ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

ΣΧΟΛΗ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ ΚΑΙ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ

ΤΜΗΜΑ ΔΙΟΙΚΗΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ

(πρώην Τμήμα Διοίκησης Επιχειρήσεων – Μεσολόγγι)

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΑΠΟ ΤΡΑΠΕΖΙΚΑ ΔΕΔΟΜΕΝΑ ΔΙΑΜΕΣΩ ΑΥΤΟΜΑΤΟΠΟΙΗΜΕΝΗΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΜΕ ΧΡΗΣΗ ΤΟΥ ΕΡΓΑΛΕΙΟΥ WEKA

ΔΙΟΝΥΣΙΑ ΜΠΙΡΜΠΙΛΗ Α.Μ 15197

Τριμελής Επιτροπή:
Ήρα Αντωνόπουλου, Καθηγήτρια
Δημήτρης Παπαδόπουλος, Επίκουρος Καθηγητής
Μαρία Ρήγκου, Επίκουρη Καθηγήτρια (Επιβλέπουσα)

ΜΕΣΟΛΟΓΓΙ 2021

Πίνακας Περιεχομένων

Εισαγωγή

Η ανάπτυξη συνδέεται στενά με την τεχνολογία. Το στάδιο ανάπτυξης που έφτασε ο άνθρωπος θα μπορούσε να ήταν δυνατό χωρίς την πρόοδο της τεχνολογίας. Η ριζική αλλαγή και η πρόοδος στην οικονομία, όπως παρατηρούμε σήμερα, είναι το αποτέλεσμα της σύγχρονης τεχνολογίας. Η τεχνολογία επέφερε αποτελεσματικότητα και ποιότητα στον κατασκευαστικό τομέα. Η τεχνολογική πρόοδος έχει μειώσει τον κίνδυνο που ενέχονται στις μεταποιητικές επιχειρήσεις. Υπήρξε τεράστια βελτίωση στον τομέα της υγείας στον κόσμο, όχι μόνο η μέση ηλικία των ανθρώπων έχει αυξηθεί, αλλά και το ποσοστό θνησιμότητας έχει επίσης μειωθεί σημαντικά. Αυτό θα μπορούσε να είναι δυνατό μόνο λόγω της τεχνολογικής προόδου στον τομέα της υγείας. Ίσως δεν υπάρχει κανένα πεδίο της ανθρώπινης ζωής που δεν έχει επηρεαστεί από την τεχνολογία. Η γεωργία, η βιομηχανία, το επάγγελμα, η υγεία, η εκπαίδευση, η τέχνη, οι πολιτικές διαδικασίες, η ψυχαγωγία, οι θρησκευτικές δραστηριότητες και οι καθημερινές δραστηριότητες υπόκεινται στην τεχνολογία.

Όμως, είναι σημαντικό να έχουμε κατά νου ότι η τεχνολογική πρόοδος έχει επηρεάσει την ανθρώπινη ζωή τόσο θετικά όσο και αρνητικά. Όχι μόνο ότι η ζωή έχει γίνει εύκολη και άνετη, υπάρχουν επίσης ενδείξεις για αρκετές απειλές για τη ζωή και την κοινωνία στο μέλλον λόγω της χρήσης / κακής χρήσης της σύγχρονης τεχνολογίας. Η φύση και η έκταση της ανάπτυξης που έχει βιώσει η ανθρώπινη κοινωνία τώρα κατευθύνεται προς κρίσεις στο μέλλον. Η βιωσιμότητα της ανάπτυξης αμφισβητείται σήμερα. Αυτό συνέβη μόνο λόγω της παράλογης χρήσης της τεχνολογίας. Έχει συζητηθεί εδώ ως προς το πώς η ανάπτυξη - οικονομική και κοινωνική - λαμβάνει χώρα με την πρόοδο της τεχνολογίας, αλλά όχι χωρίς να αφήνει μια ουλή για να απειλήσει την ανθρώπινη κοινωνία. Η ανάπτυξη της τεχνολογίας, η οποία από μόνη της είναι συμπωματική της ανάπτυξης, επέφερε όχι μόνο οικονομική ανάπτυξη αλλά και ριζικές αλλαγές στην κοινωνική και πολιτιστική σφαίρα της κοινωνίας.

Τα τυπικά δεδομένα είναι ένας κατάλογος τιμών «κανονικού χρόνου» για διαφορετικά στοιχεία εργασιών ή για κινήσεις λεπτών που εμπλέκονται σε διαφορετικές εργασίες. Αυτός ο κατάλογος καταρτίζεται με τη σύνταξη των χρονοδιαγραμμάτων ενός αριθμού τυπικών στοιχείων . Η ανάγκη προετοιμασίας ενός τέτοιου καταλόγου χρόνου ή τυπικών δεδομένων προέκυψε επειδή (σε έναν κλάδο), γενικά, παρόμοια στοιχεία ή κινήσεις εμπλέκονται σε πολλές θέσεις εργασίας. (Για παράδειγμα, οι τρύπες διάτρησης είναι ένα κοινό χαρακτηριστικό πολλών εργασιών μηχανοστασίου). Εάν πρέπει να διεξαχθεί μελέτη χρόνου για κάθε νέα εργασία, είναι φυσικά σπατάλη η επαναφορά εκείνων των στοιχείων της νέας εργασίας που είναι κοινά με τις προηγούμενες θέσεις εργασίας.

1.Βιβλιογραφική Ανασκόπηση

1.1Συστήματα Ανάλυσης Δεδομένων

1.1.1Τα Δεδομένα

Στα στατιστικά στοιχεία, τα ονομαστικά δεδομένα (επίσης γνωστά ως ονομαστική κλίμακα) είναι ένας τύπος δεδομένων που χρησιμοποιείται για την επισήμανση μεταβλητών χωρίς να παρέχει καμία ποσοτική τιμή. Είναι η απλούστερη μορφή κλίμακας μέτρου. Σε αντίθεση με τα κανονικά δεδομένα , τα ονομαστικά δεδομένα δεν μπορούν να ταξινομηθούν και δεν μπορούν να μετρηθούν. Εν αντιθέσει με τα δεδομένα διαστήματος ή αναλογίας, τα ονομαστικά δεδομένα δεν μπορούν να χειριστούν χρησιμοποιώντας τους διαθέσιμους μαθηματικούς τελεστές. Έτσι, το μόνο μέτρο της κεντρικής τάσης για τέτοια δεδομένα είναι ο τρόπος.

Τα ονομαστικά δεδομένα μπορούν να είναι ποιοτικά και ποσοτικά. Ωστόσο, οι ποσοτικές ετικέτες δεν έχουν αριθμητική τιμή ή σχέση (π.χ., αριθμός αναγνώρισης). Από την άλλη πλευρά, διάφοροι τύποι ποιοτικών δεδομένων μπορούν να αναπαρασταθούν σε ονομαστική μορφή. Μπορεί να περιλαμβάνουν λέξεις, γράμματα και σύμβολα. Τα ονόματα ατόμων, φύλου και εθνικότητας είναι μόνο μερικά από τα πιο κοινά παραδείγματα ονομαστικών δεδομένων. Τα ονομαστικά δεδομένα μπορούν να αναλυθούν χρησιμοποιώντας τη μέθοδο ομαδοποίησης. Οι μεταβλητές μπορούν να ομαδοποιηθούν σε κατηγορίες και για κάθε κατηγορία μπορεί να υπολογιστεί η συχνότητα ή το ποσοστό. Τα δεδομένα μπορούν επίσης να παρουσιαστούν οπτικά,

όπως χρησιμοποιώντας ένα γράφημα πίτας. Αν και τα ονομαστικά δεδομένα δεν μπορούν να αντιμετωπιστούν χρησιμοποιώντας μαθηματικούς τελεστές, μπορούν ακόμη να αναλυθούν χρησιμοποιώντας προηγμένες στατιστικές μεθόδους. Για παράδειγμα, ένας τρόπος για την ανάλυση των δεδομένων είναι μέσω δοκιμών υπόθεσης (Diebold,2012).

Για ονομαστικά δεδομένα, ο έλεγχος υπόθεσης μπορεί να πραγματοποιηθεί χρησιμοποιώντας μη παραμετρικές δοκιμές, όπως η δοκιμή chi-squared . Η δοκιμή chi-squared στοχεύει να προσδιορίσει εάν υπάρχει σημαντική διαφορά μεταξύ της αναμενόμενης συχνότητας και της παρατηρούμενης συχνότητας των δεδομένων τιμών Τα ποσοτικά δεδομένα αφορούν αριθμούς και πράγματα που μπορείτε να μετρήσετε αντικειμενικά: διαστάσεις όπως ύψος, πλάτος και μήκος. Θερμοκρασία και υγρασία. Τιμές Περιοχή και όγκος. Τα ποιοτικά δεδομένα αναφέρονται σε χαρακτηριστικά και περιγραφές που δεν μπορούν να μετρηθούν εύκολα, αλλά μπορούν να παρατηρηθούν υποκειμενικά - όπως μυρωδιές, γεύσεις, υφές, ελκυστικότητα και χρώμα. Σε γενικές γραμμές, όταν μετράτε κάτι και του δίνετε μια αριθμητική τιμή, δημιουργείτε ποσοτικά δεδομένα. Όταν ταξινομείτε ή κρίνετε κάτι, δημιουργείτε ποιοτικά δεδομένα. Μέχρι εδώ καλά. Αλλά αυτό είναι μόνο το υψηλότερο επίπεδο δεδομένων: υπάρχουν επίσης διαφορετικοί τύποι ποσοτικών και ποιοτικών δεδομένων(Berente & Seidel, 2014). Σε τέτοιες περιπτώσεις είναι πάντοτε οικονομικό να χρησιμοποιούνται τα δεδομένα που είχαν ήδη χρονομετρηθεί και συγκεντρωθεί, που ονομάζεται Τυπικά δεδομένα. Μόλις τα τυπικά δεδομένα είναι έτοιμα, πρέπει να αναφέρετε τα στοιχεία εργασίας ή τις λεπτές κινήσεις μιας λειτουργίας, να διαβάσετε τους χρόνους τους από τον τυπικό κατάλογο δεδομένων και να τα προσθέσετε. Ο συνολικός χρόνος που προκύπτει είναι μια εκτίμηση του κανονικού χρόνου για μια εργασία που μπορεί να μετατραπεί σε κανονικό χρόνο με την προσθήκη κατάλληλων επιδομάτων. Τα τυπικά δεδομένα (Macrodata) βασίζονται σε στοιχεία μιας εργασίας, είναι επίσης γνωστά ως "Element Standard Data" και συλλέγονται για μια αντιπροσωπευτική ομάδα στοιχείων με μακροσκοπικές μεθόδους. Είναι για οικογένειες θέσεων εργασίας και δίνει κανονικό χρόνο για διάφορα στοιχεία θέσεων εργασίας (Diebold,2012).

Οι τιμές χρόνου προέρχονται από τις πραγματικές μετρήσεις χρονομέτρου (ή άλλων) μετρήσεων των εργασιών (εντός της οικογένειας εργασίας) που πραγματοποιήθηκαν προηγουμένως. Αυτός ο τύπος δεδομένων περιορίζεται σε συγκεκριμένες λειτουργίες όπως η κατεργασία σε τόρνο κ.λπ. Οι πράξεις χωρίζονται σε στοιχεία. πότε είναι τότε, χρονοδιάγραμμα για τη λήψη ενός συστήματος δεδομένων που δείχνει τον κανονικό χρόνο του στοιχείου για οποιαδήποτε και όλες τις εργασίες

(ολοκληρώθηκαν σε αυτόν τον τόρνο αλλά) με διαφορετικά μεγέθη, υλικά, τροφοδοσία, ταχύτητα, βάθη κοπής και μέθοδος κράτησης της εργασίας κ.λπ. Έτσι, τα συγκεντρωμένα μεγάλα δεδομένα βοηθούν σημαντικά στο συγχρονισμό μιας νέας εργασίας, χωρίς να χρειάζεται πλέον μελέτη χρόνου. Αυτό μειώνει σημαντικά το χρόνο και την εργασία που απαιτείται για την εύρεση του κανονικού χρόνου για μια νέα εργασία .

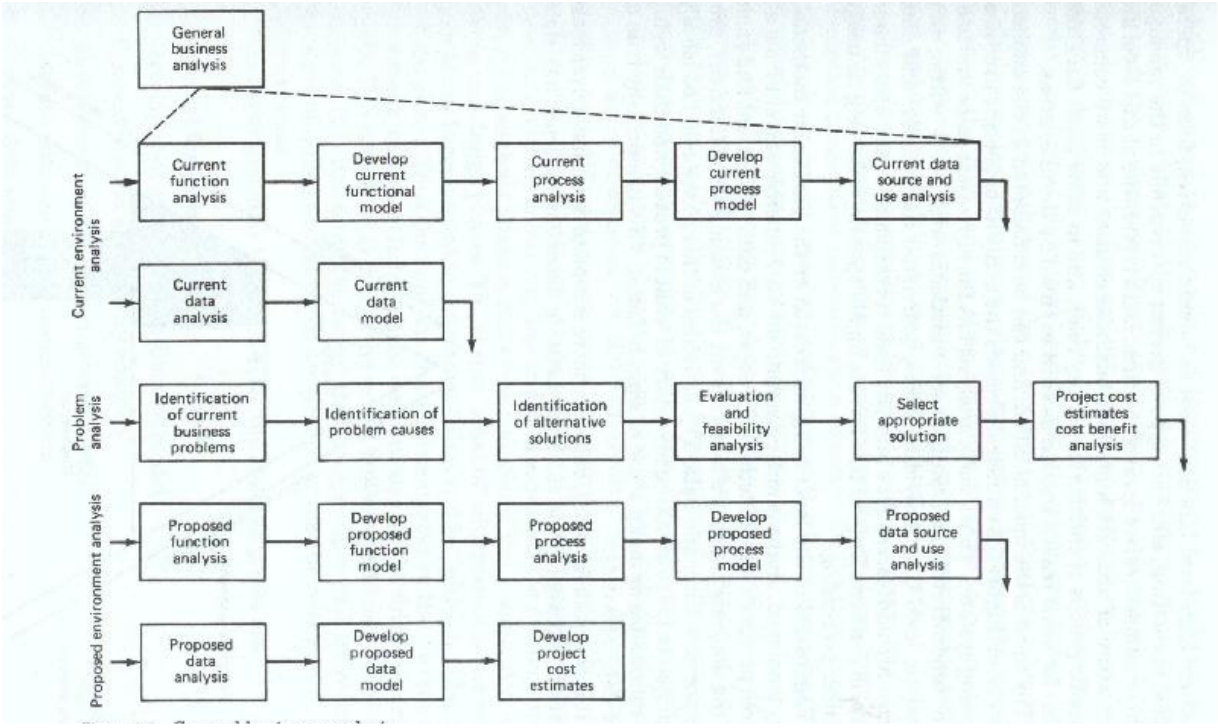
Τα καθολικά τυπικά δεδομένα (Microdata) βασίζονται σε κινήσεις λεπτών (δηλαδή, therbligs-reach, carry, hold κ.λπ.) που εμπλέκονται σε μια λειτουργία και συλλέγονται με μικροσκοπικές μεθόδους. Οι μέθοδοι, βασίζονται στην αρχή ότι όλες οι εργασίες αποτελούνται από πολύ λίγες κινήσεις που ονομάζονται therbligs ή με άλλα λόγια, όλες οι εργασίες μπορούν να χωριστούν σε therbligs.Τα μικροδεδομένα συγκεντρώνουν τον κανονικό χρόνο για έναν κύκλο εργασίας ή μια εργασία αναλύοντας τους βασικούς τύπους κινήσεων (therbligs). Αυτή η ανάλυση πραγματοποιείται από πλαίσιο σε πλαίσιο μελέτης της ταινίας του κύκλου εργασίας που καταγράφεται από κάμερα ταινίας (Micromotion Analysis). Το σύστημα MTM (Method-Time-Measurement) και Work factor είναι παραδείγματα καθολικών τυπικών δεδομένων.Τα μακροδεδομένα ασχολούνται με (μεγάλα) στοιχεία και μικροδεδομένα με (λεπτά) κινήσεις. Τα μακροδεδομένα συλλέγονται από τη μελέτη χρόνου (ας πούμε τη μελέτη χρονόμετρου), ενώ τα μικροδεδομένα είναι το αποτέλεσμα μελέτης και ανάλυσης μικροκινημάτων. αλλά και οι δύο οδηγούν σε κανονικό χρόνο για έναν κύκλο εργασίας (Berente & Seidel,2014).

1.1.2Συστήματα Ανάλυσης Δεδομένων

Ανάλυση συστημάτων, και επομένως αυτές οι δραστηριότητες δεδομένων που σχετίζονται με αυτό είναι ουσιαστικά μια φάση συγκέντρωσης γεγονότων. Εξετάζει τι συμβαίνει στον υπό μελέτη επιχειρηματικό τομέα, αναπτύσσει τεκμηρίωση του τρέχοντος περιβάλλοντος που με τη σειρά του λειτουργεί ως θεμέλιο ή σημείο εκκίνησης για οποιαδήποτε περαιτέρω ανάλυση ή επανασχεδιασμό δραστηριοτήτων. Δεδομένου ότι ο σχεδιασμός συστημάτων είναι ουσιαστικά ένας επανασχεδιασμός του τρέχοντος περιβάλλοντος, πραγματοποιώντας αλλαγές για τη βελτίωση της τρέχουσας αποτελεσματικότητας, αποδοτικότητας και ικανότητας, καθώς και για τη διόρθωση των ελαττωμάτων που μπορεί να υπάρχουν και την προσθήκη ικανότητας που λείπει, η ομάδα ανάλυσης πρέπει να έχει ένα ακριβές σημείο εκκίνησης. Υπάρχουν πολλά επίπεδα

επιχειρηματικής δραστηριότητας. Αυτά τα επίπεδα επιχειρηματικής δραστηριότητας μπορούν να κατηγοριοποιηθούν κατά προσέγγιση σε επίπεδο στρατηγικού (ή προγραμματισμού), σε επίπεδο διαχείρισης (ή παρακολούθησης και ελέγχου) και σε επιχειρησιακό επίπεδο. Κάθε ένα από αυτά τα επίπεδα αντιπροσωπεύει ένα διαφορετικό επίπεδο προοπτικής για την εταιρεία, καθένα από αυτά απαιτεί διαφορετικό επίπεδο συγκέντρωσης δεδομένων και διαφορετικά είδη δεδομένων. Με διαφορετικό τρόπο, ο καθένας έχει διαφορετικές ανάγκες δεδομένων και διαφορετικούς τρόπους προβολής αυτών των δεδομένων. Κάθε επίπεδο χρησιμοποιεί επίσης δεδομένα με διαφορετικούς τρόπους.

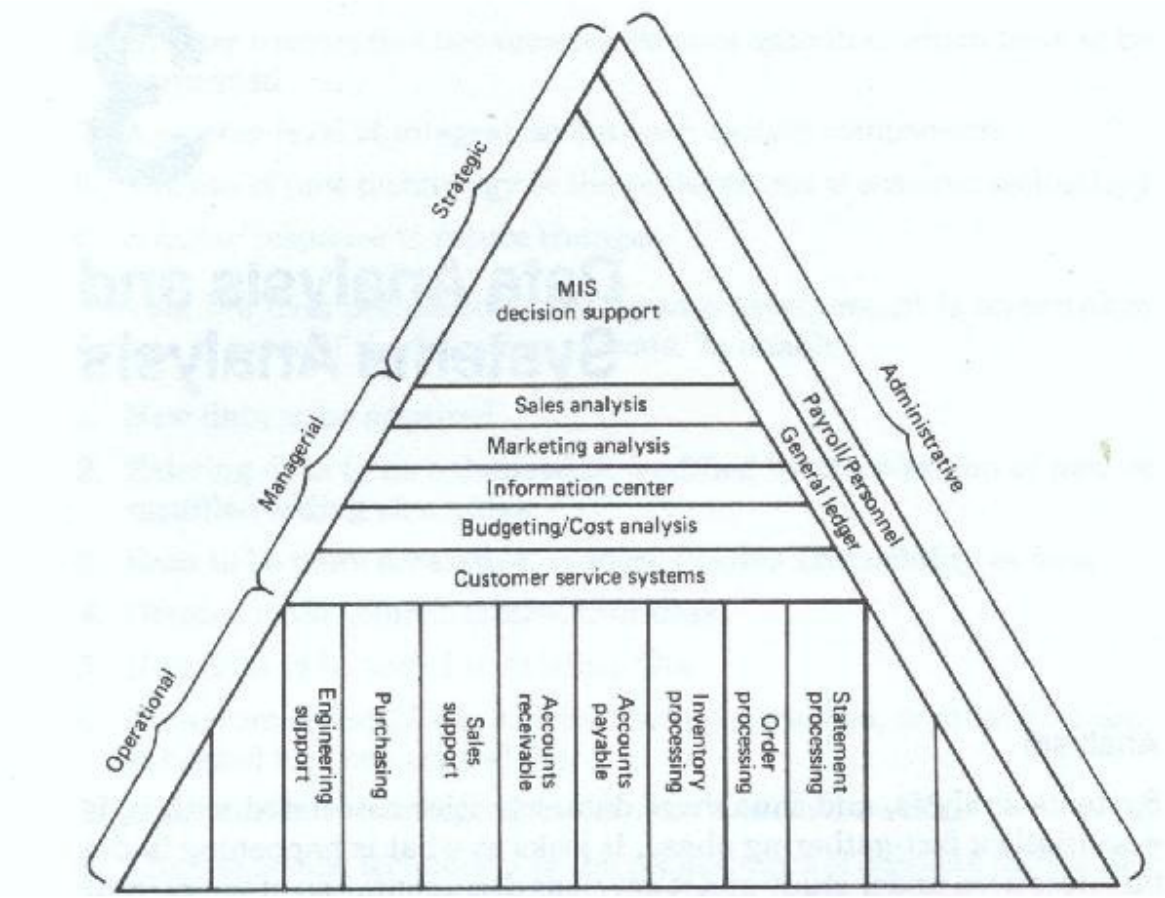
Οι περισσότερες μεθοδολογίες αναγνωρίζουν ορισμένες από αυτές τις διαφορές και επιτρέπουν την ανάλυση σε πολλαπλά επίπεδα. Οι περισσότερες μεθοδολογίες δεν επιτρέπουν τις διαφορές στην προοπτική δεδομένων ή στη χρήση δεδομένων. Σχεδόν όλοι επιτρέπουν ανάλυση πολλαπλών επιπέδων. Αν και μπορεί να υπάρχουν πολλά επίπεδα, ή διαδοχικά πιο λεπτομερείς επαναλήψεις της ανάλυσης, για πρακτικούς σκοπούς οι περισσότερες συζητήσεις περιορίζονται σε τρία (Berente & Seidel, 2014).



Η πρώτη φέρει συνήθως το γενικό ή επιχειρηματικό επίπεδο περιβαλλοντικής ανάλυσης και

επικεντρώνεται στις λειτουργίες, τις διαδικασίες και τα μοντέλα δεδομένων σε ολόκληρη την επιχείρηση. Αυτό έχει κληθεί από ορισμένους, επιχειρηματική ανάλυση. Αυτή η ανάλυση σε αυτό το επίπεδο έχει το ευρύτερο πεδίο όσον αφορά τον αριθμό των λειτουργιών που καλύπτονται και εξετάζει τα υψηλότερα επίπεδα της εταιρείας. Αυτό το επίπεδο αντιστοιχεί στο στρατηγικό επίπεδο της εταιρείας. Αυτό το επίπεδο δεν αντιμετωπίζει τα δεδομένα ως τέτοια, αλλά μάλλον εξετάζει υψηλότερα επίπεδα αφαίρεσης δεδομένων - συνήθως εκείνα τα άτομα, μέρη, πράγματα και ιδέες που φέρουν την ένδειξη επιχειρηματικές οντότητες.

