ΣΧΟΛΗ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ ΚΑΙ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ

ΤΜΗΜΑ ΔΙΟΙΚΗΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ

ΠΠΣ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ ΜΕΣΟΛΟΓΓΙ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

«ΜΕΛΕΤΗ ΤΕΧΝΙΚΩΝ ΚΑΙ ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΚΟΙΝΩΝΙΚΑ ΔΙΚΤΥΑ»

ΝΤΖΩΡΗ ΔΗΜΗΤΡΑ

Μεσολόγγι 2020

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ

ΣΧΟΛΗ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ ΚΑΙ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ

ΤΜΗΜΑ ΔΙΟΙΚΗΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ

ΠΠΣ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ ΜΕΣΟΛΟΓΓΙ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

«ΜΕΛΕΤΗ ΤΕΧΝΙΚΩΝ ΚΑΙ ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΚΟΙΝΩΝΙΚΑ ΔΙΚΤΥΑ»

ΝΤΖΩΡΗ ΔΗΜΗΤΡΑ

Επιβλέπων καθηγητής ή καθηγήτρια

Γριβοκωστοπούλου Φωτεινή

Μεσολόγγι 2020

UNIVERSITY OF PATRAS

SCHOOL OF ECONOMICS & BUSINESS

DEPARTMENT OF MANAGEMENT SCIENCE AND TECHNOLOGY

FORMER DEPARTMENT OF BUSINESS ADMINISTRATION AT MESSOLONGHI

THESIS

«ΜΕΛΕΤΗ ΤΕΧΝΙΚΩΝ ΚΑΙ ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΚΟΙΝΩΝΙΚΑ ΔΙΚΤΥΑ»

ΝΤΖΩΡΗ ΔΗΜΗΤΡΑ

Messolonghi 2020

Η έγκριση της πτυχιακής εργασίας από το Τμήμα Διοικητικής Επιστήμης & Τεχνολογίας του Πανεπιστημίου Πατρών δεν υποδηλώνει απαραιτήτως και αποδοχή των απόψεων του συγγραφέα εκ μέρους του Τμήματος.

ΠΡΟΛΟΓΟΣ

Η ψηφιακή εποχή που είτε ηθελημένα είτε όχι διανύουμε μας καθιστά τμήμα μιας ακούσιας ή εκούσιας συνεχόμενης και καθημερινής ανταλλαγής γνώσεων και πληροφοριών μεταξύ μας και μεταξύ ατόμων σε όλον τον κόσμο. Μέσα από όλη αυτή την παγκόσμια νέα πραγματικότητα μπορούν να προκύψουν και νέες δυνατότητες χρησιμοποίησης όλης αυτής της τεράστιας ποσότητας πληροφοριών στην κατεύθυνση της εκπλήρωσης αναπάντητων ερωτημάτων μέσω της συλλογής, επεξεργασίας και ανάλυσης δεδομένων που η στατιστική αδυνατεί να εκμεταλλευθεί. Στην παρακάτω εργασία αναλύονται οι τεχνικές εξόρυξης δεδομένων μαζί με τις παραμέτρους, τον αντίκτυπο και την σημασία τους για την εξαγωγή πολύτιμων πληροφοριών.

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΞΩΦΥΛΛΟ 1

ΕΞΩΦΥΛΛΟ 2

ΠΡΟΛΟΓΟΣ

ΠΕΡΙΛΗΨΗ

ABSTRACT

ΠΕΡΙΕΧΟΜΕΝΑ

ΣΥΝΤΟΜΟΓΡΑΦΙΕΣ

ΕΙΣΑΓΩΓΗ

ΓΕΝΙΚΟ ΜΕΡΟΣ

ΕΙΔΙΚΟ ΜΕΡΟΣ

ΒΙΒΛΙΟΓΡΑΦΙΑ

ΠΕΡΙΛΗΨΗ

Η εξόρυξη δεδομένων (data mining) είναι μια νέα πραγματικότητα που πλέον επιτρέπει στις επιχειρήσεις, τους οργανισμούς, ακόμη και τους ερασιτέχνες με πηγαίο ενδιαφέρον για μελέτη των τάσεων, να διερευνήσουν πολλές πτυχές της καθημερινότητας από την ιατρική μέχρι τις καταναλωτικές συνήθειες και την χρηματοπιστωτική ικανότητα των δανειοληπτών. Συνιστά ένα αξιόχρηστο εργαλείο που μπορεί να μετατρέψει άναρχα και μη δομημένα δεδομένα σε πολύτιμες πληροφορίες στην κατεύθυνση των σκοπών της αύξησης της επιχειρηματικότητας κ.α. Επομένως, πρόκειται για μια καινοτομία που πρέπει να καταστεί περισσότερο γνωστή στους ενδιαφερόμενους και οι παράμετροί της να αποσαφηνιστούν κατά το μέγιστο δυνατό. Όπως κάθε άλλη καινοτομία, εμπεριέχει κι αυτή τον κίνδυνο υπερβολής, υπέρβασης της δεοντολογίας και καταστρατήγησης των προσωπικών δεδομένων εκατοντάδων χιλιάδων καταναλωτών, επομένως οι συνιστώσες της χρήζουν μέτρου και επισταμένης προσοχής.

Η παρακάτω εργασία - εφαλτήριο για την συγγραφή της οποίας αποτέλεσε το πηγαίο ερευνητικό ενδιαφέρον και η προτροπή των αξιότιμων καθηγητών μου - συνιστά μια δευτερογενή έρευνα που εκπονήθηκε με την μεθοδολογία της βιβλιογραφικής ανασκόπησης των πιο προσφάτως δημοσιευμένων μελετών της τελευταίας δεκαετίας.

Στο γενικό μέρος περιγράφεται η εξόρυξη δεδομένων και οι συνιστώσες της, οι αλγόριθμοι υπό τους οποίους εκτελείται μέχρι την λήψη των προσδοκώμενων αποτελεσμάτων, ενώ παράλληλα συγκρίνεται και αντιπαραβάλλεται με την στατιστική ανάλυση.

Στο ειδικό μέρος, αναλύονται οι τεχνικές και τα δημοφιλέστερα εργαλεία εξόρυξης δεδομένων από τα κοινωνικά δίκτυα με κυριότερο το weka, που συνιστά ένα κοινόχρηστο λογισμικό το οποίο με την βοήθεια ενός νευραλγικά και διαδοχικά εκτελούμενων συνόλου αλγορίθμων, πραγματοποιεί την εξόρυξη δεδομένων από κάθε έναν ξεχωριστό άνθρωπο που γνωρίζει μόνο τις βασικές παραμέτρους της εξόρυξης δεδομένων και τον καθιστά ικανό να εξάγει πολύτιμα συμπεράσματα για τις ανθρώπινες τάσεις.

Τέλος, ακολουθούν συμπεράσματα και υποδείξεις για μελλοντική και συνεχιζόμενη έρευνα σε αυτό το ενδιαφέρον και πολλά υποσχόμενο πεδίο που καλείται εξόρυξη δεδομένων.

Λέξεις κλειδιά : data_mining, algorithms, knowledge, social_media, weka.

ABSTRACT

Data mining is a new reality that now allows businesses, organizations and even amateurs interested in studying trends to explore many aspects of everyday life from medicine to consumer habits and borrowers' financial ability. It is a valuable tool that can turn anonymous and unstructured data into valuable information for the purposes of entrepreneurship growth and more. Therefore, this is an innovation that needs to be made known to stakeholders and its parameters clarified as much as possible. Like all other innovations, it encompasses the embezzlement, overreaching and misappropriation of the personal data of hundreds of thousands of consumers, so its elements need moderation and careful attention.

The following work which is inspired by a deep irresistible research interest and the encouragement of my honorable teachers - is a secondary research based on the methodology of the bibliographic review of the most recent published studies of the last decade.

The general section describes the data mining and its components, the algorithms under which they are performed until the expected results are obtained, and are compared with the statistical analysis.

In the specific section, weka is analyzed as a common and free software, which, with the help of a basic and sequential executable set of algorithms, extracts data from every single person who knows only the basic parameters of the data mining and is capable of drawing valuable conclusions on human tendencies.

Finally, we provide conclusions and suggestions for future and ongoing research in this interesting and highly underlying field called data mining

ΕΙΣΑΓΩΓΗ

ΚΕΦΑΛΑΙΟ 1. **ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ**

Η εποχή που διανύουμε σηματοδοτείται από την αθρόα χρησιμοποίηση του διαδικτύου από τους ανθρώπους όλων των ηλικιών, γενεών και κοινωνικο-οικονομικού υπόβαθρου σε όλο τον κόσμο. Έτσι, η διακίνηση αξιόχρηστης πληροφορίας μέσα από το διαδίκτυο συνιστά μια νέα πραγματικότητα με τεράστια δυναμική για την ανάλυσή της και την εξαγωγή νευραλγικών συμπερασμάτων. Η διαθεσιμότητα των διαδικτυακών δεδομένων, οδήγησε στην ανάπτυξη διαφόρων μεθόδων με σκοπό αυτά να συγκεντρώνονται στρατηγικά και να εισάγονται σε αποθήκες δεδομένων. Οι νέες μέθοδοι ανάλυσης που αναπτύσσονται και αναβαθμίζονται διαρκώς, αποσκοπούν στην προαγωγή της κατανόησής μας για τα διαθέσιμα δεδομένα.

1.1 ΟΡΙΣΜΟΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Η εξόρυξη δεδομένων (data mining), αδρά ορίζεται ως η διαδικασία που αποσκοπεί στη δημιουργία γνώσης από δεδομένα και παρουσιάζει συγκεκριμένα ευρήματα στον χρήστη. Η δημιουργία γνώσης μέσα από την εξόρυξη δεδομένων μπορεί να μεταφραστεί ως η ανακάλυψη νέων και μη τετριμμένων προτύπων, σχέσεων και τάσεων σε δεδομένα χρήσιμα για τον εκάστοτε χρήστη (Fayyad, Piatetsky-Shapiro & Smyth, 1996).

1.2 ΠΕΔΙΑ ΕΦΑΡΜΟΓΗΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Οι μέθοδοι εξόρυξης δεδομένων και μηχανικής μάθησης (machine learning) από την επιστήμη της πληροφορικής παρουσιάζουν μια μεγάλη ποικιλία ικανών και αξιόπιστων μεθόδων που έχουν επιλύσει αποτελεσματικά διάφορα προβλήματα σε διάφορους τομείς. Η ανάπτυξή τους, έχει πυροδοτήσει την ανάπτυξη ενός πεδίου έρευνας σε πολλούς και διακριτούς τομείς που χρήζουν περεταίρω ανάλυσης στον επιστημονικό ή εμπορικό τομέα (Bellinger, Mohomed Jabbar, Zaïane & Osornio-Vargas, 2017).

Τα σύνολα δεδομένων (data sets) υφίστανται κατά παραγγελία και κατ'επίκληση τα τελευταία είκοσι χρόνια συλλογή και ανάλυση από εξειδικευμένες εταιρείες. Έτσι, μπορούν να διερευνηθούν :

- Οι τάσεις του αγοραστικού κοινού

Τα σούπερ μάρκετ πάντα παρακολουθούσαν τον τρόπο με τον οποίο οι άνθρωποι προβαίνουν σε αγορές, ωστόσο τα τελευταία χρόνια ο βαθμός στον οποίο τα σύνολα δεδομένων συλλέγονται και αναλύονται από τον τομέα του λιανικού εμπορίου έχει επιταχυνθεί. Η Tesco, αποτελεί ένα παράδειγμα εταιρείας που κατέχει πλειοψηφική συμμετοχή στην εξόρυξη δεδομένων και την ανάλυσή τους για μεγάλο αριθμό κολοσσιαίων εταιρειών όπως η Coca-Cola, η Mars, η Vodafone κ.α. Με την βοήθεια του συστήματος Tesco Clubcard, τα δεδομένα που συλλέγονται χρησιμοποιούνται για την πρόβλεψη της αγοραστικής και καταναλωτικής τάσης, ακόμη και τον τρόπο πληρωμής (μετρητοίς, με προπληρωμένη/χρεωστική/πιστωτική κάρτα) που χρησιμοποιεί η κάθε ηλικιακή ομάδα. Όπως αναφέρεται, τα παραπάνω προκάλεσαν μεγάλη αύξηση των λειτουργικών κερδών των εταιρειών που κάνουν χρήση των υπηρεσιών της εν λόγω εταιρείας κατά 32% το 2018 (Agarwal, 2014).

- Οι πολιτικοί και οικονομικοί χειρισμοί από τους κυβερνητικούς παράγοντες

Μετά την έλευση και εδραίωση της εξόρυξης δεδομένων, κατέστη δυνατή η πρόσβαση σε οικονομικά στοιχεία για την κυβερνητική οικονομική διαχείριση σε πολύ μεγαλύτερη κλίμακα, με την δημοσίευση των στοιχείων δαπανών COINS που συμπληρώνεται με αναλυτικές δαπάνες άνω των £500 από την τοπική κυβέρνηση. Έκτοτε, έχουν εμφανιστεί αρκετοί οργανισμοί (OpenlyLocal και Auditor Armchair) που αποσκοπούν να δώσουν μια σαφή εικόνα στο κοινό σχετικά με τον τρόπο με τον οποίο η κυβέρνηση ξοδεύει χρήματα. Παρόλο που οι παραπάνω φορείς λειτουργούν σε μέχρι στιγμής σε σχετικά μικρή κλίμακα, έχουν επιτύχει πολλά σε σύντομο χρονικό διάστημα. Έχοντας ως τελικό στόχο τον εντοπισμό της διαδρομής των εσόδων που προέρχονται από τους φορολογούμενους, το Datastore και το Data.gov.uk προωθούν και παρέχουν προσβασιμότητα στα δεδομένα οικονομικής διαχείρισης από την κυβέρνηση και η δράση αναμένεται να συνχρηματοδοτηθεί από την ίδια την κυβέρνηση μελλοντικά.

- Στοιχεία που αφορούν πολέμους κι επιδρομές

Το ημερολόγιο πολέμου Wikileaks είναι το πιο ολοκληρωμένο σύνολο δεδομένων σχετικά με έναν πόλεμο που κυκλοφόρησε ποτέ. Σε αυτό, συμπεριλαμβάνονται εκθέσεις για τους θανάτους πολιτών, τις επιθέσεις και τις επιδρομές που έχουν πραγματοποιηθεί μαζί με τις απώλειες που έχουν επιφέρει αποδεικνύοντας τη ματαιότητα του πολέμου ανά δεκαετία αναφοράς του πολέμου.

- Αποτελεσματικοί τρόποι προσέγγισης του αγοραστικού κοινού (διαφήμιση)

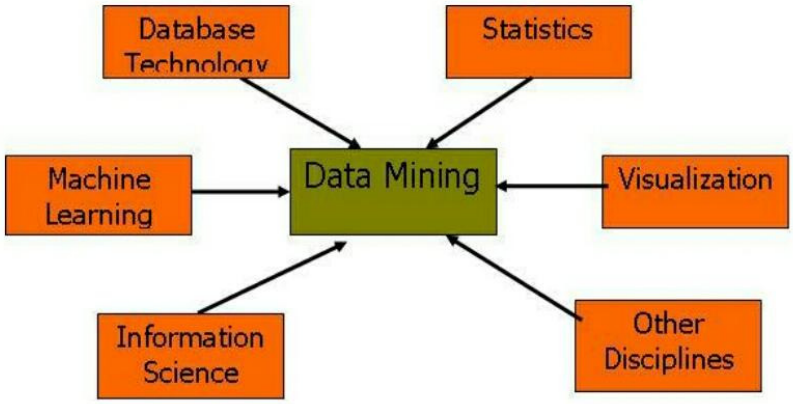
Καμία ιστορία σχετικά με τη χρήση δεδομένων που αλλάζουν τη ζωή των ανθρώπων θα μπορούσε να παραλείψει μια αναφορά της Google. Η Google βασίζεται σε μαθηματικά μοντέλα και εισάγει περισσότερα δεδομένα για την αύξηση των εσόδων και την επιτυχία της. Η Google κατέκτησε τον κόσμο της διαφήμισης με απλά εφαρμοσμένα μαθηματικά γνώριζε τίποτα για τον πολιτισμό και τις συμβάσεις της διαφήμιση μέσω προηγμένων και καινοτόμων αναλυτικών εργαλείων, όπως η εξόρυξη δεδομένων (Anderson, 2017).

Όπως προβλέπεται, ο ψηφιακός κόσμος στον οποίο πλέον είμαστε καθολικά εντεταγμλενοι, μπορεί να αποτελέσει έναν «εγγράψιμο κόσμο» όπου τα φυσικά υποκείμενα όπως οι άνθρωποι διαθέτουν ένα ψηφιακό σήμα που λειτουργεί ως εικονική ταυτότητα που μπορεί να προσδιορίζει οποιοδήποτε άτομο χρησιμοποιεί συστηματικά το διαδίκτυο. Αυτό θα μπορούσε να σημαίνει ότι τα tweets και οι ενημερώσεις σχετικά με την κατάσταση των χρηστών αναφορικά με τα πάντα, θα μπορούσαν να συμβάλουν στην δημιουργία μιας συλλογικής βάσης δεδομένων της οποίας οι αναλυτικές δυνατότητες θα μπορούσαν να είναι ατελείωτες.

ΓΕΝΙΚΟ ΜΕΡΟΣ

ΚΕΦΑΛΑΙΟ 2. ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Η εξόρυξη δεδομένων, ως κεντρικό συστατικό στη ευρύτερη διαδικασία της Ανακάλυψης Γνώσης σε βάσεις δεδομένων (Knowledge Discovery in Databases - KDD), τυγχάνει ευρύτατης εφαρμογής όχι μόνο στην επιστήμη των υπολογιστών (περιοδικά & συνέδρια), αλλά και στον επιστημονικό και επιχειρησιακό κλάδο, ύστερα από την συνειδητοποίηση ότι αυτή η τεράστια ποσότητα επιστημονικών και επιχειρησιακών δεδομένων έχει τη δυνατότητα να αξιοποιηθεί ως επέκταση της επιχειρηματικής υπολογιστικής ευφυΐας. Η εξόρυξη δεδομένων χρησιμοποιεί παραδοσιακά εργαλεία ανάλυσης (π.χ στατιστικά και τα γραφικά) καθώς και εκείνα που συνδέονται με την τεχνητή νοημοσύνη (π.χ τεχνητά νευρωνικά δίκτυα) μ'έναν καινοτόμο τρόπο. Η μεγαλύτερη έμφαση μετατοπίζεται από την εξαγωγή των αποτελεσμάτων, στο ίδιο το θέμα, διαμέσου της δημιουργίας πρωταρχικών υποθέσεων. Με άλλα λόγια, κυρίαρχος στόχος της εξόρυξης δεδομένων είναι η παράθεση ερωτημάτων παρά η παραλαβή απαντήσεων. Σε κάθε περίπτωση, τα στοιχεία που αποκτώνται από την εξόρυξη δεδομένων μπορούν στη συνέχεια να επαληθευτούν με συμβατική ανάλυση (Maimon & Rokach, 2010).



Εικόνα [1]

Η σχετικά νέα διαδικασία της εξόρυξης δεδομένων εφαρμόζεται συνήθως στην εξαγωγή επωφελούς γνώσης από επιχειρηματικά δεδομένα. Ωστόσο, είναι επίσης χρήσιμη σε

[1] Πηγή Asian Journal of Applied Science and Technology (AJAST), 2018

ορισμένες επιστημονικές εφαρμογές όπου αυτή η περισσότερο εμπειρική προσέγγιση συμπληρώνει ή ολοκληρώνει την συμβατική ανάλυση δεδομένων.

2.1 ΤΥΠΙΚΕΣ ΔΙΕΡΓΑΣΙΕΣ ΤΗΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Η εξόρυξη δεδομένων ως διαδικασία περιλαμβάνει, ουσιαστικά, τη συλλογή και επιλογή δεδομένων, την προεπεξεργασία δεδομένων, την ανάλυση δεδομένων συμπεριλαμβανομένης της απεικόνισης των αποτελεσμάτων, την ερμηνεία των ευρημάτων και την εφαρμογή της γνώσης.

2.1.1 ΔΙΑΚΡΙΣΗ ΑΝΑΛΟΓΑ ΜΕ ΤΟΝ ΣΚΟΠΟ

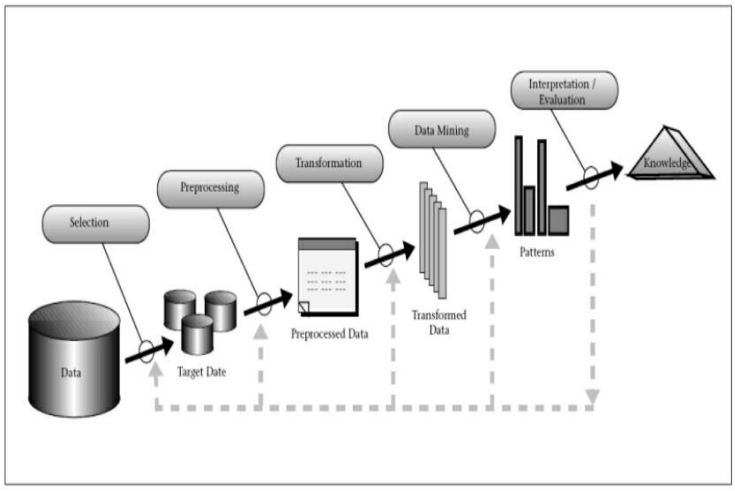
Η εξόρυξη δεδομένων, ανάλογα με τους σκοπούς που εξυπηρετεί διακρίνεται σε προβλεπτική (Predictive Data Mining) και σε περιγραφική.

