Κεφάλαιο 2
Η Εξόρυξη Δεδομένων ως Διαδικασία

Στόχοι

Στόχος του κεφαλαίου είναι η περιγραφή τριών αφηρημένων μοντέλων τα οποία ρίχνουν κάποιο φως στη διαδικασία την οποία πρέπει να εφαρμόσουμε, για να μπορέσουμε να εξάγουμε γνώση σε μορφή προτύπων ή μοντέλων από τα δεδομένα που διαθέτουμε.

Προσδοκώμενα Αποτελέσματα

Με την ολοκλήρωση της μελέτης του κεφαλαίου, ο αναγνώστης θα είναι σε θέση να

- περιγράφει σχηματικά το μοντέλο ωρίμανσης της αναλυτικής δεδομένων,
- περιγράφει σχηματικά τη διαδικασία ανακάλυψης γνώσης από βάσεις δεδομένων,
- αναφέρει τα 5 βασικά στάδια της διαδικασίας ανακάλυψης γνώσης από βάσεις δεδομένων,
- περιγράφει το πρότυπο CRICP-DM που μοντελοποιεί τη διαδικασία ανακάλυψης γνώσης σε βάσεις δεδομένων.

Έννοιες – Κλειδιά

Ανακάλυψη γνώσης από βάσεις δεδομένων
Αναλυτική δεδομένων
Εξόρυξη Δεδομένων
Πρότυπο CRISP-DM

Προαπαιτούμενες γνώσεις

Το κεφάλαιο είναι εισαγωγικό και δεν είναι απαραίτητες προηγούμενες γνώσεις.

Εισαγωγικές Παρατηρήσεις

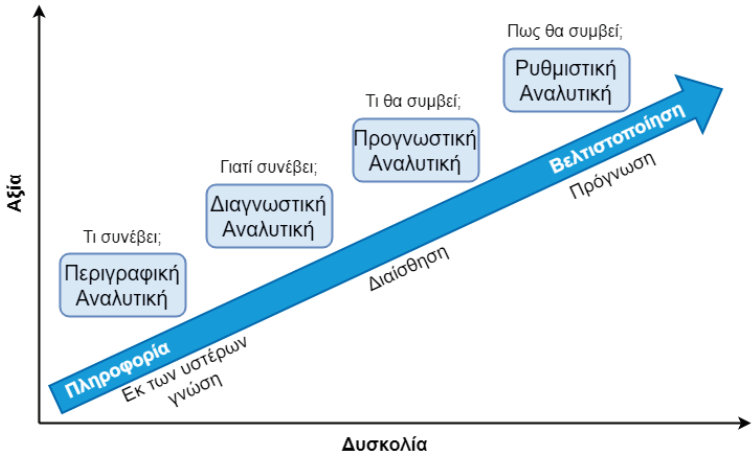
Στο παρόν κεφάλαιο δίνεται έμφαση στη διαδικασία της Εξόρυξης Γνώσης από Βάσεις Δεδομένων μέσω της οποίας παρέχεται ένας τρόπος τυποποίησης του πλαισίου με το οποίο ένας Επιστήμονας Δεδομένων ανακαλύπτει καινούργια πρότυπα ή μοτίβα, με γνώμονα την αποφυγή αστοχιών και άλλων ατελειών που αποπροσανατολίζουν από το επιθυμητό αποτέλεσμα.

