ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΘΕΜΑ: Δημιουργία Εκπαιδευτικού περιεχομένου για την «Εξόρυξη δεδομένων με χρήση τεχνικών μηχανικής μάθησης» (Educational Content for "Data Mining Using Machine Learning Techniques")

Εκπόνηση διπλωματικής εργασίας: Αναστάσιος Τσολακίδης (ΑΜ 18390283)

Επιβλέπων Καθηγητής: Χρήστος Σκουρλάς

Αθήνα, Ιούλιος 2021

UNIVERSITY OF WEST ATTICA
FACULTY OF ENGINEERING
DEPARTMENT OF INFORMATICS AND COMPUTER SCIENCE

Diploma Thesis

Educational Content for "Data Mining Using Machine Learning Techniques"

Student: Anastasios Tsolakidis
Registration Number: 18390283

Supervisor: Prof. Christos Skourlas

Athens, July 2021

Εγκρίθηκε από τριμελή εξεταστική επιτροπή

Αθήνα, Ιούλιος 2021

ΕΠΙΤΡΟΠΗ ΑΞΙΟΛΟΓΗΣΗΣ

Επιβλέπων καθηγητής:
Χρήστος Σκουρλάς

Christos Skourlas
Digitally signed by Christos Skourlas
Date: 2021.07.23 09:19:44 +03'00'

Μέλος επιτροπής:
Κλειώ Σγουροπούλου

KLEIO SGOUROPOULOU
Digitally signed by KLEIO SGOUROPOULOU
Date: 2021.07.28 12:39:03 +03'00'

Μέλος επιτροπής:
Βασίλειος Μάμαλης

Vasileios Mamalis
Digitally signed by Vasileios Mamalis
Date: 2021.07.23 22:44:26 +03'00'

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΠΤΥΧΙΑΚΗΣ/ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Ο κάτωθι υπογεγραμμένος Αναστάσιος Τσολακίδης του Γεωργίου, με αριθμό μητρώου 18390283 φοιτητής του Πανεπιστημίου Δυτικής Αττικής της Σχολής Μηχανικών του Τμήματος Μηχανικών Πληροφορικής και Υπολογιστών, δηλώνω υπεύθυνα ότι:

«Είμαι συγγραφέας αυτής της πτυχιακής/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».

«Με επιφύλαξη παντός δικαιώματος δηλώνω υπεύθυνα και γνωρίζοντας τις κυρώσεις του Ν. 2121/1993 περί Πνευματικής Ιδιοκτησίας, ότι η παρούσα πτυχιακή εργασία είναι εξ ολοκλήρου αποτέλεσμα δικής μου ερευνητικής εργασίας, δεν αποτελεί προϊόν αντιγραφής ούτε προέρχεται από ανάθεση σε τρίτους. Όλες οι πηγές που χρησιμοποιήθηκαν (κάθε είδους, μορφής και προέλευσης) για τη συγγραφή της περιλαμβάνονται στη βιβλιογραφία».

Ο Δηλών

Αναστάσιος Τσολακίδης

Copyright © Με επιφύλαξη παντός δικαιώματος. All rights reserved.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ και Αναστάσιος Τσολακίδης, Ιούλιος, 2021

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τους συγγραφείς.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τη συγγραφέα του και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις θέσεις του επιβλέποντος, της επιτροπής εξέτασης ή τις επίσημες θέσεις του Τμήματος και του Ιδρύματος.

Περίληψη

Η εξόρυξη δεδομένων τα τελευταία χρόνια εμφανίζει πολύ μεγάλη εξέλιξη και αποτελεί έναν από τους βασικούς τομείς ανάλυσης δεδομένων στο τομέα της πληροφορικής. Η χρήση των αλγορίθμων της εξόρυξης δεδομένων αυξάνεται συνεχώς καθώς είναι εφικτή η επίλυση προβλημάτων τα οποία περιέχουν μεγάλα δεδομένα και τα οποία δεν θα μπορούσαμε να διαχειριστούμε με τους παραδοσιακούς τρόπους ανάλυσης. Συνεπώς η γνώση των μεθόδων εξόρυξης γνώσης και της μηχανικής μάθησης αποτελούν απαραίτητα εφόδια για αυτούς που θέλουν να ασχοληθούν με τον τομέα της ανάλυσης δεδομένων. Οι ψηφιακές δεξιότητες και ικανότητες που απαιτούνται μπορούν να αποκτηθούν διαδικτυακά μέσω της εξ αποστάσεως εκπαίδευσης.

Η εξ αποστάσεως εκπαίδευση δίνει την δυνατότητα στους εκπαιδευόμενους να παρακολουθήσουν ασύγχρονα τις θεματικές ενότητες του μαθήματος κάνοντας χρήση των σημειώσεων και του οπτικοακουστικού υλικού. Στην παρούσα διπλωματική εργασία δημιουργήθηκε ένα διαδικτυακό μάθημα με σκοπό την ανάπτυξη των απαραίτητων γνώσεων και δεξιοτήτων για την εξόρυξη γνώσης από βάσεις δεδομένων, με χρήση σύγχρονων τεχνικών και μεθοδολογιών, Στόχος είναι η υποβοήθηση της λήψης αποφάσεων για μια εταιρία ή έναν οργανισμό. Οι εκπαιδευόμενοι έρχονται σε επαφή με τα χρήσιμα εργαλεία για το σχεδιασμό, ανάπτυξη και εφαρμογή ενός συστήματος εξόρυξης γνώσης και λήψης αποφάσεων. Με τον τρόπο αυτό καλύπτονται οι ανάγκες που προκύπτουν από τη λειτουργία πληροφοριακών συστημάτων προς την κατεύθυνση της πρόβλεψης πιθανών καταστάσεων, την κατηγοριοποίηση των δεδομένων και την εξαγωγή χρήσιμων συμπερασμάτων.

Το μάθημα περιλαμβάνει έξι (6) διδακτικές ενότητες και έχει σχεδιαστεί με γνώμονα να παρέχει τις απαραίτητες γνώσεις στους νέους επιστήμονες του χώρου της ανάλυσης δεδομένων:

- ΔΕ1: Μεθοδολογία Διαχείρισης γνώσης από βάσεις δεδομένων
- ΔΕ2: Επεξεργασία και διαχείριση δεδομένων και μεταβλητών
- ΔΕ3:Μελέτη , σχεδιασμός και εφαρμογή τεχνικών Κατηγοριοποίησης (Classification)
- ΔΕ4: Μελέτη, σχεδιασμός και εφαρμογή τεχνικών Συσταδοποίησης (clustering)
- ΔΕ5: Μελέτη, σχεδιασμός και εφαρμογή τεχνικών Πρόβλεψης (Prediction).
- ΔΕ6: Μελέτη, σχεδιασμός και εφαρμογή τεχνικών Παλινδρόμησης (Regression).

Λέξεις Κλειδιά: Εξόρυξη Δεδομένων, e-learning, Κατηγοριοποίηση, Συσταδοποίηση, Παλινδρόμηση

Abstract

Nowadays, data mining has shown great development and is one of the key areas of data analysis in the field of information technology. The use of data mining algorithms is constantly increasing as it is possible to solve problems that contain big data and complex problems which could not be managed by traditional ways of analysis. Therefore, data mining and machine learning are prerequisite digital skills and competencies for those scientists who want to deal with data analysis and these skills can be acquired through online distance learning.

Distance learning schemes enable learners to attend the course through online educational material. The goal of the online course is the following: to develop the digital skills (of the students-learners) for extracting knowledge from databases, using modern techniques and methodologies. Another goal of the course is to assist learners in making decisions for a company or an organization. In the framework of the course, Learners work with useful tools for designing, developing and implementing a knowledge mining and decision-making system which is necessary for the operation of information systems in the direction of predicting possible situations, categorizing data and extracting useful conclusions. The course includes six (6) Units and is designed to provide the necessary knowledge to new scientists in the field of data analysis:

- DE1: Database Knowledge Management Methodology
- DE2: Processing and management of data and variables
- DE3: Study, design and application of Classification techniques
- DE4: Study, design and application of clustering techniques
- DE5: Study, design and application of Prediction techniques.
- DE6: Study, design and application of Regression techniques.

Keywords: Data Mining, E-learning, Clustering, Classification, Prediction, Regression

Πίνακας περιεχομένων

1 *Εισαγωγή στη σχεδίαση περιεχομένου για την εξόρυξη δεδομένων με χρήση τεχνικών μηχανικής μάθησης*

Η Εξόρυξη Δεδομένων είναι ένας τομέας που έχει πολλά κοινά θέματα με τους τομείς της Τεχνητής Νοημοσύνης και των Συστημάτων Βάσεων Δεδομένων. Η εξόρυξη δεδομένων ορίζεται ως η "non trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" («μη τετριμένη διαδικασία αναγνώρισης έγκυρων, νέων, δυνητικά χρήσιμων και τελικά κατανοητών προτύπων στα δεδομένα») (Fayyad et al., 1996).

Σε μία περιγραφική προσέγγιση η μηχανική μάθηση ασχολείται με τον τρόπο κατασκευής προγραμμάτων υπολογιστών που βελτιώνονται αυτόματα («μαθαίνουν») με την εμπειρία. Επομένως, θα μπορούσαμε να πούμε ότι η μηχανική μάθηση επικαλύπτεται σε μεγάλο βαθμό με την εξόρυξη δεδομένων, καθώς πολλοί από τους αλγόριθμούς της μαθαίνουν από δεδομένα.

Στόχος της παρούσας εργασίας είναι η σχεδίαση ενός μαθήματος στην εξόρυξη δεδομένων και στη χρήση τεχνικών που χρησιμοποιούνται στη μηχανική μάθηση.

Περιεχόμενο μαθήματος

Η σχεδίαση περιεχομένου για το μάθημα της «Εξόρυξης δεδομένων με χρήση τεχνικών μηχανικής μάθησης» απασχόλησε και απασχολεί τη βιβλιογραφία. Ο Musicant (Musicant, 2006) πιστεύει ότι σε προπτυχιακό επίπεδο ένα μάθημα όπως αυτό παρέχει την ευκαιρία στους διδασκόμενους να αποκτήσουν ερευνητικές δεξιότητες, αλλά και να εξοικειωθούν περαιτέρω με μαθήματα όπως οι δομές δεδομένων και να κατανοήσουν σε βάθος αλγόριθμούς και έννοιες της αλγοριθμικής γενικότερα.