Το τρίτο, είναι το πιο λεπτομερές, και μπορεί να θεωρηθεί ως το επίπεδο εφαρμογής που αφορά την ανάλυση συγκεκριμένων συστημάτων επεξεργασίας χρηστών. Αυτό το επίπεδο ανάλυσης έχει τη στενότερη εμβέλεια και συνήθως επικεντρώνεται σε έναν λειτουργικό τομέα ενός χρήστη. Σε αυτό το αναλυτικό επίπεδο αντιμετωπίζονται μεμονωμένες εργασίες και μεμονωμένες εγγραφές αρχείων και στοιχεία δεδομένων. Αυτό το αναλυτικό επίπεδο αντιστοιχεί στο επιχειρησιακό επίπεδο της εταιρείας. Τα δεδομένα σε αυτό το επίπεδο είναι ιδιαίτερα συγκεκριμένα, αντικατοπτρίζοντας τη χρήση τους στην επεξεργασία χρηστών (Berente & Seidel, 2014).



Ανεξάρτητα από το επίπεδο που εξετάζεται ή τη συγκεκριμένη μεθοδολογία που χρησιμοποιείται, η ανάλυση πρέπει να εξετάσει το τρέχον περιβάλλον και από τις τρεις προοπτικές - λειτουργική, διαδικασία και δεδομένα. Ο σκοπός και οι στόχοι της φάσης ανάλυσης είναι να τεκμηριώσουν τις υπάρχουσες λειτουργίες, διαδικασίες, δραστηριότητες και δεδομένα των χρηστών. Οι δραστηριότητες αυτής της φάσης θα μπορούσαν να εξομοιωθούν χαλαρά με εκείνες της απογραφής του τι υπάρχει και στη συνέχεια προσπαθώντας να εξορθολογίσει ή να κατανοήσει αυτό που βρέθηκε. Αυτό ισχύει ιδιαίτερα για τις δραστηριότητες δεδομένων. Είναι η ανάγκη για περισσότερα, καλύτερα ή πληρέστερα δεδομένα που οδηγούν τα περισσότερα έργα επανασχεδιασμού του συστήματος. Τα περισσότερα υπάρχοντα επιχειρηματικά συστήματα χρησιμοποιούν δεδομένα που αντικατοπτρίζουν παλιές απαιτήσεις, παλιά επιχειρηματικά συστήματα ή παλιές επιχειρηματικές φιλοσοφίες. Αυτό που σχεδόν όλοι οι αναλυτές αποκαλύπτουν καθώς προχωρούν στις δραστηριότητες ανάλυσης είναι ότι πολλές από τις έννοιες που αποτελούν το θεμέλιο της εταιρείας δεν είναι σαφώς κατανοητές, ή χειρότερα έχουν πολλούς ορισμούς. Πολλά από τα δεδομένα, επειδή χρησιμοποιούνται κυρίως υπό τοπικό έλεγχο,

ανέπτυξαν τοπικούς ορισμούς, οι οποίοι με την πάροδο του χρόνου προκάλεσαν εσφαλμένη επικοινωνία μεταξύ των διαφόρων τομέων της εταιρείας. Αυτός ο τοπικός ορισμός προκάλεσε ακόμη μεγαλύτερη εσφαλμένη επικοινωνία μεταξύ των διαφόρων επιπέδων της επιχείρησης και με την πάροδο του χρόνου εμπόδισε ή απέτρεψε εντελώς τις προσπάθειες της διοίκησης στην αναδιοργάνωση.

Οι δραστηριότητες ανάλυσης δεδομένων της φάσης ανάλυσης συστήματος προσπαθούν να εντοπίσουν αυτές τις διαφορές δεδομένων κατά την προετοιμασία για την επίλυσή τους και το σχεδιασμό μιας συνεπούς δομής επικοινωνίας για την εταιρεία. Τα επιχειρηματικά δεδομένα με συνεπή σημασία και κατάλληλη οργάνωση, αποτελούν τον πυρήνα της επιχειρηματικής διαδικασίας επικοινωνίας. Η αρχική ανάλυση ή το απόθεμα προσδιορίζει όλα τα τρέχοντα δεδομένα, όλες τις τρέχουσες πηγές αυτών των δεδομένων, σε όλες τις ιδιότυπες μορφές της, και πιο συχνά από ό, τι όχι, εντοπίζει κενά στα δεδομένα της εταιρείας. Αυτά τα κενά έχουν τη μορφή απαιτήσεων χρήστη για "νέα" δεδομένα. Τα νέα δεδομένα μπορούν στην πραγματικότητα να είναι νέα δεδομένα ή, όπως συμβαίνει συχνότερα, υπάρχουν υπάρχοντα δεδομένα σε διαφορετική μορφή, διαφορετική οργάνωση, διαφορετικά επίπεδα λεπτομέρειας, διαφορετικά χρονικά πλαίσια ή σε διαφορετικούς συνδυασμούς από το τρέχον. Αυτό το απόθεμα του "τι προστίθεται" στις νέες απαιτήσεις και την ανάγκη αναθεώρησης των υπαρχόντων αρχείων και πηγών δεδομένων του χρήστη είναι το κύριο προϊόν της φάσης ανάλυσης και της πρώτης ύλης της εξέτασης και μελέτης, καθώς και φάσεις σχεδιασμού. Η φράση «τι είναι» σε σχέση με τη φάση ανάλυσης έρχεται σε αντίθεση με τη φράση «τι θα είναι» που χαρακτηρίζει τη φάση σχεδιασμού. Έτσι, η φάση ανάλυσης επικεντρώνεται στο τρέχον περιβάλλον, ενώ η φάση σχεδιασμού επικεντρώνεται στο μελλοντικό περιβάλλον. Αυτή η δήλωση ισχύει στο βαθμό που η ανάλυση εξετάζει τα τρέχοντα συστήματα. Είναι αναληθές στο βαθμό που η ανάλυση συγκεντρώνει επίσης απαιτήσεις (ανεκπλήρωτες ανάγκες) που αποτελούν μέρος ή ενδέχεται να γίνουν μέρος του μελλοντικού συστήματος (Berente & Seidel, 2014).

Οι κύριες δραστηριότητες για τη φάση ανάλυσης είναι:

1. Ανάλυση τρέχουσας λειτουργίας

2. Ανάπτυξη του μοντέλου τρέχουσας λειτουργίας

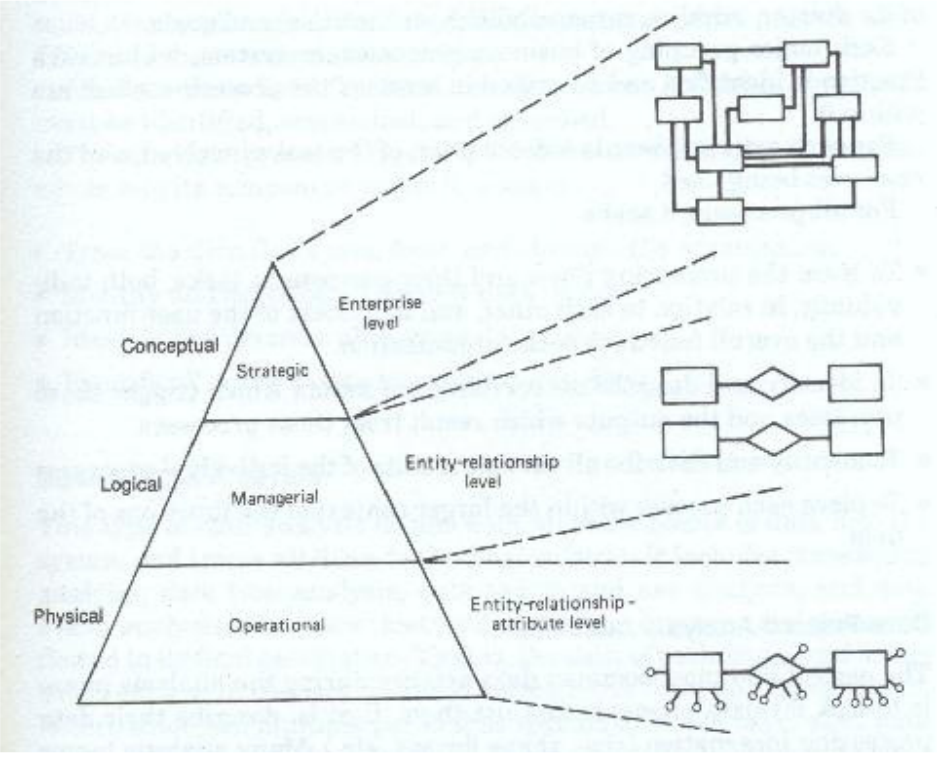
3. Ανάλυση τρέχουσας διαδικασίας και δραστηριότητας

4. Ανάπτυξη του μοντέλου τρέχουσας διαδικασίας

5. Ανάλυση πηγής δεδομένων και χρήσης

6. Ανάλυση τρέχοντων δεδομένων

7. Ανάπτυξη του μοντέλου τρέχοντος δεδομένων



Κάθε μία από τις παραπάνω σημαντικές δραστηριότητες περιλαμβάνει είτε διαδικασία είτε δεδομένα είτε και τα δύο. Προφανώς τα τελευταία τρία είναι δραστηριότητες ανάλυσης που σχετίζονται με δεδομένα. Ανάλογα με τη μεθοδολογία που επιλέχθηκε, η τρίτη και η τέταρτη δραστηριότητα θα μπορούσαν επίσης να θεωρηθούν δραστηριότητες ανάλυσης δεδομένων. Αυτό ισχύει ιδιαίτερα εάν η μέθοδος διαγράμματος ροής δεδομένων έχει επιλεγεί έναντι της μεθόδου

αποσύνθεσης της διαδικασίας. Κάθε μία από αυτές τις δραστηριότητες επικεντρώνεται σε κάποια πτυχή του τρέχοντος περιβάλλοντος ή σε κάποια πτυχή των απαιτήσεων των χρηστών και η κάθε μία αναπτύσσει κάποιο προϊόν (είτε μια αφήγηση είτε ένα γραφικό μοντέλο. Ο σχεδιασμός οποιουδήποτε νέου συστήματος πρέπει να βασίζεται στην κατανόηση του παλαιού συστήματος. Η μόνη εξαίρεση σε αυτό είναι όπου δεν υπάρχει παλιό σύστημα. Είναι η μετάβαση από λογαριασμό σε προσανατολισμό προς τους πελάτες (μια νέα φιλοσοφία ή ένας τρόπος για να κοιτάξετε την επιχείρηση) που οδήγησε πολλές μεγάλες τράπεζες, ασφαλιστικές εταιρείες και χρηματιστηριακές εταιρείες σε μεγάλα αναπτυξιακά έργα να επανασχεδιάσουν και να επαναπροσδιορίσουν τα συστήματά τους. Αυτός ο τύπος σημαντικής αλλαγής προκάλεσε αυτές τις εταιρείες να απαιτήσουν την εκτέλεση εντελώς νέων συνόλων δραστηριοτήτων. Δραστηριότητες με τις οποίες η εταιρεία δεν έχει προηγούμενη εμπειρία.

Πολλές από αυτές τις εταιρείες, ωστόσο, είχαν πολύ δύσκολο χρόνο στην ανάπτυξη αυτών των συστημάτων, ειδικά στα στάδια ανάλυσης και σχεδιασμού. Αυτή η δυσκολία προέκυψε επειδή πολλοί από αυτούς δεν μπορούσαν ή δεν κατάλαβαν ποιοι ή ποιοι ήταν οι πελάτες τους. Αυτό δεν θα μπορούσε να δημιουργήσει έναν ορισμό ενός πελάτη έτσι ώστε να μπορεί να αναγνωρίσει και να περιγράψει έναν. Πολλές από τις εταιρείες είχαν πολλούς διαφορετικούς τύπους πελατών με μη καθορισμένες διαφορές μεταξύ τους. Για παράδειγμα, πολλά μεσιτικά γραφεία δεν μπορούσαν να κάνουν διάκριση μεταξύ των πελατών λιανικής και των θεσμικών πελατών. Πολλοί ασφαλιστικοί οίκοι δεν μπορούσαν να κάνουν διάκριση μεταξύ πελατών και ασφαλισμένων.Το μεγαλύτερο μέρος των αναλυτικών δραστηριοτήτων σε αυτές τις εταιρείες επικεντρώθηκε στον προσδιορισμό του τρόπου και του πού αποκτήθηκαν τα δεδομένα των πελατών αυτήν τη στιγμή και όχι στο τι ήταν ένας πελάτης. Δηλαδή, όχι σε τι είδους άτομα ή οργανισμούς αποτελούσαν τη βάση πελατών, αλλά σε τι είδους πληροφορίες χρειάστηκε η εταιρεία για να αποκτήσει και να διατηρήσει σχετικά με αυτούς τους πελάτες. Χωρίς κατανόηση του τι ήταν ο πελάτης, πολλά από τα σχέδια συστημάτων που προέκυψαν ήταν ανεπαρκή για τις ανάγκες της εταιρείας (Cheng, Qin & Rusu,2012).

Δραστηριότητες σχετικές με τη διαδικασία

- Κατά τη διάρκεια της φάσης ανάλυσης, κάθε λειτουργία περιγράφεται λεπτομερώς ως προς τον χάρτη, την αποστολή, τις ευθύνες, τις αρχές και τους στόχους της.

- Κάθε σημαντική ομαδοποίηση επιχειρηματικών διαδικασιών ή συστήματος, μέσα σε κάθε συνάρτηση προσδιορίζεται και περιγράφεται με όρους των διαδικασιών που βρίσκονται μέσα σε αυτήν.

- Για κάθε δραστηριότητα υπάρχει μια περιγραφή των σχετικών εργασιών και των πόρων που χρησιμοποιούνται.

Για όλες τις διαδικασίες επιδιώκει:

- για τον εντοπισμό των ροών επεξεργασίας και των εργασιών τους, τόσο μεμονωμένα, σε σχέση μεταξύ τους, και στο πλαίσιο της λειτουργίας χρήστη και των συνολικών λειτουργιών του οργανισμού.

- να προσδιορίσει και να περιγράψει όλες τις τρέχουσες εισόδους δεδομένων που ενεργοποιούν αυτές τις διαδικασίες και τις εξόδους που προκύπτουν από αυτές τις διαδικασίες

- για τον προσδιορισμό και την περιγραφή όλων των τρεχόντων αποτελεσμάτων των επιμέρους διαδικασιών για να τοποθετήσει κάθε διαδικασία με το ευρύτερο πλαίσιο των λειτουργιών της εταιρείας.

Δραστηριότητες ανάλυσης που σχετίζονται με δεδομένα

Η ευκολότερη και πιο κοινή δραστηριότητα δεδομένων κατά τη φάση ανάλυσης είναι η αναζήτηση στοιχείων δεδομένων και η καταχώρισή τους. Αυτό περιγράφει την ενσάρκωση επεξεργασίας δεδομένων τους (μέγεθος, μορφή σχήματος κ.λπ.). Πολλές ομάδες ανάλυσης καταλήγουν σε μεγάλες λίστες αυτών των στοιχείων δεδομένων. Αυτά τα στοιχεία, συλλέχθηκαν από αρχείο και αναφορές. Αυτή η δραστηριότητα, αν και έχει κάποια αξία δεν έχει νόημα, εκτός εάν η ομάδα ανάλυσης αφιερώσει επίσης το χρόνο, (πολύς χρόνος) για να καθορίσει κάθε στοιχείο (ή όπως συμβαίνει συχνότερα, η περίπτωση προσδιορίζει κάθε έναν από τους πολλαπλούς και μερικές φορές αντικρουόμενους ορισμούς κάθε στοιχείου) και για να προσδιορίσετε ποια στοιχεία είναι στην πραγματικότητα τα ίδια, ανεξάρτητα από την ετικέτα που τους επισυνάπτεται. Τα

περισσότερα παλαιότερα συστήματα, τόσο χειροκίνητα όσο και αυτοματοποιημένα, έχουν τα δικά τους αρχεία δεδομένων. Αυτά τα αρχεία είναι πολλά ηλεκτρονικά ή χαρτί, ή κάποιος συνδυασμός και των δύο (Cheng, Qin & Rusu,2012).

Αυτά τα αρχεία περιέχουν δεδομένα, σχεδιασμένα για αυτό το επιχειρηματικό σύστημα, και ονομάζονται σύμφωνα με την ιδιοτροπία, τη λογική ή ακόμη και τα πρότυπα της ομάδας που είναι υπεύθυνη για το σχεδιασμό τους. Μεμονωμένα, τα δεδομένα σε αυτά τα αρχεία είναι συνεπή και συνήθως ελάχιστα περιττά, ωστόσο σε συνδυασμό με αρχεία από άλλα συστήματα, τα αποτελέσματα είναι χάος. Σε όλες σχεδόν τις περιπτώσεις, δεν υπάρχουν σαφείς ορισμοί αυτών των στοιχείων δεδομένων, και χωρίς σαφείς ορισμούς, δεν υπάρχει τρόπος να καθοριστεί πότε τα στοιχεία είναι ίδια ή διαφορετικά. Σε πολλές περιπτώσεις χωρίς σαφείς ορισμούς, δεν υπάρχει τρόπος να προσδιοριστεί τι αντιπροσωπεύει το περιεχόμενο του στοιχείου δεδομένων ή ακόμη και τι πρέπει να αντιπροσωπεύει. Εντός της φάσης ανάλυσης, όλες οι πηγές δεδομένων και πληροφοριών πρέπει να προσδιοριστούν και όλες οι χρήσεις στις οποίες χρησιμοποιούνται αυτά τα δεδομένα και πληροφορίες πρέπει επίσης να προσδιοριστούν. Όλες οι φόρμες δεδομένων, οι αναφορές και τα αρχεία πρέπει να ταυτοποιηθούν και να περιγραφθεί το περιεχόμενό τους. Όλες οι ροές δεδομένων μεταξύ συναρτήσεων, διαδικασιών, δραστηριοτήτων και εργασιών πρέπει να εντοπιστούν, να ερευνηθούν και να περιγραφούν (Cheng , Qin & Rusu,2012).

Αυτός ο τύπος ανάλυσης δεδομένων ξεκινά με όλα τα σημεία εισόδου δεδομένων στο σύστημα και εντοπίζει όλες τις ροές μέχρι τις τελικές εξόδους. Περιλαμβάνει ανάλυση συναλλαγών, ανάλυση ροής δεδομένων, ανάλυση πηγής δεδομένων και χρήσης και ανάλυση συμβάντων δεδομένων. Για αυτές τις μεθόδους, κάθε είσοδος δεδομένων στο σύστημα μεταφέρεται στον τελικό του προορισμό. Δηλαδή, η διαδρομή κάθε εισόδου και μερικές φορές κάθε στοιχείο δεδομένων εντοπίζεται μέσω της εταιρείας και τεκμηριώνεται. Όπου υπάρχουν πολλές διαδρομές που μπορούν να λάβουν συγκεκριμένα δεδομένα, κάθε διαδρομή τεκμηριώνεται και τεκμηριώνονται οι συνθήκες υπό τις οποίες αυτή η διαδρομή είναι ή πρέπει να ακολουθηθεί. Αυτός ο τύπος ανάλυσης και τεκμηρίωσης διαφέρει από τις παραδοσιακές τεχνικές δομημένης ανάλυσης, καθώς και διαγράμματα ροής δεδομένων που είναι στην πραγματικότητα διαγράμματα αποσύνθεσης ροής δεδομένων. Εξετάζονται μετασχηματισμοί και χειρισμοί δεδομένων (χρησιμοποιώντας διαγράμματα ροής δεδομένων) και τεκμηριώνονται τα αποτελέσματα. Αυτός ο τύπος ανάλυσης είναι συνήθως αριστερά προς τα δεξιά, καθώς οι είσοδοι απεικονίζονται συνήθως

ως εισερχόμενοι στα αριστερά και βγαίνοντας από τα δεξιά. Οι τεχνικές ανάλυσης ροής δεδομένων χρησιμοποιούνται για την απεικόνιση της ροής μέσω διαδοχικών επιπέδων αποσύνθεσης της διαδικασίας, φτάνοντας τελικά στο επίπεδο εργασιών μονάδας (Demchenko, de Laat & Membrey, 2014).

Οι αναλυτές που εκτελούν αυτήν την ανάλυση πρέπει να προσέχουν να συμπεριλάβουν στην τεκμηρίωσή τους όλα τα αντίγραφα που αποτελούνται από έντυπα εισαγωγής και αναφορές και όλα τα σημεία της εταιρείας ότι τα δεδομένα που περιέχονται σε φόρμες και αναφορές εισάγονται ξανά σε άλλες αναφορές και προσωπικούς υπολογιστές. Αυτός ο τύπος δραστηριότητας είναι ιδιαίτερα διαδεδομένος στα γραφεία προσωπικού όπου η ανάλυση της απόδοσης της εταιρείας πραγματοποιείται για τη διοίκηση της εταιρείας. Η διαδρομή κάθε αντιγράφου κάθε φόρμας πολλαπλών τμημάτων πρέπει να εντοπιστεί και να τεκμηριωθεί, μαζί με τον λόγο για κάθε αντίγραφο, και πρέπει να ληφθούν προσεκτικές σημειώσεις σχετικά με το πότε αυτές οι φόρμες και οι αναφορές και τα αντίγραφα σχολιάζονται με πρόσθετα Η επικύρωση αυτών των μεθόδων απαιτεί από τον αναλυτή και τον χρήστη να εργάζονται πίσω από τις εξόδους στις εισόδους. Για να επιτευχθεί αυτό, κάθε έξοδος ή στοιχείο αποθήκευσης (στοιχεία δεδομένων που αποθηκεύονται σε τρέχοντα αρχεία - επίσης μια έξοδος), εντοπίζεται μέσω των τεκμηριωμένων μετασχηματισμών και διαδικασιών στην τελική πηγή τους. Η επικύρωση εξόδου προς εισαγωγή δεν απαιτεί τη χρήση όλων των εισόδων δεδομένων, ωστόσο όλα τα στοιχεία εξόδου ή αποθηκευμένα δεδομένα που χρησιμοποιούνται πρέπει να έχουν την τελική πηγή εισόδου και να έχουν μία μόνο είσοδο (Demchenko, de Laat , Membrey,2014).

- Η επικύρωση εισόδου σε έξοδο λειτουργεί με αντίστροφο τρόπο. Εδώ, κάθε στοιχείο εισαγωγής εντοπίζεται μέσω των διαδικασιών και των μετασχηματισμών του στο τελικό αποτέλεσμα.

- Επικύρωση εξόδου σε είσοδο ή είσοδος σε έξοδο, η ανάλυση αναζητά στοιχεία δεδομένων πολλαπλής προέλευσης και στοιχεία δεδομένων που αποκτήθηκαν αλλά δεν χρησιμοποιήθηκαν.

