ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΔΥΤ. ΕΛΛΑΔΑΣ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ & ΟΙΚΟΝΟΜΙΑΣ

ΤΜΗΜΑ ΛΟΓΙΣΤΙΚΗΣ & ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΣΤΙΣ ΕΠΙΧΕΙΡΗΣΕΙΣ

ΓΡΑΜΜΕΝΝΟΥ ΛΕΥΚΗ Α.Μ.14489

ΚΑΡΑΝΙΚΑΣ ΑΝΑΣΤΑΣΙΟΣ Α.Μ. 13799

ΝΕΡΗΣ ΒΑΣΙΛΕΙΟΣ Α.Μ. 15065

ΕΙΣΗΓΗΤΗΣ:

ΠΑΠΑΣΤΕΡΓΙΟΥ ΘΩΜΑΣ

ΜΕΣΟΛΟΓΓΙ 2015

Επισήμανση

Οι διαπιστώσεις, τα αποτελέσματα, τα συμπεράσματα και οι πιθανές προτάσεις της παρούσας Πτυχιακής Εργασίας, εκτός των αναφορών που σημαίνονται ως λήμματα, αποτελούν προσωπικές θεωρητικές ή εμπειρικές διαπιστώσεις της ομάδας των φοιτητών που την επιμελήθηκαν και δεν απηχούν κατ' ανάγκη τη γνώμη του εισηγητή εκπαιδευτικού, ή του Εκπαιδευτικού Προσωπικού του Τμήματος Λογιστικής & Χρηματοοικονομικής ή του Τ.Ε.Ι. Δυτ. Ελλάδας.

Πρόλογος – Εισαγωγή

Κατά τη σημερινή εποχή, η λήψη αποφάσεων αποτελεί καθοριστικό παράγοντα οποιασδήποτε επιστήμης, αλλά και της καθημερινής ζωής γενικότερα. Οι αποφάσεις αυτές λαμβάνονται με βάση τα δεδομένα που έχουν στη διάθεσή τους οι υπεύθυνοι και τα οποία συσσωρεύονται σε όλο και μεγαλύτερες βάσεις δεδομένων. Η μεγάλη αύξηση τόσο των δεδομένων όσο και των βάσεων δεδομένων έκανε επιτακτική την ανάπτυξη της εξόρυξης δεδομένων.

Η εξόρυξη δεδομένων αποτελεί μία πολύ σημαντική διαδικασία κατά τη σημερινή εποχή, αφού, μέσω αυτής, θεωρείται ότι μπορούν να εξαχθούν σημαντικά αποτελέσματα, τα οποία μπορούν να ληφθούν υπόψη κατά τον σχεδιασμό και τη λήψη αποφάσεων. Η παρούσα διπλωματική εργασία αποτελεί μία πλήρη περιγραφή της διαδικασίας της εξόρυξης δεδομένων, όπως και των αποτελεσμάτων που εξάγονται από αυτήν. Συνεπώς, σκοπός της διπλωματικής αυτής είναι να παρουσιαστεί η πορεία της επιστήμης της εξόρυξης δεδομένων μέσω των διαφόρων μεθοδολογιών που έχουν αναπτυχθεί, αφού πρώτα προσδιοριστεί ο ορισμός που συνδέεται με αυτήν. Περαιτέρω, διενεργούμε τη διασύνδεση της εξόρυξης δεδομένων με την ηθική δεοντολογία και τους κώδικες που την διέπουν. Μία ακόμη διασύνδεση της επιστήμης της εξόρυξης δεδομένων είναι με τις επιχειρήσεις και τον τρόπο που αυτές λαμβάνουν αποφάσεις συνδυάζοντας όλη την διαθέσιμη πληροφόρηση.

Στο επόμενο κεφάλαιο δίνεται ακριβώς ο ορισμός της εξόρυξης δεδομένων, όπως και τα αίτια που συνδέονται με την ανάπτυξη του συγκεκριμένου κλάδου. Σε αυτό το κεφάλαιο συνδέεται η εξόρυξη δεδομένων με την ανακάλυψη γνώσης από τα δεδομένα, η οποία αποτελεί και το κύριο αποτέλεσμα της εξόρυξης δεδομένων. Έτσι, εκτός από την ανακάλυψη γνώσης στα δεδομένα παρουσιάζονται και οι παράγοντες που επηρεάζουν τα αποτελέσματα και την ποιότητά τους. Σε αυτό το κεφάλαιο, επίσης, παρουσιάζεται και μία εξελεγκτική πορεία στη διαδικασία της εξόρυξης δεδομένων, η οποία αφορά στην εξόρυξη κειμένου, αλλά και οι προκλήσεις που σχετίζονται με την εξόρυξη δεδομένων.

Στη συνέχεια, παρουσιάζονται όλες οι μεθοδολογίες που χρησιμοποιούνται κατά τη διαδικασία της εξόρυξης δεδομένων. Στις μεθοδολογίες αυτές συμπεριλαμβάνονται μεθοδολογίες, οι οποίες χαρακτηρίζονται ως πιο απλές και μπορούν να εφαρμοστούν σε σχεδόν όλα τα είδα μεταβλητών, όπως και πιο σύνθετες μεθοδολογίες, οι οποίες μπορεί να συνδυάζουν στοιχεία από περισσότερες από μία μεθοδολογίες. Σε αυτό το κεφάλαιο, επίσης, παρουσιάζεται και η διαδικασία που ακολουθείται κατά την προετοιμασία της διαδικασίας της εξόρυξης δεδομένων.

Στο τρίτο κεφάλαιο γίνεται η διασύνδεση ανάμεσα στην ηθική και την διαδικασία της εξόρυξης δεδομένων. Η διασύνδεση αυτή κρίνεται ως επιτακτική, αφού, όπως θα δούμε στη συνέχεια, η εξόρυξη δεδομένων οδηγεί σε κάποια αποτελέσματα που συνδέονται με τη λήψη αποφάσεων, οι οποίες με τη σειρά τους επηρεάζουν τους ανθρώπους και την κοινωνία γενικότερα. Επιπλέον, γίνεται και η διασύνδεση με την χρήση των προσωπικών δεδομένων, τα οποία στην πραγματικότητα πρέπει να είναι απόρρητα και σε αρκετές περιπτώσεις προέρχονται με έμμεσο τρόπο από τους καταναλωτές.

Στο τέταρτο κεφάλαιο, συνδέουμε την έννοια της εξόρυξης δεδομένων με τις επιχειρήσεις. Πιο συγκεκριμένα, διασυνδέουμε την έννοια και την διαδικασία της εξόρυξης δεδομένων με τους διάφορους τομείς της επιχείρησης και τον τρόπο με τον οποίο λαμβάνονται αποφάσεις, οι οποίες βασίζονται σε αυτά τα δεδομένα. Σκοπός μας είναι να δείξουμε τον ειδικότερο τρόπο με τον οποίο η λειτουργία, η αποτελεσματικότητα και η κερδοφορία των επιχειρήσεων μπορεί να επηρεαστεί από την διαδικασία της εξόρυξης δεδομένων. Γι’ αυτό το λόγο παρουσιάζουμε και αποτελέσματα σχετικών ερευνών, αλλά και τους τρόπους με τους οποίους πραγματοποιήθηκε η εξόρυξη μέσω αυτών. Επιπλέον, προσδιορίζουμε τον τρόπο με τον οποίο αποθηκεύονται τα δεδομένα στις βάσεις των επιχειρήσεων, σε συνεργασία με τα λογισμικά συστήματα που έχουν αναπτυχθεί από μεγάλες εταιρείες παγκοσμίως.

Το πέμπτο κεφάλαιο, αποτελεί την συζήτηση που προκύπτει από την παραπάνω ανάλυση. Στην πραγματικότητα, σκοπός του έκτου κεφαλαίου είναι η διασύνδεση όλων όσων έχουν ειπωθεί στις προηγούμενες ενότητες και στη συνέχεια η διατύπωση προτάσεων που σχετίζονται με την επιστήμη της εξόρυξης δεδομένων γενικά, αλλά και ειδικά όταν αυτή αφορά τις επιχειρήσεις. Γι’ αυτό το λόγο εστιάζουμε κυρίως σε

προτάσεις, οι οποίες σχετίζονται με την επίλυση των προβλημάτων που έχουν αναφερθεί στα προηγούμενα κεφάλαια.

Το τελευταίο κεφάλαιο αυτής της εργασίας αφορά στα συμπεράσματα που έχουν εξαχθεί από την παραπάνω ανάλυση. Σε αυτό το κεφάλαιο επί της ουσίας διενεργούμε μία περίληψη των όσων έχουν ειπωθεί στα προηγούμενα κεφάλαια με σκοπό την συνοπτική περιγραφή όλων των αποτελεσμάτων που έχουν βρεθεί.

Περιεχόμενα

1. Ορισμός της Εξόρυξης Δεδομένων

Σε αυτό το κεφάλαιο, διενεργούμε μία επισκόπηση της θεωρίας που συνδέεται άμεσα με την εξόρυξη δεδομένων έτσι όπως έχει αυτή αποτυπωθεί στη διεθνή βιβλιογραφία. Το κεφάλαιο αυτό αποτελεί επί της ουσίας έναυσμα για την ουσιαστικότερη κατανόηση της όλης διαδικασίας, η οποία θα ακολουθήσει στα επόμενα κεφάλαια. Αξίζει να σημειώσουμε ότι στην δεκαετία 2000 – 2011 γράφτηκαν περισσότερα από 10.000 άρθρα με θέμα την εξόρυξη δεδομένων, ενώ όπως σημειώνουν οι Liao et al. (2012) η ανάπτυξη της τεχνολογίας βοήθησε σε αυτή την διαδικασία. Αρχικά, παρουσιάζουμε τα αίτια, τον ορισμό και τους τομείς στους οποίους χρησιμοποιείται η εξόρυξη δεδομένων. Στη συνέχεια, συνδέουμε την εξόρυξη δεδομένων με την εξαγωγή/ανακάλυψη γνώσης από τις βάσεις δεδομένων. Ακολουθεί η περιγραφή των αποτελεσμάτων που προκύπτουν από αυτές, ενώ δείχνουμε και την εξελικτική τάση του τομέα της εξόρυξης δεδομένων, όπως και τις προκλήσεις που συνδέονται με αυτήν.

1.1 Αίτια και Ορισμός της Εξόρυξης Δεδομένων

Η εξόρυξη δεδομένων έχει προκύψει ως μια επιτακτική ανάγκη για την παραγωγή νέας γνώσης η οποία προκύπτει από πληροφορίες, οι οποίες έχουν συσσωρευτεί κατά το παρελθόν και μπορούν να οδηγήσουν στην βελτίωση της υφιστάμενης κατάστασης. Στην πραγματικότητα, η εξόρυξη δεδομένων σχετίζεται άμεσα με την απόσπαση πληροφοριών από μεγάλες βάσεις δεδομένων, οι οποίες συνδυάζονται μέσω της στατιστικής με σκοπό την εξαγωγή ασφαλών συμπερασμάτων.

Στα οικονομικά, η ανάγκη αυτή έγινε πιο εμφανής όταν όλο και περισσότερες επιχειρήσεις ξεκίνησαν να διατηρούν πολύ μεγάλες βάσεις δεδομένων στις οποίες καταγράφονταν όλες οι σχετιζόμενες με την δραστηριότητά τους πληροφορίες. Ως αποτέλεσμα, οι βάσεις των δεδομένων έφτασαν να αποθηκεύουν εκατομμύρια bytes δεδομένων, μέσα στα οποία εμπεριέχονταν όλες οι απαραίτητες πληροφορίες για την ανάλυση του υπό εξέταση φαινομένου.

Το μεγάλο μέγεθος των πληροφοριών έχει προκύψει ως αποτέλεσμα της συνεχόμενης νέας πληροφορίας, η οποία εισρέει σε ήδη μεγάλες βάσεις δεδομένων. Το μέγεθος των βάσεων δεδομένων μπορεί να αυξηθεί με δύο τρόπους: α) μέσω της αύξησης του μεγέθους του δείγματος και β) μέσω της αύξησης του αριθμού των τομέων/κλάδων/επιστημονικών πεδίων που μία μεταβλητή μπορεί να καλύψει (Fayyad et al., 1996). Κατά συνέπεια από το μέγεθος των βάσεων δεδομένων και μόνον, θεωρείται ότι η διαδικασία αναζήτησης και επεξεργασίας των απαραίτητων δεδομένων που αποσκοπούν στην εξαγωγή γνώσης είναι ιδιαίτερα δύσκολη και χρονοβόρα. Παράλληλα, αυτές οι βάσεις δεδομένων μπορεί να είναι τόσο πολύπλοκες από τη στιγμή που εμπεριέχουν πολλά δεδομένα, τα οποία μπορούν εύκολα να συγχέονται και μεταξύ τους.

Η δυνατότητα συλλογής δεδομένων, ωστόσο, έχει αυξηθεί σημαντικά κατά τα τελευταία χρόνια μέσω της εξέλιξης της τεχνολογίας, η οποία μπόρεσε να απλοποιήσει τη διαδικασία αυτή και να δημιουργήσει ένα σημαντικό αριθμό συνεχώς εξελισσόμενων και προσιτών συστημάτων, τα οποία έκαναν ακόμη πιο αναγκαία την ανάπτυξη του τομέα της εξόρυξης δεδομένων (Chen et al., 1996).

Γενικότερα, η εξόρυξη δεδομένων έχει εντοπιστεί στην αναζήτηση προτύπων στα υπό εξέταση προβλήματα και μεταξύ των τομέων, στους οποίους εφαρμόζεται περιλαμβάνονται και οι ακόλουθοι τομείς:

- Ιατρική: όπου περιλαμβάνονται η Γενετική και η Βιοτεχνολογία,
- Γεωγραφία: όπου αναλύονται όλα τα πιθανά δεδομένα με σκοπό την καλύτερη ανάπτυξη της περιοχής ή την λεπτομερέστερη καταγραφή των περιβαλλοντικών στοιχείων της,
- Ανάλυση εικόνας: όπου εμπεριέχονται όλα τα στοιχεία που μπορούν να προβούν στη λεπτομερέστερη καταγραφή μιας εικόνας εκ των οποίων προκύπτουν δεδομένα προς ανάλυση,
- Αστρονομία: όπου διενεργούνται πράξεις για την καλύτερη εξερεύνηση και κατανόηση του κλάδου αυτού,
- Τηλεπικοινωνίες: όπου η εξόρυξη δεδομένων εστιάζει κυρίως στην πρόληψη κακόβουλων λογισμικών ή στην πρόληψη απατών που μπορεί να διενεργηθούν μέσω αυτών,

- Οικονομικά: όπου εστιάζει στη μελέτη μικροοικονομικών ή/και μακροοικονομικών μεγεθών με σκοπό την καλύτερη αποτελεσματικότητα.

Διερευνώντας λίγο περισσότερο τον τομέα των οικονομικών εντοπίζουμε ότι η εξόρυξη δεδομένων μπορεί να χρησιμοποιηθεί σε τομείς όπως ο σχεδιασμός της παραγωγικής διαδικασίας, των προϊόντων, των μηχανημάτων που σχετίζονται με την παραγωγική διαδικασία, των πρώτων υλών και των υλικών γενικότερα, των διαδικασιών, των αποθεμάτων, των πωλήσεων, του μάρκετινγκ, των πατεντών, της ανάπτυξης των στρατηγικών προώθησης των προϊόντων και των υπηρεσιών, της χρηματοοικονομικής και ελεγκτικής, του εμπορίου, της επενδυτικής δραστηριότητας αλλά και τους τρόπους με τους οποίους αλληλεπιδρούν μεταξύ τους τα παραπάνω μεγέθη.

Ο Naisbitt (1982) επισημαίνει ότι μέσω της διαδικασίας της εξόρυξης δεδομένων προσπαθούμε να επιτύχουμε την παραγωγή νέας γνώσης από μία πληθώρα πληροφοριών. Η έννοια της εξόρυξης δεδομένων είναι πάρα πολύ σημαντική εάν λάβουμε υπόψη και τον παράγοντα ότι τα δεδομένα πλέον μπορούν να βρίσκονται σε περισσότερες από μία μορφές (για παράδειγμα εικόνες, ήχος, αριθμοί, κείμενο κλπ.) και μάλιστα ότι πολλές φορές αυτές οι διαφορετικές μορφές μπορούν να αλληλεπιδρούν και να συσχετίζονται μεταξύ τους. Γι' αυτό το λόγο, η ανάγκη για μια αυτόματη ανάλυση και η εξεύρεση των κατάλληλων εργαλείων για την εξόρυξη δεδομένων αποτελούν μία πολύ σημαντική διαδικασία κατά τα τελευταία χρόνια.

Οι Bagga and Singh (2012) θεωρούν ότι ο κύριος στόχος της διαδικασίας της εξόρυξης δεδομένων είναι η εξαγωγή αποτελεσμάτων από τον χρήστη, η οποία βασίζεται στα λιγότερα εισαγόμενα δεδομένα και την μικρότερη δυνατή προσπάθεια από τον χρήστη. Σε αυτό το σημείο οφείλουμε να κάνουμε ένα σημαντικό διαχωρισμό ανάμεσα στη γνώση και την πληροφορία. Η γνώση αποτελεί το παραγόμενο προϊόν της καλύτερης και σωστότερης ανάλυσης της πληροφορίας. Αντίθετα, η πληροφορία αναφέρεται στις τιμές των δεδομένων, όπως και τις εξεταζόμενες σχέσεις, οι οποίες μπορεί να έχουν ή όχι παρατηρηθεί κατά το παρελθόν ανάμεσα σε αυτά.

Σύμφωνα με τους Frawley et al. (1991) η διαδικασία της εξόρυξης δεδομένων ορίζεται ως η διαδικασία κατά την οποία εξάγονται χρήσιμες πληροφορίες από μία βάση δεδομένων η οποία έχει δημιουργηθεί με παλαιότερες πληροφορίες. Η εξόρυξη δεδομένων, δηλαδή, αφορά στην διερεύνηση όλων των τεχνικών και των διαδικασιών που αποσκοπούν στην εύρεση νέας και πιθανόν χρήσιμης γνώσης από τις βάσεις δεδομένων. Εναλλακτικά, όπως την ορίζουν οι Elovici and Braha (2003), η εξόρυξη δεδομένων είναι η διαδικασία κατά την οποία αποκτούνται έγκυρες, προηγουμένως άγνωστες και κατανοητές πληροφορίες από μεγάλες βάσεις δεδομένων με σκοπό την βελτίωση των αποφάσεων των επιχειρήσεων. Ο Ahmed (2004), συνδέοντας την εξόρυξη δεδομένων με την επιχειρηματικότητα και το μάνατζμεντ, υποστηρίζει ότι η εξόρυξη δεδομένων οδηγεί στη δημιουργία νέων πληροφοριών οι οποίες στη συνέχεια μπορούν να χρησιμοποιηθούν για την περαιτέρω ανάπτυξη μιας εταιρείας, αλλά και την διατήρηση της θέσης της στην οικονομία. Ο Cook (2005) υποστηρίζει ότι βασικός σκοπός της εξόρυξης δεδομένων είναι η πρόβλεψη κάποιων καταστάσεων και όχι η απλή περιγραφή τους.

Μπορούμε, δηλαδή, να διαπιστώσουμε ότι ως λύση στο πρόβλημα της σύνθετης μορφής των δεδομένων και της περαιτέρω ανάλυσής τους προτάθηκε η τεχνική της εξόρυξης δεδομένων (data mining). Σύμφωνα με τον Larose (2005), η εξόρυξη δεδομένων αποτελεί τη διαδικασία κατά την οποία επιτυγχάνεται η εξαγωγή γνώσης. Η εξαγωγή γνώσης έχει προκύψει ως αποτέλεσμα της εξέτασης των συσχετίσεων μεταξύ των υπό εξέταση μεταβλητών, των προτύπων, των τάσεων, όπως και της επεξεργασίας των μεταβλητών μέσω αλγορίθμων, οι οποίοι στηρίζονται στις βασικές αρχές της στατιστικής. Πιο συγκεκριμένα, η εξόρυξη δεδομένων στοχεύει στην πρόβλεψη μελλοντικών καταστάσεων, οι οποίες επηρεάζουν τη διαδικασία λήψης αποφάσεων, μέσω της συλλογής και επεξεργασίας των διαφορετικών δεδομένων.

Οι Fayyad et al. (1996), όπως και οι Berry and Linoff (1999), διακρίνουν δύο μεγάλες κατηγορίες στην διαδικασία της εξόρυξης δεδομένων. Η πρώτη κατηγορία είναι προσανατολισμένη/κατευθυνόμενη στην επιβεβαίωση των υπό εξέταση υποθέσεων και αποσκοπεί στο ιδανικό «ταίριασμα» των δεδομένων. Σε αυτή την κατηγορία, τα αποτελέσματα συνήθως παρουσιάζονται με τη μορφή διαγραμμάτων, της κατάταξης, της εκτίμησης και της πρόβλεψης. Η δεύτερη κατηγορία είναι προσανατολισμένη στην εξεύρεση νέων κανόνων και προτύπων που στοχεύουν στην περιγραφή και την πρόβλεψη των μελλοντικών καταστάσεων. Σε αυτή την περίπτωση, η παλινδρόμηση,

η ομαδοποίηση και η συσταδοποίηση. Οι Braha and Shmilovici (2003), εκτός από αυτές τις δύο κατηγορίες, εντοπίζουν και μία τρίτη, η οποία αφορά στον εντοπισμό ασυνήθιστων προτύπων στα δεδομένα.

1.2 Η Ανακάλυψη της Γνώσης στις Βάσεις Δεδομένων (Knowledge Discovery in Databases – KDDs)

Όπως αναφέραμε και παραπάνω η παραγωγή γνώσης από τα δεδομένα αποτελεί το τελικό προϊόν μιας διαδικασίας, ενώ τα δεδομένα δεν αποτελούν παρά την πληροφορία που χρησιμοποιείται για την παραγωγή αυτής της γνώσης. Με αυτόν τον τρόπο ορίζεται και η ανακάλυψη γνώσης στις βάσεις δεδομένων (Knowledge Discovery in Databases – KDDs) ως η διαδικασία εντοπισμού νέας και πιθανώς χρήσιμης γνώσης από ένα σύνολο δεδομένων, μέσω της οποίας εντοπίζονται πρότυπα στα δεδομένα.

Οι Fayyad et al. (1996) υποστηρίζουν ότι εξαιτίας της πληθώρας δεδομένων, έχει εντοπιστεί η ανάγκη θέσπισης και δημιουργίας νέων θεωριών που να οδηγούν στην ανακάλυψη γνώσης από τις βάσεις δεδομένων. Έτσι, η διαδικασία εστιάζει στην εξόρυξη δεδομένων μέσω της χρήσης δεδομένων χαμηλής ποιότητας. Η κύρια διαφορά ανάμεσα στην εξόρυξη δεδομένων και την ανακάλυψη γνώσης από τις βάσεις δεδομένων, είναι, ότι η δεύτερη, αναφέρεται στο σύνολο της διαδικασίας κατά την οποία προκύπτει γνώση από τα δεδομένα, ενώ η εξόρυξη δεδομένων σε ένα μόνο βήμα της συνολικής διαδικασίας που εστιάζει στη χρήση συγκεκριμένων μεθοδολογιών που αποσκοπούν στην εξαγωγή συγκεκριμένων προτύπων. Η συνολική διαδικασία βασίζεται κατά κύριο λόγο σε κανόνες της στατιστικής από τη στιγμή που μέσω αυτής μπορεί να ποσοτικοποιηθεί η αβεβαιότητα, και η μόνη διαφορά έγκειται στο γεγονός ότι στην ανακάλυψη γνώσης από τις βάσεις δεδομένων αυτοματοποιείται η ανάλυση των υποθέσεων.

Όπως σημειώνουν οι Elovici and Braha (2003) συνήθως η εξόρυξη δεδομένων προσδιορίζεται από την απόκτηση νέων πληροφοριών. Ωστόσο, σημειώνουν ότι όταν η εξόρυξη δεδομένων εστιάζει στην εξαγωγή πληροφοριών και όχι γνώσης, τότε

θεωρείται ότι αυτή εντοπίζει μόνο κάποια πρότυπα στα δεδομένα. Με αυτό τον τρόπο διαπιστώνουν ότι η ανακάλυψη γνώσης από τις βάσεις δεδομένων αποτελείται από τα ακόλουθα πέντε στάδια:

- Πλήρης κατανόηση του κλάδου στον οποίο χρησιμοποιείται η εξόρυξη δεδομένων και ακριβής προσδιορισμός των στόχων για τον οποίο αυτή πραγματοποιείται,
- Επιλογή και έλεγχος του συνόλου των δεδομένων που χρησιμοποιούνται,
- Μετασχηματισμός των δεδομένων προκειμένου αυτά να είναι κατάλληλα για επεξεργασία, αλλά και καθαρισμός των δεδομένων από ελλιπή στοιχεία ή λανθασμένη πληροφορία,
- Εξόρυξη των δεδομένων μέσω των προαποφασισθέντων προτύπων, δηλαδή μέσω της χρήσης του κατάλληλου αλγορίθμου και μοντέλου,
- Παρουσίαση των αποτελεσμάτων της διαδικασίας εξόρυξης δεδομένων.

Όπως καταλαβαίνουμε, η ανακάλυψη γνώσης από τις βάσεις δεδομένων επικεντρώνεται στο σύνολο της διαδικασίας της παραγωγής γνώσης και περιλαμβάνει τον τρόπο με τον οποίο τα δεδομένα βρίσκονται και αξιολογούνται, την αποτελεσματική χρήση των αλγορίθμων, την ερμηνεία των αποτελεσμάτων και γενικότερα την αλληλεπίδραση ανθρώπων και μηχανημάτων.

Αντίστοιχα, οι Brachman and Anand (1996) παρουσιάζοντας τα βήματα της εξόρυξης δεδομένων δίνουν έμφαση στη διαδραστική φύση της διαδικασίας αυτής στα οικονομικά και αναγνωρίζουν τα ακόλουθα βήματα:

- Πλήρης κατανόηση του τομέα στον οποίο διεξάγεται η ανακάλυψη γνώσης από τις βάσεις δεδομένων,
- Πλήρης κατανόηση της προγενέστερης γνώσης που σχετίζεται τον στόχο της ανακάλυψης γνώσης από τις βάσεις δεδομένων,
- Δημιουργία του σετ των δεδομένων,

- Ξεκαθάρισμα των δεδομένων, όπου αφαιρούνται ελλιπείς και λανθασμένες πληροφορίες,
- Μετασχηματισμός των δεδομένων προκειμένου να είναι κατάλληλα προς επεξεργασία,
- Ταίριασμα του στόχου της διαδικασίας της ανακάλυψης της γνώσης δεδομένων με μία συγκεκριμένη μεθοδολογία εξόρυξης δεδομένων,
- Περαιτέρω επεξηγηματική ανάλυση των δεδομένων και της μεθοδολογίας εξόρυξης που χρησιμοποιείται σε συνδυασμό με την προς εξέταση υπόθεση,
- Εξόρυξη των δεδομένων μέσω της εύρεσης του κατάλληλου προτύπου που περιγράφει καλύτερα τις σχέσεις μεταξύ των μεταβλητών,
- Ερμηνεία των αποτελεσμάτων, η οποία μπορεί να κάνει την όλη διαδικασία να επιστρέψει στα προηγούμενα βήματα προκειμένου να επιβεβαιωθούν τα αποτελέσματα ή να διορθωθούν τυχόν πιθανά λάθη,
- Χρήση της γνώσης που ανακαλύφθηκε ή ενσωμάτωση αυτής της γνώσης σε κάποια άλλη διαδικασία με σκοπό την παραγωγή περαιτέρω γνώσης.