Προτείνει μια καλά ορισμένη σειρά μαθημάτων εξόρυξης δεδομένων που θα βασίζονται:

1) στη μελέτη ερευνητικών άρθρων που θα αποτελούν το «πρωτεύον υλικό μελέτης» ανάγνωσης για το μάθημα και
2) στη χρήση και υλοποίηση αλγορίθμων για τις εργασίες του μαθήματος (assignments and project)

Πιστεύει ότι ένα τέτοιο μάθημα πρέπει να παρέχεται στους φοιτητές χωρίς προϋποθέσεις πέρα από κάποιες βασικές γνώσεις δομών δεδομένων. Επιπλέον, πιστεύει ότι το μάθημα αυτό «επιτρέπει στους μαθητές να βιώσουν τόσο εφαρμοσμένη όσο και θεωρητική εργασία σε μια σειρά αντικειμένων που καλύπτουν πολλαπλούς τομείς της επιστήμης των υπολογιστών» (‘allows students to experience both applied and theoretical work in a discipline that straddles multiple areas of computer science”). Επισημαίνεται ότι ο συγγραφέας περιγράφει με λεπτομέρεια κάθε πτυχή της σχεδίασης και της διδασκαλίας του μαθήματος.

Οι King και Satyanarayana (2013) εστιάζουν στη διδασκαλία της εξόρυξης δεδομένων και πως επηρεάζεται από τα δεδομένα μεγάλης κλίμακας (big data). Το άρθρο τους παρέχει μία λεπτομερή περιγραφή της μελέτης, των εργασιών και των εργαλείων (assignments, project, tools) που θα μπορούσε ο καθηγητής-εισηγητής του μαθήματος να χρησιμοποιήσει για να φτιάξει ένα παρόμοιο μάθημα.

Παρέχονται τα παρακάτω:

1) ένα σύνολο βασικών θεμάτων που πρέπει να περιλαμβάνει ένα τέτοιο μάθημα,
2) ένα σύνολο δημοφιλών εργαλείων και γλωσσών που χρησιμοποιούνται για την εξόρυξη δεδομένων,
3) μια λίστα πηγών για σύνολα δεδομένων του «πραγματικού κόσμου» που μπορούν να είναι χρήσιμα για εργασίες και έργα, και
4) μια συζήτηση για τις προκλήσεις που αντιμετωπίζει ο καθηγητής, συζήτηση που βασίζεται στην ερευνητική δραστηριότητα των συγγραφέων, την εμπειρία εφαρμογών στη βιομηχανία και την ανάπτυξη και τη διδασκαλία ενός μαθήματος.

Παράδειγμα περιγραφής στόχων μαθήματος (Goal Description)

Ενδεικτικά αντιγράφουμε την περιγραφή των στόχων ενός μαθήματος που προτείνεται στη βιβλιογραφία:

G1 Gain an understanding of what data mining is all about.
G2 Be able to perform the data preparation tasks and understand the implications.
G3 Demonstrate an understanding of the alternative knowledge representations such as rules, decision trees, decision tables, and Bayesian networks.
G4 Demonstrate an understanding of the basic machine learning algorithmic methods that support knowledge discovery.
G5 Be able to evaluate what has been learned through the application of the appropriate statistics.
G6 Be able to discuss alternative data mining implementations and what might be most appropriate for a given data mining task.
G7 Become proficient in the use of a set of data mining tools.

(πηγή Cecil Schmidt (2011))

Πρότυπα

Σημαντική πηγή για τη σχεδίαση του μαθήματος αποτελούν οι οδηγίες της επιτροπής της ACM SIGKDD Curriculum Committee:

"Data Mining Curriculum: A Proposal (Version 1.0).

Intensive Working Group of ACM SIGKDD Curriculum Committee: Soumen Chakrabarti, Martin Ester, Usama Fayyad, Johannes Gehrke, Jiawei Han, Shinichi Morishita, Gregory Piatetsky-Shapiro, Wei Wang April 30, 2006".

Στη βιβλιογραφία επισημαίνεται ότι η πρόταση της επιτροπής αποτελεί ένα υπερσύνολο που όμως είναι σημαντικό να έχουμε υπόψη όταν σχεδιάζουμε ένα τέτοιο μάθημα. Από τη μελέτη της σχετικής βιβλιογραφίας δεν βρήκαμε κάποιο μάθημα που να συμμορφώνεται πλήρως στις οδηγίες αυτές.

1.1 Εισαγωγή στην Εξόρυξη δεδομένων

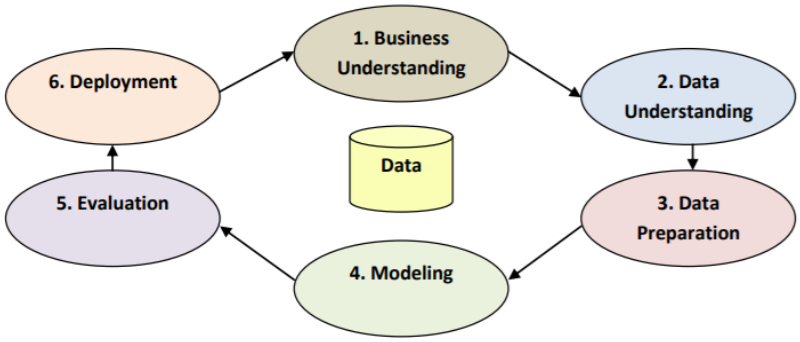
Η εξέλιξη των επιστημών, σε συνδυασμό με την ταχεία διείσδυση των εφαρμογών κοινωνικής δικτύωσης και των έξυπνων συσκευών έχουν οδηγήσει σε μια καταιγίδα δεδομένων με τα οποία ο μέσος χρήστης έρχεται σε επαφή. Η διαχείριση όλων αυτών των δεδομένων με σκοπό την εξαγωγή χρήσιμων συμπερασμάτων αποτελεί αντικείμενο της εξόρυξης δεδομένων.

Η εξόρυξη δεδομένων, έχει εξελιχθεί σε ένα από τους σημαντικότερους κλάδους της επιστήμης της πληροφορικής καθώς διερευνά την εφαρμογή μοντέλων ανάλυσης χρησιμοποιώντας αλγορίθμους μηχανικής μάθησης. Μέσα από τη χρήση των μοντέλων επιχειρείται η μελέτη των δεδομένων και η πρόβλεψη των πιθανών καταστάσεων που μπορεί να περιέλθουν. Η Εξόρυξη δεδομένων μπορεί να χρησιμοποιηθεί

- Για την ανάλυση μεγάλου όγκου δεδομένων, όπου με τους παραδοσιακούς τρόπους θα ήταν δύσκολο έως αδύνατο.
- Από επιστημονικές ομάδες γιατί βοηθούνται μέσω της ανάλυσης των δεδομένων στην μελέτη και έγκυρη διάγνωση φαινομένων.

- Μια εταιρία προκειμένου να μελετήσει τους πελάτες της και να σχεδιάσει πιο αποδοτικές στρατηγικές διαφήμισης με σκοπό να αυξήσει τα έσοδα και να μειώσει τα κόστη.
- Για να αναλυθούν οι τάσεις που παρατηρούνται.
- Για να εντοπιστούν ακραίες συμπεριφορές οι οποίες ενδεχομένως να οδηγούν σε φαινόμενα απάτης.
- Για να εφαρμοστούν προγνωστικές διαδικασίες σχετικά με την κατάσταση που μπορεί να περιέλθει ένα αντικείμενο ή μια οντότητα.
- Για να πραγματοποιηθεί έλεγχος Υποθέσεων
- Για να εντοπιστούν πιθανές ομάδες εντός ενός συνόλου πραγμάτων
- Για να εντοπιστούν συσχετίσεις μεταξύ πραγμάτων

Η πλέον δημοφιλής μεθοδολογία που ακολουθείται για την ανάλυση των προβλημάτων εξόρυξης γνώσης ονομάζεται CRISP-DM (CRoss-Industry Standard Process for Data Mining).



Τα έξι στάδια της μεθοδολογίας CRISP-DM

1. **<u>Business Understanding.</u>**

Το πρώτο στάδιο όταν ξεκινάμε μια διαδικασία ανάλυσης είναι να κατανοήσουμε το πρόβλημα και όλες του τις παραμέτρους προκειμένου να οδηγηθούμε σε ασφαλή συμπεράσματα. Είναι πολύ σημαντικό να γνωρίζουμε τον χώρο για τον οποίο θέλουμε να αναλύσουμε τα δεδομένα, το πρόβλημα για το οποίο θα πρέπει να

βρούμε απαντήσεις καθώς και τα δεδομένα που έχουμε στην διάθεση μας το οποίο είναι και το επόμενο στάδιο της διαδικασίας της ανάλυσης.

2. **Data Understanding**

Όσο σημαντικό είναι να γνωρίζουμε τον χώρο τον οποίο καλούμαστε να αναλύσουμε εξίσου σημαντικό είναι να γνωρίζουμε και τα δεδομένα. Η γνώση των δεδομένων μας βοηθάει στο να αποφύγουμε την εξαγωγή εσφαλμένων συμπερασμάτων και στην σωστή ερμηνεία και χρήση των μεταβλητών.

3. **Data Preperation**

Τα δεδομένα που θα πρέπει να χρησιμοποιήσουμε δεν είναι πολλές σε μορφή επεξεργάσιμη ή σε μορφή που να μπορούμε άμεσα να τα χρησιμοποιήσουμε. Επίσης τα δεδομένα αποτελούν το βασικό συστατικό της διαδικασίας ανάλυσης, οπότε όλα τα δεδομένα που έχουμε στην διάθεση μας ανεξαρτήτως μορφής θα πρέπει να τα χρησιμοποιήσουμε. Για αυτό το λόγο ένα ακόμα σημαντικό στάδιο της ανάλυσης αποτελεί το στάδιο της προετοιμασίας των δεδομένων το οποίο περιλαμβάνει όλες εκείνες τις ενέργειες που θα πρέπει να κάνουμε έτσι ώστε να μπορούμε να χρησιμοποιήσουμε τα δεδομένα μας κατά την διάρκεια της ανάλυσης.