2.1 Μοντέλο Ωρίμανσης της Αναλυτικής Δεδομένων

Όπως αναφέραμε σε γενικές γραμμές στην Εισαγωγή, η Επιστήμη των Δεδομένων έχει σαν στόχο τη μελέτη των βασικών αρχών που διέπουν την ανεύρεση (συνήθως χρησιμοποιείται ο όρος ανακάλυψη) μοντέλων ή προτύπων που είτε περιγράφουν κάποιο ενδιαφέρον φαινόμενο, είτε εξηγούν το λόγο για τον οποίο συνέβη κάποιο φαινόμενο, είτε προβλέπουν τι μπορεί να συμβεί, είτε τέλος κατά μία έννοια «συνταγογραφούν» (μετάφραση από τον όρο prescribe) τον τρόπο με τον οποίο μπορεί να επιτευχθεί μελλοντικά κάποιος επιθυμητός στόχος. Οι τέσσερεις παραπάνω εκδοχές των πιθανών αποτελεσμάτων μίας διαδικασίας εξόρυξης δεδομένων έχουν αποτυπωθεί σε ένα γράφημα, που έχει δημιουργήσει η εταιρία συμβούλων Gartner (βλέπε Σχήμα 2.1) και παρουσιάζει την αξία που αποκομίζεται σε σχέση με το επίπεδο δυσκολίας κάποια συγκεκριμένης εργασίας ή την προσπάθεια που διοχετεύουν οι επιστήμονες στην εν λόγω εργασία.

Για να είμαστε σε θέση να αναρριχηθούμε επάνω σε αυτή την υποθετική κλίμακα της αξίας των αποτελεσμάτων, από τα εννοιολογικά πιο εύκολα, στα πιο σύνθετα και περίπλοκα, και κυρίως για να έχουμε ένα αναπαραγώγιμο αποτέλεσμα, έχει δείξει τόσο η θεωρία όσο και η πράξη, ότι χρειαζόμαστε μία καλά ορισμένη διαδικασία με διακριτές φάσεις, τις οποίες εάν ακολουθήσουμε, θα είμαστε σε κάποιο βαθμό σίγουροι ότι θα έχουμε αποφύγει συχνά επαναλαμβανόμενα λάθη, που τις περισσότερες φορές προεξοφλούν την αποτυχία τέτοιων εργασιών. Για αυτούς από τους αναγνώστες, που έχουν ένα υπόβαθρο στην Μηχανική του Λογισμικού, η διαδικασία της ανακάλυψης νέων μοντέλων κάθε είδους, μοιάζει σε κάποιο

βαθμό με τα μοντέλα και τις μεθοδολογίες ανάπτυξης λογισμικού, εάν και υπάρχουν και αρκετές διαφορές με αυτά, για τις οποίες πρέπει να είναι κάποιος ενήμερος, και να μην τις συγχέει.



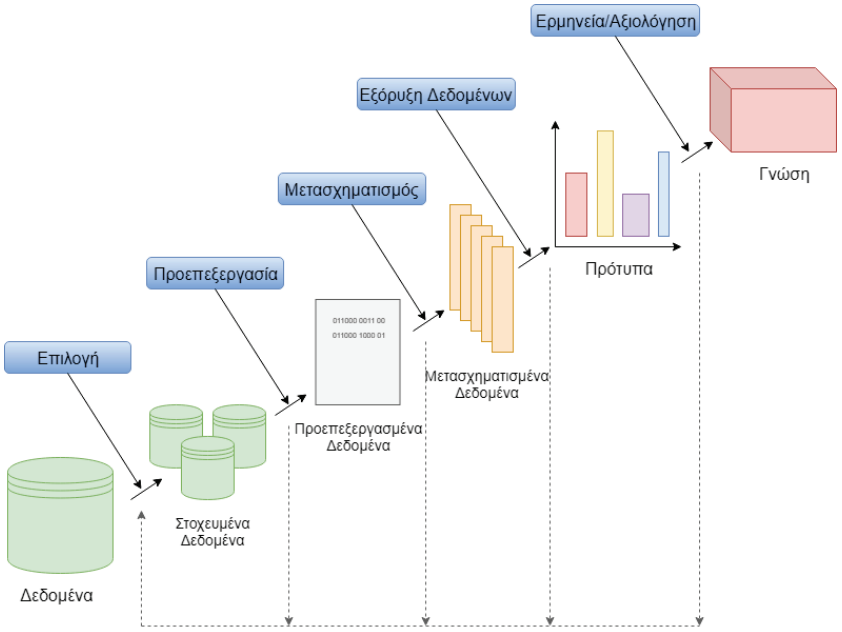
Σχήμα 2.1 *Μοντέλο Ορίμανσης της Αναλυτικής των Δεδομένων*

