου και αυτό οφείλεται στον περιορισμό που έχει το κάθε μοντέλο σχετικά με του τύπο δεδομένων που μπορεί να διαχειριστεί. Στην περίπτωση που μελετάμε, για παράδειγμα, ο αλγόριθμος Naïve Bayes δεν λειτουργεί με συνεχείς τιμές αλλά μόνο με διακριτές τιμές. Έτσι στο καινούργιο μοντέλο που δημιουργήσαμε, ο αλγόριθμος Δέντρων Απόφασης, ο οποίος δεν έχει κάποιο περιορισμό στην χρήση συνεχών τιμών, είχε σαν είσοδο τις στήλες με συνεχείς τιμές *Age* και

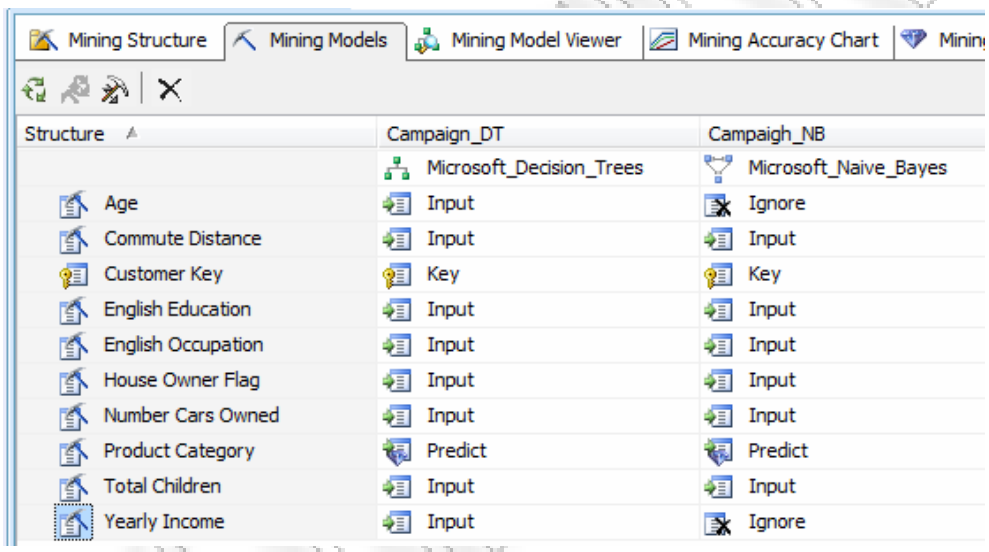
[4] Η τεχνική Naïve Bayes είναι μια απλή τεχνική εξόρυξης δεδομένων, η οποία δεν υποθέτει συσχετίσει μεταξύ των μεταβλητών εισόδου και της τιμές πρόβλεψης. Ο αλγόριθμος Naïve Bayes απλώς συγκρίνει την μεταβλητή προς πρόβλεψη με την κάθε μεταβλητή εισόδου του μοντέλου (σύγκριση σε ζευγάρια) και στο τέλος υπολογίζει το κατά πόσο η μια μεταβλητή εισόδου επηρεάζει το αποτέλεσμα.

Yearly Income, τις οποίες δεν θα μπορεί να διαχειριστεί πλέον το νέο μοντέλο Naïve Bayes που εισάγουμε. Έτσι εμφανίζεται το παρακάτω μήνυμα προειδοποίησης για τον τύπο των δεδομένων αυτών.



Εικόνα 13: Προειδοποίηση για τύπο δεδομένων των στηλών από το νέο Μοντέλο

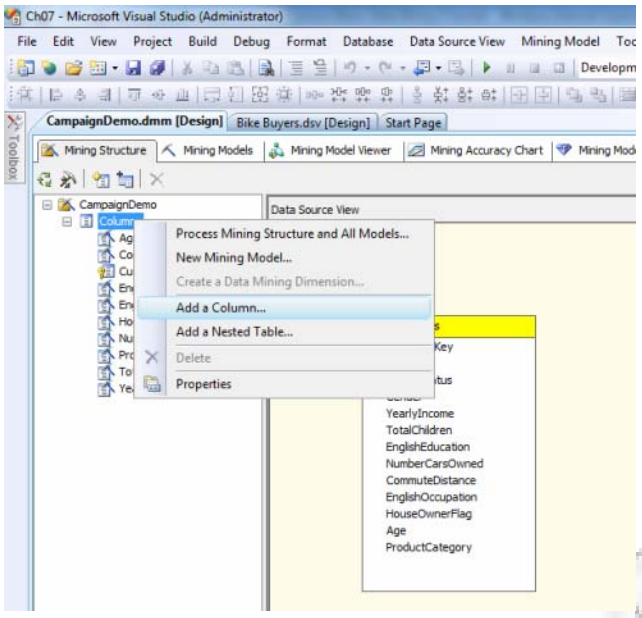
Αυτό που μπορούμε να κάνουμε είναι να πατήσουμε **Yes** προκειμένου το καινούργιο μοντέλο να αγνοήσει (Ignore) τις στήλες με τις συνεχείς τιμές. Το αποτέλεσμα της εισαγωγής του νέου μοντέλου το βλέπουμε στην παρακάτω εικόνα.



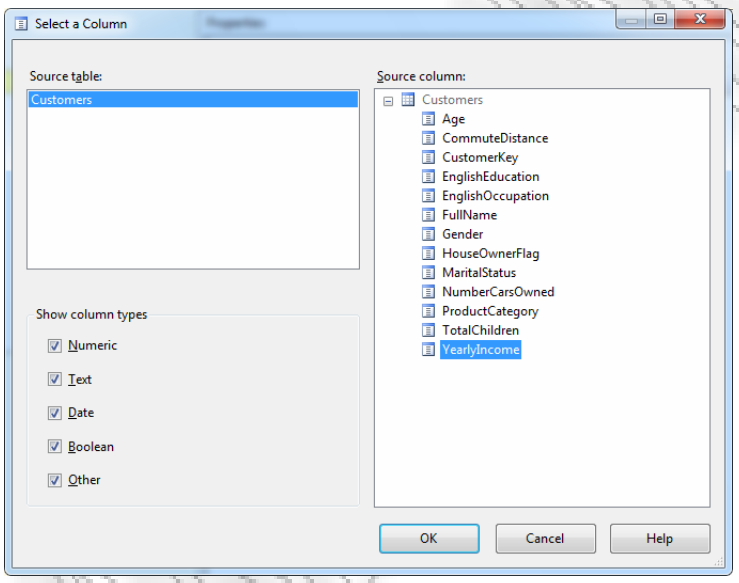
Τι γίνεται όμως στην περίπτωση που μια στήλη με συνεχείς τιμές είναι σημαντική και θέλουμε να την χρησιμοποιήσουμε στο μοντέλο Naïve Bayes που δημιουργήσαμε; Για παράδειγμα το εισόδημα του πελάτη (*Yearly Income*) είναι μια σημαντική παράμετρος που θέλουμε να λάβουμε υπόψη στην ανάλυση μας. Το BIDS δίνει την δυνατότητα στον χρήση να δημιουργήσει μια διακριτή στήλη μέσα από μια στήλη με συνεχείς τιμές. Η διαδικασία αυτή ονομάζεται και «διακριτοποίηση» (**Discretization**).

Δημιουργία Διακριτής Στήλης (Discrete Column)

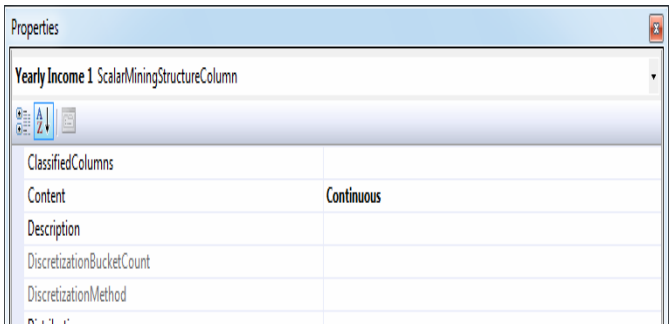
Για να δημιουργήσουμε μια καινούργια διακριτή στήλη, πηγαίνουμε στο tab Mining Structure και στο δέντρο του πίνακα δεδομένων της δομής επιλέγουμε δεξί κλικ και πατάμε **Add a Column**.



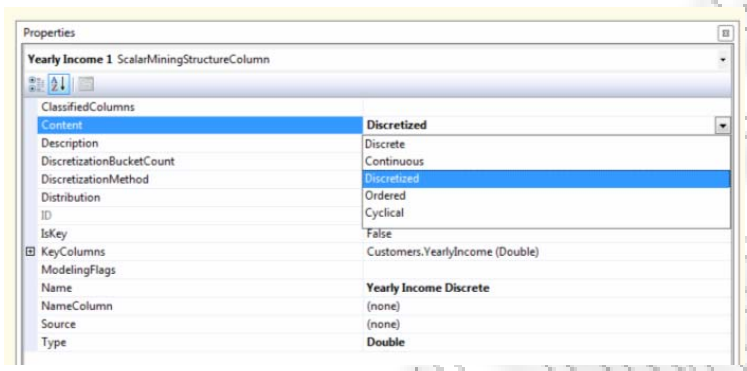
Ανοίγει το παράθυρο επιλογής του πίνακα (Source table) από τον οποίο θα προέλθει η καινούργια στήλη και πατάμε την στήλη (Source Column) η οποία θα «τροφοδοτήσει» την νέα μας στήλη. Για παράδειγμα επιλέγουμε *YearlyIncome* και πατάμε **OK**.



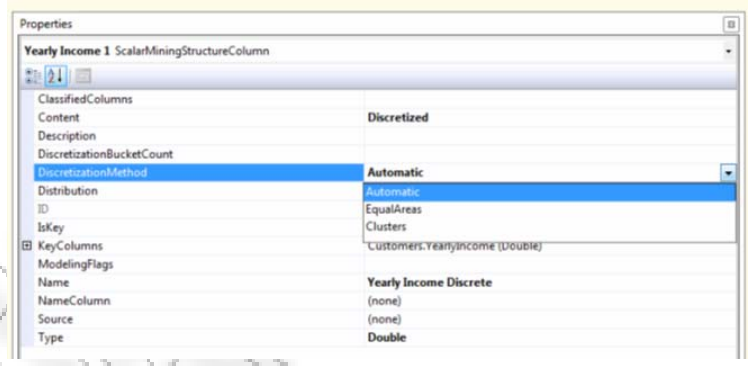
Η καινούργια στήλη εισάγεται στον πίνακα δεδομένων της δομής που δουλεύουμε. Στο παράθυρο Ιδιοτήτων αλλάζουμε το όνομα της καινούργιας στήλης. Για παράδειγμα δίνουμε το όνομα *Yearly Income Discrete* για να ξεχωρίζουμε ότι η στήλη αυτή περιέχει διακριτές τιμές. Στο μήνυμα επιβεβαίωσης που εμφανίζεται μετά την μετονομασία, πατάμε **Yes** για να εφαρμοστεί η αλλαγή της στήλης σε ολόκληρη την δομή δεδομένων, έτσι ώστε από εδώ και στο εξής να χρησιμοποιείται η στήλη με το νέο όνομα που της δώσαμε.



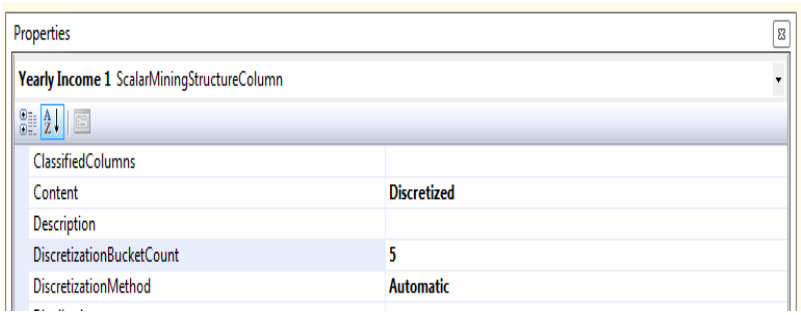
Στην συνέχεια, θα πρέπει να αλλάξουμε το περιεχόμενο της στήλη. Στο πεδίο **Content** από Continuous αλλάζουμε την τιμή σε **Discretized**.



Στο πεδίο **DiscretizationMethod** (η μέθοδος που θα χρησιμοποιηθεί για να δημιουργήσει την καινούργια διακριτή τιμή) επιλέγουμε **Automatic** (οι άλλες επιλογές είναι EqualAreas, ο χωρισμός δηλαδή των συνεχών τιμών σε ίσες περιοχές και Clusters, όπου ορίζει ο χρήστης τον ακριβή αριθμό των ομάδων, στις οποίες θα γίνει ο διαχωρισμός). Η αυτόματη επιλογή χρησιμοποιεί στατιστική ανάλυση των τιμών για να τις χωρίσει σε ομάδες, τον αριθμό των οποίων ορίζει ο χρήστης.

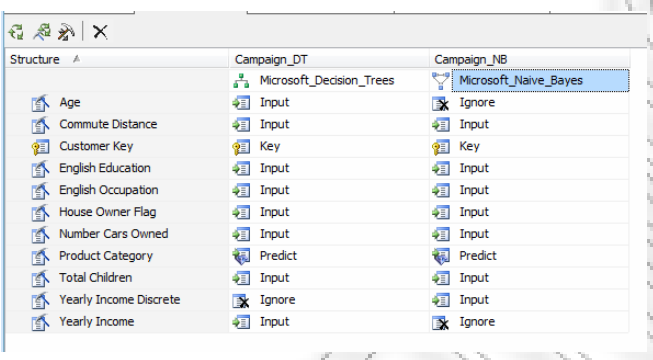


Στην επιλογή **DiscretizationBucketCount** ορίζουμε τον μέγιστο αριθμό των ομάδων που θα δημιουργηθούν. Δίνουμε την τιμή 5.



Εικόνα 14: Δημιουργία Διακριτής Στήλης

Αφού δημιουργήσουμε την καινούργια στήλη με διακριτές τιμές σχετικά με το εισόδημα του πελάτη, στο tab Mining Models θα αλλάξουμε την χρήση της νέας στήλης *Yearly Income Discrete* σε Input για τον αλγόριθμο Naïve Bayes. Στον αλγόριθμο των Δέντρων Απόφασης δεν χρειάζεται να χρησιμοποιήσουμε την καινούργια στήλη με τις διακριτές τιμές του εισοδήματος του πελάτη, αφού ήδη χρησιμοποιεί την στήλη Yearly Income με τις συνεχείς τιμές. Η μορφή των δυο μοντέλων παρουσιάζεται στην παρακάτω εικόνα:

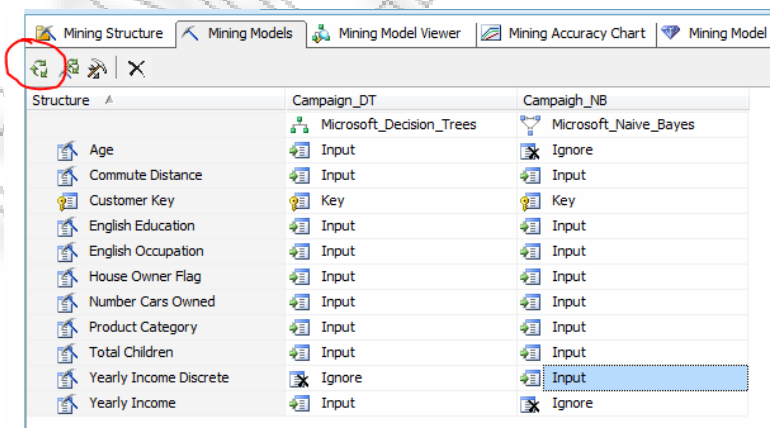


Structure	Campaign_DT	Campaign_NB
	Microsoft_Decision_Trees	Microsoft_Naive_Bayes
Age	Input	Ignore
Commute Distance	Input	Input
Customer Key	Key	Key
English Education	Input	Input
English Occupation	Input	Input
House Owner Flag	Input	Input
Number Cars Owned	Input	Input
Product Category	Predict	Predict
Total Children	Input	Input
Yearly Income Discrete	Ignore	Input
Yearly Income	Input	Ignore

Εκτέλεση της Δομής Δεδομένων

Αφού δημιουργήσουμε τα μοντέλα, το επόμενο βήμα είναι η εκτέλεσή τους (**deploy**). Με την εκτέλεση των μοντέλων, στην ουσία «ανεβάζουμε» τα μοντέλα που κατασκευάσαμε στο BIDS στον Analysis Server που «τρέχει» στον υπολογιστή μας προκειμένου να δημιουργήσει τα μοντέλα και να τα χρησιμοποιήσει για την ανάλυση των δεδομένων.

Στο tab Mining Models επιλέγουμε **Deploy**

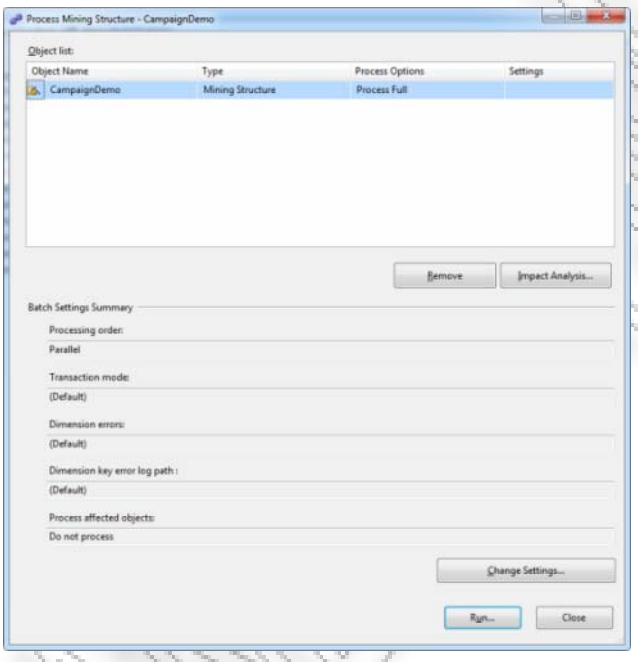


Structure	Campaign_DT	Campaign_NB
	Microsoft_Decision_Trees	Microsoft_Naive_Bayes
Age	Input	Ignore
Commute Distance	Input	Input
Customer Key	Key	Key
English Education	Input	Input
English Occupation	Input	Input
House Owner Flag	Input	Input
Number Cars Owned	Input	Input
Product Category	Input	Input
Total Children	Input	Input
Yearly Income Discrete	Ignore	Input
Yearly Income	Input	Ignore

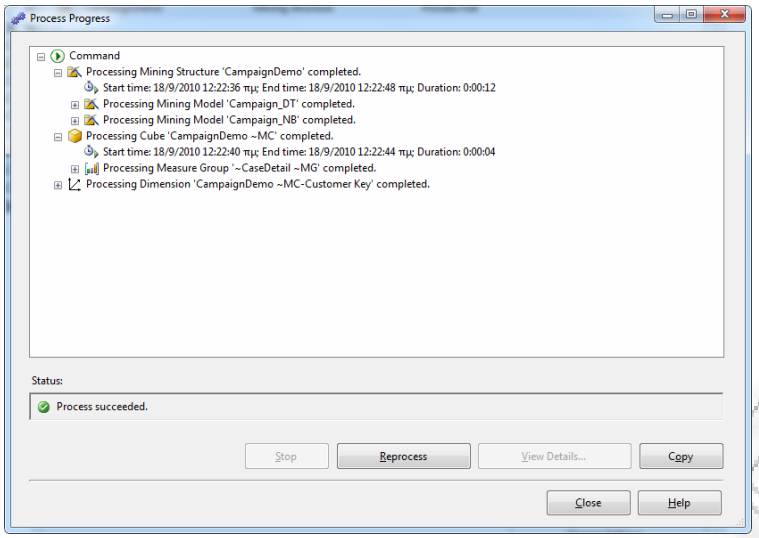
Αν η αποστολή (deployment) των μοντέλων στον Analysis Server επιτύχει, παίρνουμε το παρακάτω μήνυμα.

