ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ
ΠΡΟΗΓΜΕΝΕΣ ΤΕΧΝΟΛΟΓΙΕΣ ΥΠΟΛΟΓΙΣΤΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Μεταπτυχιακή Διπλωματική Εργασία

Έρευνα και επισκόπηση αλγορίθμων μηχανικής μάθησης για πολλαπλή κατηγοριοποίηση με χρήση του εργαλείου Weka και εφαρμογή στην εκπαίδευση

Συγγραφέας
Γεώργιος Μήτρου
ΑΜ: 21014

Επιβλέπων
Φοίβος Μυλωνάς

Αθήνα, Φεβρουάριος 2023

UNIVERSITY OF WEST ATTICA
SCHOOL OF ENGINEERING
DEPARTMENT OF INFORMATICS AND COMPUTER ENGINEERING
ADVANCED COMPUTING SYSTEMS

Diploma Thesis

Research and overview of machine learning algorithms for multiple classification using the Weka tool and application in education

Student name and surname: George Mitrou
Registration Number: 21014

Supervisor name and surname:
Foivos Mylonas

Athens, February 2023

ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ
ΠΡΟΗΓΜΕΝΕΣ ΤΕΧΝΟΛΟΓΙΕΣ ΥΠΟΛΟΓΙΣΤΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Έρευνα και επισκόπηση αλγορίθμων μηχανικής μάθησης για πολλαπλή κατηγοριοποίηση με χρήση του εργαλείου Weka και εφαρμογή στην εκπαίδευση

Μέλη Εξεταστικής Επιτροπής συμπεριλαμβανομένου και του Εισηγητή

Η μεταπτυχιακή διπλωματική εργασία εξετάστηκε επιτυχώς από την κάτωθι Εξεταστική Επιτροπή:

Α/Α	ΟΝΟΜΑ ΕΠΩΝΥΜΟ	ΒΑΘΜΙΔΑ/ΙΔΙΟΤΗΤΑ	ΨΗΦΙΑΚΗ ΥΠΟΓΡΑΦΗ
1.	**Φοίβος Μυλωνάς**	Αναπληρωτής Καθηγητής	Φοίβος-Απόστολος Μυλωνάς - Ψηφιακά υπογεγραμμένο 2023.02.12 20:47:20 +02'00'
2.	**Ιωάννης Βογιατζής**	Καθηγητής	Ioannis Voyiatzis Digitally signed by Ioannis Voyiatzis Date: 2023.02.23 18:48:54 +02'00'
3.	**Χρήστος Τρούσσας**	Επίκουρος Καθηγητής	Christos Troussas Digitally signed by Christos Troussas Date: 2023.02.13 10:35:25 +02'00'

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Ο/η κάτωθι υπογεγραμμένος Μήτρου Γεώργιος του Αναστασίου, με αριθμό μητρώου 21014 φοιτητής του Προγράμματος Μεταπτυχιακών Σπουδών Προηγμένες Τεχνολογίες Υπολογιστικών Συστημάτων του Τμήματος Μηχανικών Πληροφορικής και Υπολογιστών της Σχολής Μηχανικών του Πανεπιστημίου Δυτικής Αττικής, δηλώνω ότι:

«Είμαι συγγραφέας αυτής της μεταπτυχιακής 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 μένα 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».

**Επιθυμώ την απαγόρευση πρόσβασης στο πλήρες κείμενο της εργασίας μου μέχρι …… και έπειτα από αίτηση μου στη Βιβλιοθήκη και έγκριση του επιβλέποντα καθηγητή.*

Ο/Η Δηλών/ούσα

GEORGIOS MITROU
Digitally signed by GEORGIOS MITROU
Date: 2023.02.10 14:11:16 +02'00'

*** Ονοματεπώνυμο /Ιδιότητα**

Ψηφιακή Υπογραφή Επιβλέποντα
(Υπογραφή)

