ΣΥΝΘΕΤΙΚΗ ΕΡΓΑΣΙΑ

ΜΕΡΟΣ Β: ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

Στόχος της εργασίας είναι η εφαρμογή και η αξιολόγηση μεθόδων εξόρυξης δεδομένων με τον SQL Server Business Intelligence στη βάση δεδομένων της εταιρίας **FoodMart.** Η FoodMart είναι μια μεγάλη αλυσίδα παντοπωλείων με πωλήσεις στις Ηνωμένες Πολιτείες, το Μεξικό, και τον Καναδά. Πιο συγκεκριμένα στόχος είναι η ανάλυση όλων των πωλήσεων των προϊόντων της εταιρίας και της αγοραστικής συμπεριφοράς των πελατών της κατά τη διάρκεια του **1997**. (στη βάση δεδομένων η διάσταση του χρόνου αποθηκεύεται στο επίπεδο ημέρας μιας συναλλαγής με την χρήση του πεδίου time_id). Χρησιμοποιώντας τα στοιχεία που αποθηκεύονται στην βάση δεδομένων της επιχείρησης, πρέπει να κτιστεί μια πολυδιάστατη δομή δεδομένων (ένας κύβος) για να υπάρχουν γρήγοροι χρόνοι απόκρισης της βάσης, όταν θέτουν ερωτήματα σε αυτήν οι εμπορικοί αναλυτές της εταιρείας. Επίσης, για την ανάλυση των πωλήσεων και τη διεξαγωγή συμπερασμάτων θα πρέπει να εφαρμοστούν δέντρα απόφασης, ομαδοποιήσεις και κανόνες συσχέτισης σύμφωνα με τις παρακάτω οδηγίες. Πριν προχωρήσετε στις παρακάτω ενέργειες βεβαιωθείτε ότι έχετε κάνει την εισαγωγή της βάσης δεδομένων FoodMart στον SQL Server και την κατάλληλη προετοιμασία που περιγράφεται στο Κεφάλαιο 6 του Εργαστηριακού Οδηγού (ενότητες 6.2, 6.5).

1. (α) Δημιουργήστε έναν κύβο δεδομένων για την βάση FoodMart με τα εξής στοιχεία:

Πίνακας γεγονότων: Sales_fact_1997
Πίνακες διαστάσεων: Product, Time By Day, Store, Customer
Μετρήσεις στον πίνακα γεγονότων: store_sales, store_cost και unit_sales.

Να καθοριστούν στον κύβο οι ακόλουθες ιεραρχίες των διαστάσεων: Στη διάσταση Time By Day: The Year → Quarter → The Month → Week Of Year → The Day. Στη διάσταση Store: Store Country → Store State → Store City → Store Name. Στη διάσταση Product: Product Family → Product Category → Product Subcategory → Product Name.

(Παραδοτέο είναι το σχήμα του κύβου [απεικόνιση χιονονιφάδας])

(β) Στη συνέχεια να εκτελέσετε τα παρακάτω ερωτήματα στον κύβο, να εμφανίσετε τα αποτελέσματά τους στον Pivot Table του Browser και να δημιουργήσετε τα αντίστοιχα Pivot Charts στο Excel:

I. Ποιες ήταν οι πωλήσεις της εταιρείας για το 3^{ο} τετράμηνο του 1997 ανά πολιτεία των USA;
II. Ποια 6 προϊόντα σημείωσαν τις μεγαλύτερες πωλήσεις το μήνα Δεκέμβριο του 1997;
III. Σε ποιες πολιτείες, περιοχές και σε ποια συγκεκριμένα μαγαζιά έγιναν οι περισσότερες πωλήσεις για όλο το 1997;

(Παραδοτέα είναι τα Pivot Tables και τα διαγράμματα του Excel)

2. Προκειμένου να επαναπροσδιοριστεί το πρόγραμμα Κάρτας Μέλους, το εμπορικό τμήμα θέλει να αναλύσει τις συναλλαγές πωλήσεων και να ανακαλύψει τα πρότυπα μεταξύ των δημογραφικών πληροφοριών των πελατών (φύλλο, συζυγική κατάσταση, ετήσιο εισόδημα, αριθμός παιδιών στο σπίτι, αριθμός αυτοκινήτων, εκπαίδευση) και της κάρτας που αυτοί χρησιμοποίησαν. Με αυτήν την γνώση, οι κάρτες θα επαναπροσδιοριστούν βασισμένες στα χαρακτηριστικά των πελατών που τις χρησιμοποίησαν. Να δημιουργηθεί ένα μοντέλο για να εκπαιδευτούν τα στοιχεία των πωλήσεων χρησιμοποιώντας δέντρα απόφασης ώστε να βρεθούν τα κύρια στοιχεία που συντελούν στην επιλογή μιας κάρτας μέλους (π.χ. χρυσή κάρτα μέλους, ασημένια κάρτα μέλους κλπ.). Η κύρια διάσταση (πίνακας case) που θα χρησιμοποιηθεί θα είναι ο πίνακας των Πελατών (customer), ενώ η κύρια ιδιότητα που θα διερευνηθεί θα είναι αυτή της κάρτας μέλους (member_card). Κατόπιν να επιλεχθεί ένας κατάλογος δημογραφικών χαρακτηριστικών (ως input) από τα οποία ο αλγόριθμος θα καθορίσει τα πρότυπα. Ενδεικτικά αναφέρονται τα παρακάτω χαρακτηριστικά: φύλλο, συζυγική κατάσταση, ετήσιο εισόδημα, αριθμός παιδιών στο σπίτι, αριθμός αυτοκινήτων και εκπαίδευση. Πιο συγκεκριμένα:

(α) Δημιουργήστε δύο διαφορετικά δέντρα απόφασης για πρόβλεψη της ιδιότητας Member Card από τις ιδιότητες: Gender, Marital Status, Num Cars Owned, Num Children At Home, Yearly Income, Education. Για το πρώτο δέντρο αφήστε τις προκαθορισμένες ρυθμίσεις για την εκτέλεση του αλγορίθμου, ενώ για το δεύτερο προσπαθήστε να το δημιουργήσετε ώστε να είναι όσο το δυνατόν καλύτερο (να προσδιορίζει δηλαδή ακριβέστερα τους κατόχους των καρτών) αλλά και πιο αποδοτικό, αλλάζοντας τις παραμέτρους και ακολουθώντας στρατηγικές που αναπτύσσονται στον εργαστηριακό οδηγό.

(Παραδοτέα είναι οι απεικονίσεις των δύο δένδρων καθώς και οι ρυθμίσεις που χρησιμοποιήσατε)

(β) Αξιολογήστε τα δέντρα απόφασης που δημιουργήσατε και προτείνετε εκείνο το μοντέλο που πετυχαίνει ακριβέστερο προσδιορισμό των κατόχων ασημένιας κάρτας.

(Παραδοτέα είναι για το κάθε δέντρο: το Score και το Prediction Percentage του, ο Classification Matrix, και το Lift Chart στο Silver στο οποίο θα απεικονίζονται οι 3 καμπύλες: ideal, random, decision tree model)

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

(α) Δημιουργείστε με τον αλγόριθμο k-Means δύο διαφορετικά μοντέλα ομαδοποίησης με τρεις ομάδες πελατών βάσει των δημογραφικών τους χαρακτηριστικών. Ενδεικτικά αναφέρονται τα παρακάτω δημογραφικά χαρακτηριστικά που μπορείτε να επιλέξτε και να χρησιμοποιήσετε ως input: φύλλο (Gender), συζυγική κατάσταση (Marital Status), ετήσιο εισόδημα (Yearly Income), αριθμός παιδιών στο σπίτι (Num Children At Home), αριθμός αυτοκινήτων (Cars Owned) και εκπαίδευση (Education). Προσοχή στο ότι η επιλογή όλων των χαρακτηριστικών μαζί δεν συνεπάγεται και την δημιουργία του καλύτερου μοντέλου.

(Παραδοτέα είναι οι επιλογές χαρακτηριστικών και οι ρυθμίσεις που κάνατε)

(β) Για τα δύο μοντέλα που δημιουργήσατε δώστε τα αποτελέσματα σύγκρισης μεταξύ των ομάδων τους: Cluster1 και Cluster2, Cluster2 και Cluster3, και Cluster1 και Cluster3.

(Παραδοτέα είναι οι 3 πίνακες Cluster Discrimination για κάθε μοντέλο)

(γ) Αξιολογήστε τα δύο μοντέλα που δημιουργήσατε και προσδιορίστε ποιο τελικά προτείνετε καθώς επίσης και τους λόγους της απόφασης σας.

(Παραδοτέα είναι για το κάθε μοντέλο: Score, Prediction Percentage και Lift Chart)

4. Θέλουμε να βρούμε συσχετίσεις μεταξύ των ιδιοτήτων των πελατών. Από την διάσταση customer του προηγούμενου κύβου (πίνακας case), δημιουργήστε κανόνες συσχέτισης (Association Rules) από τις εξής ιδιότητες (οι οποίες θα είναι και input και predictable): City, Education, Gender, Houseowner, Marital Status. Για support, confidence και λοιπές παραμέτρους αφήστε τις default τιμές.

(α) Αναφέρετε τους 5 κανόνες με την μεγαλύτερη τιμή Probability

(Παραδοτέα είναι οι 5 κανόνες που ανακαλύψατε με τις τιμές Probability και Importance τους)

(β) Αποτυπώστε το Dependency Network με τις ισχυρές και τις πιο ισχυρές συσχετίσεις

(Παραδοτέα είναι τα δύο αντίστοιχα διαγράμματα)

(γ) Σχολιάστε τα αποτελέσματα των συσχετίσεων.

Τα παραδοτέα να καταγραφούν και να ενσωματωθούν σε μία συνολική αναφορά (αρχείο word ή pdf) την οποία πρέπει να παραδώσετε το αργότερο μέχρι τις 20 Ιουνίου 2017.