ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΔΥΤ. ΕΛΛΑΔΑΣ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ & ΟΙΚΟΝΟΜΙΑΣ

ΤΜΗΜΑ ΛΟΓΙΣΤΙΚΗΣ & ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗΣ

Π Τ Υ Χ Ι Α Κ Η Ε Ρ Γ Α Σ Ι Α

Εργαλεία Επιχειρηματικής Νοημοσύνης

Δάλλας Σπυρίδων (16055)

ΜΕΣΟΛΟΓΓΙ 2016

ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΔΥΤ. ΕΛΛΑΔΑΣ

Περιεχόμενα

1 Εισαγωγή

1.1 Ανάγκη ανάλυσης μεγάλου όγκου δεδομένων

Η έννοια των μεγάλων δεδομένων υπάρχει και απασχολεί ερευνητές και επαγγελματίες εδώ και δεκαετίες. Οι περισσότεροι οργανισμοί κατανοούν πλέον ότι αν κατορθώσουν να συλλέξουν όλα τα δεδομένα που ρέουν στις δραστηριότητές τους, μπορούν να εφαρμόσουν αναλυτικές μεθόδους και να αντλήσουν σημαντική αξία από τη διαδικασία. Αλλά ακόμη και από τη δεκαετία του 1950, δεκαετίες πριν οποιοσδήποτε αναφέρει τον όρο "μεγάλα δεδομένα," οι επιχειρήσεις χρησιμοποιούσαν τις βασικές αναλυτικές μεθόδους (ουσιαστικά αριθμούς σε υπολογιστικά φύλλα που αποδελτιώθηκαν και υπολογίστικαν με το χέρι) για να αποκαλύψουν ιδέες και τάσεις.

Τα νέα οφέλη που η ανάλυση μεγάλων δεδομένων φέρνει στο τραπέζι, ωστόσο, είναι η ταχύτητα και η αποτελεσματικότητα. Ενώ πριν από λίγα χρόνια μια επιχείρηση θα συγκέντρωνε πληροφορίες, θα έτρέχε την ανάλυση που θα έφερνε στο φως πληροφορίες που θα μπορούσαν να χρησιμοποιηθούν για τις μελλοντικές αποφάσεις, σήμερα η επιχείρηση μπορεί να εντοπίσει τις ιδέες για την άμεση λήψη αποφάσεων. Η ικανότητα να λειτουργούν ταχύτερα - και να παραμείνουν ευέλικτες - δίνει στους οργανισμούς ένα ανταγωνιστικό πλεονέκτημα που δεν είχαν στο παρελθόν.

Η ανάλυση μεγάλων δεδομένων βοηθά τους οργανισμούς να αξιοποιήσουν τα δεδομένα τους και να τα χρησιμοποιούν για να εντοπίσουν νέες ευκαιρίες. Αυτό, με τη σειρά του, οδηγεί σε έξυπνες επιχειρηματικές κινήσεις, πιο αποτελεσματική λειτουργία, υψηλότερα κέρδη και πιο ευτυχισμένοι πελάτες. Στην έκθεσή του Big Data in Big Companies,ο Διευθυντής Ερευνών της IIA Tom Davenport πέρασε από τη διαδικασία της συνέντευξης περισσότερες από 50 επιχειρήσεις ώστε να κατανοήσει πώς χρησιμοποιούν τα μεγάλα δεδομένα. Κατέληξε ότι οι περισσότερες προσπάθθησαν να αντλήσουν αξία με τους ακόλουθους τρόπους:

Α) **Μείωση κόστους**.Οι τεχνολογίες διαχείρισης μεγάλων δεδομέων, όπως η Hadoop και αναλυτικά εργαλεία που λειτουργούν στο cloud, παρουσιάζουν σημαντικά πλεονεκτήματα κόστους, όταν πρόκειται για την αποθήκευση μεγάλου όγκου δεδομένων - συν το ότι μπορούν να εντοπιστούν πιο αποτελεσματικοί τρόποι για την επιχειρηματική δραστηριότητα.

Β) **Ταχύτερη και καλύτερη λήψη αποφάσεων**. Με την ταχύτητα του Hadoop και των in-memory analytics, και σε συνδυασμό με την ικανότητα να αναλύουν νέες πηγές δεδομένων, οι επιχειρήσεις είναι σε θέση να αναλύσει τις πληροφορίες άμεσα - και να λαμβάνουν αποφάσεις με βάση τις πληροφορίες.

Γ) **Νέα προϊόντα και υπηρεσίες**. Με τη δυνατότητα να εκτιμήσει τις ανάγκες και την ικανοποίηση των πελατών μέσω μέσω της ανάλυσης δεδομένων, δημιουργείται και η δύναμη να δώσει στους πελάτες αυτό ακριβώς που θέλουν. Επισημαίνεται ότι με την ανάλυση μεγάλων δεδομένων, όλο και περισσότερες εταιρείες προσπαθούν να δημιουργήσουν νέα προϊόντα που θα είναι ικανά να καλύψουν τις ανάγκες των πελατών τους.[1]

1.2 Η ταυτότητα των χρηστών

Σκεφτείτε μια επιχείρηση που στηρίζεται στην ανάληψη γρήγορων και ευέλικτων αποφάσεων ώστε να παραμείνουν ανταγωνιστικές, και το πιθανότερο να στηρίζονται στα μεγάλα δεδομένα για την λειτουργία της επιχείρησης. Παρακάτω ακολουθούν παραδείγματα επιχειρήσεων σε διαφορερικούς τομείς της οικονομικής δραστηριότητας και το πώς θα μπορούσαν να χρησιμοποιήσουν την τεχνολογία:

1.2.1 Τουρισμός και Φιλοξενία

Διατήρηση των πελατών ευχαριστημένοι είναι το κλειδί για την ταξιδιωτική βιομηχανία και το ξενοδοχείο, αλλά η ικανοποίηση του πελάτη μπορεί να είναι δύσκολο να μετρηθεί - ειδικά σε εύθετο χρόνο. Θέρετρα και τα καζίνο, για

[1] Θεοδωρίδης Γ., Πελέκης Ν. (2011): Εξόρυξη Γνώσης από Δεδομένα - Συσταδοποίηση, Ομάδα Διαχείρισης Δεδομένων Πανεπιστήμιο Πειραιώς

παράδειγμα, έχουν μόνο ένα μικρό παράθυρο ευκαιρίας για να γυρίσει μια εμπειρία των πελατών που πρόκειται νότια γρήγορα. Big analytics δεδομένων δίνει σε αυτές τις επιχειρήσεις τη δυνατότητα να συλλέξει τα δεδομένα των πελατών, ισχύουν analytics και αμέσως εντοπίσει πιθανά προβλήματα πριν να είναι πολύ αργά.

1.2.2 Υπηρεσίες υγείας

Τα μεγάλα δεδομένα είναι πολύ κοινά στον χώρο της υγείας και της υγειονομικής περίθαλψης. Ιατρικοί φακέλοι ασθενών, προγράμματα υγείας, πληροφορίες ασφάλισης και άλλα είδη πληροφοριών που μπορεί να είναι δύσκολα διαχειρίσιμα- αλλά είναι γεμάτα από ιδέες όταν εφαρμόζεται η ανάλυση. Γι 'αυτό η τεχνολογία ανάλυσης μεγάλων δεδομένων είναι τόσο σημαντικό για τη φροντίδα της υγείας. Με την γρήγορη ανάλυση μεγάλου όγκου πληροφοριών - τόσο δομημένων και αδόμητων – οι παρόχοι υγειονομικής περίθαλψης μπορεί να προσφέρουν σωτήριες διαγνώσεις ή επιλογές θεραπείας σχεδόν αμέσως.

1.2.3 Κυβέρνηση

Ορισμένες κυβερνητικές υπηρεσίες αντιμετωπίζουν την πρόκληση των σφιχτών προϋπολογισμών χωρίς όμως συμβιβασμούς στην ποιότητα ή την παραγωγικότητα. Αυτό είναι ιδιαίτερα προβληματικό στις υπηρεσίες επιβολής του νόμου, οι οποίοι αγωνίζονται να διατηρήσουν χαμηλά ποσοστά εγκληματικότητας με σχετικά λιγοστούς πόρους. Και αυτός είναι ο λόγος που πολλοί οργανισμοί χρησιμοποιούν ανάλυση μεγάλων δεδομένων. Η τεχνολογία βελτιώνει τον επιχειρησιακό τους κλάδο, ενώ δίνει στους οργανισμούς μια πιο ολιστική άποψη της εγκληματικής δραστηριότητας.

1.2.4 Λιανεμπορίο

Η εξυπηρέτηση πελατών έχει εξελιχθεί τα τελευταία χρόνια, καθώς οι περισσότερο υποψιασμένοι αγοραστές αναμένουν οι λιανοπωλητές να καταλάβει ακριβώς αυτό που χρειάζονται, όταν το χρειάζονται. Η ανάλυση μεγάλων δεδομένων βοηθάει τις

εταιρείες λιανικής να πληρούν αυτές τις απαιτήσεις. Οπλισμένοι με ατελείωτες ποσότητες δεδομένων από προγράμματα πιστότητας πελατών, αγοραστικές συνήθειες και άλλες πηγές, οι λιανοπωλητές δεν έχουν μόνο μια σε βάθος κατανόηση των πελατών τους, μπορούν επίσης να προβλέψουν τις τάσεις, να προτείνουν νέα προϊόντα - και να ενισχύσουν την κερδοφορία.[2]

1.3 Πώς λειτουργεί και βασικές τεχνολογίες

Δεν υπάρχει μία ενιαία τεχνολογία που περικλείει την ανάλυση μεγάλων δεδομένων. Στην πραγματικότητα υπάρχουν διάφορα εργαλεία που συνεργαζόμενα βοηθούν να αντλήσεις τις περισσότερες πληροφορίες από τα διαθέσιμα δεδομένα τεχνολογία.

Διαχείριση δεδομένων (data management). Τα δεδομένα πρέπει να είναι υψηλής ποιότητας πριν να γίνει εφικτή η αξιόπιστη ανάλυσή τους. Με τα δεδομένα να διαρρέουν συνεχώς μέσα και έξω από έναν οργανισμό, είναι σημαντική η δημιουργία επαναλαμβανόμενων διαδικασιών για να δημιουργηθούν και να διατηρηθούν πρότυπα ποιότητας των δεδομένων. Όταν τα δεδομένα είναι αξιόπιστα, οι οργανώσεις θα πρέπει να καθιερώσουν ένα κύριο πρόγραμμα διαχείρισης δεδομένων που θα εξισορροπήσει τη λειτουργία στο σύνολο της επιχείρησης.

Εξόρυξη δεδομένων (data mining). Οι τεχνολογίες εξόρυξης δεδομένων βοηθούν στην εξέταση και επεξεργασία μεγάλων ποσοτήτων δεδομένων ώστε να καταστούν εμφανή επαναλαμβανόμενα πρότυπα στα δεδομένα - και αυτή η πληροφορία μπορεί να χρησιμοποιηθεί για περαιτέρω ανάλυση στην προσπάθεια απάντησης πολύπλοκων επιχειρηματικών αποφάσεων. Με το λογισμικό εξόρυξης δεδομένων, είναι δυνατό ένα «κοσκίνισμα» όλων των χαοτικών και επαναλαμβανόμενων θορύβων στα δεδομένα, να εντοπιστεί το τι είναι σχετικό και να χρησιμοποιούν τις

[2] Θεοδωρίδης Γ., Πελέκης Ν. (2011): Εξόρυξη Γνώσης από Δεδομένα - Συσταδοποίηση, Ομάδα Διαχείρισης Δεδομένων Πανεπιστήμιο Πειραιώς

πληροφορίες αυτές για να εκτιμηθούν πιθανά αποτελέσματα, και στη συνέχεια να επιταχυνθεί ο ρυθμός της λήψη τεκμηριωμένων αποφάσεων.

Hadoop. Αυτό το λογισμικό ανοιχτού κώδικα μπορεί να αποθηκεύσει μεγάλες ποσότητες δεδομένων και να εκτελέσετε εφαρμογές σε συστάδες του hardware της μονάδας. Έχει καταστεί βασική τεχνολογία για την επιχειρηματική δραστηριότητα λόγω της συνεχούς αύξησης του όγκου των δεδομένων και ποικιλιών, και το διανεμόμενο υπολογιστικό μοντέλο του επεξεργάζεται μεγάλα δεδομένα γρήγορα. Ένα επιπλέον πλεονέκτημα είναι ότι το πλαίσιο ανοιχτού κώδικα Hadoop είναι δωρεάν και χρησιμοποιεί και αποθηκεύει μεγάλες ποσότητες δεδομένων.
In memory analytics. Με την ανάλυση των δεδομένων από τη μνήμη του συστήματος (αντί σκληρού δίσκου), είναι δυνατή η άμεση άντληση πληροφορίων από τα δεδομένα και η ενέργεια σε αυτές (αποφάσεις, ιδέες). Αυτή η τεχνολογία είναι σε θέση να αφαιρέσει αναλυτικές λανθάνουσες επεξεργασίες για τη δοκιμή νέων σεναρίων και τη δημιουργία μοντέλων. Δεν αποτελεί απλά ένα εύκολο εργαλείο μέσω του οποίου οι οργανισμοί μπορούν να παραμείνουν ευέλικτες και να κάνουν καλύτερες επιχειρηματικές αποφάσεις, αλλά τους δίνει επίσης τη δυνατότητα να τρέξουν επαναληπτικά και διαδραστικά σενάρια.

Προγνωστική Ανάλυση (Predictive analytics). Η τεχνολογία predictive analytics χρησιμοποιεί δεδομένα, στατιστικούς αλγόριθμους και τεχνικές μηχανικής μάθησης για να προσδιορίσει την πιθανότητα μελλοντικών αποτελεσμάτων με βάση τα ιστορικά δεδομένα. Ασχολείται με την παροχή μιας καλύτερης εκτίμησης για το μέλλον, ώστε οι οργανισμοί να αισθάνονται περισσότερο σίγουροι για την καλύτερη δυνατή επιχειρηματική απόφαση. Μερικές από τις πιο κοινές εφαρμογές των προγνωστικών αναλύσεων περιλαμβάνουν στην ανίχνευση απάτης, τον επιχειρησιακό κίνδυνο και το μάρκετινγκ.

Εξόρυξη κειμένου (text mining). Με εργαλεία εξόρυξης κειμένου, είναι δυνατή η ανάλυση δεδομένων κειμένου από το διαδίκτυο, τα πεδία σχολίων,από βιβλία και άλλες πηγές που βασίζονται σε κείμενο, για να αναδείξει ιδέες που δεν είναι εύκολα παρατηρούμενες. Η εξόρυξη κειμένου χρησιμοποιεί μηχανική μάθηση ή

φυσική τεχνολογία επεξεργασίας γλώσσας για να «χτενίσει» έγγραφα - emails, blogs, Twits, έρευνες, κλπ – ώστε να προσφέρει επικουρικό έργο στην ανάλυση μεγάλων δεδομένων και να ανακαλυφθούν θεματικές και συσχετίσεις.

2 Εισαγωγή στην Επιχειρηματική Ευφυΐα

Το παρόν Κεφάλαιο προσφέρει μια εισαγωγή στην Επιχειρηματική Ευφυΐα (ΕΕ). Αρχικά, γίνεται αναφορά στην άνθηση που παρουσιάζει η ΕΕ τα τελευταία χρόνια, και δίνονται ο ορισμός της ΕΕ καθώς και ο ορισμός των Συ στημάτων Επιχειρηματικής Ευφυΐας (ΣΕΕ). Στη συνέχεια, γίνεται αναλυτική παρουσίαση των επιχειρηματικών και τεχνολογικών αίτιων, τα οποία καθόρισαν την ανάγκη ύπαρξης, τη δυνατότητα υλοποίησης και την πρόσφατη ραγδαία ανάπτυξη των ΣΕΕ. Παρουσιάζεται η πυραμίδα των ΣΕΕ. και γίνεται συνοπτική αναφορά στα επίπεδα που την απαρτίζουν, από το βασικό επίπεδο των πηγών δεδομένων, μέχρι το τελικό επίπεδο της λήψης αποφάσεων. Ακολούθως, παρατίθενται τα οφέλη που προσφέρει η ΕΕ αλλά και τα σχετικά προβλήματα, οι κίνδυνοι και οι ανασχετικοί παράγοντες.

Η ΕΕ γνωρίζει σήμερα πολλά πεδία εφαρμογής στη σύγχρονη επιχείρηση. Στο παρόν Κεφάλαιο παρουσιάζονται τα κυριότερα πεδία εφαρμογής, όπως η Διοίκηση Επιχειρησιακής Απόδοσης, η χρηματοοικονομική ανάλυση και διαχείριση, οι πωλήσεις, το μάρκετινγκ, η διαχείριση της εφοδιαστικής αλυσίδας κλπ. Τέλος, γίνεται παρουσίαση των σημαντικότερων παρόχων λογισμικού και υπηρεσιών ΕΕ, καθώς και των βασικών προϊόντων τους..

2.1 Η Επιχειρηματική Ευφυΐα

Αποτελεί κοινό τόπο ότι το επιχειρηματικό περιβάλλον στην αρχή του 21ου αιώνα μπορεί να χαρακτηριστεί πλούσιο, τόσο σε νέες δυνατότητες και ευκαιρίες όσο και σε δυσκολίες που ανέκυψαν από την πρόσφατη οικονομική κρίση. Για την επιτυχή ανταπόκριση των επιχειρήσεων σε αυτές τις νέες προκλήσεις, απαιτείται αναβάθμιση των διοικητικών πρακτικών και βελτίωση των διαδικασιών λήψης αποφάσεων. Προαπαιτούμενο για βελτιωμένες αποφάσεις είναι η βαθιά κατανόηση και γνώση του περιβάλλοντος, αλλά και της ίδιας της επιχείρησης, καθώς και η έγκαιρη και ουσιαστική πληροφόρηση. Έχει ειπωθεί επανειλημμένως ότι η πληρο- φορία είναι ένα από τα πολυτιμότερα κεφάλαια ενός οργανισμού.

Τα παραπάνω μπορούν να αποτελέσουν μια καταρχήν εξήγηση του γεγονότος ότι η Επιχειρηματική Ευφυΐα βρίσκεται τον τελευταίο καιρό στο επίκεντρο του

ενδιαφέροντος του επιχειρηματικού κόσμου. Τα αποτελέσματα της έκθεσης του οίκου Gartner είναι εξόχως αποκαλυπτικά. Ο οίκος Gartner πραγματοποιεί κάθε χρόνο μια έρευνα με στόχο να εντοπίσει τις τεχνολογικές και επιχειρηματικές προτεραιότητες των μεγάλων επιχειρήσεων. Στην έρευνα συμμετέχουν περισσότεροι από 2000 διευθύνοντες σύμβουλοι πολύ μεγάλων επιχειρή- σεων, οι οποίοι αντιπροσωπεύουν δεκάδες επιχειρηματικούς κλάδους καθώς και δεκάδες χώρες. Στη σχετική έκθεση των ετών 2012[3] και 2015[4] η Επιχειρηματική Ευφυΐα βρίσκεται στην κορυφαία θέση του καταλόγου των τεχνολογικών προτεραιοτήτων.

Το ισχυρό ενδιαφέρον του επιχειρηματικού κόσμου για Συστήματα Επιχειρηματικής Ευφυΐας έχει προκαλέσει τη δημιουργία μιας αντίστοιχης, πολύδυναμικής αγοράς. Σύμφωνα με μια ανάλυση αγοράς της IDC, την οποία δημοσιοποιεί στην ιστοθέση της η εταιρεία συστημάτων επιχειρηματικής ευφυΐας SAS, το συνολικό ύψος εσόδων από πωλήσεις λογισμικού αναλυτικής των επιχειρήσεων ανήλθε το έτος 2013 στο ποσό των 37 δισεκατομμυρίων δολαρίων, παρουσιάζοντας αύξηση σε σχέση με το έτος 2012 της τάξης του 8%. Στην ίδια έκθεση αναφέρεται ότι ο ετήσιος ρυθμός αύξησης μέχρι το έτος 2018 αναμένεται να είναι της τάξης του 9%.

Ο όρος Επιχειρηματική Ευφυΐα (Business Intelligence) δεν είναι πρόσφατος. Πρωτοεμφανίζεται το 1865 στο βιβλίο “Cyclopædia of commercial and business anecdotes” του Devens (1865)[5]. Ο Devens χρησιμοποιεί αυτόν τον όρο για να αναφερθεί στον τρόπο με τον οποίο ο τραπεζίτης Sir Henry Furnese αξιοποιούσε πληροφορίες νωρίτερα από τους ανταγωνιστές του, έτσι ώστε να επιτύχει αύξηση των κερδών του. Η επόμενη εμφάνιση του όρου καταγράφεται το 1958 σε τίτλο άρθρου του Luhn (1958) σε περιοδικό της IBM[6].

Στη σύγχρονη βιβλιογραφία ο αναγνώστης θα συναντήσει διαφοροποιημένους ορισμούς της Επιχειρημα- τικής Ευφυΐας. Στο παρόν σύγγραμμα, θα ορίσουμε την

[3] Gartner Executive Programs’ Worldwide Survey of More Than 2,300 CIOs Shows Flat IT Budgets in 2012, but IT Organizations Must Deliver on Multiple Priorities, 2013 («Evtm_219_ CIOtop10[3].pdf,” n.d.)

[4] http://www.gartnerinfo.com/cios9/ CIOLeadershipForum2015Profile. pdf, 2015

[5] Devens, M. (1865). Cyclopædia of commercial and business anecdotes. New York, NY: D. Appleton and Company. Evtm_219_CIOtop10[3].pdf (n.d.). http://www.gartnerinfo.com/sym23/ evtm_219_CIOtop10%5B3%5D.pdf

[6] Luhn, H. P. (1958). A Business Intelligence System. IBM Journal of Research and Development, 2(4), 314- 319

Επιχειρηματική Ευφυΐα ως ένα σύνολο από μεθόδους ανάλυσης, τεχνολογίες, ικανότητες και στρατηγικές, οι οποίες στόχο έχουν την επεξεργασία των διαθέσιμων δεδομένων και την εξαγωγή χρήσιμης πληροφορίας από αυτά, για την υποστήριξη της διαδικασίας λήψης επιχειρηματικών αποφάσεων. Ένας άλλος συγγενής, αν και όχι ταυτόσημος όρος, ο οποίος γνωρίζει ιδιαίτερη διάδοση τον τελευταίο καιρό είναι «Αναλυτική των Επιχειρήσεων» (Business Analytics). Η Επιχειρηματική Ευφυΐα επιτρέπει σε έναν οργανισμό να μαθαίνει, να αντιλαμβάνεται καταστάσεις και συμβάντα, να σκέφτε- ται αφαιρετικά, να προβλέπει τάσεις και μελλοντικά συμβάντα, να σχεδιάζει και να καινοτομεί. Η παραγόμενη πληροφορία μετουσιώνεται σε γνώση που αξιοποιείται από τα διοικητικά στελέχη, ώστε να δρομολογήσουν κατάλληλες δράσεις, που θα οδηγήσουν στον καθορισμό και την επίτευξη επιχειρηματικών στόχων, με τρόπο αποτελεσματικό και αποδοτικό.

Τα συστήματα Επιχειρηματικής Ευφυΐας είναι εξειδικευμένα πληροφοριακά συστήματα, τα οποία προ- σφέρουν ποιοτική πληροφορία. Η πληροφορία βασίζεται σε ποιοτικά και συγκεντρωτικά δεδομένα, τα οποία συνδυάζονται με λογισμικό ικανό να διεξάγει κατάλληλες αναλύσεις. Η βελτίωση της ποιότητας της πληροφο- ρίας οφείλεται στις δυνατότητες αυτών των συστημάτων, τα οποία επιτρέπουν την ταχύτερη πρόσβαση στην πληροφορία, την ευκολότερη υποβολή ερωτημάτων στο σύστημα και τη σύνταξη αναφορών, την προχωρημέ- νη ανάλυση των δεδομένων, καθώς και τη βελτίωση της ποιότητας των δεδομένων. Οι τελικοί αποδέκτες του προϊόντος των συστημάτων ΕΕ, οι οποίοι πολλές φορές αναφέρονται στη βιβλιογραφία ως «εργάτες γνώσης», τροφοδοτούνται έγκαιρα με γνώση που χρησιμοποιούν για τη λήψη αποφάσεων[7].

Πρόδρομοι των σύγχρονων Συστημάτων Επιχειρηματικής Ευφυΐας μπορούν να θεωρηθούν τα Συστήματα Υποστήριξης Αποφάσεων (ΣΥΑ). Τα ΣΥΑ, τα οποία καθιερώθηκαν ως πεδίο συστηματικής έρευνας τη δεκαετία του 1970, στηρίζονται κυρίως στη χρήση μοντέλων. Κάνοντας χρήση των μοντέλων, ο χρήστης μπορεί να πειραματιστεί με διάφορα σενάρια, όπως π.χ. τι θα συμβεί εάν μεταβληθεί κάποια συνθήκη εισόδου (ανάλυση what-if) ή να καθορίσει το επιθυμητό αποτέλεσμα και

[7] Ferrari, A. (2011). Business Intelligence Systems, Uncertainty in Decision-Making and Effectiveness of Organizational Coordination. In A. Carugati & C. Rossignioli (Eds.), Emerging Themes in Information Systems and Organization Studies (pp. 155-167). Berlin: Springer – Verlag

να αναζητήσει τις αναγκαίες συνθήκες εισόδου (αναζήτηση στόχου). Οι Αποθήκες Δεδομένων (Data Warehouse) και οι τεχνικές OLAP (OnLine Analytical Processing) αποτέλεσαν τον επόμενο σταθμό στην ιστορία της Επιχειρηματικής Ευφυΐας. Στις Αποθήκες Δεδομένων συγκεντρώνονται δεδομένα που είναι διάσπαρτα σε διάφορες πηγές. Τα δεδομένα αυτά, αφού υπο- στούν επεξεργασία ώστε να αντιμετωπιστούν διάφορα προβλήματα, αποθηκεύονται σε συγκεντρωτική μορφή (πχ πωλήσεις ανά μήνα ή ανά κατηγορία προϊόντος). Με τις τεχνικές OLAP ο χρήστης μπορεί να προβάλει και να αναλύσει τα δεδομένα σε διάφορα επίπεδα γενίκευσης (π.χ. πωλήσεις ανά μήνα ή ανά τρίμηνο ή ανά έτος). Στη σημερινή εποχή ένας νέος κλάδος της Πληροφορικής, η Εξόρυξη Δεδομένων, έρχεται να δώσει νέα ώθηση στην Επιχειρηματική Ευφυΐα. Η Εξόρυξη Δεδομένων (Data Mining) ή Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων (Knowledge Discovery in Databases) στοχεύει στην ανακάλυψη γνώσης που είναι κρυμμένη σε μεγάλους όγκους δεδομένων. Οι τεχνικές Εξόρυξης Δεδομένων δεν απαιτούν τον προκαθορισμό μοντέλων. Αντιθέτως, τα μοντέλα προκύπτουν από την επεξεργασία των δεδομένων. Επίσης, τα μοντέλα μπορούν να χρησιμοποιηθούν για τη διατύπωση προβλέψεων[8].

2.2 Γιατί Επιχειρηματική Ευφυΐα;

Όπως διατυπώθηκε και παραπάνω, η Επιχειρηματική Ευφυΐα βρίσκεται στο επίκεντρο του ενδιαφέροντος των σύγχρονων μεγάλων επιχειρήσεων. Οι κυριότερες αιτίες γι’ αυτό το γεγονός είναι οι ακόλουθες:

2.2.1 Λήψη Επιχειρηματικών Αποφάσεων σε συνθήκες αβεβαιότητας

Η λήψη αποφάσεων είναι μια από τις σημαντικότερες ευθύνες της διοίκησης μιας επιχείρησης. Ο ισχυρισμός αυτός, αν και έκδηλα προφανής, στοιχειοθετείται με σαφήνεια στις εργασίες επιστημόνων οι οποίοι ασχο- λούνται με τη διοίκηση επιχειρήσεων. Ο Fayol (1949) υποστηρίζει ότι η διοίκηση ενός οργανισμού εκτελεί εργασίες πρόβλεψης και κατάστρωσης σχεδίων, οργάνωσης των δομών και

[8] Sabherwal, R., & Beccera – Fernandez, I. (2010). Business Intelligence. Hoboken, NJ: John Wiley and Sons Inc

διάθεσης υλικών και ανθρωπίνων πόρων, διοίκησης των δραστηριοτήτων και του προσωπικού, συντονισμού, ενοποίησης και εναρμόνισης πρα- κτικών και τέλος, ελέγχου συμφωνίας με καθορισμένες πρακτικές και πολιτικές[9]. Ο Mintzberg ασκεί κριτική στον Fayol και ορίζει ότι η διοίκηση επιτελεί τρεις βασικούς ρόλους: διαπροσωπικούς, πληροφορια- κούς και ρόλους λήψης αποφάσεων[10].

Οι αποφάσεις που λαμβάνονται στα πλαίσια της λειτουργίας ενός οργανισμού ποικίλουν ως προς τον βαθμό αβεβαιότητας. Αποφάσεις που σχετίζονται με ζητήματα καθημερινής λειτουργίας είναι συνήθως σχετικά απλές και τυποποιημένες. Θα μπορούσε να πει κανείς ότι είναι περισσότερο διαδικασίες και λιγότερο αποφά- σεις. Μια απόφαση για αναπαραγγελία νέων εμπορευμάτων, όταν τα αποθέματα ξεπεράσουν το χαμηλότερο επιτρεπτό όριο, είναι μια απλή απόφαση καθημερινής λειτουργίας. Τέτοιες αποφάσεις μπορούν να τυποποι- ηθούν και να ληφθούν ακόμα και αυτόματα, με τη χρήση κατάλληλου λογισμικού. Άλλες αποφάσεις όμως, που αφορούν ευρύτερα τμήματα του οργανισμού ή, ακόμα περισσότερο, που αφορούν ζητήματα στρατηγικού προσανατολισμού είναι πολύ πιο περίπλοκες. Για παράδειγμα, η απόφαση μιας επιχείρησης να παράξει ένα πρωτοποριακό προϊόν, το οποίο δημιουργεί μια νέα κατηγορία προϊόντων, είναι ιδιαίτερα απαιτητική. Θα πρέπει να συνεκτιμηθούν οι καταναλωτικές τάσεις, οι προτιμήσεις και ανάγκες των πελατών, ο προσανατολισμός των τεχνολογικών εξελίξεων, η δυναμική που δημιουργεί το νέο προϊόν στην αγορά, οι πιθανές αντιδράσεις των ανταγωνιστών, οι πιθανές αντιδράσεις συνεργατών, οι οποίοι ενδεχομένως να θιγούν από μια τέτοια κίνηση της εταιρείας, τα χαρακτηριστικά που πρέπει να έχει το νέο προϊόν, το κόστος της επένδυσης και τα αναμενόμενα οικονομικά οφέλη, η τιμή του νέου προϊόντος ώστε η πώληση του να είναι εφικτή, καθώς και πολλά άλλα ζητήματα. Η απόφαση της Apple να λανσάρει το iPod είναι μια χαρακτηριστική τέτοια περίπτωση. Προφανώς, αποφάσεις αυτής της εμβέλειας και αυτού του τύπου είναι ιδιαίτερα περίπλοκες, καθώς υπεισέρχεται μεγάλος βαθμός αβεβαιότητας σε σχέση με πολλά ζητήματα[11].