Φοίβος-Απόστολος Μυλωνάς - Ψηφιακά υπογεγραμμένο
2023.02.12 20:47:45 +02'00'

**** Εάν κάποιος επιθυμεί απαγόρευση πρόσβασης στην εργασία για χρονικό διάστημα 6-12 μηνών (embargo), θα πρέπει να υπογράψει ψηφιακά ο/η επιβλέπων/ουσα καθηγητής/τρια, για να γνωστοποιεί ότι είναι ενημερωμένος/η και συναινεί. Οι λόγοι χρονικού αποκλεισμού πρόσβασης περιγράφονται αναλυτικά στις πολιτικές του Ι.Α. (σελ. 6):***
https://www.uniwa.gr/wp-content/uploads/2021/01/%CE%A0%CE%BF%CE%BB%CE%B9%CF%84%CE%B9%CE%BA%CE%B5%CC%81%CF%82_%CE%99%CE%B4%CF%81%CF%85%CE%BC%CE%B1%CF%84%CE%B9%CE%BA%CE%BF%CF%85%CC%81_%CE%91%CF%80%CE%BF%CE%B8%CE%B5%CF%84%CE%B7%CF%81%CE%B9%CC%81%CE%BF%CF%85_final.pdf

Στη μνήμη των γονέων μου

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον καθηγητή κύριο Φοίβο Μυλωνά για την ευκαιρία που μου έδωσε να ασχοληθώ με ένα τόσο ενδιαφέρον θέμα, για τη βοήθειά του καθώς και για την υπομονή του.

Ακόμα, θα ήθελα να ευχαριστήσω και την οικογένειά μου για την κατανόηση και τη συμπαράστασή τους.

Περίληψη

Η παρούσα διπλωματική εργασία προσεγγίζει ένα θέμα που απασχολεί την ερευνητική κοινότητα, την εφαρμογή αλγορίθμων μηχανικής μάθησης στην εκπαίδευση. Γίνεται εφαρμογή τεχνικών και μεθόδων από το πεδίο της μηχανικής μάθησης με στόχο την πρόβλεψη της απόδοσης των μαθητών, στόχος που επιπλέον παρουσιάζει ενδιαφέρον για την εκπαιδευτική και τη μαθητική κοινότητα.

Αρχικά, γίνεται μια καταγραφή του επιστημονικού πεδίου της μηχανικής μάθησης παρουσιάζοντας και τα τρία είδη μηχανικής μάθησης, τη μάθηση με επίβλεψη, τη μάθηση χωρίς επίβλεψη και τη μάθηση με ενίσχυση. Πραγματοποιήθηκε εκτενέστερη αναφορά στη μάθηση με επίβλεψη και ειδικότερα στους σημαντικότερους αλγορίθμους κατηγοριοποίησης – ταξινόμησης (classification) που παρουσιάστηκαν αναλυτικά.

Καταγράφονται προηγούμενες έρευνες που δείχνουν ότι η χρήση αλγορίθμων κατηγοριοποίησης υλοποιείται για την πρόβλεψη της απόδοσης των μαθητών, αλλά και για άλλους μαθησιακούς στόχους με ποικίλα ποσοστά επιτυχίας.

Στη συνέχεια γίνεται παρουσίαση του λογισμικού εξόρυξης γνώσης Weka, που αποτελεί ένα δημοφιλές εργαλείο έρευνας και εφαρμογής αλγορίθμων μηχανικής μάθησης. Οι αλγόριθμοι του Weka εκπαιδεύτηκαν σε ένα σύνολο εκπαιδευτικών δεδομένων που αντλήθηκε από το Διαδίκτυο. Τα δεδομένα χωρίστηκαν σε 2 και σε 5 κλάσεις και εκτελέστηκαν οι αλγόριθμοι κατηγοριοποίησης. Σύμφωνα με τα αποτελέσματα των πειραμάτων οι μέθοδοι κατηγοριοποίησης μπορούν να εφαρμοστούν επιτυχώς, καθώς η απόδοση τους κυμάνθηκε από 80% έως 95% περίπου.