Επεξεργασία της Δομής

Με την επιτυχή αποστολή της δομής στον server, ανοίγει το παράθυρο για την επεξεργασία της δομής. Ουσιαστικά στο βήμα αυτό, ο server θα πάρει τα δεδομένα εκπαίδευσης που έχουμε ορίσει για την δομή και θα εκτελέσει τους αλγορίθμους του κάθε μοντέλου που περιέχει η δομή για να πάρουμε τα αποτελέσματα της Εξόρυξης Γνώσης που επιθυμούμε. Η επεξεργασία της δομής, ανάλογα με το πλήθος των δεδομένων που έχουμε στην διάθεσή μας και με την πολυπλοκότητα του κάθε αλγορίθμου Εξόρυξης Γνώσης που χρησιμοποιείται, ενδέχεται να διαρκέσει από μερικά δευτερόλεπτα έως μερικά λεπτά.



Επιλέγουμε το κουμπί **Run** για να επεξεργαστούμε την δομή Εξόρυξης Γνώσης που έχουμε δημιουργήσει



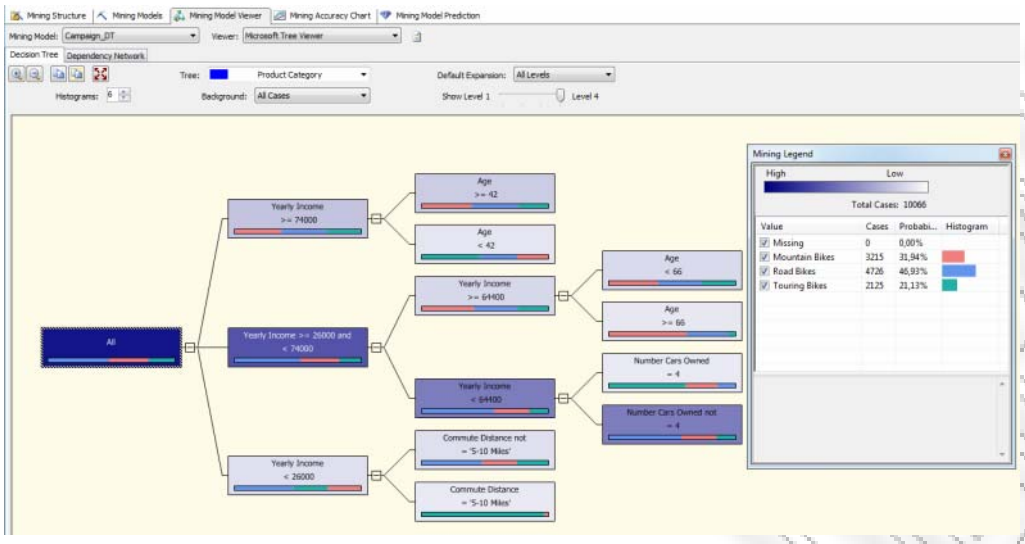
Αν επιτύχει και η διαδικασία επεξεργασία των μοντέλων στον Analysis Server, εμφανίζεται στο παράθυρο Process Progress στο πεδίο Status το μήνυμα **Process Succeeded**. Αυτό σημαίνει ότι τα μοντέλα Εξόρυξης Γνώσης δημιουργήθηκαν με επιτυχία στον server και μπορούμε να δούμε τα αποτελέσματα των μοντέλων αυτών μέσα από τους Viewers που είναι διαθέσιμοι από το BIDS στο tab Mining Model Viewer.

5. Εξερεύνηση του Μοντέλου

Ένα από τα βασικότερα ζητήματα που πρέπει να διαχειριστεί μια εφαρμογή διαχείρισης λύσεων Εξόρυξης Γνώσης είναι η οπτικοποίηση των αποτελεσμάτων της εκπαίδευσης των διαφόρων αλγορίθμων στους χρήστες. Τα αποτελέσματα της ανάλυσης των δεδομένων μετά από την εκπαίδευση του κάθε αλγορίθμου πρέπει να εμφανίζονται με ένα απλό και κατανοητό τρόπο στους χρήστες για να βλέπουν τα αποτελέσματα και να επιλέγουν τον αλγόριθμο που τους ταιριάζει. Το BIDS παρέχει αρκετά εργαλεία οπτικοποίησης των αποτελεσμάτων (viewers) για την κάθε τεχνική Εξόρυξης Γνώσης που χρησιμοποιεί ο χρήστης. Για κάθε αλγόριθμο, στον designer που παρουσιάζουμε, δημιουργείται το tab **Mining Model Viewer**, στο οποίο περιέχονται οπτικοποιημένα τα αποτελέσματα της εκπαίδευσης και της εκτέλεσης του κάθε αλγορίθμου.

Δέντρο Απόφασης

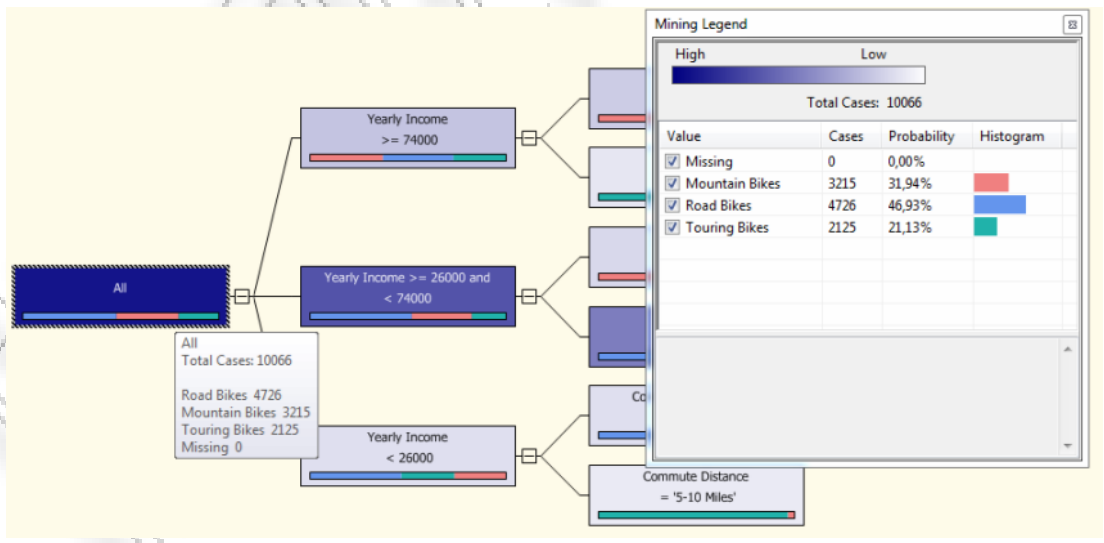
Για τον αλγόριθμο Δέντρων Απόφασης, που παρουσιάζουμε στο tab Mining Model Viewer εμφανίζεται το Δέντρο Απόφασης που έχει κατασκευαστεί μετά την επεξεργασία του μοντέλου. Για κάθε χαρακτηριστικό που προβλέπει ο αλγόριθμος (στο παράδειγμα μας για την στήλη *ProductCategory*) δημιουργείται και ένα διαφορετικό δέντρο απόφασης.



Εικόνα 15: Decision Trees – Decision Tree Viewer

Το δέντρο απόφασης χρησιμοποιείται κυρίως για Κατηγοριοποίηση. Κάθε κόμβος περιγράφει μια κύρια κατηγορία με πληροφορίες για τους πελάτες. Στην επιλογή Background, επιλέγουμε το φίλτρο που μπορούμε να εφαρμόσουμε στο δέντρο. Η προεπιλεγμένη τιμή είναι η *All Cases*, όπου βλέπουμε τα αποτελέσματα για όλες τις περιπτώσεις όσον αφορά τις κατηγορίες ποδηλάτων της εταιρείας. Στο φίλτρο επίσης μπορούμε να επιλέξουμε μια συγκεκριμένη κατηγορία ποδηλάτων, για να δούμε ποιοι παράγοντες την επηρεάζουν.

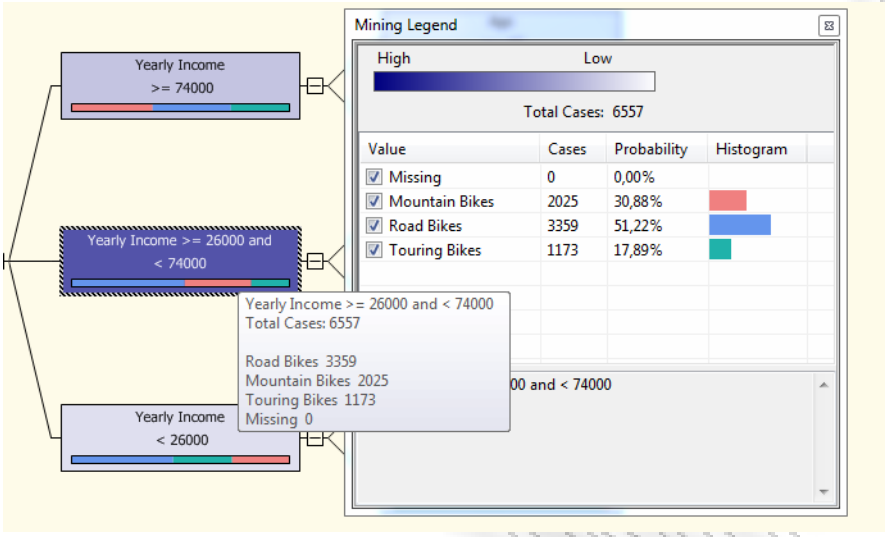
Το χρώμα του κάθε κόμβου είναι ένας πολύ σημαντικός οπτικός δείκτης πληροφοριών, γιατί δείχνει το πλήθος των εγγραφών (cases) που βρίσκεται στον κόμβο αυτό (δείχνει με λίγα λόγια τον πληθυσμό του κάθε κόμβου). Όσο πιο σκοτεινό είναι το χρώμα του κόμβου, τόσο μεγαλύτερος ο πληθυσμός του κόμβου. Πάντα ο πρώτος κόμβος του δέντρου (η ρίζα του) έχει το σκοτεινότερο χρώμα. Αν κουνήσουμε τον δείκτη του ποντικού πάνω από ένα κόμβο, εμφανίζεται ένα tooltip, που δείχνει της πληροφορίες του κόμβου. Η ίδια πληροφορία εμφανίζεται και στο υπόμνημα (legend) του κόμβου.



Για παράδειγμα, για τον πρώτο κόμβο (ρίζα) του δέντρου βλέπουμε ότι περιέχει συνολικά 10.066 εγγραφές (cases) αγοράς κατηγοριών ποδηλάτων από τους πελάτες, από τις

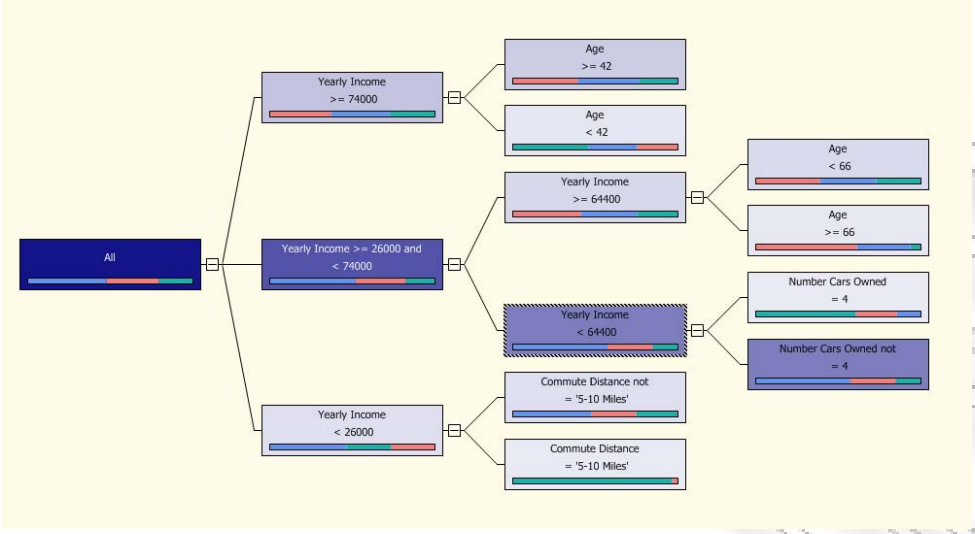
οποίες οι 4.726 αφορούν στην αγορά Road Bikes, οι 3.125 στην αγορά Mountain Bikes και οι 2.215 στην αγορά Touring Bikes (οι τρείς συνολικά κατηγορίες ποδηλάτων της εταιρείας).

Το πρώτο πράγμα που μπορούμε να παρατηρήσουμε από το δέντρο απόφασης είναι ότι η βασικότερη αιτία που επηρεάζει την επιλογή του πελάτη είναι το εισόδημα (*Yearly Income*) αφού σε αυτό το χαρακτηριστικό γίνεται ο πρώτος διαχωρισμός. Μάλιστα, η βασικότερη κατηγορία διαχωρισμού είναι το εύρος εισοδήματος μεταξύ 26.000 και 74.000 (είναι ο κόμβος με το πιο σκούρο χρώμα άρα και με τον μεγαλύτερο πληθυσμό.

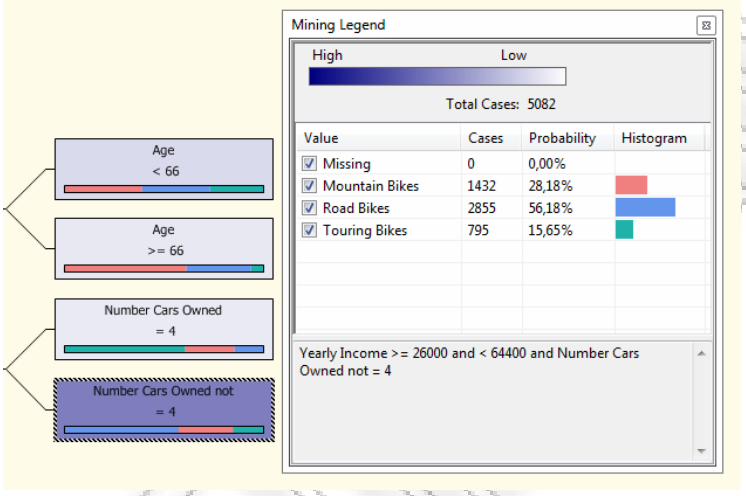


Πράγματι, αν πατήσουμε πάνω στον κόμβο βλέπουμε ότι περιέχει 6.557 cases από τα οποία το μεγαλύτερο ποσοστό πελατών (οι 3.359) έχουν αγοράσει Road Bikes. Η ίδια πληροφορία φαίνεται σε κάθε κόμβο οπτικά και από τα χρώματα της κάθε κατηγορίας, που περιέχει ο κόμβος. Έτσι για το παράδειγμα μας, η μπλε κατηγορία (Road Bikes) είναι μεγαλύτερη, ύστερα έρχεται η κόκκινη κατηγορία (Mountain Bikes) και στο τέλος η πράσινη κατηγορία (Touring Bikes).

Στην συνέχεια στο επίπεδο 3 του δέντρου απόφασης βλέπουμε ότι για τον προηγούμενο κόμβο (τον μεσαίο κόμβο) η διάσπαση γίνεται πάλι στο εισόδημα του πελάτη (*Yearly Income* < 64.000) και στην συνέχεια στο επίπεδο 4 του δέντρου ο διαχωρισμός γίνεται πάνω στον αριθμό των αυτοκινήτων (*Number Cars Owned*) του πελάτη.



Αν πατήσουμε στον σκούρο κόμβο του επιπέδου 4, βλέπουμε ότι για ένα πλήθος 5.082 περιπτώσεων, οι 2.855 έχουν αγοράσει Road Bikes.



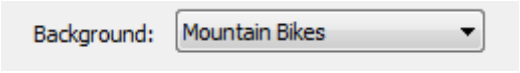
Επίσης παρατηρούμε στο υπόμνημα του δέντρου και τον αντίστοιχο κανόνα που υπάρχει σε κάθε κόμβο. Για παράδειγμα, για τον επιλεγμένο κόμβο που αναφέρουμε, ο κανόνας είναι ο παρακάτω:

```
Yearly Income >= 26000 and < 64400 and Number Cars
Owned not = 4
```

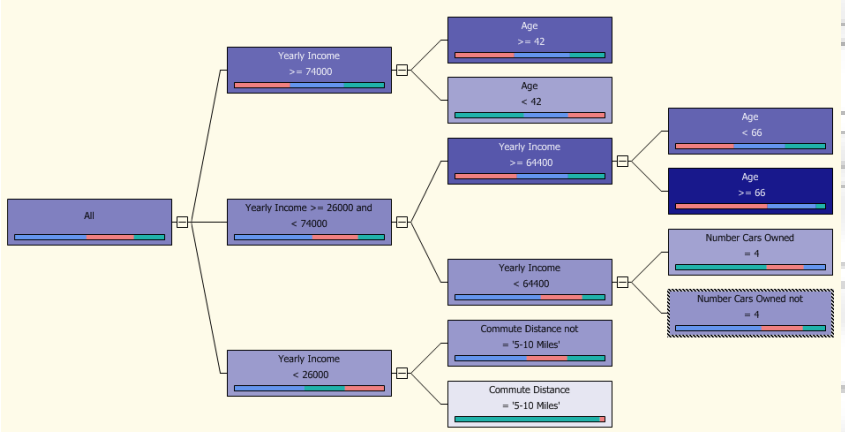
Ο κανόνας αυτός σημαίνει ότι οι πελάτες που έχουν ετήσιο εισόδημα από 26.000 μέχρι 63.999 και δεν έχουν 4 αυτοκίνητα είναι πιο πιθανό να αγοράσουν Road Bikes (με μια πιθανότητα, όπως βλέπουμε στο υπόμνημα 56.18%). Αυτή η ομάδα πελατών είναι μια αρκετά καλή ομάδα για να απευθυνθεί το τμήμα πωλήσεων της εταιρείας και να τους κάνει μια προσφορά για αγορά Road Bikes.