1.3 Αποτελέσματα της Διαδικασίας της Εξόρυξης Δεδομένων και Εξέλιξη της Πορείας τους

Αφού διενεργηθεί η εξόρυξη των δεδομένων μας, απαιτείται στη συνέχεια η καταγραφή των αποτελεσμάτων που προέρχονται από αυτή. Η καταγραφή αυτή θα πρέπει να είναι όσο το δυνατόν πιο κατανοητή προκειμένου να μπορεί να χρησιμοποιηθεί στο μέλλον η παραγόμενη γνώση είτε ως αυτοτελής είτε ως πρώτη ύλη για την παραγωγή νέας γνώσης.

Τα αποτελέσματα της διαδικασίας της εξόρυξης των δεδομένων συνδέονται άμεσα με την ποιότητα των δεδομένων που έχει χρησιμοποιηθεί. Τα αποτελέσματα, συνολικά, της διαδικασίας της εξόρυξης των δεδομένων επηρεάζονται, είτε θετικά είτε αρνητικά, από τους ακόλουθους παράγοντες:

- Ακραίες τιμές των δεδομένων: οι οποίες μπορεί να οδηγήσουν σε λανθασμένα συμπεράσματα εάν τους δοθεί μεγαλύτερη βαρύτητα από την προβλεπόμενη καθώς αναφέρονται σε περιστάσεις, οι οποίες μπορεί πλέον να μην ισχύουν,
- Δεδομένα που ενώ έπρεπε να χρησιμοποιηθούν, δεν χρησιμοποιήθηκαν εν τέλει: δηλαδή αναφερόμαστε στην περίπτωση που έχουμε θέσει εκτός του συνόλου των δεδομένων μας μεταβλητές που θα μπορούσαν να εξηγήσουν σημαντικά κάποιο φαινόμενο,
- Δεδομένα που σχετίζονται κυκλικά με άλλα δεδομένα: αναφερόμαστε στην περίπτωση όπου τα δεδομένα σχετίζονται με άλλα δεδομένα αφού πολλές φορές είναι παράγωγα πράξεων των ίδιων των δεδομένων,
- Πιθανά λάθη που έγιναν κατά την κωδικοποίηση των δεδομένων: αφορά σε λάθη που έγιναν όταν είτε οι απαντήσεις δεν κωδικοποιήθηκαν με τον κατάλληλο τρόπο προκειμένου να είναι διαθέσιμα προς επεξεργασία είτε έγιναν λάθη κατά την κωδικοποίηση,
- Τον χρησιμοποιούμενο αλγόριθμο εξόρυξης των δεδομένων: είναι πιθανό να εξάγονται διαφορετικά αποτελέσματα ανάλογα με την χρησιμοποιούμενη μεθοδολογία, αφού πολλές φορές ο ίδιος αλγόριθμος δεν είναι ο καταλληλότερος για όλες τις περιπτώσεις.

Περαιτέρω, οι Feelders et al. (2000) υποστηρίζουν ότι η εξόρυξη δεδομένων επηρεάζεται από τους ακόλουθους τέσσερις ακόλουθους παράγοντες:

- Λόγω της αυξημένης ανάγκης της όσο πιο δυνατόν επαρκούς και ακριβούς πληροφόρησης είναι πολύ πιθανό οι επιχειρήσεις να επενδύουν σημαντικά ποσά σε αυτή και να δίνουν αυξημένη προσοχή στο σύνολο της διαδικασίας,
- Όσο πιο συχνή είναι η διαδικασία της εξόρυξης δεδομένων (υπάρχουν περιπτώσεις όπου αυτή εφαρμόζεται ακόμη και σε καθημερινή βάση), τόσο μεγαλύτερη συνολοκλήρωση απαιτείται με τα υπάρχοντα συστήματα,
- Λόγω της απουσίας αρχικού σχεδιασμού της διαδικασίας της μελέτης, από θέμα στατιστικής απόψεως η διαδικασία είναι αμφιλεγόμενη, αφού μπορεί ο αλγόριθμος που θα χρησιμοποιηθεί να μην είναι ο κατάλληλος και να

χρειαστεί επανάληψη του συνόλου της διαδικασίας ή να δοθούν λάθος αποτελέσματα τα οποία να επηρεάσουν αρνητικά τη λήψη αποφάσεων,

- Τα αποτελέσματα παρουσιάζονται και ερμηνεύονται ανάλογα με την τεχνική της εξόρυξης που χρησιμοποιήθηκε. Αυτό σημαίνει ότι τα αποτελέσματα θα μπορούσαν να είναι πολύ πιο διαφορετικά αν χρησιμοποιούταν κάποια άλλη διαδικασία από αυτή που επιλέγεται τελικά.

Σύμφωνα με την έως τώρα ανάλυσή μας, έχει γίνει αντιληπτό ότι η εξόρυξη δεδομένων προέρχεται από αριθμητικά στοιχεία, τα οποία απεικονίζουν την πορεία των υπό εξέταση μεταβλητών και ότι μέσω των κατάλληλων ανά περίπτωση αλγορίθμων, επιτυγχάνεται η σωστή εξαγωγή γνώσης. Ωστόσο, πέρα από την εξόρυξη των δεδομένων έχει παρατηρηθεί ότι έχουν δημιουργηθεί κατά τα τελευταία χρόνια και άλλες μορφές εξαγωγής γνώσης. Μία από αυτές είναι και η εξόρυξη κειμένου (text mining).

Ο Hearst (2003) υποστηρίζει ότι η εξόρυξη κειμένου εστιάζει στην εξαγωγή χρήσιμης γνώσης από κείμενα γραμμένα με λέξεις και όχι από βάσεις δεδομένων. Η διαδικασία εξαγωγής γνώσης σε αυτή την περίπτωση είναι ακόμη πιο δύσκολη καθώς το κείμενο αυτό θα πρέπει να διαβαστεί από ανθρώπους, οι οποίοι θα πρέπει να είναι σε θέση να μπορούν να καταλάβουν πλήρως όλες τις λέξεις και όλα τα νοήματα και μηνύματα του κειμένου. Αυτό σημαίνει, ότι σε αυτή την περίπτωση τα δεδομένα δεν μπορούν να αναλυθούν μέσω ενός κατάλληλου αλγορίθμου, κάτι το οποίο είναι εξαιρετικά δύσκολο ως διαδικασία.

Οι Zhang and Zhou (2004), αναγνωρίζοντας τις δυσκολίες της συγκεκριμένης επιστήμης και την όλο εξελισσόμενη πορεία της, εντοπίζουν τις προκλήσεις που θεωρούν ότι θα σχετιστούν με αυτήν κατά το μέλλον:

- Επιλογή των μεθοδολογιών και των παραμέτρων: όσο οι βάσεις δεδομένων μεγαλώνουν και γίνονται πιο περίπλοκες, τόσο πιο πιθανό είναι οι υπάρχουσες μεθοδολογίες να μην επαρκούν για την εξόρυξη και την ανάλυση όλο και πιο περίπλοκων και δύσκολων στον εντοπισμό σχέσεων,

- Επάρκεια και παρουσίαση: τα αποτελέσματα μπορεί να μην είναι αντιπροσωπευτικά της πραγματικότητας όταν οι μεθοδολογίες θεωρούνται ανεπαρκείς ή οι συνθήκες του περιβάλλοντος είναι πολύ ρευστές (αβεβαιότητα),
- Εξόρυξη κειμένου: αποτελεί ένα μέρος της εξελιγκτικής διαδικασίας της εξόρυξης δεδομένων που μπορεί να δώσει χρήσιμα αποτελέσματα αλλά είναι ιδιαίτερα δύσκολο να προσδιοριστεί και να ερμηνευθεί,
- Συνολοκλήρωση των πολλών τεχνικών εξόρυξης δεδομένων: όσο οι βάσεις δεδομένων γίνονται πιο περίπλοκες και συνδέονται μεταξύ τους απαιτείται και ο συνδυασμός των επιμέρους μεθοδολογιών με σκοπό την εξαγωγή καλύτερων αποτελεσμάτων,
- Ετερογενείς πηγές: αφορά στο γεγονός ότι τα δεδομένα δεν αντλούνται πάντοτε από την ίδια βάση και μπορεί να μην μπορούν στην πραγματικότητα να συνδυαστούν μεταξύ τους από τη στιγμή που μπορεί ακόμα και οι κλίμακες μέτρησης να διαφέρουν μεταξύ τους.

Παράλληλα, εκτός από τα προβλήματα αυτά, θα μπορούσαμε να επισημάνουμε και κάποια ακόμη σημαντικά ζητήματα/προβλήματα που σχετίζονται με την διαδικασία της εξόρυξης δεδομένων. Ένα από τα σημαντικότερα ζητήματα που θεωρούμε ότι εμποδίζει την ευκολότερη ανάπτυξη του κλάδου είναι το γεγονός ότι πολλές φορές τόσο τα δεδομένα όσο και η γνώση που μπορεί να παραχθεί από αυτά αφορούν μόνον συγκεκριμένα προβλήματα/ζητήματα με αποτέλεσμα η γνώση αυτή να μην μπορεί να γενικευτεί και να μην υπάρχει η δυνατότητα αξιοποίησής της από άλλους τομείς ακόμη κι εάν αυτή μπορούσε να έχει πολύ σημαντικές θετικές επιπτώσεις σε αυτούς. Επιπλέον, εξαιτίας του μεγάλου όγκου των δεδομένων που εμπεριέχονται στις βάσεις και της πολυπλοκότητας που τις χαρακτηρίζει, μόνον οι αναλυτές που έχουν ένα ισχυρό στατιστικό και μαθηματικό υπόβαθρο είναι σε θέση να μπορέσουν να αναλύσουν τα δεδομένα αυτά και να παράγουν γνώση από αυτά. Με αυτόν τον τρόπο αποκλείονται από το σύνολο της διαδικασίας όσοι δεν έχουν τη γνώση για να κάνουν αυτές τις αναλύσεις και κατά συνέπεια μπορεί ένα σημαντικό μέρος των δεδομένων που υπάρχει σε συγκεκριμένες μόνον βάσεις (π.χ. ιδιωτικές βάσεις δεδομένων των επιχειρήσεων) να μένουν ανεκμετάλλευτες, ενώ εμπεριέχουν σημαντική πληροφόρηση.

Άμεσα συνδεδεμένο με το παραπάνω πρόβλημα είναι και η περίπτωση όπου οι αναλυτές δεν είναι ούτε πλήρως ενημερωμένοι ούτε πλήρως εξοικειωμένοι με τις δυνατότητες της ανάλυσης που μπορεί να επιφέρει η τεχνολογία. Αυτό πολλές φορές μπορεί να οφείλεται τόσο στην συνεχώς ανανεωτική τάση της τεχνολογίας όσο και στην περιορισμένη ενημέρωσή τους σχετικά με αυτήν. Επίσης, η εξόρυξη δεδομένων μπορεί να επικριθεί αρκετές φορές επειδή πολλές φορές τα αποτελέσματα που προκύπτουν μπορεί να είναι τόσο απλά που να μην απαιτούνταν καν η χρήση κάποιου συγκεκριμένου αλγορίθμου.

Η εξόρυξη δεδομένων θεωρείται μία πολύ χρήσιμη διαδικασία ιδίως όταν τα δεδομένα μας αφορούν στην ύπαρξη ακραίων τιμών. Αυτό σημαίνει ότι οι ακραίες τιμές, οι οποίες σε πολλές περιπτώσεις δεν ακολουθούν τη γενική κατανομή των υπολοίπων δεδομένων μας αν και θεωρείται ότι λειτουργούν ως θόρυβος στα δεδομένα μας, χρησιμεύουν ιδιαίτερα κατά τον εντοπισμό ων ακραίων φαινομένων, τα οποία συνήθως δεν συναντώνται διαρκώς.

2. Οι Μεθοδολογίες Εξόρυξης Δεδομένων

Κατά τα τελευταία χρόνια και μετά τη σημαντική αύξηση του όγκου των δεδομένων, η οποία προκάλεσε και σημαντική αύξηση των ίδιων των βάσεων, αναπτύχθηκαν και πολλές νέες μεθοδολογίες προκειμένου να απλοποιηθεί το σύνολο της διαδικασίας. Σε αυτή την ενότητα παρουσιάζουμε τις μεθοδολογίες εξόρυξης δεδομένων. Ωστόσο, πριν την παρουσίαση των μεθοδολογιών παρουσιάζονται τα κύρια βήματα της διαδικασίας της εξόρυξης δεδομένων, όπως και τα στάδια που εφαρμόζονται στα δεδομένα πριν από την εφαρμογή της διαδικασίας αυτής. Οφείλουμε να σημειώσουμε ότι τα βήματα αυτά είναι διαφορετικά από αυτά που παρουσιάστηκαν στο προηγούμενο κεφάλαιο και αφορούν στην εξαγωγή γνώσης από τις βάσεις δεδομένων, αφού η εξόρυξη δεδομένων αποτελεί όπως είδαμε ένα βήμα αυτής της διαδικασίας.

Οι Wu et al. (2003) υποστηρίζουν ότι η εξόρυξη δεδομένων εστιάζει στην καταλληλότερη και πιο ακριβή χρήση των ήδη υπαρχόντων αλγορίθμων με σκοπό την παραγωγή γνώσης από αυτά. Γι' αυτό το λόγο έχουν διαχωρίσει τα εξής τέσσερα βήματα σε αυτή τη διαδικασία:

- Καθάρισμα των δεδομένων (Data Cleaning): είναι η διαδικασία κατά την οποία εντοπίζονται, διορθώνονται ή αφαιρούνται οι λανθασμένες ή ελλιπείς πληροφορίες από τη βάση δεδομένων με σκοπό να εξαχθούν κάποια ασφαλή συμπεράσματα,
- Προσδιορισμός των χαρακτηριστικών των δεδομένων που πρόκειται να χρησιμοποιηθούν για την κατασκευή του υποδείγματος (Feature Construction/Extraction): είναι ο τρόπος κατά τον οποίο περιγράφεται με ακρίβεια το σύνολο των δεδομένων που θα χρησιμοποιηθεί στο υπόδειγμα,
- Επιλογή του αλγορίθμου και προσδιορισμός των παραμέτρων του υποδείγματος προς εκτίμηση (Algorithm and Parameter Selection): επιλέγεται ο κατάλληλος αλγόριθμος εξόρυξης δεδομένων, σύμφωνα με τον ερευνητή, και καθορίζονται οι παράμετροι που περιλαμβάνονται στο υπόδειγμα και πρέπει να εκτιμηθούν,

- Ερμηνεία και εγκυρότητα των εξαγόμενων αποτελεσμάτων (Interpretation and Validation): ερμηνεύονται τα αποτελέσματα που εξήχθησαν και ελέγχεται εάν -τα αποτελέσματα αυτά είναι συμβατά με τους περιορισμούς της ανάλυσης και εάν αυτά τα αποτελέσματα μπορούν όντως να προβλέψουν το υπό εξέταση πρόβλημα.

Αντίστοιχα, και οι Fayyad et al. (1996), σημείωσαν ότι όλες οι μεθοδολογίες εξόρυξης δεδομένων αποτελούνται από τα εξής τρία στοιχεία:

- Απεικόνιση και περιγραφή του μοντέλου που χρησιμοποιείται: επί της ουσίας αναφερόμαστε στη γλώσσα που χρησιμοποιείται για την περιγραφή των προτύπων και την πλήρη κατανόηση από την πλευρά του ερευνητή των υποθέσεων του μοντέλου που σχετίζονται άμεσα με τη συγκεκριμένη μεθοδολογία,
- Αξιολόγηση του μοντέλου που χρησιμοποιείται: αναφέρεται σε μέτρα (measures) που εξετάζουν την προβλεπτική ικανότητα του μοντέλου, τη χρησιμότητά του στο συγκεκριμένο τομέα και τη δυνατότητα προσαρμογής του στα δεδομένα και την υφιστάμενη κατάσταση,
- Εκτίμηση των παραμέτρων του μοντέλου: αναφέρεται στην διαδικασία βελτιστοποίησης των μέτρων αξιολόγησης του μοντέλου με σκοπό την εκτίμηση των παραμέτρων.

Αντίστοιχα, οι Feelders et al. (2000) υποστηρίζουν ότι προκειμένου να εξαχθούν χρήσιμα αποτελέσματα ή γνώση, όπως χαρακτηριστικά αναφέρουν, θα πρέπει να ακολουθηθούν τα ακόλουθα δυναμικά βήματα:

- Προσδιορισμός των πραγματικά χρήσιμων ερωτήσεων/προβλημάτων που πρέπει να αναλυθούν,
- Επιλογή μεταξύ των διαθέσιμων δεδομένων αυτών που πρέπει πραγματικά να χρησιμοποιηθούν προκειμένου να απαντηθούν τα ερωτήματα,
- Προσδιορισμός των χαρακτηριστικών που αναμένουμε να έχουν τα αποτελέσματά μας,
- Ερμηνεία των αποτελεσμάτων της συνολικής διαδικασίας.

Επιπρόσθετα, σημειώνουν ότι κατά την εφαρμογή της διαδικασίας της εξόρυξης δεδομένων θα πρέπει σε αρκετές περιπτώσεις να επαναξιολογούνται προηγούμενα αποτελέσματα από την ανάλυση ίδιων ή παρόμοιων δεδομένων ή ακόμη και να ελέγχονται και τα δεδομένα που χρησιμοποιήθηκαν σε αντίστοιχες περιπτώσεις. Τα αποτελέσματα της εξόρυξης δεδομένων μπορούν να χρησιμοποιηθούν είτε ως τελικά και άρα να θεωρηθούν παράγωγο προϊόν της γνώσης είτε ως εισροές μιας άλλης διαδικασίας/ενός άλλου συστήματος.

Όπως γίνεται αντιληπτό από την έως τώρα περιγραφή, ανεξάρτητα από την επιλεχθείσα μεθοδολογία, εφαρμόζεται πάντα μία συγκεκριμένη διαδικασία με σκοπό την επιλογή της μεθοδολογίας που όντως μπορεί να αποδώσει τα καλύτερα δυνατά και πιο ακριβή αποτελέσματα. Το ερώτημα που εξετάζεται δηλαδή σε όλες τις περιπτώσεις είναι εάν η επιλεχθείσα μεθοδολογία μπορεί να αποδώσει και τα αναμενόμενα αποτελέσματα.

Ωστόσο, έχει εντοπιστεί ένας σημαντικός αριθμός ανασταλτικός αριθμός παραγόντων κατά τη διαδικασία της εξόρυξης δεδομένων, οι οποίοι θεωρείται ότι μπορούν να επηρεάσουν τόσο την ίδια τη διαδικασία όσο τα αποτελέσματα που προκύπτουν από αυτή. Η κυριότερη αδυναμία σύμφωνα με τους Fayyad et al. (1996) είναι ότι οι βάσεις δεδομένων είναι πολύ μεγάλες με αποτέλεσμα να απαιτείται η χρήση αρκετά δύσκολων αλγορίθμων, προκειμένου να μπορέσει να εξαχθεί γνώση από αυτές. Επίσης, τα δεδομένα μπορεί να είναι αρκετά στενά συνδεδεμένα μεταξύ τους με αποτέλεσμα η υπό εξέταση μεταβλητή να μπορεί να μπορεί να ερμηνευθεί από πολλές επιμέρους μεταβλητές. Ένας ακόμη ανασταλτικός παράγοντας είναι και το γεγονός ότι τα δεδομένα μπορεί να είναι ελλιπή ή και να παρουσιάζουν υψηλά επίπεδα θορύβου. Ο τελευταίος ανασταλτικός παράγοντας που παρουσιάζεται στην ίδια μελέτη αφορά στο ότι τα δεδομένα αρκετές φορές μπορεί να χρειάζεται να συνδυαστούν και με άλλα συστήματα προκειμένου να εξαχθούν ασφαλή συμπεράσματα κι έτσι η διαδικασία να γίνεται ακόμη πιο περίπλοκη. Οι Feelders et al. (2000) προσδιορίζουν ακόμη έναν ανασταλτικό παράγοντα στη διαδικασία αυτή ο οποίος αφορά στο ότι η εξόρυξη δεδομένων θα πρέπει να γίνεται από άτομα τα οποία χαρακτηρίζονται ως ειδικοί, αφού αυτοί έχουν την ικανότητα να συνδυάζουν με ιδιαίτερα αποτελεσματικό τρόπο τα δεδομένα μεταξύ τους αλλά και να εντοπίσουν πολύ λεπτά και ιδιαίτερα πρότυπα.

Για να αποφασιστεί εάν τελικά, το υπόδειγμα/μεθοδολογία που χρησιμοποιείται στην εξέταση του προβλήματος, χρησιμοποιείται ένα δοκιμαστικό δείγμα (training set), το οποίο εμπεριέχει τιμές τις οποίες ο ερευνητής γνωρίζει εκ των προτέρων. Μέσω των υπολοίπων χρησιμοποιούμενων μεταβλητών, ο ερευνητής προσπαθεί να εξετάσει μέσω αυτού του δείγματος, εάν τα αποτελέσματα που προέκυψαν από το νέο μοντέλο ανταποκρίνονται στην πραγματικότητα. Η διαδικασία αυτή επιτυγχάνεται με τη χρήση ενός ελέγχου, ο οποίος καθορίζεται εκ των προτέρων και δείχνει την ακρίβεια της προβλεπτικής ικανότητας του υποδείγματος. Πιο συγκεκριμένα, περιγράφονται τα δεδομένα μέσω της συλλογής στατιστικών στοιχείων (όπως η μέση τιμή, η τυπική απόκλιση ή οι συντελεστές συσχέτισης), στη συνέχεια δημιουργείται ένα υπόδειγμα, το οποίο ελέγχει τις σχέσεις που όντως υπάρχουν μεταξύ των μεταβλητών και τέλος ελέγχεται εάν αυτό το υπόδειγμα μπορεί όντως να προβλέψει τις σχέσεις μεταξύ των μεταβλητών, τα πρότυπα μεταξύ των υπό εξέταση δεδομένων και γενικότερα τις μελλοντικές καταστάσεις. Όπως μπορεί να γίνει άμεσα αντιληπτό, η διαδικασία της εξόρυξης δεδομένων μπορεί να χρησιμοποιήσει τεχνικές, που προέρχονται από διάφορους επιστημονικούς κλάδους, μεταξύ των οποίων η επιστήμη των υπολογιστών, η στατιστική, η γενετική, η φυσική και η τεχνητή νοημοσύνη (Ahmed, 2004).

Για την ακριβέστερη πρόβλεψη αυτών των μελλοντικών καταστάσεων έχουν δημιουργηθεί αρκετές μεθοδολογίες εξόρυξης δεδομένων, οι οποίες παρουσιάζονται στη συνέχεια (Fayyad et al.,1996; Shaw et al., 2001). Συνοπτικά αναφέρουμε ότι οι υπό εξέταση μεθοδολογίες είναι:

- Τεχνητά νευρωνικά δίκτυα (Artificial Neural Networks),
- Δέντρα αποφάσεων (Decision Trees),
- Γενετικοί αλγόριθμοι (Genetic Algorithms),
- Επαγωγή αποτελεσμάτων μέσω διαφόρων κανόνων απόφασης (Rule Induction),
- Μέθοδος του κοντινότερου γείτονα (Nearest Neighbor Method),
- Απεικόνιση των δεδομένων (Data Visualization),
- Μοντέλα πρόβλεψης (Predictive Models),
- Συσχετίσεις (Correlations),

- Κατάταξη (Classification),
- Συσταδοποίηση (Clustering),
- Παλινδρόμηση (Regression),
- Περιληπτική παρουσίαση (Descriptives),
- Ανίχνευση μεταβολών και αποκλίσεων (Deviation Analysis).

Όπως θα δούμε στη συνέχεια, όλοι οι αλγόριθμοι της διαδικασίας της εξόρυξης δεδομένων αποτελούνται από κάποιες συναρτήσεις αποτελεσμάτων, κάποια μοντέλα και κάποια μέθοδο βελτιστοποίησης. Έτσι αφού προσδιοριστεί ο στόχος, γίνεται η αναπαράσταση του μοντέλου που θα χρησιμοποιηθεί και αποφασίζεται η συνάρτηση μέσω της οποίας γίνεται η αξιολόγηση. Στη συνέχεια, εφαρμόζεται η διαδικασία της βελτιστοποίησης και διευκρινίζεται ο τρόπος με τον οποίο διαχειρίζονται τα δεδομένα, όπως και οι παράμετροι των μοντέλων. Αυτή η διαδικασία μπορεί να γίνει αντιληπτή και από το παρακάτω διάγραμμα, το οποίο μπορεί να γενικευτεί σε όλες τις εξεταζόμενες μεθοδολογίες.

Στόχος ⇨ Αναπαράσταση ⇨ Συνάρτηση αξιολόγησης ⇨Βελτιστοποίηση ⇨Διαχείριση ⇨Παράμετροι μοντέλων

Στη συνέχεια, παρουσιάζουμε όλες τις μεθοδολογίες εξόρυξης δεδομένων που έχουν παρουσιαστεί στη διεθνή βιβλιογραφία και τα κύρια χαρακτηριστικά τους.

2.1 Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks)

Τα Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks) αποτελούν μία πιο εξειδικευμένη μορφή εξόρυξης δεδομένων που χρησιμοποιεί στοιχεία από την γενετική και ένας σημαντικός αριθμός διαφορετικών ειδών μεταβλητών μπορεί να επεξεργαστεί μέσω αυτών. Ωστόσο, ένα από τα κυριότερα μειονεκτήματα των Τεχνητών Νευρωνικών Δικτύων είναι ότι δεν μπορεί να επεξεργαστεί κατηγορικές

μεταβλητές. Στην πραγματικότητα αυτή η μεθοδολογία στηρίζεται στη βιολογία και ειδικότερα στη γενετική και η κατάλληλη πληροφορία εξάγεται μέσω μιας διαδικασίας εκπαίδευσης (training) κατά την οποία τα δεδομένα προσομοιώνονται στην υπό εξέταση κατάσταση (Mitchie et al., 1994; Quinlan, 1996; Weiss and Kulikowski, 1991).