Λέξεις κλειδιά

Μηχανική μάθηση, κατηγοριοποίηση, ταξινόμηση, πρόβλεψη, απόδοση μαθητών, Weka

Abstract

This diploma thesis approaches a topic that concerns the research community, the application of machine learning algorithms in education. Techniques and methods from the field of machine learning are applied with the aim of predicting student performance, a goal that is also of interest to the educational and student community.

First, an overview of the scientific field of machine learning is given, presenting all three types of machine learning, supervised learning, unsupervised learning, and reinforcement learning. A more extensive reference was made to supervised learning and in particular to the most important classification algorithms that were presented in detail.

Previous research is documented showing that the use of categorization algorithms is implemented to predict student performance, but also for other learning objectives with varying success rates.

Weka data mining software is then presented, which is a popular tool for research and application of machine learning algorithms. Weka's algorithms were trained on a training data set from the Internet. The data were divided into 2 and 5 classes and the classification algorithms were performed. According to the results of the experiments, the classification methods can be applied successfully, as their performance ranged from 80% to 95% approximately.

Keywords

Machine learning, classification, prediction, student’s performance, Weka

Περιεχόμενα

1. Εισαγωγή 16
   1.1 Αντικείμενο της Διπλωματικής Εργασίας 18
   1.2 Δομή της Διπλωματικής Εργασίας 18
2. Μηχανική Μάθηση 19
   2.1 Εισαγωγή 19
   2.2 Ιστορική αναδρομή 20
   2.3 Είδη μηχανικής μάθησης 20
3. Μάθηση με επίβλεψη 21
   3.1 Εισαγωγή 21
   3.2 Γραμμική Παλινδρόμηση 23
   3.3 Μέθοδοι κατηγοριοποίησης 25
      3.3.1 Δέντρα απόφασης / ταξινόμησης 25
         3.3.1.1 Ο αλγόριθμος ID3 28
         3.3.1.2 Οι αλγόριθμοι C4.5 και C5.0 30
      3.3.2 Κατηγοριοποίηση με βάση τους κοντινότερους γείτονες 32
      3.3.3 Κατηγοριοποιητής Κανόνων 34
      3.3.4 Κατηγοριοποιητές Bayes 35
         3.3.4.1 Ο αφελής κατηγοριοποιητής Bayes 36
         3.3.4.2 Δίκτυα Bayes 37
      3.3.5 Μηχανές διανυσμάτων υποστήριξης 39
      3.3.6 Τεχνητά Νευρωνικά δίκτυα 42
4. Μάθηση χωρίς επίβλεψη 46
   4.1 Εισαγωγή 46
   4.2 Ανακάλυψη Κανόνων Συσχέτισης 47
   4.3 Ομαδοποίηση 48
5. Μάθηση με ενίσχυση 50
6. Εφαρμογές της μηχανικής μάθησης 52
7. Βιβλιογραφική επισκόπηση 54
8. WEKA 57
   8.1 Εισαγωγή 57
   8.2 Αλγόριθμοι και δυνατότητες του Weka. 58
   8.3 Η μορφή δεδομένων του WEKA. 58
   8.4 Το περιβάλλον διεπαφής (GUI) του Weka. 60
   8.5 Ο Explorer του Weka. 67
      8.5.1 Η καρτέλα Preprocess 67
      8.5.2 Η καρτέλα Classify 71
         8.5.2.1 Κριτήρια Εκτίμησης Αλγορίθμων 76
      8.5.3 Η καρτέλα Cluster 79