Ανάλυση πηγής δεδομένων για χρήση

Αυτή η μέθοδος ανάλυσης δεδομένων προσεγγίζεται σε επίπεδο στοιχείου δεδομένων και αγνοεί τα συγκεκριμένα έγγραφα που μεταφέρουν τα δεδομένα. Η λογική για αυτόν τον τύπο ανάλυσης δεδομένων είναι ότι τα δεδομένα, αν και αρχικά συγκεντρωτικά σε έγγραφα τείνουν να διασκορπίζονται, ή να θραύονται εντός των ροών δεδομένων της εταιρείας. Αντίθετα, μία φορά μέσα στις ροές δεδομένων της εταιρείας, τα δεδομένα τείνουν να συγκεντρώνονται με διαφορετικούς τρόπους. Δηλαδή, τα δεδομένα συγκεντρώνονται σε διαφορετικές συλλογές και από πολλές διαφορετικές πηγές για διάφορους σκοπούς επεξεργασίας. Ορισμένα δεδομένα χρησιμοποιούνται για σκοπούς αναφοράς και μερικά δημιουργούνται ως αποτέλεσμα διαφόρων σταδίων επεξεργασίας και διαφόρων μετασχηματισμών. Το αποτέλεσμα είναι ένας ιστός δεδομένων που μπορεί να χαρτογραφηθεί ανεξάρτητα από τις ροές επεξεργασίας. Τα μοντέλα ροής δεδομένων και τα μοντέλα δεδομένων από τις φάσεις ανάλυσης δεδομένων είναι ιδιαίτερα χρήσιμα εδώ. Ο αναλυτής πρέπει να είναι βέβαιος ότι θα παραπέμψει και επικυρώσει το μοντέλο δεδομένων από το υπάρχον σύστημα (Demchenko, de Laat & Membrey,2014).

Επικύρωση της Ανάλυσης

Αν και δεν είναι μια καθορισμένη φάση στον κύκλο ζωής της ανάπτυξης του συστήματος, η επικύρωση αυτού του σταδίου της ανάλυσης δεδομένων είναι αρκετά σημαντική ώστε αξίζει ξεχωριστή επεξεργασία.

Η ολοκληρωμένη αναλυτική τεκμηρίωση πρέπει να επικυρωθεί για να διασφαλιστεί ότι:

- Όλα τα μέρη συμφωνούν ότι οι όροι όπως παρουσιάζονται στην τεκμηρίωση αντιπροσωπεύουν με ακρίβεια το περιβάλλον

- Ότι τα παραγόμενα έγγραφα περιέχουν δηλώσεις που είναι πλήρεις, ακριβείς και ξεκάθαρες.

- Η επικύρωση των προϊόντων της φάσης ανάλυσης πρέπει να αφορά τις δύο ίδιες πτυχές του περιβάλλοντος: Δεδομένα και επεξεργασία δεδομένων.

Ένα σύστημα είναι ένα πολύπλοκο σύνολο. Τα παλιά συστήματα δεν είναι μόνο περίπλοκα, αλλά σε πολλές περιπτώσεις, είναι ένα συνονθύλευμα διεργασιών, διαδικασιών και εργασιών που συναρμολογήθηκαν με την πάροδο του χρόνου και οι οποίες ενδέχεται να μην ταιριάζουν πλέον σε ένα συνεκτικό σύνολο. Πολλές φορές προέκυψαν ανάγκες που απαιτούσαν τη δημιουργία προσωρινών διαδικασιών για την επίλυση ενός συγκεκριμένου προβλήματος. Με την πάροδο του χρόνου θεσπίζονται αυτές οι πρόχειρες διαδικασίες. Μεμονωμένα μπορεί να έχουν νόημα, και μπορεί ακόμη και να λειτουργήσουν, ωστόσο, στο ευρύτερο πλαίσιο του οργανισμού, είναι ανακριβείς, ελλιπείς και μπερδεμένες.

Οι περισσότεροι οργανισμοί έρχονται αντιμέτωποι με πολλά από αυτά τα συστήματα που είναι τόσο παλιά, και τόσο επιδιορθωμένα, ελλιπή, περίπλοκα και χωρίς έγγραφα που κανείς δεν κατανοεί πλήρως όλες τις περιπλοκές και τα προβλήματα που ενυπάρχουν σε αυτά, πολύ λιγότερο έχει μια εικόνα επισκόπησης ολόκληρου του συνόλου. Η αναπαράσταση του περιβάλλοντος όπως απεικονίζεται από τον αναλυτή μπορεί να είναι η πρώτη φορά που οποιοσδήποτε χρήστης βλέπει το σύνολο των λειτουργικών λειτουργιών. Εάν το περιβάλλον είναι ιδιαίτερα μεγάλο ή περίπλοκο, μπορεί να χρειαστεί τόσο ο χρήστης όσο και ο αναλυτής για να επικυρώσουν την ανάλυση όπως έκανε για να την δημιουργήσουν, αν και αυτό είναι πιθανώς ακραίο. Η επικύρωση επιδιώκει να διασφαλίσει ότι έχουν επιτευχθεί οι στόχοι της ανάλυσης και ότι τα έγγραφα που περιγράφουν τα συστατικά μέρη του συστήματος είναι πλήρη και ακριβή. Προσπαθεί επίσης να διασφαλίσει ότι για κάθε σημαντικό συστατικό, τα αναγνωρισμένα εξαρτήματα εξαρτημάτων αναδημιουργούν ή είναι συνεπή με το σύνολο. Για να χρησιμοποιήσετε μια αναλογία, η διαδικασία ανάλυσης συστήματος και σχεδιασμού συστήματος είναι παρόμοια με τη διάσπαση μιας σπασμένης συσκευής, την επισκευή του ελαττωματικού εξαρτήματος και την επανασύνδεσή της ξανά. Είναι εύκολο να αποσυναρμολογήσετε τη συσκευή, κάπως πιο δύσκολο να απομονώσετε το ελαττωματικό τμήμα και να το επισκευάσετε, και πιο δύσκολο να ξανασυναρμολογήσετε όλα τα κομμάτια ξανά ώστε να λειτουργεί. Το τελευταίο ισχύει ιδιαίτερα όταν το σχήμα για τη συσκευή

λείπει, είναι ελλιπές ή χειρότερα ανακριβές. Εάν κατά τη διάρκεια της διαδικασίας επισκευής αναμένεται νέα, βελτιωμένα ή υποκατάστατα ανταλλακτικά ή εξαρτήματα να αντικαταστήσουν ορισμένα από τα υπάρχοντα ανταλλακτικά ή εάν κάποια από τα υπάρχοντα ανταλλακτικά αφαιρούνται επειδή δεν χρειάζονται πλέον, η έλλειψη κατάλληλης τεκμηρίωσης γίνεται πρόβλημα σχεδόν αδύνατο να ξεπεραστεί. Η τεκμηρίωση που δημιουργήθηκε ως αποτέλεσμα της ανάλυσης είναι παρόμοια με τη δημιουργία ενός σχηματικού σχήματος καθώς αποσυναρμολογείτε τη συσκευή, η διαδικασία επικύρωσης είναι παρόμοια με τη διασφάλιση ότι όλα τα κομμάτια λογιστικοποιούνται στο σχηματικό που δημιουργήσατε. Ωστόσο, επειδή σκοπεύετε να επισκευάσετε και να βελτιώσετε, η τεκμηρίωση που δημιουργείτε δεν πρέπει να περιγράφει μόνο πού ταιριάζει κάθε στοιχείο, αλλά γιατί είναι εκεί, ποια είναι η λειτουργία του και τι εάν υπάρχουν προβλήματα με τον τρόπο που αρχικά κατασκευάστηκε και κατασκευάστηκε (Wu , Zhu, Wu & Ding, 2014).

Η επικύρωση δεν πρέπει μόνο να διασφαλίζει ότι η υπάρχουσα διαδικασία έχει τεκμηριωθεί σωστά, αλλά και ότι, στο μέτρο του δυνατού, έχουν τεκμηριωθεί οι λόγοι ή το σκεπτικό πίσω από τις διαδικασίες. Η επικύρωση των προϊόντων ανάλυσης επιδιώκει να διασφαλίσει ότι ο επαληθευτής βλέπει τι υπάρχει και όχι τι πρέπει να υπάρχει. Η ανάλυση συστημάτων, από τη φύση της, λειτουργεί για τον προσδιορισμό, τον ορισμό και την περιγραφή των διαφόρων συστατικών στοιχείων του συστήματος. Κάθε δραστηριότητα και κάθε έρευνα επιδιώκει να εντοπίσει και να περιγράψει ένα συγκεκριμένο κομμάτι. Το κομμάτι μπορεί να είναι μακρο ή μικρό, αλλά είναι ένα κομμάτι. Αν και είναι συνήθως απαραίτητο να δημιουργηθούν μοντέλα επισκόπησης, αυτά τα επίπεδα επισκόπησης, σε επιχειρησιακό και λειτουργικό επίπεδο, επιδιώκουν μόνο να δημιουργήσουν ένα πλαίσιο ή κατευθυντήριες γραμμές για το κρέας της ανάλυσης, αυτό που επικεντρώνεται στα επιχειρησιακά καθήκοντα. Είναι η λεπτομέρεια στα επιχειρησιακά επίπεδα που μπορούν να επικυρωθούν. Η διαδικασία επικύρωσης τόσο των δεδομένων όσο και της διαδικασίας λειτουργεί σε αυτό το επίπεδο. Κάθε δραστηριότητα, κάθε έξοδος και κάθε συναλλαγή, που προσδιορίζονται στα χαμηλότερα επίπεδα, πρέπει να εντοπίζονται από το τελικό σημείο στο υψηλότερο επίπεδο συγκέντρωσης ή στο σημείο έναρξης. Στην ανάλυση, ο αναλυτής συγκεντρώνει γεγονότα και επιδιώκει να συγκεντρώσει μια εικόνα του τρέχοντος περιβάλλοντος (Wu, Zhu & Wu,2014).

Κατά την επικύρωση, ο αναλυτής ξεκινά με μια κατανόηση του περιβάλλοντος και των εικόνων ή των μοντέλων που έχει κατασκευάσει. Ο στόχος εδώ όμως είναι να καθοριστεί:

- αν η κατανόηση του περιβάλλοντος από τον αναλυτή είναι πλήρης και σωστή
- εάν οι απεικονίσεις του τρέχοντος περιβάλλοντος ταιριάζουν με αυτό που υπάρχει στην πραγματικότητα και ταιριάζουν με την κατανόηση του περιβάλλοντος από τον χρήστη
- Πρέπει να γίνει κατανοητό ότι τα έγγραφα ανάλυσης αντιπροσωπεύουν συνδυασμό τόσο γεγονότων όσο και απόψεων. Είναι επίσης πολύ υποκειμενικά. Βασίζονται σε συνέντευξη, παρατήρηση και αντίληψη. Η επικύρωση επιδιώκει να διασφαλίσει ότι η αντίληψη και η υποκειμενικότητα δεν έχουν παραμορφώσει τα γεγονότα.

Η δημιουργία διαγραμματικών μοντέλων, σε επίπεδο λειτουργικής, διεργασίας και δεδομένων, διευκολύνει σε μεγάλο βαθμό τη διαδικασία επικύρωσης. Όπου αυτά τα μοντέλα έχουν αντληθεί από τις αναλυτικές πληροφορίες και όπου συμπληρώνονται από λεπτομερείς αφηγήσεις, η διαδικασία επικύρωσης μπορεί να μειωθεί σε δύο στάδια.

Στάδιο πρώτο - Παραπομπή των διαγραμμάτων στις αφηγήσεις, για να διασφαλιστεί ότι:

- ο καθένας λέει το ίδιο πράγμα
- κάθε σχήμα στο διάγραμμα έχει μια αντίστοιχη αφήγηση και το αντίστροφο
- Τα διαγράμματα δεν περιέχουν μη τερματισμένες ροές, ότι δεν υπάρχουν αποσυνδεδεμένα σχήματα, δεν υπάρχουν αμφιλεγόμενες συνδέσεις, ότι όλα τα σχήματα και όλες οι συνδέσεις φέρουν σαφή και πλήρη σήμανση και παραπέμπουν στις συνοδευτικές αφηγήσεις τα διαγράμματα είναι συνεπή μέσα τους, δηλαδή ότι τα διαγράμματα δεδομένων περιέχουν μόνο δεδομένα, τα διαγράμματα διεργασίας περιέχουν μόνο διεργασίες και ότι τα μοντέλα λειτουργιών περιέχουν μόνο συναρτήσεις. Κάθε διάγραμμα φέρει σαφή σήμανση και ότι έχει παρασχεθεί ένας θρύλος που προσδιορίζει την έννοια

κάθε συμβόλου που χρησιμοποιείται όταν η πολυπλοκότητα του περιβάλλοντος χρήστη είναι τέτοια που τα μοντέλα πρέπει να κατατμηθούν σε πολλά μέρη, καθένα από τα μέρη πρέπει να έχει συνεπείς ετικέτες, τίτλους και θρύλους, οι σύνδεσμοι μεταξύ των μερών πρέπει να είναι συνεπείς στις εμπρός και πίσω αναφορές τους και τα ονόματα των Τα σχήματα που εμφανίζονται σε πολλά μέρη πρέπει να είναι συνεπή.

Δεύτερο στάδιο - Διασταυρούμενη παραπομπή στα μοντέλα. Αυτό περιλαμβάνει τη διασφάλιση ότι:

- οι διαδικασίες αναφέρονται πίσω στις λειτουργίες του κατόχου τους, και οι συναρτήσεις αναφέρονται στις διαδικασίες των συστατικών τους
- οποιεσδήποτε σχέσεις που εντοπίζονται μεταξύ οντοτήτων δεδομένων έχουν μια αντίστοιχη διαδικασία που τη συλλαμβάνει και τη διατηρεί
- Όλα τα δεδομένα που προσδιορίζονται ως μέρος του μοντέλου δεδομένων της εταιρείας έχουν μια αντίστοιχη διαδικασία που τη συλλαμβάνει, την επικυρώνει, τη διατηρεί, τη διαγράφει και τα χρησιμοποιεί.
- όλες οι προβολές επεξεργασίας των δεδομένων λογίζονται στα μοντέλα δεδομένων.
- Οι αναφορές σε δεδομένα ή διαδικασίες εντός των μεμονωμένων μοντέλων είναι συνεπείς μεταξύ των μοντέλων
- Όλα τα δεδομένα που αναμένονται από τις διάφορες διαδικασίες λογίζονται στα μοντέλα δεδομένων.

2.Big Data

2.1Ορισμός

Το μέγεθος των δεδομένων που παράγονται και μοιράζονται μεταξύ των επιχειρήσεων, των δημόσιων διοικήσεων πολλών βιομηχανικών και μη κερδοσκοπικών τομέων και της επιστημονικής έρευνας έχει αυξηθεί ανυπολόγιστα . Αυτά τα δεδομένα περιλαμβάνουν περιεχόμενο κειμένου (π.χ. δομημένο, ημι-δομημένο και αδόμητο), σε περιεχόμενο πολυμέσων (π.χ. βίντεο, εικόνες, ήχο) σε πολλαπλές πλατφόρμες (π.χ. επικοινωνίες από μηχανή σε μηχανή, ηλεκτρονικά συστήματα και το Διαδίκτυο των πραγμάτων.. Οι Dobre και Xhafa αναφέρουν ότι κάθε μέρα όλος ο κόσμος παράγει περίπου 2,5 quintillion bytes δεδομένων (δηλαδή 1 exabyte ισούται με 1 bytes quintillion ή 1 exabyte ισούται με 1 δισεκατομμύριο gigabytes), με το 90% αυτών των δεδομένων να παράγεται στον κόσμο να είναι αδόμητο.Οι Gantz και Reinsel ισχυρίζονται ότι έως το 2020 θα έχουν δημιουργηθεί, μιμηθεί και καταναλωθεί πάνω από 40 Zettabytes (ή 40 τρισεκατομμύρια gigabytes) δεδομένων. Με αυτή τη συντριπτική πλειάδα πολύπλοκων και ετερογενών δεδομένων που χύνονται από οποιαδήποτε στιγμή, οποιαδήποτε στιγμή, και οποιαδήποτε συσκευή, υπάρχει αναμφισβήτητα μια εποχή Big Data - ένα φαινόμενο που επίσης αναφέρεται ως Απορροή Δεδομένων . Το δυναμικό της BD είναι εμφανές καθώς έχει συμπεριληφθεί στις 10 κορυφαίες στρατηγικές τεχνολογικές τάσεις της Gartner για το 2013 και στις 10 κορυφαίες τάσεις της τεχνολογίας για τα επόμενα πέντε χρόνια. Είναι τόσο ζωτικής σημασίας όσο η νανοτεχνολογία και ο κβαντικός υπολογισμός στην παρούσα εποχή. Στην ουσία, το BD είναι το τεκμήριο της ανθρώπινης ταυτότητας καθώς και της συλλογικής νοημοσύνης που παράγεται και διανέμεται κυρίως μέσω του τεχνολογικού περιβάλλοντος, όπου σχεδόν τίποτα και όλα μπορούν να τεκμηριωθούν, να μετρηθούν και να ληφθούν ψηφιακά, μεταβάλλοντας έτσι σε δεδομένα.Σύμφωνα με την έννοια της επεξεργασίας δεδομένων και τις συνεχώς αυξανόμενες τεχνολογικές εξελίξεις, οι υποστηρικτές υποστηρίζουν ότι στο μέλλον θα δημιουργηθεί και θα μοιραστεί η πλειοψηφία των δεδομένων μέσω μηχανών, καθώς οι μηχανές επικοινωνούν μεταξύ

τους μέσω δικτύων δεδομένων. Ανεξάρτητα από το πού προέρχεται από και το μοιράζεται, με την πραγματικότητα του BD έρχεται η πρόκληση να το αναλύσουμε με τρόπο που φέρνει μεγάλη αξία . Με τόσο μεγάλη αξία που διαμένουν στο εσωτερικό του, η BD θεωρείται ως το σημερινό ψηφιακό πετρέλαιο συμπεριλαμβανομένης της νέας πρώτης ύλης του 21ου αιώνα. Η κατάλληλη επεξεργασία και διαχείριση δεδομένων θα μπορούσε να εκθέσει νέες γνώσεις και να διευκολύνει την έγκαιρη ανταπόκριση στις αναδυόμενες ευκαιρίες και προκλήσεις . Ωστόσο, η αύξηση των δεδομένων σε όγκους στον ψηφιακό κόσμο φαίνεται να εξουδετερώνει την πρόοδο των πολλών υφιστάμενων υπολογιστικών υποδομών. Οι καθιερωμένες τεχνολογίες επεξεργασίας δεδομένων, όπως η βάση δεδομένων και η αποθήκη δεδομένων, καθίστανται ανεπαρκείς, δεδομένης της ποσότητας των δεδομένων που παράγει σήμερα ο κόσμος. Η τεράστια ποσότητα δεδομένων πρέπει να αναλυθεί με επαναληπτικό και χρονικά ευαίσθητο τρόπο. Με τη χρήση προηγμένων τεχνολογιών ανάλυσης BD (π.χ. βάσεις δεδομένων NoSQL, BigQuery, MapReduce, Hadoop, WibiData και Skytree) μπορούν να επιτευχθούν καλύτερες γνώσεις για τη βελτίωση των επιχειρηματικών στρατηγικών και της διαδικασίας λήψης αποφάσεων σε κρίσιμους τομείς όπως η υγειονομική περίθαλψη, , συμβόλαια ενέργειας, και πρόβλεψη φυσικής καταστροφής, για να αναφέρουμε μόνο μερικά (Wu , Zhu & Wu,2014).

Όπως είναι προφανές, πολλά έχουν γραφτεί για το φαινόμενο της BD. Η πλειοψηφία των αντικειμένων ακαδημαϊκής έρευνας που ανασκοπήθηκαν είναι αναλυτικά στη φύση που είτε εστιάζει στη χρήση πειραμάτων, προσομοιώσεων, αλγορίθμων ή τεχνικών μαθηματικών μοντέλων για την αντιμετώπιση των δεδομένων. Ανεξάρτητα από την ερευνητική τους προσέγγιση, αυτά τα άρθρα παρουσιάζουν τα δεδομένα ως πηγή που όταν διοχετεύεται κατάλληλα, επεξεργάζεται και αναλύεται, έχει τη δυνατότητα να παράγει νέες γνώσεις, προτείνοντας έτσι καινοτόμες και πρακτικές γνώσεις για τις επιχειρήσεις . Υπάρχει ένας διαρκώς αυξανόμενος λόγος που προσφέρει τόσο μεγάλες ευκαιρίες όσο και μεγάλες προκλήσεις μέσω της πληθώρας πηγών από διαφορετικούς τομείς. από επιχειρήσεις έως επιστήμες. Για παράδειγμα, οι ευκαιρίες περιλαμβάνουν τη δημιουργία αξίας , πλούσια επιχειρησιακή πληροφόρηση για επιχειρηματικές αποφάσεις με καλύτερη επίγνωση και υποστήριξη στην ενίσχυση της προβολής και της ευελιξίας της αλυσίδας εφοδιασμού.

Από την άλλη πλευρά, οι προκλήσεις είναι σημαντικές όπως η πολυπλοκότητα ολοκλήρωσης των δεδομένων, η έλλειψη εξειδικευμένων προσωπικών και επαρκών πόρων θέματα ασφάλειας και προστασίας της ιδιωτικής ζωής , ανεπαρκής υποδομή και ασήμαντη αρχιτεκτονική αποθήκης

δεδομένων και συγχρονισμός μεγάλων δεδομένων . Οι υποστηρικτές όπως οι Sandhu και Sood αντιλαμβάνονται ότι η δυνητική αξία της BD δεν μπορεί να ανακαλυφθεί με απλή στατιστική ανάλυση. Οι Zhang, Liu et αl. υποστηρίζουν αυτήν την προοπτική και δηλώνουν ότι για να αντιμετωπιστούν οι προκλήσεις της BD, η προηγμένη BDA απαιτεί εξαιρετικά αποτελεσματικές, κλιμακούμενες και ευέλικτες τεχνολογίες για την αποτελεσματική διαχείριση σημαντικών ποσών δεδομένων - ανεξάρτητα από τον τύπο της μορφής δεδομένων.

Αν και τα οφέλη της πληθώρας δεδομένων είναι πραγματικά και ουσιαστικά, εξακολουθούν να υπάρχουν πολλές προκλήσεις που πρέπει να αντιμετωπιστούν για την πλήρη αξιοποίηση του δυναμικού τους. Ορισμένες από αυτές τις προκλήσεις είναι συνάρτηση των χαρακτηριστικών της BD, μερικές από τις υπάρχουσες μεθόδους και μοντέλα ανάλυσης, και μερικά μέσα από τους περιορισμούς του σημερινού συστήματος επεξεργασίας δεδομένων. Σωζόμενες μελέτες σχετικές με τη διαχείριση δεδομένων και τις προκλήσεις τους έχουν δοθεί προσοχή στις δυσκολίες κατανόησης της έννοιας των δεδομένων , τη λήψη αποφάσεων από ό, τι τα δεδομένα παράγονται και συλλέγονται , τα θέματα της προστασίας της ιδιωτικής ζωής και δεοντολογικές εκτιμήσεις σχετικές με την εξόρυξη τέτοιων δεδομένων . Ο Tole υποστηρίζει ότι η οικοδόμηση μιας βιώσιμης λύσης για μεγάλα και πολύπλευρα δεδομένα είναι μια πρόκληση που οι επιχειρήσεις συνεχώς μαθαίνουν και στη συνέχεια εφαρμόζουν νέες προσεγγίσεις. Για παράδειγμα, ένα από τα μεγαλύτερα προβλήματα σχετικά με το BD είναι το υψηλό κόστος της υποδομής. Ο εξοπλισμός υλικού είναι πολύ ακριβός ακόμη και με τη διαθεσιμότητα τεχνολογιών υπολογιστικού νέφους (Wu, Zhu & Wu,2014).