---

[9] Fayol, H. (1949). General and Industrial Management. London, UK: Pitman.
[10] Mintzberg, H. (1990). Mintzberg on Management: Inside our Strange World of Organizations. New York, NY: Free Press
[11] Information Week. (n.d.). Retrieved 27 December, 2014, from http://www.informationweek.com/software.asp

Εκτός του γεγονότος ότι οι αποφάσεις στρατηγικού προσανατολισμού είναι από τη φύση τους περίπλοκες και απαιτούν τη διαχείριση του ρίσκου ή της αβεβαιότητας, το σύγχρονο επιχειρηματικό περιβάλλον είναι ιδιαίτερα απαιτητικό, με αποτέλεσμα η λήψη αποφάσεων να καθίσταται ακόμα δυσκολότερη. Μερικοί παράγοντες που αυξάνουν τον βαθμό πολυπλοκότητας είναι οι ακόλουθοι:

- Το εξωτερικό περιβάλλον είναι ασταθές και μεταβάλλεται με μεγάλη ταχύτητα.
- Ο ρυθμός λειτουργίας έχει εντατικοποιηθεί, με αποτέλεσμα οι αποφάσεις να λαμβάνονται υπό την πίεση του χρόνου.
- Έχει διαπιστωθεί αύξηση του ανταγωνισμού ποσοτικά αλλά και ποιοτικά.
- Οι επιχειρήσεις γιγαντώνονται και διασπείρονται γεωγραφικά, με αποτέλεσμα να καθίσταται δυσκολότερη η διαχείριση τους.
- Το ανθρώπινο δυναμικό είναι ποιοτικά αναβαθμισμένο και διαθέτει υψηλή εξειδίκευση και αυξημένες δυνατότητες.
- Η απορρύθμιση κανονιστικών διατάξεων επιτρέπει στις επιχειρήσεις μεγαλύτερη ευελιξία κινήσεων, με αποτέλεσμα να αυξάνεται το πλήθος των εναλλακτικών λύσεων.
- Ο ρυθμός παροχής πληροφοριών είναι καταιγιστικός. Η δυνατότητα παροχής πρωτόγνωρα ποιοτικής πληροφόρησης είναι παρούσα.

Τα διοικητικά στελέχη των επιχειρήσεων, κατά τη λήψη αποφάσεων, χρησιμοποιούν τη γνώση τους σχετικά με τον τομέα τους και το αντικείμενο τους, τη διοικητική τους εμπειρία και τα υποκειμενικά στοιχεία του χαρακτήρα τους και τέλος τις διαθέσιμες πληροφορίες. Για τον λόγο αυτό, η παροχή κατάλληλης πληροφόρησης αποτελεί καθοριστικό παράγοντα για τη λήψη επιτυχημένων αποφάσεων. Κατάλληλη πληροφόρηση σημαίνει ότι δίνεται η σωστή πληροφορία στο σωστό άτομο την αναγκαία χρονική στιγμή. Βελτιωμένες αποφάσεις και κατ' επέκταση βελτιωμένο μάνατζμεντ μπορούν να αυξήσουν τις επιδόσεις της επιχείρησης και να της εξασφαλίσουν το ανταγωνιστικό πλεονέκτημα. Τα συστήματα Επιχειρηματικής Ευφυΐας συμβάλλουν σε αυτήν την κατεύθυνση,

προσφέροντας πληροφόρηση και μειώνοντας τον βαθμό αβεβαιότητας κατά τη λήψη αποφάσεων[12].

2.2.2 Οι προκλήσεις της παγκοσμιοποίησης

Στην εποχή της παγκοσμιοποίησης το επιχειρηματικό περιβάλλον άλλαξε και εξακολουθεί να αλλάζει με ταχύτατους ρυθμούς. Η παγκοσμιοποιημένη οικονομία προκάλεσε την ανάπτυξη και ολοκλήρωση παγκόσμιων αγορών. Οι επιχειρήσεις πλέον δραστηριοποιούνται και ανταγωνίζονται σε παγκόσμια κλίμακα. Ο περιορισμός των συνοριακών δασμών και η απορρύθμιση των προστατευτικών μέτρων επιτρέπει σε ξένες επιχειρήσεις να εισέλθουν ευκολότερα σε εγχώριες αγορές. Η άρση των εμποδίων και ο περιορισμός του κόστους εισόδου αυξάνει το πλήθος των ανταγωνιστών. Το τελικό αποτέλεσμα είναι η ένταση του ανταγωνισμού, τόσο ποσοτικά όσο και ποιοτικά.

Οι σημερινές επιχειρήσεις είναι διασκορπισμένες σε πολλές χώρες. Το γεγονός αυτό αυξάνει την πολυπλοκότητα τους και καθιστά δυσκολότερη την παρακολούθηση και τη διοίκηση τους. Επίσης, η επιχειρηματική δραστηριοποίηση σε παγκόσμια κλίμακα περιλαμβάνει και την αντιμετώπιση προβλημάτων, που ανακύπτουν από τις διαφορετικές κουλτούρες. Η πρόσληψη του σημαινομένου μιας διαφημιστικής εκστρατείας μπορεί να είναι τελείως διαφορετική σε ανθρώπους διαφορετικών πολιτισμών. Μια εικόνα, η οποία για έναν καταναλωτή δυτικών κοινωνιών είναι ελκυστική, μπορεί να θεωρηθεί κακόγουστη ή και προσβλητική σε μια κοινωνία του ανατολικού κόσμου. Επίσης, το εργατικό δυναμικό πολυεθνικών επιχειρήσεων, το οποίο έχει διαφορετικές θρησκείες και κουλτούρες, μπορεί να αντιδράσει διαφορετικά σε εργασιακές πολιτικές ενθάρρυνσης και παρακίνησης των εργαζομένων[13].

Τα νέα κανάλια επικοινωνίας και κυρίως το διαδίκτυο επιτρέπουν τη διάχυση της πληροφορίας σε παγκόσμια κλίμακα. Ο καταναλωτής της σημερινής οικονομίας

[12] Ferrari, A. (2011). Business Intelligence Systems, Uncertainty in Decision-Making and Effectiveness of Organizational Coordination. In A. Carugati & C. Rossignioli (Eds.), Emerging Themes in Information Systems and Organization Studies (pp. 155-167). Berlin: Springer – Verlag

[13] Saran, C. (2012). Almost a Third of BI Projects Fail to Deliver on Business Objectives Computer Weekly. http://www.computerweekly.com/news/2240113585/Almost-a-third-of-BI-projectsfail-to-deliver-on-business-objectives

είναι καλύτερα πληροφορημένος, διαθέτει μόρφωση και δεξιότητες χειρισμού νέων τεχνολογιών, έχει υψηλό εισόδημα και για τους λόγους αυτούς έχει και υψηλότερες απαιτήσεις. Η ανταπόκριση στις υψηλές απαιτήσεις των σύγχρονων πελατών αποτελεί νέα πρόκληση για τις επιχειρήσεις.

Μία άλλη σημαντική παράμετρος του σημερινού επιχειρηματικού περιβάλλοντος είναι η ανάδυση των πάλλε ποτέ αναπτυσσόμενων χωρών και η καθιέρωση τους ως πρωταγωνιστικές δυνάμεις, με οικονομικά μεγέθη συγκρίσιμα με αυτά των παραδοσιακά ανεπτυγμένων δυτικών κοινωνιών. Η καταναλωτική άνθηση αυτών των κοινωνιών προσφέρει νέες επιχειρηματικές ευκαιρίες.

Όλοι οι παραπάνω παράγοντες συμβάλλουν στη διαμόρφωση ενός επιχειρηματικού περιβάλλοντος ιδιαίτερα σύνθετου και αβέβαιου. Για την αντιμετώπιση των αυξημένων προκλήσεων της παγκοσμιοποίησης χρειάζεται ιδιαίτερα αποτελεσματική διοίκηση. Η αναβάθμιση των διοικητικών πρακτικών περιλαμβάνει ως βασική συνιστώσα και τη βελτίωση των διαδικασιών λήψης αποφάσεων. Η τροφοδότηση με ποιοτική, δηλαδή ακριβή, σαφή, σχετική με το εξεταζόμενο ζητούμενο και έγκαιρη πληροφορία, επιτρέπει τη λήψη καλύτερων αποφάσεων[14].

2.2.3 Η οικονομική κρίση και οι νέες κανονιστικές διατάξεις

Οι απαρχές του 21ου αιώνα σημαδεύτηκαν από μια δριμύτατη οικονομική κρίση. Η κρίση πρωτοεμφανίστηκε στην αγορά ακινήτων των ΗΠΑ και στη συνέχεια εξελίχθηκε σε τραπεζική κρίση. Το αποτέλεσμα ήταν η χρεοκοπία εκατοντάδων αμερικανικών τραπεζών και η διάσωση άλλων. Πολύ σύντομα, η κρίση πέρασε τον Ατλαντικό ωκεανό και παρουσιάστηκε και στην Ευρωπαϊκή Ένωση, προκαλώντας προβλήματα στον τραπεζικό τομέα αλλά και στην πιστοληπτική ικανότητα κρατών. Ορισμένα κράτη, μεταξύ των οποίων και η Ελλάδα, οδηγήθηκαν σε προγράμματα δανειοδότησης ελεγχόμενα από θεσμούς, όπως το Διεθνές Νομισματικό Ταμείο και η Ευρωπαϊκή Κεντρική Τράπεζα.

[14] SAP. (2015). Business Intelligence Tools | BI & Analytics | SAP. http://go.sap.com/solution/platform-technology/business-intelligence.html

Σε μια προσπάθεια θωράκισης του χρηματοπιστωτικού συστήματος και αντιμετώπισης ατελειών που ανέδειξε η οικονομική κρίση, αρμόδιοι φορείς ενεργοποιήθηκαν για τη θέσπιση ενός νέου κανονιστικού πλαισίου για τη λειτουργία των τραπεζών. Επιδιώκοντας τη μείωση της μόχλευσης, η συνθήκη Βασιλεία ΙΙΙ ορίζει νέους κανόνες που αφορούν στην κεφαλαιακή επάρκεια των τραπεζών, στα τεστ αντοχής και σε κινδύνους σχετικούς με τη ρευστότητα. Σύμφωνα με τα νέες διατάξεις, οι τράπεζες είναι υποχρεωμένες να συντάσσουν και να κοινοποιούν πλήθος αναφορών σχετικά με τα οικονομικά τους στοιχεία. Για την εργασία αυτή απαιτείται η συγκέντρωση, ενοποίηση και επεξεργασία πολλών δεδομένων και η παραγωγή κατάλληλης πληροφορίας. Εξειδικευμένα συστήματα μπορούν να αναλάβουν την αποτελεσματική εκτέλεση αυτών των εργασιών και να διασφαλίσουν την κανονιστική συμμόρφωση (regulatory compliance)[15].

2.2.4 Διαθεσιμότητα δεδομένων

Στη σημερινή εποχή, κάθε επιχείρηση διαθέτει μηχανογραφικό σύστημα, με το οποίο καταγράφει δεδομέ- να για τις συναλλαγές και τις λοιπές δραστηριότητες της. Τα Συστήματα Σχεδιασμού Επιχειρησιακών Πόρων (Enterprise Resources Planning (ERP)), τα οποία αποτελούν τη βασική πλατφόρμα μηχανοργάνωσης των σημερινών επιχειρήσεων, επιτρέπουν την παρακολούθηση των συναλλαγών σε όλες τις λειτουργικές περιοχές της αλυσίδας αξίας ενός οργανισμού, μέσα από ένα ενιαίο περιβάλλον. Άλλα συστήματα παρακολούθησης συναλλαγών, που γνωρίζουν ιδιαίτερη διάδοση, είναι τα Συστήματα Διαχείρισης Εφοδιαστικής Αλυσίδας (Supply Chain Management (SCM)) και τα Συστήματα Διαχείρισης Σχέσεων Πελατών (Customer Relationship Management (CRM)). Όλα αυτά τα συστήματα καταγράφουν καθημερινά, σε σχεσιακές βάσεις, τεράστιους όγκους δεδομένων, που αφορούν τις δραστηριότητες της επιχείρησης. Η παραγωγή και καταγραφή δεδομένων εντείνεται περαιτέρω, με τη χρήση διαφόρων συσκευών όπως barcode readers, συστήματα ετικε- τών RFID, συστήματα GPS, κάμερες κλπ.

[15] Saran, C. (2012). Almost a Third of BI Projects Fail to Deliver on Business Objectives Computer Weekly. http://www.computerweekly.com/news/2240113585/Almost-a-third-of-BI-projectsfail-to-deliver-on-business-objectives

Οι εταιρικές ιστοθέσεις είναι μια άλλη πηγή παραγωγής και καταγραφής δεδομένων. Οι σύγχρονες επιχειρήσεις επιθυμούν να έχουν παρουσία στον παγκόσμιο ιστό. Οι ιστοθέσεις τους, οι οποίες σε ορισμένες περιπτώσεις είναι κανονικές πύλες (portals), χρησιμοποιούνται καθημερινά από διάφορους χρήστες όπως υπαλλήλους της εταιρείας, προμηθευτές, συνεργάτες και πελάτες. Η χρήση της ιστοθέσης από τους επισκέπτες της παράγει δεδομένα. Τα δεδομένα αυτά, σε αντίθεση με τα δεδομένα των συστημάτων παρακολούθησης συναλλαγών τα οποία είναι δομημένα, είναι κατά κανόνα αδόμητα και μπορούν να αφορούν σχόλια πελατών για τα προϊόντα της επιχείρησης ή το ρεύμα κλικ των επισκεπτών της ιστοθέσης[16].

Πέρα από τα δεδομένα που παράγονται από τα μηχανογραφικά συστήματα των επιχειρήσεων, είναι διαθέσιμα και πολλά δεδομένα, τα οποία προέρχονται από εξωτερικές πηγές. Τρίτοι φορείς, όπως κρατικές υπηρεσίες, μέσα ενημέρωσης, τράπεζες και άλλες επιχειρήσεις, μπορεί να προσφέρουν σημαντική πληροφόρηση. Επίσης, μια τεράστια και διαρκώς αυξανόμενη δεξαμενή δεδομένων είναι το Web 2.0. Ιστοθέσεις κοινωνικής δικτύωσης, blogs, wikis και γενικώς ιστοθέσεις το περιεχόμενο των οποίων παράγεται από τους χρήστες του δικτύου, επιτρέπουν την ελεύθερη έκφραση των ανθρώπων και την καταγραφή των απόψεων τους. Κατάλληλη επεξεργασία των στοιχείων αυτών μπορεί να αποκαλύψει καταναλωτικές τάσεις και επιχειρηματικές ευκαιρίες.

Συμπερασματικά, η σύγχρονη επιχείρηση έχει στη διάθεση της τεράστιους όγκους εσωτερικών και εξωτερικών δεδομένων. Τα δεδομένα αυτά μπορεί να είναι διάσπαρτα σε διάφορες πηγές και να περιέχουν ελλιπή ή και αντιφατικά στοιχεία. Ταυτόχρονα όμως, περιέχουν και πληροφορία πολύτιμη για την επιχείρηση. Ένας σύγχρονος όρος, που περιγράφει την υπερσυσσώρευση των δεδομένων και αναφέρεται στις τεχνικές επεξεργασίας τους και στη δυνατότητα εύρεσης πληροφορίας σε αυτά, είναι «Big Data». Τα συστήματα Επιχειρηματικής Ευφυΐας στοχεύουν ακριβώς στη συγχώνευση και επεξεργασία, τόσο των εσωτερικών όσο και των εξωτερικών δεδομένων, και στην ανακάλυψη πολύτιμης πληροφορίας που θα χρησιμοποιηθεί για τη λήψη αποφάσεων.

[16] SAP. (2015). Business Intelligence Tools | BI & Analytics | SAP. http://go.sap.com/solution/platform-technology/business-intelligence.html

2.2.5 Νέες τεχνολογίες και μέθοδοι ανάλυσης

Η ανάλυση των δεδομένων και η εξαγωγή συμπερασμάτων από αυτά γινόταν παλαιότερα αποκλειστικά με χρήση στατιστικών μεθόδων. Αργότερα, η πολυδιάστατη ανάλυση, με χρήση Αποθηκών Δεδομένων και κύβων, εμπλούτισε το φάσμα των διαθέσιμων τεχνικών. Κοινό χαρακτηριστικό και στις δύο παραπάνω περιπτώσεις είναι ότι ο χρήστης διατυπώνει εκ των προτέρων υποθέσεις και στη συνέχεια ελέγχει την ισχύ τους αναλύοντας τα δεδομένα.

Στη σημερινή εποχή ένας νέος κλάδος της Πληροφορικής, η Εξόρυξη Δεδομένων, προσφέρει πρωτόγνωρες δυνατότητες για την επεξεργασία των δεδομένων και την ανακάλυψη της γνώσης. Κατ' αρχήν, η Εξόρυξη Δεδομένων ασχολείται με την επεξεργασία μεγάλου όγκου δεδομένων, δίνοντας απαντήσεις σε σχετικά προβλήματα. Δεύτερον, ακολουθεί μια ολιστική προσέγγιση και παρέχει μεθοδολογίες για όλα τα στάδια της ανακάλυψης γνώσης, από την αρχική συγκέντρωση και προεπεξεργασία των δεδομένων μέχρι και την οπτικοποίηση των προτύπων και τη διατύπωση των τελικών συμπερασμάτων. Αντιμετωπίζονται προβλήματα όπως οι χαμένες τιμές, ο θόρυβος, ο κατάλληλος μετασχηματισμός των δεδομένων κλπ. Τρίτον, οι μέθοδοι της επεξεργασίας των δεδομένων δεν προέρχονται μόνο από τη Στατιστική. Η Εξόρυξη Δεδομένων κάνει ευρύτατη χρήση μεθόδων οι οποίες προέρχονται από την Τεχνητή Νοημοσύνη, τη Μηχανική Μάθηση και την Αναγνώριση Προτύπων. Έρευνες έχουν αποδείξει ότι οι νέες αυτές μέθοδοι μπορούν να δώσουν καλύτερα αποτελέσματα από τις παραδοσιακές στατιστικές μεθόδους. Επίσης, η ανάλυση Κανόνων Συσχέτισης είναι μια νέα μέθοδος επεξεργασίας, η οποία προέρχεται απ' ευθείας από την Εξόρυξη Δεδομένων. Τέταρτον, πολλές από τις παραπάνω μεθόδους δεν απαιτούν την εκ των προτέρων διατύπωση υποθέσεων. Αντιθέτως, τα μοντέλα προκύπτουν απευθείας από τα δεδομένα, με κατάλληλη επεξεργασία. Τέλος, οι νέες μέθοδοι δίνουν τη δυνατότητα προγνωστικής ανάλυσης, δηλαδή την επεξεργασία ιστορικών στοιχείων και τη διατύπωση προβλέψεων για το μέλλον[17].

[17] Scheps, S. (2007). Business Intelligence for Dummies. Hoboken, NJ: Willey Publishing Inc.

Από τα παραπάνω, καθίσταται σαφές ότι ο σύγχρονος αναλυτής έχει πλέον στη διάθεση του βελτιωμένες μεθόδους για να επεξεργαστεί τους τεράστιους όγκους των αποθηκευμένων δεδομένων και να αντλήσει πληροφόρηση, πολύτιμη για τη λήψη αποφάσεων. Συμπερασματικά, η φύση της διαδικασίας λήψης επιχειρηματικών αποφάσεων, κυρίως σε στρατηγικό επίπεδο, η οποία περιλαμβάνει τη διαχείριση της αβεβαιότητας, σε συνδυασμό με τις νέες προκλήσεις της παγκοσμιοποιημένης οικονομίας και της πρόσφατης οικονομικής κρίσης, έθεσαν επιτακτικά την ανάγκη ποιοτικής και έγκαιρης πληροφόρησης. Ταυτόχρονα, η μαζική εφαρμογή της πληροφορικής πρόσφερε τα αναγκαία δεδομένα, ενώ οι νέες μεθοδολογίες ανάλυσης έδωσαν τη δυνατότητα της επεξεργασίας τους και την εξαγωγή της χρήσιμης πληροφορίας. Οι παραπάνω παράγοντες είναι αυτοί που συνέβαλαν στην άνθιση της Επιχειρηματικής Ευφυΐας.

2.3 Δομικά Επίπεδα Συστημάτων Επιχειρηματικής Ευφυΐας

Τα συστήματα Επιχειρηματικής Ευφυΐας είναι δομημένα σε μια σειρά από επάλληλα επίπεδα, τα οποία συ- γκροτούν μια πυραμίδα. Στη βάση της πυραμίδας βρίσκονται τα αρχικά ακατέργαστα δεδομένα, ενώ στην κορυφή της βρίσκεται η λήψη των τελικών αποφάσεων. Κάθε μετάβαση από ένα επίπεδο σε κάποιο ανώτερο, αυξάνει τη δυνατότητα υποστήριξης επιχειρηματικών αποφάσεων.

2.3.1 Πηγές Δεδομένων

Στη βάση της πυραμίδας βρίσκονται οι πηγές των αρχικών δεδομένων. Τα δεδομένα αυτά προέρχονται κυρίως από συστήματα παρακολούθησης συναλλαγών, όπως πχ τα συστήματα ERP, και από εταιρικές βάσεις δεδομένων. Άλλες πρόσθετες πηγές δεδομένων είναι οι εταιρικοί δικτυακοί servers, εσωτερικά έγγραφα ή και εξωτερικές πηγές. Τα δεδομένα αυτά μπορεί να είναι σημαντικά για την καθημερινή λειτουργία της επιχείρησης, είναι όμως ακατάλληλα για τη λήψη αποφάσεων. Η πληροφορία ότι το ταμείο No. 3 ενός υποκαταστήματος super market εξέδωσε απόδειξη για την πώληση ενός κουτιού καφέ, μια συγκεκριμένη ημέρα και ώρα, είναι σημαντική για το λογιστήριο και την αποθήκη, είναι όμως

αδιάφορη για τη διοίκηση. Αυτό που ενδιαφέρει τη διοίκηση είναι οι συγκεντρωτικές πωλήσεις καφέ, σε μια γεωγραφική περιοχή και σε μια χρονική περίοδο. Τα λειτουργικά δεδομένα είναι υπερβολικά αναλυτικά και για τον λόγο αυτό, ακατάλληλα για επεξεργασία και εξαγωγή συμπερασμάτων. Επίσης, τα δεδομένα αυτά είναι διάσπαρτα σε διάφορες πηγές και πρέπει να ενοποιηθούν. Τέλος, τα δεδομένα μπορεί να έχουν διαφόρων ειδών προβλήματα, τα οποία πρέπει να αντιμετωπιστούν[18].

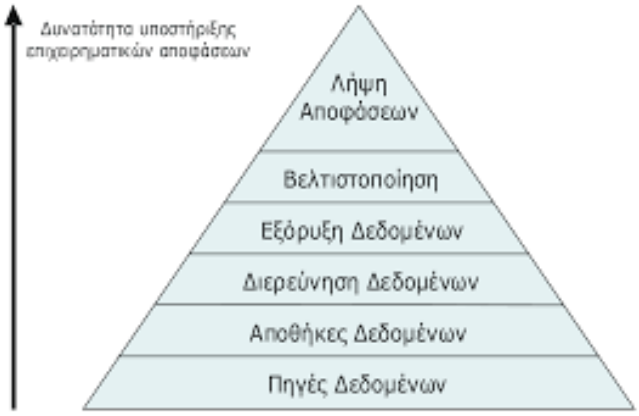
2.3.2 Αποθήκες Δεδομένων

Το επόμενο επίπεδο είναι αυτό των Αποθηκών Δεδομένων. Πρόκειται για βάσεις δεδομένων που περιέχουν τα ενοποιημένα, συγκεντρωτικά και καθαρά δεδομένα. Αυτά τα δεδομένα θα χρησιμοποιηθούν για την ανάλυση και την εξαγωγή συμπερασμάτων. Οι εργασίες εξαγωγής, μετασχηματισμού και φόρτωσης των δεδομένων στις Αποθήκες, γνωστές και ως εργασίες ETL (Extract, Transform, Load), εκτελούνται σε τακτά χρονικά διαστήματα. Στα πλαίσια των εργα- σιών αυτών, επιλέγονται καταρχήν τα λειτουργικά δεδομένα που είναι σχετικά με την ανάλυση που πρέπει να πραγματοποιηθεί. Οι Αποθήκες Δεδομένων είναι θεματικά προσανατολισμένες, επικεντρώνονται δηλαδή σε θεματικές περιοχές, όπως πχ πελάτες ή προμηθευτές. Για τον λόγο αυτό, πρέπει να περιληφθούν τα σχετικά δεδομένα και να αποκλειστούν τα μη σχετικά. Επίσης τα δεδομένα πρέπει να συνολικοποιηθούν σύμφωνα με θέματα που ενδιαφέρουν τη διοίκηση, όπως πχ πωλήσεις ανά περιοχή ή ανά χρονική περίοδο ή ανά κατηγορία προϊόντος, καθώς επίσης και να οριστεί ο βαθμός λεπτομέρειας ή γενίκευσης, όπως πχ πωλήσεις ανά εβδομάδα ή ανά μήνα ή ανά τρίμηνο[19].

[18] Scheps, S. (2007). Business Intelligence for Dummies. Hoboken, NJ: Willey Publishing Inc.

[19] Sabherwal, R., & Beccera – Fernandez, I. (2010). Business Intelligence. Hoboken, NJ: John Wiley and Sons Inc

Εικόνα 1 Η πυραμίδα Συστημάτων Επιχειρηματικής Ευφυΐας



Πηγή: Scheps, S. (2007). Business Intelligence for Dummies. Hoboken, NJ: Willey Publishing Inc, σελ 34

2.3.3 Διερεύνηση Δεδομένων

Το τρίτο επίπεδο περιλαμβάνει εργασίες αρχικής επεξεργασίας των δεδομένων. Στο στάδιο αυτό ο χρήστης υποβάλλει ερωτήματα (queries) στη βάση δεδομένων, λαμβάνει απαντήσεις και συντάσσει αναφορές. Στις αναφορές μπορεί να περιλαμβάνονται αριθμητικές τιμές αλλά και πίνακες και γραφήματα. Τα γραφήματα μπορούν να αποδώσουν με πιο παραστατικό και ευχάριστο τρόπο την πληροφορία. Γενικώς οι μέθοδοι οπτικοποίησης βοηθούν στην καλύτερη παράθεση και κατανόηση των δεδομένων. Στο στάδιο αυτό μπορεί να γίνει και μια αρχική στατιστική επεξεργασία των δεδομένων. Μπορούν για παράδειγμα να υπολογίζονται μέσοι όροι, τυπικές αποκλίσεις κλπ. Χαρακτηριστικό αυτού του επιπέδου είναι ότι ο χρήστης, σύμφωνα με το σκεπτικό του, αναπτύσσει εκ των προτέρων υποθέσεις και στη συνέχεια χρησιμοποιεί τα εργαλεία ανάλυσης για να επιβεβαιώσει ότι οι υποθέσεις του υποστηρίζονται από τα δεδομένα.

2.3.4 Εξόρυξη Δεδομένων

Στο τέταρτο στάδιο εκτελείται υψηλού επιπέδου ανάλυση των δεδομένων, με τη χρήση των πιο εξελιγμένων τεχνικών. Χρησιμοποιούνται προχωρημένες στατιστικές μέθοδοι, αλλά και μέθοδοι που προέρχονται από την Τεχνητή Νοημοσύνη και τη Μηχανική Μάθηση. Οι μέθοδοι κατηγοριοποίησης (classification) επιτρέπουν την

πρόβλεψη της κατηγορίας στην οποία ανήκει ένα αντικείμενο με βάση τα χαρακτηριστικά του. Η πρόβλεψη χρεοκοπίας και η εκτίμηση πιστοληπτικής ικανότητας είναι χαρακτηριστικά παραδείγματα εφαρμογής τεχνι- κών κατηγοριοποίησης. Μέθοδοι ανάλυσης συστάδων (cluster analysis) επιτρέπουν τον εντοπισμό ομάδων ομοειδών αντικειμένων. Ανάλυση συστάδων μπορεί να εφαρμοστεί σε μελέτες τμηματοποίησης της αγοράς, εύρεσης δηλαδή ομάδων πελατών με ομοειδή χαρακτηριστικά. Οι κανόνες συσχέτισης είναι πολύ χρήσιμοι για την ανάλυση του καταναλωτικού καλαθιού (market basket analysis), την εύρεση δηλαδή προϊόντων που πωλούνται συχνά μαζί. Η πληροφορία αυτή μπορεί να είναι χρήσιμη για τη διαμόρφωση των ραφιών σε super market. Ένα χαρακτηριστικό που συναντάται συχνά στις μεθόδους αυτού του επιπέδου είναι ότι ο χρήστης δεν χρειάζεται να διατυπώσει δικές του αρχικές υποθέσεις. Οι αλγόριθμοι επεξεργάζονται τα δεδομένα και εξάγουν την πληροφορία απευθείας από αυτά. Συχνά το αποτέλεσμα είναι ένα μοντέλο. Για παράδειγμα ένα δένδρο απόφασης μπορεί να περιγράφει τα χαρακτηριστικά των αγοραστών μιας κατηγορίας προϊόντων, πχ τετρακίνητων αυτοκινήτων. Ο αλγόριθμος θα διαβάσει τα στοιχεία των πωλήσεων, θα εντοπίσει τα κοινά χαρακτηριστικά των καταναλωτών του συγκεκριμένου προϊόντος και θα κατασκευάσει ένα μοντέλο από κανόνες της μορφής εάν-τότε, οι οποίοι θα περιγράφουν ποιοι αγοράζουν το προϊόν και με ποια πιθανότητα. Ο χρήστης δεν χρειάζεται να διατυπώσει καμία αρχική υπόθεση[20].

2.3.5 Βελτιστοποίηση

Η λήψη αποφάσεων είναι μια διαδικασία επιλογής. Οι αναλύσεις που πραγματοποιήθηκαν στα χαμηλότερα επίπεδα αποφέρουν μια σειρά ενδεχόμενων λύσεων. Ο αποφασίζων καλείται να επιλέξει μια από τις πολλές εναλλακτικές λύσεις. Ως προς το πλήθος των πιθανών λύσεων, τα προβλήματα χωρίζονται σε τρεις κατηγορίες. Τα διχότομα προβλήματα μπορούν να έχουν δύο δυνατές λύσεις, πχ έγκριση του δανείου ή απόρριψη της αίτησης. Τα προβλήματα πολλαπλών λύσεων μπορούν να έχουν έναν περιορισμένο αριθμό ενδεχόμενων λύ- σεων. Η

[20] Scheps, S. (2007). Business Intelligence for Dummies. Hoboken, NJ: Willey Publishing Inc.

επιλογή ενός προμηθευτή μέσα από ένα σύνολο υποψήφιων προμηθευτών είναι τέτοιου είδους πρόβλημα. Τέλος, υπάρχουν προβλήματα απεριόριστου αριθμού ενδεχόμενων λύσεων. Αντικείμενο των εργασιών αυτού του επιπέδου είναι ο εντοπισμός της βέλτιστης λύσης. Υπάρχουν διάφορες μέθοδοι για την επιλογή της βέλτιστης απόφασης. Μεταξύ άλλων, επιλέγουμε να αναφέρουμε τον Γραμμικό Προγραμματισμό και τις ευρετικές μεθόδους (heuristics)[21].

