ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΠΑΤΡΑΣ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΙΜΙΑΣ

ΤΜΗΜΑ ΛΟΓΙΣΤΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΣΤΗ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗ ΑΝΑΛΥΣΗ

ΤΡΙΑΝΤΑΦΥΛΛΟΣ ΓΙΟΡΤΖΟΓΛΟΥ Α.Μ. 6602

ΕΙΣΗΓΗΤΗΣ ΚΩΤΣΙΑΝΤΗΣ ΣΩΤΗΡΗΣ

ΠΑΤΡΑ 2011

ΠΕΡΙΛΗΨΗ

Στο πρώτο κεφάλαιο της παρούσας πτυχιακής εργασίας ασχοληθήκαμε με την εξόρυξη γνώσης, τα αποτελέσματα της, τους στόχους της, το λογισμικό της εξόρυξης τη διαδικασία εξόρυξης γνώσης από μια αποθήκη δεδομένων καθώς και τις φάσεις προετοιμασίας, υλοποίησης και υπολογισμού των δεδομένων.

Στο δεύτερο κεφάλαιο παρουσιάσαμε τις γενικές αρχές εξόρυξης δεδομένων, στους στόχους και τις διαδικασίες της εξόρυξης τα δέντρα απόφασης, τις μεθόδους που στηρίζονται στους κανόνες απόφασης καθώς και την ομαδοποίηση στην εξόρυξη δεδομένων.

Στο τρίτο κεφάλαιο επικεντρώθηκε στην ανάλυση δεδομένων με το knowledgeseeker, παρουσιάστηκε η διαδικασία εξαγωγής και εισαγωγής δεδομένων, τα χαρακτηριστικά τους καθώς και ο αλγόριθμος για την επιλογή των διασπάρσεων των δέντρων ενώ αναλύονται οι ρυθμίσεις Bonferroni, η δημιουργία ταμπλό και κανόνων και ο ορισμός κόστους – μεγιστοποίηση κέρδους.

Στο τέταρτο και τελευταίο κεφάλαιο παρουσιάζεται η χρήση WEKA και μελετώνται οι περιπτώσεις CREDIT-A και GERMAN CREDIT.

Τέλος παρατίθενται συμπεράσματα σχετική βιβλιογραφία και αναφορά σε διαδικτυακές πηγές.

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΡΟΛΟΓΟΣ

Η εξόρυξη γνώση από δεδομένα είναι μια νέα και δυναμική τεχνολογία που βοηθάει τις επιχειρήσεις να επικεντρωθούν στην σημαντική πληροφορία που βρίσκεται μέσα στις αποθήκες δεδομένων τους, αναζητώντας κρυμμένα πρότυπα και ανακαλύπτοντας πληροφορίες που οι ειδικοί μπορεί να χάσουν ή να παραβλέψουν. Τα τελευταία χρόνια έχει αναπτυχθεί πλήθος αλγορίθμων της εξόρυξης δεδομένων, οι οποίοι ακολουθούν διαφορετικές μεθοδολογικές προσεγγίσεις, ενώ ταυτόχρονα παρουσιάζουν σημαντική ποικιλία εφαρμογών. Η εξόρυξη γνώσης είναι μία διαδικασία που σαν ρόλο έχει να εφαρμόζει μεθόδους ανάλυσης με μεγάλο όγκο δεδομένων. Πρόκειται για μία πολύ πρόσφατη τεχνολογία που βοηθάει τους μάνατζερ να εστιάζουν μόνο στα πιο σημαντικά δεδομένα από τις αποθήκες δεδομένων τους. Σκοπός αυτού του εργαλείου είναι η ανακάλυψη που είναι οι πιο χρήσιμες για τις επιχειρήσεις.

Λόγω της νέας τεχνολογίας που χρησιμοποιεί μπορεί και προβλέπει τάσεις και συμπεριφορές ώστε να παίρνονται κάθε φορά οι σωστές αποφάσεις. Αναγνωρίζει επίσης τις μορφές των δεδομένων και μ' αυτό τον τρόπο αποκαλύπτει την ύπαρξη ενός γεγονότος. Ταξινομεί τα δεδομένα και βελτιστοποιεί όλους τους πόρους που έχει στα χέρια της μία εταιρεία.

ΕΙΣΑΓΩΓΗ

ΟΙ ΝΕΕΣ ΤΕΧΝΟΛΟΓΙΕΣ ΣΤΟΝ ΧΩΡΟ ΤΩΝ ΒΑΣΕΩΝ ΔΕΔΟΜΕΝΩΝ ΣΤΗΝ ΥΠΗΡΕΣΙΑ ΤΩΝ ΣΥΣΤΗΜΑΤΩΝ ΥΠΟΣΤΗΡΙΞΗΣ ΑΠΟΦΑΣΕΩΝ

Στις αρχές της δεκαετίας του 1990 τρία πανίσχυρα εργαλεία εμφανίστηκαν στην περιοχή της ανάπτυξης συστημάτων υποστήριξης αποφάσεων. Το πρώτο νέο εργαλείο ήταν οι αποθήκες δεδομένων. Τα δυο επόμενα που ακολούθησαν ήταν η επεξεργασία δεδομένων σε πραγματικό χρόνο (OLAP) και η εξόρυξη γνώσης. Ο παρακάτω πίνακας αναπαριστά τα χαρακτηριστικά τεχνολογικά βήματα της κάθε εποχής.

Εξελικτικό βήμα	Επιχειρηματική Ερώτηση	Βοηθητικές Τεχνολογίες	Κατασκευαστές προϊόντων	Χαρακτηριστικά
Συλλογή Δεδομένων (1960)	«Ποια ήταν τα συνολικά μου έσοδα τα τελευταία 5 χρόνια;»	Υπολογιστές, ταινίες, δισκέτες	IBM, CDC	Αναδρομική, στατική ανάκτηση δεδομένων
Πρόσβαση σε δεδομένα (1980)	«Ποιες ήταν οι πωλήσεις μου στην Πάτρα τον τελευταίο Μάρτιο;»	Σχεδιαστικές βάσεις δεδομένων (RDBMS), γλώσσα SQL, ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Αναδρομική, δυναμική ανάκτηση δεδομένων σε επίπεδο εγγραφής
Αποθήκες Δεδομένων & Υποστήριξη Αποφάσεων (1990)	«Ποιες ήταν οι πωλήσεις μου στην Πάτρα τον τελευταίο Μάρτιο; Βάση αυτών παίρνω αποφάσεις για την Αθήνα»	Επεξεργασία σε πραγματικό χρόνο (OLAP), πολυδιάστατες βάσεις δεδομένων, αποθήκες δεδομένων	Pilot, Comshare, Arbor, Cognos, Microstrategy	Αναδρομική, δυναμική ανάκτηση δεδομένων σε πολλαπλά επίπεδα
Εξόρυξη γνώσης (Ανερχόμενος τομέας σήμερα)	«Ποιες είναι οι πιθανές πωλήσεις του επόμενου μήνα στην Αθήνα;»	Ανεπτυγμένοι αλγόριθμοι, πολυεπεξεργαστικά υπολογιστικά συστήματα, μεγάλες βάσεις δεδομένων	Pilot, Lockheed, IBM, SGI	Προφητική ανάκτηση πληροφορίας

Στον παραπάνω πίνακα φαίνεται ότι ο τομέας της εξόρυξης γνώσης είναι ανερχόμενος. Σε αυτό το αποτέλεσμα όμως δε φτάσαμε τυχαία. Οι επιχειρήσεις θεώρησαν πολύ σημαντικό να διαθέτουν απαντήσεις σε ερωτήσεις όπως αυτή που φαίνεται στον πίνακα. Αυτή ακριβώς η ανάγκη επέβαλε την ανάπτυξη συστημάτων υποστήριξης αποφάσεων βασισμένα σε εξόρυξη γνώσης.

Ας θεωρήσουμε μια εταιρία που κατασκευάζει υποδήματα και αναλύσουμε τις δικές της ανάγκες ώστε να δούμε αν αυτή χρειάζεται αυτές τις τεχνολογίες για να βρει απαντήσεις στα ερωτήματα της. Η εταιρία αυτή πουλάει τα προϊόντα της με δυο τρόπους. Είτε κατευθείαν στους πελάτες, είτε μέσω μεταπωλητών. Οι ειδικοί του τμήματος μάρκετινγκ της εταιρίας χρειάζεται να εξάγουν τις παρακάτω πληροφορίες από το «βουνό» πληροφοριών της εταιρίας:

- τις πέντε μεγαλύτερες αυξήσεις σε πωλήσεις στην κατηγορία νέων προϊόντων για τα περασμένα χρόνια,
- τις συνολικές πωλήσεις σε υποδήματα στη Νέα Υόρκη τον τελευταίο μήνα ανά προϊόν παραγωγής,
- τις πενήντα πόλεις με τον μεγαλύτερο αριθμό «καλών» πελατών, ένα εκατομμύριο πελάτες που αποτελούν τους πιο πιθανούς αγοραστές του νέου τύπου Walk-On-Air.

Για να βρεθούν οι απαντήσεις σε αυτά είναι σαφές ότι δεν αρκεί μια απλή ανάγνωση των δεδομένων που διαθέτει η εταιρία. Χρειάζεται μια διαφορετική προσέγγιση διαχείρισης των δεδομένων ώστε να προκύψουν πληροφορίες που ουσιαστικά είναι κρυμμένες.

Έτσι λοιπόν η ανάγκη θα την ωθούσε στην επέκταση ή ακόμα και αλλαγή του πιθανού υπάρχοντος συστήματος υποστήριξης αποφάσεων. Ο στόχος είναι να βρεθούν απαντήσεις σε σύνθετα ερωτήματα. Την λύση μπορεί να την παρέχει η παρακάτω διαδικασία.

- δημιουργία αποθήκης δεδομένων

- Εφαρμογή OLAP πράξεων
- Εφαρμογή αλγορίθμων εξόρυξης γνώσης πάνω στην αποθήκη δεδομένων

Αυτή τη διαδικασία ακολουθεί μια εταιρία που επιθυμεί να προσδώσει στο σύστημα υποστήριξης αποφάσεων της δυνατότητες εξόρυξης γνώσης. Φαίνεται λοιπόν ότι η πολυπόθητη εξόρυξη γνώσης δεν μπορεί να εφαρμοστεί αμέσως στα δεδομένα της εταιρίας.

Αποθήκη δεδομένων(data warehouse) : Περιλαμβάνει δεδομένα που συσσωρεύονται εκεί από τις βάσεις δεδομένων της επιχείρησης και συχνά το μέγεθος τους φτάνει τα gigabytes ή ακόμα και terabytes. Τυπικά η αποθήκη δεδομένων συντηρείται ξεχωριστά από τις βάσεις δεδομένων του οργανισμού γιατί οι απαιτήσεις των εφαρμογών ανάλυσης δεν συμπίπτουν με τις δυνατότητες των βάσεων δεδομένων. Οι αποθήκες δεδομένων εξυπηρετούν τα συστήματα υποστήριξης αποφάσεων γιατί παρέχουν ιστορικά, ομαδοποιημένα και συγκεντρωτικά δεδομένα αντί για λεπτομερείς εγγραφές.

Υπάρχουν εμφανείς διαφορές μεταξύ των κλασικών βάσεων δεδομένων και των αποθηκών δεδομένων. Στην αποθήκη δεδομένων καταλήγουν κατάλληλα επεξεργασμένα δεδομένα των επιμέρους βάσεων δεδομένων, διαφοροποιώντας την έτσι ως προς το περιεχόμενο των πληροφοριών, πολλές φορές μάλιστα αυτά τα δεδομένα αποτελούν δομημένες πληροφορίες και όχι απλά μια καταγραφή απλών στοιχείων – πράγματα που εμφανίζονται στις απλές βάσεις δεδομένων.

Επειδή όμως η κατασκευή μιας αποθήκης δεδομένων μπορεί να διαρκέσει πολλά χρόνια, μερικοί οργανισμοί αντί αυτών κτίζουν τα λεγόμενα data marts που περιλαμβάνουν πληροφορίες για κάποια συγκεκριμένα τμήματα. Έτσι μπορεί ένα data mart μπορεί να ανήκει στο

τμήμα Μάρκετινγκ, ένα άλλο στο Λογιστήριο. Όλα αυτά μαζί αποτελούν την κεντρική αποθήκη δεδομένων. Μια ακόμα σημαντική παράμετρος είναι και ο τρόπος υλοποίησης της αποθήκης δεδομένων. Αν δηλαδή θα βασίζεται στο σχεσιακό ή το πολυδιάστατο μοντέλο(ROLAP εναντίον MOLAP). Η επιλογή παίζει ρόλο στην απόδοση της αποθήκης δεδομένων όχι όμως και στις δυνατότητες που αυτή μπορεί να προσφέρει.

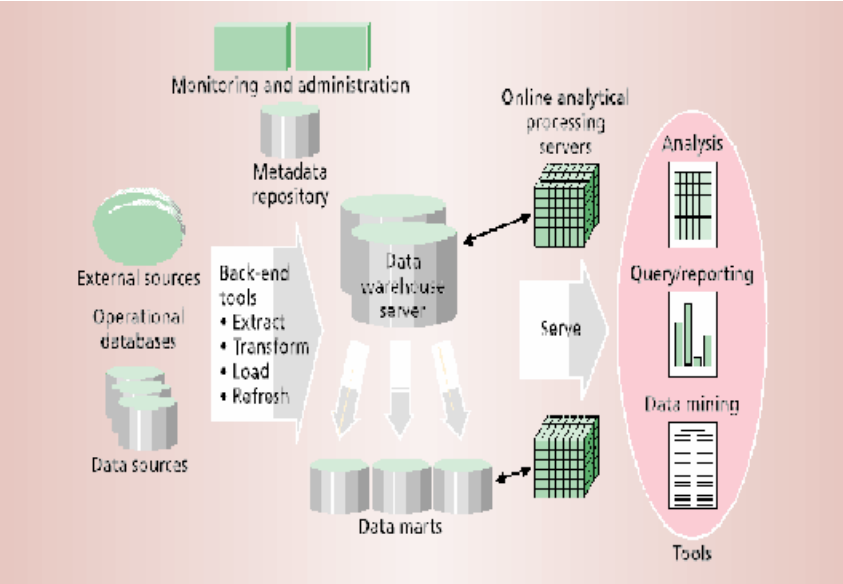
OLAP : Οι αποθήκες δεδομένων παρέχουν τη δυνατότητα για Συνεχή Αναλυτική Επεξεργασία (On-Line Analytical Processing – OLAP) των δεδομένων περιέχοντας ιστορικά και συγκεντρωτικά δεδομένα χρήσιμα για υποστήριξη αποφάσεων.

Η ανάπτυξη και η εξέλιξη της συνεχούς αναλυτικής διαδικασίας (OLAP) οφείλεται κυρίως σε δύο λόγους: Στη ραγδαία αύξηση των ποσοτήτων των δεδομένων και την ταυτόχρονη ανάγκη για ταχεία ανάλυση τους. Με τα εργαλεία OLAP παρέχονται περισσότερες δυνατότητες από αυτές που οι απλές ερωτήσεις και οι αναφορές(reports) μπορούν να δώσουν. Βοηθούν τους αναλυτές, τους μάνατζερ και τα υψηλόβαθμα στελέχη των επιχειρήσεων στη ταχεία πρόσβαση και πολυδιάστατη επεξεργασία των δεδομένων τους με σκοπό τη παρουσίαση και τη λύση των προβλημάτων της επιχείρησης στις πραγματικές τους διαστάσεις.

Βοηθούν τον χρήστη να δημιουργεί αναλύσεις μέσα από πολλαπλές ερωτήσεις του τύπου “what-if” και έτσι να μοντελοποιεί το σενάριο του. Οι εφαρμογές OLAP έχουν γίνει συνώνυμα με τη πολυδιάστατη παρουσίαση των δεδομένων. Αυτή η πολυδιάστατη παρουσίαση ενισχύεται και υποστηρίζεται από τις πολυδιάστατες βάσεις δεδομένων παρέχοντας έτσι στις OLAP εφαρμογές τη βάση για τον υπολογισμό και την ανάλυση των δεδομένων. Η ανάγκη για πολυδιάστατη ανάλυση

αναδεικνύει τις αποθήκες δεδομένων ως την κύρια πηγή άντλησης πληροφοριών.

Εξόρυξη Γνώσης: Αφού λοιπόν έχουμε δημιουργήσει την αποθήκη δεδομένων και έχουμε εκμεταλλευτεί τις δυνατότητες που προσφέρει η τεχνολογία OLAP, μπορούμε να προχωρήσουμε ακόμα ένα βήμα και να ψάξουμε με εξελιγμένους αλγόριθμους για κρυμμένη πληροφορία που βρίσκεται στην αποθήκη δεδομένων. Η εξόρυξη γνώσης από αποθήκη δεδομένων είναι ότι πιο σύγχρονο χρησιμοποιούν οι αναλυτές σήμερα.



Η αρχιτεκτονική ενός συστήματος υποστήριξης Αποφάσεων που αποτελείται από τρία μέρη: έναν data warehouse server,εργαλεία ανάλυσης και εξόρυξης γνώσης καθώς και back – end εργαλεία για την αποθήκη δεδομένων

Τώρα λοιπόν που υπάρχει μια σχετική εξοικείωση με τους μέχρι τώρα άγνωστους όρους μπορούμε να δούμε καλύτερα το πώς μπορεί ένα ΣΥΑ να στηρίζεται σε μια αποθήκη δεδομένων που μπορεί να αποτελείται από πολλά data marts και η οποία γεμίζει με στοιχεία που προέρχονται μετά από επεξεργασία των βάσεων δεδομένων της εταιρίας ή από άλλες

εξωτερικές πηγές(από το internet για παράδειγμα). Θέματα που έχουν να κάνουν με φυσική διαχείριση της αποθήκης και των μεταδεδομένων της δεν θα μας απασχολήσουν. Παρατηρούμε όμως ότι η αποθήκη δεδομένων μπορεί να εξυπηρετήσει τόσο την εφαρμογή OLAP πράξεων καθώς επίσης και την εξόρυξη γνώσης.

Η παραπάνω συνοπτική παρουσίαση ασφαλώς δεν έχει αγγίξει βαθύτερα θέματα των τεχνολογιών αυτών. Έγινε μια πρώτη εισαγωγή για να μπορεί ο αναγνώστης να παρακολουθήσει την ακόλουθη ιεραρχική(από χαμηλό σε υψηλότερο επίπεδο) υλοποίηση ενός ΣΥΑ.

ΚΕΦΑΛΑΙΟ 1
ΤΟ ΕΡΓΑΛΕΙΟ: Η ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ

1.1 ΕΙΣΑΓΩΓΗ

Είδαμε λοιπόν με πιο τρόπο μπορούν οι σχετικές με τις αποθήκες δεδομένων τεχνολογίες να παίξουν πρωταγωνιστικό ρόλο στο χώρο της υποστήριξης αποφάσεων. Υπάρχουν όμως κατηγορίες αποφάσεων που για να ληφθούν σωστά απαιτούν να έχει ο αποφασίζων στη διάθεση του απαντήσεις και πληροφορίες που δεν είναι εύκολο να προκύψουν με τις τεχνολογίες που μέχρι τώρα αναλύθηκαν. Αυτές τις κατηγορίες αποφάσεων καλείται να υποστηρίξει η εξόρυξη γνώσης.

Η εξόρυξη γνώσης, δηλαδή η διαδικασία εφαρμογής μεθόδων ανάλυσης σε μεγάλο όγκο δεδομένων, είναι μια πολύ ισχυρή νέα τεχνολογία που μπορεί να βοηθήσει τις εταιρίες να εστιάσουν μόνο στα πιο σημαντικά δεδομένα των αποθηκών δεδομένων τους. Τα εργαλεία εξόρυξης γνώσης δίνουν τη δυνατότητα στο χρήστη να προφητεύει μελλοντικές συμπεριφορές και ροπές επιτρέποντας έτσι στις επιχειρήσεις να παίρνουν κατευθυνόμενες από τη γνώση αποφάσεις. Το αποτέλεσμα είναι να μπορούν τα εργαλεία αυτά να απαντήσουν σε επιχειρηματικές ερωτήσεις που παραδοσιακά απαιτούσαν πολύ χρόνο ανάλυσης. Οι περισσότερες εταιρίες ήδη συλλέγουν και επεξεργάζονται τεράστιες ποσότητες δεδομένων. Οι τεχνικές εξόρυξης γνώσης μπορούν να αναπτυχθούν γρήγορα χωρίς να χρειάζονται αλλαγές στην υλικοτεχνική υποδομή και ως σκοπό έχουν την αξιοποίηση των πηγών πληροφοριών.

1.2 ΤΑ ΑΠΟΤΕΛΕΣΜΑΤΑ ΤΗΣ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ

Η εφαρμογή των μεθόδων εξόρυξης γνώσης αποσκοπεί στην ανακάλυψη πληροφοριών που είναι πολύ χρήσιμες για τις επιχειρήσεις. Πληροφορίες για **συσχετίσεις** όπως «όταν ένας πελάτης αγοράζει βίντεο τότε αγοράζει επίσης κάποια άλλη ηλεκτρονική συσκευή» ή για **τυποποιημένες μορφές** όπως «ο πελάτης που θα ψωνίσει περισσότερο από δύο φορές σε περίοδο εκπτώσεων είναι πιθανό να αγοράσει τουλάχιστο μία φορά κατά τη διάρκεια των Χριστουγέννων» αποτελούν πραγματικό θησαυρό για τους διοικούντες που μπορούν έτσι να αποφασίσουν για διάφορα θέματα λειτουργίας της επιχείρησης τους, όπως είναι το ωράριο, το ύψος και η διάρκεια των εκπτώσεων και η τοποθέτηση των πραγμάτων μέσα στα καταστήματα αν βέβαια μιλάμε για εμπορικού τύπου επιχειρήσεις. Τέτοιες πληροφορίες μπορούν επίσης να χρησιμοποιηθούν για τον προγραμματισμό χρήσης πρόσθετων αποθηκευτικών χώρων, για τον συνδυασμό διαφόρων πραγμάτων(βίντεο-ηλεκτρική σκούπα για παράδειγμα) στις διαφημίσεις ή για τη σχεδίαση ανάλογα την εποχή διαφορετικών στρατηγικών μάρκετινγκ.

1.3 ΟΙ ΣΤΟΧΟΙ ΤΗΣ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ

Παρακάτω αναλύονται αυτά που μπορεί να προσφέρει η εξόρυξη. Τις δυνατότητες αυτές καλείται να εκμεταλλευτεί το μάνατζμεντ της εταιρίας ή ενός οργανισμού και να προχωρήσει σε αποφάσεις που θα μετατρέψουν τη γνώση σε χειροπιαστά αποτελέσματα. Αν το πετύχει τότε οι αρχές της επιχειρηματικής νοημοσύνης, που αποτελούν και την κεντρική ιδέα των συστημάτων υποστήριξης αποφάσεων, εφαρμόζονται και είναι σίγουρο ότι τα οφέλη θα είναι μεγάλα.