Επιπλέον, για να ταξινομηθούν τα δεδομένα, έτσι ώστε να μπορούν να κατασκευαστούν πολύτιμες πληροφορίες, απαιτούνται συχνά ανθρώπινες αναλύσεις. Ενώ οι τεχνολογίες πληροφορικής που απαιτούνται για τη διευκόλυνση αυτών των δεδομένων διατηρούν το ρυθμό τους, οι ανθρώπινοι εμπειρογνώμονες και οι ηγέτες των επιχειρηματικών ταλέντων που απαιτούν τη μόχλευση των δεδομένων καθυστερούν, αυτό αποδεικνύεται μια ακόμη μεγάλη πρόκληση.

Όπως αναφέρουν οι Akerkar και Zicari , οι μεγάλες προκλήσεις της σχεδιάσης και εφαρμογής δεδομένων μπορούν να ομαδοποιηθούν σε τρεις κύριες κατηγορίες, με βάση τον κύκλο ζωής των δεδομένων: Προκλήσεις σε δεδομένα, διαδικασίες και διαχείριση:

- Οι προκλήσεις της διαδικασίας σχετίζονται με τη σειρά τεχνικών: πώς να συλλαμβάνει δεδομένα, πώς να ενσωματώνει δεδομένα, πώς να μετασχηματίζει τα δεδομένα, πώς να επιλέγει το σωστό μοντέλο για ανάλυση και πώς να παρέχει τα αποτελέσματα. Οι προκλήσεις της διαχείρισης αφορούν, για παράδειγμα, την προστασία της ιδιωτικής ζωής, την ασφάλεια, τη διακυβέρνηση και τις ηθικές πτυχές

- Για να διευκολυνθεί η λήψη αποφάσεων βάσει τεκμηρίων, οι οργανισμοί χρειάζονται αποτελεσματικές μεθόδους για τη διεκπεραίωση μεγάλων ποσοτήτων διαφόρων δεδομένων σε σημαντικές κατανοήσεις . Οι δυνατότητες χρήσης των δεδομένων είναι ατελείωτες, αλλά περιορίζονται από τη διαθεσιμότητα τεχνολογιών, εργαλείων και δεξιοτήτων. Σύμφωνα με τους Labrinidis και Jagadish, η BDA αναφέρεται σε μεθόδους που χρησιμοποιούνται για την εξέταση και την επίτευξη της διάνοιας από τα μεγάλα σύνολα δεδομένων. Η δυνητική τιμή του BD επιλύεται απλά όταν αξιοποιείται η διαδικασία λήψης αποφάσεων κίνησης. Οι εκτεταμένες έρευνες έχουν δείξει ότι οι επιχειρήσεις μπορούν να επιτύχουν ουσιαστική αξία και ανταγωνιστικό πλεονέκτημα από τη λήψη αποτελεσματικών αποφάσεων βάσει δεδομένων. Οι Davenport και Dyche υπογραμμίζουν ότι οι μεγάλες οργανώσεις συγκεντρώνουν τακτικά BD και εκμεταλλεύονται τα αναλυτικά στοιχεία για τη στήριξη στη λήψη αποφάσεων στο πλαίσιο των συνήθων διαδικασιών τους και οι MME είναι αυτοί που επί του παρόντος αγωνίζονται να ενισχύσουν τις αποφάσεις της ανώτατης διοίκησης προσθέτοντας περισσότερα δεδομένα για τη διαδικασία ανάλυσης. Η ευθυγράμμιση των ανθρώπων, της τεχνολογίας και των οργανωτικών πόρων για να γίνουν μια εταιρεία με βάση δεδομένα είναι προβληματική. Δεδομένου ότι η μέθοδος διαχείρισης δεδομένων μπορεί να ενισχύσει τη λήψη αποφάσεων και να αυξήσει την οργανωτική παραγωγή, αυτό είναι δυνατό όταν χρησιμοποιείται μια επιλογή αναλυτικών μεθόδων για την εξαγωγή της αίσθησης από τα δεδομένα.

Οι περιγραφικές αναλύσεις εξετάζουν λεπτομερώς τα δεδομένα και τις πληροφορίες για τον προσδιορισμό της τρέχουσας κατάστασης μιας επιχειρηματικής κατάστασης, με τρόπο που καθιστά προφανείς τις εξελίξεις, τα πρότυπα και τις εξαιρέσεις, με τη μορφή τυποποιημένων εκθέσεων, εκθέσεων ad hoc και προειδοποιήσεων. Οι αναλυτικές αναλύσεις αφορούν την ανίχνευση δεδομένων για την πιστοποίηση / απόρριψη επιχειρηματικών προτάσεων, για

παράδειγμα, αναλύσεων σε στατιστικά στοιχεία, στατιστικής ανάλυσης, ανάλυσης παραγόντων. Οι προγνωστικές αναλύσεις αφορούν την πρόβλεψη και τη στατιστική μοντελοποίηση για τον προσδιορισμό των μελλοντικών δυνατοτήτων.Οι προδιαγραφικές αναλύσεις αφορούν τη βελτιστοποίηση και τις τυχαίες δοκιμές για να εκτιμηθεί ο τρόπος με τον οποίο οι επιχειρήσεις ενισχύουν τα επίπεδα των υπηρεσιών τους μειώνοντας παράλληλα τα έξοδα. Οι προληπτικές αναλύσεις αφορούν την ικανότητα λήψης προληπτικών ενεργειών σε γεγονότα που ενδέχεται να επηρεάσουν ανεπιθύμητα την οργανωτική απόδοση, για παράδειγμα, εντοπίζοντας τους πιθανούς κινδύνους και προτείνουν στρατηγικές μετριασμού πολύ νωρίτερα στο χρόνο. Πολλοί μελετητές υποστηρίζουν ότι αυτοί οι τύποι αναλυτικών μεθόδων υποστηρίζουν τη βελτίωση της λήψης αποφάσεων και των οργανωτικών επιδόσεων καθιστώντας τα πάντα πιο διαφανή και ποσοτικοποιήσιμα, ενώ παράλληλα ανακαλύπτουν περαιτέρω ασυνέπειες καθώς και πιθανές ανησυχίες και ευκαιρίες (Diebold,2012).

Η ανάλυση μεγάλων δεδομένων είναι η συχνά περίπλοκη διαδικασία εξέτασης μεγάλων και ποικίλων συνόλων δεδομένων ή μεγάλων δεδομένων , για να αποκαλυφθούν πληροφορίες - όπως κρυμμένα μοτίβα, άγνωστοι συσχετισμοί, τάσεις αγοράς και προτιμήσεις πελατών - που μπορούν να βοηθήσουν τους οργανισμούς να λάβουν ενημερωμένες επιχειρηματικές αποφάσεις. Σε ευρεία κλίμακα, δεδομένα και τεχνολογίες και τεχνικές παρέχουν ένα μέσο για την ανάλυση συνόλων δεδομένων και την εξαγωγή συμπερασμάτων σχετικά με αυτά που βοηθούν τις επιχειρήσεις να προβαίνουν σε ενημερωμένες επιχειρηματικές αποφάσεις. Τα ερωτήματα επιχειρηματικής ευφυΐας (BI) απαντούν σε βασικές ερωτήσεις σχετικά με τις λειτουργίες και τις επιδόσεις των επιχειρήσεων. Το Big data analytics είναι μια μορφή προηγμένων αναλυτικών στοιχείων , η οποία περιλαμβάνει σύνθετες εφαρμογές με στοιχεία όπως μοντέλα πρόβλεψης , στατιστικούς αλγόριθμους και τι γίνεται αν η ανάλυση τροφοδοτείται από συστήματα ανάλυσης υψηλής απόδοσης. Η ανάλυση μεγάλων δεδομένων είναι η χρήση προηγμένων αναλυτικών τεχνικών έναντι πολύ μεγάλων, διαφορετικών συνόλων δεδομένων που περιλαμβάνουν δομημένα, ημι-δομημένα και μη δομημένα δεδομένα, από διαφορετικές πηγές και σε διαφορετικά μεγέθη από terabyte έως zettabytes (Diebold,2012).

Τα μεγάλα δεδομένα είναι ένας όρος που εφαρμόζεται σε σύνολα δεδομένων των οποίων το μέγεθος ή ο τύπος υπερβαίνει την ικανότητα των παραδοσιακών σχεσιακών βάσεων δεδομένων να συλλάβουν, να διαχειρίζονται και να επεξεργάζονται τα δεδομένα με χαμηλό λανθάνοντα

χρόνο. Τα μεγάλα δεδομένα έχουν ένα ή περισσότερα από τα ακόλουθα χαρακτηριστικά: υψηλή ένταση, υψηλή ταχύτητα ή υψηλή ποικιλία. Η τεχνητή νοημοσύνη (AI), τα κινητά, τα κοινωνικά και το Διαδίκτυο των πραγμάτων (IoT) οδηγούν την πολυπλοκότητα των δεδομένων μέσω νέων μορφών και πηγών δεδομένων. Για παράδειγμα, τα μεγάλα δεδομένα προέρχονται από αισθητήρες, συσκευές, βίντεο / ήχο, δίκτυα, αρχεία καταγραφής, εφαρμογές συναλλαγών, ιστούς και μέσα κοινωνικής δικτύωσης - μεγάλο μέρος των οποίων δημιουργήθηκε σε πραγματικό χρόνο και σε πολύ μεγάλη κλίμακα. Η ανάλυση μεγάλων δεδομένων επιτρέπει σε αναλυτές, ερευνητές και επιχειρηματικούς χρήστες να λαμβάνουν καλύτερες και ταχύτερες αποφάσεις χρησιμοποιώντας δεδομένα που προηγουμένως δεν ήταν προσβάσιμα ή αχρησιμοποίητα. Οι επιχειρήσεις μπορούν να χρησιμοποιήσουν προηγμένες τεχνικές αναλυτικών στοιχείων, όπως αναλυτικά κείμενα, μηχανική εκμάθηση, αναλυτικά στοιχεία πρόβλεψης, εξόρυξη δεδομένων, στατιστικά στοιχεία και επεξεργασία φυσικής γλώσσας για να αποκτήσουν νέες πληροφορίες από πηγές δεδομένων που δεν έχουν αξιοποιηθεί προηγουμένως ανεξάρτητα ή μαζί με υπάρχοντα εταιρικά δεδομένα (Diebold,2012).

2.2Ιστορία

Το 90% των διαθέσιμων δεδομένων δημιουργήθηκε τα τελευταία δύο χρόνια και ο όρος Big Data ήταν γύρω στο 2005, όταν κυκλοφόρησε από την O'Reilly Media το 2005. Ωστόσο, η χρήση των Big Data και η ανάγκη κατανόησης όλων των διαθέσιμων τα δεδομένα ήταν πολύ περισσότερο. Στην πραγματικότητα, τα πρώτα αρχεία χρήσης δεδομένων για τον εντοπισμό και τον έλεγχο των επιχειρήσεων χρονολογούνται από 7.000 χρόνια πριν, όταν εισήχθη η λογιστική στη Μεσοποταμία προκειμένου να καταγραφεί η ανάπτυξη των καλλιεργειών και των κοπαδιών. Οι λογιστικές αρχές συνέχισαν να βελτιώνονται και το 1663 ο Τζον Γκράουντκατέγραψε και εξέτασε όλες τις πληροφορίες σχετικά με τους ρόλους θνησιμότητας στο Λονδίνο. Ήθελε να αποκτήσει μια κατανόηση και να χτίσει ένα σύστημα προειδοποίησης για τη συνεχιζόμενη πανούκλα. Στο πρώτο καταγεγραμμένο αρχείο της ανάλυσης στατιστικών δεδομένων, συγκέντρωσε τα ευρήματά του στο βιβλίο Natural and Political Observations Made on the Bills of Mortality, το οποίο παρέχει μεγάλες πληροφορίες για τις αιτίες του θανάτου τον δέκατο έβδομο αιώνα. Λόγω της δουλειάς του, ο Γκράουντ μπορεί να θεωρηθεί ο πατέρας των

στατιστικών. Από εκεί και πέρα, οι λογιστικές αρχές βελτιώθηκαν, αλλά δεν συνέβη τίποτα θεαματικό. Μέχρι τον 20ο ^{ου} αιώνα ξεκίνησε η εποχή της πληροφορίας. Η πρώτη ανάμνηση των σύγχρονων δεδομένων είναι από το 1887 όταν ο Herman Hollerith εφηύρε μια υπολογιστική μηχανή που θα μπορούσε να διαβάσει τρύπες που έχουν διατρυπηθεί σε χάρτινες κάρτες για να οργανώσει δεδομένα απογραφής. Το πρώτο μεγάλο έργο δεδομένων δημιουργήθηκε το 1937 και διατάχθηκε από τη διοίκηση του Franklin D. Roosevelt στις ΗΠΑ. Αφού ο νόμος περί κοινωνικής ασφάλισης έγινε νόμος το 1937, η κυβέρνηση έπρεπε να παρακολουθεί τη συνεισφορά από 26 εκατομμύρια Αμερικανούς και περισσότερους από 3 εκατομμύρια εργοδότες . Η IBM ανέλαβε τη σύμβαση να αναπτύξει μηχανή ανάγνωσης καρτών διάτρησης για αυτό το τεράστιο έργο τήρησης βιβλίων. Η πρώτη μηχανή επεξεργασίας δεδομένων εμφανίστηκε το 1943 και αναπτύχθηκε από τους Βρετανούς για να αποκρυπτογραφήσει τους Ναζί κώδικες κατά τη διάρκεια του Β 'Παγκοσμίου Πολέμου. Αυτή η συσκευή, με την ονομασία Colossus , έψαξε μοτίβα σε αναχαιτισμένα μηνύματα με ρυθμό 5.000 χαρακτήρων ανά δευτερόλεπτο. Με αυτόν τον τρόπο μειώνοντας το έργο από εβδομάδες σε μόνο ώρες (Diebold,2012).

Το 1952 η Εθνική Υπηρεσία Ασφαλείας (NSA) έχει δημιουργηθεί και εντός 10 ετών συνάπτει περισσότερους από 12.000 κρυπτολόγους . Αντιμετωπίζουν υπερφόρτωση πληροφοριών κατά τον Ψυχρό Πόλεμο καθώς αρχίζουν να συλλέγουν και να επεξεργάζονται αυτόματα σήματα πληροφοριών. Το 1965 η Ενωμένη Κυβέρνηση αποφάσισε να κατασκευάσει το πρώτο κέντρο δεδομένων για να αποθηκεύσει πάνω από 742 εκατομμύρια φορολογικές δηλώσεις και 175 εκατομμύρια σύνολα δακτυλικών αποτυπωμάτων μεταφέροντας όλα αυτά τα αρχεία σε μαγνητική ταινία υπολογιστή που έπρεπε να αποθηκευτεί σε μία μόνο τοποθεσία. Το έργο αργότερα αποσύρθηκε από φόβο για το «Big Brother», αλλά είναι γενικά αποδεκτό ότι ήταν η αρχή της εποχής ηλεκτρονικής αποθήκευσης δεδομένων.

Το 1989 Βρετανός επιστήμονας υπολογιστών Tim Berners-Lee επινόησε τελικά τον Παγκόσμιο Ιστό. Ήθελε να διευκολύνει την ανταλλαγή πληροφοριών μέσω ενός συστήματος «υπερκειμένου». Λίγο θα μπορούσε να ξέρει αυτή τη στιγμή τον αντίκτυπο της εφεύρεσής του. Από τη δεκαετία του '90 η δημιουργία δεδομένων ενισχύεται καθώς όλο και περισσότερες συσκευές συνδέονται στο Διαδίκτυο. Το 1995 κατασκευάστηκε ο πρώτος υπερ-υπολογιστής , ο οποίος μπόρεσε να κάνει όσο το δυνατόν περισσότερη δουλειά σε ένα δευτερόλεπτο από ότι μια

αριθμομηχανή που χειρίζεται ένα άτομο μπορεί να κάνει σε 30.000 χρόνια. Το 2005 ο Roger Mougalas από την O'Reilly Media επινόησε τον όρο Big Data για πρώτη φορά, μόνο ένα χρόνο μετά τη δημιουργία του όρου Web 2.0. Αναφέρεται σε ένα μεγάλο σύνολο δεδομένων που είναι σχεδόν αδύνατο να διαχειριστεί και να επεξεργαστεί χρησιμοποιώντας παραδοσιακά εργαλεία επιχειρηματικής ευφυΐας.Το 2005 είναι επίσης η χρονιά που δημιουργήθηκε το Hadoop από το Yahoo! χτισμένο πάνω από το MapReduce της Google . Στόχος του ήταν να ευρετηριάσει ολόκληρο τον Παγκόσμιο Ιστό και στις μέρες μας το Hadoop ανοιχτού κώδικα χρησιμοποιείται από πολλούς οργανισμούς για να ξεπεράσει τεράστιες ποσότητες δεδομένων. Καθώς όλο και περισσότερα κοινωνικά δίκτυα αρχίζουν να εμφανίζονται και το Web 2.0 ξεκινά, όλο και περισσότερα δεδομένα δημιουργούνται καθημερινά. Οι καινοτόμες νεοσύστατες επιχειρήσεις αρχίζουν αργά να ανακαλύπτουν αυτό το τεράστιο όγκο δεδομένων και επίσης οι κυβερνήσεις αρχίζουν να εργάζονται σε έργα Big Data. Το 2009 η ινδική κυβέρνηση αποφάσισε να κάνει σάρωση ίριδας, δακτυλικό αποτύπωμα και φωτογραφία όλων των 1,2 δισεκατομμυρίων κατοίκων. Όλα αυτά τα δεδομένα αποθηκεύονται στη μεγαλύτερη βιομετρική βάση δεδομένων στον κόσμο.Το 2010 ο Eric Schmidt μιλά στο συνέδριο Techonomy στη λίμνη Tahoe στην Καλιφόρνια και δηλώνει ότι «υπήρχαν 5 exabytes πληροφοριών που δημιουργήθηκαν από ολόκληρο τον κόσμο μεταξύ της αυγής του πολιτισμού και του 2003. Τώρα το ίδιο ποσό δημιουργείται κάθε δύο ημέρες».Το 2011 η έκθεση McKinsey για τα Big Data:Τα επόμενα σύνορα για την καινοτομία, τον ανταγωνισμό και την παραγωγικότητα, δηλώνουν ότι το 2018 μόνο οι ΗΠΑ θα αντιμετωπίσουν έλλειψη 140.000 - 190.000 επιστημόνων δεδομένων καθώς και 1,5 εκατομμύρια διαχειριστές δεδομένων (Diebold,2012).

Τα τελευταία χρόνια, υπήρξε μια τεράστια αύξηση στις εκκινήσεις Big Data, όλες προσπαθώντας να αντιμετωπίσουν το Big Data και βοηθώντας τους οργανισμούς να κατανοήσουν τα Big Data και όλο και περισσότερες εταιρείες υιοθετούν και κινούνται αργά προς το Big Data. Ωστόσο, παρόλο που φαίνεται ότι το Big Data υπάρχει εδώ και πολύ καιρό, στην πραγματικότητα το Big Data φτάνει μέχρι το Διαδίκτυο το 1993 . Η μεγάλη επανάσταση Big Data είναι ακόμα μπροστά μας, οπότε πολλά θα αλλάξουν τα επόμενα χρόνια.

2.3 Μέθοδοι

Όσον αφορά τη μεθοδολογία, η ανάλυση μεγάλων δεδομένων διαφέρει σημαντικά από την παραδοσιακή στατιστική προσέγγιση του πειραματικού σχεδιασμού. Το Analytics ξεκινά με δεδομένα. Κανονικά μοντελοποιούμε τα δεδομένα με τρόπο που να εξηγεί μια απάντηση. Οι στόχοι αυτής της προσέγγισης είναι η πρόβλεψη της συμπεριφοράς απόκρισης ή η κατανόηση του τρόπου με τον οποίο οι μεταβλητές εισόδου σχετίζονται με μια απόκριση. Κανονικά σε στατιστικά πειραματικά σχέδια, αναπτύσσεται ένα πείραμα και ανακτώνται δεδομένα ως αποτέλεσμα. Αυτό επιτρέπει τη δημιουργία δεδομένων με τρόπο που μπορεί να χρησιμοποιηθεί από ένα στατιστικό μοντέλο, όπου ισχύουν ορισμένες παραδοχές όπως η ανεξαρτησία, η κανονικότητα και η τυχαιοποίηση (Mayer-Schonberger & Cukier,2013).

Σε μεγάλη ανάλυση δεδομένων, μας παρουσιάζονται τα δεδομένα. Δεν μπορούμε να σχεδιάσουμε ένα πείραμα που ικανοποιεί το αγαπημένο μας στατιστικό μοντέλο. Σε εφαρμογές μεγάλης κλίμακας αναλυτικών στοιχείων, απαιτείται μεγάλη εργασία (συνήθως το 80% της προσπάθειας) μόνο για τον καθαρισμό των δεδομένων, ώστε να μπορεί να χρησιμοποιηθεί από ένα μοντέλο μηχανικής μάθησης. Δεν έχουμε μια μοναδική μεθοδολογία για να ακολουθήσουμε πραγματικές εφαρμογές μεγάλης κλίμακας. Κανονικά, μόλις καθοριστεί το επιχειρηματικό πρόβλημα, απαιτείται ένα στάδιο έρευνας για το σχεδιασμό της μεθοδολογίας που θα χρησιμοποιηθεί. Ωστόσο, οι γενικές οδηγίες είναι σχετικές για να αναφερθούν και να ισχύουν για σχεδόν όλα τα προβλήματα. Ένα από τα πιο σημαντικά καθήκοντα στο big data analytics είναι η στατιστική μοντελοποίηση , που σημαίνει εποπτευόμενα και μη εποπτευόμενα προβλήματα ταξινόμησης ή παλινδρόμησης. Μόλις τα δεδομένα καθαριστούν και προεπεξεργαστούν, είναι διαθέσιμα για μοντελοποίηση, θα πρέπει να ληφθεί μέριμνα για την αξιολόγηση διαφορετικών μοντέλων με λογικές μετρήσεις απώλειας και, στη συνέχεια, μετά την εφαρμογή του μοντέλου, θα πρέπει να αναφέρονται περαιτέρω αξιολόγηση και αποτελέσματα. Ένα κοινό μειονέκτημα στην προγνωστική μοντελοποίηση είναι η απλή εφαρμογή του μοντέλου και ποτέ η μέτρηση της απόδοσής του (Mayer-Schonberger & Cukier,2013).