Ένα άλλο ζήτημα που μπορεί να αντιμετωπιστεί είναι εάν το τμήμα μάρκετινγκ θέλει να μάθει πληροφορίες για τους πελάτες που ενδιαφέρονται για Mountain Bikes. Αυτό είναι πολύ

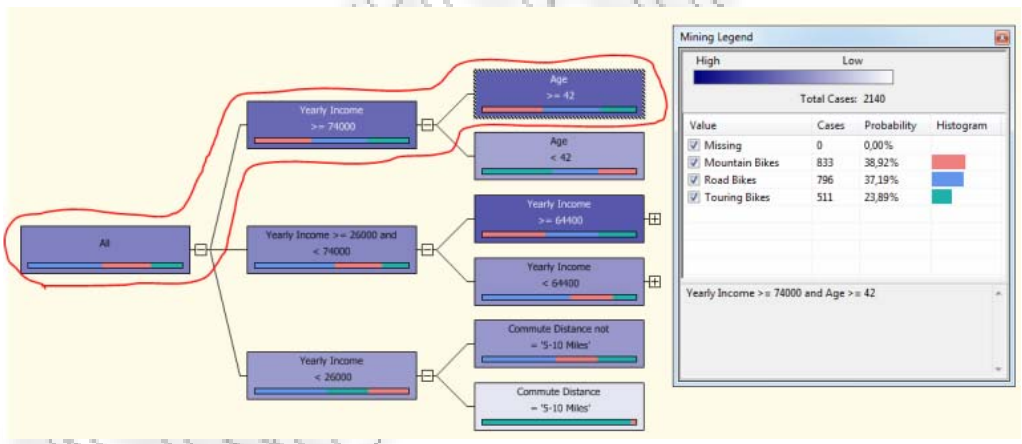
εύκολο να απαντηθεί από το Δέντρο Απόφασης. Αρκεί ο χρήστης στο πεδίο Background να επιλέξει από την λίστα τα Mountain Bikes.



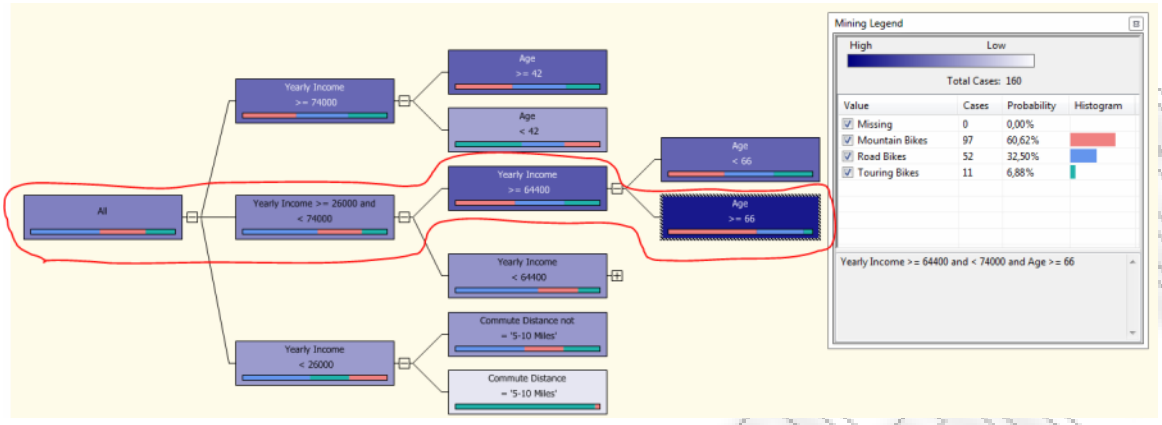
Το δέντρο απόφασης αλλάζει μορφή και παρουσιάζεται παρακάτω:



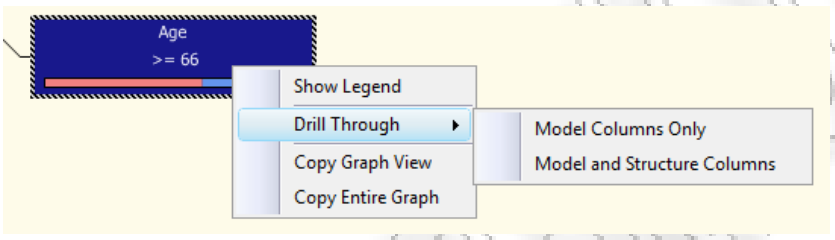
Από το δέντρο παρατηρούμε ότι υπάρχουν 2 πιθανές ομάδες πελατών για αγορά Mountain Bike. Η πρώτη ομάδα περιέχει πελάτες που έχουν εισόδημα μεγαλύτερο ή ίσο με 74.000 και έχουν ηλικία μεγαλύτερη ή ίση των 42 ετών.



Η άλλη πιθανή ομάδα είναι οι πελάτες που έχουν εισόδημα μεγαλύτερο ή ίσο των 64.000 και είναι μεγαλύτεροι ή ίσοι των 66 ετών.



Τέλος, μια χρήσιμη επιλογή είναι η προβολή των πελατών που ανήκουν σε κάθε κόμβο του Δέντρου Απόφασης. Αυτό επιτυγχάνεται μέσω της επιλογής Allow drill through, που είχαμε επιλέξει κατά την δημιουργία του μοντέλου Εξόρυξης Γνώσης. Κάνοντας δεξί κλικ σε ένα κόμβο στην επιλογή Drill Through μπορούμε να επιλέξουμε να δούμε τα δεδομένα του κόμβου.



Η επιλογή **Model Columns Only** δείχνει μόνο τις στήλες που ανήκουν στο μοντέλο Δέντρων Απόφασης που δουλεύουμε ενώ η επιλογή **Model and Structure Columns** δείχνει και τις υπόλοιπες στήλες της δομής Εξόρυξης Γνώσης, που ανήκει το μοντέλο (που ενδεχομένως να μην εμφανίζονται στο τρέχον μοντέλο). Ένα παράδειγμα της επιλογής Drill Through φαίνεται παρακάτω:

Drill Through

Cases Classified to:

Yearly Income >= 64400 and < 74000 and Age >= 66

Age	Commute Dist...	Customer Key	English Educati...	English Occup...	House Owner ...	Number Cars ...	Product Categ...	Total Children	Yearly Income
75	5-10 Miles	Noah Coleman	Graduate Degr...	Management	1	2	Mountain Bikes	4	70000
75	5-10 Miles	Noah Coleman	Graduate Degr...	Management	1	2	Road Bikes	4	70000
75	1-2 Miles	Jon Luo	Graduate Degr...	Management	1	2	Mountain Bikes	4	70000
75	1-2 Miles	Jon Luo	Graduate Degr...	Management	1	2	Road Bikes	4	70000
74	1-2 Miles	Orlando E. Vaz...	Graduate Degr...	Management	1	2	Mountain Bikes	4	70000
74	1-2 Miles	Orlando E. Vaz...	Graduate Degr...	Management	1	2	Road Bikes	4	70000
73	1-2 Miles	Jade E. Bailey	Graduate Degr...	Management	0	2	Mountain Bikes	4	70000
73	1-2 Miles	Jade E. Bailey	Graduate Degr...	Management	0	2	Road Bikes	4	70000
67	1-2 Miles	Ricardo C. Nath	Graduate Degr...	Management	1	1	Touring Bikes	2	70000
66	1-2 Miles	Garrett S. Kelly	Bachelors	Management	0	2	Mountain Bikes	4	70000
66	1-2 Miles	Garrett S. Kelly	Bachelors	Management	0	2	Road Bikes	4	70000

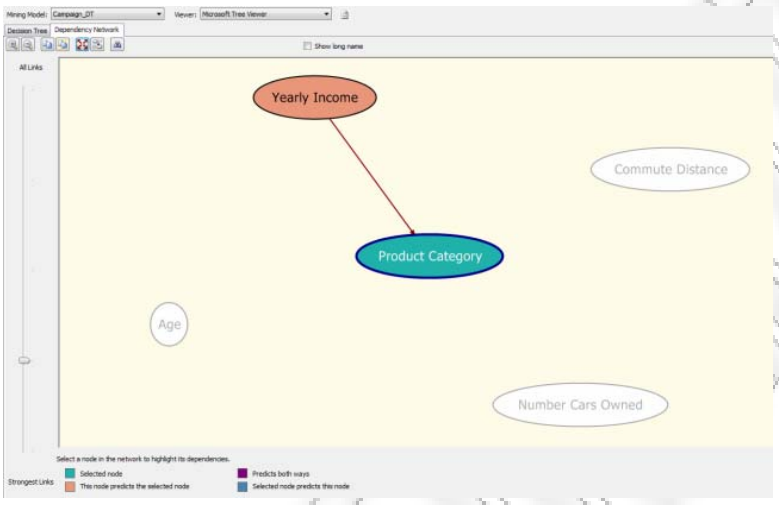
Query execution completed with 11 rows fetched

Εικόνα 16: Drill through Δεδομένων

Όπως παρατηρούμε από τον παραπάνω πίνακα, στην στήλη *CustomerKey* δεν εμφανίζεται το κλειδί του πελάτη αλλά το όνομα του, όπως ακριβώς το είχαμε ορίσει σε προηγούμενα βήματα.

Dependency Network

Στην περίπτωση που ένα πρόβλημα της εταιρείας, που επιχειρούμε να το επιλύσουμε με την τεχνική των Δέντρων Απόφασης περιέχει ένα αρκετά μεγάλο αριθμό χαρακτηριστικών, είναι αρκετά δύσκολο να δούμε και να κατανοήσουμε τις συσχετίσεις μεταξύ των χαρακτηριστικών αυτών και πώς το ένα επηρεάζει το άλλο. Την λύση στο ζήτημα αυτό μας την δίνει το Dependency network, το οποίο ακριβώς δείχνει τις συσχετίσεις μεταξύ των χαρακτηριστικών. Για παράδειγμα, στο παρακάτω διάγραμμα βλέπουμε τα χαρακτηριστικά του προβλήματος μας (δεδομένα εισόδου και δεδομένα προς πρόβλεψη) και τις συσχετίσεις που υπάρχουν μεταξύ τους.



Εικόνα 17: Decision Trees – Dependency Network

Πατώντας πάνω σε ένα χαρακτηριστικό, βλέπουμε τις συσχετίσεις του με τα υπόλοιπα χαρακτηριστικά. Για να δούμε τις ισχυρότερες συνδέσεις του επιλεγμένου χαρακτηριστικού με τα υπόλοιπα (λέγοντας συνδέσεις εννοούμε αν το επιλεγμένο χαρακτηριστικό προβλέπει το άλλο ή αν υπάρχουν άλλα χαρακτηριστικά που προβλέπουν το επιλεγμένο) μπορούμε να χρησιμοποιήσουν το φίλτρο (slider) που υπάρχει στην αριστερή πλευρά του διαγράμματος. Αν κουνήσουμε τον slider προς τα πάνω βλέπουμε όλες τις συνδέσεις με το επιλεγμένο χαρακτηριστικό ενώ όσο κατεβάζουμε τον slider προς τα κάτω, εμφανίζονται οι πλέον ισχυρότερες συνδέσεις.

Για παράδειγμα, στο πρόβλημα που περιγράφουμε, έχουμε επιλέξει το *Product Category* και όπως βλέπουμε, το χαρακτηριστικό εκείνο που το επηρεάζει περισσότερο είναι το *Yearly Income*. Επίσης, όπως παρατηρούμε από το υπόμνημα του διαγράμματος (το κείμενο στο κάτω μέρος του διαγράμματος) το επιλεγμένο χαρακτηριστικό είναι το πράσινο και το *Yearly Income*, που έχει πάρει ένα κόκκινο χρώμα σημαίνει ότι χρησιμοποιείται για να προβλέψει το επιλεγμένο χαρακτηριστικό, δηλαδή το *Product Category*. Για το λόγο αυτό η γραμμή που συνδέει τον κόμβο *Yearly Income* με το *Product Category* έχει το βέλος προς το *Product Category*. Αν η σχέση ήταν αμφίδρομη, δηλαδή και τα δύο χαρακτηριστικά πρόβλεπαν το ένα το άλλο, η γραμμή θα είχε βέλη και στις δύο άκρες της.

Naïve Bayes

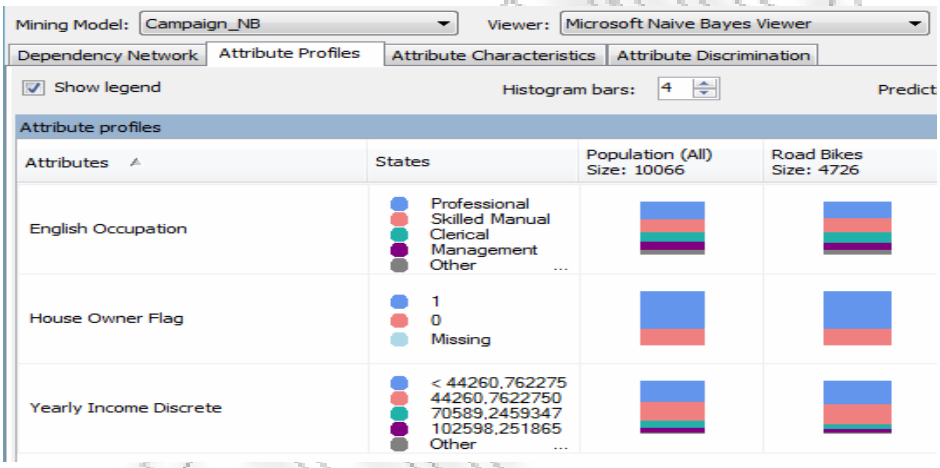
Στην συνέχεια θα περιγράψουμε τα διαγράμματα που παρέχονται από το BIDS για τον αλγόριθμο Naïve Bayes. Στο tab **Mining Model Viewer**, αν αλλάξουμε στο πεδίο **Mining Model:** την επιλογή σε **Campaign_NB** (που είναι το μοντέλο Εξόρυξης Γνώσης που δημιουργήσαμε με την χρήση του αλγορίθμου Naïve Bayes) βλέπουμε τα διαγράμματα που αντιστοιχούν στον επιλεγμένο αλγόριθμο.

Το πρώτο διάγραμμα που εμφανίζεται είναι στο tab Dependency Network, το οποίο έχει την ίδια μορφή και την ίδια λειτουργικότητα με το Dependency Network που εμφανίζεται για τα Δέντρα Απόφασης.

TAB: Attribute Profiles Diagram

Το tab Attribute Profiles περιέχει ένα διάγραμμα που δείχνει την ταξινόμηση των δεδομένων εισόδου για κάθε κατηγορία τιμών της μεταβλητής προς πρόβλεψη. Το διάγραμμα έχει τόσες στήλες όσες και οι διαφορετικές τιμές της στήλης πρόβλεψης, συν μια επιπλέον στήλη για όλο την πληθυσμό (η στήλη Population (All)) και μια για τις ελλιπείς τιμές (Missing). Αν παρατηρήσουμε τις στήλες, βλέπουμε ότι ο πληθυσμός έχει 10.066 εγγραφές από τις οποίες, οι πελάτες έχουν αγοράσει 4.726 Road Bikes, 2.125 Touring Bikes και 3.125 Mountain Bikes ενώ δεν υπάρχουν ελλιπείς τιμές.

Το διάγραμμα δείχνει τα χαρακτηριστικά εισόδου του αλγορίθμου σε γραμμές και στην στήλη States δείχνει τις διαφορετικές τιμές από το κάθε χαρακτηριστικό.

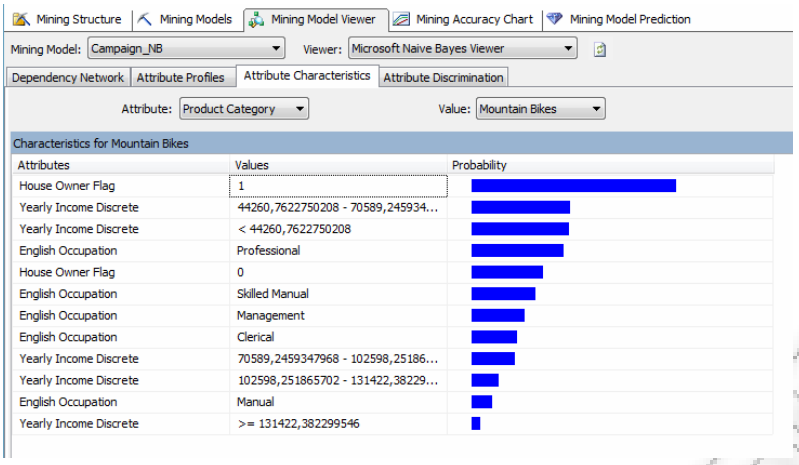


Εικόνα 18: Attribute Profiles Diagram

Αν πατήσουμε σε κάθε κελί του διαγράμματος, και έχουμε ανοιχτό το παράθυρο του υπομνήματος του διαγράμματος (επιλογή Show/Hide Legend με δεξί κλικ πάνω στο διάγραμμα) βλέπουμε την κατανομή των τιμών του χαρακτηριστικού εισόδους ως προς τον πληθυσμό (Population) αν το κελί ανήκει στην αντίστοιχη στήλη ή ως προς την τιμή της μεταβλητής προς πρόβλεψη, αν το κελί ανήκει σε μία από τις στήλης με τις διαφορετικές τιμές του χαρακτηριστικού πρόβλεψης του μοντέλου μας.

TAB: Attribute Characteristics

Το διάγραμμα στο tab Attribute Characteristics παρουσιάζει με λεπτομέρεια για κάθε τιμή του χαρακτηριστικού προς πρόβλεψη τις τιμές των χαρακτηριστικών εισόδου στον αλγόριθμο, που επηρεάζουν το αποτέλεσμα.

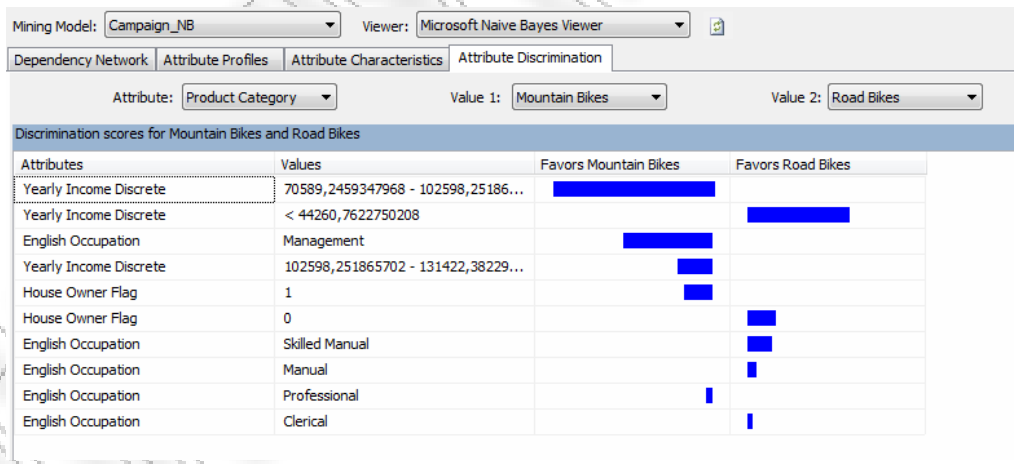


Εικόνα 19: Attribute Characteristics

Για παράδειγμα, για την τιμή Mountain Bikes του χαρακτηριστικού *Product Category*, βλέπουμε ότι οι περισσότεροι πελάτες που αγόρασαν ποδήλατο αυτής της κατηγορίας έχουν δικό τους σπίτι (*House Owner flag* = 1) και ένα εισόδημα μικρότερο από 44.260 (*Yearly Income Discrete*) ενώ σαν επάγγελμα θεωρούνται Professional (*English Occupation*). Αλλάζοντας την επιλογή **Value** για κάθε τιμή του χαρακτηριστικού προς πρόβλεψη, ο χρήστης μπορεί να παρατηρήσει τις διαφορετικές τιμές των χαρακτηριστικών εισόδου που επηρεάζουν το επιλεγμένο αποτέλεσμα (ποιοι παράγοντες με λίγα λόγια επηρεάζουν την αγορά ποδηλάτου από την επιλεγμένη κατηγορία).