Τα Τεχνητά Νευρωνικά Δίκτυα αποτελούνται από επιμέρους δίκτυα, τα οποία συνδέονται μεταξύ τους μέσω κόμβων για τους οποίους ισχύει ότι έχουν διαφορετικές μεταξύ τους βαρύτητες, και παράλληλα, αποτελούνται από διαφορετικά επίπεδα, διαφορετικές εισόδους και διαφορετικές εξόδους. Ένα από τα κύρια χαρακτηριστικά των Τεχνητών Νευρωνικών Δικτύων είναι και το ότι αυτά μπορούν να αποτελούνται ακόμη και από κρυφά επίπεδα, μέσα στα οποία μπορούν να υπάρχουν περαιτέρω κόμβοι. Ο αριθμός αυτών των επιπρόσθετων κόμβων θα πρέπει να είναι τέτοιος που να εξασφαλίζει την ακριβή κατηγοριοποίηση των καταστάσεων και να μην είναι ούτε πολύ μεγάλος ούτε και πολύ μικρός.

Η κατηγοριοποίηση που μπορεί να γίνει μέσω των Τεχνητών Νευρωνικών Δικτύων επηρεάζεται από την βαρύτητα του κάθε κόμβου, η οποία υπολογίζεται με έναν πολύ συγκεκριμένο τρόπο, όπως και από την αλληλεπίδραση των κόμβων μεταξύ τους. Ο υπολογισμός των βαρών και των συναρτήσεων ενεργοποίησης του κάθε κόμβου γίνεται μέσω της χρήσης ενός δοκιμαστικού δείγματος (training set) με σκοπό την εξαγωγή ασφαλέστερων συμπερασμάτων. Ο υπολογισμός τόσο των βαρών όσο και των συναρτήσεων ενεργοποίησης του κάθε κόμβου επαναλαμβάνεται τόσες φορές όσες απαιτείται προκειμένου να επιτευχθεί τόσο η σωστή είσοδος όσο και η σωστή έξοδος.

2.2 Δέντρα Αποφάσεων (Decision Trees)

Τα Δέντρα Αποφάσεων αποτελούν μία αρκετά συνηθισμένη τακτική εξόρυξης δεδομένων από τη στιγμή που χαρακτηρίζονται από σχετική απλότητα σε σύγκριση με άλλες μεθοδολογίες, ενώ και η παρουσίασή τους είναι αρκετά εύκολη και κατανοητή. Αυτό συμβαίνει γιατί δεν απαιτείται μεγάλη υπολογιστική δύναμη και τα αποτελέσματα μπορούν να προκύψουν είτε εάν έχει χρησιμοποιηθεί φυσική είτε

προγραμματιστική γλώσσα. Παράλληλα, στα Δέντρα Αποφάσεων μπορούν να χρησιμοποιηθούν τόσο συνεχείς όσο και διακριτές μεταβλητές.

Ένα από τα κύρια μειονεκτήματά τους είναι το ότι δεν μπορούν να δώσουν ιδιαίτερα εξειδικευμένα αποτελέσματα. Αυτό συμβαίνει αφού σε αρκετές περιπτώσεις είναι ιδιαίτερα δύσκολο να ληφθούν υπόψη όλα τα κλαδιά λόγω των περίπλοκων υπολογισμών, κι έτσι το πεδίο εξέτασης περιορίζεται αρκετά.

Τα Δέντρα Αποφάσεων αποτελούνται από διαφορετικά σενάρια, τα οποία ενσωματώνονται σε διαφορετικά κλαδιά, τα οποία με τη σειρά τους επηρεάζουν την τελική απόφαση ανάλογα με το κλαδί που επιλέγεται. Σε κάθε αρχικό κλαδί, αντιστοιχεί και ένα διαφορετικό σενάριο δηλαδή, το οποίο μπορεί να αποτελείται και από επιπρόσθετα κλαδιά (συνθήκες του σεναρίου), τα οποία μπορούν να επηρεάσουν με τη σειρά τους τα αποτελέσματα του κάθε σεναρίου ή ακόμη και να διακόψουν τη διαδικασία εξόρυξης δεδομένων εάν είναι τέτοιος ο σχεδιασμός τους.

2.3 Μέθοδος του Κοντινότερου Γείτονα (Nearest Neighbor Method)

Η Μέθοδος του Κοντινότερου Γείτονα αποτελεί μία ακόμη πιο απλή μεθοδολογία εξόρυξης δεδομένων και στηρίζεται στη μελέτη παρόμοιων περιπτώσεων με την εξεταζόμενη κατάσταση. Με αυτό τον τρόπο τα αποτελέσματα εξάγονται μέσω της μελέτης αυτών των καταστάσεων, ενώ στη συνέχεια αυτές οι καταστάσεις ταξινομούνται με σκοπό να βρεθεί όντως αυτή που είναι πιο κοντά στην εξεταζόμενη.

Πιο συγκεκριμένα, χρησιμοποιείται ένα σύνολο δεδομένων και ένα μέτρο, το οποίο καθορίζει την απόσταση μεταξύ των διαφόρων περιπτώσεων. Ωστόσο, το δυσκολότερο κομμάτι σε αυτή τη μεθοδολογία είναι η απόφαση για τον αριθμό των κοντινότερων γειτόνων. Αυτό σημαίνει ότι θα πρέπει να αποφασιστεί μία τιμή για αυτόν τον παράγοντα που να καθορίζει τους απαιτούμενους κοντινότερους γείτονες.

2.4 Συσχετίσεις (Correlations)

Οι συσχετίσεις εξετάζουν τον τρόπο με τον οποίο συνδέονται οι διάφορες μεταβλητές μεταξύ τους. Οι συσχετίσεις μπορεί να εντοπιστούν ανάμεσα στις εξεταζόμενες μεταβλητές, αλλά και να συνδέσουν όλες ή κάποιες από τις εξεταζόμενες μεταβλητές με κάποιο άλλο γεγονός. Η συσχέτιση επί της ουσίας μετράει το πόσο δύο μεταβλητές κυμαίνονται ή όχι προς την ίδια κατεύθυνση.

Το σημαντικότερο πρόβλημα που αντιμετωπίζεται κατά τον υπολογισμό των συσχετίσεων είναι ότι είναι ότι προκειμένου να υπολογιστούν αυτές θα πρέπει οι εξεταζόμενες μεταβλητές να ανήκουν στην ίδια κατηγορία. Δηλαδή, για να υπολογιστεί η συσχέτιση πρέπει οι μεταβλητές να ανήκουν στον ίδιο τύπο, για παράδειγμα εάν είναι συνεχής (διακριτή) η μία μεταβλητή θα πρέπει και η άλλη να είναι συνεχής (διακριτή).

Οι συσχετίσεις επί της ουσίας ανήκουν στην κατηγορία της μοντελοποίηση; εξαρτήσεων, όπου περιγράφονται οι στατιστικά σημαντικές σχέσεις ανάμεσα στις εξεταζόμενες μεταβλητές, δηλαδή εντοπίζονται τα στοιχεία που έχουν κοινά χαρακτηριστικά μεταξύ τους. Οι σχέσεις αυτές εντοπίζονται σε δύο επίπεδα. Σε πρώτο επίπεδο, γίνεται η γραφική απεικόνιση που δείχνει την εξάρτηση που υπάρχει ανάμεσα στις μεταβλητές και σε δεύτερο επίπεδο, η σχέση αυτή ποσοτικοποιείται μέσω ενός μέτρου.

2.5 Ακολουθίες - Επαγωγή Αποτελεσμάτων μέσω Διάφορων Κανόνων Απόφασης (Rule Induction)

Οι ακολουθίες συνδέονται άμεσα με την μεθοδολογία της Επαγωγής Αποτελεσμάτων μέσω Διάφορων Κανόνων Απόφασης. Επί της ουσίας αναφερόμαστε στην παρατήρηση ότι όταν ένα φαινόμενο εμφανιστεί, τότε κάποιο άλλο θα ακολουθήσει, δηλαδή ότι κάποιες πληροφορίες ή κάποια γεγονότα μπορούν να προκύψουν ως αποτέλεσμα άλλων.

Η Επαγωγή Αποτελεσμάτων μέσω Διαφόρων Κανόνων Απόφασης, βασίζεται στη λογική της επαγωγικής διαδικασίας. Δηλαδή εξετάζεται το ερώτημα σχετικά με το ποιο θα είναι το αποτέλεσμα μιας εάν συμβεί κάποιο γεγονός. Πιο συγκεκριμένα, η μεθοδολογία βασίζεται στη λογική του «εάν» συμβεί κάτι «τότε» τι θα ακολουθήσει. Αφού εξεταστούν όλα τα σενάρια, τότε τα αποτελέσματα κατατάσσονται ανάλογα με τη σημαντικότητά τους και τη συνάφειά τους με το εξεταζόμενο θέμα και επιλέγεται κάθε φορά το καταλληλότερο.

2.6 Κατάταξη (Classification)

Σύμφωνα με τους Weiss and Kulikowski (1991) και Hand (1981) η κατάταξη παρουσιάζει τα αποτελέσματα που εξάγονται από τα εξεταζόμενα δεδομένα σε μία ή περισσότερες τάξεις με σκοπό την εξαγωγή ασφαλέστερων, σωστότερων και πιο χρήσιμων αποφάσεων.

Τα δεδομένα σε αυτή την περίπτωση κατατάσσονται σε ομάδες (κλάσεις) αφού πρώτα έχουν χωριστεί σε άλλες ομάδες. Δηλαδή, τα υπό εξέταση δεδομένα κατατάσσονται σε πρώτο στάδιο σε ομάδες και στη συνέχεια τα αποτελέσματα που προκύπτουν από αυτά κατατάσσονται εκ νέου σε ομάδες. Οι Elovici and Braha (2003) προτείνουν ένα υπόδειγμα, στο οποίο οι κατατάξεις, $Y = \{y_1, \dots, y_n\}$, προκύπτουν μέσω κάποιων κατηγοριών $S = \{s_1, \dots, s_n\}$, οι οποίες χαρακτηρίζονται από τις αντίστοιχες πιθανότητες που είναι $\pi = (\pi_1, \dots, \pi_{ns})$. Σε αυτή την περίπτωση, η εξόρυξη δεδομένων έχει ως αντικείμενο την παρατήρηση της τιμής $X = x$, η οποία έχει προκύψει ως αποτέλεσμα της κατάταξης $Y = \{y_1, \dots, y_n\}$. Έτσι, προκύπτει ένας πίνακας, ο οποίος συμπεριλαμβάνει όλες τις πιθανές τιμές λαμβάνοντας υπόψη όλες τις πιθανότητες και τις εναλλακτικές, και από τον οποίο προκύπτει το τελικό αποτέλεσμα.

2.7 Συσταδοποίηση (Clustering)

Η συσταδοποίηση μοιάζει αρκετά με την διαδικασία της κατάταξης. Ωστόσο, σε αντίθεση με την κατηγοριοποίηση, εδώ τα δεδομένα ταξινομούνται σε μεγάλες ομάδες ανάλογα με το βαθμό ομοιογένειας που παρουσιάζουν μεταξύ τους, έχοντας ως σκοπό την εξέταση των επιμέρους ζητημάτων, τα οποία με τη σειρά τους επηρεάζουν το κεντρικό ερώτημα (Jain and Dubes, 1988; Titterington et al. 1985). Εφόσον η ομαδοποίηση θεωρηθεί ως επιτυχής, τότε τα αποτελέσματα που θα προκύψουν θα θεωρηθούν πιο ακριβή.

Οι ομάδες (clusters) σε αυτή την περίπτωση δημιουργούνται με τέτοιο τρόπο έτσι ώστε να επιτυγχάνεται ταυτόχρονα η ομοιότητα των χαρακτηριστικών των δεδομένων εντός της ίδιας ομάδας, αλλά και η διαφορετικότητα μεταξύ των επιμέρους ομάδων. Όπως καταλαβαίνουμε, ο υπολογισμός του αριθμού των ομάδων και η ομοιότητα των χαρακτηριστικών είναι αρκετά δύσκολο να επιτευχθεί.

Ένας τρόπος υπολογισμού των ομάδων είναι μία τεχνική αυτόματης εύρεσής τους. Το κυριότερο πλεονέκτημα είναι ότι αυτή αποτελεί μία σχετικά εύκολη διαδικασία, για την οποία δεν απαιτείται η ανθρώπινη παρέμβαση και μπορεί να εφαρμοστεί ακόμη και από κάποιον που δεν θεωρείται ότι είναι ο κύριος γνώστης του αντικειμένου. Επίσης, είναι πιο εύκολο ακόμη και σε δεδομένα διαφορετικών τύπων να δημιουργηθούν ομάδες. Αυτό συμβαίνει γιατί δεν απαιτείται κάποια προγενέστερη επεξεργασία των δεδομένων ούτε και ο διαχωρισμός των μεταβλητών σε ανεξάρτητες και εξαρτημένες, αλλά ούτε και σε δεδομένα εισόδου ή εξόδου, όπως συμβαίνει στα Τεχνητά Νευρωνικά Δίκτυα.

Ωστόσο, όπως και στις προηγούμενες περιπτώσεις και σε αυτή την περίπτωση υπάρχουν κάποια σημαντικά μειονεκτήματα. Το κυριότερο μειονέκτημα είναι ότι είναι αρκετά πιθανό να γίνει λάθος εκτίμηση στο μέτρο των αποστάσεων για την κάθε ομάδα, αφού εάν δεν επιλεγεί από τον χρήστη ο σωστός αριθμός των ομάδων, τότε είναι πολύ πιθανό να οδηγηθούμε σε λανθασμένα αποτελέσματα. Ένα ακόμη σημαντικό μειονέκτημα αυτής της μεθοδολογίας είναι ότι όσο αυξάνονται τα χαρακτηριστικά των υπό κατηγοριοποίηση μεταβλητών τόσο πιο δύσκολη είναι αυτή η διαδικασία και σε αυτό το πρόβλημα η λύση έρχεται μέσω της διαδικασίας αυτόματης εύρεσης των ομάδων. Άλλο ένα πρόβλημα που αντιμετωπίζουμε είναι η

επιλογή του αριθμού των τάξεων. Προκειμένου να επιλεχθεί ο σωστός αριθμός των clusters θα πρέπει να δοκιμαστούν διάφορες τιμές και να επιλεγεί η τιμή που θα δίνει ομογενείς τάξεις με ίσες μεταξύ τους αποστάσεις. Είναι επίσης πολύ πιθανό να βρεθούν περισσότερες από μία ιδανικές τιμές.

2.8 Παλινδρόμηση (Regression)

Η παλινδρόμηση αποτελεί μία διαδικασία κατά την οποία η υπό εξέταση μεταβλητή συνδέεται με άλλες επεξηγηματικές μεταβλητές, οι οποίες θεωρείται ότι μπορούν να ερμηνεύσουν, αλλά και να προβλέψουν την εξαρτημένη μεταβλητή. Πιο συγκεκριμένα, στην παλινδρόμηση θεωρείται ότι σε μία τυχαία μεταβλητή (εξαρτημένη), ο τρόπος συμπεριφοράς της, καθορίζεται από τον τρόπο συμπεριφοράς κάποιας άλλης ή κάποιων άλλων τυχαίων μεταβλητών (ανεξάρτητες). Αυτό που επιθυμούμε είναι μέσω αυτής της διαδικασίας να προσδιορίσουμε τη σχέση που υπάρχει μεταξύ των μεταβλητών με σκοπό να εξαχθούν εκτιμήσεις για τη μελλοντική πορεία της υπό εξέταση μεταβλητή. Η παλινδρόμηση αφορά στην εκτίμηση ενός υποδείγματος που αφορά είτε διαστρωματικά δεδομένα είτε χρονολογικές σειρές.

Η παλινδρόμηση λαμβάνει τη μορφή:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots \ldots + \beta_n x_{nt} + \varepsilon_t$$

Όπου y_t είναι η εξαρτημένη μεταβλητή, $x_{1t}, \ldots, x_{nt}$ είναι οι ανεξάρτητες μεταβλητές, $\beta_0, \ldots, \beta_n$ είναι οι συντελεστές που πρέπει να εκτιμηθούν και ε_t τα κατάλοιπα του υποδείγματος. Επί της ουσίας, τα κατάλοιπα αποτελούνται από όλες τις υπόλοιπες μεταβλητές, οι οποίες δεν συμπεριλαμβάνονται στο υπόδειγμα και αναφέρονται στο κομμάτι της συνολικής ερμηνευτικής ικανότητας του υποδείγματος, το οποίο δεν προσδιορίζεται. Συνεπώς, στόχος του ερευνητή είναι και να καταφέρει να ελαχιστοποιήσει την τιμή των καταλοίπων προκειμένου να εξασφαλίσει τη μεγαλύτερη ερμηνευτικότητα του υποδείγματος. Η συνηθέστερη μέθοδος εκτίμησης των συντελεστών του υποδείγματος είναι αυτή των ελαχίστων τετραγώνων (Least Squares Method). Αφού εκτιμηθούν οι παράμετροι του υποδείγματος, στη συνέχεια

ελέγχεται η στατιστική σημαντικότητά τους. Η υπό έλεγχο υπόθεση αφορά στη μηδενική υπόθεση ότι κάποια από τις παραμέτρους ισούται με το μηδέν έναντι της εναλλακτικής ότι είναι διαφορετική από το μηδέν:

$$H_0: \beta_0 = \beta_1 = \cdots \ldots = \beta_{nt} = 0$$

$$H_1: \beta_0 \neq 0 \text{ ή } \beta_1 \neq 0 \text{ ή } \ldots\ldots \beta_{nt} \neq 0$$

Στη συνέχεια ελέγχεται η συνολική ερμηνευτική ικανότητα του υποδείγματος μέσω της στατιστικής ελέγχου F, αλλά και το κατά πόσο η συνολική μεταβλητότητα του υποδείγματος ερμηνεύεται μέσω των ανεξάρτητων μεταβλητών και όχι μέσω των καταλοίπων. Αυτή η μέτρηση γίνεται μέσω του συντελεστή προσδιορισμού (R^2), ο οποίος υποδηλώνει το ποσοστό της μεταβλητότητας του υποδείγματος που εξηγείται από την παλινδρόμηση και είναι ίσος με:

$$R^2 = \frac{RSS}{TSS}$$

Όπου RSS είναι η μεταβλητότητα του υποδείγματος που ερμηνεύεται από τις ανεξάρτητες μεταβλητές του υποδείγματος και TSS η συνολική μεταβλητότητα του υποδείγματος. Ο συντελεστής προσδιορισμού λαμβάνει τιμές από 0 – 1, και οι μεγαλύτερες τιμές δείχνουν και την καλύτερη ερμηνευτική ικανότητα της μεταβλητότητας του υποδείγματος.

Συνολικά, μπορούμε να πούμε ότι με βάση και το διάγραμμα 1 ο στόχος σε αυτή τη μεθοδολογία είναι η εφαρμογή της παλινδρόμησης, η οποία αποτυπώνεται με βάση μία εξίσωση η οποία πρέπει να εκτιμηθεί μέσω μιας συνάρτησης αξιολόγησης (π.χ. Μέθοδος των Ελαχίστων Τετραγώνων), όπου στη συνέχεια εκτιμώνται οι παράμετροι μέσω μιας διαδικασίας βελτιστοποίησης και προκύπτουν οι συντελεστές του υποδείγματος.

2.9 Λοιπές Μεθοδολογίες Εξόρυξης Δεδομένων

Εκτός των μεθοδολογιών που αναλύθηκαν παραπάνω υπάρχουν κάποιες ακόμη μεθοδολογίες που αναλύονται σε αυτή την ενότητα, οι οποίες είναι: οι Γενετικοί Αλγόριθμοι (Genetic Algorithms), η Απεικόνιση των Δεδομένων (Data Visualization) και τα Μοντέλα Πρόβλεψης (Predictive Models).

Οι Γενετικοί Αλγόριθμοι βασίζονται, όπως προκύπτει και από την ονομασία τους, στην επιστήμη της Γενετικής. Τα αποτελέσματα προκύπτουν μέσω της διαδικασίας της φυσικής επιλογής των δεδομένων και της μεταξύ τους μετάλλαξης. Μία ακόμη μεθοδολογία εξόρυξης δεδομένων είναι η γραφική Απεικόνιση των Δεδομένων, μέσω της οποίας οι όποιες σχέσεις και τάσεις μεταξύ των υπό εξέταση μεταβλητών εμφανίζονται μέσω της εικόνας των δεδομένων. Δηλαδή, προκειμένου να εξηγηθεί ένα υπό εξέταση πρόβλημα/ζήτημα δημιουργούνται απλά γραφικές αναπαραστάσεις των δεδομένων. Η περιληπτική παρουσίαση των αποτελεσμάτων των δεδομένων περιλαμβάνει μεμονωμένα στοιχεία, τα οποία περιγράφουν την τάση μεταξύ των εξεταζόμενων μεταβλητών μέσω του υπολογισμού απλών μέτρων, όπως ο μέσος όρος και η τυπική απόκλιση (Agrawal et al., 1996).

Αντίστοιχα, εντοπίζεται ακόμη μία μεθοδολογία, η οποία ανιχνεύει μεταβολές και αποκλίσεις, εντοπίζοντας τις μεγάλες και ασυνήθιστες μεταβολές στις τιμές των δεδομένων σε σχέση με τις προηγούμενες μετρήσεις. Επιπρόσθετα, οι Elovici and Braha (2003) προτείνουν ένα συνδυαστικό σχήμα προκειμένου να προβούν στην αποτελεσματικότερη και πιο σωστή εξόρυξη δεδομένων. Σε αυτό το σχήμα συνδυάζουν ένα Τεχνητό Νευρωνικό Δίκτυο με ένα Δέντρο Αποφάσεων. Για το συνδυασμό αυτό χρησιμοποιούν δύο διαφορετικές μεθοδολογίες, η πρώτη βασίζεται στο καρτεσιανό γινόμενο και η δεύτερη στην κατηγοριοποίηση του κάθε συστήματος. Κύρια προϋπόθεση και των δύο μεθοδολογιών είναι και τα δύο σχήματα να μην δίνουν χειρότερα αποτελέσματα από τις επιμέρους μεθοδολογίες.

3. Εξόρυξη Δεδομένων και Ηθική

Η διαδικασία της εξόρυξης δεδομένων πολλές φορές επιδέχεται σκληρή κριτική εξαιτίας του γεγονότος ότι τα δεδομένα και τα αποτελέσματα που προκύπτουν από αυτά μπορεί να συνδέονται άμεσα με την προσωπική κρίση και τις πτυχές της ζωής των μεμονωμένων ατόμων. Σε αυτό το κεφάλαιο, στόχος μας είναι να ελέγξουμε βασιζόμενοι στη διεθνή βιβλιογραφία τον τρόπο με τον οποίο οι δύο αυτές έννοιες – ερμηνείες συνδέονται μεταξύ τους, αφού πρώτα δώσουμε τον ορισμό της ηθικής.

3.1 Ορισμός της Ηθικής

Η ηθική αποτελεί περισσότερο μία έννοια φιλοσοφική παρά έναν επιστημονικό κλάδο. Γι' αυτό τον λόγο, δεν μπορεί να αποδοθεί ένας γενικά αποδεκτός ορισμός για την ηθική, αφού θεωρείται ότι αποτελεί μία πολυδιάστατη έννοια. Ωστόσο, μπορούμε να πούμε ότι η ηθική ορίζεται όταν απαντάται το ερώτημα σχετικά με το ποιες πράξεις των ατόμων είναι κοινά αποδεκτές και θεωρούνται ως σωστές και το ποιες είναι μη αποδεκτές και θεωρούνται λανθασμένες.

Στο ίδιο συμπέρασμα καταλήγουν και οι Robert and Iles (2014), ορίζοντας την ηθική ως την φιλοσοφία που περιλαμβάνει την αντίληψη της κοινωνίας, και των ατόμων γενικότερα, τόσο της καλής όσο και της κακής συμπεριφοράς.

Η ηθική, γενικότερα, επηρεάζεται από την κοινωνία και τις αντιλήψεις που επικρατούν σε αυτήν, και οι οποίες μπορεί να θεωρηθούν ότι εξελίσσονται και αλλάζουν διαρκώς. Σε αυτή την αλλαγή συνδράμουν παράγοντες που σχετίζονται με το γενικότερο περιβάλλον της κοινωνίας στην οποία αυτή η ηθική υπάρχει και συμπεριλαμβάνονται μεταξύ άλλων η τεχνολογία, ο πολιτισμός και το γεωγραφικό περιβάλλον μιας κοινωνίας. Οι Robert and Iles (2014) σημειώνουν επιπλέον ότι η αντίληψη για την ηθική και τα ζητήματα που σχετίζονται με αυτήν μπορούν να είναι διαφορετικά τόσο σε επίπεδο χωρών όσο και ανάμεσα στα άτομα που κατοικούν στην ίδια χώρα.

Ωστόσο, θα πρέπει να αναφέρουμε ότι υπάρχουν κάποιοι κανόνες που διέπουν την ηθική γενικότερα σε παγκόσμιο επίπεδο. Πιο συγκεκριμένα, αρχές όπως η ελευθερία

του ατόμου, αφορούν σε έννοιες που σχετίζονται με την ηθική και έχουν καθιερωθεί ακόμη και μέσω διεθνών συμβάσεων.

3.2 Σύνδεση της Ηθικής με την Εξόρυξη Δεδομένων

Ο Seltzer (2005), υποστηρίζει, ότι η εξόρυξη δεδομένων αποτελεί στην πραγματικότητα ακόμη μία στατιστική διαδικασία, η οποία οδηγεί στην εξαγωγή κάποιων αποτελεσμάτων και ότι όπως όλες οι επιστήμες θα πρέπει και αυτή να είναι ηθικά ουδέτερη. Ωστόσο, όπως σημειώνουν οι Fule and Roddick (2004), η διαδικασία της εξόρυξης των δεδομένων και η ανακάλυψη γνώσης στα δεδομένα οδηγούν στην εξαγωγή κάποιων συμπερασμάτων, τα οποία με τη σειρά τους επηρεάζουν τη λήψη αποφάσεων, οι οποίες στη συνέχεια επηρεάζουν την ανθρώπινη ζωή είτε με άμεσο είτε με έμμεσο τρόπο. Από τη στιγμή που αυτές οι αποφάσεις επηρεάζουν τα άτομα, γίνεται εύκολα αντιληπτό ότι οι δύο αυτές διαδικασίες αποτελούν ένα σημαντικό ηθικό ζήτημα. Περαιτέρω, ο Zarsky (2003) υποστηρίζει ότι η εξόρυξη δεδομένων είναι άμεσα συνδεδεμένη με την ηθική από τη στιγμή που πλέον τα δεδομένα συλλέγονται μέσω της διαρκούς παρακολούθησης, η οποία προκύπτει και ως αποτέλεσμα της εξελισσόμενης τεχνολογίας.