2.3.6 Λήψη απόφασης

Στο κορυφαίο επίπεδο της πυραμίδας γίνεται η λήψη της οριστικής απόφασης. Στο σημείο αυτό, είναι σημαντικό να τονιστεί ότι όλες οι μέθοδοι και τα συστήματα που αναφέρονται παραπάνω, έχουν στόχο την υποβοήθηση ενός ανθρώπου στη λήψη της απόφασης και όχι την αυτοματοποιημένη λήψη απόφασης από έναν υπολογιστή. Πρόκειται ουσιαστικά για εργαλεία ανάλυσης δεδομένων και παραγωγής πληροφοριών. Η τελική απόφαση λαμβάνεται από άνθρωπο, ο οποίος φέρει και την ευθύνη για αυτήν την απόφαση. Ο άνθρωπος, όταν λαμβάνει μια απόφαση, διευκολύνεται στην εργασία του εάν χρησιμοποιήσει περίτεχνα εργαλεία, τα οποία θα του προσφέρουν κατάλληλη πληροφόρηση. Την πληροφόρηση αυτή θα τη χρησιμοποιήσει σε συνδυασμό με τη δική του λογική, τη γνώση και τις ικανότητες του. Πέρα όμως από αυτά, ο άνθρωπος διαθέτει και άλλες ικανότητες και ιδιότητες, τις οποίες μπορεί να επιστρατεύσει. Τέτοιες είναι η φαντασία, το ένστικτο, η διαίσθηση καθώς και πλευρές του χαρακτήρα του[22].

2.4 Οφέλη και Περιορισμοί της Επιχειρηματικής Ευφυΐας

Τα Συστήματα Επιχειρηματικής Ευφυΐας αξιοποιούν τεχνολογίες της Πληροφορικής για να επεξεργαστούν δεδομένα, να παράξουν πληροφορία και να συνδράμουν τη διοίκηση στον έλεγχο και την καλύτερη λειτουργία ενός οργανισμού. Όπως κάθε τεχνολογική λύση, μπορούν να προσφέρουν πολλά οφέλη, ταυτόχρονα όμως υπόκεινται σε περιορισμούς.

---

[21] Sabherwal, R., & Beccera – Fernandez, I. (2010). Business Intelligence. Hoboken, NJ: John Wiley and Sons Inc

[22] Price Waterhouse Coopers. (2007). Guide to Performance Indicators.

2.4.1 Οφέλη της Επιχειρηματικής Ευφυΐας

Τα βασικά οφέλη που προσφέρουν τα συστήματα Επιχειρηματικής Ευφυΐας είναι τα ακόλουθα:

- Καλύτερη κατανόηση πελατών, αγορών, ανταγωνιστών, προμηθειών και πόρων. Η κατάλληλη οργάνωση των δεδομένων και τα εξελιγμένα εργαλεία πληροφορικής δίνουν πρωτόγνωρες δυνατότητες στην εμβάθυνση όλων των παραπάνω ζητημάτων.
- Τροφοδότηση της διοίκησης με τη σωστή πληροφόρηση, την κατάλληλη στιγμή και με τον κατάλληλο τρόπο. Τα συστήματα της Ε.Ε. μπορούν να αναδείξουν την ουσιαστική πληροφορία. Ταυτόχρονο και βασικό μέλημα όμως είναι και η έγκαιρη πληροφόρηση.
- Βελτίωση της ποιότητας των αποφάσεων. Η αναβαθμισμένη και έγκαιρη πληροφόρηση επιτρέπει στη διοίκηση του οργανισμού να λάβει βελτιωμένες αποφάσεις.
- Συμβολή στη διαμόρφωση των στρατηγικών στόχων. Τα συστήματα Ε.Ε. απευθύνονται κυρίως στα υψηλά ή και κορυφαία στελέχη των επιχειρήσεων. Στο επίπεδο αυτό λαμβάνονται οι στρατηγικές αποφάσεις. Η διοίκηση αξιοποιεί τα συστήματα ΕΕ για την άντληση ποιοτικής πληροφόρησης και τον καθορισμό των στρατηγικών στόχων.
- Επίτευξη συγκριτικού πλεονεκτήματος. Η εξασφάλιση συγκριτικού πλεονεκτήματος αποτελεί μόνιμη επιδίωξη κάθε επιχείρησης. Η βελτίωση των αποφάσεων και μέσω αυτού η αύξηση της αποτελεσματικότητας και αποδοτικότητας της διοίκησης, καθώς και ο καθορισμός σωστών στρατηγικών στόχων, μπορούν να αποτελέσουν το συγκριτικό πλεονέκτημα και να οδηγήσουν σε αυξημένη ανταγωνιστικότητα.
- Δυνατότητες αύξησης της κερδοφορίας, μείωσης του κόστους και βελτίωσης της αποδοτικότητας. Η βελτίωση της πληροφόρησης σχετικά με τη διαχείριση της εφοδιαστικής αλυσίδας μπορεί να βοηθήσει στη συμπίεση του κόστους, ενώ η κατανόηση των αγορών μπορεί να αυξήσει τις πωλήσεις και τα κέρδη. Γενικώς,

επιτυχημένα συστήματα ΕΕ συμβάλλουν στην αύξηση των επιδόσεων και της κερδοφορίας.

- Αύξηση της πιθανότητας πρόβλεψης συμβάντων και επιχειρηματικών ευκαιριών. Η βαθύτερη κατανόηση της αγοράς επιτρέπει τον εντοπισμό επιχειρηματικών ευκαιριών. Επιπλέον, οι μέθοδοι προγνωστικής ανάλυσης (predictive analytics) επεξεργάζονται ιστορικά δεδομένα και επιτρέπουν τη διατύπωση προβλέψεων.
- Μεγαλύτερη αξιοποίηση των δεδομένων και αύξηση της απόδοσης της επένδυσης σε τεχνολογίες πληροφορικής. Οι σημερινές επιχειρήσεις έχουν επενδύσει εκατομμύρια ευρώ σε πληροφοριακά συστήματα. Τα δεδομένα αυτών των συστημάτων μπορούν να αποδειχθούν πολύτιμη πηγή πρόσθετης, μη συμβατικής πληροφόρησης, εάν αξιοποιηθούν με τη χρήση της Επιχειρηματικής Ευφυΐας. Με τον τρόπο αυτό, οι επενδύσεις πληροφορικής αποδίδουν πρόσθετους καρπούς[23].

2.4.2 Περιορισμοί της Επιχειρηματικής Ευφυΐας

Η ανάπτυξη συστημάτων Επιχειρηματικής Ευφυΐας έχει να αντιμετωπίσει διάφορους ανασχετικούς παράγο- ντες, προβλήματα και ενδεχόμενους κινδύνους:

- Κόστος απόκτησης και λειτουργίας Αποθηκών Δεδομένων και συστημάτων ΕΕ. Απαιτούνται επενδύσεις σε υλικό, λογισμικό και τεχνογνωσία. Επίσης οι εργασίες ETL είναι χρονοβόρες, δύσκολες και δαπανηρές. Όλα τα παραπάνω επιφέρουν ένα όχι ευκαταφρόνητο κόστος, το οποίο πρέπει να αναλάβει η επιχείρηση.
- Χαμηλή ποιότητα δεδομένων. Το πρόβλημα αυτό είναι ένα από τα σημαντικότερα στην ανάπτυξη συστημάτων ΕΕ. Τα αρχικά δεδομένα είναι διάσπαρτα, ανομοιογενή, ελλιπή και πιθανώς λανθασμένα ή αντιφατικά. Τροφοδότηση του συστήματος με προβληματικά δεδομένα θα οδηγήσει σε εσφαλμένη πληροφόρηση. Όπως χαρακτηριστικά λέγεται «garbage in, garbage out».

[23] Price Waterhouse Coopers. (2007). Guide to Performance Indicators.

● Ζητήματα συμβατότητας με τα υπάρχοντα συστήματα. Τα συστήματα ΕΕ λειτουργούν επί δεδομένων άλλων συστημάτων. Τα συστήματα αυτά μπορεί να είναι πολλά, διαφορετικά, και πιθανότατα δεν έχει ληφθεί εκ των προτέρων καμία πρόνοια για ενοποίηση των δεδομένων τους. Μπορεί να εμφανιστούν προβλήματα συμβατότητας, τόσο μεταξύ των βασικών συστημάτων όσο και μεταξύ αυτών και του συστήματος ΕΕ.

● Πιθανή ύπαρξη επιφυλάξεων, δυσπιστίας και μη συνεργασίας από την πλευρά των στελεχών. Η ανάπτυξη συστημάτων Ε.Ε. επιφέρει αλλαγές σε λειτουργίες των οργανισμών. Έχει παρατηρηθεί ότι τέτοιες αλλαγές μπορεί να προκαλέσουν τις επιφυλάξεις και τη δυσπιστία των εμπλεκόμενων στελεχών. Είναι πολύ σημαντικό, τα ανώτατα στελέχη της διοίκησης να εφαρμόσουν πολιτικές διαχείρισης της αλλαγής (change management) και να επιληφθούν τέτοιων προβλημάτων.

● Προβλήματα επικοινωνίας και συνεννόησης μεταξύ των στελεχών και των ειδικών πληροφορικής. Τα στελέχη της επιχείρησης και οι ειδικοί της πληροφορικής έχουν ο καθένας τη δική του οπτική γωνία. Τα στελέχη επικεντρώνονται στα επιχειρησιακά ζητήματα, ενώ οι ειδικοί πληροφορικής στα τεχνικά. Αυτό μπορεί να προκαλέσει προβλήματα συνεννόησης. Ειδικά στα συστήματα ΕΕ, όπου τα επιχειρησιακά ζητήματα παίζουν βαρύνοντα ρόλο, το πρόβλημα αυτό μπορεί να ενταθεί.

● Ανάγκη ειδικά εκπαιδευμένου προσωπικού. Πρέπει να προσληφθεί νέο προσωπικό, αλλά κυρίως πρέπει τα στελέχη να μάθουν να χρησιμοποιούν, με τον βέλτιστο τρόπο, τα νέα αυτά συστήματα.

● Κίνδυνος υπερβολικής και άκριτης εμπιστοσύνης στο σύστημα ΕΕ και συνακόλουθης επανάπαυσης. Έχει ήδη τονιστεί ότι ο τελικός υπεύθυνος για τη λήψη των αποφάσεων είναι ο άνθρωπος. Συστήματα ευφυούς ανάλυσης των δεδομένων και κυρίως συστήματα ικανά να διατυπώνουν προβλέψεις, μπορεί μετά από κάποιον χρόνο να εμπνεύσουν υπερβολική εμπιστοσύνη στους χρήστες τους. Τα στελέχη δεν πρέπει να επαναπαύονται στις προβλέψεις του συστήματος, και πρέπει να αντιμετωπίζουν την πληροφόρηση στη βάση της δικής τους υποκειμενικής κρίσης.

● Πολλές περιπτώσεις αποτυχίας σε έργα ΕΕ. Τα έργα Επιχειρηματικής Ευφυΐας έχουν να αντιμετωπίσουν πολλές προκλήσεις. Ως αποτέλεσμα αυτού του γεγονότος

καταγράφεται μεγάλο ποσοστό αποτυχίας έργων επιχειρηματικής ευφυΐας. Σύμφωνα με τον Saran (2012), ο οποίος επικαλείται πηγές του οίκου Gartner, λιγότερο από το 30% των έργων ΕΕ επιτυγχάνει τους σκοπούς του[24].

2.5 Η Επιχειρηματική Ευφυΐα στην Πράξη

Δεδομένου ότι κάθε δραστηριότητα μιας επιχείρησης απαιτεί τη λήψη αποφάσεων, η Επιχειρηματική Ευφυΐα μπορεί να βρει αντίστοιχες δυνατότητες εφαρμογής. Υπό την έννοια αυτή τα πεδία εφαρμογής της Επιχειρη- ματικής Ευφυΐας στη σύγχρονη επιχείρηση μπορεί να είναι εξαιρετικά ποικίλα και θεωρητικά απεριόριστα. Στο σημείο αυτό θα επιχειρήσουμε μια χοντρική κατηγοριοποίηση και παρουσίαση των συνηθέστερων πεδίων εφαρμογής.

2.5.1 Διοίκηση Επιχειρησιακής Απόδοσης Σύμφωνα με το γλωσσάριο του οίκου Gartner,

η Διοίκηση Επιχειρησιακής Απόδοσης (ΔΕΑ) (Corporate Performance Management (CPM)) ("CPM (corporate performance management) - Gartner IT Glossary," n.d.) είναι ένα σύνολο μεθοδολογιών, μετρικών, διαδικασιών και συστημάτων, τα οποία επιτρέπουν στα διευθυντικά στελέχη ενός οργανισμού να ελέγχουν και να διαχειρίζονται την απόδοση του.

Στη σύγχρονη εποχή, η ΔΕΑ υλοποιείται με τη χρήση κατάλληλου λογισμικού. Εφαρμογές κατάλληλες για ΔΕΑ αντιστοιχούν στρατηγική πληροφορία στα επιχειρησιακά σχέδια και παράγουν συγκεντρωτικά αποτελέσματα. Οι εφαρμογές αυτές ολο- κληρώνονται με τις διαδικασίες σχεδιασμού και ελέγχου του οργανισμού. Απαραίτητο στοιχείο για τη ΔΕΑ είναι οι λεγόμενοι Κύριοι Δείκτες Επιδόσεων (ΚΔΕ) (Key Performance Indicators (KPI)). Οι ΚΔΕ είναι καλά καθορισμένοι δείκτες, οι οποίοι αποτυπώνουν την επίδοση του οργανισμού σε σχέση με κάποια δραστηριότητα του. Οι δραστηριότητες αυτές συνηθέστερα αφορούν την εκπλήρωση κάποιου στρατηγικού στόχου ή σχετίζονται με παράγοντες που είναι ζωτικής σημασίας για τον οργανισμό. Οι επιχειρήσεις χρησιμοποιούν τους

[24] CPM (Corporate Performance Management) – Gartner IT Glossary. (n.d.). http://www.gartner.com/it-glossary/cpm-corporate-performance-management

ΚΔΕ για να ελέγχουν και να μετρούν τον βαθμό επίτευξης στρατηγικών και επιχειρησιακών στόχων.

Ο καθορισμός των κατάλληλων ΚΔΕ δεν είναι μια τετριμμένη εργασία και διαφέρει από επιχείρηση σε επιχείρηση. Οι ΚΔΕ μπορεί να αναφέρονται σε διάφορες δραστηριότητες και λειτουργίες, όπως πχ τις πωλήσεις και τη διαφήμιση, την παραγωγή και τη διοίκηση της εφοδιαστικής αλυσίδας, τα χρηματοοικονομικά και την κερδοφορία, τη διαχείριση ανθρωπίνων πόρων, τη διαχείριση του επιχειρηματικού κινδύνου κλπ. Ένα κρίσιμο ερώτημα είναι το ποιες προϋποθέσεις καθιστούν έναν δείκτη επίδοσης «κύριο». Σε ορισμένες περιπτώσεις η χρήση ΚΔΕ επιβάλλεται από κανονιστικές διατάξεις που διέπουν τη λειτουργία των επιχειρήσεων, όπως ο βρετανικός Companies Act 2006. Εταιρείες συμβούλων και πάροχοι λογισμικού Επιχειρηματικής Ευφυΐας μπορούν να συνδράμουν έναν οργανισμό στη δύσκολη εργασία της επιλογής των κατάλληλων ΚΔΕ. Η Price Waterhouse Coopers (2007) έχει εκδώσει οδηγό για τον καθορισμό ΚΔΕ, στον οποίο παρέχονται οδηγίες για την επιλογή, το περιεχόμενο και τον τρόπο παρουσίασης των ΚΔΕ[25].

Οι τιμές των ΚΔΕ αντιπαραβάλλονται με προκαθορισμένους στόχους. Τα διευθυντικά στελέχη ορίζουν αρχικά τις τιμές στόχους και στη συνέχεια συγκρίνουν τις τρέχουσες τιμές των ΚΔΕ με τους στόχους. Εάν διαπιστώσουν ότι υπάρχουν υστερήσεις, θα αναζητήσουν τα αίτια και θα προβούν στις αναγκαίες ενέργειες για τη θεραπεία του προβλήματος. Επίσης, μπορεί να αναθεωρήσουν τις τιμές στόχους. Με τον τρόπο αυτό ελέγχουν αλλά και ρυθμίζουν τις επιδόσεις του οργανισμού. Η Επιχειρηματική Ευφυΐα σχετίζεται άμεσα με τη Διοίκηση Επιχειρησιακής Απόδοσης.

Τα συστήματα Επιχειρηματικής Ευφυΐας οφείλουν να συγκεντρώνουν και να προεπεξεργάζονται όλα τα δεδομένα που σχετίζονται με τους ΚΔΕ, να προβαίνουν στον υπολογισμό των τιμών με ταχύτητα και αποτελεσματικότητα και να παρουσιάζουν τα αποτελέσματα με τρόπο κατανοητό. Η παραγόμενη πληροφορία πρέπει να είναι ορθή, έγκαιρη, ουσιαστική και να αποκαλύπτει την πραγματική κατάσταση του υπό διερεύνηση ζητήματος,

[25] Price Waterhouse Coopers. (2007). Guide to Performance Indicators.

2.5.2 Χρηματοοικονομική ανάλυση και διαχείριση

Αντικείμενο είναι ο σχεδιασμός και η παρακολούθηση των χρηματοοικονομικών ροών. Τα στελέχη παρακολουθούν την πορεία των εσόδων και εξόδων της επιχείρησης. Αναλύονται τα εισπρακτέα, τα πληρωτέα και η κατάσταση των αποθεμάτων. Καθίσταται δυνατή η εύκολη σύνταξη χρηματοοικονομικών καταστάσεων με τρέχοντα στοιχεία, ώστε τα στελέχη να εκτιμούν την επίδοση της επιχείρησης. Επίσης, γίνεται σύγκριση με τα μεγέθη του προϋπολογισμού ώστε, αν διαπιστωθούν αποκλίσεις, να ληφθούν οι αναγκαίες μέριμνες. Η διαδικασία ενημέρωσης σε περίπτωση αποκλίσεων μπορεί να είναι και αυτοματοποιημένη.

Αναλυτικότερα, τα συστήματα Επιχειρηματικής Ευφυΐας, για την ανάλυση των χρηματοοικονομικών μεγεθών, παρακολουθούν τα πάγια της επιχείρησης σε όλο τον κύκλο ζωής τους από την απόκτηση μέχρι την απόσβεση. Επίσης, ελέγχουν την κερδοφορία συνολικά, αλλά και ειδικότερα ανά χρονική περίοδο, περιοχή, πελάτες, κατηγορία προϊόντων κλπ. ώστε να εντοπίζονται με αυτόν τον τρόπο τάσεις, δυναμικές και ευκαιρίες. Η παρακολούθηση των εισπρακτέων και πληρωτέων λογαριασμών επιτρέπει την καλύτερη διαχείριση του κεφαλαίου κίνησης και τον έλεγχο των κινδύνων που αφορούν τις απαιτήσεις. Τα τρέχοντα στοιχεία συγκρίνονται με ιστορικά στοιχεία προηγούμενων ετών και με τιμές στόχους, ώστε να παρέχεται πληρέστερη εικόνα για την πορεία της επιχείρησης και τις χρηματοοικονομικές της επιδόσεις[26].

2.5.3 Πωλήσεις

Τα Συστήματα Επιχειρηματικής Ευφυΐας διευκολύνουν την παρακολούθηση και τον έλεγχο του κρίσιμου τομέα των πωλήσεων, δίνοντας έτσι τη δυνατότητα στις επιχειρήσεις να ανταγωνιστούν αποτελεσματικότερα μέσα στις αγορές. Αναλύονται τα στοιχεία του αγωγού πωλήσεων, από το στάδιο των αρχικών επαφών με τους εν' δυνάμει πελάτες μέχρι την τελική πώληση. Τα στοιχεία αυτά συγκρίνονται με τις τιμές στόχους και εκτιμάται η πορεία των πωλήσεων, ώστε να ληφθούν κατάλληλα μέτρα σε περίπτωση που υπάρχει υστέρηση. Η ανάλυση του αγωγού πωλήσεων

[26] Price Waterhouse Coopers. (2007). Guide to Performance Indicators.

μπορεί να αναδείξει και νέες ευκαιρίες. Επίσης η ανάλυση των ιστορικών και άλλων στοιχείων επιτρέπει την ακριβέστερη πρόβλεψη του ύψους των μελλοντικών πωλήσεων. Ένας άλλος σχετικός τομέας είναι αυτός της διαχείρισης του δυναμικού του τμήματος πωλήσεων. Η ανάλυση των στοιχείων μπορεί να γίνει σε διάφορα επίπεδα που να φθάνουν μέχρι τις ατομικές επιδόσεις των πωλητών. Η διοίκηση εντοπίζει τα ισχυρά σημεία αλλά και τις αδυναμίες και στη συνέχεια αξιοποιεί αυτήν την πληροφόρηση και προβαίνει στις αναγκαίες δράσεις, ώστε να επιτευχθεί η διάχυση των βέλτιστων πρακτικών και η αντιμετώπιση προβλημάτων.

2.5.4 Marketing

Η επεξεργασία των στοιχείων που αφορούν τους πελάτες και η άντληση πολύτιμης σχετικής πληροφορίας είναι από τα σημαντικότερα και αποδοτικότερα πεδία εφαρμογής της Επιχειρηματικής Ευφυΐας. Βασικός στόχος είναι η κατανόηση της αγοραστικής συμπεριφοράς των καταναλωτών και η αναγνώριση των αναγκών και των προτιμήσεων τους. Οι πληροφορίες αυτές επιτρέπουν την προώθηση των πωλήσεων και την αξιοποίηση νέων ευκαιριών. Επιπλέον, με τη χρήση των τεχνικών Επιχειρηματικής Ευφυΐας μπορεί να γίνει πολύ επιτυχημένη ανάλυση τμηματοποίησης της αγοράς, εντοπισμός δηλαδή συνόλων πελατών με ομοειδή χαρακτηριστικά και καταναλωτική συμπεριφορά. Αυτή η πληροφορία αξιοποιείται με τη διοργάνωση στοχευμένων διαφημιστικών εκστρατειών. Η αξιολόγηση των αποτελεσμάτων διαφημιστικών εκστρατειών είναι ένας ακόμα τομέας που διευκολύνεται με τη χρήση της Επιχειρηματικής Ευφυΐας. Επιλεγμένες διαφημιστικές δράσεις αποτιμώνται σε σχέση με το κόστος τους και τα οφέλη που απέφεραν, και γίνεται σύγκριση των πραγματικών αποτελεσμάτων με τα προϋπολογισμένα μεγέθη. Με τον τρόπο αυτό επιτυγχάνεται βελτιστοποίηση των διαφημιστικών πρακτικών[27].

[27] Scheps, S. (2007). Business Intelligence for Dummies. Hoboken, NJ: Willey Publishing Inc.

2.5.5 Διαχείριση Εφοδιαστικής Αλυσίδας.

Αντικείμενο είναι η καλύτερη διαχείριση της Εφοδιαστικής Αλυσίδας με την παραγωγή και διάχυση των κατάλληλων πληροφοριών. Γίνεται αποτελεσματικός έλεγχος των επιπέδων των αποθεμάτων, σε συνδυασμό με τις ανάγκες σε υλικά απαραίτητα για την παραγωγή προϊόντων. Εντοπίζονται έγκαιρα και αντιμετωπίζονται ελλείψεις και καθυστερήσεις σε παραγγελίες, ώστε να μην επιβραδύνεται η παραγωγή. Με τον τρόπο αυτό γίνεται καλύτερος έλεγχος της ροής των προϊόντων, αυξάνεται η ικανοποίηση του πελάτη με την έγκαιρη πα- ράδοση και μειώνονται οι ακυρώσεις και οι επιστροφές. Η Επιχειρηματική Ευφυΐα βρίσκει εφαρμογή επίσης στην επιλογή προμηθευτών. Αναλύονται τα ιστορικά στοιχεία των προμηθευτών σχετικά με την ποιότητα των προϊόντων και υπηρεσιών, τους χρόνους παράδοσης, τη συνέπεια, τις τιμολογιακές πολιτικές και τις εκ- πτώσεις και προσφορές τους κλπ. Επίσης, μπορεί να αξιοποιηθούν και εξωτερικά στοιχεία σχετικά με τους υποψήφιους προμηθευτές που να αφορούν την επιχειρηματική δυναμική τους, τη χρηματοοικονομική τους κατάσταση κλπ.[28]

2.5.6 Διαχείριση Ανθρωπίνων Πόρων

Ζητήματα στελέχωσης της επιχείρησης με ανθρώπινο δυναμικό, αμοιβών και παραγωγικότητας περιλαμβάνονται στα τυπικά αντικείμενα που καλύπτονται από τα συστήματα Επιχειρηματικής Ευφυΐας. Η διοίκηση μπορεί ευκολότερα να διαχειριστεί θέματα μισθοδοσίας όπως αμοιβές, φόρους, ασφαλιστικές εισφορές, υπερωρίες κλπ. Επίσης, επιτυγχάνεται καλύτερος έλεγχος της παραγωγικότητας με υπολογισμό του παραγωγικού και μη παραγωγικού χρόνου, χρόνους προσέλευσης και αποχώρησης, εντοπισμός των πλέον παραγωγικών εργαζομένων και των ταλέντων, καθώς και ο σχεδιασμός πολιτικών για τη συγκράτηση και εξέλιξη των ταλαντούχων εργαζομένων.

Καθίσταται ευκολότερος ο σχεδιασμός και η σύγκριση διαφορετικών πλάνων, για την κάλυψη των αναγκών σε εργατικό δυναμικό με εναλλακτικούς τρόπους, όπως πρόσληψη μόνιμου ή εποχιακού προσωπικού, πλήρους ή μερικής απασχόλησης,

[28] Scheps, S. (2007). Business Intelligence for Dummies. Hoboken, NJ: Willey Publishing Inc.

υπερωρίες, εσωτερική κινητικότητα κλπ. Τα προγράμματα διαχείρισης ανθρωπίνων πόρων μπορούν να ποσοτικοποιηθούν και να συγκριθούν ως προς τις οικονομικές και λειτουργικές επιπτώσεις τους. Επιτυγχάνεται η πρόβλεψη των αναγκών σε εργατικό δυναμικό με ανάλυση στοιχείων για συνταξιοδοτήσεις, αποχωρήσεις, επαναπροσλήψεις, απολύσεις κλπ.[29]

2.5.7 Χρηματοπιστωτικός τομέας

Ο τομέας των χρηματοοικονομικών υπηρεσιών, δηλαδή των τραπεζών και των ασφαλειών, βρέθηκε στο επίκεντρο της πρόσφατης οικονομική κρίσης. Προέκυψε λοιπόν η ανάγκη για στενότερη επιτήρηση και έλεγχο των χρηματοπιστωτικών ιδρυμάτων. Οι νέες κανονιστικές διατάξεις που διέπουν τη λειτουργία τους (Βασιλεία ΙΙΙ κλπ.) επιβάλλουν αυστηρούς όρους καθώς και τη δημοσίευση πλήθους αναφορών σχετικά με τα διαθέσιμα κεφάλαια τους, τις συναλλαγές τους, τις εσωτερικές διαδικασίες, τους πελάτες τους κλπ. Στόχος είναι, τόσο η καλύτερη διαχείριση του επιχειρησιακού κινδύνου (operational risk management) όσο και η αντιμετώπιση του οικονομικού εγκλήματος, όπως πχ του «πλυσίματος χρήματος» και της διαφθοράς. Τα πρόστιμα και τα ποσά για αποζημιώσεις πελατών και επενδυτών που μπορεί να προέρθουν από ανεπαρκή διαχείριση του ρίσκου, είναι δυνατόν σήμερα να ανέρχονται στο ύψος δισεκατομμυρίων ευρώ.

Για την εξυπηρέτηση των παραπάνω στόχων και επιδιώξεων χρειάζεται συγκέντρωση επιπλέον δεδομένων, κατάλληλη ενοποίηση τους και ιδιαίτερα αποτελεσματική ανάλυση και αξιοποίηση τους. Τα συστήματα Επιχειρηματικής Ευφυΐας έχουν ακριβώς αυτό το αντικείμενο και είναι τα πλέον κατάλληλα για την ικανοποίηση αυτών των απαιτήσεων. Οι μεθοδολογίες που προσφέρει η Εξόρυξη Δεδομένων είναι ιδιαίτερα ικανές να δίνουν λύσεις σε προβλήματα, όπως η εκτίμηση της πιστοληπτικής ικανότητας των πελατών, η διαχείριση του κινδύνου, η αντιμετώπιση του οικονομικού εγκλήματος και ο εντοπισμός παραποιημένων χρηματοοικο- νομικών καταστάσεων. Επιπλέον, η οργανωμένη και συγκεντρωτική

[29] Scheps, S. (2007). Business Intelligence for Dummies. Hoboken, NJ: Willey Publishing Inc.

διαχείριση των δεδομένων διευκολύνει τη σύνταξη των αναφορών (reports) που απαιτούνται από τη νομοθεσία.

Πρέπει να γίνει κατανοητό ότι η παραπάνω παρουσίαση πεδίων εφαρμογής της Επιχειρηματικής Ευφυΐας είναι ενδεικτική και όχι εξαντλητική. Κατασκευαστές λογισμικού Επιχειρηματικής Ευφυΐας παρέχουν διαφορετικά προϊόντα και προσφέρουν ποικίλες λύσεις για διάφορα πεδία εφαρμογής. Οι ιστοθέσεις κατασκευαστών λογισμικού, όπως η ιστοθέση της Oracle ("Oracle Business Intelligence Applications," n.d.)[30] και η ιστοθέση της SAP (SAP, 2015), περιέχουν αναλυτικές παρουσιάσεις λογισμικών για εξειδικευμένα πεδία εφαρμογής.

2.6 Πάροχοι λογισμικού και υπηρεσιών Επιχειρηματικής Ευφυΐας

Ως συνέπεια της απαίτησης του επιχειρηματικού κόσμου για λύσεις συστημάτων Επιχειρηματικής Ευφυΐας υψηλού επιπέδου, έχει δημιουργηθεί μια αντίστοιχη μεγάλη αγορά με κύκλο εργασιών της τάξης δισεκατομμυρίων ευρώ. Στην αγορά αυτή δραστηριοποιούνται γνωστές και πολύ μεγάλες εταιρείες πληροφορικής, εταιρείες εξειδικευμένες στο λογισμικό στατιστικής ανάλυσης, εταιρείες που πρωτοστατούσαν στον χώρο των βάσεων δεδομένων και κατασκευαστές συστημάτων ERP. Μεταξύ αυτών, εξέχουσα θέση στην προσφορά συστημάτων Επιχειρηματικής Ευφυΐας κατέχουν οι ακόλουθες:

2.6.1 SAS

Η SAS (Statistical Analysis System) είναι μια εταιρεία, που από την ίδρυση της ασχολήθηκε με το λογισμικό στατιστικής ανάλυσης. Σήμερα αποτελεί έναν από τους σημαντικότερους παρόχους συστημάτων Επιχειρηματικής Ευφυΐας. Προσφέρει λογισμικό αναλυτικής των επιχειρήσεων (Business Intelligence and Analytics) με προχωρημένα εργαλεία οπτικοποίησης, εύκολης ανάλυσης, αυξημένες δυνατότητες χρήσης φορητών συσκευών, καθώς και εργαλεία συνεργασίας. Λογισμικό για τη διαχείριση των πελατών και του μάρκετινγκ (Customer Intelligence) αναλύει

[30] Oracle Business Intelligence Applications. (n.d.). http://www.oracle.com/technetwork/middleware/biapplications/overview/index.html

καταναλωτικές συμπεριφορές, διευκολύνει την προσωπική στόχευση και επιτρέπει τον σχεδιασμό και αποτίμηση των διαφημιστικών εκστρατειών. Εξειδικευμένο λογισμικό ασφάλειας και αντιμετώπισης απάτης (Fraud and Security Intelligence) ανιχνεύει εκ των προτέρων δόλιες πληρωμές με χρήση κανόνων, μεθόδων εντοπισμού ανωμαλιών και προγνωστικής ανάλυσης, και διασφαλίζει τη συμμόρφωση με κανονιστικές διατάξεις, ελέγχοντας συναλλαγές για παράνομες δραστηριότητες. Λογισμικό διαχείρισης επιδόσεων (Performance Management) επιτρέπει τον συνδυασμένο έλεγχο επίτευξης βραχυπρόθεσμων και στρατηγικών στόχων, διευκολύνει τον εντοπισμό ευκαιριών και κινδύνων και την κατανόηση των πηγών κόστους και παραγόμενης αξίας. Το λογισμικό για τη διαχείριση του ρίσκου (Risk Management) ασχολείται με θέματα ιδίων κεφαλαίων, με διαχείριση του πιστωτικού κινδύνου και με δοκιμές αντοχής πιστωτικών ιδρυ- μάτων. Τέλος, παρέχεται εξειδικευμένο λογισμικό για τη διαχείριση της εφοδιαστικής αλυσίδας (Supply Chain Intelligence). Οι επιχειρηματικές λύσεις που προσφέρει η SAS αντιμετωπίζουν ζητήματα όπως η διαχείριση των δεδομένων (Data Management), η ανάλυση δεδομένων μεγάλου όγκου (Big Data) και η λειτουργία σε περιβάλλον υπολογιστικού νέφους (SAS Cloud Analytics). Συνολικά η εταιρεία διαθέτει περισσότερα από 200 προϊόντα. Ιδιαίτερη μνεία γίνεται στο SAS Enterprise Miner, λογισμικό εξόρυξης δεδομένων για επιχειρήσεις με αυξημένες δυνατότητες περιγραφικής και προγνωστικής μοντελοποίησης.

