ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Εξόρυξη Δεδομένων και εκπαιδευτικά δεδομένα
(Data Mining and Educational Data)

Ιωάννης Κωνσταντινίδης

161076

Επιβλέπων καθηγητής: Χρήστος Σκουρλάς

ΑΙΓΑΛΕΩ, ΙΟΥΛΙΟΣ 2021

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Ο κάτωθι υπογεγραμμένος Ιωάννης Κωνσταντινίδης, με αριθμό μητρώου 161076 φοιτητής του Πανεπιστημίου Δυτικής Αττικής της Σχολής Μηχανικών του Τμήματος Μηχανικών Πληροφορικής και Υπολογιστών, δηλώνω υπεύθυνα ότι:

«Είμαι συγγραφέας αυτής της διπλωματικής εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από εμένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος. Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου».

Christos Skourlas
Digitally signed by Christos Skourlas
Date: 2021.07.23 09:27:24 +03'00'

Vasileios Mamalis
Digitally signed by Vasileios Mamalis
Date: 2021.07.23 22:43:06 +03'00'

KLEIO SGOUROPOULOU
Digitally signed by KLEIO SGOUROPOULOU
Date: 2021.07.28 12:42:41 +03'00'

Ο Δηλών

Ιωάννης Κωνσταντινίδης

Υπογραφή

ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ και ΥΠΟΛΟΓΙΣΤΩΝ

Η παρούσα διπλωματική εργασία παρουσιάστηκε

από τον Ιωάννη Κωνσταντινίδη (161076) στις 21/7/2021

Εγκρίθηκε από τριμελή εξεταστική επιτροπή

Επιβλέπων καθηγητής:
Χρήστος Σκουρλάς

Μέλος επιτροπής:
Κλειώ Σγουροπούλου

Μέλος επιτροπής:
Βασίλειος Μάμαλης

ΠΕΡΙΛΗΨΗ

Η ανάπτυξη των Τεχνολογιών Πληροφορικής και Επικοινωνιών σε συνδυασμό με την ανάπτυξη του Διαδικτύου έχει αυξήσει σημαντικά την ποσότητα δεδομένων σε πληθώρα πεδίων, συμπεριλαμβανομένης της Εκπαίδευσης. Η υγειονομική κρίση και η εκτεταμένη χρήση διαδικασιών Εκπαίδευσης από Απόσταση οδήγησε σε περαιτέρω αύξηση εκπαιδευτικών δεδομένων. Ο τομέας της Εκπαίδευσης μπορεί να ωφεληθεί, με την αξιοποίηση των δεδομένων που παράγονται από τη διαδικασία της τηλεκπαίδευσης, μετά από επεξεργασία τους με χρήση τεχνικών και εργαλείων εξόρυξης δεδομένων. Έτσι, δίνεται η δυνατότητα για εξαγωγή πολύτιμης γνώσης που μπορεί να συμβάλει στην λήψη αποφάσεων σχετικά με την εκπαιδευτική διαδικασία. Στην παρούσα εργασία, χρησιμοποιήθηκε ανωνυμοποιημένα δεδομένα από το προπτυχιακό μάθημα «Βάσεις Δεδομένων» του τμήματος Μηχανικών Πληροφορικής και Υπολογιστών του Πανεπιστημίου Δυτικής Αττικής. Η επεξεργασία των δεδομένων έγινε με το εργαλείο RapidMiner Studio και χρησιμοποιήθηκαν τεχνικές clustering και δέντρων απόφασης. Με την οπτικοποίηση των αποτελεσμάτων παρατηρήθηκαν δυνητικά χρήσιμες συμπεριφορές των φοιτητών σε δύο κατευθύνσεις: α) συμμετοχή φοιτητών σε ασύγχρονη τηλεκπαίδευση μέσω της πλατφόρμας eClass και β) συμμετοχή φοιτητών σε σύγχρονη τηλεκπαίδευση μέσω της πλατφόρμας Microsoft Teams.

Λέξεις κλειδιά: Εξόρυξη Δεδομένων, Μηχανική Μάθηση, RapidMiner, K-Means, Δέντρα Απόφασης, Εκπαίδευση από Απόσταση

ABSTRACT

The development of Information and Communication Technology, and the growth of the Internet have significantly increased the amount of data in numerous fields including education. The health crisis and the application of distance learning procedures in Higher Education led to further increase of educational data. Data generated by the distance learning procedures, and the use of data mining techniques and tools could be beneficiary for the education sector. The extraction of valuable information and knowledge can assist in making decisions related to the educational process. In the present diploma thesis, anonymous data was extracted from the undergraduate course "Databases" of the Department of Informatics and Computer Engineering of the University of West Attica. These data were processed applying the RapidMiner Studio tool. Clustering and decision tree techniques were used. Visualization of the results offered potentially useful conclusions related to the behaviors observed in two directions: a) students' participation through asynchronous distance learning which is based on the open eClass platform, and b) students' participation through synchronous distance learning, which is based on the Microsoft Teams platform.

Keywords: Data mining, Machine Learning, RapidMiner, K-Means, Decision trees, Distance Learning

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ ΚΑΙ ΕΙΚΟΝΩΝ

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω θερμά τους επιβλέποντες καθηγητές μου κ. Χρήστο Σκουρλά και κ. Αναστάσιο Τσολακίδη και τον υποψήφιο διδάκτορα κ. Κωνσταντίνο Χύτα για την πολύτιμη βοήθεια τους και την καθοδήγηση τους στην εκπόνηση και τη συγγραφή της παρούσας εργασίας. Τους ευχαριστώ για την ευκαιρία να συνεργαστώ μαζί τους, να αποκομίσω πολύτιμες γνώσεις και να διευρύνω τις γνώσεις μου πάνω στον τομέα της Εξόρυξης Δεδομένων. Θα ήθελα επίσης να ευχαριστήσω τη Διοίκηση και όλους τους καθηγητές του τμήματος Πληροφορικής και των Υπολογιστών του Πανεπιστημίου Δυτικής Αττικής για όλες τις γνώσεις και την υποστήριξη που μου προσέφεραν τα πέντε χρόνια φοίτησης μου στο τμήμα. Τέλος θα ήθελα να ευχαριστήσω την οικογένεια μου και τους συμφοιτητές μου για την στήριξη τους σε όλη τη διάρκεια των σπουδών μου.

1. ΕΙΣΑΓΩΓΗ

Η υγειονομική κρίση των τελευταίων χρόνων επέβαλε μεγάλες αλλαγές στην καθημερινότητα των ανθρώπων σε παγκόσμιο επίπεδο. Η διαδικτυακή κίνηση αυξήθηκε κατακόρυφα εξαιτίας των lockdown, παράγοντας τεράστιο όγκο δεδομένων. Αυτά τα δεδομένα αποτελούν την πρώτη ύλη, μετά από επεξεργασία, για την παραγωγή πολύτιμης γνώσης σε πολλά πεδία, συμπεριλαμβανομένης της εκπαίδευσης.

1.1 ΣΚΟΠΟΣ ΤΗΣ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Σκοπός της παρούσας εργασίας είναι η η διερεύνηση της χρήσης αλγορίθμων εξόρυξης δεδομένων σε εκπαιδευτικά δεδομένα με σκοπό την μελέτη της συμπεριφοράς των φοιτητών και πως αυτή έχει επηρεαστεί από την απομακρυσμένη διεξαγωγή των μαθημάτων. Η ανάπτυξη της εργασίας χωρίζεται σε δύο μέρη: α) συζήτηση της βιβλιογραφίας, και β) πειραματικό μέρος, δηλαδή η εφαρμογή αλγόριθμων εξόρυξης δεδομένων σε εκπαιδευτικά δεδομένα με χρήση του εργαλείου RapidMiner Studio.

1.2 ΘΕΜΑΤΙΚΗ ΠΑΡΟΥΣΙΑΣΗ ΚΕΦΑΛΑΙΩΝ

Η διάρθρωση της παρούσας εργασίας καταγράφεται παρακάτω:

Στο 2^{ο} κεφάλαιο γίνεται παρουσίαση της εξόρυξης δεδομένων. Παρουσιάζονται οι πηγές, κάποια από τα βασικότερα πεδία εφαρμογής και δημοφιλή εργαλεία της εξόρυξης δεδομένων.

Στο 3^{ο} κεφάλαιο παρουσιάζεται η διαδικασία της τηλεκπαίδευσης και οι δύο συνιστώσες της, η σύγχρονη και ασύγχρονη εκπαίδευσης, που χρησιμοποιήθηκαν στο ΠΑΔΑ κατά το ακαδημαϊκό έτος 2019-2020.

Στο 4ο κεφάλαιο παρουσιάζονται τα είδη μηχανικής μάθησης και ευρέως χρησιμοποιούμενες μέθοδοι. Επίσης, παρουσιάζονται οι αλγόριθμοι που χρησιμοποιούνται στο πειραματικό μέρος και η χρησιμότητα των εργαλείων οπτικοποίησης στην εξόρυξη δεδομένων.

Στο 5ο κεφάλαιο παρουσιάζεται το πειραματικό μέρος της εργασίας. Αρχικά, παρουσιάζεται η δομή του αρχείου δεδομένων και το εργαλείο που χρησιμοποιήθηκε.
Καταγράφεται η διαδικασία εφαρμογής των αλγόριθμων στα δεδομένα, και γίνεται αναφορά στις χρήσιμες οπτικοποιήσεις που παρέχονται από το εργαλείο εξόρυξης δεδομένων.

Τέλος, στο 6ο κεφάλαιο παρουσιάζονται τα συμπεράσματα.

2. ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

2.1 ΓΕΝΙΚΑ ΓΙΑ ΤΗΝ ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

Τα τελευταία χρόνια, όλο και περισσότεροι άνθρωποι σε κάθε γωνιά του πλανήτη έχουν τη δυνατότητα πρόσβασης στο διαδίκτυο. Η δυνατότητα αυτή οδήγησε με τη σειρά της σε ραγδαία αύξηση δημιουργίας ιστοτόπων και εφαρμογών. Καθημερινά δεκάδες petabytes δεδομένων μεταδίδονται στα παγκόσμια τηλεπικοινωνιακά δίκτυα.

Επιστημονικές και μηχανικές πρακτικές παράγουν συνεχώς τεράστιο όγκο δεδομένων. Κάποιες από αυτές είναι η τηλεπισκόπηση, η μέτρηση διεργασιών, τα επιστημονικά πειράματα, οι επιδόσεις συστήματος, οι μηχανικές παρατηρήσεις και η παρακολούθηση του περιβάλλοντος. Η ιατρική και η βιομηχανία υγείας, επίσης, αποτελούν πηγή δεδομένων από ιατρικά αρχεία και δεδομένα ασθενών. Καθημερινά δισεκατομμύρια αναζητήσεις σε μηχανές αναζήτησης επεξεργάζονται δεκάδες petabytes δεδομένων. Τα μέσα κοινωνικής δικτύωσης παράγουν καθημερινά πολυμεσικό υλικό και δεδομένα που αφορούν τις προτιμήσεις και τις ανάγκες των χρηστών τους.

Η λίστα των πηγών που συμβάλλουν στην τεράστια αύξηση δεδομένων είναι ατελείωτη. Όλα αυτά μας έχουν οδηγήσει στην «εποχή των δεδομένων», όπου το πλήθος των διαθέσιμων δεδομένων είναι τεράστιο και αυξάνεται εκθετικά κάθε μέρα. Ο τεράστιος όγκος δεδομένων, που συσσωρεύεται σε βάσεις δεδομένων και αποθήκες δεδομένων (data warehouses) δεν μπορεί να αξιοποιηθεί όπως είναι.

Έχει προκύψει επείγουσα ανάγκη για νέα εργαλεία και τεχνικές που μπορούν να συμβάλουν αποτελεσματικά στην αξιοποίηση των τεράστιων ποσοτήτων δεδομένων, μετατρέποντας τα σε χρήσιμες πληροφορίες και γνώσεις. Όμως, πρέπει αρχικά να γίνουν κάποιες ενέργειες, για να δομηθούν κατάλληλα τα δεδομένα, ώστε στη συνέχεια να μπορεί να εξαχθεί από τα δεδομένα χρήσιμη και αξιοποιήσιμη πληροφορία. Αυτή η αναγκαιότητα οδήγησε στη γέννηση της εξόρυξης δεδομένων.

Η εξόρυξη δεδομένων, επίσης γνωστή ως ανακάλυψη γνώσης από δεδομένα (KDD), είναι η εξαγωγή μοτίβων που αντιπροσωπεύουν γνώση που αποθηκεύεται ή συλλαμβάνεται σιωπηρά σε μεγάλες βάσεις δεδομένων, αποθήκες δεδομένων, στον Ιστό και σε τεράστια αποθετήρια πληροφοριών ή ροές δεδομένων.

Η εξόρυξη δεδομένων εμφανίστηκε στα τέλη της δεκαετίας του 1980, έκανε μεγάλα βήματα κατά τη διάρκεια της δεκαετίας του 1990 και συνεχίζει να αναπτύσσεται ραγδαία μέχρι και σήμερα.

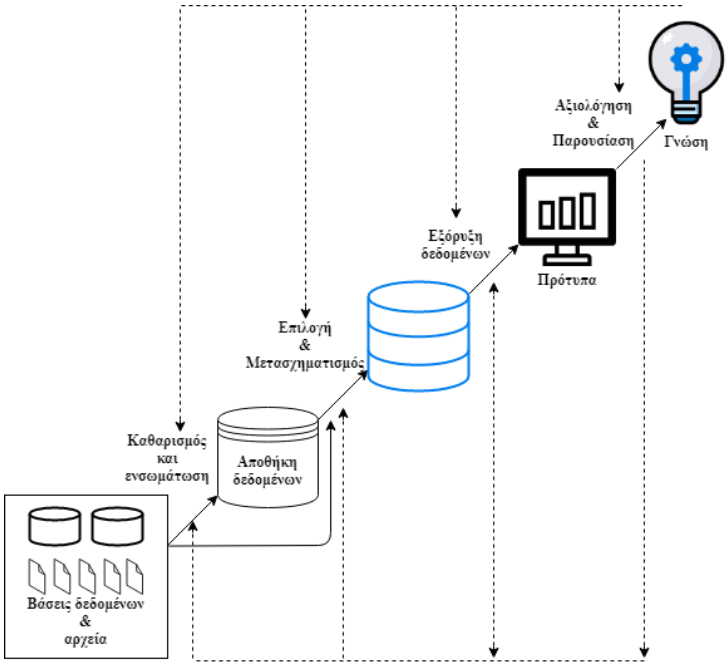
2.2 ΒΑΣΙΚΑ ΒΗΜΑΤΑ

Η διαδικασία της Εξόρυξης Γνώσης είναι μια αμφίδρομη διαδικασία που εκκινούν από τη συλλογή δεδομένων και φτάνουν σε πιο πρακτικό επίπεδο, στην αξιοποίηση των αποτελεσμάτων μέσα από επαναλαμβανόμενα στάδια. Η διαδικασία επαναλαμβάνεται, διότι οι χρήστες πολύ συχνά δεν έχουν εκ των προτέρων καθαρή εικόνα για το ποια πληροφορία παρουσιάζει ενδιαφέρον. Επιπλέον, αρκετά συχνά μέσω της εξαγωγής των πρώτων συμπερασμάτων προκύπτουν νέα

ερωτήματα. Ακόμη, ενδέχεται τα αποτελέσματα της ανάλυσης των δεδομένων να μην οδηγήσουν σε αξιοποιήσιμα συμπεράσματα, με αποτέλεσμα να πρέπει να επανασχεδιαστεί η διαδικασία.

Η διαδικασία ανακάλυψης γνώσεων αποτελείται από συγκεκριμένα στάδια (Σχήμα 1) με επανάληψη των ακόλουθων βημάτων [Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)]:

1. Καθαρισμός δεδομένων (απομάκρυνση θορύβου και ασυνεπή δεδομένα)
2. Ενσωμάτωση δεδομένων (υπάρχει δυνατότητα για συνδυασμό πολλών πηγών δεδομένων)
3. Επιλογή δεδομένων (όπου τα δεδομένα που σχετίζονται με την εργασία ανάλυσης ανακτώνται από τη βάση δεδομένων)
4. Μετασχηματισμός δεδομένων (όπου τα δεδομένα μετασχηματίζονται και ενοποιούνται σε μορφές κατάλληλες για εξόρυξη με τη διεξαγωγή περιληπτικών ή συγκεντρωτικών πράξεων)
5. Εξόρυξη δεδομένων (βασική διαδικασία όπου εφαρμόζονται έξυπνες μέθοδοι για την εξαγωγή μοτίβων δεδομένων)
6. Αξιολόγηση προτύπων (για τον προσδιορισμό των πραγματικά ενδιαφερόντων προτύπων που αντιπροσωπεύουν τη γνώση βάσει των μέτρων ενδιαφέροντος)
7. Παρουσίαση γνώσης (όπου χρησιμοποιούνται τεχνικές οπτικοποίησης και αναπαράστασης γνώσης για την παρουσίαση των εξορύξεων) αξιοποιήσιμη στους χρήστες)



Σχήμα 2.1. Στάδια εξόρυξης δεδομένων

Στα πρώτα τέσσερα βήματα υλοποιούνται διαφορετικές μορφές προεπεξεργασίας δεδομένων, όπου τα δεδομένα προετοιμάζονται για εξόρυξη. Στο 5ο βήμα γίνεται εφαρμογή αλγορίθμων για την εξαγωγή δεδομένων. Στα δύο τελευταία στάδια, τα ενδιαφέροντα μοτίβα παρουσιάζονται στον χρήστη για κατανόηση και αξιολόγηση (με χρήση μεθόδων οπτικοποίησης και αναπαράστασης γνώσης) και μπορούν να αποθηκευτούν ως νέες γνώσεις στη βάση γνώσεων. Η προηγούμενη προβολή είναι ουσιώδης, επειδή αποκαλύπτει κρυμμένα μοτίβα για αξιολόγηση.

Οι τύποι γνώσης που ανακαλύπτονται κατά την εξόρυξη δεδομένων είναι οι εξής:

- Κανόνες Συσχέτισης : Συσχετίζουν την ύπαρξη ενός συνόλου αντικειμένων με το εύρος τιμών ενός άλλου συνόλου μεταβλητών.
- Ιεραρχίες Ταξινόμησης : Ξεκινώντας από ένα υπάρχον σύνολο γεγονότων ή συναλλαγών γίνεται προσπάθεια να δημιουργηθεί μια ιεραρχία κλάσεων.
- Ακολουθιακά πρότυπα : Αναζητείται μια ακολουθία ενεργειών ή γεγονότων.

- Κατηγοριοποίηση και κατάτμηση - Ένα σύνολο γεγονότων ή αντικειμένων μπορεί να διαμεριστεί σε σύνολα παρόμοιων στοιχείων.
- Πρότυπα σε χρονοσειρές : Γίνεται προσπάθεια να εντοπισθούν ομοιότητες στις θέσεις μιας ακολουθίας δεδομένων που λαμβάνονται σε τακτά χρονικά διαστήματα.

2.3 ΠΗΓΕΣ ΔΕΔΟΜΕΝΩΝ

Σε αυτό το υποκεφάλαιο θα παρουσιαστούν οι βασικότερες πηγές δεδομένων.

2.3.1 ΓΕΝΙΚΑ

Η εξόρυξη δεδομένων μπορεί να εφαρμοστεί σε οποιοδήποτε είδος δεδομένων, υπό την προυπόθεση τα δεδομένα να έχουν νόημα για μια εφαρμογή-στόχο.

Οι βασικότερες μορφές δεδομένων για εφαρμογές εξόρυξης είναι :

α. δεδομένα βάσης δεδομένων (Ενότητα 2.3.2),

β. δεδομένα αποθήκης δεδομένων (Ενότητα 2.3.3).

Η εξόρυξη δεδομένων μπορεί επίσης, να εφαρμοστεί σε άλλες μορφές δεδομένων (π.χ. συναλλαγές, ροές δεδομένων, δεδομένα παραγγελίας / ακολουθίας, δεδομένα γραφήματος ή δικτύου, χωρικά δεδομένα, δεδομένα κειμένου, δεδομένα πολυμέσων και το WWW). Μια επισκόπηση αυτών των δεδομένων παρουσιάζεται στην Ενότητα 2.3.4.

Η εξόρυξη δεδομένων θα συνεχίσει σίγουρα να περιλαμβάνει νέους τύπους δεδομένων καθώς εμφανίζονται.

2.3.2 ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ

Ένα σύστημα βάσης δεδομένων, βασίζεται σε ένα Σύστημα Διαχείρισης Βάσεων Δεδομένων (DBMS), και αποτελείται από μια συλλογή αλληλένδετων δεδομένων, γνωστών ως βάση δεδομένων, και ένα σύνολο προγραμμάτων λογισμικού για τη διαχείριση και την πρόσβαση στα δεδομένα. Ο πιο ευρέως χρησιμοποιούμενος τρόπος συλλογής και αποθήκευσης δεδομένων σε ένα πληροφοριακό σύστημα είναι οι σχεσιακές βάσεις δεδομένων. Παραδείγματα δημοφιλών DBMS είναι: Mysql, Oracle, SQL Server.

Μια σχεσιακή βάση δεδομένων είναι μια συλλογή πινάκων, ο καθένας από τους οποίους έχει ένα μοναδικό όνομα. Κάθε πίνακας αποτελείται από ένα σύνολο χαρακτηριστικών (στήλες ή πεδία). Συνήθως, αποθηκεύει ένα μεγάλο σύνολο εγγραφών. Κάθε καταχώρηση σε έναν σχεσιακό πίνακα αντιπροσωπεύει ένα αντικείμενο που αναγνωρίζεται από ένα μοναδικό κλειδί και περιγράφεται από ένα σύνολο τιμών.

Τα μοντέλα σημασιολογικών δεδομένων κατασκευάζονται συνήθως για να κατανοήσουμε τα δεδομένα μας και με την εφαρμογή κανόνων οδηγούν στη σχεδίαση σχεσιακών βάσεων δεδομένων. Παράδειγμα μοντέλου είναι το Entity Relationsip Data Model (μοντέλο ER). Το μοντέλο δεδομένων ER «περιγράφει» τα δεδομένα της βάσης δεδομένων ως ένα σύνολο οντοτήτων και τις σχέσεων τους.

Η πρόσβαση σε σχεσιακά δεδομένα μπορεί να αποκτηθεί είτε χρησιμοποιώντας ερωτήματα βάσης δεδομένων γραμμένα σε σχεσιακές γλώσσες ερωτημάτων (όπως SQL) είτε με βοήθεια από γραφικές διεπαφές χρήστη. Ένα δεδομένο ερώτημα μετατρέπεται σε ένα σύνολο σχεσιακών λειτουργιών, όπως σύνδεση, επιλογή, προβολή και στη συνέχεια βελτιστοποιείται για

αποτελεσματική επεξεργασία. Ένα ερώτημα επιτρέπει την ανάκτηση συγκεκριμένων υποομάδων των δεδομένων.

2.3.3 ΑΠΟΘΗΚΕΣ ΔΕΔΟΜΕΝΩΝ

Η αποθήκη δεδομένων (Data Warehouse) είναι ένα αποθετήριο πληροφοριών που συλλέγονται από πολλές ετερογενείς πηγές (σχεσιακή βάση, αρχείο csv, log files κ.α.) και αποθηκεύονται σε ένα ενοποιημένο σχήμα.

Οι αποθήκες δεδομένων κατασκευάζονται μέσω μιας διαδικασίας καθαρισμού, ολοκλήρωσης, μετατροπής, φόρτωσης και περιοδικής ανανέωσης δεδομένων. Για την επεξεργασία τους χρησιμοποιείται η τεχνική ETL (Extract Transform Load). Η ETL είναι μια διαδικασία που εξάγει, μετασχηματίζει και φορτώνει δεδομένα από πολλές πηγές σε αποθήκη δεδομένων ή άλλο ενοποιημένο αποθετήριο δεδομένων.

Με σκοπό να διευκολυνθεί η λήψη αποφάσεων, τα δεδομένα οργανώνονται γύρω από σημαντικά θέματα (π.χ. πελάτης, αντικείμενο, προμηθευτής και δραστηριότητα) σε μια αποθήκη δεδομένων. Τα δεδομένα αποθηκεύονται, για να παρέχουν πληροφορίες από ιστορική άποψη, όπως τους τελευταίους 6 έως 12 μήνες και συνήθως συνοψίζονται.

Μια αποθήκη δεδομένων μοντελοποιείται συνήθως από μια πολυδιάστατη δομή δεδομένων, που ονομάζεται κύβος δεδομένων, στην οποία κάθε διάσταση αντιστοιχεί σε ένα χαρακτηριστικό ή ένα σύνολο χαρακτηριστικών στο σχήμα και κάθε κελί αποθηκεύει την τιμή κάποιου συνολικού μέτρου. Μια αποθήκη δεδομένων μπορεί να αποτελείται από έναν ή περισσότερους κύβους.

2.3.4 ΑΛΛΑ ΕΙΔΗ ΔΕΔΟΜΕΝΩΝ

Σε γενικές γραμμές, κάθε εγγραφή σε μια βάση δεδομένων συναλλαγών καταγράφει μια συναλλαγή, όπως αγορά ενός πελάτη, κράτηση πτήσης ή κλικ ενός χρήστη σε μια ιστοσελίδα. Μια συναλλαγή συνήθως περιλαμβάνει έναν μοναδικό αριθμό ταυτότητας συναλλαγής (transaction ID) και μια λίστα με τα στοιχεία που αποτελούν τη συναλλαγή, όπως τα είδη που αγοράστηκαν στη συναλλαγή. Μια βάση δεδομένων συναλλαγών μπορεί να έχει επιπρόσθετους πίνακες, οι οποίοι περιέχουν κι'άλλες πληροφορίες που σχετίζονται με τις συναλλαγές, όπως περιγραφή στοιχείων, πληροφορίες σχετικά με τον πωλητή ή το υποκατάστημα και ούτω καθεξής.