4. **Modeling**

Τα μοντέλα ανάλυσης αποτελούν την εφαρμογή των αλγορίθμων στα δεδομένα με σκοπό την εξαγωγή κάποιων συμπερασμάτων. Τα μοντέλα μπορούν να μας προβλέψουν νέες τιμές ή να κατηγοριοποιήσουν τις ήδη υπάρχουσες.

5. **Evaluation**

Το επόμενο στάδιο μετά την δημιουργία του μοντέλου αποτελεί η αξιολόγηση του. Καταυτό τον τρόπο πρίν από την οριστική εφαρμογή του μοντέλου, θα πρέπει να αξιολογηθούν οι απαντήσεις που δίνει το μοντέλο έτσι ώστε να γνωρίζουμε τον βαθμό που αυτές είναι αποδεκτές και αν μπορούμε να τις χρησιμοποιήσουμε με ασφάλεια.

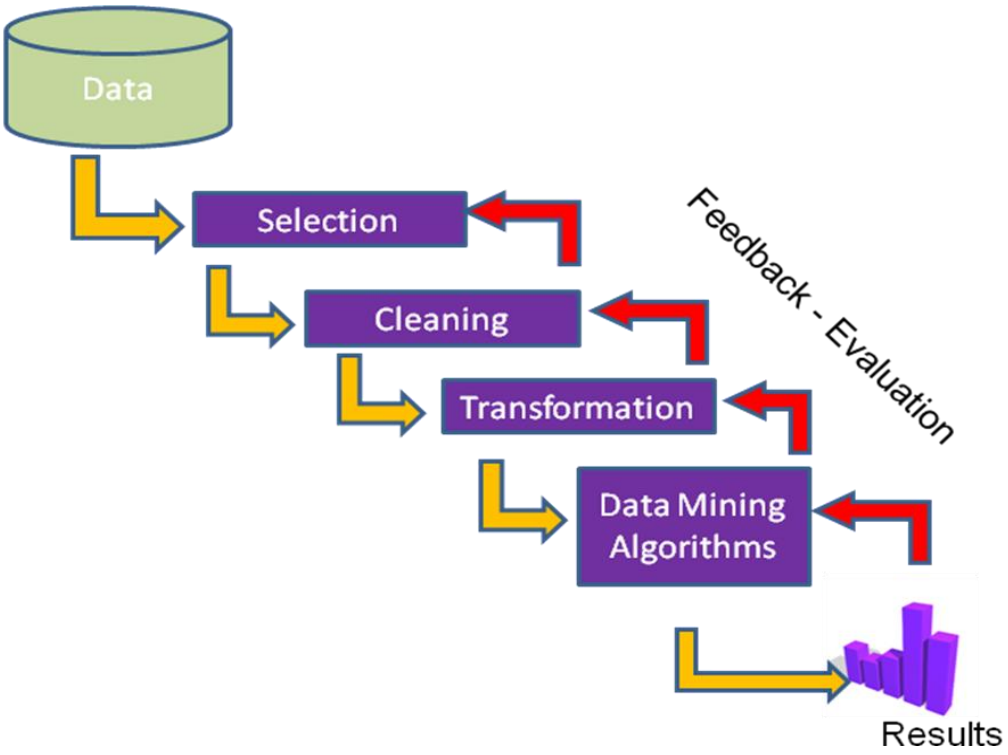
6. **Deployment**

Το τελευταίο στάδιο, με την προϋπόθεση ότι έχουν ολοκληρωθεί επιτυχώς όλα τα προηγούμενα βήματα είναι η εφαρμογή του μοντέλου με σκοπό την χρήση από τους αναλυτές.

Η διαδικασία της Εξόρυξης Γνώσης από την στιγμή που ερχόμαστε σε επαφή με τα αρχικά δεδομένα μέχρι και το στάδιο της εξαγωγής των συμπερασμάτων είναι μια αμφίδρομη διαδικασία με πολλά επαναλαμβανόμενα στάδια. Πιο αναλυτικά τα στάδια αυτά περιλαμβάνουν

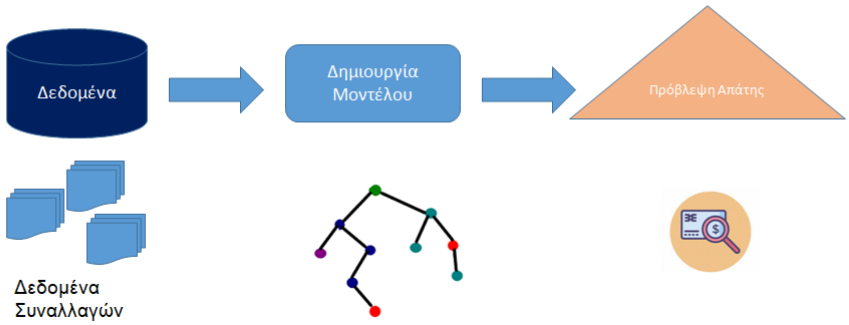
- Επιλογή των δεδομένων.
- Προετοιμασία και μετασχηματισμός των δεδομένων (data cleaning, reduction & transformation).
- Επιλογή του κατάλληλου αλγορίθμου.
- Αξιολόγηση των αποτελεσμάτων.
- Εξαγωγή συμπερασμάτων.
- Προβολή των αποτελεσμάτων.

Το γεγονός ότι πολλές φορές θα πρέπει να επανέρθουμε σε προηγούμενα βήματα της διαδικασίας προκειμένου να αλλάξουμε κάποιες παραμέτρους ή κάποιες παραδοχές κατά της διαδικασία της ανάλυσης οφείλεται στο γεγονός ότι οι χρήστες συχνά δεν έχουν εκ των προτέρων καθαρή εικόνα για το ποια πληροφορία είναι ενδιαφέρουσα. Επίσης μέσω της παραγωγής των πρώτων συμπερασμάτων προκύπτουν νέα ερωτήματα τα οποία πολλές φορές μας οδηγούν στον ανασχεδιασμό της διαδικασίας ή την δημιουργία νέων μοντέλων. Τέλος υπάρχουν περιπτώσεις όπου τα αποτελέσματα της ανάλυσης των δεδομένων δεν οδηγούν σε χρήσιμα συμπεράσματα, με σκοπό τον εκ νέου επανασχεδιασμό της διαδικασίας.



Οι βασικές κατηγορίες αλγορίθμων που έχουμε στην διάθεση μας κατά την διάρκεια της εξόρυξης γνώσης είναι

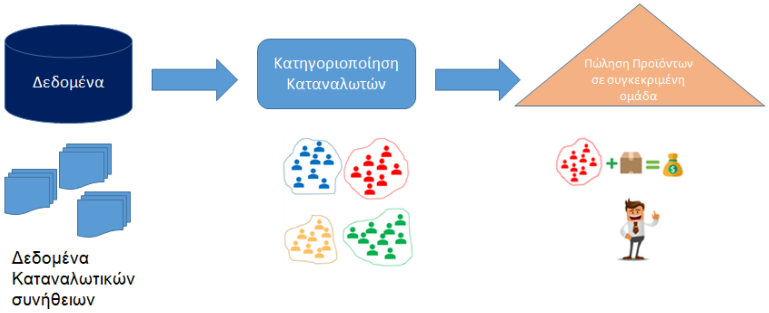
- Ταξινόμηση (Classification), είναι η μέθοδος μέσω της οποίας τα αντικείμενα ταξινομούνται σε ένα προκαθορισμένο σύνολο κατηγοριών με βάση τα χαρακτηριστικά τους.



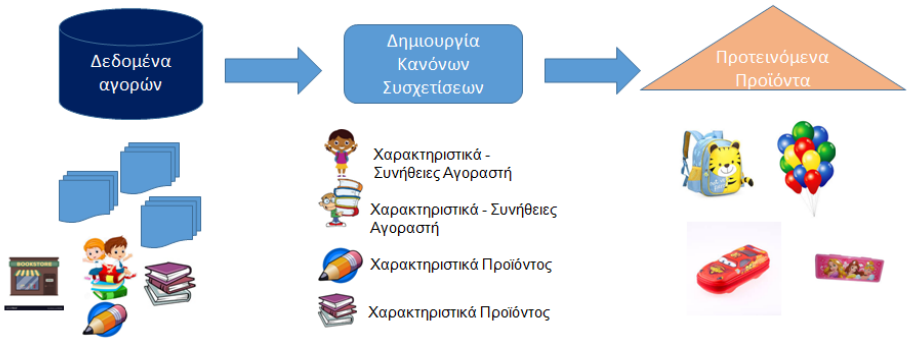
- Παλινδρόμηση (Regression), σε αυτή την μέθοδο τα δεδομένα μοντελοποιούνται χρησιμοποιώντας γραμμικές συναρτήσεις, και οι άγνωστες παράμετροι υπολογίζονται με βάση το μοντέλο που έχει δημιουργηθεί

.

- Συσταδοποίηση (Clustering), είναι η διαδικασία εκείνη κατά την οποία κατηγοριοποιούμε ένα σύνολο από «αντικείμενα», σε συγκεκριμένες ομάδες. Όμοια αντικείμενα καταχωρούνται στην ίδια ομάδα. Την ομοιότητα μεταξύ των αντικειμένων μπορούμε να την μετρήσουμε με διαφόρων ειδών αλγορίθμων. Οι πιο γνωστοί από αυτούς είναι η ευκλείδειος απόσταση και η απόσταση Manhattan.



- Κανόνες Συσχετίσεων (Association Rules) , σε αυτή την μέθοδο μελετώνται και αναλύονται τα δεδομένα με σκοπό την εξαγωγή συσχετίσεων μεταξύ των αντικειμένων.



Ταξινόμηση αλγορίθμων σε αλγόριθμους Επιβλεπόμενης Μάθησης και μη Επιβλεπόμενης Μάθησης.