2.2 Διαδικασία Ανακάλυψης Γνώσης από Βάσεις Δεδομένων

Σε κάθε περίπτωση, η διαδικασία της ανακάλυψης γνώσης από βάσεις δεδομένων έχει προτυποποιηθεί και η προτυποποίηση αυτή έχει οδηγήσει στο γνωστό πρότυπο Cross-Industry Standard Process for Data Mining (CRISP-DM) το οποίο αναπτύχθηκε το 1996 από μία ομάδα εταιριών (βλέπε (Larose, 2005)). Το πρότυπο αυτό αποτελεί μία μετεξέλιξη, ή τουλάχιστον φαίνεται να βασίζεται σε μία αναπαράσταση των βημάτων της ανακάλυψης γνώσης από βάσεις δεδομένων, η οποία παρουσιάζεται στο Σχήμα 2.2.

Το σημαντικό στοιχείο που μπορεί να παρατηρήσει κανείς στο Σχήμα 2.2 είναι το γεγονός της ύπαρξης συγκεκριμένων βημάτων που διαδέχεται το ένα το άλλο, όπως επίσης και οι διακεκομμένες γραμμές οπισθοδρόμησης που οδηγούν σε προηγούμενες φάσεις της διαδικασίας. Η ουσία των γραμμών αυτών έχει να κάνει με το γεγονός ότι δεν νομοτελειακά σίγουρο πόσες επαναλήψεις θα χρειαστούν για

να έχει κάποιος επιστήμονας δεδομένων ένα σαφές αποτέλεσμα, το οποίο να πληροί τις προϋποθέσεις των χαρακτηριστικών που προδιαγράφει η θεωρία ως προς την ποιότητα, την καινοτομία, την εφαρμοσιμότητα της παραχθείσας γνώσης. Επίσης ισοδύναμα μπορούμε να πούμε ότι η εφαρμογή του συνόλου των βημάτων εφάπαξ δεν προεξοφλεί την επίλυση ενός προβλήματος.

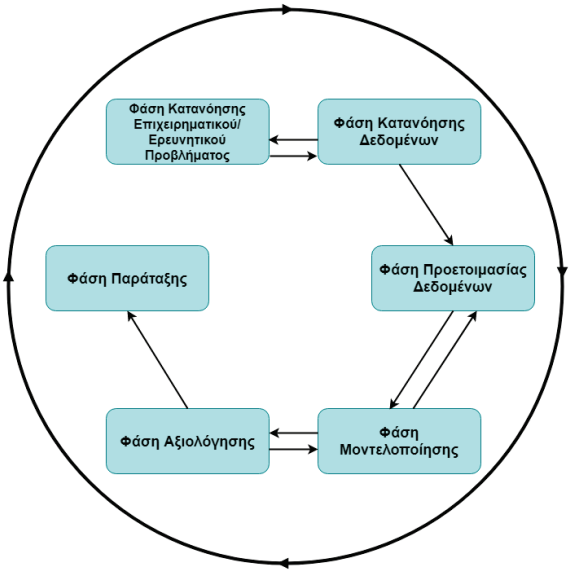


Σχήμα 2.2 *Αναπαράσταση Διαδικασίας Ανακάλυψης Γνώσης από Βάσεις Δεδομένων*

2.3 Το CRISP-DM Μοντέλο

Το CRISP-DM πρότυπο που αναφέραμε μόλις προηγουμένως μεταθέτει το πλαίσιο της διαδικασίας που φαίνεται να εστιάζει μόνο στα δεδομένα, στην επίλυση ενός προβλήματος με την χρήση των δεδομένων. Το CRISP-DM πρότυπο φαίνεται στο Σχήμα 2.3.

Το CRISP-DM μοντέλο αναδεικνύει πρωτίστως την ουσία της επαναληψιμότητας της διαδικασίας της εξόρυξης γνώσης, ενώ υποδεικνύει ότι η διαδικασία χρειάζεται να προσαρμοστεί στο συγκεκριμένο πρόβλημα, ακολουθώντας τις μεταβάσεις εκείνες, όσες φορές αυτό χρειαστεί, που απαιτούνται κατά περίπτωση.