Εκτός από τα σχεσιακά δεδομένα βάσης δεδομένων και τα δεδομένα αποθήκης δεδομένων, υπάρχουν πολλοί άλλοι τύποι δεδομένων, οι οποίοι έχουν ευέλικτες μορφές, δομές και διαφορετικές σημασιολογικές έννοιες.

Τύποι δεδομένων σαν αυτούς συναντώνται σε πολλές εφαρμογές: δεδομένα συναλλαγών (π.χ. αγορά ενός πελάτη ή κλικ ενός χρήστη σε μια ιστοσελίδα), δεδομένα που σχετίζονται με το χρόνο ή ακολουθίες (π.χ. ιστορικά αρχεία ή δεδομένα ανταλλαγής αποθεμάτων), ροές δεδομένων (π.χ., παρακολούθηση βίντεο και δεδομένα αισθητήρων, τα οποία είναι συνεχώς μεταδιδόμενα), χωρικά δεδομένα (π.χ. χάρτες), δεδομένα σχεδιασμού μηχανικής (π.χ. σχεδιασμός κτιρίων ή ολοκληρωμένων κυκλωμάτων), δεδομένα υπερκειμένου και πολυμέσων, δικτυωμένα δεδομένα (π.χ. κοινωνικά δίκτυα και δίκτυα πληροφοριών) και το Web (ένα τεράστιο, ευρέως διανεμημένο αποθετήριο πληροφοριών που διατίθεται από το Διαδίκτυο).

Αυτές οι εφαρμογές δημιουργούν νέες προκλήσεις, όπως σχετικά με τη διαχείριση δεδομένων που μεταφέρουν ειδικές δομές (π.χ. ακολουθίες, δέντρα, γραφήματα και δίκτυα), συγκεκριμένη

σημασιολογία, όπως περιεχόμενο εικόνας, ήχου και βίντεο και συνδεσιμότητα και τον τρόπο που υλοποιούνται μοτίβα που φέρουν πλούσιες δομές και σημασιολογία.

2.4 ΠΕΔΙΑ ΕΦΑΡΜΟΓΗΣ

Σε αυτό το υποκεφάλαιο θα παρουσιαστούν κάποια από τα πεδία εφαρμογής της εξόρυξης δεδομένων.

2.4.1 ΓΕΝΙΚΑ

Στο επίκεντρο της εξόρυξης δεδομένων, από την αρχή υπήρξαν οι εφαρμογές έχοντας σημειώσει μεγάλες επιτυχίες σε πολλές περιπτώσεις. Με την εξέλιξη της εξόρυξης δεδομένων, νέοι τομείς επωφελούνται από την παραγώμενη γνώση. Για παράδειγμα, ο αθλητισμός τα τελευταία χρόνια, πέρα από τις στατιστικές αναλύσεις για την ανάλυση του παιχνιδιού του αντιπάλου λαμβάνει βιομετρικά στοιχεία από τους ίδιους τους παίκτες της ομάδας την ώρα του αγώνα, καταγράφοντας ανά πάσα στιγμή τις ενέργειες που κάνανε και την κατάσταση τους.

Είναι πρακτικά αδύνατο να παρουσιαστούν όλες οι εφαρμογές όπου η εξόρυξη δεδομένων παίζει κρίσιμο ρόλο. Παρακάτω παρουσιάζονται κάποιοι τομείς, στους οποίους η εξόρυξη δεδομένων βρίσκει εφαρμογή, όπως οι τηλεπικοινωνίες, η ιατρική, ο παγκόσμιος ιστός, η οικονομία και η εκπαίδευση.

2.4.2 ΤΗΛΕΠΙΚΟΙΝΩΝΙΕΣ

Ο τομέας των τηλεπικοινωνιών ήταν ένας από τους πρώτους που υιοθέτησε την τεχνολογία εξόρυξης δεδομένων. Ο μεγάλος όγκος από τηλεπικοινωνιακά δεδομένα, τα οποία παράγονται από την τεράστια βάση των πελατών τους, όπως ο τύπος κλήσης, η τοποθεσία του καλούντος και του κληθέντος, ο χρόνος κλήσης και η διάρκεια, αποτελούν ευκαιρία εκμετάλλευσης για την καλύτερη εξυπηρέτηση των χρηστών του δικτύου.

Σε επίπεδο απόδοσης, εφαρμόζονται τεχνικές εξόρυξης δεδομένων για την εξισορρόπηση του φορτίου του συστήματος και της κίνησης των δεδομένων. Επιπλέον, δίνεται η δυνατότητα για παραγωγή γνώσης και σε επίπεδο marketing. Με χρήση των δεδομένων μπορεί να δημιουργηθεί ένα προφίλ των πελατών. Με βάση αυτά τα προφίλ οι πελάτες μπορούν να ομαδοποιηθούν. Ανάλογα με την ομάδα στην οποία κατατάσσεται ένας υπάρχων ή νέος πελάτης, να προσφέρεται το αντίστοιχο πακέτο επικοινωνίας.

Ο έντονος ανταγωνισμός, καθώς και ο μεταβαλλόμενος χαρακτήρας της βιομηχανίας, σε συνδυασμό με την τεράστια παράγωγη δεδομένων, διασφαλίζει πως η εξόρυξη δεδομένων θα συμβάλει σημαντικά στο μέλλον του τομέα των τηλεπικοινωνιών.

2.4.3 ΙΑΤΡΙΚΗ

Με την ραγδαία ανάπτυξη κλάδων της Ιατρικής, όπως η γενετική και η βιοϊατρική, αποδεικνύεται η χρησιμότητα της εξόρυξης δεδομένων στην Ιατρική.

Στον τομέα της γενετικής, στόχος είναι η κατανόηση και η χαρτογράφηση της σχέσης μεταξύ της μεταβολής των ακολουθιών του ανθρώπινου DNA και της προδιάθεσης κάποιας ασθένειας. Η εξόρυξη δεδομένων είναι ένα ανεκτίμητο εργαλείο, το οποίο μπορεί να συμβάλλει αποτελεσματικά τόσο στη βελτίωση της διάγνωσης όσο και της πρόληψης και κατ' επέκταση στην θεραπεία ασθενειών. Ένας από τους κύριους στόχους, συνυφασμένος με την ανάλυση του DNA, είναι η σύγκριση ποικίλων ακολουθιών και η αναζήτηση ομοιοτήτων μεταξύ των δεδομένων του DNA. Η σύγκριση γίνεται μεταξύ της γονιδιακής ακολουθίας υγιών και βλαβερών ιστών, με σκοπό την εύρεση διαφορών.

Στον κλάδο της βιοϊατρικής, τα εργαλεία οπτικοποίησης διαδραματίζουν σημαίνοντα ρόλο. Τα εργαλεία αυτά παρέχουν τη δυνατότητα παρουσίασης πολύπλοκων δομών γονιδίων σε γράφους, δένδρα και αλυσίδες. Μέσω της οπτικής αναπαράστασης επιτυγχάνεται καλύτερη κατανόηση αυτών των δομών συμβάλλοντας στην εξερεύνηση των δεδομένων.

Κλασικά παραδείγματα εφαρμογής της εξόρυξης δεδομένων είναι η πρόβλεψη επιληπτικής κρίσης, μέσα από την ανάλυση δεδομένων που συλλέχθηκαν από τη μαγνητική MRI των ασθενών και η κατηγοριοποίηση καρκίνου σε καλοήθη ή κακοήθη.

Τα σύγχρονα νοσοκομεία είναι καλά εξοπλισμένα με συστήματα παρακολούθησης και άλλες συσκευές συλλογής δεδομένων ,που εξυπηρετούν στη συλλογή και αποθήκευση τους σε πληροφοριακά συστήματα.

Οι μεγάλες συλλογές ιατρικών δεδομένων αποτελούν ανεκτίμητο πόρο που δυνητικά νέες και χρήσιμες γνώσεις μπορούν να προκύψουν μέσω της εξόρυξης δεδομένων.

2.4.4 ΠΑΓΚΟΣΜΙΟΣ ΙΣΤΟΣ

Τα τελευταία χρόνια με την ραγδαία αύξηση των φορητών συσκευών, ο Παγκόσμιος Ιστός αποτελεί τον δημοφιλέστερο τρόπο επικοινωνίας και διάδοσης πληροφοριών. Καθημερινά δημιουργούνται νέες ιστοσελίδες, ηλεκτρονικά καταστήματα, ερευνητικά άρθρα, εκπαιδευτικό περιεχόμενο και λογισμικό.

Για να κατανοήσουμε το μέγεθος της και τη σημασία της συνεισφοράς της ανάπτυξης του διαδικτύου, θα πρέπει να αντιληφθούμε πως ο όγκος της πληροφορίας που υπάρχει μέχρι τώρα στο διαδίκτυο είναι αδύνατο να μετρηθεί με ακρίβεια. Οι σελίδες που είναι δημοσιευμένες στον Παγκόσμιο Ιστό ανέρχονται στα 4.2 δισεκατομμύρια, δημοσιευμένες σε 1,7 δισεκατομμύρια sites. Η μηχανή αναζήτησης της Google πραγματοποιεί 5.6 δισεκατομμύρια αναζητήσεις ημερησίως, με κάθε ερώτημα στην μηχανή αναζήτησης να μην ξεπερνά σε χρόνο τα δυο δευτερόλεπτα.

Αυτό αποτελεί τεράστια επιτυχία της εξόρυξης δεδομένων, αρχικά γιατί η αναζήτηση σε τόσο μεγάλο όγκο δεδομένων γίνεται σε πολύ σύντομο χρόνο και δεύτερον γιατί τα πρώτα αποτελέσματα σε κάθε ερώτημα κατά κανόνα είναι τα πιο χρήσιμα. Έτσι, ο χρήστης λαμβάνει γρήγορα και εύκολα μόνο της ουσιώδη πληροφορία που θέλει.

Η εξόρυξη δεδομένων από το περιεχόμενο του διαδικτύου θεωρείται μια επέκταση της λειτουργίας που εκτελείται με τις παραδοσιακές μηχανές αναζήτησης, οι οποίες βασίζονται σε λέξεις-κλειδιά. Στην εξόρυξη βασικό ρόλο παίζουν τα προγράμματα αναζήτησης ιστού (crawlers).

Η εξόρυξη δεδομένων στο διαδίκτυο μπορεί να διακριθεί σε τρεις κατηγορίες:

- Εξόρυξη από τα περιεχόμενα (Web content mining)
- Εξόρυξη από τα δεδομένα χρήσης (Web usage mining)

- Εξόρυξη από τα δεδομένα δομής (Web structure mining).

Η εξόρυξη από τα περιεχομένα του Ιστού εξετάζει το περιεχόμενο των ιστοσελίδων. Το περιεχόμενο περιλαμβάνει τόσο κείμενο όσο και γραφικά δεδομένα. Η εξόρυξη περιεχομένου μοιάζει με την διαδικασία που εκτελείται από τις βασικές τεχνικές της ανάκτησης πληροφοριών προχωρώντας όμως περισσότερο από την απλή χρήση λέξεων κλειδιών για την αναζήτηση.

Η εξόρυξη που βασίζεται στα δεδομένα χρήσης περιλαμβάνει τεχνικές και εργαλεία για την παρακολούθηση των αιτημάτων των χρηστών.

Τέλος, η εξόρυξη από τα δεδομένα δομής χρησιμοποιεί τη θεωρία γράφων για την ανάλυση μιας ιστοσελίδας, τη δομή της σύνδεσης των ιστοσελίδων μεταξύ τους, καθώς και την δομή των δένδρων για την ανάλυση και την περιγραφή του HTML και XML πηγαίου κώδικα των ιστοσελίδων.

2.4.5 ΟΙΚΟΝΟΜΙΑ

Η οικονομία αποτελεί ακόμα ένα τομέα στον οποίο οι τεχνικές εξόρυξης δεδομένων βρίσκουν εφαρμογή. Αξιόπιστα και υψηλής ποιότητας οικονομικά δεδομένα συλλέγονται κατά κύριο λόγο από τράπεζες και άλλους οικονομικούς φορείς. Η συλλογή τους έχει σαν στόχο την κατανόηση και βελτίωση δεδομένων, τη δημιουργία, την εκτίμηση και την ανάπτυξη μοντέλων.

Η ανάλυση των οικονομικών δεδομένων όμως, έχει ως στόχους τη βέλτιστη επιλογή κατά τη λήψη αποφάσεων και την πρόβλεψη τόσο κινήσεων (π.χ. μετοχές) όσο και προβλημάτων. Τυπικά παραδείγματα οικονομικών προβλημάτων αποτελούν η εκτίμηση του πιστωτικού κινδύνου, η πτώχευση εταιρειών και η πρόβλεψη της εταιρικής απόδοσης. Η συλλογή δεδομένων από

διάφορους οικονομικούς οργανισμούς, όπως οι τράπεζες, το χρηματιστήριο, φορολογικές αρχές, εξειδικευμένα γραφεία λογιστηρίων και ελεγκτών, συγκεντρώνονται σε αποθήκες δεδομένων.

Μια άλλη εφαρμογή της εξόρυξης δεδομένων σχετίζεται με την πρόβλεψη αποπληρωμής δανείων και την πολιτική πίστωσης πελατών. Προσδιορίζοντας το επίπεδο εισοδήματος του πελάτη, την κατάσταση αποπληρωμής βάσει του εισοδήματος, καθώς και το πιστωτικό ιστορικό του και άλλα χαρακτηριστικά, η τράπεζα μπορεί να καθορίσει την πολιτική δανείου βάσει ενός σχετικά χαμηλού κινδύνου. Οι τεχνικές συσταδοποίησης και ταξινόμησης χρησιμοποιούνται από τα χρηματοπιστωτικά ιδρύματα. Μέσω της ομαδοποίησης διαφορετικών πελατών με κοινά χαρακτηριστικά, συνδέουν νέους πελάτες με υπάρχουσες ομάδες και τους προσφέρουν κοινά οφέλη.

Τέλος, μια επιπλέον εφαρμογή της εξόρυξης δεδομένων στην οικονομία είναι ο εντοπισμός πιθανών απατών ή/και παραποιημένων δεδομένων. Αναλύοντας οικονομικές συναλλαγές και εξάγοντας κάποια μοτίβα, η καταγραφή ενός «ασυνήθιστου» γεγονότος μπορεί να αποτελέσει δείκτη για πιθανή παρατυπία ή σφάλμα και να ενεργοποιήσει μια διαδικασία ταυτοποίησης του πραγματικού πελάτη. Η τεχνολογία οπτικοποίησης βοηθά στην παρουσίαση δεδομένων σε διαφορετικές μορφές, όπως γραφήματα με βάση συγκεκριμένα χαρακτηριστικά. Εξετάζοντας δεδομένα από διαφορετικές οπτικές γωνίες, οι οικονομικοί οργανισμοί έχουν τη δυνατότητα να διακρίνουν τους πελάτες που προσπαθούν να διαπράξουν παράνομες πράξεις. Μάλιστα, μπορούν να βοηθήσουν στον εντοπισμό απάτης και εγκληματικών πράξεων μετά από λεπτομερείς έρευνες για αυτές τις ύποπτες περιπτώσεις.

2.4.6 ΕΚΠΑΙΔΕΥΣΗ

Τέλος, η εφαρμογή στην Εκπαίδευση (στα εκπαιδευτικά δεδομένα) της Εξόρυξη Δεδομένων αποτελεί έναν αναδυόμενο τομέα. Σκοπός είναι η ανάπτυξη μεθόδων και η ανακάλυψη γνώσεων με πρώτη ύλη τα δεδομένα από το εκπαιδευτικό περιβάλλον. Η μεγάλη διάδοση της τηλεκπαίδευσης σε παγκόσμιο επίπεδο έχει συνεισφέρει στη δημιουργία νέων δεδομένων για παραγωγή γνώσης.

Οι στόχοι του τομέα αυτού αφορούν την πρόβλεψη της μελλοντικής συμπεριφοράς των φοιτούντων και επιπλέον αφορούν προβλέψεις για τις επιδόσεις των μαθητών. Μελετώνται δεδομένα και αποτελέσματα της εκπαιδευτικής διαδικασίας, και στοιχεία για την υποστήριξη και προώθηση της επιστημονικής γνώσης που σχετίζεται με τη μάθηση.

Αυτή η γνώση αποτελεί χρήσιμο εργαλείο για τους εκπαιδευτικούς οργανισμούς όσον αφορά τη λήψη αποφάσεων σχετικά με το εκπαιδευτικό υλικό αλλά και με τις εκπαιδευτικές μεθόδους. Έτσι, με την αξιοποίηση της νεοαποκτηθείσας γνώσης δίνεται η δυνατότητα για ανάπτυξη νέων εκπαιδευτικών τεχνικών.

2.5 ΕΡΓΑΛΕΙΑ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Σε αυτό το υποκεφάλαιο θα παρουσιαστούν κάποια από τα πιο διαδεδομένα εργαλεία εξόρυξης δεδομένων.

2.5.1 RAPIDMINER

Το RapidMiner είναι ένα εργαλείο ανάλυσης εξόρυξης δεδομένων που χρησιμοποιείται για την ανάλυση δεδομένων και την υποστήριξη διαφόρων τεχνικών εξόρυξης δεδομένων (Hofmann & Klinkenberg, 2013).

Χρησιμοποιείται σε πλήθος εφαρμογών στη βιομηχανία, στην έρευνα, στην εκπαίδευση και στην ανάπτυξη εφαρμογών. Περιέχει πληθώρα προγραμμάτων μάθησης για την ομαδοποίηση, την ταξινόμηση και την ανάλυση και παράγει visualizations και reports. Σημαντικό πλεονέκτημα αποτελεί το γεγονός πως δεν απαιτείται η συγγραφή κώδικα.

Επιπλέον, υποστηρίζει σχεδόν όλες τις τιμές δεδομένων, πράγμα που σημαίνει ότι οι χρήστες μπορούν να εισαγάγουν πληροφορίες από μια ποικιλία πηγών δεδομένων, προς εξέταση και ανάλυση εντός της εφαρμογής. Χρησιμοποιείται σε πληθώρα βιομηχανιών, όπως στις επικοινωνίες, στην ενέργεια, στα οικονομία κ.α.

Το RapidMiner προσφέρει ανανεώσιμη μονοετή εκπαιδευτική άδεια μέσω του εκπαιδευτικού προγράμματος του (RapidMiner Educational License Program) για φοιτητές και καθηγητές.

Εικόνα 2.1. Λογότυπο rapidminer

2.5.2 ORANGE

Ένα διαδεδομένο εργαλείο εξόρυξης δεδομένων για προγραμματιστές Python είναι το Orange, ένα ισχυρό εργαλείο ανοιχτού κώδικα.

Η Python είναι μια γενική γλώσσα προγραμματισμού υψηλού επιπέδου και χρησιμοποιείται ευρέως σε όλους τους επιστημονικούς κλάδους, όπως γενικός προγραμματισμός, ανάπτυξη ιστού, ανάπτυξη λογισμικού, ανάλυση δεδομένων, μηχανική μάθηση κ.λπ. Η Python προσφέρει έναν τεράστιο αριθμό βιβλιοθηκών επέκτασης, εκ των οποίων αρκετές σχετίζονται με τη μηχανική μάθηση.

Το Orange σχεδιάστηκε στα τέλη της δεκαετίας του 1990 και είναι από τα παλαιότερα τέτοια εργαλεία. Επικεντρώνεται στην απλότητα και τη διαδραστικότητα μέσω scripting και τη σχεδίαση βάσει στοιχείων (components). Παρέχει πληθώρα λύσεων σχετικά με τις οπτικοποιήσεις δεδομένων και διαθέτει εργαλεία που ονομάζονται widgets και βοηθούν στην αποκάλυψη κρυφών δεδομένων. Οι χρήστες έχουν την δυνατότητα να χρησιμοποιήσουν το Orange ως βιβλιοθήκη της Python σχετικά με το χειρισμό των δεδομένων και την τροποποίηση των widget.

Το Orange διανέμεται δωρεάν υπό την άδεια GPL και υποστηρίζεται με πολλά online σεμινάρια. Η ανάπτυξη και η συντήρηση του υλοποιείται στο Εργαστήριο Βϊοπληροφορικής του Τμήματος Ηλεκτρονικών Υπολογιστών και πληροφορικής, στο Πανεπιστήμιο της Λιουμπλιάνα της Σλοβενίας.

Εικόνα 2.2. Λογότυπο orange

2.5.3 R

Η R αποτελεί τόσο μια γλώσσα προγραμματισμού όσο και ένα περιβάλλον λογισμικού. Χρησιμοποιείται ευρέως, για στατιστικό υπολογισμό, για τη δημιουργία γραφικών αναπαραστάσεων και για επεξεργασία και ανάλυση δεδομένων στη διαδικασία εξόρυξης δεδομένων. Εκτός από την εξόρυξη δεδομένων, παρέχει τεχνικές γραφικών και στατιστικής γραμμικών και μη γραμμικών μοντέλων, ταξινόμηση, κλασικούς στατιστικούς ελέγχους, ανάλυση χρονοσειρών, ομαδοποίηση και άλλα.

Η υλοποίηση της έχει βασιστεί στη γλώσσα προγραμματισμού S, η οποία αναπτύχθηκε από τον John Chambers, στο Bell Labs. Οι Ross Ihaka και Robert Gentleman από το Πανεπιστήμιο του Ώκλαντ της Νέας Ζηλανδίας δημιούργησαν την R.

Τα τελευταία χρόνια η δημοφιλία της αυξάνεται διαρκώς για αρκετούς λόγους, όπως η ευκολία στην εκμάθηση της, η συμβατότητα της με τα πιο ευρέως χρησιμοποιούμενα λειτουργικά συστήματα (Linux, Mac OS και Windows), ο μεγάλος αριθμός έτοιμων πακέτων, τα καλογραμμένα εγχειρίδια χρήστη και φυσικά το γεγονός πως είναι δωρεάν.

Εικόνα 2.3. Λογότυπο R

2.5.4 WEKA

Το WEKA (Waikato Environment for Knowledge Analysis) είναι μια σουίτα λογισμικού για μηχανική μάθηση και Εξόρυξη Δεδομένων. Η ανάπτυξη του υλοποιήθηκε στο Πανεπιστήμιο του Waikato της Ν. Ζηλανδίας.

Είναι γραμμένο σε java και διανέμεται δωρεάν υπό την άδεια GNU General Public License επιτρέποντας στους χρήστες να χρησιμοποιούν ελεύθερα και να τροποποιούν το λογισμικό. Είναι ένα από τα πιο δημοφιλή προγράμματα εξόρυξης δεδομένων. Έχει χρησιμοποιηθεί σε μεγάλο αριθμό επιστημονικών εργασιών και πολλά βιβλία εξόρυξης δεδομένων αναφέρονται σε αυτό.

Η δημοφιλία του οφείλεται στα ειδικά χαρακτηριστικά του και στις δυνατότητες που προσφέρει. Συγκεκριμένα:

- Περιέχει αρκετές μεθόδους ταξινόμησης, παλινδρόμησης, ανάλυσης συστάδων και κανόνων συσχέτισης. Παρέχει επίσης, λειτουργίες προεπεξεργασίας δεδομένων και εργαλεία οπτικοποίησης.

- Είναι λογισμικό ανοιχτού κώδικα. Αυτό σημαίνει ότι ο πηγαίος κώδικας είναι διαθέσιμος στο κοινό. Οι χρήστες με γνώσεις προγραμματισμού μπορούν να τροποποιήσουν και να αναπτύξουν αλγόριθμους.
- Είναι γραμμένο σε Java, και το γεγονός αυτό επιτρέπει να εγκατασταθεί σε διαφορετικές πλατφόρμες υλικού και λογισμικού.
- Έχει μια γραφική διεπαφή χρήστη που επιτρέπει στους τελικούς χρήστες χωρίς γνώση προγραμματισμού να χρησιμοποιούν το λογισμικό.

Το WEKA διατίθεται σε δύο διαφορετικές εκδόσεις· μια που απευθύνεται σε χρήστες και μια σε προγραμματιστές.

Εικόνα 2.4. Λογότυπο WEKA

3. ΗΛΕΚΤΡΟΝΙΚΗ ΜΑΘΗΣΗ

3.1 ΕΙΣΑΓΩΓΗ

Η ανάπτυξη και η εξέλιξη του διαδικτύου, των δικτύων και των τηλεπικοινωνιών, τα τελευταία χρόνια προσφέρουν καινούριες δυνατότητες. Καθημερινά αναπτύσσονται νέες τεχνολογίες στη βιομηχανία, στις επιστήμες και στην εκπαίδευση. Οι νέες τεχνολογίες έχουν συμβάλει καθοριστικά προς όφελος της εκπαιδευτικής διαδικασίας, με την τηλεκπαίδευση να αποτελεί ένα από τα κύρια νέα μέσα.