Όπως γίνεται αντιληπτό, η εξόρυξη δεδομένων έχει ιδιαίτερα σημαντικές κοινωνικές επιπτώσεις από την στιγμή που τα δεδομένα που χρησιμοποιούνται είναι αυστηρά προσωπικά και πολλές φορές έχουν προκύψει με έμμεσο τρόπο. Αυτό σημαίνει ότι εάν κάποιος καταναλωτής, για παράδειγμα, πραγματοποιεί τις αγορές του μέσω διαδικτύου, τότε μέσω της φόρμας συμπλήρωσης της παραγγελίας του είναι πάρα πολύ πιθανό να αντλούνται στοιχεία όπως το φύλο, η ηλικία, ο αριθμός τηλεφώνου ή η ηλεκτρονική του διεύθυνση, αλλά και οι καταναλωτικές του συνήθειες. Με αυτό τον τρόπο, μπορεί για λαμβάνονται στοιχεία και κάθε φορά που ο καταναλωτής επισκέπτεται τη συγκεκριμένη σελίδα, να εμφανίζονται μηνύματα που να τον προτρέπουν στην αγορά συγκεκριμένων προϊόντων.

Ωστόσο, ο Zarsky (2003), υποστηρίζοντας, όπως αναφέραμε και παραπάνω, ότι τα δεδομένα συλλέγονται μέσω της διαρκούς παρακολούθησης, επισημαίνει ότι αυτή η

παρακολούθηση είναι ανούσια εάν τα δεδομένα αυτά δεν καταγραφούν σε κάποια βάση και δεν αναλυθούν. Όπως αναφέρει η καταγραφή των δεδομένων είναι μία απλή διαδικασία από την στιγμή που η τεχνολογία μπορεί να βοηθήσει σε αυτή την κατεύθυνση, αλλά η ανάλυση μπορεί να χαρακτηριστεί ως ιδιαίτερα δύσκολη από την στιγμή που ο όγκος των δεδομένων είναι τόσο μεγάλος που εμποδίζει το σύνολο της διαδικασίας.

Τον διαχωρισμό των ηθικών ζητημάτων που σχετίζονται με την ηθική σε άμεσα και έμμεσα έκαναν και οι Robert and Iles (2014). Μάλιστα, στα άμεσα ζητήματα συμπεριέλαβαν παράγοντες όπως η απώλεια της μυστικότητας των προσωπικών δεδομένων, η λανθασμένη ερμηνεία των πληροφοριών και η πιθανώς προβληματική ανώνυμη πληροφορία, ενώ στα έμμεσα ζητήματα τον τρόπο συλλογής των δεδομένων, όπως και το ερώτημα σχετικά με το σε ποιον ανήκει η κάθε πληροφορία, που αποθηκεύεται και για ποιους λόγους χρησιμοποιείται με αυτό τον τρόπο.

Ο Zarsky (2003) διαπιστώνει ότι η εξόρυξη δεδομένων παραβιάζει την ηθική των ατόμων, αλλά και τον ιδιωτικό/προσωπικό χαρακτήρα των στοιχείων μέσω των ακόλουθων τρόπων:

- Διάκριση των καταναλωτών σε επιμέρους κατηγορίες/ομάδες: στην πραγματικότητα, οι καταναλωτές ανάλογα με τις προηγούμενες προτιμήσεις τους (αγορές) αλλά και τα προσωπικά/δημογραφικά χαρακτηριστικά τους χωρίζονται από τις επιχειρήσεις σε διαφορετικές ομάδες με σκοπό να εστιάζουν σε διαφορετικά καταναλωτικά κοινά τα προϊόντα τους. Μία τέτοια διάκριση μπορεί να έχει αποτελέσματα όπως η άδικη και υποτιμητική συμπεριφορά από τους πωλητές από τη στιγμή που αυτοί θα τείνουν να εξυπηρετήσουν με τον καλύτερο δυνατό τρόπο μόνον όσους από αυτούς ανήκουν σε κάποια από τις ομάδες – στόχους. Σε αυτό το σημείο μπορούμε να αναφέρουμε ότι πρόκειται για μία αποτυχία της αγοράς, από τη στιγμή που υπάρχει ασύμμετρη πληροφόρηση μεταξύ των δύο ευρύτερων ομάδων με αποτέλεσμα την απώλεια σε όρους συνολικής ευημερίας,
- Χειραγώγηση των καταναλωτών και παρέμβαση στην αυτονομία τους: όπου οι καταναλωτές με αυτόν τον τρόπο μπορεί να οδηγηθούν στην κατανάλωση συγκεκριμένων προϊόντων ακόμη κι εάν αυτά δεν είχαν την πρόθεση να τα

αγοράσουν. Με αυτόν τον τρόπο δημιουργείται ένας φαύλος κύκλος που ωθεί τα άτομα σε συγκεκριμένα προϊόντα, με αποτέλεσμα σημαντικές επιπτώσεις στο σύνολο της κοινωνίας. Οι επιπτώσεις αυτές συναντώνται ξανά στα πλαίσια της απώλειας ευημερίας, όπου το μεγαλύτερο μέρος της απώλειας αυτής εντοπίζεται στην πλευρά του καταναλωτή,

- Κατάχρηση και κακή χρήση των δεδομένων που συλλέγονται: σε αυτή την περίπτωση τα δεδομένα που έχουν συλλεχθεί χρησιμοποιούνται με σκοπό μόνο το κέρδος από την πλευρά των επιχειρήσεων με αποτέλεσμα τα άτομα να οδηγούνται στην προσβολή της προσωπικότητάς τους αλλά και την γενικότερη απομόνωση όσων από αυτά δεν ανήκουν στην μεγάλη μάζα των καταναλωτών που αποτελούν την ομάδα στόχο,
- Τραγωδία των λαθών: όπως σε κάθε περίπτωση έτσι και κατά τη συλλογή δεδομένων και ιδίως όταν αυτά αφορούν σε προσωπικά στοιχεία, είναι πολύ πιθανό είτε αυτά να συλλεχθούν με τον λάθος τρόπο είτε να αναλυθούν με ακατάλληλο αλγόριθμο. Αυτό θα έχει ως αποτέλεσμα ξανά την απώλεια σε όρους ευημερίας για το σύνολο της κοινωνίας (τόσο τους καταναλωτές όσο και τις επιχειρήσεις).

Ο Seltzer (2005) θεωρεί ότι από την εφαρμογή της ηθικής στη διαδικασία της εξόρυξης δεδομένων προκύπτουν τρία σημαντικά ζητήματα που σχετίζονται με την εφαρμογή της στη διαδικασία αυτή, τα οποία είναι:

- Καταλληλότητα και Εγκυρότητα: αναφερόμαστε στην εφαρμογή της κατάλληλης μεθοδολογίας εξόρυξης δεδομένων, η οποία μπορεί να οδηγήσει στην παραγωγή έγκυρων αποτελεσμάτων,
- Εμπιστευτικότητα και Μυστικότητα σε σχέση με τα προσωπικά δεδομένα: όπου τα προσωπικά δεδομένα που αφορούν ένα άτομο θα πρέπει να διαχειρίζονται με τέτοιο τρόπο έτσι ώστε να εξασφαλίζεται η μυστικότητα σχετικά με τις απαντήσεις/παρατηρήσεις που δόθηκαν από τα άτομα,
- Εφαρμογή της διαδικασίας προς όφελος της κοινωνίας και αποφυγή πράξεων που μπορούν να την βλάψουν: δηλαδή αναφερόμαστε στην εφαρμογή της διαδικασίας με τέτοιο τρόπο έτσι ώστε η κοινωνία να είναι σε θέση να επωφελείται από τα αποτελέσματα και να μην ζημιώνεται.

Όπως αναφέραμε και σε προηγούμενο κεφάλαιο, η συσσώρευση των δεδομένων μπορεί να επιφέρει σημαντικά αποτελέσματα στις επιχειρήσεις αφού μπορεί να επηρεάσει τις αποφάσεις, οι οποίες στη συνέχεια επηρεάζουν τη λειτουργία τους, την αποτελεσματικότητά τους και τα κέρδη τους. Ο Busovsky (2007) υποστήριξε ότι η δύναμη των δεδομένων μεγαλώνει διαρκώς από τη στιγμή που οι επιχειρήσεις αντιλαμβάνονται πλέον τη δύναμη της πληροφορίας η οποία εξάγεται από άμορφα και ακατέργαστα δεδομένα που συσσωρεύονται σε μεγάλες βάσεις δεδομένων. Όπως αναφέρει, υπάρχουν αρκετά ηθικά ζητήματα που σχετίζονται με την εξόρυξη δεδομένων. Ένα από αυτά σχετίζεται με την μυστικότητα/ιδιωτικότητα που πρέπει να υπάρχει στα δεδομένα όταν αυτά χρησιμοποιούνται ακόμη και για σκοπούς όπως η εθνική ασφάλεια και η αποφυγή μελλοντικών τρομοκρατικών επιθέσεων. Σε αυτή την κατεύθυνση κυβερνήσεις των ΗΠΑ, συνδύασαν δεδομένα που αντλήθηκαν τόσο από ιδιωτικές όσο και δημόσιες βάσεις. Ωστόσο, το ερώτημα που τίθεται σε αυτή την περίπτωση είναι εάν τα άτομα προτίθενται να χάσουν μέρος της ιδιωτικής τους ζωής προς όφελος της ασφάλειας τόσο των ίδιων όσο και του συνόλου της κοινωνίας.

Από τη στιγμή που η ηθική περιλαμβάνει ένα σύνολο κανόνων και αξιών που καθορίζουν τη συμπεριφορά των ατόμων και των επιχειρήσεων με στόχο το κοινωνικό όφελος, μπορούμε εύκολα να αντιληφθούμε ότι η δημοσιοποίηση μιας αρνητικής γνώσης μπορεί να οδηγήσει σε πολλαπλά αρνητικά αποτελέσματα, τα οποία οδηγούν είτε σε απώλειες ευημερίας είτε ακόμη και σε επιβλαβείς και ανεπανόρθωτες συνέπειες για το σύνολο της κοινωνίας (Fule and Roddick, 2004). Ωστόσο, για να αποφευχθούν τέτοιες αρνητικές συνέπειες πολλές φορές είναι απαραίτητο να άρεται η μυστικότητα των προσωπικών δεδομένων, κάτι το οποίο δεν είναι πάντοτε αποδεκτό από το σύνολο των ατόμων της κοινωνίας.

Ο λόγος για τον οποίο δεν είναι πάντοτε αποδεκτή η άρση του απορρήτου των προσωπικών δεδομένων, είναι, όπως, καταλαβαίνουμε, το γεγονός ότι η ευαισθησία του κάθε ατόμου ως προς την ηθική πλευρά της εξόρυξης δεδομένων δεν μπορεί να μετρηθεί με μία κοινή κλίμακα από τη στιγμή που αποτελεί καθαρά υποκειμενικό ζήτημα. Αυτό συμβαίνει γιατί μπορεί όλα τα άτομα να μην αντιμετωπίζουν με τον ίδιο τρόπο την κουλτούρα της κοινωνίας στην οποία ζουν.

Οι Estivill – Castro et al. (1999) υποστήριξαν ότι πολλές φορές τα αποτελέσματα της εξόρυξης δεδομένων είτε δεν είναι ακριβή είτε δεν πληρούν την προϋπόθεση του

σεβασμού της ιδιωτικής ζωής και των προσωπικών δεδομένων των ατόμων. Σε αυτή την κατεύθυνση παρουσιάζουν μία σειρά μελετών που σχετίζονται με αυτό τα ζήτημα που είτε υποστηρίζουν αυτή την άποψη είτε όχι. Οι Clifton et al. (2002), οι Wahlstrom et al. (2006) και ο Okur (2008) σημείωσαν, ότι η διατήρηση της μυστικότητας/ιδιωτικότητας των προσωπικών δεδομένων μπορεί να επιτευχθεί, ακόμα και όταν αυτά χρησιμοποιούνται στην εξέταση ενός φαινομένου, μέσω των ακόλουθων τρόπων:

- Ασφαλής ανταλλαγή των δεδομένων ανάμεσα στους φορείς που τα κατέχουν (δημόσιους ή ιδιωτικούς), ακόμη και ανάμεσα στις ίδιες τις επιχειρήσεις χωρίς να ενισχύεται ο ανταγωνισμός μεταξύ τους: δηλαδή, ακόμη και όταν οι επιχειρήσεις ανταλλάσουν δεδομένα μεταξύ τους, η ανταλλαγή αυτή θα πρέπει να γίνεται με γνώμονα το κοινωνικό όφελος και χωρίς να χρησιμοποιούνται αυτά τα δεδομένα με σκοπό την «εξολόθρευση» των ανταγωνιστών,
- Ανωνυμία των ιδιωτικών/προσωπικών δεδομένων: δηλαδή, όταν τα άτομα δίνουν μέρος των προσωπικών τους στοιχείων, θα πρέπει να εξασφαλίζεται ότι αυτά χρησιμοποιούνται αθροιστικά με άλλα όμοια στοιχεία και δεν μπορούν να αποδοθούν σε αυτά με κάποιον τρόπο. Η μυστικότητα των προσωπικών στοιχείων είναι ένας παράγοντας ο οποίος είναι ιδιαίτερα δύσκολος να επιτευχθεί ειδικά από τη στιγμή που τα άτομα επικοινωνούν και διαπραγματεύονται σε καθημερινή βάση μεταξύ τους,
- Έλεγχος προσβασιμότητας στα δεδομένα: αναφέρονται, δηλαδή, στον περιορισμό των ατόμων που έχουν πρόσβαση στα δεδομένα με σκοπό την αποφυγή της διαρροής των πληροφοριών,
- Ακρίβεια των δεδομένων: τα δεδομένα θα πρέπει να είναι ολοκληρωμένα και να ισχύουν με σκοπό να προστατευθούν τα άτομα και το σύνολο της κοινωνίας από την ατελή πληροφόρηση, η οποία θα είναι προϊόν των δεδομένων χαμηλής ποιότητας,
- Στερεότυπα που προκύπτουν από την ανάλυση των δεδομένων: αναφέρεται στο γεγονός ότι ακόμη κι αν εξαχθούν κάποια αποτελέσματα θα πρέπει να αντιμετωπιστούν με σύνεση αφού είναι πιθανό να μην ισχύουν για το σύνολο της κοινωνίας.

Οι Feelders et al. (2000) υποστηρίζοντας ξανά ότι η εξόρυξη δεδομένων συνδέεται με την ηθική και ειδικότερα την αρχή της προστασίας των προσωπικών δεδομένων υποστηρίζουν ότι πρέπει να υπάρχει γνώση της νομοθεσίας που σχετίζεται με αυτό το ζήτημα. Η νομοθεσία αφορά στους ακόλουθους τομείς:

- Η διαδικασία πρέπει να είναι αυτοματοποιημένη και να ελαχιστοποιείται η εμπλοκή του ατόμου με σκοπό την όσο μεγαλύτερη γίνεται ανωνυμία,
- Τα δεδομένα συλλέγονται με πλήρη επίγνωση του ερωτώμενου και δεν αποσπώνται με έμμεσο τρόπο από αυτόν,
- Ο ερωτώμενος έχει το δικαίωμα να επέμβει και να διορθώσει οποιαδήποτε ανακρίβεια σχετίζεται με τα στοιχεία που έχει ο ίδιος δώσει,
- Οι φορείς που συλλέγουν τα δεδομένα υποχρεούνται να προσδιορίσουν εξαρχής τον σκοπό συλλογής των δεδομένων προς αποφυγή ανακριβειών,
- Ο ερωτώμενος πρέπει να ενημερωθεί σχετικά με τη χρήση ή την απομάκρυνση των στοιχείων του από τη βάση δεδομένων,
- Τα στοιχεία/δεδομένα μπορούν να χρησιμοποιηθούν από κάποιο τρίτο μέρος μόνο εάν ο ερωτώμενος ενημερωθεί σχετικά,
- Παρέχεται πλήρης προστασία στα δεδομένα από τους φορείς που τα συλλέγουν προς αποφυγή ενδεχόμενων υποκλοπών.

4. Επιχειρήσεις και Εξόρυξη Δεδομένων

Σε αυτό το κεφάλαιο, αναλύουμε τον τρόπο με τον οποίο μία επιχείρηση μπορεί να επηρεαστεί από την εξόρυξη δεδομένων σε βασικές λειτουργίες της. Μεταξύ των βασικότερων τομέων που επηρεάζονται σημαντικά από την εξόρυξη δεδομένων στις επιχειρήσεις, περιλαμβάνονται μεταξύ άλλων η παραγωγική διαδικασία, το μάνατζμεντ, η παρακολούθηση των συναλλαγών, οι στρατηγικές μάρκετινγκ και γενικότερα οποιαδήποτε στρατηγική της επιχείρησης που μπορεί να επηρεάσει τη λειτουργία, την αποτελεσματικότητα και την κερδοφορία της. Επίσης, επισημαίνουμε την σημαντικότητα που διαδραματίζει η εξόρυξη δεδομένων σε μία επιχείρηση μέσω παραδειγμάτων που προέρχονται από την διεθνή βιβλιογραφία. Στο πρώτο υποκεφάλαιο, περιγράφουμε την διαδικασία της εξόρυξης δεδομένων σε μία επιχείρηση βασιζόμενοι και στα όσα είπαμε στα προηγούμενα κεφάλαια, ενώ στη συνέχεια παρουσιάζουμε τα ειδικά χαρακτηριστικά στοιχεία της εξόρυξης δεδομένων στις επιχειρήσεις καθώς και κάποιες από τις επιπτώσεις τους.

4.1 Διαδικασία της Εξόρυξης Δεδομένων σε μία Επιχείρηση

Όπως σε όλες τις επιστήμες και τις περιπτώσεις που εξετάσαμε παραπάνω, έτσι και στις επιχειρήσεις, η εξόρυξη δεδομένων είναι μία εξίσου σημαντική διαδικασία, ιδίως εάν αναλογιστούμε ότι επί της ουσίας οι σχέσεις και τα πρότυπα που μελετώνται, αλλά και η γνώση που προκύπτει από αυτά μπορούν είτε να επαληθευθούν είτε όχι στον πραγματικό κόσμο. Η εξόρυξη δεδομένων, συνήθως, χρησιμοποιείται σε τομείς όπως η πρόβλεψη των μελλοντικών τάσεων για τις αγοραστικές συνήθειες των καταναλωτών, την πορεία των εσόδων και των κοστών, το μάρκετινγκ, η παρακολούθηση των συναλλαγών και γενικότερα όλους τους τομείς των επιχειρήσεων, οι οποίοι συνδέονται με την διαδικασία λήψης αποφάσεων.

Η διαδικασία της εξόρυξης δεδομένων σε μία επιχείρηση μπορεί να περιγραφεί ως μία διαδραστική διαδικασία, στην οποία επιδρούν ταυτόχρονα πολλοί παράγοντες

μεταξύ τους. Στο αρχικό στάδιο της διαδικασίας διατυπώνεται με ακρίβεια το υπό εξέταση πρόβλημα. Για παράδειγμα, το αρχικό πρόβλημα μιας επιχείρησης, θα μπορούσε να είναι η δυνατότητα πρόβλεψης με ακρίβεια της πιθανότητας της αγοράς ενός συγκεκριμένου προϊόντος από τους καταναλωτές ή επιλογή για την αποτελεσματικότερη μέθοδο παραγωγής. Κατά συνέπεια, ο σωστός εντοπισμός των καταναλωτών μπορεί, θεωρητικά, να οδηγήσει στην ώθηση περισσότερων καταναλωτών προς την αγορά του συγκεκριμένου προϊόντος, από τη στιγμή που αυτοί έχουν την δυνατότητα να επιλέξουν ανάμεσα σε πολλά άλλα ομοιογενή προϊόντα που παράγονται από άλλες επιχειρήσεις. Στην δεύτερη περίπτωση η αποτελεσματικότερη μέθοδος παραγωγής μπορεί τόσο να αυξήσει το συνολικό επίπεδο της παραγωγής των προϊόντων και των υπηρεσιών όσο και την μείωση του κόστους παραγωγής. Έτσι, η επιχείρηση, θα μπορούσε, για παράδειγμα, να συγκεντρώσει στοιχεία δημογραφικά, γεωγραφικά, εισοδηματικά, τις τάσεις της αγοράς των προηγούμενων ετών με σκοπό να ελέγξει ποιοι καταναλωτές είχαν προβεί σε αγορές κατά το παρελθόν. Με αυτόν τον τρόπο εστιάζει στους πιο πιθανούς μελλοντικούς καταναλωτές της και στον εντοπισμό της καλύτερης μεθόδου παραγωγής.

Το δεύτερο στάδιο της εξόρυξης δεδομένων στις επιχειρήσεις αφορά στην προετοιμασία των δεδομένων προς ανάλυση. Πιο συγκεκριμένα, τα δεδομένα αυτά συνήθως λαμβάνουν δύο μορφές, είτε αριθμητική είτε κειμένου, όπου τα δεύτερα κωδικοποιούνται με αριθμούς προκειμένου να μπορέσουν να αναλυθούν στη συνέχεια. Τα δεδομένα περιγράφουν συνήθως στοιχεία όπως το φύλο, η ηλικία, η κατηγορία και το εισόδημα του πελάτη, αλλά και στοιχεία όπως οι πωλήσεις, τα έσοδα και τα έξοδα των επιχειρήσεων. Πριν ξεκινήσει η ουσιώδης διαδικασία της εξόρυξης, τα δεδομένα συλλέγονται, ομαδοποιούνται, καθαρίζονται, κατηγοριοποιούνται και στη συνέχεια κωδικοποιούνται κατάλληλα. Η κατάλληλη προετοιμασία των δεδομένων, η οποία μπορεί να χαρακτηρίζεται και από κάποιους μετασχηματισμούς αυτών, έχει ως αποτέλεσμα και την καλύτερη ποιότητα της εξόρυξης. Με αυτόν τον τρόπο σε αυτό το επίπεδο προσδιορίζεται και το βέλτιστο σετ των μεταβλητών. Το βέλτιστο σετ των μεταβλητών είναι πολύ πιθανό, πολλές φορές, να είναι αρκετά μειωμένο σε σχέση με το αρχικό, αφού με αυτόν τον τρόπο μπορεί να μειωθεί τόσο ο θόρυβος των καταλοίπων όσο και να αυξηθεί η ποιότητα και το περιεχόμενο της πληροφορίας. Δηλαδή, ενώ χρησιμοποιούνται οι μεταβλητές

που θεωρείται ότι επηρεάζουν περισσότερο την υπό εξέταση μεταβλητή, η συσχέτιση αυτών των μεταβλητών μεταξύ τους είναι η μικρότερη δυνατή με σκοπό την μείωση της πολυσυγραμμικότητας αλλά και την μεγαλύτερη παροχή πληροφορίας. Με αυτόν τον τρόπο μειώνεται και ο όγκος της επαναλαμβανόμενης πληροφορίας, η οποία σε άλλες περιπτώσεις θα μπορούσε να δημιουργήσει προβλήματα κατά την ανάλυση. Ανάλογα με το υπό εξέταση πρόβλημα, όπως αντιλαμβανόμαστε, χρησιμοποιείται και το κατάλληλο σύνολο δεδομένων μέσω της χρήσης κάποιας τεχνικής.

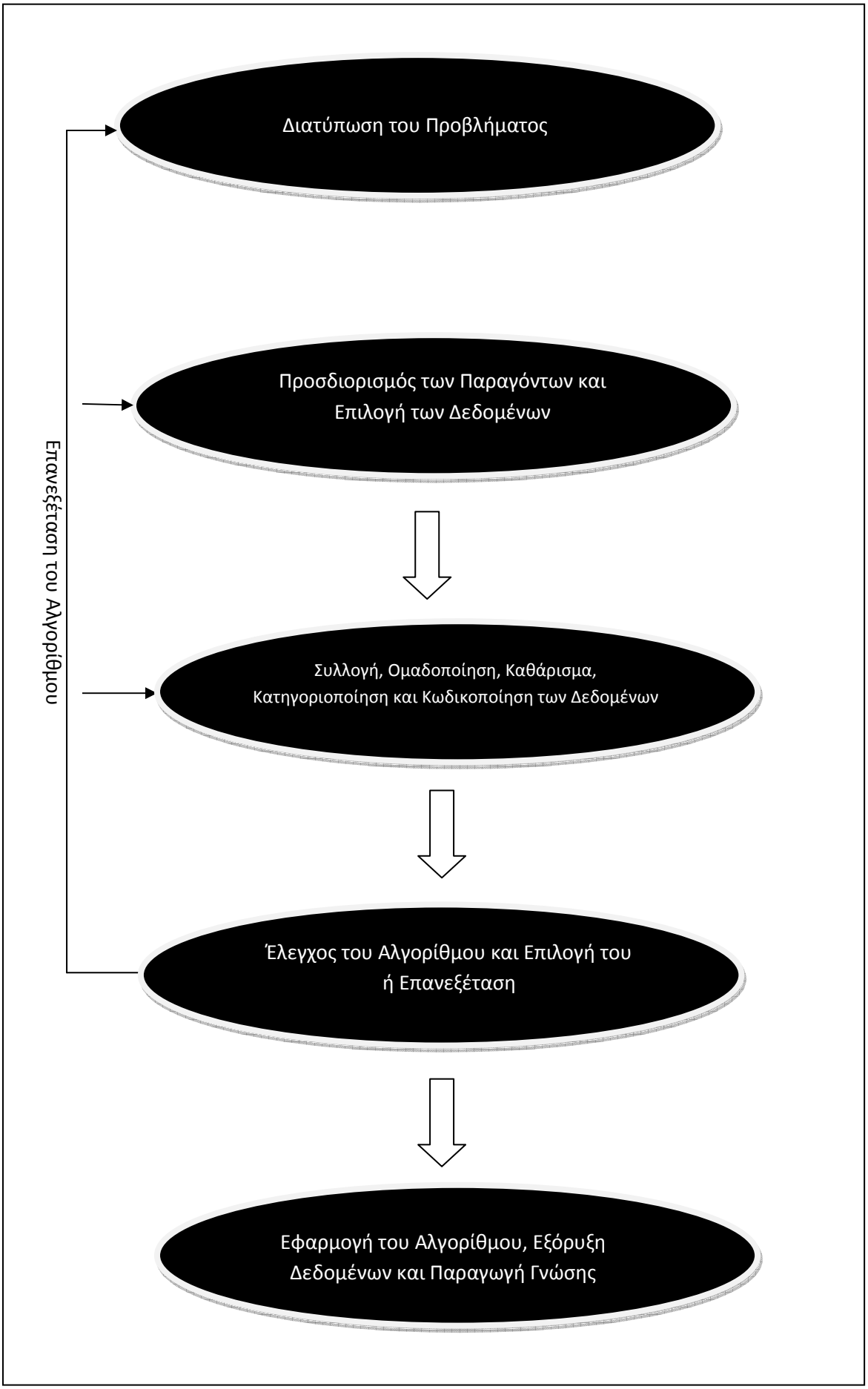
Το τρίτο στάδιο της διαδικασίας της εξόρυξης δεδομένων αποτελεί και το ουσιαστικότερο τμήμα της, όπου επιλέγονται όλα τα απαραίτητα στοιχεία για την έναρξη της εξόρυξης. Αρχικά, επιλέγεται ο αλγόριθμος που σύμφωνα με τον ερευνητή ανταποκρίνεται στην καλύτερη ερμηνεία και ανάλυση των δεδομένων που έχουν καθοριστεί στο προηγούμενο στάδιο, και θεωρείται ότι είναι αυτός που μπορεί να δώσει λύση στο υπό εξέταση πρόβλημα. Οι παράμετροι που πρέπει να εκτιμηθούν, με βάση τον αλγόριθμο που χρησιμοποιείται επιλέγονται με βάση την πληροφόρηση που έχουμε για το συγκεκριμένο πρόβλημα που εξετάζεται, την πληροφόρηση για το συνολικότερο περιβάλλον μέσα στο οποίο εντάσσεται αυτό πρόβλημα, αλλά και τα χαρακτηριστικά των δεδομένων που χρησιμοποιούνται ως εισροές στο σύστημα.