- Επιβλεπόμενης Μάθησης (Supervised Learning), όταν μας δίνεται ένα σύνολο δεδομένων με τις αντίστοιχες ομάδες (κλάσεις) κάθε εγγραφής. Στόχος είναι η δημιουργία ενός μοντέλου, το οποίο όταν θα δέχεται νέα

δεδομένα να μπορεί να τα κατηγοριοποιεί σε κάποια από τις προϋπάρχουσες κλάσεις.

- Μη Επιβλεπόμενης Μάθησης (Unsupervised Learning),όταν μας δίνεται ένα σύνολο δεδομένων, χωρίς όμως τις αντίστοιχες κλάσεις κάθε εγγραφής. Οπότε έχουμε δεδομένα χωρίς να γνωρίσουμε σε ποια κλάση ανήκουν. Στόχος είναι η ανάλυση αυτών των δεδομένων προκειμένου να ανακαλύψουμε κάποια ενδεχομένως ενδιαφέροντα στοιχεία στα δεδομένα.

Στα επόμενα κεφάλαια θα ακολουθήσει η περιγραφή του διαδικτυακού μαθήματος καθώς και το προτεινόμενο σχέδιο του μαθήματος το οποίο θα περιλαμβάνει

- Τα γενικά Χαρακτηριστικά του Προγράμματος
  - Αντικείμενο και στόχοι του Προγράμματος Δια Βίου Εκπαίδευσης
  - Ομάδα-Στόχος του Προγράμματος
  - Μαθησιακά αποτελέσματα του Προγράμματος
  - Εκπαιδευτικές και διδακτικές μέθοδοι του Προγράμματος
- Το αναλυτικό περιεχόμενο του Μαθήματος
  - Διδακτικές ενότητες (ΔΕ) του Προγράμματος
  - Αναλυτικό Περιεχόμενο ΔΕ του Προγράμματος

2 Γενικά Χαρακτηριστικά του Μαθήματος

2.1 Αντικείμενο και στόχοι του Μαθήματος

Το Μάθημα στην εξόρυξη δεδομένων με χρήση τεχνικών μηχανικής μάθησης» έχει σκοπό την ανάπτυξη των απαραίτητων γνώσεων και δεξιοτήτων για την εξόρυξη γνώσης από βάσεις δεδομένων, με χρήση σύγχρονων τεχνικών και μεθοδολογιών με στόχο την λήψη αποφάσεων για μια εταιρία ή έναν οργανισμό.

Το μάθημα σκοπεύει να παρέχει σε αυτούς που θα το παρακολουθήσουν τα χρήσιμα εργαλεία για το σχεδιασμό, ανάπτυξη και εφαρμογή ενός συστήματος εξόρυξης γνώσης και λήψης αποφάσεων που προκύπτουν από την λειτουργία πληροφοριακών συστημάτων προς την κατεύθυνση της πρόβλεψης πιθανών καταστάσεων, την κατηγοριοποίηση των δεδομένων και την εξαγωγή χρήσιμων συμπερασμάτων.

2.2 Ομάδα-Στόχος του Προγράμματος

Το μάθημα θα απευθύνεται σε:

- Αποφοίτους όλων των επιστημονικών πεδίων οι οποίοι επιθυμούν να απασχοληθούν με τον χώρο της ανάλυσης δεδομένων
- Στελέχη επιχειρήσεων και οργανισμών που επιθυμούν να χρησιμοποιήσουν τις τεχνικές της εξόρυξης δεδομένων

- Φοιτητές ή απόφοιτοι ΑΕΙ που επιθυμούν να αποκτήσουν γνώσεις σχετικά με την ανάλυση δεδομένων.
- Αποφοίτους Λυκείου που επιθυμούν να αποκτήσουν γνώσεις στον τομέα της ανάλυσης δεδομένων

2.3 Μαθησιακά αποτελέσματα του Προγράμματος

Τα μαθησιακά αποτελέσματα που επιδιώκονται από το προτεινόμενο πρόγραμμα είναι τα κάτωθι:

- **Επίπεδο Γνώσεων**

Οι εκπαιδευόμενοι θα είναι σε θέση:

  - Να μπορούν να αναλύσουν ένα πρόβλημα εφαρμόζοντάς τις βασικές αρχές διαχείρισης γνώσης με σκοπό την επίλυση του.
  - Να προσδιορίσουν τον τρόπο επεξεργασίας και διαχείριση των δεδομένων.
  - Να εφαρμόζουν τους κατάλληλος αλγόριθμούς εξόρυξης γνώσης σε συνάρτηση με την φύση του προβλήματος
  - Να αξιολογούν τις λύσεις που βρίσκουν .

- **Επίπεδο Δεξιοτήτων και Ικανοτήτων**

Οι εκπαιδευόμενοι θα είναι σε θέση:

  - Να εφαρμόζουν αποτελεσματικά τις βασικές αρχές διαχείρισης γνώσης και εξόρυξης δεδομένων
  - Να συλλέγουν και να επεξεργάζεται δεδομένα
  - Να μοντελοποιούν διαδικασίες ανάλυσης
  - Να εφαρμόζουν τεχνικές ανάλυσης βασισμένοι σε σύγχρονες μεθόδους
  - Να ερμηνεύουν και να αξιολογούν τα αποτελέσματα.

- **Επίπεδο Στάσεων και Συμπεριφορών**

Οι εκπαιδευόμενοι θα μπορούν:

- Να αναγνωρίζουν όλες τις ιδιαιτερότητες των προβλημάτων που καλούνται να επιλύσουν προκειμένου να οδηγούνται σε ασφαλή συμπεράσματα.

2.4 Εκπαιδευτικές και διδακτικές μέθοδοι του Προγράμματος

Στο προτεινόμενο πρόγραμμα θα εφαρμοστεί η Εξ αποστάσεως εκπαίδευση με την μέθοδο ασύγχρονης τηλεκπαίδευσης. Το εκπαιδευτικό υλικό του προγράμματος θα διατίθεται σταδιακά, ανά διδακτική ενότητα. Κατά την εξέλιξη κάθε θεματικής ενότητας θα αναρτώνται όλο το απαραίτητο υλικό για την διεξαγωγή της ενότητας. Επίσης στο τέλος κάθε διδακτικής ενότητας θα υπάρχουν τα αντίστοιχα τεστ με ερωτήσεις τύπου «Σωστό-Λάθος» , πολλαπλής επιλογής, ανάπτυξης ή επίλυσης προβλημάτων (μέσω upload στο σύστημα), όπου ο εκπαιδευόμενος θα πρέπει να διατυπώσει και να επισυνάψει την απάντησή του.

3 *Αναλυτικό περιεχόμενο του Μαθήματος*

3.1 Διδακτικές ενότητες (ΔΕ) του Μαθήματος

Το παρόν μάθημα περιλαμβάνει έξι (6) Διδακτικές Ενότητες και έχει σχεδιαστεί με γνώμονα να παρέχει τις απαραίτητες γνώσεις στους νέους επιστήμονες του χώρου της ανάλυσης δεδομένων:

3.1.1 ΔΕ1: Μεθοδολογία Διαχείρισης γνώσης από βάσεις δεδομένων

ΤΙΤΛΟΣ ΔΕ	Μεθοδολογία Διαχείρισης γνώσης από βάσεις δεδομένων
ΚΩΔΙΚΟΣ ΔΕ	1
ΜΑΘΗΣΙΑΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ	Σκοπός της διδακτικής ενότητας είναι η εισαγωγή στη διαχείριση γνώσης από βάσεις δεδομένων παρουσιάζοντας την μεθοδολογία του “Knowledge Discovery from Databases”. Πλέον των παραπάνω θα μελετηθεί ο σχεδιασμός, η διαχείριση μιας βάσης δεδομένων. Επίσης θα παρουσιαστεί το εργαλείο «RapidMiner» το οποίο θα χρησιμοποιηθεί κατά την διάρκεια του μαθήματος. Μετά το πέρας της Διδακτικής Ενότητας οι εκπαιδευόμενοι θα διαθέτουν: • Βασικές Αρχές Εξόρυξης Γνώσης • Ανάλυση μεθοδολογίας Knowledge Discovery from

	Databases (KDD) • Ανάπτυξη γνώσεων και δεξιοτήτων σχετικά με την σχεδιασμό και διαχείριση Βάσεων δεδομένων

3.1.2 ΔΕ2: Επεξεργασία και διαχείριση δεδομένων και μεταβλητών

ΤΙΤΛΟΣ ΔΕ	Επεξεργασία και διαχείριση δεδομένων και μεταβλητών
ΚΩΔΙΚΟΣ ΔΕ	2
ΜΑΘΗΣΙΑΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ	Σκοπός της διδακτικής ενότητας είναι η μελέτη του τρόπου επεξεργασίας και διαχείρισης δεδομένων. Πιο συγκεκριμένα θα μελετηθούν περιπτώσεις με ελλείπει στοιχεία ή ασυνεπή δεδομένα. Επίσης θα παρουσιαστούν τεχνικές και μεθοδολογίες διαχείρισης των μεταβλητών (μετατροπή σε άλλη μορφή, μείωση αριθμού...). Μετά το πέρας της Διδακτικής Ενότητας οι εκπαιδευόμενοι θα μπορούν να: • Κατανοούν τις βασικές αρχές επεξεργασίας δεδομένων • Εφαρμόζουν τεχνικές διαχείρισης δεδομένων σε περιπτώσεις ελλιπών (missing data) ή ασυνεπών δεδομένων (inconsistent data) • Εφαρμόζουν τεχνικές διαχείρισης των κριτηρίων / μεταβλητών σε σχέση με τον τύπο τους ή το πλήθος. • Εφαρμόζουν τεχνικές συσχετίσεων των μεταβλητών

3.1.3 ΔΕ3: Μελέτη, σχεδιασμός και εφαρμογή Κανόνων Συσχετίσεων (Association Rules)

ΤΙΤΛΟΣ ΔΕ	Μελέτη, σχεδιασμός και εφαρμογή τεχνικών Κατηγοριοποίησης (Association Rules)
ΚΩΔΙΚΟΣ ΔΕ	3
ΜΑΘΗΣΙΑΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ	Σκοπός της διδακτικής ενότητας είναι η μελέτη και εφαρμογή των βασικών μοντέλων για εξαγωγή κανόνων συσχετίσεων. Οι αλγόριθμοι θα παρουσιαστούν μέσω παραδειγμάτων και τα αποτελέσματα τους θα αξιολογηθούν με σκοπό την επιλογή του βέλτιστου αλγορίθμου. Μετά το πέρας της Διδακτικής Ενότητας οι εκπαιδευόμενοι θα μπορούν να: • Προετοιμάζουν και μετατρέπουν τα δεδομένα στην απαραίτητη μορφή προκειμένου να εκτελεστεί ο αλγόριθμος. • Εφαρμόζουν τους αλγορίθμους των κανόνων συσχετίσεων. • Χτίζουν μοντέλα ανάλυσης. • Εφαρμόζουν μοντέλα ανάλυσης. • Ερμηνεύουν και να αξιολογούν τα αποτελέσματα.