3.2 E-LEARNING

Σήμερα, σε παγκόσμιο επίπεδο γίνεται έρευνα και ανάπτυξη στο τομέα της ηλεκτρονικής μάθησης (e-learning), με σκοπό την παροχή ίσων ευκαιριών μάθησης σε όλους. Αυτό επιβάλλει τη χρήση τόσο τεχνολογιών όσο και ψηφιακών μέσων στην εκπαιδευτική διαδικασία, χωρίς χρονικές και χωρικές δεσμεύσεις.

Η τηλεκπαίδευση - e-learning είναι η διαδικασία εκμάθησης η οποία διαφοροποιείται από την απλή μάθηση στα εξής:

α. Η εκπαίδευση εκτελείται μέσα από σύγχρονες τεχνολογίες (π.χ. υπολογιστής με χρήση διαδικτύου κ.α.)

β. Παρέχεται βοήθεια και υποστήριξη από τον εκπαιδευτή με διάφορους τρόπους

Η δυνατότητα αυτής της επικοινωνίας με τον εκπαιδευτή διαφοροποιεί το e-learning από την εκπαίδευση μέσω ενός e-book ή ενός CD. Βασικός πυλώνας του e-learning είναι η αλληλεπίδραση

μεταξύ μαθητών και μαθητών-εκπαιδευτών, όπως συμβαίνει στη δια ζώσης εκπαίδευση, στο πλαίσιο όμως μιας εικονικής τάξης. Η διαδικασία της διδασκαλίας μπορεί να υλοποιηθεί τόσο με σύγχρονο όσο με ασύγχρονο τρόπο.

Στην διδασκαλία με ασύγχρονο τρόπο, οι εκπαιδευόμενοι έχουν τη δυνατότητα να εργαστούν με το υλικό του μαθήματος απομακρυσμένα και στον χρόνο της επιλογής τους. Επιπλέον, έχουν τη δυνατότητα (ασύγχρονης) επικοινωνίας και ανταλλαγής απόψεων τόσο μεταξύ τους, όσο και με τους διδάσκοντες. Κάποια από τα εργαλεία που χρησιμοποιούνται για ασύγχρονη εκπαίδευση είναι η πλατφόρμα Open eClass και το web-based σύστημα διαχείρισης μαθημάτων Moodle.

Αντίθετα, στη σύγχρονη διδασκαλία, καθένας από τους συμμετέχοντες συμμετέχει από το χώρο του, μέσω διαδικτύου, σε μια εικονική τάξη με χρήση οπτικοακουστικού εξοπλισμού. Κατά τη διάρκεια της διδασκαλίας, συνηθίζεται να παρέχεται κάποια υπηρεσία ανταλλαγής μηνυμάτων. Στη σύγχρονη επικοινωνία χρησιμοποιούνται συνήθως πλατφόρμες τηλεδιάσκεψης. Παρέχονται πολλές λύσεις τόσο επί πληρωμή (π.χ. MS Teams) όσο και δωρεάν (π.χ. BigBlueButton, Jitsi).

Φυσικά, το e-learning χρησιμοποιείται σε διάφορους τομείς, εκτός της εκπαίδευσης. Ενδεικτικά αναφέρονται ιδιωτικές εταιρείες, Επιμελητήρια (π.χ. βιομηχανίες, βιοτεχνίες, κλπ.) και Δημόσιοι Φορείς.

3.3 OPEN ECLASS

Η πλατφόρμα Open eClass (http://www.openeclass.org/) είναι ένα ολοκληρωμένο Σύστημα Διαχείρισης Ηλεκτρονικών Μαθημάτων για την ηλεκτρονική οργάνωση, αποθήκευση και παρουσίαση του εκπαιδευτικού υλικού. Βασίζεται στη φιλοσοφία του λογισμικού ανοικτού

κώδικα και διανέμεται ελεύθερα. Το Ακαδημαϊκό Διαδίκτυο GUNET έχει την ευθύνη για το σχεδιασμό, την υποστήριξη, τη διάθεση και τη συντήρηση της πλατφόρμας.

Βασική επιδίωξη του Open eClass είναι η ενσωμάτωση των νέων τεχνολογιών και η εποικοδομητική χρήση του Διαδικτύου στην εκπαιδευτική διαδικασία. Η πλατφόρμα έχει τη δυνατότητα να προσαρμοστεί σε όλες τις οθόνες (H/Y, tablet, smartphone). Επίσης διατίθεται και ως εφαρμογή για φορητές συσκευές.

Είναι συμβατή με διεθνή πρότυπα ηλεκτρονικής μάθησης (SCORM, IMSCP). Με αυτό τον τρόπο εξασφαλίζεται η επαναχρησιμοποίηση, η προσβασιμότητα, η ανθεκτικότητα του εκπαιδευτικού υλικού στις τεχνολογικές μεταβολές, καθώς και η διαλειτουργικότητα μεταξύ των συστημάτων ηλεκτρονικής μάθησης. Έχει πολλές δυνατότητες, όπως δημιουργία απεριόριστων μαθημάτων, αναφορές επίδοσης, στατιστικά, ενώ υποστηρίζει σύγχρονη και ασύγχρονη επικοινωνία και δημιουργία αντιγράφων ασφαλείας. Χρησιμοποιείται από πλήθος ακαδημαϊκών ιδρυμάτων και εκπαιδευτικών οργανισμών με συμμετοχή χιλιάδων μαθητών/φοιτητών.

Εικόνα 3.1. Λογότυπο eClass

3.4 MICROSOFT TEAMS

Το Microsoft Teams είναι πλατφόρμα επικοινωνίας που παρέχει η Microsoft επί πληρωμή. Το Teams αποτελεί μέρος του Microsoft Office 365. Προσφέρει δυνατότητες συνομιλίας και τηλεδιάσκεψης, αποθήκευσης εγγράφων και ενοποίηση με άλλες εφαρμογές. Παρέχει δυνατότητες όπως κοινή χρήση οθόνης, ανταλλαγή μηνυμάτων, δημιουργία ομάδας, τηλεδιασκέψεις, ψηφιακό πίνακα ελεύθερης γραφής κ.ά. Καθ 'όλη την πανδημία του κοροναϊού, η πλατφόρμα άρχισε να χρησιμοποιείται ευρέως, καθώς αρκετές συναντήσεις έχουν πλέον μετακινηθεί σε εικονικό περιβάλλον.

Εικόνα 3.2. Λογότυπο MS Teams

4. ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

4.1 ΕΙΣΑΓΩΓΗ

Η μηχανική μάθηση αποτελεί ένα υποπεδίο της επιστήμης των υπολογιστών. Αναπτύχθηκε μέσω της μελέτης της αναγνώρισης προτύπων και της υπολογιστικής θεωρίας μάθησης στην τεχνητή νοημοσύνη. Η μηχανική μάθηση διερευνά τον τρόπο με τον οποίο οι υπολογιστές μαθαίνουν (ή βελτιώνουν την απόδοσή τους) με βάση τα δεδομένα.

Ένας βασικός τομέας έρευνας είναι πως τα προγράμματα υπολογιστών θα μάθουν να αναγνωρίζουν αυτόματα πολύπλοκα μοτίβα και πως θα λαμβάνουν έξυπνες αποφάσεις βάσει δεδομένων. Για παράδειγμα, ένα τυπικό πρόβλημα μηχανικής μάθησης είναι να προγραμματιστεί ένας υπολογιστής, έτσι ώστε να αναγνωρίζει αυτόματα χειρόγραφους ταχυδρομικούς κώδικες σε μηνύματα ηλεκτρονικού ταχυδρομείου μετά την εκμάθηση μέσα από ένα σύνολο παραδειγμάτων. Στην εξόρυξη δεδομένων υπάρχει μεγάλη ποικιλία μεθόδων, οι οποίες κατηγοριοποιούνται σύμφωνα με κάποια χαρακτηριστικά τους, όπως παρουσιάζονται παρακάτω στο κεφάλαιο.

Πλέον ο όρος Επιστήμη των Δεδομένων (Data Science) έχει έρθει, για να αντικαταστήσει προγενέστερους όρους όπως Εξόρυξη Δεδομένων ή Ανακάλυψη Γνώσης από Βάσεις Δεδομένων. Αυτοί οι τρείς όροι περιγράφουν μία ημι-αυτοματοποιημένη διαδικασία, σκοπός της οποίας είναι η ανάλυση μεγάλου όγκου δεδομένων που σχετίζονται με ένα συγκεκριμένο πρόβλημα, συνήθως εμπορικού ή επιστημονικού ενδιαφέροντος και χρησιμοποιούνται για τη δημιουργία προτύπων σε τομείς όπως η Στατιστική, η Μηχανική Μάθηση και η Αναγνώριση Προτύπων.

Αξίζει να σημειωθεί πως στην Επιστήμη των Δεδομένων η ειδοποιός διαφορά βρίσκεται στο σημείο της εκπαίδευσης, στο οποίο γίνεται με τη χρήση δεδομένων. Αντίθετα, σε άλλες μορφές εκπαίδευσης χρησιμοποιείται ένας δάσκαλος ή κάποιος ειδικός μεταφέροντας τη γνώση.

4.2 ΕΙΔΗ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Η μηχανική μάθηση είναι ένα ραγδαία εξελισσόμενο θέμα. Υπάρχουν διάφορα είδη μηχανικής μάθησης, με τα κυριότερα να είναι :

- Επιβλεπόμενη (supervised)
- Μη επιβλεπόμενη (unsupervised)
- Ημι-εποπτευόμενη (semi-supervised)
- Ενεργή (active)

Σχήμα 4.1. Είδη μηχανικής μάθησης

4.2.1 ΕΠΙΒΛΕΠΟΜΕΝΗ

Η επιβλεπόμενη μάθηση είναι βασικά συνώνυμη με την ταξινόμηση. Η επίβλεψη της μάθησης προέρχεται από ένα σύνολο επισημασμένων παραδειγμάτων στο σύνολο δεδομένων εκπαίδευσης (dataset). Το σύνολο δεδομένων περιέχει τις αντίστοιχες ομάδες (κλάσεις) κάθε εγγραφής. Ο στόχος είναι να δημιουργηθεί ένα μοντέλο που, όταν λαμβάνει τα νέα δεδομένα, να μπορεί να τα ταξινομεί σε μια από τις προϋπάρχουσες τάξεις.

Για παράδειγμα, στο πρόβλημα αναγνώρισης ταχυδρομικού κώδικα, ένα σύνολο χειρόγραφων εικόνων ταχυδρομικού κώδικα και αντίστοιχων μηχαναγνώσιμων μεταφράσεων χρησιμοποιούνται ως δείγματα εκπαίδευσης για την επίβλεψη της μάθησης του μοντέλου ταξινόμησης.

4.2.2 ΜΗ ΕΠΙΒΛΕΠΟΜΕΝΗ

Η μη επιβλεπόμενη μάθηση είναι ουσιαστικά συνώνυμη με την ομαδοποίηση. Η διαδικασία εκμάθησης δεν παρακολουθείται, επειδή τα εισαγωγικά παραδείγματα δεν επισημαίνονται ως τάξεις (κλάσεις). Οπότε, έχουμε δεδομένα χωρίς να γνωρίζουμε σε ποια τάξη ανήκουν. Συνήθως, χρησιμοποιούνται συστάδες, για να ανακαλυφθούν τάξεις στα δεδομένα. Στόχος είναι η ανάλυση αυτών των δεδομένων, προκειμένου να ανακαλύψουμε κάποια ενδεχομένως ενδιαφέροντα στοιχεία στα δεδομένα.

Για παράδειγμα, η μη εποπτευόμενη μέθοδος μάθησης μπορεί να εισαγάγει ένα σύνολο χειρόγραφων ψηφιακών εικόνων. Ας υποθέσουμε ότι βρίσκει 10 σύνολα δεδομένων. Αυτές οι

ομάδες μπορούν να αντιστοιχούν σε 10 μεμονωμένους αριθμούς από 0 έως 9. Ωστόσο, δεδομένου ότι τα δεδομένα εκπαίδευσης δεν φέρουν ετικέτα, το μοντέλο εκπαίδευσης δεν μπορεί να μας «περιγράψει» τη σημασιολογική σημασία των ανακαλυφθέντων ομάδων.

4.2.3 ΗΜΙ-ΕΠΙΒΛΕΠΟΜΕΝΗ ΜΑΘΗΣΗ

Η ημι-επιβλεπόμενη μάθηση είναι ένας υβριδικός τύπος τεχνικής μηχανικής μάθησης που χρησιμοποιεί τόσο επισημασμένα παραδείγματα όσο μη, όταν μαθαίνει ένα μοντέλο. Στο στάδιο της «προπόνησης» συνδυάζει μια μικρή ποσότητα δεδομένων με ετικέτα με μεγάλη ποσότητα δεδομένων χωρίς ετικέτα.

Σε μια προσέγγιση, επισημασμένα παραδείγματα χρησιμοποιούνται, για να μάθουν το μοντέλο τάξης και μη επισημασμένα παραδείγματα χρησιμοποιούνται για τη βελτίωση των ορίων τάξης. Για προβλήματα με δύο τάξεις, μπορούν να θεωρηθούν όλα τα παραδείγματα που ανήκουν σε μια κατηγορία ως θετικά παραδείγματα και όλα τα παραδείγματα που ανήκουν στην άλλη κατηγορία ως αρνητικά παραδείγματα.

4.2.4 ΕΝΕΡΓΗ

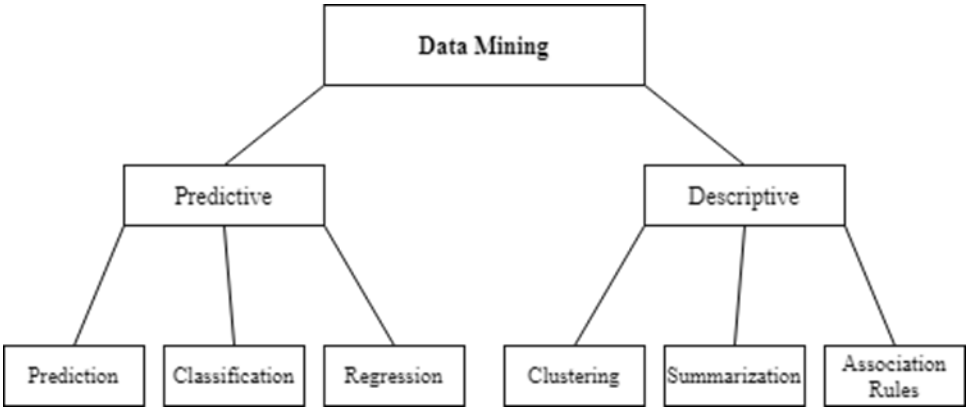
Η ενεργή εκμάθηση είναι μια μέθοδος μηχανικής μάθησης που επιτρέπει στους χρήστες να διαδραματίσουν ενεργό ρόλο στη διαδικασία μάθησης. Αυτή η μέθοδος ενδέχεται να απαιτήσει από τον χρήστη να επισημάνει ένα παράδειγμα, το οποίο μπορεί να προέρχεται από ένα σύνολο παραδειγμάτων χωρίς ετικέτα ή να συντίθενται από ένα πρόγραμμα εκμάθησης. Λαμβάνοντας υπόψη τον περιορισμό του αριθμού των παραδειγμάτων που μπορούν να ζητηθούν, ο στόχος είναι

η βελτιστοποίηση της ποιότητας του μοντέλου μέσω της ενεργητικής απόκτησης γνώσεων από ανθρώπους.

4.3 ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΜΕΘΟΔΩΝ

Υπάρχουν διάφορες μέθοδοι εξόρυξης δεδομένων. Ανάλογα με τον τύπο των δεδομένων και τον τύπο των γνώσεων που εξάγονται, χωρίζονται σε διαφορετικές κατηγορίες. Οι βασικότερες μέθοδοι εξόρυξης δεδομένων είναι:

- Συσταδοποίηση (clustering)
- Κατηγοριοποίηση (classification)
- Παλινδρόμηση (regression)
- Συσχέτιση κανόνων (association rules)



Σχήμα 4.2. Μέθοδοι μηχανικής μάθησης

4.3.1 ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

Η κατηγοριοποίηση ή ταξινόμηση (classification) είναι η διαδικασία εύρεσης μοντέλων (ή συναρτήσεων) που περιγράφουν και διακρίνουν κατηγορίες δεδομένων ή έννοιες. Πρόκειται για προγνωστική μέθοδο και αποτελεί μια από τις πιο κοινές τεχνικές εξόρυξης δεδομένων.

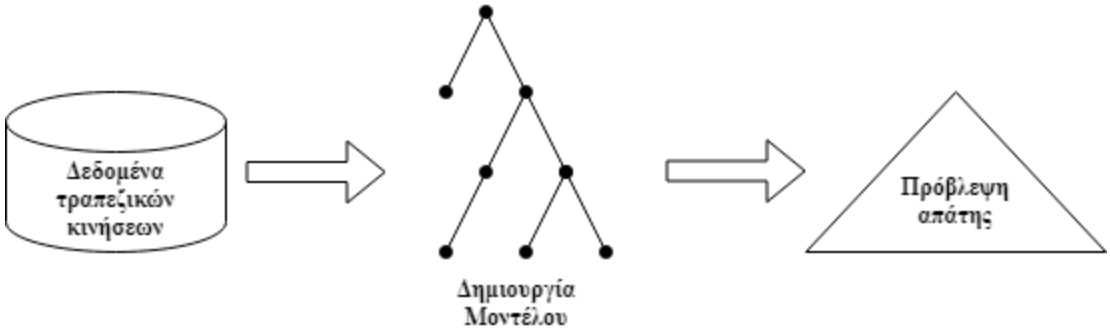
Η διαδικασία στοχεύει στη δημιουργία ενός μοντέλου γνωστό ως κατηγοριοποιητής (classifier). Το μοντέλο δημιουργείται με βάση την ανάλυση ενός συνόλου δεδομένων εκπαίδευσης (δηλαδή αντικειμένων δεδομένων με γνωστές ετικέτες τάξης). Αυτό το μοντέλο χρησιμοποιείται για την πρόβλεψη της ετικέτας κλάσης ενός αντικειμένου, του οποίου η ετικέτα κλάσης είναι άγνωστη. Ουσιαστικά, η λειτουργία της διαδικασίας είναι η κατασκευή ενός μοντέλου, που να μπορεί να εφαρμοστεί στα μη ταξινομημένα δεδομένα, ώστε να τα ταξινομήσει.

Τα παράγωγα μοντέλα μπορούν να εκφραστούν σε διάφορες μορφές, όπως κανόνες ταξινόμησης (π.χ. κανόνες IF-THEN), δέντρα αποφάσεων, μαθηματικοί τύποι ή νευρωνικά δίκτυα. Ένα δέντρο αποφάσεων είναι μια δομή δέντρου που μοιάζει με ένα διάγραμμα ροής, στον οποίο κάθε κόμβος αντιπροσωπεύει μια δοκιμή σε μια τιμή χαρακτηριστικού, κάθε κλάδος αντιπροσωπεύει ένα αποτέλεσμα δοκιμής και τα φύλλα αντιπροσωπεύουν μια κατανομή τάξης. Τα δέντρα αποφάσεων μπορούν εύκολα να μετατραπούν σε κανόνες ταξινόμησης.

Όταν χρησιμοποιείται για ταξινόμηση, ένα νευρικό δίκτυο είναι συνήθως μια συλλογή μονάδων επεξεργασίας που μοιάζουν με νευρώνες με σταθμισμένες συνδέσεις μεταξύ των μονάδων. Υπάρχουν πολλοί τρόποι για τη δημιουργία ενός μοντέλου ταξινόμησης, όπως η ταξινόμηση naïve Bayes, η ταξινόμηση k-πλησιέστερου γείτονα κ.ά..

Η κατηγοριοποίηση συχνά συγχέεται με την πρόβλεψη. Στην πρώτη μέθοδο το αποτέλεσμα που είναι επιθυμητό, είναι να προβλεφθεί η κλάση των δειγμάτων. Αυτή μπορεί να πάρει διαφορετικές τιμές από ένα περιορισμένο σετ. Αντίθετα, όπως θα φανεί στη συνέχεια, όταν χρησιμοποιείται η παλινδρόμηση για προβλέψεις, η μεταβλητή στόχος μπορεί να είναι οποιοσδήποτε πραγματικός αριθμός.

Η κατηγοριοποίηση ασχολείται με διακριτά αποτελέσματα (π.χ. ναι ή όχι, ήλιος ή βροχή). Κάποια ενδεικτικά παραδείγματα είναι η επιλογή του περιεχομένου που θα εμφανίζεται σε μια ιστοσελίδα, η πρόβλεψη περίπτωσης απάτης σε τραπεζικές συναλλαγές και ο εντοπισμός spam e-mails.



Σχήμα 4.3. Παράδειγμα κατηγοριοποίησης για πρόβλεψη απάτης

4.3.2 ΠΑΛΙΝΔΡΟΜΗΣΗ

Η παλινδρόμηση (regression) είναι μια διαδικασία που χρησιμοποιείται για την πρόβλεψη αριθμητικών δεδομένων που λείπουν ή δεν είναι διαθέσιμα και όχι για την διακριτή ετικέτα κλάσης, όπως συμβαίνει στην ταξινόμηση. Πρόκειται για μια ακόμη προγνωστική μέθοδο. Σκοπός της διαδικασίας είναι να εκπαιδεύσει μια συνάρτηση που απεικονίζει ένα αντικείμενο σε μια

πραγματική μεταβλητή. Κάποιες ανεξάρτητες μεταβλητές χρησιμοποιούνται για την πρόβλεψη των τιμών μιας εξαρτημένης μεταβλητής.

Η ανάλυση παλινδρόμησης είναι μια στατιστική μέθοδος που χρησιμοποιείται συχνότερα για αριθμητική πρόβλεψη, αν και υπάρχουν και άλλες μέθοδοι. Επίσης, περιλαμβάνει τον προσδιορισμό των τάσεων διανομής με βάση τα διαθέσιμα δεδομένα. Υπάρχει περίπτωση να απαιτηθεί ταξινόμηση και παλινδρόμηση πριν από την ανάλυση συνάφειας, η οποία επιχειρεί να εντοπίσει χαρακτηριστικά που σχετίζονται σημαντικά με τη διαδικασία ταξινόμησης και παλινδρόμησης. Αυτά τα χαρακτηριστικά θα επιλεγούν για τη διαδικασία κατάταξης και παλινδρόμησης, ενώ άλλα χαρακτηριστικά που πιθανόν να μην σχετίζονται μπορούν να αποκλειστούν από την εξέταση.

Η εφαρμογή της προϋποθέτει ότι τα σχετικά δεδομένα ταιριάζουν με κάποια γνωστά είδη συναρτήσεων (π.χ. γραμμική, πολυωνυμική κ.ά.) Κατόπιν, γίνεται καθορισμός της συνάρτησης που μοντελοποιεί τα δεδομένα με βέλτιστο τρόπο. Η κύρια διαφορά της παλινδρόμησης με την κατηγοριοποίηση, όπως αναφέραμε είναι πως παίρνει συνεχείς τιμές.

Διάσημο παράδειγμα παλινδρόμησης αποτελεί η εκτίμηση τιμής πώλησης μιας κατοικίας σύμφωνα με τα χαρακτηριστικά του, όπως τετραγωνικά μέτρα, περιοχή, δωμάτια κ.ά..

Σχήμα 4.4. Παράδειγμα πρόβλεψης πωλήσεων με χρήση παλινδρόμησης

4.3.3 ΣΥΣΤΑΔΟΠΟΙΗΣΗ

Η συσταδοποίηση (clustering) έχοντας ένα σύνολο δεδομένων, έχει ως στόχο της τη δημιουργία συστάδων (clusters), δηλαδή ομάδων, οι οποίες θα περιέχουν όμοια ή παρεμφερή δείγματα. Είναι μια περιγραφική μέθοδος. Αντίθετα με όσα έχουν αναφερθεί ως τώρα στην κατηγοριοποίηση και την παλινδρόμηση, η συσταδοποίηση αναλύει αντικείμενα δεδομένων χωρίς ετικέτες τάξης. Στην συσταδοποίηση δεν υπάρχουν προκαθορισμένες κατηγορίες και παραδείγματα.

Τα αντικείμενα ομαδοποιούνται με βάση την αρχή της μεγιστοποίησης της ομοιότητας εντός της κλάσης και της ελαχιστοποίησης της ομοιότητας μεταξύ κλάσεων. Οι συστάδες αντικειμένων σχηματίζονται έτσι, ώστε τα αντικείμενα μέσα σε ένα σύμπλεγμα να έχουν υψηλή ομοιότητα σε σύγκριση μεταξύ τους, αλλά και να είναι μάλλον ανόμοια με τα αντικείμενα σε άλλες συστάδες. Έτσι, κάθε σύμπλεγμα που σχηματίζεται μπορεί να θεωρηθεί ως μια κατηγορία αντικειμένων, από την οποία μπορούν να προκύψουν κανόνες.

Οι κατηγορίες που θα προκύψουν είναι πιθανό να είναι αμοιβαία αποκλειόμενες και εξαντλητικές ή να έχουν μία πιο σύνθετη αναπαράσταση(π.χ. ιεραρχικές και επικαλυπτόμενες). Η συσταδοποίηση μπορεί επίσης να διευκολύνει τον σχηματισμό ταξινόμησης, δηλαδή την οργάνωση των παρατηρήσεων σε μια ιεραρχία τάξεων που ομαδοποιούν παρόμοια γεγονότα.

Παραδείγματα εφαρμογής της τεχνικής της συσταδοποίησης είναι η ομαδοποίηση μετοχών με παρόμοια διακύμανση τιμών, οι εικόνες, τα κείμενα κ.ά..