- **Πρόβλεψη τάσεων και συμπεριφορών**. Δηλαδή η προσπάθεια ανακάλυψης κάποιων μελλοντικών συμπεριφορών ώστε να παρθούν οι κατάλληλες αποφάσεις με σκοπό τη μεγιστοποίηση του κέρδους ή την πρόληψη δυσμενών καταστάσεων. Τα αποτελέσματα αυτού του είδους εξόρυξης μπορεί να είναι η πρόβλεψη για το που θα φτάσουν οι πωλήσεις ενός καταστήματος για μια συγκεκριμένη χρονική περίοδο και το αν το κλείσιμο μιας γραμμής παραγωγής προϊόντων θα ενεργούσε θετικά σε ότι αφορά αυτές (τις πωλήσεις). Σε επιστημονικό επίπεδο, η μελέτη παλαιότερων σεισμικών ακολουθιών μπορεί ίσως να οδηγήσει στην πρόβλεψη, με υψηλά ποσοστά επιτυχίας, σεισμικής δραστηριότητας.
- **Αναγνώριση**. Οι τυποποιημένες μορφές ανάμεσα στα δεδομένα μπορούν να χρησιμοποιηθούν για να αποκαλύψουν την ύπαρξη ενός γεγονότος, μια δραστηριότητας. Για παράδειγμα οι εισβολείς στη προσπάθεια να σπάσουν ένα σύστημα ασφαλείας μπορούν να αναγνωριστούν από τα προγράμματα που εκτέλεσαν, τα αρχεία που προσπέλασαν και τον χρόνο που απασχόλησαν την CPU.
- **Ταξινόμηση**. Η εξόρυξη γνώσης μπορεί να διαχωρίσει έτσι τα δεδομένα ώστε να προκύψουν διαφορετικές κλάσεις ή κατηγορίες βάση κάποιων παραμέτρων. Για παράδειγμα οι πελάτες ενός super-market μπορούν να χωριστούν σε κατηγορίες, όπως φίλοι-των-εκπτώσεων, παρορμητικοί, πιστοί-κανονικοί, και σπάνιοι πελάτες. Αυτή η κατηγοριοποίηση μπορεί να χρησιμοποιηθεί στην ανάλυση των πωλήσεων ώστε να μπορεί για παράδειγμα ο μάνατζερ να λάβει αποφάσεις για να προσελκύσει σε μεγαλύτερο βαθμό κάποια από τις παραπάνω κατηγορίες.
- **Βελτιστοποίηση**. Ένας τελικός στόχος της εξόρυξης γνώσης μπορεί να είναι και η βέλτιστη χρήση περιορισμένων πόρων όπως είναι ο χρόνος, ο χώρος, το χρήμα ή τα υλικά και η μεγιστοποίηση, κάτω από ορισμένους περιορισμούς, κάποιων «ποσοτήτων» όπως είναι οι πωλήσεις

ή τα κέρδη. Έτσι σε ότι αφορά τουλάχιστον αυτό το στόχο, η εξόρυξη γνώσης έχει κοινά στοιχεία με την επιχειρησιακή έρευνα που επίσης ασχολείται με θέματα βελτιστοποίησης κάτω από περιορισμούς.

1.4ΟΙ ΚΥΡΙΟΤΕΡΕΣ ΤΕΧΝΙΚΕΣ

Τα εργαλεία εξόρυξης γνώσης συνήθως δίνουν τη δυνατότητα στον χρήστη να επιλέξει ποια τεχνική-αλγόριθμο θέλουν να εφαρμόσουν. Παρακάτω γίνεται μια σύντομη γνωριμία με αυτούς τους αλγόριθμους.

- **Τα νευρωνικά δίκτυα:** Μη γραμμικά, προφητικά και μπορούν να εκπαιδευτούν. Σε ότι αφορά τη δομή μοιάζουν στα βιολογικά νευρωνικά δίκτυα.
- **Δένδρα απόφασης:** Δενδρικές δομές που αναπαριστούν σύνολα απόφασης. Αυτές οι αποφάσεις γεννούν κανόνες για τη ταξινόμηση ενός συνόλου δεδομένων.
- **Γενετικοί αλγόριθμοι:** Τεχνικές βελτιστοποίησης που χρησιμοποιούν διαδικασίες όπως γενετικοί συνδυασμοί, μετάλλαξη.
- **Επαγωγή κανόνα:** Η εξαγωγή χρήσιμων, και με στατιστική σημασία , if-then κανόνων από τα δεδομένα.

1.5ΠΩΣ ΛΕΙΤΟΥΡΓΕΙ Η ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ

Πως ακριβώς μπορεί να μας πει η εξόρυξη γνώσης πράγματα που δεν ξέρουμε ή που θα συμβούν στο μέλλον; Η τεχνική που χρησιμοποιείται για να επιτευχθούν αυτά λέγεται μοντελοποίηση. Με άλλα λόγια η σκέψη του κτισίματος ενός μοντέλου για μια κατάσταση όπου γνωρίζουμε την απάντηση και στη συνέχεια η εφαρμογή του σε μια άλλη που δεν τη ξέρουμε. Για παράδειγμα, αν αναζητούσαμε μια βυθισμένη ισπανική γαλέρα στην ανοικτή θάλασσα το πρώτο πράγμα που ίσως σκεφτόμασταν

θα ήταν να ερευνήσουμε όλες τις περασμένες περιπτώσεις εύρεσης ισπανικών θησαυρών από άλλους. Ίσως λοιπόν να παρατηρούσαμε ότι αυτά τα πλοία στην πλειονότητα τους βρέθηκαν στις ακτές Βερμούδα και ότι υπήρχαν κάποιες βέβαιες πορείες που ακολουθούσαν οι καπετάνιοι των πλοίων αυτών εκείνη την εποχή. Αυτές οι ομοιότητες σημειώνονται και κτίζεται ένα μοντέλο που περιλαμβάνει τα χαρακτηριστικά που είναι κοινά στις τοποθεσίες αυτών των βυθισμένων θησαυρών. Με αυτό το μοντέλο αρχίζει το ψάξιμο σε περιοχές που δείχνει αυτό ότι είναι πιθανό να υπήρξε μια παρόμοια κατάσταση στο παρελθόν. Αν το μοντέλο είναι καλό ο θησαυρός θα βρεθεί.

Επομένως τη σκέψη του κτισίματος μοντέλων την είχαν οι άνθρωποι εδώ και πολύ καιρό και σίγουρα πριν την έλευση των υπολογιστών και της τεχνολογίας της εξόρυξης γνώσης. Πάντως αυτό που συμβαίνει στους υπολογιστές δεν διαφέρει πολύ από τον τρόπο με τον οποίο οι άνθρωποι κτίζουν μοντέλα. Οι υπολογιστές φορτώνονται με πληροφορίες για μια ποικιλία καταστάσεων ενώ μια απάντηση είναι γνωστή. Τότε το λογισμικό εξόρυξης γνώσης τρέχει πάνω σε αυτά τα δεδομένα και ξεχωρίζει εκείνα τα χαρακτηριστικά που πρέπει να συμπεριληφθούν στο μοντέλο. Όταν τελειώσει η διαδικασία κτισίματος μπορεί το μοντέλο να χρησιμοποιηθεί σε παρόμοιες καταστάσεις που όμως η απάντηση δεν είναι γνωστή.

Εξόρυξη Γνώσης = Μοντελοποίηση

Είμαστε γνώστες μιας κατάστασης -1-

Φτιάχνουμε πάνω σε αυτή ένα μοντέλο -2-

Το εφαρμόζουμε σε μια άλλη κατάσταση που δεν γνωρίζουμε -3-

Η φιλοσοφία της εξόρυξης γνώσης.

Για παράδειγμα ας υποθέσουμε ότι βρισκόμαστε στη θέση του διευθυντή μάρκετινγκ μιας εταιρίας τηλεπικοινωνιών και θέλουμε να αποκτήσουμε μερικούς πελάτες που κάνουν τηλεφωνήματα μεγάλων αποστάσεων. Βρισκόμαστε δηλαδή αντιμέτωποι με ένα πρόβλημα απόφασης, σε ποιους να απευθυνθούμε. Θα μπορούσαμε να ταχυδρομήσουμε με τυχαίο τρόπο κουπόνια στο γενικό πληθυσμό όπως θα μπορούσαμε να ταξιδεύουμε στις θάλασσες ψάχνοντας για βυθισμένους θησαυρούς. Πάντως σε καμιά από τις δυο περιπτώσεις δεν θα είχαμε τα επιθυμητά αποτελέσματα. Αντί αυτού θα μπορούσαμε να χρησιμοποιήσουμε την εμπειρία της εταιρίας που βρίσκεται αποθηκευμένη στις βάσεις δεδομένων και να κτίσουμε ένα μοντέλο.

Ο διευθυντής μάρκετινγκ έχει πρόσβαση σε πολλές πληροφορίες σχετικές με τους πελάτες μας: την ηλικία τους, το φύλο τους, το αν είναι καλοί πληρωτές, το πόσα τηλεφωνήματα μεγάλων αποστάσεων κάνουν. Το καλό είναι ότι υπάρχουν πληροφορίες και για τους πιθανούς πελάτες της εταιρίας: την ηλικία τους, το φύλο τους, το πόσο γρήγορα θα πληρώνουν κτλ. Το πρόβλημα είναι ότι δεν γνωρίζουμε πόσο πολύ θα κάνουν χρήση τηλεφωνημάτων σε απομακρυσμένες περιοχές. Επειδή θέλουμε αυτούς που κάνουν πολλά τέτοια τηλεφωνήματα μπορούμε να το πετύχουμε αυτό κτίζοντας ένα μοντέλο.

Ένα απλό μοντέλο που θα ταίριαζε σε μια τηλεπικοινωνιακή εταιρία είναι το παρακάτω:

98% των πελατών που έχουν λογαριασμό μεγαλύτερο

από 60.000$ το χρόνο δαπανούν περισσότερα

από 80$ το μήνα για τηλεφωνήματα σε μακρινές περιοχές

Αυτό το μοντέλο θα μπορούσε να εφαρμοστεί στα δεδομένα των πιθανών πελατών και να δοθεί απάντηση στο πρόβλημα απόφασης. Αφού γίνει αυτό θα ξέρει σε ποιους να απευθυνθεί η εταιρία.

1.6 ΛΟΓΙΣΜΙΚΟ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ

Η εξόρυξη γνώσης είναι κατά κάποιο τρόπο μια επέκταση της στατιστικής με κάποια στοιχεία τεχνητής νοημοσύνης και μηχανική μάθηση(machine learning). Όπως και η στατιστική, η εξόρυξη γνώσης δεν αποτελεί επιχειρηματική λύση. Είναι απλά μια τεχνολογία. Για παράδειγμα φανταστείτε ότι πρέπει από ένα κατάλογο εμπόρων λιανικής να αποφασιστεί σε ποιους θα σταλούν πληροφορίες για κάποιο νέο προϊόν. Η πληροφορία που αναζητείται από την διαδικασία εξόρυξης γνώσης περιλαμβάνεται σε βάσεις ιστορικών δεδομένων προηγούμενων συναλλαγών με τους πελάτες και στα χαρακτηριστικά των πελατών όπως η ηλικία, ο ταχυδρομικός τους κώδικας, το αν αποκρίθηκαν στο παρελθόν. Το λογισμικό εξόρυξης γνώσης θα χρησιμοποιήσει αυτές τις πληροφορίες από το παρελθόν για να χτίσει ένα μοντέλο συμπεριφοράς πελάτη που θα μπορεί να χρησιμοποιηθεί για να προβλέψουμε ποιοι πελάτες θα ήταν πιθανό να ανταποκριθούν στο νέο προϊόν. Ένας διευθυντής μάρκετινγκ κάνοντας χρήση αυτής της πληροφορίας μπορεί να επιλέξει μόνο τους πελάτες που είναι πιο πιθανό να ανταποκριθούν. Το λογισμικό της επιχείρησης μπορεί τότε να τροφοδοτήσει με τα αποτελέσματα της απόφασης τα κατάλληλα «σημεία επαφής» (τηλεφωνικά κέντρα, web servers, e-mails κτλ) ώστε οι κατάλληλοι πελάτες να λαμβάνουν τις κατάλληλες πληροφορίες.

1.7 Η ΔΙΑΔΙΚΑΣΙΑ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ ΑΠΟ ΜΙΑ ΑΠΟΘΗΚΗ ΔΕΔΟΜΕΝΩΝ

Σε αυτή την παράγραφο θα ασχοληθούμε με πιο τεχνικά θέματα των συστημάτων υποστήριξης αποφάσεων ώστε ο αναγνώστης να έχει συνολική εικόνα του θεμάτων με τα οποία ασχολείται η εργασία. Πιο

συγκεκριμένα θα δούμε αναλυτικά τα στάδια που μεσολαβούν μέχρι να είναι η δυνατή η ανάλυση και η ερμηνεία των αποτελεσμάτων. Η ανακάλυψη γνώσης – η διαδικασία καθορισμού και επίτευξης ενός σκοπού μέσω επαναληπτικής εξόρυξης γνώσης – τυπικά αποτελείται από τρεις φάσεις:

- Προετοιμασία των δεδομένων,
- Υλοποίηση και αποτίμηση του μοντέλου και
- Ανάπτυξη του μοντέλου

1.7.1 Η ΦΑΣΗ ΤΗΣ ΠΡΟΕΤΟΙΜΑΣΙΑΣ ΤΩΝ ΔΕΔΟΜΕΝΩΝ

Στη φάση της προετοιμασία των δεδομένων, ο αναλυτής προετοιμάζει ένα σύνολο δεδομένων που περιλαμβάνει αρκετές πληροφορίες για να κτιστεί ένα σωστό μοντέλο σε ακόλουθες φάσεις. Προσδιορίζοντας αυτές τις απαραίτητες πληροφορίες για μια εταιρία, ένα αποτελεσματικό μοντέλο θα μπορούσε να προβλέψει αν υπάρχει πιθανότητα να αγοράσει κάποιος πελάτης προϊόντα που διαφημίζονται σε ένα νέο κατάλογο. Επειδή οι προβλέψεις βασίζονται σε παράγοντες που πιθανότατα επηρεάζουν τις αγορές των πελατών, ένα μοντέλο συνόλου δεδομένων θα μπορούσε να περιλάμβανε όλους τους πελάτες που ανταποκρίθηκαν σε καταλόγους μέσω e-mails, ταχυδρομείων κτλ τα τελευταία τρία χρόνια, τις δημογραφικές πληροφορίες τους, τα δέκα πιο ακριβά προϊόντα που αγόρασε κάθε πελάτης και πληροφορίες για τους καταλόγους από τους οποίους έγιναν οι αγορές.

Η προετοιμασία των δεδομένων μπορεί να περιλαμβάνει πολύπλοκες ερωτήσεις με τεράστια αποτελέσματα-απαντήσεις. Για παράδειγμα στην υποθετική εταιρία που αναφέρθηκε και στα προηγούμενα παραδείγματα, η προετοιμασία του μοντέλου περιλαμβάνει joins μεταξύ του πίνακα των

πελατών και του πίνακα των πωλήσεων καθώς επίσης και τον προσδιορισμό των δέκα κορυφαίων προϊόντων για κάθε πελάτη. Όλα τα θέματα που έχουν να κάνουν με την αποτελεσματική επεξεργασία ερωτήσεων υποστήριξης αποφάσεων σχετίζονται με το περιβάλλον της εξόρυξης γνώσης.

Η εξόρυξη γνώσης τυπικά περιλαμβάνει επαναληπτικό κτίσιμο μοντέλων πάνω σε ένα ήδη προετοιμασμένο σύνολο δεδομένων και στη συνέχεια την ανάπτυξη ενός ή περισσότερων μοντέλων. Επειδή το κτίσιμο των μοντέλων σε μεγάλα σύνολα δεδομένων μπορεί να είναι δαπανηρό, οι αναλυτές συχνά εργάζονται επαναληπτικά με δείγματα συνόλων δεδομένων.

1.7.2 Η ΦΑΣΗ ΤΗΣ ΥΛΟΠΟΙΗΣΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΜΟΣ ΤΟΥ ΜΟΝΤΕΛΟΥ

Μόνο όταν έχει αποφασιστεί ποιο μοντέλο θα αναπτυχθεί, κτίζει ο αναλυτής το μοντέλο πάνω στο συνολικά προετοιμασμένο σύνολο δεδομένων. Ο σκοπός της φάσης της υλοποίησης είναι ο εντοπισμός των τυποποιημένων μορφών που καθορίζουν ένα χαρακτηριστικό-στόχο(target attribute). Ένα παράδειγμα τέτοιου χαρακτηριστικού-στόχου σε ένα σύνολο δεδομένων θα μπορούσε να ήταν το αν αγόρασε ένας πελάτης τουλάχιστον ένα προϊόν από ένα περασμένο κατάλογο.

Μερικές κλάσεις μοντέλων εξόρυξης γνώσης βοηθούν την πρόβλεψη τόσο ρητά καθορισμένων όσο και κρυφών χαρακτηριστικών. .υο σημαντικά θέματα που επηρεάζουν την επιλογή του μοντέλου είναι η ακρίβεια του και η αποτελεσματικότητα του αλγορίθμου κατασκευής του μοντέλου πάνω σε μεγάλα σύνολα δεδομένων. Από στατιστικής πλευράς η ακρίβεια των περισσότερων μοντέλων βελτιώνεται με το πλήθος των δεδομένων που χρησιμοποιούνται, οπότε οι αλγόριθμοι που επηρεάζουν

τα μοντέλα εξόρυξης πρέπει να κάνουν αποτελεσματική και κλιμακωτή επεξεργασία μεγάλων συνόλων δεδομένων σε ένα λογικό χρονικό διάστημα.

ΚΕΦΑΛΑΙΟ 2
ΓΕΝΙΚΕΣ ΑΡΧΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

2.1 ΟΡΙΣΜΟΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Η εξόρυξη δεδομένων είναι μια διαδικασία εξαγωγής κρυμμένης πληροφορίας από μεγάλες βάσεις δεδομένων. Ένας πιο αυστηρός και τυπικός ορισμός της εξόρυξης δεδομένων αναφέρει :

«*Εξόρυξη δεδομένων είναι η διαδικασία εξαγωγής υπονοούμενης και εν πολλοίς άγνωστης, αλλά ενδεχομένως χρήσιμης γνώσης, υπό την μορφή συσχετίσεων, προτύπων και τάσεων, μέσω της εξέτασης, ανάλυσης και επεξεργασίας βάσεων δεδομένων, συνδυάζοντας και χρησιμοποιώντας τεχνικές από την μηχανική μάθηση, την αναγνώριση προτύπων, την στατιστική, τις βάσεις δεδομένων και την οπτικοποίηση*».

Παρά το γεγονός ότι υπάρχει μια γενικότερη συμφωνία ότι στόχος της εξόρυξης δεδομένων είναι η ανακάλυψη νέας και χρήσιμης πληροφορίας σε βάσεις δεδομένων, τα μέσα για την επίτευξη του στόχου αυτού, ποικίλουν σε πολύ υψηλό βαθμό. Η εξόρυξη δεδομένων περιλαμβάνει ένα ευρύ πεδίο υπολογιστικών μεθόδων που μεταξύ άλλων περιλαμβάνουν, την *στατιστική ανάλυση* (*statistical analysis*), τα *δέντρα αποφάσεων* (*decision trees*), τα *νευρωνικά δίκτυα* (*neural networks*), την

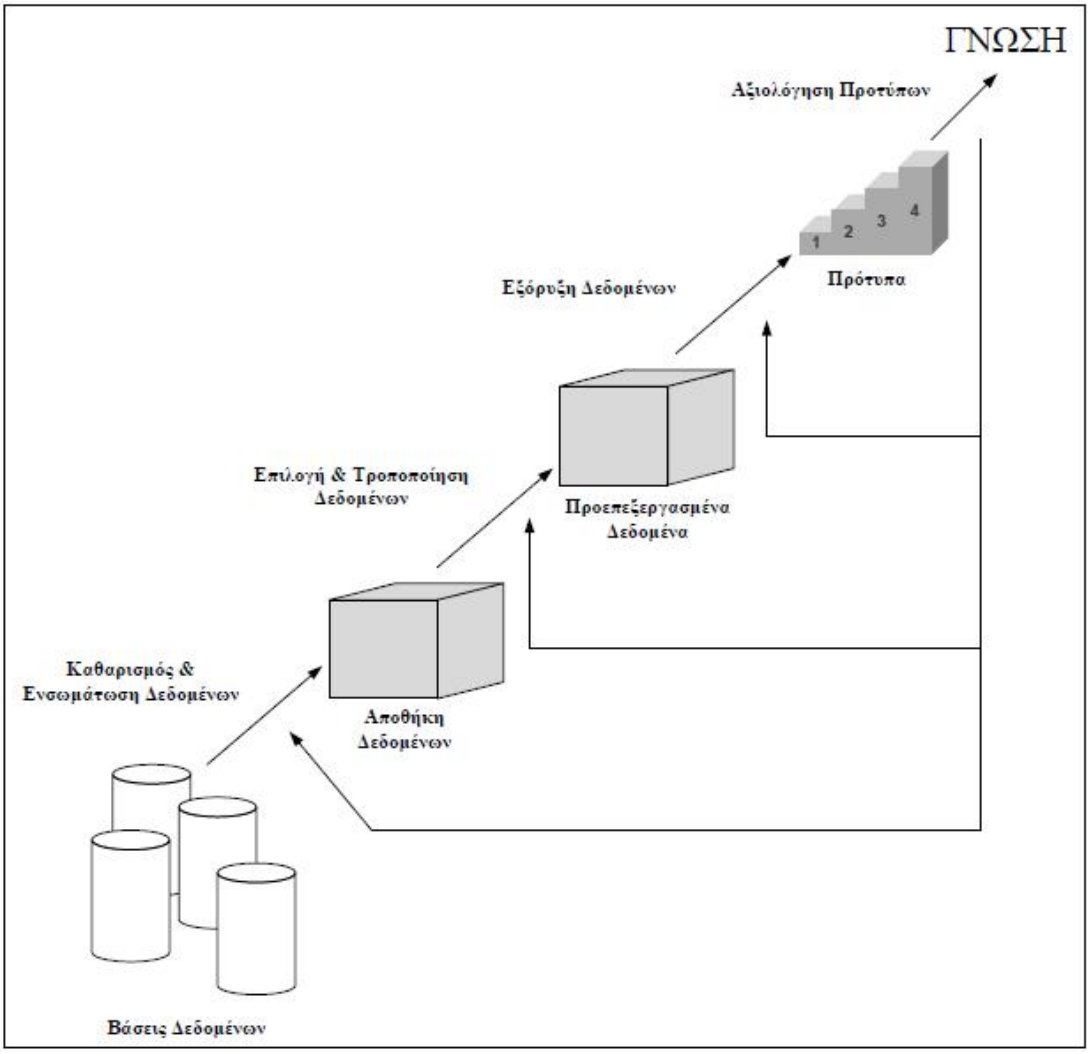
εξαγωγή κανόνων (*rule induction*) και την *γραφική οπτικοποίηση* (*graphic*

visualization). Τέτοιες μέθοδοι χρησιμοποιούνται για την εύρεση συσχετίσεων, προτύπων και δομών, σε μεγάλες και διαρκώς αυξανόμενες βάσεις δεδομένων. Ειδικά η εύρεση προτύπων είναι ένα ιδιαίτερα σημαντικό εξαγόμενο της εξόρυξης δεδομένων, αυτά δε περιγράφονται,

μέσω σχέσεων μεταξύ των *χαρακτηρικών* (*attributes*) των βάσεων δεδομένων. Αξίζει επίσης να σημειώσουμε ότι η εξόρυξη δεδομένων δεν εξειδικεύεται σε ένα μόνο τύπο δεδομένων. Ωστόσο, οι αλγόριθμοι της, τόσο σε θεωρητικό όσο και σε πρακτικό επίπεδο, μπορεί να διαφέρουν εφαρμοζόμενοι σε διαφορετικά είδη δεδομένων.

2.2 ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΑΝΕΥΡΕΣΗ ΓΝΩΣΗΣ

Υπάρχει μια γενικότερη σύγχυση ανάμεσα στους όρους «*Εξόρυξη Δεδομένων*» και «*Ανεύρεση γνώσης στις βάσεις δεδομένων*» (*Knowledge discovery in databases*). Σε πολλές περιπτώσεις οι δύο όροι ταυτίζονται, ενώ στην πραγματικότητα η εξόρυξη δεδομένων αποτελεί τμήμα της ανεύρεσης γνώσης, συγκροτώντας τον πυρήνα αυτής. Προκειμένου λοιπόν να κατανοηθεί καλύτερα η εξόρυξη δεδομένων, θα γίνει μια πολύ σύντομη αναφορά στην διαδικασία της ανεύρεσης γνώσης. Η ανεύρεση γνώσης είναι μια επαναληπτική διαδικασία που αποτελείται από μια σειρά βημάτων, τα οποία οδηγούν από την συλλογή των δεδομένων, στην ανακάλυψη και εξαγωγή χρήσιμης γνώσης από αυτά.