TAB: Attributes Discrimination

Το tab Attribute Discrimination παρουσιάζει ένα διάγραμμα στο οποίο ο χρήστης μπορεί να συγκρίνει τις διαφορετικές τιμές του χαρακτηριστικού προς πρόβλεψη μεταξύ τους ανά ζευγάρι. Έτσι για κάθε επιλεγμένο ζεύγος βλέπουμε ποια τιμή των χαρακτηριστικών εισόδου επηρεάζει περισσότερο ή λιγότερο (και πόσο επηρεάζει) τα μέλη του ζευγαριού.



Εικόνα 20: Attributes Discrimination

Για το παράδειγμα που παρουσιάζουμε, μπορούμε να χρησιμοποιήσουμε το διάγραμμα attributes discrimination για να συγκρίνουμε την κατηγορία Mountain Bikes (πεδίο **Value 1**) με την κατηγορία Road Bikes (πεδίο **Value 2**). Έτσι βλέπουμε ότι την κατηγορία Mountain Bikes την προτιμούν πελάτες με υψηλό ετήσιο εισόδημα (πάνω από 70.589 στην τιμή *Yearly Income*

Discrete) και σαν κατηγορία εργασίας (*English Occupation*) εντάσσονται στις Management και Professional. Αντίθετα, οι πελάτες που αγοράζουν Road Bikes έχουν χαμηλότερο ετήσιο εισόδημα (μικρότερο από 44.260) ενώ επαγγελματικά εντάσσονται στις κατηγορίες Skilled Manual και Manual.

Οι χρήστες αλλάζοντας τις επιλογές στα πεδία Value 1 και Value 2 μπορούν να δουν στο διάγραμμα για κάθε ζεύγος όλες εκείνες τις τιμές των χαρακτηριστικών που επηρεάζουν το την αγορά ποδηλάτων από κάθε κατηγορία.

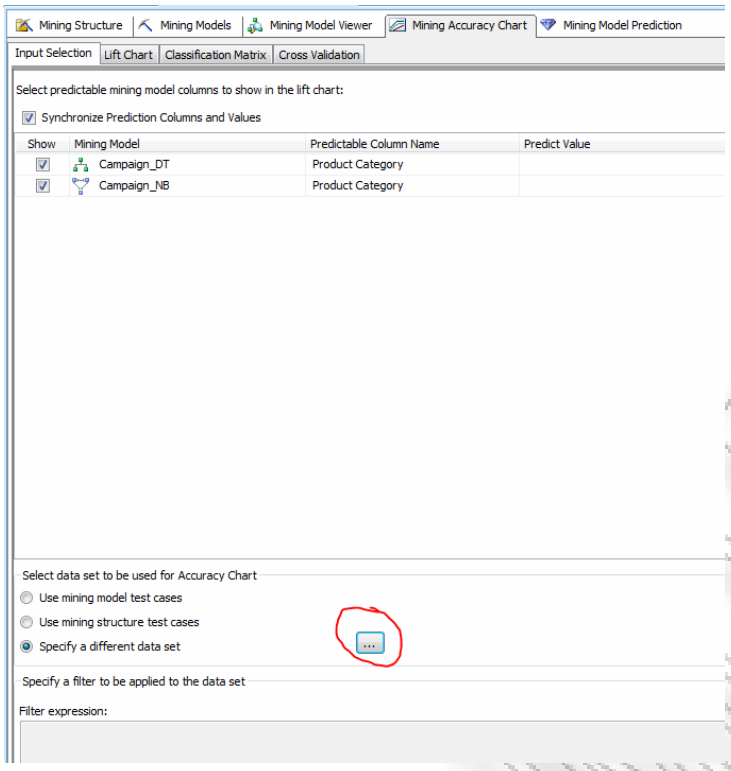
6. Αποτίμηση του Μοντέλου

Το επόμενο βήμα μετά την δημιουργία των μοντέλων Εξόρυξης Γνώσης και την προβολή των αποτελεσμάτων της, μέσα από τους διάφορους viewers, που παρέχονται από το BIDS, είναι η αποτίμηση των αποτελεσμάτων των μοντέλων. Αν και πολλοί χρήστες μπορεί να αποτιμήσουν τα αποτελέσματα των μοντέλων Εξόρυξης Γνώσης κοιτώντας απλώς τους viewers, ο καλύτερος τρόπος αποτίμησης των αποτελεσμάτων είναι μέσα από το BIDS, από το οποίο παρέχονται διάφορα διαγράμματα και τεχνικές αξιολόγησης και αποτίμησης των μοντέλων που έχουν δημιουργηθεί. Η αποτίμηση των μοντέλων γίνεται για να επαληθευθεί το αποτέλεσμα που παράγει το κάθε μοντέλο έτσι ώστε να γίνουν οι όποιες διορθώσεις απαιτούνται στο καθένα (για παράδειγμα αλλαγή των τιμών κάποιων παραμέτρων ενός αλγορίθμου ή η προσθαφαίρεση στηλών εισόδου στον αλγόριθμο). Επίσης η αποτίμηση των μοντέλων μπορεί να μας βοηθήσει να επιλέξουμε το μοντέλο εκείνο που παράγει το καλύτερο αποτέλεσμα (όπως στην περίπτωση του δικού μας παραδείγματος που έχουμε δημιουργήσει δυο διαφορετικά μοντέλα για την επίλυση του ίδιου προβλήματος).

Η αποτίμηση των μοντέλων Εξόρυξης Γνώσης, όντας μέρος της CRISP-DM διαδικασίας βοηθάει τους χρήστες να αξιολογήσουν τα μοντέλα που έχουν δημιουργήσει και στην συνέχεια να επιλέξουν το μοντέλο (ή τα μοντέλα) που δίνει τα καλύτερα αποτελέσματα και να προχωρήσουν στην εγκατάσταση του μοντέλου στον οργανισμό (για παράδειγμα ενσωμάτωση του μοντέλου σε μια εφαρμογή του οργανισμού ή χρήση του μοντέλου μέσα από το Microsoft Excel κλπ). Οι χρήστες επίσης, μετά την αξιολόγηση των μοντέλων μπορεί να ανακαλύψουν ότι δεν έχουν λάβει υπόψη τους όλα τα δεδομένα και όλες τις πληροφορίες που έχουν στην διάθεση τους, έτσι μπορούν να γυρίσουν στο πρώτο βήμα της κατανόησης της επιχείρησης και να κάνουν τις απαραίτητες αλλαγές (για παράδειγμα εισαγωγή νέων στηλών δεδομένων στο μοντέλο που παρέχουν περισσότερες πληροφορίες και στη συνέχεια να εκτελέσουν τον κύκλο δημιουργίας των μοντέλων Εξόρυξης Γνώσης από την αρχή).

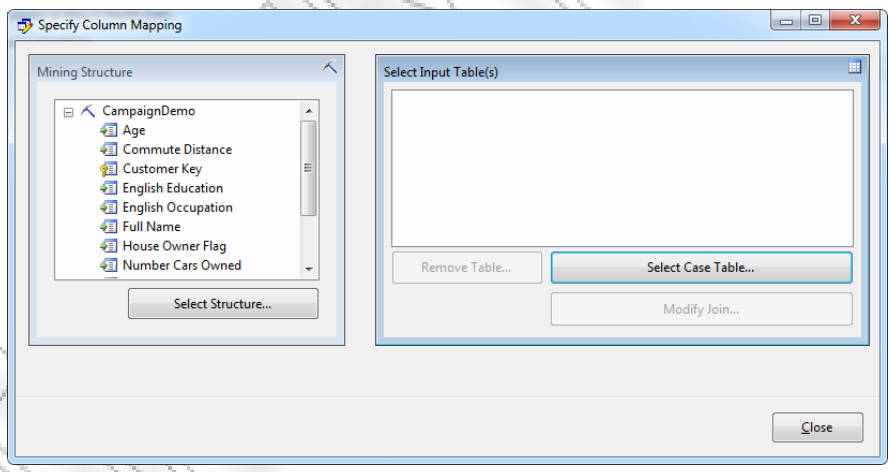
Mining Accuracy Chart

Το Mining Accuracy Chart βοηθάει τους χρήστες να γνωρίσουν κατά πόσο ένα μοντέλο Εξόρυξης Γνώσης προβλέπει τις τιμές αποτελεσματικά. Το πρώτο βήμα που πρέπει να γίνει είναι να επιλέξει ο χρήστης το σύνολο των δεδομένων που θα χρησιμοποιηθεί για την αξιολόγηση των αποτελεσμάτων. Στο πρώτο tab Input Selection, επιλέγουμε το σύνολο δεδομένων. Οι επιλογές που έχουμε εδώ είναι να χρησιμοποιήσουμε το σύνολο εκπαίδευσης που έχει χρησιμοποιηθεί για την εκπαίδευση και την δημιουργία του μοντέλου. Αυτός ίσως να μην είναι και ο καλύτερος τρόπος αποτίμησης του μοντέλου, γιατί πολλές φορές έχει παρατηρηθεί το γεγονός της υπέρ-εκπαίδευσης του μοντέλου στις περιπτώσεις που τα δεδομένα έχουν πολύ ακραίες τιμές (για παράδειγμα πολύ μικρές τιμές για μια στήλη – έτσι το μοντέλο έχει εκπαιδευτεί για τις τιμές αυτές και στην περίπτωση που στο μέλλον έρθει μια μεγάλη τιμή να μην μπορεί να την αξιολογήσει σωστά το μοντέλο). Έτσι επειδή το μοντέλο μπορεί να έχει εκπαιδευτεί πάνω σε συγκεκριμένα δεδομένα, η χρήση των ίδιων δεδομένων για την αποτίμηση του να δίνει πάντα το καλύτερο αποτέλεσμα. Για το λόγο αυτό, συνήθως δεν χρησιμοποιούμε το σύνολο δεδομένων εκπαίδευσης του μοντέλου αλλά χρησιμοποιούμε ένα σύνολο που περιέχει όλες τις περιπτώσεις δεδομένων. Στη επιλογή **specify a different data set** και πατώντας το  επιλέγουμε το σύνολο αυτό.

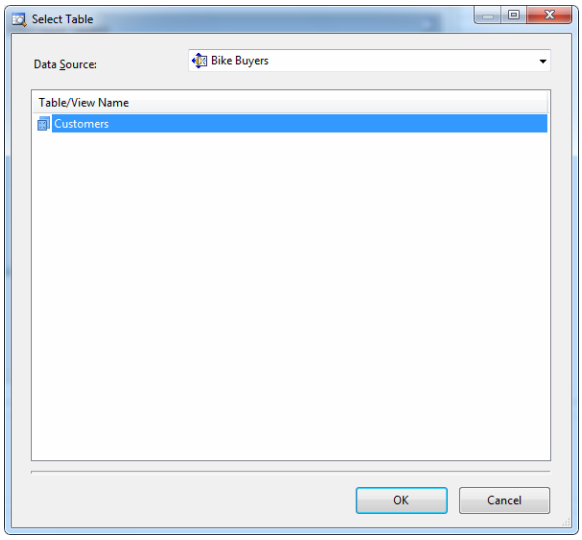


Εικόνα 21: Επιλογή συνόλου δεδομένων για αποτίμηση μοντέλου

Στο παράθυρο που ανοίγει, βλέπουμε τις στήλες της δομής Εξόρυξης Γνώσης που έχουμε δημιουργήσει. στο πεδίο **Select Input table** επιλέγουμε τον πίνακα δεδομένων, που θα χρησιμοποιηθεί για την αποτίμηση των μοντέλων.

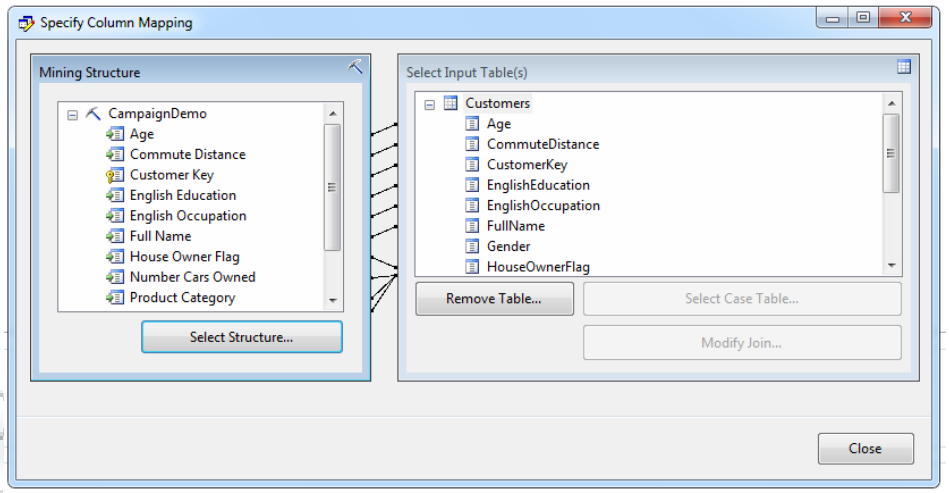


Πατώντας στην επιλογή **Select Case Table** ανοίγει τον παράθυρο για την επιλογή του Case Table που θα χρησιμοποιηθεί. Για χάριν ευκολίας της παρουσίασης, θα χρησιμοποιήσουμε το ίδιο σύνολο δεδομένων **Customers** που χρησιμοποιήσαμε για την δημιουργία των μοντέλων. Σε κάθε άλλη περίπτωση όμως και σε πραγματικά έργα Εξόρυξης ετοιμάζεται συγκεκριμένα για τον σκοπό της αποτίμηση των μοντέλων Εξόρυξης Γνώσης ένα συγκεκριμένο σύνολο δεδομένων, το οποίο επιλέγουμε στο βήμα αυτό.

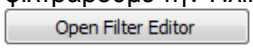


Επιλογή του πίνακα Customers και πατάμε **OK**

Ο αλγόριθμος έχει αντιστοιχίσει αυτόματα τις στήλες ανάλογα με τα ονόματα τους. Στην περίπτωση που θέλει να αντιστοιχίσει ο χρήστης μια στήλη, μπορεί να επιλέξει μια στήλη του μοντέλου και να την «σύρει» με το ποντίκι πάνω στη στήλη του πίνακα δεδομένων, που επιθυμεί να αντιστοιχίσει. Για παράδειγμα, επειδή έχουμε δημιουργήσει την στήλη *Yearly Income Discrete*, με τις διακριτές τιμές από τις συνεχείς τιμές της στήλης *Yearly Income*. Έτσι μπορούμε να επιλέξουμε με το ποντίκι την στήλη *Yearly Income Discrete* από το πλαίσιο Mining Structure και να την σύρουμε στην στήλη *Yearly Income* του πίνακα Customers στο πλαίσιο Select Input Table.

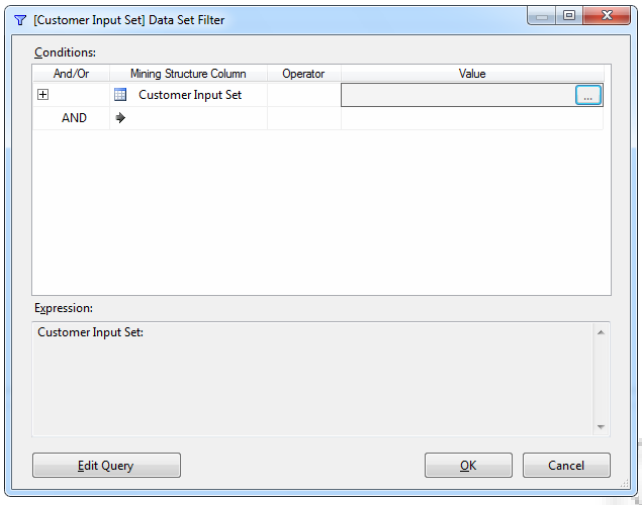


Εικόνα 22: Αντιστοίχιση στηλών δεδομένων του Μοντέλου με τις στήλες του Πίνακα

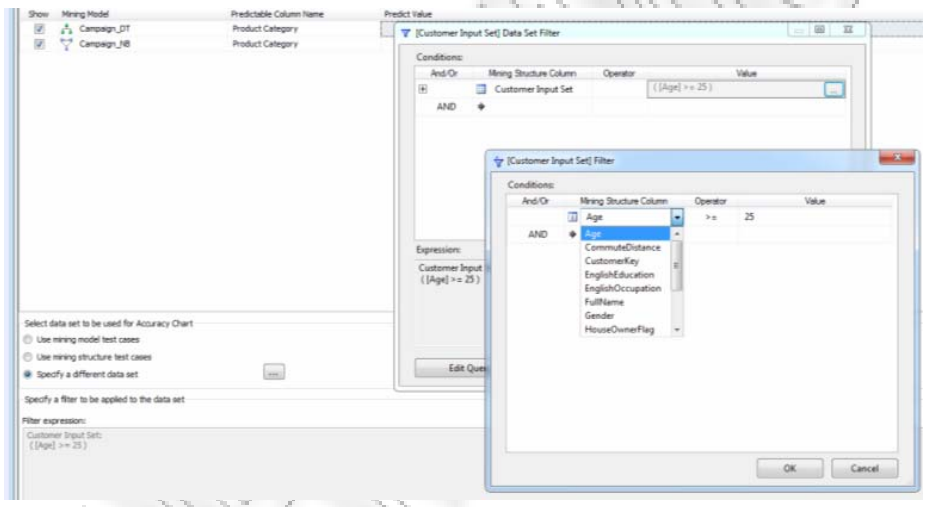
Επίσης ο χρήστης έχει την δυνατότητα να ορίσει φίλτρα στα δεδομένα που θα χρησιμοποιήσει για τον έλεγχο του μοντέλου. Στο συγκεκριμένο παράδειγμα, μπορούμε να φιλτράρουμε την Ηλικία (Age) των πελατών, έτσι ώστε είναι μεγαλύτερη από το 25. Πατάμε στο 

και ανοίγει το παράθυρο Data Set Filter. Πατάμε μέσα στο πεδίο Mining

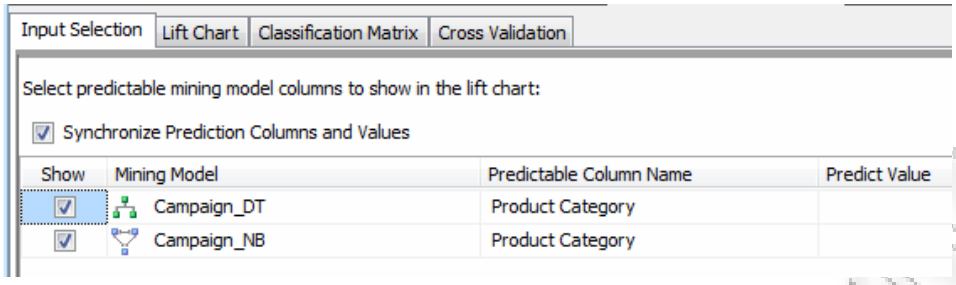
Structure Column και επιλέγουμε το **Customer Input Set**. Στη συνέχεια πατάμε στο πεδίο Value και επιλέγουμε το [...] για να ανοίξει το παράθυρο **Filter**.