Σχήμα 4.5. Παράδειγμα ομαδοποίησης καταναλωτών

4.3.4 ΣΥΣΧΕΤΙΣΗ

Η τεχνική της εξαγωγής κανόνων συσχέτισης (Mining Association Rules) θεωρείται μία από τις πιο σημαντικές διαδικασίες στην Εξόρυξη Δεδομένων. Προκάλεσε ιδιαίτερο ενδιαφέρον, επειδή οι σχετικοί κανόνες παρέχουν ένα συνοπτικό τρόπο έκφρασης δυνητικά χρήσιμων πληροφοριών, που είναι εύκολα κατανοητές από τον τελικό χρήστη.

Η διαδικασία της εξαγωγής κανόνων συσχέτισης στοχεύει στην ανακάλυψη προτύπων που περιγράφουν σημαντικές αλληλεξαρτήσεις μεταξύ των διαφόρων πεδίων - χαρακτηριστικών ενός συνόλου δεδομένων. Οι κανόνες συσχέτισης ανακαλύπτουν κρυφούς συσχετισμούς μεταξύ των χαρακτηριστικών του συνόλου δεδομένων. Οι συσχετισμοί εκφράζονται με τη μορφή $A \rightarrow B$, όπου τα A και B είναι σύνολα, τα οποία αναφέρονται στα χαρακτηριστικά του συνόλου δεδομένων προς ανάλυση. Δεδομένης της εμφάνισης του χαρακτηριστικού του συνόλου A, ο κανόνας συσχέτισης ενισχύει την πρόβλεψη εμφάνισης των χαρακτηριστικών του συνόλου B.

Ένα χαρακτηριστικό παράδειγμα είναι ο εντοπισμός συσχετίσεων στις αγοραστικές συνήθειες των καταναλωτών. Συγκεντρώνοντας όλη την πληροφορία σχετικά με τις αγορές των πελατών μιας εταιρείας αλλά και των προϊόντων, μπορεί να αυξηθούν οι πωλήσεις. Αυτό συμβαίνει γιατί μέσω συσχετίσεων γίνεται να προταθούν προϊόντα, τα οποία συγκεντρώνουν μεγάλη πιθανότητα πώλησης στον συγκεκριμένο πελάτη.

Σχήμα 4.6. Παράδειγμα στρατηγικής marketing με χρήση κανόνων συσχέτισης

4.4 ΑΛΓΟΡΙΘΜΟΙ

Σε αυτό το υποκεφάλαιο παρουσιάζονται οι αλγόριθμοι οι οποίοι, χρησιμοποιήθηκαν για την εξόρυξη δεδομένων στο πειραματικό μέρος της εργασίας.

4.4.1 K-MEANS

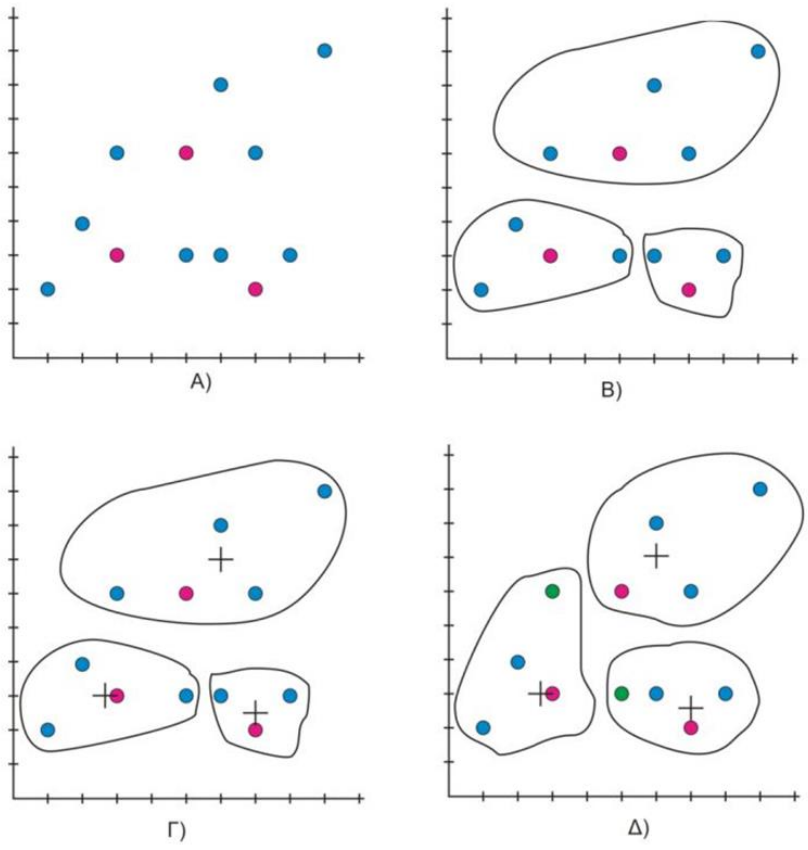
Ο αλγόριθμος k-means είναι ένας αλγόριθμος συσταδοποίησης. Ο αλγόριθμος εκκινεί με k σημεία, τα οποία δίνονται από τον χρήστη και υποδηλώνει τον αριθμό των ομάδων που θα δημιουργηθούν. Τα σημεία αυτά καλούνται κεντροειδή της συστάδας, αποτελούν αντιπροσωπευτικό δείγμα της ομάδας και δηλώνουν το κέντρο βάρους της.

Ο αλγόριθμος K-means αποτελείται από τα παρακάτω βήματα:

1. Αρχικά ο χρήστης καθορίζει τον αριθμό των clusters (k).
2. Επιλέγονται τυχαία K αντικείμενα, τα οποία αποτελούν και τα πρώτα κέντρα των clusters.
3. Στην συνέχεια κάθε αντικείμενο εντάσσεται στο cluster, του οποίου το κέντρο είναι πλησιέστερα του.
4. Για τον υπολογισμό της απόστασης συνήθως χρησιμοποιείται κάποια μετρική (π.χ. Ευκλείδεια απόσταση, απόσταση Manhattan).
5. Κάθε φορά που εντάσσεται ένα νέο αντικείμενο τα κέντρα επαναϋπολογίζονται.

Μεγάλο πλεονέκτημα του k-means αποτελεί το γεγονός πως απαιτείται λιγότερος χρόνος σε σχέση με τις ιεραρχικές μεθόδους.

Η αρχικοποίηση των κέντρων των συστάδων είναι παράγοντας μέγιστης σημασίας για την αποτελεσματικότητα του k-means. Παρότι το συγκεκριμένο βήμα φαίνεται απλό και ασήμαντο εκ πρώτης όψεως, μια «λανθασμένη» αρχικοποίηση μπορεί να οδηγήσει σε κακής ποιότητας συστάδες στην πορεία. Ένα ακόμη μειονέκτημα του αλγόριθμου είναι πως δεν υπάρχει αυτοματοποιημένος τρόπος επιλογής του αριθμού των clusters. Η σωστή επιλογή απαιτεί εμπειρία και γνώση του αντικειμένου και μελέτη μέσα από πειραματισμούς με χρήση εργαλείων οπτικοποίησης. Τα σύγχρονα εργαλεία προσφέρουν κάποια εργαλεία για αυτόματο υπολογισμό των clusters (π.χ. στο RapidMiner επιλογή X-means), χωρίς όμως να επιστρέφει τα βέλτιστα αποτελέσματα σε κάθε περίπτωση.



Εικόνα 4.1. Παράδειγμα δημιουργίας cluster

4.4.2 ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ

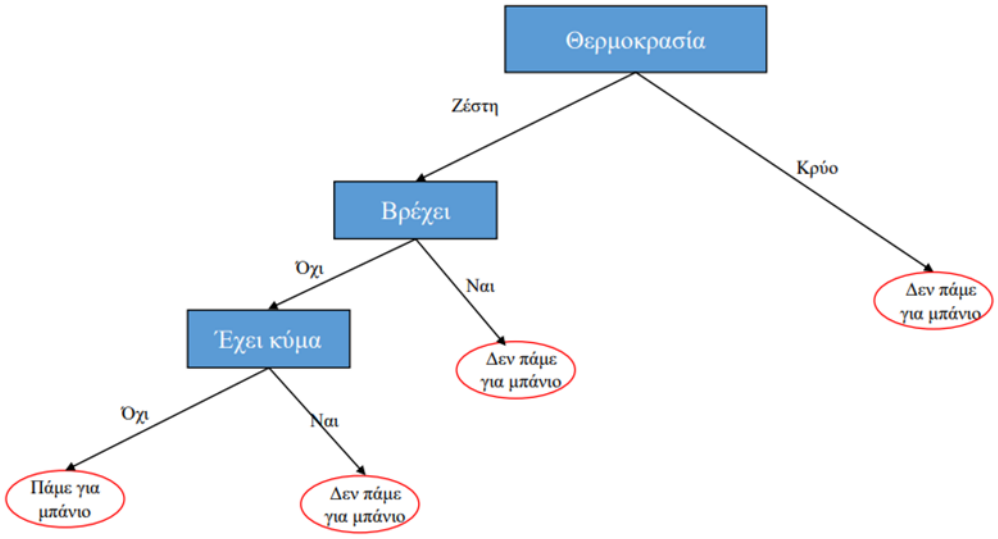
Ένα από τα πρώτα και δημοφιλέστερα μοντέλα κατηγοριοποίησης είναι τα δέντρα απόφασης. Η δημοφιλία τους οφείλεται στο γεγονός ότι πρόκειται για μια μια απλή μορφή αναπαράστασης κανόνων, ευρέως διαδεδομένων και ευκόλα κατανοητών από τον άνθρωπο.

Τα δέντρα απόφασης αναπαριστούν ένα μοντέλο πρόβλεψης, το οποίο χτίζεται μέσα από μια σειρά Boolean αποφάσεων του τύπου ΝΑΙ/ΟΧΙ, μεγαλύτερο/μικρότερο κ.τ.λ.. Ένα δέντρο αποτελείται από ενδιάμεσους κόμβους και φύλλα. Οι κόμβοι στο τελευταίο επίπεδο, οι οποίοι δεν έχουν παιδιά, ονομάζονται εσωτερικοί. Τα δέντρα αποφάσεων αναπαρίστανται ως εξής:

- κάθε εσωτερικός κόμβος αντιστοιχεί σε ένα χαρακτηριστικό
- κάθε σύνδεση μεταξύ κόμβων αντιστοιχεί σε μια συνθήκη ή τιμή για το χαρακτηριστικό του γονικού κόμβου
- κάθε φύλλο λαμβάνει το όνομα μιας κλάσης

Λαμβάνοντας αποφάσεις σύμφωνα με τις σύμφωνα με τις παραμέτρους οδηγούμαστε σε ένα συμπέρασμα. Τα βασικά στάδια ενός αλγορίθμου δέντρου αποφάσεων είναι τα παρακάτω:

- Επιλογή του κατάλληλου χαρακτηριστικού, προκειμένου να αποτελέσει την κορυφή του δέντρου.
- Επανάληψη της διαδικασίας στο επόμενο επίπεδο για κάθε παιδί του κόμβου που προκύπτει.
- Η διαδικασία τερματίζει όταν συμβεί κάτι από τα παρακάτω:
  α. Όλες οι εγγραφές καταλήγουν στην ίδια παράμετρο.
  β. Δεν υπάρχουν υποσύνολα που περιέχουν περισσότερες από μία κατηγορίες
  γ. Έχουν χρησιμοποιηθεί όλα τα χαρακτηριστικά



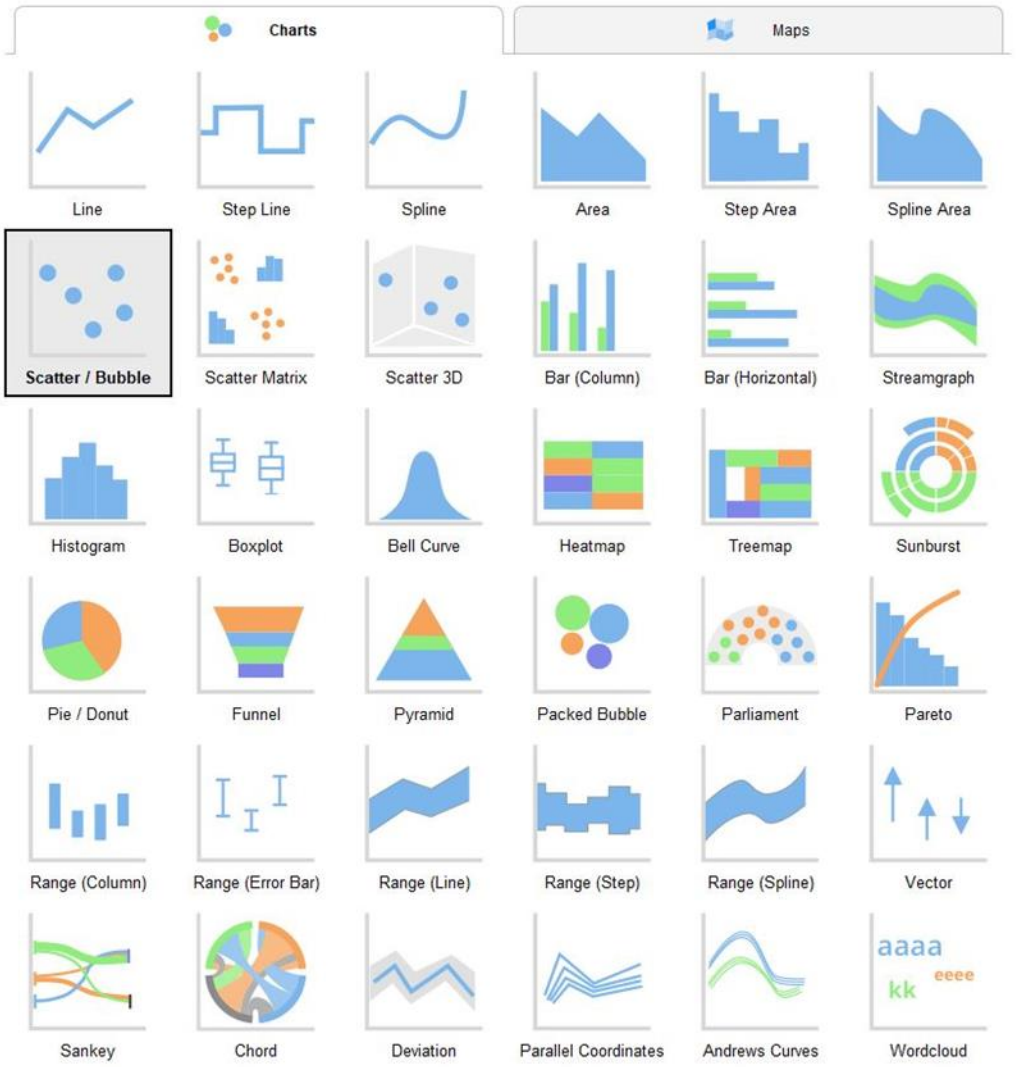
Εικόνα 4.2. Παράδειγμα δημιουργίας δέντρου απόφασης

4.5 ΟΠΤΙΚΟΠΟΙΗΣΗ

Με τον όρο οπτικοποίηση, σχετικά με τα δεδομένα, γίνεται αναφορά στη χρήση γραφικών, κίνησης, τρισδιάστατης απεικόνισης και άλλων εργαλείων πολυμέσων για την αναπαράσταση δεδομένων. Οι επιστήμονες χρησιμοποιούν απλές εφαρμογές οπτικοποίησης, για πολλά χρόνια (ραβδόγραμμα, γραφήματα διασποράς, γραφήματα πίτας κ.λ.π.) με σκοπό να εμφανίσουν τα δεδομένα της δουλειάς τους. Όμως με τη βοήθεια νέων τεχνολογιών, αρχών οπτικοποίησης και δυναμικών εφαρμογών και μεγάλων ποσοτήτων δεδομένων δίνεται η δυνατότητα δημιουργίας πιο προηγμένων οπτικοποιήσεων.

Η οπτικοποίηση δεδομένων συνήθως βοηθά όχι μόνο στην καλύτερη κατανόηση των ίδιων των δεδομένων, αλλά και στην καλύτερη κατανόηση των πιθανών συσχετίσεων μεταξύ τους. Όμως, η οπτικοποίηση είναι διαθεσιμη μόνο για συγκεκριμένο αριθμό διαστάσεων. Επομένως, για σύνολα

δεδομένων με πολλά χαρακτηριστικά μπορούμε να έχουμε οπτικοποιημένο μόνο ένα μέρος τους. Τέλος, είναι απαραίτητο η διαδικασία να συνοδεύεται από αντίστοιχους στατιστικούς ελέγχους για τη διασφάλιση της εγκυρότητας της.



Εικόνα 4.3. Εργαλεία οπτικοποίησης του λογισμικού RapidMiner

5. ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΜΕ ΧΡΗΣΗ ΤΟΥ ΕΡΓΑΛΕΙΟΥ RAPIDMINER

5.1. ΕΙΣΑΓΩΓΗ ΣΤΟ ΠΕΙΡΑΜΑΤΙΚΟ ΜΕΡΟΣ

Για τις ανάγκες της παρούσας εργασίας χρησιμοποιήθηκε πακέτο ανωνυμοποιημένων δεδομένων από το προπτυχιακό μάθημα «Βάσεις Δεδομένων» του τμήματος Μηχανικών Πληροφορικής και Υπολογιστών του Πανεπιστημίου Δυτικής Αττικής.

Για τη διαδικασία της εξόρυξης δεδμένων χρησιμοποιήθηκε το εργαλείο RapidMiner. Τα δεδομένα επεξεργάστηκαν με χρήση του αλγόριθμου K-means και με την τεχνική των δέντρων απόφασης. Επιπλέον, χρησιμοποιήθηκε και το εργαλείο (τελεστής, operator) Correlation Matrix του RapidMiner για την εξαγωγή χρήσιμων συμπερασμάτων.

5.2. ΔΟΜΗ ΑΡΧΕΙΟΥ ΔΕΔΟΜΕΝΩΝ

Το αρχείο δεδομένων, το οποίο περιέχει 256 εγγραφές, χρησιμοποιήθηκε για το πειραματικό μέρος της εργασίας και αποτελείται από τα παρακάτω χαρακτηριστικά:

- ID : το id που αποτελεί μοναδικό χαρακτηριστικό κάθε εγγραφής.
- Rows : ο αριθμός συνδέσεων (login) στην πλατφόρμα του eClass.
- AccessNum : ο αριθμός επισκέψεων διαφόρων ενοτήτων εντός του μαθήματος.
- AccessTime : ο συνολικός χρόνος σύνδεσης στην πλατφόρμα του eClass σε δευτερόλεπτα.

- Meetings : ο αριθμός συμμετοχών σε τηλεδιάσκεψη του MS Teams. Ο αριθμός αυξάνει μόνο κατά την πρώτη σύνδεση στην τηλεδιάσκεψη του μαθήματος.
- Sessions : ο αριθμός συνδέσεων σε τηλεδιάσκεψη του MS Teams. Ο αριθμός αυξάνει με κάθε σύνδεση στην τηλεδιάσκεψη του μαθήματος, δηλαδή αυξάνει και με τις επανασυνδέσεις.
- Duration : ο συνολικός χρόνος σύνδεσης σε τηλεδιασκέψεις του μαθήματος στην πλατφόρμα MS Teams σε δευτερόλεπτα.
- GradeLab : ο βαθμός του εργαστηριακού μέρους του μαθήματος. Οι τιμές που λαμβάνει το συγκεκριμένο χαρακτηριστικό είναι από 0 έως και 3. Ο λόγος που συμβαίνει αυτό είναι πως το εργαστηριακό μέρος συμβάλλει στον τελικό βαθμό του μαθήματος (GradeFinal) με ποσοστό 30%.
- GradeCourse : ο βαθμός του θεωρητικού μέρους του μαθήματος. Οι τιμές που λαμβάνει το συγκεκριμένο χαρακτηριστικό είναι από 0 έως και 7, για τον ίδιο λόγο με το εργαστηριακό.
- GradeFinal : ο τελικός βαθμός του μαθήματος. Ο βαθμός προκύπτει από το άθροισμα των δύο μερών του μαθήματος.

Για την παρούσα εργασία δεν υπήρξε ανάγκη για μετασχηματισμό ή καθαρισμό δεδομένων εφόσον δεν υπήρχαν κενές ή λανθασμένες τιμές. Το dataset μετασχηματίστηκε σε μορφή csv από xls, προκειμένου να εισαχθεί στο πρόγραμμα.

Παρατηρούμε επίσης πως το dataset, πλην του χαρακτηριστικού ID, περιέχει τρεις υποομάδες εντός του. Η πρώτη αποτελείται από τα χαρακτηριστικά Rows, AccessNum και AccessTime και αφορά το ασύγχρονο τμήμα της εκπαίδευσης μέσω της πλατφόρμας του eClass. Η δεύτερη αποτελείται από τα χαρακτηριστικά Meetings, Sessions και Duration και αφορά το σύγχρονο

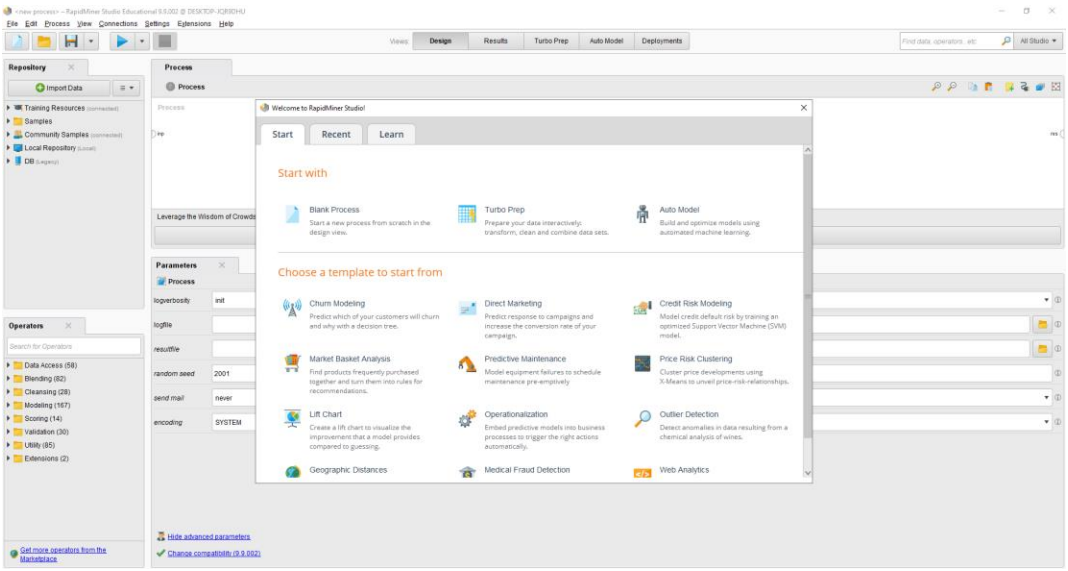
τμήμα της εκπαίδευσης μέσω της πλατφόρμας MS Teams. Τέλος, τα χαρακτηριστικά GradeLab, GradeCourse και GradeFinal παρουσιάζουν τη βαθμολογία του μαθήματος.

ID	Rows	AccessNum	AccessTime	Meetings	Sessions	Duration	GradeLab	GradeCourse	GradeFinal
1	1	1	900	0	0	0	4	6	10
2	16	44	11972	1	1	4268	0	5	5
3	3	8	2729	0	0	0	0	5	5
4	13	50	6867	2	4	3952	0	0	0
5	26	66	12978	0	0	0	0	0	0
6	2	2	901	0	0	0	0	0	0
7	168	729	135735	0	0	0	2.8	4.7	7.5
8	42	123	18769	0	0	0	0	0	0
9	32	87	14691	1	1	25	0	0	0
10	6	6	2816	0	0	0	0	0	0

Εικόνα 5.1. Τμήμα του αρχείου δεδομένων

5.3. ΤΟ ΛΟΓΙΣΜΙΚΟ RAPIDMINER

Όπως αναφέρθηκε και στο δεύτερο κεφάλαιο, το εργαλείο ανάλυσης και εξόρυξης δεδομένων RapidMiner χρησιμοποιείται για την ανάλυση δεδομένων και την υποστήριξη διαφόρων τεχνικών εξόρυξης δεδομένων. Επιπλέον, προσφέρει πληθώρα προγραμμάτων μάθησης για την ομαδοποίηση, την ταξινόμηση και την ανάλυση και παράγει visualizations και reports, χωρίς να απαιτείται η συγγραφή κώδικα. Κατά την έναρξη του λογισμικού εμφανίζεται η οθόνη:



Εικόνα 5.2. Αρχική οθόνη του rapidminer

Βασικότερες προβολές του χώρου εργασίας του περιβάλλοντος είναι η προβολή σχεδίασης (Design) και η προβολή αποτελεσμάτων (Results). Στην προβολή σχεδίασης καθορίζονται τα δεδομένα εισόδου, τα βήματα της ανάλυσης και η έξοδος των αποτελεσμάτων. Στην προβολή αποτελεσμάτων διατίθεται μια σειρά από εργαλεία διερεύνησης των αποτελεσμάτων μέσω πολλαπλών οπτικοποιήσεων (γραφήματα, πίνακες κ.α.).