Βήματα της διαδικασίας ανεύρεσης γνώσης στις βάσεις δεδομένων

Υπάρχουν τα ακόλουθα βήματα:

Καθαρισμός δεδομένων (Data cleaning): Στο βήμα αυτό, αφαιρούνται από τη βάση δεδομένων, αυτά που παράγουν θόρυβο, καθώς και τα άσχετα δεδομένα.

Ενσωμάτωση δεδομένων (Data integration): Στο βήμα αυτό, δεδομένα από πολλές διαφορετικές πηγές (συχνά ανομοιογενή), ενσωματώνονται σε μια βάση δεδομένων.

Επιλογή δεδομένων (Data selection): Από το σύνολο των διαθέσιμων δεδομένων, επιλέγονται εκείνα που είναι σχετικά με την ανάλυση που θα ακολουθήσει.

Τροποποίηση δεδομένων (Data transformation): Τα επιλεγμένα δεδομένα τροποποιούνται ώστε η μορφή τους να είναι κατάλληλη για την διαδικασία της εξόρυξης.

Εξόρυξη δεδομένων (Data Mining): Είναι το σημαντικότερο από τα βήματα της διαδικασίας, στο οποίο ποικίλες εξελιγμένες τεχνικές χρησιμοποιούνται για την εξαγωγή δυνητικά χρήσιμων προτύπων.

Αξιολόγηση προτύπων (Pattern evaluation): Στο βήμα αυτό, αναγνωρίζονται χρήσιμα πρότυπα που αναπαριστούν γνώση, βάσει συγκεκριμένων *μέτρων αξιολόγησης* (*evaluation measures*).

Αναπαράσταση γνώσης (Knowledge representation): Στο τελικό αυτό βήμα, η γνώση που ανακαλύφθηκε, παρουσιάζεται στον χρήστη, βοηθώντας τον να καταλάβει και να ερμηνεύσει τα αποτελέσματα της εξόρυξης δεδομένων.

Πολλές φορές κάποια από τα παραπάνω βήματα συνδυάζονται. Για παράδειγμα, τα βήματα του καθαρισμού και της ενσωμάτωσης των δεδομένων, μπορούν να υλοποιηθούν μαζί, με στόχο την δημιουργία μια αποθήκης δεδομένων. Με την ίδια λογική μπορούν να συνδυαστούν τα βήματα της επιλογής και της τροποποίησης των δεδομένων. Η παραπάνω περιγραφή καθιστά σαφές ότι η εξόρυξη δεδομένων είναι διαδικασία-κλειδί για την ανεύρεση γνώσης. Παρόλα αυτά, δεν καταλαμβάνει παρά

μόνο ένα μικρό μέρος της όλης προσπάθειας, με δεδομένη την πολυπλοκότητα της τελευταίας. Αξίζει να σημειωθεί ότι ο χρήστης, εκμεταλλευόμενος την επαναληπτική μορφή της διαδικασίας ανεύρεσης γνώσης, έχει την δυνατότητα να τροποποιήσει τα μέτρα αξιολόγησης, να τελειοποιήσει την διαδικασία της εξόρυξης, να επιλέξει νέα δεδομένα, να τροποποιήσει περαιτέρω τα ήδη υπάρχοντα δεδομένα ή να ενσωματώσει

στη βάση νέα από καινούργιες πηγές, με τελικό στόχο την εξαγωγή διαφορετικών, πιο κατάλληλων αποτελεσμάτων.

2.3 ΣΤΟΧΟΙ ΚΑΙ ΔΙΑΔΙΚΑΣΙΕΣ ΤΗΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Η εξόρυξη δεδομένων έχει σαν βασικούς της στόχους την εφαρμογή τεχνικών ***πρόβλεψης*** (*prediction*) και ***περιγραφής*** (*description*) σε μεγάλες βάσεις δεδομένων). Ειδικότερα:

_ Η ***πρόβλεψη*** περιλαμβάνει την χρήση μερικών μεταβλητών ή χαρακτηριστικών μιας βάσης δεδομένων για την πρόβλεψη άγνωστων ή μελλοντικών τιμών χρήσιμων μεταβλητών. Με άλλα λόγια, οι *διαδικασίες πρόβλεψης της εξόρυξης δεδομένων* (***predictive data mining tasks***), προσπαθούν να κάνουν εκτιμήσεις βγάζοντας συμπεράσματα από τα διαθέσιμα δεδομένα.

_ Η ***περιγραφή*** επικεντρώνεται στην ανακάλυψη προτύπων και αναπαριστά τα δεδομένα μιας πολύπλοκης βάσης δεδομένων με όσο το δυνατό πιο κατανοητό και αξιοποιήσιμο τρόπο. Με άλλα λόγια, οι *περιγραφικές διαδικασίες της εξόρυξης δεδομένων* (***descriptive data mining tasks***) περιγράφουν τις γενικές ιδιότητες των υπαρχόντων διαθέσιμων δεδομένων.

Οι βασικότερες από τις *διαδικασίες* (***tasks***) της εξόρυξης δεδομένων, μέσω των οποίων επιτυγχάνονται οι παραπάνω στόχοι της *πρόβλεψης* και της *περιγραφής*, είναι οι εξής:

- **Ταξινόμηση**

Η διαδικασία της *ταξινόμησης* (***classification***) περιλαμβάνει την οργάνωση ενός συνόλου από *αντικείμενα* (***objects***) που περιγράφονται από ένα σύνολο *χαρακτηριστικών* (***attributes***), σε μια σειρά από προκαθορισμένες *κλάσεις* (***classes***), χρησιμοποιώντας *μεθόδους μάθησης με επίβλεψη* (***supervised learning methods***). Οι τεχνικές της ταξινόμησης

χρησιμοποιούν κατά κανόνα ένα *σύνολο εκπαίδευσης* (***training set***), όπου όλα τα αντικείμενα είναι ήδη συνδεδεμένα με γνωστές κλάσεις. Ο αλγόριθμος ταξινόμησης «μαθαίνει» από αυτό το σύνολο, χρησιμοποιώντας την μάθηση αυτή για την κατασκευή ενός μοντέλου. Το μοντέλο αυτό στην συνέχεια ταξινομεί νέα αντικείμενα στις κατάλληλες κλάσεις.

- **Ομαδοποίηση**

Η *ομαδοποίηση* (***clustering***) αφορά τον *διαχωρισμό* (***partition***) των αντικειμένων μιας βάσης δεδομένων σε μη συνδεδεμένες μεταξύ τους και ομοιογενείς ομάδες, κατά τέτοιο τρόπο ώστε αντικείμενα του συνόλου που ανήκουν σε μια ομάδα, να είναι πιο όμοια μεταξύ τους, παρά με τα αντικείμενα που ανήκουν σε διαφορετικές ομάδες. Ένα ιδιαίτερο χαρακτηριστικό της ομαδοποίησης, σε αντίθεση με την ταξινόμηση, είναι ότι η δομή και το πλήθος των ομάδων είναι καταρχάς άγνωστα, καθορίζονται δε από τον εκάστοτε αλγόριθμο ομαδοποίησης. Αυτοί οι αλγόριθμοι ομαδοποίησης βασίζονται στο σύνολο τους στην αρχή της μεγιστοποίησης *της ομοιότητας ανάμεσα στα αντικείμενα την ίδιας ομάδας* (***intra-class similarity***) και την ταυτόχρονη αρχή της ελαχιστοποίησης *της ομοιότητας μεταξύ των αντικειμένων διαφορετικών ομάδων* (***inter-class similarity***). Αξίζει να σημειωθεί ότι η ερμηνεία των ομάδων που προκύπτουν από την ανωτέρω διαδικασία καθορίζεται από τον εκάστοτε χρήστη.

- **Ανάλυση Συσχέτισης**

Η *ανάλυση συσχέτισης* (***association analysis***) έχει σαν βασικό της στόχο την ανακάλυψη κρυμμένων συσχετίσεων μεταξύ των χαρακτηριστικών μιας βάσης δεδομένων. Με άλλα λόγια, η παραπάνω ανάλυση ψάχνει να βρει κανόνες για την ποσοτικοποίηση των σχέσεων

μεταξύ δύο ή περισσοτέρων χαρακτηριστικών μιας βάσης δεδομένων. Οι κανόνες αυτοί ονομάζονται *κανόνες συσχέτισης* (***association rules***), και έχουν την μορφή «If *A* then *B*». Οι κανόνες συσχέτισης χαρακτηρίζονται από το *κατώφλι στήριξης* (***support threshold***), που αναγνωρίζει τα στοιχεία (π.χ. χαρακτηριστικά) των βάσεων δεδομένων που εμφανίζονται συχνά σε αυτά, καθώς και το *κατώφλι εμπιστοσύνης* (***confidence threshold***), που είναι η *υπό συνθήκη πιθανότητα* (***conditional probability***) ένα στοιχείο να εμφανίζεται σε μια διαδικασία όταν ένα άλλο στοιχείο εμφανίζεται επίσης. Αξίζει να σημειωθεί ότι η ανάλυση συσχέτισης είναι γνωστή στον επιχειρηματικό κόσμο σαν *ανάλυση συνάφειας* (***affinity analysis***) με πολλές εφαρμογές.

- **Παλινδρόμηση**

Η *παλινδρόμηση* (***regression***) είναι η παλαιότερη και η πλέον γνωστή στατιστική τεχνική που υλοποιείται εντός των πλαισίων της εξόρυξης δεδομένων. Συγκεκριμένα η παλινδρόμηση, χρησιμοποιώντας μια βάση αριθμητικών δεδομένων, αναπτύσσει μια μαθηματική σχέση που ταιριάζει στα δεδομένα αυτά. Στην συνέχεια, η μαθηματική αυτή σχέση χρησιμοποιείται για την πρόβλεψη μελλοντικής συμπεριφοράς, εφαρμόζοντας σε αυτήν νέα αριθμητικά δεδομένα. Ο βασικός περιορισμός της συγκεκριμένης τεχνικής είναι ότι εφαρμόζεται καλά μόνο σε συνεχή ποσοτικά δεδομένα (όπως π.χ. βάρος, ταχύτητα ή ηλικία). Αντίθετα, η παλινδρόμηση δεν λειτουργεί καλά με κατηγορικά δεδομένα.

2.4 Η ΤΑΞΙΝΟΜΗΣΗ ΣΤΗΝ ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

Η ταξινόμηση είναι μια από τις σημαντικότερες διαδικασίες μέσω των οποίων επιτυγχάνεται η επίτευξη των στόχων της πρόβλεψης και της περιγραφής, στην εξόρυξη δεδομένων. Λαμβάνοντας υπόψη ότι η εξόρυξη δεδομένων αποτελεί εφαρμογή του ευρύτερου πεδίου της μηχανικής μάθησης, στην παρούσα παράγραφο θα περιγραφούν συνοπτικά τόσο οι κυριότερες μέθοδοι ταξινόμησης στη μηχανική μάθηση, και κατ'επέκταση στην εξόρυξη δεδομένων, όσο και σημαντικοί αλγόριθμοι των μεθόδων αυτών. Αξίζει να σημειωθεί ότι οι μέθοδοι και οι αλγόριθμοι που ακολουθούν αφορούν την ταξινόμηση δεδομένων όλων των ειδών (κατηγορικά, αριθμητικά και μεικτά).

2.4.1 ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ

Τα *δέντρα απόφασης* (*decision trees*) είναι μια από τις πιο σημαντικές και ευρύτατα διαδεδομένες μεθόδους για την ταξινόμηση δεδομένων. Είναι δομές που ταξινομούν τα αντικείμενα μιας βάσης δεδομένων βάσει των τιμών των χαρακτηριστικών αυτών, κατασκευάζονται δε με βάση ένα σύνολο εκπαίδευσης, το οποίο περιλαμβάνει προ-ταξινομημένα δεδομένα. Κάθε κόμβος του δέντρου αναπαριστά ένα χαρακτηριστικό ενός αντικειμένου που πρόκειται να ταξινομηθεί, ενώ κάθε κλαδί που ξεκινά από τον κόμβο αυτό αντιστοιχεί

σε μια από τις πιθανές τιμές του χαρακτηριστικού τις οποίες ο κόμβος μπορεί να λάβει. Επιπλέον, ένα φύλλο αντιστοιχεί σε μια από τις προκαθορισμένες κλάσεις της διαδικασίας της ταξινόμησης.

2.4.2 ΜΕΘΟΔΟΙ ΠΟΥ ΣΤΗΡΙΖΟΝΤΑΙ ΣΤΟΥΣ ΚΑΝΟΝΕΣ ΑΠΟΦΑΣΗΣ

Μια πολύ σημαντική ιδιότητα των δέντρων απόφασης, είναι η ικανότητα μετατροπής τους σε ένα σύνολο *κανόνων απόφασης* (***decision rules***). Συγκεκριμένα, δημιουργείται ένας ξεχωριστός κανόνας για κάθε μονοπάτι που ξεκινά από την κορυφή του δέντρου και καταλήγει σε ένα φύλλο που αναπαριστά μια κλάση. Επιπλέον, τα περισσότερα από τα άλλα είδη τυποποίησης των εξαγομένων των αλγορίθμων της εξόρυξης δεδομένων, όπως οι *λίστες απόφασης* (***decision lists***), τα *προς τα κάτω αναπτυσσόμενα σύνολα κανόνων* (***ripple down rule sets***), τα *επαγωγικά λογικά προγράμματα* (***inductive logic programs***) ή τα *νευρωνικά δίκτυα* (***neural networks***), μπορούν επίσης να μετατραπούν σε κανόνες. Ειδικά για την μετατροπή των τελευταίων σε κανόνες απόφασης, η διεθνής βιβλιογραφία είναι ιδιαιτέρως πλούσια.

Ωστόσο, αξίζει να σημειωθεί ότι οι κανόνες απόφασης μπορούν επιπλέον να εξαχθούν και απ'ευθείας από το σύνολο εκπαίδευσης μιας βάσης δεδομένων, μέσω μιας σειράς αλγορίθμων ταξινόμησης, οι οποίοι βασίζονται στους κανόνες απόφασης (***rule-based methods***). Στόχος των παραπάνω αλγορίθμων είναι η εξαγωγή του μικρότερου δυνατού συνόλου κανόνων απόφασης που είναι συνεπές με τα υπό εκπαίδευση δεδομένα. Οι εξαχθέντες κανόνες απόφασης έχουν την γενική μορφή «If *A* Then *B*», με το «If» κομμάτι να αποτελεί ένα συνδυασμό ζευγών από

τιμές χαρακτηριστικών, αναπαριστώντας τις επαρκείς συνθήκες για την εφαρμογή - ανάθεση της τιμής της κλάσης που περιγράφεται στο «Then» κομμάτι του κανόνα, στο υπό ταξινόμηση αντικείμενο της βάσης δεδομένων.

2.4.3 ΤΕΧΝΗΤΑ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ

Πέρα από τις μεθόδους ταξινόμησης που βασίζονται στα δέντρα και τους κανόνες απόφασης, τα *τεχνητά νευρωνικά δίκτυα* (***artificial neural networks***) είναι επίσης μια διαδεδομένη μέθοδος ταξινόμησης. Συγκεκριμένα, είναι μια δομή που αποτελείται από ένα δίκτυο *νευρώνων* (***neurons***) οι οποίοι συνδέονται μεταξύ τους. Η πιο διαδεδομένη κατηγορία νευρωνικών δικτύων είναι τα λεγόμενα *δίκτυα πρόσθιας τροφοδότησης* (***feed-forward neural networks***), τα οποία επιτρέπουν την κίνηση των δεδομένων μόνο προς μια κατεύθυνση, δηλαδή από μια είσοδο προς μια έξοδο. Δίκτυα που σχηματίζουν κυκλικές δομές ονομάζονται *ανατροφοδοτούμενα νευρωνικά δίκτυα* (***recurrent neural networks***).

Οι νευρώνες ενός δικτύου χωρίζονται σε τρείς βασικές κατηγορίες: (1) τους *νευρώνες εισόδου* (***input neurons***), οι οποίοι δέχονται τις πληροφορίες που θα υποστούν επεξεργασία, (2) τους *νευρώνες εξόδου* (***output neurons***), στους οποίους καταλήγουν τα αποτελέσματα της παραπάνω επεξεργασίας, και (3) τους *ενδιάμεσους νευρώνες*, οι οποίοι βρίσκονται μεταξύ των νευρώνων εισόδου και εξόδου. Οι τελευταίοι εναλλακτικά ονομάζονται και *κρυφοί νευρώνες* (***hidden neurons***). Ουσιαστικά, οι νευρώνες σε ένα δίκτυο είναι αφενός ένα σύνολο εισερχόμενων τιμών και των αντίστοιχων βαρών τους, και αφετέρου μια συνάρτηση που αθροίζει τα παραπάνω βάρη, αντιστοιχώντας τα αποτελέσματα σε ένα νευρώνα εξόδου.

2.4.4 ΣΤΑΤΙΣΤΙΚΕΣ ΜΕΘΟΔΟΙ

Οι *στατιστικές μέθοδοι* (*statistical methods*) ταξινόμησης χαρακτηρίζονται από το γεγονός ότι χρησιμοποιούν μοντέλα πιθανότητας, τα οποία αντί για μια απλή ταξινόμηση ενός αντικειμένου που ανήκει σε μια βάση δεδομένων, δίνουν την πιθανότητα το αντικείμενο αυτό να ανήκει σε κάθε μια από τις κλάσεις της διαδικασίας της ταξινόμησης. Τα πιο συνηθισμένα στατιστικά μοντέλα ταξινόμησης

ορίζονται βάσει της *θεωρίας του Bayes* (Cheeseman & Stutz; 1996). Πρόκειται για τον *αφελή ταξινομητή Bayes* αφενός, και τα *δίκτυα Bayes* αφετέρου.

Ο αφελής ταξινομητής Bayes (*Naïve Bayes classifier*) χρησιμοποιήθηκε για πρώτη φορά στο πεδίο της μηχανικής μάθησης από τους Cestnik et al (1987). Υποθέτει ότι η παρουσία (ή απουσία) ενός συγκεκριμένου χαρακτηριστικού μιας κλάσης είναι ανεξάρτητη από την παρουσία (ή απουσία) κάθε άλλου χαρακτηριστικού. Η υπόθεση αυτή ονομάζεται *υπό συνθήκη ανεξαρτησία* (*conditional independence*).

Ένα *δίκτυο Bayes* (*Bayes network*) (Jensen, 1996) είναι ένα γραφικό μοντέλο που βασίζεται σε πιθανότητες, λαμβάνοντας υπόψη το σύνολο των μεταβλητών του μοντέλου και τις μεταξύ τους εξαρτήσεις. Πιο συγκεκριμένα, ένα δίκτυο Bayes είναι ένας *κατευθυνόμενος μη κυκλικός γράφος* (*directed acyclic graph*), κάθε κόμβος του οποίου αντιπροσωπεύει ένα αντικείμενο X , ενώ κάθε τόξο αντιπροσωπεύει τις

μεταξύ των αντικειμένων εξαρτήσεις, υπό την μορφή πιθανοτήτων (*probabilistic dependencies*).

2.4.5 ΜΕΘΟΔΟΙ ΜΑΘΗΣΗΣ ΚΑΤΑ ΠΕΡΙΠΤΩΣΗ

Μια άλλη διαδεδομένη μέθοδος ταξινόμησης είναι η *μέθοδος της μάθησης κατά περίπτωση* (***instance-based learning***). Οι αλγόριθμοι της μεθόδου αυτής, είναι *αλγόριθμοι αναβλητικής μάθησης* (***lazylearning algorithms***). Αυτό σημαίνει ότι στους αλγορίθμους αυτούς, η γενίκευση πέρα από τα δεδομένα της εκπαίδευσης καθυστερεί μέχρις ότου γίνει μια

πρώτη ταξινόμηση, σε αντίθεση με τους *αλγορίθμους έγκαιρης μάθησης* (***eagerlearning algorithms***), όπως οι αλγόριθμοι δέντρων απόφασης ή νευρωνικών δικτύων, στους οποίους το μοντέλο προσπαθεί πρώτα να γενικεύσει τα δεδομένα εκπαίδευσης και μετά να ταξινομήσει νέα δεδομένα. Κατ'επέκταση, οι αλγόριθμοι αναβλητικής μάθησης έχουν μικρότερη υπολογιστική πολυπλοκότητα στην φάση της εκπαίδευσης σε σχέση με τους αλγορίθμους έγκαιρης μάθησης, αλλά μεγαλύτερη πολυπλοκότητα στην φάση της ταξινόμησης.

Ένας από τους πλέον διαδεδομένους αλγόριθμους μάθησης κατά περίπτωση, είναι ο αλγόριθμος του ***k-*** *Κοντινότερου Γείτονα* (***k-Nearest Neighbor*** ή ***kNN***). Ο αλγόριθμος αυτός βασίζεται στην αρχή που υποστηρίζει ότι τα αντικείμενα μιας βάσης δεδομένων βρίσκονται σε εγγύτητα με άλλα αντικείμενα που έχουν περεμφερείς ιδιότητες. Αν κάθε ένα από τα αντικείμενα αυτά είναι προσκολημένα σε μια κλάση, τότε ο καθορισμός της κλάσης στην οποία θα ανατεθεί ένα μη ταξινομημένο αντικείμενο, γίνεται μέσα από την παρατήρηση των κλάσεων στις οποίες είναι αντιστοιχισμένα τα κοντινότερα σε αυτό αντικείμενα.

2.5 Η ΟΜΑΔΟΠΟΙΗΣΗ ΣΤΗΝ ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

Η ομαδοποίηση είναι μια από τις πιο χρήσιμες και ευρύτατα εφαρμοζόμενες διαδικασίες εξόρυξης δεδομένων. Χαρακτηριστικές εφαρμογές της υφίστανται στους τομείς των επιχειρήσεων, της ιατρικής, της βιολογίας, της εξόρυξης στον παγκόσμιο ιστό, της χωροθέτησης κ.α. (Berry & Linoff, 2004). Όπως προ-αναφέρθηκε, η ομαδοποίηση αφορά τον *διαχωρισμό* (*partition*) των αντικειμένων μιας βάσης δεδομένων σε μη συνδεδεμένες μεταξύ τους ομοιογενείς ομάδες, οι οποίες είναι καταρχάς άγνωστες όσον αφορά το πλήθος και την δομή τους, κατά τέτοιο τρόπο ώστε αντικείμενα του συνόλου που ανήκουν σε μια ομάδα, να είναι πιο όμοια μεταξύ τους, παρά με τα αντικείμενα που ανήκουν σε διαφορερικές ομάδες.

Οι αλγόριθμοι ομαδοποίησης μπορούν να ταξινομηθούν με βάση:

- **Τον τύπο δεδομένων που εισάγονται στον αλγόριθμο:** οι αλγόριθμοι ομαδοποίησης κατηγοριοποιούνται με γνώμονα την αριθμητική, κατηγορική ή μεικτή (αριθμητική και κατηγορική) φύση των υπό εξέταση βάσεων δεδομένων.
- **Τη θεωρία και τις έννοιες στις οποιές βασίζεται η ανάλυση ομάδας (cluster analysis):** Οι αλγόριθμοι ομαδοποίησης κατηγοριοποιούνται βάσει του τρόπου διαχειρισμού της αβεβαιότητας στο ζήτημα της άλληλοεπικάλυψης μεταξύ των ομάδων.
- **Τη μέθοδο που καθορίζει την ομαδοποίηση της εκάστοτε βάσης δεδομένων:** Οι αλγόριθμοι κατηγοριοποιούνται βάσει του τρόπου με τον οποίο γίνεται η ομαδοποίηση των υπό εξέταση δεδομένων.