Στο παράθυρο Filter που ανοίγει, στο πεδίο Mining Structure Columns επιλέγουμε την στήλη **Age**, στο πεδίο Operator επιλέγουμε το **>=** και στο πεδίο Value γράφουμε **25**.



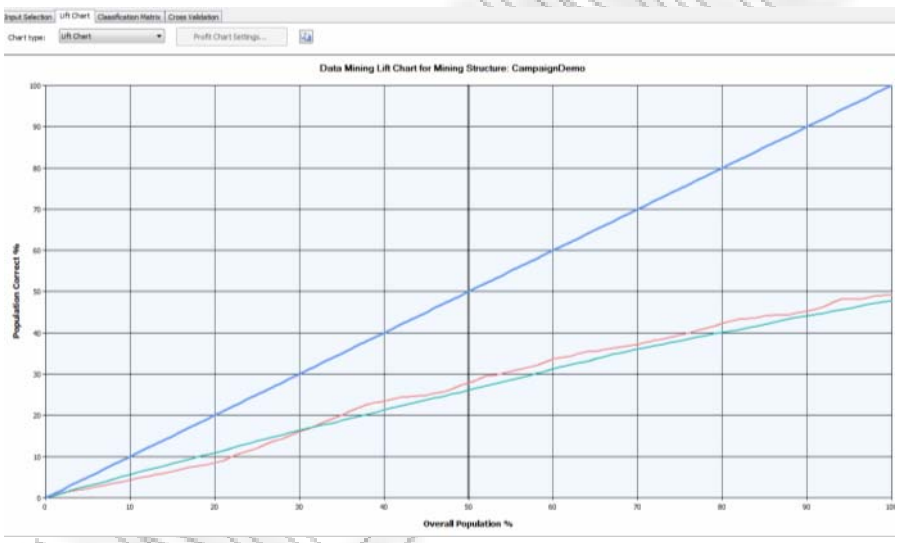
Έχουμε καθορίσει όλες τις επιλογές που χρειάζεται το BIDS για να συνεχίσει στην διαδικασία της αξιολόγησης των μοντέλων. Τα μοντέλα, που θέλουμε να χρησιμοποιηθούν στην αξιολόγηση τα επιλέγουμε στην λίστα που υπάρχει με τα διαθέσιμα μοντέλα, επιλέγοντας ή όχι το πεδίο **Show**.



Να σημειώσουμε ότι στο πεδίο **Predict Value** προς το παρόν δεν επιλέξαμε τίποτα. Την χρησιμότητα του πεδίου αυτού θα την δούμε στη συνέχεια.

Lift Chart

Το επόμενο βήμα είναι να πατήσουμε στο **tab Lift Chart** για να δούμε τα αποτελέσματα της αξιολόγησης των μοντέλων. Το lift chart συγκρίνει την ακρίβεια για τις προβλέψεις όλων των διαφορετικών τιμών του χαρακτηριστικού προς πρόβλεψη (του χαρακτηριστικού Product Category στη συγκεκριμένη περίπτωση) (ή για μια συγκεκριμένη προβλεπόμενη τιμή – το πεδίο Predict Value που αναφέραμε) για όλα τα μοντέλα Εξόρυξης Γνώσης που έχουμε συμπεριλάβει στην ανάλυση. Η σύγκριση γίνεται ως προς σε μια μέση πρόβλεψη των τιμών (που πραγματοποιεί το σύστημα) αλλά και ως προς την βέλτιστη πρόβλεψη (που επίσης πραγματοποιεί το σύστημα). Το lift chart φαίνεται στο παρακάτω διάγραμμα.

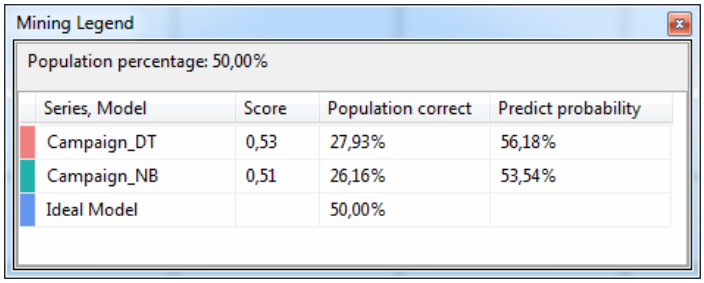


Εικόνα 23: Lift Chart

Οι άξονες του lift chart είναι οι εξής: ο κάθετος άξονας περιέχει το ποσοστό ορθότητας της πρόβλεψης του αποτελέσματος για το κάθε μοντέλο ενώ ο οριζόντιος άξονας περιέχει το ποσοστό του πληθυσμού που απαιτείται για να γίνει η πρόβλεψη.

Το lift chart περιέχει τρείς γραμμές. Η πρώτη γραμμή (η μπλε) δείχνει το ιδεατό αποτέλεσμα (το ιδεατό μοντέλο) ενώ οι υπόλοιπες δυο γραμμές αναφέρονται στα δύο μοντέλα Εξόρυξης Γνώσης, που έχουμε προσθέσει στο διάγραμμα. Στον κάθετο άξονα υπάρχει και μια κάθετη σε αυτόν γραμμή η οποία τέμνει τις γραμμές του διαγράμματος για την προβολή πληροφοριών. Όσο πιο κοντά βρίσκεται η γραμμή ενός μοντέλου στο ιδεατό αποτέλεσμα, τόσο καλύτερο είναι το μοντέλο αυτό.

Η κάθετη γραμμή όπως είπαμε δείχνει πληροφορίες για τα σημεία που τέμνει τις γραμμές του διαγράμματος. Η προεπιλεγμένη θέση της κάθετης γραμμής είναι για το 50% του πληθυσμού. Οι πληροφορίες φαίνονται στο υπόμνημα του διαγράμματος ή αν τοποθετήσει ο χρήστης τον δείκτη του ποντικιού στις τομές των αντίστοιχων γραμμών.

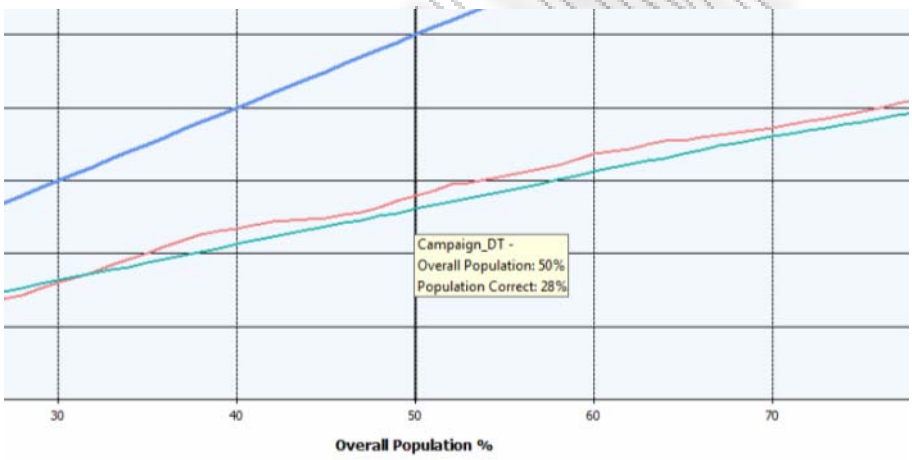


Από το υπόμνημα του διαγράμματος βλέπουμε τα παρακάτω:

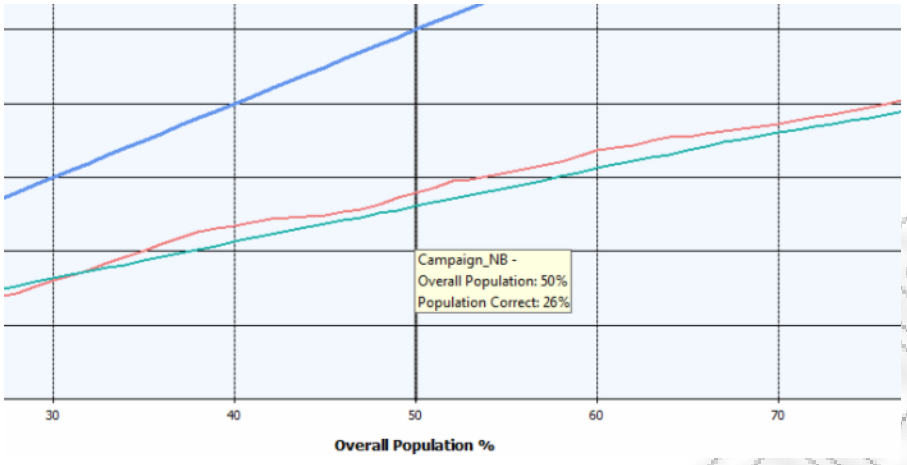
Για το ιδεατό μοντέλο, χρειάζεται το 50% του πληθυσμού (Population percentage: 50,00%) για να έχουμε ποσοστό επιτυχίας 50%. Αυτό είναι και το νόημα του ιδεατού μοντέλου, ότι όσους πελάτες χρησιμοποιήσουμε για πρόβλεψη, η πρόβλεψη θα είναι σωστή.

Για το μοντέλο των Δέντρων Απόφασης (Campaign_DT) στο 50% του πληθυσμού, θα έχουμε σωστά αποτελέσματα περίπου 28% ενώ για το μοντέλο Naïve Bayes (Campaign_NB) θα έχουμε σωστά επιτυχίας περίπου 26%.

Τις ίδιες πληροφορίες βλέπουμε αν τοποθετήσουμε τον δείκτη του ποντικιού στα σημεία τομής της κάθετης γραμμής με τις γραμμές των μοντέλων Δέντρων Απόφασης και Naïve Bayes.



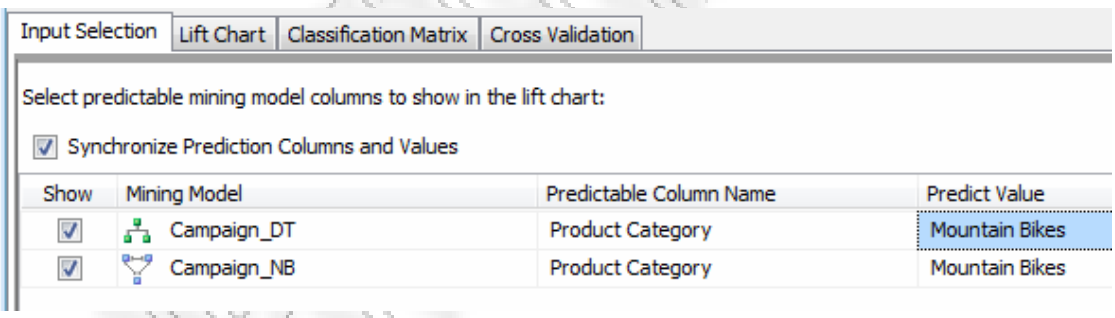
Το μοντέλο των Δέντρων Απόφασης, για το 50% του πληθυσμού υπολογίζει σωστά περίπου το 28% ενώ το μοντέλο του Naïve Bayes περίπου το 26%.



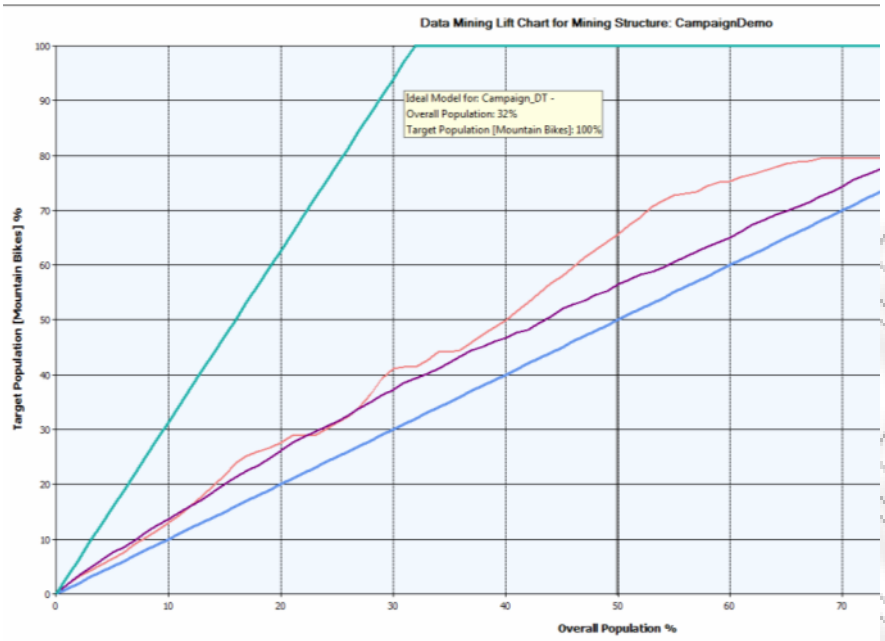
Ο χρήστης μπορεί κάνοντας κλικ πάνω στο διάγραμμα να τοποθετήσει την κάθετη γραμμή σε οποιοδήποτε σημείο του διαγράμματος και να δει τις πληροφορίες. Για παράδειγμα, μπορεί να πατήσει στον αριθμό **30** του οριζόντιου άξονα για να δει την επίδοση των μοντέλων στο 30% του πληθυσμού.

Πρόβλεψη μιας συγκεκριμένης Κατηγορίας

Τι γίνεται όμως στην περίπτωση που θέλουμε να δούμε την αξιολόγηση της πρόβλεψης των μοντέλων για μια συγκεκριμένη τιμή του χαρακτηριστικού. Για παράδειγμα, πόσο σωστά προβλέπουν οι αλγόριθμοι την αγορά ποδηλάτου από την κατηγορία Mountain Bikes, έτσι ώστε να προωθηθεί η κατηγορία αυτή από το τμήμα μάρκετινγκ της εταιρείας; Αυτό μπορεί να επιτευχθεί από το tab Input Selection, επιλέγοντας στο πεδίο Predict Value την τιμή Mountain Bikes του χαρακτηριστικού *ProductCategory* που προβλέπουν οι αλγόριθμοι.



Στην συνέχεια, αν επιλέξουμε το tab Lift Chart, βλέπουμε το παρακάτω διάγραμμα:



Εικόνα 24: Πρόβλεψη συγκεκριμένη τιμής

Στο διάγραμμα πλέον βλέπουμε 4 γραμμές. Η πράσινη γραμμή δείχνει το ιδεατό αποτέλεσμα (Ideal Model) (είναι η γραμμή που βρίσκεται πάνω από όλες τις άλλες), η μπλε γραμμή που δείχνει τα αποτελέσματα ενός εσωτερικού μοντέλου που πραγματοποιεί τυχαίες προβλέψεις (Random Guess Model) (είναι η γραμμή που βρίσκεται κάτω από τις υπόλοιπες γραμμές) και οι υπόλοιπες 2 γραμμές των μοντέλων Δέντρων Απόφασης και Naïve Bayes.

Από το διάγραμμα παρατηρούμε ότι για το ιδανικό μοντέλο, θα πρέπει να επικοινωνήσουμε με περίπου το 32% των πελατών για να καταλάβουμε ότι θα είναι αγοραστές της κατηγορίας Mountain bike (είναι το σημείο που η πράσινη γραμμή φτάνει στο 100% του κάθετου άξονα του διαγράμματος Target Population.

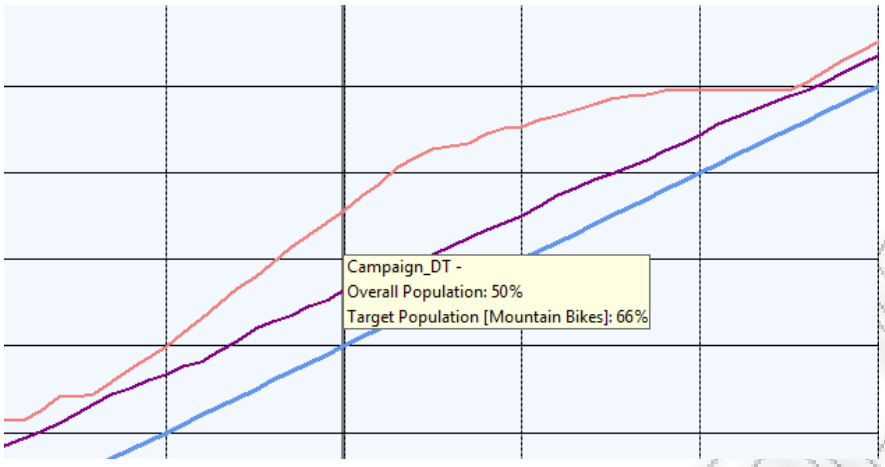
Στο υπόμνημα του διαγράμματος βλέπουμε τις παρακάτω πληροφορίες:

Mining Legend

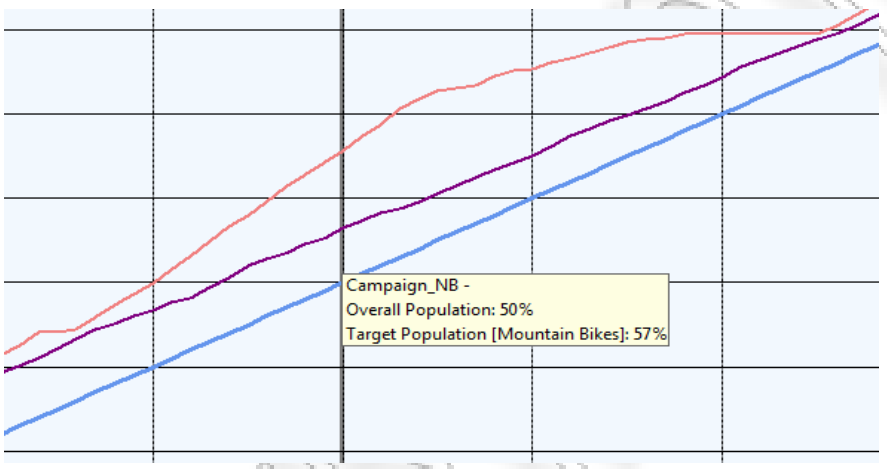
Population percentage: 50,00%

Series, Model	Score	Target population	Predict probability
Campaign_DT	0,69	65,69%	28,18%
Campaign_NB	0,65	56,52%	29,31%
Random Guess Model		50,00%	
Ideal Model for: Campaign_DT, Campaign_NB		100,00%	

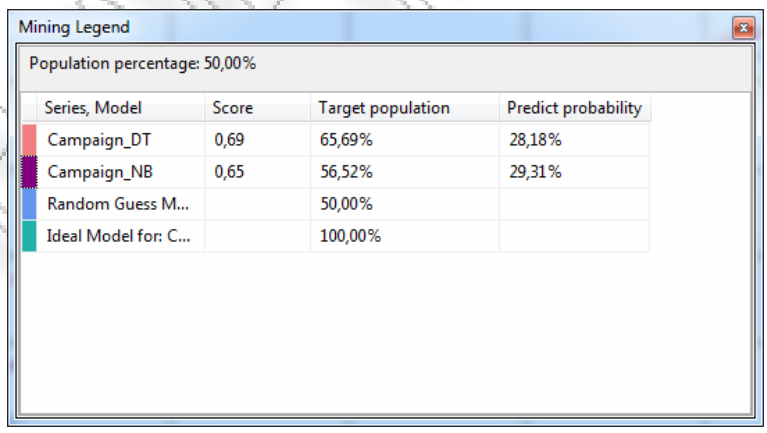
Το μοντέλο των Δέντρων Απόφασης είναι καλύτερο, γιατί για το 50% του πληθυσμού προβλέπει την αγορά της κατηγορίας Mountain Bike κατά 66% περίπου. Την ίδια πληροφορία βλέπουμε αν τοποθετήσουμε τον δείκτη του ποντικού πάνω στην γραμμή του μοντέλου Δέντρων Απόφασης.