Σχήμα 2.3 *Το CRISP-DM μοντέλο*

Ακολουθώντας τις διάφορες φάσεις, βλέπουμε ότι η διαδικασία ξεκινάει από τη κατανόηση του προβλήματος που καλείται να λύσει μία ομάδα επιστημόνων-αναλυτών. Σε αυτό το σημείο πρέπει να γίνει σαφές το γεγονός ότι οι ομάδες που εμπλέκονται στην επίλυση ενός επιχειρηματικού προβλήματος, δεν έχουν όλες το ίδιο υπόβαθρο, ή δεν γνωρίζουν εξίσου καλά το πρόβλημα. Στην πράξη αυτό που συμβαίνει είναι ότι σε μία επιχείρηση, για παράδειγμα η ομάδα Μάρκετινγκ θέλει να τρέξει μία διαφημιστική καμπάνια για την προσέλκυση πελατών, και ζητάει τη βοήθεια της ομάδας των Επιστημόνων Δεδομένων η οποία μπορεί αναλύοντας ιστορικά δεδομένα, να βγάλει κάποια πολύτιμα πρότυπα και συμπεράσματα, για το συγκεκριμένο κομμάτι του αγοραστικού κοινού που θα πρέπει να προσεγγίσει η ομάδα Μάρκετινγκ, όπως επίσης και τον τρόπο με τον οποίο θα πρέπει αυτό το πληθυσμιακό κομμάτι να προσεγγιστεί, για να έχει στο τέλος της ημέρας η εταιρία το μέγιστο δυνατό κέρδος βάσει του συγκεκριμένου προϋπολογισμού που διατίθεται για την συγκεκριμένη καμπάνια.

Πολλές φορές τα δεδομένα που βοηθούν στην επίλυση ενός προβλήματος είναι πολλά. Επίσης τα δεδομένα αυτά μπορεί να είναι διασκορπισμένα μέσα στην εταιρία, σε βάσεις δεδομένων άλλων ομάδων, όπως για παράδειγμα στο τμήμα πωλήσεων για παράδειγμα. Υπάρχει επίσης η πιθανότητα κάποια από τα δεδομένα που

φαίνεται να απαιτεί η εφαρμογή να μην υπάρχουν καν στην εταιρία. Υπάρχει η πιθανότητα τα δεδομένα αυτά να πρέπει να τα πάρει η εταιρία από αλλού, ή να τα αναζητήσει στο διαδίκτυο, ή ακόμα και να πληρώσει για αυτά. Μία προσέγγιση στην επίλυση ενός επιχειρηματικού προβλήματος με τη χρήση της Επιστήμης των Δεδομένων θεωρεί ότι τα δεδομένα, όπως και οι δεξιότητες της ομάδας των Επιστημόνων Δεδομένων, αποτελούν πραγματικό κεφάλαιο στρατηγικής σημασίας για την εταιρία, γεγονός που υποδηλώνει πως μερικές φορές θα πρέπει να δαπανηθούν χρήματα προκειμένου κάποια συγκεκριμένα δεδομένα να αποκτηθούν στην εταιρία. Οπουδήποτε πάντως και εάν βρίσκονται αυτά τα δεδομένα, τα δεδομένα αυτά τις περισσότερες των περιπτώσεων δεν έχουν συλλεχθεί ή δημιουργηθεί με γνώμονα την ανάγκη της εταιρίας να κτίσει μοντέλα από αυτά. Αυτό σημαίνει ότι τα δεδομένα αυτά θα πρέπει να κατανοηθούν, να καθαριστούν πολλές φορές από άχρηστες πληροφορίες που δυσκολεύουν το έργο της εξόρυξης δεδομένων, να μειωθούν πολλές από άσχετες προς το στόχο πληροφορίες, και ίσως να μετασχηματιστούν σε κάποια άλλη μορφή που θα κάνει πιο εύκολη την επεξεργασία τους. Αναφέραμε στο Σχήμα 2.2 κάποια από αυτά τα στάδια με μεγαλύτερη λεπτομέρεια. Οι παραπάνω δύο φάσεις της CRISP-DM διαδικασίας αποτελούν την κατανόηση και την προετοιμασία των δεδομένων.