Αφού επιλεχθεί ο κατάλληλος αλγόριθμος για την διαδικασία της εξόρυξης δεδομένων, στη συνέχεια, ελέγχεται αυτός ο αλγόριθμος ως προς την αποτελεσματικότητά του. Ο έλεγχος αυτός αφορά στο διαχωρισμό των δεδομένων σε δύο μικρότερα μέρη, όπου το ένα χρησιμοποιείται για την εκτίμηση των παραμέτρων και την πρόβλεψη του υπό εξέταση μεγέθους (in – sample) και το δεύτερο στην επαλήθευση του μεγέθους του υπό εξέταση προβλήματος (out – of – sample). Εάν βρεθεί ότι ο συγκεκριμένος αλγόριθμος δεν προβλέπει με ακρίβεια το υπό εξέταση μέγεθος στο out – of – sample κομμάτι, τότε ο ερευνητής επιστρέφει σε κάποιο από τα προηγούμενα στάδια προκειμένου να εντοπίσει τα σημεία που πρέπει να διορθωθούν για να πραγματοποιηθεί η σωστή εξόρυξη δεδομένων. Έτσι μπορεί να επιστρέψει τόσο στο στάδιο κατά το οποίο επιλέγονται τα δεδομένα όσο και στο στάδιο στο οποίο επιλέγεται ο κατάλληλος αλγόριθμος. Κάποιες φορές, ωστόσο, είναι πολύ πιθανό, ο ερευνητής να αναγκάζεται να επιστρέψει στο πρώτο στάδιο κατά το οποίο προσδιορίζει το υπό εξέταση ζήτημα. Εάν, όμως, βρεθεί ότι ο συγκεκριμένος αλγόριθμος είναι σε θέση να προβλέψει με ακρίβεια την εξεταζόμενη μεταβλητή, τότε εφαρμόζεται η συγκεκριμένη μεθοδολογία στο σύνολο των δεδομένων με σκοπό

την εξόρυξη δεδομένων. Οφείλουμε, να αναφέρουμε, ωστόσο, ότι σε αυτό το στάδιο πέρα από τον υπό εξέταση αλγόριθμο μελετώνται ο τομέας στον οποίο συνεισφέρει αυτός ο συγκεκριμένος αλγόριθμος, το εξωτερικό περιβάλλον του υπό εξέταση προβλήματος που μπορεί να χρειαστεί να αναλυθεί σε έναν βαθμό, αλλά και οι επιπρόσθετοι περιορισμοί που μπορεί να ισχύουν.

Το τελευταίο στάδιο της εξόρυξης των δεδομένων σε μία επιχείρηση αφορά στην παραγωγή γνώσης από τα δεδομένα (KDD), όπου τα συμπεράσματα που εξάγονται μπορούν να χρησιμοποιηθούν είτε άμεσα για την επιλογή της κατάλληλης στρατηγικής από την επιχείρηση, είτε να αποθηκευτούν και να χρησιμοποιηθούν μελλοντικά από αυτήν. Σε αυτό το στάδιο, επίσης, είναι πολύ πιθανό να χρειαστεί να γίνουν κάποιοι μετασχηματισμοί στα αποτελέσματα προκειμένου αυτά να είναι άμεσα διαθέσιμα προς ανάλυση και για την κατάλληλη λήψη αποφάσεων.

Το διάγραμμα που ακολουθεί περιλαμβάνει το σύνολο της διαδικασίας που παρουσιάστηκε σε αυτή την υποενότητα και έχει ως σκοπό την καλύτερη κατανόηση της διαδικασίας της εξόρυξης δεδομένων στις επιχειρήσεις.


Διατύπωση του Προβλήματος
Προσδιορισμός των Παραγόντων και Επιλογή των Δεδομένων
Συλλογή, Ομαδοποίηση, Καθάρισμα, Κατηγοριοποίηση και Κωδικοποίηση των Δεδομένων
Έλεγχος του Αλγορίθμου και Επιλογή του ή Επανεξέταση
Εφαρμογή του Αλγορίθμου, Εξόρυξη Δεδομένων και Παραγωγή Γνώσης
Επανεξέταση του Αλγορίθμου

4.2 Ειδικά Χαρακτηριστικά της Εξόρυξης Δεδομένων στις Επιχειρήσεις

Οι Marbán et al. (2009) υποστηρίζουν ότι λόγω της μεγάλης αλληλεπίδρασης των διάφορων τομέων στις επιχειρήσεις και των όλο και πιο πολλών και πολύπλοκων δεδομένων που συσσωρεύονται και αποθηκεύονται αναποτελεσματικά, έχουν παρατηρηθεί φαινόμενα που σχετίζονται με καθυστερήσεις στην εκτέλεση των πλάνων από τις επιχειρήσεις, χαμηλή παραγωγικότητα και κατά συνέπεια αποτυχία ολοκλήρωσης των στόχων. Γι' αυτό το λόγο υποστηρίζουν ότι στις αρχές του 1990 υπήρξε μία πολύ έντονη τάση της ανάπτυξης των αλγορίθμων και των λογισμικών που χρησιμοποιούνται για την εξόρυξη δεδομένων με σκοπό την καλυτέρευση αυτών των μεγεθών. Γενικότερα, στις επιχειρήσεις υπεύθυνοι για την εξόρυξη δεδομένων θεωρείται ότι είναι οι αναλυτές, οι οποίοι χρησιμοποιούν όλο και πιο πολύπλοκες στατιστικές τεχνικές. Ωστόσο, η χρήση των ηλεκτρονικών υπολογιστών, το διαδίκτυο και οι βάσεις δεδομένων που έχουν δημιουργηθεί από τις επιχειρήσεις βοηθούν στην καλύτερη μελέτη αυτής της τεχνικής. Για παράδειγμα εταιρείες που δραστηριοποιούνται στον τομέα του εμπορίου, όπως η Wal – Mart Stores ελέγχουν μέσω των διαδικασιών αυτών τις πωλήσεις τους, την συμπεριφορά των καταναλωτών και τα επίπεδα των πωλήσεων ανά κατάστημα. Με αυτό τον τρόπο έχουν τη δυνατότητα να καθορίσουν αποτελεσματικά το κόστος τους, αλλά και να προσδιορίσουν με ακρίβεια το επίπεδο της παραγωγής τους.

Οι Bagga and Singh (2012) επίσης υποστηρίζουν ότι η εξόρυξη δεδομένων βοηθά στην λήψη αποφάσεων των επιχειρήσεων. Ειδικότερα, διαπιστώνουν ότι στο εμπόριο για παράδειγμα η εξόρυξη δεδομένων βοηθά σε τομείς όπως ο προσδιορισμός της συμπεριφοράς του καταναλωτή, η αλληλεπίδραση των επιχειρήσεων μεταξύ τους, οι επιπτώσεις της διαφήμισης στις πωλήσεις. Ωστόσο, οι Ghani and Soares (2006) υποστηρίζουν ότι η εξόρυξη δεδομένων στις επιχειρήσεις είναι μία ιδιαίτερα δύσκολη διαδικασία από την στιγμή που υπάρχει σημαντική δυσκολία πρόσβασης σε μεγάλες και νέες πηγές, ενώ και σε ορισμένους τομείς μόνον ειδικοί μπορούν να εισέλθουν. Αυτό το ζήτημα, όπως υποστηρίζουν εξακολουθεί να υπάρχει ακόμη και σήμερα και ενώ τα περισσότερα δεδομένα βρίσκονται σε δημόσιες πηγές. Όπως υποστηρίζουν

περαιτέρω, αυτό το κενό προσπαθεί να καλυφθεί ανάμεσα στην ακαδημαϊκή και επιχειρηματική κοινότητα.

Παράλληλα, οι Bose and Mahapatra (2001) σημειώνουν ότι η εξόρυξη δεδομένων αποτελεί μία πολύ σημαντική διαδικασία στις επιχειρήσεις από τη στιγμή που αυτές έρχονται αντιμέτωπες με τον οξύ ανταγωνισμό και αναγκάζονται να βρουν τρόπους προκειμένου είτε να αυξήσουν είτε να διατηρήσουν το μερίδιό τους στην αγορά μέσω της συνολικής οργάνωσης της στρατηγικής τους. Όπως υποστηρίζουν, η μεθοδολογία που ακολουθείται από αυτές είναι:

- Επιλογή των δεδομένων που πρέπει να αναλυθούν με σκοπό την εξαγωγή των επιθυμητών συμπερασμάτων: τα δεδομένα αυτά αφορούν συνήθως σε ιστορικά στοιχεία που έχουν συσσωρευτεί σε μία ή περισσότερες βάσεις δεδομένων. Μάλιστα σε πολλές περιπτώσεις οι βάσεις δεδομένων μπορεί να μην είναι δημόσιες, αλλά να κρατούνται από τις ίδιες τις επιχειρήσεις και να έχουν μόνον οι ίδιες πρόσβαση σε αυτές. Με αυτόν τον τρόπο μπορούν να χρησιμοποιηθούν τόσο μόνον τα δεδομένα που έχουν συλλεχθεί από την επιχείρηση όσο και να συνδυαστούν τα δεδομένα από τις διάφορες πηγές,
- Ξεκαθάρισμα των δεδομένων και σχετική προεργασία: σε αυτό το στάδιο ελέγχεται εάν τα δεδομένα είναι επαρκή και ακριβή, εάν λείπουν παρατηρήσεις από αυτά κλπ, ενώ στη συνέχεια γίνεται μία πρώτη επεξεργασία προκειμένου αυτά να χρησιμοποιηθούν στη συνέχεια. Η προεργασία αυτή αφορά συνήθως στον μετασχηματισμό των δεδομένων (π.χ. υπολογισμός αποδόσεων και τάσεων), ο οποίος οδηγεί και στην καλύτερη ανάλυσή τους,
- Αναζήτηση προτύπων μεταξύ των μεταβλητών: τα πρότυπα αυτά αφορούν στις σχέσεις που υπάρχουν ανάμεσα στις μεταβλητές και τον τρόπο με τον οποίο αυτές συσχετίζονται ή επηρεάζουν η μία την άλλη,
- Μετατροπή των προτύπων σε επιχειρηματικό σχέδιο (business plan): αφού εντοπιστούν τα όποια πρότυπα μεταξύ των εξεταζόμενων μεταβλητών πρέπει να γίνει η μετατροπή τους σε ένα επιχειρηματικό σχέδιο της επιχείρησης που αντιπροσωπεύει και την στρατηγική της επιχείρησης με σκοπό την επίτευξη των στόχων της,

- Εξαγωγή γνώσης από το σύνολο της διαδικασίας: όπου αφού προσδιοριστούν τα δεδομένα, τα πρότυπα που τα χαρακτηρίζουν και διαμορφωθεί ο στρατηγικός σχεδιασμός της επιχείρησης, τότε παράγεται κάποια γνώση η οποία μπορεί να χρησιμοποιηθεί είτε άμεσα είτε να αποθηκευτεί και να χρησιμοποιηθεί μεταγενέστερα σε κάποια άλλη διαδικασία.

Οι συνηθέστερες μεθοδολογίες εξόρυξης δεδομένων στις επιχειρήσεις είναι η επαγωγή κανόνων, τα τεχνητά νευρωνικά δίκτυα, οι μελέτες περιπτώσεων και ο υπολογισμός των συσχετίσεων. Η επαγωγή κανόνων προέρχεται ως το αποτέλεσμα που προκύπτει από την σχέση μεταξύ αιτίας και αιτιατού και απαντά επί της ουσίας στο ερώτημα εάν γίνει κάτι, τότε θα ακολουθήσει κάτι άλλο. Τα τεχνητά νευρωνικά δίκτυα είναι μία λίγο πιο περίπλοκη διαδικασία αφού όπως εξηγήσαμε και στο δεύτερο κεφάλαιο, στηρίζεται στη γενετική και την βιολογία ενώ οι χρησιμοποιούμενοι αλγόριθμοι πρέπει να εκπαιδεύονται με σκοπό να δώσουν ακριβέστερα αποτελέσματα. Οι μελέτες περιπτώσεων, συνήθως, συνδυάζονται με την μελέτη των αντιδράσεων της ίδιας της επιχείρησης ή άλλων των επιχειρήσεων σε άλλες αντίστοιχες καταστάσεις. Σε αυτή την περίπτωση, οι αποφάσεις λαμβάνονται με γνώμονα το τι είχε γίνει σε άλλες αντίστοιχες περιπτώσεις.

Ο πιο συνηθισμένος, όμως, τρόπος υπολογισμού των προτύπων, αλλά και παραγωγής γνώσης στο εσωτερικό μιας επιχείρησης είναι η χρήση των κανόνων συσχέτισης. Πιο συγκεκριμένα, οι κανόνες συσχέτισης μπορούν να χρησιμοποιηθούν ευκολότερα από τη στιγμή που μέσω αυτών προσδιορίζεται ο τρόπος με τον οποίο διάφορα προϊόντα, υπηρεσίες ή οι διαφορετικοί τομείς μιας επιχείρησης μπορούν να αλληλεπιδρούν μεταξύ τους. Υπάρχουν συνολικά δύο είδη κανόνων συσχέτισης στις επιχειρήσεις. Ο πρώτος, είναι ο πιο εύκολος τόσο στην παρατήρησή του όσο και στην εξήγησή του και αφορά σε απλές διεργασίες οι οποίες μπορούν να παρατηρηθούν σχετικά εύκολα. Για παράδειγμα, κάποια προϊόντα αγοράζονται συχνότερα μαζί, οπότε θα ήταν σχετικά καλύτερο να τοποθετούνται σε κοντινά ράφια μεταξύ τους προκειμένου να αγοραστούν ταυτόχρονα ή εργασίες που διενεργούνται ταυτόχρονα κατά την παραγωγική διαδικασία, θα ήταν προτιμότερο να γίνονται στον ίδιο χώρο. Ο δεύτερος κανόνας είναι αυτός που είναι αντιληπτός για κάθε επιχείρηση και μπορεί να θεωρηθεί και ως ασήμαντος για την χάραξη της στρατηγικής της επιχείρησης.

Χαρακτηριστικό παράδειγμα είναι η είναι αναγκαστική αγορά ενός προϊόντος με την χρήση μιας υπηρεσίας (π.χ. τηλεφωνική συσκευή και σύνδεση τηλεφωνίας, ηλεκτρονικός υπολογιστής και σύνδεση στο διαδίκτυο).

Όλες αυτές οι μεθοδολογίες, κάποιες από αυτές που αναλύθηκαν σε προηγούμενα κεφάλαια αλλά μπορούν να χρησιμοποιηθούν και στην περίπτωση των επιχειρήσεων, καθώς και όσες θα παρουσιαστούν στη συνέχεια μπορούν να οδηγήσουν στην καλύτερη λήψη αποφάσεων. Κατά την λήψη αποφάσεων από τις επιχειρήσεις, λαμβάνονται υπόψη, σύμφωνα με τους Battiti and Passerini (2010), οι εξής σημαντικοί παράγοντες:

- Περιορισμένη Ορθολογικότητα: οι λαμβάνοντες τις αποφάσεις στις επιχειρήσεις, δεν έχουν την δυνατότητα να επεξεργαστούν πολλές πολύπλοκες πληροφορίες. Ως αποτέλεσμα είναι πάρα πολύ πιθανό να μην μπορεί να επιλεχθεί η πραγματικά βέλτιστη λύση (απόφαση), η οποία στην πραγματικότητα θα αντιστοιχεί στην απόφαση που θα λαμβανόταν εάν και μόνο εάν αναλύονταν όλοι οι επιμέρους παράγοντες,
- Περιορισμένη Πληροφόρηση: κάθε επιχείρηση και κάθε άτομο μπορεί να έχει πρόσβαση σε έναν σημαντικό αριθμό πληροφοριών αλλά όχι στο σύνολο της πληροφόρησης. Για μία επιχείρηση, αυτό πρακτικά μπορεί να σημαίνει ότι ενώ έχει πρόσβαση στη γενική πληροφόρηση (εφημερίδες, ισολογισμοί, ίντερνετ, κλπ), δεν έχει τη δυνατότητα να έχει πρόσβαση σε παράγοντες όπως οι στρατηγικές των ανταγωνιστριών επιχειρήσεων, ούτε και στο σύνολο των καταναλωτών, αφού είναι πρακτικά αδύνατο τόσο να γνωρίζει τις συνήθειες όλων των εν δυνάμει καταναλωτών της, αλλά και τις στρατηγικές που ακολουθούνται από τις ανταγωνίστριες επιχειρήσεις,
- Μάθηση κατά την Επανάληψη και την Εργασία: όπου πολλές φορές, η λήψη αποφάσεων πρέπει να χαρακτηρίζεται από την εμπειρία που έχουν αποκτήσει οι ιθύνοντες κατά το παρελθόν σε παρόμοιες καταστάσεις. Με αυτόν τον τρόπο μπορούν να αποφευχθούν σημαντικά λάθη και παραλείψεις, τα οποία σε αντίθετη περίπτωση μπορεί να μην ήταν δυνατό να εντοπιστούν,
- Ποιοτική Κριτική σε σχέση με την πολυπλοκότητα των ερωτημάτων: όπου στην πραγματικότητα οι περισσότερες από τις αποφάσεις που πρέπει να

ληφθούν χαρακτηρίζονται από μεγάλη πολυπλοκότητα με αποτέλεσμα να είναι πολύ δύσκολη η απάντησή τους. Αυτό σημαίνει ότι σε αρκετές περιπτώσεις είναι πολύ δύσκολο να συνδυαστούν όλα τα απαραίτητα δεδομένα προκειμένου να απαντηθεί ένα και μόνο ερώτημα. Για παράδειγμα, εάν θέλουμε να εξετάσουμε την αντίδραση των ανταγωνιστών στην είσοδο ενός νέου προϊόντος στην αγορά από την δική μας επιχείρηση, τότε θα πρέπει να αναλυθούν ταυτόχρονα τόσο η συμπεριφορά των καταναλωτών, η συμπεριφορά της κάθε μίας ανταγωνίστριας επιχείρησης, η αντίδραση της δικής μας επιχείρησης αλλά και άλλοι τοπικοί παράγοντες που σχετίζονται με αυτή την διαδικασία, αλλά και το γενικότερο περιβάλλον στο οποίο εμπίπτει η επιχείρηση,

- Αβεβαιότητα ως προς τα αποτελέσματα: ακόμη και εάν γίνει η οποιαδήποτε ανάλυση και ληφθεί η καλύτερη σύμφωνα με το σύνολο των δεδομένων απόφαση, είναι πάρα πολύ πιθανό αυτή η απόφαση να μην αντιστοιχεί στην καλύτερη δυνατή αφού μπορεί να αλλάξει η παραμικρή παράμετρος που δεν είχε ληφθεί υπόψη με αποτέλεσμα να εντοπιστεί ότι αυτή δεν αποτελεί πλέον την καλύτερη εναλλακτική,
- Ανεπάρκεια των δεδομένων και των αποφάσεων: πολλές φορές τα δεδομένα που πρέπει να χρησιμοποιηθούν για να ληφθεί μία απόφαση χαρακτηρίζονται ως ανεπαρκή από την στιγμή που μπορεί να μην υπάρχουν ακριβώς τα απαιτούμενα δεδομένα που πρέπει να χρησιμοποιηθούν για την εξέταση μίας μεταβλητής, αλλά να χρησιμοποιούνται παρεμφερή με αυτήν την μεταβλητή δεδομένα. Ως αποτέλεσμα, μπορεί και η απόφαση να χαρακτηριστεί ως ανεπαρκής από την στιγμή που με αυτόν τον τρόπο δεν θα ανταποκρίνεται στην πραγματική λύση του προβλήματος.

Οι παράγοντες αυτοί λαμβάνονται υπόψη κατά την λήψη αποφάσεων με σκοπό τη μεγιστοποίηση της χρησιμότητας των επιχειρήσεων, η οποία μπορεί να μεταφραστεί και στην αυξημένη κερδοφορία.

4.3 Αποθήκευση των Δεδομένων και Σχετικά Συστήματα

Όπως αντιλαμβανόμαστε από την έως τώρα ανάλυσή μας, τα δεδομένα που αναλύονται διαδραματίζουν καθοριστικό ρόλο στις αποφάσεις που λαμβάνονται από τις επιχειρήσεις. Το γεγονός αυτό επισημαίνουν και οι Pathak et al. (2013), εξηγώντας πως ακόμη και εάν το γενικό περιβάλλον προσφέρεται για την ύπαρξη δεδομένων, τα δεδομένα αυτά δεν είναι πάντοτε σίγουρο ότι οδηγούν στην επίλυση των προβλημάτων των επιχειρήσεων και την επίτευξη των στόχων τους. Γι' αυτό το λόγο κρίνεται επιτακτική η ανάγκη για την αποθήκευση των δεδομένων, η οποία βοηθά στην επάρκεια και την πληρότητά τους. Η αποθήκη των δεδομένων αποτελεί στην πραγματικότητα την ένωση των δεδομένων που προέρχονται από διαφορετικές βάσεις, η οποία στο σύνολό της ενημερώνεται ανά τακτά χρονικά διαστήματα, προκειμένου να εξασφαλιστεί η σταθερότητα και η εγκυρότητα των δεδομένων διαχρονικά. Η αποθήκη των δεδομένων διαφέρει από τις βάσεις αφού αυτή αποτελεί επί τη ουσίας μία περιληπτική πληροφόρηση των βάσεων δεδομένων, ενώ διατηρείται και χωριστά από αυτές. Μέσω της χρήσης της μπορούν να δοθούν σε πολύ συντομότερο χρονικό διάστημα απαντήσεις σχετικά με τα ερωτήματα των επιχειρήσεων, ενώ δεν ξοδεύονται άσκοπα και αλόγιστα χρήματα.

Προκειμένου η αποθήκευση των δεδομένων να γίνει με αποτελεσματικό τρόπο, θα πρέπει να ακολουθηθούν κάποια συγκεκριμένα στάδια, τα οποία αναλύονται ως:

- Εύρεση και Συλλογή Ακατέργαστων Δεδομένων: είναι το στάδιο κατά το οποίο εντοπίζονται τα δεδομένα από διάφορες εξωτερικές ή μη πηγές, που θα αποτελέσουν την πηγή που θέλουμε να δημιουργήσουμε,
- Δημιουργία Ενοποιημένης Πηγής Δεδομένων: είναι το στάδιο κατά το οποίο ενοποιούνται και οργανώνονται τα ακατέργαστα δεδομένα που συλλέχθηκαν κατά το προηγούμενο στάδιο,
- Δημιουργία της Αποθήκης Δεδομένων: προκύπτει ως το αποτέλεσμα των διάφορων ενοποιημένων βάσεων δεδομένων, που κατασκευάστηκαν κατά το προηγούμενο στάδιο.

Αφού κατασκευαστεί η επιθυμητή βάση δεδομένων, ακολουθεί η ανάλυση των δεδομένων αυτών μέσω των οποίων οδηγούμαστε στη λήψη αποφάσεων. Η ανάλυση των δεδομένων γίνεται και κατά την διαδικασία της εξόρυξης, η οποία σύμφωνα με τους Pathak et al. (2013), πρέπει να προσδιορίζεται και να καθορίζεται και από τον στόχο της επιχείρησης. Έτσι, ο στόχος της επιχείρησης πρέπει να χαρακτηρίζεται από μοναδικότητα και να προσδιορίζεται με σαφήνεια. Αυτό σημαίνει ότι ο στόχος της επιχείρησης και για τον οποίο διενεργείται η εξόρυξη των δεδομένων πρέπει να είναι γνωστός εκ των προτέρων, ενώ και τα κύρια χαρακτηριστικά του πρέπει να έχουν προσδιοριστεί πλήρως. Αφού προσδιοριστεί ο στόχος της επιχείρησης, θα πρέπει να επιλεχθούν και τα κατάλληλα δεδομένα – κατάλληλες μεταβλητές που μπορούν να οδηγήσουν στην επίτευξή του. Τα δεδομένα αυτά θα πρέπει να είναι επαρκή για την επεξεργασία τους και να βασίζονται σε πρόσφατα στοιχεία, ενώ θα πρέπει να ελέγχονται διαρκώς για την εγκυρότητά τους.

Στη συνέχεια, ακολουθεί η προετοιμασία των δεδομένων. Σε αυτό το στάδιο γίνονται οι κατάλληλοι μετασχηματισμοί που τα κάνουν να είναι κατάλληλα προς επεξεργασία. Για παράδειγμα, θα πρέπει να ληφθούν αποφάσεις σχετικά με τον πιθανό υπολογισμό αποδόσεων, την αντιμετώπιση ελλιπών δεδομένων ή ακραίων τιμών, καθώς και την κατανομή που αυτά θα ακολουθούν. Παράλληλα, πρέπει να γίνει και η κατάλληλη αξιολόγηση των δεδομένων, όπου αυτά εξετάζονται ως προς την δομή, την διάρθρωσή τους και τα γενικότερα χαρακτηριστικά τους. Αφού διενεργηθούν όλα τα παραπάνω στάδια, στη συνέχεια επιλέγεται η μέθοδος με την οποία θεωρούμε ότι μπορούμε να επιτύχουμε την καλύτερη ανάλυση των δεδομένων σε συνάφεια με την επίτευξη του στόχου της επιχείρησης.