Αντίθετα, το μοντέλο Naïve Bayes για το 50% του πληθυσμού προβλέπει περίπου το 57%.



Είναι σημαντικό επίσης να γνωρίζουμε ότι το Lift Chart δείχνει την πιθανότητα της πρόβλεψης μιας κατάστασης (για παράδειγμα την αγορά της κατηγορίας Mountain Bikes από τους πελάτες) και δεν δείχνει απαραίτητα την ορθότητα της πρόβλεψης. Με λίγα λόγια δείχνει την πιθανότητα να αγοράσει ένας πελάτης ή όχι το προϊόν μιας συγκεκριμένης κατηγορίας, δεν δείχνει ότι ο πελάτης θα αγοράσει σίγουρα το προϊόν αυτό.

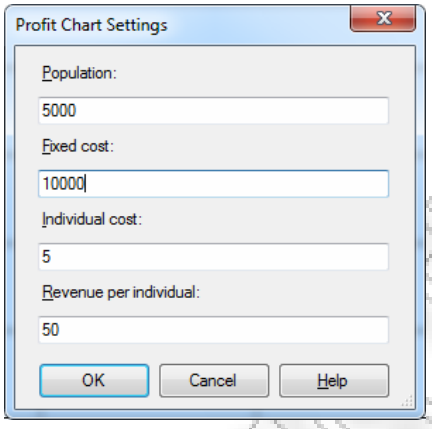


- **Score**: μια μετρική που συγκρίνει τα Μοντέλα Εξόρυξης Γνώσης. Επιλέγουμε το μοντέλο με το μεγαλύτερο score
- **Target Population**: για το 50% του πληθυσμού (Population Percentage 50%) δείχνει το ποσοστό που προβλέπει το κάθε μοντέλο (για παράδειγμα 66% περίπου τα Δέντρα Απόφασης και 56.5% το μοντέλο Naïve Bayes).
- **Predict Probability**: δείχνει την πιθανότητα, που έχει το κάθε μοντέλο να προβλέψει σωστά το ποσοστό του πληθυσμού, που εμφανίζεται στην στήλη Target Population.

Profit Chart

Με το Profit Chart εισάγουμε μια οικονομική διάσταση στην ανάλυση μας για να αξιολογήσουμε πόσο κέρδος μπορεί να προσφέρει το κάθε μοντέλο. Στο Profit chart μπορούμε να εισάγουμε ορισμένα οικονομικά στοιχεία και να υπολογίσουμε το κέρδος που περιμένουμε από το κάθε μοντέλο που αναλύουμε.

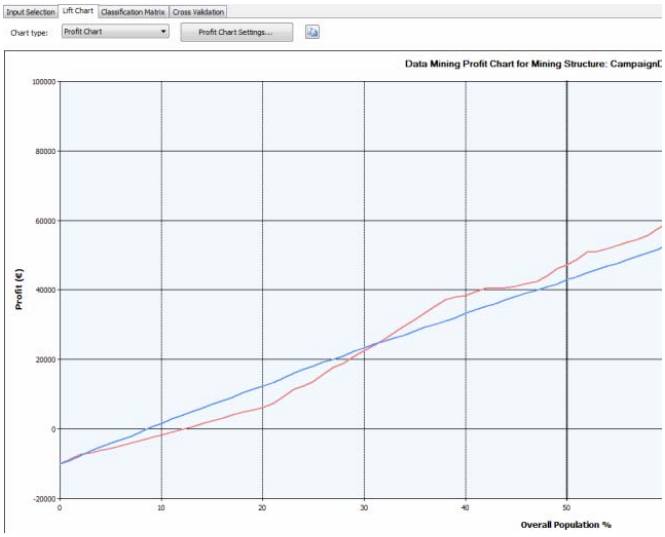
Για να προβάλλουμε το Profit Chart, στην λίστα **Chart type** επιλέγουμε **Profit Chart**. Με την επιλογή, ανοίγει το παρακάτω παράθυρο ρυθμίσεων:



Για το παράδειγμα που περιγράφουμε, έστω ότι το τμήμα μάρκετινγκ έχει υπολογίσει ορισμένα αριθμητικά στοιχεία για την διαφημιστική εκστρατεία, που πρόκειται να ξεκινήσει. Στην προκειμένη περίπτωση, θέτουμε τα παρακάτω στοιχεία.

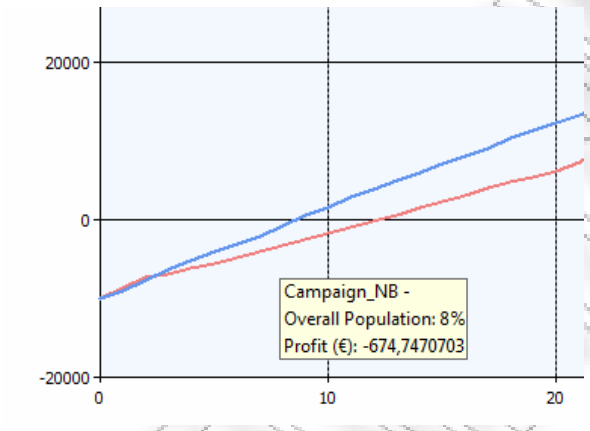
- **Population**: ο αριθμός του πληθυσμού, στον οποίο θα απευθυνθεί η εκστρατεία, έστω 5.000 πελάτες.
- **Fixed Cost**: το σταθερό κόστος της διαφημιστικής εκστρατείας, έστω 5.000 €.
- **Individual Cost**: το επιπλέον κόστος επικοινωνίας για τον κάθε μεμονωμένο πελάτη, έστω 5€
- **Revenue per Individual**: Προσδοκώμενο έσοδο από τον κάθε πελάτη, έστω 50€.

Το Profit Chart που δημιουργείται παρουσιάζεται παρακάτω:

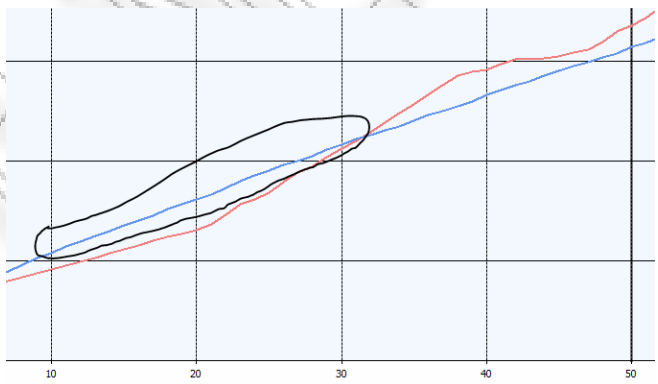


Εικόνα 25: Profit Chart

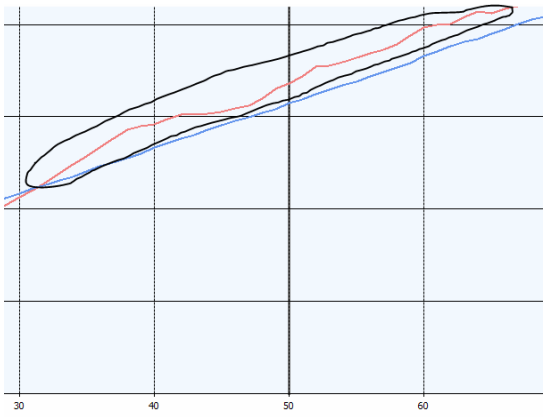
Από το Profit Chart βλέπουμε ότι η εταιρεία θα έχει ζημιά (αρνητικό κέρδος) στην περίπτωση που επικοινωνήσει με λιγότερους από το 10% του συνόλου των πελατών της.



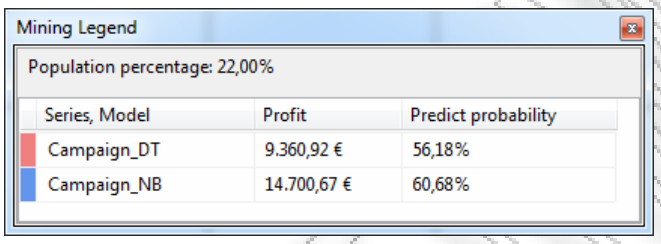
Στην συνέχεια παρατηρούμε ότι για το εύρος μεταξύ 10% και 30% του πληθυσμού, το μοντέλο Naïve Bayes (η μπλε γραμμή) προσφέρει μεγαλύτερο κέρδος.



Από το 30% του πληθυσμού και παραπάνω, το μοντέλο των Δέντρων Απόφασης προσφέρει και το μεγαλύτερο κέρδος.



Επίσης, από το Legend του Profit Chart βλέπουμε για παράδειγμα, ότι για το 50% του πληθυσμού το μοντέλο των Δέντρων Απόφασης παράγει ένα κέρδος των 47.314 €. Στην στήλη Predict Probability βλέπουμε και την αντίστοιχη πιθανότητα να παράγει το κάθε μοντέλο το αναμενόμενο κέρδος του.

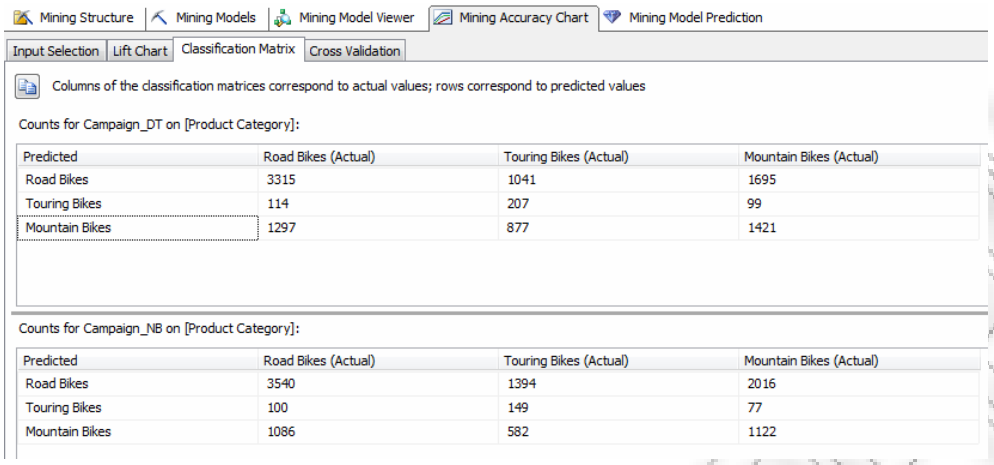


Με την χρήση των Profit Charts, οι χρήστες μπορούν να υλοποιήσουν What-If σενάρια για να υπολογίσουν το μέγιστο κέρδος (μέγιστη επιστροφή απόδοσης) από την επένδυση στην εκάστοτε διαφημιστική εκστρατεία, που πρόκειται να υλοποιήσει η εταιρεία.

Classification Matrix

Τα αποτελέσματα των διαγραμμάτων Lift και Profit charts προσφέρουν, όπως είδαμε σημαντικές πληροφορίες για την εγκυρότητα των μοντέλων Εξόρυξης Γνώσης. Σε αρκετές όπως περιπτώσεις, μπορεί να απαιτείται πιο λεπτομερής πληροφόρηση, ιδίως στις περιπτώσεις εκείνες που οι λανθασμένες αποφάσεις να κοστίζουν ακριβά τόσο για την εταιρεία όσο και για τους πελάτες. Μια πιο λεπτομερείς πληροφόρηση παρέχεται μέσα από το Classification Matrix, που προσφέρει το BIDS.

Στο **tab Classification Matrix**, παρουσιάζεται ένας πίνακας ο οποίος δείχνει τις διαφορετικές τιμές από το προβλεπόμενο χαρακτηριστικό που παράγει το κάθε μοντέλο και τις πραγματικές τιμές (αυτές που όντως πραγματοποιήθηκαν όπως προκύπτουν από τα πραγματικά στοιχεία του πίνακα δεδομένων). Για το παράδειγμά μας, εμφανίζονται οι παρακάτω πίνακες:



Εικόνα 26: Classification Matrix

Οι στήλες του πίνακα παρουσιάζουν τις πραγματικές τιμές ενώ οι γραμμές του πίνακα τις προβλεπόμενες τιμές, για την κάθε κατηγορία. Για παράδειγμα, βλέπουμε ότι για την κατηγορία Road bikes (πρώτη γραμμή του πίνακα) έγινε πρόβλεψη για 3.315 αγορές και πραγματοποιήθηκαν όντως 3.315 αγορές από την κατηγορία αυτή, ενώ για την ίδια κατηγορία προϊόντος, τελικά αγοράστηκαν 1.041 από την κατηγορία Touring bikes και 1.695 από την κατηγορία Mountain Bikes. Μια σημαντική πληροφόρηση δίνεται από την διαγώνιο του πίνακα, η οποία περιέχει τις σωστές προβλέψεις του κάθε μοντέλου.

Έτσι παρατηρούμε, ότι για το μοντέλο των Δέντρων Απόφασης, έγιναν σωστές προβλέψεις για 3.315 + 207 + 1.421 = 4.943 προϊόντα ενώ για το μοντέλο Naïve Bayes έγιναν σωστές προβλέψεις για 3.540 + 149 + 1.122 = 4.811 προϊόντα.

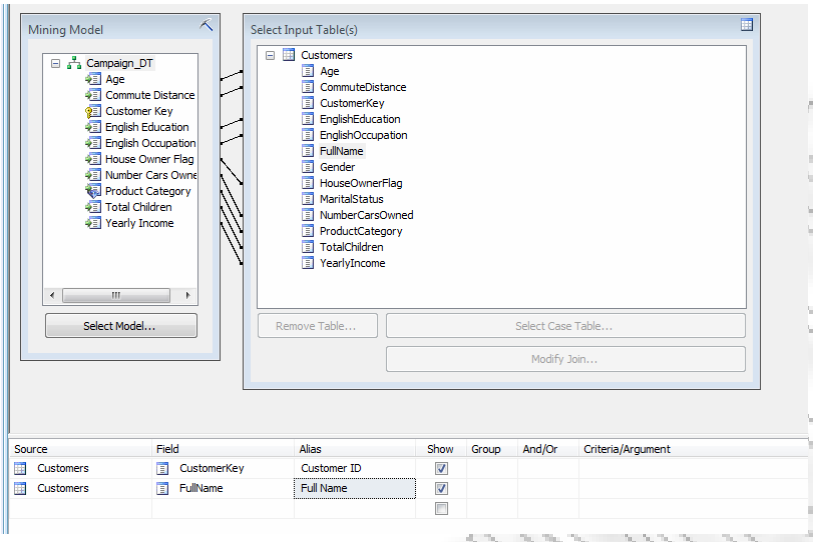
Δημιουργία Προβλέψεων

Η δυνατότητα δημιουργίας προβλέψεων με την χρήση των μοντέλων Εξόρυξης Γνώσης είναι μια πολύ σημαντική λειτουργία που παρέχεται από το BIDS. Οι προβλέψεις γίνονται πάνω σε καινούργια δεδομένα (για παράδειγμα γίνεται πρόβλεψη αγοράς προϊόντων για καινούργιους πελάτες που έχουν συγκεκριμένα χαρακτηριστικά) με την χρήση εντολών DMX. Το περιβάλλον του BIDS δίνει την δυνατότητα στους χρήστες να σχεδιάσουμε ερωτήματα DMX μέσα από ένα γραφικό περιβάλλον (editor) όπως ακριβώς δίνεται η δυνατότητα σχεδιασμού ερωτημάτων SQL από τους αντίστοιχους editors του SQL Server Management Studio. Για να πραγματοποιηθεί μια πρόβλεψη, μπορούμε να χρησιμοποιήσουμε την DMX εντολή SELECT μέσα από το BIDS.

Για να σχεδιάσουμε τις DMX εντολές για τις προβλέψεις, θα χρησιμοποιήσουμε (για χάριν ευκολίας) τα ίδια δεδομένα που χρησιμοποιήσαμε για την εκπαίδευση του μοντέλου. Η πρόβλεψη γίνεται μέσα από το **tab Mining Model Prediction**, με την χρήση του editor για να δημιουργήσουμε μέσα από ένα γραφικό περιβάλλον τα αντίστοιχα DMX ερωτήματα. Έστω ότι θέλουμε να προβλέψουμε αν ένας πελάτης θα αγοράσει ποδήλατο από την κατηγορία Mountain Bikes.