Μετά την κατανόηση και την προετοιμασία των δεδομένων, τα δεδομένα πρέπει να αναλυθούν με κάποιο συγκεκριμένο στόχο, για να μπορέσουμε να προχωρήσουμε στην παραγωγή των μοντέλων που θα είναι σε θέση να μας βοηθήσουν να λύσουμε το συγκεκριμένο επιχειρηματικό πρόβλημα. Για παράδειγμα σε αυτή τη φάση μπορεί το ζητούμενο να είναι η επαγωγή ενός μοντέλου Απλοϊκού Κατηγοριοποιητή Bayes, για την δημιουργία του οποίου είναι απαραίτητος ο υπολογισμός κάποιων εκ των προτέρων και υπό συνθήκη πιθανοτήτων. Αυτές οι πληροφορίες μπορούν να υπολογιστούν μόνο από ιστορικά δεδομένα, για τα οποία ο στόχος θα είναι γνωστός.

Μετά την παραγωγή των μοντέλων θα πρέπει να περάσουμε στην αξιολόγηση των μοντέλων αυτών, χρησιμοποιώντας κάποιες μετρικές και διαστήματα εμπιστοσύνης που μας ενημερώνουν για το πόσο ενδιαφέροντα είναι τα μοντέλα, πόσο αντιπροσωπευτικά για τα δεδομένα που συνοψίζουν, όπως επίσης πόση εμπιστοσύνη θα πρέπει να έχουμε σε αυτά. Η φάση της αξιολόγησης των μοντέλων μπορεί να γίνει με διάφορους τρόπους, ένας από τους πιο διαδεδομένους όπως θα συζητήσουμε στη συνέχεια είναι το λεγόμενο cross-validation.

Στο τελικό στάδιο θα πρέπει να γίνει η κατανόηση των παραχθέντων μοντέλων και η αξιολόγησή τους με γνώμονα το πώς μπορούν να χρησιμοποιηθούν στην πράξη για να μπορέσουμε να έχουμε κάποιο όφελος από αυτά. Άλλες φορές πάλι θα πρέπει να δούμε πως ταιριάζουν με άλλα μοντέλα που ήδη χρησιμοποιούνται στην καθημερινή λειτουργία της εταιρίας, και να αποφασίσουμε με ποιο τρόπο μπορεί να γίνει πιθανή συγχώνευσή τους με αυτά.

Σύνοψη

Στο κεφάλαιο αυτό κάναμε παρουσιάσαμε τη διαδικασία αναλυτικής των δεδομένων και ανακάλυψης γνώσης από βάσεις δεδομένων, εισάγοντας έτσι τον αναγνώστη στις βασικές έννοιες και διαδικασίες που ακολουθούνται για την εξόρυξη πληροφορίας και την εξαγωγή συμπερασμάτων. Το πρότυπο CRISP-DM μοντελοποιεί τη διαδικασία ανακάλυψης γνώσης, εστιάζοντας στα δεδομένα.

Ο αναγνώστης μπορεί να ανατρέξει σε πλούσια βιβλιογραφία για να εμβαθύνει και να εμπλουτίσει τις γνώσεις του. Ενδεικτικά αναφέρουμε τα βιβλία (Larose, 2005), (Provost & Fawcett, 2013), (Tan, Steinbach, Karpatne, & Kumar, 2018), (Zaki & Wagner, 2014). Μια πιο πρακτική προσέγγιση μπορείτε να βρείτε στα (Frank, Hall, & Witten, 2016) και (Siegel, 2016). Σημαντική πηγή αποτελούν και τα πρακτικά του διεθνούς συνεδρίου KDD'96 (Simoudis, Han, & Fayyad, 1996) αφιερωμένο στην ανακάλυψη γνώσης και στην εξόρυξη δεδομένων. Επί της ευκαιρίας, αναφέρουμε τα βιβλία (Kabacoff, 2011), (Matloff, 2011), (Zhao, 2013) και (Williams, 2011) που αφορούν τον προγραμματισμό σε γλώσσα R και το ολοκληρωμένο περιβάλλον Ruttle, που θα χρησιμοποιήσουμε στα Κεφάλαια που ακολουθούν για την υλοποίηση των τεχνικών και των μεθόδων που θα περιγράψουμε.