Η επιλογή της επιθυμητής μεθοδολογίας ακολουθείται από την προετοιμασία του επιθυμητού υποδείγματος/μοντέλου όπου χρησιμοποιείται ένα συγκεκριμένο σετ δεδομένων για το οποίο κατασκευάζεται ένα συγκεκριμένο υπόδειγμα. Αφού κατασκευαστεί αυτό το υπόδειγμα αξιολογείται ως προς την προβλεπτική και ερμηνευτική του ικανότητα και ξενικά ουσιαστικά η διαδικασία της εξόρυξης δεδομένων. Τα αποτελέσματα που προκύπτουν από το επιλεχθέν υπόδειγμα/την επιλεχθείσα μεθοδολογία αξιολογούνται ως προς την ακρίβειά τους και την ανταπόκρισή τους στους στόχους της επιχείρησης. Προκειμένου, ωστόσο, να μπορέσουν τα αποτελέσματα αυτά να χρησιμοποιηθούν κατά την διαδικασία λήψης αποφάσεων πρέπει να ετοιμαστεί μία έγγραφη αναφορά των αποτελεσμάτων. Η

έγγραφη αναφορά αποτελεί στην ουσία μία σύντομη περιγραφή του συνόλου της διαδικασίας που ακολουθήθηκε, καθώς και των αποτελεσμάτων που προέκυψαν από αυτήν. Αυτή η διαδικασία αποτελεί και το σημαντικότερο στοιχείο για την διαδικασία λήψης αποφάσεων της επιχείρησης, όπως και για την μετέπειτα πορεία της. Το τελευταίο στάδιο αυτής της διαδικασίας αποτελείται από τον συνδυασμό όλων των διαθέσιμων στοιχείων προς την εύρεση της κατάλληλης λύσης. Σε αυτό το στάδιο, τα στοιχεία από όλα τα στάδια αναλύονται με σκοπό την ελαχιστοποίηση της λήψης αναποτελεσματικών αποφάσεων, οι οποίες θα βασίζονται είτε σε λανθασμένα συμπεράσματα είτε σε λανθασμένες υποθέσεις είτε ακόμη και στην επιλογή λανθασμένων ή ακατάλληλων δεδομένων.

Για την καλύτερη ανάλυση, αλλά και οργάνωση – αποθήκευση των δεδομένων, μεγάλες εταιρείες έχουν κατασκευάσει τα κατάλληλα λογισμικά. Μεταξύ αυτών των εταιρειών περιλαμβάνονται η IBM, η Microsoft και η Oracle, για τις οποίες αναλύονται κάποια βασικά τους χαρακτηριστικά, αλλά και η λογική που ακολουθούν για το σύνολο της διαδικασίας της εξόρυξης δεδομένων στη συνέχεια.

Πριν προβούμε στην ανάλυση των επιπτώσεων των συστημάτων που αναπτύχθηκαν για την οργάνωση – αποθήκευση των δεδομένων από ορισμένες εταιρείες, οφείλουμε να αναφέρουμε ότι οι αποθήκες δεδομένων γενικότερα στις επιχειρήσεις χρησιμοποιούν δεδομένα από πολλαπλές, ετερογενείς βάσεις, οι οποίες οργανώνονται στο σύνολό τους με τον ίδιο τρόπο και διευκολύνουν με αυτόν τον τρόπο την διαδικασία λήψης αποφάσεων. Έτσι, οι αποθήκες δεδομένων μπορούν να θεωρηθούν ως οι πλέον κατάλληλες μέθοδοι για την επαλήθευση ή μη συγκεκριμένων υποθέσεων, οι οποίες, όμως, λειτουργούν συμπληρωματικά στην διαδικασία της εξόρυξης. Η σημαντικότητά τους έγκειται στο γεγονός ότι αυτές στην πραγματικότητα περιλαμβάνουν τη συλλογή, την διόρθωση, τον καθαρισμό και τον μετασχηματισμό του συνόλου των διαφορετικών δεδομένων με σκοπό να χρησιμοποιηθούν στην εξόρυξη, η οποία στοχεύει στην εύρεση της κρυφής γνώσης.

Η IBM (2011) υποστήριξε ότι ο βασικός στόχος κάθε επιχείρησης είναι η δυνατότητα διατήρησης του μεριδίου της στην αγορά, το οποίο μπορεί να εξασφαλιστεί μέσω της κατάλληλης αξιοποίησης του ανταγωνιστικού της πλεονεκτήματος. Για την διατήρηση αυτού του ανταγωνιστικού πλεονεκτήματος κρίνεται ως επιτακτική η ανάγκη για την εξόρυξη δεδομένων με σκοπό την αύξηση της αξίας της επιχείρησης.

Περαιτέρω, εντόπισε ότι στην δεκαετία από το 2000 έως το 2010, η ανάλυση τόσο των εσωτερικών όσο και των εξωτερικών παραγόντων των επιχειρήσεων εντασσόταν μέσα στις τέσσερις κορυφαίες τεχνολογικές τάσεις. Γι' αυτό το λόγο η ύπαρξη δεδομένων κρίνεται ως ένα επιτακτικό και συμπληρωματικό στοιχείο για την ανάλυση οποιουδήποτε τομέα των επιχειρήσεων, ενώ και οι επιχειρήσεις αποτελούν ένα απαραίτητο συστατικό για την ανάλυση των δεδομένων. Η IBM (2011) διαχωρίζει τις εξής τάσεις στην εξόρυξη δεδομένων:

- Ολοκληρωμένες Αναλύσεις: πρόκειται για αποφάσεις που σχετίζονται άμεσα με την εξόρυξη δεδομένων και προκύπτουν ως αποτέλεσμα της βελτιστοποίησης των χρησιμοποιούμενων αλγορίθμων,
- Δεδομένα και Χρόνος ανάλυσής τους: όσο αυξάνονται τα δεδομένα αυξάνεται και ο όγκος των βάσεων στις οποίες αποθηκεύονται. Αυτό έχει ως αποτέλεσμα την ανάγκη της ανάλυσης των δεδομένων σε συντομότερο χρονικό διάστημα με σκοπό να διατηρηθεί η θέση της επιχείρησης στην αγορά,
- Μέσα Κοινωνικής Δικτύωσης: θεωρείται ότι μέσω των μέσων κοινωνικής δικτύωσης μπορεί να επηρεαστεί η συμπεριφορά των καταναλωτών και να καθοδηγηθεί από αυτά,
- Χρόνος και Χωρικές Διαστάσεις: αναφερόμαστε στην περίπτωση όπου πρέπει να ισχύει ο κανόνας για τον σωστό χρόνο στο σωστό χώρο. Κατά συνέπεια σε αυτή την περίπτωση αναφερόμαστε στις συνθήκες του μικροοικονομικού σχεδιασμού και της προστασίας των ιδιωτικών δεδομένων.

Η Oracle (2008) αποτελεί μία από τις εταιρείες που παρέχουν λογισμικά συστήματα για την καλύτερη λειτουργία των επιχειρήσεων. Στην έκθεσή της αναφέρει ότι η Telephonica O2 Germany Gmbh & Co., η οποία αποτελεί μέλος του ομίλου Telephonica S.A., μπόρεσε μέσω του παρεχόμενου συστήματος να επιτύχει τόσο την αποθήκευση των δεδομένων που απαιτούνται για την ομαλή της λειτουργία, την ίδια στιγμή που τα δεδομένα αυτά μπορούσαν να προέρχονται από πολλαπλές πηγές. Με αυτόν τον τρόπο μπόρεσε να έχει στη διάθεσή της μία πληθώρα δεδομένων, τα οποία οδήγησαν στην καθοδήγηση της στρατηγικής του μάνατζμεντ με αποτέλεσμα την

ικανοποίηση των πελατών και την καθιέρωση άριστων σχέσεων ανάμεσα στην επιχείρηση και τους πελάτες.

Γενικότερα, η Oracle (2010), επεσήμανε ότι μπόρεσε χάρη στην ανάπτυξη των συστημάτων της, να βοηθήσει τις επιχειρήσεις σε τομείς όπως:

- Το λιανικό εμπόριο: αφού δόθηκε η ευκαιρία στις επιχειρήσεις να κατηγοριοποιήσουν τους πελάτες τους και να προσδιορίσουν το προφίλ των καταναλωτών με αποτέλεσμα να διατεθούν τα κατάλληλα προϊόντα και να διεξαχθούν οι σωστότερες διαφημιστικές καμπάνιες,
- Τον τραπεζικό τομέα: όπου τα πιστωτικά ιδρύματα βοηθήθηκαν σε τομείς όπως η στόχευση συγκεκριμένων πελατών προκειμένου να οδηγηθούν σε υψηλότερα επίπεδα κερδοφορίας,
- Τον τομέα της ασφάλειας: όπου προσδιορίζονται παράγοντες κινδύνου και πιθανές διαφθορές που θα μπορούσαν να οδηγήσουν σε γενικότερη χειροτέρευση της οικονομίας,
- Τον τομέα της υγείας: όπου παρακολουθούνται τόσο οι ασθενείς με σκοπό τον τρόπο εξέλιξης των διαφόρων ασθενειών, αλλά και την πρόβλεψη για την μελλοντική πορεία της υγείας τους. Σε αυτόν τον τομέα επίσης παρακολουθούνται οι συμπεριφορές των γιατρών και των νοσηλευτών, καθώς και η πιθανότητα ανακάλυψης νέων φαρμάκων για την αποτελεσματικότερη αντιμετώπιση των ασθενειών,
- Τον τομέα των τηλεπικοινωνιών: όπου εντοπίζονται τόσο οι ευκαιρίες για την εισαγωγή στην αγορά νέων προϊόντων ή υπηρεσιών, όσο και η εξασφάλιση των προϊόντων με σκοπό την αποφυγή ανεπιθύμητων εισβολών και χάκινγκ,
- Τον δημόσιο τομέα: όπου τα συστήματα αυτά εφαρμόζονται σε τομείς όπως η καταπολέμηση της διαφθοράς στην φορολογία, η μείωση της εγκληματικότητας, αλλά και η αύξηση της εθνικής ασφάλειας.

Η Microsoft (Microsoft Corporation, 2007) επίσης προσφέροντας λογισμικά αποθήκευσης και επεξεργασίας των δεδομένων, μπόρεσε να βοηθήσει την British Telecom, μία από τις μεγαλύτερες εταιρείες παρόχων επικοινωνίας, στην μείωση των

λειτουργικών της κοστών και την επίτευξη οικονομιών κλίμακας στο εσωτερικό της. Με αυτόν τον τρόπο, μπόρεσε να προσφέρει καλύτερες διαδικτυακές υπηρεσίες, ενώ μειώθηκε ο φόρτος εργασίας των υπαλλήλων και βελτιώθηκε το προφίλ της εταιρείας στα μάτια των καταναλωτών.

Η Microsoft (ISL, 2009), αντίστοιχα, βασιζόμενη στα λογισμικά που ανέπτυξε μπόρεσε να προσφέρει βοήθεια στους εξής τομείς:

- Τον τρόπο με τον οποίο νέα προϊόντα αντικαθιστούν παλιά προϊόντα: με αυτόν τον τρόπο θεωρήθηκε ότι ήταν δυνατό να προβλεφθεί με μεγαλύτερη ακρίβεια η διάρκεια ζωής ενός προϊόντος, καθώς και ο χρόνος κατά τον οποίο αυτό το προϊόν έπρεπε είτε να αντικατασταθεί είτε να βελτιωθεί,
- Τον τρόπο με τον οποίο καθορίζεται και αλλάζει η συμπεριφορά του καταναλωτή: όπου εξετάζονταν παράγοντες τόσο μικροοικονομικοί ή μακροοικονομικοί όσο και προσωπικοί ή ψυχολογικοί με σκοπό να γίνουν αντιληπτοί οι λόγοι για τους οποίους ένας καταναλωτής θα οδηγούταν στην επιλογή κατανάλωσης συγκεκριμένων μόνο προϊόντων και τον αποκλεισμό κάποιων άλλων,
- Την καλύτερη ανάλυση του συνόλου της αγοράς μέσω του εντοπισμού από μεριάς των επιχειρήσεων των επικερδών ομάδων των καταναλωτών: επί της ουσίας εξετάζεται η κύρια στρατηγική των επιχειρήσεων, όπου περιλαμβάνεται και το μάρκετινγκ,
- Την πρόβλεψη για την αλληλεπίδραση των αποθεμάτων και των πωλήσεων μεταξύ τους: πρόκειται για ένα στάδιο στο οποίο μπορεί να γίνει καλύτερη διαχείριση του κόστους λειτουργίας της επιχείρησης, ενώ μπορούν να αντιμετωπιστούν περιπτώσεις έλλειψης προϊόντων σε περιόδους ξαφνικής αυξημένης ζήτησης,
- Την εξερεύνηση και περαιτέρω εξήγηση των ήδη υπαρχόντων δεδομένων: όπου τα δεδομένα που ήδη υπάρχουν στην επιχείρηση μπορούν να αναλυθούν περαιτέρω προκειμένου να εξεταστούν κάποια συγκεκριμένα μεγέθη, ενώ δεν θεωρείται απαραίτητο να βρεθούν πάντοτε νέα,
- Την χρήση προγενέστερης γνώσης, η οποία για κάποιον λόγο έχει αποθηκευτεί αλλά δεν έχει χρησιμοποιηθεί: πρόκειται για την περίπτωση όπου

προηγούμενες αναλύσεις δεν εφαρμόστηκαν ποτέ στην πράξη και σε αρκετές περιπτώσεις έχουν ξεχαστεί από την επιχείρηση την στιγμή που θα μπορούσαν να επιφέρουν σημαντικά οφέλη για αυτήν.

4.4 Επιπτώσεις της Εξόρυξης Δεδομένων στις Επιχειρήσεις

Αφού εξετάσαμε τα κύρια χαρακτηριστικά της διαδικασίας της εξόρυξης δεδομένων στις επιχειρήσεις, αλλά και τα συστήματα που έχουν αναπτυχθεί ως αποτέλεσμα της αυξημένης ανάγκης των επιχειρήσεων για ενοποιημένες βάσεις, οι οποίες βοηθούν στην διαδικασία λήψης αποφάσεων. Όπως καταλαβαίνουμε από την έως τώρα ανάλυση, ένας από τους σημαντικότερους τομείς, στους οποίους χρησιμοποιείται η εξόρυξη δεδομένων στο εσωτερικό μιας επιχείρησης είναι το μάνατζμεντ και το μάρκετινγκ (Ahmed, 2004). Κύριος σκοπός της εξόρυξης δεδομένων σε αυτόν τον τομέα είναι η ανακάλυψη και διερεύνηση των όποιων σχέσεων/προτύπων υπάρχουν ανάμεσα στους εσωτερικά προσδιορισμένους παράγοντες της επιχείρησης (όπως η τιμή του προϊόντος και ο τρόπος διάθεσής του) και τους εξωτερικούς παράγοντες της επιχείρησης (όπως η γενικότερη οικονομική κατάσταση της αγοράς στην οποία δραστηριοποιείται η επιχείρηση και τα δημογραφικά χαρακτηριστικά των υποψηφίων πελατών). Με αυτόν τον τρόπο μπορεί να μετρηθεί ο βαθμός ικανοποίησης του πελάτη από την επιχείρηση, που αντιστοιχεί και στον εντοπισμό του βαθμού της αλληλεπίδρασης πελάτη – επιχείρησης, με σκοπό να σχεδιαστεί με αποτελεσματικότερο τρόπο η στρατηγική μάρκετινγκ.

Πιο συγκεκριμένα, θεωρούμε ότι μέσω της εξόρυξης δεδομένων στο μάρκετινγκ μπορούν να προσδιοριστούν οι ομάδες στις οποίες οφείλει να στοχεύσει (target groups) η ομάδα μάρκετινγκ. Με αυτόν τον τρόπο, θεωρείται ότι είναι πολύ πιο πιθανό να αντιμετωπιστούν μελλοντικά προβλήματα στις επιχειρήσεις όπως η αβεβαιότητα ως προς τις πληρωμές από συγκεκριμένους πελάτες, αλλά και η σταθερή ή αυξανόμενη πορεία της ζήτησης των προϊόντων της.

Επιπλέον, άλλα σημαντικά ερωτήματα που απαντώνται μέσω της εξόρυξης δεδομένων στις επιχειρήσεις μπορούν να σχετίζονται με διάφορους άλλους λόγους.

Μία περίπτωση συναντάται όταν οι πελάτες μπορεί να αποφασίσουν να μην προβούν στην κατανάλωση ενός προϊόντος από μία συγκεκριμένη επιχείρηση, αλλά να αγοράσουν το ίδιο προϊόν από μία άλλη (χαρακτηριστικό παράδειγμα αποτελούν οι συνδέσεις ίντερνετ). Άλλη μία περίπτωση είναι η εξέταση των σταυροειδών πωλήσεων όπου εξετάζεται ποιο άλλο προϊόν θα αγόραζε ο πελάτης μαζί με το προϊόν που ήδη αγόρασε. Σε αυτή την περίπτωση, δηλαδή, η επιχείρηση επιδιώκει να προσελκύσει τον πελάτη στην αγορά κάποιου ακόμη προϊόντος μαζί με αυτό που ήδη έχει αγοράσει. Παρεμφερής με αυτή την περίπτωση είναι και η κατηγοριοποίηση των πελατών ανά ομάδες (τμηματοποίηση), όπου εξετάζονται τα ειδικά χαρακτηριστικά τους ως ένα μεγαλύτερο σύνολο με σκοπό να επιτευχθούν οι υψηλότερες πωλήσεις ανά κατηγορία.

Πέρα από τα ζητήματα που σχετίζονται αποκλειστικά με τον πελάτη, οι επιχειρήσεις μέσω της εξόρυξης στοχεύουν και στην επίλυση προβλημάτων που σχετίζονται καθαρά με την λειτουργία τους. Για παράδειγμα η διαχείριση των κινδύνων μιας επιχείρησης είναι ένας πρωταρχικός στόχος της από την στιγμή που οποιαδήποτε επιχειρηματική/επενδυτική δραστηριότητα εμπεριέχει ρίσκο. Ένα ακόμη παράδειγμα είναι οι προβλέψεις για το ύψος των πωλήσεων μιας επιχείρησης όπου σε αυτό το στάδιο επιδιώκεται να γίνει μία σωστή διαχείριση των αποθεμάτων, των πρώτων υλών αλλά και των συνολικών προϊόντων της.

Οι Islam and Abedin (2013), ήταν μεταξύ αυτών που εξέτασαν τις επιπτώσεις της διαδικασίας της εξόρυξης δεδομένων στα συστήματα των βάσεων δεδομένων που σχετίζονται με το μάνατζμεντ (Relational Database Management System – RDBMS). Επεσήμαναν το γεγονός ότι η εξόρυξη δεδομένων μπορεί να βοηθήσει στην εύρεση σημαντικών προτύπων ή χρήσιμων πληροφοριών που είτε οδηγούν άμεσα είτε υποβοηθούν την διαδικασία λήψης αποφάσεων σε μία επιχείρηση. Υποστήριξαν μάλιστα ότι αυτή η διαδικασία μπορεί να οδηγήσει στη διατήρηση και ενδυνάμωση του ανταγωνιστικού πλεονεκτήματος της επιχείρησης. Ωστόσο, όπως ανέφεραν, ακόμη και εάν οι επιχειρήσεις έχουν την δυνατότητα πρόσβασης σε πληθώρα δεδομένων, είναι πολύ πιθανό να μην έχουν τα επιθυμητά αποτελέσματα και σε αρκετές περιπτώσεις να είναι ζημιωμένες. Αυτό το γεγονός μπορεί να εξηγηθεί από την παραδοχή ότι ενώ υπάρχουν πολλά δεδομένα και συνεχίζουν να συσσωρεύονται και άλλα, δεν υπάρχει ούτε η κατάλληλη στρατηγική ούτε και η κατάλληλη γνώση προκειμένου αυτά να μετατραπούν σε τέτοια μορφή που να επιτρέπουν την λήψη

αποφάσεων. Γι' αυτό το λόγο επισημαίνουν ότι για τα δεδομένα θα πρέπει να εξασφαλίζεται ασφάλεια, έλεγχος ως προς την προσβασιμότητα, ακρίβεια, εγκυρότητα και συνάφεια μεταξύ τους.

Την λύση στο πρόβλημα που εντόπισαν οι Islam and Abedin (2013), φαίνεται να έδωσε ο Satalkar (2009), ο οποίος μέσω της χρήσης των θεωρητικά κατάλληλων συστημάτων για την εξόρυξη και την ανάλυση των δεδομένων, εντόπισε τα εξής σημαντικά πλεονεκτήματα:

- Η αποθήκευση της πληροφορίας δίνει μία πολύ σημαντική πηγή δεδομένων, η οποία μπορεί να είναι διαθέσιμη σε οποιαδήποτε μελλοντική χρονική στιγμή. Με αυτόν τον τρόπο, η επιχείρηση δεν είναι απαραίτητο να προβαίνει στην αναζήτηση νέων δεδομένων κάθε φορά που προκύπτει η ανάγκη για την επίλυση ενός προβλήματος, αλλά μπορεί να εξασφαλίζει χρόνο μέσω της χρήσης της πληροφορίας που έχει συλλέξει κατά τα προηγούμενα στάδια,
- Ο προσδιορισμός των αποτελεσμάτων που μπορεί να έχει μία κίνηση της επιχείρησης, μπορεί να οδηγήσει στην αποτίμηση των πραγματικών σχέσεων της επιχείρησης με το γενικότερο περιβάλλον και να απλοποιήσει αυτές τις διαδικασίες. Έτσι, οι επιχειρήσεις μπορούν να αποφύγουν κινήσεις, οι οποίες μπορεί να είναι ζημιογόνες ή μπορεί να θεωρηθούν ως επιβλαβείς για την φήμη της,
- Ο προσδιορισμός των πραγματικά χρήσιμων δεδομένων για την επίτευξη των στόχων της επιχείρησης μπορεί να οδηγήσει στη σημαντική μείωση του κόστους αποθήκευσης των δεδομένων, ενώ μπορούν να προσδιοριστούν με πολύ πιο εύκολο τρόπο και οι όποιες αλλαγές. Παράλληλα, με αυτόν τον τρόπο αποκλείονται δεδομένα που δυσχεραίνουν την ανάλυση,
- Το σύστημα διατήρησης και επεξεργασίας των δεδομένων αυτών μπορεί να βασίζεται σε οποιοδήποτε λογισμικό επιλέγει η επιχείρηση ως το καταλληλότερο, ενώ τα δεδομένα αυτά μπορούν να χρησιμοποιηθούν σε οποιονδήποτε τομέα της επιχείρησης.

Σε σχέση με την περίπτωση του ηλεκτρονικού εμπορίου, οι Chen et al. (2012), υποστηρίζουν ότι όταν αναλύεται η περίπτωση του ηλεκτρονικού εμπορίου από την πλευρά των επιχειρήσεων, τότε αν και τα δεδομένα που συλλέγονται είναι λιγότερο δομημένα, συμπεριλαμβάνουν στην πραγματικότητα πλούσια πληροφόρηση για την συμπεριφορά του καταναλωτή. Όλη η ανάλυση που έχει διενεργηθεί έως αυτό το σημείο βασίζεται στο γεγονός ότι τα δεδομένα που χρησιμοποιούνται είτε είναι σε αριθμητική μορφή είτε μετατρέπονται σε αριθμούς. Ωστόσο, πλέον η εξόρυξη δεδομένων στις επιχειρήσεις δεν μπορεί να επέλθει μόνο μέσω της ανάλυσης των δεδομένων η οποία αναφέρεται σε μία καθαρά στατιστική ανάλυση. Αντίθετα, πρέπει να λαμβάνεται υπόψη και η διαδικασία της εξόρυξης κειμένου (text mining), στην οποία συμπεριλαμβάνονται σημαντικές πληροφορίες τόσο για το εσωτερικό όσο και για το εξωτερικό περιβάλλον της επιχείρησης. Γι' αυτό το λόγο, προσδιορίζει ότι οι αναλυτές μιας επιχείρησης θα πρέπει να κατηγοριοποιούνται ως εξής:

- Αναλυτές μεγάλων δεδομένων, όπου η διαδικασία της εξόρυξης βασίζεται στην στατιστική ανάλυση των υπό εξέταση μεγεθών,
- Αναλυτές κειμένων, όπου οι αναλυτές προσπαθούν να εξάγουν γνώση μέσω της ανάλυσης των κειμένων,
- Αναλυτές διαδικτύου, όπου η γνώση εξάγεται από όλη την πληροφόρηση που αντλείται από διαδικτυακούς τόπους, οι οποίοι περιλαμβάνουν πληροφόρηση τόσο για την ίδια την επιχείρηση όσο και για το σύνολο της οικονομίας,
- Αναλυτές δικτύων, όπου αναλύονται τα διάφορα δίκτυα που έχουν αναπτυχθεί τόσο στο εσωτερικό όσο και στο εξωτερικό περιβάλλον της επιχείρησης, με σκοπό την αποτελεσματικότερη λειτουργικότητα της επιχείρησης.

Οι Elovici and Braha (2003) με σκοπό να βελτιώσουν τα αποτελέσματα της εξόρυξης δεδομένων, πρότειναν δύο τρόπους εξόρυξης δεδομένων που μπορούν να βοηθήσουν στην διαδικασία λήψης αποφάσεων μιας επιχείρησης, στηριζόμενοι στην παραδοχή ότι με το πέρασμα του χρόνου κάθε επιχείρηση συσσωρεύει σε βάσεις δεδομένων όλο και πιο λεπτομερή και ακριβή στοιχεία, οι οποίες λόγω του όγκου τους μπορεί να είναι και πιο δύσκολες στην διαχείριση. Τα δεδομένα που χρησιμοποιούνται σε αυτή την περίπτωση αφορούν στο σχεδιασμό των προϊόντων, την διαδικασία παραγωγής

των προϊόντων, τις πωλήσεις, την προώθηση των προϊόντων και τις γενικότερες τάσεις που επικρατούν στις επιχειρήσεις. Για την εξόρυξη δεδομένων, χρησιμοποιούν τον συνδυασμό δύο μεθοδολογιών για την εξόρυξη δεδομένων, ένα νευρωνικό δίκτυο και ένα δέντρο αποφάσεων με σκοπό να εξαχθούν καλύτερα αποτελέσματα σχετικά με την λήψη αποφάσεων των επιχειρήσεων υπό την προϋπόθεση ότι το συνδυαστικό σύστημα δεν θα αποφέρει χειρότερα αποτελέσματα από τα επιμέρους συστήματα. Τα αποτελέσματα δείχνουν ότι η προτεινόμενη συνδυαστική μεθοδολογία είναι πιο αποτελεσματική από τις επιμέρους μεθοδολογίες.