2.6.2 IBM

Η IBM, εταιρεία σταθμός στην ιστορία της πληροφορικής, έχει αναπτύξει πολύπλευρη δραστηριότητα στον τομέα του υλικού και του λογισμικού, και έχει εισάγει ριζοσπαστικά καινοτόμα προϊόντα, μεταξύ των οποίων και το περιβόητο IBM Personal Computer, το οποίο αποτέλεσε πρότυπο για τους μελλοντικούς προσωπικούς υπολογιστές (PCs). Η IBM διαθέτει μακροχρόνια εμπειρία στον τομέα της τεχνητής νοημοσύνης και έχει να επιδείξει διάφορα πρωτοποριακά σχετικά προϊόντα όπως ο υπολογιστής Deep Blue, ο οποίος νίκησε τον παγκόσμιο πρωταθλητή σκακιού Kasparov και το σύστημα Watson, το οποίο το 2011 αντιμετώπισε στο τηλεοπτικό κουίζ Jeopardy προηγούμενους νικητές. Επιπλέον,

πρόσφατα, με μια σειρά εξαγορών, η IBM απόκτησε διάφορες εταιρείες το αντικείμενο των οποίων άπτεται των συστημάτων Επιχειρηματικής Ευφυΐας. Τέτοιες περιπτώσεις είναι η εταιρεία συστημάτων Επιχειρηματικής Ευφυΐας και διαχείρισης επίδοσης Cognos, η εταιρεία στατιστικού λογισμικού SPSS, η εταιρεία αποθηκών δεδομένων Netezza, καθώς και πολλές άλλες.

Σήμερα η IBM θεωρείται ένας από τους μεγαλύτερους παρόχους συστημάτων Επιχειρηματικής Ευφυΐας και προσφέρει έναν μακρύ κατάλογο σχετικών προϊόντων και λύσεων. Το λογισμικό IBM SPSS χρησιμοποιείται για διαχείριση δεδομένων, στατιστική ανάλυση, εξόρυξη δεδομένων και κειμένου, βελτιστοποίηση αποφάσεων και συνεργασία. Το IBM Cognos προσφέρει dashboards, scorecards, what-if σενάρια, εργαλεία για σχεδιασμό, προϋπολογισμό και πρόβλεψη, διαχείριση επίδοσης, προχωρημένα εργαλεία οπτικοποίησης, αυτοματοποιημένα εργαλεία για σύνταξη χρηματοοικονομικών αναφορών και πολλά άλλα. Το νέο σύστημα Watson Analytics προσφέρει εξελιγμένη ανάλυση των επιχειρηματικών δεδομένων για τον έλεγχο υποθέσεων και απάντηση ερωτημάτων, καθώς επίσης και βελτιωμένα εργαλεία οπτικοποίησης.

Το λογισμικό OpenPages έχει αντικείμενο τη διαχείριση του ρίσκου, τη συμμόρφωση με τις νέες κανονιστικές διατάξεις, την αυτοματοποίηση των διαδικασιών χρηματοοικονομικών ελέγχων και τη διευκόλυνση των διαδικασιών εσωτερικού ελέγχου. Το λογισμικό IBM Algorithmics απευθύνεται σε χρηματοοικονομικούς οργανισμούς και προσφέρει λύσεις διαχείρισης ρίσκου για πιστώσεις και ρευστότητα, διαχείρισης κεφαλαίου και υποθηκών, διαχείρισης χαρτοφυλακίου και επενδυτικών αποφάσεων. Η IBM συμμετέχει πρωταγωνιστικά στη διαμόρφωση των νέων τάσεων. Αξιοποιώντας τις τεχνολογίες κινητής υπολογιστικής, προσφέρει μέσω κινητών συσκευών πληροφόρηση σε οποιοδήποτε σημείο. Προϊόντα προσφέρονται υπό το σχήμα «Λογισμικό ως υπηρεσία» (Software As A Service) σε περιβάλλον υπολογιστικού νέφους (Cloud computing).

2.6.3 ORACLE[31]

Η Oracle, πασίγνωστη για την ηγετική της παρουσία στον χώρο των βάσεων δεδομένων, δραστηριοποιείται σήμερα και στον χώρο του υλικού υπολογιστών, κυρίως μετά την εξαγορά της Sun Microsystems, αλλά και στον χώρο του λογισμικού επιχειρησιακών συστημάτων, προσφέροντας λύσεις σχεδιασμού επιχειρησιακών πόρων (ERP), διαχείρισης εφοδιαστικής αλυσίδας (SCM) και διαχείρισης σχέσεων πελατών (CRM). Επίσης, θεωρείται ένας από τους κορυφαίους σύγχρονους παρόχους συστημάτων Επιχειρηματική Ευφυΐας και κάτοχος του μεγαλύτερου τμήματος της σχετικής αγοράς. Η πλατφόρμα Enterprise Business Intelligence περιλαμβάνει εξελιγμένα εργαλεία ανάλυσης, δημιουργίας αναφορών, υποβολής ερωτημάτων, dashboards και scorecards, πράξεων OLAP, ειδοποίησης σε πραγματικό χρόνο κλπ. Το λογισμικό Oracle Essbase είναι ένας ισχυρός server πολυδιάστατης ανάλυσης και πράξεων OLAP, που επιτρέπει τη γρήγορη ανάπτυξη σύνθετων επιχειρηματικών μοντέλων και τη διεξαγωγή αναλύσεων what-if. Η πλατφόρμα Oracle Advanced Analytics συνδυάζει τη βάση δεδομένων της Oracle με δύο ισχυρότατα εργαλεία ανάλυσης, το Oracle Data Mining για εξόρυξη δεδομένων και προγνωστικές αναλύσεις, καθώς επίσης και με την ελεύθερη γλώσσα προγραμματισμού R, η οποία χρησιμοποιείται για στατιστικές αναλύσεις και εξόρυξη δεδομένων. Το σύστημα Oracle Exalytics συνίσταται σε μια ολοκληρωμένη λύση, που συνδυάζει υψηλότατης ποιότητας υλικό υπολογιστών (hardware), κορυφαίο λογισμικό Επιχειρηματικής Ευφυΐας και τεχνολογία βάσεων δεδομένων in-memory, συστήματα βάσεων δεδομένων δηλαδή, που λειτουργούν πρωτίστως στην κύρια μνήμη του υπολογιστή, εξασφαλίζοντας πολύ μεγαλύτερη ταχύτητα. Ως προς τις επιχειρηματικές λύσεις που παρέχει η Oracle, αυτές καλύπτουν όλα τα πεδία εφαρμογής που αναφέρονται στο υποκεφάλαιο 'Η Επιχειρηματική Ευφυΐα στην Πράξη', δηλαδή χρηματοοικονομική διοίκηση, πωλήσεις, μάρκετινγκ, διαχείριση εφοδιαστικής αλυσίδας, διαχείριση ανθρωπίνων πόρων, χρηματοπιστωτικός τομέας, καθώς και πολλές επιπλέον, όπως διαχείριση ρίσκου και κανονιστική συμμόρφωση, διαχείριση χαρτοφυλακί- ου, διαχείριση κοινωνικών

[31] Oracle Business Intelligence Applications. (n.d.). http://www.oracle.com/technetwork/middleware/biapplications/overview/index.html

σχέσεων κλπ. Ως κυρίαρχη δύναμη στον χώρο των βάσεων δεδομένων, η Oracle διαθέτει εξαιρετική τεχνογνωσία σε ζητήματα διαχείρισης δεδομένων, τεχνογνωσία την οποία αξιοποιεί και στον νέο χώρο του Big Data. Μια σειρά από εργαλεία και εφαρμογές δίνουν προωθημένες λύσεις σε ζητήματα Big Data. Επίσης, η Oracle τα τελευταία χρόνια έχει εξαγοράσει πολλές εταιρείες που ασχολούνταν με το υπολογιστικό νέφος, εξασφαλίζοντας έτσι σημαντική παρουσία και σε αυτόν τον χώρο.

2.6.4 SAP[32]

Η SAP είναι μια ευρωπαϊκή εταιρεία που κυριαρχεί στον χώρο των συστημάτων Σχεδιασμού Επιχειρησιακών Πόρων (Enterprise Resources Planning), και είναι ένας από τους μεγαλύτερους παραγωγούς λογισμικού παγκοσμίως. Το 2007 η SAP εξαγόρασε την Business Objects, μια γαλλική εταιρεία εξειδικευμένη στα συστήματα Επιχειρηματικής Ευφυΐας, εντείνοντας την παρουσία της σε αυτόν τον χώρο, και σήμερα θεωρείται μια από τις πρωταγωνίστριες δυνάμεις.

Υπό τον τίτλο SAP Business Objects, η εταιρεία προσφέρει μια σειρά από σουίτες εφαρμογών Επιχειρηματικής Ευφυΐας. Το SAP Business Objects BI platform περιλαμβάνει εργαλεία για πρόσβαση σε δεδομένα διαφόρων κατασκευαστών (IBM, Oracle, Teradata κλπ.), εργαλεία για την αποτελεσματική σύνταξη αναφορών με δυνατότητες επεξεργασίας Big Data και ενσωμάτωσης αναφορών σε εφαρμογές, εργαλεία για τη δημιουργία ισχυρών διαδραστικών dashboards, λογισμικό για την αποτελεσματική και γρήγορη απάντηση επιχειρηματικών ερωτήσεων καθώς και λύσεις κινητής υπολογιστικής που διανέμουν πληροφόρηση σε φορητές συσκευές. Η έκδοση Analytics Edition συνδυάζει την ολοκλήρωση και διαχείριση δεδομένων με εξελιγμένο λογισμικό Επιχειρηματικής Ευφυΐας. Κάνοντας χρήση προχωρημένων αναλυτικών μεθόδων επιτρέπει την αναγνώριση τάσεων και εξαιρέσεων, την αξιοποίηση επιχειρηματικών ευκαιριών και την έγκαιρη αντιμετώπιση κινδύνων. Η έκδοση OLAP edition προσφέρει εργαλεία πολυδιάστατης ανάλυσης. Το λογισμικό

[32] SAP. (2015). Business Intelligence Tools | BI & Analytics | SAP. http://go.sap.com/solution/platform-technology/business-intelligence.html CIOLeadershipForum2015Profile.pdf. (2015). http://www.gartnerinfo.com/cios9/CIOLeadershipForum2015Profile.pdf

SAP Crystal Reports έχει αντικείμενο τη δημιουργία καλαίσθητων αναφορών με δυνατότητα επεξεργασίας δεδομένων από διάφορες πηγές, ενώ το SAP Lumira περιλαμβάνει εξελιγμένα εργαλεία οπτικοποίησης. Τα συστήματα Επιχειρηματικής Ευφυΐας της SAP δίνουν δυνατότητες προγνωστικής ανάλυσης και προσφέρουν λύσεις για τη διαχείριση και έλεγχο της επίδοσης της επιχείρησης, καθώς και για τον έλεγχο του ρίσκου και την κανονιστική συμμόρφωση.

2.6.5 Microsoft[33]

Η Microsoft, ο μεγαλύτερος κατασκευαστής λογισμικού παγκοσμίως ως προς τα έσοδα, είναι ευρύτερα γνωστή κυρίως για το λειτουργικό σύστημα Windows και τη σουίτα εφαρμογών αυτοματισμού γραφείου MS Office. Επίσης, η παιχνιδομηχανή Xbox και τα tablets Microsoft Surface είναι πολύ γνωστά προϊόντα hardware. Στον μακρύ κατάλογο προϊόντων λογισμικού της εταιρείας περιλαμβάνονται και εφαρμογές για επιχειρήσεις, όπως συστήματα ERP και λογισμικό Επιχειρηματικής Ευφυΐας. Δύο προϊόντα της, η βάση δεδομένων SQL Server και το Microsoft Office, ειδικότερα η εφαρμογή φύλλων εργασίας Excel και το πρόγραμμα δημιουργίας παρουσιάσεων Power Point, έπαιξαν σημαντικό ρόλο στην καθιέρωση της ως ένας από τους βασικούς παρόχους λογισμικού Επιχειρηματικής Ευφυΐας.

Η βάση δεδομένων SQL Server και ειδικότερα η έκδοση Business Intelligence, προσφέρει ένα περιβάλλον Επιχειρηματικής Ευφυΐας που επιτρέπει την ταχεία και διαδραστική διερεύνηση και οπτικοποίηση των δεδο- μένων, τη συγχώνευση δομημένων και αδόμητων δεδομένων και την ταχεία ανάλυση τους με τη χρήση της εγκατεστημένης στη μνήμη αναλυτικής μηχανής (analytics engine). Ο SQL Server Analysis Services δίνει τη δυνατότητα δημιουργίας πολυδιάστατων μοντέλων, και περιλαμβάνει εργαλεία οπτικοποίησης και σύνταξης αναφορών. Επίσης περιλαμβάνονται εργαλεία εξόρυξης δεδομένων για τη διεξαγωγή προγνωστικών αναλύσεων. Τα εργαλεία αυτά είναι διαθέσιμα ως add-ins του Excel αλλά και μέσω του SQL Server Development Tools για πιο περίτεχνες αναλύσεις. Η πλατφόρμα ανάπτυξης εφαρμογών Microsoft Azure προσφέρει λογισμικό μηχανικής μάθησης

[33] www.microsoft.au
www.msazure.com

για την εξόρυξη δεδομένων και τη διατύπωση προβλέψεων, συνδυασμένο με μια φιλική προς τον χρήστη διεπαφή. Το Azure υποστηρίζει και τη γλώσσα R.

Μεγάλη βαρύτητα δίνει η Microsoft στο υπολογιστικό νέφος και το Big Data. Όλες οι ιστοσελίδες της εταιρείας που αναφέρονται στα συστήματα Επιχειρηματικής Ευφυΐας, τονίζουν με έμφαση τις δυνατότητες αξιοποίησης του νέφους και της λειτουργίας του λογισμικού στα πλαίσια του. Το Microsoft Data Warehouse επιτρέπει τη διαχείριση εξωτερικών δεδομένων μεγάλου όγκου. Τα δομημένα επιχειρηματικά δεδομένα μπο- ρούν εύκολα να συνδυαστούν με αδόμητα δεδομένα από το Hadoop, ώστε να αποτελέσουν μια ολοκληρωμένη βάση πληροφόρησης. Το νέο Office 365, λογισμικό βασισμένο στο νέφος, περιλαμβάνει το Power BI, ένα εύχρηστο περιβάλλον κατάλληλο για εργασίες Επιχειρηματικής Ευφυΐας, προσαρμόσιμες στις μεταβαλλόμενες ανάγκες του χρήστη. Η Microsoft αξιοποιεί και τη βαθιά τεχνογνωσία της στον αυτοματισμό γραφείου. Το Share Point προσφέρει ένα ελκυστικό περιβάλλον για τη δημιουργία και διανομή αναφορών και dashboards. Το Excel, το οποίο στο παρελθόν χρησιμοποιήθηκε κατά κόρον από επιχειρηματικά στελέχη για τη διεξαγωγή αναλύσεων, ενισχύεται με δυνατότητες εξόρυξης δεδομένων. Το ευρύτατα διαδεδομένο Microsoft Office αποτελεί χρήσιμη πλατφόρμα για σύνταξη αναφορών. Ακόμα και τρίτοι κατασκευαστές συστημάτων Επιχειρηματικής Ευφυΐας, όπως η Oracle και η SAP[34], τονίζουν τη δυνατότητα του λογισμικού τους να συνδεθεί με τα προγράμματα του Office και να ενσωματώσει λειτουργικότητες και αποτελέσματα σε φύλλα εργασίας του Excel, σε παρουσιάσεις του Power Point και σε έγγραφα του Word.

2.6.6 Qlik[35]

Η Qlik είναι μια εταιρεία παραγωγής λογισμικού εξειδικευμένη στα συστήματα Επιχειρηματικής Ευφυΐας. Ιδρύθηκε το 1993 στη Σουηδία και γνώρισε ταχύτατη ανάπτυξη. Σήμερα είναι μια διεθνής εταιρεία με δεκάδες χιλιάδες πελάτες σε

[34] SAP. (2015). Business Intelligence Tools | BI & Analytics | SAP. http://go.sap.com/solution/platform-technology/business-intelligence.html
CIOLeadershipForum2015Profile.pdf. (2015). http://www.gartnerinfo.com/cios9/CIOLeadershipForum2015Profile.pdf

[35] www.qlik.com

περισσότερες από 100 χώρες. Τα βασικά προγράμματα της εταιρείας είναι το QlikView και το QlikSense. Το QlikView είναι μια πλατφόρμα για την ανάπτυξη εφαρμογών Επιχειρηματικής Ευφυΐας. Το λογισμικό διαθέτει μια σειρά από ιδιότητες που το καθιστούν αποτελεσματικό και ελκυστικό. Προβλέπεται διαχείριση των δεδομένων μέσα στη μνήμη ώστε να αυξάνεται η ταχύτητα επεξεργασίας. Υπάρχει δυνατότη- τα χρήσης του μέσα από internet browsers με τη χρήση κατάλληλων plug-ins. Επίσης αξιοποιείται η κινητή υπολογιστική και η εφαρμογή είναι προσβάσιμη μέσα από κινητές συσκευές όπως tablets και smartphones. Με το QlikView Desktop ο χρήστης μπορεί να αποκτά πρόσβαση σε δεδομένα, να εκτελεί αναλύσεις και να σχεδιάζει αναφορές και dashboards. Το QlikView Workbench είναι ένα plug in για Microsoft Visual Studio, που επιτρέπει την εύκολη ανάπτυξη εφαρμογών για την επέκταση των λειτουργιών του QlikView. Το πρόγραμμα μπορεί να έχει πρόσβαση σε μεγάλους όγκους δεδομένων μέσα από πηγές συμβατές με πρότυπα όπως το ODBC και το XML. Επίσης το πρόγραμμα μπορεί να συνδεθεί με λογισμικά άλλων κατασκευαστών όπως το SAP ERP, το Salesforce και το Informatica.

Το QlikSense είναι μια εφαρμογή οπτικοποίησης δεδομένων και δημιουργίας αναφορών. Ο χρήστης μπορεί με διαδραστικό και εύκολο τρόπο να διερευνά τα δεδομένα, να υποβάλλει ερωτήσεις και να κατασκευάζει dashboards. Το λογισμικό είναι ικανό να συνδυάζει δεδομένα από πολλαπλές πηγές. Επίσης, είναι προσβάσιμο από φορητές συσκευές και προσαρμόζεται αυτόματα σε αυτές. Έχουν προβλεφθεί ιδιαίτερες λειτουργικότητες που διευκολύνουν τη συνεργασία και τη διανομή των αναλύσεων και των πληροφοριών σε ομάδες. Έμφαση έχει δοθεί στην ευχρηστία και την προσαρμοστικότητα του λογισμικού, ώστε κάθε χρήστης να μπορεί να το χειριστεί σύμφωνα με τις επιθυμίες και τις ανάγκες του.

3 Τεχνικές Data Mining

3.1 ΟΡΙΣΜΟΣ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ ΚΑΙ ΔΕΔΟΜΕΝΩΝ

Η εξόρυξη γνώσης από δεδομένα (data mining) ή πιο απλά η εξόρυξη γνώσης είναι μια νέα δυναμική τεχνολογία που βοηθάει τις επιχειρήσεις να εστιάσουν στην σημαντική πληροφορία που βρίσκεται μέσα στις αποθήκες δεδομένων τους (data warehouses). Οι τεχνικές της είναι σε θέση να αναζητήσουν και να βρουν γρήγορα και λεπτομερειακά βάσεις δεδομένων για την αναζήτηση κρυμμένων προτύπων (patterns). Έτσι λοιπόν μπορούμε να πούμε ότι η εξόρυξη γνώσης είναι μια διαδικασία εξαγωγής κρυμμένης πληροφορίας από μεγάλες βάσεις δεδομένων.

«Εξόρυξη δεδομένων είναι η διαδικασία εξαγωγής υπονοούμενης και εν πολλοίς άγνωστης αλλά ενδεχομένως χρήσιμης γνώσης υπό την μορφή συσχετίσεων προτύπων και τάσεων, μέσω της εξέτασης ανάλυσης και επεξεργασίας βάσεων δεδομένων, συνδυάζοντας και χρησιμοποιώντας τεχνικές από την μηχανική μάθηση, την αναγνώριση προτύπων, την στατιστική, τις βάσεις δεδομένων και την οπτικοποίηση.»[36].

Παρά το γεγονός ότι υπάρχει μια γενικότερη συμφωνία ότι ο στόχος της εξόρυξης δεδομένων είναι η ανακάλυψη νέας και χρήσιμης πληροφορίας σε βάσεις δεδομένων, τα μέσα για την επίτευξη του στόχου αυτού ποικίλουν σε πολύ υψηλό βαθμό. Η εξόρυξη γνώσης περιλαμβάνει ένα ευρύ πεδίο υπολογιστικών μεθόδων που μεταξύ άλλων περιλαμβάνουν, την στατιστική ανάλυση (statistical analysis), τα δένδρα αποφάσεων (decision trees), τα νευρωνικά δίκτυα (neural networks), την εξαγωγή κανόνων (rule induction) και την γραφική οπτικοποίηση (graphic visualization). Τέτοιες μέθοδοι χρησιμοποιούνται για την εύρεση συσχετίσεων, προτύπων και δομών σε μεγάλες και διαρκώς αυξανόμενες βάσεις δεδομένων. Ειδικά η εύρεση εργαλείων είναι ένα ιδιαίτερα σημαντικό εξαγόμενο της εξόρυξης δεδομένων μέσω σχέσεων μεταξύ των χαρακτηριστικών των βάσεων δεδομένων.

3.2 ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΑΝΕΥΡΕΣΗ ΓΝΩΣΗΣ

Η εξόρυξη γνώσης βοηθά τις σύγχρονες εταιρείες να εστιάζουν στα πιο σημαντικά στοιχεία από τις αποθήκες δεδομένων τους. Με άλλα λόγια είναι η διαδικασία

[36] Piatetsky-Shapiro & Frawley, 1991

εφαρμογής μεθόδων ανάλυσης σε μεγάλο όγκο δεδομένων. Ο χρήστης των εργαλείων εξόρυξης μπορεί να προβλέψει μελλοντικές συμπεριφορές και συνήθειες, ώστε οι εταιρίες να παίρνουν επιτυχημένες αποφάσεις. Συνειδητοποιούμε ότι οι τεχνικές εξόρυξης γνώσης αναπτύσσονται γρήγορα, δίχως αλλαγές στην υποδομή και με μοναδικό στόχο την αξιοποίηση των επεξεργασμένων δεδομένων. Στη διεθνή βιβλιογραφία υπάρχει μια γενικότερη σύγχυση ανάμεσα στους όρους «Εξόρυξη Γνώσης» (Data mining) και «Ανεύρεση γνώσης στις βάσεις δεδομένων» (Knowledge discovery in databases, KDD). Σε πολλές περιπτώσεις αξίζει να σημειωθεί ότι οι δύο αυτοί όροι ταυτίζονται, ενώ στην πραγματικότητα η εξόρυξη δεδομένων αποτελεί τμήμα της ανεύρεσης γνώσης, συγκροτώντας το πυρήνα αυτής[37] (Zaiane, 1999). Προκειμένου λοιπόν να κατανοηθεί καλύτερα η εξόρυξη δεδομένων, θα γίνει μια σύντομη αναφορά στη διαδικασία της ανεύρεσης γνώσης.

Η ανεύρεση γνώσης είναι μια επαναληπτική διαδικασία που αποτελείται από μια σειρά βημάτων, τα οποία οδηγούν από τη συλλογή των δεδομένων στην ανακάλυψη και εξαγωγήχρήσιμης πληροφορίας από αυτά.

Τα βήματα από τα οποία αποτελείται η διαδικασία ανεύρεσης γνώσης είναι τα ακόλουθα:

♣ Καθαρισμός δεδομένων (Data cleaning): Στο βήμα αυτό, αφαιρούνται από τη βάση δεδομένων αυτά που παράγουν θόρυβο, δηλαδή όλα εκείνα τα στοιχεία που μπορούν να επηρεάσουν ή και να διαστρεβλώσουν το αποτέλεσμα.

♣ Ενσωμάτωση δεδομένων (Data integration): Σε αυτό το βήμα τα δεδομένα που έχουν συλλεχθεί, πολλές φορές ανομοιογενή και από πολλές διαφορετικές πηγές, ενσωματώνονται σε μια κοινή βάση δεδομένων

Επιλογή δεδομένων (Data selection): Από όλα εκείνα τα δεδομένα που έχουμε στη διάθεση μας, επιλέγονται προσεκτικά εκείνα που είναι σχετικά και χρήσιμα για την ανάλυση που θα ακολουθήσει.

[37] Jessica Enright and Jonathan Klippenstein (2004), “Tanagra: An Evaluation”- URL: http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput695-04/work/A2-reports/tanagra.pdf

♣ Τροποποίηση δεδομένων (Data transformation): Τα δεδομένα που έχουμε επιλέξει δέχονται τις απαραίτητες τροποποιήσεις έτσι ώστε η μορφή τους να είναι κατάλληλη για την διαδικασία της εξόρυξης.

♣ Εξόρυξη δεδομένων (Data mining): Είναι το σημαντικότερο από τα βήματα της διαδικασίας και αυτό γιατί στο συγκεκριμένο στάδιο, ποικίλες εξελιγμένες τεχνικές χρησιμοποιούνται για την εξαγωγή δυνητικά χρήσιμων προτύπων.

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

♣ Αναπαράσταση γνώσης (Knowledge representation): Στο τελικό αυτό στάδιο, η γνώση που έχει ανακαλυφθεί παρουσιάζεται στον χρήστη, βοηθώντας τον έτσι να κατανοήσει και να ερμηνεύσει τα αποτελέσματα της εξόρυξης δεδομένων. Πολλές φορές κάποια από τα παραπάνω βήματα μπορούν να συνδυαστούν μεταξύ τους για το καλύτερο δυνατό αποτέλεσμα. Για παράδειγμα, τα βήματα του καθαρισμού και της ενσωμάτωσης των δεδομένων, μπορούν να υλοποιηθούν μαζί με στόχο την δημιουργία μια αποθήκης δεδομένων. Με την ίδια λογική μπορούν να συνδυαστούν και τα βήματα της επιλογής και τροποποίησης των δεδομένων. Από τα παραπάνω λοιπόν συμπεραίνουμε ότι η εξόρυξη δεδομένων είναι μια διαδικασία- κλειδί για την ανεύρεση γνώσης. Παρόλα αυτά, δεν καταλαμβάνει παρά μόνο ένα μικρό μέρος της όλης προσπάθειας, δεδομένου της πολυπλοκότητας της. Σε αυτό το σημείο αξίζει να σημειωθεί ότι ο χρήστης, εκμεταλλευόμενος την επαναληπτική μορφή της διαδικασίας ανεύρεσης γνώσης, έχει την δυνατότητα να τροποποιήσει τα μέτρα αξιολόγησης, να τελειοποιήσει την διαδικασία της εξόρυξης, να επιλέξει νέα δεδομένα, να τροποποιήσει περαιτέρω τα ήδη υπάρχοντα ή να ενσωματώσει στη βάση νέα από καινούργιες πηγές, με τελικό στόχο την εξαγωγή διαφορετικών και ακόμη πιο κατάλληλων αποτελεσμάτων.

3.3 ΣΤΟΧΟΙ ΤΗΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Οι μέθοδοι εξόρυξης γνώσης στοχεύουν στην ανακάλυψη στοιχειών που θα είναι χρήσιμα για τους οργανισμούς και τις επιχειρήσεις. Πληροφορίες για τυποποιημένες μορφές όπως για παράδειγμα, ότι υπάρχουν πελάτες που θα

ψωνίσουν περισσότερο από δύο φορές σε περίοδο εκπτώσεων ή προσφορών, ή είναι πιθανό να αγοράσουν τουλάχιστον μια φορά κατά την διάρκεια των εορταστικών ημερών, Πάσχα και Χριστουγέννων, είτε για συσχετίσεις όπως όταν ένας πελάτης αγοράζει dvd player τότε πιθανότατα να αγοράσει και κάποια άλλη ηλεκτρονική συσκευή, μπορεί να αποτελέσουν καθοριστικούς παράγοντες για την λήψη αποφάσεων όσον αφορά τη λειτουργία μιας εμπορικής επιχείρησης. Αυτό συμβαίνει επειδή μπορεί να ληφθούν αποφάσεις σχετικά με το ωράριο, το ύψος και τη διάρκεια των εκπτώσεων, ακόμη και για την τοποθέτηση των προϊόντων μέσα στα καταστήματα

Παράλληλα τέτοιου είδους πληροφορίες χρησιμοποιούνται για τον προγραμματισμό χρήσης πρόσθετων αποθηκευτικών χώρων ή και για τον σχεδιασμό διαφορετικών στρατηγικών μάρκετινγκ. Τα στελέχη της επιχείρησης, που είναι υπεύθυνα για την λήψη των αποφάσεων εκμεταλλεύονται τις δυνατότητες της εξόρυξης γνώσης και μετατρέπουν τις γνώσεις σε επιτυχή αποτελέσματα. Παρακάτω περιγράφονται και αναλύονται οι στόχοι της εξόρυξης δεδομένων. Η εξόρυξη δεδομένων έχει λοιπόν σαν βασικούς της στόχους την εφαρμογή τεχνικών πρόβλεψης και συμπεριφοράς τάσεων (prediction), την αναγνώριση, την περιγραφή (description) σε μεγάλες βάσεις δεδομένων[38], καθώς επίσης την ταξινόμηση και την βελτιστοποίηση των πόρων της. Ειδικότερα:

-Πρόβλεψη: Περιλαμβάνει την χρήση μερικών μεταβλητών ή χαρακτηριστικών μιας βάσης δεδομένων για την πρόβλεψη άγνωστων ή μελλοντικών τιμών χρήσιμων μεταβλητών. Με άλλα λόγια, οι διαδικασίες πρόβλεψης της εξόρυξης δεδομένων (predictive data mining tasks), προσπαθούν να κάνουν εκτιμήσεις βγάζοντας συμπεράσματα από τα διαθέσιμα δεδομένα. Η προσπάθεια πρόβλεψης μελλοντικών συμπεριφορών έχει ως στόχο να ληφθούν αποφάσεις που να μεγιστοποιούν το κέρδος και να προλαμβάνουν δυσάρεστες καταστάσεις. Τα αποτελέσματα της εξόρυξης μπορεί να είναι πληροφορίες σχετικές με το ύψος των πωλήσεων ενός καταστήματος για μια συγκεκριμένη χρονική περίοδο, αλλά και αν το κλείσιμο μιας γραμμής παραγωγής θα είχε θετική επίδραση στις πωλήσεις.