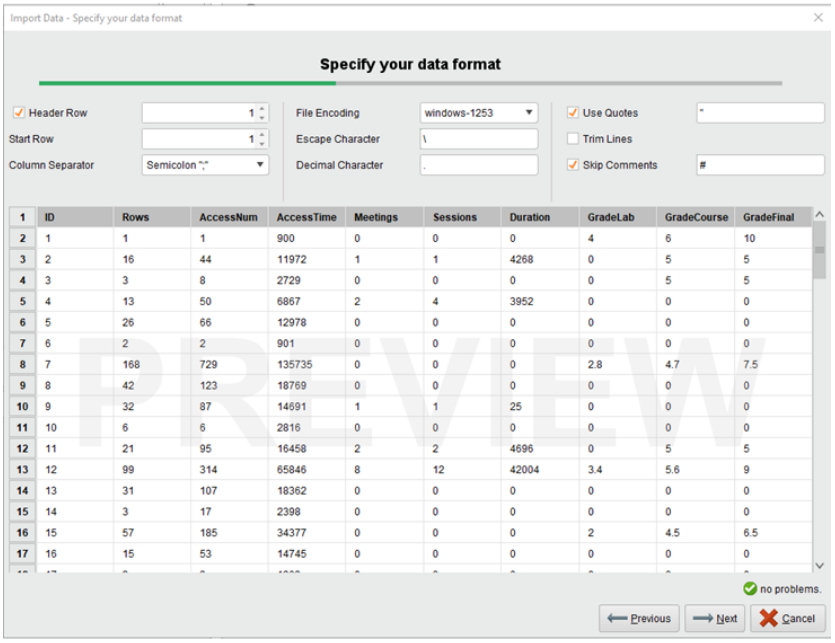
Τα βασικά στοιχεία του περιβάλλοντος του είναι η περιοχή τελεστών (Operators) και η περιοχή διαδικασίας (Process). Στην περιοχή τελεστών υπάρχουν διάφορα αντικείμενα, τα οποία έχουνε είσοδο/εισόδους και έξοδο/εξόδους και συνδέονται μεταξύ τους μέσω γραμμών. Οι τελεστές είναι αναπαραστάσεις αλγόριθμων, φίλτρα, εργαλεία κ.α. και μεταφέρονται στην περιοχή διαδικασίας με drag and drop. Στην περιοχή διαδικασίας τα δεδομένα δίνονται ως είσοδος και συνδέονται με διάφορους operators, με σκοπό την παραγωγή αποτελεσμάτων. Οι πιο πολλοί τελεστές είναι παραμετροποιήσιμοι, έτσι ώστε να είναι προσαρμόσιμοι σε κάθε πρόβλημα για την εύρεση βέλτιστης λύσης.

5.4. ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

Στο πειραματικό μέρος της εργασίας χρησιμοποιήθηκαν φίλτρα σχετικά με τις εγγραφές, παρ'ότι δεν ήταν απαραίτητο, αφού τα δεδομένα δεν χρειάζονταν μετασχηματισμό. Κρίθηκε σκόπιμο όμως, για λόγους πληρότητας, να προστεθούν φίλτρα (π.χ. για κενές τιμές πεδίων) και να γίνει κανονικοποίηση των δεδομένων με σκοπό την ευκολότερη εξαγωγή συμπερασμάτων.

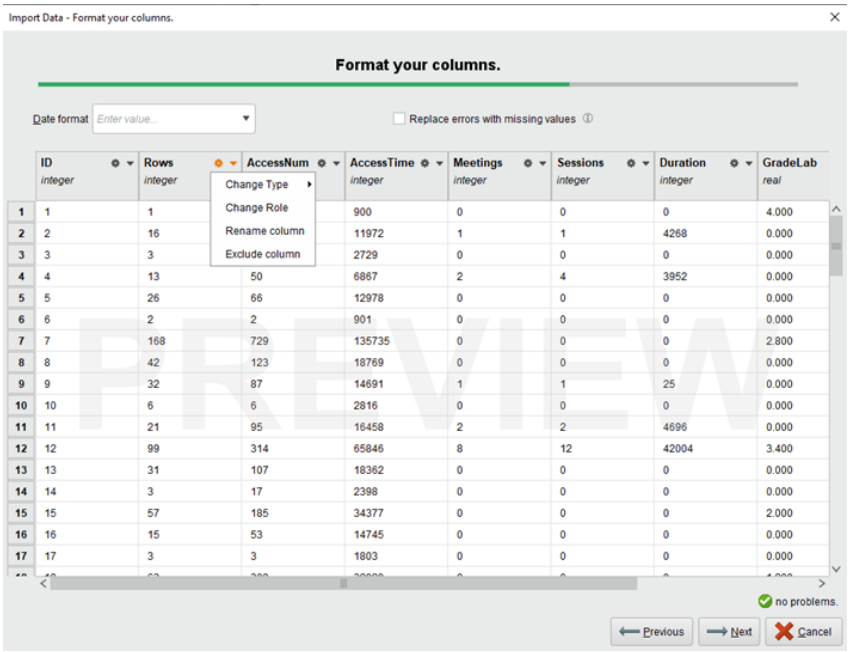
5.4.1. ΕΙΣΑΓΩΓΗ ΔΕΔΟΜΕΝΩΝ

Η εισαγωγή των εκπαιδευτικών δεδομένων από το αρχείο csv γίνεται από το αρχικό περιβάλλον εργασίας επιλέγοντας File -> Import Data. Μέσω της επιλογής browse, εντοπίζουμε το αρχείο και το επιλέγουμε. Στη συνέχεια, εμφανίζεται παράθυρο καθορισμού της μορφής των δεδομένων, με επιλογές σχετικά με την σωστή είσαγωγή των δεδομένων (από ποια γραμμή ξεκινούν τα δεδομένα, ποιο σύμβολο χρησιμοποιείται σαν διαχωριστής γραμμών κ.τ.λ.), όπως φαίνεται στην ακόλουθη εικόνα:



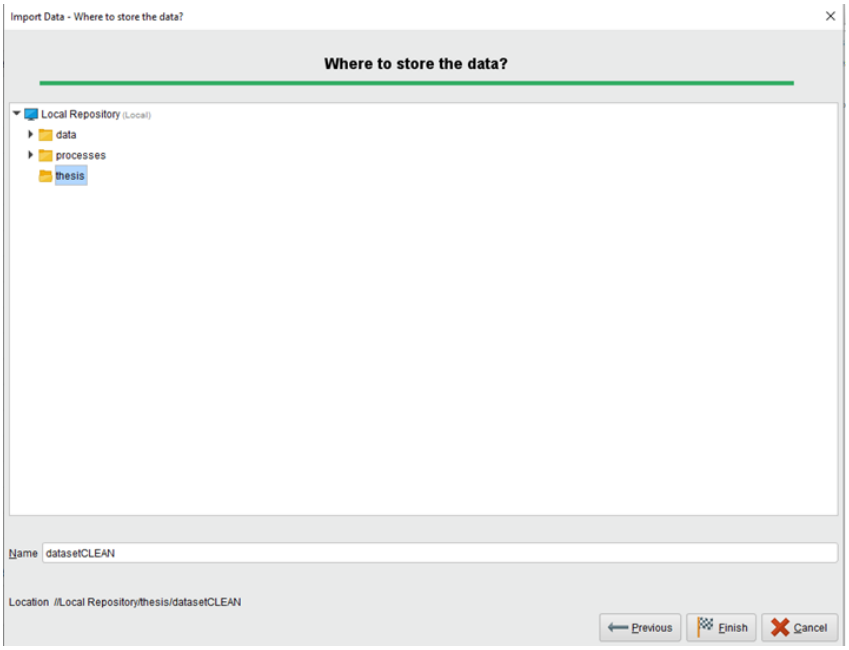
Εικόνα 5.3. Οθόνη καθορισμού μορφής εισαγωγής δεδομένων

Πατώντας Next εμφανίζεται παράθυρο, το οποίο δίνει τη δυνατότητα μορφοποίησης των στηλών (αλλαγή τύπου δεδομένων, αλλαγή ονομασίας κ.α.).



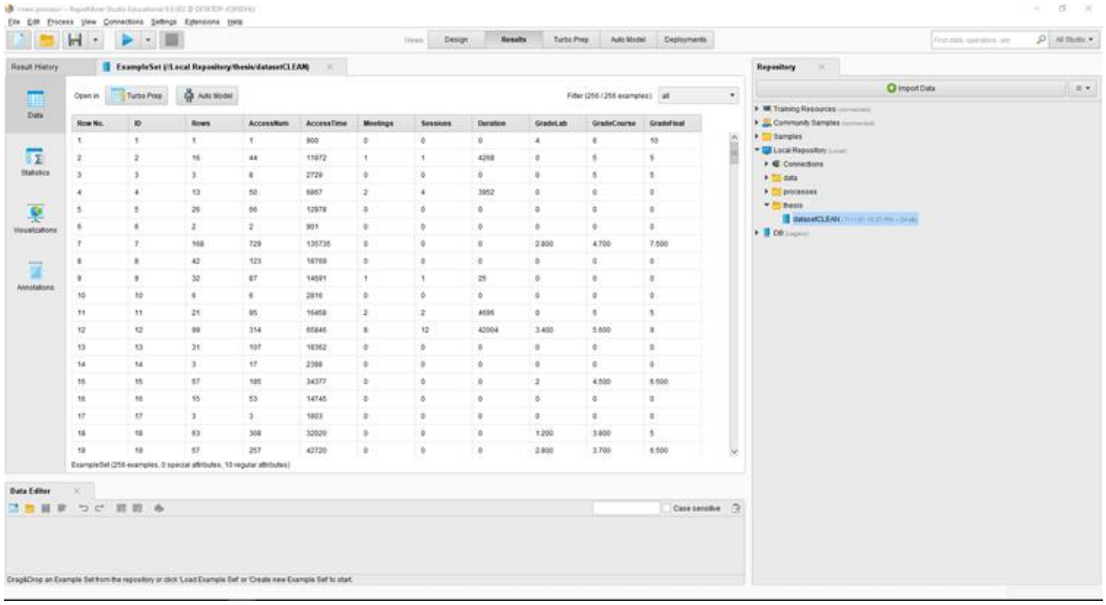
Εικόνα 5.4. Οθόνη καθορισμού μορφής εισαγωγής δεδομένων (2)

Τέλος, επιλέγεται η τοποθεσία αποθήκευσης των δεδομένων και επιλέγεται το Finish.



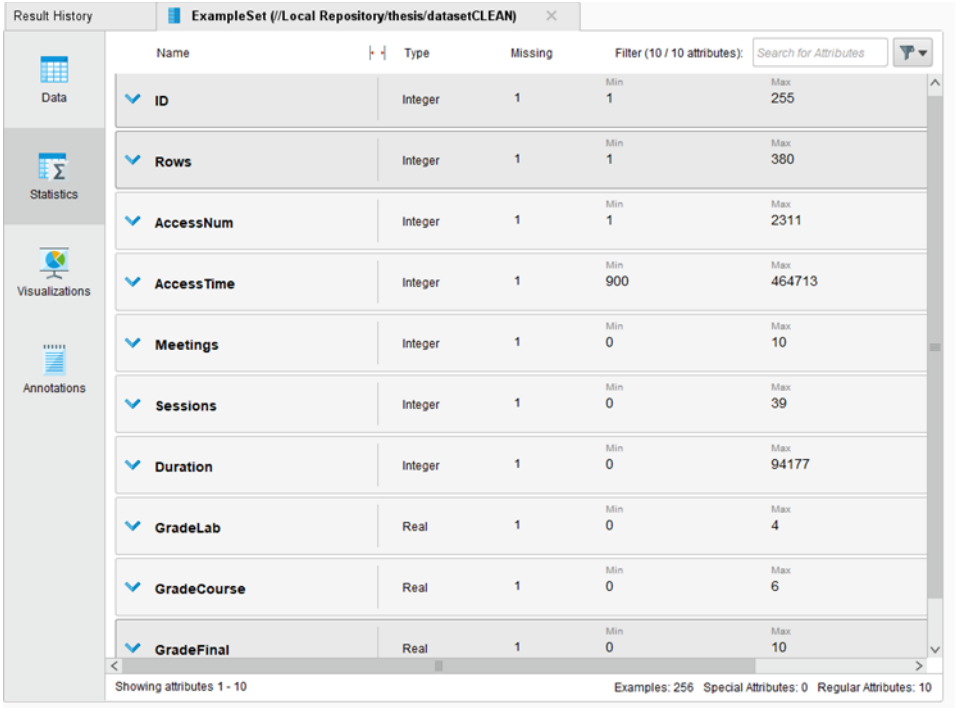
Εικόνα 5.5. Οθόνη αποθήκευσης δεδομένων

Τα δεδομένα είναι πλέον διαθέσιμα για επεξεργασία, όπως φαίνεται από την προβολή αποτελεσμάτων που ανοίγει.



Εικόνα 5.6. Οθόνη προβολής δεδομένων

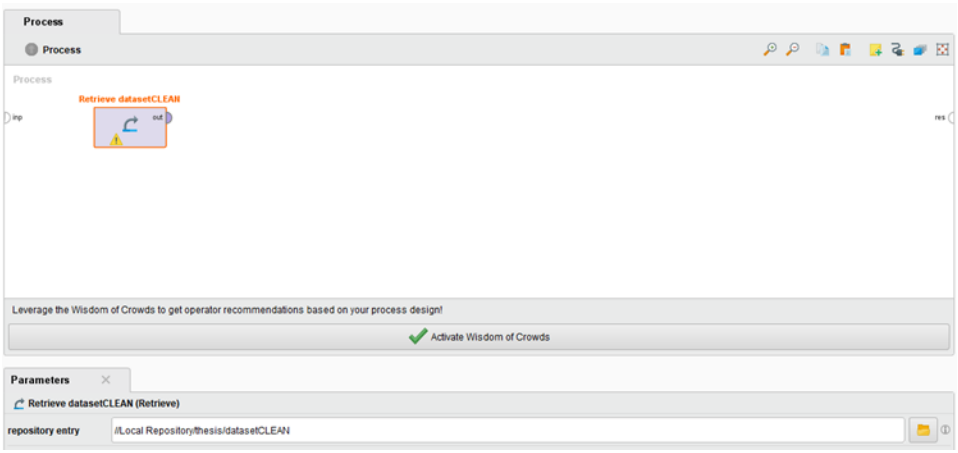
Άξια αναφοράς είναι η καρτέλα των στατιστικών, η οποία από προεπιλογή παρέχει πληροφορίες σχετικά με το σύνολο των εγγραφών, το σύνολο και τον τύπο των χαρακτηριστικών (attributes) και τις μέγιστες και ελάχιστες τιμές τους.



Εικόνα 5.7. Οθόνη προβολής στατιστικών δεδομένων

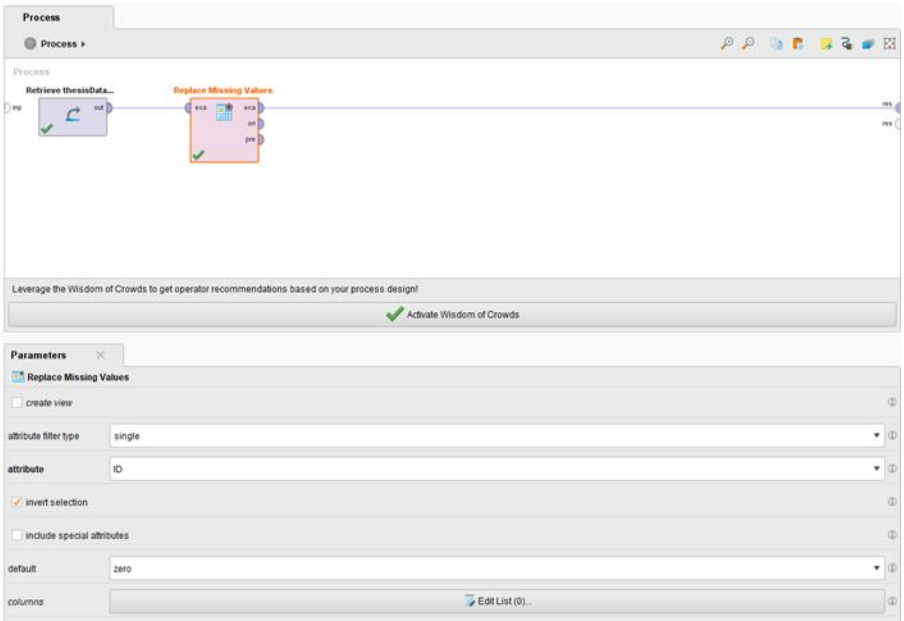
5.4.2. ΕΦΑΡΜΟΓΗ K-MEANS

Με τα δεδομένα έτοιμα προς χρήση, έπεται η διαδικασία μοντελοποίησης. Η διαδικασία υλοποιείται μέσω επιλογής του κατάλληλου αντικειμένου και της μεταφοράς του μέσω drag and drop στην περιοχή διαδικασίας. Αρχικά, εισάγονται τα δεδομένα από την περιοχή του αποθετηρίου, όπως φαίνεται στην ακόλουθη εικόνα:



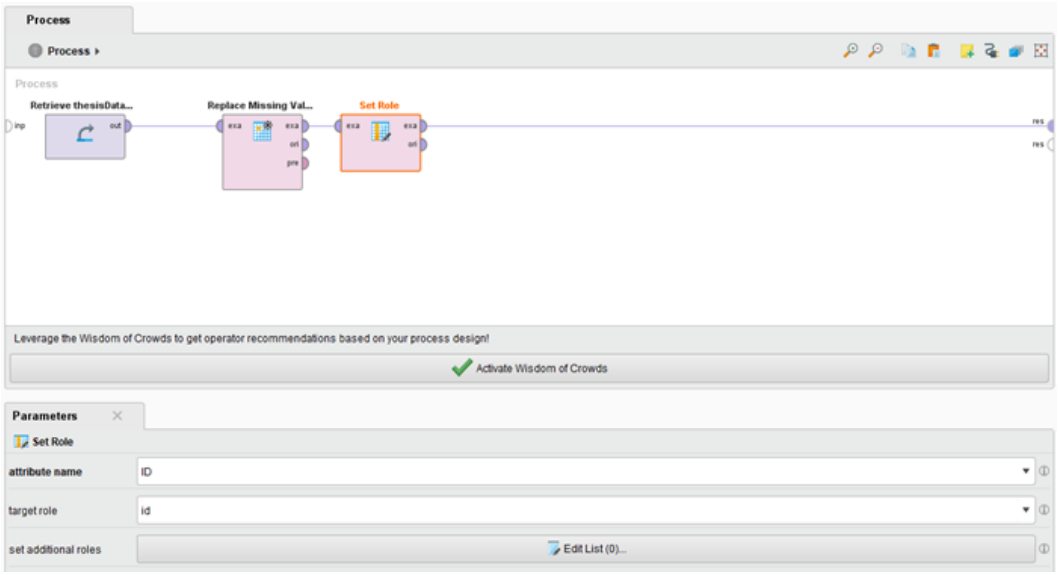
Εικόνα 5.8. Εισαγωγή dataset στο rapidminer

Επόμενο βήμα είναι η προσθήκη αντικειμένου, του οποίου η λειτουργία είναι να μηδενίζει τα κενά πεδία των χαρακτηριστικών. Η επιλογή αυτή έγινε επειδή το κενό πεδίο σημαίνει πως δεν υπήρξε καμία ενέργεια από μέρους του συμμετέχοντα. Επίσης, ο μηδενισμός δεν εφαρμόζεται στο ID. Επιλέγεται το ID με default τιμή μηδέν και με την επιλογή invert selection εφαρμόζεται σε όλα τα χαρακτηριστικά πλην του ID.



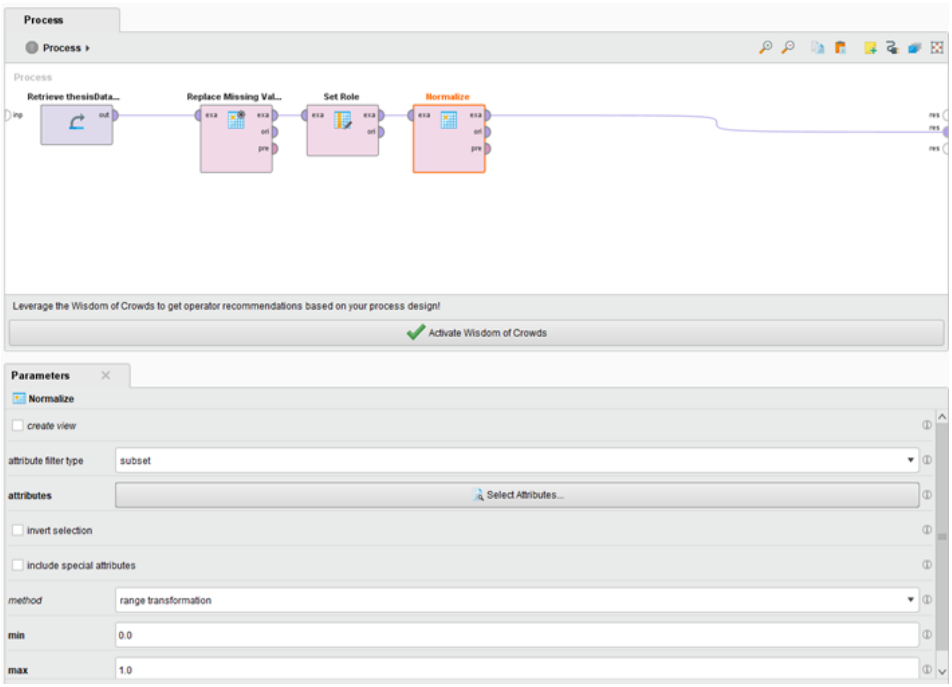
Εικόνα 5.9. Εισαγωγή αντικειμένου Replace Missing Values

Κατόπιν, καθορίζουμε το ρόλο του ID με το αντικείμενο Set Role. Μέσα από το αντικείμενο ορίζεται το χαρακτηριστικό ως τύπος id, προκειμένου να μην ληφθεί υπόψιν στην παραγωγή των αποτελεσμάτων, όπως φαίνεται στην εικόνα που ακολουθεί:



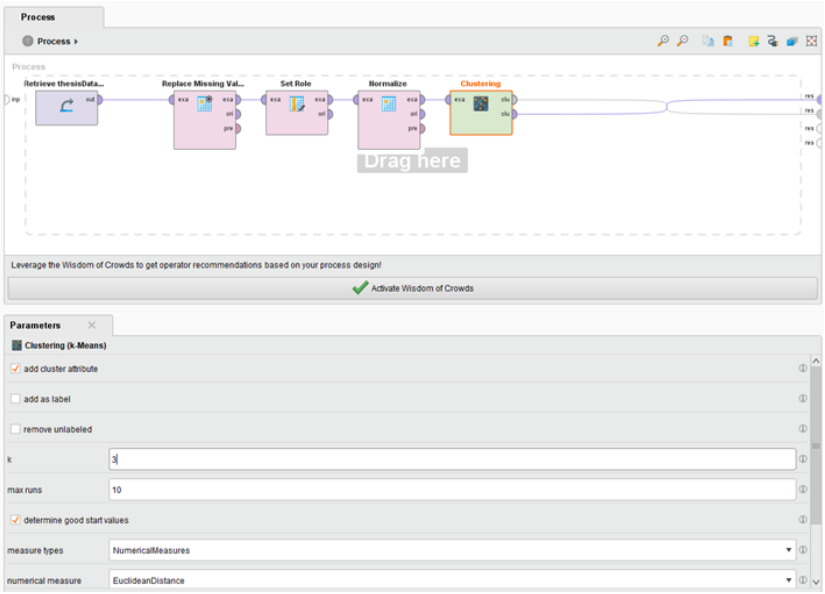
Εικόνα 5.10. Εισαγωγή αντικειμένου Set Role

Στη συνέχεια, επιλέγεται αντικείμενο, το οποίο κανονικοποιεί τα δεδομένα. Συγκεκριμένα, υλοποιείται κανονικοποίηση των τιμών σε ένα εύρος από 0 έως 1, για ευκολία στην κατανόηση αποτελεσμάτων. Το συγκεκριμένο βήμα είναι προαιρετικό. Το αντικείμενο και η παραμετροποίηση του φαίνονται στην ακόλουθη εικόνα:



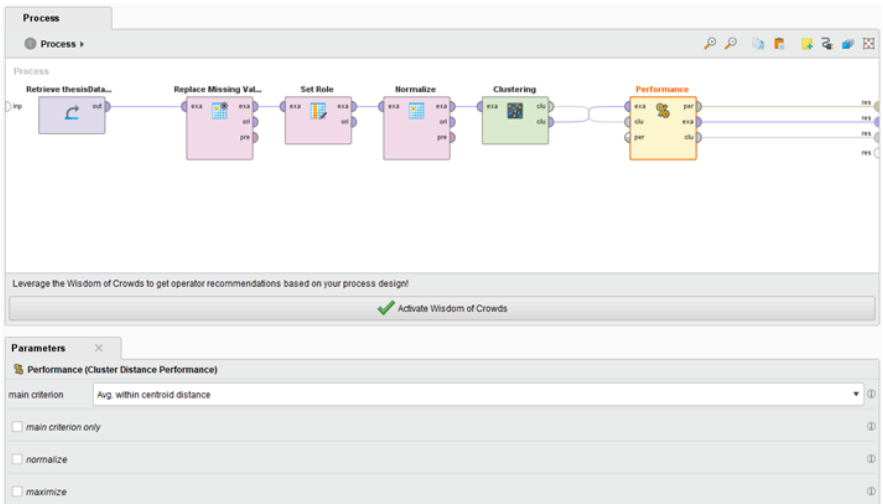
Εικόνα 5.11. Εισαγωγή αντικειμένου Normalize