- Η προβλεπτική είναι μια κατηγορία εξόρυξης δεδομένων που εστιάζει στην ανακάλυψη μιας σχέσης μεταξύ ανεξάρτητων και εξαρτημένων μεταβλητών. Η προβλεπτική εξόρυξη δεδομένων μπορεί να χρησιμοποιηθεί για την πρόβλεψη αμιγών τιμών που βασίζονται σε πρότυπα στα δεδομένα και εφαρμόζεται συνήθως με στόχο τον εντοπισμό στατιστικής σχέσης ή ενός μοντέλου νευρωνικού δικτύου που μπορεί να χρησιμοποιηθεί για την πρόβλεψη κάποιου ενδιαφέροντος αποτελέσματος. Οι προβλεπτικές τεχνικές ανάλυσης δεδομένων μπορούν να χρησιμοποιηθούν, για παράδειγμα, στις πωλήσεις με στόχο την πρόβλεψη του μελλοντικού κέρδους βάσει της προηγούμενης δραστηριότητας πωλήσεων.
- Η περιγραφική εξόρυξη δεδομένων περιγράφει ένα σύνολο δεδομένων με έναν σύντομο αλλά περιεκτικό τρόπο αναδεικνύοντας ενδιαφέροντα χαρακτηριστικά των δεδομένων χωρίς να λαμβάνεται υπόψιν κάποιος προκαθορισμένος στόχος. Οι περιγραφικές τεχνικές δεν προβλέπουν κάποια τιμή, αλλά εστιάζουν περισσότερο στην εγγενή δομή, τις σχέσεις, τη διασύνδεση κλπ. των δεδομένων. Αυτές οι μέθοδοι λαμβάνουν τα δεδομένα που δίνονται και αναδεικνύουν την συσχέτιση (Asian Journal of Applied Science and Technology -AJAST, 2018).

2.1.2 ΦΑΣΕΙΣ ΤΗΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Η διαδικασία της αποκάλυψης γνώσης μέσω της εξόρυξης δεδομένων περιλαμβάνει συνήθως επτά φάσεις από την αρχή μέχρι το τέλος.

- Φάση 1 - Ενσωμάτωση δεδομένων / Συλλογή δεδομένων από πηγές
- Φάση 2 - Επιλογή χρήσιμων δεδομένων
- Φάση 3 - Καθαρισμός δεδομένων / Απαλλαγή δεδομένων από σφάλματα, ελλείπουσες τιμές, ασυμβίβαστα δεδομένα
- Φάση 4 - Μετασχηματισμός δεδομένων / Κανονικοποίηση, εξομάλυνση, μετατροπή σε άλλες μορφές, περισσότερο κατάλληλες για την εξόρυξη δεδομένων
- Φάση 5 - Εξόρυξη δεδομένων / εφαρμογή τεχνικών εξόρυξης με σκοπό την ανακάλυψη κρυμμένων μοτίβων
- Φάση 6 - Αξιολόγηση / παρουσίαση μοτίβου / Οπτικοποίηση και αφαίρεση περιττών σχεδίων
- Φάση 7 - Ανακάλυψη γνώσης / χρησιμοποίηση της γνώσης για λήψη αποφάσεων (Bengio, Buhmann, Embrechts, & Zurada, 2012)



Εικόνα [2]

2.2 ΙΣΤΟΡΙΚΗ ΕΞΕΛΙΞΗ ΤΗΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

[2] Πηγή Bharati & Ramageri, 2010

Η εξόρυξη δεδομένων εξελίχθηκε σημαντικά σε σχέση με το αρχικό εγχείρημα και τις αρχικές προσδοκίες. Ειδικότερα, οι ιστορικές ρίζες της εξόρυξης δεδομένων μπορούν να εντοπιστούν σε τρεις διαδρομές.

- Το παλαιότερο στοιχείο της εξόρυξης δεδομένων αντιστοιχεί στα μονοπάτια της κλασσικής στατιστικής. Η εξόρυξη δεδομένων αποτελεί προέκταση και απότοκο της στατιστικής, χωρίς την οποία δεν θα υπήρχε κανένας τρόπος μέτρησης των δεδομένων. Η στατιστική συνιστά το παλαιότερο και πιο σημαντικό μέρος της εξόρυξης δεδομένων.
- Μια άλλη σχέση συνάφειας της εξόρυξης δεδομένων είναι εκείνη με την τεχνητή νοημοσύνη (artificial intelligence). Η τεχνητή νοημοσύνη επικεντρώνεται στις τεχνικές που βασίζονται στην εμπειρία για την ανακάλυψη της γνώσης, επιχειρώντας να εφαρμόσει διαδικασίες ανθρώπινης σκέψης στα διάφορα στατιστικά προβλήματα.
- Μια τελευταία και πιο ρεαλιστική πορεία της εξόρυξης δεδομένων αποτελεί περισσότερο ο συνδυασμός των δύο προηγούμενων που αδρά αντιστοιχεί στην μηχανική μάθηση και απεικονίζει διάφορες τεχνικές βασισμένες στην εμπειρία μαζί με την προηγμένη γνώση από την στατιστική ανάλυση (Zurada, 1992).

Ο όρος εξόρυξη δεδομένων εισήχθη για πρώτη φορά τη δεκαετία του 1960, όταν πρωτοεμφανίσθηκαν οι δυνατότητες συλλογής δεδομένων. Η εξόρυξη δεδομένων ως μέσο για την ανεύρεση βασικών πληροφοριών μέσα από μεγάλες συλλογές δεδομένων, μοιραία κατέστη δυνατή όταν επιτεύχθηκε για πρώτη φορά η δημιουργία δισκετών, σκληρών δίσκων και υπολογιστών με δυνατότητες συλλογής και αποθήκευσης μεγάλου όγκου δεδομένων. Η δεκαετία του 1980 έφερε τις πραγματικές βάσεις δεδομένων σε ευρύτερη χρήση. Αυτό επέτρεψε την απρόσκοπτη πρόσβαση σε δεδομένα μέσω της δημιουργίας μιας ειδικά σχεδιασμένης για την ανάλυση δεδομένων γλώσσας προγραμματισμού (standard language for storing-SQL). Την δεκαετία του 1990 καθιερώθηκε η χρήση των αποθηκών δεδομένων (Berry & Lindoff, 1997). Αυτή ήταν η χρονική περίοδος κατά την οποία πολλά από τα δεδομένα εξόρυξης που βλέπουμε σήμερα αναπτύχθηκαν εκτενώς.

2.3 ΚΡΙΤΗΡΙΑ ΚΑΤΑΛΛΗΛΟΤΗΤΑΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Η τεχνολογία εξόρυξης δεδομένων θεωρείται κατάλληλη εάν:

- Το επιχειρησιακό πρόβλημα είναι αδόμητο
- Η ακριβής πρόβλεψη είναι πιο σημαντική από την εξήγηση
- Τα δεδομένα περιλαμβάνουν το μείγμα διαστημάτων, ονομαστικών, σειριακών, μετρήσεων και μεταβλητών κειμένου και ο ρόλος και ο αριθμός των μη αριθμητικών μεταβλητών είναι ουσιώδεις
- Μεταξύ αυτών των μεταβλητών υπάρχουν πολλές άσχετες και περιττές ιδιότητες
- Η σχέση μεταξύ των μεταβλητών μπορεί να είναι μη γραμμική
- Τα δεδομένα είναι εξαιρετικά ετερογενή, με μεγάλο ποσοστό υπερβολικών τιμών, σημεία μόχλευσης και ελλείπουσες αξίες
- Το μέγεθος του δείγματος είναι σχετικά μεγάλο
- Σημαντικές είναι οι σημαντικές μελέτες / έργα marketing και πωλήσεων (Wielenga, 2007).

2.4 ΔΙΑΦΟΡΟΠΟΙΗΣΗ ΜΕΤΑΞΥ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ & ΣΤΑΤΙΣΤΙΚΗΣ ΑΝΑΛΥΣΗΣ

Συνοψίζοντας, οι διαφορές μεταξύ της εξόρυξης δεδομένων και της στατιστικής είναι διακριτές και αναφέρονται αδρά στον παρακάτω πίνακα (Brusilovsky, 2009).

ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ	ΣΤΑΤΙΣΤΙΚΗ
Η εξόρυξη δεδομένων έχει σχεδιαστεί για να ασχολείται με τα μη δομημένα δεδομένα και ειδικότερα για την επίλυση μη δομημένων επιχειρηματικών προβλημάτων	Η στατιστική ανάλυση έχει σχεδιαστεί για την αντιμετώπιση δομημένων δεδομένων προκειμένου να επιλυθούν δομημένα προβλήματα: Τα αποτελέσματα είναι ανεξάρτητα από το λογισμικό και τον ερευνητή
Το συμπέρασμα αντανακλά τις υπολογιστικές ιδιότητες του αλγόριθμου εξόρυξης δεδομένων στο χέρι	Το συμπέρασμα αντανακλά τη δοκιμασία των στατιστικών υποθέσεων

Το συμπέρασμα αντανακλά τις υπολογιστικές ιδιότητες και την ικανότητα απεικόνισης του αλγορίθμου εξόρυξης κειμένου στο χέρι	Τα αποτελέσματα εξαρτώνται από λογισμικό και ερευνητή (απουσία προτύπων εφαρμογής)

Πίνακας 1 *διαφορές εξόρυξης δεδομένων &στατιστικής ανάλυσης*

2.5 ΤΟΜΕΙΣ ΕΦΑΡΜΟΓΗΣ

Καθώς η σημασία της εξόρυξης δεδομένων συνεχίζει να αυξάνεται, έχει χρησιμοποιηθεί σε πολλούς τομείς εφαρμογών όπως στις πωλήσεις / μάρκετινγκ, στον τομέα των χρηματοοικονομικών, της ασφάλισης, των τηλεπικοινωνιών, της ανίχνευσης απάτης, της ικανότητας χρηματοδότησης, στον εκπαιδευτικό τομέα, στην ιατρική κ.ο.κ. Μερικοί από τους πιο αξιοσημείωτους τομείς εφαρμογής της εξόρυξης δεδομένων παρατίθενται παρακάτω :

- Εξόρυξη δεδομένων στον τομέα της εκπαίδευσης

Η Εκπαιδευτική Εξόρυξη Δεδομένων (Educational Data Mining - EDM) είναι ένα αναδυόμενο διεπιστημονικό πεδίο έρευνας. Η εκπαιδευτική εξόρυξη δεδομένων αναφέρεται σε τεχνικές και εργαλεία που εφαρμόζονται στις πληροφορίες που προέρχονται από εκπαιδευτικά περιβάλλοντα που σχετίζονται με τους φοιτητές και τις μαθησιακές τους δραστηριότητες και διερευνούν επιστημονικά ερωτήματα εντός του πλαισίου της εκπαιδευτικής έρευνας. Η κατανόηση των μαθησιακών συμπεριφορών των μαθητών και των σπουδαστών είναι πολύ χρήσιμη ως προς την ανάδειξη των νευραλγικών προσαρμογών που μπορούν να εισαχθούν (μέσω της ανάλυσης των δεδομένων του φοιτητή) για την πρόβλεψη των αποτελεσμάτων (Pushpam & Gnana, 2017)

- Εξόρυξη δεδομένων στην ιατρική

Ένα τεράστιο ηλεκτρονικό αρχείο υγείας (Electronic Health Record) είναι διαθέσιμο προς στατιστική ανάλυση και πρόσφατα προς εξόρυξη δεδομένων στον ιατρικό τομέα. Η ακρίβεια θεωρείται ένας εξαιρετικά σημαντικός παράγοντας κατά το χειρισμό αυτών των EHRs που σχετίζονται με την υγεία των ασθενών. Η εξόρυξη δεδομένων, μπορεί να παράγει πληροφορίες που μπορεί να είναι χρήσιμες για όλους τους ενδιαφερόμενους φορείς στην

υγειονομική περίθαλψη, μεταξύ άλλων ασθενών με την ταυτοποίηση αποτελεσματικών θεραπειών και των καλύτερων πρακτικών (Pushpam & Gnana, 2017).

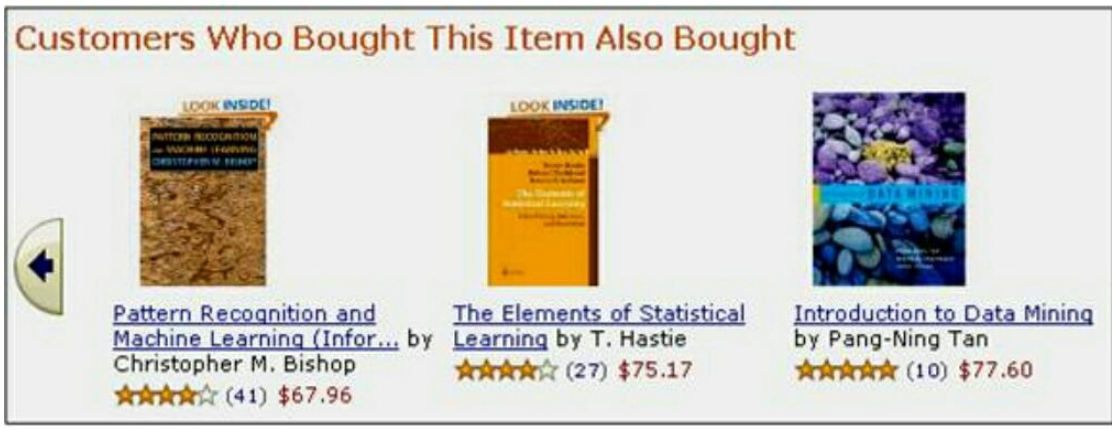
- Εξόρυξη δεδομένων στη γεωργία

Η γεωργία είναι ο πυρήνας της ανθρώπινης ζωής. Πρόσφατα, προκύπτουν αρκετά προβλήματα στον τομέα αυτό. Η εφαρμογή της εξόρυξης δεδομένων στον τομέα της γεωργίας είναι ένας νέος και ελπιδοφόρος τομέας έρευνας. Η ιδέα της χρήσης μοτίβων ενσωματωμένων στον τεράστιο όγκο που προκύπτει από την ανάλυση των δεδομένων παρέχει λύσεις σε πολύπλοκα γεωργικά προβλήματα και την βέλτιστη πρόβλεψη των μελλοντικών τάσεων στις γεωργικές διεργασίες. Για παράδειγμα, οι παράμετροι για το νερό του εδάφους μπορούν να εκτιμηθούν σε έναν συγκεκριμένο τύπο εδάφους γνωρίζοντας τη συμπεριφορά των παρόμοιων τύπων εδάφους (Milovic & Radojevic, 2015).

- Εξόρυξη δεδομένων στο μάρκετινγκ

Λόγω της ταχείας διακύμανσης της αξίας του νομίσματος της συμπεριφοράς των δυναμικών πελατών, είναι πολύ δύσκολο να ληφθούν επενδυτικές αποφάσεις στις επιχειρήσεις, χωρίς την εξόρυξη γνώσεων. Επίσης, η χρηματιστηριακή αγορά αυτογενώς δημιουργεί έναν τεράστιο και αξιοποιήσιμο όγκο δεδομένων. Αυτή η αναγνώριση προσέλκυσε από νωρίς τους ερευνητές να εφαρμόσουν εξόρυξη σε αυτά τα δεδομένα και να βρουν πρότυπα για την πρόβλεψη της πιθανότητας αγοράς ενός προϊόντος με άμεση σύνδεση της με τη μελλοντική τάση ανάπτυξης των επιχειρήσεων (Vijayalakshmi, Mahalakshmi & Magesh, 2013).

Example: Amazon.com purchase suggestion



- Εξόρυξη δεδομένων σε κοινωνικά μέσα

Η ανάλυση των δεδομένων των κοινωνικών μέσων και των μέσων μαζικής ενημέρωσης εξάγει πρότυπα, συσχετισμούς ή τάσεις από ακατέργαστα δεδομένα κοινωνικών μέσων (π.χ. η συχνότητα κι ο τρόπος χρήσης των μέσων ενημέρωσης, οι ηλεκτρονικές συμπεριφορές, η κοινή χρήση περιεχομένου, οι συνδέσεις μεταξύ ατόμων, η ηλεκτρονική συμπεριφορά αγοράς, κ.λ.π). Αυτά τα πρότυπα παρέχουν έγκυρες πληροφορίες σε εταιρείες, κυβερνήσεις και μη κερδοσκοπικούς οργανισμούς, με σκοπό να σχεδιάσουν εξατομικευμένες στρατηγικές ή να τολμήσουν την εισαγωγή νέων προγραμμάτων / προϊόντων / υπηρεσιών (Fernando et.al, 2014).

2.6 ΔΙΑΦΟΡΟΠΟΙΗΣΗ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ

2.6.1 Ανάλυση επιστημονικών δεδομένων

Οι τεχνικές εξόρυξης δεδομένων είναι πολύτιμες όταν παρατηρείται έλλειψη γενικών θεωριών που να μπορούν να εφαρμοστούν σε μεγάλες ποσότητες δεδομένων που πλαισιώνονται από θορυβώδες υπόβαθρο. Η προσέγγιση αυτή, προσδοκεί να αποκτήσει με αυτόματο τρόπο μια θεωρητική γενίκευση από τα δεδομένα με μέσα επαγωγής, εξάγοντας εμπειρικά μοντέλα και μαθαίνοντας από παραδείγματα. Το θεωρητικό αποτέλεσμα, ακόμη κι αν δεν είναι θεμελιώδες, μπορεί να δώσει μια καλή κατανόηση της φυσικής διαδικασίας και να αποδώσει μεγάλη πρακτική χρησιμότητα (Read, 2012).

2.6.2 Ανάλυση επιχειρησιακών δεδομένων

Οι δημοφιλείς εμπορικές εφαρμογές της τεχνολογίας εξόρυξης δεδομένων μπορούν να είναι άμεσες και να καθορίζουν την κατευθυνόμενη ηλεκτρονική αλληλογραφία, την αξιολόγηση της δανειοληπτικής ικανότητας, την πρόβλεψη της ρευστότητας, τη διαπραγμάτευση μετοχών, την ανίχνευση ηλεκτρονικής απάτης και την πελατειακή διαχείριση. Είναι στενά συνδεδεμένη με την αποθήκευση δεδομένων στην οποία μεγάλες εταιρικές βάσεις δεδομένων κατασκευάζονται για εφαρμογές υποστήριξης αποφάσεων. Οι βάσεις δεδομένων είναι συχνά πολυδιάστατες δομές που χρησιμοποιούνται για τη λεγόμενη διαδικτυακή αναλυτική επεξεργασία (on-line analytical processing - OLAP).

2.6.3 Ανάλυση δικτύου πληροφοριών

Με την ανάπτυξη της Google και άλλων αποτελεσματικών μηχανών αναζήτησης, η ανάλυση πληροφοριών κατέστη ένα σημαντικό ερευνητικό σύνορο, με ευρείες εφαρμογές στην ανάλυση κοινωνικών δικτύων, την περιγραφή και κατηγοριοποίηση της κοινότητας των χρηστών του δικτύου, την εξόρυξη πληροφοριών σχετικά με τρομοκρατικές ομάδες, την ανάλυση δικτύων υπολογιστών και την ανίχνευση διαδικτυακής εισβολής. Ωστόσο, η έρευνα στο δίκτυο πληροφοριών πρέπει να υπερβαίνει τα ρητά διαμορφωμένα και ομοιογενή δίκτυα και να εμβαθύνει περισσότερο σε σιωπηρά σχηματισμένα, ετερογενή και πολυδιάστατα δίκτυα πληροφοριών (Han & Gao, 2009).

Η εξόρυξη γνώσης μέσα από το διαδίκτυο επιτρέπει την αναζήτηση μοντέλων δεδομένων μέσω της εξόρυξης περιεχομένου των δεδομένων, της εξόρυξης δομών και της εξόρυξης δεδομένων χρήσης.

- Η εξόρυξη περιεχομένου χρησιμοποιείται για την εξέταση δεδομένων που συλλέγονται από τις μηχανές αναζήτησης.
- Η εξόρυξη δομών χρησιμοποιείται για την εξέταση δεδομένων που σχετίζονται με τη δομή μιας συγκεκριμένης τοποθεσίας Web.
- Η εξόρυξη χρήσης χρησιμοποιείται για την εξέταση δεδομένων που σχετίζονται με το συγκεκριμένο χρήστη, το πρόγραμμα περιήγησης καθώς και τα δεδομένα που συλλέγονται από τις φόρμες που ενδέχεται να έχει υποβάλει ο χρήστης κατά τις συναλλαγές στο διαδίκτυο.