2.4 Βιβλιογραφικές Πηγές

Frank, E., Hall, M. A., & Witten, I. H. (2016). *Data Mining: Practical Machine Learning Tools and Techniques* (4th ed.). Morgan Kaufmann.

Kabacoff, R. I. (2011). *R in Action*. Manning.

Larose, D. T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*. Wiley Blackwell. https://doi.org/10.1002/0471687545

Matloff, N. S. (2011). *The art of R programming : tour of statistical software design*. No Starch Press.

Provost, F., & Fawcett, T. (2013). *Data science for business : [what you need to know about data mining and data-analytic thinking]*. O'Reilly.

Siegel, E. (2016). *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*. Wiley. Retrieved from https://www.predictiveanalyticsworld.com/book/overview.php

Simoudis, E., Han, J., & Fayyad, U. (1996). Proceedings of the Second Knowledge Discovery and Data Mining Conference Contents. Association for the Advancement of Artificial Intelligence.

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2018). *Introduction to data mining* (2nd ed.). Pearson.

Williams, G. J. (2011). *Data mining with Rattle and R : the art of excavating data for knowledge discovery*. Springer.

Zaki, M. J., & Wagner, M. J. (2014). *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press.

Zhao, Y. (2013). *R and Data Mining*. Elsevier Inc. https://doi.org/10.1016/C2011-0-06686-3

Κεφάλαιο 3
Περιβάλλοντα και Εργαλεία

Στόχοι

Στόχος του κεφαλαίου είναι η εισαγωγή του αναγνώστη στη γλώσσα προγραμματισμού R, χρησιμοποιώντας είναι τα πιο δημοφιλή γραφικά περιβάλλοντα και εργαλεία για την ανάπτυξη προγραμμάτων σε R. Περιγράφονται στοιχειώσεις εντολές, τύπους δεδομένων, βασικές πράξεις, δομές ελέγχου και επανάληψης, καθώς και βασικές εργασίες που μπορεί να πραγματοποιήσει ο προγραμματιστής με χρήση της R. Επίσης, γίνεται αναφορά σε πακέτα (βιβλιοθήκες) της R που περιλαμβάνουν συναρτήσεις ιδιαίτερα χρήσιμες για την διαχείριση και ανάλυση των δεδομένων και για την οπτική αναπαράσταση τους.

Προσδοκώμενα Αποτελέσματα

Με την ολοκλήρωση της μελέτης του κεφαλαίου, ο αναγνώστης θα είναι σε θέση να

- Εγκαθιστά την πρόσφατη έκδοση της R και να περιηγείται στο γραφικό της περιβάλλον
- Συντάσσει και να εκτελεί βασικές εντολές της R
- Δημιουργεί και να εκτελεί αρχεία δέσμης εντολών της R
- Αναζητά βοήθεια για τον τρόπο χρήσης συναρτήσεων της R
- Ορίζει αντικείμενα διαφόρων κλάσεων και να εκτελεί πράξεις μεταξύ τους
- Γνωρίζει και χρησιμοποιεί βασικές δομές ελέγχου και επανάληψης

- Καλεί έτοιμες συναρτήσεις αλλά και να δημιουργεί τις δικές του συναρτήσεις
- Χρησιμοποιεί αναδρομικές συναρτήσεις επαναληπτικές συναρτήσεις
- Εκτελεί βασικές λειτουργείες που αφορούν στη διαχείριση συνόλων δεδομένων
- Παράγει ακολουθίες δεδομένων και τυχαιοποιημένα δεδομένα
- Εγκαθιστά και να φορτώνει πακέτα της R και χρησιμοποιεί της συναρτήσεις τους
- Γνωρίζει κάποια δημοφιλή γραφικά περιβάλλοντα για την ανάπτυξη προγραμμάτων σε R