Βάσει της λογικής του τελευταίου από τα παραπάνω τρία κριτήρια, οι μέθοδοι και οι αλγόριθμοι ομαδοποίησης μπορούν να κατηγοριοποιηθούν ως ακολούθως:

Ø Ιεραρχικοί (hierarchical) αλγόριθμοι ομαδοποίησης: Οι αλγόριθμοι αυτοί χωρίζονται σε *συσσωρευτικούς* (***agglomerative***) και σε *διαιρετικούς* (***divisive***). Η βασική τους λογική είναι η διάσπαση μιας βάσης δεδομένων που περιλαμβάνει ***n*** αντικείμενα, σε πολλά επίπεδα από ομάδες, οι οποίες παριστάνονται μέσω μιας *δενδρικής μορφής* (*dendrogram*).

Ø Αλγόριθμοι ομαδοποίησης βασισμένοι στην βελτιστοποίηση της συνάρτησης κόστους: Οι αλγόριθμοι αυτοί σχηματίζουν τις ομάδες τους μέσω της βελτιστοποίησης ενός κριτηρίου ομαδοποίησης. Οι ομάδες αυτές είναι ομοιογενείς και μη επικαλυπτόμενες, τα δε αντικειμενά τους ανήκουν αποκλειστικά σε αυτές. Υπάρχουν δύο υποκατηγορίες αλγοριθμων, οι *επαναληπτικοί αλγόριθμοι διαχωρισμού* (***iterative partitioning algorithms***) τους οποίους θα δούμε αναλυτικότερα, και οι *πιθανοθεωρητικοί αλγόριθμοι ομαδοποίησης* (***probabilistic clustering algorihtms***).

Ø Τεχνικές Density ή Mode-Seeking: Στις τεχνικές αυτές, οι ομάδες θεωρούνται ως περιοχές του χώρου με υψηλή πυκνότητα σε αντικείμενα, οι οποίες χωρίζονται από άλλες τέτοιες περιοχές, από τμήματα του χώρου χαμηλής πυκνότητας σε αντικείμενα. Η αναζήτηση περιοχών υψηλής πυκνότητας γίνεται μέσω κάποιου τοπικού κριτηρίου ομαδοποίησης.

Ø Ανταγωνιστικοί αλγόριθμοι μάθησης: Οι συγκεκριμένοι αλγόριθμοι μάθησης παράγουν ομαδοποιήσεις και συγκλίνουν στην πιο «λογική» από αυτές, σύμφωνα με ένα μέτρο απόστασης.

Ø Branch & Bound αλγόριθμοι ομαδοποίησης: Οι αλγόριθμοι αυτής της κατηγορίας παρέχουν *ολικά βέλτιστες* ομαδοποιήσεις χωρίς την εξέταση όλων των πιθανών ομαδοποιήσεων, για ένα συγκεκριμένο πληθος ***m*** ομάδων και για κάποιο προκαθορισμένο κριτήριο. Η υπολογιστική πολυπλοκότητά τους ωστόσο είναι πολύ μεγάλη.

ΚΕΦΑΛΑΙΟ 3
ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΜΕ ΤΟ KNOWLEDGESEEKER

3.1 ΕΙΣΑΓΩΓΗ

Το Knowledge Seeker είναι ένα υπολογιστικό πακέτο που επεξεργάζεται και αναλύει δεδομένα (data mining tool) και επιτρέπει στους χρήστες του να επεξεργαστούν και να καταλάβουν τις σχέσεις που υπάρχουν μεταξύ των μεταβλητών σε ένα σύνολο δεδομένων. Παρουσιάζει μία πλήρη ανάλυση γραμμικών και μη γραμμικών σχέσεων και εκθέτει τα αποτελέσματα γραφικά με στατιστικά δέντρα αποφάσεων, χρησιμοποιώντας ποσοστά ή μέσους όρους δίνοντας σημαντικές πληροφορίες. Υποστηρίζεται ότι η ανάλυση που παρουσιάζει θα διαρκούσε πολλές μέρες από ένα στατιστικό για να ολοκληρωθεί. Είναι φιλικό με το χρήστη και εύχρηστο ακόμη και σε μη στατιστικούς. Ψάχνει για τις σχέσεις που θα καθορίσει ο χρήστης, έχει τη δυνατότητα γρήγορων ελέγχων υποθέσεων και μπορεί επίσης να εκφράσει τις σχέσεις που έχει βρει με τη μορφή κανόνων. Αυτοί οι κανόνες βοηθούν για τις προβλέψεις, τον προγραμματισμό ή τις διαγνώσεις που μπορεί να γίνουν. Παρέχει τέλος, μία μέθοδο για τον προκαθορισμό της διάταξης και της τυποποίησης των δεδομένων.

Το υπολογιστικό αυτό πακέτο εξόρυξης δεδομένων μπορεί να δώσει απαντήσεις σε σημαντικές ερωτήσεις έχοντας στη διάθεσή του μεγάλο πλήθος από δεδομένα. Αρχικά επεξεργάζεται αυτόματα όλα τα δεδομένα, συνοψίζει τα στατιστικά σημαντικά πεδία και τις σχέσεις μεταξύ τους και παρουσιάζει τα αποτελέσματα με δέντρα αποφάσεων παρέχοντας ταυτόχρονα έλεγχο αξιοπιστίας και εγκυρότητας αυτών. Έτσι όχι μόνο ανακαλύπτει και υπογραμμίζει τις σχέσεις των δεδομένων, αλλά εξασφαλίζει και επιβεβαιώνει την εγκυρότητά τους.

Μπορεί να χρησιμοποιηθεί για την ανάλυση δεδομένων από βάσεις δεδομένων, λογιστικά ή στατιστικά φύλλα ή από κειμενογράφους, σε υπολογιστικά συστήματα μεγάλης ισχύος ή μικροϋπολογιστές. Χρησιμοποιείται λοιπόν σε ανάλυση αγοράς, όπου καθορίζει τους παράγοντες που επηρεάζουν τις πωλήσεις των προϊόντων (γεωγραφικούς, τιμές, χαρακτηριστικά πελάτη), σε έλεγχο ποιότητας, όπου αναγνωρίζει τους σημαντικούς παράγοντες για ελαττωματικά προϊόντα, σε θέματα υγείας, όπου εξετάζει τα δεδομένα για να ανακαλύψει συνδυασμένα αποτελέσματα που συμβάλλουν στην υγεία και την αρρώστια, σε θέματα διοικητικής ανάλυσης, όπου καθορίζει τους παράγοντες που επηρεάζουν το μισθό σε μεγάλο δείγμα εργαζομένων και καθορίζει πως σχετίζονται, σε θέματα επιστημονικής έρευνας, όπου αναλύει αποτελέσματα πειραμάτων και καθορίζει τους παράγοντες που επηρεάζουν την έρευνα και σε θέματα εξυπηρέτησης πελατών, όπου ανακαλύπτει προβλήματα στο χώρο παραγωγής πριν γίνουν επιδημικά.

3.2 ΕΙΣΑΓΩΓΗ ΔΕΔΟΜΕΝΩΝ (IMPORT)

Ο χρήστης έχει δυνατότητα εισαγωγής δεδομένων στο Knowledge Seeker IV από βάσεις δεδομένων, λογιστικά φύλλα και άλλα στατιστικά πακέτα, όπως από τα: dBase II (*.dbf), Paradox (*.db), Sawtooth (*.alp), SmartWare (*.db), SAS (*.tpt, *.ssd, *.sd2), SPSS (*.sav, *.por), Gauss (*.dat), Excel (*.xls), Lotus (*.w??), QuattroPro (*.wql), Splus (*.*), Stata (*.dta), Systat (*.sys) και παλαιότερες εκδόσεις του Knowledge Seeker (*.fmt). Όμως, και στην περίπτωση που το Knowledge Seeker δεν υποστηρίζει απευθείας κάποιο άλλο πρόγραμμα διαχείρισης δεδομένων, μπορούμε να εισάγουμε δεδομένα στο πρώτο υπό τη μορφή αρχείου κειμένου χαρακτήρων ASCII, αρκεί το δεύτερο να μπορεί να εξάγει δεδομένα σε αυτή τη μορφή. Ακόμα, παρέχεται η δυνατότητα

επεξεργασίας ODBC και SQL βάσεων δεδομένων, καθώς και επεξεργασία δεδομένων από το clipboard. Όλη η διαδικασία εισαγωγής δεδομένων γίνεται μέσω ενός αρκετά φιλικού οδηγού που ζητά από το χρήστη τον καθορισμό της πηγής δεδομένων, των πεδίων, το μέγεθος, τον τύπο και το χαρακτήρα διαχωρισμού τους, της εξαρτημένης μεταβλητής (Dependent Variable, DV), κ.α. Έχουμε επίσης τη δυνατότητα, από τον οδηγό εισαγωγής δεδομένων, να επιλέξουμε τόσο δείγμα δεδομένων από το συνολικό πληθυσμό, όσο και τον τύπο κωδικοποίησης του προς εισαγωγή αρχείου. Εάν το μέγεθος της βάσης δεδομένων είναι πολύ μεγάλο μπορούμε να μειώσουμε τον αριθμό των παρατηρήσεων επιλέγοντας δείγμα. Σε αυτή την περίπτωση, η δειγματοληψία δεν αλλοιώνει το αποτέλεσμα της ανάλυσης κυρίως στα "πάνω" επίπεδα του δέντρου αποφάσεων. Παίρνοντας λοιπόν ως δείγμα το 25% των παρατηρήσεων της βάσεως, το πρόγραμμα εκτιμά αυτό το ποσοστό επιλογής με τέτοιο τρόπο ώστε ο πραγματικός αριθμός των εγγραφών να ποικίλει.

Σε περίπτωση που θέλουμε να εισάγουμε δεδομένα από λογιστικά φύλλα, αυτά πρέπει να είναι στοιχισμένα πάνω αριστερά σε κάθε κελί του φύλλου. Οι διαφορετικές εγγραφές (παρατηρήσεις) πρέπει να είναι διατεταγμένες σε γραμμές και τα πεδία (μεταβλητές) σε στήλες. Η πρώτη εγγραφή (γραμμή) θα πρέπει να περιέχει τα ονόματα των πεδίων. Εάν δεν τα περιέχει το Knowledge Seeker δίνει αριθμητικό όνομα στα πεδία. Επίσης, μέσα στη βάση δεδομένων δεν πρέπει να υπάρχουν κενές γραμμές ή στήλες. Σε περίπτωση τέτοιων κενών το Knowledge Seeker δεν μπορεί να αναγνωρίσει τα όρια (μέγεθος) των δεδομένων. Κατά την εισαγωγή δεδομένων από εφαρμογές που υποστηρίζουν πολλαπλή επεξεργασία λογιστικών φύλλων εισάγεται μόνο το πρώτο φύλλο.

Στο Knowledge Seeker μπορούμε να πραγματοποιήσουμε εισαγωγή δεδομένων από το στατιστικό πακέτο SAS. Το πρόγραμμα μπορεί να διαβάσει αρχεία του SAS με επέκταση (*.tpt), που είναι συμβατά σχεδόν με όλα τα λειτουργικά συστήματα (portable file). Επίσης διαβάζει αρχεία του SAS έκδοση 6.0x με επέκταση (*.ssd) που είναι συμβατά με πλατφόρμες Windows και Sun, καθώς και αρχεία SAS έκδοση 6.11 που είναι συμβατά μόνο με πλατφόρμες Windows.

Στο Knowledge Seeker μπορούμε να πραγματοποιήσουμε εισαγωγή δεδομένων από το στατιστικό πακέτο SPSS. Το SPSS αποθηκεύει αρχεία σε δυαδική μορφή (binary format). Για να εισαχθεί αρχείο δεδομένων του SPSS στο Knowledge Seeker πρέπει πρώτα αυτό να έχει εξαχθεί από το SPSS σε μορφή αρχείου με επέκταση (*.por, portable file). Τέτοιου είδους αρχεία είναι γραμμένα υπό τη μορφή κειμένου χαρακτήρων ASCII.

3.3 ΕΞΑΓΩΓΗ ΔΕΔΟΜΕΝΩΝ (EXPORT)

Το Knowledge Seeker IV μπορεί να εξάγει δεδομένα σε μορφή αρχείου κειμένου (titled – no titled, delimited), καθώς και σε μορφή αρχείου έτοιμο προς χρήση από τις εφαρμογές dBase III, SAS (ver. 6.12) και Excel. Μπορούν να εξαχθούν όλα τα δεδομένα ή μέρος αυτών (δεδομένα που αντιστοιχούν σε συγκεκριμένο κόμβο).

3.4 ΔΕΝΤΡΑ ΑΠΟΦΑΣΕΩΝ

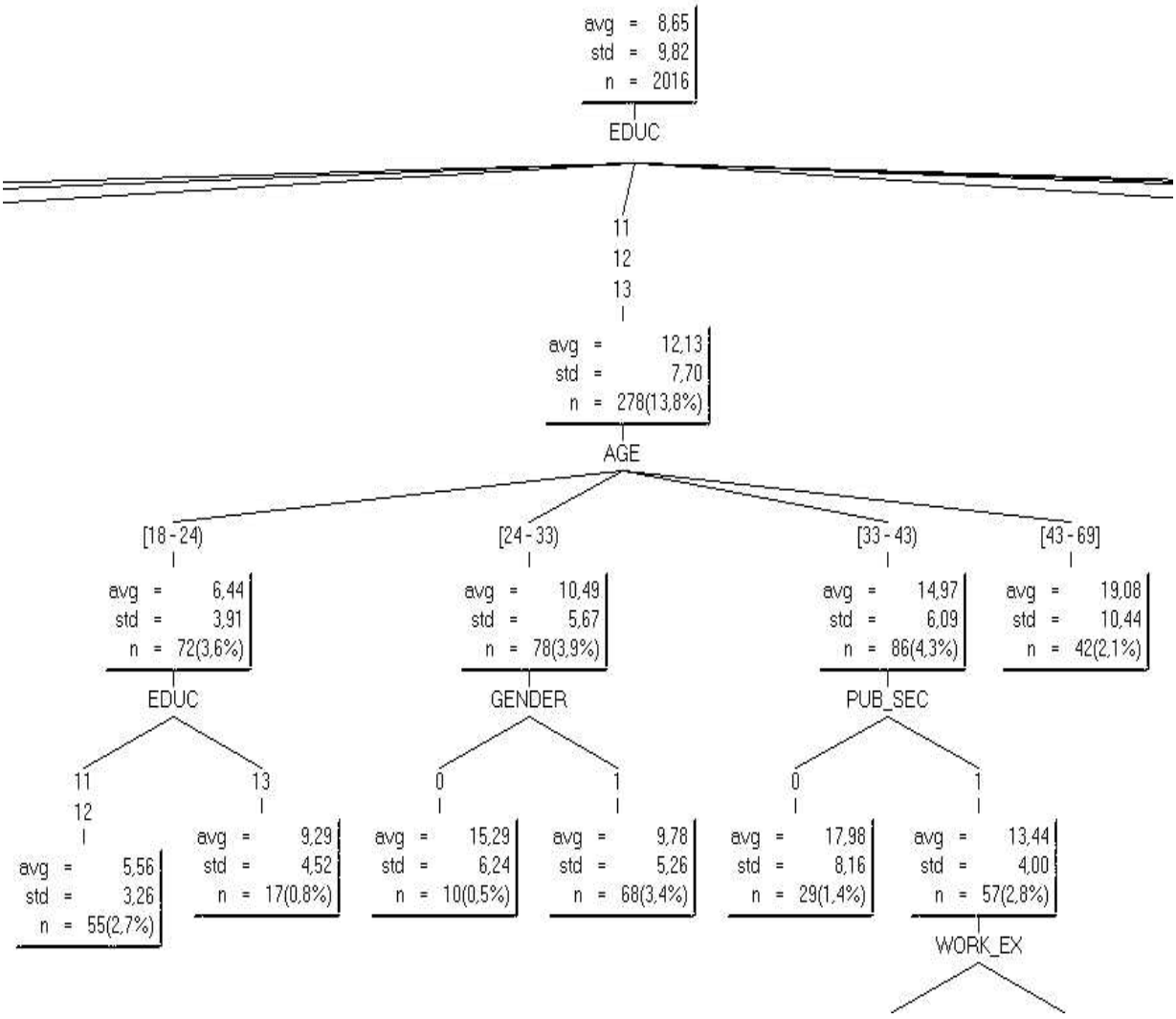
Όπως ήδη αναφέραμε το Knowledge Seeker χρησιμοποιεί τη μέθοδο της ανάλυσης με δέντρα αποφάσεων. Τα δέντρα αποφάσεων αναπτύχθηκαν κυρίως για την εξουδετέρωση κάποιων προβλημάτων που είχαν προκύψει με τη χρήση της πολυμεταβλητής παλινδρόμησης στην

Εξόρυξη Δεδομένων. Με αφορμή τέτοια προβλήματα οι Morgan και Sonquist ανέπτυξαν μία στατιστική τεχνική ανάλυσης δεδομένων, τον «αυτόματο αλληλεπιδρών ανιχνευτή», την AID (Automatic Interactive Detector) μέθοδο, με την οποία ανακαλύπτονται ακόμη και «κρυμμένες» σχέσεις των μεταβλητών. Τα δέντρα, έτσι των αποφάσεων, παρέχουν αυτές τις σχέσεις των μεταβλητών και δημιουργούν ένα μοντέλο αυτών που είναι έγκυρο αλλά και εύκολο να το ερμηνεύσει κανείς.

Μετά την εισαγωγή των δεδομένων μπορούμε να δημιουργήσουμε το δέντρο, το οποίο θα δώσει πληροφορίες για την ανάλυσή μας. Ο πρώτος κόμβος ή ρίζα του δέντρου, που εμφανίζεται αρχικά περιέχει τις τιμές της εξαρτημένης μεταβλητής που έχει καθοριστεί από το χρήστη. Προχωρώντας το Knowledge Seeker ψάχνει για όλες τις πιθανές σχέσεις των άλλων μεταβλητών με τη ρίζα και τις παρουσιάζει, αρχίζοντας από τις πιο σημαντικές. Εξετάζει όλα τα πεδία που μπορούν να χρησιμοποιηθούν για να περιγράψει την εξαρτημένη μεταβλητή και επιλέγει αυτά που μπορούν καλύτερα να εξηγήσουν ή να προβλέψουν τις μεταβλητότητες στη μεταβλητή. Κάθε μία από αυτές τις μεταβλητές είναι ανεξάρτητη με τις υπόλοιπες και με τον ίδιο τρόπο μπορεί να χωριστεί σε κατηγορίες, στις οποίες συνεχίζεται η ίδια διαδικασία ώστε να έχουμε περισσότερες πληροφορίες.

Στο παράδειγμα που θα χρησιμοποιήσουμε για την παρακάτω ανάλυση, δουλεύουμε με δεδομένα από τον χώρο της εργασίας, που περιέχει πληροφορίες για το εισόδημα των εργαζομένων, το παρελθόν τους όσον αφορά την εργασία τους καθώς και πληροφορίες για την εκπαιδευτική και οικογενειακή τους κατάσταση. Θα προσδιορίσουμε ένα προφίλ χαμηλόμισθων και υψηλόμισθων ανθρώπων όσο το δυνατό καθαρό και κατανοητό. Τα συμπεράσματα θα προκύψουν από τις τιμές των μεταβλητών που θα καθορίσουμε και τις σχέσεις μεταξύ τους και θα πρέπει να είναι ουσιώδη και έγκυρα.

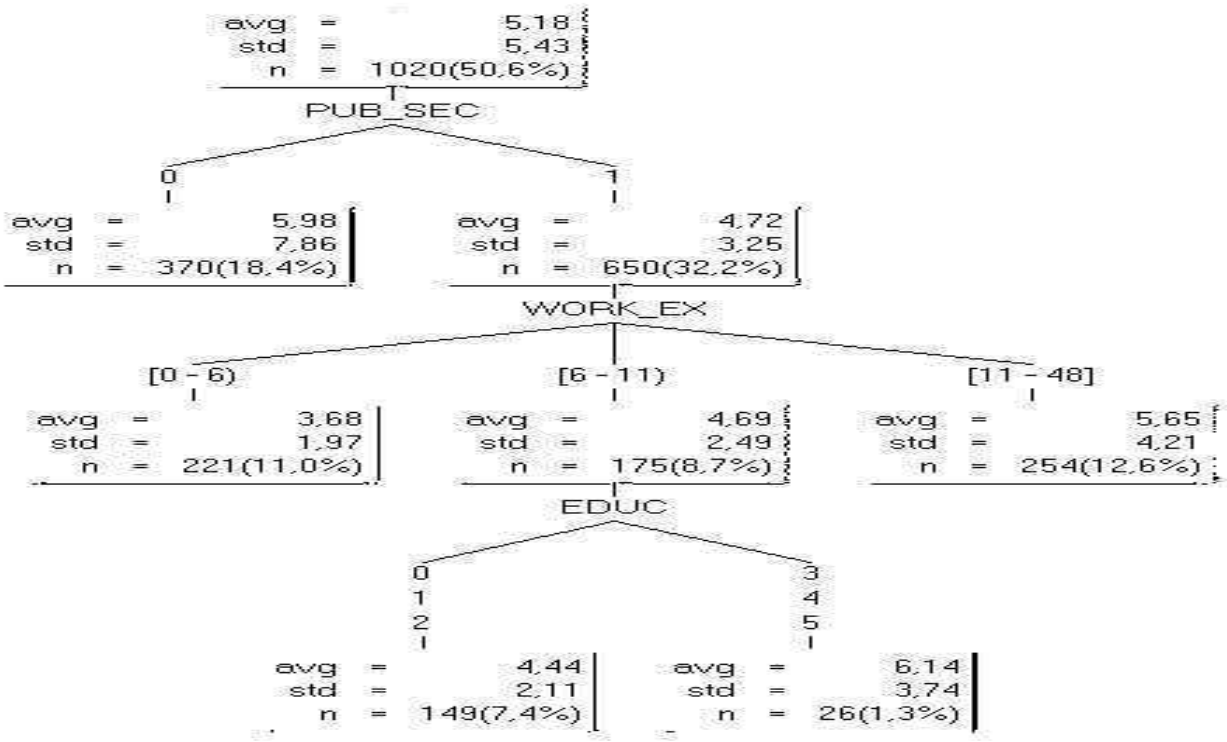
Έχουμε ορίσει δηλαδή, ως ***Wage*** το ωριαίο εισόδημα του ερωτηθέντα, ***Age*** την ηλικία (σε χρόνια) κάθε ερωτηθέντα, ***Work_ex*** την προϋπηρεσία (σε χρόνια), ***Gender*** την εικονική (dummy) μεταβλητή που παίρνει τιμές 0 και 1, αν ο ερωτούμενος είναι θηλυκού ή αρσενικού γένους αντίστοιχα, ***Pub_sec*** την εικονική μεταβλητή με τιμές 0 και 1, αν ο ερωτούμενος εργάζεται στον ιδιωτικό ή δημόσιο τομέα αντίστοιχα, ***Educ*** το επίπεδο εκπαίδευσης (σε χρόνια), ***Fath_ed*** το επίπεδο εκπαίδευσης (σε χρόνια) του πατέρα και ***Moth_ed*** το επίπεδο εκπαίδευσης (σε χρόνια) της μητέρας του ερωτηθέντα.