Η εξόρυξη γνώσης μέσα από το διαδίκτυο ως διαδοχική εφαρμογή εξόρυξης προτύπων, ασχολείται με την εύρεση προτύπων πλοήγησης χρηστών στο από τον παγκόσμιο ιστό με την εξαγωγή γνώσεων από τα αρχεία καταγραφής ιστού. Ένα παράδειγμα εφαρμογής διαδοχικών εξορυκτικών προτύπων σε αυτό το θέμα θα ήταν η αναζήτηση ενός μοτίβου στους ιστότοπους που επισκέπτονται οι χρήστες μιας συγκεκριμένης ηλικιακής ομάδας ή ενός φύλου (Asian Journal of Applied Science and Technology- AJAST, 2018)

ΕΙΔΙΚΟ ΜΕΡΟΣ

ΚΕΦΑΛΑΙΟ 3. ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΜΕΣΑ ΑΠΟ ΚΟΙΝΩΝΙΚΑ ΔΙΚΤΥΑ

3.1 ΔΥΝΑΜΙΚΗ ΚΟΙΝΩΝΙΚΩΝ ΔΙΚΤΥΩΝ

Σήμερα, η χρήση των κοινωνικών δικτύων αυξάνεται ασταμάτητα και γρήγορα. Αξιοσημείωτο είναι το γεγονός ότι αυτά τα δίκτυα έχουν γίνει μια ουσιαστική βάση για μη δομημένα δεδομένα που ανήκουν σε πλήθος τομέων, συμπεριλαμβανομένων των επιχειρήσεων, των κυβερνήσεων και της υγείας. Η αυξανόμενη εξάρτηση από τα κοινωνικά δίκτυα απαιτεί τεχνικές εξόρυξης δεδομένων που είναι πιθανό να διευκολύνουν τη μεταρρύθμιση των μη δομημένων δεδομένων και να τα τοποθετήσουν σε ένα συστηματικό πρότυπο.

Αναμφισβήτητα, ο κόσμος συμπυκνώνεται σε ψηφιακές μικρές ή μεγαλύτερες κοινότητες, λόγω της απτής επιρροής των κοινωνικών μέσων. Συνδέει ανθρώπους από διαφορετικά μέρη του κόσμου, ηλικίες και εθνικότητες και τους επιτρέπει να μοιραστούν τις απόψεις, τις εμπειρίες, τα συναισθήματα, τα χόμπι, εικόνες και βίντεο. Το γεγονός αυτό, έδωσε την ευκαιρία στους δημόσιους και ιδιωτικούς οργανισμούς από όλους τους τομείς για να προωθήσουν, να επωφεληθούν, να αναλύσουν, να μάθουν και να βελτιώσουν τις οργανώσεις τους με βάση τα δεδομένα που παρέχονται στα κοινωνικά μέσα. Έτσι, η σημασία των κοινωνικών μέσων για τον ακαδημαϊκό κόσμο και την βιομηχανία είναι αρκετά εμφανείς στο ποσό της έρευνας που διεξάγεται από αυτούς τους δύο τομείς, αναζητώντας απαντήσεις σε καίρια και επίκαιρα ερωτήματα που ολοένα γεννώνται (Injadat, Salo & Nassif, 2016).

Η δομή των δεδομένων των κοινωνικών μέσων είναι ανοργάνωτη και εμφανίζεται με διάφορες μορφές όπως: κείμενο, φωνή, εικόνες και βίντεο. Επιπλέον, τα κοινωνικά μέσα παρέχουν ένα τεράστιο ποσό συνεχών δεδομένων σε πραγματικό χρόνο που καθιστά παραδοσιακά στατιστικές μεθόδους ακατάλληλες για την ανάλυση αυτών των μαζικών δεδομένων. Ως εκ τούτου, οι τεχνικές εξόρυξης δεδομένων μπορούν να διαδραματίσουν σημαντικό ρόλο στην αντιμετώπιση αυτού του προβλήματος (Chen & Chiang, 2012).

Παρά τον μεγάλο αριθμό εμπειρικών ερευνών σχετικά με τις τεχνικές εξόρυξης δεδομένων και τα κοινωνικά μέσα ενημέρωσης, ένας περιορισμένος αριθμός μελετών

συγκρίνετε τις τεχνικές εξόρυξης δεδομένων όσον αφορά την ακρίβεια, την απόδοση και την καταλληλότητα. Για παράδειγμα, παρατηρήθηκε ότι το η ακρίβεια ορισμένων τεχνικών μάθησης μηχανών υπολογίζεται με διάφορες μεθόδους που καθιστούν δύσκολη την εύρεση απαντήσεων στο καταλληλότητα των τεχνικών εξόρυξης δεδομένων (Injadat, Salo & Nassif, 2016).

Εικόνα [3]

Σύνοψη των πιο διαδεδομένων κοινωνικών μέσων

Σήμερα, η χρήση των κοινωνικών δικτύων αυξάνεται γρήγορα και διαδοχικά. Τα κοινωνικά δίκτυα κερδίζουν μαζική δημοτικότητα ως μέσα διάδοσης πληροφοριών και επιπλέον διευκόλυνση των κοινωνικών αλληλεπιδράσεων (Salloum, Al-Emran & Shaalan, 2016). Οι ιστότοποι κοινωνικών μέσων επιτρέπουν στους χρήστες να επικοινωνούν μεταξύ τους μέσω διαφόρων εργαλείων όπως συζητήσεις, φόρουμ, σχόλια κλπ. Αυτό έχει ως αποτέλεσμα την εκμάθηση και την ανταλλαγή πληροφοριών μεταξύ των χρηστών. Στην ψηφιακή αυτή αρένα κυριαρχούν τα μη δομημένα δεδομένα (με όγκο έως 80%) σε σύγκριση με τα δομημένα (με όγκο λιγότερο από 20%) (Chakraborty & Krishna, 2014). Τα δίκτυα αυτά, έχουν μετατραπεί σε μια ουσιαστική βάση για μη δομημένα δεδομένα που ανήκουν σε πλήθος τομέων, συμπεριλαμβανομένων των επιχειρήσεων, των κυβερνήσεων και της υγείας.

[3] Πηγή Adedoyin-Olowe, Gaber & Stahl, 2014

Η αυξανόμενη τάση προς τα κοινωνικά δίκτυα ενέπνευσε την ανάπτυξη τεχνικών εξόρυξης δεδομένων που πιθανότατα να οδηγήσουν στη μετατροπή των μη δομημένων δεδομένων με συστηματικό τρόπο (Injadat, Salo & Nassif, 2016).

Ωστόσο, η φύση των πληροφοριών σχετικά με αυτούς τους ιστοτόπους κοινωνικής δικτύωσης μπορούν να κατηγοριοποιηθούν ως άναρχη και ασαφής. Πολύ περισσότερο, κατά τις τακτικές καθημερινές συζητήσεις η χρήση των greeklish, η παραμελλημένη ορθογραφία, γραμματική και δομή των προτάσεων μπορεί να προκαλέσει διάφορα είδη σημασιολογικών αμφισημιών καθιστώντας δύσκολη την ανάλυση και την εξαγωγή δεδομένων μέσα από τέτοια σύνολα (Salloum, Al-Emran & Shaalan, 2016).

Επιπλέον, πολλές ιστοσελίδες κοινωνικής δικτύωσης συμπεριλαμβανομένων των δημοφιλών Facebook, Instagram και Tweeter, βρίθουν σε γραπτά κείμενα, αναρτήσεις, συνομιλίες και σχόλια στα οποία οι χρήστες επισυνάπτουν διαφορετικούς συνδέσμους ιστοτόπων μέσα στα κείμενά τους και ως εκ τούτου, η εξαγωγή λογικών και ακριβών αποτελεσμάτων μέσα από τα κοινωνικά δίκτυα χαρακτηρίζεται από σχετική περιπλοκότητα.

Δεδομένου οτι η εξόρυξη δεδομένων απαιτεί τεράστια σύνολα δεδομένων προς εξόρυξη αξιόλογων μοτίβων, οι ιστότοποι των κοινωνικών δικτύων μοιάζουν ιδανικοί ειδικά όταν πρόκειται για έκφραση απόψεων / συναισθημάτων (Cortizo et al., 2009).

Το Twitter έχει αποδειχθεί ότι είναι η πιο συχνά χρησιμοποιούμενη microblogging εφαρμογή σήμερα. Με περίπου 500 εκατομμύρια εκτιμώμενους εγγεγραμμένους χρήστες, από το 2012, το Twitter έχει εξελιχθεί σε ένα αξιόπιστο μέσο έκφρασης συναισθήματος αλλά και γνώμης των χρηστών. Είναι επίσης ένα πολύ διαδεδομένο μέσο ενημέρωσης με τα δεδομένα του (ευρέως γνωστά ως tweets) να μπορούν να παρομοιαστούν ως «ειδήσεις σε πραγματικό χρόνο». Τα tweets που δημοσιεύονται στο διαδίκτυο περιλαμβάνουν ειδήσεις, σημαντικά γεγονότα και θέματα τοπικού, εθνικού ή παγκόσμιου ενδιαφέροντος. Τέλος, διαφορετικά γεγονότα / περιστατικά δημοσιεύονται σε πραγματικό χρόνο σε όλο τον κόσμο καθιστώντας το εν λόγω κοινωνικό μέσο ικανό να παράγει σύνολα δεδομένων σε πολύ μικρό χρόνο (Adedoyin-Olowe, Gaber & Stahl, 2014).

Στην αναζήτηση στο διαδίκτυο, ο χρήστης συνήθως αναζητά κάτι που έχει δημοσιοποιηθεί και έχει συσταθεί από άλλο πρόσωπο ή φορέα. Το ζήτημα είναι να απομονωθούν όλα τα δεδομένα που έστω και φαινομενικά είναι ασύνδετα και ανεξάρτητα με το ερευνητικό ερώτημα με στόχο την ανάδειξη, την κατηγοριοποίηση (Classification), την ομαδοποίηση (συσταδοποίηση- clustering), την συσχέτιση (Association) και την των σχετικών δεδομένων.

Εικόνα [4]

3.2 ΤΥΠΙΚΕΣ ΔΙΕΡΓΑΣΙΕΣ ΤΗΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

3.2.1 Κατηγοριοποίηση (Classification)

Η κατηγοριοποίηση (Classification), που συχνά αναφέρεται και ως ταξινόμηση, είναι η διαδικασία μέσα από την οποία ιδέες και αντικείμενα αναγνωρίζονται, διαφοροποιούνται και γίνονται κατανοητά (Kalmegh, 2015). Πρόκειται για μια κοινή εποπτευόμενη προσέγγιση (μάθηση όπου υπάρχει ένα σετ κατάρτισης σωστά προσδιορισμένων διαθέσιμων παρατηρήσεων) με συγκεκριμένη καταλληλότητα όταν το σύνολο δεδομένων (ή τμήμα τους) έχει ετικέτες (labels). Η ετικέτα δεδομένων είναι ένα στατικό μέρος ενός διαγράμματος, μιας αναφοράς ή άλλης δυναμικής διάταξης που αποτελεί αναπόσπαστο μέρος της αναφοράς και της ανάπτυξης εφαρμογών και ορίζει τις πληροφορίες στο στοιχείο γραμμής (Tsoumakas,

[4] Πηγή https://data-flair.training/blogs/data-mining-techniques/

Katakis & Vlahavas, 2010). Οι αλγόριθμοι ομαδοποίησης ξεκινούν με ένα σύνολο δεδομένων εκπαίδευσης που περιλαμβάνει ετικέτες για κάθε ένα στοιχείο δεδομένων. Ο αλγόριθμος (που είναι γνωστός ως ταξινομητής) εκπαιδευόμενος από τα δεδομένα εκπαίδευσης, δομεί ένα μοντέλο που θα κατηγοριοποιήσει αυτόματα τα νέα στοιχεία δεδομένων σε μια από τις διακριτές τάξεις που παρέχονται με τα δεδομένα εκπαίδευσης. Οι κανόνες ταξινόμησης και τα δέντρα αποφάσεων αποτελούν παραδείγματα εποπτευόμενων τεχνικών ταξινόμησης Η ταξινόμηση, ως αλγόριθμος εξόρυξης δεδομένων, δημιουργεί μια αλληλουχία για τον τρόπο προσδιορισμού της εξόδου. Το δέντρο που δημιουργεί είναι ένα δέντρο όπου κάθε διαδοχικός κόμβος αντιπροσωπεύει ένα σημείο όπου η απόφαση πρέπει να ληφθεί με βάση τις εισροές έως ότου φτάσει σε ένα φύλλο που αντιστοιχεί στην προβλεπόμενη έξοδο (Kalmegh, 2015).

3.2.2 Ομαδοποίηση (clustering)

Η ομαδοποίηση (clustering) είναι μια κοινή τεχνική εξόρυξης δεδομένων με σκοπό να βρεθούν ομάδες παρόμοιων εγγράφων σε μια συλλογή εγγράφων χωρίς εποπτεία που είναι χρήσιμη στην διαχείριση συνόλων δεδομένων χωρίς ετικέτες. Σε αντίθεση με τους αλγορίθμους ταξινόμησης, οι αλγόριθμοι ομαδοποίησης δεν εξαρτώνται από την επισήμανση των δεδομένων εκπαίδευσης για την ανάπτυξη ενός μοντέλου. Αντ 'αυτού, οι αλγόριθμοι ομαδοποίησης καθορίζουν ποια στοιχεία στα δεδομένα είναι παρόμοια μεταξύ τους με βάση την ομοιότητα. Η ομοιότητα που καθορίζεται μέσω μιας συνάρτησης, μπορεί να θεωρηθεί ως η απόσταση για ορισμένα σύνολα αριθμητικών δεδομένων, αλλά συχνά σε δεδομένα που σχετίζονται με τα κοινωνικά δίκτυα, οι τεχνικές εξόρυξης πρέπει να εφαρμόζονται σε δεδομένα που αμιγώς προκύπτουν από το κείμενο. Στην περίπτωση αυτή, οι τεχνικές ομαδοποίησης χρησιμοποιούν λέξεις-κλειδιά για να αναπαραστήσουν ένα έγγραφο και το μέτρο ομοιότητας χρησιμοποιείται για να διακρίνει πόσο παρόμοιο είναι ένα στοιχείο δεδομένων με κάποιο άλλο (Anick & Vaithyanathan, 1997)

3.2.3 Πρόβλεψη (Prediction)

Η τεχνική παλινδρόμησης μπορεί να προσαρμοστεί για την πρόβλεψη. Η ανάλυση παλινδρόμησης μπορεί να χρησιμοποιηθεί για το μοντέλο της σχέσης μεταξύ μιας ή περισσοτέρων ανεξάρτητων μεταβλητών και εξαρτημένων μεταβλητών. Στην εξόρυξη δεδομένων, οι ανεξάρτητες μεταβλητές είναι γνωστές πληροφορίες και οι εξαρτημένες μεταβλητές είναι κάτι που θέλουμε να προβλέψουμε. Δυστυχώς, πολλά προβλήματα που

συναντώνται στον πραγματικό κόσμο δεν υπόκεινται σε απλή πρόβλεψη. Για παράδειγμα, ο όγκος πωλήσεων, το απόθεμα, οι τιμές και τα ποσοστά αποτυχίας του προϊόντος είναι πολύ δύσκολο να προβλεφθούν επειδή μπορούν να εξαρτώνται από πολύπλοκες αλληλεπιδράσεις πολλαπλών μεταβλητών πρόβλεψης. Συνεπώς, πιο σύνθετες τεχνικές (π.χ., λογιστική παλινδρόμηση, δέντρα αποφάσεων ή νευρωνικά δίχτυα) μπορεί να είναι απαραίτητες για την πρόβλεψη μελλοντικών τιμών. Οι ίδιοι τύποι μοντέλων μπορούν συχνά να χρησιμοποιηθούν τόσο για παλινδρόμηση όσο και για ταξινόμηση. Για παράδειγμα, το CART (δένδρο ταξινόμησης και παλινδρόμησης) συνιστά έναν αλγόριθμο αποφάσεων που μπορεί να χρησιμοποιηθεί για την κατασκευή και των δύο ταξινομικών δέντρων (για να ταξινομήσει την κατηγορική απάντηση μεταβλητές) και δέντρα παλινδρόμησης (για την πρόβλεψη μεταβλητών συνεχούς απόκρισης). Τα νευρωνικά δίκτυα μπορούν επίσης να δημιουργήσουν τα μοντέλα ταξινόμησης και παλινδρόμησης.

Τύποι μεθόδων παλινδρόμησης

- Γραμμική παλινδρόμηση
- Πολλαπλασιαστική γραμμική παλινδρόμηση
- Μη γραμμική παλινδρόμηση
- Πολλαπλασιαστική μη γραμμική παλινδρόμηση (Bharati & Ramageri, 2010)

3.2.4 Συσχέτιση (Association)

Ο κανόνας της συσχέτισης

Η συσχέτιση είναι συνήθως η εύρεση συνδέσεων που εντοπίζονται συχνά μεταξύ μεγάλων συνόλων δεδομένων. Αυτός ο τύπος ευρημάτων βοηθά τις επιχειρήσεις να λαμβάνουν ορισμένες αποφάσεις, όπως ο σχεδιασμός ενός καταλόγου, το cross marketing και την ανάλυση συμπεριφοράς αγορών του πελάτη. Οι αλγόριθμοι του κανόνα της συσχέτισης πρέπει να είναι σε θέση να παράγουν κανόνες με διάστημα εμπιστοσύνης με τιμές μικρότερες από το 1. Ωστόσο, ο αριθμός των πιθανών κανόνων συσχέτισης για ένα δεδομένο σύνολο δεδομένων είναι γενικά πολύ μεγάλο και ένα μεγάλο ποσοστό των κανόνων είναι συνήθως μικρής αξίας (αν υπάρχει).

Τύποι κανόνων συσχέτισης

- Κανόνας πολυεπίπεδης σύνδεσης
- Πολυδιάστατος κανόνας σύνδεσης
- Ποσοτικός κανόνας σύνδεσης (Bharati & Ramageri, 2010)

3.2.5 Παλινδρόμηση (Regression)

Η παλινδρόμηση είναι μια τεχνική εξόρυξης δεδομένων που χρησιμοποιείται για την πρόβλεψη αριθμών από σύνολα δεδομένων που έχουν γνωστές τιμές. Η παλινδρόμηση είναι μια τεχνική πρόβλεψης εξόρυξης δεδομένων. Παραδείγματα καταστάσεων στις οποίες μπορεί να εφαρμοστεί η παλινδρόμηση είναι: οι πωλήσεις, η απόσταση, η θερμοκρασία, κλπ. Η παλινδρόμηση θα μπορούσε να χρησιμοποιηθεί για την πρόβλεψη της αξίας ενός σπιτιού με βάση την τοποθεσία, τον αριθμό των δωματίων κ.λπ. με την παρατήρηση των δεδομένων του παρελθόντος για τις κατοικίες και την μεταβολή της μεταπωλητικής τους αξίας με την πάροδο του χρόνου. Η γνωστή τιμή για το παράδειγμα αυτό, συνιστά η αξία του σπιτιού.

3.2.6 Διαδοχικά μοτίβα (Sequential Patterns)

Η διαδοχική εξόρυξη προτύπων ανακαλύπτει συχνές ακολουθίες ως μοτίβα σε μια βάση δεδομένων αλληλουχίας. Τα μη επεξεργασμένα πρότυπα χρησιμοποιούνται για περαιτέρω ανάλυση για την αναγνώριση των σχέσεων μεταξύ των δεδομένων. Η διαδοχική ανάλυση προτύπου είναι μια τεχνική πρόβλεψης εξόρυξης δεδομένων. Η βάση δεδομένων ακολουθιών συνιστά μια βάση δεδομένων που αποθηκεύει έναν αριθμό και καταγράφει ως ακολουθίες τα εντοπισμένα συμβάντα που μπορεί ή δεν μπορεί να έχουν μια ορισμένη έννοια του χρόνου.

Αλγόριθμοι δευτερογενούς εξόρυξης μοτίβων

Ένας αξιόπιστος αλγόριθμος εξόρυξης μοτίβων ακολουθιών θα πρέπει να παρέχει αποδεκτά μέτρα απόδοσης όπως χαμηλό χρόνο εκτέλεσης και χαμηλή χρήση της μνήμης όταν εξορύσσεται με χαμηλές ελάχιστες τιμές στήριξης και αυτές πρέπει να είναι κλιμακωτές.

Υπάρχουν τρεις κατηγορίες στις οποίες εμπίπτουν οι βασικές τεχνικές εξόρυξης προτύπων.