3.1.4 ΔΕ4: Μελέτη, σχεδιασμός και εφαρμογή τεχνικών Συσταδοποίησης (clustering)

ΤΙΤΛΟΣ ΔΕ	ΔΕ4: Μελέτη, σχεδιασμός και εφαρμογή τεχνικών Συσταδοποίηση (clustering)
ΚΩΔΙΚΟΣ ΔΕ	4
ΜΑΘΗΣΙΑΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ	Σκοπός της διδακτικής ενότητας είναι η μελέτη και εφαρμογή των βασικών αλγορίθμων συσταδοποίησης. Οι αλγόριθμοι θα παρουσιαστούν μέσω παραδειγμάτων και τα αποτελέσματα τους θα αξιολογηθούν με σκοπό την επιλογή του βέλτιστου αλγορίθμου. Μετά το πέρας της Διδακτικής Ενότητας οι εκπαιδευόμενοι θα μπορούν να: • Προετοιμάζουν και μετατρέπουν τα δεδομένα στην απαραίτητη μορφή προκειμένου να εκτελεστεί ο αλγόριθμος. • Εφαρμόζουν τους αλγορίθμους συσταδοποίησης. • Χτίζουν μοντέλα ανάλυσης. • Εφαρμόζουν μοντέλα ανάλυσης. Ερμηνεύουν και να αξιολογούν τα αποτελέσματα.

3.1.5 ΔΕ5: Μελέτη, σχεδιασμός και εφαρμογή τεχνικών Πρόβλεψης (Prediction).

ΤΙΤΛΟΣ ΔΕ	Μελέτη, σχεδιασμός και εφαρμογή τεχνικών Πρόβλεψης (Prediction).
ΚΩΔΙΚΟΣ ΔΕ	5
ΜΑΘΗΣΙΑΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ	Σκοπός της διδακτικής ενότητας είναι η μελέτη και εφαρμογή των βασικών μοντέλων πρόβλεψης. Οι αλγόριθμοι θα παρουσιαστούν μέσω παραδειγμάτων και τα αποτελέσματα τους θα αξιολογηθούν με σκοπό την επιλογή του βέλτιστου μοντέλου. Μετά το πέρας της Διδακτικής Ενότητας οι εκπαιδευόμενοι θα μπορούν να: • Προετοιμάζουν και μετατρέπουν τα δεδομένα στην απαραίτητη μορφή προκειμένου να εκτελεστεί ο αλγόριθμος. • Εφαρμόζουν τους αλγορίθμους πρόβλεψης. • Χτίζουν μοντέλα ανάλυσης. • Εφαρμόζουν μοντέλα ανάλυσης. • Ερμηνεύουν και να αξιολογούν τα αποτελέσματα.

3.1.6 ΔΕ6: Μελέτη, σχεδιασμός και εφαρμογή τεχνικών Παλινδρόμησης (Regression).

ΤΙΤΛΟΣ ΔΕ	Μελέτη, σχεδιασμός και εφαρμογή τεχνικών Παλινδρόμηση (Regression).
ΚΩΔΙΚΟΣ ΔΕ	6
ΜΑΘΗΣΙΑΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ	Σκοπός της διδακτικής ενότητας είναι η μελέτη και εφαρμογή των βασικών μοντέλων παλινδρόμησης (γραμμική - λογιστική). Οι αλγόριθμοι θα παρουσιαστούν μέσω παραδειγμάτων και τα αποτελέσματα τους θα αξιολογηθούν με σκοπό την επιλογή του βέλτιστου αλγορίθμου. Μετά το πέρας της Διδακτικής Ενότητας οι εκπαιδευόμενοι θα μπορούν να: • Προετοιμάζουν και μετατρέπουν τα δεδομένα στην απαραίτητη

	μορφή προκειμένου να εκτελεστεί ο αλγόριθμος. • Εφαρμόζουν τους αλγορίθμους παλινδρόμησης. • Χτίζουν μοντέλα ανάλυσης. • Εφαρμόζουν μοντέλα ανάλυσης. • Ερμηνεύουν και να αξιολογούν τα αποτελέσματα.

3.2 Περιεχόμενο Διδακτικών ενοτήτων

3.2.1 ΔΕ1: Μεθοδολογία Διαχείρισης γνώσης από βάσεις δεδομένων

Η πρώτη διδακτική ενότητα αποτελείται από 2 υποενότητες

- Βασικές Αρχές Εξόρυξης Γνώσης , όπου αναλύονται οι βασικές αρχές εξόρυξης γνώσης
- Σχεδιασμός και διαχείριση Αποθηκών και Βάσεων δεδομένων, όπου παρουσιάζονται ο σχεδιασμός και η διαχείριση βάσεων δεδομένων και αποθηκών δεδομένων

3.2.1.1 Βασικές Αρχές Εξόρυξης Γνώσης.

Σε αυτή την υποενότητα δίνεται ο ορισμός του πεδίου της εξόρυξης γνώσης και αναλύονται οι βασικές αρχές της καθώς και τα πεδία εφαρμογής της. Στην συνέχεια επιχειρείται μια σύντομη περιγραφή των επιμέρους σταδίων της εξόρυξης γνώσης όπως είναι αυτό του καθορισμού των δεδομένων, της προετοιμασίας των δεδομένων καθώς επίσης παρουσιάζονται και τα διάφορα είδη αλγορίθμων. Τέλος μέσω παραδειγμάτων πραγματοποιείται η παρουσίαση των αλγορίθμων της κατηγοριοποίησης, των κανόνων συσχετίσεων, της πρόβλεψης μέσω δέντρων αποφάσεων και της γραμμικής παλινδρόμησης. Στην συνέχεια ακολουθούν ενδεικτικές διαφάνειες από το εκπαιδευτικό υλικό.

Βασικές Αρχές Εξόρυξης Γνώσης

Τι είναι η εξόρυξη δεδομένων (Data Mining)

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Wikipedia

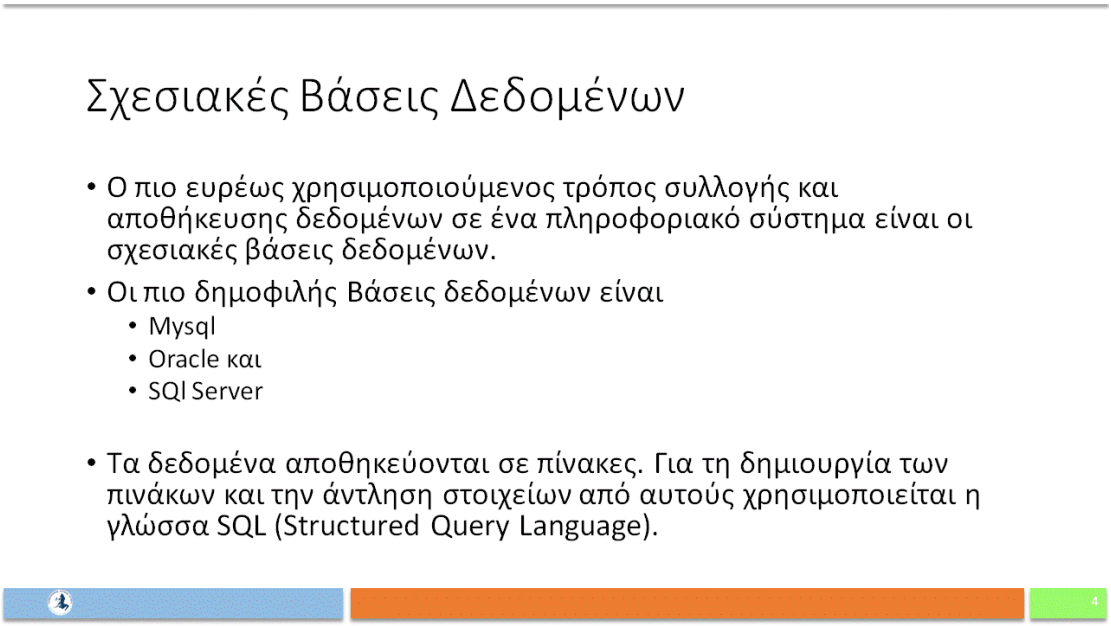
3.2.1.2 Σχεδιασμός και διαχείριση Αποθηκών και Βάσεων δεδομένων

Σε αυτή την υποενότητα παρουσιάζονται τα πιο γνωστά είδη συλλογής και αποθήκευσης δεδομένων σε ένα πληροφοριακό σύστημα. Τα είδη ποικίλουν από σχεσιακές βάσεις δεδομένων μέχρι αποθήκες δεδομένων «Data Warehouse». Επίσης παρουσιάζεται η μεθοδολογία ETL (Extract Transform Load) και πως αυτή εφαρμόζεται σε ένα σύστημα όπου χρησιμοποιούνται αποθήκες δεδομένων. Στην συνέχεια περιγράφονται οι κύβοι καθώς επίσης και τα datamart μαζί με τα διάφορα είδη αρχιτεκτονικής τους.