2.4Προοπτικές

Το Big Data είναι το κυνήγι νοήματος από έναν ωκεανό δεδομένων. Μέχρι να γίνουν διαθέσιμα εργαλεία όπως το Hadoop και το NoSQL, δεν ήταν πρακτικό να αντλήσουμε μεγάλη ορατότητα από μη δομημένα δεδομένα και σίγουρα όχι πολύ νόημα από τα μέσα κοινωνικής δικτύωσης. Τώρα με αυτά τα εργαλεία, μπορούμε να παρέχουμε τάξη στο χάος και να εξετάσουμε πιο προσεκτικά τα δεδομένα. Μια πεποίθηση σχετικά με την ανάλυση Big Data είναι ότι δεν χρειάζεται να κατανοήσουμε την αιτία συσχέτισης που μπορεί να βρούμε στα δεδομένα. Απλώς κατανοώντας τις σχέσεις μεταξύ παραγόντων που γίνονται εμφανείς στα δεδομένα, μπορούμε να βρούμε χρήσιμες πληροφορίες. Ωστόσο, η συσχέτιση δεν είναι αιτιώδης. Μπορεί να μην καταλάβουμε γιατί σχετίζονται δύο παράγοντες, αλλά είναι ακόμη χρήσιμο να κατανοήσουμε τη συσχέτιση.

Για να προωθήσουμε την κατανόησή μας πέρα από τη συσχέτιση, ένα βήμα πιο κοντά στην κατανόηση της αιτιώδους συνάφειας είναι η Εξόρυξη Διαδικασίας. Η διαδικασία εξόρυξης φαίνεται πέρα από τη συσχέτιση με την περαιτέρω βελτίωση των συσχετίσεων στα δεδομένα. Πράγματι, η διαδικασία εξόρυξης ανά Wil van der Aalst του Πανεπιστημίου Τεχνολογίας στο Αϊντχόβεν, θα έδειχνε ότι εξετάζοντας μια πιο δομημένη άποψη των σχέσεων δεδομένων, μπορούμε να ανακαλύψουμε διαδικασίες στα δεδομένα. Με βάση μια προοπτική εξόρυξης διεργασιών μπορούμε να εντοπίσουμε διαδικασίες, να εντοπίσουμε σημεία συμφόρησης της διαδικασίας και εξετάζοντας τι κάνουν πραγματικά οι άνθρωποι, ενδεχομένως να βελτιώσουν τις διαδικασίες. Τέλος, μπορούμε να προβλέψουμε τα αποτελέσματα με βάση τις διαδικασίες που βρίσκουμε δοκιμάζοντας με δεδομένα πραγματικών γεγονότων.

Η διαδικασία εξόρυξης είναι διαφορετική από, αλλά σχετίζεται με, ανάλυση δεδομένων. Και οι δύο προσεγγίσεις ξεκινούν με δεδομένα συμβάντων. Η διαφορά είναι ότι η διαδικασία εξόρυξης ξεκινά με τη χαρτογράφηση μιας αλυσίδας συμβάντων για τη δημιουργία, βελτίωση και δοκιμή μοντέλων που ταιριάζουν στα δεδομένα. Στη συνέχεια, μπορείτε να χρησιμοποιήσετε το μοντέλο που προτείνει αυτή η τεχνική για να αναζητήσετε σημεία συμφόρησης με υπάρχουσες διαδικασίες. Με τη δοκιμή με πραγματικά δεδομένα συμβάντων βλέπετε τον τρόπο με τον οποίο συμβαίνουν πραγματικά τα γεγονότα, όχι πώς πρέπει να συμβούν.Με οποιαδήποτε μέθοδο δημιουργίας ενός μοντέλου (Play-Out ή Play-In) επαναλάβετε τα πραγματικά δεδομένα στο μοντέλο χρησιμοποιώντας δεδομένα καταγραφής συμβάντων, ώστε να μπορείτε να αξιολογήσετε τα πλεονεκτήματα και τις αδυναμίες των διαφόρων μοντέλων. Δοκιμάζοντας ένα μοντέλο με το Replay, μπορείτε να δείξετε αποκλίσεις από τα αναμενόμενα αποτελέσματα και

να συντονίσετε ανάλογα το μοντέλο. Η ΒΗ κλασική εξόρυξη μεγάλων δεδομένων ξεκινά με δεδομένα. Η διαδικασία εξόρυξης ξεκινά επίσης με δεδομένα. Η διαφορά είναι η αναζήτηση συσχέτισης στην εξόρυξη δεδομένων και η αναζήτηση διεργασιών στο Process Mining. Η εξόρυξη δεδομένων ενδέχεται να παράγει ένα απλό σύνολο βασικών δεικτών απόδοσης (KPI). Ο Δρ van der Aalst υποστηρίζει ότι το απλό KPI θα οδηγήσει σε προβλήματα επειδή είναι πολύ απλοϊκό. Η κατανόηση του τρόπου συνεργασίας των ανθρώπων για την ολοκλήρωση μιας εργασίας είναι πιο ενημερωτική. Το KPI μπορεί να εμφανίζει αποκλίσεις, αλλά ο χρήστης μπορεί να μην έχει ιδέα από πού προέρχεται η απόκλιση ή πώς να επαναφέρει τη διαδικασία παραγωγική (Mayer-Schonberger & Cukier,2013).

Χρησιμοποιώντας τη διαδικασία εξόρυξης μπορείτε να εντοπίσετε σημεία συμφόρησης ή μη παραγωγικά βήματα διαδικασίας μέσω ελέγχου συμμόρφωσης για να βεβαιωθείτε ότι αυτό που νομίζετε ότι συμβαίνει, πραγματικά συμβαίνει. Η Εξόρυξη Δεδομένων μπορεί να είναι είτε εποπτευόμενη μάθηση με επισημασμένα δεδομένα είτε μη εποπτευόμενη μάθηση με δεδομένα χωρίς ετικέτα. Η μη επιτηρούμενη μάθηση βασίζεται συχνά στην ανακάλυψη συστάδων ή μοτίβων. Η εποπτευόμενη μάθηση συχνά χρησιμοποιεί παλινδρόμηση για την ανάλυση των σχέσεων. Υπάρχουν πολλά εργαλεία για να βοηθήσετε στην Εξόρυξη Δεδομένων και το μεγαλύτερο μέρος της προσοχής στον κλάδο έχει επικεντρωθεί στην εξόρυξη δεδομένων.

Το μέγεθος των δεδομένων που παράγονται και μοιράζονται οι επιχειρήσεις, οι δημόσιες διοικήσεις πολλοί βιομηχανικοί και μη κερδοσκοπικοί τομείς, καθώς και η επιστημονική έρευνα, έχει αυξηθεί πάρα. Αυτά τα δεδομένα περιλαμβάνουν περιεχόμενο κειμένου (δηλ. Δομημένο, ημι-δομημένο καθώς και μη δομημένο), σε περιεχόμενο πολυμέσων (π.χ. βίντεο, εικόνες, ήχος) σε πολλές πλατφόρμες (π.χ. επικοινωνίες από μηχανή σε μηχανή, ιστότοποι κοινωνικών μέσων, δίκτυα αισθητήρων, κυβερνο-φυσικά συστήματα και Διαδίκτυο των πραγμάτων [IoT]). Οι Dobre και Xhafa αναφέρουν ότι κάθε μέρα ο κόσμος παράγει περίπου 2,5 quintillion byte δεδομένων (δηλ. 1 exabyte ισούται με 1 quintillion bytes ή 1 exabyte ισούται με 1 δισεκατομμύριο gigabyte), με το 90% αυτών των δεδομένων που παράγονται στον κόσμο να είναι αδόμητα.Οι Gantz και Reinsel ισχυρίζονται ότι έως το 2020, περισσότερα από 40 Zettabytes (ή 40 τρισεκατομμύρια gigabytes) δεδομένων θα έχουν δημιουργηθεί, απομιμηθεί και καταναλωθεί. Με αυτήν τη συντριπτική ποσότητα πολύπλοκων και ετερογενών δεδομένων που προέρχονται από οποιοδήποτε σημείο, οποτεδήποτε και οποιαδήποτε συσκευή, υπάρχει αναμφισβήτητα μια εποχή Big Data - ένα φαινόμενο που αναφέρεται επίσης ως Data Deluge . Το δυναμικό της BD είναι

εμφανές καθώς έχει συμπεριληφθεί στις 10 κορυφαίες τάσεις στρατηγικής τεχνολογίας της Gartner για το 2013 και στις κορυφαίες 10 τάσεις τεχνολογίας για τα επόμενα πέντε χρόνια. Είναι τόσο ζωτικής σημασίας όσο η νανοτεχνολογία και ο κβαντικός υπολογιστής στη σημερινή εποχή. Στην ουσία, το BD είναι το τεχνούργημα του ανθρώπινου ατόμου καθώς και η συλλογική νοημοσύνη που δημιουργείται και μοιράζεται κυρίως μέσω του τεχνολογικού περιβάλλοντος, όπου σχεδόν οτιδήποτε και όλα μπορούν να τεκμηριωθούν, να μετρηθούν και να καταγραφούν ψηφιακά, και με αυτόν τον τρόπο να μετατραπούν σε δεδομένα - μια διαδικασία που Οι Mayer-Schönberger και Cukier αναφέρονται επίσης ως δεδομένα (Mayer-Schonberger & Cukier,2013).

.Σύμφωνα με την έννοια των δεδομένων και τις ολοένα αυξανόμενες τεχνολογικές εξελίξεις, οι υποστηρικτές υποστηρίζουν ότι στο μέλλον η πλειονότητα των δεδομένων θα δημιουργηθεί και θα μοιραστεί μέσω μηχανών, καθώς οι μηχανές επικοινωνούν μεταξύ τους μέσω δικτύων δεδομένων. Ανεξάρτητα από το πού δημιουργείται και κοινοποιείται το BD, με την πραγματικότητα του BD έρχεται η πρόκληση της ανάλυσής του με τρόπο που φέρνει τη Μεγάλη Αξία . Με τόσο μεγάλη αξία να βρίσκεται μέσα, το BD έχει θεωρηθεί ως το Digital Oil, συμπεριλαμβανομένης της Νέας Πρώτης Ύλης του 21ου αιώνα. Η κατάλληλη επεξεργασία και διαχείριση δεδομένων θα μπορούσε να εκθέσει νέες γνώσεις και να διευκολύνει την άμεση ανταπόκριση σε αναδυόμενες ευκαιρίες και προκλήσεις. Ωστόσο, η αύξηση των δεδομένων σε όγκους στον ψηφιακό κόσμο φαίνεται να επιταχύνει την πρόοδο των πολλών υφιστάμενων υποδομών υπολογιστών. Οι καθιερωμένες τεχνολογίες επεξεργασίας δεδομένων, για παράδειγμα βάση δεδομένων και αποθήκη δεδομένων, καθίστανται ανεπαρκείς, δεδομένου του όγκου δεδομένων που παράγει ο κόσμος. Ο τεράστιος όγκος δεδομένων πρέπει να αναλυθεί με επαναληπτικό, καθώς και με χρονικά ευαίσθητο. Με τη διαθεσιμότητα προηγμένων τεχνολογιών ανάλυσης BD (π.χ. βάσεις δεδομένων NoSQL, BigQuery, MapReduce, Hadoop, WibiData και Skytree), οι πληροφορίες μπορούν να επιτευχθούν καλύτερα για να βελτιώσουν τις επιχειρηματικές στρατηγικές και τη διαδικασία λήψης αποφάσεων σε κρίσιμους τομείς όπως η υγειονομική περίθαλψη, η οικονομική παραγωγικότητα , ενεργειακά συμβόλαια και πρόβλεψη φυσικής καταστροφής, για να αναφέρουμε μόνο λίγα (Mayer-Schonberger & Cukier,2013).

3. Μια κανονιστική προοπτική των Big Data

3.1 Προκλήσεις και αναλυτικές μέθοδοι

Η έννοια του μεγάλου είναι προβληματική για να εντοπιστεί, κυρίως επειδή ένα σύνολο δεδομένων που φαίνεται να είναι τεράστιο σήμερα σχεδόν σίγουρα θα φαίνεται μικρό στο εγγύς μέλλον (MIT Technology Review, 2013). Προσθέτοντας την πολυπλοκότητα του ίδιου του BD, ορισμένοι επαγγελματίες υποστηρίζουν ότι τα τεράστια σύνολα δεδομένων δεν είναι πάντα περίπλοκα και τα μικρά σύνολα δεδομένων είναι πάντα απλά, υπογραμμίζοντας έτσι ότι η πολυπλοκότητα ενός συνόλου δεδομένων είναι ένας σημαντικός παράγοντας για να καθοριστεί εάν είναι μεγάλο dataset.

3.2 Μεγάλες προκλήσεις δεδομένων

Αν και τα οφέλη του BD είναι πραγματικά και ουσιαστικά, παραμένει μια πληθώρα προκλήσεων που πρέπει να αντιμετωπιστούν για να αξιοποιηθεί πλήρως το δυναμικό του BD. Μερικές από αυτές τις προκλήσεις είναι συνάρτηση των χαρακτηριστικών του BD, κάποιες, από τις υπάρχουσες μεθόδους ανάλυσης και μοντέλα και μερικές, μέσω των περιορισμών του τρέχοντος συστήματος επεξεργασίας δεδομένων (Jin, Wah, Cheng, & Wang, 2015). Εκτενείς μελέτες σχετικά με τις προκλήσεις του BD έχουν δώσει προσοχή στις δυσκολίες κατανόησης της έννοιας του BD (Hargittai, 2015), στη λήψη αποφάσεων σχετικά με τα δεδομένα που παράγονται και συλλέγονται (Crawford, 2013), θέματα απορρήτου (Lazer et al., 2009) και ηθικά ζητήματα που σχετίζονται με την εξόρυξη τέτοιων δεδομένων (Boyd & Crawford, 2012).Ο Tole (2013) υποστηρίζει ότι η οικοδόμηση μιας βιώσιμης λύσης για μεγάλα και πολύπλευρα δεδομένα είναι μια πρόκληση που οι επιχειρήσεις μαθαίνουν συνεχώς και στη συνέχεια εφαρμόζουν νέες προσεγγίσεις. Για παράδειγμα, ένα από τα μεγαλύτερα προβλήματα σχετικά με την BD είναι το υψηλό κόστος της υποδομής (Wang & Wiebe, 2014). Ο εξοπλισμός υλικού είναι πολύ ακριβός ακόμη και με τη διαθεσιμότητα τεχνολογιών υπολογιστικού νέφους.

Επιπλέον, για να ταξινομήσετε δεδομένα, έτσι ώστε να μπορούν να κατασκευαστούν πολύτιμες πληροφορίες, απαιτείται συχνά ανθρώπινη ανάλυση. Ενώ οι τεχνολογίες υπολογιστών που απαιτούνται για τη διευκόλυνση αυτών των δεδομένων συμβαδίζουν, η ανθρώπινη τεχνογνωσία

και ταλέντα που απαιτούν οι ηγέτες των επιχειρήσεων να αξιοποιήσουν το BD καθυστερούν, αυτό αποδεικνύεται ότι είναι μια άλλη μεγάλη πρόκληση. Όπως ανέφεραν οι Akerkar (2014) και Zicari (2014) , οι ευρείες προκλήσεις του BD μπορούν να ομαδοποιηθούν σε τρεις κύριες κατηγορίες, με βάση τον κύκλο ζωής των δεδομένων: δεδομένα, διαδικασίες και προκλήσεις διαχείρισης:

- Οι προκλήσεις δεδομένων σχετίζονται με τα χαρακτηριστικά των ίδιων των δεδομένων (π.χ. όγκος δεδομένων, ποικιλία, ταχύτητα, ακρίβεια, μεταβλητότητα, ποιότητα, ανακάλυψη και δογματισμός).
- Οι προκλήσεις της διαδικασίας σχετίζονται με τη σειρά τεχνικών: πώς να συλλέξετε δεδομένα, πώς να ενσωματώσετε δεδομένα, πώς να μετατρέψετε δεδομένα, πώς να επιλέξετε το σωστό μοντέλο για ανάλυση και πώς να παρέχετε τα αποτελέσματα.
- Οι διαχειριστικές προκλήσεις καλύπτουν για παράδειγμα το απόρρητο, την ασφάλεια, τη διακυβέρνηση και τις ηθικές πτυχές.

Για να διευκολύνουν τη λήψη αποφάσεων βάσει τεκμηρίων, οι οργανισμοί χρειάζονται αποτελεσματικές μεθόδους για την επεξεργασία μεγάλου όγκου ανάμεικτων δεδομένων σε ουσιαστικές γνώσεις (Gandomi & Haider, 2015). Οι δυνατότητες χρήσης του BD είναι ατελείωτες αλλά περιορίζονται από τη διαθεσιμότητα τεχνολογιών, εργαλείων και δεξιοτήτων που διατίθενται για το BDA. Σύμφωνα με τους Labrinidis and Jagadish (2012) , το BDA αναφέρεται σε μεθόδους που χρησιμοποιούνται για την εξέταση και την επίτευξη γνώσης από τα μεγάλα σύνολα δεδομένων. Έτσι, το BDA μπορεί να θεωρηθεί ως υπο-διαδικασία σε ολόκληρη τη διαδικασία της εξαγωγής πληροφοριών από το BD. Είναι βέβαιο ότι για να μπορέσει η BD να πραγματοποιήσει τους στόχους της και τις υπηρεσίες προόδου στο επιχειρηματικό περιβάλλον, απαιτεί τα σωστά εργαλεία και προσεγγίσεις να αναλυθούν και να ταξινομηθούν αποτελεσματικά και αποτελεσματικά (Al Nuaimi, Al Neyadi, Mohamed, & Al-Jaroodi, 2015). Η πιθανή τιμή του BD επιλύεται απλά όταν αξιοποιείται στη διαδικασία λήψης αποφάσεων. Εκτεταμένες ερευνητικές μελέτες έχουν δείξει ότι ουσιαστική αξία και ανταγωνιστικό πλεονέκτημα μπορούν να επιτευχθούν από τις επιχειρήσεις από τη λήψη αποτελεσματικών αποφάσεων βάσει δεδομένων (Davenport & Harris, 2007). Όμως, το BDA είναι πιο

μπερδεμένο από την απλή ανίχνευση, ταξινόμηση, κατανόηση και αναφορά δεδομένων. Ντάβενπορτ και Ντίτσε (2013)υπογραμμίστε ότι οι μεγάλοι οργανισμοί συγκεντρώνουν τακτικά BD και εκμεταλλεύονται αναλυτικά στοιχεία για υποστήριξη στη λήψη αποφάσεων ως μέρος των συνηθισμένων διαδικασιών τους, και οι ΜΜΕ είναι εκείνες που αγωνίζονται επί του παρόντος να βελτιώσουν τις αποφάσεις ανώτερης διαχείρισης, προσθέτοντας περισσότερα δεδομένα για τη διαδικασία ανάλυσης. Η ευθυγράμμιση των ανθρώπων, της τεχνολογίας και των οργανωτικών πόρων για να γίνει μια εταιρεία βάσει δεδομένων είναι προβληματική (Weill & Ross, 2009). Δεδομένου BD μπορεί να βελτιώσει τη λήψη αποφάσεων και να αυξήσει την παραγωγή οργανισμού Αυτό είναι εφικτό όταν χρησιμοποιείται μια επιλογή αναλυτικών μεθόδων για την εξαγωγή λογικής από τα δεδομένα, όπως:

- η περιγραφική ανάλυση εξετάζει τα δεδομένα και τις πληροφορίες για να καθορίσει την τρέχουσα κατάσταση μιας επιχειρηματικής κατάστασης με τρόπο που οι εξελίξεις, τα μοτίβα και οι εξαιρέσεις γίνονται εμφανείς, με τη μορφή τυποποιημένων αναφορών, ειδικών αναφορών και ειδοποιήσεων (Joseph & Johnson, 2013).
- Η διερευνητική ανάλυση αφορά την έρευνα δεδομένων για την πιστοποίηση / απόρριψη επιχειρηματικών προτάσεων, για παράδειγμα, αναλυτικές αναλύσεις δεδομένων, στατιστική ανάλυση, ανάλυση παραγόντων (Bihani & Patil, 2014).
- Η προγνωστική ανάλυση ασχολείται με την πρόβλεψη και τη στατιστική μοντελοποίηση για τον προσδιορισμό των μελλοντικών δυνατοτήτων (Waller & Fawcett, 2013).
- τα προδιαγραφικά analytics αφορούν τη βελτιστοποίηση και τον τυχαιοποιημένο έλεγχο για να εκτιμήσουν πώς οι επιχειρήσεις βελτιώνουν τα επίπεδα των υπηρεσιών τους μειώνοντας παράλληλα τα έξοδα (Joseph & Johnson, 2013).

- Το προληπτικό analytics αφορά την ικανότητα προληπτικών ενεργειών σε γεγονότα που ενδέχεται να επηρεάσουν ανεπιθύμητα την οργανωτική απόδοση, για παράδειγμα, τον εντοπισμό των πιθανών κινδύνων και τη σύσταση στρατηγικών μετριασμού πολύ μπροστά στο χρόνο (Szongott, Henne, & von Voigt, 2012).
- Οι υποστηρικτές υποστηρίζουν ότι αυτοί οι τύποι αναλυτικών μεθόδων υποστηρίζουν τη βελτιωμένη λήψη αποφάσεων και την οργανωτική απόδοση, καθιστώντας τα πάντα πιο

ημιδιαφανή και ποσοτικά, ενώ αποκαλύπτουν περαιτέρω ασυνέπειες, καθώς και πιθανές ανησυχίες και ευκαιρίες.

4. Τεχνητή Νοημοσύνη στα Συστήματα

4.1 Ορισμός

Η τεχνητή νοημοσύνη είναι ένας συνδυασμός πολλών διαφορετικών τεχνολογιών που συνεργάζονται για να επιτρέπουν στις μηχανές να αισθάνονται, να κατανοούν, να ενεργούν και να μαθαίνουν με ανθρώπινα επίπεδα νοημοσύνης. Ίσως γι 'αυτό φαίνεται ότι ο ορισμός της τεχνητής νοημοσύνης όλων είναι διαφορετικός: Η τεχνητή νοημοσύνη δεν είναι μόνο ένα πράγμα. Τεχνολογίες όπως η μηχανική εκμάθηση και η επεξεργασία φυσικών γλωσσών αποτελούν μέρος του τοπίου της τεχνητής νοημοσύνης. Ο καθένας εξελίσσεται στη δική του πορεία και, όταν εφαρμόζεται σε συνδυασμό με δεδομένα, αναλυτικά στοιχεία και αυτοματισμούς, μπορεί να βοηθήσει τις επιχειρήσεις να επιτύχουν τους στόχους τους, είτε βελτιώνουν την εξυπηρέτηση πελατών είτε βελτιστοποιούν την αλυσίδα εφοδιασμού. Μερικοί προχωρούν ακόμη περισσότερο για να ορίσουν την τεχνητή νοημοσύνη ως «στενή» και «γενική» τεχνητή νοημοσύνη. Τα περισσότερα από αυτά που βιώνουμε στην καθημερινή μας ζωή είναι στενή τεχνητή νοημοσύνη, η οποία εκτελεί μία μόνο εργασία ή ένα σύνολο στενά συνδεδεμένων εργασιών. Τα παραδείγματα περιλαμβάνουν:

- Εφαρμογές καιρού
- Ψηφιακοί βοηθοί
- Λογισμικό που αναλύει δεδομένα για τη βελτιστοποίηση μιας δεδομένης επιχειρησιακής λειτουργίας

Με τη σωστή εφαρμογή, το στενό AI έχει τεράστια δύναμη μετασχηματισμού - και συνεχίζει να επηρεάζει τον τρόπο με τον οποίο εργαζόμαστε και ζούμε σε παγκόσμια κλίμακαΕνώ οι μηχανές μπορούν να εκτελέσουν κάποιες εργασίες καλύτερα από τους ανθρώπους (π.χ. επεξεργασία

δεδομένων), το πλήρως συνειδητοποιημένο όραμα της γενικής τεχνητής νοημοσύνης δεν υπάρχει ακόμη. Αυτός είναι ο λόγος για τον οποίο η συνεργασία ανθρώπου-μηχανής είναι ζωτικής σημασίας - στον σημερινό κόσμο, η τεχνητή νοημοσύνη παραμένει επέκταση των ανθρώπινων δυνατοτήτων και όχι αντικατάστασης (Russell & Norvig,2004).