Στο επόμενο βήμα εισάγετε το αντικείμενο που υλοποιεί τον αλγόριθμο k-means. Στην παραμετροποίηση του ορίζεται ο αριθμός των cluster στο πεδίο k, ο αριθμός των μέγιστων επαναλήψεων του k-means σχετικά με τα σημεία αρχικοποίησης του αλγόριθμου στο πεδίο max runs και η παραμετροποίηση της μετρικής της απόστασης. Ο τύπος των δεδομένων ορίζεται ως αριθμητικός και ως μετρική ορίζεται η ευκλείδια απόσταση. Η διαδικασία της σχεδίασης μέχρι αυτό το βήμα και η παραμετροποίηση φαίνεται στην εικόνα που ακολουθεί:



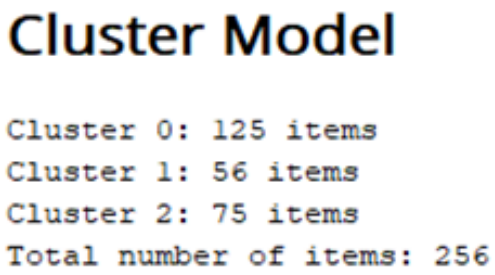
Εικόνα 5.12. Εισαγωγή αντικειμένου Clustering

Τελευταίο βήμα πριν την εκτέλεση του παραδείγματος, είναι η προσθήκη αντικειμένου απόδοσης και συγκεκριμένα cluster distance performance. Το συγκεριμένο αντικείμενο μετράει την απόδοση του μοντέλου με κριτήριο την απόσταση μεταξύ των κεντροειδών, δηλαδή την απόσταση μεταξύ των κέντρων των cluster. Ο τελικός σχεδιασμός της διεργασίας και η παραμετροποίηση του αντικειμένου performance φαίνονται στην εικόνα που ακολουθεί:



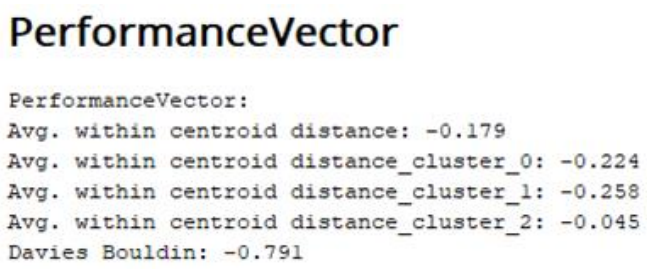
Εικόνα 5.13. Εισαγωγή αντικειμένου Performance

Θα μελετηθούν διάφοροι αριθμοί cluster εκκινώντας από τους τρεις, προκειμένου να επιλεχθεί ο βέλτιστος. Για k=3 τα αποτελέσματα μοιράστηκαν στους cluster, όπως φαίνεται στην ακόλουθη εικόνα:



Εικόνα 5.14. Διαμοιρασμός αντικειμένων ανά cluster για k=3

Οι αποστάσεις μεταξύ των cluster είναι:



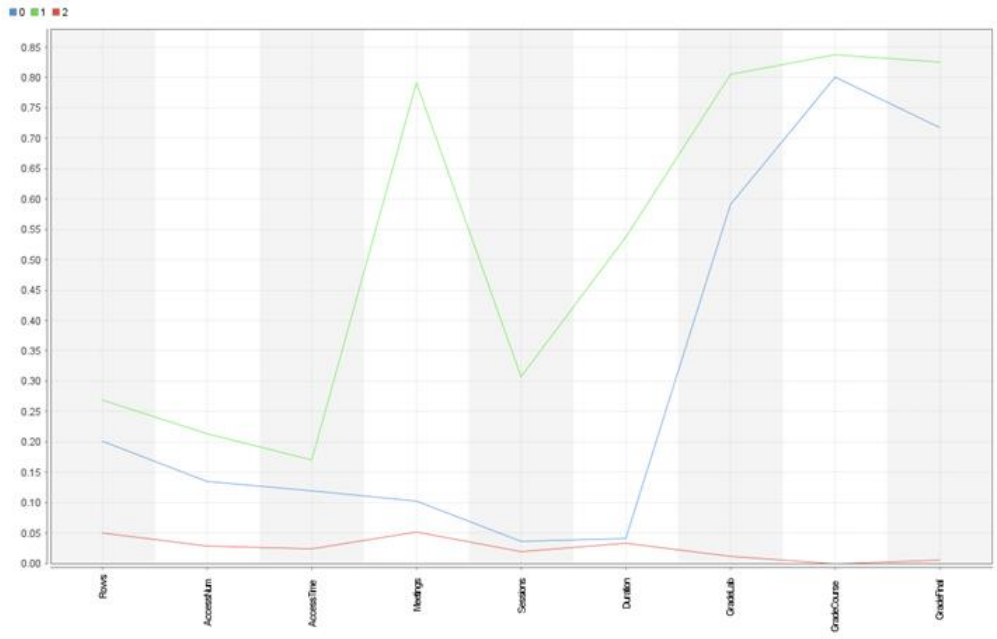
Εικόνα 5.15. Πίνακας απόδοσης για k=3

Ο πίνακας κεντροειδών είναι ο παρακάτω:

Attribute	cluster_0	cluster_1	cluster_2
Rows	0.201	0.270	0.050
AccessNum	0.134	0.213	0.029
AccessTime	0.119	0.170	0.024
Meetings	0.103	0.791	0.052
Sessions	0.036	0.308	0.019
Duration	0.040	0.538	0.034
GradeLab	0.591	0.805	0.012
GradeCourse	0.801	0.838	0
GradeFinal	0.717	0.825	0.005

Εικόνα 5.16. Πίνακας κεντροειδών για k=3

Οι ομάδες οπτικοποιούνται μέσω γραφήματος στην εικόνα που ακολουθεί:



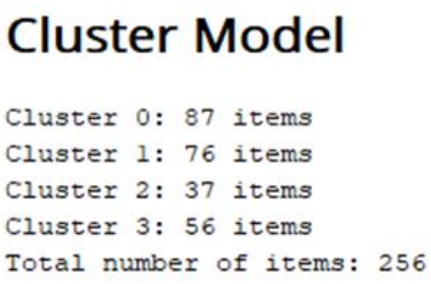
Εικόνα 5.17. Γράφημα απόδοσης ανά χαρακτηριστικό για k=3

Παρατηρήσεις:

- Ο cluster 0 έχει υπερδιπλάσιες εγγραφές από τον cluster 1.
- Υπάρχουν αρκετές εγγραφές ανά ομάδα για να δικαιολογηθεί ο αριθμός των cluster.
- Η μέση τιμή μεταξύ των cluster είναι αποδεκτή.
- Ο cluster 2 είναι πολύ πιο συμπαγής από τους άλλους δύο.
- Ο cluster 0 περιέχει τους φοιτητές με μέσο όρο βαθμολογίας 7.17, ο cluster 1 περιέχει τους φοιτητές με μέσο όρο 8.25 και ο cluster 2 τους φοιτητές με μέσο όρο 0.005.
- Οι φοιτητές του cluster 0 σε σχέση με τους φοιτητές του cluster 1 επισκέφτηκαν λιγότερο την πλατφόρμα του eClass (0.134 < 0.213), παρακολούθησαν σημαντικά λιγότερα meetings μέσω του MS Teams (0.103<0.791) και είχαν μικρότερη μέση βαθμολογία στο εργαστηριακό μέρος του μαθήματος (0.591<0.805).

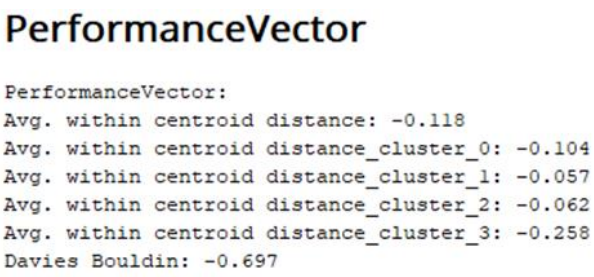
- Οι αποκλίσεις ανάμεσα στα cluster υποδεικνύουν πως δεν είναι ο ιδανικός αριθμός ομάδων για τα δεδομένα.

Για k=4 τα αποτελέσματα μοιράστηκαν στους cluster, όπως φαίνεται στην ακόλουθη εικόνα:



Εικόνα 5.18. Διαμοιρασμός αντικειμένων ανά cluster για k=4

Οι αποστάσεις μεταξύ των cluster είναι:



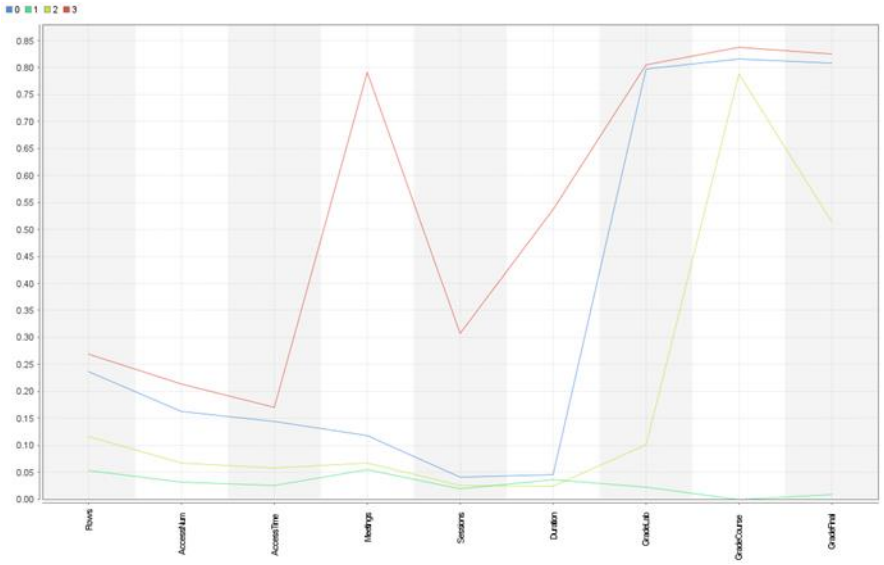
Εικόνα 5.19. Πίνακας απόδοσης για k=4

Ο πίνακας κεντροειδών είναι ο παρακάτω:

Attribute	cluster_0	cluster_1	cluster_2	cluster_3
Rows	0.237	0.053	0.117	0.270
AccessNum	0.162	0.031	0.067	0.213
AccessTime	0.144	0.026	0.057	0.170
Meetings	0.117	0.054	0.068	0.791
Sessions	0.040	0.020	0.025	0.308
Duration	0.046	0.036	0.024	0.538
GradeLab	0.797	0.022	0.101	0.805
GradeCourse	0.816	0	0.788	0.838
GradeFinal	0.808	0.009	0.514	0.825

Εικόνα 5.20. Πίνακας κεντροειδών για k=4

Οι ομάδες οπτικοποιούνται μέσω γραφήματος στην εικόνα που ακολουθεί:



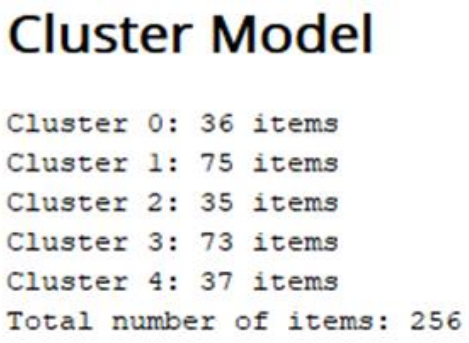
Εικόνα 5.21. Γράφημα απόδοσης ανά χαρακτηριστικό για k=4

Παρατηρήσεις:

- Οι εγγραφές είναι πιο ισομοιρασμένες σε σχέση με τα προηγούμενα αποτελέσματα.
- Η μέση απόσταση μειώθηκε από -0.179 σε -0.118, ενώ πλέον δύο cluster παρουσιάζουν μεγάλη συμπύκνωση.
- Ο δείκτης Davies-Bouldin παρουσίασε βελτίωση από -0.791 σε -0.697.
- Ο cluster 0 περιέχει φοιτητές, οι οποίοι πέτυχαν υψηλή βαθμολογία συμμετέχοντας στο μάθημα αποκλειστικά με ασύγχρονη τηλεκπαίδευση μέσω του eClass, με ελάχιστη συμμετοχή στη σύγχρονη τηλεκπαίδευση μέσω του MS Teams. Αυτό έχει σαν αποτέλεσμα μια μικρή διαφοροποίηση στον τελικό βαθμό σε σχέση με τους φοιτητές του cluster 3 (0.808<0.825), οι οποίοι παρακολούθησαν 0.791 meetings έναντι των φοιτητών του cluster 0 που παρακολούθησαν 0.117.

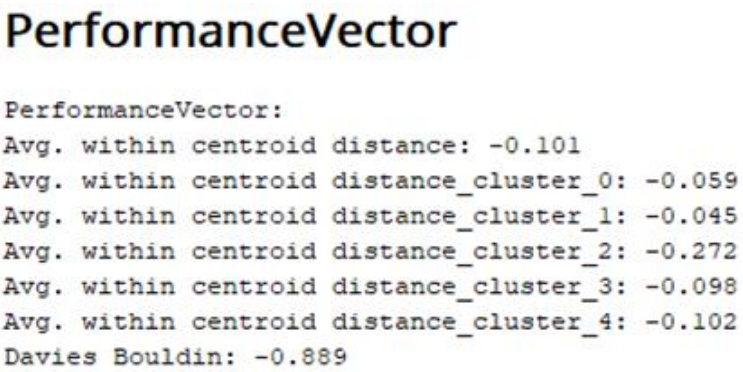
- Ο cluster 1 περιέχει φοιτητές, οι οποίοι είχαν ελάχιστη συμμετοχή, τόσο στο eClass, όσο και στο MS Teams, το οποίο αποτυπώνεται στον τελικό βαθμό (0.009).
- Ο cluster 2 περιέχει τους φοιτητές, οι οποίοι πέτυχαν οριακά προβιβάσιμο βαθμό (0.514). Αυτή η ομάδα έχει σχεδόν διπλάσιες τιμές σχετικά με την πρόσβαση στο eClass, σε σχέση με τους φοιτητές του cluster 1, αλλά παρόμοιες τιμές σε σχέση με τη συμμετοχή στο MS Teams.

Για k=5 τα αποτελέσματα μοιράστηκαν στους cluster, όπως φαίνεται στην ακόλουθη εικόνα:



Εικόνα 5.22. Διαμοιρασμός αντικειμένων ανά cluster για k=5

Οι αποστάσεις μεταξύ των cluster είναι:



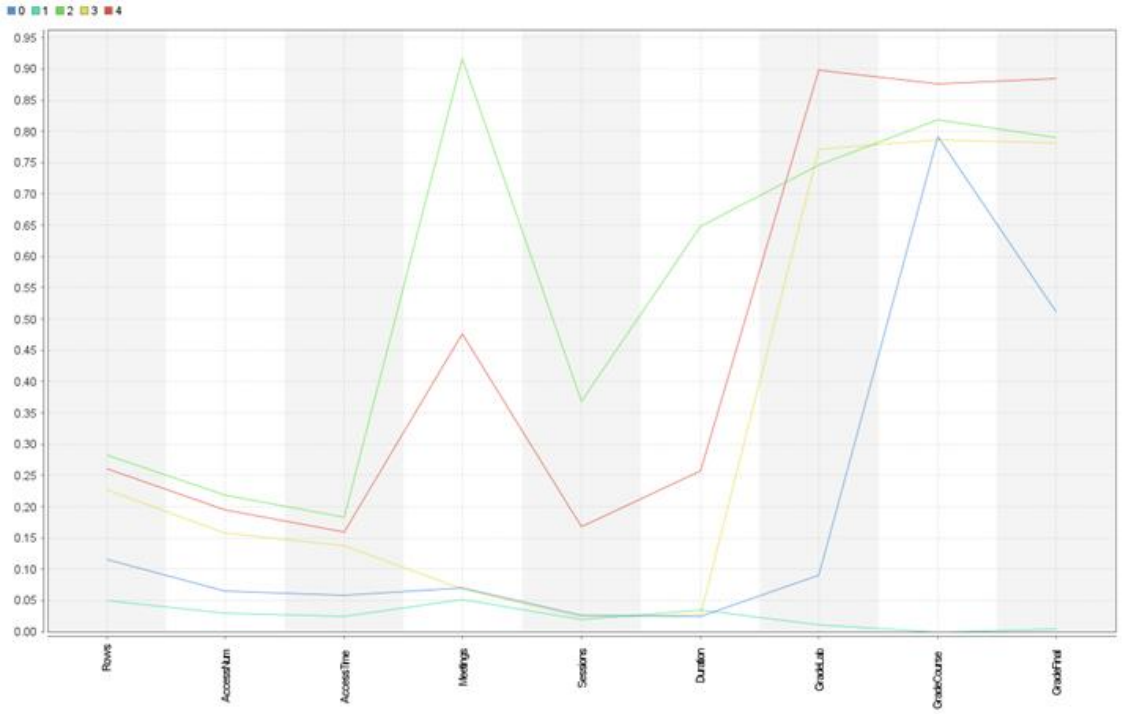
Εικόνα 5.23. Πίνακας απόδοσης για k=5

Ο πίνακας κεντροειδών είναι ο παρακάτω:

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
Rows	0.116	0.050	0.282	0.227	0.260
AccessNum	0.065	0.029	0.219	0.158	0.194
AccessTime	0.057	0.024	0.183	0.137	0.159
Meetings	0.069	0.052	0.917	0.068	0.476
Sessions	0.026	0.019	0.369	0.023	0.168
Duration	0.025	0.034	0.649	0.028	0.257
GradeLab	0.091	0.012	0.746	0.772	0.898
GradeCourse	0.791	0	0.819	0.787	0.877
GradeFinal	0.511	0.005	0.789	0.781	0.885

Εικόνα 5.24. Πίνακας κεντροειδών για k=5

Οι ομάδες οπτικοποιούνται μέσω γραφήματος στην εικόνα που ακολουθεί:



Εικόνα 5.25. Γράφημα απόδοσης ανά χαρακτηριστικό για k=5

Παρατηρήσεις:

- Οι εγγραφές είναι πιο ισομοιρασμένες σε σχέση με όλα τα προηγούμενα αποτελέσματα.
- Η μέση απόσταση μειώθηκε από -0.118 σε -0.101, με τέσσερις από τους πέντε cluster να εμφανίζουν μεγάλη συνοχή.

- Ο δείκτης Davies-Bouldin παρουσίασε αύξηση από -0.697 σε -0.889. Πρέπει να ληφθεί υπόψιν στην επιλογή του βέλτιστου αριθμού των cluster παρακάτω.
- Με μια ματιά στον πίνακα κεντροειδών, φαίνεται πως η ομάδα του cluster 2 με την ομάδα του cluster 3 έχουν ελάχιστη διαφορά στην τελική βαθμολογία, με μόνη διαφορά την συμμετοχή σε meetings του μαθήματος.
- Τα αποτελέσματα δεν προσφέρονται για επιπλέον παρατηρήσεις σε σχέση με τους τέσσερις cluster.

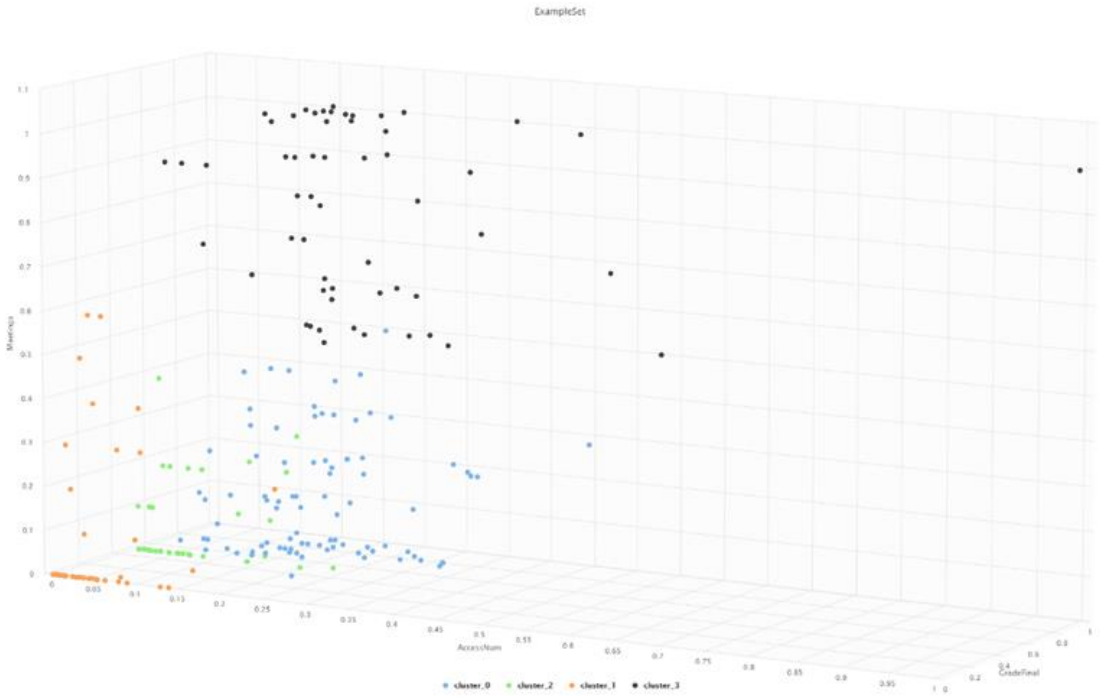
Για την επιλογή του βέλτιστου αριθμού cluster λαμβάνουμε υπόψιν διάφορες τιμές, όπως την απόσβεση της απόστασης μεταξύ των κεντροειδών. Τα αποτελέσματα σε παράθεση είναι:

- Για k=3: Avg. within centroid distance: -0.179
- Για k=4: Avg. within centroid distance: -0.118
- Για k=5: Avg. within centroid distance: - 0.101

Όπως είναι αναμενόμενο η μέση απόσταση φθίνει με την αύξηση των cluster. Όμως, η μετάβαση από τέσσερις σε πέντε εμφανίζει ελάχιστη διαφορά. Ακόμα ένας δείκτης είναι ο Davies-Bouldin, που παρουσιάζει την βέλτιστη τιμή του (-0.697) για k=4. Τέλος, τα αποτελέσματα για k=4 προσφέρονται για περισσότερα συμπεράσματα σε σχέση με τους τρεις cluster, ενώ το παράδειγμα με τους πέντε δεν προσφέρει νέαερμηνεύσιμα στοιχεία. Επομένως, επιλέγεται ως βέλτιστη η επιλογή των τεσσάρων. Μέσα από τις οπτικοποιήσεις που παρέχει το RapidMiner είναι δυνατόν να γίνουν πιο σαφή τα συμπεράσματα.

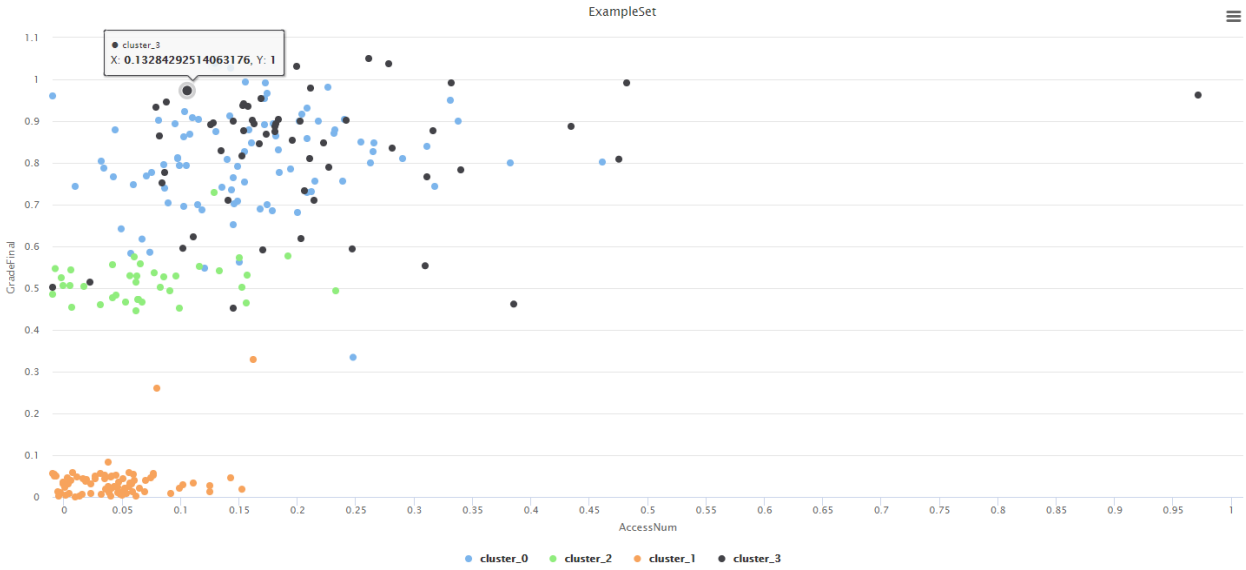
Στο ακόλουθο σχήμα παρατηρείται η σχέση μεταξύ σύγχρονης (AccessNum), ασύγχρονης (Meetings) και τελικού βαθμού (GradeFinal). Το συγκεκριμένο γράφημα επιβεβαιώνει τη συμβολή του eClass στην τελική βαθμολογία, εφόσον κανένας φοιτητής που πέτυχε προβιβάσιμη

βαθμολογία δεν έχει χαμηλό αριθμό επισκέψεων στην πλατφόρμα. Αντίθετα, παρατηρούμε κάποιους φοιτητές με ικανοποιητικό αριθμό παρακολούθησης στα meeting του μαθήματος, οι οποίοι δεν πέτυχαν προβιβάσιμο βαθμό.