Σχεδιασμός και διαχείριση Αποθηκών και Βάσεων δεδομένων

Οι πιο συνηθισμένες πηγές δεδομένων είναι οι παρακάτω

- Σχεσιακές ΒΔ
- Αποθήκες δεδομένων
- Από Αρχεία



3.2.2 ΔΕ2: Επεξεργασία και διαχείριση δεδομένων και μεταβλητών

Η διδακτική ενότητα «Επεξεργασία και διαχείριση δεδομένων και μεταβλητών» αποτελείται από 4 υποενότητες

- Επεξεργασία και διαχείριση δεδομένων και μεταβλητών, όπου παρουσιάζονται οι βασικές τεχνικές επεξεργασίας και διαχείρισης των δεδομένων και μεταβλητών
- Προετοιμασία των δεδομένων, όπου παρουσιάζεται αναλυτικά η τεχνική προετοιμασίας των δεδομένων
- Καθαρισμός Δεδομένων, όπου παρουσιάζεται αναλυτικά η τεχνική του καθορισμού των δεδομένων
- Μετασχηματισμός Δεδομένων , όπου παρουσιάζεται αναλυτικά η τεχνική του μετασχηματισμού των δεδομένων

3.2.2.1 Επεξεργασία και διαχείριση δεδομένων και μεταβλητών

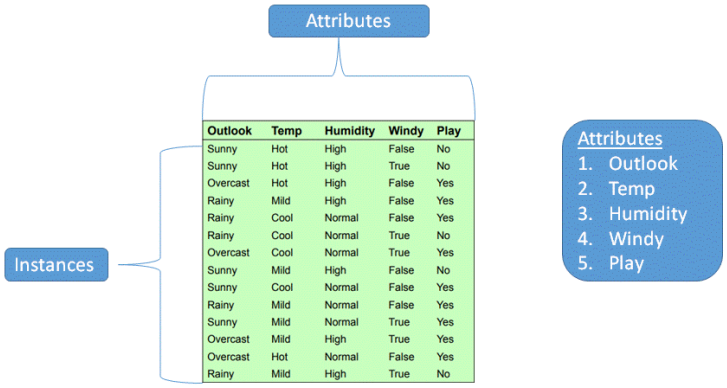
Σε αυτή την ενότητα παρουσιάζεται οι βασικές τεχνικές επεξεργασίας και διαχείρισης των δεδομένων και μεταβλητών. Επίσης επεξηγούνται βασικές έννοιες όπως είναι τα attribute, τα instances καθώς και τα διάφορα είδη αρχείων που μπορούμε να χρησιμοποιήσουμε. Στην συνέχεια παρουσιάζονται τα παρακάτω είδη μεταβλητών

- Numeric,
- Nominal,

- Ordinal,
- Interval ,
- Ratio και
- Binary.

Επεξεργασία και διαχείριση δεδομένων και μεταβλητών

Περιγραφή των δεδομένων



3.2.2.2 Προετοιμασία των δεδομένων

Σε αυτή την ενότητα περιγράφεται αναλυτικά η τεχνική προετοιμασίας των δεδομένων. Πιο συγκεκριμένα παρουσιάζονται

- Ο καθαρισμός των Δεδομένων (**cleaning**)

- Ελλιπείς τιμές (missing values)
- Μη σωστά δεδομένα – ακραίες τιμές (outliers)
- Δεδομένα με προβλήματα συσχέτισης(Inconsistent data)
- Διπλοεγγραφές

- Η σύνδεση δεδομένων από διαφορετικές πηγές (βάσεις δεδομένων, αρχεία,…) (**integration**)
- Ο Μετασχηματισμός των δεδομένων (**transformation**)

Προετοιμασία των δεδομένων

Ελλιπείς τιμές (missing values)

	Gender	Race	Birth_Ye...	Marital_...	Years_o...	Hours_P...	Preferre...	Preferre...	Preferre...	Read_N...	Online_...	Online_...	Face
2	M	White	1972	M	8	1	Firefox	Google	Yahoo	Y	N	N	Y
3	M	Hispanic	1981	S	14	2	Chrome	Google	Hotmail	Y	N	N	Y
4	F	African A...	1977	S	6	2	Firefox	Yahoo	Yahoo	Y	Y		Y
5	F	White	1961	D	8	6	Firefox	Google	Hotmail	N	Y	N	N
6	M	White	1954	M	2	3	Internet ...	Bing	Hotmail	Y	Y	N	Y
7	M	African A...	1982	D	15	4	Internet ...	Google	Yahoo	Y	N	Y	N
8	M	African A...	1981	D	11	2	Firefox	Google	Yahoo		Y		Y
9	M	White	1977	S	3	3	Internet ...	Yahoo	Yahoo	Y			Y
10	F	African A...	1969	M	6	2	Firefox	Google	Gmail	N	Y	N	N
11	M	White	1987	S	12	1	Safari	Yahoo	Yahoo	Y		Y	Y
12	F	Hispanic	1959	D	12	5	Chrome	Google	Gmail	Y	N	N	Y

Ελλιπείς τιμές

3.2.2.3 *Καθαρισμός Δεδομένων*

Σε αυτή την ενότητα περιγράφεται αναλυτικά η τεχνική του καθαρισμού των δεδομένων. Πιο συγκεκριμένα παρουσιάζονται τεχνικές επίλυσης των παρακάτω προβλημάτων:

- Ελλιπείς τιμές (missing values)
- Μη σωστά δεδομένα (outliers)
- Διπλοεγγραφές

Καθαρισμός Δεδομένων

Ελλιπείς τιμές (missing values)

Προκειμένου να διορθώσουμε τις εγγραφές που παρουσιάζουν ελλιπείς τιμές μπορούμε να εφαρμόσουμε μια από τις παρακάτω μεθόδους

Αντικατάσταση τιμής με βάση

- τον ΜΟ των υπολοίπων τιμών
- Τη μικρότερη τιμή
- Τη μεγαλύτερη τιμή
- Να βάλουμε μια δική μας τιμή
- Να βάλουμε το μηδέν.

missing_values

3.2.2.4 *Μετασχηματισμός Δεδομένων*

Σε αυτή την ενότητα περιγράφεται αναλυτικά η τεχνική του μετασχηματισμού των δεδομένων. Πιο συγκεκριμένα παρουσιάζονται τεχνικές επίλυσης των παρακάτω προβλημάτων:

- Μείωση των δεδομένων (data reduction)
- Μετατροπή των δεδομένων (data transformation)

Μετασχηματισμός των Δεδομένων

Data reduction

Μείωση στηλών (Correlation)

- **Συσχέτιση** δύο τυχαίων μεταβλητών ορίζουμε τη συναρτησιακή σχέση εξάρτησης της μίας μεταβλητής ως προς την άλλη. Αν οι μεταβλητές είναι δύο τότε έχουμε απλή συσχέτιση, ενώ αν είναι περισσότερες έχουμε την πολλαπλή συσχέτιση.
- Ορίζουμε τον **Συντελεστή Συσχέτισης** με σκοπό να μετρήσουμε το βαθμό συσχέτισης των δυο μεταβλητών. Όταν ο συντελεστής συσχέτισης ισούται με 1 υποδηλώνεται η απόλυτη συσχέτιση των δυο μεταβλητών ενώ με 0 υποδηλώνεται ότι δεν υπάρχει συσχέτιση μεταξύ των μεταβλητών.

Correlation

3.2.3 ΔΕ3: Μελέτη, σχεδιασμός και εφαρμογή Κανόνων Συσχετίσεων

Η διδακτική ενότητα «Μελέτη, σχεδιασμός και εφαρμογή Κανόνων Συσχετίσεων» αποτελείται από 2 υποενότητες

- Εισαγωγή στους Αλγορίθμους εξαγωγής Κανόνων Συσχετίσεων.
- Παράδειγμα Κανόνων Συσχετίσεων.

3.2.3.1 Εισαγωγή στους Αλγορίθμους εξαγωγής Κανόνων Συσχετίσεων

Σε αυτή την ενότητα περιγράφεται αναλυτικά η τεχνική των κανόνων συσχετίσεων, η οποία βασίζεται στην ανακάλυψη και διατύπωση σχέσεων που υπάρχουν ανάμεσα στα δεδομένα (X → Y). Οι σχέσεις αυτές προκύπτουν από τη συχνή εμφάνιση ζεύγους τιμών. Δύο ποσοτικά μεγέθη καθορίζουν πόσο ισχυρός είναι ο κανόνας

1. Support (s). Το ποσοστό των συναλλαγών (επί του συνόλου των συναλλαγών) που περιέχουν και το X και το Y.
2. Confidence (c). Είναι η πιθανότητα εμφάνισης του Y, όταν εμφανίζεται το X.

Κανόνες Συσχετίσεων

Κανόνες Συσχετίσεων - Ορισμοί

• Ένας κανόνας ορίζεται ως έξης

Χ→Υ, όπου τα Χ και Υ είναι αντικείμενα του συνόλου

π.χ. {Μπλοκ ζωγραφικής}→{Μαρκαδόροι}. Τα άτομα που αγόρασαν μπλοκ ζωγραφικής αγόρασαν επίσης και μαρκαδόρους.

Κανόνες Συσχετίσεων - Ορισμοί

Παράδειγμα

Συναλλαγή	Προϊόντα
1	Α-Β-Γ
2	Γ-Δ
3	Α-Β
4	Α-Β-Δ
5	Α-Δ
6	Β-Γ

Κανόνας :Α->Β

- Το ζεύγος Α-Β εμφανίζεται 3 φορές.
- Οι συνολικές συναλλαγές είναι 6.
- Το Α εμφανίζεται 4 φορές ενώ το Β 3.
- Η υποστήριξη του κανόνα είναι 3/6 , 50%
- Η εμπιστοσύνη του κανόνα είναι 3/4,

3.2.3.2 Παράδειγμα Κανόνων Συσχετίσεων.

Σε αυτή την ενότητα παρουσιάζονται παραδείγματα εφαρμογής των κανόνων συσχετίσεων.

Παράδειγμα – Κανόνων Συσχετίσεων

Πρόβλημα

Η Μαρία είναι διευθύντρια σε ένα σχολικό συγκρότημα το οποίο έχει πολλές και μεγάλες ανάγκες που δυστυχώς όμως δεν υπάρχουν οι απαραίτητοι πόροι για να ικανοποιηθούν.