- Apriori-based (AprioriAll, το GSP, το PSP και το SPAM)
- Ανάπτυξη μοτίβων (FreeSpan, PrefixSpan, WAP-mine και FS-Miner)
- Κλαδέματος (LAPIN, HVSM και DISC-all)

3.2.7 Οπτικοποίηση εξόρυξης δεδομένων (Visual Data Mining)

Η οπτικοποίηση της εξόρυξης δεδομένων συνδυάζει τις παραδοσιακές μεθόδους εξόρυξης με τεχνικές απεικόνισης πληροφοριών. Ο χρήστης συμμετέχει άμεσα στη διαδικασία εξερεύνησης. Επωφελείται από τους αυτόματους υπολογισμούς και τις δυνατότητες της ανθρώπινης αντίληψης για την εξαγωγή δομών από εικόνες. Ένα οπτικό σύστημα εξόρυξης δεδομένων πρέπει να έχει απλότητα, αξιοπιστία, επαναχρησιμοποίηση, διαθεσιμότητα και ασφάλεια. Η απλότητα είναι η πιο σημαντική όταν δημιουργείτε οπτικά δεδομένα εξόρυξης. Η διεπαφή πρέπει να είναι εύκολη στη χρήση και τα δεδομένα που εμφανίζονται πρέπει να ερμηνεύονται εύκολα. Αυτά τα συστήματα θα πρέπει να βοηθήσουν στην καθοδήγηση του χρήστη μέσω της διαδικασίας ανακάλυψης της γνώσης. Οι προγραμματιστές πρέπει να κατασκευάσουν διασυνδέσεις με τρόπο που να επιτρέπει ακριβείς οπτικές παρουσιάσεις των δεδομένων όταν χρησιμοποιούνται από τον άνθρωπο για την ερμηνεία των δεδομένων.

Ένας αλγόριθμος εξόρυξης δεδομένων στη μηχανική μάθηση δημιουργεί ένα μοντέλο αναλύοντας δεδομένα εισόδου και εξάγοντας συγκεκριμένους τύπους, μοτίβα ή συσχετισμούς. Αυτή η παραγωγή αναλύεται μέσω πολλών επαναλήψεων με σκοπό να βρεθούν οι βέλτιστες παράμετροι για τη δημιουργία ενός μοντέλου. Αυτές οι παράμετροι εφαρμόζονται στη συνέχεια σε ολόκληρα σύνολα δεδομένων για την εξαγωγή ενεργών μοτίβων. Η επιλογή του κατάλληλου αλγόριθμου για μια συγκεκριμένη ανάλυση συχνά αποτελεί μια πρόκληση, καθώς διαφορετικοί αλγόριθμοι μπορούν να χρησιμοποιηθούν για την εκτέλεση της ίδιας εργασίας, ενώ κάθε αλγόριθμος παράγει διαφορετικά αποτελέσματα

και μερικοί αλγόριθμοι μπορούν παράγουν περισσότερα από ένα είδη αποτελεσμάτων. Ένας μεγάλος αριθμός αλγορίθμων είναι διαθέσιμοι στο πεδίο έρευνας για να ικανοποιήσουν τις ανάγκες των χρηστών. Μερικοί από τους κορυφαίους αλγόριθμους είναι :

1) Μηχανές Διανυσμάτων Υποστήριξης - ΜΔΥ (Support Vector Machines - SVM)

2) Δίκτυα Bayes (Bayesian Networks - BN)

3) Δένδρα απόφασης (Decision Tree - DT)

4) Αλγόριθμος C4.5

5) Αλγόριθμος K-Nearest Neighbors (KNN)

6) Αλγόριθμος k-means

7) Αλγόριθμος Apriori

8) Αλγόριθμος Expectation-Maximization (EM)

9) Αλγόριθμος PageRank

10) Αλγόριθμος AdaBoost

11) Αλγόριθμος Naive Bayes

12) Αλγόριθμος CART (Pushpam, Amali & Gnana, 2017).

3.3 ΤΕΧΝΙΚΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ ΑΠΟ ΤΑ ΚΟΙΝΩΝΙΚΑ ΜΕΣΑ

Βάση της σπουδαιότητας και της δημοφιλίας που έχουν λάβει τα τελευταία χρόνια τα κοινωνικά δίκτυα, είναι σημαντικό να αξιολογείται και να αποκωδικοποιείται η στάση, η άποψη και το συναίσθημα που εκφράζουν οι χρήστες των κοινωνικών δικτύων. Η σημασία των τεχνικών εξόρυξης δεδομένων από τα κοινωνικά δίκτυα, έγκειται στην μετατροπή των παραπάνω σε χρήσιμες πληροφορίες που αξιοποιούνται από το μάρκετινγκ και άλλες

ειδικότητες των επιχειρήσεων, της ιατρικής έρευνας και των βιομηχανιών. Οι τεχνικές αυτές, είναι ικανές να χειριστούν τα τρία κυρίαρχα ερευνητικά ζητήματα με παραμέτρους των δεδομένων των κοινωνικών δικτύων όπως α) το μέγεθος, β) τον θόρυβο και γ) τον δυναμισμό (Adedoyin-Olowe, Gaber & Stahl, 2014).

Παρακάτω, εξετάζονται αναλυτικά οι τεχνικές εξόρυξης δεδομένων που χρησιμοποιούνται επί του παρόντος για την ανάλυση δεδομένων από κοινωνικά δίκτυα.

3.3.1 Μη εποπτευόμενη ταξινόμηση

Ένας απλός, μη εποπτευόμενος αλγόριθμος μπορεί να χρησιμοποιηθεί για να αξιολογήσει μια κριτική ως «αποδεκτή» ή «μη αποδεκτή» μέσω της εξόρυξης φράσεων που περιλαμβάνουν επίθετα ή επιρρήματα (προσθήκη ετικετών). Ο σημασιολογικός προσανατολισμός κάθε φράσης μπορεί να προσεγγιστεί με την χρήση του αλγόριθμου PMI-IR (Pointwise Mutual Information and Information Retrieval) και στη συνέχεια η κριτική ταξινομείται βάση του μέσου σημασιολογικού προσανατολισμού της. Η συνάφεια του τίτλου, του σώματος και των σχολίων μιας ανάρτησης ιστολογίου έχει παρομοίως χρησιμοποιηθεί για την ομαδοποίηση σε σημαντικές ομάδες χρηστών με την βοήθεια λέξεων-κλειδιών. Σε αυτήν την περίπτωση οι λέξεις-κλειδιά έπαιξαν πολύ σημαντικό ρόλο. Οι αλγόριθμοι EM και LDA έχουν χρησιμοποιηθεί ευρέως για την ομαδοποίηση φράσεων ανά κατηγορίες (Aggarwal & Liu, 2008).

Έτσι, διακρίνεται το λεξικό συναισθηματικών λέξεων (που αντιστοιχεί σε μια λίστα με τις κοινές λέξεις που ενισχύουν τις τεχνικές εξόρυξης δεδομένων όταν χρησιμοποιούνται συναισθήματα στο έγγραφο), ο ορισμός και η σύνοψη της γνώμης (βασικές τεχνικές για την αναγνώριση της γνώμης, αναλύοντας τις πολικότητες του και τον βαθμό του συναισθήματος) και ο προσανατολισμός των συναισθημάτων και η εξαγωγή της γνώμης (πρότυπα αξιολόγησής για την ταξινόμηση της διάθεσης ή την απομόνωση άσχετων ή παραπλανητικών κριτικών) (Adedoyin-Olowe, Gaber & Stahl, 2014).

3.3.2 Ημι-εποπτευόμενη ταξινόμηση

Η ημι-εποπτευόμενη ταξινόμηση είναι μια στοχοθετημένη δραστηριότητα, που σε αντίθεση με την χωρίς επίβλεψη μπορεί να αξιολογηθεί συγκεκριμένα. Στην ημι-εποπτευόμενη μάθηση, εισήχθη η ανίχνευση της πολικότητας ως ημι-εποπτευόμενο

πρόβλημα ετικετών σε ένα γράφημα. Κάθε κόμβος αντιπροσωπεύει λέξεις των οποίων η πολικότητα πρόκειται να ανακαλυφθεί (Adedoyin-Olowe, Gaber & Stahl, 2014).

3.3.3 Εποπτευόμενη ταξινόμηση

Οι τεχνικές ταξινόμησης είναι εποπτευόμενες τεχνικές μάθησης που σε γενικές γραμμές χρησιμοποιούνται όταν η οργάνωση δεδομένων έχει ήδη εκτλεστεί. Ωστόσο, δεδομένου ότι τα κοινωνικά δίκτυα είναι δυναμικοί ιστότοποι, ο χρόνος μπορεί να μην είναι ουσιαστικός στην περίπτωση διεύρυνσης συμπεριφοράς/επιρροής μιας ομάδας, καθώς αυτά τα χαρακτηριστικά αλλάζουν από καιρό σε καιρό. Ένας αλγόριθμος εποπτευόμενης μάθησης επιστρατεύει τον συνδυασμό πολλαπλών βάσεων γεγονότων για την επισήμανση ορισμένων αναρτήσεων που έχουν παρόμοιους ή μη σημασιολογικούς προσανατολισμούς ((Adedoyin-Olowe, Gaber & Stahl, 2014).

Οι Injadat, Salo & Nassif (2016), ανέδειξαν μέσα από την μελέτη τους 19 τεχνικές/αλγόριθμους ταξινόμησης εξόρυξης δεδομένων που εφαρμόζονται ευρέως από ερευνητές στον τομέα των κοινωνικών μέσων. Ο κατάλογος αυτών των τεχνικών παρατίθεται παρακάτω.

- AdaBoost
- Τεχνητά νευρωνικά δίκτυα (ANN)
- Apriori
- Δίκτυα Bayesian (BN)
- Δέντρα απόφασης (DT)
- Αλγόριθμος βάση πυκνότητας (DBA)
- Αλγόριθμος Fuzzy
- Γενετικοί αλγόριθμοι (GA)
- Ιεραρχική ομαδοποίηση (HC)
- K-Means
- k-nearest Neighbors (k-NN)
- Γραμμική διακριτική ανάλυση (LDA)
- Γραμμική παλινδρόμηση (Lin-R)
- Λογιστική παλινδρόμηση (LR)
- Αλυσίδες Markov
- Μέγιστη εντροπία (ME)
- Novel
- Μηχανές διανυσμάτων υποστήριξης (SVM)

- Μέθοδοι τύπου Wrapper

Μεταξύ των παραπάνω αλγορίθμων, οι SVM, BN, και DT είναι οι πιο εφαρμοσμένες τεχνικές στην περιοχή της εξόρυξης γνώσης μέσα από τα κοινωνικά δίκτυα.

Οι Μηχανές διανυσμάτων υποστήριξης (SVM), τα δίκτυα Bayesian (BN) και τα δέντρα απόφασης (DT) είναι οι πιο ευρέως εφαρμοσμένες τεχνικές στον τομέα των κοινωνικών μέσων με ποσοστό 51%. Ο αριθμός των τεχνικών εξόρυξης δεδομένων που υιοθετήθηκαν από ερευνητές στον τομέα των κοινωνικών μέσων αυξήθηκαν δραματικά το 2012 και το 2014 σε 39 και 35 τεχνικές αντίστοιχα. Ο αριθμός μειώθηκε ελαφρά σε 24 τεχνικές το 2013. Επιπλέον, αξίζει να αναφέρουμε ότι πολλές νέες τεχνικές έχουν προκύψει από το 2012 έως τις αρχές του 2015 με συνολικό αριθμό 12 νέων τεχνικών (Injadat, Salo & Nassif, 2016).

Τεχνική ταξινόμησης/ αλγόριθμος μηχανών διανυσμάτων υποστήριξης (SVM)

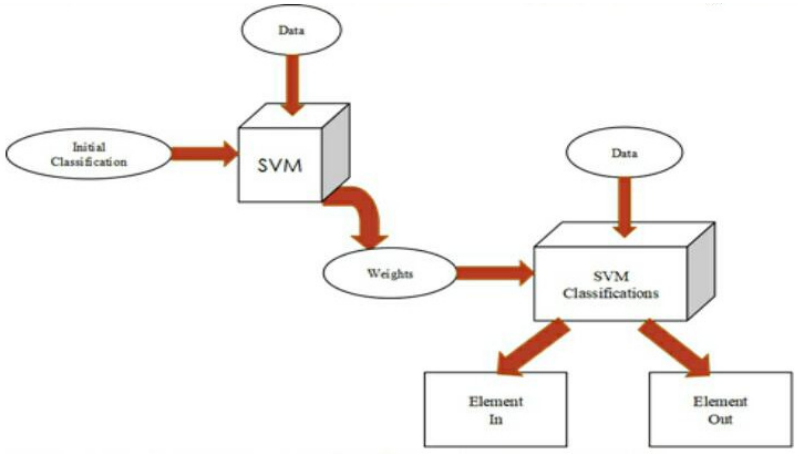
Η τεχνική SVM ταξινομεί τα δεδομένα σε 2 κατηγορίες χρησιμοποιώντας την έννοια του υπερεπιπέδου (hyperplane) N-διαστάσεων βέλτιστης διαίρεσης των δεδομένων σε δύο κατηγορίες. Για μια απλή εργασία ταξινόμησης με μόλις 2 δυνατότητες, το υπερεπίπεδο μπορεί να είναι μια γραμμή,

$$y = m\,x + \beta.$$

Ένα απλό παράδειγμα επεξήγησης της τεχνικής είναι η χρωματική ταξινόμηση μερικών ανακατεμένων μπάλων κόκκινου και μπλε χρώματος σε ένα τραπέζι. Σε αυτήν την περίπτωση μπορούν να χωριστούν με το ραβδί. Όταν προστίθεται μια νέα μπάλα στο τραπέζι, γνωρίζοντας ποια πλευρά του ραβδιού είναι η μπάλα, μπορεί να προβλεφθεί το χρώμα της. Στο παραπάνω παράδειγμα, οι μπάλες αντιπροσωπεύουν σημεία δεδομένων και τα κόκκινα και μπλε χρώματα αντιπροσωπεύουν τις 2 κλάσεις. Το ραβδί αντιπροσωπεύει το απλούστερο υπερεπίπεδο που είναι μια γραμμή. Πρόκειται για έναν τύπο εποπτευόμενης μάθησης που συνιστά μια σημαντική εξέλιξη των αλγορίθμων μηχανικής μάθησης για την ταξινόμηση δεδομένων (Yadanar, Htun & Soe, 2018 ; Pushpam & Jayanthi, 2017)

Ο αλγόριθμος που υιοθετήθηκε στην τεχνική SVM βασίζεται στη θεωρία της στατιστικής μάθησης και στη διάσταση Vapnik-Chervonenkis [VC] που εισήγαγαν οι

Vladimir Vapnik και Alexey Chervonenkis. Ακόμη και χωρίς επιλογή χαρακτηριστικών, η απόδοση του SVM μπορεί να είναι πολύ αποτελεσματική. Η αρχιτεκτονική του SVM φαίνεται στο παρακάτω σχήμα (Yadanar, Htun & Soe, 2018).



Σχήμα 1

Αρχιτεκτονική της SVM (Yadanar, Htun & Soe, 2018)

Σύμφωνα με την βιβλιογραφία, έχουν αναφερθεί πέντε διαδοχικά βήματα για την εξόρυξη δεδομένων από τα κοινωνικά δίκτυα μέσω της SVM :

Βήμα 1: συλλογή των γραπτών κειμένων (πρόκειται για τη συλλογή ενός αριθμού σχολίων των χρηστών ενός κοινωνικού δικτύου όπως π.χ το Facebook).

Βήμα 2: ανάσυρση των σχολίων σε πραγματικό χρόνο από το επιλεγμένο κοινωνικό δίκτυο μέσω του εργαλείου άντλησης που συνδέεται με τον διακομιστή.

Βήμα 3: Ταξινόμηση: Το εργαλείο ταξινόμησης ταξινομεί τα συλλεχθέντα σχόλια σε υποκατηγορίες (π.χ θετικά, αρνητικά και ουδέτερα).

Βήμα 4: Επεξεργασία μέσω SVM (οι αλγόριθμοι δημιουργούν αρχεία κειμένου και κατόπιν επεξεργασίας, παρέχουν το ποσοστό ακρίβειας για τον έλεγχο της ταξινόμησης που στη συνέχεια υπόκειται σε περεταίρω προβλέψεις και ανάλυση).

Βήμα 5: Ανάλυση των αποτελεσμάτων που παρατίθενται σε μορφή γραφήματος και εξάγονται συμπεράσματα για την απόδοση της ταξινόμησης (Yadanar, Htun & Soe, 2018).

Τεχνική ταξινόμησης/ αλγόριθμος δικτύων Bayesian (Belief)

Τα δίκτυα Bayesian (που επίσης αποκαλούνται Belief) συνιστούν μια πολύ σημαντική τεχνική εξόρυξης δεδομένων που αποτελείται από δύο μέρη, το κατευθυνόμενο ακυκλικό γράφημα G, με κόμβους (χαρακτηριστικά) και τόξα (άμεσες εξαρτήσεις) και τους πίνακες πιθανότητας υπό όρους για κάθε κόμβο. Ο ταξινομητής Bayes επιτυγχάνει το βέλτιστο αποτέλεσμα εφαρμόζοντας τη θεωρία πιθανοτήτων. Τα δίκτυα Bayesian αντιπροσωπεύουν γεγονότα και αιτιώδεις σχέσεις μεταξύ τους ως πιθανότητες υπό όρους που περιλαμβάνουν τυχαίες μεταβλητές. Δεδομένων των τιμών ενός υποσυνόλου αυτών των μεταβλητών (μεταβλητές αποδεικτικών στοιχείων), υπολογίζονται οι πιθανότητες ενός άλλου υποσυνόλου μεταβλητών (μεταβλητές ερωτήματος). Ωστόσο, οι προσεγγίσεις Bayesian δεν μπορούν να ξεπεράσουν την ανάγκη εκτίμησης πιθανότητας από το σύνολο δεδομένων εκπαίδευσης. Είναι αξιοσημείωτο ότι σε ορισμένες περιπτώσεις, όπου η απόφαση βασίζεται σαφώς σε ορισμένα κριτήρια ή το σύνολο δεδομένων έχει υψηλό βαθμό τυχαιότητας, οι προσεγγίσεις Bayesian δεν συνιστούν μια αποτελεσματική τεχνική (Pushpam & Jayanthi, 2017).

Τα δίκτυα Bayesian χρησιμοποιούνται ως μέθοδος βελτίωσης της ανάλυσης κοινωνικών δικτύων και να αποδείξουν την αποτελεσματικότητα αυτής της προσέγγισης σε ένα εφαρμοσμένο σύστημα. Τα δίκτυα Bayesian επιτρέπουν στον χρήστη να εκτελεί εργασίες εξόρυξης δεδομένων που περιλαμβάνουν ελλιπή δεδομένα, κόμβους με αβέβαια χαρακτηριστικά και ασαφείς σχέσεις. Επιτρέπουν επίσης στο χρήστη να συνάγει νέες σχέσεις μεταξύ κόμβων που δεν αποκαλύπτονται στα αρχικά δεδομένα και να εντοπίσει κόμβους στο δίκτυο που παρουσιάζουν ιδιαίτερο ενδιαφέρον λόγω των χαρακτηριστικών και των σχέσεων τους. Αυτές οι δυνατότητες κάνουν τα δίκτυα Bayesian ένα ισχυρό εργαλείο στη διεξαγωγή ανάλυσης δεδομένων μέσω των κοινωνικών δικτύων (Koelle et al., 2006).

Τεχνική ταξινόμησης/ αλγόριθμος δένδρων αποφάσεων (DT)

Το δέντρο αποφάσεων συνιστά μια πολύ χρήσιμη τεχνική ταξινόμησης και παλινδρόμησης των δεδομένων. Χαρακτηρίζονται από ευελιξία και γίνονται εύκολα κατανοητά. Έτσι, η πρόβλεψη μιας κατηγορηματικής τιμής (π.χ κόκκινο, πράσινο, πάνω, κάτω) ή μιας συνεχούς τιμής (π.χ 2.9, 3.4 κ.λπ.) η τα δένδρα αποφάσεων δύνανται να διαχειριστούν και τα δύο προβλήματα. Χρειάζονται μόνο ένας πίνακας εκ του οποίου δημιουργείται απευθείας ένας ταξινομητής χωρίς την εισαγωγή κάποιας εργασίας σχεδιασμού εκ των προτέρων (Desai & Patil, 2015).

Ο ταξινομητής δέντρων απόφασης διαιρεί το σύνολο δεδομένων σε μικρότερο υποσύνολο, βάση διαφόρων κριτηρίων. Το δέντρο αποφάσεων είναι ένας ταξινομητής που κατασκευάζει δομή δέντρου με κόμβους και τόξα. Οι ριζικοί και οι εσωτερικοί κόμβοι επισημαίνονται με ερώτηση. Το βέλος αντιπροσωπεύει την απάντηση στη σχετική ερώτηση. Κάθε τερματικός κόμβος δείχνει μια πρόβλεψη για μια λύση στο πρόβλημα / την τιμή της μεταβλητής-στόχου. Το δέντρο αποφάσεων προβλέπει πληροφορίες με τη μορφή κανόνων που είναι εκφράσεις υπό συνθήκη if-then-else. Αυτό το αποτέλεσμα εξηγεί τις αποφάσεις που οδηγούν στην πρόβλεψη (Neelamegam, 2013).