[38] Fayyad U., Piatetsky-Shaprio G., Smyth P. and Uthurusamy R., (1996): Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge

Συγχρόνως σε επιστημονικό επίπεδο, η μελέτη παλαιότερων σεισμικών φαινόμενων ίσως να οδηγούσε στην πρόβλεψη σεισμικής δραστηριότητας.

-Αναγνώριση: Σε αυτή τη φάση οι τυποποιημένες μορφές των δεδομένων χρησιμοποιούνται για να δείξουν την ύπαρξη μιας δραστηριότητας ή ενός γεγονότος.

-Περιγραφή: Είναι η διαδικασία η οποία επικεντρώνεται στην ανακάλυψη προτύπων και αναπαριστά τα δεδομένα μιας πολύπλοκης βάσης δεδομένων με όσο το δυνατό πιο κατανοητό και αξιοποιήσιμο τρόπο. Με άλλα λόγια, οι περιγραφικές διαδικασίες της εξόρυξης δεδομένων (descriptive data mining tasks) περιγράφουν τις γενικές ιδιότητες των υπαρχόντων διαθέσιμων δεδομένων.

-Ταξινόμηση: Σε αυτό το στάδιο έχουμε διαχωρισμό των στοιχείων, με αποτέλεσμα να προκύπτουν διαφορετικές κατηγορίες ή κλάσεις. Για παράδειγμα, οι πελάτες ενός σούπερ μάρκετ είναι δυνατόν να χωριστούν σε παρορμητικούς, πιστούς ή αλλιώς όπως θα λέγαμε κανονικούς, σπάνιους και σε φίλους των εκπτώσεων και προσφορών. Κατά την ανάλυση των πωλήσεων αυτή η κατηγοριοποίηση χρησιμοποιείται για να ληφθούν αποφάσεις, ώστε να προσελκυστούν περισσότεροι πελάτες ανεξαρτήτως κατηγορίας.

-Βελτιστοποίηση: Μεταξύ των άλλων σκοπός της εξόρυξης γνώσης είναι η βέλτιστη χρήση κάποιων πόρων κάτω από περιορισμούς. Τέτοιοι πόροι μπορεί να είναι ο χρόνος, ο χώρος, το χρήμα και η μεγιστοποίηση κάποιων μεγεθών, όπως είναι τα κέρδη είτε οι πωλήσεις. Σε αυτή την περίπτωση η εξόρυξη γνώσης έχει κοινά σημεία με την επιχειρησιακή έρευνα[39].

3.4 ΔΙΑΔΙΚΑΣΙΑ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ

Η διαδικασία ανακάλυψης γνώσης από βάσεις δεδομένων (KDD) συνήθως ορίζεται από τα εξής στάδια:

1. Συλλογή
2. Προ επεξεργασία
3. Μετασχηματισμός

[39] Fayyad U., Piatetsky-Shaprio G., Smyth P. and Uthurusamy R., (1996): Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge

4. Εξόρυξη δεδομένων
5. Ερμηνεία και Αξιολόγηση

Υπάρχουν και παραλλαγές για τον ορισμό των σταδίων αυτών σύμφωνα και με το Cross Industry Standard Process for Data Mining (CRISP-DM) όπου τα στάδια έχουν ως εξής:

1. Κατανόηση Θέματος
2. Κατανόηση δεδομένων
3. Προετοιμασία δεδομένων
4. Μοντελοποίηση
5. Αξιολόγηση
6. Ανάπτυξη

Στην πιο απλοποιημένη διαδικασία της διαχωρίζεται στα εξής στάδια:

1. Προ-επεξεργασία
2. Εξόρυξη δεδομένων
3. Επικύρωση αποτελέσματος.

Σε αυτό το σημείο αξίζει να κάνουμε μια αναφορά σε δύο βασικά στάδια που περιλαμβάνει η διαδικασία εξόρυξης γνώσης, αυτό της προ-επεξεργασίας και της μοντελοποίησης.

3.4.1 ΠΡΟ-ΕΠΕΞΕΡΓΑΣΙΑ

Πριν την εφαρμογή των αλγορίθμων εξόρυξης δεδομένων, το ερευνώμενο σύνολο αυτών πρέπει να συναρμολογείται. Καθώς η εξόρυξη δεδομένων μπορεί να αποκαλύψει μόνο τα πρότυπα που πράγματι εμφανίζονται στα δεδομένα, το φάσμα αυτών που ερευνούμε, πρέπει να είναι αρκετά ευρύ για να περιέχει αυτά τα πρότυπα προκειμένου να προκύψει σε ένα αποδεκτό χρονικό διάστημα. Η προ επεξεργασία είναι απαραίτητη για την ανάλυση πολλών παραγόντων-συνόλων πριν την εξόρυξη δεδομένων. Έτσι το ερευνώμενο σύνολο καθαρίζεται. Το καθάρισμα

δεδομένων διαγράφει τις παρατηρήσεις που περιέχουν θόρυβο και αυτές με ελλιπή δεδομένα.

Η εξόρυξη δεδομένων περιλαμβάνει έξι κατηγορίες:

-Ανίχνευση ανωμαλιών (Anomaly detection): Ο προσδιορισμός ασυνήθιστων εγγραφών δεδομένων, που μπορεί να παρουσιάζουν κάποιο ενδιαφέρον ή λάθη στα δεδομένα που απαιτούν περαιτέρω έρευνα

-Κατηγοριοποίηση: Είναι η διαδικασία γενίκευσης γνωστών δομών για την εφαρμογή της πάνω σε νέα δεδομένα. Για παράδειγμα, ένα πρόγραμμα ηλεκτρονικού ταχυδρομείου ενδέχεται να προσπαθήσει να χαρακτηρίσει ένα μήνυμα ηλεκτρονικού ταχυδρομείου ως νόμιμο ή spam.

-Συσταδιοποίηση: Πρόκειται για τη διαδικασία ανακάλυψης ομάδων και δομών στα δεδομένα που είναι «παρόμοια» κατά κάποιο τρόπο, χωρίς να χρησιμοποιούνται γνωστές δομές στα δεδομένα.

-Ανάλυση συσχέτισης (Μοντέλο αλληλεξάρτησης): Αναζητήσεις για σχέσεις μεταξύ των μεταβλητών. Για παράδειγμα, ένα σούπερ μάρκετ μπορεί να συλλέξει δεδομένα που αφορούν της αγοραστικές συνήθεις των πελατών του. Χρησιμοποιώντας τους κανόνες συσχέτισης, μπορεί να υπολογίσει ποια προϊόντα αγοράζονται συνήθως μαζί και να χρησιμοποιήσει αυτή την πληροφορία για αγοραστικούς σκοπούς προς όφελος των πελατών του και του ίδιου.

-Παλινδρόμηση: Προσπαθεί να βρει μία συνάρτηση που μοντελοποιεί τα δεδομένα με το λιγότερο δυνατό λάθος.

-**Σύνοψη**: Παρέχει μια συμπαγέστερη αναπαράσταση των δεδομένων, συμπεριλαμβάνοντας την οπτικοποίηση και την παραγωγή κανόνων.

3.4.2 ΜΟΝΤΕΛΟΠΟΙΗΣΗ

Η τεχνική που εφαρμόζεται για να μάθουμε από την εξόρυξη γνώσης πληροφορίες που δεν γνωρίζουμε ή που θα συμβούν στο μέλλον ονομάζεται μοντελοποίηση. Δηλαδή η κατασκευή ενός μοντέλου για μια κατάσταση που γνωρίζουμε την απάντηση και στη συνέχεια η εφαρμογή του σε μια άλλη που δεν ξέρουμε. Για παράδειγμα, αν αναζητούσαμε μια βυθισμένη ισπανική γαλέρα στην ανοικτή θάλασσα το πρώτο πράγμα που ίσως σκεφτόμασταν θα ήταν να

ερευνήσουμε όλες τις περασμένες περιπτώσεις εύρεσης ισπανικών θησαυρών από άλλους. Ίσως λοιπόν να παρατηρούσαμε ότι αυτά τα πλοία στην πλειονότητα τους βρέθηκαν στις ακτές Βερμούδα και ότι υπήρχαν κάποιες βέβαιες πορείες που ακολουθούσαν οι καπετάνιοι των πλοίων αυτών εκείνη την εποχή. Αυτές οι ομοιότητες σημειώνονται και κτίζεται ένα μοντέλο που περιλαμβάνει τα χαρακτηριστικά που είναι κοινά στις τοποθεσίες αυτών των βυθισμένων θησαυρών. Με αυτό το μοντέλο αρχίζει το ψάξιμο σε περιοχές που δείχνει αυτό ότι είναι πιθανό να υπήρξε μια παρόμοια κατάσταση στο παρελθόν. Αν το μοντέλο είναι καλό ο θησαυρός θα βρεθεί. Η σκέψη κτισίματος μοντέλων από τους ανθρώπους υπήρχε αρκετό καιρό πριν από την τεχνολογία της εξόρυξης γνώσης. Η διαδικασία που ακολουθείται είναι να φορτώνονται οι υπολογιστές με στοιχεία για πολλές καταστάσεις ενώ μια απάντηση είναι γνωστή. Έπειτα το λογισμικό εξόρυξης γνώσης τρέχει πάνω σε αυτό τα δεδομένα και διαλέγει τα πιο χαρακτηριστικά που θα συμπεριληφθούν στο μοντέλο. Όταν τελειώσει η κατασκευή του μοντέλου είναι δυνατό να χρησιμοποιηθεί σε παρόμοιες καταστάσεις που δεν γνωρίζουμε την απάντηση

Για παράδειγμα ας υποθέσουμε ότι βρισκόμαστε στη θέση του διευθυντή μάρκετινγκ μιας εταιρίας τηλεπικοινωνιών και θέλουμε να αποκτήσουμε μερικούς πελάτες που κάνουν τηλεφωνήματα μεγάλων αποστάσεων. Βρισκόμαστε δηλαδή αντιμέτωποι με ένα πρόβλημα απόφασης, σε ποιους να απευθυνθούμε. Θα μπορούσαμε να ταχυδρομήσουμε με τυχαίο τρόπο κουπόνια στο γενικό πληθυσμό όπως θα μπορούσαμε να ταξιδεύουμε στις θάλασσες ψάχνοντας για βυθισμένους θησαυρούς. Πάντως σε καμιά από τις δυο περιπτώσεις δεν θα είχαμε τα επιθυμητά αποτελέσματα. Αντί αυτού, θα μπορούσαμε να χρησιμοποιήσουμε την εμπειρία της εταιρίας που βρίσκεται αποθηκευμένη στις βάσεις δεδομένων και να κτίσουμε ένα μοντέλο. Ο διευθυντής μάρκετινγκ έχει πρόσβαση σε πολλές πληροφορίες σχετικές με τους πελάτες όπως η ηλικία τους, το φύλο, αν είναι καλοί πληρωτές, το πόσα τηλεφωνήματα μεγάλων αποστάσεων κάνουν. Το πρόβλημα είναι ότι δεν γνωρίζουμε πόσο πολύ θα κάνουν χρήση τηλεφωνημάτων σε απομακρυσμένες περιοχές. Επειδή θέλουμε αυτούς που κάνουν πολλά από αυτά τα τηλεφωνήματα, μπορούμε να το πετύχουμε αυτό κτίζοντας ένα μοντέλο. Ένα τέτοιο απλό μοντέλο που θα ταίριαζε σε μια τηλεπικοινωνιακή εταιρία είναι το παρακάτω:

Για παράδειγμα το 98% των πελατών που έχουν λογαριασμό μεγαλύτερο από 6000 ευρώ το χρόνο δαπανούν περισσότερα από 80 ευρώ το μήνα για τηλεφωνήματα σε μακρινές περιοχές. Αυτό το μοντέλο θα μπορούσε να εφαρμοστεί στα δεδομένα των πιθανών πελατών και να δοθεί απάντηση στο πρόβλημα απόφασης. Αφού γίνει αυτό θα ξέρει σε ποιους θα πρέπει να απευθυνθεί η εταιρεία.

Η εξόρυξη γνώσης με άλλα λόγια είναι μια επέκταση της στατιστικής με κάποια στοιχεία τεχνητής νοημοσύνης και μηχανικής μάθησης(machine learning). Η εξόρυξη γνώσης είναι μια τεχνολογία και όπως και η στατιστική δεν αποτελεί επιχειρηματική λύση. Είναι μόνο μια τεχνολογία. Για παράδειγμα, σε περίπτωση που έχουμε ένα κατάλογο εμπόρων λιανικής πώλησης και πρέπει να αποφασιστεί ποιοι από αυτούς θα ενημερωθούν για ένα νέο προϊόν. Η εξόρυξη γνώσης αναζητά την πληροφορία που βρίσκεται μέσα στις βάσεις δεδομένων προηγούμενων συναλλαγών με τους πελάτες καθώς και σε χαρακτηριστικά αυτών, όπως αν ανταποκρίθηκαν στο παρελθόν, η ηλικία τους, η διεύθυνση τους κλπ. Το λογισμικό της εξόρυξης γνώσης χρησιμοποιεί αυτά τα στοιχεία για να κατασκευάσει ένα μοντέλο συμπεριφοράς του πελάτη έτσι ώστε αυτό να χρησιμοποιηθεί για να προβλεφθεί ποιοι πελάτες θα ανταποκριθούν στο νέο προϊόν. Επομένως ένα στέλεχος του τμήματος marketing μπορεί να επιλέξει τους πιθανούς πελάτες. Αντιλαμβανόμαστε ότι το λογισμικό της εταιρείας έχει την δυνατότητα να τροφοδοτεί τα κατάλληλα σημεία επαφής (web servers, τηλεφωνικά κέντρα, e-mails κλπ) με τις αποφάσεις έτσι ώστε οι πελάτες να παίρνουν τις πληροφορίες που χρειάζονται.

Παρακάτω βλέπουμε τα στάδια που μεσολαβούν μέχρι να είναι δυνατή η ερμηνεία και η ανάλυση των αποτελεσμάτων. Άρα η ανακάλυψη γνώσης, ή αλλιώς η διαδικασία καθορισμού και επίτευξης ενός σκοπού μέσω επαναληπτικής εξόρυξης γνώσης, αποτελείται από τα εξής τρία στάδια:

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

Αρχικά ο αναλυτής προετοιμάζει ένα σύνολο στοιχείων για να κτιστεί ένα σωστό μοντέλο στις επόμενες φάσεις. Στοχεύοντας τις αναγκαίες πληροφορίες για μια επιχείρηση, ένα σωστό μοντέλο θα προβλέπει τι πιθανότητα υπάρχει ο πελάτης να αγοράσει προϊόντα από έναν νέο κατάλογο. Οι προβλέψεις βασίζονται σε παράγοντες που επιδρούν τις αγορές των πελατών και γι' αυτό ένα μοντέλο συνόλου δεδομένων θα έπρεπε να περιέχει όλους τους πελάτες που ανταποκρίθηκαν σε καταλόγους μέσω ταχυδρομείων, e-mails κλπ. τα τελευταία 4 χρόνια, τα 8 πιο ακριβά προϊόντα που αγόρασε κάθε πελάτης, τις δημογραφικές πληροφορίες τους, και στοιχεία για τους καταλόγους που έγιναν οι αγορές. Συνειδητοποιούμε ότι πολύπλοκες ερωτήσεις με μεγάλες απαντήσεις περιλαμβάνονται στην προετοιμασία των δεδομένων.

Για παράδειγμα για την εταιρία που αναφερθήκαμε προηγουμένως, η προετοιμασία του μοντέλου έχει συνδέσεις μεταξύ του πίνακα πωλήσεων και του πίνακα πελατών, καθώς και για τον προσδιορισμό των 8 κορυφαίων προϊόντων για κάθε πελάτη. Επομένως η αποτελεσματική επεξεργασία ερωτήσεων υποστήριξης αποφάσεων σχετίζονται με το περιβάλλον της εξόρυξης γνώσης. Η εξόρυξη γνώσης περιλαμβάνει την επαναληπτική κατασκευή μοντέλων πάνω σε ένα σύνολο δεδομένων που έχει προετοιμαστεί και εν συνεχεία στην ανάπτυξη ενός ή περισσοτέρων μοντέλων. Εκτός των άλλων οι αναλυτές - ειδικοί εργάζονται με επαναληπτικό τρόπο σε δείγματα συνόλων δεδομένων, επειδή το κτίσιμο των μοντέλων σε μεγάλα σύνολα δεδομένων είναι αρκετά ακριβό. Ο αναλυτής κατασκευάζει το μοντέλο πάνω στο σύνολο δεδομένων, αφού όμως πρώτα έχει αποφασιστεί ποιο μοντέλο θα χρησιμοποιηθεί.

Στη φάση της υλοποίησης εντοπίζονται οι τυποποιημένες μορφές που ορίζουν ένα χαρακτηριστικό-στόχος (target attribute). Αν και μερικές κλάσεις μοντέλων εξόρυξης γνώσης συμβάλλουν σημαντικά στην πρόβλεψη τόσο κρυφών χαρακτηριστικών όσο και φανερά καθορισμένων, κρίνονται αναγκαία, για την επιλογή του μοντέλου τα χαρακτηριστικά της ακρίβειας και της αποτελεσματικότητας του αλγορίθμου κατασκευής του μοντέλου σε μεγάλα σύνολα δεδομένων. Αξιοπρόσεκτο είναι το γεγονός ότι από στατιστικής πλευράς η ακρίβεια των περισσότερων μοντέλων βελτιώνεται με το πλήθος των δεδομένων που χρησιμοποιούνται

3.5 ΜΕΘΟΔΟΙ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ ΚΑΙ ΔΕΔΟΜΕΝΩΝ[40]

Οι βασικότερες από τις μεθόδους της εξόρυξης δεδομένων, μέσω των οποίων επιτυγχάνονται οι στόχοι που αναφέραμε προηγουμένως, είναι οι εξής:

¬ Κατηγοριοποίηση

¬ Συσταδιοποίηση

¬ Ανάλυση Συσχέτισης

¬ Παλινδρόμηση

3.5.1 ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

Η διαδικασία της κατηγοριοποίησης, ή αλλιώς ταξινόμησης (classification) περιλαμβάνει την οργάνωση ενός συνόλου αντικειμένων (objects) που περιγράφονται από ένα σύνολο χαρακτηριστικών (attributes), σε μια σειρά από προκαθορισμένες κλάσεις (classes), χρησιμοποιώντας μεθόδους μάθησης με επίβλεψη (supervised learning methods). Οι τεχνικές της ταξινόμησης ή αλλιώς κατηγοριοποίησης χρησιμοποιούν κατά κανόνα ένα σύνολο εκπαίδευσης (training set), όπου όλα τα αντικείμενα είναι ήδη συνδεδεμένα με γνωστές κλάσεις. Ο αλγόριθμος ταξινόμησης μαθαίνει από αυτό το σύνολο, χρησιμοποιώντας την μάθηση αυτή για την κατασκευή ενός μοντέλου και το μοντέλο αυτό στην συνέχεια ταξινομεί νέα αντικείμενα στις κατάλληλες κλάσεις[41]. Άρα μπορούμε να πούμε ότι η κατηγοριοποίηση μαθαίνει σε μία λειτουργία να χαρτογραφεί ή πιο απλά να ταξινομεί ένα στοιχείο δεδομένων σε μία από τις διάφορες προκαθορισμένες κατηγορίες. Η κατηγοριοποίηση πρόκειται ίσως για την πιο δημοφιλή τεχνική με πλήθος εφαρμογών στην αναγνώριση προτύπων και εικόνας σε διάφορους κλάδους.

Στην πράξη μια διαδικασία κατηγοριοποίησης μπορεί να οριστεί ως η εκτέλεση δύο συγκεκριμένων βημάτων:

1. Δημιουργία μοντέλου βασιζόμενου σε δεδομένα εκπαίδευσης

[40] http://axon.cs.byu.edu/~martinez/classes/478/readings/DataPrep.pdf
[41] Jessica Enright and Jonathan Klippenstein (2004), “Tanagra: An Evaluation”- URL: http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput695-04/work/A2-reports/tanagra.pdf

2. Εφαρμογή του μοντέλου στο σύνολο των δεδομένων

Αν και βάσει του επιστημονικού ορισμού το δεύτερο από τα παραπάνω βήματα είναι αυτό της κατηγοριοποίησης, το πρώτο είναι το βήμα που απαιτεί και την μεγαλύτερη προσπάθεια. Η εργασία της κατηγοριοποίησης χαρακτηρίζεται από έναν καλά καθορισμένο ορισμό των κατηγοριών και το σύνολο που χρησιμοποιείται για την εκπαίδευση του μοντέλου αποτελείται από προ κατηγοριοποιημένα παραδείγματα.

Η επόμενη εικόνα δείχνει μία γενική προσέγγιση επίλυσης προβλήματος κατηγοριοποίησης. Αρχικά, πρέπει να δοθεί ένα training set το οποίο περιέχει εγγραφές των οποίων οι ετικέτες κατηγορίας είναι σωστές. Το training set χρησιμοποιείται για να φτιάξει το μοντέλο ταξινόμησης, το οποίο μετέπειτα εφαρμόζεται στο test set, όπου περιέχει εγγραφές των οποίων οι ετικέτες κατηγορίας είναι άγνωστες. Η διαδικασία που ακολουθείται δηλαδή έχει να κάνει με την παραγωγή της test set με άγνωστες ετικέτες από το αρχικό training set οι οποίες πρέπει να προβλεφθούν από κάποιον αλγόριθμο με όσο το δυνατό μεγαλύτερη επιτυχία. Σκοπός της παρούσας μελέτης είναι οι βελτιωμένες προβλέψεις των test set που περιλαμβάνουν δεδομένα επιχειρήσεων εισηγμένων στο ΧΑΑ.

Η αξιολόγηση της απόδοσης ενός μοντέλου ταξινόμησης βασίζεται στον αριθμό των εγγραφών του test set που προβλέφθηκαν σωστά ή λάθος από τον ταξινομητή. Για να είναι ευκολότερη η σύγκριση των αποδόσεων διαφορετικών μοντέλων χρησιμοποιούνται δύο δείκτες επίδοσης, η ακρίβεια (accuracy) και η αποτίμηση του σφάλματος (error rate).

Έτσι τελικά ο ταξινομητής με τη μεγαλύτερη ακρίβεια και το μικρότερη αποτίμηση σφάλματος είναι ορθότερος και πιο αποτελεσματικός, δηλαδή μπορεί και κάνει καλύτερες προβλέψεις.

Στην παρακάτω εικόνα έχουμε έναν απλό διαχωρισμό των στοιχείων δανείου σε δύο περιοχές κατηγοριών. Η τράπεζα πιθανώς να θελήσει να χρησιμοποιήσει τις περιοχές ταξινόμησης για να αποφασίσει, εάν θα δοθεί δάνειο ή όχι στους μελλοντικούς υποψηφίους

Ένα απλό γραμμικό όριο κατηγοριοποίησης για το σύνολο των στοιχείων δανείου. Η διαμορφωμένη περιοχή δείχνει την κατηγορία (απόρριψη-έγκριση) και όχι το δάνειο.

Η εργασία της κατηγοριοποίησης χαρακτηρίζεται από έναν καλά καθορισμένο ορισμό των κατηγοριών και το σύνολο που χρησιμοποιείται για την εκπαίδευση του μοντέλου αποτελείται από προ κατηγοριοποιημένα παραδείγματα. Η βασική εργασία είναι να δημιουργηθεί ένα μοντέλο το οποίο θα μπορούσε να εφαρμοστεί για να οργανώσει δεδομένα που δεν έχουν ακόμα κατηγοριοποιηθεί. Στις περισσότερες περιπτώσεις, υπάρχει ένα περιορισμένος αριθμός κατηγοριών και εμείς θα πρέπει να αναθέσουμε κάθε εγγραφή στην κατάλληλη κατηγορία. Για αυτό το σκοπό χρησιμοποιούνται κάποιες τεχνικές, τις οποίες μπορούμε να κατατάξουμε σε δύο βασικές κατηγορίες. Η πρώτη χρησιμοποιεί τα λεγόμενα δέντρα απόφασης (decision trees) ενώ η δεύτερη τα νευρωνικά δίκτυα (neural networks).

Γενικά μπορούμε να πούμε ότι οι αλγόριθμοι κατηγοριοποίησης μπορούν να διαχωριστούν στις ακόλουθες κατηγορίες, κάποιες από τις οποίες θα αναλύσουμε στη συνέχεια:

- Στατιστικοί Αλγόριθμοι
- Αλγόριθμοι Απόστασης
- Δένδρα Απόφασης
- Νευρωνικά Δίκτυα
- Αλγόριθμοι Κανόνων

3.5.1.1 ΔΕΝΔΡΑ ΑΠΟΦΑΣΗΣ

Τα δέντρα απόφασης (decision trees) είναι μια από τις πιο σημαντικές και ευρύτατα διαδεδομένες μεθόδους για την ταξινόμηση δεδομένων. Σύμφωνα με τους Quinlan (1986, 1987, 1993) και Murphy (1998), τα δέντρα απόφασης είναι δομές που ταξινομούν τα αντικείμενα μιας βάσης δεδομένων βάσει των τιμών των χαρακτηριστικών αυτών. Η κατασκευή του από την άλλη βασίζεται σε ένα σύνολο εκπαίδευσης, το οποίο περιλαμβάνει προ-ταξινομημένα δεδομένα.

Η ταξινόμηση ενός νέου αντικειμένου μέσω ενός δέντρου απόφασης ακολουθεί κάποια βήματα. Ξεκινώντας από την ρίζα του δέντρου (αρχικός κόμβος) και εξετάζοντας τα χαρακτηριστικά που καθορίζονται από τον κόμβο αυτό, προσδιορίζονται διαδοχικά οι εσωτερικοί κόμβοι του δέντρου που πρέπει να ακολουθηθούν, έως ότου καταλήξουμε σε ένα συγκεκριμένο φύλλο. Πολλά διαφορετικά φύλλα μπορούν να οδηγούν στην ίδια ταξινόμηση, αλλά κάθε φύλλο κάνει την ταξινόμηση αυτή για διαφορετικό λόγο. Σε κάθε εσωτερικό κόμβο, εξετάζεται αν το προς ταξινόμηση αντικείμενο ικανοποιεί τον συγκεκριμένο κόμβο. Η έκβαση της εξέτασης αυτής καθορίζει το κλαδί που θα ακολουθηθεί στην συνέχεια, καθώς και τον επόμενο κόμβο. Η κλάση στην οποία θα ταξινομηθεί το νέο αντικείμενο αντιστοιχεί σε ένα από τα φύλλα του δέντρου απόφασης, είναι δε αυτή του τελικού κόμβου (Mitchell, 1997).

Μερικά από τα κρίσιμα ζητήματα που αφορούν τους αλγόριθμους δημιουργίας δένδρων απόφασης ή κατηγοριοποίησης είναι τα ακόλουθα:

-**Η επιλογή των γνωρισμάτων διάσπασης**: Τα γνωρίσματα του συνόλου δεδομένων που θα επιλεγούν για τη δημιουργία του δένδρου είναι κρίσιμης σημασίας. Αναμφισβήτητα, κάποια γνωρίσματα είναι σημαντικότερα από κάποια άλλα και η επιλογή των καταλληλότερων είναι πολλές φορές όχι μόνο θέμα εξέτασης των δεδομένων εκπαίδευσης αλλά και εμπεριστατωμένης άποψης ειδικών πάνω στη φύση των δεδομένων.

-**Η διάταξη των γνωρισμάτων διάσπασης**: Εκτός από το ποια γνωρίσματα είναι πλέον κατάλληλα, κρίσιμη απόφαση είναι και η διάταξη των καταλληλότερων. Μια λάθος διάταξη γνωρισμάτων διάσπασης μπορεί να σημαίνει τον επανέλεγχο γνωρισμάτων αρκετές φορές.

-**Οι διασπάσεις**: Γνωρίσματα με μικρό πεδίο τιμών οδηγούν σε φανερό αριθμό διασπάσεων, σε αντίθεση με γνωρίσματα συνεχών πεδίων όπου ο αριθμός των διασπάσεων κάθε άλλο παρά εύκολος μπορεί να θεωρηθεί.

-**Η δομή του δένδρου**: Η δομή του δένδρου απόφασης ασφαλώς και παίζει σημαντικό ρόλο στη δεύτερη από τις δύο φάσεις κατηγοριοποίησης, αυτή της εφαρμογής του δένδρου πάνω στις πλειάδες της βάσης δεδομένων. Ισοζυγισμένα

δένδρα λίγων επιπέδων ασφαλώς και βοηθούν στην αποδοτικότερη κατηγοριοποίηση.

-**Τα κριτήρια του τερματισμού**: Η δημιουργία του δένδρου απόφασης ολοκληρώνεται όταν τα δεδομένα κατηγοριοποιούνται με απόλυτη ακρίβεια. Αυτό ωστόσο κρύβει κινδύνους στη δημιουργία μεγάλων δένδρων. Από αυτό συμπεραίνουμε ότι ο συμβιβασμός μεταξύ ακρίβειας της κατηγοριοποίησης και αποδοτικότητας του δένδρου είναι απαραίτητος.

-**Τα δεδομένα εκπαίδευσης**: Η δημιουργία του δένδρου βασίζεται αποκλειστικά στα δεδομένα εκπαίδευσης. Μικρό σύνολο τέτοιων δεδομένων ίσως οδηγήσει σε δένδρο μη κατάλληλο για το σύνολο των δεδομένων που διαθέτουμε. Μεγάλο σύνολο δεδομένων εκπαίδευσης μπορεί να προκαλέσει υπερ-προσπαρμογή.

-**Το κλάδεμα του δένδρου**: Η ολοκλήρωση της δημιουργίας ενός δένδρου απόφασης πολλές φορές απαιτεί την αφαίρεση περιττών συγκρίσεων ή και τη διαγραφή ολόκληρων κλαδιών με στόχο την καλύτερη απόδοση της κατηγοριοποίησης. Η φάση του κλαδέματος έχει ως σκοπό τη βέλτιστη απόδοση του δένδρου.

Οι αλγόριθμοι ταξινόμησης που βασίζονται στα δέντρα απόφασης, περιλαμβάνουν δύο διακριτές φάσεις:

1. **Τη φάση οικοδόμησης (building phase)**: Σε αυτή την πρώτη φάση, η οποία χρίζει μεγαλύτερης έρευνας και προσπάθειας, το σύνολο των δεδομένων εκπαίδευσης χωρίζεται πολλές φορές, έως ότου όλα τα αντικείμενα σε ένα τμήμα του ανωτέρω συνόλου να ανήκουν στην ίδια κλάση.