8.5.4 Η καρτέλα Associate 79
8.5.5 Η καρτέλα Select attributes..... 80
8.5.6 Η καρτέλα Visualize..... 80
9. Εκτέλεση αλγορίθμων..... 83
9.1 Δεδομένα..... 83
9.2 Προεπεξεργασία των δεδομένων..... 83
9.3 Αποτελέσματα αλγορίθμων κατηγοριοποίησης σε δεδομένα 5 κλάσεων 85
9.4 Αποτελέσματα αλγορίθμων κατηγοριοποίησης σε δεδομένα 2 κλάσεων 92
10. Συμπεράσματα..... 96
Βιβλιογραφία – Αναφορές..... 98

Κατάλογος Εικόνων

Εικόνα 1: Εξόρυξη δεδομένων σε συστήματα εκπαίδευσης 17
Εικόνα 2: Μάθηση με επίβλεψη 21
Εικόνα 3: Παράδειγμα απλής γραμμικής παλινδρόμησης 24
Εικόνα 4: Σφάλμα στην απλή γραμμική παλινδρόμηση 24
Εικόνα 5: Γενική αναπαράσταση δέντρου ταξινόμησης 26
Εικόνα 6: Διαφορετικά δέντρα προερχόμενα από το ίδιο σύνολο δεδομένων 27
Εικόνα 7: Προσδιορισμός κατηγορίας με βάση τους 3 και 7 κοντινότερους γείτονες 33
Εικόνα 8: Υπό συνθήκη ανεξάρτητες μεταβλητές 38
Εικόνα 9: Δίκτυο Bayes με τον αντίστοιχο Πίνακα Υπό Συνθήκη Πιθανοτήτων 39
Εικόνα 10: Υπερεπίπεδα στον $\mathbb{R}^2$ και στον $\mathbb{R}^3$ 40
Εικόνα 11: Βέλτιστο υπερεπίπεδο και διανύσματα υποστήριξης 41
Εικόνα 12: Μοντέλο Τεχνητού Νευρώνα 43
Εικόνα 13: Γράφος νευρωνικού δικτύου 44
Εικόνα 14: Μάθηση χωρίς επίβλεψη 46
Εικόνα 15:Ομαδοποίηση δεδομένων 48
Εικόνα 16: Μάθηση με ενίσχυση 50
Εικόνα 17: Η σελίδα του Weka στο πανεπιστήμιο Waikato 57
Εικόνα 18: Το logo του προγράμματος και το πουλί weka 58
Εικόνα 19: Δομή αρχείου .arff 59
Εικόνα 20: Η κεντρική οθόνη του προγράμματος Weka 61
Εικόνα 21: Ο Weka Explorer 61
Εικόνα 22: Η οθόνη αποτελεσμάτων του Experimenter 62
Εικόνα 23: Παράδειγμα χρήσης του KnowledgeFlow 63
Εικόνα 24: Η κύρια οθόνη του περιβάλλοντος Workbench 64
Εικόνα 25:Το περιβάλλον γραμμής εντολών (SimpleCLI) του Weka 65
Εικόνα 26: Ο package manager του Weka 65
Εικόνα 27: Το σύνολο δεδομένων όπως εμφανίζεται στην εφαρμογή ArffViewer 66
Εικόνα 28: Υποστηριζόμενοι τύποι αρχείων του ArffViewer 66
Εικόνα 29: Η καρτέλα Preprocess 67
Εικόνα 30: Το μενού επιλογών των διαθέσιμων φίλτρων του Weka 69
Εικόνα 31: Επιλογές του φίλτρου Attribute selection 70
Εικόνα 32: Οπτικοποίηση του συνόλου δεδομένων 71
Εικόνα 33: Οι αλγόριθμοι κατηγοριοποίησης του Weka 72
Εικόνα 34: Πλαίσιο πληροφοριών για τον αλγόριθμο DecisionStump. 72