4.2 Ιστορία

Οι πνευματικές ρίζες της τεχνητής νοημοσύνης και η έννοια των ευφυών μηχανών βρίσκονται στην ελληνική μυθολογία. Έξυπνα αντικείμενα εμφανίζονται στη λογοτεχνία από τότε, με πραγματικές (και δόλιες) μηχανικές συσκευές που έχουν αποδειχθεί ότι συμπεριφέρονται με κάποιο βαθμό νοημοσύνης. Μερικά από αυτά τα εννοιολογικά επιτεύγματα αναφέρονται παρακάτω στην ενότητα "Αρχαία ιστορία".Μετά τη διάθεση των σύγχρονων υπολογιστών, μετά τον Β 'Παγκόσμιο Πόλεμο, κατέστη δυνατή η δημιουργία προγραμμάτων που εκτελούν δύσκολες πνευματικές εργασίες. Από αυτά τα προγράμματα, κατασκευάζονται γενικά εργαλεία που έχουν εφαρμογές σε μια μεγάλη ποικιλία καθημερινών προβλημάτων. Μερικά από αυτά τα υπολογιστικά ορόσημα παρατίθενται παρακάτω στην ενότητα "Σύγχρονη ιστορία".Η σύγχρονη ιστορία της τεχνητής νοημοσύνης ξεκινά με την ανάπτυξη ηλεκτρονικών υπολογιστών αποθηκευμένου προγράμματος. Για μια σύντομη περίληψη, δείτε το Genius and Tragedy at Dawn of Computer Age By ALICE RAWSTHORN, NY Times (25 Μαρτίου 2012), μια ανασκόπηση του βιβλίου του ιστορικού της τεχνολογίας Τζορτζ Ντίσον «Turing's Cathedral: The Origins of the Digital Universe» (Russell & Norvig,2004).

4.3 Συνεισφορά Τεχνητής Νοημοσύνης στην Ανάλυση Δεδομένων

Η εισαγωγή της τεχνητής νοημοσύνης, της αυτοματοποίησης και της αφήγησης δεδομένων στον κόσμο των αναλυτικών στοιχείων δεν είχε μόνο άμεσο αντίκτυπο στους τελικούς χρήστες των αναλυτικών στοιχείων, αλλά και στους ανθρώπους που εργάζονται στον τομέα. Ενώ πολλοί αναλυτές φοβούνται ότι θα αντικατασταθούν από αυτοματοποίηση και τεχνητή νοημοσύνη, πιστεύω ότι ο ρόλος του αναλυτή δεδομένων θα αυξηθεί σε σημασία για την επιχείρηση και το εύρος των απαιτούμενων δεξιοτήτων. Οι αναλυτές δεδομένων έχουν ξοδέψει παραδοσιακά ένα

σημαντικό μέρος του χρόνου τους κάνοντας συνηθισμένες και επαναλαμβανόμενες εργασίες, όπως προετοιμασία δεδομένων για ανάλυση, δημιουργία αναφορών και ταμπλό και στη συνέχεια χρήση αυτών για μη αυτόματη αναζήτηση σημαντικών αλλαγών στα δεδομένα τους. Με τα παραδοσιακά εργαλεία αναλυτικής και επιχειρηματικής ευφυΐας, οι αναλυτές απλά δεν μπορούν να διερευνήσουν κάθε συνδυασμό ή παραλλαγή των δεδομένων τους. Και αν βρουν κάτι ενδιαφέρον, πώς καθορίζουν εάν είναι στατιστικά σχετικό και έχει ουσιαστικό όφελος για την επιχείρηση; Η εισαγωγή της αυτόματης ανακάλυψης δεδομένων αντιμετωπίζει αυτά τα ζητήματα. Μειώνει το χρόνο για να βρει πληροφορίες, αφήνοντας στη συνέχεια πολύ περισσότερο χρόνο για τους αναλυτές να προσθέσουν αξία ερμηνεύοντας τα ευρήματά τους. Αυτό θα απαιτήσει από τους αναλυτές να καταλάβουν τις επιχειρήσεις. Ο ρόλος του αναλυτή δεδομένων περιλαμβάνει σήμερα ένα ευρύ φάσμα δραστηριοτήτων διαχείρισης και ανάλυσης δεδομένων. Αυτά περιλαμβάνουν την προμήθεια, την προετοιμασία, τον καθαρισμό και τη μοντελοποίηση δεδομένων, στη συνέχεια τη δημιουργία αναφορών και πινάκων ελέγχου για την ανάλυση της επιχείρησης για την υποστήριξη της λήψης αποφάσεων. Από όλες αυτές τις δραστηριότητες, η πραγματική αξία για την επιχείρηση είναι εκείνες οι δραστηριότητες που σχετίζονται με τον εντοπισμό κρίσιμων αλλαγών ή τάσεων που επηρεάζουν την επιχείρηση και την ερμηνεία αυτών των πληροφοριών για να προσδιοριστεί ο πιθανός αντίκτυπος στην επιχείρηση. Το δίλημμα που αντιμετωπίζουν οι επιχειρηματικοί αναλυτές είναι ότι, αν και η ερμηνεία είναι η πιο πολύτιμη δραστηριότητα που αναλαμβάνουν, είναι εκείνο που ξοδεύουν το λιγότερο χρόνο. Οι περισσότεροι αναλυτές δεδομένων ξοδεύουν μόνο το 20 τοις εκατό του χρόνου τους στην πραγματική ανάλυση δεδομένων και το 80 τοις εκατό του χρόνου τους κάνοντας εργασίες με μικρό επιχειρηματικό όφελος όπως η εύρεση, ο καθαρισμός και η μοντελοποίηση δεδομένων, το οποίο είναι εξαιρετικά αναποτελεσματικό και προσθέτει μικρή αξία στην επιχείρηση (Adadi, & Berrada,2018).

Δεν είναι απλώς αποτελεσματική η προετοιμασία δεδομένων. Τα παραδοσιακά εργαλεία για ανάλυση δεδομένων και οπτικοποίηση απαιτούν μια εντελώς χειροκίνητη προσέγγιση για την ανακάλυψη δεδομένων. Οι χρήστες πρέπει να επιλέξουν από μια μεγάλη γκάμα πεδίων και φίλτρων και, στη συνέχεια, δεδομένα τεμαχίων κατά την αναζήτηση μοτίβων, αλλαγών στις τάσεις και ανωμαλιών. Αυτή η χειροκίνητη διαδικασία είναι απίστευτα χρονοβόρα και είναι πολύ επιρρεπής σε ανθρώπινα λάθη και προκατάληψη, ειδικά στον σημερινό κόσμο πλούσιο σε δεδομένα. Το αποτέλεσμα? Ο προσδιορισμός των κρίσιμων αλλαγών στα επιχειρηματικά

δεδομένα είναι τυχαίο και όχι κάτι που θα συμβεί με βεβαιότητα. Αυτό δημιουργεί κίνδυνο για ηγέτες επιχειρήσεων που θέλουν βεβαιότητα στα δεδομένα που χρησιμοποιούν για τη λήψη αποφάσεων. Το AI και ο αυτοματισμός υπόσχονται να αλλάξουν ριζικά αυτό το παράδειγμα. Εφαρμοσμένη στην ανάλυση και την επιχειρηματική ευφυΐα, πολλές από τις κουραστικές και χρονοβόρες διαδικασίες θα γίνουν από μηχανήματα. Η έξυπνη προετοιμασία δεδομένων που χρησιμοποιεί μηχανική εκμάθηση για τον εξορθολογισμό των διαδικασιών προφίλ δεδομένων, αντιστοίχησης και καθαρισμού θα μειώσει σημαντικά το χρόνο που αφιερώνουν οι αναλυτές για την προετοιμασία δεδομένων για ανάλυση. Αυτό σε συνδυασμό με την ανακάλυψη δεδομένων βάσει τεχνητής νοημοσύνης, η οποία εφαρμόζει μια σειρά εξελιγμένων αλγορίθμων στα δεδομένα, θα μειώσει την χρονοβόρα εξερεύνηση δεδομένων και την ανακάλυψη σχετικών επιχειρηματικών πληροφοριών (Adadi, & Berrada, 2018).

Ωστόσο, αυτές οι εξελίξεις δεν σημαίνουν ότι η AI θα αντικαταστήσει τον αναλυτή δεδομένων. Το AI είναι ιδανικό για αυτοματοποίηση, αλλά έχει θεμελιώδεις περιορισμούς. Τα μηχανήματα δεν μπορούν να κατανοήσουν το πλαίσιο. Μόνο οι άνθρωποι έχουν την ικανότητα να συγκρίνουν δεδομένα με πολύπλοκους όρους όπως το οργανωτικό περιβάλλον, οι παράγοντες της εξωτερικής αγοράς, η δυναμική των πελατών και πολλά άλλα. Για παράδειγμα, η ικανότητα εύρεσης νοήματος σε μια πτωτική τάση των πωλήσεων προϊόντων βάσει της ανεκδοτικής αύξησης του μάρκετινγκ από έναν ανταγωνιστή είναι πολύ περισσότερο από ό, τι μπορεί να επεξεργαστεί η AI, αλλά είναι σχετικά απλό για έναν άνθρωπο να το κάνει (Raedt, Kersting, Natarajan, & Poole,2016).

Το αποτέλεσμα αυτής της αλλαγής θα δει τους αναλυτές δεδομένων να ξοδεύουν πολύ περισσότερο χρόνο κάνοντας ό, τι δεν μπορούν οι μηχανές - παρέχοντας πλαίσιο και ερμηνεία δεδομένων. Οι αναλυτές δεδομένων θα ανυψωθούν σε σχέση με τους σημαντικούς επιχειρηματικούς εταίρους, όπου θα χρησιμοποιήσουν τις δεξιότητες τους για την ανάγνωση δεδομένων για να βοηθήσουν την επιχείρηση να ερμηνεύσει τα δεδομένα, να προσαρμόσει τα στοιχεία που ανακαλύφθηκαν και να πει συναρπαστικές ιστορίες με αυτά τα δεδομένα. Το αποτέλεσμα αυτού θα είναι ότι οι αναλυτές δεδομένων σήμερα πρέπει να γίνουν πολύ πιο κατανοητοί στις επιχειρήσεις και να αναπτύξουν τις δεξιότητές τους για να αναπτύξουν αφηγήσεις. Αυτό δεν σημαίνει ότι οι επαναλαμβανόμενες εργασίες αναλυτών δεδομένων δεν θα εξαφανιστούν. Για αναλυτές δεδομένων των οποίων ο πρωταρχικός στόχος είναι η προετοιμασία

δεδομένων και η κατασκευή πινάκων ελέγχου, ο χρόνος τους θα έρθει νωρίτερα και όχι αργότερα. Ωστόσο, οι οργανισμοί θα βασίζονται σε μεγαλύτερο βαθμό σε εκείνους με τις δεξιότητες για να παρέχουν πληροφορίες σχετικά με το τι σημαίνει τα δεδομένα. Οι αναλυτές δεδομένων θα βασίζονται σε εργαλεία που βασίζονται σε τεχνολογία AI που διευκολύνουν τις συνηθισμένες πτυχές της εργασίας τους, έτσι ώστε να μπορούν να ξοδεύουν περισσότερο χρόνο σε πολύτιμες δραστηριότητες όπως η ερμηνεία δεδομένων και η αφήγηση. Ως αποτέλεσμα, θα μπορούν να παρέχουν ουσιαστική ανάλυση στην επιχείρηση για να λαμβάνουν καλύτερες αποφάσεις βάσει δεδομένων (Raedt, Kersting, Natarajan, & Poole,2016).

5. Αυτοματοποιημένη Μηχανική Μάθηση σε τραπεζικά δεδομένα

5.1 WEKA

Καθώς έχει υπάρχει η ανάγκη για όλο και πιο ευρύτερη χρήση μηχανικής μάθησης, παρατηρήθηκε το φαινόμενο της έλλειψης ανθρώπινου δυναμικού που να μπορεί να χρησιμοποιήσει περίπλοκα λογισμικά η να γράψει κώδικα σε κάποια γλώσσα προγραμματισμού όπως η R ή η Python για να αναπτυχθούν εφαρμογές που θα χρησιμοποιήσουν μηχανική μάθηση προς εξόρυξη δεδομένων. Ακόμα και έμπειροι μηχανικοί λογισμικού θα πρέπει να εξοικειωθούν με τι βιβλιοθήκες μηχανικής μάθησης και τη θεωρία για τη σωστή χρήση τους. Για να αντιμετωπισθεί αυτό το πρόβλημα εμφανίσθηκαν μια σειρά από εργαλεία είτε εμπορικού λογισμικού είτε ανοιχτού λογισμικού τα οποία μπορούν να λάβουν σαν είσοδο ένα αρχείο δεδομένων σε διάφορες μορφές όπως text αρχείο η excel αρχείο αρχείου και να δημιουργήσουν για εμάς με αυτοματοποιημένο τρόπο μοντέλα. Η είσοδος δεδομένων δεν περιορίζεται σε αρχεία αλλά μπορεί να υπάρξει και διασύνδεση με βάσεις δεδομένων, με δεδομένα στο υπολογιστικό

νέφος ακόμα με ροές δεδομένων από το Internet of Things. Δηλαδή δεδομένα από πολυποίκιλους και σε μεγάλο πλήθος αισθητήρες.

Ένα από τα πιο γνωστά λογισμικά αυτοματοποιημένης αυτοματοποιημένης μάθησης είναι το WEKA (Waikato Environment for Knowledge Analysis) (Weka 2021). Το WEKA αναπτύχθηκε από το πανεπιστήμιο Waikato της Νέας Ζηλανδίας και είναι εργαλείο ανοιχτού λογισμικό με άδεια GNU General Public Licence. Το πανεπιστήμιο έχει εκδώσει και το βιβλίο "Data Mining: Practical Machine Learning Tools and Techniques" το οποίο εξηγεί την χρήση του WEKA σε βάθος και αναλύει όλα τα μοντέλα που μπορείς να εκπαιδεύσεις με αυτό καθώς και τις παραμέτρους των μοντέλων. Το WEKA στην τωρινή του έκδοση είναι ανεπτυγμένο σε JAVA. Η μεγάλη χρησιμότητα του WEKA έγκειται στο ότι μπορούμε με τη χρήση οπτικού περιβάλλοντος να εφαρμόσουμε βασικούς μετασχηματισμούς στα δεδομένα μας και εν συνεχεία να εκπαιδεύσουμε μοντέλα. Επιπλέον μας βοηθά να επιλέξουμε τις ανεξάρτητες μεταβλητές με τις οποίες θα χτισθούν τα μοντέλα μας (feature selection). Επιπλέον παρέχει διασύνδεση με πλήθος βάσεων δεδομένων και πρόσφατα έχουν προστεθεί και δυνατότητες AI δια μέσω της βιβλιοθήκης Deeplearning4j.

Με το WEKA μπορούμε να κάνουμε με αυτοματοποιημένο τρόπο ένα πλήθος λειτουργιών μηχανικής μάθησης όπως επεξεργασία δεδομένων, μετασχηματισμούς δεδομένων , εκπαίδευση μοντέλου, εκτίμηση ακρίβειας, παραγωγή διαγραμμάτων.

Ένα άλλο πολύ χρήσιμο χαρακτηριστικό που μας παρέχει το WEKA είναι η εξαγωγή των μοντέλων σε αρχεία τα οποία εν συνεχεία μπορούν να χρησιμοποιηθούν μέσω γλωσσών προγραμματισμού όπως η JAVA για να μας δώσουν προβλέψεις και εκτιμήσεις σε νέα άγνωστα δεδομένα. Με αυτό τον τρόπο μπορούμε να χρησιμοποιήσουμε την πλατφόρμα του WEKA για να δημιουργήσουμε εύκολα και γρήγορα τα μοντέλα μας και μετά να τα παραδώσουμε σε άλλες ομάδες προγραμματιστών για να τα περάσουν σε παραγωγικά συστήματα όπου θα μπορούν να μας δώσουν εκτιμήσεις προβλέψεων real time.

Επιπλέον η λειτουργικότητα του WEKA μπορεί να επεκταθεί μέσω πακέτων τα οποία μπορούμε να αναπτύξουμε μόνοι μας ή να κατεβάσουμε δωρεάν από το ιντερνετ.

5.2 Δεδομένα και πρόβλημα προς επίλυση

Ένας από τους πρώτους χώρους όπου υιοθετήθηκαν οι τεχνικές μηχανικής μάθησης ήταν ο τραπεζικός χώρος. Αυτό συνέβη για μια σειρά από λόγους. Α) Οι τράπεζες έχουν ήδη πολλά δεδομένα διαθέσιμα. Β)Αυτά τα δεδομένα είναι ήδη δεδομένα και αποθηκευμένα σε βάσεις τραπεζικών λογισμικών Γ) Υπήρξαν οι πόροι για να εφαρμοστούν οι τεχνικές εξόρυξης δεδομένων. Δ) Ο ανταγωνισμός στα τραπεζικά προϊόντα είναι μεγάλος λόγω της περιορισμένης δυνατότητας διαφοροποίησης των προϊόντων. Για αυτούς τους λόγους και καθώς για τις διευθύνσεις ανάπτυξης και προώθησης προϊόντων οποιαδήποτε πληροφορία μπορούν να μάθουν από τα δεδομένα τους για τους τραπεζικούς πελάτες συνιστά ανταγωνιστικό πλεονέκτημα υπήρξε ευρύτατη διάδοση της μηχανικής μάθησης.

Στη δική μας περίπτωση έχουμε δεδομένα από γνωστής τράπεζας τα οποία αφορούν προωθητικές ενέργειες προς προσέλκυση πελατών για άνοιγμα νέας προθεσμιακής κατάθεσης. Η τράπεζα που μας ανέθεσε την έρευνα επιθυμεί να μάθει τα χαρακτηριστικά των πελατών που τελικά άνοιξαν κατάθεση και να δημιουργήσει μοντέλο που θα προβλέπει πόσο πιθανό είναι άλλα άτομα τα οποία θα προσεγγισθούν να ανοίξουν επίσης νέα προθεσμιακή κατάθεση.. Κατά αυτό τον τρόπο μπορεί να κάνει στοχευμένα τις επόμενες ενέργειές της και να μεγαλώσει το μερίδιο αγοράς της με πιο αποτελεσματικό τρόπο αλλά να μειώσει το κόστος προσέλκυσης πελατών.

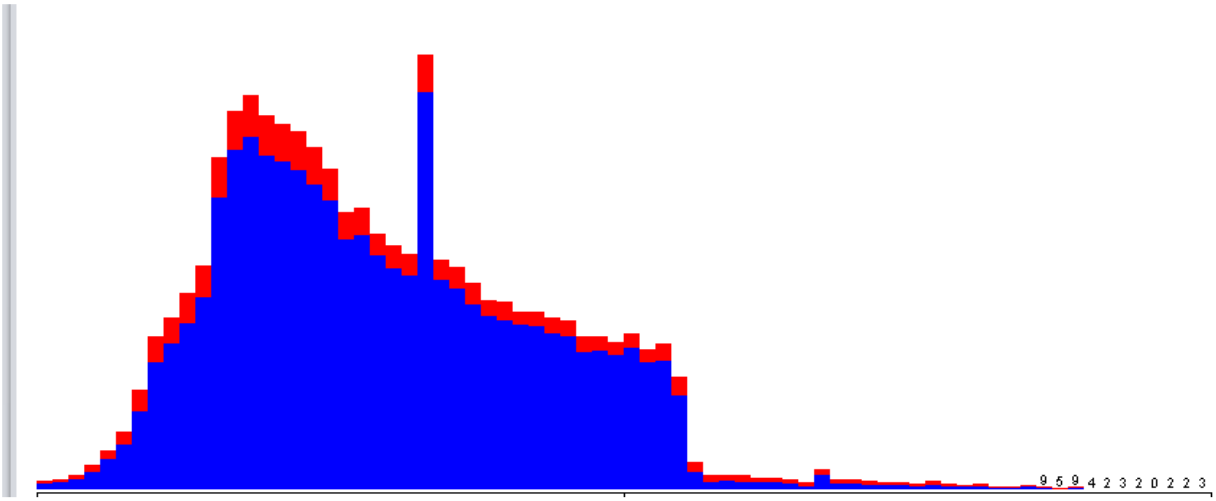
Τα δεδομένα μας είναι σε αρχείο τύπου arff και αφορούν 45.211 ενέργειες για το άνοιγμα προθεσμιακής κατάθεσης. Έχουμε 17 πεδία από τα οποία το τελευταίο δείχνει αν ο πελάτης άνοιξε ή όχι την προθεσμιακή κατάθεση. Δηλαδή έχουμε 16 ανεξάρτητες μεταβλητές και μία εξαρτημένη. Η εξαρτημένη μεταβλητή yes/no μας δείχνει αν τελικά ο πελάτης άνοιξε η όχι λογαριασμό και είναι binary.

5.3 Μελέτη των δεδομένων

Ανοίγουμε στο weka το αρχείο Bank-full.arff και παρατηρούμε τις μεταβλητές. Σαν πρώτο βήμα πρέπει να μελετήσουμε τα δεδομένα μας για να τα κατανοήσουμε. Η ορθή κατανόηση των δεδομένων μας βοηθά να εκτιμήσουμε αν τηρούνται οι θεωρητικές προϋποθέσεις για κάθε μοντέλο και επίσης να μας δώσει μια αρχική εποπτική εικόνα για τις συσχετίσεις που υπάρχουν.

Ξεκινάμε με την μεταβλητή ηλικία.

Η ηλικία έχει μια συγκέντρωση στις μέσες ηλικίες όπως περιμένουμε από τραπεζικούς πελάτες



Συνεχίζουμε με το επάγγελμα. Στα επαγγέλματα έχουμε 12 διαφορετικές κατηγορίες

Name: job
Missing: 0 (0%) Distinct: 12

No.	Label	Count
1	management	9458
2	technician	7597
3	entrepreneur	1487
4	blue-collar	9732
5	unknown	288
6	retired	2264
7	admin.	5171
8	services	4154
9	self-employed	1579
10	unemployed	1303
11	housemaid	1240
12	student	938

Για την μεταβλητή (marital status)που μας δείχνει την οικογενειακή κατάσταση έχουμε 3 διαφορετικές κατηγορίες με πιο διαδεδομένη το παντρεμένος/η

Name: marital
Missing: 0 (0%) Distinct: 3

No.	Label	Count
1	married	27214
2	single	12790
3	divorced	5207

Η εκπαίδευση (education) είναι κατηγορική μεταβλητή με 4 διαφορετικές τιμές.