Οι Battiti and Passerini (2010) συνδέουν και αυτοί την επιστήμη της εξόρυξης δεδομένων με την λήψη αποφάσεων, επισημαίνοντας ότι υπάρχει σημαντικός βαθμός αλληλεπίδρασης μεταξύ ανθρώπων και υπολογιστών, αφού η γνώση που λαμβάνεται από τα δεδομένα που συλλέγονται από ανθρώπους ενσωματώνεται σταδιακά στο υπολογιστικό λειτουργικό με αποτέλεσμα να παραχθεί η επιθυμητή γνώση. Χρησιμοποιούν για αυτή την διαδικασία ένα νευρωνικό δίκτυο που βασίζεται στην διαδικασία Kernel με σκοπό να «εκπαιδεύεται» η χρησιμότητα σταδιακά μέσω της διαδικασίας της ανατροφοδότησης, με σκοπό την επιλογή της καλύτερης εναλλακτικής, η οποία βελτιστοποιεί την χρησιμότητα. Το δίκτυο που προτείνουν χρησιμοποιεί διαδικασίες μάθησης, οι οποίες βασίζονται στη λειτουργία του εγκεφάλου. Παράλληλα χρησιμοποιούν δύο διαφορετικές προσεγγίσεις για την βελτιστοποίηση του αλγορίθμου, την μη διαδραστική και την διαδραστική. Η μη διαδραστική βασίζεται στο γεγονός ότι ο διευθυντής ή οποιοσδήποτε λαμβάνει αποφάσεις οφείλει να διαμορφώσει εκ των προτέρων τις προτιμήσεις του σχετικά με τις εναλλακτικές αποφάσεις, ενώ ο αλγόριθμος αντιπροσωπεύει εκ των υστέρων ένα βέλτιστο σετ αποφάσεων κατά Pareto. Η διαδραστική προσέγγιση βασίζεται στο γεγονός ότι χρησιμοποιείται μία τακτική ανατροφοδότησης όπου το άτομο διαμορφώνει εκ νέου την άποψη του σε κάθε στάδιο της λήψης της από την στιγμή που λαμβάνεται διαρκώς νέα πληροφόρηση. Ο αλγόριθμος για την εξόρυξη δεδομένων στην δεύτερη περίπτωση χαρακτηρίζεται από τα ακόλουθα στάδια:

- Εκμάθηση των συναρτήσεων χρησιμότητας μέσω σχετικών παραδειγμάτων που προέρχονται από την επιλογή των δεδομένων και την λήψη αποφάσεων,

- Παρέμβαση κατά την διαδικασία της λήψης αποφάσεων μέσω της άσκησης κριτικής μεταξύ των διαφόρων εναλλακτικών κατατάσσοντας τα ανταγωνιστικά πλεονεκτήματα των διαφορετικών μεταξύ τους αποφάσεων και στη συνέχεια προσδιορίζοντας τις όποιες διαφορές μεταξύ των εναλλακτικών λύσεων,
- Διόρθωση κατά την διαδικασία της εξόρυξης μέσω της διαρκούς ανατροφοδότησης, όπου λαμβάνονται υπόψη χαρακτηριστικά όπως τα ανεπαρκή ή ανακριβή δεδομένα, αλλά και η πιθανότητα τα δεδομένα ή οι αποφάσεις να είναι αντικρουόμενες μεταξύ τους,
- Η χρησιμότητα υπολογίζεται διαρκώς σε κάθε στάδιο της εκμάθησης με σκοπό να βρεθεί η βέλτιστη λύση της.

Η εκμάθηση του αλγορίθμου όπως είναι αναμενόμενο περιλαμβάνει και αυτή κάποια στάδια με σκοπό να εξαχθεί η ασφαλέστερη πληροφόρηση. Αρχικά, επιλέγεται ένα σετ δεδομένων, τα οποία θεωρούνται ως τα καλύτερα σύμφωνα με ένα συγκεκριμένο κριτήριο, στα οποία κατά την πρώτη εφαρμογή, η χρησιμότητα υπολογίζεται με τυχαίο τρόπο. Στη συνέχεια, συλλέγονται και άλλα παρόμοια δεδομένα με σκοπό να προστεθούν σε αυτά τα παραδείγματα και να εξαχθούν ασφαλέστερα αποτελέσματα. Η λύση σύμφωνα με την Kernel μεθοδολογία ακολουθεί και εκπαιδεύει τη συνάρτηση της χρησιμότητας σε όλα τα παραδείγματα με σκοπό την επιλογή της καλύτερης εναλλακτικής. Η μεθοδολογία τους φαίνεται να υπερτερεί σε σχέση με τις υπόλοιπες χρησιμοποιούμενες.

Στα ίδια αποτελέσματα κατέληξαν και οι Huang et al. (2005), οι οποίοι ανέπτυξαν και αυτοί έναν αλγόριθμο που βασίζεται στα νευρωνικά δίκτυα και την διαδικασία της ανατροφοδότησης με σκοπό την καλύτερη εξαγωγή συμπερασμάτων. Περαιτέρω, οι Adomavicius and Tuzhilin (2005), εντόπισαν ότι η επιστήμη της εξόρυξης δεδομένων στον επιχειρηματικό κόσμο μπορεί να αποφέρει συνέπειες όπως η καλύτερη μελέτη των συστάδων επιχειρήσεων, η τμηματοποίηση της αγοράς και η εξόρυξη γραφημάτων που μπορούν να απεικονίσουν τόσο την σημερινή όσο και την μελλοντική κατάσταση της επιχείρησης.

Οι Fountain et al. (2000) υποστηρίζουν ότι για μία επιχείρηση, η διαδικασία της εξόρυξης δεδομένων είναι καθοριστική από την στιγμή που μπορεί να επιφέρει πολύ

μεγάλη και σημαντική μείωση στα λειτουργικά της κόστη, αλλά και να οδηγήσει σε σημαντική αύξηση τόσο στην παραγωγή όσο και στα κέρδη της. Μάλιστα, υποστηρίζουν ότι αυτή η διαδικασία είναι ακόμα πιο σημαντική για τις επιχειρήσεις που δραστηριοποιούνται στον τομέα της τεχνολογίας από την στιγμή που τα προϊόντα αυτά χαρακτηρίζονται από μικρή διάρκεια ζωής. Εξετάζοντας την περίπτωση της κατασκευής και λειτουργίας κυκλωμάτων στην εταιρεία Hewlett Packard (HP), διαπιστώνουν ότι η εξόρυξη δεδομένων μπορεί να οδηγήσει στην γρηγορότερη ανατροφοδότηση και κατασκευή συστημάτων όταν εντοπίζεται κάποιο πρόβλημα, το οποίο μπορεί να διακόψει την λειτουργία του.

Οι Braha et al. (2007) επικεντρώνοντας, επίσης, το ενδιαφέρον τους στον κλάδο της τεχνολογίας, επίσης, διαπιστώνουν ότι, οι όσο πιο δυνατόν ακριβείς προβλέψεις της απόδοσης της βιομηχανικής διαδικασίας και των τάσεων της, έχουν ως αποτέλεσμα την συνολική μείωση στα κόστη λειτουργίας μιας επιχείρησης. Γι' αυτό το λόγο χρησιμοποιούν ένα θεωρητικό μοντέλο κατηγοριοποίησης των επιμέρους αποφάσεων, όπου στην ουσία οι προτεινόμενες αποφάσεις κατατάσσονται σε σειρά και επιλέγεται αυτή η οποία μπορεί να χαρακτηριστεί ως η καλύτερη ανάμεσα στις εναλλακτικές. Συνοπτικά, ο αλγόριθμος που χρησιμοποιούν αποτελείται από τα ακόλουθα βήματα:

- Ο αλγόριθμος δημιουργεί «σήματα» για τις μη εύκολα αντιληπτές καταστάσεις που υπάρχουν στον πραγματικό κόσμο,
- Από το σετ των δυνητικών αποφάσεων επιλέγεται αυτή η οποία έχει τα περισσότερα οφέλη σύμφωνα με τον κανόνα απόφασης. Πιο συγκεκριμένα, στο προς επίλυση πρόβλημα «ενσωματώνεται» ένα αντίστοιχο προηγούμενο παράδειγμα, στο οποίο ο ερευνητής παρατηρεί τον τρόπο που είχαν ληφθεί κατά το παρελθόν οι αποφάσεις, την κατάταξη που είχε λάβει αυτή η εναλλακτική και τα αποτελέσματα που είχαν επέλθει,
- Τέλος, χρησιμοποιείται ο κανόνας απόφασης που έχει προσδιοριστεί εξαρχής προκειμένου να επιλεχθεί η καλύτερη δυνατή λύση.

Προκειμένου να βελτιωθεί η κερδοφορία των επιχειρήσεων, πολλές τεχνικές έχουν εφαρμοστεί. Ωστόσο, η αβεβαιότητα που διέπει την ποιότητα των εξαγόμενων προϊόντων αλλά και των πρώτων υλών, την ορθή χρήση του εξοπλισμού και του γενικότερου οικονομικού περιβάλλοντος, μπορεί να οδηγήσει σε αβέβαια αποτελέσματα. Σε αυτό το σημείο, οι Tobin et al. (2000), θεωρούν ότι πρέπει να χρησιμοποιούνται αυτοματοποιημένα συστήματα και διαδικασίες με σκοπό την αποτελεσματική καταγραφή και επεξεργασία των δεδομένων που έχουν συλλεχθεί, τα οποία μπορούν να οδηγήσουν στην ακριβή πρόβλεψη αυτής της πορείας.

Επίσης πέρα από τους Braha et al. (2007), ο Shalak (1995) είναι μεταξύ των ερευνητών που πρότειναν την μεθοδολογία της κατάταξης για την λήψη αποφάσεων από τις επιχειρήσεις, αλλά προσδιορίζοντας ότι τα καλύτερα αποτελέσματα μπορούν να επιτευχθούν όταν συνδυάζονται διάφορες τάξεις μεταξύ τους. Η λογική σε αυτόν τον συνδυασμό εντοπίζεται στο γεγονός ότι κάθε στοιχείο προς ανάλυση καθορίζει τα αποτελέσματα σε όλες τις τάξεις, και η μόνη διαφορά που εντοπίζεται είναι ότι αυτό το στοιχείο συμμετέχει με διαφορετική βαρύτητα σε κάθε τάξη. Η σύνθετη πρόβλεψη προκύπτει ως το προϊόν της σύνθετης κατάταξης, η οποία ενσωματώνει τις προβλέψεις όλων των μεμονωμένων τάξεων σε μία. Η σύνθεση αυτή μπορεί να προκύψει ως ο μέσος σταθμικός όρος όλων των τάξεων, ή ως η σύνθεση μόνον των ανεξάρτητων τάξεων με σκοπό την εξαγωγή του συνόλου της πληροφορίας. Ο πιο συνηθισμένος τρόπος υπολογισμού της σύνθετης πρόβλεψης κατά συνέπεια είναι η καρτεσιανή κατάταξη σύμφωνα με την οποία δύο ή περισσότερες ανεξάρτητες μεταξύ τους τάξεις συνδυάζονται σε μία, η οποία κατατάσσει τα στοιχεία που προέρχονται από τις επιμέρους τάξεις.

Γενικότερα για τις επιχειρήσεις μπορούμε να πούμε ότι το σχήμα που ακολουθείται περιγράφεται ως:

- Σε πρώτο στάδιο εντοπίζονται όλες οι πηγές δεδομένων που μπορούν να χρησιμοποιηθούν για την ανάλυση του προβλήματος της επιχείρησης και οι οποίες μπορεί να περιλαμβάνουν στοιχεία από προηγούμενες αναλύσεις, τα αρχεία της επιχείρησης και άλλες βάσεις δεδομένων,

- Σε δεύτερο στάδιο γίνεται η ομαδοποίηση/ενοποίηση όλων των δεδομένων που έχουν εντοπιστεί κατά το πρώτο στάδιο και δημιουργείται η ενιαία αποθήκη των δεδομένων μας,
- Σε τρίτο στάδιο διενεργείται μία αρχική στατιστική επεξεργασία των δεδομένων με σκοπό την καλύτερη προετοιμασία για την έναρξη της εξόρυξης δεδομένων,
- Σε τέταρτο στάδιο πραγματοποιείται η εξόρυξη των δεδομένων με τη χρήση κάποιου εκ των αλγορίθμων που περιγράφηκαν παραπάνω και παράγεται η επιθυμητή γνώση,
- Σε πέμπτο στάδιο ακολουθεί η διαγραμματική απεικόνιση επί της ουσίας των αποτελεσμάτων που έχουν εξαχθεί κατά το στάδιο της εξόρυξης,
- Σε έκτο στάδιο ακολουθεί η λήψη αποφάσεων, η οποία αποτελεί και το ουσιαστικότερο τμήμα της διαδικασίας της εξόρυξης των δεδομένων, αφού αντιστοιχεί και στην στρατηγική που ακολουθείται από την επιχείρηση.


Πηγές Δεδομένων (Αρχεία, Προηγούμενες Αναλύσεις, Βάσεις Δεδομένων)
Αποθήκη Δεδομένων (δεδομένα από όλες τις πηγές)
Εξερεύνηση των Δεδομένων (Στατιστική Ανάλυση και Σχετικές Αναφορές)
Εξόρυξη Δεδομένων (Παραγωγή Γνώσης)
Παρουσίαση των Δεδομένων (Γραφικές Αναπαραστάσεις κλπ)
Λήψη Αποφάσεων

5. Εξόρυξη Δεδομένων στις Επιχειρήσεις: Ποια αναμένουμε ότι θα είναι η Μελλοντική της Πορεία

Από την ανάλυση που έχει διενεργηθεί έως αυτό το σημείο μπορούμε να διαπιστώσουμε ότι η εξόρυξη δεδομένων γενικότερα ορίζεται ως η διερεύνηση και η ανάλυση των πρωτογενών δεδομένων που έχουν ως σκοπό την ανάδειξη συγκεκριμένων δομών και προτύπων ανάμεσα στις υπό εξέταση μεταβλητές. Επίσης, διαπιστώσαμε ότι για τις επιχειρήσεις, η εξόρυξη δεδομένων έχει ως βασικό στόχο την ενίσχυση του ανταγωνιστικού της πλεονεκτήματος με σκοπό τόσο την διατήρηση όσο και την ενίσχυση του μεριδίου τους στην αγορά.

Με το πέρασμα των ετών και στηριζόμενοι στην παραδοχή ότι η οικονομία χαρακτηρίζεται από έντονα στοιχεία αβεβαιότητας, η ανάγκη για την εξόρυξη δεδομένων στις επιχειρήσεις είναι ακόμη μεγαλύτερη. Ο όγκος των δεδομένων έχει την τάση να αυξάνεται σε πολύ σημαντικό βαθμό δυσχεραίνοντας την ανάλυση, η οποία πλέον μπορεί να επιτευχθεί μόνον μέσω της διαθεσιμότητας των κατάλληλων τεχνολογικών υποδομών και λογισμικών. Μάλιστα, αναμένουμε ότι όσο βελτιώνεται η υπάρχουσα τεχνολογία και η καταγραφή των δεδομένων γίνεται όλο και πιο εύκολα, τόσο μεγαλύτερος θα είναι και ο όγκος των δεδομένων, ενώ και οι βάσεις θα χαρακτηρίζονται από μεγαλύτερη πολυπλοκότητα.

Για τις επιχειρήσεις παρατηρήσαμε ότι η ανάγκη για την εξόρυξη δεδομένων στις επιχειρήσεις οφείλεται κυρίως στον μεγάλο ανταγωνισμό που αυτές αντιμετωπίζουν διαρκώς και την διαρκώς αυξανόμενη νέα πληροφόρηση που εισρέει σε αυτές. Όσο αυξάνεται ο αριθμός των επιχειρήσεων που δραστηριοποιούνται σε έναν κλάδο και η παγκοσμιοποίηση στην πραγματικότητα εκμηδενίζει τις αποστάσεις μεταξύ των επιχειρήσεων που δραστηριοποιούνται ακόμα και σε διαφορετικές μεταξύ τους χώρες, η εξόρυξη των δεδομένων θεωρούμε ότι μπορεί να επιφέρει πολύ σημαντικά αποτελέσματα. Πιο συγκεκριμένα, μέσω της εξόρυξης θεωρούμε ότι μπορούν να κατασκευαστούν τα πλέον αποτελεσματικά επιχειρηματικά σχέδια πάνω στα οποία θα βασίζονται οι μελλοντικές κινήσεις των επιχειρήσεων.

Για παράδειγμα, ιδίως για τις επιχειρήσεις που δραστηριοποιούνται στον τομέα του ηλεκτρονικού εμπορίου οι αποστάσεις μεταξύ των επιχειρήσεων είναι μηδενικές και ο ανταγωνισμός μπορεί να είναι ακόμα και παγκόσμιος. Ένα χαρακτηριστικό

παράδειγμα είναι η περίπτωση αγορών προϊόντων τεχνολογίας από το διαδίκτυο, όπου πολλές φορές οι καταναλωτές μπορεί να αποφασίσουν να αγοράσουν το ίδιο προϊόν ακόμη και από το εξωτερικό. Η επιλογή τους αυτή προφανώς σχετίζεται με την τιμή του προϊόντος (υποθέτοντας ότι πρόκειται για ένα ομοιογενές προϊόν που πωλείται σε όλες τις χώρες), η οποία μπορεί να είναι σημαντικά μικρότερη σε κάποια άλλη χώρα από την στιγμή που η επιχείρηση μπορεί να έχει αναπτύξει τρόπους ελαχιστοποίησης του κόστους, οι οποίοι μπορεί να σχετίζονται με την ύπαρξη εσωτερικών οικονομιών κλίμακας, που έχουν προκύψει ως αποτέλεσμα της εξόρυξης δεδομένων. Γι' αυτό το λόγο θεωρούμε ότι η εξόρυξη δεδομένων μπορεί να διαδραματίσει σημαντικό ρόλο στην μείωση αυτού του κόστους. Δηλαδή, η εξόρυξη να μην περιοριστεί σε δεδομένα τοπικού χαρακτήρα αλλά να συμπεριλάβει δεδομένα σχετικά και με τον εξωτερικό τομέα. Με αυτόν τον τρόπο λαμβάνεται ακόμη μεγαλύτερο μέρος πληροφοριών, και άρα η εξόρυξη είναι πιο αποτελεσματική ως προς τα συμπεράσματα που βγαίνουν.

Στο μέλλον, θεωρούμε ότι η ανάπτυξη του κλάδου της εξόρυξης δεδομένων πρέπει να αντιμετωπίσει τα βασικά προβλήματα που αντιμετωπίζονται σύμφωνα με την διεθνή βιβλιογραφία. Πιο συγκεκριμένα σε αυτό το κεφάλαιο στόχος μας είναι να κάνουμε κάποιες προτάσεις σχετικά με την επίλυση αυτών των προβλημάτων, τα οποία είναι:

- Η περιορισμένη ορθολογικότητα των επιχειρηματιών και λαμβανόντων τις αποφάσεις,
- Η περιορισμένη πληροφόρηση,
- Η αβεβαιότητα για την έκβαση των αποτελεσμάτων,
- Η ανεπάρκεια των χρησιμοποιούμενων αλγορίθμων,
- Τα λάθη που γίνονται κατά την κωδικοποίηση των δεδομένων,
- Η αποφυγή μη συμπερίληψης των απαραίτητων δεδομένων,
- Η αποφυγή προσβολής της προσωπικότητας του ατόμου και της παραβίασης των ιδιωτικών τους δικαιωμάτων.

5.1 Αντιμετώπιση της Περιορισμένης Ορθολογικότητας

Η περιορισμένη ορθολογικότητα εντοπίζεται στο γεγονός ότι όλα τα άτομα ακόμα και εάν έχουν διαθέσιμη όλη την απαραίτητη πληροφόρηση, δεν έχουν την δυνατότητα να την κρίνουν με ακρίβεια από την στιγμή που δεν μπορούν να επεξεργαστούν το σύνολό της ταυτόχρονα. Πρακτικά, το πρόβλημα αυτό θα μπορούσε να διορθωθεί, αν όχι να ξεπεραστεί, σε ένα σημαντικό βαθμό μέσω της απλοποίησης της πληροφορίας. Δηλαδή, η εξόρυξη δεδομένων να εξακολουθήσει να βασίζεται στις βάσεις δεδομένων που υπάρχουν είτε σε ιδιωτικές είτε σε δημόσιες πηγές, αλλά αυτές να είναι κατασκευασμένες με πολύ πιο απλό τρόπο προκειμένου να μπορεί ο λαμβάνων τις αποφάσεις να επεξεργαστεί το σύνολο της πληροφορίας. Η απλοποίηση των βάσεων δεδομένων μπορεί να επιτευχθεί μέσω του όσο πιο ακριβούς εντοπισμού των παραγόντων που σχετίζονται με το πρόβλημα προς επίλυση, αλλά και την καλύτερη δόμηση των ίδιων των βάσεων. Αυτό πρακτικά σημαίνει ότι ο τομέας αποθήκευσης των δεδομένων θα πρέπει να χαρακτηρίζεται τόσο από την λεπτομερή καταγραφή του κάθε στοιχείου όσο και από την κατάλληλη οργάνωσή τους στο χώρο. Δηλαδή, τα δεδομένα θα θέλαμε να είναι δομημένα με σειρά και να ενημερώνονται διαρκώς προκειμένου η εξόρυξη να βασίζεται στα πιο πρόσφατα στοιχεία και όχι στοιχεία που μπορεί να αφορούσαν το συγκεκριμένο πρόβλημα κάποια χρόνια πριν. Ακόμη κι αν η IBM, η Oracle και η Microsoft, έχουν κάνει σημαντικά βήματα προς αυτόν τον τομέα δεν μπορούμε να πούμε με ακρίβεια ότι τα αποτελέσματα είναι τα απολύτως ικανοποιητικά, από την στιγμή που οι επιχειρήσεις εξακολουθούν να αντιμετωπίζουν σημαντικά προβλήματα.

5.2 Αντιμετώπιση της Περιορισμένης Πληροφόρησης

Το δεύτερο πολύ σημαντικό πρόβλημα σχετίζεται με την περιορισμένη πληροφόρηση. Η περιορισμένη πληροφόρηση προκύπτει εξαιτίας του γεγονότος ότι ακόμη και εάν το διαδίκτυο, οι εφημερίδες, οι ισολογισμοί μπορούν να δώσουν μία σαφή εικόνα της επιχείρησης αλλά και του γενικότερου περιβάλλοντος στο οποίο αυτή δραστηριοποιείται, δεν υπάρχει η δυνατότητα αξιοποίησης του συνόλου της

πληροφορίας. Αυτό το πρόβλημα αντιμετωπίζεται από την στιγμή που οι επιχειρήσεις δεν έχουν την δυνατότητα πρόσβασης σε αρχεία κρυφά άλλων επιχειρήσεων (ανταγωνιστριών και μη), ούτε και στον σχεδιασμό των αποφάσεων που λαμβάνονται από ανώτατα όργανα, όπως οι κυβερνήσεις ή τα επιμελητήρια ή τα συνδικάτα κλπ. Προφανώς, η πληροφόρηση αυτή δεν πρόκειται ποτέ να γίνει δημόσια από την στιγμή που εάν διαρρεύσει τότε θα χαθεί το ανταγωνιστικό πλεονέκτημα των επιχειρήσεων και αυτές θα βρεθούν σε δυσχερέστερη θέση. Ωστόσο, η κάθε επιχείρηση θα πρέπει να είναι σε θέση να παρακολουθεί τις κινήσεις των άλλων επιχειρήσεων σε τακτά χρονικά διαστήματα και να μπορεί να προβλέψει ως έναν βαθμό τις μελλοντικές τους κινήσεις, σχεδιάζοντας παράλληλα τις δικές της αντιδράσεις σε αυτές. Με αυτόν τον τρόπο πέρα από τον χρησιμοποιούμενο αλγόριθμο εξόρυξης δεδομένων, θεωρούμε ότι η επιχείρηση θα πρέπει να είναι σε θέση να κατασκευάσει και σενάρια σχετικά με τις κινήσεις των ανταγωνιστών της, τις σχεδιαζόμενες οικονομικές πολιτικές αλλά και την γενικότερη κατάσταση της οικονομίας και του περιβάλλοντος. Παράλληλα, η επιχείρηση θα πρέπει να σχεδιάζει και τις δικές κινήσεις ως προς την αντιμετώπιση της οποιασδήποτε κίνησης με σκοπό να διατηρήσει τόσο το ανταγωνιστικό της πλεονέκτημα όσο και την θέση της στην αγορά.

5.3 Αντιμετώπιση της Αβεβαιότητας των Αποτελεσμάτων

Η ανάλυση των εναλλακτικών σεναρίων που προτείναμε παραπάνω με σκοπό την αντιμετώπιση του προβλήματος της περιορισμένης πληροφόρησης σχετίζεται άμεσα και με την αντιμετώπιση του προβλήματος της αβεβαιότητας των αποτελεσμάτων που προκύπτουν από την εξόρυξη δεδομένων. Με την κίνηση της επιχείρησης να κατασκευαστούν εναλλακτικά σενάρια ανάλογα με τις αντιδράσεις των ανταγωνιστών και του γενικότερου περιβάλλοντός της, μπορεί να υπάρξει μεγαλύτερη βεβαιότητα ως προς την επίτευξη των στόχων της από την στιγμή που θα έχει αξιοποιηθεί στο μεγαλύτερο δυνατό βαθμό η διαθέσιμη πληροφορία.

5.4 Αντιμετώπιση της Ανεπάρκειας των Αλγορίθμων

Ένα ακόμη σημαντικό πρόβλημα που συναντάται είναι αυτό της ανεπάρκειας των διαθέσιμων αλγορίθμων. Το πρόβλημα προκύπτει από το γεγονός ότι η πληροφορία πλέον είναι όλο και πιο σύνθετη με αποτέλεσμα να μην μπορεί να αναλυθεί πλήρως από όλους τους αλγορίθμους. Παράλληλα, διαπιστώνουμε ότι οι τεχνικές που έχουν αναπτυχθεί για την εξόρυξη δεδομένων βασίζονται σε δύο διαφορετικούς κλάδους. Πρώτος είναι ο κλάδος της στατιστικής, όπου όπως μπορούμε να παρατηρήσουμε η πλειοψηφία των μεθοδολογιών που παρουσιάστηκαν στο δεύτερο κεφάλαιο βασίζονται σε αυτήν (για παράδειγμα οι συσχετίσεις, η παλινδρόμηση και ο υπολογισμός αποκλίσεων). Ο δεύτερος κλάδος που χρησιμοποιείται είναι αυτός της τεχνολογίας μέσω των ηλεκτρονικών υπολογιστών, όπου παρατηρούμε ότι κάποιες μεθοδολογίες στηρίζονται σε αυτόν τον κλάδο μεταξύ των οποίων συμπεριλαμβάνονται τα τεχνητά νευρωνικά δίκτυα και οι γενετικοί αλγόριθμοι. Όπως αντιλαμβανόμαστε κάθε ένας από αυτούς τους δύο κλάδους συσχετίζεται άμεσα με την ύπαρξη βάσεων δεδομένων στις οποίες εμπεριέχεται όλο το διαθέσιμο σύνολο της πληροφορίας.