2. **Τη φάση κλαδέματος (pruning phase)**: Έπειτα, αφού έχει ήδη δημιουργηθεί το δέντρο απόφασης, οι περισσότεροι αλγόριθμοι εκτελούν τη φάση του κλαδέματος, περικόπτοντας κάποιους από τους κόμβους, προκειμένου αφενός να αποτραπούν επικαλύψεις, και αφετέρου το δέντρο να έχει υψηλότερη ακρίβεια

ταξινόμησης. Τα πλεονεκτήματα από τη χρήση δένδρων αποφάσεων κατηγοριοποίησης είναι πολλά και παρατίθενται παρακάτω:

i. Τα δένδρα απόφασης είναι εύκολα στη χρήση και αποτελεσματικά, με κανόνες κατανοητούς και βατούς ως προς την ερμηνεία τους.

ii. Δένδρα απόφασης μπορούν να κατασκευαστούν και για τα δεδομένα με πολλά γνωρίσματα.

iii. Λειτουργούν πάρα πολύ καλά σε μεγάλες βάσεις δεδομένων λόγω του γεγονότος ότι το μέγεθος του δένδρου είναι ανεξάρτητο από το μέγεθος της βάσης.

iv. Η ευρωστία που επιδεικνύουν αναφορικά με το θόρυβο που ενδέχεται να παρουσιαστεί στα δεδομένα που απαρτίζουν το χώρο του προβλήματος.

v. Η ανοχή στην απουσία τιμών (missing values), σε κάποια χαρακτηριστικά του σώματος εκπαίδευσης.

vi. Η χρήση ακόμα και συνεχών (μη διακριτών) χαρακτηριστικών και η προσέγγιση μη διακριτών συναρτήσεων στόχου, μέσω εξειδικευμένων τεχνικών που αναλαμβάνουν τη διακριτοποίηση τους (discretization), τη διαδικασία δηλαδή της μετατροπής συνεχών αριθμητικών χαρακτηριστικών σε κατηγορικά.

vii. Η δυνατότητα μεταφοράς του παραγόμενου μοντέλου από δένδρο απόφασης σε ένα σύνολο κανόνων, προς διευκόλυνση της κατανόησής του.

Δε λείπουν ωστόσο και τα μειονεκτήματα από τη χρήση δένδρων απόφασης μερικά από τα οποία είναι:

i. Δε χειρίζονται εύκολα δεδομένα, τα γνωρίσματα των οποίων αποτελούνται από συνεχείς τιμές.
ii. ii. Υπάρχει η πιθανότητα υπερ-προσπαρμογής ενός δένδρου στα σύνολα δεδομένων εκπαίδευσης.

3.5.1.2 ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ

Πέρα από τις μεθόδους ταξινόμησης που βασίζονται στα δέντρα και τους κανόνες απόφασης, τα τεχνητά νευρωνικά δίκτυα (artificial neural networks) είναι επίσης μια διαδεδομένη μέθοδος ταξινόμησης[42] .

Συγκεκριμένα, είναι μια δομή που αποτελείται από ένα δίκτυο νευρώνων (neurons) οι οποίοι συνδέονται μεταξύ τους και αποτελούν τα δομικά στοιχεία του δικτύου. Κάθε τέτοιος κόμβος δέχεται ένα σύνολο αριθμητικών εισόδων από διαφορετικές πηγές (είτε από άλλους νευρώνες, είτε από το περιβάλλον), επιτελεί έναν υπολογισμό με βάση αυτές τις εισόδους και παράγει μία έξοδο. Η εν λόγω έξοδος είτε κατευθύνεται στο περιβάλλον, είτε τροφοδοτείται ως είσοδος σε άλλους νευρώνες του δικτύου. Η πιο διαδεδομένη κατηγορία νευρωνικών δικτύων είναι τα λεγόμενα δίκτυα πρόσθιας τροφοδότησης (feed-forward neural networks), τα οποία επιτρέπουν την κίνηση των δεδομένων μόνο προς μια κατεύθυνση, δηλαδή από μιαείσοδο προς μια έξοδο και έχουμε και τα δίκτυα που σχηματίζουν κυκλικές δομές τα οποία ονομάζονται ανατροφοδοτούμενα νευρωνικά δίκτυα (recurrent neural networks)[43].

Τα νευρωνικά δίκτυα είναι μία προσέγγιση ανάπτυξης και εκτίμησης μαθηματικών δομών. Οι μέθοδοι αυτοί είναι αποτελέσματα ακαδημαϊκών ερευνών με στόχο την μοντελοποίηση συστημάτων μάθησης. Τα νευρωνικά δίκτυα έχουν την ικανότητα να εξάγουν κάποιο συμπέρασμα από πολύπλοκα ή μη ακριβή δεδομένα και μπορούν να χρησιμοποιηθούν για να εξάγουν πρότυπα και να προσδιορίζουν τάσεις οι οποίες είναι πολύπλοκες για να προσδιοριστούν από ανθρώπους ή από άλλες υπολογιστικές τεχνικές. Ένα εκπαιδευμένο νευρωνικό δίκτυο μπορεί να αντιμετωπιστεί ως ένας ειδικός για την κατηγορία της πληροφορίας που του δόθηκε να αναλύσει. Έτσι μπορεί να χρησιμοποιηθεί για να κάνει κάποιες προβλέψεις, όταν προκύψουν κάποιες νέες περιπτώσεις. Τα νευρωνικά δίκτυα χρησιμοποιούν ένα σύνολο από στοιχεία επεξεργασίας (κόμβους) ανάλογους με τους νευρώνες στο ανθρώπινο μυαλό. Τα στοιχεία αυτά διασυνδέονται μεταξύ τους σε ένα δίκτυο το οποίο μπορεί να αναγνωρίζει πρότυπα μέσα σε ένα σύνολο δεδομένων μόλις αυτά

[42] Michie et al, ο.π. 1995, Kotsiantis, 2007
[43] Ρίζος Αν., ό.π. 2004

παρουσιαστούν μέσα στα δεδομένα, δηλαδή το δίκτυο μπορεί να μαθαίνει από την εμπειρία όπως ακριβώς κάνουν και οι άνθρωποι. Αυτό διακρίνει τα νευρωνικά δίκτυα από τα παραδοσιακά προγράμματα υπολογιστών, τα οποία απλά ακολουθούν οδηγίες σύμφωνα με μία καλά ορισμένη σειρά.

Το κύριο χαρακτηριστικό των νευρωνικών δικτύων είναι η εγγενής ικανότητα μάθησης. Ως μάθηση μπορεί να οριστεί η σταδιακή βελτίωση της ικανότητας του δικτύου να επιλύει κάποιο πρόβλημα όπως για παράδειγμα η σταδιακή προσέγγιση μίας συνάρτησης. Η μάθηση επιτυγχάνεται μέσω της εκπαίδευσης μιας επαναληπτικής διαδικασίας σταδιακής προσαρμογής των παραμέτρων του δικτύου, σε τιμές κατάλληλες ώστε να επιλύεται με επαρκή επιτυχία το προς εξέταση πρόβλημα. Αφού ένα δίκτυο εκπαιδευτεί, οι παράμετροί του συνήθως παγώνουν στις κατάλληλες τιμές και έπειτα είναι σε λειτουργική κατάσταση. Το ζητούμενο είναι το λειτουργικό δίκτυο να χαρακτηρίζεται από μία ικανότητα γενίκευσης. Αυτό σημαίνει ότι πρέπει να δίνει ορθές εξόδους για εισόδους καινοφανείς και διαφορετικές από αυτές με τις οποίες εκπαιδεύτηκε.

Σε ένα νευρωνικό δίκτυο πρόσθιας τροφοδότησης, τα κύρια βήματα για την κατασκευή ενός μοντέλου ταξινόμησης, είναι τα εξής[44]:

♣ Η αναγνώριση των χαρακτηριστικών εισόδου και εξόδου

♣ Η κατασκευή ενός δικτύου με την κατάλληλη τοπολογία

♣ Η επιλογή του σωστού συνόλου εκπαίδευσης το οποίο περιλαμβάνει δεδομένα που είναι ορισμένα ανά ζεύγη

♣ Η εκπαίδευση του δικτύου στην οποία τα δεδομένα εισέρχονται στο νευρωνικό δίκτυο ένα ένα. Το νευρωνικό δίκτυο μαθαίνει συγκρίνοντας τα αποτελέσματα ταξινόμησης ενός αντικειμένου με την γνωστή πραγματική ταξινόμηση αυτού. Τα λάθη από την αρχική ταξινόμηση του πρώτου αντικειμένου χρησιμοποιούνται για να διορθωθεί το δίκτυο μέσω της τροποποίησης των συναρτήσεων των νευρώνων. Η παραπάνω διαδικασία είναι επαναληπτική. Η επαναληπτική φύση ωστόσο της διαδικασίας εκπαίδευσης σημαίνει ότι ένα νευρωνικό δίκτυο είναι αρκετά αργό.

[44] Aggarwal & Yu, 1999, Βαζιργιάννης & Χαλκίδη, 2003

♣ Ο έλεγχος του δικτύου χρησιμοποιώντας ένα σύνολο ελέγχου, το οποίο είναι ανεξάρτητο από το σύνολο εκπαίδευσης.

Οι νευρώνες ενός δικτύου χωρίζονται σε τρείς βασικές κατηγορίες:

1) **Τους νευρώνες εισόδου** (input neurons): οι οποίοι δέχονται τις πληροφορίες που θα υποστούν επεξεργασία

2) **Τους νευρώνες εξόδου** (output neurons): στους οποίους καταλήγουν τα αποτελέσματα της παραπάνω επεξεργασίας

3) **Τους ενδιάμεσους νευρώνες**: οι οποίοι βρίσκονται μεταξύ των νευρώνων εισόδου και εξόδου. Οι τελευταίοι εναλλακτικά ονομάζονται και κρυφοί νευρώνες (hidden neurons).

Ουσιαστικά, οι νευρώνες σε ένα δίκτυο είναι αφενός ένα σύνολο εισερχόμενων τιμών και των αντίστοιχων βαρών τους και αφετέρου μια συνάρτηση που αθροίζει τα παραπάνω βάρη, αντιστοιχώντας τα αποτελέσματα σε ένα νευρώνα εξόδου[45].

Καταλήγοντας αξίζει να σημειώσουμε ότι εν πολλοίς η εκπαίδευση ενός νευρωνικού δικτύου βασίζεται στον υπολογισμό των τιμών των βαρών που προ αναφέρθηκαν. Ο πιο γνωστός αλγόριθμος, μεταξύ άλλων[46], στον οποίο βασίζεται ο παραπάνω υπολογισμός, είναι ο αλγόριθμος ανάστροφης μετάδοσης (back propagation algorithm)[47]. Άλλες προσεγγίσεις που χρησιμοποιούνται για την εκπαίδευση των νευρωνικών δικτύων, με κύριο στόχο την βελτίωση των χρονικών τους επιδόσεων, είναι αυτές των Weigend et al (1990) και Yam & Chow (2001). Επιπλέον, για την εκπαίδευση των νευρωνικών δικτύων μπορούν να χρησιμοποιηθούν τόσο γενετικοί αλγόριθμοι όσο και στατιστικές μέθοδοι.

Στο παρακάτω σχήμα παραθέτουμε μία χαρακτηριστική απεικόνιση ενός νευρωνικού δικτύου, όπου διακρίνονται οι νευρώνες εισόδου, οι κρυμμένοι νευρώνες και οι νευρώνες εξόδου όπως περιγράφτηκαν παραπάνω.

[45] Aggarwal & Yu, 1999, ο.π.
[46] Neocleous & Schizas, 2002
[47] Rumelhart et al, 1986

1.5.1.3 ΜΕΘΟΔΟΙ ΠΟΥ ΣΤΗΡΙΖΟΝΤΑΙ ΣΤΟΥΣ ΚΑΝΟΝΕΣ ΑΠΟΦΑΣΗΣ

Στη παρούσα εργασία δε θα αναλύσουμε τους αλγόριθμους που εξάγουν κανόνες απόφασης, παρακάτω όμως θα αναφερθούμε γενικά, παρουσιάζοντας τους πιο βασικούς. Μια πολύ σημαντική ιδιότητα λοιπόν των δέντρων απόφασης, είναι η ικανότητα μετατροπής τους σε ένα σύνολο κανόνων απόφασης (decision rules)[48] . Συγκεκριμένα, δημιουργείται ένας ξεχωριστός κανόνας για κάθε μονοπάτι που ξεκινά από την κορυφή του δέντρου και καταλήγει σε ένα φύλλο που αναπαριστά μια κλάση. Επιπλέον, τα περισσότερα από τα άλλα είδη τυποποίησης των εξαγομένων των αλγορίθμων της εξόρυξης δεδομένων, όπως οι λίστες απόφασης (decision lists), τα προς τα κάτω αναπτυσσόμενα σύνολα κανόνων (ripple down rule sets), τα επαγωγικά λογικά προγράμματα (inductive logic programs) ή τα νευρωνικά δίκτυα (neural networks), μπορούν επίσης να μετατραπούν σε κανόνες. Ειδικά για την μετατροπή των τελευταίων σε κανόνες απόφασης, η διεθνής βιβλιογραφία είναι ιδιαίτερα πλούσια[49]. Ωστόσο, αξίζει να σημειωθεί ότι οι κανόνες απόφασης μπορούν επιπλέον να εξαχθούν και απ' ευθείας από το σύνολο εκπαίδευσης μιας βάσης δεδομένων, μέσω μιας σειράς αλγορίθμων ταξινόμησης, οι οποίοι βασίζονται στους κανόνες απόφασης (rule-based methods).

Στόχος των παραπάνω αλγορίθμων είναι η εξαγωγή του μικρότερου δυνατού συνόλου κανόνων απόφασης που είναι συνεπές με τα υπό εκπαίδευση δεδομένα. Οι εξαχθέντες κανόνες απόφασης έχουν την γενική μορφή «If A Then B», με το «If» κομμάτι να αποτελεί ένα συνδυασμό ζευγών από τιμές χαρακτηριστικών, αναπαριστώντας τις επαρκείς συνθήκες για την εφαρμογή-ανάθεση της τιμής της κλάσης που περιγράφεται στο «Then» κομμάτι του κανόνα, στο υπό ταξινόμηση αντικείμενο της βάσης δεδομένων. Ένας αλγόριθμος που βασίζεται στους κανόνες απόφασης, πρέπει να παράγει κανόνες οι οποίοι έχουν υψηλές ικανότητες πρόβλεψης και ταυτόχρονα υψηλή αξιοπιστία. Σημαντικό ρόλο σε αυτό διαδραματίζουν συνήθως μηχανισμοί, που είτε καθιστούν πολύ εξειδικευμένους κανόνες πιο γενικούς, σε μια ξεχωριστή φάση κλαδέματός τους, είτε σταματούν την διαδικασία εξειδίκευσης των κανόνων μέσω της χρήσης μέτρων ποιότητας. Αυτά τα μέτρα ποιότητας, χρησιμοποιούνται τόσο στην διαδικασία εξαγωγής των κανόνων

[48] Quinlan M. Δ., 1993, Extravagant Business systems and Algorithms, Business Intelligence Review
[49] Xiao Zhou, 2004

όσο και στην διαδικασία ταξινόμησης του εκάστοτε αλγορίθμου. Αφενός, στην διαδικασία εξαγωγής των κανόνων, ένα μέτρο αξιολόγησης της ποιότητάς τους μπορεί να χρησιμοποιηθεί σαν κριτήριο της διαδικασίας εξειδίκευσης ή και γενίκευσης των κανόνων, αφετέρου στην διαδικασία της ταξινόμησης, μια τιμή ενός μέτρου αξιολόγησης ποιότητας μπορεί να αντιστοιχιστεί σε κάθε κανόνα, για την επίλυση συγκρούσεων στην περίπτωση που πολλοί κανόνες ταυτόχρονα ικανοποιούν το προς ταξινόμηση αντικείμενο.

Οι An & Cercone (2000) αναφέρονται αναλυτικά στα σημαντικότερα από τα μέτρα αξιολόγησης της ποιότητας των κανόνων. Αξιόλογες αναφορές στα παραπάνω μέτρα γίνονται επίσης, μεταξύ άλλων, και από τους Lavrac et al (1999), Stefanowski & Vanderpooten (2001), Flach & Lavrac (2003) και Tsumoto (2003). Στην διεθνή βιβλιογραφία υπάρχει ένας πολύ μεγάλος αριθμός αλγορίθμων ταξινόμησης που βασίζονται στους κανόνες απόφασης. Αναλυτική αναφορά σε αυτούς γίνεται στον Furnkranz (1999). Ένας από τους σημαντικότερους αλγορίθμους που βασίζεται στους κανόνες απόφασης είναι ο αλγόριθμος RIPPER (Cohen, 1995), ο οποίος διαμορφώνει κανόνες μέσα από μια συνεχή διαδικασία ανάπτυξης (growing) και κλαδέματος (pruning).

Στην διάρκεια της πρώτης φάσης, οι δημιουργηθέντες κανόνες είναι πιο συνεπτυγμένοι, με στόχο την καλύτερη δυνατή προσαρμογή τους στα δεδομένα του συνόλου εκπαίδευσης, ενώ στην δεύτερη φάση συμβαίνει ακριβώς το αντίθετο, με στόχο την καλύτερη απόδοση του αλγορίθμου σε νέα δεδομένα. Άλλοι σημαντικοί αλγόριθμοι είναι αυτοί της οικογένειας AQ, ο αλγόριθμος PART καθώς και ο CN2. Ειδικά ο αλγόριθμος CN2 είναι από τους πιο σημαντικούς αλγόριθμους που βασίζονται σε κανόνες. Βασισμένος στην «If A Then B» μορφή των κανόνων, χρησιμοποιεί μια συνάρτηση για τον τερματισμό της διαδικασίας κατασκευής τους, βάσει μιας εκτίμησης για τον θόρυβο που εμπεριέχεται στα δεδομένα. Το εξαγόμενο αποτέλεσμα του CN2 είναι ένα σύνολο διατεταγμένων «If A Then B» κανόνων, γνωστό και ως λίστα απόφασης (decision list) (Rivest, 1987). Αξίζει ακόμα να αναφέρουμε τον αλγόριθμο CL2, ο οποίος εξάγει κανόνες απόφασης χρησιμοποιώντας διαδικασίες ομαδοποίησης.

3.5.2 ΣΥΣΤΑΔΙΟΠΟΙΗΣΗ

Η συσταδιοποίηση ή αλλιώς ομαδοποίηση (clustering) αφορά τον διαχωρισμό (partition) των αντικειμένων μιας βάσης δεδομένων σε μη συνδεδεμένες μεταξύ τους και ομοιογενείς ομάδες, κατά τέτοιο τρόπο ώστε τα αντικείμενα του συνόλου που ανήκουν σε μια ομάδα, να είναι πιο όμοια μεταξύ τους, παρά με τα αντικείμενα που ανήκουν σε διαφορετικές ομάδες[50]. Ένα ιδιαίτερο χαρακτηριστικό της ομαδοποίησης, σε αντίθεση με την κατηγοριοποίηση, είναι ότι η δομή και το πλήθος των ομάδων είναι καταρχάς άγνωστα και καθορίζονται δε από τον εκάστοτε αλγόριθμο συσταδιοποίησης[51]. Αυτοί οι αλγόριθμοι βασίζονται στο σύνολο τους στην αρχή της μεγιστοποίησης της ομοιότητας ανάμεσα στα αντικείμενα την ίδιας ομάδας (intra-class similarity) και την ταυτόχρονη αρχή της ελαχιστοποίησης της ομοιότητας μεταξύ των αντικειμένων διαφορετικών ομάδων (inter-class similarity). Αξίζει να σημειωθεί ότι η ερμηνεία των ομάδων που προκύπτουν από την ανωτέρω διαδικασία καθορίζεται από τον εκάστοτε χρήστη[52].

Από τον παραπάνω ορισμό προκύπτει άμεσα και η βασική διαφορά μεταξύ κατηγοριοποίησης και συσταδιοποίησης. Στην κατηγοριοποίηση ο αριθμός και η ουσία των συστάδων αποτελεί πληροφορία εκ των προτέρων γνωστή. Εξαιτίας αυτού, στη συσταδιοποίηση εφαρμόζεται πάντα μη εποπτευόμενη μάθηση, εν αντιθέση με την κατηγοριοποίηση όπου λόγω της πρότερης γνώσης των κλάσεων κάνουμε χρήση της εποπτευόμενης μάθησης. Στην συσταδιοποίηση δεν υπάρχουν προκαθορισμένες κατηγορίες ομαδοποίησης αλλά οι εγγραφές συγκεντρώνονται σε ομάδες με βάση το κριτήριο που θέτει ο χρήστης για κάθε συστάδα όπως για παράδειγμα, η ομαδοποίηση πελατών που αγοράζουν παρόμοια αγαθά. Σκοπός είναι η δημιουργία συστάδων με όσο το δυνατόν περισσότερα κοινά χαρακτηριστικά εντός της εκάστοτε ομάδας, ενώ ταυτόχρονα η μία ομάδα από την άλλη θα πρέπει να διαφοροποιείται ικανοποιητικά ώστε να μη συγχέονται. Δηλαδή θα πρέπει να δημιουργηθούν διακριτές ομάδες με βάση ξεκάθαρα χαρακτηριστικά που περιγράφουν την κάθε ομάδα και την κάνουν να ξεχωρίζει από τις υπόλοιπες.

[50] Larose, H. D., (2004), MDI technologies, MIT Press
[51] Zaiane L. M. (1999), Technologies in Big Data, Oxford University Press
[52] Berry & Linoff, (2004), http://axon.cs.byu.edu/~martinez/classes/478/readings/DataPrep.pdf

Μερικά βασικά ζητήματα που προκύπτουν στην συσταδιοποίηση είναι τα παρακάτω:

-**Ο χειρισμός των ακραίων σημείων**: Πρόκειται στην ουσία για δεδομένα που στην πράξη δεν ανήκουν σε καμία συστάδα. Μπορούν από μόνα τους να θεωρηθούν ως ξεχωριστές συστάδες κάτι που καθιστά κάθε προσπάθεια συσταδιοποίησης φτωχή.

-**Τα δυναμικά δεδομένα**: Αυτά μπορεί να υπάρχουν σε βάσεις δεδομένων και τα οποία καθιστούν και τις ίδιες τις συστάδες δυναμικά μεταβαλλόμενες στο χρόνο.

-**Το είδος των δεδομένων που χρησιμοποιούνται**: Στη προκειμένη περίπτωση δεν είμαστε ακόμα σε θέση να έχουμε καμιά περαιτέρω πληροφορία όσον αφορά τα γνωρίσματα των ομάδων.

-**Η μοναδικότητα της λύσης του προβλήματος**: Πολλές φορές δε γίνεται να επιτευχθεί σε πολλά από τα προβλήματα συσταδιοποίησης, μιας και το ακριβές πλήθος των ομάδων που απαιτούνται δεν είναι και τόσο εύκολο να προσδιοριστεί.

Σε αυτό το σημείο πρέπει να αναφέρουμε ότι σημαντικό ρόλο παίζει και η απόσταση μεταξύ των συστάδων, η οποία θα πρέπει να ορίζεται καταλλήλως. Μια αντιπροσωπευτική λίστα μεθόδων υπολογισμού αποστάσεων είναι αυτή που περιγράφεται παρακάτω:

1) **Απόσταση απλού συνδέσμου**: Η μικρότερη απόσταση μεταξύ δύο στοιχείων των δύο συστάδων

2) **Απόσταση πλήρους συνδέσμου**: Η μεγαλύτερη απόσταση μεταξύ δύο στοιχείων των δύο συστάδων

3) **Μέση απόσταση**: Η μέση απόσταση μεταξύ των στοιχείων των δύο συστάδων

4) **Απόσταση κέντρων βάρους**: Η απόσταση μεταξύ των κέντρων βάρους των δύο συστάδων.

Η συσταδιοποίηση διακρίνεται σε τρεις βασικές μεθόδους:

1. Μέθοδοι διαχωρισμού (partitioning methods): Δημιουργούν ομάδες από ένα δεδομένο αρχικό σύνολο αντικειμένων με κάθε ομάδα να αντιπροσωπεύει ένα

cluster και να ικανοποιούνται οι εξής δύο συνθήκες: (α) κάθε cluster περιέχει τουλάχιστον ένα αντικείμενο και (β) κάθε αντικείμενο ανήκει σε ένα μόνο cluster.

2. Ιεραρχικές μέθοδοι (hierarchical methods): Διασπούν το αρχικό σύνολο δεδομένων δημιουργώντας μια ιεραρχική δομή από clusters και διακρίνονται σε agglomerative (bottom-up) ή divisive (top-down) ανάλογα με τον τρόπο που γίνεται η διάσπαση.

3. Μέθοδοι βασισμένες σε μοντέλα (model-based methods): Υποθέτουν ότι καθένα από τα clusters περιγράφεται από ένα μαθηματικό μοντέλο και εντοπίζουν τα αντικείμενα που ανήκουν σε κάθε cluster, ώστε να ικανοποιούν το αντίστοιχο μοντέλο.

Η επιλογή του κριτηρίου για μία σωστή διαδικασία συσταδοποίησης απαιτεί:

- **Επιλογή χαρακτηριστικών γνωρισμάτων**: Ο στόχος είναι να επιλεγούν τα καταλληλότερα γνωρίσματα στα οποία πρόκειται να εφαρμοστεί η συσταδοποίηση ώστε να επιτυγχάνεται η βέλτιστη ομοιογένεια σε κάθε συστάδα. Έτσι η προεπεξεργασία των δεδομένων πριν την εφαρμογή της διαδικασίας συσταδοποίησης κρίνεται απαραίτητη.

-**Επιλογή αλγορίθμων συσταδοποίησης**: Σε αυτό το στάδιο γίνεται η επιλογή ενός αλγορίθμου που θα οδηγήσει σε ένα καλό σχήμα συσταδοποίησης για ένα σύνολο δεδομένων. Για τη επιλογή του αλγορίθμου χρησιμοποιείται το μέτρο γειτνίασης και το κριτήριο συσταδοποίησης τα οποία ορίζουν απόλυτα τον αλγόριθμο, καθώς επίσης και η δυνατότητά του να καθορίσει ένα σχήμα συσταδοποίησης που να προσαρμόζεται στο συγκεκριμένο σύνολο δεδομένων

Οι αλγόριθμοι συσταδιοποίησης μπορούν να ταξινομηθούν στις ακόλουθες κατηγορίες:

-**Επικύρωση αποτελεσμάτων**: Σε αυτή τη φάση αξιολογούνται τα αποτελέσματα του αλγορίθμου συσταδοποίησης σύμφωνα με κατάλληλα κριτήρια ορθότητας συσταδοποίησης και τεχνικές. Παράδειγμα ενός τέτοιου κριτηρίου είναι η σύγκριση των αποτελεσμάτων της ανάλυσης με κάποια ήδη γνωστά αποτελέσματα ή η

σύγκριση των αποτελεσμάτων δύο διαφορετικών συσταδοποιήσεων. Η ποιότητα της συσταδοποίησης εξαρτάται από την ομοιότητα (δηλαδή μεγάλη ομοιότητα εντός της συστάδας - μικρή ομοιότητα μεταξύ των συστάδων) και την μέθοδο υλοποίησης της συσταδοποίησης.

-**Ερμηνεία των αποτελεσμάτων**: Αποτελεί το τελευταίο στάδιο της διαδικασίας συσταδοποίησης, όπου οι αναλυτές καλούνται να εξάγουν γνώση από τις παραχθείσες συστάδες, συνδυάζοντας κι άλλα στοιχεία, αναλύσεις, με σκοπό το καλύτερο και εγκυρότερο αποτέλεσμα. Μια μέθοδος συσταδοποίησης είναι καλή αν παράγει συστάδες καλής ποιότητας δηλαδή συστάδες με μεγάλη ομοιότητα εντός της συστάδας.

3.5.3 ΑΝΑΛΥΣΗ ΣΥΣΧΕΤΙΣΗΣ

Η ανάλυση συσχέτισης (association analysis) έχει σαν βασικό της στόχο την ανακάλυψη κρυμμένων συσχετίσεων μεταξύ των χαρακτηριστικών μιας βάσης δεδομένων. Με άλλα λόγια, η παραπάνω ανάλυση ψάχνει να βρει κανόνες για την ποσοτικοποίηση των σχέσεων μεταξύ δύο ή περισσοτέρων χαρακτηριστικών μιας βάσης δεδομένων[53]. Οι κανόνες αυτοί ονομάζονται κανόνες συσχέτισης (association rules), και έχουν την μορφή «If A then B»[54]. Οι κανόνες συσχέτισης χαρακτηρίζονται από το κατώφλι στήριξης (support threshold), που αναγνωρίζει τα στοιχεία των βάσεων δεδομένων που εμφανίζονται συχνά σε αυτά, καθώς και το κατώφλι εμπιστοσύνης (confidence threshold), που είναι η υπό συνθήκη πιθανότητα (conditional probability) ένα στοιχείο να εμφανίζεται σεμια διαδικασία όταν ένα άλλο στοιχείο εμφανίζεται επίσης[55]. Αξίζει να σημειωθεί ότι η ανάλυση συσχέτισης είναι γνωστή στον επιχειρηματικό κόσμο σαν ανάλυση συνάφειας (affinity analysis) με πολλές εφαρμογές[56].

[53] Larose, 2004, ο.π.
[54] Agrawal et al, 1996 ο.π.
[55] Zaiane, 1999, ο.π.
[56] Berry & Linoff, (2004), http://axon.cs.byu.edu/~martinez/classes/478/readings/DataPrep.pdf

3.5.4 ΠΑΛΙΝΔΡΟΜΗΣΗ

Η παλινδρόμηση (regression) είναι η παλαιότερη και η πλέον γνωστή στατιστική τεχνική που υλοποιείται εντός των πλαισίων της εξόρυξης δεδομένων. Κύριος σκοπός εδώ είναι η πρόβλεψη της τιμής μιας μεταβλητής μελετώντας τις τιμές που είχε στο παρελθόν. Συγκεκριμένα, η παλινδρόμηση χρησιμοποιώντας μια βάση αριθμητικών δεδομένων, αναπτύσσει μια μαθηματική σχέση που ταιριάζει στα δεδομένα αυτά. Στην συνέχεια, η μαθηματική αυτή σχέση χρησιμοποιείται για την πρόβλεψη μελλοντικής συμπεριφοράς, εφαρμόζοντας σε αυτήν νέα αριθμητικά δεδομένα. Ο βασικός περιορισμός της συγκεκριμένης τεχνικής είναι ότι εφαρμόζεται καλά μόνο σε συνεχή ποσοτικά δεδομένα (βάρος, ταχύτητα ή ηλικία). Αντίθετα, η παλινδρόμηση δεν λειτουργεί καλά με κατηγορικά δεδομένα [57].