Μερικά από τα χαρακτηριστικά των δένδρων αποφάσεων είναι ο κόμβος απόφασης ο οποίος καθορίζει μια δοκιμή σε ένα μοναδικό χαρακτηριστικό, ο τερματικός κόμβος που υποδεικνύει την τιμή του χαρακτηριστικού στόχου και τα Arc/edge που αντιστοιχούν στον διαχωρισμό ενός χαρακτηριστικού. Τα δέντρα απόφασης ταξινομούν τις περιπτώσεις ή τα παραδείγματα ξεκινώντας από τη ρίζα του δέντρου και αναπτύσσοντας το μέχρι τον τερματικό κόμβο. Προφανώς, τα μεγαλύτερα δέντρα είναι συνήθως λιγότερο ακριβή από τα μικρότερα δέντρα (Desai & Patil, 2015).

Η Εντροπία αποτελεί ένα μέτρο ομοιογένειας του συνόλου των περιπτώσεων. Όταν η εντροπία είναι 0, το αποτέλεσμα θεωρείται «σίγουρο» και αντιθέτως, η εντροπία είναι μέγιστη όταν υπάρχει αβεβαιότητα του συστήματος (ή οποιοδήποτε αποτέλεσμα είναι εξίσου δυνατό).

Ο αλγόριθμος δέντρων αποφάσεων έχει από πολλούς ερευνητές θεωρηθεί περισσότερο κατάλληλος για εξόρυξη δεδομένων μέσα από κοινωνικά δίκτυα, καθώς είναι ευέλικτος και αποδίδει ακριβέστερα αποτελέσματα σε σύγκριση με άλλες τεχνικές (Desai & Patil, 2015). Κατά την εξόρυξη δεδομένων μέσα από κοινωνικά δίκτυα, τα ακολουθούμενα βήματα περιλαμβάνουν (Deepa & JeenMarseline, 2019) :

1. Εκ των προτέρων δημιουργία του ριζικού κόμβου

2. Υπολογισμός της εντροπίας με την τρέχουσα κατάσταση H(S)

3. Υπολογισμός της εντροπίας για κάθε χαρακτηριστικό σε σχέση με το χαρακτηριστικό x H (S, x) Η μέγιστη τιμή του χαρακτηριστικού επιλέγεται σε σχέση με το IG(S, x)

5. Αφαιρούνται τα χαρακτηριστικά από το σύνολο χαρακτηριστικών που προσφέρει το υψηλότερο IG

6. Επανάληψη της διαδικασίας έως ότου το δέντρο απόφασης να έχει όλους τους τερματικούς κόμβους

Τεχνική ταξινόμησης/ αλγόριθμος Naïve Bayes

Ο αλγόριθμος Naïve Bayes χρησιμοποιεί δεσμευμένες πιθανότητες (conditional probabilities) μετρώντας την εμφάνιση τιμών και συνδυασμούς τιμών στα ιστορικά δεδομένα (πιθανολογική μέθοδος) Το Naïve Bayes είναι επίσης μια αποτελεσματική τεχνική πρόβλεψης καιρού εξόρυξης. Ο αλγόριθμος Naïve Bayes είναι μια από τις τρεις τεχνικές εποπτευόμενης μάθησης για την ανάλυση συναισθημάτων που σε συνδυασμό με την δυαδική λέξη-κλειδί μπορεί να παράγει έναν μονοδιάστατο βαθμό συναισθήματος εδραιωμένου σε tweets από το twitter (Adedoyin-Olowe, Gaber & Stahl, 2014).

Άλλα εξίσου αποτελεσματικά και αποδεκτά από την επιστημονική κοινότητα εργαλεία εξόρυξης γνώσεων είναι τα παρακάτω :

- Salford Systems Tools (CART, Random Forest, MARS, TreeNet)
- SAS Enterprise Miner/Text Miner
- SPSS Clementine
- Megaputer Intelligence PolyAnalyst (Brusilovsky, 2009).

3.3.4 Εξόρυξη κειμένου

Η βαθμολογία διαστάσεων αποτελεί μια αριθμητική αξιολόγηση σε σχέση με το επίπεδο ικανοποίησης που απεικονίζεται στα συγκεντρωμένα σχόλια. Με την χρήση φράσεων και τροποποιητών τους (π.χ καλό προϊόν, εξαιρετική τιμή), κάθε διάσταση εξάγεται και ταξινομείται χρησιμοποιώντας πιθανολογική λανθάνουσα σημασιολογική ανάλυση (pLSA) (Adedoyin-Olowe, Gaber & Stahl, 2014). Στην εξόρυξη κειμένου επιστρατεύονται μη δομημένες αναρτήσεις που εξετάζονται με στόχο ανάδειξης δομής και νοημάτων που βρίσκονται κρυμμένα μες στο κείμενο. Η εξόρυξη κειμένου χρησιμοποιεί ειδικές Natural Language Processing (NLP) τεχνικές υπό αυστηρούς περιορισμούς με στόχο την αποφυγή περίπλοκων διαδικασιών και διαφέρει ως προς την εξόρυξη δεδομένων, καθώς αναδεικνύει πληροφορίες που βρίσκονται αποθηκευμένες ως μια μη δομημένη συλλογή εγγράφων κειμένου (Παπαστεργίου, 2005).

3.4 ΕΡΓΑΛΕΙΑ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ ΑΠΟ ΚΟΙΝΩΝΙΚΑ ΔΙΚΤΥΑ

Τα εργαλεία RapidMiner, Weka, R και KNIME είναι εργαλεία ανάλυσης δεδομένων τα οποία χρησιμοποιούνται ευρέως από πολλούς οργανισμούς για την εξόρυξη και ανάλυση των δεδομένων που προέρχονται από διάφορες πηγές, συμπεριλαμβανομένων των κοινωνικών δικτύων (Dwivedi, Kasliwal & Soni, 2016).

3.4.1 RapidMiner

Το RapidMiner είναι ένα εργαλείο για τη διεξαγωγή εξόρυξης δεδομένων με διαφορετικές περιοχές εφαρμογών και σχημάτων βελτιστοποίησης παραμέτρων. Ένα από τα κύρια χαρακτηριστικά του RapidMiner είναι η προηγμένη ικανότητά του να προγραμματίζει την εκτέλεση σύνθετων ροών εργασίας, όλα μέσω μιας οπτικοποιημένης διεπαφής που εξασφαλίζει ευχρησία για τον χρήστη, χωρίς να δημιουργεί την ανάγκη απόκτησης εξελιγμένων δεξιοτήτων προγραμματισμού (Jovanovic et al., 2014). Είναι ένα λογισμικό πολλαπλών πλατφορμών που υποστηρίζει περίπου είκοσι δύο μορφές αρχείων. Περιέχει περισσότερα από 100 προγράμματα μάθησης για την ταξινόμηση παλινδρόμησης και την ανάλυση ομαδοποίησης (Dwivedi, Kasliwal & Soni, 2016).

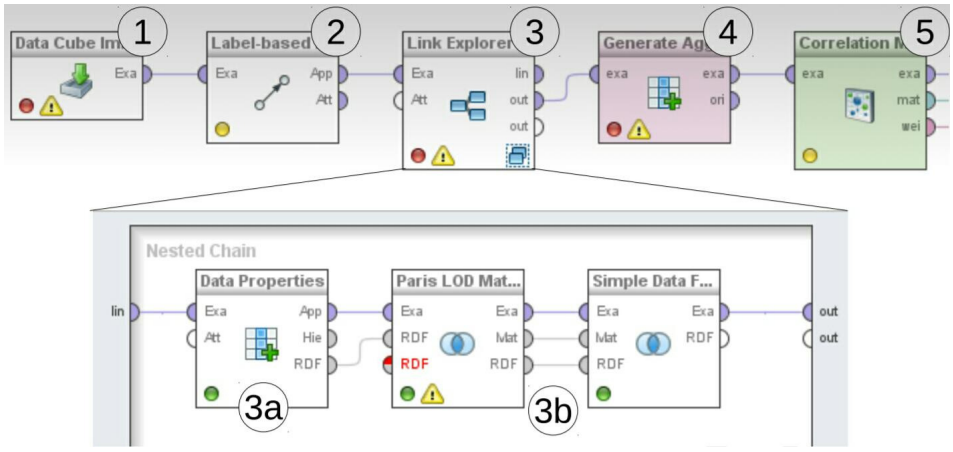
Το RapidMiner είναι μια πλατφόρμα εξόρυξης δεδομένων που ως προϊόν λογισμικού ανοιχτού κώδικα, αναπτύχθηκε το 2001 από τους Ralf Klinkenberg, Ingo Mierswa και Simon Fischer στη Μονάδα Τεχνητής Νοημοσύνης του Τεχνικού Πανεπιστημίου του Ντόρτμουντ και αξιοποιήθηκε από μια εταιρεία περιορισμένης ευθύνης που ονομάζεται Rapid-I, και εδρεύει στο Ντόρτμουντ της Γερμανίας (North, 2012). Οι προηγούμενες εκδόσεις του RapidMiner έχουν διανεμηθεί με την άδεια ανοιχτού κώδικα AGPL. Από τις 15 Μαΐου 2015, είναι διαθέσιμη η έκδοση Open Source, ενώ το RapidMiner 6 κυκλοφόρησε το 2013. Στην ετήσια δημοσκόπηση του 2014 και 2013, το RapidMiner ανακηρύχθηκε το πιο δημοφιλές λογισμικό ανάλυσης δεδομένων και έλαβε μία από τις ισχυρότερες βαθμολογίες ικανοποίησης των χρηστών (Dwivedi, Kasliwal & Soni, 2016).

Το RapidMiner είναι μια πλατφόρμα λογισμικού που παρέχει ένα ολοκληρωμένο περιβάλλον για μηχανική εκμάθηση, εξόρυξη κειμένων, εξόρυξη δεδομένων, επιχειρηματικά αναλυτικά και προγνωστικά αναλυτικά στοιχεία. Χρησιμοποιείται στην εκπαίδευση, την έρευνα και κατάρτιση, την ανάπτυξη εφαρμογών, στην γρήγορη και οικονομική πρωτοτυποποίηση, για βιομηχανικές και επαγγελματικές εφαρμογές. Το RapidMiner μπορεί να υποστηρίξει όλα τα βήματα της διαδικασίας εξόρυξης δεδομένων, τα οποία μπορεί να περιλαμβάνουν

οπτικοποίηση αποτελεσμάτων, βελτιστοποίηση και επικύρωση (Dwivedi, Kasliwal & Soni, 2016).

Στην εν λόγω πλατφόρμα, η εξόρυξη και η ανάλυση δεδομένων σχεδιάζονται από χρήσιμα εργαλεία: τα αποθετήρια (Repositories) και τους χειριστές (Operators). Η περιοχή αποθετηρίων είναι το εκείνη, στην οποία η χρήστης διασυνδέεται σε κάθε σύνολο δεδομένων. Στην περιοχή των χειριστών βρίσκονται όλα τα εργαλεία εξόρυξης δεδομένων, τα οποία δεν διαφέρουν από τα κοινά μοντέλα και εργαλεία διαχείρισης δεδομένων που έχουν ήδη προαναφερθεί (North, 2012).

Οι χειριστές, ως στοιχειώδη δομικά στοιχεία, εκτελούν μια συγκεκριμένη ενέργεια που αφορά τα δεδομένα (π.χ. φόρτωση και αποθήκευση, μετατροπή ή εξαγωγή συμπερασμάτων). Ο χρήστης μπορεί να συνθέσει μια διαδικασία από τους χειριστές με την τοποθέτησή τους σε έναν καμβά και την καλωδίωση των θυρών εισόδου και εξόδου τους. Η επέκταση Open Data Linked RapidMiner προσθέτει ένα σύνολο τελεστών στο RapidMiner, οι οποίες μπορούν να χρησιμοποιηθούν σε διαδικασίες εξόρυξης δεδομένων και σε συνδυασμό με ενσωματωμένους χειριστές RapidMiner, καθώς και με άλλους χειριστές. Οι τελεστές στην επέκταση εμπίπτουν σε διαφορετικές κατηγορίες: εισαγωγή δεδομένων, σύνδεση δεδομένων, δημιουργία χαρακτηριστικών, αντιστοίχιση σχήματος και επιλογή υποσυνόλου χαρακτηριστικών (Ristoski, Bizer & Paulheim, 2015).



Εικόνα [5]

[5] Πηγή https://rapidminer.com/products/cloud/

Επισκόπηση της διαδικασίας που χρησιμοποιείται στο τρέχον παράδειγμα, συμπεριλαμβανομένης της ένθετης δευτερεύουσας διαδικασίας στον τελεστή εξερεύνησης συνδέσμων

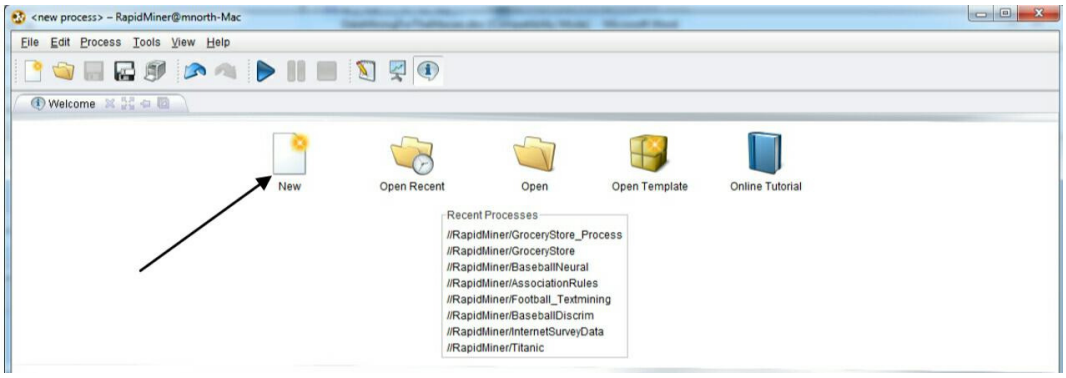
Το RapidMiner αποτελεί ένα ιδανικό συμπλήρωμα του OpenOffice που (North, 2012) :

- παρέχει συγκεκριμένες λειτουργίες εξόρυξης δεδομένων (όπως δέντρα αποφάσεων και κανόνες συσχέτισης)
- Το RapidMiner είναι εύκολο στην εγκατάσταση και λειτουργεί σε σχεδόν οποιονδήποτε υπολογιστή.
- Η έκδοση του εν λόγω λογισμικού για το κοινό, παρέχεται για δωρεάν εγκατάσταση και χρήση

Κατά την εισαγωγή δεδομένων, μέσω του RapidMiner παρέχονται τελεστές εισαγωγής για διαφορετικές μορφές δεδομένων (π.χ. Excel, CSV, XML). Η επέκταση Open Data παρέχει δύο επιπρόσθετους τελεστές εισαγωγής (SPARQL και Data Cube Importer). Μέσω του Linked Data Cube Explorer (LDCX) επιτρέπεται στο χρήστη η δημιουργία και χρήση ενός συγκεντρωτικού πίνακα με τα επιλεγμένα δεδομένα (Ristoski, Bizer & Paulheim, 2015).

Αναφορικά με την σύνδεση δεδομένων, στο συγκεκριμένο πρόγραμμα, αρκετές φορές προ-απαιτείται η σύνδεση τοπικών δεδομένων που δεν είναι RDF (π.χ. δεδομένα σε ένα αρχείο CSV ή μια βάση δεδομένων) με άλλα δεδομένα από το cloud. Για τον λόγο αυτό, εφαρμόζονται διαφορετικοί τελεστές (Pattern-based linker, Label-based linker, Lookup linker, SameAs linker) σύνδεσης των δεδομένων (Ristoski, Bizer & Paulheim, 2015).

Το συγκεκριμένο εργαλείο εξόρυξης δεδομένων, κατά την διάρκεια της προ-επεξεργασίας διαθέτει την δυνατότητα διαχείρισης των ελλιπών τιμών (Missing data), υπό την έννοια των δεδομένων που λείπουν ή που δεν υπάρχουν σε ένα σύνολο δεδομένων, ωστόσο η τιμή τους δεν είναι ίδια με το μηδέν ή κάποια άλλη τιμή. Αντίθετα, κατά την επεξεργασία των δεδομένων, αυτά είναι κενά και η τιμή τους είναι άγνωστη. Ανάλογα με τον στόχο της εξόρυξης δεδομένων, τα δεδομένα αυτά μπορούν να παραμείνουν ως έχουν ή να αντικατασταθούν με κάποια άλλη τιμή (North, 2012).



Εικόνα[6]Η αρχική οθόνη του RapidMiner

Μετά τον καθοδηγούμενο χαρακτηρισμό του τύπου των δεδομένων, το σύνολο δεδομένων είναι διαθέσιμο για χρήση στο RapidMiner. Για την έναρξη της εξόρυξης, ο φάκελος που έχει δημιουργηθεί οδηγείται στο πλαίσιο των βασικών αναλύσεων μέσω της διαδικασίας drag and drop.

Η εξόρυξη δεδομένων μπορεί να είναι συγκεχυμένη καικουραστική, ειδικά όταν τα σύνολα δεδομένων είναι εξαιρετικά και αχρείαστα μεγάλα. Η κατάλληλη διαχείρισή τους (μέσω φίλτρων) μπορεί να μειώσει τις παρατηρήσεις που περιέχουν ανεπιθύμητα δεδομένα (ή ελλείποντα δεδομένα) βάση ενός χαρακτηριστικού, όπως επίσης τα δεδομένα μπορούν να περικοπούν και να αναλυθούν σε ένα μικρότερο υποσύνολο. Αυτό μπορεί να μειώσει σημαντικά τον χρόνο επεξεργασίας κατά τη δοκιμή ενός μοντέλου με βάση αν απαντάει στις αρχικές ερωτήσεις βάση των οποίων διενεργείται η εξόρυξη.

Στο RapidMiner διατίθεται επιπλέον η δυνατότητα διαχείρισης των ασυνεπών (ασύνδετων) δεδομένων. Τα ασυνεπή δεδομένα διαφέρουν από τα ελλείποντα. Αυτά, παρουσιάζονται όταν μια τιμή υπάρχει, ωστόσο δεν είναι έγκυρη ή σημαντική. Αντίστοιχα, σε αριθμητικά δεδομένα ενδέχεται επίσης να προκύψουν ανακριβή δεδομένα, αλλά και στατιστικά ακραία που μπορούν επίσης να θεωρηθούν ασυνεπή και να απαιτούν διαχείριση. Αν και ο καθαρισμός των δεδομένων αυτών μπορεί να είναι μια επίπονη και κουραστική διαδικασία, δύναται να επηρεάσει σφόδρα τη χρησιμότητα των αποτελεσμάτων εξόρυξης δεδομένων, έτσι αυτή η λεγόμενη «μείωση των δεδομένων» είναι σημαντικό να προηγείται και να εκτελείται με προσοχή και σωστή κρίση ως προς τη λεπτομέρεια.

Η εξόρυξη δεδομένων με τις αλληλοσυνδεόμενες πλευρές της ταξινόμησης και της πρόβλεψης που διαθέτει, έχει απόλυτη εξάρτηση από την συσχέτιση. Αντίθετα από τις

[6] Πηγή North, 2012

στατιστικές δοκιμασίες, στην εξόρυξη δεδομένων δεν συνάγεται η αιτιότητα χρησιμοποιώντας μετρήσεις συσχέτισης, ούτε χρησιμοποιούνται συντελεστές συσχέτισης για την πρόβλεψη της τιμής ενός χαρακτηριστικού βάσει ενός άλλου. Ωστόσο γρήγορα μπορούν να αναδειχθούν μέσω της συσχέτισης γενικές τάσεις σε σύνολα δεδομένων και μπορεί να προβλεφθεί η ένταση της παρατήρησης μεταβολής ενός χαρακτηριστικού σε συνδυασμό με ένα άλλο. Η συσχέτιση μπορεί να είναι ένας γρήγορος και εύκολος τρόπος παρατήρησης της αλληλεπίδρασης διαφόρων στοιχείων ενός δεδομένου προβλήματος.

3.4.2 R

Η εξόρυξη και ανάλυση δεδομένων επίσης παρέχεται μέσω του εργαλείου R που παρέχεται ως δωρεάν λογισμικό για περιβάλλον στατιστικών υπολογιστών και γραφικών. Το συγκεκριμένο εργαλείο, αναπτύχθηκε από την R Core Team συνιστώντας ένα λογισμικό πολλαπλών πλατφορμών, γραμμένο σε γλώσσα προγραμματισμού S. Το εργαλείο R διατίθεται ως λογισμικό ανοιχτού κώδικα, που υποστηρίζει μια ποικίλλουσα γραμμή εντολών και ειδικό στατιστικό πακέτο. Υπάρχουν εκατοντάδες επιπλέον ελεύθερα διαθέσιμα πακέτα, τα οποία παρέχουν κάθε είδους εξόρυξη δεδομένων, μηχανική μάθηση και στατιστικές τεχνικές. Το εργαλείο R επιτρέπει στους στατιστικούς να εκτελούν περίπλοκες εργασίες και αναλύσεις χωρίς να γνωρίζουν προγραμματισμό. Ένα από τα πλεονεκτήματα του R είναι η ευκολία με την οποία μπορούν να παραχθούν καλοσχεδιασμένα και ποιοτικά μοντέλα, συμπεριλαμβανομένων μαθηματικών συμβόλων και τύπων όπου απαιτείται (Seefeld & Linder, 2007).