Η βασική-εξαρτημένη μεταβλητή που χρησιμοποιείται γι'αυτή την ανάλυση είναι το ωριαίο εισόδημα των εργαζομένων, ***Wage.*** Σκοπός μας

είναι να προσδιορίσουμε τους παράγοντες που προβλέπουν υψηλό ή χαμηλό εισόδημα αντίστοιχα. Στην αρχή, στη ρίζα του δέντρου εμφανίζονται βασικές πληροφορίες, από τις οποίες βρίσκουμε τις σχέσεις που μπορούν να εξηγήσουν υψηλό ή χαμηλό εισόδημα στη βάση δεδομένων. Όπως βλέπουμε από το παραπάνω σχήμα από τις 2016 εγγραφές, το μέσο εισόδημα ενός εργαζομένου είναι $8,65 ανά ώρα, με τυπική απόκλιση $9,82. Για να μελετήσουμε πως σχετίζεται το εισόδημα με τα άλλα πεδία προχωράμε στον επόμενο κόμβο. Στον επόμενο κόμβο εμφανίζεται η ανεξάρτητη μεταβλητή, το επίπεδο εκπαίδευσης. Βλέπουμε ότι η μεταβλητή ***Educ*** διασπάται σε οχτώ κλαδιά, τους ανθρώπους με εκπαίδευση 0,1,2,3,4,5 χρόνια στο πρώτο, με 6 χρόνια στο δεύτερο, με 8,9,10 στο τρίτο και προχωρώντας με 22 και 23 χρόνια στο τελευταίο. Ένα απλό συμπέρασμα που εξάγουμε βλέποντας έτσι το δέντρο, είναι ότι όσο αυξάνεται το επίπεδο εκπαίδευσης αυξάνεται και το ωραίο εισόδημα, αφού οι εργαζόμενοι με μέχρι πέντε χρόνια σπουδές παίρνουν ένα μέσο ωριαίο εισόδημα $5.18, μικρότερο και από το γενικό μέσο εισόδημα, ενώ αυτοί που έχουν 22 ή 23 χρόνια εκπαίδευσης παίρνουν $49.44. Είναι ευδιάκριτο όμως, ότι είναι πολύ μικρό το ποσοστό των εργαζομένων που έχουν περισσότερα από 15 χρόνια σπουδές (6.4%), ενώ το 50.6% έχουν 0-5 χρόνια σπουδών.

Με τον ίδιο τρόπο μπορούμε να συνεχίσουμε την ανάλυσή μας, εξετάζοντας για παράδειγμα πως μεταβάλλεται το εισόδημα της μιας κατηγορίας των ανθρώπων με μέχρι πέντε χρόνια σπουδές (1020) σε σχέση με μια άλλη ανεξάρτητη μεταβλητή, π.χ. την ***Pub_sec***, αν δηλαδή εργάζονται στο δημόσιο ή ιδιωτικό τομέα.



Αν κοιτάξουμε ολόκληρο το δέντρο θα παρατηρήσουμε ότι ένας ακόμη πολύ σημαντικός παράγοντας για το εισόδημα των εργαζομένων είναι η προϋπηρεσία. Ένα σημαντικό ποσοστό 13,3% με έξι χρόνια σπουδές έχει 0-24 χρόνια προϋπηρεσίας. Συνεχίζοντας, όσο αυξάνεται το επίπεδο εκπαίδευσης, μειώνεται η προϋπηρεσία, αφού μόνο το ποσοστό των 1,2% έχει εργαστεί 17,75 έως 48 χρόνια, ενώ έχει 14 με 15χρόνια σπουδών. Χαρακτηριστικό είναι επίσης, όπως βλέπουμε και από το παραπάνω σχήμα, ότι στη πρώτη κατηγορία ανθρώπων με 0-5 χρόνια σπουδές ο παράγοντας της προϋπηρεσίας λαμβάνεται και εδώ σοβαρά υπόψη, αλλά μόνο γι' αυτούς που εργάζονται στο δημόσιο τομέα. Οι ιδιωτικοί υπάλληλοι αυτής της κατηγορίας έχουν ένα σταθερό μέσο εισόδημα $5.96, ενώ στους δημοσίους μεταβάλλεται ανάλογα με την προϋπηρεσία που έχουν. Εντύπωση επιπλέον μας προκαλεί το γεγονός ότι στους δημόσιους υπαλλήλους αυτής της κατηγορίας με προϋπηρεσία 6-11 χρόνια παρεμβάλλεται πάλι ο παράγοντας της εκπαίδευσης και ανάλογα αν είναι πολλά ή λίγα τα χρόνια των σπουδών τους έχουν μεγάλο ή μικρό ωριαίο εισόδημα αντίστοιχα. Στη μεσαία κατηγορία

ανθρώπων με 11-13 χρόνια εκπαίδευσης, όπως μπορούμε να δούμε από το πρώτο σχήμα, παρεμβάλλεται η μεταβλητή της ηλικίας, με τους εργαζομένους μεγαλύτερης ηλικίας να έχουν υψηλότερο εισόδημα, κάτι που αναμενόταν. Αλλά και στην κατηγορία αυτών με ηλικία 24-33 παρεμβάλλεται η μεταβλητή του φύλου και έχουμε τις γυναίκες, παρότι είναι μικρό το ποσοστό τους (0.5%), να έχουν υψηλότερο εισόδημα από τους άντρες.

Από τα παραπάνω, βλέπουμε λοιπόν δύο πολύ σημαντικούς παράγοντες να επηρεάζουν το ωριαίο εισόδημα των εργαζομένων, το επίπεδο εκπαίδευσης και την προϋπηρεσία. Είδαμε ότι υπάρχει αύξηση του εισοδήματος, τόσο με την αύξηση της πρώτης μεταβλητής, όσο και με της δεύτερης. Όμως ο παράγοντας της εκπαίδευσης μπορούμε να πούμε ότι είναι πιο ισχυρός, αφού μπορεί να παρατηρηθεί ότι το μεγαλύτερο εισόδημα των εργαζομένων με τη μεγαλύτερη προϋπηρεσία είναι $21.36, ενώ με τα περισσότερα χρόνια εκπαίδευσης είναι $49.44, αισθητά πιο υψηλό.

Παρότι λοιπόν το Knowledge Seeker εμφανίζει την πιο σημαντική μεταβλητή στο δέντρο αμέσως μετά τη ρίζα, υπάρχουν και άλλα κλαδιά – μεταβλητές, που μπορούμε να επιλέξουμε, με τη μελέτη των οποίων μπορούμε να βγάλουμε εξίσου σημαντικά αποτελέσματα. Μπορούμε έτσι να το δημιουργήσουμε εμείς το δέντρο, αντί να το βρει μόνο του το πρόγραμμα αυτόματα, ώστε να ελέγχουμε και να δημιουργούμε όπως χρειάζεται την κατηγοριοποίηση των μεταβλητών. Ακόμη, όμως και όταν η ανάπτυξη του δέντρου εκτελείται αυτόματα, μπορούμε να καθορίσουμε τον ελάχιστο αριθμό εγγραφών σε κάθε κόμβο, να τον αναγνωρίσουμε πριν παρουσιαστεί το δέντρο, οπότε αν χρειαστεί να τον αλλάξουμε, καθώς και να δώσουμε ένα άνω φράγμα στον αριθμό των τιμών για κάθε δοσμένη ανεξάρτητη μεταβλητή. Το Knowledge Seeker, έτσι δεν επιχειρεί να διαιρέσει ένα κόμβο που περιέχει λιγότερες εγγραφές από

αυτές που θα έχουν προσδιοριστεί και αν μια μεταβλητή υπερβαίνει τον αριθμό των τιμών που θα έχει δοθεί ως άνω φράγμα τότε αυτή αγνοείται. Συνήθως, ο μέγιστος αυτός αριθμός τιμών είναι οχτώ φορές ο αριθμός των κατηγοριών σε κάθε ομάδα.

Οι μεταβλητές που χρησιμοποιούνται μπορεί να είναι κατηγορικές, δηλαδή οι τιμές τους να είναι κατηγοριοποιημένες ή συνεχείς με τιμές αριθμητικές. Οι πληροφορίες που αναγράφονται στο δέντρο αλλάζουν ανάλογα αν η μεταβλητή είναι κατηγορική ή συνεχής. Μία αναλυτική παρουσίαση αυτών για την κατηγορική περιέχει: το μέγεθος του δείγματος (αριθμός εγγραφών), τον αριθμό και τα ποσοστά από αυτά σε όλες τις κατηγορίες της μεταβλητής ,το συντελεστή εμπιστοσύνης ,το αποτέλεσμα από χ^2-ελέγχους και την τιμή της p-συνάρτησης. Για τη συνεχή μεταβλητή, περιέχει την τιμή της τυπικής απόκλισης και του μέσου όρου των τιμών της μεταβλητής ,το μέγεθος του δείγματος, το συντελεστή εμπιστοσύνης, τα αποτελέσματα με τους βαθμούς ελευθερίας του F-ελέγχου, καθώς και την τιμή της p- συνάρτησης. Αυτό φαίνεται εξάλλου και από τα σχήματα στο συγκεκριμένο παράδειγμα, αφού η μεταβλητή *Wage* είναι συνεχής μεταβλητή.

Για κάθε σχέση που αναγνωρίζεται από το Knowledge Seeker και παρουσιάζεται με μορφή κόμβων στο δέντρο, ελέγχεται η αξιοπιστία της. Χρησιμοποιούνται λοιπόν στατιστικές ρυθμίσεις για την αποφυγή ψευδών σχέσεων και την εξισορρόπηση της μεροληψίας στο βαθμό εμπιστοσύνης που έχει οριστεί κάθε φορά, λόγω των αλγορίθμων αναζήτησης που περιέχει το πακέτο. Όταν όμως δημιουργούμε εμείς το δέντρο αποφάσεων δεν χρησιμοποιούνται τέτοιοι αλγόριθμοι, επομένως δεν είναι απαραίτητες τέτοιες ρυθμίσεις. Έτσι εξηγείται και το φαινόμενο που μπορεί να παρουσιαστεί όταν το ίδιο δέντρο με τις ίδιες διασπάσεις και άρα και τις ίδιες σχέσεις των μεταβλητών παρουσιάζεται σε

διαφορετικό επίπεδο εμπιστοσύνης, αν το έχει δημιουργήσει ο χρήστης ή αυτόματα το Knowledge Seeker.

3.5 ΤΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΤΩΝ ΔΕΔΟΜΕΝΩΝ

Το πρώτο πράγμα που θα πρέπει να γίνει για να μπορεί να γίνει η ανάλυση και να χρησιμοποιηθούν τα δεδομένα είναι να δηλώσουμε την εξαρτημένη μεταβλητή. Το πεδίο που μας ενδιαφέρει περισσότερο να μελετήσουμε και να εξηγήσουμε μας παρέχει τις τιμές της εξαρτημένης μεταβλητής. Στο παράδειγμά μας θεωρούμε τη μεταβλητή ***Wage*** ως την εξαρτημένη, γιατί θέλουμε να μελετήσουμε πως μεταβάλλεται το εισόδημα των εργαζομένων και κατά πόσο εξαρτάται από άλλους παράγοντες όπως είναι η ηλικία, η προϋπηρεσία, το προηγούμενο εισόδημα και από διάφορους άλλους.

Έπειτα για τη διευκόλυνσή μας στην παρουσίαση των δεδομένων το Knowledge Seeker μας επιτρέπει να αλλάξουμε τα δεδομένα και τη μορφή που είναι στη βάση δεδομένων, καθώς και τα πεδία που έχουν χρησιμοποιηθεί και πως οι τιμές τους παρουσιάζονται.

Χρησιμοποιώντας το όνομα του πεδίου που βρίσκεται στη βάση δεδομένων μπορούμε να το αλλάξουμε στο πεδίο του αρχείου δεδομένων που έχουμε εισάγει, ώστε να δηλώνει περισσότερες πληροφορίες. Για παράδειγμα, στο όνομα της μεταβλητής ***Fath_educ*** μπορούμε να αντιστοιχίσουμε το όνομα ***Father's education***. Αν κάποιο πεδίο δε μας ενδιαφέρει μπορούμε να το αγνοήσουμε.

Κάθε μεταβλητή, τώρα, καθορίζεται από τον τύπο των δεδομένων (συνεχή ή διακριτά) να είναι συνεχής (continuous) ή κατηγορική (categorigal). Γενικά ένα κατηγορικό πεδίο δεν περιέχει αριθμητικές τιμές, ενώ σε ένα συνεχές μπορούν να γίνουν αριθμητικές πράξεις στις τιμές του. Μπορούμε βέβαια να μετατρέψουμε μια διακριτή μεταβλητή

σε κατηγορική, αντιστοιχώντας μία τιμή σε πολλές. Για παράδειγμα αν έχουμε σε ένα πεδίο τις τιμές 1, 2, 3, μπορούμε να τις αντιστοιχίσουμε στην τιμή «χαμηλό επίπεδο». Αν τα δεδομένα είναι συνεχή (αριθμητικές τιμές), για να είναι ευκολοχείριστα κατά τη δημιουργία δέντρων αποφάσεων, οι τιμές τους χωρίζονται σε ευδιάκριτα διαστήματα τιμών, κάθε ένα από τα οποία περιέχει τον ίδιο αριθμό υποθέσεων. Τα διαστήματα αυτά μπορούν να δημιουργηθούν αυτόματα από το Knowledge Seeker ή να οριστούν από τον χρήστη, δηλώνοντας τα όρια για το διαχωρισμό αυτών. Τα διαστήματα έτσι θα παράγουν συνεχή δεδομένα σε ένα πιο εύχρηστο αριθμό ξεχωριστών κατηγοριών.

Εκτός από τη χρήση διαστημάτων, για να είναι τα δεδομένα πιο εύχρηστα και κατανοητά μπορούν να ομαδοποιηθούν (grouping), ώστε να μπορούμε να διαχειριστούμε καλύτερα τις διαφορετικές τιμές σε κάθε πεδίο. Αν δεν ομαδοποιηθούν, κάθε κατηγορία – πεδίο θα αναπαριστάνει ένα ξεχωριστό κλαδί στο δέντρο αποφάσεων. Με αυτή τη μορφή τα δεδομένα μπορούν να χρησιμοποιηθούν για την προσομοίωση των ID3 machine learning algorithms ή για την παραγωγή των πινάκων δεδομένων. Αν ομαδοποιηθούν, η ομαδοποίηση γίνεται συνήθως διατάσσοντας (ή βάζοντας σε κάποια τάξη) τις διαφορετικές τιμές των δεδομένων και έχοντας τις όμοιες τιμές μαζί. Στις μεθόδους (κατά συστάδες και καθολική) που χρησιμοποιούνται για την ομαδοποίηση αυτή των τιμών σε διαχωρισμένα πεδία, τα οποία παρουσιάζονται σαν κλαδιά στο δέντρο αποφάσεων, θα αναφερθούμε αναλυτικότερα παρακάτω.

Ένας άλλος τρόπος χειρισμού δεδομένων είναι η συγχώνευση (merging) των τιμών των δεδομένων σε ομάδες, ορίζοντας πάντα το βαθμό εμπιστοσύνης για να γίνει αυτή. Ο βαθμός που συνίσταται είναι ο 0.05. Αν είναι πολύ μικρός της τάξης του 0.001 θα υπάρχει 99.9% ορθότητα στις κατηγορίες, αλλά θα υπάρχουν πολύ λιγότερα

αποτελέσματα από αυτές, κάτι που θα δυσκόλευε την περαιτέρω ανάλυση. Αν είναι αρκετά μεγάλος, της τάξης του 0.2 θα υπάρχουν περισσότερα κλαδιά (αποτελέσματα), όμως μικρότερη εμπιστοσύνη (80%). Γι αυτό προτιμάται α=0.05.

Όταν τέλος, μεταξύ των τιμών υπάρχουν τιμές που λείπουν (missing values), αυτές ή θα αποκλειστούν εντελώς από την ανάλυση, μειώνοντας έτσι το μέγεθος του δέντρου αποφάσεων αφού κάθε εγγραφή με τιμή που λείπει θα παραλείπεται, ή θα χειριστούν όπως κάθε άλλη τιμή στο δέντρο. Στη δεύτερη περίπτωση, όταν γίνεται η ομαδοποίηση με σειρά ταξινόμησης θα τοποθετούνται στην αρχή ή στο τέλος ανάλογα ,αλλιώς (χωρίς ταξινόμηση) θα βρίσκονται στην κατηγορία που ταιριάζουν περισσότερο.

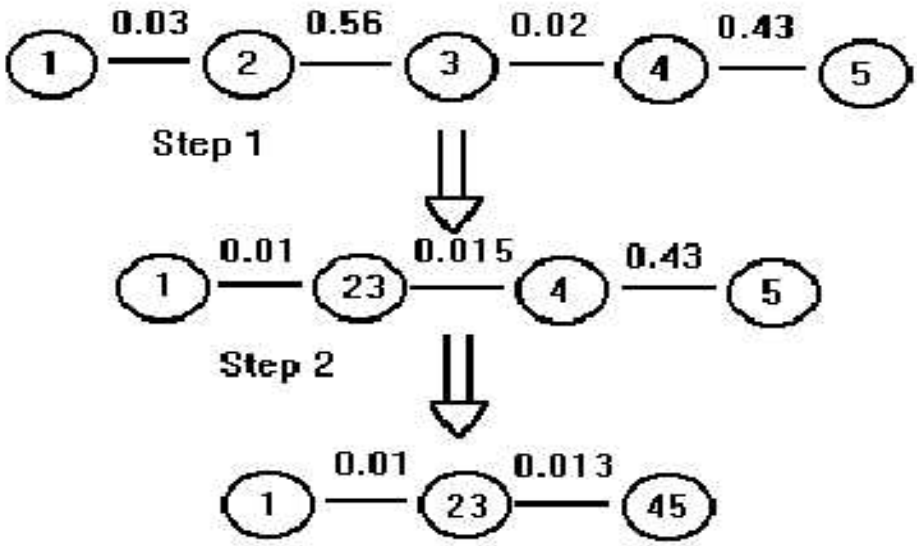
3.6 ΔΙΑΜΟΡΦΩΣΗ ΤΩΝ ΔΕΝΤΡΩΝ

3.6.1 ΑΛΓΟΡΙΘΜΟΣ ΓΙΑ ΤΗΝ ΕΠΙΛΟΓΗ ΤΩΝ ΔΙΑΣΠΑΣΕΩΝ ΤΩΝ ΔΕΝΤΡΩΝ

Το Knowledge Seeker χρησιμοποιεί δύο μεθόδους, που μπορούν να επιλεγούν από το χρήστη, για τον προσδιορισμό των κλαδιών των δέντρων αποφάσεων, δηλαδή των σχέσεων μεταξύ της εξαρτημένης και των υπόλοιπων μεταβλητών. Αυτές είναι η μέθοδος κατά συστάδες (cluster) και η καθολική μέθοδος (exhaustive). Και οι δύο μέθοδοι δημιουργούν μία διάσπαση για κάθε διαιρεμένο πεδίο – μεταβλητή.

Η μέθοδος κατά συστάδες ομαδοποιεί τα δεδομένα, βάζοντας τις παρόμοιες ή με ίδιες ιδιότητες τιμές του πεδίου μαζί. Οι τιμές που βρίσκονται στη ίδια συστάδα έχουν την ίδια στατιστική σχέση με την εξαρτημένη μεταβλητή. Οι τιμές που βρίσκονται σε διαφορετικές συστάδες έχουν διαφορετική στατιστική σχέση με την εξαρτημένη

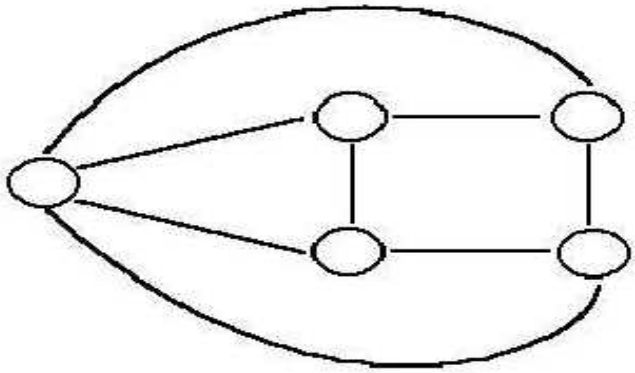
μεταβλητή. Οι συστάδες δημιουργούνται με έναν αυξανόμενο και επαναληπτικό τρόπο. Σε κάθε αύξηση οι δύο πιο όμοιες τιμές συγχωνεύονται. Το παρακάτω διάγραμμα ροής δείχνει όλη τη διαδικασία της συσταδοποίησης για ένα διατεταγμένο πεδίο με πέντε τιμές και ενώ έχει τεθεί βαθμός εμπιστοσύνης 0.05. Στο συγκεκριμένο παράδειγμα η μέθοδος αυτή είναι μία διάσπαση με τρία βήματα. Οι κύκλοι αναπαριστούν τις τιμές, ενώ οι γραμμές που τους ενώνουν δηλώνουν πιθανές συγχωνεύσεις.



Η εμπιστοσύνη που θέτουμε ελέγχει τελικά το επίπεδο ομοιότητας μέσα σε κάθε συστάδα, αφού ένας μεγάλος βαθμός εμπιστοσύνης σημαίνει υψηλό επίπεδο ομοιότητας. Αν για παράδειγμα αυξήσουμε το συντελεστή εμπιστοσύνης από 0.05 σε 0.2, θα παραχθούν περισσότερες συστάδες σε μία διάσπαση για επιεικέστερο έλεγχο διάκρισης των αποτελεσμάτων σε λιγότερες τιμές, αφού οι ίδιες έχουν ομαδοποιηθεί.

Η καθολική μέθοδος βρίσκει τις πιο αξιόπιστες σχέσεις, ψάχνοντας όλους τους δυνατούς συνδυασμούς ομαδοποίησης των τιμών.

Χρησιμοποιεί την επαναληπτική μέθοδο συγχώνευσης που χρησιμοποιούσε και η προηγούμενη και σταματάει όταν οι συστάδες που θα μείνουν είναι δύο. Για παράδειγμα όπως φαίνεται στο παρακάτω σχήμα το διατεταγμένο πεδίο πέντε τιμών μετά από τρία βήματα συγχώνευσης, καταλήγει σε διάσπαση δύο τιμών. Αυτή η μέθοδος απλά επιλέγει την πιο αξιόπιστη διάσπαση, επιλέγοντας από όλες τις διασπάσεις αυτή με το μεγαλύτερο επίπεδο εμπιστοσύνης.



Η διάσπαση λοιπόν παράγεται από την ομαδοποίηση των τιμών των πεδίων, οι οποίες είναι σχετικές. Η μέθοδος κατά συστάδες βρίσκει συστάδες, ώστε να μεγιστοποιείται η ομοιότητα των τιμών μέσα σε αυτές και η ανομοιότητα μεταξύ αυτών. Βρίσκει έτσι πιο φυσικά πρότυπα από την καθολική μέθοδο. Η δεύτερη καθορίζει τις συστάδες ώστε να μεγιστοποιείται η στατιστική εμπιστοσύνη. Βρίσκει έτσι πιο ισχυρές και αξιόπιστες σχέσεις. Και οι δύο μέθοδοι βρίσκουν δυνατές σχέσεις μεταξύ της εξαρτημένης μεταβλητής και των υπολοίπων μεταβλητών. Η καθολική μέθοδος, χρησιμοποιώντας τις ιδιότητες διάσπασης του υπολογιστή, τις πιο έγκυρες στατιστικά σχέσεις αυτών, ενώ η πρώτη μέθοδος, προσομοιώνοντας τη χειροκίνητη διαδικασία που θα έκανε ένας αναλυτής για να ομαδοποιήσει τις τιμές των πεδίων, παράγει τις πιο περιεκτικές σχέσεις αυτών.

Η μέθοδος κατά συστάδες, όταν εφαρμόζεται σε κατηγορική εξαρτημένη μεταβλητή αναφέρεται ως τεχνική CHAID (Chi-squared Automatic Interaction Detection) [28].Όταν η εξαρτημένη μεταβλητή είναι συνεχής η μέθοδος κατά συστάδες είναι όμοια με την τεχνική CART (Classification and Regression Trees) [5]. Η καθολική μέθοδος απευθύνεται κυρίως σε αδυναμίες της προηγούμενης μεθόδου [4]. Στη στατιστική ορολογία το μειονέκτημα της πρώτης μεθόδου είναι ότι είναι σε μεγάλο βαθμό συντηρητική (conservative), αφού μπορεί να παρουσιάσει υψηλό σφάλμα τύπου 1. Στη δεύτερη μέθοδο δε συμβαίνει το ίδιο. Έρευνες του Monte Carlo επιβεβαιώνουν ότι η καθολική μέθοδος δεν είναι καθόλου συντηρητική και δεν παρουσιάζει ψευδές σχέσεις, δηλαδή ελέγχει τα σφάλματα τύπου 1 και 2.