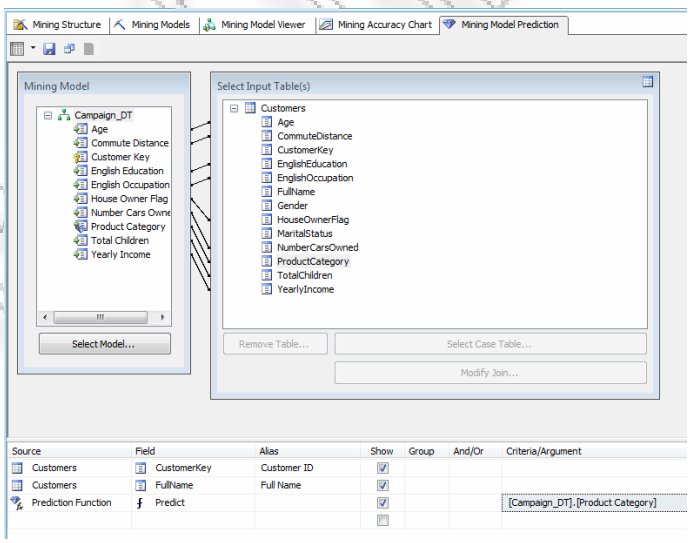
Το πρώτο βήμα που κάνουμε είναι να επιλέξουμε ένα συγκεκριμένο μοντέλο Εξόρυξης Γνώσης που θέλουμε να χρησιμοποιήσουμε για την πρόβλεψη (στο παράδειγμα μας θα χρησιμοποιήσουμε μοντέλο των Δέντρων Απόφασης αφού μέσα από την διαδικασία αποτίμησης των μοντέλων είδαμε ότι προσφέρει τα καλύτερα αποτελέσματα) και το σύνολο των δεδομένων που θα χρησιμοποιηθεί για την. Στο tab Mining Model Prediction, πατάμε στην επιλογή **Select Case Table** και από το παράθυρο που ανοίγει επιλέγουμε τον πίνακα **Customers**. Με την επιλογή του πίνακα, αντιστοιχούνται αυτόματα οι στήλες δεδομένων του μοντέλου με τις στήλες του πίνακα δεδομένων (η αντιστοίχιση, όπως έχουμε τονίσει γίνεται κυρίως βάσει του ονόματος των στηλών).

Το δεύτερο βήμα είναι να προβλέψουμε την τιμή της στήλης *CustomerKey*, να δούμε δηλαδή αν κάποιος (νέος) πελάτης με ένα συγκεκριμένο κωδικό θα αγοράσει κάποια κατηγορία ποδηλάτου. Για το λόγο αυτό θα σύρουμε με το ποντίκι την στήλη *CustomerKey* από τον πίνακα δεδομένων (Input Table) στην πρώτη γραμμή του grid, στην στήλη **Source** και σαν alias γράφουμε **Customer ID**. Επίσης, σέρνουμε την στήλη *FullName* στην δεύτερη γραμμή και δίνουμε alias **Full Name** (για να δείξουμε στα αποτελέσματα και το όνομα του πελάτη).

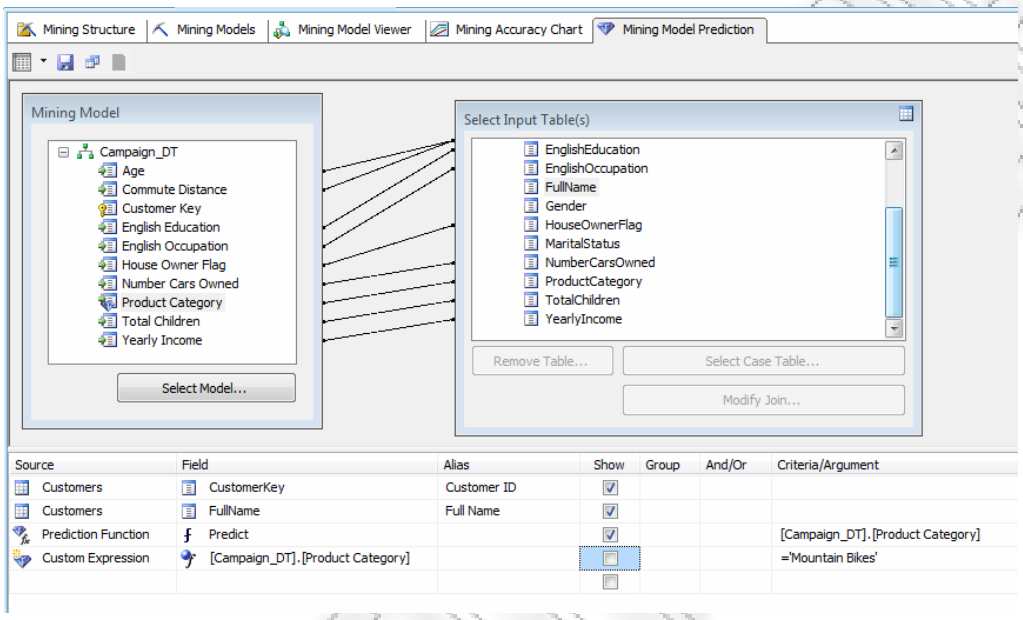


Το τρίτο βήμα είναι να γράψουμε τις εντολές DMX, οι οποίες θα πραγματοποιήσουν την πρόβλεψη. Η δημιουργία των εντολών θα γίνει μέσα από τον τρέχον designer. Η συνάρτηση που θέλουμε να χρησιμοποιήσουμε είναι η **Predict**. Στην τρίτη γραμμή του grid, στην στήλη **Source**, επιλέγουμε από την λίστα που ανοίγει την εντολή **Prediction Function**. Στην στήλη Field, επιλέγουμε την (πρώτη από την λίστα που εμφανίζεται) συνάρτηση **Predict**.

Στην στήλη **Criteria / Argument** εμφανίζεται το πρότυπο των παραμέτρων τις οποίες μπορούμε να χρησιμοποιήσουμε στην συνάρτηση. Στο πεδίο αυτό, θα πρέπει να ορίσουμε την στήλη δεδομένων του μοντέλου την οποία θέλουμε να προβλέψουμε. Θα χρησιμοποιήσουμε την στήλη *ProductCategory* (θέλουμε δηλαδή να προβλέψουμε ποια κατηγορία προϊόντων πρόκειται να αγοράσει ο πελάτης). Για το λόγο αυτό, σέρνουμε την στήλη *ProductCategory* από τον την λίστα των πεδίων του μοντέλου **Campaign_DT** στην στήλη Criteria / Argument.



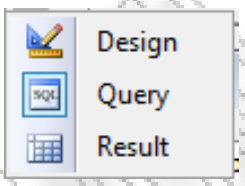
Το τέταρτο βήμα που πρέπει να κάνουμε είναι να ορίσουμε την κατηγορία προϊόντος, την οποία θέλουμε να προβλέψουμε. Στην αρχή του παραδείγματος αναφέραμε ότι θέλουμε να προβλέψουμε αν ένας πελάτης θα αγοράσει ποδήλατο από την κατηγορία Mountain Bikes. Έτσι, στην τέταρτη γραμμή του grid θα φιλτράρουμε τα δεδομένα της στήλης *ProductCategory*. Στην στήλη **Source**, επιλέγουμε από την λίστα το **Custom Expression** και σέρνουμε με το ποντίκι την στήλη *ProductCategory* από λίστα στηλών του μοντέλου Campaign_DT στην στήλη **Field**. Για να ορίσουμε το φίλτρο που θέλουμε να εφαρμόσουμε, στην στήλη **Criteria / Argument** γράφουμε **='Mountain Bikes'**.



Εικόνα 27: Σχεδιασμός DMX εντολής για Πρόβλεψη

Επίσης, δεν επιλέγουμε το **Show**, γιατί δεν θέλουμε η στήλη αυτή να εμφανίζεται στα αποτελέσματα, την θέλουμε μόνο για το φιλτράρισμα των δεδομένων.

Πατώντας στο κουμπί που βρίσκεται αριστερά του Designer, βλέπουμε τις διαθέσιμες επιλογές που έχουμε:



Η επιλογή **Design** είναι η μορφή του Designer που δουλεύουμε τώρα, δηλαδή κατασκευάζουμε το DMX ερώτημα με γραφικό τρόπο. Πατώντας στην επιλογή **Query**, μπορούμε να δούμε το ερώτημα DMX που παράγεται και είναι το παρακάτω:

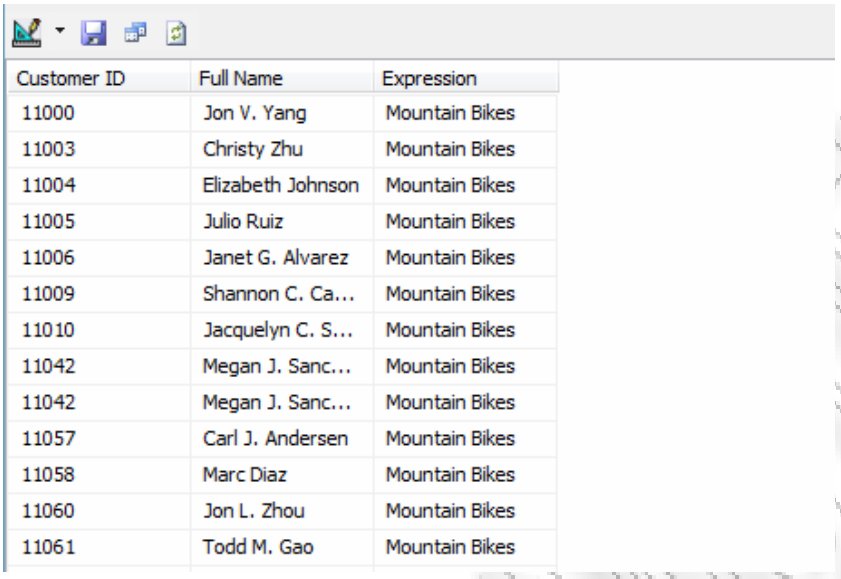
```
SELECT
  (t.[CustomerKey]) as [Customer ID],
  (t.[FullName]) as [Full Name],
  Predict([Campaign_DT].[Product Category]),
```

```
  [Campaign_DT].[Product Category] ='Mountain Bikes'
From
  [Campaign_DT]
PREDICTION JOIN
  OPENQUERY([Adventure Works DW],
    'SELECT
      [CustomerKey],
      [FullName],
      [YearlyIncome],
      [TotalChildren],
      [EnglishEducation],
      [NumberCarsOwned],
      [CommuteDistance],
      [EnglishOccupation],
      [HouseOwnerFlag],
      [Age],
      [ProductCategory]
    FROM
      (SELECT     C.CustomerKey, C.FirstName + " " + ISNULL(C.MiddleName + ". ", "") + C.LastName AS
FullName, C.MaritalStatus, C.Gender, C.YearlyIncome,
                C.TotalChildren,     C.EnglishEducation,     C.NumberCarsOwned,     C.CommuteDistance,
C.EnglishOccupation, C.HouseOwnerFlag, DATEDIFF(yy, C.BirthDate,
                GETDATE()) AS Age, CustomerFilter.Subcategory AS ProductCategory
FROM        DimCustomer AS C INNER JOIN
                  (SELECT    C.CustomerKey, PS.EnglishProductSubcategoryName AS Subcategory
                   FROM        DimCustomer AS C INNER JOIN
                               FactInternetSales AS S ON C.CustomerKey = S.CustomerKey INNER
JOIN
                               DimProduct AS P ON S.ProductKey = P.ProductKey INNER JOIN
                               DimProductSubcategory AS PS ON P.ProductSubcategoryKey =
PS.ProductSubcategoryKey
                   WHERE      (PS.ProductCategoryKey = 1)
                   GROUP BY C.CustomerKey, PS.EnglishProductSubcategoryName
                   HAVING          (COUNT(PS.ProductSubcategoryKey) = 1)) AS CustomerFilter ON
C.CustomerKey = CustomerFilter.CustomerKey) as [Customers]
    ') AS t
ON
  [Campaign_DT].[Yearly Income] = t.[YearlyIncome] AND
  [Campaign_DT].[Total Children] = t.[TotalChildren] AND
  [Campaign_DT].[English Education] = t.[EnglishEducation] AND
  [Campaign_DT].[Number Cars Owned] = t.[NumberCarsOwned] AND
  [Campaign_DT].[Commute Distance] = t.[CommuteDistance] AND
  [Campaign_DT].[English Occupation] = t.[EnglishOccupation] AND
  [Campaign_DT].[House Owner Flag] = t.[HouseOwnerFlag] AND
  [Campaign_DT].[Age] = t.[Age] AND
  [Campaign_DT].[Product Category] = t.[ProductCategory]
```

Όπως βλέπουμε χρησιμοποιείται η εντολή **PREDICTION JOIN ON** για να γίνει η σύνδεση των στηλών του μοντέλου Campaign_DT με τις στήλες που προέρχονται από την

Βάση Δεδομένων. Επίσης χρησιμοποιείται η εντολή Predict πάνω στην στήλη Product Category του μοντέλου και χρησιμοποιείται το φίλτρο για την κατηγορία Mountain Bikes.

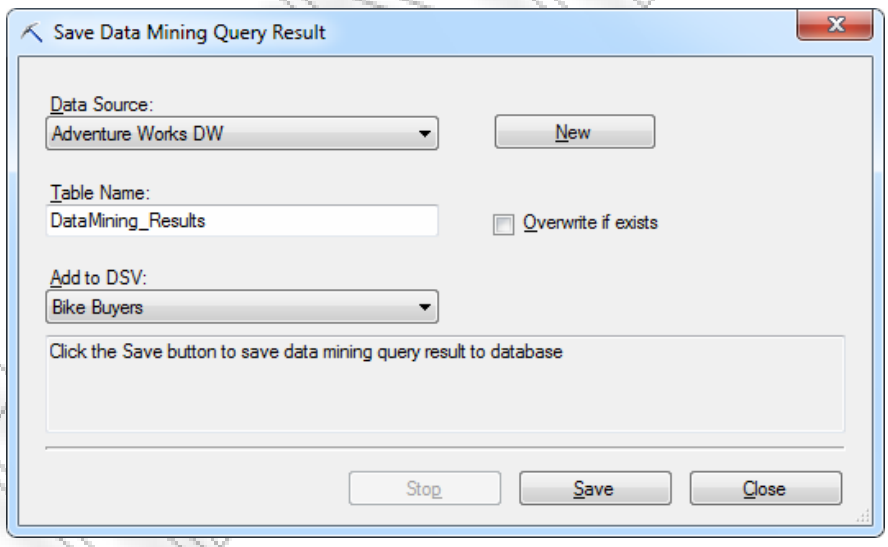
Αν πατήσουμε στην επιλογή **Result**, βλέπουμε τα αποτελέσματα της εκτέλεσης του DMX ερωτήματος.



Customer ID	Full Name	Expression
11000	Jon V. Yang	Mountain Bikes
11003	Christy Zhu	Mountain Bikes
11004	Elizabeth Johnson	Mountain Bikes
11005	Julio Ruiz	Mountain Bikes
11006	Janet G. Alvarez	Mountain Bikes
11009	Shannon C. Ca...	Mountain Bikes
11010	Jacquelyn C. S...	Mountain Bikes
11042	Megan J. Sanc...	Mountain Bikes
11042	Megan J. Sanc...	Mountain Bikes
11057	Carl J. Andersen	Mountain Bikes
11058	Marc Diaz	Mountain Bikes
11060	Jon L. Zhou	Mountain Bikes
11061	Todd M. Gao	Mountain Bikes

Για το παράδειγμα μας, το ερώτημα μας επέστρεψε 3.595 πελάτες που μπορεί να ενδιαφέρονται για την αγορά ποδηλάτου τύπου Mountain Bikes.

Πατώντας στο  (Save query results) μπορούμε να αποθηκεύσουμε τα δεδομένα σε μια σχεσιακή Βάση Δεδομένων (συνήθως στην βάση δεδομένων από την οποία έχουμε αντλήσει τα δεδομένα μας).

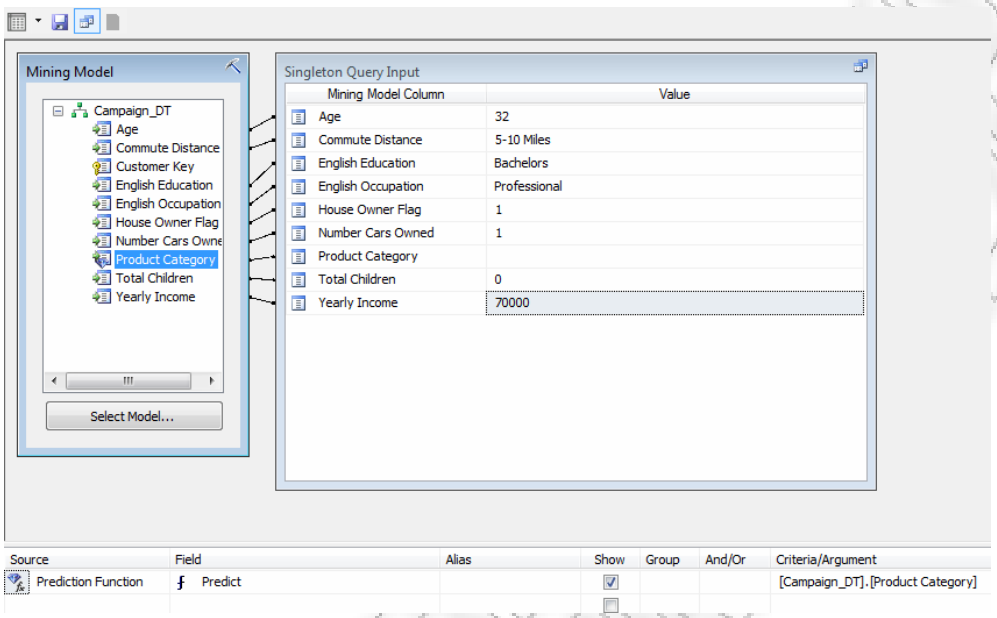


Εικόνα 28: Αποθήκευση αποτελεσμάτων πρόβλεψης

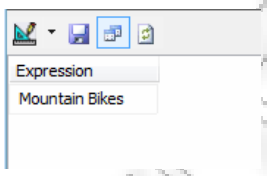
Singleton Query

Στην περίπτωση που θέλουμε να προβλέψουμε την τιμή για ένα συγκεκριμένο πελάτη (δίνοντας στα στοιχεία του) και όχι από την λίστα των υποψήφιων πελατών, επιλέγουμε το κουμπί  (**Singleton Query**).

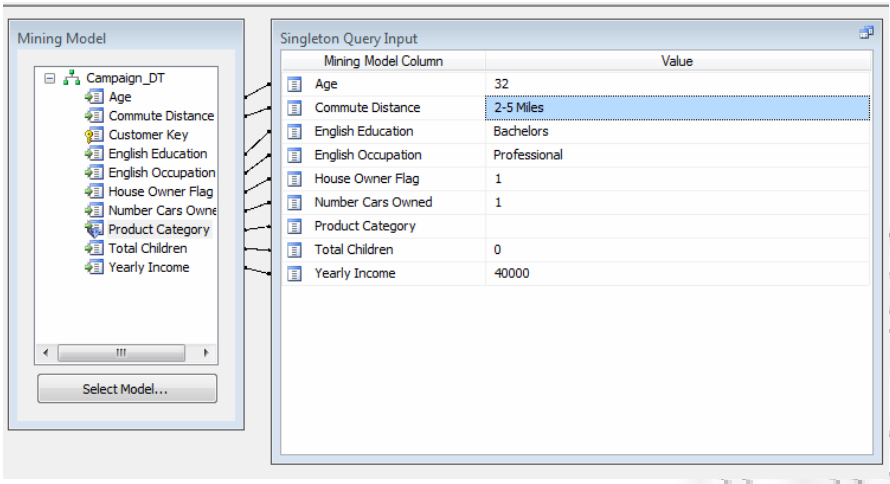
Στο παράθυρο που εμφανίζεται, μπορούμε να δώσουμε τα στοιχεία για ένα συγκεκριμένο πελάτη και να ελέγξουμε αν είναι πιθανός αγοραστής.