3.5.4.1 Η ΕΠΙΔΡΑΣΗ ΤΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΚΑΙ ΟΙ ΠΑΡΕΡΜΗΝΕΙΕΣ ΣΤΙΣ ΟΠΟΙΕΣ ΟΔΗΓΕΙ

Η πρώτη προσπάθεια για τη μελέτη της σχέσης μεταξύ δύο μεταβλητών έγινε από τον Sir Francis Galton για την μελέτη της σχέσης του ύψους των παιδιών με τους γονείς τους. Από την μελέτη αυτή προήλθε και ο όρος παλινδρόμηση (regression) που ουσιαστικά αναφέρεται στην παλινδρόμηση προς την κατεύθυνση του μέσου (regression towards the mean). Ο όρος προήλθε από την παρατήρηση του Galton ότι υπάρχει μια τάση όπου ακραίες, ως προς το μέσο τους, παρατηρήσεις της ανεξάρτητης τιμής αντιστοιχούν σε παρατηρήσεις της εξαρτημένης τιμής που δεν είναι το ίδιο ακραίες αλλά είναι πλησιέστερα προς τον μέσο τους. Με απλούστερο τρόπο μπορεί να πει κανείς ότι ακραίες παρατηρήσεις ακολουθούνται από λιγότερο ακραίες παρατηρήσεις δηλαδή αυτές που είναι πλησιέστερα προς το κέντρο. Αυτό κάνει το διάγραμμα σημείων να έχει την μορφή μπάλας του Αμερικάνικου ποδοσφαίρου. Μελετώντας αρχεία για οικογένειες, τα οποία αγόρασε, ο Galton συγκέντρωσε τα ύψη 205 ζευγαριών από γονείς και 928 ενήλικα παιδιά των γονέων αυτών. Δοθέντος ότι το μέσο ύψος των ανδρών είναι, περίπου, 8% μεγαλύτερο από ότι το μέσο ύψος των γυναικών, ο Galton πολλαπλασίασε τα ύψη των γυναικών στο δείγμα του με το συντελεστή 1.08, έτσι ώστε τα ύψη αυτά των γυναικών να γίνουν

[57] Draper & Smith, 1997

συγκρίσιμα με τα ύψη των ανδρών του δείγματος. Στη συνέχεια, για το μέσο ύψος κάθε ζευγαριού γονέων, υπολογίστηκε ο μέσος όρος έτσι ώστε να βρεθεί ένα μέσο ύψος γονέων. Τα μέσα ύψη γονέων διαιρέθηκαν στη συνέχεια σε εννέα διαστήματα. Για κάθε κατηγορία μέσου ύψους γονέων υπολογίστηκε το διάμεσο ύψος των παιδιών των γονέων που ανήκαν στην κατηγορία αυτή. Από την μελέτη των δεδομένων ο Galton παρατήρησε ότι, ασυνήθιστα υψηλοί γονείς τείνουν να έχουν παιδιά χαμηλότερα από τους ίδιους ενώ, ασυνήθιστα χαμηλοί γονείς έχουν συνήθως υψηλότερα παιδιά. Το ύψος κάθε ανθρώπου επηρεάζεται από τα γονίδια που κληρονομεί από τους γονείς του. Για διευκόλυνση της παρουσίασης, ας χαρακτηρίσουμε κάποιον ο οποίος κατά τη στιγμή της σύλληψης έχει προβλεπόμενο ύψος ενηλικίωσης με βάση τα γονίδιά του 1.72, ως ένα άτομο γονιδιακού ύψους 1.72. Δεδομένου ότι το ύψος των ανθρώπων επηρεάζεται από την διατροφή, την άσκηση, και άλλους περιβαλλοντικούς παράγοντες, το ύψος που θα έχει κάποιος στην ενηλικίωσή του δεν θα αντικατοπτρίζει με ένα τέλειο τρόπο την επίδραση των γονιδίων και επομένως δεν θα αποτελεί μία πλήρη επαλήθευση του προβλεφθέντος με βάση τα γονίδιά του ύψος κατά την παιδική ηλικία. Ένα άτομο πραγματικού ύψους 1.75 ίσως είχε ένα γονιδιακά προβλεφθέν ύψος 1.72, με την διαφορά πραγματικού και προβλεφθέντος ύψους οφειλόμενη σε θετική επίδραση περιβαλλοντικών παραγόντων. Αντίθετα, κάποιος με προβλεφθέν γονιδιακό ύψος 1.78 μπορεί να έχει πραγματικό ύψος στην ενηλικίωση 1.75 εξαιτίας αρνητικών επιδράσεων περιβαλλοντικών παραγόντων. Η πρώτη περίπτωση συμβαίνει συχνότερα απ' ότι η δεύτερη γι' αυτό και τα παρατηρούμενα ύψη παιδιών εξαιρετικά υψηλών γονέων αποτελούν, συνήθως, μια υπέρβαση των γονιδιακών υψών των παιδιών αυτών[58].

Η προηγηθείσα επιχειρηματολογία δεν συνεπάγεται ότι όλοι οι άνθρωποι θα έχουν σε κάποια μελλοντική στιγμή το ίδιο ύψος. Αν συνέβαινε κάτι τέτοιο θα μπορούσε κανείς να αντιστρέψει την επιχειρηματολογία παρατηρώντας ότι πάρα πολύ υψηλοί άνθρωποι έχουν γονείς κάπως χαμηλότερους από αυτούς ενώ πάρα πολύ χαμηλοί άνθρωποι έχουν κάπως υψηλότερους γονείς. Μήπως αυτό συνεπάγεται ότι τα ύψη των ανθρώπων αποκλίνουν; Ούτε το ένα συμβαίνει ούτε το άλλο. Τα ύψη των

[58] Thomas Smith, Galton and Regression, Statistic Review, 3:4 1998

ανθρώπων ούτε συγκλίνουν ούτε αποκλίνουν. Θα υπάρχουν πάντοτε εξαιρετικά υψηλοί και εξαιρετικά χαμηλοί άνθρωποι. Αυτό που θα πρέπει να αντιληφθούμε είναι ότι τα ύψη των ανθρώπων επηρεάζονται από τυχαίους παράγοντες και ότι, για ανθρώπους που είναι εξαιρετικά υψηλοί οι τυχαίοι παράγοντες επηρέασαν θετικά το ύψος τους και το έκαναν μεγαλύτερο από ότι αναμενόταν με βάση το γονιδίαμα τους. Η παρερμηνεία αυτή είναι μια λανθασμένη συλλογιστική, και οφείλεται στο φαινόμενο της παλινδρόμησης προς την κατεύθυνση του μέσου, (regression towards the mean) είναι δε ακριβώς η παρερμηνεία της προσωρινής φύσης μιας ακραίας παρατήρησης και ο χαρακτηρισμός της ως τάσης. Η κατάσταση που προκύπτει αποδίδεται στην επίδραση της παλινδρόμησης (regression effect).

3.5.4.2 ΠΑΡΑΔΕΙΓΜΑΤΑ ΕΠΙΔΡΑΣΗΣ ΤΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗ ΣΕ ΔΙΑΦΟΡΟΥΣ ΤΟΜΕΙΣ.

1.**Τεστ ευφυΐας**.

Σε πολλά σχολεία στο εξωτερικό και κυρίως στις Η.Π.Α, διαμορφώνονται προσχολικά προγράμματα για να ενισχύσουν το IQ των παιδιών. Τα παιδιά που συμμετέχουν στο πρόγραμμα κάνουν το IQ τεστ όταν ξεκινούν το πρόγραμμα και το επαναλαμβάνουν όταν ολοκληρώσουν το πρόγραμμα. Και στις δύο περιπτώσεις τα αποτελέσματα είναι γύρω στο 100 με τυπική απόκλιση γύρω στο 15. Τα στοιχεία δείχνουν ότι τέτοια προγράμματα δεν έχουν κάποιο ιδιαίτερο αποτέλεσμα. Μια παρατήρηση όμως που μπορεί να κάνει κάποιος που θα κοιτάξει περισσότερο τα δεδομένα δείχνει κάτι που προκαλεί έκπληξη. Τα παιδιά τα οποία είχαν απόδοση κάτω από το μέσο στο τεστ πριν αρχίσουν το πρόγραμμα πέτυχαν μία μέση βελτίωση περίπου 5 μονάδων στο τεστ που έδωσαν στο τέλος του προγράμματος. Αντιστρόφως όμως, τα παιδιά εκείνα που απέδωσαν πάνω από το μέσο όρο στο αρχικό τεστ έχασαν, κατά μέσο όρο, περίπου 5 μονάδες στο τελικό τεστ. Θα μπορούσε κανείς να οδηγηθεί στο συμπέρασμα ότι το πρόγραμμα αυτό οδηγεί τελικά σε εξισορρόπηση της ευφυΐας των παιδιών; Ή ότι τα ευφυέστερα παιδιά, επειδή παίζουν με παιδιά μικρότερης ευφυΐας, καταλήγουν να οδηγούν τις δύο αυτές κατηγορίες στην ίδια κατάσταση και οι διαφορές να εξαφανίζονται; Φυσικά δεν συμβαίνει τίποτα από αυτά. Και εδώ έχουμε τη χαρακτηριστική περίπτωση του

φαινομένου της επίδρασης της παλινδρόμησης (regression effect) σύμφωνα με το οποίο, σε όλες τις περιπτώσεις εξετάσεων, το γκρουπ με τη χαμηλότερη απόδοση σε μια πρώτη εξέταση, κατά μέσο όρο, θα αποδώσει καλύτερα σε μια δεύτερη εξέταση και το γκρουπ με την υψηλότερη απόδοση, κατά μέσο όρο, θα αποδώσει χαμηλότερα σε μια δεύτερη εξέταση. Η εσφαλμένη αντίληψη ότι η επίδραση της παλινδρόμησης οφείλεται σε κάτι σημαντικό και όχι απλώς στην διάχυση (spread) των παρατηρήσεων γύρω από την γραμμή είναι αυτό που ονομάζεται παρερμηνεία της παλινδρόμησης (regression fallacy)[59].

Μια άλλη διάσταση της παλινδρόμησης προς την κατεύθυνση του μέσου στα τεστ ευφυΐας (IQ tests) είναι η εξής: Σύμφωνα με μία μελέτη που έγινε στην Αμερική παιδιά ηλικία τεσσάρων ετών με IQ 120 συνήθως, όταν ενηλικιωθούν, επιτυγχάνουν σκορ στο IQ τεστ περίπου 110. Παρομοίως, παιδιά τεσσάρων ετών με IQ σκορ 70 έχουν ένα μέσο σκορ στο IQ τεστ όταν ενηλικιωθούν 85. Αυτό δεν συνεπάγεται ότι θα υπάρχουν λιγότεροι ενήλικες απ' ότι παιδιά με πολύ υψηλά ή πολύ χαμηλά αποτελέσματα στο IQ τεστ. Παρότι όσοι άνθρωποι ξεκινούν στην παιδική ηλικία με υψηλό ή χαμηλό IQ σκορ, συνήθως, θα παλινδρομήσουν προς την κατεύθυνση του μέσου, οι θέσεις τους θα παρθούν (θα αντικατασταθούν) από άλλους οι οποίοι στην παιδική τους ηλικία θα έχουν IQ σκορ πλησιέστερα προς τον μέσο.

2.Εκπαίδευση.

Ένα παράδειγμα λανθασμένης ερμηνείας φαινομένων που οφείλονται στην παλινδρόμηση προς την κατεύθυνση του μέσου εμφανίζεται στην αξιολόγηση των φοιτητών. Έχει παρατηρηθεί ότι οι φοιτητές εκείνοι οι οποίοι έχουν τους υψηλότερους βαθμούς στις εξετάσεις προόδου συνήθως, δεν αποδίδουν εξίσου καλά στην τελική εξέταση ενώ, εκείνοι οι οποίοι έχουν χαμηλή βαθμολογία στην εξέταση προόδου, πολλές φορές βελτιώνουν την απόδοσή τους στην τελική εξέταση. Θα μπορούσε αυτό να εκληφθεί ως ένδειξη ότι η απόδοση των φοιτητών συγκλίνει προς μια ανησυχητική μετριότητα με τους ασθενείς φοιτητές να βελτιώνονται και τους καλούς φοιτητές να χειροτερεύουν ή αντιστρέφοντας το

[59] Thomas Smith, Galton and Regression, Statistic Review, 3:4 1998

προηγούμενο επιχείρημα, το γεγονός ότι αυτοί που πέτυχαν την υψηλότερη βαθμολογία στην τελική εξέταση δεν απέδωσαν εξίσου καλά στην εξέταση προόδου σημαίνει ότι η απόδοση αποκλίνει από τον μέσο; Και στις δύο περιπτώσεις η απάντηση είναι αρνητική. Η εξαιρετικά υψηλή απόδοση σε οποιαδήποτε εξέταση εμπεριέχει και έναν παράγοντα καλής τύχης ενώ η χαμηλή απόδοση έναν παράγοντα ατυχίας. Οι φοιτητές εκείνοι που πέτυχαν την υψηλότερη βαθμολογία σε οποιαδήποτε εξέταση είναι, κυρίως, φοιτητές πάνω από το μέσο όρο που πέτυχαν εξαιρετικά υψηλή βαθμολογία γιατί τα θέματα των εξετάσεων ήταν θέματα που, εξαιτίας της καλής προετοιμασίας τους, είχαν την ευχέρεια να απαντήσουν. Είναι περισσότερο πιθανό ότι οι φοιτητές αυτοί είναι καλοί φοιτητές που απέδωσαν εξαιρετικά καλά από το ενδεχόμενο να ήταν εξαιρετικά καλοί φοιτητές που είχαν μια άσχημη μέρα. Όσοι επιτυγχάνουν τις υψηλότερες βαθμολογίες σε μια οποιαδήποτε εξέταση είναι πολύ πιθανό ότι δεν απέδωσαν εξίσου καλά στην προηγούμενη εξέταση και δεν θα αποδώσουν το ίδιο καλά στην επόμενη εξέταση. Η παλινδρόμηση προς την κατεύθυνση του μέσου μπορεί να θεωρηθεί κι ως μια περίπτωση κακής χρήσης διαθέσιμων δεδομένων. Αν για την αξιολόγηση φοιτητών σε μια εξέταση επιλέξουμε με τυχαίο τρόπο φοιτητές, η μέση βαθμολογία τους θα αποτελεί μια αμερόληπτη εκτίμηση του μέσου του πληθυσμού. Εάν όμως, μετά την εξέταση, ξεχωρίσουμε τους φοιτητές εκείνους που απέδωσαν εξαιρετικά καλά, αυτοί βέβαια δεν αποτελούν ένα τυχαίο δείγμα, αφού έχουν επιλεγεί ακριβώς επειδή είχαν τις υψηλότερες βαθμολογίες. Σε οποιοδήποτε δείγμα οι υψηλότερες τιμές αποτελούν μια υπερεκτίμηση (overestimate) του μέσου του πληθυσμού. Για να έχουμε αμερόληπτες εκτιμήσεις θα πρέπει να έχουμε ένα τυχαίο δείγμα που δεν στηρίζεται στα αποτελέσματα αυτά καθαυτά.

3.Στρατιωτικό.

Ένας εκπαιδευτής πιλότων παρατήρησε ότι πολύ καλές προσγειώσεις συνήθως, ακολουθούνται από προσγειώσεις που δεν είναι εξίσου καλές, ενώ μέτριες προσγειώσεις ακολουθούνται, συνήθως από καλύτερες. Υποπίπτοντας στην λανθασμένη προσέγγιση που οφείλεται στην παρερμηνεία της παλινδρόμησης στην κατεύθυνση του μέσου ο εκπαιδευτής ισχυρίστηκε ότι η ακολουθία αυτή συμβαίνει

γιατί συνήθιζε να επαινεί τις καλές προσγειώσεις και να κριτικάρει έντονα τις μέτριες. Για το λόγο αυτό έβγαλε το συμπέρασμα, σε αντίθεση από την κοινά αποδεκτή άποψη με βάση την έρευνα για την μαθησιακή διδασκαλία, ότι ο έπαινος έχει αρνητικά αποτελέσματα στην προσπάθεια ενώ η έντονη κριτική έχει θετικά αποτελέσματα.

4.Οικονομικό:

Ένα χαρακτηριστικό παράδειγμα του προβλήματος στον τομέα των οικονομικών δίνεται στο βιβλίο με τον προκλητικό τίτλο Ο Θρίαμβος της Μετριότητας στις Επιχειρήσεις (The Triumph of Mediocrity in Business)[60]. Ο συγγραφέας ανακάλυψε ότι επιχειρήσεις με εξαιρετικά υψηλά κέρδη σε κάθε δεδομένη χρονιά έχουν χαμηλότερα κέρδη την επόμενη χρονιά ενώ επιχειρήσεις με πολύ χαμηλά κέρδη, εν γένει επιτυγχάνουν καλύτερα αποτελέσματα το επόμενο έτος. Με αυτές τις ενδείξεις κατέληξε στο συμπέρασμα ότι οι ισχυρές επιχειρήσεις γίνονται ασθενέστερες ενώ οι ασθενείς γίνονται ισχυρότερες με αποτέλεσμα σύντομα να γίνουν όλες οι επιχειρήσεις μεσαίου μεγέθους. Η τελείως λανθασμένη προσέγγιση του συγγραφέα είναι προφανής. Ο διάσημος στατιστικός Harold Hotelling εξήγησε το λάθος αυτό ως εξής: «Οι αποδόσεις των επιχειρήσεων με ακραίες αποδόσεις τείνουν, συχνά, προς την κατεύθυνση του κέντρου ενώ εκείνες με μεσαίες αποδόσεις σε ένα σύνολο τείνουν προς τα άκρα. Μερικές, βελτιώνουν την απόδοσή τους ενώ άλλες χειροτερεύουν. Ο μέσος των κερδών του αρχικού συνόλου των επιχειρήσεων που βρισκόταν στο κέντρο είναι ενδεχόμενο, επομένως, να επιδείξει κάποια μικρή μεταβολή δοθέντος ότι, θετικές και αρνητικές αποκλίσεις ακυρώνονται στην διαδικασία υπολογισμού του μέσου, ενώ για ένα σύνολο με ακραίες αποδόσεις η μόνη δυνατή κίνηση είναι προς την κατεύθυνση του κέντρου».

[60] Horace Secrist, 1985, The Triumph of Mediocrity in Business, Northwestern Univerisity

4 Λογισμικό data mining Ανοικτού Κώδικα

Σκοπός του κεφαλαίου είναι η μελέτη και η παρουσίαση λογισμικών ανοιχτού κώδικα για το χώρο του data miming. Τα προγράμματα αυτά είναι τα παρακάτω: Carrot2, Orange, RapidMiner, Weka, Rattle και Tanagra. Λογισμικό Ανοιχτού Κώδικα είναι το λογισμικό που είναι διαθέσιμο σε μορφή πρωτογενούς κώδικα: ο πρωτογενής κώδικας και κάποια άλλα δικαιώματα που συνήθως επιφυλάσσονται για τους κατόχους των πνευματικών δικαιωμάτων, παρέχονται στο πλαίσιο μίας ανοιχτού κώδικα άδειας χρήσης που επιτρέπει στους χρήστες να μελετήσουν, αλλάξουν, βελτιώσουν και πολλές φορές να διανείμουν το λογισμικό[61].

Για να υιοθετήσουμε ένα λογισμικό ανοιχτού κώδικα (ή οποιοδήποτε άλλο λογισμικό) πρέπει να κατανοήσουμε την άδεια χρήσης και τους περιορισμούς που μας θέτει. Σε αντίθεση με άδειες κλειστού κώδικα, που έχουν ως στόχο να περιορίσουν τα δικαιώματα του χρήστη, το λογισμικό ανοιχτού κώδικα δίνει τη δυνατότητα χρήσης του λογισμικού, όπως επιθυμεί ο εκάστοτε χρήστης. Οι πιο συνηθισμένες άδειες χρήσης λογισμικών ανοιχτού κώδικα περιλαμβάνουν τα GPL, LGPL, BSD, NPL και MPL[62].

Τις περασμένες δεκαετίες, προϊόντα ανοιχτού κώδικα όπως το GNU/Linux, Apache, BSD, MySQL και OpenOffice κατόρθωσαν μεγάλη επιτυχία, αποδεικνύοντας ότι ένα λογισμικό ανοιχτού κώδικα μπορεί να είναι τόσο εύρωστο, ή και πιο πολύ, όσο ένα εμπορικό, κλειστού κώδικα λογισμικό[63]. Το λογισμικό ανοιχτού κώδικα αντιπροσωπεύει μια νέα τάση στο τομέα της εξόρυξης δεδομένων, στην εκπαίδευση και στις βιομηχανικές εφαρμογές, ιδιαίτερα σε μικρές και μεγάλες επιχειρήσεις. Με λογισμικά ανοιχτού κώδικα, μία επιχείρηση μπορεί εύκολα να ξεκινήσει ένα έργο εξόρυξης δεδομένων, χρησιμοποιώντας τη πιο πρόσφατη τεχνολογία[64].

---

[61]

[62] Xiaojun Chen et al. (2007), “A Survey of Open Source Data Mining Systems” - URL: http://togaware.redirectme.net/papers/pakdd07.pdf

[63] Xiaojun Chen et al. (2007), “A Survey of Open Source Data Mining Systems” - URL: http://togaware.redirectme.net/papers/pakdd07.pdf

[64] Xiaojun Chen et al. (2007), “A Survey of Open Source Data Mining Systems” - URL: http://togaware.redirectme.net/papers/pakdd07.pdf

4.1 Λογισμικά προς ανάλυση

Σε αυτή την εργασία θα αναλύσουμε εφτά λογισμικά ανοιχτού κώδικα για τον τομέα της εξόρυξης δεδομένων και θα περιγράψουμε τη διεπιφάνεια χρήστη και χρήση τους, τα οποία είναι τα ακόλουθα: Carrot2, Orange, RapidMiner, Weka, Rattle και Tanagra. Θα αναλύσουμε τα χαρακτηριστικά, τη λειτουργικότητα τους, τη δυνατότητα πρόσβασης σε δεδομένα και τη λειτουργικότητα εξόρυξης δεδομένων, συμπεριλαμβανομένων της διεπιφάνειας χρήστη και επεκτασιμότητας τους.

4.1.1 Carrot2

Το Carrot² είναι μία ανοιχτού κώδικα μηχανή αναζήτησης για συσταδοποίηση αποτελεσμάτων (data clustering). Μπορεί αυτόματα να συσταδοποιήσει μικρές συλλογές εγγράφων, όπως για παράδειγμα αποτελέσματα αναζήτησης ή αποσπάσματα εγγράφων, σε θεματικές κατηγορίες. Πλην των δύο ειδικευμένων αλγορίθμων συσταδοποίησης αποτελεσμάτων αναζήτησης, το Carrot2 προσφέρει έτοιμα προς χρήση συστατικά (components) για την απόκτηση αποτελεσμάτων αναζήτησης από διάφορες πηγές. Το Carrot2 είναι γραμμένο σε Java και διανέμεται από τη BSD[65].

Η αρχική έκδοση του Carrot2 υλοποιήθηκε το 2001 από τον David Weiss ως μέρος του MSc του, για να επικυρώσει την εφαρμογή του αλγόριθμου STC (Suffix Tree Clustering) σε συσταδοποιημένα αποτελέσματα αναζήτησης στα πολωνικά. Το 2003, προστέθηκαν κάποιοι ακόμα αλγόριθμοι συσταδοποίησης αποτελεσμάτων αναζήτησης, συμπεριλαμβανομένου του Lingo, ενός αλγόριθμου ειδικά σχεδιασμένου για τη συσταδοποίηση αποτελεσμάτων αναζήτησης. Ενώ ο πηγαίος κώδικας του Carrot2 ήταν διαθέσιμος από το 2002, ήταν το 2006 όταν η έκδοση 1.0 κυκλοφόρησε επίσημα. Την ίδια χρονιά, κυκλοφόρησε και η έκδοση 2.0, με βελτιωμένη διεπιφάνεια χρήστη και εκτεταμένη σειρά εργαλείων. Το 2009, με την έκδοση 3.0 έγιναν και σημαντικές βελτιώσεις στην ποιότητα συσταδοποίησης, πιο απλοποιημένη διεπιφάνεια προγραμματισμού εφαρμογών (API-application programming interface) και νέα εφαρμογή διεπιφάνειας χρήστη (GUI-graphical user

[65] URL: http://en.wikipedia.org/wiki/Carrot2

interface) για συντονισμένη συσταδοποίηση σύμφωνα με την πλατφόρμα Eclipse Rich Client[66] [3].

4.1.1.1 Εφαρμογές

Το Carrot² μπορεί να κληθεί μέσω πολλών APIs, όπως [3]: ¬ Java API ¬ C# / .NET API ¬ Άλλες πλατφόρμες 1.4.1.2 Υποστηριζόμενα Εργαλεία Το Carrot² προσφέρει ένα αξιόλογο αριθμό εργαλείων που μπορούν να χρησιμοποιηθούν για γρήγορη εφαρμογή συσταδοποίησης σε δεδομένα, περαιτέρω συντονισμό αποτελεσμάτων καθώς και χρήση της μηχανής Carrot² για εφαρμογή συσταδοποίησης σαν απομακρυσμένη υπηρεσία [3]: ¬ Carrot2 Document Clustering Workbench ¬ Carrot2 Document Clustering Server ¬ Carrot2 Command Line Interface ¬ Carrot2 Web Application

4.1.1.2 Απαιτήσεις Λογισμικού

Όλες οι εφαρμογές Carrot2 απαιτούν περιβάλλον Java έκδοσης 1.6.0 ή μεταγενέστερης. Ο Carrot2 Document Clustering Workbench διανέμεται για εκδόσεις Windows, Linux 32-bit και 64-bit και για Mac OS x86[67]. Το πακέτο Carrot2 C# API απαιτεί το .NET Framework έκδοσης 3.5 ή μεταγενέστερης και δεν χρειάζεται περιβάλλον Java

4.1.2 Orange

Η Orange είναι μία βιβλιοθήκη αντικειμένων πυρήνα και ρουτινών της C++ και περιλαμβάνει μία μεγάλη ποικιλία κάποιων βασικών και κάποιων όχι τόσο βασικών αλγορίθμων μηχανικής εκμάθησης και εξόρυξης δεδομένων. Επιπλέον, περιέχει ρουτίνες για εισαγωγή και χειρισμό δεδομένων.

Επίσης, το περιβάλλον της επιτρέπει τη δημιουργία κώδικα για γρήγορη προτυποποίηση νέων αλγορίθμων και έλεγχο συστημάτων. Είναι μία συλλογή από

[66] URL: http://en.wikipedia.org/wiki/Carrot2
[67] Ο.π.

υπό-προγράμματα (modules) Python, τα οποία βρίσκονται στη βασική βιβλιοθήκη και υλοποιούν κάποια λειτουργία για την οποία ο χρόνος εκτέλεσης δεν είναι σημαντικός και που γίνεται πιο εύκολα με Python παρά με C++.

Η Orange συμπεριλαμβάνει επίσης ένα μεγάλο αριθμό widget (γραφικά συστατικά) γραφικών που χρησιμοποιούν μεθόδους της βιβλιοθήκης πυρήνα και των υπό-προγραμμάτων της Orange. Με τη χρήση του οπτικού προγραμματισμού, τα widget μπορούν να συγκεντρωθούν σε μία εφαρμογή με ένα εργαλείο οπτικού προγραμματισμού που ονομάζεται Orange Canvas.

Όλα αυτά μαζί συνθέτουν την Orange, ένα περιεκτικό, βασισμένο σε συστατικά πλαίσιο για μηχανική εκμάθηση και εξόρυξη δεδομένων, προορισμένο για έμπειρους χρήστες και ερευνητές στη μηχανική εκμάθηση που θέλουν να αναπτύξουν τους δικούς τους αλγόριθμους χρησιμοποιώντας όσο το δυνατόν περισσότερο κώδικα γίνεται, αλλά και για αρχάριους, που μπορούν να απολαύσουν ένα παντοδύναμο, αλλά ταυτόχρονα ευκολόχρηστο περιβάλλον οπτικού προγραμματισμού[68].

Η Orange παρέχει ένα ποικιλόμορφο περιβάλλον για τους υπεύθυνους ανάπτυξης, ερευνητές και τους εξασκούντες στην εξόρυξη δεδομένων. Χάρη στη Python, μία γλώσσα συγγραφής σεναρίων (scripting language) νέας γενιάς και περιβάλλοντος προγραμματισμού, τα σενάρια (scripts) σας για την εξόρυξη δεδομένων είναι μεν απλά, αλλά ισχυρά. Για ακόμη πιο γρήγορη προτυποποίηση, η Orange υιοθετεί μία προσέγγιση βασισμένη σε συστατικά: η μέθοδος ανάλυσης μπορεί να υλοποιηθεί είτε σαν απλή στοίβαξη τούβλων LEGO, είτε με τη χρήση ενός υπάρχοντος αλγορίθμου και αντικατάσταση κάποιων βασικών συστατικών του με άλλα, καινούργια. Ότι είναι τα συστατικά της Orange για τη συγγραφή σεναρίων, είναι και τα widgets της Orange για τον οπτικό προγραμματισμό. Τα widgets χρησιμοποιούν ένα ειδικά σχεδιασμένο μηχανισμό επικοινωνίας για μεταβαλλόμενα αντικείμενα όπως σύνολα δεδομένων, λίστες χαρακτηριστικών, τεχνικές εκμάθησης, ταξινομητές, και άλλα, επιτρέποντας την εύκολη κατασκευή αρκετά πολύπλοκων

[68]Demsar J, Zupan B (2004), “Orange: From Experimental Machine Learning to Interactive Data Mining”, White Paper (www.ailab.si/orange), Faculty of Computer and Information Science, University of Ljubljana - URL: http://orange.biolab.si/wp/orange-leaflet-visual.pdf

σχημάτων εξόρυξης δεδομένων, χρησιμοποιώντας προσεγγίσεις και τεχνικές τελευταίας τεχνολογίας. Η βασική αρχή στην Orange δεν είναι να καλύψει κάθε μέθοδο και άποψη στη μηχανική εκμάθηση και εξόρυξη δεδομένων, αλλά να καλύψει σε βάθος και σχολαστικά αυτές που υλοποιούνται, δημιουργώντας τες από επαναχρησιμοποιήσιμα συστατικά τα οποία έμπειροι χρήστες μπορούν να αλλάξουν ή και να αντικαταστήσουν με καινούργια

4.1.2.1 Ιστορικό ¬

- Το 1996, το Πανεπιστήμιο της Λιουμπλιάνα και το Ινστιτούτο Josef Stefan άρχισε την ανάπτυξη του ML*, ενός πλαισίου μηχανικής εκμάθησης σε C++.

- Το 1997, αναπτύχθηκαν σύνδεσμοι Python για το ML*, που μαζί με Python υπό-προγράμματα σχημάτισαν ένα κοινό πλαίσιο, την Orange. ¬ Κατά τη διάρκεια των επόμενων χρόνων, οι πιο σημαντικοί αλγόριθμοι εξόρυξης δεδομένων και μηχανικής εκμάθησης αναπτύχθηκαν είτε σε C++ (για ταχύτητα) είτε σε Python (για ευελιξία).

¬ Το 2002, σχεδιάστηκαν τα πρώτα πρότυπα για τη δημιουργία μιας ευπροσάρμοστης γραφικής διεπιφάνειας χρήστη, χρησιμοποιώντας PMW (windows manager) Python mega widgets.

¬ Το 2003, η γραφική διεπιφάνεια χρήστη ξανασχεδιάστηκε και ξανά αναπτύχθηκε για τη Qt πλατφόρμα, χρησιμοποιώντας PyQt Python συνδέσμους. Το πλαίσιο οπτικού προγραμματισμού καθορίστηκε, και η ανάπτυξη των widgets ξεκινάει.

¬ Το 2005, δημιουργούνται επεκτάσεις για την ανάλυση δεδομένων στον τομέα της βίο-πληροφορικής.

¬ Το 2008, αναπτύσσονται πακέτα εγκατάστασης για Mac OS X DMG χρησιμοποιώντας ως βάση το Fink (πρόγραμμα για εύκολη εγκατάσταση προγραμμάτων σε Mac).

¬ Το 2009, δημιουργούνται και παραμένουν σε χρήση πάνω από 100 widgets.

¬ Από το 2009, η Orange είναι σε έκδοση 2.0 beta και η επίσημη ιστοσελίδα παρέχει πακέτα εγκατάστασης μετά από καθημερινό κύκλο μεταγλώττισης.