Έννοιες – Κλειδιά

ακέραιος (integer)
ακολουθία
αναδρομική συνάρτηση
αναζήτηση (searching)
αντικείμενο (object)
αριθμητικός (arithmetic)
αρχείο δέσμης εντολών (script)
ατομική κλάση αντικειμένων
βιβλιοθήκη (library)
γεννήτρια ψευδοτυχαίων αριθμών
γραφική διεπαφή χρήστη (GUI)
δειγματοληψία
διάνυσμα
διανυσματοποίηση
δομή επανάληψης
δομή ελέγχου
ειδική τιμή
εξαναγκασμός (coercion)
επαναληπτική συνάρτηση
κατανομή
κλάση (class)
κονσόλα
λίστα (list)
λογικός (logical)
πακέτο (package)
παράγοντας (factor)
πλαίσιο δεδομένων (data frame)
συνάρτηση
σύνθετος (complex)
σύνολο δεδομένων (data set)
ταξινόμηση (sorting)
τυχαιοποιημένα δεδομένα
τελεστής (operator)
υποσύνολο δεδομένων
χαρακτήρας (character)
χαρακτηριστικό (attribute)
χώρος εργασίας (workspace)
dplyr
ggplot2
Jupiter Notebook

mlr
OpenML
R
Rattle
Rstudio

Προαπαιτούμενες γνώσεις

Το περιεχόμενο του κεφαλαίου είναι αυτόνομο και δεν απαιτεί προηγούμενη γνώση.

Εισαγωγικές Παρατηρήσεις

Το παρόν κεφάλαιο αποτελεί μια σύντομη εισαγωγή στα πιο δημοφιλή γραφικά περιβάλλοντα και εργαλεία της R. Πρόκειται για γραφικές διεπαφές χρήστη (GUIs) που παρέχουν διευκολύνσεις στη σύνταξη και εκτέλεση εντολών της R αλλά πακέτα που παρέχουν ισχυρές συναρτήσεις για την επεξεργασία, ανάλυση και γραφική αναπαράσταση των δεδομένων. Θα ήταν παράληψη να αναφέρουμε σε αυτό το σημείο στοιχειώσεις έννοιες και βασικές εντολές της R που αφορούν τύπους δεδομένων και πράξεις μεταξύ αυτών, δομές προγραμματισμού και χρήση συναρτήσεων. Πλούσια είναι η βιβλιογραφία στην οποία μπορεί κανείς να ανατρέξει προκειμένου να εμβαθύνει στον προγραμματισμό στην R. Ενδεικτικά αναφέρουμε τις διαδικτυακές πηγές (Venables & Smith, 1997), και (Peng, n.d.)

Μελέτη Περίπτωσης

Για την παρουσίαση των παραδειγμάτων του κεφαλαίου θα χρησιμοποιήσουμε το σύνολο δεδομένων airquality, το οποίο είναι ένα από τα διαθέσιμα σύνολα δεδομένων που είναι διαθέσιμο στη διανομή της R (είναι ενσωματωμένο στο πακέτο `datasets`). Περιλαμβάνει συνολικά 153 παρατηρήσεις για 6 χαρακτηριστικά που αφορούν την ημερήσια ποιότητα του ατμοσφαιρικού αέρα στη Νέα Υόρκη: μέση τιμή όζοντος (Ozone), Ηλιακή ακτινοβολία (Solar.R), μέση ταχύτητα ανέμου (Wind), μέγιστη θερμοκρασία (Temp), μήνας (Month) και ημέρα (Day). Περισσότερες πληροφορίες και μεταδεδομένα μπορείτε να βρείτε στον σύνδεσμο https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/airquality.html.

Η εισαγωγή (ανάγνωση) του συνόλου δεδομένων στην R γίνεται με την εντολή