Στον designer δίνουμε τα στοιχεία του πελάτη, που θέλουμε να ελέγξουμε και πατάμε **Results**, για να δούμε τα αποτελέσματα της πρόβλεψης:

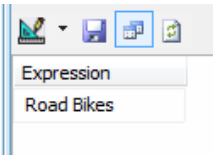


Όπως βλέπουμε, το μοντέλο πρόβλεψε ότι ο πελάτης πιθανόν να ενδιαφέρεται για Mountain Bikes. Αν αλλάξουμε τα δεδομένα του ερωτήματος, για παράδειγμα αν καταχωρήσουμε μικρό εισόδημα για τον πελάτη και μικρή απόσταση από το κέντρο της πόλης,



Εικόνα 29: Καταχώρηση τιμών για Singleton Query

Το αποτέλεσμα της πρόβλεψης είναι διαφορετικό:



Το αντίστοιχο DMX ερώτημα που παράγεται από τα παραπάνω είναι το:

```
SELECT
  Predict([Campaign_DT].[Product Category])
From
  [Campaign_DT]
NATURAL PREDICTION JOIN
(SELECT 32 AS [Age],
  '2-5 Miles' AS [Commute Distance],
  'Bachelors' AS [English Education],
  'Professional' AS [English Occupation],
  '1' AS [House Owner Flag],
  1 AS [Number Cars Owned],
  0 AS [Total Children],
  40000 AS [Yearly Income]) AS t
```

Το συγκεκριμένο ερώτημα μπορούμε να το αποθηκεύσουμε και να το τρέχουμε στον Analysis Server αλλάζοντας τις τιμές των παραμέτρων. Επίσης, μπορούμε το ερώτημα να το χρησιμοποιήσουμε σε μια custom εφαρμογή, που έχουμε δημιουργήσει για να υποστηρίξουμε τις ανάγκες του οργανισμού.

6. Επίλογος

Συμπεράσματα

Στην παρούσα εργασία παρουσιάσαμε την έννοια της Επιχειρηματικής Ευφυΐας (Business Intelligence) και τις πληροφορίες που χρειάζονται για την χρήση των εφαρμογών BI στα διάφορα επίπεδα λειτουργίας σε ένα οργανισμό. Η προσφορά μιας εφαρμογής BI σε έναν οργανισμό συνοψίζεται στην έγκαιρη παροχή σωστών και εμπλουτισμένων πληροφοριών, έτσι ώστε να υποστηριχτεί η αποτελεσματική λήψη αποφάσεων για την υλοποίηση της στρατηγικής του οργανισμού. Για να επιτύχει όμως ένα έργο BI σε ένα οργανισμό, πρέπει να υπάρχει σωστή συνεργασία μεταξύ των αναλυτών, των εργαζομένων και των στελεχών, που θα λάβουν τις αποφάσεις, έτσι ώστε να καθοριστούν σωστά τα δεδομένα που θα χρησιμοποιηθούν και οι δείκτες απόδοσης που θα δημιουργηθούν για να μετρήσουν την απόδοση του οργανισμού. Παρουσιάσαμε επίσης το γεγονός, ότι στην πληθώρα των δεδομένων, που υπάρχουν σε ένα οργανισμό, υπάρχει πολύτιμη γνώση που είναι κρυμμένη και μπορεί να ανακαλυφθεί με τις τεχνικές Εξόρυξης Γνώσης (Data Mining), για την ανακάλυψη τάσεων, προτύπων στα δεδομένα και των συσχετίσεων μεταξύ τους. Οι τεχνικές Εξόρυξης Δεδομένων που υπάρχουν μπορούν να χρησιμοποιηθούν για την επίλυση αρκετών επιχειρηματικών προβλημάτων.

Οι λύσεις που παρέχονται από διάφορες εταιρείες για την δημιουργία λύσεων BI και συγκεκριμένα για χρήση τεχνικών Εξόρυξης Γνώσης είναι αρκετές, μια από τις οποίες που παρουσιάσαμε προσφέρονται από την Microsoft μέσα από τα Analysis Services (SSAS) του SQL Server (2005/2008). Παρουσιάσαμε τα κυριότερα χαρακτηριστικά των βασικών αλγορίθμων Εξόρυξης Γνώσης που μπορούν να υλοποιηθούν από τα SSAS (Decision Trees, Clustering, Association Rules) και την ανάπτυξη μιας εφαρμογής χρήσης των μοντέλων Εξόρυξης Γνώσης μέσα από το Business Intelligence Development Studio (BIDS) και τα SSAS του SQL Server. Σκοπός της παρουσίασης της ανάπτυξης της εφαρμογής είναι να δείξουμε τα βασικά βήματα, που μπορεί να ακολουθήσει ένας χρήστης για την επίλυση ενός επιχειρηματικού προβλήματος, μέσα από το περιβάλλον BIDS ακολουθώντας την μεθοδολογία ανάπτυξης έργων Εξόρυξης Γνώσης CRISP-DM. Μέσα από το περιβάλλον BIDS του Visual Studio αλλά και με την δυνατότητα παρουσίασης των αποτελεσμάτων της Εξόρυξης Γνώσης στο περιβάλλον του Office Excel και αναφορές που δημιουργούνται από τους χρήστες μέσω των Reporting Services, παρέχεται η δυνατότητα στους χρήστες να επιλύσουν οι ίδιοι συγκεκριμένα επιχειρηματικά προβλήματα και να εξάγουν τα αποτελέσματα που επιθυμούν. Υποστηρίζεται έτσι ο στόχος για το Προσωπικό BI, όπου ο κάθε χρήστης να μπορεί να επιλύσει τα δικά του προβλήματα, έχοντας πρόσβαση σε συγκεκριμένες πληροφορίες που τον αφορούν, άμεσα και την στιγμή της λήψης της απόφασης, για την υλοποίηση της στρατηγικής του οργανισμού.

Δυσκολίες – Προβλήματα

Ένα από τα προβλήματα που αντιμετωπίσαμε κατά την διάρκεια υλοποίησης της παρούσας εργασίας ήταν η εξεύρεση ενός συνόλου δεδομένων για την ανάπτυξη του παραδείγματος εφαρμογής των τεχνικών Εξόρυξης Γνώσης μέσα από το BIDS. Τα πραγματικά δεδομένα, από επιχειρήσεις είναι αρκετά δύσκολο να βρεθούν ενώ πολλές τεχνικές Εξόρυξης Γνώσης (για παράδειγμα της Κατηγοριοποίησης, της Ομαδοποίησης ή των Κανόνων Συσχέτισης) χρειάζονται ένα αρκετά μεγάλο αριθμό περιπτώσεων (cases) για να μπορούν να βγάλουν χρήσιμα αποτελέσματα. Στην διάθεσή μας είχαμε ένα σύνολο δεδομένων για στοιχεία πελατών της εταιρείας ΒΙΟΣΕΡ (εταιρεία παραγωγή ορών και παρεντερικών διαλυμάτων) αλλά ο αριθμός των δεδομένων ήταν πολύ μικρός, ενώ το περιεχόμενο των δεδομένων ήταν τέτοιο (στοιχεία

πωλήσεων για νοσοκομεία και φαρμακεία) που οι συσχετίσεις μεταξύ των δεδομένων ήταν προφανής, χωρίς την ανάγκη χρήσης τεχνικών Εξόρυξης Γνώσης.

Θεωρήσαμε ότι στόχος της εργασίας (όσον αφορά το πρακτικό τμήμα της) είναι να παρουσιαστεί ο τρόπος χρήσης των τεχνικών Εξόρυξης Γνώσης μέσα από το περιβάλλον SSAS του SQL Server και τα βήματα που πρέπει να ακολουθήσει ο χρήστης για να υλοποιήσει ένα έργο (project) Εξόρυξης Γνώσης με το Business Intelligence Development Studio και όχι να επιλυθεί ένα συγκεκριμένο πρόβλημα Εξόρυξης Γνώσης και να παρουσιαστούν τα αποτελέσματα της επίλυσης του αλγορίθμου που θα χρησιμοποιηθεί. Για το λόγο αυτό, αποφασίστηκε να χρησιμοποιηθεί η Βάση Δεδομένων Adventure Works, που παρέχεται από το site Codeplex και χρησιμοποιείται για την εκμάθηση των λειτουργιών του SQL Server. Το πρόβλημα που παρουσιάστηκε προς επίλυση, η δημιουργία μιας διαφημιστικής εκστρατείας για την προώθηση ενός ποδηλάτου, μπορεί να φαντάζει απλό, αλλά μέσα από αυτή την απλότητα μπορέσαμε να παρουσιάσουμε τα κυριότερα βήματα που μπορεί να ακολουθήσει ο χρήστης για να επιλύσει ένα πρόβλημα με την χρήση τεχνικών Εξόρυξης Γνώσης.

Μελλοντικές Επεκτάσεις

Η παρουσίαση των αποτελεσμάτων ενός έργου Επιχειρηματικής Ευφυΐας και πιο συγκεκριμένα ενός έργου BI για την επίλυση επιχειρηματικών προβλημάτων με την χρήση τεχνικών Εξόρυξης Γνώσης είναι από τα πλέον σημαντικότερα θέματα που πρέπει να αντιμετωπιστούν κατά την διάρκεια ανάπτυξης του έργου. Ο στόχος των εφαρμογών που χρησιμοποιούν τεχνικές Εξόρυξης Γνώσης, ανάμεσα τους και τα Analysis Services του SQL Server είναι να μπορούν οι ίδιοι οι χρήστες της εφαρμογής να βλέπουν μια εικόνα των αποτελεσμάτων (με χρήση διαγραμμάτων, πινάκων, κανόνων συσχέτισης κλπ) για να μπορούν να κατανοήσουν το αποτέλεσμα την ανάλυσης, χωρίς την παρουσία κάποιου ειδικού του χώρου της Εξόρυξης Γνώσης. Είδαμε ότι μέσα από το περιβάλλον του BIDS οι χρήστες μπορούν να δουν διάφορα διαγράμματα σχετικά με τα αποτελέσματα των τεχνικών Εξόρυξης Γνώσης, που έχουν χρησιμοποιήσει για την επίλυση των προβλημάτων τους. Τι γίνεται όμως στην περίπτωση που οι χρήστες των τεχνικών Εξόρυξης Γνώσης ή καλύτερα τα στελέχη του οργανισμού που θέλουν να λάβουν απαντήσεις σχετικά με συγκεκριμένα επιχειρηματικά προβλήματα για να λάβουν αποφάσεις για την εφαρμογή της στρατηγικής του οργανισμού, δεν έχουν πρόσβαση στο περιβάλλον των Analysis Services του SQL Server ή στο περιβάλλον του BIDS αλλά έχουν μάθει να δουλεύουν με το Excel ή κάποια άλλη εφαρμογή του Microsoft Office;

Την λύση την δίνει ο SQL Server με την δυνατότητα προβολής των αποτελεσμάτων της Εξόρυξης Γνώσης στο Excel μέσα από την χρήση συγκεκριμένων επεκτάσεων για το Excel (Add-ins). Πολλές εφαρμογές του Office (εκδόσεις 2007 και μεγαλύτερες) έχουν σχεδιαστεί για να λειτουργούν σαν «πελάτες» Εξόρυξης Γνώσης (data mining clients) για την παρουσίαση των αποτελεσμάτων της ανάλυσης των δεδομένων και της εφαρμογής τεχνικών Εξόρυξης Γνώσης μέσα από τα Analysis Services του SQL Server. Οι χρήστες μπορούν να εγκαταστήσουν τα Data Mining Add-ins που παρέχονται από τον SQL Server στο περιβάλλον του Office για να μπορούν να βλέπουν τα αποτελέσματα ή να εκτελούν κάποιες τεχνικές Εξόρυξης Γνώσης μέσα από το Excel ή το Visio. Προϋπόθεση όμως για να δουλεύουν τα data mining add-ins είναι να τρέχουν σε ένα analysis server για να μπορεί να δέχεται τα αιτήματα της ανάλυσης και να επιστρέφει τα αποτελέσματα. Μια επέκταση της παρούσας εργασίας είναι να παρουσιαστούν όλες εκείνες οι δυνατότητες που παρέχονται στους χρήστες έτσι ώστε να μπορούν να υλοποιήσουν διάφορες τεχνικές Εξόρυξης Γνώσης μέσα από το Excel. Μπορεί επίσης να παρουσιαστεί ολόκληρη η λειτουργικότητα που παρέχεται μέσα από το Excel και το Visio έτσι ώστε ο χρήστης να δουλέψει όχι μόνο με τεχνικές Εξόρυξης Γνώσης αλλά με μια ολοκληρωμένη εφαρμογή Business Intelligence χρησιμοποιώντας τις δυνατότητες του Excel και των αντίστοιχων add-ins. Να παρουσιαστεί, με λίγα λόγια η δυνατότητα που παρέχεται στους χρήστες να εκτελέσουν OLAP λειτουργίες (δημιουργία διαστάσεων, κύβων, χρήση pivot tables και pivot charts) και να χρησιμοποιήσουν τεχνικές Εξόρυξης Γνώσης για την παρουσίαση των αποτελεσμάτων του οργανισμού από πολλές οπτικές γωνίες, μέσα από το γνώριμο – σε πολλούς χρήστες – περιβάλλον του Excel.

Μια άλλη επέκταση της παρούσας εργασίας, που δεν έχει αναφερθεί σε βάθος, είναι η παρουσίαση των δυνατοτήτων που παρέχονται από τον SQL Server για την δημιουργία ολοκληρωμένων αναφορών μέσω των Reporting Services. Το περιβάλλον σχεδίασης αναφορών μέσα από το BIDS και των reporting services του SQL Server είναι ένα ολοκληρωμένο σύστημα σχεδιασμού και εγκατάστασης πλούσιων και λεπτομερών αναφορών, για την παρουσίαση των αποτελεσμάτων της ανάλυσης σε ένα έργο BI και όχι μόνο. Η μορφή των αναφορών σχεδιάζεται πλήρως από τον χρήστη και η χρήση πινάκων και διαγραμμάτων στις αναφορές καθιστούν τα reporting services ένα εξαιρετικά χρήσιμο εργαλείο παρουσίασης αποτελεσμάτων. Δεν είναι τυχαίο το γεγονός ότι το μεγαλύτερο μέρος των έργων BI που σχεδιάζονται μέσα από τον SQL Server με τα Analysis Services χρησιμοποιεί και τα Reporting Services για την δημιουργία και χρήση αναφορών.

Τέλος, μπορεί να παρουσιαστεί η δυνατότητα χρήσης των αποτελεσμάτων Εξόρυξης Γνώσης και πιο συγκεκριμένα οι εντολές DMX που έχουν σχεδιαστεί για την δημιουργία μοντέλων Εξόρυξης Γνώσης, εκπαίδευσης τους και χρήσης τους για την δημιουργία προβλέψεων σε μια ολοκληρωμένη εφαρμογή, που θα έχει κατασκευαστεί για τον σκοπό αυτό. Πιο συγκεκριμένα, μπορεί να σχεδιαστεί μια εφαρμογή πάνω στο .NET Framework της Microsoft να παρουσιαστούν οι δυνατότητες του ADOMD.NET προκειμένου να επικοινωνήσει η εφαρμογή .NET με τα Analysis Services του SQL Server. Οι βιβλιοθήκες εντολών του ADOMD.NET δίνουν την δυνατότητα στους προγραμματιστές των εφαρμογών να χρησιμοποιήσουν τα μοντέλα Εξόρυξης Γνώσης, που έχουν δημιουργηθεί στην εφαρμογή, που θα αναπτυχθεί για το σκοπό αυτό.

Βιβλιογραφία

Βιβλιογραφικές Αναφορές

[1] "Data Mining: Concepts and Techniques", Jiawei Han, University of Illinois at Urbana-Champaign, Morgan Kaufmann Publishers

[2] "Data Mining: Concepts, Models, Methods, and Algorithms", Mehmed Kantardzic, John Wiley & Sons, 2003

[3] "Foundations of SQL Server 2005 Business Intelligence", Lynn Langit, apress

[4] "Delivering Business Intelligence with Microsoft SQL Server 2008", Brian Larson, McGraw-Hill

[5] "Practical Business Intelligence with SQL Server 2005", John C. Hancock, Roger Toren, Addison Wesley Professional

[6] "Data Mining with Microsoft SQL Server 2008", Jamie MacLennan, ZhaoHui Tang, Bogdan Crivat, Wiley Publishing, Inc. 2009

[7] "Applied Microsoft Analysis services 2005 and Microsoft Business Intelligence Platform, a guide to the leading OLAP Platform", Teo Lachev, Prologika Press 2005

[8] "Smart Business Intelligence Solutions with SQL Server 2008", Lynn Langit, (Kevin S. Goff, Davide Mauri, Sahil Malik, John Welch), Microsoft Press, 2009

[9] "Data Mining, Εισαγωγικά και Προηγμένα Θέματα Εξόρυξης Γνώσης από Δεδομένα", Margaret H. Dunham, Επιμέλεια Ελληνικής Έκδοσης Β. Βερύκιος & Γ. Θεοδωρίδης, Εκδόσεις Νέων Τεχνολογιών

[10] "Αποθήκες Δεδομένων και Εξόρυξη Γνώσης", Γ. Θεοδωρίδης, Ν. Πελέκης, Σημειώσεις Πανεπιστημίου Πειραιώς

[11] "Data Mining Techniques in CRM, Inside customer segmentation", K. Tsiptis, A. Chorianopoulos, Wiley, 2009

[12] "Building Data Mining Applications for CRM", A. Berson, S. Smith, K. Thearling, Mc Graw Hill, 2000

[13] "Overview of Microsoft SQL Server 2008 Business Intelligence", .NET Training Notes, N. Jacobs, J. Postelmans (www.u2u.net), 2010

Αναφορές στο Διαδίκτυο

[1] http://www.crisp-dm.org (CRoss Industry Standard Process for Data Mining)

[2] http://technet.microsoft.com/en-gb/library/ms175595.aspx (Data Mining Algorithms (Analysis Services - Data Mining))

[3] http://social.msdn.microsoft.com/forums/en-US/sqldatamining/threads/ (Data Mining Forum)

[4] http://www.sqlserverdatamining.com (SQL Server Data Mining)

[5] http://www.sqldataminingbook.com/dmbook/ (Data Mining with Microsoft SQL Server 2008 (blog))

[6] http://msdn.microsoft.com/en-us/library/ms132058.aspx (Data Mining Extensions (DMX) Reference)

[7] http://jamiemaclennan.blogspot.com/

[8] http://sqlblog.com/

[9] http://www.bogdancrivat.net/dm/

[10] http://www.businessintelligence.com/

[11] http://www.codeplex.com/