3.4.3 KNIME

Το KNIME είναι μια πλατφόρμα ανάλυσης δεδομένων, αναφορών και ενοποίησης που ενσωματώνει διάφορα στοιχεία για μηχανική μάθηση και εξόρυξη δεδομένων. Αναπτύχθηκε από την KNIME.com AG, το 2004 από μια ομάδα προγραμματιστών στο Πανεπιστήμιο της Konstanz, με δυνατότητα λειτουργίας σε πολλαπλά λειτουργικά συστήματα (Linux, OS X, Windows) (Berthold et al., 2009). Αρχικός στόχος της ανάπτυξης του συγκεκριμένου εργαλείου ήταν η δημιουργία μιας αρθρωτής, επεκτάσιμης και ανοιχτής πλατφόρμας επεξεργασίας δεδομένων που επιτρέπει την εύκολη ενσωμάτωση διαφορετικών μονάδων

φόρτωσης, επεξεργασίας, μετασχηματισμού, ανάλυσης και οπτικής εξερεύνησης των δεδομένων χωρίς να εστιάζεται σε συγκεκριμένη περιοχή εφαρμογών. Η βασική έκδοση του Knime ενσωματώνει περισσότερους από 100 κόμβους προ-επεξεργασίας, επεξεργασίας και καθορισμού, μοντελοποίησης, ανάλυσης και εξόρυξης δεδομένων καθώς και διάφορες διαδραστικές προβολές, όπως γραφικές αναπαραστάσεις των δεδομένων τύπου (scatter plot), παράλληλες συντεταγμένες και άλλα (Kataria, 2013).

Σύνοψη και αξιολόγηση εργαλείων εξόρυξης δεδομένων από κοινωνικά δίκτυα

Όλα τα προαναφερόμενα εργαλεία, παρουσιάζουν ορισμένα κοινά χαρακτηριστικά όπως ο τύπος ανοιχτού κώδικα και η γλώσσα java. Όλα είναι επεκτάσιμα και υποστηρίζουν τόσο δομημένα όσο και μη δομημένα δεδομένα. Έχουν ισχυρή οπτικοποίηση και είναι εύκολα ως προς την εκμάθηση και την χρήση. Ειδικότερα, τα RapidMiner και KNIME χρησιμοποιούν το περιβάλλον εργασίας χρήστη για το σχεδιασμό αναλυτικών διεργασιών (Kalpana & Bansal, 2014). Παρόλο που υπάρχουν πολλές διαφορετικές προσεγγίσεις στα αναλυτικά στοιχεία, όλες μοιράζονται ένα αναγνωρίσιμο παρόμοιο σύνολο βημάτων.

Τα βήματα αυτά είναι (Witten, 2005) :

- Συλλογή & Απόκτηση για την εξαγωγή δεδομένων προέλευσης
- Αποθήκευση δεδομένων προέλευσης σε αποθήκη δεδομένων, βάση συσχέτισης
- Καθαρισμός δεδομένων (διόρθωση ανωμαλιών και ασυνεπειών και ομαλοποίηση της σύνταξης των δεδομένων)
- Ενσωμάτωση που αφορά την ευθυγράμμιση των δεδομένων είτε σε υπάρχοντα σύνολα δεδομένων είτε σε ένα κοινό λεξιλόγιο
- Ανάλυση δεδομένων, προκειμένου να δημιουργηθούν περιγραφικά ή προγνωστικά μοντέλα
- Αναπαράσταση και οπτικοποίηση για την δημιουργία αναφορών και διαγραμμάτων που απεικονίζουν τα μοντέλα που μπορούν να χρησιμοποιηθούν από ένα ευρύτερο κοινό.

Εργαλείο	**Γλώσσα προγρ/μού**	**Λειτουργία**	**Πλεονεκτήματα**	**Περιορισμοί χρήσης**
Weka	Java	Γενική εξόρυξη δεδομένων, προ-επεξεργασία ταξινόμηση, ομαδοποίηση	Εύχρηστο, υποστηρίζει διαφορετικούς τύπους πακέτων δεδομένων (ARFF, CSV, C4.5, binary)	Φτωχή αναπαράσταση των αποτελεσμάτων, ακατάλληλο για μεγάλα πακέτα δεδομένων
RapidMiner	Java	Γενική εξόρυξη δεδομένων, προεπεξεργασία, οπτικοποίηση, προβλεπτικότητα	δεν απαιτείται κωδικοποίηση, πλούσια εργαλειοθήκη, πλήρες πακέτο	Μειωμένη ικανότητα καταμερισμού, προαπαιτούμενη γνώση διαχείρισης βάσεων δεδομένων
Orange C++,	Python, Qt framework	Γενική εξόρυξη δεδομένων, προ-επεξεργασία, ταξινόμηση, μοντελοποίηση, οπισθοδρόμηση, ομαδοποίηση	Εύκολο στην εκμάθηση, δυναμικό περιβάλλον, υποστήριξη οπτικού προγραμματισμού, γλώσσα σεναρίων Python	Περιορισμένες αναφερόμενες δυνατότητες, αδυναμία στην κλασσική στατιστική
R	C, Fortran, R	Εξόρυξη δεδομένων και στατιστική	Εύχρηστο περιβάλλον και προηγμένο στατιστικό και αναλυτικό λογισμικό	Αδύναμη διαχείριση μνήμης και χαμηλή ταχύτητα
KNIME	Java	Εξόρυξη και ανάλυση δεδομένων και κειμένου	Εύκολη επέκταση και προσθήκη συστατικών πρόσθετων δυνατοτήτων	Περιορισμένες μετρήσεις σφαλμάτων, κακή βελτιστοποίηση παραμέτρων

			(plug-in), ισχυρό εργαλείο με GUI	

Πίνακας 1 *Γενική σύγκριση χαρακτηριστικών των εργαλείων εξόρυξης δεδομένων* (μετασχ. από Pushpam & Jayanthi, 2017)

ΚΕΦΑΛΑΙΟ 4. ΠΕΡΙΓΡΑΦΗ ΤΟΥ ΠΡΟΓΡΑΜΜΑΤΟΣ WEKA

4.1 ΠΕΡΙΓΡΑΦΗ ΤΟΥ ΛΟΓΙΣΜΙΚΟΥ

Πέρα από τους προαναφερόμενους αλγόριθμους, στην πραγματικότητα κυκλοφορούν και προγράμματα ευρείας και κοινής χρήσης που επιτρέπουν σε οποιονδήποτε το επιθυμεί, με ελάχιστους πόρους και εξειδίκευση να εξαγάγει συμπεράσματα μέσω ανάλυσης δεδομένων (π.χ που αφορούν τον πληθυσμό σε τοπικό επίπεδο ή σε κοινωνικο-οικονομικό επίπεδο) για την ανακάλυψη γνώσης χρησιμοποιώντας ένα ισχυρό εργαλείο εξόρυξης δεδομένων, που ονομάζεται Weka (Jagtap, 2013).

Το Weka (Waikato Environment for Knowledge Analysis) είναι ένα δημοφιλές λογισμικό μηχανικής εκμάθησης γραμμένο σε γλώσσα προγραμματισμού Java, που αναπτύχθηκε στο Πανεπιστήμιο του Waikato στη Νέα Ζηλανδία. Το Weka είναι ένα λογισμικό ελεύθερης χρήσης που διατίθεται υπό Γενική Άδεια Δημόσιας Χρήσης (General Public License -GNU). Ο πίνακας εργασίας του προγράμματος Weka περιέχει μια συλλογή εργαλείων και αλγορίθμων κατάλληλων για ανάλυση δεδομένων και μια πρότυπη μοντελοποίηση, μαζί με γραφικά διεπαφών που διατίθενται στον χρήστη για διευκολυμένη πρόσβαση σε όλο το εύρος της λειτουργικότητας (http://www.cs.waikato.ac.nz/ml/weka/).

Το Weka συνιστά μια συλλογή αλγορίθμων μηχανικής μάθησης για την επίλυση πραγματικών προβλημάτων εξόρυξης δεδομένων που λειτουργεί σε σχεδόν οποιαδήποτε πλατφόρμα. Οι αλγόριθμοι μπορούν είτε να εφαρμοστούν απευθείας σε ένα σύνολο δεδομένων είτε να φορτώνονται από τον εξατομικευμένα κώδικα Java του κάθε χρήστη.

Η αρχική έκδοση του προγράμματος (μη Java) σχεδιάστηκε κυρίως ως εργαλείο ανάλυσης δεδομένων από τον αγροτικό τομέα, ωστόσο η πιο πρόσφατη έκδοση που βασίζεται στην γλώσσα Java (Weka 3) και παραδόθηκε στο κοινό το 1997, χρησιμοποιείται σήμερα σε πολλούς τομείς, ιδίως για τους σκοπούς της έρευνας στην εκπαίδευση.

4.2 ΠΛΕΟΝΕΚΤΗΜΑΤΑ

Τα πλεονεκτήματα του Weka περιλαμβάνουν:

- Ελεύθερη πρόσβαση βάσει της γενικής δημόσιας άδειας GNU
- Φορητότητα, δεδομένου ότι υλοποιείται πλήρως στη γλώσσα προγραμματισμού Java και επομένως λειτουργεί σχεδόν σε οποιαδήποτε σύγχρονη πλατφόρμα υπολογιστών
- Μια ολοκληρωμένη συλλογή τεχνικών προεπεξεργασίας και μοντελοποίησης δεδομένων
- Ευκολία στη χρήση λόγω των γραφικών διεπαφών χρήστη

4.3 ΤΥΠΙΚΕΣ ΔΙΕΡΓΑΣΙΕΣ ΤΟΥ ΠΡΟΓΡΑΜΜΑΤΟΣ WEKA

Το πρόγραμμα Weka διαθέτει αρκετές τυπικές διεργασίες εξόρυξης δεδομένων. Πιο συγκεκριμένα, υποστηρίζει την προεπεξεργασία δεδομένων, την ομαδοποίηση, την ταξινόμηση, την παλινδρόμηση, την οπτικοποίηση και την επιλογή χαρακτηριστικών. Βασική προϋπόθεση χρήσης του Weka είναι η ενιαία διαθεσιμότητα των δεδομένων το καθένα εκ των οποίων περιγράφεται ως σετ που αποτελείται από έναν σταθερό αριθμό μεταβλητών (κανονικές, αριθμητικές ή ονομαστικές μεταβλητές ή χαρακτηριστικά). Η Weka παρέχει πρόσβαση σε βάσεις δεδομένων SQL χρησιμοποιώντας Java Database Connectivity και μπορεί να επεξεργαστεί το αποτέλεσμα που επιστρέφεται από ένα ερώτημα βάσης δεδομένων. Δεν είναι ικανή για πολυτομεακή εξόρυξη δεδομένων, αλλά υπάρχει ξεχωριστό λογισμικό για τη μετατροπή μιας συλλογής συνδεδεμένων πινάκων βάσης δεδομένων σε έναν

ενιαίο πίνακα που είναι κατάλληλος για επεξεργασία χρησιμοποιώντας το Weka (Sunita & Lobo, 2011).

Το σημείο εισόδου στην πλατφόρμα του Weka είναι ο επιλογέας Weka GUI. Σε αυτό το περιβάλλον, ο χρήστης μπορεί να επιλέξει την έναρξη των εργασιών που τον ενδιαφέρουν

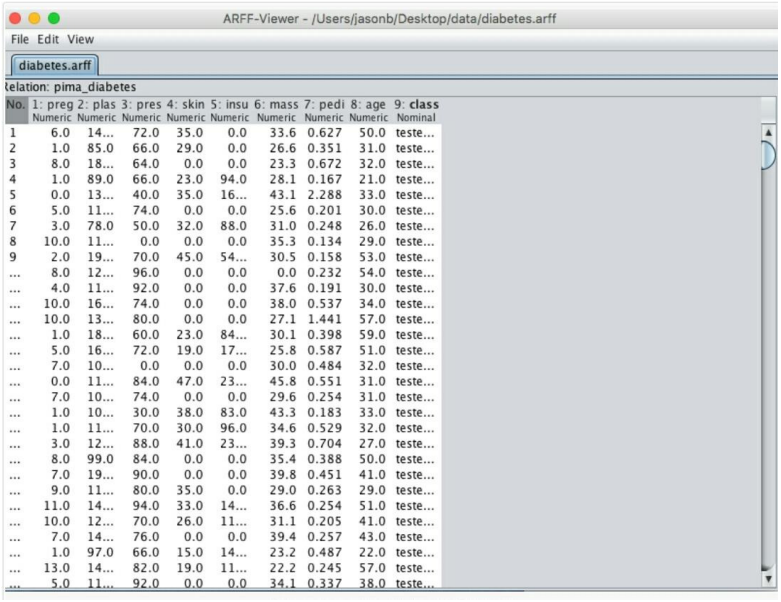
σε ένα συγκεκριμένο περιβάλλον Weka.

Εκτός από την πρόσβαση στα βασικά εργαλεία του Weka, διαθέτει επίσης ένα πλήθος πρόσθετων εργαλείων που παρέχονται στο μενού. Τα δύο σημαντικότερα βοηθητικά εργαλεία είναι :

1. Ο Διαχειριστής πακέτων που επιτρέπει την περιήγηση κι εγκατάσταση πρόσθετων (π.χ νέων αλγορίθμων) για τρίτους σε περιβάλλον Weka.

2. Το ARFF-Viewer που επιτρέπει την φόρτωση και μετατροπή συνόλων δεδομένων και την αποθήκευσή τους σε μορφή ARFF, η οποία είναι απαραίτητη για την ανάλυση οποιουδήποτε συνόλου δεδομένων.

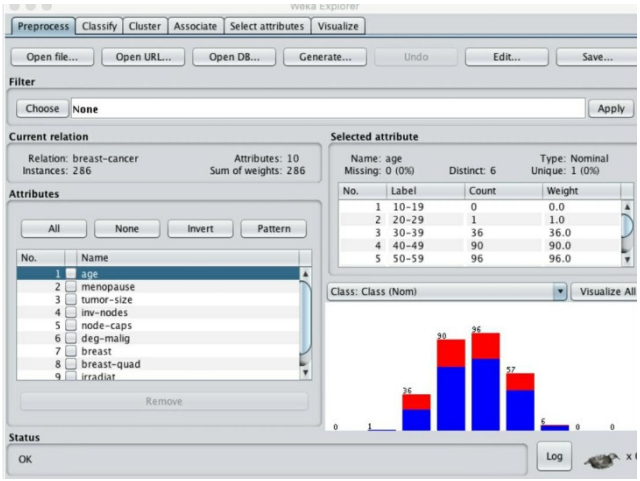
Weka Explorer



Ο Weka Explorer είναι μια ειδικά σχεδιασμένη εφαρμογή για την ανάλυση του το σύνολο δεδομένων μέσω μηχανικής εκμάθησης. Χωρίζεται σε 6 καρτέλες, κάθε μία με μια συγκεκριμένη λειτουργία:

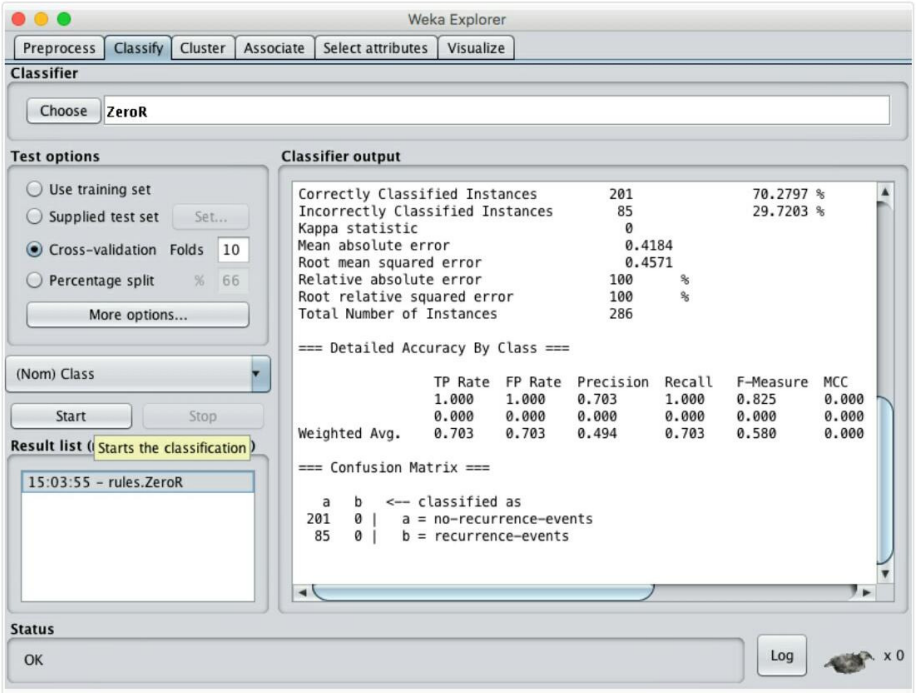
4.3.1 Προεπεξεργασία

Η καρτέλα προεπεξεργασίας προορίζεται για τη επιτυχή φόρτωση του συνόλου δεδομένων και την εφαρμογή φίλτρων για τη μετατροπή των δεδομένων σε μια φόρμα που εκθέτει καλύτερα τη δομή του προβλήματος στις διαδικασίες μοντελοποίησης. Παρέχει επίσης ορισμένα συνοπτικά στατιστικά στοιχεία σχετικά με τα φορτωμένα δεδομένα. Η φόρτωση ενός τυποποιημένου συνόλου δεδομένων στην εγκατάσταση στο Weka, εάν τα δεδομένα βρίσκονται στην συμβατή με το πρόγραμμα μορφή είναι το πρώτο βήμα για την μαζική τους επεξεργασία.



4.3.2 Ταξινόμηση

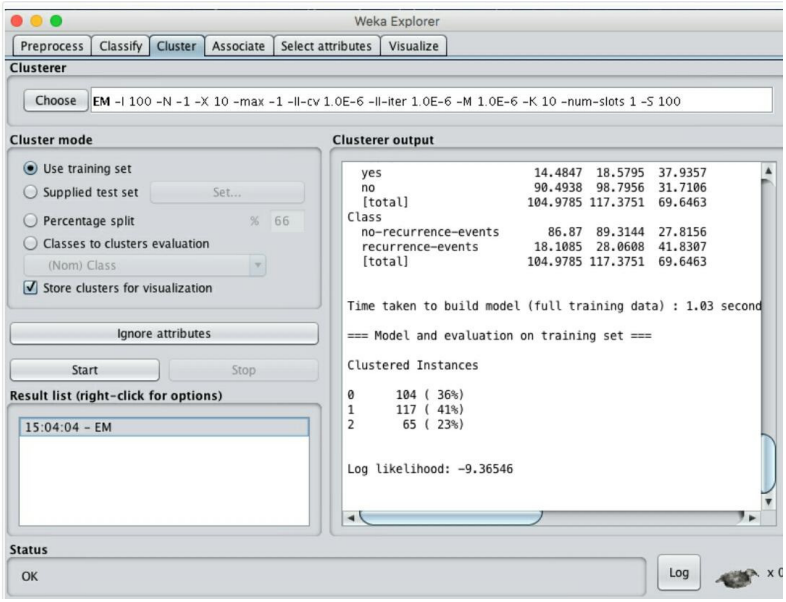
Η καρτέλα ταξινόμησης είναι για την εκπαίδευση και την αξιολόγηση της απόδοσης των διαφορετικών αλγορίθμων μηχανικής μάθησης στο πρόβλημα ταξινόμησης ή παλινδρόμησης. Οι αλγόριθμοι χωρίζονται σε ομάδες, τα αποτελέσματα διατηρούνται σε μια λίστα αποτελεσμάτων και συνοψίζονται στην κύρια απόδοση ταξινομητή. Η επιλογή του ταξινομητή ZeroR στο σύνολο δεδομένων οδηγεί στην σύνοψη των αποτελεσμάτων.