Εικόνα 35: Οι παράμετροι του αλγορίθμου RandromTree 73
Εικόνα 36: Cross-validation 74
Εικόνα 37: Αποτελέσματα εκτέλεσης του Naïve Bayes 75
Εικόνα 38: Πίνακας σύγχυσης 76
Εικόνα 39: Παράδειγμα καμπύλης ROC 77
Εικόνα 40: Η καρτέλα Cluster 79
Εικόνα 41: Η καρτέλα Visualize 81
Εικόνα 42: Διάγραμμα διασποράς δύο μεταβλητών 82
Εικόνα 43: Διαδικασία εξόρυξης δεδομένων 83
Εικόνα 44: Οι ρυθμίσεις του φίλτρου Discretize 85
Εικόνα 45: Αποτελέσματα καρτέλας Select Attributes 89
Εικόνα 46: Αποτελέσματα καρτέλας Select Attributes 94

Κατάλογος Πινάκων

Πίνακας 1: Απόδοση αλγορίθμων στο σύνολο των δεδομένων 88
Πίνακας 2: Απόδοση αλγορίθμων στο νέο σύνολο δεδομένων που περιλαμβάνει μόνο τα σημαντικά χαρακτηριστικά 91
Πίνακας 3: Αποτελέσματα αλγορίθμων σε δεδομένα 2 κλάσεων 93
Πίνακας 4: Αποτελέσματα μετά τη διαγραφή του χαρακτηριστικού lunch 95

1. Εισαγωγή

Η αλματώδης τεχνολογική πρόοδος και οι εφαρμογές της τεχνολογίας στη ζωή της ανθρωπότητας έχουν αλλάξει ριζικά την καθημερινότητα όλων. Οι ηλεκτρονικές συσκευές (ηλεκτρονικοί υπολογιστές, tablets, κινητά, smartwatches κ.ά.) αποτελούν βοηθό στις καθημερινές εργασίες και για πολλούς αχώριστο και απαραίτητο συνεργάτη. Το Διαδίκτυο έχει διαμορφώσει τον τρόπο που οι άνθρωποι επικοινωνούν και εργάζονται. Η χρήση του Διαδικτύου είναι πλέον συνεχής.

Η εισβολή της τεχνολογίας δεν θα μπορούσε να απουσιάζει από τον κεφαλαιώδους σημασίας τομέα της εκπαίδευσης. Το τοπίο στην εκπαίδευση έχει τροποποιηθεί. Τα τελευταία χρόνια η παραδοσιακή διδασκαλία έχει κατ' ελάχιστον εμπλουτιστεί με τεχνολογικά μέσα, ενώ είναι σαφές πως εν μέρει ή ολικά έχει αντικατασταθεί από σύγχρονα εκπαιδευτικά συστήματα. Στην Αμερική ήδη από το 2010, περισσότεροι από έξι εκατομμύρια φοιτητές, περίπου το ένα τρίτο των φοιτητών της τριτοβάθμιας εκπαίδευσης, είχαν πραγματοποιήσεις εγγραφές σε μαθήματα που παρέχονταν πλήρως μέσω του Διαδικτύου (Picciano 2012).

Ως συνέπεια των παραπάνω, οι εκπαιδευτικές συμπεριφορές και ανάγκες των μαθητών έχουν τροποποιηθεί. Η μελέτη των νέων συνθηκών και, γενικά, η βελτίωση της παρεχόμενης εκπαίδευσης, είναι ένα πεδίο που έχει τραβήξει την προσοχή πολλών ερευνητών.