Name: education
Missing: 0 (0%) Distinct: 4

No.	Label	Count
1	tertiary	13301
2	secondary	23202
3	unknown	1857
4	primary	6851

Η μεταβλητή default μας δείχνει αν κάποιος πελάτης είχε κηρύξει χρεωκοπία στο παρελθόν. Πολλοί λίγοι είχαν προηγούμενη χρεωκοπία

Selected attribute

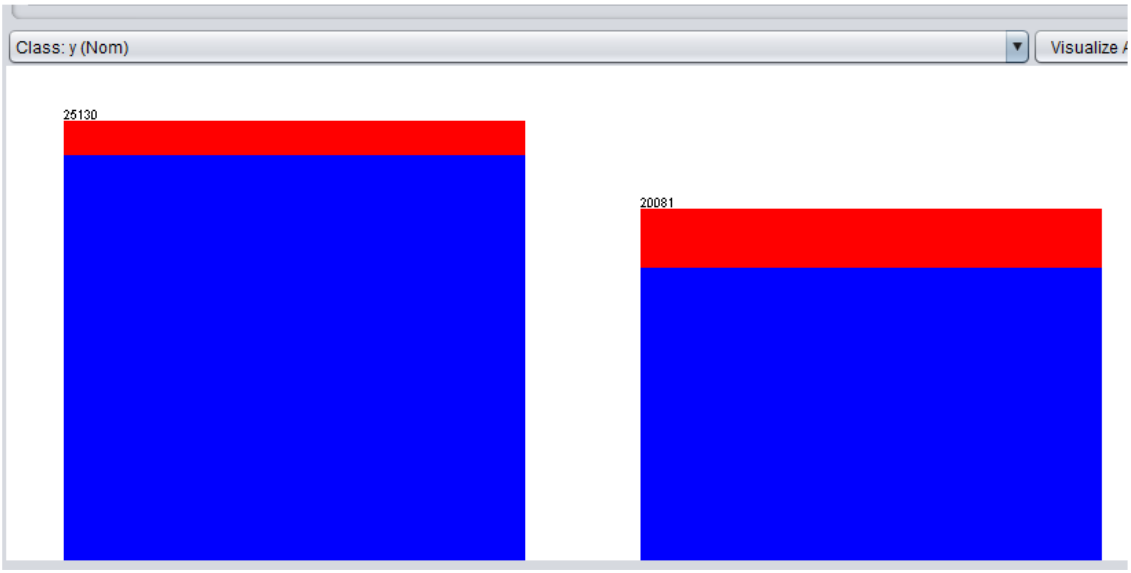
Name: default
Missing: 0 (0%) Distinct: 2

No.	Label	Count
1	no	44396
2	yes	815

Ακολούθως εξετάζουμε την μεταβλητή υπόλοιπο λογαριασμού η οποία αναμένουμε να είναι πολύ σημαντική. Παρατηρούμε μεγάλη συγκέντρωση στα υπόλοιπα λογαριασμών

Σαν επόμενο βήμα επιλέγουμε να δούμε πως κατανέμεται η εξαρτημένη μεταβλητή σε σχέση με κάποιες από τις ανεξάρτητες. Εξετάζουμε δηλαδή τις τιμές που πήρε η εξαρτημένη βάσει των τιμών μιας ανεξάρτητης. Αυτό το κάνουμε με την χρήση είτε πινάκων είτε διαγραμμάτων.

Ως προς την σχέση εξαρτημένης μεταβλητής και της ανεξάρτητης housing (ιδιόκτητη κατοικία) έχουμε μια ισορροπία όσον αφορά το yes/no



Στην περίπτωση της μεταβλητής loan που μας δείχνει αν ο πελάτης έχει ή όχι κάποιο δάνειο. έχουμε συγκέντρωση στο no

Name: loan
Missing: 0 (0%) Distinct: 2

No.	Label	Count
1	no	37967
2	yes	7244

Η μεταβλητή contact μας δείχνει τον τρόπο με τον οποίο προσεγγίσαμε κάποιο πελάτη. Δηλαδή αν έγινε μέσω κλήσης σε σταθερό ή κινητό. Εχουμε και περιπτώσεις όπου το τμήμα marketing δεν έχει καταγράψει τον τρόπο επαφής. Αυτό θα μπορούσε να είναι πρόβλημα στο μοντέλο μας καθώς υπάρχει κρυμμένη πληροφορία

Selected attribute

Name: contact
Missing: 0 (0%) Distinct: 3

No.	Label	Count
1	unknown	13020
2	cellular	29285
3	telephone	2906

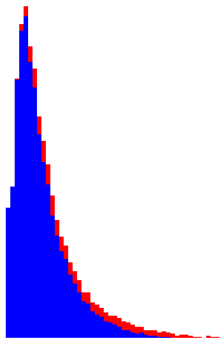
Day και month παίρνουν τις τιμές 1-31 και Jan-Dec ως αναμενόμενο

Το duration μας δείχνει πόσο κράτησε η επικοινωνία με τον πελάτη και έχει μέση διάρκεια 258 δευτερόλεπτα με μία συγκέντρωση προς τα αριστερά

Name: duration
Missing: 0 (0%) Distinct: 1573

Statistic	Value
Minimum	0
Maximum	4918
Mean	258.163
StdDev	257.528

lass: y (Nom)

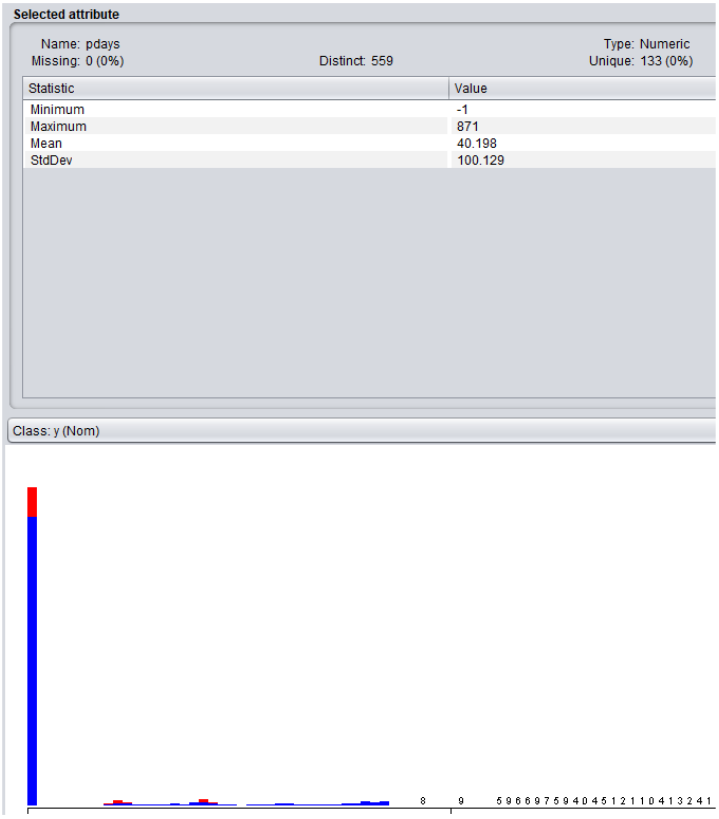


Για το campaign που μας δείχνει τους κωδικούς διαφορετικών προωθητικών ενεργειών έχουμε 48 διακριτές τιμές.

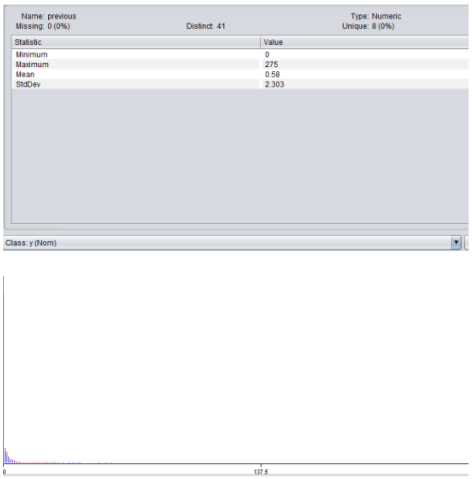
Name: campaign
Missing: 0 (0%) Distinct: 48

Statistic	Value
Minimum	1
Maximum	63
Mean	2.764
StdDev	3.098

Το pdays είναι imbalanced



Imbalanced είναι και το previous



Στο poutcome που μας δείχνει το αποτέλεσμα προηγούμενης προωθητικής ενέργειας στον ίδιο πελάτη. κυριαρχεί το unkown value

Selected attribute

Name: poutcome
Missing: 0 (0%)
Distinct: 4
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	unknown	36959	36959.0
2	failure	4901	4901.0
3	other	1840	1840.0
4	success	1511	1511.0

Τέλος όσον αφορά την εξαρτημένη μεταβλητή κυριαρχεί το no (ο πελάτης δεν άνοιξε προθεσμιακή) Υπάρχει μια σχετική ανισορροπία το οποίο ίσως δυσκολέψει το μοντέλο μας να μάθει.

Selected attribute

Name: y
Missing: 0 (0%)
Distinct: 2

No.	Label	Count
1	no	39922
2	yes	5289

5.4 Αλγόριθμοι ταξινόμησης

Η ταξινόμηση είναι δηλαδή η κατηγοριοποίηση παρατηρήσεων σε κλάσεις είναι μία από τις ποιο συχνά χρησιμοποιούμενες τεχνικές στο χώρο της μηχανικής μάθησης. Το ζητούμενο είναι να μάθει το μοντέλο μας από εκπαιδευτικά παραδείγματα ποιες ιδιότητες παρουσιάζουν οι παρατηρήσεις που ανήκουν σε κάθε κλάση με σκοπό να μπορέσουμε να ταξινομήσουμε και νέα άγνωστα δεδομένα.

Υπάρχει μία πληθώρα αλγορίθμων που προσπαθούν να επιλύσουν αυτό το πρόβλημα. Εμείς θα επικεντρωθούμε σε κάποιους αλγόριθμους που μας παρέχει το WEKA και συγκεκριμένα στους J48, RandomForest, RandomTree και RepTree.

J48

Ο J48 είναι η υλοποίηση στο WEKA του αλγορίθμου C4.5 που ανέπτυξε ο Ross Quinian. Στην ουσία είναι ένας αλγόριθμος που ανήκει στην κατηγορία των δέντρων αποφάσεων. Στα δέντρα αποφάσεων έχουμε ένα σπάσιμο των δεδομένων σε κάθε κλαδί στηριζόμενοι στην θεωρία της εντροπίας. Σε κάθε κόμβο του δέντρου προκαλούμε έναν διαχωρισμό των δεδομένων βάσει του ποια ιδιότητα των δεδομένων μας διαχωρίζει καλύτερα τις κατηγορίες. Η μεταβλητή η οποία επιλέγεται κάθε φορά είναι αυτή που μας δίνει το μεγαλύτερο information gain. Είναι από τους πιο συχνά χρησιμοποιούμενους αλγορίθμους σύμφωνα με τους δημιουργούς του WEKA.

Random Forest

Τα τυχαία δάση αποφάσεων ανήκουν και αυτά στην μεγάλη κατηγορία των αλορίθμων που προέρχονται από δέντρα αποφάσεων αλλά σε αυτή την περίπτωση έχουμε μια μεγάλη διαφοροποίηση καθώς ανήκουν στον χώρο των μοντέλων που χαρακτηρίζουμε ως ensemble learning. Αναπτύχθηκαν από τον Tin Kam Ho και αντιμετωπίζουν το πρόβλημα του overfitting. Το πρόβλημα που έχουν όμως είναι ότι είναι δύσκολο να κατανοήσουμε τα μοντέλα που παράγουν. Τα random forest επιτελούν αυτό που ονομάζουμε bootstrap aggregating.

Random Tree

Τα random trees είναι ένας αλγόριθμος λιγότερο περίπλοκος σε σχέση με τα Random Forest. Και σε αυτή την περίπτωση χρησιμοποιούμε κάποια αρχικούς τυχαίους κόμβους για το σπάσιμο στο

επόμενο επίπεδο και βλέπουμε πιο τυχαίο αποτέλεσμα τα πήγε καλύτερα και το κρατάμε. Είναι πιο γρήγορα σε εκτέλεση από τα random forest.

RepTree

Είναι άλλη μία περίπτωση δέντρων αποφάσεων. Το θεωρητικό τους υπόβαθρο έχει να κάνει με το σπάσιμο σε κόμβους βάσει του information gain αλλά προσπαθεί να μειώσει το βάθος των επιπέδων επιλέγοντας να θυσιάσει λίγη ακρίβεια με σκοπό να πετύχει μικρότερα μοντέλα που είναι πιο εύκολα όσον αφορά την κατανόησή τους.

Σημαντική όμως για κάθε αλγόριθμο που επιλέγουμε να τρέξουμε είναι και η παραμετροποίηση του. Όταν λέμε παραμετροποίηση εννοούμε τον καθορισμό παραμέτρων οι οποίες τροποποιούν την συμπεριφορά του αλγορίθμου και τις αποφάσεις που αυτός παίρνει. Θα χρησιμοποιήσουμε συγκεκριμένα τις παραμέτρους percentage split και cross validation. Το percentage split αναφέρεται στο ποσοστό των παρατηρήσεων των οποίων αφήνουμε εντός των εκπαιδευτικών δειγμάτων για να μάθει ο αλγόριθμος την αλήθεια. Το υπόλοιπο ποσοστό χρησιμοποιείται για την εκτίμηση της απόδοσης του αλγορίθμου.

Αντίστοιχη λειτουργία έχει και η παράμετρος cross validation. Στην απλούστερη περίπτωση ,leave one out cross validation , κρατάμε μία παρατήρηση σαν τεστ δεδομένο και εκπαιδεύουμε τον αλγόριθμο με τις υπόλοιπες παρατηρήσεις. Αυτό το κάνουμε όμως πολλές φορές κρατώντας κρυμμένη κάθε φορά διαφορετική παρατήρηση. Σε περίπτωση K cross validation αντί για μία κρατάμε εκτός K παρατηρήσεις.

5.5 Εκτέλεση αλγορίθμων με διαφορετικές παραμέτρους.

Η διαδικασία της εύρεσης του βέλτιστου μοντέλου είναι επαναληπτική. Δηλαδή δεν τρέχουμε μια φορά ένα μοντέλο και σταματάμε. Επιλέγουμε την χρήση διάφορων αλγορίθμων αλλά και διαφορετικούς τρόπους και τιμές παραμετροποίησης για να βρούμε το βέλτιστο μοντέλο.

Χωρίς να προκαλέσουμε κάποια μεταβολή στα δεδομένα εισόδου τρέχουμε τους αλγόριθμους classification J48, RandomTree και κάνουμε παραμετροποίηση:

Α) με percentage split 55%,60%, 65%, 70%,75%, 80%, 85%

B) με cross validation για K:5,6,7,8,9,10,11,12

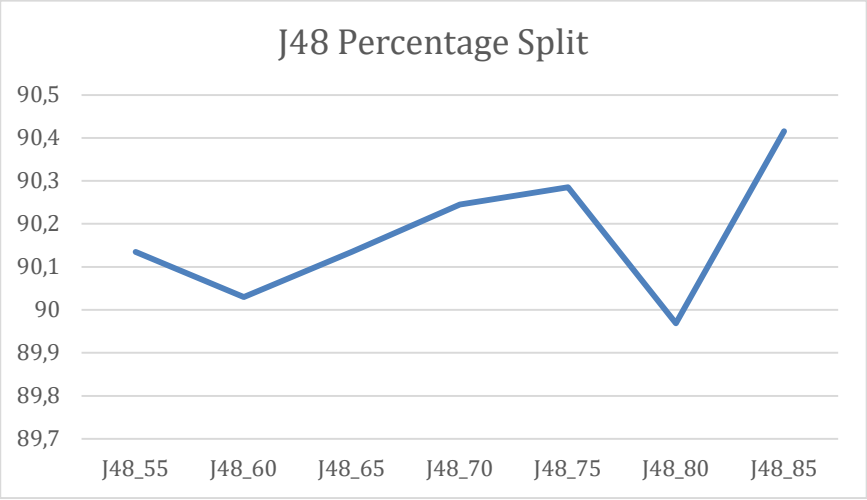
Παρατηρούμε ότι σε αυτή τη φάση ο J48 παράγει μέγεθος δέντρου 1716 με 1168 φύλλα

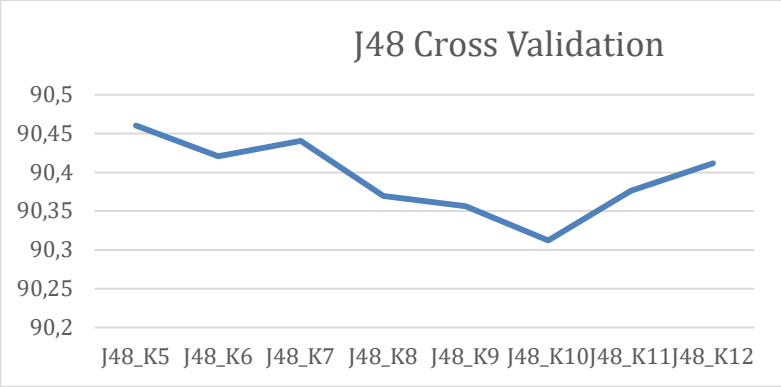
Καταγράφουμε για κάθε αλγόριθμο και κάθε διαφορετική παράμετρο τρεξίματος πόσες εγγραφές ταξινομήθηκαν σωστά (percentage of Correctly Classified Instances)

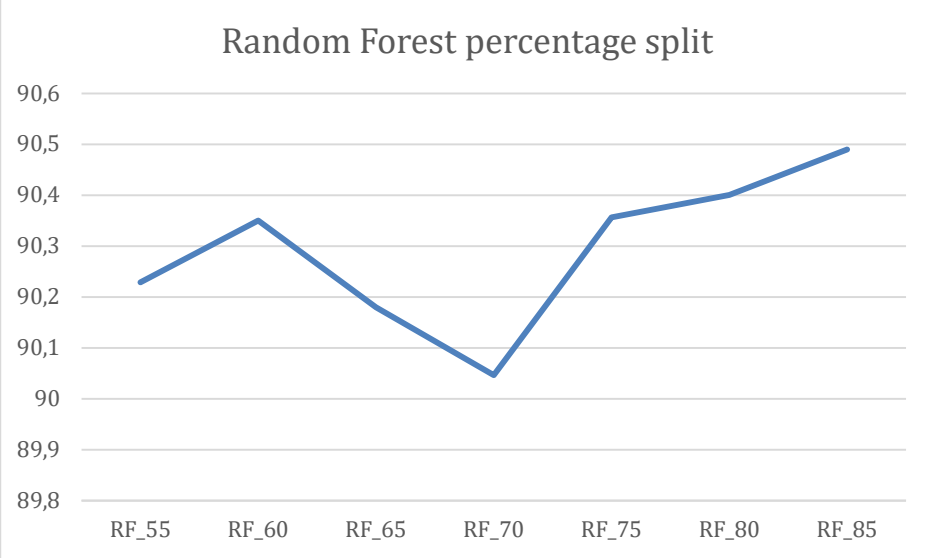
J48_55	90.1352	J48_K5	90.4603	RF_55	90.2286	RF_K5	90.3961
J48_60	90.0299	J48_K6	90.4205	RF_60	90.3506	RF_K6	90.416
J48_65	90.1352	J48_K7	90.4404	RF_65	90.1795	RF_K7	90.4625
J48_70	90.2455	J48_K8	90.3696	RF_70	90.0464	RF_K8	90.3984
J48_75	90.2858	J48_K9	90.3563	RF_75	90.3565	RF_K9	90.478
J48_80	89.969	J48_K10	90.3121	RF_80	90.4004	RF_K10	90.3895
J48_85	90.4158	J48_K11	90.3762	RF_85	90.4895	RF_K11	90.4957
		J48_K12	90.4116			RF_K12	90.447

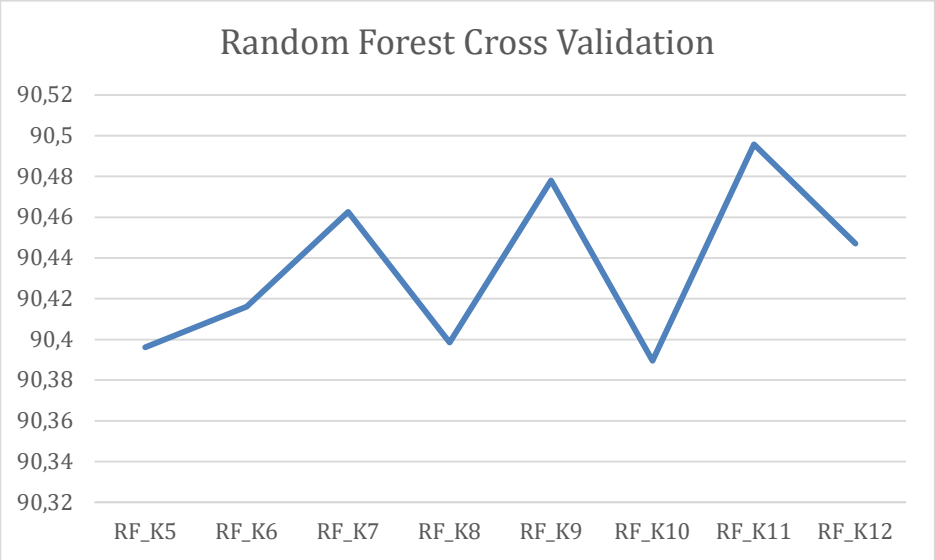
RanTr_55	87.6186	RanTr_K5	87.231	RepTr_55	89.7174	RepTr_K5	89.9936
RanTr_60	87.4364	RanTr_K6	87.7663	RepTr_60	89.9359	RepTr_K6	90.0577
RanTr_65	87.3862	RanTr_K7	87.6114	RepTr_65	89.6234	RepTr_K7	90.2568
RanTr_70	87.3774	RanTr_K8	87.565	RepTr_70	89.8842	RepTr_K8	90.0489
RanTr_75	86.9592	RanTr_K9	87.534	RepTr_75	90.1177	RepTr_K9	90.1705
RanTr_80	87.1267	RanTr_K10	87.2996	RepTr_80	89.6925	RepTr_K10	90.2236
RanTr_85	87.8502	RanTr_K11	87.3128	RepTr_85	90.0029	RepTr_K11	90.0776
		RanTr_K12	87.795			RepTr_K12	90.0599

Παρατηρούμε ότι ο J48 και ο RandomForest είχαν την καλύτερη απόδοση. Οπότε τους μελετάμε περεταίρω για να δούμε τι ρόλο έπαιξαν οι διαφορετικοί παράμετροι


J48 Percentage Split
90,5
90,4
90,3
90,2
90,1
90
89,9
89,8
89,7
J48_55
J48_60
J48_65
J48_70
J48_75
J48_80
J48_85


J48 Cross Validation
90,5
90,45
90,4
90,35
90,3
90,25
90,2
J48_K5
J48_K6
J48_K7
J48_K8
J48_K9
J48_K10
J48_K11
J48_K12


Random Forest percentage split
90,6
90,5
90,4
90,3
90,2
90,1
90
89,9
89,8
RF_55
RF_60
RF_65
RF_70
RF_75
RF_80
RF_85



Σαν γενικό συμπέρασμα μπορούμε να δούμε ότι όσο αυξάνουμε το percentage split βελτιώνεται η απόδοση ενώ στο cross validation αρκεί μία μέση τιμή.