Αν ενδιαφερόμαστε για μία εκτενή και λεπτομερή ανάλυση μπορεί να χρησιμοποιηθεί μία διαδικασία σε δύο φάσεις, ώστε να χρησιμοποιούνται και οι δύο μέθοδοι. Στην πρώτη φάση η καθολική μέθοδος βρίσκει αξιόπιστες σχέσεις, οι οποίες μπορούν να βελτιωθούν στη δεύτερη φάση με τη μέθοδο κατά συστάδες. Αν όμως θέλουμε να κάνουμε μία γρήγορη ανάλυση των δεδομένων μας, η μέθοδος κατά συστάδες είναι επαρκής για τις περισσότερες εφαρμογές.

3.6.2 ΚΡΙΤΗΡΙΟ ΑΝΑΖΗΤΗΣΗΣ ΤΗΣ ΔΙΑΣΠΑΣΗΣ ΤΩΝ ΔΕΝΤΡΩΝ

Το Knowledge Seeker χρησιμοποιεί ως κριτήριο για την επιλογή της διάσπασης το ρυθμιζόμενο επίπεδο εμπιστοσύνης. Αυτή η μέθοδος είναι μια στατιστική προσέγγιση, αφού μειώνει την πιθανότητα παραγωγής τυχαίων αποτελεσμάτων και βάζει όλα τα πεδία σε μια κοινή στατιστική βάση. Μη διατεταγμένα, ομαδοποιημένα πεδία και πεδία με πολλές τιμές

έχουν μία μεροληπτική πιθανότητα να επιλεχθούν ως κλαδιά ,εκτός αν χρησιμοποιηθούν ρυθμίσεις στατιστικών ελέγχων. Έτσι, επιβεβαιώνεται η ακεραιότητα των δέντρων ταξινόμησης που μας παρέχει το Knowledge Seeker.

Ένα δείγμα δεδομένων είναι απλά μια εικόνα των απόψεων του πληθυσμού, από το οποίο πήραμε το δείγμα. Οι κλασικοί έλεγχοι εμπιστοσύνης είναι σχεδιασμένοι να δηλώνουν το βαθμό στον οποίο τα αποτελέσματα του δείγματος αντιπροσωπεύουν τον πληθυσμό. Αυτοί οι έλεγχοι εμπιστοσύνης υποθέτουν ότι υπάρχει ένας έλεγχος ,ένα δείγμα και ένας πληθυσμός. Βασιζόμενος κανείς σε αυτές τις υποθέσεις και με όριο λάθους 5%, θα περίμενε να είναι λάθος πέντε φορές στις εκατό. Υπάρχει λοιπόν ένας αριθμός ρυθμίσεων που μπορούν να γίνουν για να ρυθμιστεί το επίπεδο ορίου για έναν έλεγχο εμπιστοσύνης. Το knowledge seeker χρησιμοποιεί τέτοιες ρυθμίσεις και αναφέρονται ως ρυθμίσεις Bonferroni.

3.6.3 ΦΙΛΤΡΑ ΩΣ ΟΡΙΑ (FILTER THRESHOLD)

Όταν γίνεται μία οποιαδήποτε ανάλυση δεδομένων, υπάρχει ο κίνδυνος να πάρουμε κάποια στοιχεία κατά τύχη, με αποτέλεσμα να πάρουμε ενδεχομένως παραπλανητικά αποτελέσματα. Για να μειωθεί αυτή η πιθανότητα των κατά τύχη στοιχείων, το Knowledge Seeker παρέχει τρία “είδη” φίλτρων τα οποία καθορίζουν πόσο αυστηρή μπορεί να γίνει η ανάλυση των δεδομένων. Με τα όρια που μπορούμε να θέσουμε, ρυθμίζουμε το βαθμό εμπιστοσύνης που χρησιμοποιείται στους ελέγχους για την επιθυμητή αξιοπιστία στις σχέσεις.

Η ρύθμιση της «πρόβλεψης» (prediction) έχει αρκετά καλή αξιοπιστία, της τάξης του 95%. Παρέχει εμπιστοσύνη ώστε τα αποτελέσματα που

παίρνουμε να είναι αξιόπιστα σε μία μεγάλη κλίμακα για τα συμπεράσματά μας. Είναι έτσι σχεδόν απίθανο να παρουσιάσουμε παραπλανητικά ή τυχαία αποτελέσματα. Όταν χρειαζόμαστε εξαιρετικά υψηλό βαθμό εμπιστοσύνης, χρησιμοποιούμε τη ρύθμιση της «απόφασης» (decision), που παρέχει εγκυρότητα σε κάθε μέρος του δέντρου αποφάσεων και η πιθανότητα για λανθασμένα αποτελέσματα είναι ένα προς εκατό. Η ρύθμιση της «εξερεύνησης» (exploration) δίνει τη δυνατότητα για μία γρήγορη εξέταση των δεδομένων, για να αναγνωριστούν σημαντικά αποτελέσματα που θα θέλαμε να εξετάσουμε περισσότερο, τα οποία όμως είναι αξιόπιστα με επίπεδο λάθους 20%.

3.6.3 ΡΥΘΜΙΣΕΙΣ BONFERRONI

Όπως έχουμε αναφέρει οι ρυθμίσεις Bonferroni είναι αυτές που καθορίζουν το επίπεδο εμπιστοσύνης όταν ελέγχεται η εγκυρότητα των σχέσεων. Αυτές οι ρυθμίσεις είναι όμοιες με αυτές στο κριτήριο αναζήτησης διασπάσεων του δέντρου, με τη διαφορά ότι απευθύνονται και σε ρυθμίσεις που εξισορροπούν τον αριθμό των περιττών πεδίων των δεδομένων. Δίνεται έτσι η δυνατότητα, η ανάλυση που βασίζεται σε κατηγοριοποιημένες ομάδες πεδίων να γίνεται με υψηλό επίπεδο εμπιστοσύνης. Με αυτές τις ρυθμίσεις μπορεί να διατηρηθεί η ποσοστιαία αναλογία λάθους (α) πρώτου βαθμού, δηλαδή η πιθανότητα να βρεθεί αξιόπιστη διάσπαση σε κάποιο κόμβο του δέντρου, ενώ στην πραγματικότητα δεν υπάρχει τέτοια σχέση μεταξύ των μεταβλητών, σε επιτρεπτά όρια, όπως 0.05%.

3.7 ΔΗΜΙΟΥΡΓΙΑ ΤΑΜΠΛΟ (GENERATE CROSSTABLE)

Με αυτή την εντολή ο χρήστης μπορεί να απεικονίσει τις πληροφορίες όλου ή μέρους του στατιστικού δέντρου αποφάσεων που έχει παράγει υπό τη μορφή πίνακα κειμένου. Στον πίνακα φαίνονται οι ιδιότητες των πεδίων (μεταβλητών) που διασπούν (split) τον αμέσως προηγούμενο κόμβο (για παράδειγμα φαίνεται εάν η μεταβλητή είναι συνεχής ή κατηγορική, εάν η ομαδοποίηση έχει γίνει με βάση τη διάταξη ή όχι, εάν η μεταβλητή έχει παραλειπόμενες τιμές, κ.α.). Επίσης παρουσιάζονται οι συχνότητες του αριθμού των παρατηρήσεων που εμφανίζονται σε κάθε κόμβο του δέντρου, παρουσιάζονται οι συστάδες , στις οποίες χωρίζεται η μεταβλητή, που στη συνέχεια θα παράγουν την επόμενη διάσπαση και τα επί τις εκατό ποσοστά των κόμβων που επιλέχτηκαν για τη δημιουργία του ταμπλό, τοποθετημένα σε γραμμές και στήλες. Ανά γραμμές παρουσιάζονται ποσοστά που αφορούν τον τρέχον κόμβο, ενώ ανά στήλες ποσοστά που αφορούν απογόνους του. Τέλος, καταγράφονται στατιστικά μεγέθη που υπολογίστηκαν για να προσδιορισθεί η διάσπαση, όπως το επίπεδο εμπιστοσύνης, η τιμή της στατιστικής συνάρτησης ελέγχου, οι βαθμοί ελευθερίας, κ.α.

3.8 ΔΗΜΙΟΥΡΓΙΑ ΚΑΝΟΝΩΝ (GENERATE RULES)

Το Knowledge Seeker μπορεί να μετατρέψει το παραγόμενο στατιστικό δέντρο αποφάσεων, ή μέρος αυτού, σε μια ακολουθία κανόνων (rules) ή δηλώσεων της SQL. Ένα τέτοιο αποτέλεσμα – έξοδος (output) μπορεί να χρησιμοποιηθεί στη δημιουργία συστημάτων διεξαγωγής πληροφοριών από ένα σύνολο δεδομένων με βάση μια σειρά κανόνων.

Το Knowledge Seeker μπορεί να δημιουργήσει κανόνες γενικής ή αναλυτικής μορφής, για το SPSS, για το SAS, τη Visual Basic και την Prolog. Σε αυτή την περίπτωση το δέντρο μετατρέπεται σε μια σειρά IF…THEN δηλώσεων ή άλλων εντολών γλώσσας προγραμματισμού, που εκφράζουν τις διασπάσεις και γενικότερα τις πληροφορίες του δέντρου. Η έξοδος στέλνεται σε ένα κειμενογράφο, όπου εκεί ο χρήστης έχει τη δυνατότητα να τροποποιήσει και να αποθηκεύσει το αρχείο.

Το Knowledge Seeker μπορεί να δημιουργήσει κανόνες υπό τη μορφή SQL δηλώσεων. Οι κανόνες αυτοί επιλέγουν εγγραφές (παρατηρήσεις) που αφορούν ένα συγκεκριμένο – ιδιαίτερο γκρουπ (κόμβο) από ένα μεγαλύτερο σύνολο δεδομένων. Για παράδειγμα, φανταστείτε ότι αναλύεται η επιτυχημένη αποστολή συγκεκριμένου μηνύματος μέσο ηλεκτρονικού ταχυδρομείου σε 1000 ανθρώπους. Τότε ανακαλύπτεται ένα γκρουπ ανθρώπων που πιθανώς το έλαβαν. Οι SQL κανόνες μπορούν να δημιουργήσουν μια ερώτηση (query) που θα επιλέξει ανθρώπους από μια διαφορετική βάση δεδομένων που θα είχαν την ίδια πιθανότητα λήψης αυτού του μηνύματος.

3.9 ΕΥΡΕΣΗ ΣΥΝΕΙΣΦΟΡΑΣ (LEVERAGE) ΚΑΙ ΠΙΝΑΚΑΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ (GAINS CHART)

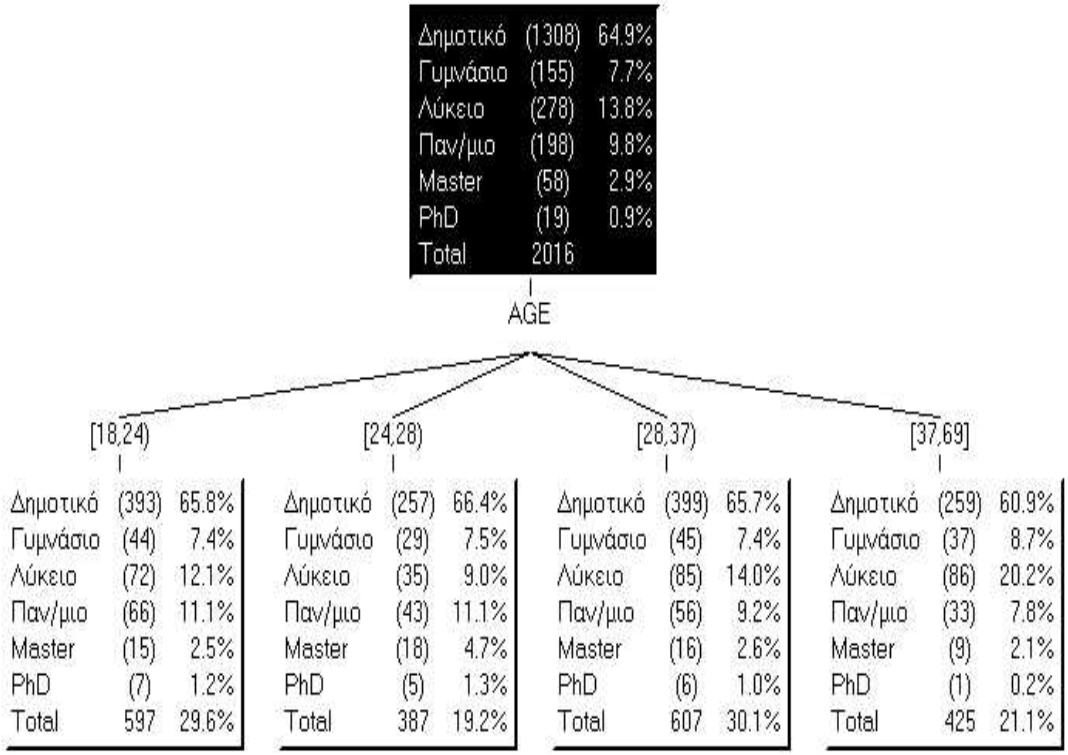
Η εντολή "leverage" υπολογίζει τη σχετική συνεισφορά συγκεκριμένων τιμών σε μια κατηγορική εξαρτημένη μεταβλητή (η διεργασία αυτή δεν εφαρμόζεται σε συνεχή εξαρτημένη μεταβλητή) σε σχέση με το συνολικό αριθμό εμφάνισης αυτών των τιμών στο σύνολο δεδομένων μας. Η λειτουργία αυτή βασίζεται στην Αρχή Βελτιστοποίησης του Pareto. Με βάση αυτή την αρχή, το 80% της επίδρασης μιας συγκεκριμένης τιμής παράγεται από το 20% της εμφάνισής της στο σύνολο των δεδομένων. Αναλογικά, μπορούμε να

βρούμε έναν ή περισσότερους κόμβους (κατά τη διάσπαση μίας μεταβλητής σε κατηγορίες δημιουργούνται περισσότεροι από ένα κόμβο) στο δέντρο αποφάσεων που να περιέχουν το 20% των συνολικών παρατηρήσεων των δεδομένων, το οποίο συνεισφέρει κατά 80% στην επίδραση των τιμών αυτών. Τέτοιου είδους διαδικασίες είναι πολύ χρήσιμες σε μια μεγάλη γκάμα εφαρμογών, όπως από την αναγνώριση των χαρακτηριστικών του καλύτερου πελάτη μιας επιχείρησης έως την αναγνώριση των χαρακτηριστικών ανθρώπων που ανήκουν σε ομάδα ύψιστου κινδύνου υγείας.

Για την καλύτερη κατανόηση, θα χρησιμοποιήσουμε το σύνολο δεδομένων μας, με τρόπο τέτοιο που θα μας επιτρέψει να παρουσιάσουμε τη χρησιμότητα της διεργασίας αυτής. Ας υποθέσουμε λοιπόν ότι έχουμε ορίσει ως εξαρτημένη μεταβλητή την κατηγορική μεταβλητή ***Educ***. Αυτή έχει κατηγοριοποιηθεί, για τις ανάγκες του παραδείγματος, σε έξι κατηγορίες με τους εξής χαρακτηρισμούς:

Κατηγορία	*Χρόνια Εκπαίδευσης*	*Επίπεδο Εκπαίδευσης (Χαρακτηρισμός)*
1	0 – 6	Δημοτικό
2	8 – 10	Γυμνάσιο
3	11 – 13	Λύκειο
4	14 – 17	Πανεπιστήμιο
5	18 – 19	Master
6	20 – 23	PhD

Στη συνέχεια, καθοδηγούμε το πρόγραμμα να βρει σχέσεις μεταξύ της *Educ* και της συνεχούς μεταβλητής *Age*, διασπώντας τη δεύτερη σε τέσσερις κατηγορίες. Έτσι έχουμε το ακόλουθο δέντρο αποφάσεων:



Ας επιλέξουμε την κατηγορία "Master" της εξαρτημένης μεταβλητής *Educ* για την περαιτέρω ανάλυσή μας. Χρησιμοποιώντας την εντολή εύρεσης συνεισφοράς "leverage" θα προσπαθήσουμε να αναγνωρίσουμε το 20% των παρατηρήσεων που συνεισφέρουν 80% στην απόκτηση μεταπτυχιακού τίτλου σπουδών Master. Με αυτό τον τρόπο επιδιώκουμε την εύρεση των χαρακτηριστικών των ανθρώπων που βρίσκονται σε επίπεδο επιστημονικής κατάρτισης Master.

Εφαρμόζοντας λοιπόν την εντολή "leverage" για την κατηγορία "Master" στο πιο πάνω δέντρο παίρνουμε τον ακόλουθο πίνακα αποτελεσμάτων:

Group	*Size of Group*	*Number of Responses*	*Cumulative Responses*	*Response Rate*	*Cumulative Rate*	*Cumulative Lift*
AGE [24,28)	387	18	18	4.7%	4.7%	162
AGE [28,37)	607	16	34	2.6%	3.4%	119
AGE [18,24)	597	15	49	2.5%	3.1%	107
AGE [37,69]	425	9	58	2.1%	2.9%	100
TOTAL	2016	58		2.9%		

Η στήλη "size of group" δείχνει τον αριθμό των συνολικών παρατηρήσεων κάθε τελικού κόμβου. Η στήλη "number of responses" δείχνει τον αριθμό των παρατηρήσεων, που αντιστοιχούν στην επιλεγείσα κατηγορία κατά την ανάλυση, κάθε κόμβου. Η στήλη "cumulative responses" δείχνει τον αθροιστικό αριθμό των παρατηρήσεων της επιλεγείσας κατηγορίας από κόμβο σε κόμβο. Η στήλη "response rate" δείχνει το επί τοις εκατό ποσοστό των παρατηρήσεων της επιλεγείσας κατηγορίας σε σχέση με το συνολικό αριθμό των παρατηρήσεων του κόμβου. Αυτά τα ποσοστά επί του συνόλου θα ισούνται με το ποσοστό της επιλεγείσας κατηγορίας σε όλες τις παρατηρήσεις της ανάλυσης. Η στήλη "cumulative rate" δείχνει το ποσοστό κατά μέσο όρο της επιλεγείσας κατηγορίας διαμέσου όλων των κόμβων. Ανταποκρίνεται στο ποσοστό του αθροιστικού αριθμού των παρατηρήσεων της επιλεγείσας κατηγορίας στο συνολικό αριθμό παρατηρήσεων της ανάλυσης. Η στήλη "cumulative lift" δείχνει την αναλογία των κατηγοριών καθώς κινούμαστε στο δέντρο από πάνω προς

τα κάτω, από κόμβο σε κόμβο. Υπολογίζεται σε σχέση με το συνολικό επί τοις εκατό ποσοστό που αντιπροσωπεύει η επιλεγείσα κατηγορία σε ολόκληρο το σύνολο δεδομένων.

Αναλύοντας τον πίνακα αποτελεσμάτων που προέκυψε, παρατηρούμε ότι η ομάδα δεδομένων που σχετίζεται περισσότερο με την κατηγορία "Master" της εξαρτημένης μεταβλητής ***Educ*** είναι άτομα ηλικίας 24-28 ετών (πληροφορίες για την οποία έχουμε στην πρώτη γραμμή του πίνακα). Στην ομάδα αυτή ανήκουν 18 παρατηρήσεις εκ των 58 συνολικών παρατηρήσεων που ανήκουν στην κατηγορία (Master) και στις τέσσερις ομάδες ηλικίας. Από τις 387 συνολικές παρατηρήσεις της ομάδας ανθρώπων ηλικίας 24-28, 18 (ή το 4.7%) ανήκουν στην κατηγορία "Master". Η σχετική συνεισφορά αυτού του κόμβου στην κατηγορία "Master" σε σύγκριση με τον συνολικό αριθμό υποθέσεων που λαμβάνονται υπόψη σε όλες τις παρατηρήσεις του κόμβου, παρουσιάζεται στην τελευταία στήλη του πίνακα, που εδώ έχει τιμή 162%. Αυτή είναι η αναλογία του επί τοις εκατό ποσοστού της κατηγορίας "Master" (18 από τις 58 παρατηρήσεις ή περίπου το 4.7%) προς το εκατοστιαίο ποσοστό του αριθμού των παρατηρήσεων της ομάδας ανθρώπων ηλικίας 24-28 που αντιπροσωπεύει το συνολικό αριθμό παρατηρήσεων σε όλα τα δεδομένα (387 από τις 2016 παρατηρήσεις ή περίπου το 2.9%). Αυτό γεννά την αναλογία $\frac{18/58}{387/2016} \cong 1.62$ ή ένα "lift" 162%.

Εύκολα διαπιστώνουμε ότι το συγκεκριμένο παράδειγμα ξεφεύγει του 80/20 κανόνα του Pareto. Και αυτό γιατί το 4.7% (ποσοστό ανθρώπων ηλικίας 24-28 που κατέχουν μεταπτυχιακό τίτλο σπουδών Master) προέρχεται από το 2.9% των παρατηρήσεων. Μοιάζει λοιπόν περισσότερο με κανόνα 50/25. Η αναφορά του όμως σε αυτό το σημείο

της παρούσας εργασίας είναι σκόπιμη, μια και βοηθάει στην παρουσίαση της διεργασίας εύρεσης συνεισφοράς.

3.10 ΣΥΝΘΗΚΕΣ ΒΑΡΟΥΣ

Το Knowledge Seeker χρησιμοποιεί συνθήκες βάρους, ως ένα βήμα στην προεργασία των δεδομένων για να «ζυγίσει» υποθέσεις ή στατιστικά αποτελέσματα με δύο τρόπους:

Με δειγματοληψία βάρους (sampling weights). Η δειγματοληψία αυτή χρησιμοποιείται για να εξασφαλίσει ότι οι παρατηρήσεις του δείγματος συμπεριλαμβάνονται στην ανάλυση σε αναλογία τέτοια, που είναι σχετική με την κατανομή των παρατηρήσεων στον πληθυσμό από τον οποίο έχει επιλεγεί το δείγμα. Για παράδειγμα αν ο αριθμός των αντρών υπερκαλύπτει το δείγμα σε σχέση με την πραγματική κατανομή αντρών και γυναικών στον πληθυσμό, οι συνθήκες βάρους μπορούν να χρησιμοποιηθούν για να προσαρμόσουν τη συμβολή κάθε γένους στην ανάλυση, ώστε τα αποτελέσματα να αντιπροσωπεύουν την αληθινή αναλογία. Τέτοια δείγματα είναι τα στρωματοποιημένα.

Με συχνότητες βάρους (frequency weights). Αυτές χρησιμοποιούνται ως μία συνθήκη βάρους για να ελεγχθεί πόσες φορές εμφανίζεται η δοσμένη υπόθεση – παρατήρηση στην ανάλυση. Μπορεί να χρησιμοποιηθεί όταν εμφανίζονται πολλές ξεχωριστές παρατηρήσεις, οι οποίες βρίσκονται στη βάση δεδομένων ως μία παρατήρηση. Τότε οι παρατηρήσεις που έχουν τις ίδιες τιμές, αντικαθίστανται από μία σταθμική παρατήρηση, η οποία αντιπροσωπεύει τις συχνότητες των παρατηρήσεων στα δεδομένα. Αυτές οι συχνότητες βάρους εμφανίζονται για παράδειγμα σε πειραματικές έρευνες, όπου ο αριθμός ίδιων συνθηκών έχουν συλλεχθεί ως επαναλαμβανόμενες παρατηρήσεις.