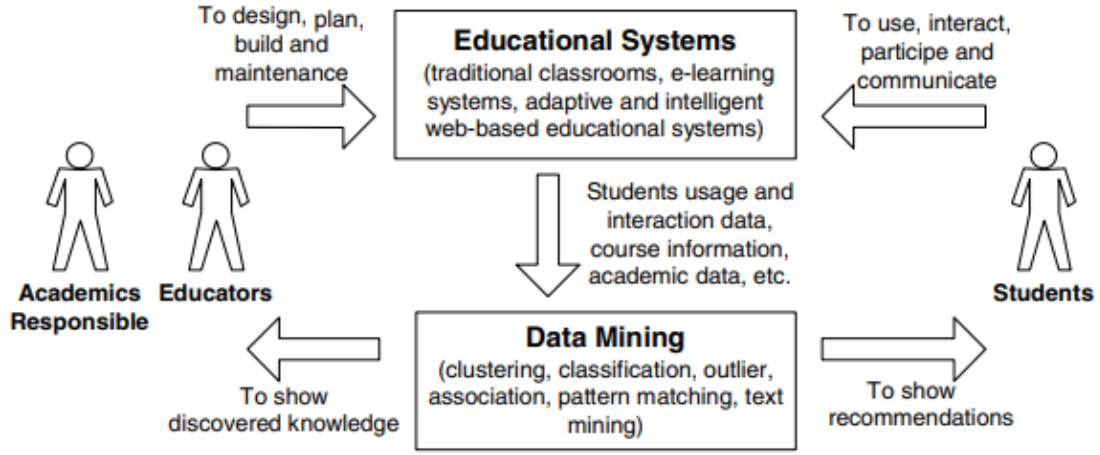
Γενικά, η μηχανική μάθηση μπορεί να βρει πολλές εφαρμογές στον τομέα της εκπαίδευσης. Ένα εκπαιδευτικό ίδρυμα έχει συχνά πολλές και ποικίλες πηγές πληροφοριών: τις κλασσικές βάσεις δεδομένων με πληροφορίες για τους μαθητές, τους καθηγητές, το πρόγραμμα κ.ά., Διαδικτυακές πληροφορίες από τις σχετικές με τα προσφερόμενα μαθήματα ιστοσελίδες και βάσεις δεδομένων πολυμέσων. Ακόμα, υπάρχουν πολλές διαφορετικές ομάδες ανθρώπων που έχουν άμεσο ενδιαφέρον και δημιουργούν διαφορετικές ανάγκες εξόρυξης πληροφοριών. Για παράδειγμα οι διαχειριστές και ακαδημαϊκοί υπεύθυνοι χρειάζονται πληροφορίες σχετικές με την εγγραφή και αποδοχή των μαθητών στα προγράμματα. Από την άλλη μεριά οι μαθητές ενδιαφέρονται για τη βέλτιστη επιλογή των μαθημάτων βασιζόμενοι σε προβλέψεις για την πορεία τους στα επιλεγμένα μαθήματα. Είναι λοιπόν φανερό, από το πλήθος των πληροφοριών και των αναγκών ότι ένα ολοκληρωμένο σύστημα μηχανικής μάθησης το οποίο θα είναι σε θέση να καλύψει αυτές τις ειδικές ανάγκες, θα έχει μεγάλη ζήτηση και εφαρμογή (Κωτσιαντής 2012).

Η χρήση τεχνικών μηχανικής μάθησης σε εκπαιδευτικά δεδομένα, ονομάζεται Εξόρυξη Εκπαιδευτικών Δεδομένων (Educational Data Mining) και είναι ένα ερευνητικό πεδίο που έχει σημειώσει μεγάλη ανάπτυξη τα τελευταία χρόνια. Ένας πιο λεπτομερής ορισμός όπως δίνεται στον ιστότοπο educationaldatamining.org, είναι ότι η Εξόρυξη Εκπαιδευτικών Δεδομένων είναι ένας αναδυόμενος κλάδος που ασχολείται με την ανάπτυξη μεθόδων για την εξερεύνηση των μοναδικών και ολοένα και πιο μεγάλης κλίμακας δεδομένων που προέρχονται από

εκπαιδευτικά περιβάλλοντα και τη χρήση αυτών των μεθόδων για την καλύτερη κατανόηση των μαθητών και των παραμέτρων μέσω των οποίων μαθαίνουν.