Επιλέγουμε να μελετήσουμε την καλύτερη απόδοση που πετύχαμε από J48. Αυτό ήταν 90.4603 για cross validation me K=5

Correctly Classified Instances	40898	90.4603 %
Incorrectly Classified Instances	4313	9.5397 %
Kappa statistic	0.4931	
Mean absolute error	0.1258	
Root mean squared error	0.2757	
Relative absolute error	60.9012 %	
Root relative squared error	85.7794 %	

Total Number of Instances 45211

=== Confusion Matrix ===

a b <-- classified as

38308 1614 | a = no

2699 2590 | b = yes

Βλέπουμε ότι στην περίπτωση που το αποτέλεσμα ήταν no κατατάξαμε σωστά 38.308 περιπτώσεις αλλά χάσαμε 2.699. Στην περίπτωση όμως που το αποτέλεσμα ήταν yes δηλαδή ο πελάτης όντως άνοιξε την προθεσμιακή κατάθεση κατατάξαμε σωστά 2.590 περιπτώσεις αλλά χάσαμε 1.614. Αυτό είναι αρκετά άσχημο αλλά αναμενόμενο καθώς η εξαρτημένη μεταβλητή παρουσιάζει μεγάλη ανισορροπία.

Παρατηρούμε και τα υπόλοιπα confusion matrices. Βλέπουμε ότι έχουμε το ίδιο πρόβλημα.

Επίσης επιλεκτικά διαλέγουμε έναν κανόνα προς σχολιασμό.

Αν η διάρκεια (duration) είναι μεταξύ 410 και 524 και το poutcome είναι unknown, για ηλικία <=60 αν η επαφή (contact) έγινε μέσω κινητού , αν το επάγγελμα είναι management και η εκπαίδευση άγνωστη τότε το split μας δίνει 2 περιπτώσεις που πετύχαμε θετικό (yes) outcome.

5.6 Διακριτοποίηση και νέα εκτέλεση αλγορίθμων

Έχει παρατηρηθεί ότι μια αρχική επεξεργασία των δεδομένων μπορεί να βοηθήσει τον αλγόριθμο να μάθει καλύτερα και να δημιουργήσει μοντέλα που είναι πιο αποδοτικά αλλά και έχουν και λιγότερο overfit. Μία από αυτές τις διαδικασίες είναι η διακριτοποίηση η οποία ομαδοποιεί τις τιμές των ανεξάρτητων μεταβλητών. Επιλέγουμε να επέμβουμε στις ανεξάρτητες μεταβλητές και να εφαρμόσουμε διακριτοποίηση. Επιλέγουμε από τα φίλτρα του Weka Discretize.

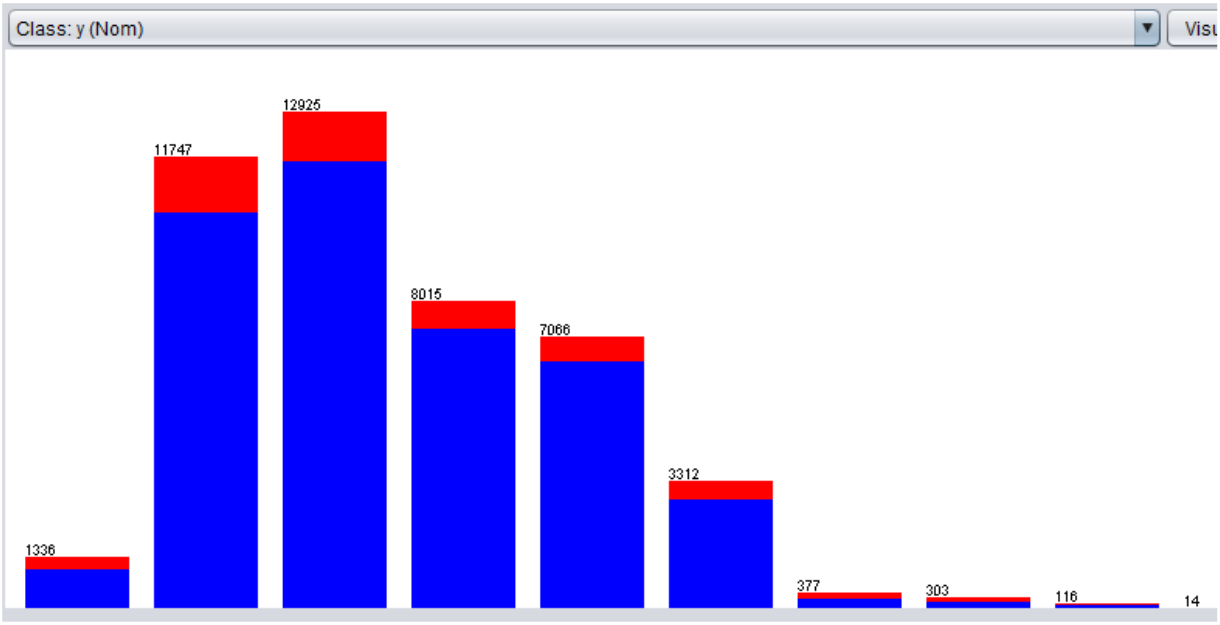
Παρατηρούμε ότι το weka ομαδοποιεί τις τιμές σε γκρουπ.

Πχ για την ηλικία έχουμε την εξής εικόνα

Selected attribute

Name: age
Missing: 0 (0%)
Distinct: 10
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	'(-inf-25.7]'	1336	1336.0
2	'(25.7-33.4]'	11747	11747.0
3	'(33.4-41.1]'	12925	12925.0
4	'(41.1-48.8]'	8015	8015.0
5	'(48.8-56.5]'	7066	7066.0
6	'(56.5-64.2]'	3312	3312.0
7	'(64.2-71.9]'	377	377.0
8	'(71.9-79.6]'	303	303.0
9	'(79.6-87.3]'	116	116.0
10	'(87.3-inf)'	14	14.0



Ξανατρέχουμε όλους τους αλγόριθμους για να δούμε πως μεταβάλλεται η απόδοση.

Καταρχάς βλέπουμε ότι όσον αφορά τον J48 έχουμε πια μέγεθος δέντρου 521 και αριθμό φύλλων 443, έχουμε δηλαδή ένα πιο μικρό δέντρο. Αυτό είναι μεγάλη βελτίωση.

J48_55	89.9926	J48_K5	89.8498	RF_55	89.0096	RF_K5	89.1022
J48_60	89.8584	J48_K6	89.8587	RF_60	88.8907	RF_K6	89.1597
J48_65	89.7561	J48_K7	89.7857	RF_65	88.884	RF_K7	89.1022
J48_70	89.6336	J48_K8	89.8188	RF_70	88.793	RF_K8	89.1354

J48_75	89.7638	J48_K9	89.8542	RF_75	88.8968	RF_K9	89.1022
J48_80	89.6262	J48_K10	89.8786	RF_80	88.841	RF_K10	89.1288
J48_85	89.4869	J48_K11	89.8808	RF_85	88.8381	RF_K11	89.1354
		J48_K12	89.8675			RF_K12	89.131

RanTr_55	87.422	RanTr_K5	87.4389	RepTr_55	89.6289	RepTr_K5	89.4981
RanTr_60	86.9608	RanTr_K6	87.6203	RepTr_60	89.8308	RepTr_K6	89.7569
RanTr_65	86.9818	RanTr_K7	87.45	RepTr_65	89.4085	RepTr_K7	89.7171
RanTr_70	86.9277	RanTr_K8	87.4831	RepTr_70	89.0732	RepTr_K8	89.7127
RanTr_75	87.3308	RanTr_K9	87.3438	RepTr_75	89.5603	RepTr_K9	89.6618
RanTr_80	87.1489	RanTr_K10	87.3571	RepTr_80	88.9847	RepTr_K10	89.7547
RanTr_85	87.715	RanTr_K11	87.6977	RepTr_85	88.9266	RepTr_K11	89.7237
		RanTr_K12	87.6203			RepTr_K12	89.7525

Παρατηρούμε ότι στους J48 και RandomForest είχαμε πτώση απόδοσης ενώ στους άλλους αλγόριθμους δεν υπήρξε ιδιαίτερη αλλαγή. Αυτό δεν είναι αυτό που περιμέναμε καθώς θα ήταν λογικό να αναμένουμε ότι το discretize θα βοηθούσε τους αλγόριθμους να μάθουν πιο εύκολα. Ίσως όμως είναι πιο robust για νέα δεδομένα.

Όσον αφορά τα confusion matrices παρατηρούμε το ίδιο πρόβλημα, ότι χάνουμε ποσοστιαία πολλές περιπτώσεις που θα έπρεπε να έχουμε κάνει classify ως yes. Ενδεικτικά παραθέτουμε τα confusion matrices για τον J48 και όλες τις τιμές percentage split

=== Confusion Matrix J48 55 ===

a b <-- classified as

7765 188 | a = no

750 339 | b = yes

=== Confusion Matrix J48 60 ===

a b <-- classified as

15555 396 | a = no

1438 695 | b = yes

=== Confusion Matrix J48 65===

a b <-- classified as

13632 332 | a = no

1289 571 | b = yes

=== Confusion Matrix J48 70===

a b <-- classified as

11615 328 | a = no

1078 542 | b = yes

=== Confusion Matrix J48 75===

a b <-- classified as

9698 265 | a = no

892 448 | b = yes

=== Confusion Matrix J48 80===

a b <-- classified as

7765 188 | a = no

750 339 | b = yes

=== Confusion Matrix J48 85===

a b <-- classified as

5801 158 | a = no

555 268 | b = yes

5.7 Feature selection και νέα εκτέλεση αλγορίθμων

Επίσης πολλή σημαντική για την μηχανική μάθηση είναι η επιλογή των ανεξαρτήτων μεταβλητών που θα χρησιμοποιήσουμε για την εκμάθηση. Περιορισμός του πλήθους των μεταβλητών οδηγεί σε πιο απλά μοντέλα που είναι εύκολο να γενικευθούν. Το WEKA παρέχει διάφορες μεθόδους για feature selection. Εμείς επιλέγουμε να χρησιμοποιήσουμε την μέθοδο CorrelationAttributeEval η οποία ελέγχει τον βαθμό συσχέτισης κάθε ανεξάρτητης μεταβλητής με την εξαρτημένη και μας τις ταξινομεί όσον αφορά αυτό τον βαθμό συσχέτισης. Επιλέγουμε να αφαιρέσουμε τα χαρακτηριστικά «default», «balance», «duration», «pdays», «previous», «poutcome» καθώς αυτά έχουν τον μικρότερο βαθμό συσχέτισης. και ξανατρέχουμε τους αλγόριθμους .

Αυτή τη φορά βλέπουμε ότι όσον αφορά τον J48 έχουμε πια μέγεθος δέντρου 383 και αριθμό φύλλων 361, έχουμε δηλαδή ένα ακόμα πιο μικρό δέντρο.

Αυτό είναι καλό γιατί το μοντέλο μας απλοποιήθηκε.

Κοιτάμε και την νέα απόδοση.

J48_55	88.5869	J48_K5	88.6709	RF_55	87.2057	RF_K5	
J48_60	88.5589	J48_K6	88.6444	RF_60	87.3092	RF_K6	
J48_65	88.6122	J48_K7	88.6576	RF_65	87.3989	RF_K7	
J48_70	88.4465	J48_K8	88.6908	RF_70	87.5323	RF_K8	
J48_75	88.3748	J48_K9	88.7328	RF_75	87.4458	RF_K9	
J48_80	88.1774	J48_K10	88.5957	RF_80	87.4253	RF_K10	
J48_85	88.1009	J48_K11	88.6244	RF_85	87.2899	RF_K11	
		J48_K12	88.6576			RF_K12	

RanTr_55	85.9179	RanTr_K5	86.4303	RepTr_55	88.2526	RepTr_K5	88.4187
RanTr_60	86.2254	RanTr_K6	86.4723	RepTr_60	88.2603	RepTr_K6	88.4165
RanTr_65	86.1223	RanTr_K7	86.5829	RepTr_65	88.4353	RepTr_K7	88.4055
RanTr_70	86.2199	RanTr_K8	86.333	RepTr_70	88.2106	RepTr_K8	88.4409
RanTr_75	86.3134	RanTr_K9	86.5851	RepTr_75	88.4102	RepTr_K9	88.4519
RanTr_80	86.4189	RanTr_K10	86.479	RepTr_80	88.0557	RepTr_K10	88.4431
RanTr_85	86.1398	RanTr_K11	86.448	RepTr_85	88.0714	RepTr_K11	88.4099
		RanTr_K12	86.5564			RepTr_K12	88.4563

Παρατηρούμε για όλους τους αλγόριθμους μια μικρή πτώση. Η πτώση είναι όμως μικρή με το κέρδος ότι χρησιμοποιήσαμε λιγότερες μεταβλητές με αποτέλεσμα πιο απλό μοντέλο. Το καλύτερο αποτέλεσμα το πήραμε από J48 και cross validation για K=9. Όσον αφορά τα confusion matrices παρατηρούμε το ίδιο πρόβλημα, ότι χάνουμε ποσοστιαία πολλές περιπτώσεις που θα έπρεπε να έχουμε κάνει classify ως yes.

5.8 Apriori αλγόριθμος

Στις προηγούμενες παραγράφους πειραματιστήκαμε με διαφορετικούς αλγορίθμους και διαφορετικές παραμετροποιήσεις. Όλες αυτές οι μέθοδοι όμως ανήκαν στην ευρύτερη ομάδα των δέντρων αποφάσεων. Σε αυτή την ενότητα θα ασχοληθούμε με τον αλγόριθμο apriori. Στον αλγόριθμό apriori για ένα σύνολο παρατηρήσεων προσπαθούμε να βρούμε κανόνες που θα προβλέπουν την εμφάνιση ενός στοιχείου Y με βάση την εμφάνιση ενός άλλου στοιχείου X. Από τις βασικότερες έννοιες στον αλγόριθμο apriori είναι η εμπιστοσύνη (confidence), η υποστήριξη (support) η ανέλκυση (lift) και η πίστη (conviction) . Η εμπιστοσύνη του κανόνα X->Y αποτελεί την δεσμευμένη πιθανότητα να εμφανισθεί το Y δεδομένου ότι έχει ήδη εμφανισθεί το X. Η υποστήριξη του κανόνα X->Y είναι το ποσοστό των παρατηρήσεων (σε σχέση με το πλήθος των δεδομένων) που περιέχουν και το X και το Y. Η ανέλκυση (lift) δείχνει πόσο είναι να προκύψει το Y όταν έχουμε το X ελέγχοντας όμως για τη συχνότητα εμφάνισης του Y. Η πίστη συγκρίνει την πιθανότητα να εμφανισθεί το X χωρίς το Y αν ήταν εξαρτημένα σε σχέση με την πραγματική συχνότητα εμφάνισης του X χωρίς ταυτόχρονη εμφάνιση του Y.

Επιλέγουμε στο WEKA να εφαρμόσουμε τον αλγόριθμο apriori.

Αυτοί είναι οι καλύτεροι κανόνες:

1. loan=no contact=cellular 24485 ==> campaign='(-inf-7.2]' 23379 <conf:(0.95)> lift:(1.01) lev:(0) [155] conv:(1.14)

2. contact=cellular 29285 ==> campaign='(-inf-7.2]' 27929 <conf:(0.95)> lift:(1.01) lev:(0) [152] conv:(1.11)

3. housing=yes 25130 ==> campaign='(-inf-7.2]' 23942 <conf:(0.95)> lift:(1) lev:(0) [106] conv:(1.09)

4. loan=no 37967 ==> campaign='(-inf-7.2]' 36033 <conf:(0.95)> lift:(1) lev:(0) [21] conv:(1.01)

5. contact=cellular y=no 24916 ==> campaign='(-inf-7.2]' 23635 <conf:(0.95)> lift:(1) lev:(0) [2] conv:(1)

6. marital=married 27214 ==> campaign='(-inf-7.2]' 25744 <conf:(0.95)> lift:(1) lev:(-0) [-68] conv:(0.95)

7. loan=no y=no 33162 ==> campaign='(-inf-7.2]' 31333 <conf:(0.94)> lift:(1) lev:(-0) [-120] conv:(0.93)

8. y=no 39922 ==> campaign='(-inf-7.2]' 37707 <conf:(0.94)> lift:(1) lev:(-0) [-158] conv:(0.93)

9. marital=married y=no 24459 ==> campaign='(-inf-7.2]' 23051 <conf:(0.94)> lift:(0.99) lev:(-0) [-148] conv:(0.89)

10. housing=yes 25130 ==> y=no 23195 <conf:(0.92)> lift:(1.05) lev:(0.02) [1004] conv:(1.52)

6. Συμπέρασμα

Η τεχνητή νοημοσύνη (AI), που μερικές φορές ονομάζεται μηχανική νοημοσύνη, είναι η ευφυΐα που «αποδεικνύεται» από τις μηχανές, σε αντίθεση με τη φυσική νοημοσύνη που εμφανίζεται από ανθρώπους και ζώα. Η επιστήμη των υπολογιστών ορίζει την έρευνα της τεχνητής νοημοσύνης ως τη μελέτη των «ευφυών πρακτόρων»: κάθε συσκευή που αντιλαμβάνεται το περιβάλλον της και αναλαμβάνει ενέργειες που μεγιστοποιούν τις πιθανότητες επιτυχίας της επίτευξης των στόχων της. Συνήθως, ο όρος «τεχνητή νοημοσύνη» χρησιμοποιείται για να περιγράψει μηχανές που μιμούνται ορισμένες «γνωστικές» λειτουργίες που οι άνθρωποι συνδέουν με άλλα ανθρώπινα μυαλά, όπως «μάθηση» και «επίλυση προβλημάτων».

Πολλές από τις προκλήσεις που περιγράφονται στους προηγούμενους ορισμούς είναι εντατικές για τον άνθρωπο και θα μπορούσαν να απαιτήσουν πολύ ανεπτυγμένες δεξιότητες στη μάθηση, τη συλλογιστική, τη λήψη αποφάσεων και την επίλυση προβλημάτων. Επομένως, η Τεχνητή Νοημοσύνη και όλες οι παρεμβαλλόμενες παραλλαγές της (μηχανική μάθηση, μηχανική γνώσης, τεχνητή συλλογιστική, οντολογίες, μέθοδοι βελτιστοποίησης κ.λπ.) σχετίζονται όλο και περισσότερο με τη μηχανική συστημάτων. Από την άλλη πλευρά, τα αναδυόμενα ευφυή συστήματα όπως τα αυτόνομα οχήματα σε όλες τις πλευρές τους (αυτοκίνητα, τρένα, υποβρύχια, αεροσκάφη, πλοία κ.λπ.) φέρνουν επανάσταση στην αντίληψή μας για τις υπηρεσίες. Αυτά τα συστήματα προσφέρουν διαφορετικούς τρόπους λειτουργίας, όπου οι μηχανές μαθαίνουν από τη

δική τους λειτουργία και, θεωρητικά, βελτιώνουν την ποιότητα της υπηρεσίας. Τα ντετερμινιστικά συστήματα δεν προτείνουν τεράστιες προκλήσεις όπως ο τρόπος με τον οποίο πρέπει να πραγματοποιηθεί η πιστοποίηση, δεδομένου ότι θα λειτουργούν διαφορετικά κατά τη διάρκεια ζωής του. Πώς να τα V&V ή ακόμα και πώς να τα διαμορφώσετε σε περίπτωση πιθανών ατυχημάτων, ηθικών πτυχών κ.λπ.

Βιβλιογραφία

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138– 52160.

Akerkar R. Big data computing CRC Press, Taylor & Francis Group, Florida, USA (2014).

Al Nuaimi E., Al Neyadi H., Mohamed N., Al-Jaroodi J. Applications of big data to smart cities

Journal of Internet Services and Applications, 6 (1) (2015), pp. 1-15.

Berente N and Seidel S (2014) Big data & inductive theory development: towards computational grounded theory? In Proceedings of the Americas Conference on Information Systems (Tiwana A and Ramesh B, Eds), Association for Information Systems, Savannah, USA.

Bihani P., Patil S.T. A comparative study of data analysis techniques International Journal of Emerging Trends & Technology in Computer Science, 3 (2) (2014), pp. 95-101.

Boyd, D. K. Crawford, Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon Information, communication & society, 15 (5) (2012), pp. 662-679.

Cheng Y, Qin C, Rusu F. GLADE: big data analytics made easy. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, 2012. pp 697–700.

Crawford K. The hidden biases of big data. Harvard Business Review Blog. Available at: http://blogs.hbr.org/2013/04/the-hidden-biases-in-big-data/ (1 April, 2013) (accessed 5 January 2016).

Demchenko Y, de Laat C, Membrey P. Defining architecture components of the big data ecosystem. In: Proceedings of the International Conference on Collaboration Technologies and Systems, 2014. pp 104–112.

Gandomi A., Haider M. Beyond the hype: Big data concepts, methods, and analytics International Journal of Information Management, 35 (2) (2015), pp. 137-144.

Davenport T.H., Harris J.G. Competing on analytics: The new science of winning

Harvard Business Press (2007).

Diebold FX. On the origin(s) and development of the term “big data”, Penn Institute for Economic Research, Department of Economics, University of Pennsylvania, Tech. Rep. 2012. [Online]. Available: http://economics.sas.upenn.edu/sites/economics.sas.upenn.edu/files/12-037.pdf.

HargittaiE. Is bigger always better? Potential biases of big data derived from social network sites, The ANNALS of the American Academy of Political and Social Science, 659 (1) (2015), pp. 63-76.

Jin, B.W. Wah, X. Cheng, Y. Wang Significance and challenges of big data research

Big Data Research, 2 (2) (2015), pp. 59-64.

Joseph R.C., Johnson N.A. Big data and transformational governmentIT Professional, 15 (6) (2013), pp. 43-48.

Labrinidis A., Jagadish H.V. Challenges and opportunities with big data Proceedings of the VLDB Endowment, 5 (12) (2012), pp. 2032-2033.

Lazer, D. A. Pentland, L. Adamic, S. Aral, A. Baraba´si, D. Brewer, ..., M. Van Alstyne

‘Computational social science’ Science, vol. 323 (no. 5915) (2009), pp. 721-723.

Mayer-Schonberger V, Cukier K. Big data: a revolution that will transform how we live, work, and think. Boston: Houghton Mifflin Harcourt; 2013.

MIT Technology Review The Big Data Conundrum: How to define it? (2013)

Available Online at

https://www.technologyreview.com/s/519851/the-big-data-conundrum-how-to-define-it/

(Accessed 19th May 2016).

Raedt, L. D., Kersting, K., Natarajan, S., & Poole, D. (2016). Statistical relational artificial intelligence: Logic, probability, and computation. Synthesis Lectures on Artificial Intelligence and Machine Learning, 10(2), 1– 189.

Russell, S., & Norvig, P. (2004). Τεχνητή νοημοσύνη. (Φ. Σκουλαρίκης, Επιμ., Τ. Άλβας, Δ. Καρτσακλής, & Φ. Σκουλαρίκης, Μεταφρ.) Αθήνα: Κλειδάριθμος.

Szongott C., Henne B., von Voigt G.Big data privacy issues in public social media 6th IEEE international conference on digital ecosystems technologies (DEST) (2012), pp. 1-6.

Tole A.A.Big data challenges Database Systems Journal, 4 (3) (2013), pp. 31-40.

Waller M.A., Fawcett S.E. Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management Journal of Business Logistics, 34 (2) (2013), pp. 77-84.

Wang Y., Wiebe V.J. Big Data Analytics on the characteristic equilibrium of collective opinions in social networks International Journal of Cognitive Informatics and Natural Intelligence (IJCINI), 8 (3) (2014), pp. 29-44.

Wu X, Zhu X, Wu G-Q, Ding W. Data mining with big data. IEEE Trans Knowl Data Eng. 2014;26(1):97–107.