Οι συνθήκες, λοιπόν, βάρους που χρησιμοποιεί το Knowledge Seeker έχουν σαν αποτέλεσμα την αλλαγή των συχνοτήτων κάποιων παρατηρήσεων. Αυτό επηρεάζει και τον υπολογισμό των X^2 και F στατιστικών μεγεθών. Όταν υπολογίζεται X^2 τιμή ο αριθμός των παρατηρήσεων αντικαθίστανται από το άθροισμα των σταθμικών παρατηρήσεων. Η X^2 είναι κατανομή με (d-1)(k-1) βαθμούς ελευθερίας – οι βαθμοί ελευθερίας είναι ανεξάρτητοι από τις συνθήκες βάρους της εξαρτημένης μεταβλητής. Οι έλεγχοι που βασίζονται σε αυτή την τιμή όταν οι σταθμικές τιμές είναι αποτελέσματα στρωματοποιημένης δειγματοληψίας είναι αρκετά συντηρητικοί.

3.11 ΟΡΙΣΜΟΣ ΚΟΣΤΟΥΣ – ΜΕΓΙΣΤΟΠΟΙΗΣΗ ΚΕΡΔΟΥΣ

Το Knowledge Seeker μπορεί να αναθέσει κόστος ή κέρδος σε απαντήσεις που αντιστοιχούν σε κατηγορίες της εξαρτημένης μεταβλητής. Για παράδειγμα αν η εξαρτημένη μεταβλητή έχει τις απαντήσεις "αγοράζω" / "δεν αγοράζω", κωδικοποιημένες με τις τιμές 1 και 0 αντίστοιχα, μπορούμε να αναθέσουμε κέρδος στην τιμή 1 και κόστος στην τιμή 0. Το συνολικό κέρδος στο συγκεκριμένο κόμβο υπολογίζεται πολλαπλασιάζοντας το κέρδος με τον αριθμό των τιμών 1 και αφαιρώντας το κόστος πολλαπλασιασμένο με τον αριθμό των τιμών 0. Μπορεί έτσι να παράγεται στη ρίζα του δέντρου το άθροισμα των κατανομών συχνοτήτων των τιμών σε κάθε απάντηση – πεδίο. Μπορούμε τελικά να μετατρέψουμε το χάσιμο σε κέρδος, αρκεί να μπορούμε να αναγνωρίσουμε το τμήμα αγοράς που έχει μπει σε στόχο και το οποίο έχει ένα σχετικά υψηλό ποσοστό σε σχέση με το κέρδος των απαντήσεων.

Ας υποθέσουμε για παράδειγμα ότι έχουμε τις κωδικοποιημένες απαντήσεις τριών τιμών, απάντηση: πληρωμή, απάντηση: όχι πληρωμή,

καμία απάντηση. Η πρώτη τιμή δίνει έσοδα $15, η δεύτερη κοστίζει $15 και η τρίτη κοστίζει $5. Τα συνολικά έσοδα / έξοδα για κάποια ανάλυση μπορούν να προσδιοριστούν από τα αποτελέσματα των πεδίων των απαντήσεων. Επιλέγοντας το πεδίο που μας ενδιαφέρει ως εξαρτημένη μεταβλητή υπολογίζεται αυτόματα το κόστος / κέρδος που μας ενδιαφέρει και παρουσιάζεται σε σχέση με την κατανομή των παρατηρήσεων.

ΚΕΦΑΛΑΙΟ 4
ΧΡΗΣΗ WEKA

Στα πλαίσια της εργασίας μας χρησιμοποιήσαμε αλγόριθμους εξόρυξης γνώσης για την έγγριση πιστωτικών καρτών βασιζόμενη σε στοιχεία όπως ηλικία, εισόδημα, πιστωτική ιστορία και ιδιοκτησία κατοικίας κλπ. Αξιολογίσαμε αρκετούς αλγόριθμους εξόρυξης γνώσης και επιλέξαμε τον καλύτερο, χρησιμοποιώντας το εργαλείο εξόρυξης δεδομένων WEKA.

4.1 ΠΕΡΙΠΤΩΣΗ: CREDIT-A

Η κάθε περίπτωση των δεδομένων μας αντιπροσωπεύει μια αίτηση για εγκαταστάσεις πιστωτικών καρτών που περιγράφονται από οκτώ «φρόνημα» και έξι συνεχή χαρακτηριστικά , με δύο περιπτώσεις απόφασης (Αποδέχομαι/ Απορρίπτω). Στο UCI Repository τα αρχικά ονόματα των χαρακτηριστικών έχουν αλλαχθεί σε σύμβολα χωρίς νόημα (A1 – A14) με την αιτιολογία ότι προστατεύουν το απόρρητο των δεδομένων. Ωστόσο τα πραγματικά ονόματα των χαρακτηριστικών είναι διαθέσιμα στη σελίδα της Rulequest Research (http:/www.rulequest.com/see5-examples.html).

Τα χαρακτηριστικά των βάσεων δεδομένων φαίνονται στον πίνακα 4.1. παρακάτω. Τα αρχικά ονόματα των χαρακτηριστικών (που μας παρέχονται από τη σελίδα της Rulequest Research) δίνονται στις παρενθέσεις. Η στήλη πεδίο δείχνει την σειρά ή την κατηγορία των πιθανών αξιών για κάθε χαρακτηριστικό. Στη στήλη τύπος κάνουμε μια διάκριση μεταξύ κατηγορικών και αριθμητικών μεταβλητών.

Χαρακτηριστικό	Πεδίο	Τύπος
Α1 (φύλο)	0, 1	Αριθμητικός
Α2(ηλικία)	13,75-80,25	Συνεχής
Α3(Παρούσα διεύθυνση)	0-28	Συνεχής
Α4(Οικογενειακή κατάσταση)	1,2,3	Αριθμητικός
Α5(Παρούσα απασχόληση)	1-14	Αριθμητικός
Α6(Παρούσα εργασιακή κατάσταση)	1-9	Αριθμητικός
Α7(Φορέας απασχόλησης)	0-28,5	Συνεχής
Α8(Άλλες επενδύσεις)	0, 1	Αριθμητικός
Α9(Τραπεζικός λογ/μός)	0, 1	Αριθμητικός
Α 10(Παρούσα τράπεζα)	0-67	Συνεχής
Α11(Αναφορά υπευθυνότητας)	0, 1	Αριθμητικός
Α12(Οικονομική αναφορά)	1,2,3	Αριθμητικός
Α13(Μηνιαία έξοδα κατοικίας)	0-2000	Συνεχής
Α14(Οικονομική ισορροπία αποταμιεύσεων)	1-100001	Συνεχής
Τάξη (Απορρίπτω/ Αποδέχομαι)	0, 1	Αριθμητικός

Πίνακας 4.1 Χαρακτηριστικά βάσεων δεδομένων

Εν συνεχεία παρουσιάζουμε τους παραγόμενους ταξινομητές και την ακρίβεια πρόγνωσης (τις σωστές προβλέψεις/το σύνολο των προβλέψεων) για κάθε εξεταζόμενο αλγόριθμο εξόρυξης γνώσης.

4.1.1. ΆΦΕΛΗΣ ΤΑΞΙΝΟΜΗΤΗΣ BAYES

=== Classifier model (full training set) ===

Naive Bayes Classifier

Class +: Prior probability = 0.45

A1: Discrete Estimator. Counts = 207 99 (Total = 306)

A2: Normal Distribution. Mean = 33.723 StandardDev = 12.7816 WeightSum = 305 Precision = 0.1910919540229885

A3: Normal Distribution. Mean = 5.9075 StandardDev = 5.4649 WeightSum = 307 Precision = 0.1308411214953271

A4: Discrete Estimator. Counts = 257 46 3 1 (Total = 307)

A5: Discrete Estimator. Counts = 257 46 3 (Total = 306)

A6: Discrete Estimator. Counts = 63 8 30 15 4 15 17 3 52 34 33 15 20 8 (Total = 317)

A7: Discrete Estimator. Counts = 170 88 26 4 3 7 3 9 2 (Total = 312)

A8: Normal Distribution. Mean = 3.4242 StandardDev = 4.1179 WeightSum = 307 Precision = 0.21755725190839695

A9: Discrete Estimator. Counts = 285 24 (Total = 309)

A10: Discrete Estimator. Counts = 210 99 (Total = 309)

A11: Normal Distribution. Mean = 4.6823 StandardDev = 6.5274 WeightSum = 307 Precision = 3.0454545454545454

A12: Discrete Estimator. Counts = 147 162 (Total = 309)

A13: Discrete Estimator. Counts = 288 6 16 (Total = 310)

A14: Normal Distribution. Mean = 164.1864 StandardDev = 161.3686 WeightSum = 301 Precision = 11.834319526627219

A15: Normal Distribution. Mean = 2027.9939 StandardDev = 7655.808 WeightSum = 307 Precision = 418.41004184100416

Class -: Prior probability = 0.55

A1: Discrete Estimator. Counts = 263 113 (Total = 376)

A2: Normal Distribution. Mean = 29.8068 StandardDev = 10.9057 WeightSum = 373 Precision = 0.1910919540229885

A3: Normal Distribution. Mean = 3.8409 StandardDev = 4.3316 WeightSum = 383 Precision = 0.1308411214953271

A4: Discrete Estimator. Counts = 264 119 1 1 (Total = 385)

A5: Discrete Estimator. Counts = 264 119 1 (Total = 384)

A6: Discrete Estimator. Counts = 76 24 13 46 8 38 23 2 28 32 7 12 36 47 (Total = 392)

A7: Discrete Estimator. Counts = 231 52 35 6 3 3 5 50 2 (Total = 387)

A8: Normal Distribution. Mean = 1.2525 StandardDev = 2.1128 WeightSum = 383 Precision = 0.21755725190839695

A9: Discrete Estimator. Counts = 78 307 (Total = 385)

A10: Discrete Estimator. Counts = 87 298 (Total = 385)

A11: Normal Distribution. Mean = 0.6043 StandardDev = 1.9863 WeightSum = 383 Precision = 3.0454545454545454

A12: Discrete Estimator. Counts = 171 214 (Total = 385)

A13: Discrete Estimator. Counts = 339 4 43 (Total = 386)

A14: Normal Distribution. Mean = 199.7986 StandardDev = 181.3128 WeightSum = 376 Precision = 11.834319526627219

A15: Normal Distribution. Mean = 186.8097 StandardDev = 675.6316 WeightSum = 383 Precision = 418.41004184100416

Time taken to build model: 0.72 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	536	77.6812 %
Incorrectly Classified Instances	154	22.3188 %
Kappa statistic	0.534	
Mean absolute error	0.2228	
Root mean squared error	0.4356	
Relative absolute error	45.0979 %	
Root relative squared error	87.6429 %	
Total Number of Instances	690	

‘Αρα η ακρίβεια του Naïve Bayes στα δεδομένα ήταν 77.68%.

4.1.2 ΔΕΝΤΡΟ ΑΠΟΦΑΣΗΣ C4.5

=== Classifier model (full training set) ===

C4.5

A9 = t

| A10 = t: + (228.0/21.0)

| A10 = f

| | A15 <= 444

| | | A7 = v

| | | | A4 = u

| | | | | A14 <= 112: + (16.57/1.57)
| | | | | A14 > 112
| | | | | | A15 <= 70: - (30.0/10.0)
| | | | | | A15 > 70: + (2.0)
| | | | A4 = y
| | | | | A13 = g: - (12.0/2.0)
| | | | | A13 = p: - (0.0)
| | | | | A13 = s: + (3.0/1.0)
| | | | A4 = l: - (0.0)
| | | | A4 = t: - (0.0)
| | | A7 = h: + (27.24/8.24)
| | | A7 = bb
| | | | A3 <= 1.375: + (5.0/1.0)
| | | | A3 > 1.375: - (9.13/1.0)
| | | A7 = j: - (1.01)
| | | A7 = n: + (0.0)
| | | A7 = z: + (0.0)
| | | A7 = dd: + (1.01/0.01)
| | | A7 = ff: - (5.05/1.0)
| | | A7 = o: + (0.0)
| | A15 > 444: + (21.0/1.0)
A9 = f
| A3 <= 0.165
| | A7 = v
| | | A2 <= 35.58: - (18.72/3.44)
| | | A2 > 35.58: + (3.6/0.16)
| | A7 = h: - (0.0)
| | A7 = bb: + (1.24/0.08)
| | A7 = j: + (1.24/0.08)

| | A7 = n: + (1.24/0.08)

| | A7 = z: - (0.0)

| | A7 = dd: - (0.0)

| | A7 = ff: - (4.96/0.64)

| | A7 = o: - (0.0)

| A3 > 0.165: - (298.0/12.0)

Number of Leaves : 30

Size of the tree : 42

Time taken to build model: 0.16 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	594	86.087 %
Incorrectly Classified Instances	96	13.913 %
Kappa statistic	0.718	
Mean absolute error	0.1924	
Root mean squared error	0.3313	
Relative absolute error	38.9417 %	
Root relative squared error	66.6637 %	
Total Number of Instances	690	

Άρα η ακρίβεια του C4.5 στα δεδομένα ήταν 86.09%.

4.1.3. ΚΑΝΟΝΕΣ ΑΠΟΦΑΣΕΩΝ ΑΠΟ ΤΟΝ ΑΛΓΟΡΙΘΜΟ RIPPER

RIPPER rules:

==========

(A9 = t) and (A15 >= 234) => class=+ (157.0/7.0)

(A9 = t) and (A10 = t) => class=+ (99.0/18.0)

(A9 = t) and (A14 <= 110) and (A15 <= 0) => class=+ (31.0/5.0)

=> class=- (403.0/50.0)

Number of Rules : 4

Time taken to build model: 0.19 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	592	85.7971 %
Incorrectly Classified Instances	98	14.2029 %
Kappa statistic	0.7136	
Mean absolute error	0.2145	
Root mean squared error	0.345	
Relative absolute error	43.4166 %	
Root relative squared error	69.4163 %	
Total Number of Instances	690	

‘Αρα η ακρίβεια του RIPPER στα δεδομένα ήταν 85.79%. Πρέπει επίσης ότι αυτή η ακρίβεια επιτεύχθηκε χρησιμοποιώντας τέσσερες μόνο κανόνες.

4.1.4. ΑΛΓΟΡΙΘΜΟΣ SMO

BinarySMO

Machine linear: showing attribute weights, not support vectors.

```
      0.0002 * (normalized) A1
+     -0.0042 * (normalized) A2
+      0.001  * (normalized) A3
+      0.3316 * (normalized) A4=u
+      0.3325 * (normalized) A4=y
+     -0.664  * (normalized) A4=l
+      0.3316 * (normalized) A5=g
+      0.3325 * (normalized) A5=p
+     -0.664  * (normalized) A5=gg
+     -0.0033 * (normalized) A6=c
+      0      * (normalized) A6=d
+     -0.0053 * (normalized) A6=cc
+     -0.0028 * (normalized) A6=i
+      0.0135 * (normalized) A6=j
+      0.001  * (normalized) A6=k
+     -0.0016 * (normalized) A6=m
+      0      * (normalized) A6=r
+     -0.0038 * (normalized) A6=q
+     -0.0038 * (normalized) A6=w
+     -0.0068 * (normalized) A6=x
+     -0.0051 * (normalized) A6=e
+     -0.0011 * (normalized) A6=aa
+      0.0191 * (normalized) A6=ff
+      0.0015 * (normalized) A7=v
```

+ 0.0004 * (normalized) A7=h
+ 0.0032 * (normalized) A7=bb
+ -0.0143 * (normalized) A7=j
+ 0.012 * (normalized) A7=z
+ 0.0019 * (normalized) A7=dd
+ -0.0063 * (normalized) A7=ff
+ 0.0015 * (normalized) A7=o
+ -0.0087 * (normalized) A8
+ 2.0008 * (normalized) A9
+ 0.0013 * (normalized) A10
+ -0.0255 * (normalized) A11
+ 0.0003 * (normalized) A12
+ 0.5003 * (normalized) A13=g
+ -1 * (normalized) A13=p
+ 0.4997 * (normalized) A13=s
+ 0.0195 * (normalized) A14
+ -0.0919 * (normalized) A15
- 2.162

Number of kernel evaluations: 146567

Time taken to build model: 3.09 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	586	84.9275 %
Incorrectly Classified Instances	104	15.0725 %
Kappa statistic	0.7003	
Mean absolute error	0.1507	
Root mean squared error	0.3882	
Relative absolute error	30.5133 %	
Root relative squared error	78.1202 %	
Total Number of Instances	690	

‘Αρα η ακρίβεια του SMO στα δεδομένα ήταν 84.93%.

4.1.5. ΑΛΓΟΡΙΘΜΟΣ BP ΓΙΑ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ

=== Classifier model (full training set) ===

Sigmoid Node 0

Inputs Weights

Threshold 1.6104258980164028

Node 2 -4.443436741637387

Sigmoid Node 1

Inputs Weights

Threshold -1.6104258980164028

Node 2 4.4434367416373854

Sigmoid Node 2

Inputs Weights

Threshold 0.770654936922311

Attrib A1 2.4686158578559754

Attrib A2 0.8708952741587124

Attrib A3 -0.7549003156668521

Attrib A4=u -1.252814575011715

Attrib A4=y 1.2731626928266144

Attrib A4=l -0.8256421775758617

Attrib A4=t -0.018649069530998742

Attrib A5=g -1.2232136702861873

Attrib A5=p 1.2793061063753015

Attrib A5=gg -0.8028974465710356

Attrib A6=c -2.1191703452723525

Attrib A6=d 1.1120027067707585

Attrib A6=cc -2.6843318698573273

Attrib A6=i -1.4713722840422685

Attrib A6=j -1.6563898371098598

Attrib A6=k 5.520794265236609

Attrib A6=m 0.1863790932329909

Attrib A6=r -0.8054759846117279

Attrib A6=q 2.164039086823643

Attrib A6=w 0.5658601329308847

Attrib A6=x -2.514722554169417

Attrib A6=e -10.62883677481209

Attrib A6=aa 3.134327320776049

Attrib A6=ff 2.0765942952825274

Attrib A7=v -4.252076488931497

Attrib A7=h -3.9745622887954943

Attrib A7=bb 4.422678836776461

Attrib A7=j -1.6640772168090898

Attrib A7=n -0.5309832459227579

Attrib A7=z -1.1535365212607271

Attrib A7=dd 1.528578732692125

Attrib A7=ff 2.030109631120246

Attrib A7=o -0.5576360965635871

Attrib A8 -2.940881884083295

Attrib A9 19.84931142649782

Attrib A10 7.595041544872093

Attrib A11 -2.2558404973487423

Attrib A12 2.0871426254550425

Attrib A13=g -1.5003977682739393

Attrib A13=p -0.7431097370472662

Attrib A13=s 1.5371222088707703

Attrib A14 5.939779629008369

Attrib A15 -1.5960341690201199

Class +

Input

Node 0

Class -

Input

Node 1

Time taken to build model: 4.39 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	593	85.942 %
Incorrectly Classified Instances	97	14.058 %
Kappa statistic	0.7157	
Mean absolute error	0.205	
Root mean squared error	0.3436	
Relative absolute error	41.5017 %	
Root relative squared error	69.147 %	
Total Number of Instances	690	

‘Αρα η ακρίβεια του BP στα δεδομένα ήταν 85.94%.

4.2 ΠΕΡΙΠΤΩΣΗ: GERMAN CREDIT

Η German Credit σειρά δεδομένων (διαθέσιμο στο ftp.ics.uci.edu/pub/mashine-learning-databesees/statlog/) περιέχει παρατηρήσεις σε 30-μεταβλητές-για 1000 αιτώντες στο παρελθόν για πίστωση. Κάθε αιτών έχει χαρακτηριστεί ως «καλός πιστωτής» (700 Περιπτώσεις) ή «κακός πιστωτής» (300 περιπτώσεις)

Νέοι αιτούντες για πίστωση μπορούν επίσης να εξεταστούν σ’ αυτές τις 30 μεταβλητές που προβλέπουν. Θέλουμε να αναπτύξουμε ένα επιτυχημένο κανόνα πίστωσης που μπορεί να χρησιμοποιηθεί για να καθορίσει αν ένας νέος αιτώντας είναι ένας καλός πιστωτικός κίνδυνος ή κακός πιστωτικός κίνδυνος, βασιζόμενοι σε αξίες για μία ή περισσότερες από τις μεταβλητές που προβλέπουν. Όλες οι μεταβλητές εξηγούνται στον πίνακα 4.2.

Α/Α	Όνομ. Μεταβλητής	Περιγραφή	Τύπος Μεταβλητής	Κωδ. Περιγραφής
1	OBS#	Αριθμός παρατήρησης	Κατηγορηματικός	Αλληλουχία αριθμών στη σειρά δεδομένων
2	CHK_ACCT	Κατάσταση ελέγχου λογ/μού	Κατηγορηματικός	0:<0 DM 1:0<=…<DM 2:=>200DM 3: μη ελενχόμενος λογ/μος
3	DURATION	Διάρκεια πίστωσης σε μήνες	Αριθμητικός	
4	HISTORY	Ιστορικό πίστωσης	Κατηγορηματικός	0: δεν έχουν παρθεί πιστώσεις 2:όλες οι πιστώσεις σ'αυτή την τράπεζα έχουν πληρωθεί 3:καθυστερήσει στην εξόφληση στο παρελθόν 4:κρίσιμος λογ/μος
5	NEW_CAR	Αιτία πίστωσης	Δυαδικός	Αυτοκίνητο(νέο):0:όχι, 1:ναι
6	USED_CAR	Αιτία πίστωσης	Δυαδικός	Αυτοκίνητο(παλιό): 0:όχι, 1:ναι
7	FURNITURE	Αιτία πίστωσης	Δυαδικός	Επίπλωση/ εξοπλησμός: 0:όχι, 1:ναι
8	RADIO/TV	Αιτία πίστωσης	Δυαδικός	Ράδιο/τηλεόραση: 0:όχι, 1:ναι
9	EDUCATION	Αιτία πίστωσης	Δυαδικός	Μόρφωση: 0:όχι, 1:ναι
10	RETRAINING	Αιτία πίστωσης	Δυαδικός	Μετεκπαίδευση: 0:όχι, 1:ναι
11	AMOUNT	Ποσό πίστωσης	Αριθμητικός	

12	SAV_ACCT	Μέση ισορροπία σε αποταμιευτικούς λογ/μους	Κατηγορημα-τικός	0:<100DM 1:100<=…<500 DM 2:500<=…<1000DM 3:=> 1000 DM 4: άγνωστος/μη αποθεματικός λογ/μος
13	EMPLOYMENT	Παρούσα εργασία	Κατηγορημα-τικός	0: άνεργος 1:< 1 χρόνο 2:1<=…<4 χρόνο 3:4<=…<7 χρόνια 4: >= 7 χρόνια
14	INSTALL_RATE	Ποσοστό δόσης επί %στο διατιθέμενο εισόδημα	Αριθμητικός	0:όχι, 1:ναι
15	MALE_DIV	Ο αιτών είναι άνδρας και διαζευγμένος	Δυαδικός	0:όχι, 1:ναι
16	MALE_SINGLE	Ο αιτών είναι άνδρας και ανύπαντρος	Δυαδικός	0:όχι, 1:ναι
17	MALE_MAR_WID	Ο αιτών είναι άνδρας και παντρεμένος ή χήρος	Δυαδικός	0:όχι, 1:ναι
18	CO-APPLICANT	Η αίτηση έχει και δεύτερο αιτούντα	Δυαδικός	0:όχι, 1:ναι
19	GUARANTOR	Ο αιτών έχει εγγυητή	Δυαδικός	0:όχι, 1:ναι
20	PRESENT_RESIDENT	Παρούσα κατοικία από πότε	Κατηγορημα-τικός	0: <=1 χρόνο 1: <…<= 2 χρόνια 2: <…<= 3 χρόνια 3: > 4 χρόνια

21	REAL ESTATE	Ο αιτών κατέχει ακίνητο	Δυαδικός	0:όχι, 1:ναι
22	PROP_UNKN_ NONE	Ο αιτών δεν έχει ιδιοκτησία(ή είναι άγνωστη)	Δυαδικός	0:όχι, 1:ναι
23	AGE	Ηλικία σε χρόνια	Αριθμητικός	0:όχι, 1:ναι
24	OTHER_ INSTALL	Ο αιτών έχει και άλλα πιστωτικά σχέδια	Δυαδικός	0:όχι, 1:ναι
25	RENT	Ο αιτών νοικιάζει	Δυαδικός	0:όχι, 1:ναι
26	OWN_RES	Ο αιτών ιδιοκατοικεί	Δυαδικός	0:όχι, 1:ναι
27	NUM_CREDITS	Αριθμός από υπάρχουσες πιστώσεις στην τράπεζα	Αριθμητικός	
28	JOB	Φύση της δουλείας	Κατηγορημα-τικός	0:άνεργος/ ανειδίκευτος/ όχι μόνιμος 1: ανειδίκευτος/ μόνιμος 2:ειδικευμένος/ μόνιμος 3:στέλεχος/ ελευθ. Επαγγελματίας/ υψηλών προσόντων εργαζόμενος/ αξιωματούχος
29	NUM_ DEPENDENTS	Αριθμός ατόμων από τα οποία εξαρτάται η μονιμοποίηση	Αριθμητικός	
30	TELEPHONE	Ο αιτών έχει αριθμό	Δυαδικός	0:όχι, 1:ναι

		τηλεφώνου στο όνομά του/ της		
31	FOREIGN	Αλλοδαπός εργοδότης	Δυαδικός	0:όχι, 1:ναι
32	RESPONSE	Το ποσοστό πίστωσης είναι καλό	Δυαδικός	0:όχι, 1:ναι

Πίνακας 4.2 Μεταβλητές που καθορίζουν αν ένας νέος αιτώντας είναι ένας καλός πιστωτικός κίνδυνος ή κακός πιστωτικός κίνδυνος.