Γνωρίζει όμως ότι πολλοί γονείς δραστηριοποιούνται σε διάφορους συλλόγους και οργανώσεις και θεωρεί πως αν καταφέρει να χρησιμοποιήσει τις γνωριμίες που έχουν οι γονείς ή πολύ περισσότερο αν καταφέρει να κάνει γονείς που ανήκουν σε διαφορετικούς συλλόγους να συνεργαστούν μεταξύ τους, τα οφέλη για το σχολικό συγκρότημα θα είναι πολύ μεγάλα.

3.2.4 ΔΕ4: Μελέτη, σχεδιασμός και εφαρμογή τεχνικών Συσταδοποίηση (clustering)

Η διδακτική ενότητα «Μελέτη, σχεδιασμός και εφαρμογή τεχνικών Συσταδοποίηση (clustering)» αποτελείται από 2 υποενότητες

- Εισαγωγή στους Αλγορίθμους Συσταδοποίηση.
- Αναλυτικά οι αλγόριθμοι Συσταδοποίησης.

3.2.4.1 Εισαγωγή στους Αλγορίθμους Clustering.

Σε αυτή την ενότητα περιγράφεται αναλυτικά η τεχνική της συσταδοποίηση των δεδομένων. Πιο συγκεκριμένα παρουσιάζονται:

- Η έννοια της συσταδοποίησης
- Οι μέθοδοι υπολογισμού της ομοιότητας μεταξύ των αντικειμένων
- Τα διάφορα είδη αλγορίθμων συσταδοποίησης
- Οι μέθοδοι εκτέλεσης της διαδικασίας συσταδοποίησης

Εξόρυξη δεδομένων με χρήση τεχνικών μηχανικής μάθησης

Ομαδοποίηση (clustering)

Τι είναι η Ομαδοποίηση (clustering)

- Σκοπός όλων των clustering αλγορίθμων αποτελεί ο διαχωρισμός ενός πλήθους σημείων ή αντικειμένων σε ομοειδείς ομάδες (clusters).
- Προς την κατεύθυνση αυτή χρησιμοποιούνται συναρτήσεις οι οποίες υπολογίζουν την απόσταση μεταξύ των σημείων.
- Με βάση την απόσταση μεταξύ των σημείων δημιουργούνται ομάδες από αντικείμενα τα οποία εμφανίζουν τη μικρότερη δυνατή απόσταση. Δηλαδή 2 αντικείμενα που ανήκουν σε διαφορετικά clusters θα πρέπει να εμφανίζουν μεγαλύτερη απόσταση σε σχέση με αυτά που βρίσκονται στο ίδιο το δικό τους cluster.

3.2.4.2 Αναλυτικά οι αλγόριθμοι Clustering.

Σε αυτή την ενότητα περιγράφεται αναλυτικά η τεχνική της ομαδοποιήσης με την χρήση αλγορίθμων ιεραρχικής και διαχωριστικής ανάλυσης συστάδων. Στους ιεραρχικούς αλγορίθμους δημιουργούνται οι πίνακες των αποστάσεων για όλα τα αντικείμενα και ενώνονται αυτά με την μικρότερη απόσταση ενώ στους διαχωριστικούς αλγορίθμους ένα σύνολο αντικειμένων διαχωρίζεται σε ένα προκαθορισμένο αριθμό clusters κατά τέτοιο τρόπο έτσι ώστε να δημιουργούνται ομάδες ομοειδών αντικειμένων..

Περιγραφή Αλγορίθμου Clustering

Αλγόριθμοι Ανάλυσης σε Συστάδες

Οι κατηγορίες αλγορίθμων που θα εξετάσουμε είναι

- Ιεραρχικής Ανάλυσης Συστάδων
- Διαχωριστικής Ανάλυσης Συστάδων

3.2.5 ΔΕ5: Μελέτη, σχεδιασμός και εφαρμογή τεχνικών Πρόβλεψης (Prediction).

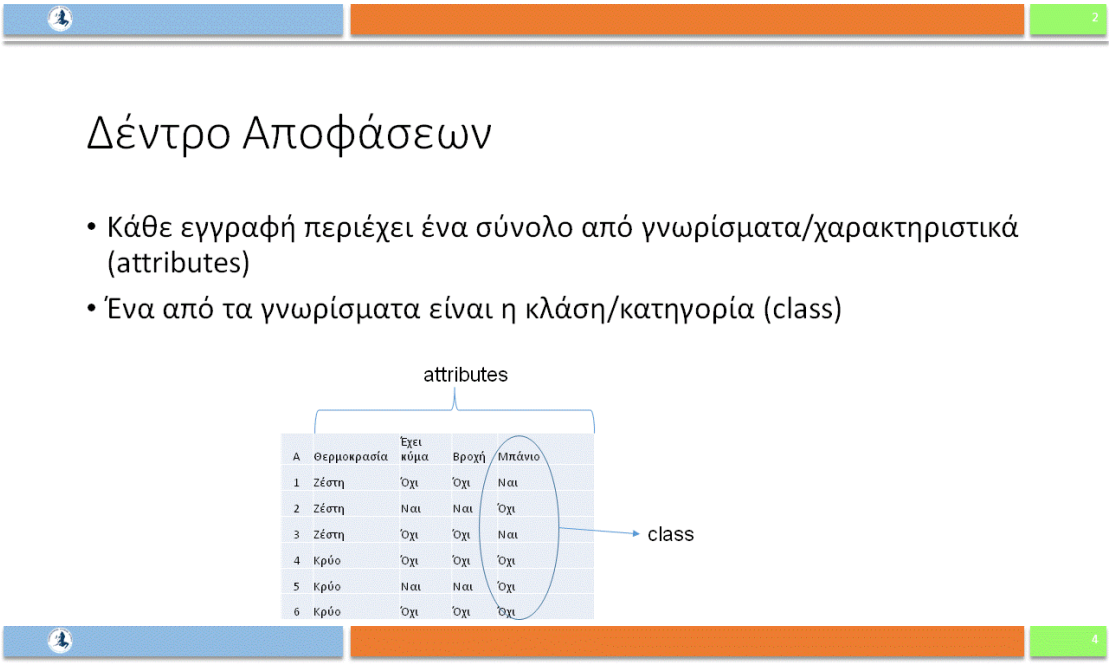
Η διδακτική ενότητα «Μελέτη, σχεδιασμός και εφαρμογή τεχνικών Πρόβλεψης» αποτελείται από 2 υποενότητες

- Εισαγωγή στους Αλγορίθμους Prediction με χρήση δέντρων αποφάσεων .
- Αναλυτικά οι αλγόριθμοι των δέντρων αποφάσεων.

3.2.5.1 Εισαγωγή στους Αλγορίθμους *Prediction* με χρήση δέντρων αποφάσεων.

Σε αυτή την ενότητα περιγράφεται η τεχνική των δέντρων αποφάσεων στην εφαρμογή μοντέλων πρόβλεψης τα οποία αποτελούν και τα δημοφιλέστερα μοντέλα κατηγοριοποίησης. Τα δέντρα αποφάσεων αναπαριστούν ένα μοντέλο πρόβλεψης το οποίο χτίζεται μέσα από μια σειρά Boolean αποφάσεων του τύπου ΝΑΙ/ΟΧΙ, μεγαλύτερο/μικρότερο,…..

Δέντρα Αποφάσεων

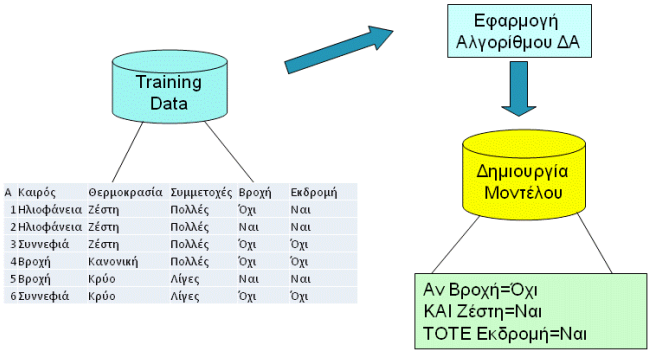


3.2.5.2 *Αναλυτικά οι αλγόριθμοι των δέντρων αποφάσεων.*

Σε αυτή την ενότητα παρουσιάζεται αναλυτικά η βασική δομή των δέντρων αποφάσεων μαζί με τα κύρια χαρακτηριστικά τους. Στην συνέχεια περιγράφονται τα δεδομένα εκπαίδευσης, τα δεδομένα επαλήθευσης των μοντέλων αλλά και οι αλγόριθμοί που χρησιμοποιούνται για την δημιουργία των δέντρων αποφάσεων. Οι αλγόριθμοι ID3 και Cart παρουσιάζονται αναλυτικά.

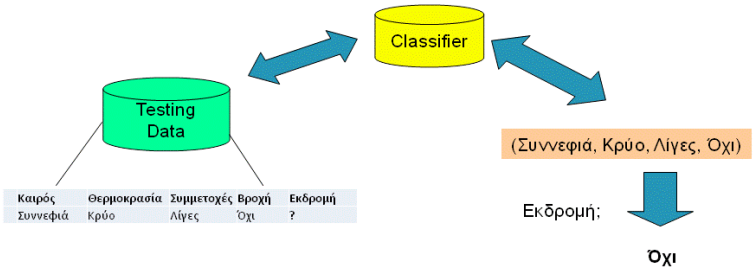
Δέντρα Αποφάσεων

ΔΑ: Κατασκευή Μοντέλου



Α	Καιρός	Θερμοκρασία	Συμμετοχές	Βροχή	Εκδρομή
1	Ηλιοφάνεια	Ζέστη	Πολλές	Όχι	Ναι
2	Ηλιοφάνεια	Ζέστη	Πολλές	Ναι	Ναι
3	Συννεφιά	Ζέστη	Πολλές	Όχι	Όχι
4	Βροχή	Κανονική	Πολλές	Όχι	Όχι
5	Βροχή	Κρύο	Λίγες	Ναι	Ναι
6	Συννεφιά	Κρύο	Λίγες	Όχι	Όχι

ΔΑ: Χρήση Μοντέλου για Πρόβλεψη



3.2.6 ΔΕ6: Μελέτη, σχεδιασμός και εφαρμογή τεχνικών Παλινδρόμησης (Regression).