¬ Το Σεπτέμβρη του 2012, δημιουργείται η έκδοση 2.5a.[69]

Απαιτήσεις Συστήματος Η Orange υποστηρίζεται σε διάφορες εκδόσεις Linux, Apple's Mac OS X και Microsoft Windows[70].

4.1.3 RapidMiner

Το RapidMiner, πρώην YALE (Yet Another Learning Environment), είναι ένα περιβάλλον για μηχανική εκμάθηση, εξόρυξη δεδομένων, προβλεπτική και επιχειρησιακή ανάλυση. Χρησιμοποιείται στην έρευνα, εκπαίδευση, ανάπτυξη εφαρμογών, γρήγορη προτυποποίηση και σε βιομηχανικές εφαρμογές. Σε μία δημοσκόπηση που έγινε από τη KDnuggets, μία εφημερίδα για την εξόρυξη δεδομένων, το RapidMiner ήρθε δεύτερο το 2009 ανάμεσα σε εργαλεία εξόρυξης δεδομένων/ανάλυσης που χρησιμοποιήθηκαν για πραγματικά έργα και πρώτο το 2010. Διανέμεται υπό την άδεια χρήσης λογισμικών ανοιχτού κώδικα AGPL, και στεγάζεται στο Source Forge από το 2004. Τέλος, είναι υλοποιημένο σε Java και τελευταία έκδοση είναι η 5.2 (2002)[71]

4.1.3.1 Ιστορικό

Το πρόγραμμα RapidMiner ξεκίνησε το 2001 από τουςRalf Klinkenberg, Ingo Mierswa και Simon Fischer στο Τμήμα Τεχνητής Νοημοσύνης στο πανεπιστήμιο του Dortmund. Το 2006 οι Ingo Mierswa και Ralf Klinkenberg ίδρυσαν την εταιρεία Rapid-I, που είναι τώρα ο κύριος συνεισφέρων στη περαιτέρω ανάπτυξη του RapidMiner καθώς και σε άλλους 30 προγραμματιστές παγκοσμίως[72].

4.1.3..2 Εφαρμογές

Το RapidMiner μπορεί να χρησιμοποιηθεί σε εξόρυξη κειμένων (text mining), εξόρυξη πολυμέσων (multimedia mining), μηχανική χαρακτηριστικών (feature

[69] URL: http://en.wikipedia.org/wiki/Orange_(software)
[70] URL: http://en.wikipedia.org/wiki/Orange_(software)
[71] URL: http://en.wikipedia.org/wiki/RapidMiner
[72] URL: http://en.wikipedia.org/wiki/RapidMiner

engineering), εξόρυξη ροής δεδομένων (data stream mining) και ανίχνευση εννοιών παρέκκλισης (tracking drifting concepts), ανάπτυξη μεθόδων συνόλου (development of ensemble methods) και επιμεριστική εξόρυξης δεδομένων.

Το RapidMiner συναντάται στο τομέα της ηλεκτρονικής, ενέργειας, πληροφορικής, φαρμακευτική και αυτοκινητιστική βιομηχανία, εμπόριο, αεροπλοΐα, τηλεπικοινωνίες, τραπεζικό και ασφαλιστικό κλάδο, στη παραγωγική διαδικασία, έρευνα αγοράς και άλλα πολλά πεδία.

4.1.3.3 Επεκτάσεις Λογισμικού

Το RapidMiner μπορεί να επεκταθεί με πρόσθετα plug-in. Περιέχει γύρω στις 15 επεκτάσεις, βελτιώνοντας έτσι την εφαρμογή του σε εξόρυξη κειμένου, επεξεργασία εικόνας, επεξεργασία χρονικών σειρών, διαδικτυακή εξόρυξη, στατιστική, οπτικοποίηση, σημασιολογία, παράλληλη επεξεργασία, αυτόματη διαδικασία σχεδιασμού και άλλα.

Πολλές από τις επεκτάσεις μπορούν να βρεθούν απευθείας στην εφαρμογή, μέσω του διαχειριστή επεκτάσεων. Οι υπόλοιπες μπορούν να ληφθούν από το διαδίκτυο από τους αντίστοιχους υπεύθυνους ανάπτυξης[73].

Για την εγκατάσταση των plug-in του RapidMiner, είναι απαραίτητο να αντιγραφούν στον υπό-κατάλογο lib/plugin του καταλόγου εγκατάστασης του RapidMiner. Το RapidMiner ελέγχει όλα τα jar αρχεία σε αυτό τον κατάλογο. Σε περίπτωση που ένα plug-in περιέχεται σε φάκελο με περισσότερα αρχεία από ένα μόνο jar αρχείο (μπορεί να είναι τεκμηρίωση ή παραδείγματα), πρέπει να γίνει είσοδος μόνο του αρχείου jar στον υπό-κατάλογο, ενώ τα υπόλοιπα έγγραφα να εισαχθούν μαζί με τα υπόλοιπα αρχεία. Για Windows, μπορεί να υπάρχει ένα εκτελέσιμο αρχείο .exe το οποίο μπορεί να χρησιμοποιηθεί για αυτόματη εγκατάσταση του plug-in στο σωστό κατάλογο. Και στις δύο περιπτώσεις, το plug-in είναι διαθέσιμο με την επόμενη επανεκκίνηση του RapidMiner[74].

[73] URL: http://www.cs.ubbcluj.ro/~gabis/ml/MLSoftware/rapidminer-4.4-tutorial.pdf
[74] URL: http://www.cs.ubbcluj.ro/~gabis/ml/MLSoftware/rapidminer-4.4-tutorial.pdf

4.1.4 Weka

Το Weka (Waikato Environment for Knowledge Analysis), που είναι μία περιεκτική οικογένεια βιβλιοθηκών Java που υλοποιεί πολλούς σύγχρονους αλγόριθμους μηχανικής εκμάθησης και εξόρυξης δεδομένων, αναπτύχθηκε στο Πανεπιστήμιο του Waikato της Νέας Ζηλανδίας. Είναι ένα δωρεάν λογισμικό και διατίθεται με άδεια χρήσης GPL (General Public License)[75].

Συνοδεύεται από ένα κείμενο πάνω στην εξόρυξη δεδομένων, το οποίο τεκμηριώνει και εξηγεί πλήρως όλους τους αλγόριθμους που περιέχει. Οι εφαρμογές που γράφονται με το Weka, μπορούν να εκτελεστούν σε οποιοδήποτε σύστημα με δυνατότητα σύνδεσης στο Internet. Αυτό επιτρέπει στους χρήστες να εφαρμόσουν τεχνικές μηχανικής εκμάθησης στα δεδομένα τους, ανεξαρτήτως του συστήματος που χρησιμοποιούν. Παρέχονται εργαλεία για προ-επεξεργασία δεδομένων, τροφοδότηση τους σε ποικίλα σχήματα εκμάθησης και ανάλυση των ταξινομητών που προκύπτουν και της αποδοτικότητας τους. Μία σημαντική πηγή για την περιήγηση στο Weka είναι η on-line τεκμηρίωση του, η οποία παράγεται αυτόματα από την πηγή[76].

Οι πρωταρχικοί μέθοδοι μάθησης στο Weka είναι οι ταξινομητές που επάγουν ένα σύνολο κανόνων ή ένα δέντρο απόφασης που μοντελοποιεί τα δεδομένα. Το Weka επίσης συμπεριλαμβάνει αλγόριθμους για κανόνες συσχέτισης και συσταδοποίηση δεδομένων. Όλες οι υλοποιήσεις έχουν μία ομοιόμορφη διεπιφάνεια γραμμής εντολών. Ένα κοινό υπό-πρόγραμμα αξιολόγησης αξιολογεί τη σχετική απόδοση αρκετών αλγορίθμων εκμάθησης όσον αφορά συγκεκριμένα σύνολα δεδομένων. Τα φίλτρα, εργαλεία για προ-επεξεργασία δεδομένων, είναι ένας άλλος σημαντικός

[75] Ian H. Witten et al. (1999) "Weka: Practical Machine Learning Tools and Techniques with Java Implementation"- URL: http://www.cs.waikato.ac.nz/~eibe/pubs/99IHW-EF-LT-MH-GH-SJCTools-Java.pdf

[76] Ian H. Witten et al. (1999) "Weka: Practical Machine Learning Tools and Techniques with Java Implementation"- URL: http://www.cs.waikato.ac.nz/~eibe/pubs/99IHW-EF-LT-MH-GH-SJCTools-Java.pdf

πόρος. Ομοίως με τα σχήματα εκμάθησης, τα φίλτρα έχουν μία τυποποιημένη διεπιφάνεια γραμμής εντολών με ένα σύνολο απλών επιλογών γραμμής εντολών[77].

Το λογισμικό είναι γραμμένο εξολοκλήρου σε Java για να διευκολύνει τη διαθεσιμότητα των εργαλείων εξόρυξης δεδομένων, ανεξαρτήτως του χρησιμοποιούμενου συστήματος. Με μια λέξη, το σύστημα είναι μια οικογένεια πακέτων ανάπτυξης Java, με το καθένα από αυτά να είναι τεκμηριωμένο ώστε να παρέχει στους υπεύθυνους ανάπτυξης δυνατότητες τελευταίας τεχνολογίας[78].

4.1.4.1 Ιστορικό

¬ Το 1993, το Πανεπιστήμιο του Waikato στη Νέα Ζηλανδία αρχίζει την ανάπτυξη της αρχικής έκδοσης του Weka (ένα μείγμα από TCL/TK, C και Make αρχεία).

¬ Το 1997, αποφασίζεται η επανάπτυξη του Weka από την αρχή με χρήση Java, συμπεριλαμβάνοντας υλοποιήσεις αλγορίθμων μοντελοποίησης.

-Το 2005, το Weka λαμβάνει το βραβείο SIGKDD (Special Interest Group on Knowledge Discovery and Data Mining).

¬ Το 2006, η εταιρεία Pentaho αποκτά την αποκλειστική άδεια χρήσης του Weka για επιχειρησιακή πληροφορία. Δημιουργεί το συστατικό εξόρυξης δεδομένων και προβλεπτικής ανάλυσης του πακέτου επιχειρησιακής πληροφορίας της Pentaho.

¬ Κατέχει τη 246η θέση στο Sourceforge.net από το 2009, με 1.566.318 λήψεις[79].

4.1.4.2 Πρόσθετα πακέτα

Το Weka έχει την έννοια πακέτου σαν μία δέσμη πρόσθετης λειτουργικότητας, ξεχωριστής από αυτή που υπάρχει στο κυρίως αρχείο weka.jar. Ένα πακέτο αποτελείται από διάφορα jar αρχεία, τεκμηρίωση, μετά-δεδομένα και πιθανώς

[77] Ian H. Witten et al. (1999) "Weka: Practical Machine Learning Tools and Techniques with Java Implementation"- URL: http://www.cs.waikato.ac.nz/~eibe/pubs/99IHW-EF-LT-MH-GH-SJCTools-Java.pdf

[78] Ian H. Witten et al. (1999) "Weka: Practical Machine Learning Tools and Techniques with Java Implementation"- URL: http://www.cs.waikato.ac.nz/~eibe/pubs/99IHW-EF-LT-MH-GH-SJCTools-Java.pdf

[79] . URL: http://en.wikipedia.org/wiki/Rattle_GUI

πηγαίο κώδικα. Υπάρχουν πολλά διαθέσιμα πακέτα για το Weka που προσθέτουν σχήματα εκμάθησης ή επεκτείνουν τη λειτουργικότητα του κυρίως συστήματος κατά κάποιο τρόπο. Πολλά από αυτά παρέχονται από την ομάδα του Weka και άλλα από τρίτους. Το Weka περιλαμβάνει δυνατότητα διαχείρισης των πακέτων και μηχανισμό για τη δυναμική φόρτωση τους στο περιβάλλον εκτέλεσης. Υπάρχει γραμμή εντολών και πακέτο διαχείρισης διεπιφάνειας χρήστη.

4.1.5 Rattle

Το Rattle είναι ένα πακέτο λογισμικού ανοιχτού κώδικα που παρέχει μία γραφική διεπιφάνεια χρήστη για εξόρυξη δεδομένων με χρήση της προγραμματιστικής στατιστικής γλώσσας R. Ο πηγαίος κώδικας είναι διαθέσιμος στο rattle.googlecode.com. Σήμερα, το Rattle χρησιμοποιείται παγκοσμίως για πληθώρα καταστάσεων. Προς το παρόν, 15 διαφορετικά κυβερνητικά τμήματα στην Αυστραλία και σε όλο τον κόσμο χρησιμοποιούν το Rattle σε δραστηριότητες εξόρυξης δεδομένων και ως στατιστικό πακέτο[80]. Σχεδιάστηκε ειδικά για να διευκολύνει τη μετάβαση από την απλή και βασική εξόρυξη δεδομένων, που υπάρχει απαραίτητα στις διεπιφάνειες χρήστη, στην εξελιγμένη ανάλυση δεδομένων, χρησιμοποιώντας μία ισχυρή στατιστική γλώσσα.

Το Rattle ενώνει μία πλειάδα πακέτων R τα οποία είναι απαραίτητα για κάποιον που ασχολείται με τον τομέα της εξόρυξης δεδομένων, αλλά συχνά δύσκολα προς χρήση για ένα αρχάριο. Δεν είναι απαραίτητη η κατανόηση της R για να ξεκινήσει κανείς να χρησιμοποιεί το Rattle, αυτό θα γίνει σιγά-σιγά, με την ολοένα και αυξανόμενη επιτήδευση στα έργα εξόρυξης δεδομένων που κάνει. Η διεπιφάνεια χρήστη του Rattle παρέχει μία πρώτη εικόνα στη δύναμη της R ως εργαλείο για την εξόρυξη δεδομένων[81].

Το Rattle χρησιμοποιείται ως λογισμικό εκμάθησης της γλώσσας R. Υπάρχει καρτέλα για καταγραφή κώδικα (Log Code tab), η οποία αναπαράγει το κώδικα R που χρησιμοποιήθηκε για οποιαδήποτε δραστηριότητα στη διεπιφάνεια χρήστη, η

[80] . URL: http://en.wikipedia.org/wiki/Rattle_GUI
[81] Graham J Williams (2009), "Rattle, a data mining GUI for R" – URL: http://journal.rproject.org/archive/2009-2/RJournal_2009-2_Williams.pdf

οποία μπορεί μετά να αντιγραφεί και επικολληθεί. Επίσης, το Rattle μπορεί να χρησιμοποιηθεί για στατιστική ανάλυση ή παραγωγή μοντέλων. Επιτρέπει στο σύνολο δεδομένων να χωριστεί σε δεδομένα εκπαίδευσης, διασταύρωσης και ελέγχου. Το σύνολο δεδομένων μπορεί να προβληθεί και επεξεργαστεί.

4.1.5.1 Πακέτα Ανάπτυξης

Το Rattle βασίζεται σε μία εκτεταμένη συλλογή πακέτων ανάπτυξης R. Αυτό είναι και μία απόδειξη της δύναμης της R. Προσφέρει στατιστική ανάλυση εις εύρος και βάθος που είναι δύσκολο να βρεθεί αλλού. Μερικά από τα πακέτα που διέπουν το Rattle είναι: ada, arules, doBy, ellipse, fBasics, fpc, gplots, Hmisc, kernlab, mice, network, party, playwith, pmml, random Forest, reshape, rggobi, RGtk2, ROCR, RODBC και rpart. Τα παραπάνω είναι διαθέσιμα από το CRAN (Comprehensive R Archive Network). Σε περίπτωση που ένα πακέτο δεν έχει εγκατασταθεί, αν ρωτήσουμε μέσω του Rattle για κάποια υποστηριζόμενη από αυτό το πακέτο λειτουργία, θα εμφανιστεί ένα μήνυμα που θα υποδεικνύει πιο πακέτο πρέπει να εγκατασταθεί[82].

4.1.5.2 Συμβατές Πλατφόρμες

Η Rattle χρησιμοποιεί τη γραφική διεπιφάνεια χρήστη Gnome, όπως αυτή παρέχεται από το πακέτο RGtk2 (Lawrence και Lang, 2006). Τρέχει σε όλα τα συστήματα, συμπεριλαμβανομένων των GNU/Linux, Macintosh OS/X και MS/Windows. Τελευταία έκδοση είναι η 2.6.5 (αναθεώρηση 656) που κυκλοφόρησε το Μάρτη του 2011. Η ίδια η διεπιφάνεια χρήστη αναπτύχθηκε με χρήση του εργαλείου δόμησης διαδραστικών διεπιφανειών Glade[83].

[82] Graham J Williams (2009), “Rattle, a data mining GUI for R” – URL: http://journal.rproject.org/archive/2009-2/RJournal_2009-2_Williams.pdf
[83] . URL: http://en.wikipedia.org/wiki/Rattle_GUI

4.1.6 Tanagra

Το Tanagra είναι ένα δωρεάν λογισμικό εξόρυξης δεδομένων, γραμμένο σε C++. Αλγόριθμοι εξόρυξης δεδομένων, διερευνητικής ανάλυσης δεδομένων, μηχανικής εκμάθησης και στατιστικής εκμάθησης, είναι όλοι διαθέσιμοι με το Tanagra. Μπορεί να χρησιμοποιηθεί σε πειραματισμούς για ακαδημαϊκές δημοσιεύσεις ή μελέτες πάνω σε πραγματικές εφαρμογές[84].

Η λειτουργία χρήστη στηρίζεται εξ' ολοκλήρου στο μοντέλο διαγράμματος ροής (stream diagram paradigm). Υπό το καθεστώς του μοντέλου διαγράμματος ροής, ο χρήστης σχεδιάζει ένα γράφο ορίζοντας τις πηγές δεδομένων και λειτουργιών που εκτελούνται πάνω στα δεδομένα. Οι διαδρομές πάνω στο γράφο μπορούν να περιγράψουν τη ροή δεδομένων μέσα από χειρισμούς και αναλύσεις. Το Tanagra απλοποιεί αυτό το μοντέλο περιορίζοντας το γράφο σε δέντρο. Αυτό σημαίνει ότι κάθε κόμβος μπορεί να έχει ένα μόνο πατέρα, δηλαδή μόνο μία πηγή δεδομένων για κάθε λειτουργία[85].

Ο πηγαίος κώδικας του έργου είναι δωρεάν και μπορεί να ληφθεί από το διαδίκτυο. Το λογισμικό μπορεί να είναι δωρεάν, αλλά το ίδιο πρέπει να συμβαίνει και για το έργο που υλοποιείται. Επίσης, υποστηρίζεται ένα μεγάλο σύνολο λειτουργιών, όπως φιλτράρισμα, συσταδοποίηση, κατηγοριοποίηση και άλλες τεχνικές ανάλυσης. Υποστηρίζονται επίσης πολλοί τύποι αρχείων και πρόσβαση σε βάσεις δεδομένων για είσοδο και έξοδο δεδομένων. Επίσης είναι γραμμένο και σε Java, κάτι που το κάνει χρήσιμο σε πολλές πλατφόρμες και ιδιαίτερα εξυπηρετικό [15].

Το Tanagra διανέμεται από το Δεκέμβριο του 2003. Μεταγλωττίζεται για WIN32 πλατφόρμα, αλλά μπορεί και να εκτελεστεί και σε άλλα συστήματα όπως το WINE με LINUX[86].

Συμπεράσματα

[84] URL: http://en.wikipedia.org/wiki/TANAGRA_(software)
[85] Jessica Enright and Jonathan Klippenstein (2004), "Tanagra: An Evaluation"- URL: http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput695-04/work/A2-reports/tanagra.pdf
[86] Jessica Enright and Jonathan Klippenstein (2004), "Tanagra: An Evaluation"- URL: http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput695-04/work/A2-reports/tanagra.pdf

Συμπεράσματα

Το Business Intelligence είναι μια έννοια που συνήθως περιλαμβάνει την παράδοση και ολοκλήρωση των σχετικών και χρήσιμων επιχειρηματικών πληροφοριών σε έναν οργανισμό. Ως εκ τούτου, οι εταιρείες χρησιμοποιούν την επιχειρηματική ευφυΐα για την ανίχνευση σημαντικών γεγονότων και να εντοπίσουν ή να παρακολουθήσουν τις επιχειρηματικές τάσεις, προκειμένου να προσαρμοστούν γρήγορα στο μεταβαλλόμενο περιβάλλον. Εάν χρησιμοποιηθεί αποτελεσματική εκπαίδευση στην επιχειρηματική ευφυΐα στον οργανισμό σας, μπορεί να βελτιωθούν οι διαδικασίες λήψης αποφάσεων σε όλα τα επίπεδα της διοίκησης και να βελτιωθούν ο τακτικές και στρατηγικές διαδικασίες διαχείρισης σας. Εδώ είναι μερικοί από τους κορυφαίους λόγους για την επένδυση σε ένα κατάλληλο σύστημα επιχειρηματικής ευφυΐας.

Ένα από τα κύρια πλεονεκτήματα της επένδυσης σε λογισμικό επιχειρηματικής ευφυΐας και εξειδικευμένο προσωπικό είναι το γεγονός ότι θα ενισχύσει την ικανότητά σας να αναλύσετε τις τρέχουσες τάσεις αγοράς των καταναλωτών. Μόλις καταλάβετε τι αγοράζουν οι καταναλωτές σας, μπορείτε να χρησιμοποιήσετε αυτές τις πληροφορίες για να αναπτύξουν προϊόντα που να ταιριάζουν με τις τρέχουσες τάσεις της κατανάλωσης και, κατά συνέπεια, να βελτιώσει την αποδοτικότητά σας, αφού θα είστε σε θέση να προσελκύσετε πολύτιμους πελάτες.

Αν θέλετε να βελτιώσετε τον έλεγχό σας πάνω από διάφορες σημαντικές διεργασίες στον οργανισμό σας, θα πρέπει να εξετάσει την επένδυση σε ένα καλό σύστημα επιχειρηματικής ευφυΐας. Το λογισμικό επιχειρηματικής ευφυΐας θα βελτιώσει την ορατότητα αυτών των διαδικασιών και να καταστεί δυνατό να προσδιοριστούν οι τομείς που χρειάζονται βελτίωση. Επιπλέον, αν έχετε αυτή τη στιγμή να αναλύσετε εκατοντάδες σελίδες από αναλυτικές περιοδικές εκθέσεις σας για να αξιολογήσετε την απόδοση των διαδικασιών του οργανισμού σας, μπορείτε να εξοικονομήσετε χρόνο και να βελτιώσετε την παραγωγικότητα, έχοντας έμπειρους αναλυτές νοημοσύνης με τη χρήση λογισμικού.

Ένα σύστημα επιχειρηματικής ευφυΐας είναι ένα αναλυτικό εργαλείο που μπορεί να σας δώσει τη διορατικότητα που χρειάζεστε για να κάνετε επιτυχημένα στρατηγικά σχέδια για τον οργανισμό σας. Αυτό οφείλεται στο γεγονός ότι ένα τέτοιο σύστημα θα είναι σε θέση να προσδιορίσει τις βασικές τάσεις και τα πρότυπα στα δεδομένα οργανώσεις σας και, κατά συνέπεια, να καταστεί ευκολότερο για σας να κάνουν σημαντικές συνδέσεις μεταξύ των διαφόρων περιοχών της επιχείρησής σας που μπορεί διαφορετικά να φαίνονται άσχετα. Ως εκ τούτου, ένα σύστημα επιχειρηματικής ευφυΐας μπορεί να σας βοηθήσει να κατανοήσετε τις επιπτώσεις των διαφόρων οργανωτικών διαδικασιών καλύτερα και να ενισχύσει την ικανότητά σας για τον εντοπισμό κατάλληλων ευκαιριών για τον οργανισμό σας, έτσι ώστε να μπορείτε να προγραμματίσετε ένα επιτυχημένο μέλλον.

Ένας από τους σημαντικότερους λόγους για τους οποίους θα πρέπει να επενδύσουν σε ένα αποτελεσματικό σύστημα επιχειρηματικής ευφυΐας είναι επειδή ένα τέτοιο σύστημα μπορεί να βελτιώσει την αποδοτικότητα στον οργανισμό σας και, ως αποτέλεσμα, την αύξηση της παραγωγικότητας. Μπορείτε να χρησιμοποιήσετε την επιχειρηματική ευφυΐα για την ανταλλαγή πληροφοριών μεταξύ των διαφόρων τμημάτων στον οργανισμό σας. Αυτό θα σας επιτρέψει να εξοικονομήσετε χρόνο σχετικά με τις διαδικασίες και τα analytics εκθέσεων. Αυτή η ευκολία στην ανταλλαγή πληροφοριών είναι πιθανό να μειώσει τις επικαλύψεις των ρόλων / καθηκόντων στο πλαίσιο του οργανισμού και να βελτιωθεί η ακρίβεια και η χρησιμότητα των δεδομένων που παράγονται από διαφορετικά τμήματα. Επιπλέον, η ανταλλαγή πληροφοριών επίσης εξοικονομεί χρόνο και βελτιώνει την παραγωγικότητα.

Προκειμένου να αποκομιστούν όλα τα οφέλη ενός αποτελεσματικού συστήματος επιχειρηματικής ευφυΐας, απαιτείται ή επένδυση σε εξειδικευμένο προσωπικό επιχειρηματικής ευφυΐας και λογισμικό.

Βιβλιογραφία/Αναφορές

Ελληνική Βιβλιογραφία

Θεοδωρίδης Γ., Πελέκης Ν. (2011): Εξόρυξη Γνώσης από Δεδομένα - Συσταδοποίηση, Ομάδα Διαχείρισης Δεδομένων, Πανεπιστήμιο Πειραιώς

Σαλατάς Ι. (2011): Υλοποίηση και εφαρμογή Τεχνητών Νευρωνικών Δικτύων για την πρόβλεψη χρονοσειρών συναλλαγματικών ισοτιμιών, Ελληνικό Ανοιχτό Πανεπιστήμιο.

Σταυλιώτης Ε. Γεράσιμος (2009): Εξόρυξη Δεδομένων και Αναγνώριση προτύπων σε κατηγορικά δεδομένα μέσω συσταδοποίησης, Ελληνικό Στατιστικό Ινστιτούτο

Κωνσταντίνος Δ. (2007): Τεχνητά Νευρωνικά Δίκτυα., Εκδόσεις Κλειδάριθμος, Αθήνα

Διεθνής Βιβλιογραφία

-CIO Leadership Forum 2015 (2015). http://www.gartnerinfo.com/cios9/CIOLeadershipForum2015Profile.pdf

-CPM (Corporate Performance Management) – Gartner IT Glossary. (n.d.). http://www.gartner.com/it-glossary/cpm-corporate-performance-management

--Demsar J, Zupan B (2004), "Orange: From Experimental Machine Learning to Interactive Data Mining", White Paper (www.ailab.si/orange), Faculty of Computer and Information Science, University of Ljubljana - URL: http://orange.biolab.si/wp/orange-leaflet-visual.pdf

-Devens, M. (1865). Cyclopædia of commercial and business anecdotes. New York, NY: D. Appleton and Company. Evtm_219_CIOtop10[3].pdf http://www.gartnerinfo.com/sym23/evtm_219_CIOtop10%5B3%5D.pdf

-Dunham M.H., (2004): Data Mining introductory and advanced topics", Prentice Hall

-Enright, J. and Klippenstein, J (2004), "Tanagra: An Evaluation"- URL: http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput695-04/work/A2-reports/tanagra.pdf

-Freitas, A.A., (1998): A Survey of Parallel Data Mining, in Proceedings of 2nd International Conference on the Practical Applications of Knowledge Discovery and Data Mining

-Friedman J. H. K., (1997): On bias, variance, 0/1-loss, and the curse-of-dimensionality, Data Mining and Knowledge Discovery

-Fayyad U., Piatetsky-Shaprio G., Smyth P. and Uthurusamy R., (1996): Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge

-Fayol, H. (1949). General and Industrial Management. London, UK: Pitman.

-Ferrari, A. (2011). Business Intelligence Systems, Uncertainty in Decision-Making and Effectiveness of Organizational Coordination, Springe

-Gartner Executive Programs' Worldwide Survey on More Than 2300 CIOs Shows flat IT Budgets in 2012, but IT Organizations Must Deliver on Multiple Priorities. (n.d.). http://www.gartner.com/newsroom/id/1897514

-Graham J Williams (2009), "Rattle, a data mining GUI for R" – URL: http://journal.rproject.org/archive/2009-2/RJournal_2009-2_Williams.pdf

-Hall A. M., Frank E., Witten H. I (2011): Data Mining, Practical Machine Learning Tools and Techniques

-Ian H. W. et al. (1999) “Weka: Practical Machine Learning Tools and Techniques with Java Implementation”- URL: http://www.cs.waikato.ac.nz/~eibe/pubs/99IHW-EF-LT-MH-GH-SJCTools-Java.pdf

-Information Management | IT Business News. (n.d.). http://www.information-management.com

-Kumar V., Steinbach M. (2006): Introduction to Data Mining, Addison Wesley.

-Kumar, Steinbach, Tan (2004): Introduction to Data Mining, University of Stanford

-Lazarevic A., (2008): Data Mining for Anomaly Detection, Tutorial at the European

-Luhn, H. P. (1958). A Business Intelligence System. IBM Journal of Research and Development, 2(4), 314- 319.

-Mintzberg, H. (1990). Mintzberg on Management: Inside our Strange World of Organizations. New York, NY: Free Press.

-Murthy S. (1998): Automatic construction of decision trees from data, A multidisciplinary survey. Data Mining and Knowledge Discovery

-Oracle Business Intelligence Applications. (n.d.). http://www.oracle.com/technetwork/middleware/biapplications/overview/index.html

-Price Waterhouse Coopers. (2007). Guide to Performance Indicators.

-Prodromidis A. and Chan P., (2000): Meta-learning in distributed data mining systems, Issues and Approaches, in Advances of Distributed Data Mining, AAAI Press.

-SAP. (2015). Business Intelligence Tools | BI & Analytics | SAP. http://go.sap.com/solution/platform-technology/business-intelligence.html

-Sabherwal, R., & Beccera – Fernandez, I. (2010). Business Intelligence. Hoboken, NJ: John Wiley and Sons Inc.

-Saran, C. (2012). Almost a Third of BI Projects Fail to Deliver on Business Objectives Computer Weekly. http://www.computerweekly.com/news/2240113585/Almost-a-third-of-BI-projectsfail-to-deliver-on-business-objectives

-Scheps, S. (2007). Business Intelligence for Dummies. Hoboken, NJ: Willey Publishing Inc.

-Witten I.H. and Frank E., (2005): Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann

-Xiaojun Chen et al. (2007), “A Survey of Open Source Data Mining Systems” - URL: http://togaware.redirectme.net/papers/pakdd07.pdf

Ιστότοποι

URL: http://en.wikipedia.org/wiki/Orange_(software)

URL: http://en.wikipedia.org/wiki/RapidMiner

URL: http://www.cs.ubbcluj.ro/~gabis/ml/MLSoftware/rapidminer-4.4-tutorial.pdf

URL: http://en.wikipedia.org/wiki/Open-source_software

URL: http://en.wikipedia.org/wiki/Weka_(machine_learning)

URL: http://jwork.org/jhepwork/

URL: http://en.wikipedia.org/wiki/JHepWork

URL: http://en.wikipedia.org/wiki/Rattle_GUI

URL: http://www.thearling.com/text/dmwhite/dmwhite.htm

URL: http://en.wikipedia.org/wiki/Carrot2

URL: http://en.wikipedia.org/wiki/TANAGRA_(software)

TDWI | Advancing all things data. (n.d.). http://tdwi.org/Home.aspx

TDAN.com, from http://www.tdan.com

-