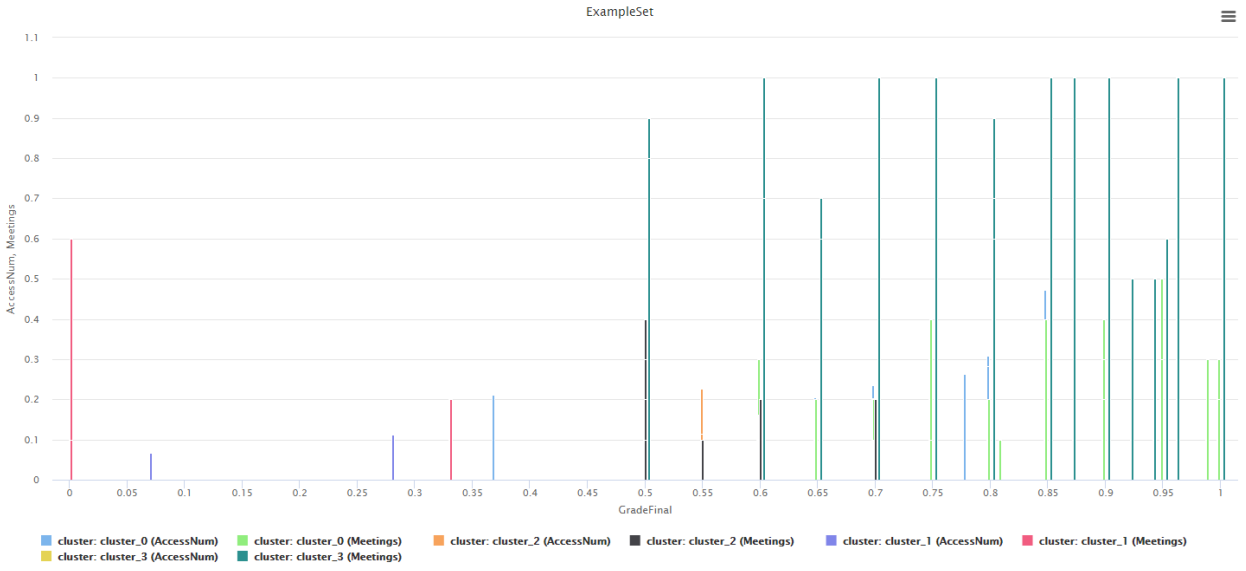
Εικόνα 5.26. Γράφημα συσχέτισης της συμμετοχής ανά είδος εκπαίδευσης με τελικό βαθμό

Το ίδιο αποτέλεσμα επιβεβαιώνεται στο ακόλουθο γράφημα:



Εικόνα 5.27. Γράφημα διασποράς για επισκέψεις στο eclass σε σχέση με τον τελικό βαθμό

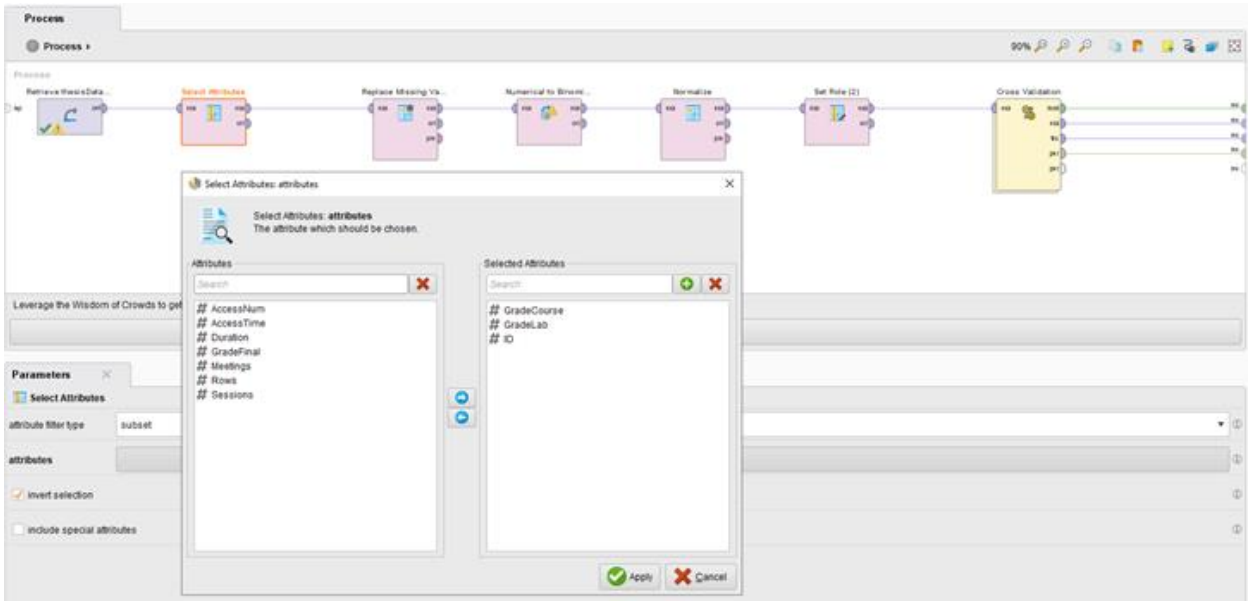
Ομοίως και για το γράφημα με τις στήλες:



Εικόνα 5.28. Γράφημα συσχέτισης τελικού βαθμού ανά συμμετοχή ανά ομάδα

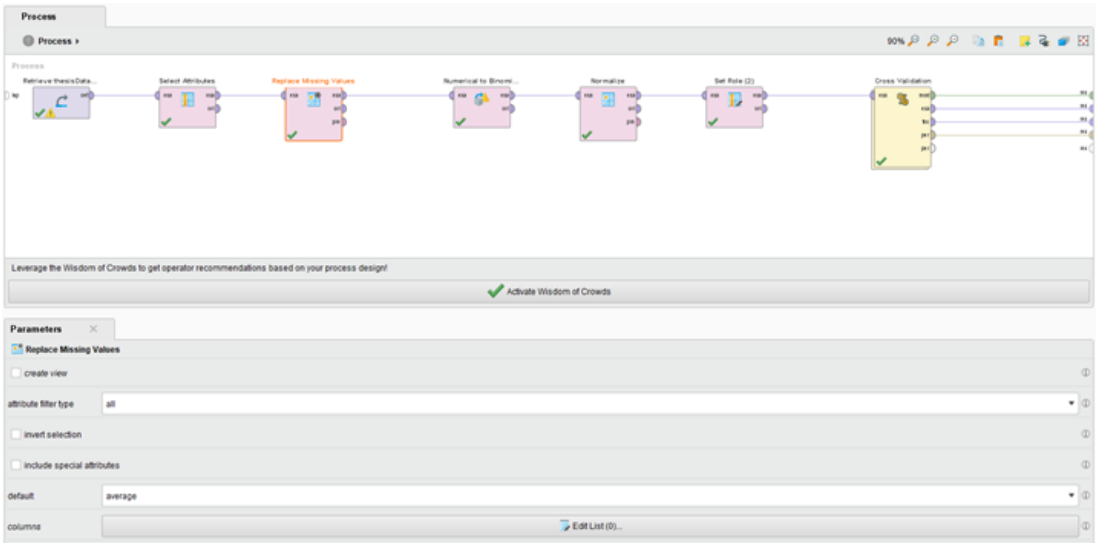
5.4.3. ΕΦΑΡΜΟΓΗ ΔΕΝΤΡΟΥ ΑΠΟΦΑΣΗΣ

Για την υλοποίηση του δέντρου απόφασης με χρήση του RapidMiner, θα πρέπει αρχικά να αφαιρεθούν τα χαρακτηριστικά ID και οι βαθμοί θεωρίας και εργαστηρίου από το dataset. Αυτό γίνεται με χρήση αντικειμένου επιλογής χαρακτηριστικών, το Select Attributes. Επιλέγεται το υποσύνολο χαρακτηριστικών, τα οποία δεν θα συμμετέχουν στο δέντρο (ID, GradeCourse, GradeLab) και με χρήση του invert selection επιτυγχάνεται το επιθυμητό αποτέλεσμα, όπως φαίνεται στην εικόνα που ακολουθεί:



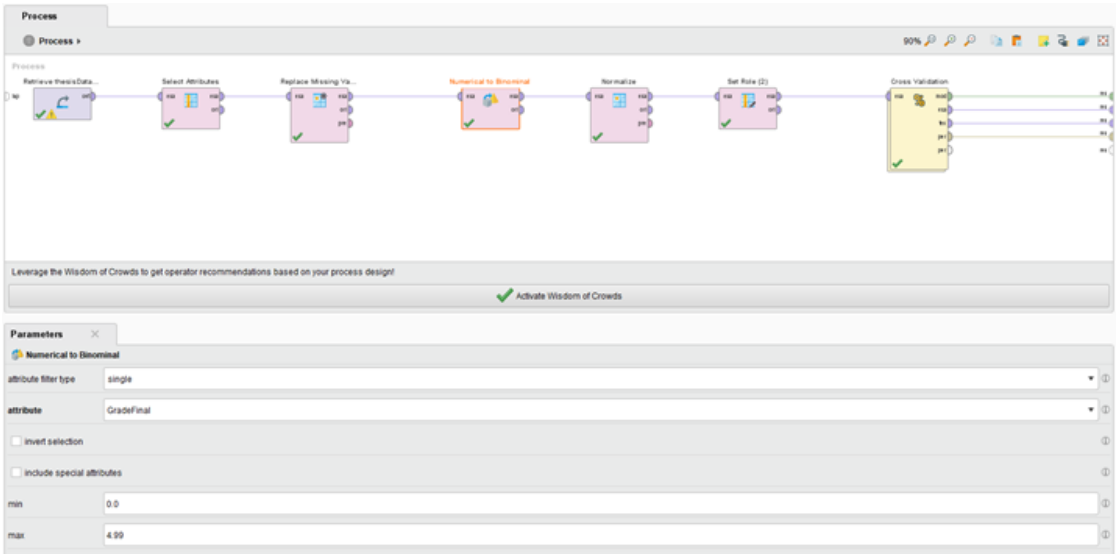
Εικόνα 5.29. Εισαγωγή αντικειμένου Select Attributes

Επόμενο βήμα, όπως στην εφαρμογή του k-means, είναι η προσθήκη αντικειμένου, του οποίου η λειτουργία είναι να μηδενίζει τα κενά πεδία των χαρακτηριστικών. Η επιλογή αυτή έγινε επειδή το κενό πεδίο σημαίνει πως δεν υπήρξε καμία ενέργεια από μέρους του συμμετέχοντα. Ο μηδενισμός εφαρμόζεται σε όλα τα πεδία, εφόσον το ID έχει αφαιρεθεί από το προηγούμενο βήμα σχεδιασμού. Οι παράμετροι του αντικειμένου εμφανίζονται στην εικόνα που ακολουθεί:



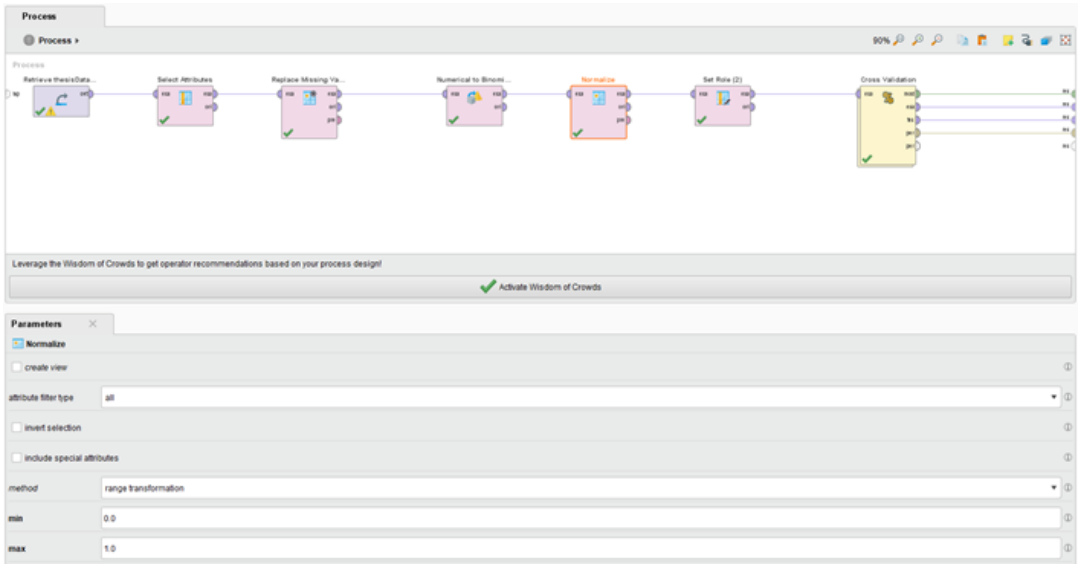
Εικόνα 5.30. Εισαγωγή αντικειμένου Replace Missing Values

Επόμενο βήμα στο σχεδιασμό του δέντρου απόφασης είναι η μετατροπή του τελικού βαθμού του μαθήματος από αριθμητική τιμή σε διαδική. Ορίζοντας τα πεδία min και max σε 0 και 4.99 αντίστοιχα, εάν η τιμή του πεδίου ανήκει στο εύρος τιμών, χαρακτηρίζεται ως ψευδής. Σε κάθε άλλη περίπτωση είναι αληθής. Η παραμετροποίηση του συγκεκριμένου αντικειμένου φαίνεται στην εικόνα που ακολουθεί:



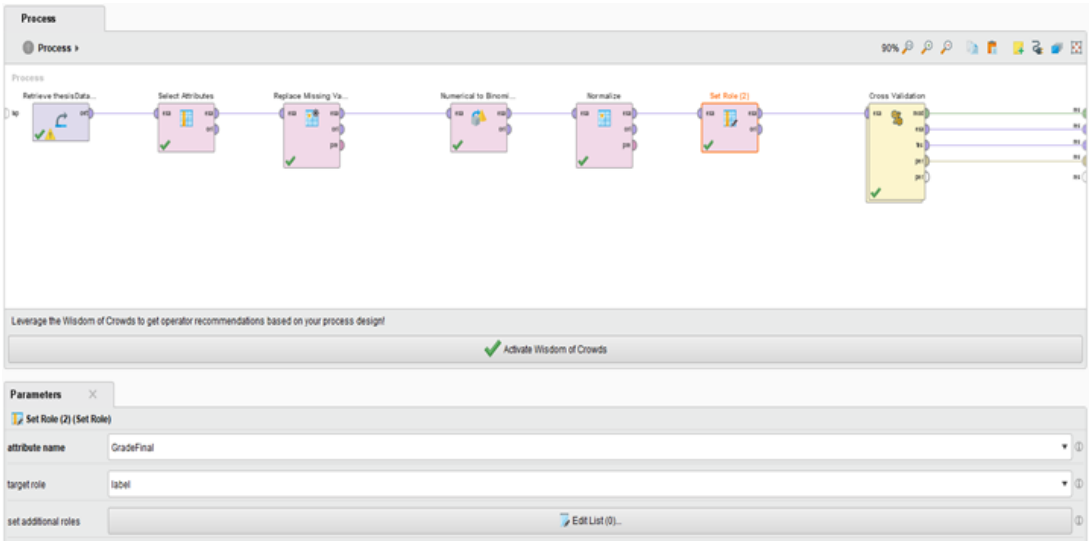
Εικόνα 5.31. Εισαγωγή αντικειμένου Numerical to Binominal

Στη συνέχεια, επιλέγεται αντικείμενο, το οποίο κανονικοποιεί τα δεδομένα. Συγκεκριμένα, υλοποιείται κανονικοποίηση των τιμών σε ένα εύρος από 0 έως 1, για ευκολία στην κατανόηση αποτελεσμάτων. Το συγκεκριμένο βήμα είναι προαιρετικό. Το αντικείμενο και η παραμετροποίηση του φαίνονται στην ακόλουθη εικόνα:



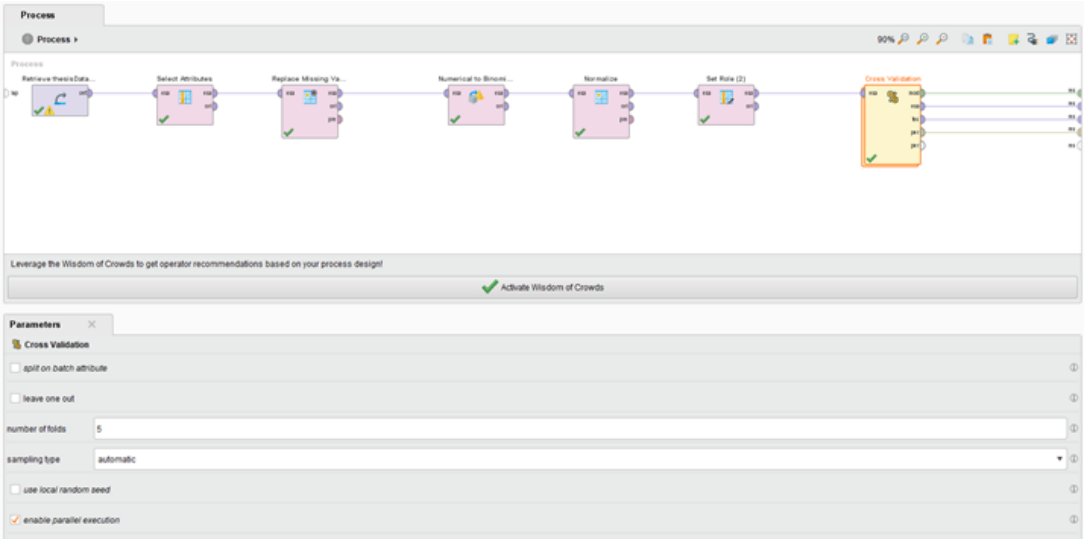
Εικόνα 5.32. Εισαγωγή αντικειμένου Normalize

Επόμενο βήμα στη δημιουργία του μοντέλου είναι ο ορισμός του τελικού βαθμού ως label. Με αυτή τη μετατροπή, ο τελικός βαθμός υποδεικνύεται ως το χαρακτηριστικό που λαμβάνουν υπόψιν τους οι τελεστές μάθησης. Αυτό υλοποιείται με το αντικείμενο Set Role και τις παραμέτρους στην ακόλουθη εικόνα:



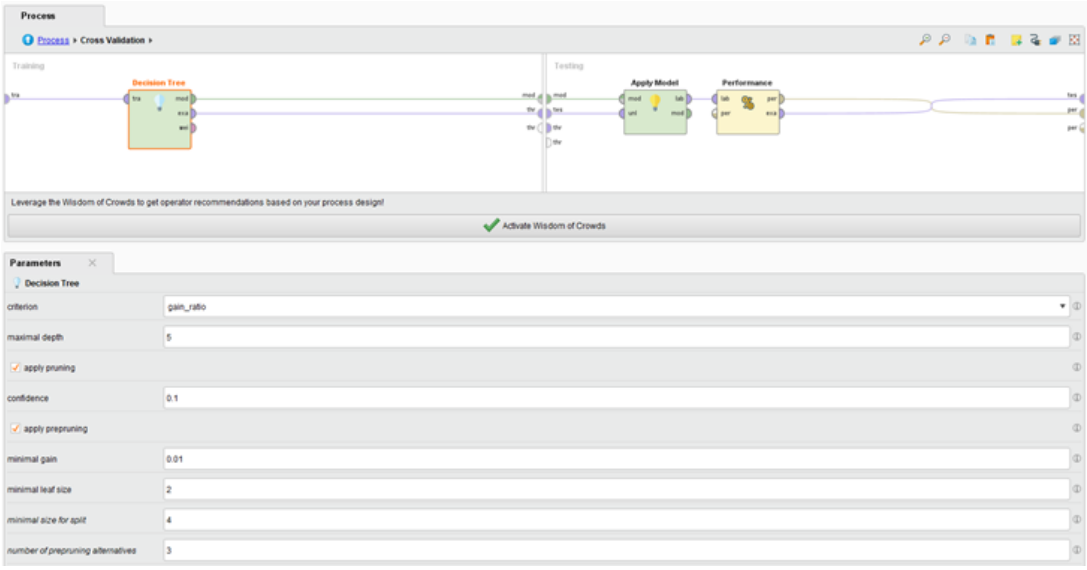
Εικόνα 5.33. Εισαγωγή αντικειμένου Set Role

Τελευταία προσθήκη στο παράθυρο του process είναι το αντικέιμενο Cross Validation. Ο συγκεκριμένος τελεστής χρησιμοποιείται κυρίως, για να εκτιμήσει με ακρίβεια την απόδοση ενός μοντέλου στην πράξη, μέσα από δυο υποεπεξεργασίες: μια υποεπεξεργασία κατάρτισης και μια υποεπεξεργασία δοκιμής. Η υποεπεξεργασία εκπαίδευσης χρησιμοποιείται για την εκπαίδευση ενός μοντέλου. Το εκπαιδευμένο μοντέλο, στη συνέχεια εφαρμόζεται στην υποεπεξεργασία δοκιμής. Η απόδοση του μοντέλου μετριέται κατά τη φάση δοκιμής. Η παραμετροποίηση του για το παράδειγμα φαίνεται στην επόμενη εικόνα, με το πεδίο number of folds να αντιπροσωπεύει τον αριθμό των subset εγγραφών που θα χρησιμοποιήσει ο τελεστής και το sampling type να ορίζει τη μέθοδο επιλογής των subset:



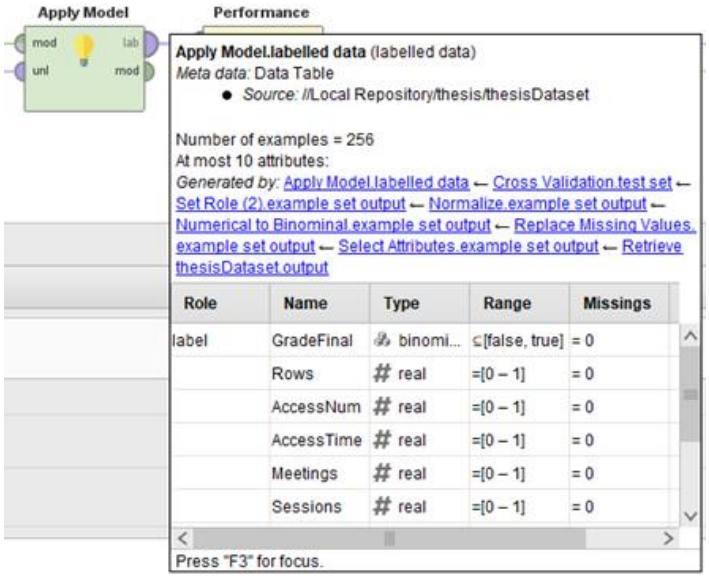
Εικόνα 5.34. Εισαγωγή αντικειμένου Cross Validation

Με διπλό click στο αντικείμενο του Cross Validation, εμφανίζεται το παράθυρο σχεδίασης με τα δύο προαναφερόμενα μέρη, το Training και το Testing. Στο Training μέρος εισάγεται το δέντρο απόφασης με την παραμετροποίηση που φαίνεται στην ακόλουθη εικόνα:



Εικόνα 5.35. Εισαγωγή αντικειμένου Decision Tree

Το αντικείμενο Decision Tree παράγει ένα μοντέλο (mod) και ένα example set (ex). Αυτά δίδονται σαν είσοδος, στο τέλος της διαδικασίας Training στο αντικείμενο Cross Validation. Κατόπιν, το Cross Vallidation μεταφέρει στο κομμάτι του Testing το μοντέλο και ένα test set, τα οποία δίδονται ως είσοδοι στο αντικείμενο Apply Model. Το αντικείμενο Apply Model παράγει LabeledData, όπως φαίνεται και στην ακόλουθη εικόνα:



Εικόνα 5.36. Εισαγωγή αντικειμένου Apply Model

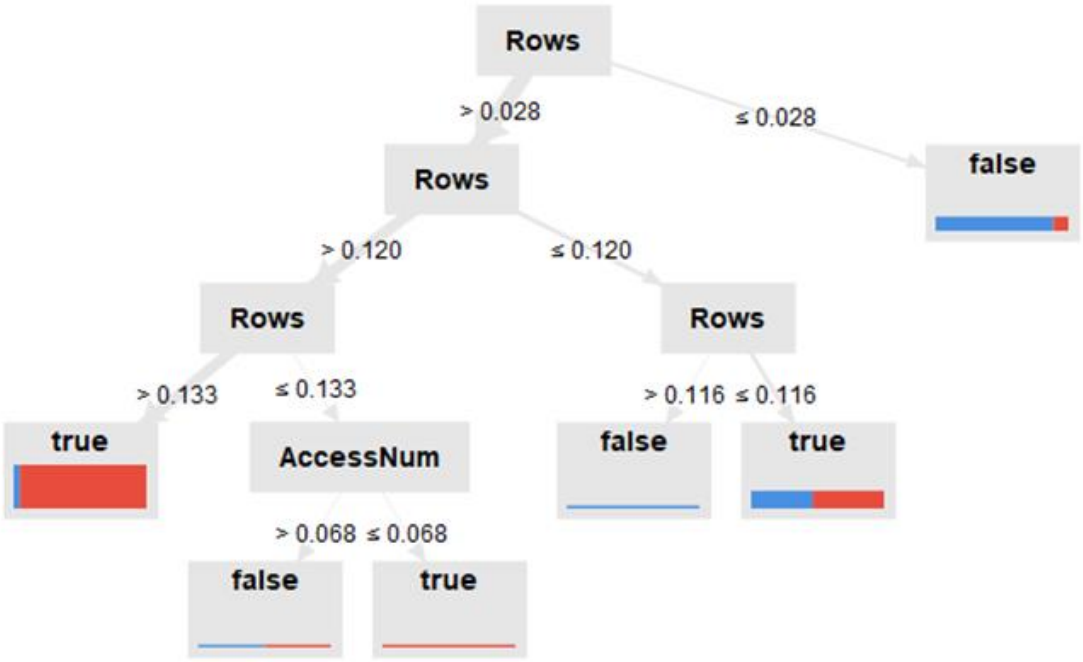
Τέλος, ένα αντικείμενο Performance χρησιμοποιείται για αξιολόγηση απόδοσης στα αποτελέσματα του μοντέλου. Με την εκτέλεση του μοντέλου λαμβάνουμε τα αποτελέσματα απόδοσης, όπως φαίνεται στην ακόλουθη εικόνα:

accuracy: 81.26% +/- 4.00% (micro average: 81.25%)

	true false	true true	class precision
pred. false	55	27	67.07%
pred. true	21	153	87.93%
class recall	72.37%	85.00%	

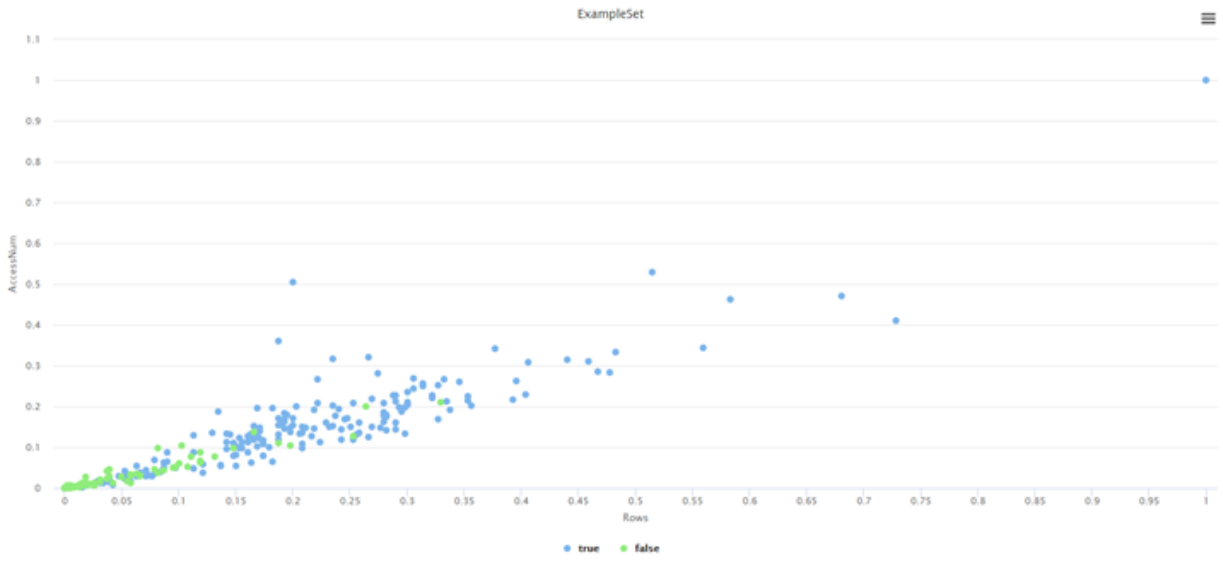
Εικόνα 5.37. Αποτελέσματα απόδοσης δέντρου απόδοσης

Το παραγόμενο δέντρο είναι στην ακόλουθη εικόνα:



Εικόνα 5.38. Παραγόμενο δέντρο απόφασης

Χρήσιμο για παρατηρήσεις είναι το ακόλουθο γράφημα:



Εικόνα 5.39. Γράφημα συσχέτισης επιτυχίας – συμμετοχής

Παρατηρήσεις:

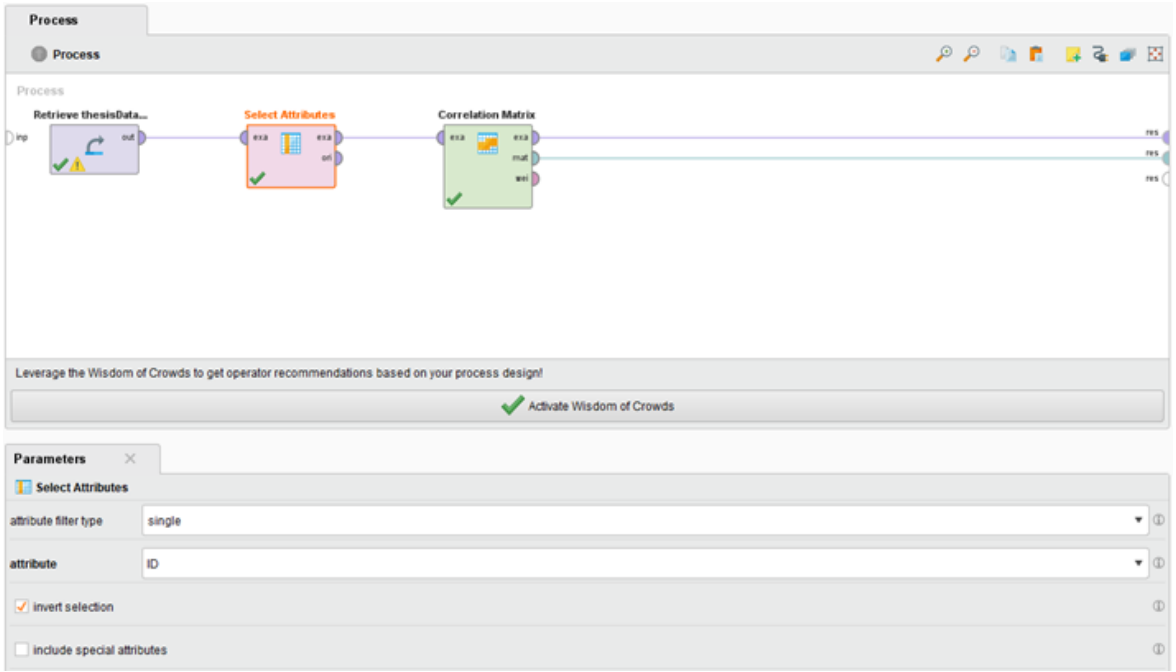
- Το παραγόμενο δέντρο περιέχει μόνο το κομμάτι της ασύγχρονης εκπαίδευσης μέσω eClass, παρότι το dataset περιέχει και το κομμάτι της σύγχρονης μέσω MS Teams.
- Το μοντέλο έχει αποδεκτά ποσοστά τόσο στο class prediction (67.07% - 87.93%) όσο και στο true/false (72.37% - 85.00%).
- Το γενικό accuracy είναι καλό, με ποσοστό 81.26% και απόκλιση +/- 4%.
- Τα συμπεράσματα του δέντρου, τα οποία είναι πιο ευανάγνωστα στο γράφημα, συμφωνούν με την εικόνα που προέκυψε από την εφαρμογή του k-means, σχετικά με την επίτευξη προβιβάσιμου βαθμού ενός φοιτητή στο μάθημα.

5.4.4. CORRELATION MATRIX

Ένας πίνακας συσχέτισης (correlation matrix) δείχνει τους συντελεστές συσχέτισης μεταξύ των μεταβλητών. Κάθε κελί του πίνακα δείχνει τη συσχέτιση μεταξύ δύο μεταβλητών, με το εύρος τιμών να είναι από 0 έως 1, όπου στο μέγιστο παρατηρείται απόλυτη συσχέτιση. Ο πίνακας συσχέτισης χρησιμοποιείται, για να συνοψίσει τα δεδομένα, ως εισαγωγή σε μια πιο προηγμένη ανάλυση και ως διαγνωστικό για προηγμένες αναλύσεις.

Στο RapidMiner, για να υλοποιηθεί ο πίνακας συσχέτισης χρειάζονται λίγα μόνο βήματα. Αρχικά, εισάγεται το dataset. Στη συνέχεια, χρησιμοποιείται ένα αντικείμενο επιλογής των χαρακτηριστικών για να μην συμπεριληφθεί το ID στον πίνακα. Όπως φαίνεται στην παρακάτω εικόνα, αφού εισαχθεί το αντικείμενο Select Attributes, επιλέγεται το ID και η επιλογή invert selection, για να το αποκλείσουμε από τα αποτελέσματα. Στη συνέχεια, προστίθεται το

αντικείμενο Correlation Matrix που υλοποιεί τον πίνακα συσχέτισης. Στην ακόλουθη εικόνα φαίνεται ολοκληρωμένη η διαδικασία:



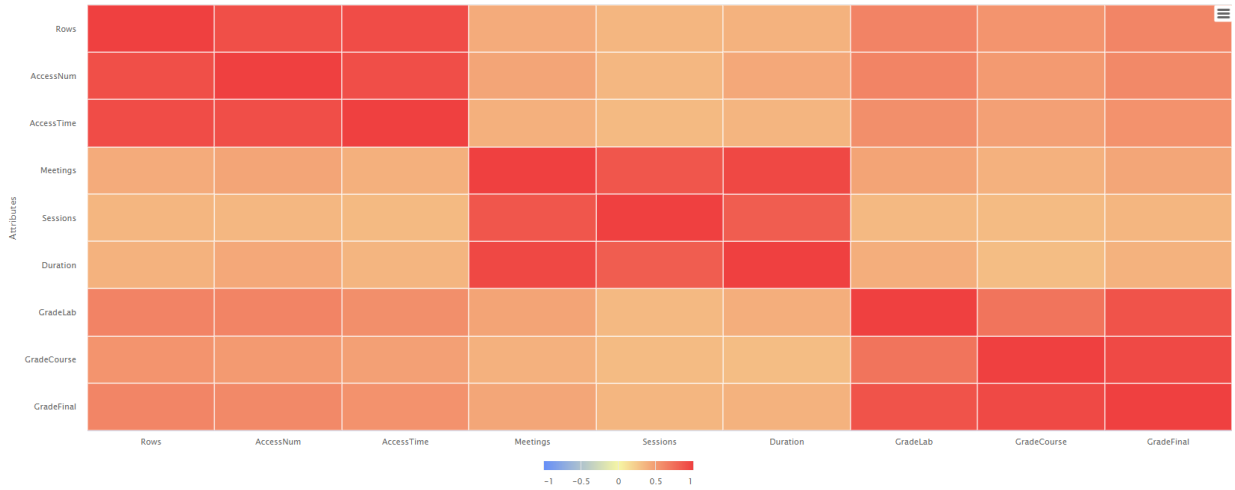
Εικόνα 5.40. Σχεδιασμός διαδικασίας Correlation Matrix

Με την εκτέλεση της διαδικασίας που σχεδιάστηκε λαμβάνονται τα παρακάτω αποτελέσματα:

Attribut...	Rows	Access...	Access...	Meetings	Sessions	Duration	GradeLab	GradeC...	GradeFi...
Rows	1	0.916	0.934	0.415	0.357	0.378	0.634	0.541	0.623
AccessN...	0.916	1	0.921	0.446	0.350	0.431	0.630	0.509	0.601
AccessTi...	0.934	0.921	1	0.389	0.336	0.359	0.566	0.475	0.552
Meetings	0.415	0.446	0.389	1	0.880	0.956	0.456	0.381	0.444
Sessions	0.357	0.350	0.336	0.880	1	0.842	0.340	0.326	0.357
Duration	0.378	0.431	0.359	0.956	0.842	1	0.399	0.317	0.378
GradeLab	0.634	0.630	0.566	0.456	0.340	0.399	1	0.718	0.898
GradeCo...	0.541	0.509	0.475	0.381	0.326	0.317	0.718	1	0.951
GradeFi...	0.623	0.601	0.552	0.444	0.357	0.378	0.898	0.951	1

Εικόνα 5.41. Correlation Matrix

Επιπλέον, το πρόγραμμα παρέχει και την εικονικοποίηση που ακολουθεί:



Εικόνα 5.42. Εικονικοποίηση Correlation Matrix

Μέσω του πίνακα παρατηρείται πως :

- Είναι ξεκάθαρες οι υποομάδες εντός των δεδομένων, όπως είχαν σχολιαστεί στο υποκεφάλαιο της δομής αρχείου δεδομένων. Η πρώτη αποτελείται από τα χαρακτηριστικά Rows, AccessNum και AccessTime και αφορά το ασύγχρονο τμήμα της εκπαίδευσης, μέσω της πλατφόρμας του eClass. Η δεύτερη αποτελείται από τα χαρακτηριστικά Meetings, Sessions και Duration και αφορά το σύγχρονο τμήμα της εκπαίδευσης μέσω της πλατφόρμας MS Teams. Στην τρίτη τα χαρακτηριστικά GradeLab, GradeCourse και GradeFinal παρουσιάζουν τη βαθμολογία του μαθήματος.
- Το ασύγχρονο κομμάτι της εκπαίδευσης μέσω eClass εμφανίζεται για άλλη μια φορά να έχει μεγαλύτερη συσχέτιση με τον τελικό βαθμό, σε σχέση με την σύγχρονη εκπαίδευση μέσω MS Teams. Αυτό είναι παρατηρήσιμο τόσο στην αριθμητική αναπαράσταση όσο και στην εικονικοποίηση.

- Υπάρχει ισχυρή συσχέτιση μεταξύ του βαθμού στο εργαστηριακό μέρος του μαθήματος και του θεωρητικού. Αυτό το αποτέλεσμα είναι αναμενόμενο, λαμβάνοντας υπόψιν τη φύση του μαθήματος των βάσεων δεδομένων, όπου πράξη και θεωρία είναι αλληλένδετα.

6. ΣΥΜΠΕΡΑΣΜΑΤΑ

Η τηλεκπαίδευση αποτελεί ένα εξαιρετικά χρήσιμο εργαλείο στον τομέα της εκπαίδευσης. Είναι ένα μέσο το οποίο, μπορεί είτε να χρησιμοποιηθεί αποκλειστικά για την εκπαιδευτική διαδικασία, εφόσον υπάρξει ανάγκη, είτε σε συνεργασία με τη διδασκαλία δια ζώσης. Με τη χρήση τεχνικών και εργαλείων εξόρυξης δεδομένων μπορεί να αντληθεί γνώση και να αναλυθούν δεδομένα που θα βοηθήσουν να αναγνωριστούν έγκαιρα προβλήματα ή να ληφθούν χρήσιμες αποφάσεις στην εκπαιδευτική διαδικασία.

Στην παρούσα εργασία μελετήθηκαν ανωνυμοποιημένα δεδομένα από το προπτυχιακό μάθημα «Βάσεις Δεδομένων» του τμήματος Μηχανικών Πληροφορικής και Υπολογιστών του Πανεπιστημίου Δυτικής Αττικής. Το μάθημα (θεωρητικό και εργαστηριακό μέρος) διδάχθηκε αποκλειστικά εξ΄αποστάσεως μέσω ασύγχρονης τηλεκπαίδευσης με χρήση της πλατφόρμας του ανοικτού eClass και μέσω σύγχρονης με χρήση της πλατφόρμας Microsoft Teams. Συνολικά χρησιμοποιήθηκαν στοιχεία από 256 φοιτητές.

Τα δεδομένα επεξεργάστηκαν με χρήση του εργαλείου RapidMiner. Εφαρμόστηκε ο αλγόριθμος K-Means και τεχνική δέντρων απόφασης. Χρησιμοποιήθηκαν οπτικοποιήσεις που παρέχει το εργαλείο RapidMine για διευκόλυνση στην εξαγωγή συμπερασμάτων.

Ο αλγόριθμος K-Means εφαρμόστηκε για τρεις, τέσσερις και πέντε συστάδες (clusters).

Για k=3 παρατηρήθηκε πως η cluster 0 περιέχει τους φοιτητές με μέσο όρο βαθμολογίας 7.17, η cluster 1 περιέχει τους φοιτητές με μέσο όρο 8.25 και η cluster 2 τους φοιτητές με μέσο όρο 0.005. Οι φοιτητές της cluster 0 σε σχέση με τους φοιτητές της cluster 1 επισκέφτηκαν λιγότερο την πλατφόρμα του eClass ($0.134 < 0.213$), παρακολούθησαν σημαντικά λιγότερα meetings μέσω

του MS Teams (0.103<0.791) και είχαν μικρότερη μέση βαθμολογία στο εργαστηριακό μέρος του μαθήματος (0.591<0.805). Όμως, οι αποκλίσεις ανάμεσα στα cluster υποδεικνύουν πως δεν είναι ο ιδανικός αριθμός ομάδων για τα δεδομένα.

Για k=4 παρατηρήθηκε πως οι εγγραφές (observations, examples) είναι πιο ισομοιρασμένες σε σχέση με τα προηγούμενα αποτελέσματα. Η cluster 0 περιέχει φοιτητές, οι οποίοι πέτυχαν υψηλή βαθμολογία συμμετέχοντας στο μάθημα αποκλειστικά με ασύγχρονη τηλεκπαίδευση μέσω του eClass, με ελάχιστη συμμετοχή στη σύγχρονη τηλεκπαίδευση μέσω του MS Teams. Αυτό έχει σαν αποτέλεσμα μια μικρή διαφοροποίηση στον τελικό βαθμό σε σχέση με τους φοιτητές της cluster 3 (0.808<0.825), οι οποίοι παρακολούθησαν 0.791 meetings έναντι των φοιτητών του cluster 0 που παρακολούθησαν 0.117. Η cluster 1 περιέχει φοιτητές, οι οποίοι είχαν ελάχιστη συμμετοχή, τόσο στο eClass, όσο και στο MS Teams, το οποίο αποτυπώνεται στον τελικό βαθμό (0.009). Ο cluster 2 περιέχει τους φοιτητές, οι οποίοι πέτυχαν οριακά προβιβάσιμο βαθμό (0.514). Αυτή η ομάδα έχει σχεδόν διπλάσιες τιμές σχετικά με την πρόσβαση στο eClass, σε σχέση με τους φοιτητές του cluster 1, αλλά παρόμοιες τιμές σε σχέση με τη συμμετοχή στο MS Teams. Επιπλέον, βελτίωση παρουσίασε τόσο η μέση απόσταση (από -0.179 σε -0.118) τόσο και ο δείκτης Davies-Bouldin (από -0.791 σε -0.697).

Για k=5 παρατηρήθηκε πως οι εγγραφές είναι πιο ισομοιρασμένες σε σχέση με όλα τα προηγούμενα αποτελέσματα και πως η μέση απόσταση μειώθηκε από -0.118 σε -0.101, με τέσσερις από τις πέντε cluster να εμφανίζουν μεγάλη συνοχή. Όμως, ο δείκτης Davies-Bouldin παρουσίασε αύξηση από -0.697 σε -0.889 και η ομάδα της cluster 2 με την ομάδα του cluster 3 έχουν ελάχιστη διαφορά στην τελική βαθμολογία, με μόνη διαφορά τη συμμετοχή σε meetings του μαθήματος. Επομένως, επιλέχθηκε ως βέλτιστη η επιλογή των τεσσάρων.

Στο δέντρο απόφασης παρατηρήθηκε πως το παραγόμενο δέντρο περιέχει μόνο το κομμάτι της ασύγχρονης εκπαίδευσης μέσω eClass, παρότι το dataset περιέχει και το κομμάτι της σύγχρονης μέσω MS Teams. Τα συμπεράσματα που προκύπτουν από το παραγώμενο δέντρο, συμφωνούν με την εικόνα που προέκυψε από την εφαρμογή του k-means, σχετικά με την επίτευξη προβιβάσιμου βαθμού ενός φοιτητή στο μάθημα.

Με τη χρήση του εργαλείου Correlation Matrix, παρατηρήθηκε πως το ασύγχρονο κομμάτι της εκπαίδευσης μέσω eClass εμφανίζεται να έχει μεγαλύτερη συσχέτιση με τον τελικό βαθμό, σε σχέση με την σύγχρονη εκπαίδευση μέσω MS Teams. Η συμβολή του eClass στην τελική βαθμολογία είναι εξέχουσας σημασίας, εφόσον κανένας φοιτητής που πέτυχε προβιβάσιμη βαθμολογία δεν έχει χαμηλό αριθμό επισκέψεων στην πλατφόρμα. Αντίθετα, παρατηρείται πως κάποιοι φοιτητές με ικανοποιητικό αριθμό παρακολούθησης στα meeting του μαθήματος, δεν πέτυχαν προβιβάσιμο βαθμό. Επίσης, εμφανίζεται ισχυρή συσχέτιση μεταξύ του βαθμού στο εργαστηριακό μέρος του μαθήματος και του θεωρητικού, αποτέλεσμα αναμενόμενο, λαμβάνοντας υπόψιν τη φύση του μαθήματος των βάσεων δεδομένων, όπου πράξη και θεωρία είναι αλληλένδετα.

Τέλος, θα πρέπει να επισημανθεί ότι όλες οι διαλέξεις στην πλατφόρμα MS-TEAMS καταγράφονται και είναι προσβάσιμες για ασύγχρονη παρακολούθηση μέσω συνδέσμων που είναι διαθέσιμοι μέσω της πλατφόρμας eClass.

Μελλοντικά θα μελετηθεί η «κίνηση» που αφορά το κατέβασμα (download) των μαγνητοσκοπημένων διαλέξεων και στοιχεία για την ασύγχρονη παρακολούθηση των διαλέξεων.

Τέλος, η έρευνα θα επεκταθεί με στοιχεία και άλλων μαθημάτων και θα χρησιμοποιηθούν και άλλες τεχνικές και αλγόριθμοι εξόρυξης δεδομένων.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Jiawei Han, Micheline Kamber, Jian Pei, (2011), Data Mining: Concepts and Techniques 3rd Edition.

2. Βερύκιος, Β., Καγκλής, Β., Σταυρόπουλος, Η., (2015), Η επιστήμη των δεδομένων μέσα από τη γλώσσα R. [ηλεκτρ. βιβλ.] Αθήνα:Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. Διαθέσιμο στο: http://hdl.handle.net/11419/2965

3. Κύρκος, Ε., (2015). Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων. [ηλεκτρ. βιβλ.] Αθήνα:Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. Διαθέσιμο στο: http://hdl.handle.net/11419/1226

4. Gkountava, Efrossini. (2013). Οπτικοποίηση δεδομένων (data visualization). 10.13140/2.1.4895.5844.

5. Hofmann, M., & Klinkenberg, R. (Eds.). (2013). RapidMiner: Data mining use cases and business analytics applications.

6. Bagga S. and Singh. G.N., (2012). Applications of Data Mining. International Journal for Science and Emerging Technologies with Latest Trends.

7. Abdulmohsen Algarni, (2016) Data Mining in Education. International Journal of Advanced Computer Science and Applications (IJACSA), 7(6). http://dx.doi.org/10.14569/IJACSA.2016.070659

8. Janez Demsar, Tomaz Curk, Ales Erjavec et al., Orange: Data Mining Toolbox in Python retrieved from https://www.jmlr.org/papers/volume14/demsar13a/demsar13a.pdf

9. Algarni, Abdulmohsen. (2016). Data Mining in Education. International Journal of Advanced Computer Science and Applications. 7. 10.14569/IJACSA.2016.070659.

10. Μάγγα, Ειρήνη (2021), Εξόρυξη Δεδομένων στα Πληροφοριακά Συστήματα της Δημόσιας Διοίκησης (Μεταπτυχιακή διπλωματική εργασία). Πανεπιστήμιο Δυτικής Αττικής. https://polynoe.lib.uniwa.gr/xmlui/handle/11400/164

11. Στρατογιάννη Ιωάννα (2019), Εφαρμογή τεχνικών εξόρυξης δεδομένων σε e-learning συστήματα (Διπλωματική εργασία). Πανεπιστήμιο Πατρών. http://hdl.handle.net/10889/13611

12. Οικονόμου Σταύρος (2020), Αλγόριθμοι Μηχανικής Μάθησης για την εξόρυξη δεδομένων, (Διπλωματική Εργασία). Ελληνικό Ανοικτό Πανεπιστήμιο. https://apothesis.eap.gr/handle/repo/48997

13. Βαμβακούσης Βασίλειος, Θεοχάρη Αικατερίνη (2019), Data Mining - Από την αποκάλυψη στην εφαρμογή (Πτυχιακή εργασία). ΤΕΙ Δυτικής Ελλάδας. http://repository.library.teimes.gr/xmlui/handle/123456789/7509

14. Τσούμας Ηλίας (2016), Συλλογή Δεδομένων και Εξόρυξη Γνώσης από Κοινωνικά Δίκτυα (Πτυχιακή εργασία). Πανεπιστήμιο Πειραιώς. https://dione.lib.unipi.gr/xmlui/handle/unipi/9467

15. https://rapidminer.com/ [Accessed June 2021].

16. https://orangedatamining.com/ [Accessed June 2021].

17. https://el.wikipedia.org/wiki/Εξ_αποστάσεως_εκπαίδευση [Accessed June 2021].

18. https://www.openeclass.org [Accessed June 2021].

19. https://en.wikipedia.org/wiki/Microsoft_Teams [Accessed June 2021].