Η διδακτική ενότητα «Μελέτη, σχεδιασμός και εφαρμογή τεχνικών Παλινδρόμησης» αποτελείται από 2 υποενότητες

- Εισαγωγή στους Αλγορίθμους Παλινδρόμησης.
- Παράδειγμα εφαρμογής του αλγόριθμου Γραμμικής Παλινδρόμησης.

3.2.6.1 Εισαγωγή στους Αλγορίθμους Παλινδρόμησης.

Σε αυτή την ενότητα πραγματοποιείται μια εισαγωγή στις βασικές έννοιες της παλινδρόμησης και παρουσιάζεται η εξίσωση που περιγράφει την συσχέτιση του Y με το X. Επίσης παρουσιάζεται η χαρακτηριστική γραφική παράσταση μαζί με τους αντίστοιχους συντελεστές σφάλματος. Τέλος παρουσιάζεται η μέθοδος των ελαχίστων τετραγώνων προκειμένου να προσδιοριστεί η ακρίβεια του σφάλματος στον υπολογισμό των προβλεπόμενων τιμών.

Ανάλυση Παλινδρόμησης

Ανάλυση Παλινδρόμησης – Παράδειγμα Υπολογισμού Ευθείας

Έστω ότι μια εταιρεία θέλει να μελετήσει την συσχέτιση ανάμεσα στον αριθμό των ωρών που απασχολήθηκαν οι εργαζόμενοι της σε ένα project και τις αμοιβές που έλαβαν.

Ο αριθμός των ωρών αποτελεί την ανεξάρτητη μεταβλητή

Χ=Αριθμός ωρών

Οι αμοιβές αποτελούν την εξαρτημένη μεταβλητή

Υ=Αμοιβές

3.2.6.2 Παράδειγμα εφαρμογής του αλγόριθμου Γραμμικής Παλινδρόμησης

Στην συγκεκριμένη ενότητα παρουσιάζονται αναλυτικά δυο παραδείγματα εφαρμογής της ανάλυση παλινδρόμησης και ο υπολογισμός της αντίστοιχης ευθείας.

Ανάλυση Παλινδρόμησης

Ανάλυση Παλινδρόμησης – Παράδειγμα Υπολογισμού Ευθείας

Έστω ότι μια εταιρεία θέλει να μελετήσει την συσχέτιση ανάμεσα στον αριθμό των ωρών που απασχολήθηκαν οι εργαζόμενοι της σε ένα project και τις αμοιβές που έλαβαν.

Ο αριθμός των ωρών αποτελεί την ανεξάρτητη μεταβλητή

Χ=Αριθμός ωρών

Οι αμοιβές αποτελούν την εξαρτημένη μεταβλητή

Υ=Αμοιβές

4 *Συμπεράσματα*

Σκοπός της διπλωματικής εργασίας είναι η δημιουργία ενός διαδικτυακού μαθήματος το οποίο θα παρέχει τα κατάλληλα μαθησιακά αποτελέσματα στους εκπαιδευόμενους προκειμένου να είναι σε θέση να απασχοληθούν στο τομέα την ανάλυσης δεδομένων και πιο συγκεκριμένα με την εξόρυξη δεδομένων.

Το υλικό που παρήχθη και παρουσιάστηκε αξιολογήθηκε στο πλαίσιο της διδασκαλίας τριών μεταπτυχιακών μαθημάτων:

1) Εξόρυξη Δεδομένων και Διαχείριση Δεδομένων Μεγάλης Κλίμακας
2) Οπτικοποίηση και Δεδομένα Μεγάλης Κλίμακας
3) Εξόρυξη Εκπαιδευτικών Δεδομένων και Ανάλυση της Μάθησης.

Για τις ανάγκες των τριών μαθημάτων η εξόρυξη δεδομένων θεωρήθηκε ότι είναι η τεχνική ανακάλυψης γνώσης ή προτύπων από ένα σύνολο δεδομένων με χρήση αλγορίθμων μηχανικής μάθησης και ότι βασίζεται στις αρχές της μεθοδολογία CRISP-DM.

Προς αυτή την κατεύθυνση, στο πρώτο μέρος παρουσιάστηκε η μεθοδολογία CRISP-DM και τα επιμέρους στάδια που θα πρέπει να ακολουθήσουμε για την ανάλυση προβλημάτων εξόρυξης γνώσης.

Επίσης πραγματοποιήθηκε μια εισαγωγή στους αλγορίθμους συσταδοποίησης, πρόβλεψης και εξαγωγής κανόνων συσχετίσεων.

Στη συνέχεια ακολουθήθηκε το προτεινόμενο σχέδιο μαθήματος. Στο πλαίσιο της διπλωματικής εργασίας έχουμε παρουσιάσει τα γενικά χαρακτηριστικά του μαθήματος και το σχέδιο μας περιλαμβάνει:

- το αντικείμενο και τους στόχους του μαθήματος,

- την ομάδα στόχο,
- τα μαθησιακά αποτελέσματα και
- τις εκπαιδευτικές και διδακτικές μεθόδους του μαθήματος.

Στην τελευταία ενότητα της παρούσας διπλωματικής έγινε η παρουσίαση των εκπαιδευτικών ενοτήτων του μαθήματος.

Το μάθημα φιλοξενείται στην πλατφόρμα τηλεκπαίδευσης του πανεπιστήμιου δυτικής αττικής. Εξόρυξη Δεδομένων με Χρήση Τεχνικών Μηχανικής Μάθησης.

Επίσης για τις ανάγκες του μαθήματος δημιουργήθηκαν 11 βίντεο τα οποία φιλοξενούνται στο YOUTUBE. Βίντεο Μαθήματος

Αξιολόγηση φοιτητών με έργα (projects)

Το βασικό στοιχείο αξιολόγησης των φοιτητών είναι η σχεδίαση, υλοποίηση και παρουσίαση ενός έργου. Τα έργα εκπονήθηκαν ατομικά ή από ομάδες δύο φοιτητών.

Για τη διαχείριση του έργου ζητήθηκε από του φοιτητές να προτείνουν και να περιγράψουν τη διαδικασία ανακάλυψης γνώσης σε βάσεις δεδομένων (KDD-knowledge discovery in databases process). Οι φοιτητές είχαν την υποχρέωση να προτείνουν τη χρήση ενός συγκεκριμένου πραγματικού συνόλου δεδομένων (dataset) ή να κατασκευάσουν ένα τέτοιο σύνολο με τα προγράμματά τους. Στο πλαίσιο των μαθημάτων παρουσιάστηκαν πηγές και παραδείγματα συνόλων δεδομένων.

Οι φάσεις διεκπεραίωσης του έργου και τα αντίστοιχα παραδοτέα είναι:

Παραδοτέο 1: Περιγραφή της μελέτης (του προς επίλυση προβλήματος) και των στόχων της. Περιγραφή του συνόλου δεδομένων. Παράθεση 2-3 βιβλιογραφικών αναφορών που συνδέονται με τη μελέτη.

Παραδοτέο 2: Με βάση τους στόχους της μελέτης, οι φοιτητές πρέπει να ανατρέξουν στη βιβλιογραφία που είναι σχετική και να συντάξουν 2-4 σελίδες επισκόπηση και αποτίμηση ενός άρθρου σχετικού με τη μελέτη και την έρευνά τους.

Παραδοτέο 3: Αναφορά στα εργαλεία και τα προγράμματα που χρησιμοποιήθηκαν και στα πειραματικά ευρήματα. Συζήτηση αποτελεσμάτων και μελλοντικές επεκτάσεις.

Παραδοτέο 4: Παρουσίαση σε PowerPoint και κατάθεση αναφοράς έργου.

Το πιο σημαντικό κριτήριο της αξιολόγησης ήταν το κατά πόσο υλοποιήθηκαν αυτά τα έργα. Σύμφωνα με τους διδάσκοντες αρκετά από τα έργα που παρουσιάστηκαν θα μπορούσαν να βελτιωθούν και να υποβληθούν σε συνέδρια. Σε κάθε περίπτωση όλα τα έργα αξιολογήθηκαν αρκετά θετικά και θετικά.

5

Βιβλιογραφία

1. Βασίλειος Σ. Βερύκιος, Βασίλειος Καγκλής. Ηλίας Κ. Σταυρόπουλος (2015), Η επιστήμη των δεδομένων μέσα από τη γλώσσα R, Κάλλιπος, ISBN: 978-960-603-394-0, Copyright©ΣΕΑΒ.
2. Γεώργιος Σταλίδης, Δημήτριος Καρδαράς (2015) Διαχείριση Δεδομένων και Επιχειρηματική Ευφυία, Κάλλιπος, ISBN: 978-960-603-398-8, Copyright©ΣΕΑΒ.
3. Ευστάθιος Κύρκος (2015) Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων, Κάλλιπος, ISBN: 978-960-603-109-0, Copyright©ΣΕΑΒ.
4. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth (1996) Advances in Knowledge Discovery and Data Mining, chapter From Data Mining to Knowledge Discovery: An Overview, pages 1–30. AAAI Press, Menlo Park, CA, 1996.
5. Han, J., Kamber, M. & Pei, J. (2012). Data Mining: Concepts and Techniques (3rd edition). Morgan Kaufmann, Elsevier
6. King, B. R., & Satyanarayana, A. (2013). Teaching data mining in the era of big data. Presented at the 120th American Society for Engineering Education Annual Conference and Exposition, Atlanta, GA.
7. Musicant DR (2006) A data minig course for computer science: primary sources and implementations. SIGCSE '06–Proceedings of the 37th SIGCSE technical symposium on computer science education. 538-542.
8. M. North, Data Mining for the Masses, 2012, ISBN: 978-0615684376 (This book is licensed under a Creative Commons Attribution 3.0 License)
9. Cecil Schmidt (2011) Lessons learned in the design of an undergraduate data mining course, Papers of the Seventeenth Annual CCSC Central Plains Conference April 8-9, 2011, The Journal of Computing Sciences in Colleges
10. https://docs.rapidminer.com/downloads/RapidMiner-v6-user-manual.pdf