Εν συνεχεία παρουσιάζουμε τους παραγόμενους ταξινομητές και την ακρίβεια πρόγνωσης για κάθε εξεταζόμενο αλγόριθμο εξόρυξης γνώσης.

4.2.1. ΆΦΕΛΗΣ ΤΑΞΙΝΟΜΗΤΗΣ BAYES

=== Classifier model (full training set) ===

Naive Bayes Classifier

Class good: Prior probability = 0.7

checking_status: Discrete Estimator. Counts = 140 165 50 349 (Total = 704)
duration: Normal Distribution. Mean = 19.1766 StandardDev = 10.9817 WeightSum = 700 Precision = 2.125
credit_history: Discrete Estimator. Counts = 16 22 362 61 244 (Total = 705)
purpose: Discrete Estimator. Counts = 146 87 124 219 9 15 29 1 9 64 8 (Total = 711)
credit_amount: Normal Distribution. Mean = 2985.6721 StandardDev = 2399.7801 WeightSum = 700 Precision = 19.754347826086956

savings_status: Discrete Estimator. Counts = 387 70 53 43 152 (Total = 705)

employment: Discrete Estimator. Counts = 40 103 236 136 190 (Total = 705)

installment_commitment: Normal Distribution. Mean = 2.92 StandardDev = 1.1273 WeightSum = 700 Precision = 1.0

personal_status: Discrete Estimator. Counts = 31 202 403 68 1 (Total = 705)

other_parties: Discrete Estimator. Counts = 636 24 43 (Total = 703)

residence_since: Normal Distribution. Mean = 2.8429 StandardDev = 1.1076 WeightSum = 700 Precision = 1.0

property_magnitude: Discrete Estimator. Counts = 223 162 231 88 (Total = 704)

age: Normal Distribution. Mean = 36.1723 StandardDev = 11.4005 WeightSum = 700 Precision = 1.0769230769230769

other_payment_plans: Discrete Estimator. Counts = 83 29 591 (Total = 703)

housing: Discrete Estimator. Counts = 110 528 65 (Total = 703)

existing_credits: Normal Distribution. Mean = 1.4243 StandardDev = 0.5843 WeightSum = 700 Precision = 1.0

job: Discrete Estimator. Counts = 16 145 445 98 (Total = 704)

num_dependents: Normal Distribution. Mean = 1.1557 StandardDev = 0.3626 WeightSum = 700 Precision = 1.0

own_telephone: Discrete Estimator. Counts = 410 292 (Total = 702)

foreign_worker: Discrete Estimator. Counts = 668 34 (Total = 702)

Class bad: Prior probability = 0.3

checking_status: Discrete Estimator. Counts = 136 106 15 47 (Total = 304)

duration: Normal Distribution. Mean = 24.8129 StandardDev = 13.3608 WeightSum = 300 Precision = 2.125

credit_history: Discrete Estimator. Counts = 26 29 170 29 51 (Total = 305)

purpose: Discrete Estimator. Counts = 90 18 59 63 5 9 23 1 2 35 6 (Total = 311)

credit_amount: Normal Distribution. Mean = 3938.1609 StandardDev = 3529.4788 WeightSum = 300 Precision = 19.754347826086956

savings_status: Discrete Estimator. Counts = 218 35 12 7 33 (Total = 305)

employment: Discrete Estimator. Counts = 24 71 105 40 65 (Total = 305)

installment_commitment: Normal Distribution. Mean = 3.0967 StandardDev = 1.0866 WeightSum = 300 Precision = 1.0

personal_status: Discrete Estimator. Counts = 21 110 147 26 1 (Total = 305)

other_parties: Discrete Estimator. Counts = 273 19 11 (Total = 303)

residence_since: Normal Distribution. Mean = 2.85 StandardDev = 1.0928 WeightSum = 300 Precision = 1.0

property_magnitude: Discrete Estimator. Counts = 61 72 103 68 (Total = 304)

age: Normal Distribution. Mean = 33.9267 StandardDev = 11.259 WeightSum = 300 Precision = 1.0769230769230769

other_payment_plans: Discrete Estimator. Counts = 58 20 225 (Total = 303)

housing: Discrete Estimator. Counts = 71 187 45 (Total = 303)

existing_credits: Normal Distribution. Mean = 1.3667 StandardDev = 0.5588 WeightSum = 300 Precision = 1.0

job: Discrete Estimator. Counts = 8 57 187 52 (Total = 304)

num_dependents: Normal Distribution. Mean = 1.1533 StandardDev = 0.3603 WeightSum = 300 Precision = 1.0

own_telephone: Discrete Estimator. Counts = 188 114 (Total = 302)

foreign_worker: Discrete Estimator. Counts = 297 5 (Total = 302)

Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	754	75.4 %
Incorrectly Classified Instances	246	24.6 %
Kappa statistic	0.3813	
Mean absolute error	0.2936	
Root mean squared error	0.4201	
Relative absolute error	69.8801 %	
Root relative squared error	91.6718 %	
Total Number of Instances	1000	

‘Αρα η ακρίβεια του NB στα δεδομένα ήταν 75.40%.

4.2.2. ΔΕΝΤΡΟ ΑΠΟΦΑΣΗΣ C4.5

=== Classifier model (full training set) ===

C4.5 pruned tree

---

checking_status = <0
| foreign_worker = yes
| | credit_history = no credits/all paid: bad (11.0/3.0)
| | credit_history = all paid: bad (9.0/1.0)
| | credit_history = existing paid
| | | other_parties = none
| | | | savings_status = <100
| | | | | job = unemp/unskilled non res: bad (1.0)

| | | | | job = unskilled resident: bad (18.0/8.0)
| | | | | job = skilled
| | | | | | own_telephone = none
| | | | | | | purpose = new car: bad (6.0/1.0)
| | | | | | | purpose = used car: good (1.0)
| | | | | | | purpose = furniture/equipment: good (16.0/8.0)
| | | | | | | purpose = radio/tv: bad (11.0/5.0)
| | | | | | | purpose = domestic appliance: bad (1.0)
| | | | | | | purpose = repairs: bad (0.0)
| | | | | | | purpose = education: bad (2.0)
| | | | | | | purpose = vacation: bad (0.0)
| | | | | | | purpose = retraining: bad (0.0)
| | | | | | | purpose = business: bad (1.0)
| | | | | | | purpose = other: bad (0.0)
| | | | | | own_telephone = yes: bad (9.0)
| | | | | job = high qualif/self emp/mgmt: good (10.0/3.0)
| | | | savings_status = 100<=X<500: bad (8.0/3.0)
| | | | savings_status = 500<=X<1000: good (1.0)
| | | | savings_status = >=1000: good (2.0)
| | | | savings_status = no known savings: bad (12.0/5.0)
| | | other_parties = co applicant: good (4.0/2.0)
| | | other_parties = guarantor: good (8.0/1.0)
| | credit_history = delayed previously: bad (7.0/2.0)
| | credit_history = critical/other existing credit: good (38.0/10.0)
| foreign_worker = no: good (12.0/2.0)
checking_status = 0<=X<200
| other_parties = none
| | credit_history = no credits/all paid: bad (9.0/1.0)
| | credit_history = all paid: bad (10.0/4.0)

| | credit_history = existing paid

| | | credit_amount <= 8858: good (70.0/21.0)

| | | credit_amount > 8858: bad (8.0)

| | credit_history = delayed previously: good (25.0/6.0)

| | credit_history = critical/other existing credit: good (26.0/7.0)

| other_parties = co applicant: bad (7.0/1.0)

| other_parties = guarantor: good (18.0/4.0)

checking_status = >=200: good (44.0/9.0)

checking_status = no checking: good (262.0/31.0)

Number of Leaves : 36

Size of the tree : 47

Time taken to build model: 0.03 seconds

=== Evaluation on training set ===

=== Summary ===

Correctly Classified Instances	775	77.5 %
Incorrectly Classified Instances	225	22.5 %
Kappa statistic	0.4022	
Mean absolute error	0.3248	
Root mean squared error	0.4032	
Relative absolute error	77.3008 %	
Root relative squared error	87.976 %	
Total Number of Instances	1000	

‘Αρα η ακρίβεια του C4.5 στα δεδομένα ήταν 77.50%.

4.2.3 ΚΑΝΟΝΕΣ ΑΠΟΦΑΣΕΩΝ ΑΠΟ ΤΟΝ ΑΛΓΟΡΙΘΜΟ RIPPER

RIPPER rules:

==========

(checking_status = <0) and (job = skilled) => class=bad (172.0/76.0)

(checking_status = 0<=X<200) and (duration >= 24) and (savings_status = <100) => class=bad (61.0/19.0)

=> class=good (767.0/162.0)

Number of Rules : 3

Time taken to build model: 0.5 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	717	71.7 %
Incorrectly Classified Instances	283	28.3 %
Kappa statistic	0.2513	
Mean absolute error	0.3781	
Root mean squared error	0.4472	
Relative absolute error	89.9974 %	
Root relative squared error	97.5906 %	
Total Number of Instances	1000	

‘Αρα η ακρίβεια του RIPPER στα δεδομένα ήταν 71.70%. Πρέπει επίσης να σημειωθεί ότι αυτή η ακρίβεια επιτεύχθηκε με την χρήση μόνο 3 κανόνων.

4.2.4. ΑΛΓΟΡΙΘΜΟΣ SMO

SMO

Classifier for classes: good, bad

BinarySMO

Machine linear: showing attribute weights, not support vectors.

0.6805 * (normalized) checking_status=<0
+ 0.3347 * (normalized) checking_status=0<=X<200
+ -0.4616 * (normalized) checking_status=>=200
+ -0.5537 * (normalized) checking_status=no checking
+ 1.6987 * (normalized) duration
+ 0.5398 * (normalized) credit_history=no credits/all paid

+ 0.6015 * (normalized) credit_history=all paid
+ -0.109 * (normalized) credit_history=existing paid
+ -0.3182 * (normalized) credit_history=delayed previously
+ -0.7141 * (normalized) credit_history=critical/other existing credit
+ 0.5673 * (normalized) purpose=new car
+ -0.5615 * (normalized) purpose=used car
+ -0.1464 * (normalized) purpose=furniture/equipment
+ -0.0798 * (normalized) purpose=radio/tv
+ 0.5456 * (normalized) purpose=domestic appliance
+ 0 * (normalized) purpose=repairs
+ 0.4441 * (normalized) purpose=education
+ -0.3951 * (normalized) purpose=retraining
+ -0.0823 * (normalized) purpose=business
+ -0.2919 * (normalized) purpose=other
+ 1.1473 * (normalized) credit_amount
+ 0.4056 * (normalized) savings_status=<100
+ 0.115 * (normalized) savings_status=100<=X<500
+ 0.1378 * (normalized) savings_status=500<=X<1000
+ -0.3775 * (normalized) savings_status=>=1000
+ -0.2809 * (normalized) savings_status=no known savings
+ 0.2887 * (normalized) employment=unemployed
+ 0.1663 * (normalized) employment=<1
+ 0.0021 * (normalized) employment=1<=X<4
+ -0.3348 * (normalized) employment=4<=X<7
+ -0.1222 * (normalized) employment=>=7
+ 0.6503 * (normalized) installment_commitment
+ 0.3335 * (normalized) personal_status=male div/sep
+ 0.1177 * (normalized) personal_status=female div/dep/mar

+ -0.3697 * (normalized) personal_status=male single
+ -0.0815 * (normalized) personal_status=male mar/wid
+ 0.0514 * (normalized) other_parties=none
+ 0.5697 * (normalized) other_parties=co applicant
+ -0.6211 * (normalized) other_parties=guarantor
+ -0.0001 * (normalized) residence_since
+ -0.2247 * (normalized) property_magnitude=real estate
+ -0.0544 * (normalized) property_magnitude=life insurance
+ -0.0795 * (normalized) property_magnitude=car
+ 0.3586 * (normalized) property_magnitude=no known property
+ -0.4191 * (normalized) age
+ 0.0697 * (normalized) other_payment_plans=bank
+ 0.159 * (normalized) other_payment_plans=stores
+ -0.2287 * (normalized) other_payment_plans=none
+ 0.3271 * (normalized) housing=rent
+ -0.0702 * (normalized) housing=own
+ -0.257 * (normalized) housing=for free
+ 0.4503 * (normalized) existing_credits
+ -0.2026 * (normalized) job=unemp/unskilled non res
+ 0.1501 * (normalized) job=unskilled resident
+ 0.1027 * (normalized) job=skilled
+ -0.0502 * (normalized) job=high qualif/self emp/mgmt
+ 0.0198 * (normalized) num_dependents
+ -0.1394 * (normalized) own_telephone
+ -0.9888 * (normalized) foreign_worker
- 1.5398

Number of kernel evaluations: 436644

Time taken to build model: 5.67 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	751	75.1 %
Incorrectly Classified Instances	249	24.9 %
Kappa statistic	0.3654	
Mean absolute error	0.249	
Root mean squared error	0.499	
Relative absolute error	59.2607 %	
Root relative squared error	108.8905 %	
Total Number of Instances	1000	

‘Αρα η ακρίβεια του SMO στα δεδομένα ήταν 75.10%.

4.2.5 ΑΛΓΟΡΙΘΜΟΣ BP ΓΙΑ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ

Sigmoid Node 0

Inputs Weights

Threshold -0.07324978292862763

Node 2 3.0044957011533477

Sigmoid Node 1

Inputs Weights

Threshold 0.07324978292862741

Node 2 -3.0044957011533455

Sigmoid Node 2

Inputs Weights

Threshold -0.1524383241253404

Attrib checking_status=<0 -7.127308909968609

Attrib checking_status=0<=X<200 -5.568840800067419

Attrib checking_status=>=200 5.536442678431966

Attrib checking_status=no checking 7.448517624737233

Attrib duration -15.869684583430612

Attrib credit_history=no credits/all paid -6.817339661283083

Attrib credit_history=all paid -3.6648681519473008

Attrib credit_history=existing paid -0.42855909699610395

Attrib credit_history=delayed previously 6.067056265086362

Attrib credit_history=critical/other existing credit 5.202479053504914

Attrib purpose=new car -4.358809242887763

Attrib purpose=used car -0.16391374319658863

Attrib purpose=furniture/equipment 0.7622694371506793

Attrib purpose=radio/tv 1.7737583840894737

Attrib purpose=domestic appliance 1.8504336534889916

Attrib purpose=repairs 2.1256495964623667

Attrib purpose=education 0.8104421363684363

Attrib purpose=vacation 0.01806673914461429

Attrib purpose=retraining 0.6635877207353198

Attrib purpose=business -2.9018486603268867

Attrib purpose=other 0.5412025537531738

Attrib credit_amount -14.471604796632404

Attrib savings_status=<100 -8.491660261965142

Attrib savings_status=100<=X<500 -3.0368882447270185

Attrib savings_status=500<=X<1000 2.1008069748721

Attrib savings_status=>=1000 6.738265810863285

Attrib savings_status=no known savings 3.058206882235268

Attrib employment=unemployed -6.519641320196984

Attrib employment=<1 -1.799287155988702

Attrib employment=1<=X<4 0.6247469663446439

Attrib employment=4<=X<7 4.375385310436861

Attrib employment=>=7 3.809204026168733

Attrib installment_commitment -5.807958324468605

Attrib personal_status=male div/sep -2.713296402011672

Attrib personal_status=female div/dep/mar 0.6876305771791159

Attrib personal_status=male single 0.6271535791111095

Attrib personal_status=male mar/wid 1.6984781745224942

Attrib personal_status=female single 0.04660287398503703

Attrib other_parties=none 1.2286702730465122

Attrib other_parties=co applicant -7.090495452923499

Attrib other_parties=guarantor 5.9829579169630716

Attrib residence_since 3.3348346927199923

Attrib property_magnitude=real estate 3.0872430162523985

Attrib property_magnitude=life insurance 1.7090352341067732

Attrib property_magnitude=car 0.584304141243327

Attrib property_magnitude=no known property -5.05061522200528

Attrib age 0.8517640853787177

Attrib other_payment_plans=bank -2.5652543834404966

Attrib other_payment_plans=stores -1.0643671748128702

Attrib other_payment_plans=none 3.7371110599404647

Attrib housing=rent -3.292778487618977

Attrib housing=own -0.5720561151656457

Attrib housing=for free 4.048407500585945

Attrib existing_credits -0.301979734122562

Attrib job=unemp/unskilled non res 2.0646391791137657

Attrib job=unskilled resident 2.424464680108377

Attrib job=skilled -0.9748219219521777

Attrib job=high qualif/self emp/mgmt -3.2410608868256805

Attrib num_dependents 2.8533458905397975

Attrib own_telephone 4.980271719159508

Attrib foreign_worker 9.605799763342299

Class good

Input

Node 0

Class bad

Input

Node 1

Time taken to build model: 9.72 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	729	72.9 %
Incorrectly Classified Instances	271	27.1 %
Kappa statistic	0.3767	
Mean absolute error	0.3226	
Root mean squared error	0.4313	
Relative absolute error	76.7887 %	
Root relative squared error	94.1153 %	
Total Number of Instances	1000	

‘Αρα η ακρίβεια του BP στα δεδομένα ήταν 72. 90%.

Όπως έχει ήδη αναφερθεί ο σκοπός της εργασίας μας ήταν να χρησιμοποίσει αλγόριθμους εξόρυξης γνώσης για την έγγριση πιστωτικών καρτών βασιζόμενη σε στοιχεία όπως ηλικία, εισόδημα, πιστωτική ιστορία και ιδιοκτησία κατοικίας κλπ. Τα δέντρα αποφάσεων αποδείχθηκε ότι έχουν την καλύτερη ακρίβεια στο πρόβλημα αυτό. Εν

συνεχεία, παρουσιάζουμε το εμπορικό πακέτο εξόρυξης γνώσης KnowledgeSeeker που επιτρέπει μεγάλη παραμετροποίηση στα δέντρα αποφάσεων.

ΣΥΜΠΕΡΑΣΜΑΤΑ

Η βελτίωση της ακρίβειας των αποτελεσμάτων είναι κεφαλαιώδους σημασίας, δεδομένου ότι αποτελεί το ουσιαστικότερο κίνητρο της συνεχιζόμενης ερευνητικής προσπάθειας της εξόρυξης δεδομένων. Καθίσταται δε ακόμα πιο σημαντική αν ληφθούν υπόψη οι επιστημονικές, οικονομικές και κοινωνικές επιπτώσεις των βελτιωμένων αποτελεσμάτων από την εφαρμογή του νέου μεθοδολογικού πλαισίου, σε

βάσεις δεδομένων που αντιστοιχούν σε προβλήματα της καθημερινής ζωής. Η συνεισφορά του πεδίου της Επιχειρησιακής Έρευνας στην γενικότερη ραγδαία ανάπτυξη της Εξόρυξης Δεδομένων ήταν και εξακολουθεί να είναι καθοριστική.

Ειδικότερα, η συνεισφορά των μεθόδων βελτιστοποίησης της επιχειρησιακής έρευνας, αγγίζει σχεδόν κάθε τμήμα των διαδικασιών της εξόρυξης δεδομένων, από την απεικόνιση των δεδομένων και την εκπαίδευσή τους, μέχρι την επιλογή του καλύτερου μοντέλου μετά την ολοκλήρωση της εκπαίδευσης.

Επιπλέον, ιδιαίτερα σημαντική είναι και η αντίστροφη σχέση. Συγκεκριμένα, η εξόρυξη δεδομένων είναι χρήσιμη σε πολλές εφαρμογές της επιχειρησιακής έρευνας, ειδικά αν χρησιμοποιηθεί σε συμπληρωματικό ρόλο σε μεθόδους βελτιστοποίησης, με στόχο την αναγνώριση μεταβλητών και τον περιορισμό του διαθέσιμου χώρου έρευνας.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Allen D., Shore L. & Griffeth R. (1999).A model of perceived organizational support.University of Memphis and Georgia

Anderson J.C. & Gerbing D.W. (1998).SEM in practice: A review and recommended two step approach. Psycological Bulletin

Freund, Y., Schapire, R., Experiments with a new boosting algorithm. *In Proceedings of the 13th International Conference on Machine Learning (ICML)*, 148-156, 1996.

Friedman, J., Hastie, T., Tibshirani, R., Additive Logistic Regression: a Statistical View Of Boosting. *Technical Report*, Stanford University, 1999.

ΔΙΑΔΙΚΤΥΑΚΕΣ ΠΗΓΕΣ

1. **http://invenio.lib.auth.gr/record/114392/files/ptuxiaki.pdf?version=1**
2. **http://oldportal.demokritos.gr/parousiaseis/PALIOURAS_140705.pdf**
3. **http://nemertes.lis.upatras.gr/dspace/bitstream/123456789/272/1/200.pdf**
4. **http://www.dardanosnet.gr/book_details.php?id=1118**
5. **www.cs.aueb.gr/.../lab-db-net.shtml**
6. **http://195.251.230.144/attachments/article/106/Papaoikonomou_Emmanuel_BI_and_DM_presentation.pdf**
7. **http://www.medinfo.cs.ucy.ac.cy/doc/Publications/PhD/MKaraolis/PhD_Mina_Karaolis.pdf**
8. **http://www.dit.hua.gr/index.php?option=com_content&view=article&id=71&Itemid=39&lang=el**
9. **http://multimine.iti.gr/seminar5%20presentations/thewritiko_meros/DPTh_Efremidis_DataMining.pdf**
10. **http://vivliothmmy.ee.auth.gr/76/**

11. kxalv.spaces.live.com/.../cns!4A16CC7BD5DDCF61!302.entry

12. http://www.esi-stat.gr/drastiriotites/TOMOS%20PRAKTIKON%20CHANION/pdf/201-210.pdf

13. http://www.engr.sjsu.edu/meirinaki/papers/Thesis-GR_final_A4.pdf

14. http://gtziralis.com/wp-content/uploads/mis_introtodatamining.pdf

15. http://www.cs.uoi.gr/~pitoura/courses/dm07/index.html