4.3.3 Ομαδοποίηση

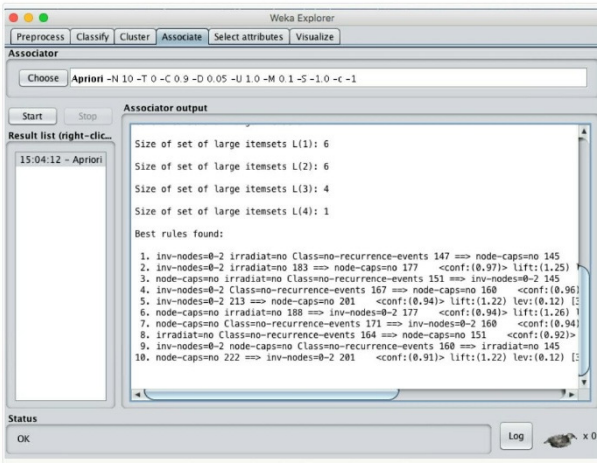
Η καρτέλα ομαδοποίησης (cluster) χρησιμεύει στην κατάρτιση και την αξιολόγηση της απόδοσης των διαφορετικών αλγόριθμων ομαδοποίησης χωρίς επίβλεψη στο σύνολο των δεδομένων που δεν φέρει ετικέτα. Όπως και στην καρτέλα Ταξινόμηση, οι αλγόριθμοι διαιρούνται σε ομάδες, τα αποτελέσματα διατηρούνται σε μια λίστα αποτελεσμάτων και συνοψίζονται στην κύρια έξοδο του Clusterer.

Η επιλογή εκτέλεσης του αλγορίθμου EM στο σύνολο δεδομένων οδηγεί στην σύνοψη των αποτελεσμάτων.



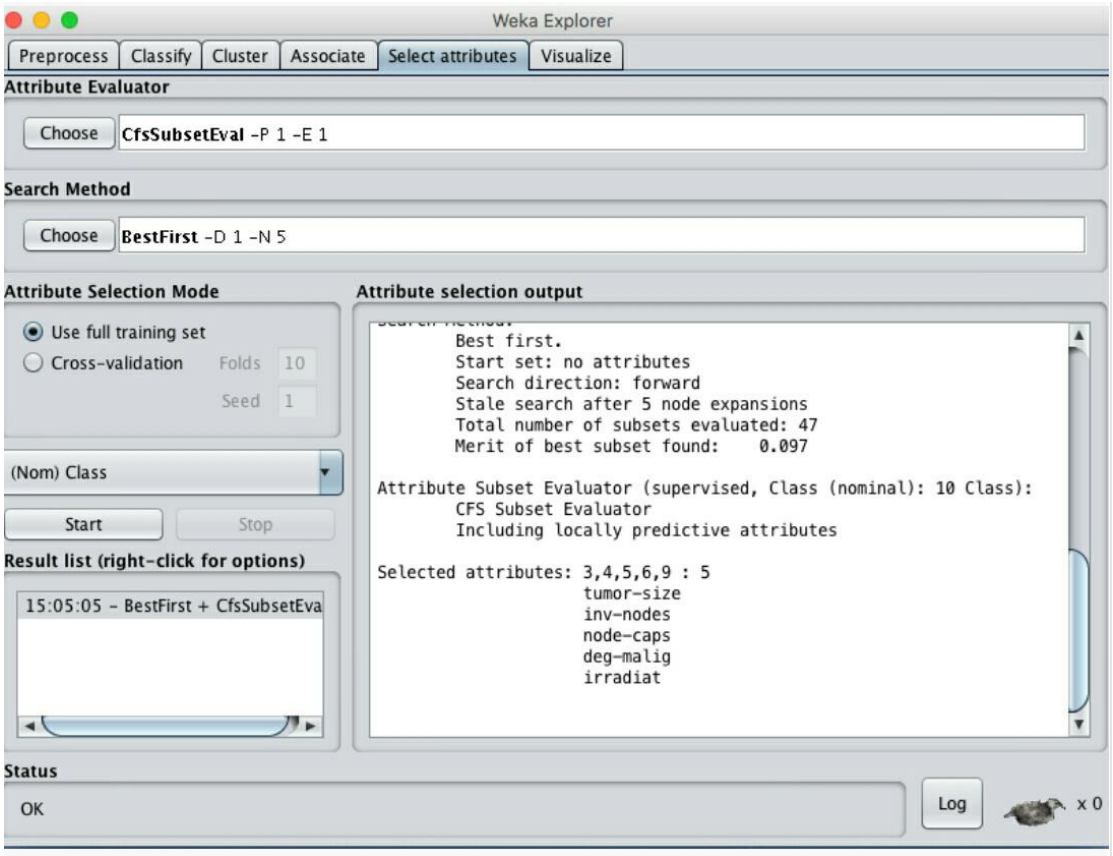
4.3.4 Συσχέτιση

Η καρτέλα της συσχέτισης (Association) εξυπηρετεί την αυτόματη εύρεση συσχετίσεων σε ένα σύνολο δεδομένων. Οι τεχνικές χρησιμοποιούνται συχνά για προβλήματα εξόρυξης δεδομένων τύπου έρευνας αγοράς και απαιτούν δεδομένα όπου όλα τα οι μεταβλητές είναι κατηγορηματικές. Η επιλογή της εφαρμογής του αλγορίθμου Apriori στο σύνολο δεδομένων οδηγεί στη σύνοψη των αποτελεσμάτων.



4.3.5 Επιλογή μεταβλητών

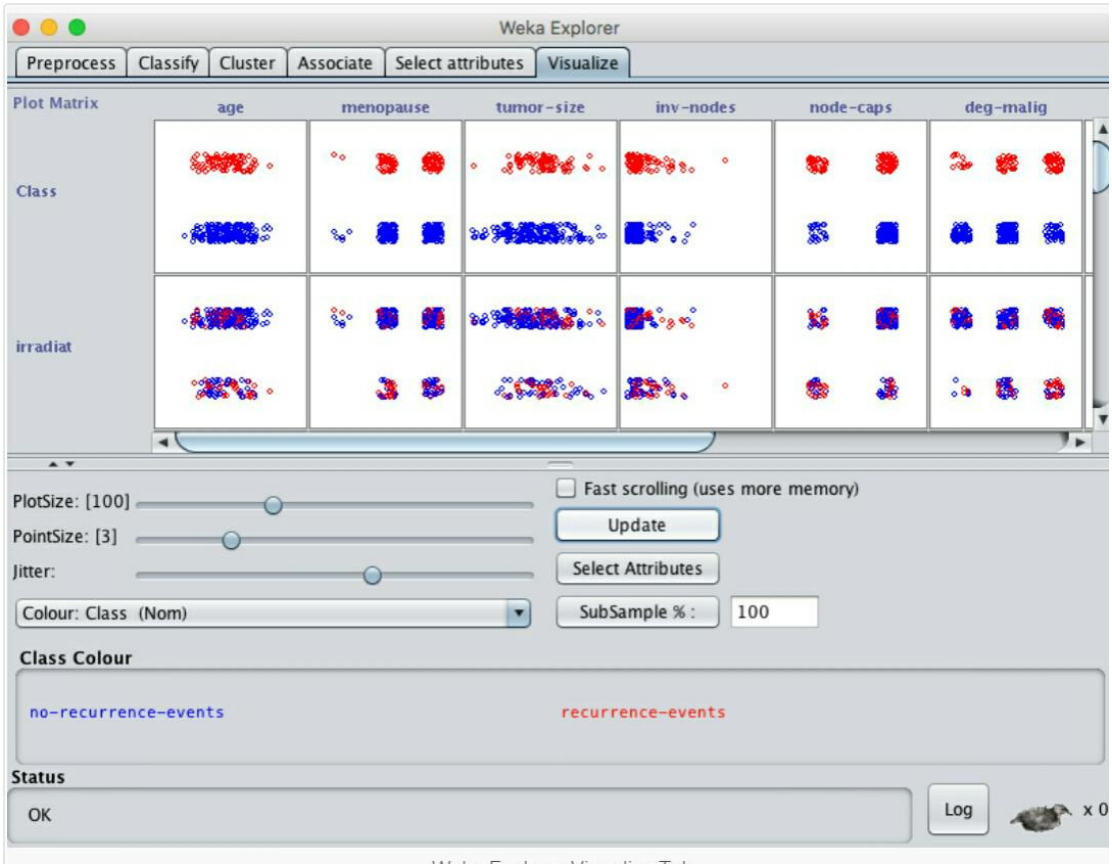
Η καρτέλα επιλογής μεταβλητών (select attributes) είναι για την εκτέλεση της επιλογής μεταβλητών στο φορτωμένο σύνολο δεδομένων και την αναγνώριση των μεταβλητών που είναι πιθανότερο να είναι σημαντικές για την ανάπτυξη ενός προγνωστικού μοντέλου. Η εκτέλεση του αλγορίθμου CfsSubsetEval μέσω της μεθόδου BestFirst στο σύνολο δεδομένων οδηγεί στη σύνοψη των αποτελεσμάτων.



4.3.6 Οπτικοποίηση

Η καρτέλα της οπτικοποίησης (visualize) είναι για την ανασκόπηση της ζεύγους για κάθε μεταβλητή που σχεδιάστηκε σε σχέση με κάθε άλλη μεταβλητή στο φορτωμένο σύνολο δεδομένων. Η συσχέτιση των μεταβλητών μπορούν να βοηθήσουν τον ερευνητή στο φιλτράρισμα, τον μετασχηματισμό και τη μοντελοποίηση δεδομένων.

Η τροποποίηση του jitter παράγει μια βελτιωμένη γραφική παράσταση των κατηγορικών μεταβλητών του φορτωμένου συνόλου δεδομένων.



Ακόμη κι αν το συγκεκριμένο λογισμικό παρουσιάζει αξιοσημείωτη ευχρησία, μερικά από τα προβλήματα συνεχίζουν να παρατηρούνται στην εξόρυξη δεδομένων μέσα από κοινωνικά δίκτυα. Τέτοια προβλήματα συνιστούν η ταχύτητα ή η εισαγωγή των δεδομένων που είναι τεράστια και η δυναμική φύση των δεδομένων που είναι επίσης απρόβλεπτη. Τα μαζικά δεδομένα που είναι διαθέσιμα στα κοινωνικά μέσα δεν είναι δομημένα, καθιστώντας απαραίτητη την ανεύρεση κατάλληλων αλγορίθμων που να προσφέρουν σωστά πρότυπα / τάσεις / συσχετισμούς που υπάρχουν μεταξύ των δεδομένων που είναι χρήσιμο για τους εκάστοτε ερευνητές και ενδιαφερόμενους. Το συγκεκριμένο πρόγραμμα εκπληρώνει πολλές από τις παραπάνω απαιτήσεις, ωστόσο διατηρεί κάποιους περιορισμούς που δύνανται να βελτιωθούν (Pushpam, Amali & Gnana, 2017).

4.4 Weka περιβάλλον πειραματισμού (Weka Experiment Environment)

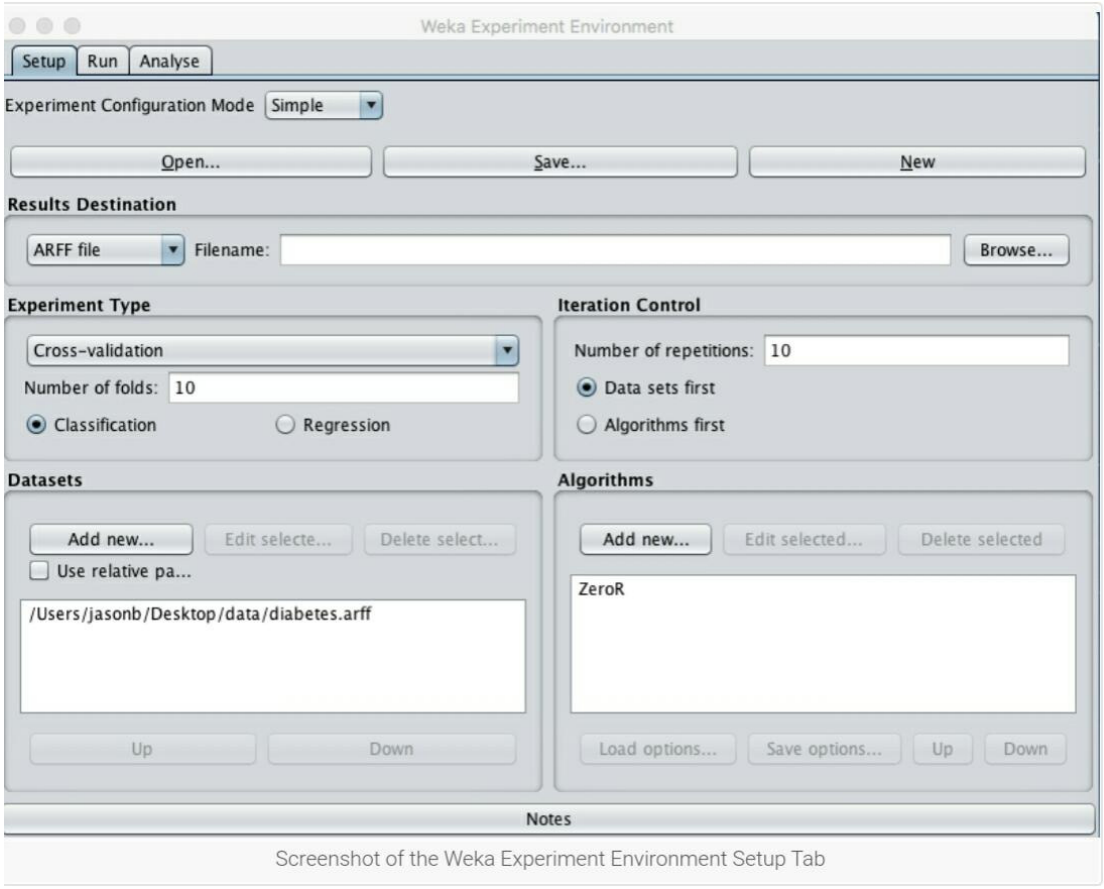
Το περιβάλλον πειραματισμού Weka έχει σχεδιαστεί για την εκπόνηση ελεγχόμενων πειραμάτων και την ανάλυση των αποτελεσμάτων που συλλέγονται.

Είναι το επόμενο βήμα μετά τη χρήση του Weka Explorer, όπου δύναται να φορτωθούν μία ή περισσότερες προβολές του συνόλου δεδομένων και μέσω μιας σειράς αλγορίθμων να σχεδιαστεί ένα πείραμα για να αναδειχθεί ο συνδυασμός που έχει ως αποτέλεσμα την καλύτερη απόδοση.

Το περιβάλλον πειραματισμού χωρίζεται σε 3 καρτέλες.

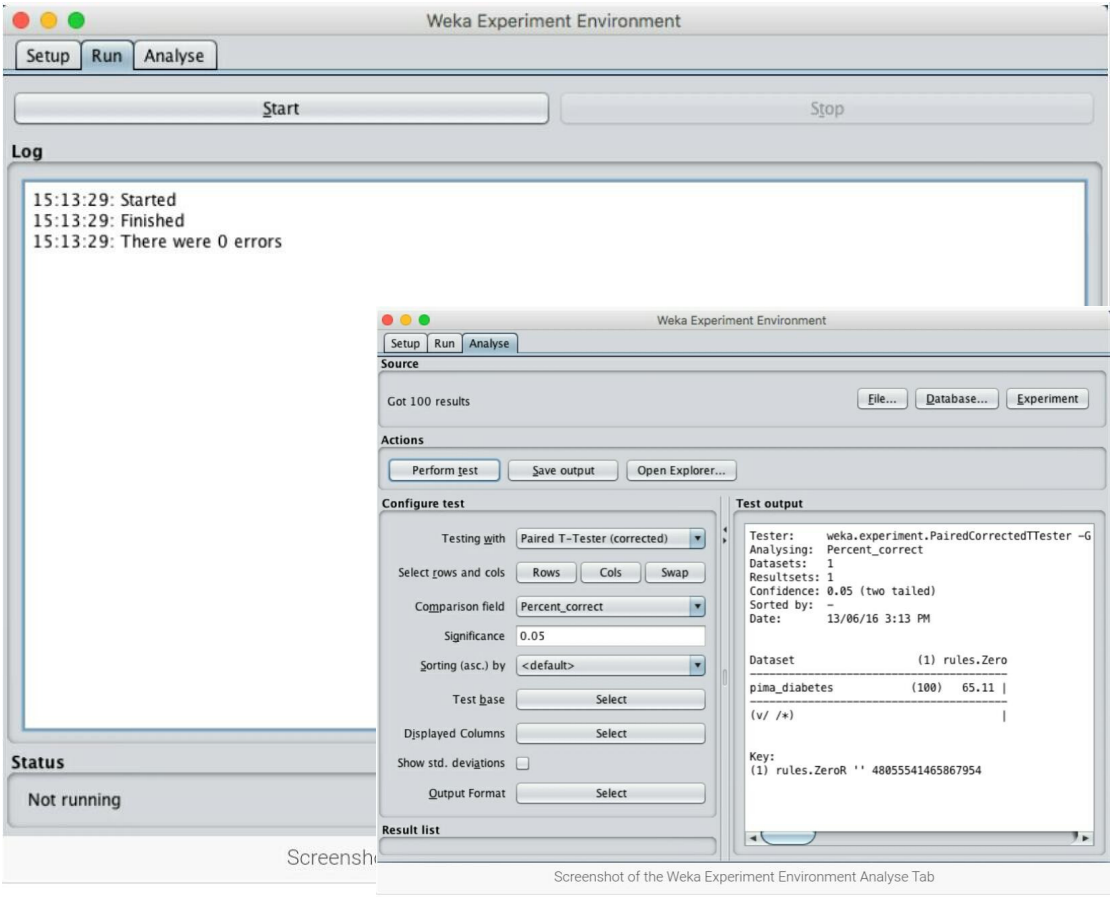
Η καρτέλα "Εγκατάσταση" είναι για το σχεδιασμό ενός πειράματος. Αυτό περιλαμβάνει το αρχείο όπου τα αποτελέσματα είναι γραμμένα, η δοκιμαστική διαμόρφωση από την άποψη του τρόπου αξιολόγησης των αλγορίθμων, των συνόλων δεδομένων σε μοντέλο και των αλγορίθμων για το μοντέλο τους. Οι ιδιαιτερότητες ενός πειράματος μπορούν να αποθηκευτούν για μεταγενέστερη χρήση και τροποποίηση.

Με την επιλογή "Νέο" δημιουργείται ένα νέο Πείραμα. Με την επιλογή "Προσθήκη νέου ..." στο παράθυρο του συνόλων δεδομένων επιλέγεται το σύνολο δεδομένων που είναι ένα αρχείο της μορφής .arff.



Screenshot of the Weka Experiment Environment Setup Tab

Η καρτέλα "Εκτέλεση" αφορά την εκτέλεση των σχεδιασμένων πειραμάτων. Τα πειράματα μπορούν να ξεκινήσουν και να σταματήσουν.



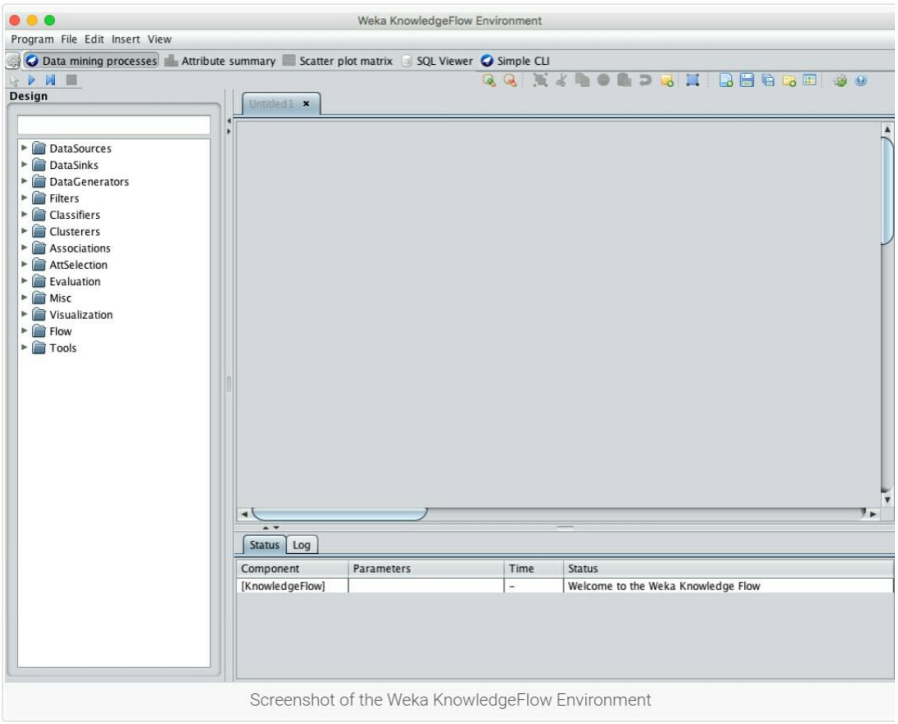
Screenshot of the Weka Experiment Environment Analyse Tab

Η καρτέλα ανάλυσης είναι για την ανάλυση των αποτελεσμάτων που συλλέγονται από ένα πείραμα. Τα αποτελέσματα μπορούν να φορτωθούν από ένα αρχείο, από τη βάση δεδομένων ή από ένα πείραμα που μόλις ολοκληρώθηκε στο εργαλείο. Μια σειρά μέτρων απόδοσης συλλέγονται από ένα δεδομένο πείραμα το οποίο μπορεί να συγκριθεί μεταξύ αλγορίθμων χρησιμοποιώντας εργαλεία όπως η στατιστική σημασία.