Διαπιστώσαμε με βάση και τις εμπειρικές μελέτες που παρουσιάστηκαν ότι ακόμη κι αν οι χρησιμοποιούμενοι αλγόριθμοι έδωσαν κάποια θετικά αποτελέσματα, η πολυπλοκότητα και η αβεβαιότητα που χαρακτηρίζει την σημερινή εποχή μπορεί να οδηγήσει σε λάθη ακόμα και τους πιο αποτελεσματικούς αλγόριθμους. Μία λύση που μπορούμε να προτείνουμε προέρχεται από το επιστημονικό πεδίο της στατιστικής και αφορά στις προβλέψεις, οι οποίες συνδυάζουν περισσότερα από ένα υποδείγματα. Σε αυτήν την περίπτωση, θεωρούμε, ότι ο συνδυασμός περισσότερων του ενός αλγορίθμου, μπορεί να επιφέρει σημαντικά αποτελέσματα. Είδαμε, ότι ήδη έχουν γίνει κάποια βήματα προς αυτή την κατεύθυνση, αλλά ακόμη μπορούν να γίνουν πολλές σημαντικές αλλαγές. Για παράδειγμα, αναμένουμε ότι σε έναν συνδυαστικό αλγόριθμο θα πρέπει να τα ποσοστά συμμετοχής του κάθε ενός αλγορίθμου να υπολογίζονται με βάση ένα κριτήριο. Το κριτήριο αυτό θα πρέπει να αποφασιστεί εκ των προτέρων προκειμένου να λυθεί και το πρόβλημα μεγιστοποίησης ή ελαχιστοποίησης. Για παράδειγμα, για μία επιχείρηση σκοπός της είναι στο σύνολό της είτε η μεγιστοποίηση των κερδών είτε η ελαχιστοποίηση του κόστους παραγωγής. Εάν όμως οι προβλέψεις που επιθυμεί να διεξάγει αφορούν το τμήμα του μάρκετινγκ,

για παράδειγμα, τότε θεωρούμε ότι μπορεί το πρόβλημα της μεγιστοποίησης να μην εστιάζει στην μεγιστοποίηση των κερδών αλλά στη μεγιστοποίηση της χρησιμότητας των καταναλωτών, η οποία θα επιφέρει σημαντικές αλλαγές στο σύνολο της επιχείρησης.

Αφού αποφασιστεί το κριτήριο με βάση το οποίο θα αποφασιστούν τα ποσοστά συμμετοχής των επιμέρους αλγορίθμων, το οποίο θα πρέπει να ελαχιστοποιεί και τα τυπικά σφάλματα του τελικού μοντέλου, θα πρέπει να επιλεχθούν οι αλγόριθμοι που θα μπουν σε αυτή την διαδικασία. Προκειμένου να αυξηθεί το σετ της πληροφόρησης μπορούμε να αφήσουμε τους αλγορίθμους μας να έχουν διαφορετικά αρχικά δεδομένα, αλλά και να επιτρέψουμε να λαμβάνουν ακόμη και αρνητικά ποσοστά συμμετοχής εάν το κριτήριό μας θεωρεί ότι κάποιος από αυτούς δεν είναι ο καταλληλότερος για την αντιμετώπιση αυτού του προβλήματος. Ταυτόχρονα, θα πρέπει να βάλουμε και την παράμετρο ότι το συνολικό παραχθέν μοντέλο, θα πρέπει να μας δίνει καλύτερα αποτελέσματα (κυρίως ως προς κάποια στατιστικά κριτήρια, αλλά και με τον διαχωρισμό του δείγματος σε in – sample και out – of – sample) από τους επιμέρους αλγορίθμους.

5.5 Αντιμετώπιση της Λάθους Κωδικοποίησης των Δεδομένων μας

Το επόμενο σημαντικό πρόβλημα που αντιμετωπίζεται κατά την διαδικασία της εξόρυξης δεδομένων είναι αυτό της λάθος κωδικοποίησης των δεδομένων μας. Αναμένουμε ότι όσο αυξάνονται οι πηγές από τις οποίες λαμβάνονται τα δεδομένα μας και αυτές είναι όλο και πιο πολύπλοκες, τόσο πιο δύσκολη θα είναι και η κωδικοποίησή τους. Για παράδειγμα, όταν μιλάμε για μεγέθη όπως οι πωλήσεις ή τα κόστη, η κωδικοποίησή τους θεωρείται πιο εύκολη από την στιγμή που αποτελούν αριθμητικά δεδομένα. Όταν όμως τα δεδομένα μας είναι σε μορφή κειμένου ή εικόνας, τότε η κωδικοποίηση δεν μπορεί να θεωρηθεί ως η βέλτιστη δυνατή, από την στιγμή που πλέον αυτή έγκειται στην προσωπική κρίση του ερευνητή. Σε αυτό το σημείο αντιμετωπίζουμε το πρόβλημα της μεροληψίας από την πλευρά του ερευνητή,

ο οποίος κατά την κρίση του μπορεί να θεωρήσει ως λιγότερο ή περισσότερο σημαντικά κάποια στοιχεία των δεδομένων. Η πρόταση που κάνουμε σε αυτή την περίπτωση είναι η ανάλυση των στοιχείων αυτών από περισσότερους από έναν ερευνητές/αναλυτές με σκοπό τον εντοπισμό των πραγματικά χρήσιμων πληροφοριών και την ελαχιστοποίηση του σφάλματος μεροληψίας. Με αυτόν τον τρόπο, η κωδικοποίηση βασίζεται σε μία συνολικότερη γενικά εικόνα από την πλευρά των ερευνητών με σκοπό την εξασφάλιση της μεγαλύτερης αντικειμενικότητας.

5.6 Αντιμετώπιση της Μη Συμπερίληψης των Κατάλληλων Δεδομένων

Η επιλογή των κατάλληλων δεδομένων, όπως είδαμε και σε προηγούμενα κεφάλαια, μπορεί να επηρεάσει σημαντικά την ποιότητα των αποτελεσμάτων της εξόρυξης δεδομένων. Έτσι, εάν δεν συμπεριληφθούν κατά την διάρκεια της εξόρυξης τα κατάλληλα και απαραίτητα δεδομένα είναι πάρα πολύ πιθανό να οδηγηθούμε σε ανεπαρκή ή λάθος αποτελέσματα.

Όπως και η κωδικοποίηση των δεδομένων, έτσι και η επιλογή των πλέον κατάλληλων δεδομένων μπορούμε να πούμε ότι είναι μία καθαρά υποκειμενική κρίση, η οποία μπορεί να επηρεαστεί από τους λαμβάνοντες τις αποφάσεις. Γι' αυτό και σε αυτή την περίπτωση προτείνουμε τα δεδομένα να επιλέγονται από περισσότερους τους ενός ερευνητές/αναλυτές με αποτέλεσμα να αυξηθεί η αμεροληψία του δείγματος.

5.7 Αντιμετώπιση της Πιθανής Προσβολής της Προσωπικότητας του Ατόμου και της Παραβίασης των Ιδιωτικών του Δικαιωμάτων

Σε αυτή την περίπτωση εξετάζουμε την ηθική περισσότερο πλευρά της διαδικασίας της εξόρυξης δεδομένων παρά την επιστημονική της χροιά. Όπως είδαμε και στο κεφάλαιο 3, η ηθική είναι ένας καθαρά υποκειμενικός παράγοντας, ο οποίος επηρεάζεται από στοιχεία όπως η κουλτούρα, η γεωγραφική θέση της χώρας, αλλά και η τεχνολογία. Παρά τα όσα βήματα έχουν διενεργηθεί κατά τα τελευταία χρόνια με βάση την νομοθεσία, πολλές φορές τα προσωπικά δεδομένα υποκλέπτονται από τους καταναλωτές μέσω της παρακολούθησης των αγορών τους (αυτή η περίπτωση αφορά κυρίως στο ηλεκτρονικό εμπόριο). Γι' αυτό το λόγο, θεωρούμε ότι πέρα από την νομοθεσία, τα προσωπικά δεδομένα και ο σεβασμός της προσωπικότητας του ατόμου, θα πρέπει να προστατεύονται και από την ίδια την επιχείρηση. Η προστασία αυτή μπορεί να επέλθει κυρίως μέσω της καλλιέργειας ενός εσωτερικού ηθικού κώδικα δεοντολογίας, ο οποίος θα στοχεύει και στην προάσπιση των δικαιωμάτων των καταναλωτών. Παράλληλα, οι καταναλωτές θα πρέπει να είναι πλήρως ενημερωμένοι για την καταγραφή των δεδομένων τους, αλλά να νιώθουν ότι αποκομίζουν και οι ίδιοι κάποια σημαντικά οφέλη από την ίδια την επιχείρηση. Τα οφέλη αυτά μπορούν να αφορούν τόσο στη βελτίωση των παρεχόμενων προϊόντων και υπηρεσιών προς αυτούς, όσο και σε κάποιες επιπρόσθετες παροχές όπως κάποιες χορηγίες ή κάποιες δωρεές ή εκδηλώσεις στις οποίες και οι ίδιοι θα έχουν τη δυνατότητα συμμετοχής.

Με βάση τα αποτελέσματα αυτού του κεφαλαίου μπορούμε να συμπεράνουμε ότι μπορούν να διενεργηθούν σημαντικές αλλαγές ακόμα στον τομέα της εξόρυξης δεδομένων, οι οποίες θα βελτιώσουν τα ήδη υπάρχοντα συστήματα ή θα δημιουργήσουν νέα.

Συμπεράσματα

Αυτή η πτυχιακή εργασία αποτέλεσε στην ουσία μία λεπτομερή εξέταση της διαδικασίας της εξόρυξης δεδομένων τόσο σε γενικό επίπεδο όσο και στην περίπτωση των επιχειρήσεων. Σκοπός μας δεν ήταν η πρακτική εφαρμογή κάποιου συγκεκριμένου αλγορίθμου εξόρυξης δεδομένων, αλλά η παρουσίαση των αλγορίθμων που έχουν χρησιμοποιηθεί στην διεθνή βιβλιογραφία καθώς και οι συνολικότερες επιπτώσεις της εξόρυξης δεδομένων. Η μελέτη αυτή κρίθηκε ως απαραίτητο εάν αναλογιστούμε και το γεγονός ότι περισσότερα από 10.000 άρθρα γράφτηκαν την περίοδο 2000 – 2011 σχετικά με την εξόρυξη δεδομένων.

Η εξέταση της επιστήμης της εξόρυξης δεδομένων θεωρείται ακόμα πιο σημαντική εάν αναλογιστούμε και το γεγονός ότι αυτή δεν αφορά στην συμπερασματική επεξεργασία ερωτημάτων, η οποία σχετίζεται μόνον με την απάντηση απλών ερωτήσεων για τις οποίες δεν απαιτείται η χρήση κάποιων βάσεων δεδομένων, ούτε όμως και με τα απλά στατιστικά προγράμματα που παράγουν κάποιες απλές στατιστικές πράξεις. Αντίθετα, αυτή αφορά στην εύρεση νέας γνώσης.

Στο πρώτο κεφάλαιο αυτής της εργασίας προσδιορίσαμε την εξόρυξη δεδομένων ως την επιστήμη που έχει ως στόχο τον εντοπισμό νέας και χρήσιμης γνώσης, η οποία εμπεριέχεται στα δεδομένα μας. Μάλιστα, κάναμε τον πολύ σημαντικό διαχωρισμό ανάμεσα στην γνώση και την πληροφορία, όπου η πρώτη έχει ως στόχο τον εντοπισμό των προτύπων ανάμεσα στις υπό εξέταση μεταβλητές, ενώ η δεύτερη αντιστοιχεί στην πραγματικότητα στα δεδομένα μας. Επιπλέον, προσδιορίσαμε ότι η εξόρυξη δεδομένων μπορεί είτε να αποσκοπεί στην επιβεβαίωση κάποιων αρχικών υποθέσεων είτε στην εξεύρεση νέων κανόνων και προτύπων είτε ακόμη και στον εντοπισμό ασυνήθιστων προτύπων μεταξύ των μεταβλητών. Ωστόσο, όπως διαπιστώσαμε η πρόβλεψη των μελλοντικών καταστάσεων, η οποία διασυνδέεται με την εξόρυξη δεδομένων, πρέπει να γίνεται με την μικρότερη δυνατή προσπάθεια από τον ερευνητή. Αυτό σημαίνει ότι οι αλγόριθμοι πρέπει να είναι τόσο αποτελεσματικοί έτσι ώστε να μπορούν να εξάγουν τα επιθυμητά αποτελέσματα από διάφορα είδη δεδομένων.

Στο δεύτερο κεφάλαιο, παρουσιάζονται οι αλγόριθμοι που έχουν χρησιμοποιηθεί σε έρευνες της διεθνούς βιβλιογραφίας. Μεταξύ αυτών περιλαμβάνονται τα τεχνητά

νευρωνικά δίκτυα, τα δέντρα αποφάσεων, ο υπολογισμός συσχετίσεων, η παλινδρόμηση, η κατάταξη, η συσταδοποίηση αλλά και ο εντοπισμός του κοντινότερου γείτονα. Όπως διαπιστώσαμε τα αποτελέσματα που προκύπτουν από τους αλγορίθμους αυτούς μπορούν είτε να χρησιμοποιηθούν ως τελική γνώση είτε ως εισροές άλλων συστημάτων. Ωστόσο, πριν καταλήξει ο ερευνητής στη χρήση ενός συγκεκριμένου αλγορίθμου για την εξόρυξη δεδομένων από μία πηγή οφείλει να προβεί σε έναν αρχικό έλεγχό του προκειμένου να εντοπίσει εάν αυτός μπορεί ή όχι να δώσει ακριβή αποτελέσματα.

Το τρίτο κεφάλαιο, αφορούσε στη διασύνδεση ενός περισσότερο φιλοσοφικού παρά επιστημονικού ορισμού, της ηθικής με την εξόρυξη δεδομένων. Η ηθική όπως διαπιστώσαμε αποτελεί περισσότερο μία υποκειμενική έννοια από τη στιγμή που το κάθε άτομο είναι διαφορετικό και δεν μπορούν όλα τα άτομα να αντιδρούν με τον ίδιο τρόπο σε όλες τις περιστάσεις. Μάλιστα, εντοπίσαμε ότι η ηθική συνδέεται άμεσα με την εξόρυξη δεδομένων, αφού η τελευταία οδηγεί στη λήψη σημαντικών αποφάσεων, οι οποίες με τη σειρά τους επηρεάζουν το σύνολο της κοινωνίας.

Στο τέταρτο κεφάλαιο, παρουσιάσαμε την περίπτωση της εξόρυξης δεδομένων στις επιχειρήσεις. Όπως είδαμε, η εξόρυξη δεδομένων στις επιχειρήσεις είναι μία ιδιαίτερα σημαντική διαδικασία από την στιγμή που επηρεάζει την λήψη αποφάσεων την ίδια στιγμή που η επιχείρηση επιδιώκει να αντιμετωπίσει τα αυξημένα επίπεδα ανταγωνισμού στην αγορά. Στις επιχειρήσεις, η ανάγκη για την εξόρυξη δεδομένων έγινε πιο άμεσα αντιληπτή όταν άρχισε να αυξάνεται ο όγκος των βάσεων δεδομένων που κατείχαν οι ίδιες και αναγκάστηκαν να προβούν σε μία πιο αποτελεσματική αποθήκευση των δεδομένων τους. Γι’ αυτό το λόγο μεγάλες εταιρείες παραγωγής λογισμικών προγραμμάτων δημιούργησαν συναφή προϊόντα, τα οποία διέθεσαν στις επιχειρήσεις και αυτές στη συνέχεια δημιούργησαν μία βάση, η οποία περιείχε έναν αυξημένο αριθμό δεδομένων προερχόμενο από πολλές διαφορετικές βάσεις.

Ο κύριος λόγος για την δημιουργία όλο και περισσότερων αλγορίθμων ήταν η ανάγκη των επιχειρήσεων να αναπτύξουν και να διατηρήσουν το ανταγωνιστικό τους πλεονέκτημα. Γι’ αυτό και στους αλγορίθμους αυτούς πολλές φορές συνδυάστηκαν τόσο αριθμητικά όσο και μη αριθμητικά στοιχεία. Ωστόσο, ένας πολύ σημαντικός ανασταλτικός παράγοντας στην εξόρυξη δεδομένων από τις επιχειρήσεις είναι η περιορισμένη πληροφόρηση στην οποία έχουν πρόσβαση οι επιχειρήσεις αλλά και η

πολυπλοκότητα των αλγορίθμων που πολλές φορές οδηγεί στην ανάγκη ανάλυσης μόνον από ειδικούς.

Το πέμπτο κεφάλαιο της εργασίας αποτελεί μία συζήτηση σχετικά με τις προκλήσεις που αναμένουμε ότι θα αντιμετωπιστούν στο μέλλον στην επιστήμη της εξόρυξης δεδομένων. Σε αυτό το κεφάλαιο, εστιάζουμε κυρίως στους παράγοντες που θεωρούμε ότι επηρεάζουν αρνητικά αυτόν τον κλάδο με σκοπό να προσπαθήσουμε να προτείνουμε κάποιες πιθανές λύσεις. Έτσι, προτείναμε για παράδειγμα την χρήση των συνδυαστικών αλγορίθμων για την πρόβλεψη μελλοντικών καταστάσεων, αλλά και την επιλογή δεδομένων από περισσότερους αναλυτές προκειμένου να ελαχιστοποιηθούν τα σφάλματα μεροληψίας.

Γενικά, μπορούμε να πούμε ότι η εξόρυξη δεδομένων αποτελεί μία επιστήμη η οποία θα συνεχίσει να απασχολεί την ακαδημαϊκή και επιχειρηματική κοινότητα και κατά τα επόμενα χρόνια και ιδίως όσο αυξάνεται η αβεβαιότητα του οικονομικού περιβάλλοντος.

Βιβλιογραφία

1. Adomavicius, G. and Tuzhilin, A. (2005), "Toward the Next Generation of Recommender Systems: A Survey of the State – of – the – Art and Possible Extensions", *IEEE Transactions on Knowledge and Data Engineering*, **17 (6)**, pp. 734 – 749.
2. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. and Verkamo, I. (1996), Fast Discovery of Association Rules *in Advances in Knowledge Discovery and Data Mining*, pp. 307 – 328, eds. Fayyad, U., Piatesky – Shapiro, P., Smyth, P. and Uthurusamy, R., Menlo Park, California AAAI Press.
3. Ahmed, S. R. (2004), "Application of data mining in retail business", *International Conference on Information Technology: Coding and Computing*, **2 (2)**, pp. 455 – 459.
4. Bagga, S. and Singh, G.N. (2012), "Applications of Data Mining", *International Journal for Science and Emerging Technologies with Latest Trends*, **1 (1)**, pp. 19 – 23.
5. Battiti, R. and Passerinim A. (2010), "Brain – Computer Evolutionary Multiobjective Optimization: A Genetic Algorithm Adapting to Decision Maker", *IEEE Transactions on Evolutionary Computation*, **14 (5)**, pp. 671 – 687.
6. Berry, N. J. A. and Linoff, G. S. (1999), *Mastering data mining: The art and science of customer relationship management*, Wiley Editions.
7. Brachman, R. and Anand, T. (1996), The process of Knowledge Discovery in Databases: A Human – Centered Approach, *in Advances in Knowledge Discovery and Data Mining*, pp. 37 – 58, eds. Fayyad, U., Piatesky – Shapiro, P., Smyth, P. and Uthurusamy, R., Menlo Park, California AAAI Press.
8. Braha, D. and Shmilovici, A. (2003), "Data mining for improving a cleaning process in the semiconductor industry", *IEEE Transactions on Semiconductor Manufacturing,* **15**, pp. 91 – 101.
9. Braha, D., Elovici, Y. and Last, M. (2007), "Theory of Actionable Data Mining with Application to Semiconductor Manufacturing Control", *International Journal of Production Research*, **45 (13)**, pp. 3059 – 3084.

10. Bose, I. and Mahapatra, R.K. (2001), "Business Data Mining – A Machine Learning Perspective", *Information and Management*, **39 (3)**, pp. 221 – 225.
11. Busovsky, B. (2007), Ethics of data mining and aggregation , Working Paper.
12. Chen, M. S., Han, J. and Yu, P. S. (1996), "Data mining: An overview from database perspective", *IEEE Transactions on Knowledge and Data Engineering*, **8 (6)**, pp. 866 – 883.
13. Chen, H., Chiang, R.H.L. and Storey, V.C. (2012), "Business Intelligence and Analytics: From Big Data to Big Impact", *MIS Quarterly*, **36 (4)**, pp. 1165 – 1188.
14. Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X. and Zhu, M. (2002), "Tools for privacy preserving data mining", *SigKDD Explorations*, **4 (2)**, pp. 28 – 34.
15. Cook, J. (2005), Ethics of data mining, RIT Scholar Works.
16. Elovici, Y. and Braha, D. (2003), "A decision – theoretic approach to data mining", *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, **33 (1)**, pp. 42 – 51.
17. Estivill – Castro, V., Brankovic, L. and Dowe, D. (1999), "Privacy in data mining", *Privacy – Law and Policy Reporter*, **6 (3)**, pp. 33 – 35.
18. Fayyad, U., Piatetsky – Shapiro, G. and Smyth, P. (1996), "The KDD process for extracting useful knowledge from volumes of data", *Communications of the ACM*, **39 (11)**, pp. 27 – 34.
19. Felders, A., Daniels, H. and Holsheimer, M. (2000), "Methodological and practical aspects of data mining", *Information & Management*, **37**, pp. 271 – 281.
20. Fountain, T., Dietterich, T.G. and Sudyka, B. (2000), Mining IC Test Data to Optimize VLSI Testing, *in Proceedings of the Sixth ACM International Conference on Knowledge Discovery and Data Mining*, pages 18 – 25.
21. Frawley, U., Piatetsky – Shapiro, G. and Matheus, C. J. (1991), *Knowledge discovery in databases*, AAAI/MIT Press.
22. Fule, P. and Roddick, J.F. (2004), Detecting privacy and ethical sensitivity in data mining results, 27th Australasian Computer Science Conference ACSC2004, Dunedin, New Zealand.
23. Ghani, R. and Soares, C. (2006), "Data Mining for Business Applications", *SIGKDD Explorations*, **8 (2)**, pp. 79 – 81.
24. Hand, D.J. (1981), *Discrimination and Classification*, Chichester, UK: Wiley.

25. Hearst, M.A. (1999), Untangling Text Data Mining, Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics.
26. Huang, H.Z., Tian, Z.G. and Zuo, M.J. (2005), "Intelligent Interactive Multiobjective Optimization Method and its Application to Reliability Optimization", *IEEE Transactions on Evolutionary Computation*, **37 (11)**, pp. 983 – 993.
27. IBM (2011), 2011 Annual Report.
28. Islam, M.S. and Abedin, M.Z. (2013), "Impacts of Data Mining on Relational Database Management System Centric Business Environments", *International Journal of Computer Applications* , **75 (3)** , pp. 21 – 27.
29. ISL (2009), Microsoft SQL Server Analysis Service Data Mining.
30. Jain, A.K. and Dubes, R.C. (1988), *Algorithms for Clustering Data*, Englewoodd Cliffs, N.J.: Prentice – Hall.
31. Liao, S.H., Chu, P.H. and Hsiao, P.Y. (2012), "Data mining techniques and applications – A decade review from 2000 to 2011", *Expert Systems with Applications*, **39**, pp. 11303 – 11311.
32. Larose, D.T. (2005), *Discovering knowledge in data: An introduction to data mining*, John Wiley & Sons, Inc.
33. Marbán, Ó., Mariscal, G. and Segovia, J. (2009), "A Data Mining & Knowledge Discovery Process Model".
34. Microsoft Corporation (2007), Microsoft Case Studies.
35. Mitchie, D., Spiegelhalter, D.J. and Taylor, G.C. (1994), *Machine Learning, Neural and Statistical Classification*, Ellis Horwood.
36. Naisbitt, J. (1982), *Megatrends*, Warner Books: New York.
37. Okur, M.C. (2008), "On ethical and legal aspects of data mining", *Journal of Yasar University*, **3 (11)**, pp. 1455 – 1461.
38. Oracle (2008), Telefonica O2 Germany Creates 35 TB Data Warehouse to Manage Data and Improve Operations.
39. Oracle (2010), Oracle Data Mining 11g Release 2.
40. Pathak, M., Singh, S. and Oberoi, S.S. (2013), "Impact of Data Warehousing and Data Mining in Decision Making", *International Journal of Computer Science and Information Technologies*, **4 (6)**, pp. 995 – 999.
41. Quinlan, J.R. (1996), *C4.4: Programs for and Neural Networks*, Cambridge University Press.

42. Robert, H. and Iles, E. (2014), Ethics of data mining: a New Zealand survey, Eastern Institute of Technology Paper.
43. Satlkar, B. (2009), Advantages of Database Management Systems.
44. Seltzer, W. (2005), The Promise and Pitfalls of Data Mining: Ethical Issues, Proceedings of the American Statistical Association, Section on Risk Analysis, Alexandria.
45. Shalak, D.B. (1995), Prototype Selection for Composite Nearest Neighbor Classifiers, Technical Report No. 95 – 74, University of Massachusetts.
46. Shaw, M. J., Subramaniam, C., Tan, G. W. and Welge, M. E. (2001), "Knowledge management and data mining for marketing", *Decision Support Systems*, **31 (1)**, pp. 127 -137.
47. Titterington, D.M., Smith, A.F.M. and Makov, U.E. (1985), *Statistical Analysis of finite – mixture distributions*, Chichester, UK: Wiley.
48. Tobin, K.W., Karnowski, T.P. and Lakhani, F. (2000), A Survey of Semiconductor Data Management Systems in Technology, in *Proceedings of SPIES's 25th Annual International Symposium on Microlithography*, Santa Clara, CA, USA, February 2000.
49. Wahlstrom, K., Roddick, J.F., Sarre, R., Estivill – Castro, V. and deVries, D. (2006), On the ethical and legal implications of data mining, Technical Report SIE – 06 – 001.
50. Weiss, S.M. and Kulikowski, C.A. (1991), *Computer systems that learn: Classification and prediction methods from statistics, neural nets, machine learning, and expert systems*, Morgan Kaufmann: San Mateo.
51. Wu, X., Yu, P. S., Piatetsky – Shapiro, G., Cercone, N., Lin, T. Y., Kotagiri, R. and Wah, B. W. (2003), "Data mining: How research meets practical development?", *Knowledge and Information Systems*, **5 (2)**, pp.248 – 261.
52. Zarsky, T.Z. (2003), "Mine your own Business!": Making the Case for the Implications of the Data Mining of Personal Information in the Forum of Public Opinion, Yale J.L. & Technology Paper No.8.
53. Zhang, D. and Zhou, L. (2004), "Discovering golden nuggets: Data mining in financial application", *IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and reviews*, **34 (4)**, pp. 513 – 522.