Με την επιλογή της παραμέτρου "Πείραμα" στο παράθυρο "Πηγή" προβάλλονται τα αποτελέσματα από το πείραμα που μόλις εκτελέστηκε. Η επιλογή "Εκτέλεση δοκιμής" περιγράφει τα αποτελέσματα της ακρίβειας ταξινόμησης για τον ενιαίο αλγόριθμο στο πείραμα.

4.5 Weka περιβάλλον KnowledgeFlow

Το Weka KnowledgeFlow Environment είναι ένα γραφικό εργαλείο ροής εργασιών για το σχεδιασμό ενός αγωγού μηχανικής μάθησης από την πηγή δεδομένων έως τη σύνοψη αποτελεσμάτων και πολλά άλλα. Μόλις σχεδιαστεί, ο αγωγός μπορεί να εκτελεστεί και να αξιολογηθεί μέσα στο εργαλείο.

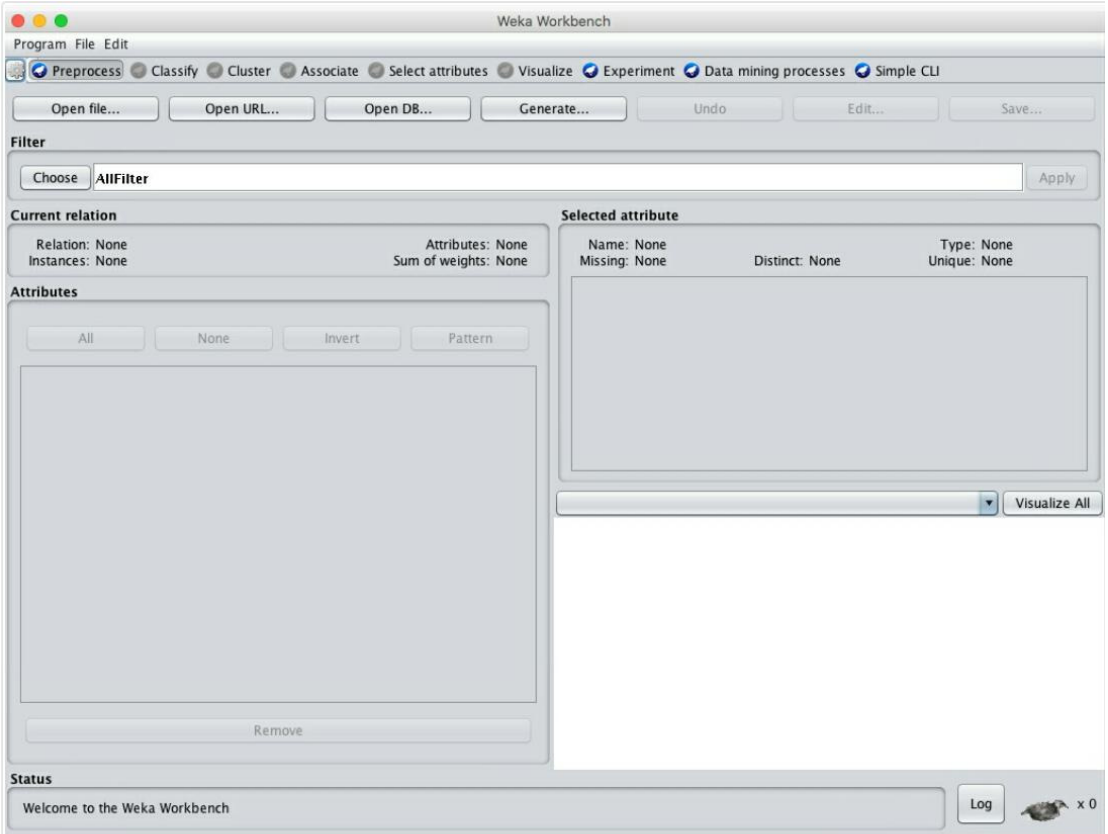


Screenshot of the Weka KnowledgeFlow Environment

Σε γενικές γραμμές, το περιβάλλον KnowledgeFlow είναι ένα ισχυρό εργαλείο το οποίο δεν ενδείκνυται για τους αρχάριους μέχρι να αποκτήσουν την σχετική εξοικείωση με το Weka Explorer και το Weka Experiment Environment.

4.6 Weka Workbench

Το Weka Workbench είναι ένα περιβάλλον που συνδυάζει όλες τις διεπαφές GUI σε μια ενιαία διεπαφή. Το συγκεκριμένο εργαλείο είναι ιδιαίτερα χρήσιμο αν ο χρήστης εναλλάσσει την χρήση δύο ή περισσότερων διαφορετικών διεπαφών, όπως του Explorer και του περιβάλλοντος πειραμάτων.

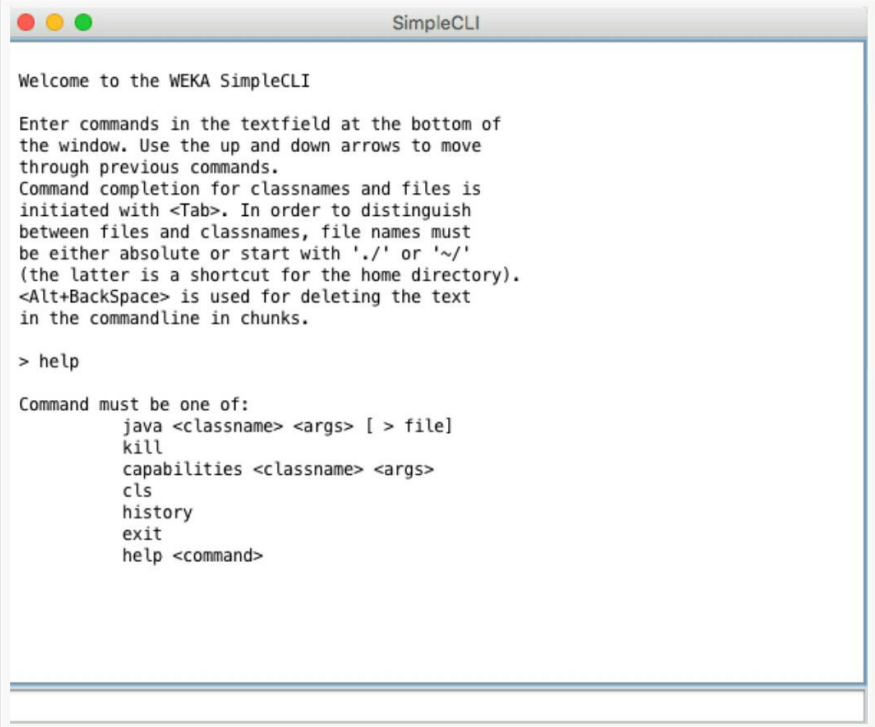


Screenshot of the Weka Workbench

4.7 Weka SimpleCLI

Το Weka μπορεί να χρησιμοποιηθεί από μια απλή διεπαφή γραμμής εντολών (CLI), επιτρέποντας στον χρήστη να δημιουργήσει μοντέλα, να εκτελέσει πειράματα και να κάνει προβλέψεις χωρίς γραφικό περιβάλλον χρήστη μέσω API από την γραμμή εντολών με παραμέτρους,

Το SimpleCLI παρέχει ένα περιβάλλον όπου οι χρήστες μπορούν να πειραματιστούν γρήγορα και εύκολα με τις εντολές διεπαφής γραμμής εντολών του Weka. Όπως και το περιβάλλον Weka KnowledgeFlow, αυτό είναι ένα ισχυρό εργαλείο που δεν προτείνεται στους αρχάριους μέχρι να έχουν καταφέρει να χρησιμοποιήσουν το Weka Explorer και το Weka Experiment Environment.



Screenshot of the Weka SimpleCLI

4.8 Weka Java API

Το Weka μπορεί επίσης να χρησιμοποιηθεί από το Java API. Το πρόγραμμα αυτό, απευθύνεται σε προγραμματιστές Java και μπορεί να είναι χρήσιμο στην ενσωμάτωση της μάθησης ή της πρόβλεψης στις προσωπικές εφαρμογές του χρήστη/προγραμματιστή. Πρόκειται και πάλι για ένα προηγμένο χαρακτηριστικό που δεν συνίσταται για αρχάριους, μέχρι να αποκτήσουν τη γνώση του Weka Explorer και του Weka Experiment Environment.

Παρά την φαινομενική πολυπλοκότητα όλων των παραπάνω, η αξιοποίηση τεχνικών εξόρυξης κειμένου σε ιστοσελίδες κοινωνικής δικτύωσης μπορεί να παρέχει σημαντικά και αξιοποιήσιμα αποτελέσματα σχετικά με τις πρακτικές επικοινωνίας μεταξύ των ανθρώπων, τις θέσεις τους απέναντι σε κοινωνικο-οικονομικά και πολιτικά ζητήματα, τα ενδιαφέροντά τους και τις προτιμήσεις τους. Η καταγραφή των αλληλεπιδράσεων των χρηστών των κοινωνικών δικτύων είναι απλώς ένα αρχικό στάδιο στην κατανόηση της συμπεριφοράς του

πελάτη από μια εντελώς προσωπική προοπτική (Markovikj, Gievska, Kosinski & Stillwell, 2013). Κεντρική συνιστώσα αποτελεί η συσχέτιση των εξαγόμενων δεδομένων που μπορούν να στοιχειοθετήσουν νέες γνώσεις ή να παρουσιάσουν νέα γεγονότα που χρήζουν περεταίρω διερεύνησης με περισσότερες στατιστικές ή εμπειρικές ερευνητικές μεθόδους (Irfan, King, Grages, Ewen, Khan et al., 2015).

ΣΥΜΠΕΡΑΣΜΑΤΑ

Η παραπάνω βιβλιογραφική ανασκόπηση επιχειρεί μια σύνοψη της δυναμικής της εξόρυξης δεδομένων, η οποία έρχεται να καλύψει τα ερευνητικά κενά της πολυχρηστικής και απαραίτητης στατιστικής ανάλυσης, συμπεριλαμβάνοντας την ανάλυση αδόμητων δεδομένων μεγάλου όγκου.

Προφανώς, τα οφέλη που προκύπτουν μέσα από την εξόρυξη δεδομένων είναι τεράστια και πολυποίκιλα. Η εξόρυξη δεδομένων εκτελούμενη από ερασιτέχνες μπορεί να απαντήσει σαφώς ορισμένα ερωτήματα, ενώ πολύ περισσότερο εκτελούμενη από επαγγελματίες μπορεί να αλλάξει το πολιτικό, κοινωνικό, πολιτισμικό και επιχειρηματικό σκηνικό. Αποτελεί ένα πολύτιμο όπλο στη φαρέτρα του ενθουσιώδη και κατάλληλα εκπαιδευμένου νέου που στοχεύει στην εκμετάλλευση του μεγαλύτερου ίσως όγκου απρόσκοπτης και προσβάσιμης γνώσης που προστίθεται καθημερινά από κάθε ξεχωριστό άτομο στα πέρατα της υφηλίου. Δεδομένου ότι το ψηφιακό αποτύπωμα του κάθε ανθρώπου, εκτός από τις περιπτώσεις καταστρατήγησης των προσωπικών δεδομένων που είναι κάτι μη αποδεκτό και πρέπει να αποφεύγεται, μπορεί να οδηγήσει σε μια τεράστιας αξίας συμπυκνωμένη και ολότελα καινούρια γνώση.

Οι ατελείωτες δυνατότητες της εξόρυξης δεδομένων πρέπει να καταστούν όσο το δυνατό περισσότερο γνωστές στους νέους με ερευνητικές και επιστημονικές αναζητήσεις. Αυτόν τον σκοπό και πολλούς ακόμη μπορεί να ενισχύσουν προγράμματα και λογισμικά ελεύθερης πρόσβασης όπως το Weka. Όπως περιγράφηκε αναλυτικά, η χρήση του είναι ιδιαιτέρως απλουστευμένη, δεν απαιτείται συνδρομή ή εφάπαξ πληρωμή από τον χρήστη και με μια στοιχειώδη καθοδήγηση μπορεί να πυροδοτήσει την δημιουργία γνώσης σε όλους τους τομείς που μπορούν να δημιουργήσουν ένα καλύτερο αύριο.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Bellinger, C., Mohomed Jabbar, M. S., Zaïane, O., & Osornio-Vargas, A. (2017). A systematic review of data mining and machine learning for air pollution epidemiology. BMC public health, 17(1), 907. doi:10.1186/s12889-017-4914-3

2. Maimon, O., & Rokach, L. (2010). Data Mining and Knowledge Discovery Handbook, 2nd ed,

3. Jiawei Han and Jing Gao. "Research Challenges for Data Mining in Science and Engineering", 09/01/2008-08/31/2009, , H. Kargupta, et al."Next Generation of Data Mining", 2009, "Chapman & Hall/CRC".

4. Injadat, M. Salo, F. & Nassif, A. (2016). Data Mining Techniques in Social Media: A Survey. Neurocomputing. 214. 10.1016/j.neucom.2016.06.045International Journal of Computing and Digital SystemsISSN (2210-142X)

5. Salloum, S., Al-Emran, M. & Shaalan, K. (2016). Mining Social Media Text: Extracting Knowledge from Facebook. Int. J. Com. Dig. Sys. 6, No.2.

6. Irfan, R., King, C. K., Grages, D., Ewen, S., Khan, S. U., Madani, S. A., ...& Tziritas, N. (2015). A survey on text mining in social networks. The Knowledge Engineering Review, 30(02), 157-170.

7. Chakraborty, G., & Krishna, M. (2014). Analysis of unstructured data: Applications of text analytics and sentiment mining. In SAS global forum(pp. 1288-2014).

8. Markovikj, D., Gievska, S., Kosinski, M., & Stillwell, D. (2013). Mining facebook data for predictive personality modeling. In Proceedings of the 7th international AAAI conference on Weblogs and Social Media (ICWSM 2013), Boston, MA, USA.

9. Tsoumakas, G. & Katakis, I. & Vlahavas, I. (2010). Mining Multi-label Data in The Data Mining and Knowledge Discovery Handbook, pp. 667-685 10.1007/978-0-387-09823-4_34.

10. Kalmegh, S. (2015). Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News. International Journal of Innovative Science, Engineering & Technology, 2(2).

11. Anick, P.G & Vaithyanathan, S. (1997). Exploiting clustering and phrases for context-based information retrieval. In ACM SIGIR Forum, 31(3), 314–323.

12. Bharati, M. & Ramageri, A., (2010). DATA MINING TECHNIQUES AND APPLICATIONS. Indian Journal of Computer Science and Engineering. 1(3), 255-257.

13. Fayyad, U; Piatetsky-Shapiro, G; Smyth, P. From data mining to knowledge discovery in databases. AI magazine 1996;17:37-54.

14. Asian Journal of Applied Science and Technology (AJAST) (Open Access Quarterly International Journal) Volume 2, Issue 2, Pages 947-953, 2018.

15. Zurada J.M. (1992), "An introduction to artificial neural networks systems", St. Paul: West Publishing.

16. Berry, J. A., Lindoff, G. (1997), Data Mining Techniques, Wiley Computer Publishing (ISBN 0-471-17980-9).

17. Bengio Y., Buhmann J. M., Embrechts M., and Zurada J.M (2012). Introduction to the special issue on neural networks for data mining and knowledge discovery. IEEE Trans. Neural Networks.

18. Injadat, Mohammadnoor & Salo, Fadi & Nassif, Ali. (2016). Data Mining Techniques in Social Media: A Survey. Neurocomputing. 214. 10.1016/j.neucom.2016.06.045.

19. H. Chen, R.H.L. Chiang, V.C. Storey, Business Intelligence and Analytics: From Big Data To Big Impact, Mis Q. 36 (2012) 1165–1188

20. S.G.S Fernando et.al "Empirical Analysis of Data Mining Techniques for Social Network Websites" in COMPUSOFT, An international journal of advanced computer technology, Volume-III, Issue-II PP:582-592, 2014.

21. Pushpam, C.Amali & Joseph, Gnana. (2017). Over view on Data Mining in Social Media. International Journal of Computer Sciences and Engineering. 5. 147-157. 10.26438/ijcse/v5i11.147157.

22. S., Vijayalakshmi & V, Mahalakshmi & S, Magesh. (2013). Knowledge discovery from consumer behavior in electronic home appliances market in Chennai by using data mining techniques. African Journal of Business Management. 7. 3332-3342. 10.5897/AJBM2013.7038

23. MiloviC, B. and V. RadojeviC, 2015. Application of data mining in agriculture. Bulg. J. Agric. Sci., 21: 26-34

24. MiloviC, B. and V. RadojeviC, 2015. Application of data mining in agriculture. Bulg. J. Agric. Sci., 21: 26-34

25. Jagtap, S B. (2013). Census Data Mining and Data Analysis using WEKA 35, ICETSTM – 2013) International Conference in "Emerging Trends in Science, Technology and Management-2013, Singapore.

26. Sunita B Aher, Lobo LMRJ, Data Mining in Educational System using Weka,

27. International Conference on Emerging Technology Trends (ICETT), Proceedings published by International Journal of Computer Applications® (IJCA) Number 3, 2011, pp-20-25.

28. Doug Wielenga (2007), Identifying and Overcoming Common Data Mining Mistakes, SAS Global Forum Paper 073-2007

29. Brusilovsky, P. (2009) Data Mining vs. Statistics, Business Intelligence Solutions [online] available at : http://www.bisolutions.us/Data-Mining-vs-Statistics.php retrieved on 12 July, 2019.

30. Agarwal, P., (2014). Benefits and Issues Surrounding Data Mining and its Application in the Retail Industry, International Journal of Scientific and Research Publications, 4(7), 115-117.

31. Injadat, M., Salo, F. & Nassif, A.B. Data Mining Techniques in Social Media: A Survey, Neurocomputing, http://dx.doi.org/10.1016/j.neucom.2016.06.045.
32. Pushpam, A. & Jayanthi, G. Overview on Data Mining in Social Media. International Journal of Computer Sciences and Engineering 2017;5(11): 2347-2693.
33. Yadanar, T., Htun, H. & Soe, NC. The Opinion Mining from Social Media by using Support Vector Machine (SVM) Algorithm. International Journal of Science, Engineering and Technology Research (IJSETR), 2018;7(9): 2278 -7798.
34. Koelle, D., Pfautz, J., Farry, M., Cox, Z., Catto, G. & Campolongo, J. Applications of Bayesian Belief Networks in Social Network Analysis In *Proceedings of 4th Bayesian Modeling Applications Workshop at the 22nd Annual Conference on Uncertainty* in AI: UAI '06.
35. North, M. (2012). Data Mining for the Masses. A Global Text Project Book. Creative Commons Attribution.ρρ. 121-134.
36. Deepa B, JeenMarseline K.S (2019). Social Media Data using Various Classification Algorithms in Datamaning. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 8(12):2278-3075.
37. S.Neelamegam."Classification algorithm in Data mining: An Overview" in International Journal of P2P Network Trends and Technology (IJPTT), Volume 4, Issue 8, PP:369 – 374, 2013.
38. Jovanovic, M., Vukicevic, M., Delibašić, B. & Suknovic, M. (2014). Using RapidMiner for Research: Experimental Evaluation of Learners.
39. Ristoski, P., Bizer, C. & Paulheim, H. Mining the Web of Linked Data with RapidMiner, Web Semantics: Science, Services and Agents on the World Wide Web (2015), http://dx.doi.org/10.1016/j.websem.2015.06.004.
40. Dwivedi, S., Kasliwal, P. & Soni, S. Comprehensive study of data analytics tools (RapidMiner, Weka, R tool, Knime), 2016 Symposium on Colossal Data Analysis and Networking (CDAN), Indore, 2016, pp. 1-8.
41. Seefeld, K. & Linder., E., Statistics Using R with Biological Examples.Boston, M.A. University of New Hampshire, Durham, NH Department of Mathematics & Statistics, 2007, pp. 6-34.
42. Berthold, M., Cebron, N., Dill, F., Gabriel, T., Kotter, T., Meinl, T., Ohl, P., Thiel, K., Wiswedel, B. KNIME – The Konstanz Information Miner" presented at University of Konstanz Nycomed Chair for Bioinformatics and Information Mining, Germany. 2009.

43. Kataria, L. Implementation of Knime-Data Mining Tool. presented at International Journal of Advanced Research in Computer Science and Software Engineering, 3(11), 2013.
44. Kalpana R. & Bansal, K. Comparative Study of Data Mining Tools, presented at International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 6, 2014.
45. Witten, I.E. Data Mining: Practical machine Learning tools and techniques, San Francisco: Morgan Kaufmann, 2005.
46. Adedoyin-Olowe, M., Gaber, M.M, Stahl, F. (2014). A Survey of Data Mining Techniques for Social Media Analysis. Pdf. [Online] Available at : https://arxiv.org/vc/arxiv/papers/1312/1312.4617v1.pdf Accesed in 28 April, 2020.
47.

Πνευματικά δικαιώματα