Σύμφωνα με τους Baker & Yacef 2009 στοχεύει στην πρόβλεψη της μαθησιακής συμπεριφοράς των μαθητών, στην ανακάλυψη νέων και βελτίωση των υπαρχόντων μαθησιακών μοντέλων, στη μελέτη της επίδρασης της εκπαιδευτικής υποστήριξης και στην ανάπτυξη της επιστημονικής γνώσης που σχετίζεται με τους εκπαιδευτές και τους μαθητές.

Στην εικόνα 1 απεικονίζεται ο κύκλος εφαρμογής τεχνικών εξόρυξης δεδομένων στα εκπαιδευτικά συστήματα.



Εικόνα 1: Εξόρυξη δεδομένων σε συστήματα εκπαίδευσης

Πηγή: Romero & Ventura 2007

1.1 Αντικείμενο της Διπλωματικής Εργασίας

Σκοπός της παρούσας διπλωματικής εργασίας είναι να παρουσιάσει τον τομέα της μηχανικής μάθησης, εστιάζοντας στη μάθηση με επίβλεψη και κυρίως στις μεθόδους κατηγοριοποίησης. Ακολούθως, να μελετήσει την απόδοση των αλγορίθμων κατηγοριοποίησης σε δεδομένα σχετικά με την εκπαίδευση ερευνώντας πιθανή συμβολή στην διδασκαλία. Μέσω της πρόβλεψης της απόδοσης των μαθητών μπορούν να υλοποιηθούν δράσεις και παρεμβάσεις, τροποποιήσεις στην εκπαιδευτική διαδικασία ώστε να επιτευχθεί κατά το δυνατόν η βελτίωση της απόδοσης του μαθητή.

Η εφαρμογή των αλγορίθμων θα γίνει με χρήση του προγράμματος Weka, το περιβάλλον και οι δυνατότητες του οποίου επίσης θα παρουσιαστούν. Πρόκειται για λογισμικό ανοικτού κώδικα που προσφέρεται δωρεάν και αποτελεί ένα από τα δημοφιλέστερα λογισμικά μηχανικής μάθησης, έχοντας μεγάλη απήχηση σε εφαρμογές αυτής της κατηγορίας.

1.2 Δομή της Διπλωματικής Εργασίας

Η παρούσα διπλωματική εργασία αναπτύσσεται σε δέκα κεφάλαια, ως εξής:

Στο παρών πρώτο κεφάλαιο γίνεται μια εισαγωγή, καταγράφεται ο στόχος της εργασίας και παρουσιάζεται η δομή της.

Το δεύτερο κεφάλαιο περιλαμβάνει μια εισαγωγή στον τομέα της μηχανικής μάθησης.

Στο τρίτο κεφάλαιο γίνεται εκτενής παρουσίαση της μάθησης με επίβλεψη και των μεθόδων της με εστιάζοντας στις μεθόδους κατηγοριοποίησης.

Στο τέταρτο κεφάλαιο παρουσιάζεται η μάθηση χωρίς επίβλεψη.

Στο πέμπτο κεφάλαιο παρουσιάζεται η μάθηση με ενίσχυση.

Στο έκτο κεφάλαιο καταγράφονται τα πεδία εφαρμογής της μηχανική μάθησης.

Στο έβδομο κεφάλαιο γίνεται καταγραφή και παρουσίαση προηγούμενων εργασιών σχετικών με τις μεθόδους κατηγοριοποίησης στην εκπαίδευση.

Στο όγδοο κεφάλαιο παρουσιάζεται το λογισμικό εξόρυξης γνώσης Weka και περιγράφονται οι λειτουργίες του.

Στο ένατο κεφάλαιο παρουσιάζεται το σύνολο δεδομένων (dataset) που χρησιμοποιήθηκε και καταγράφονται τα αποτελέσματα της εκτέλεσης των αλγορίθμων κατηγοριοποίησης.

Στο δέκατο κεφάλαιο καταγράφονται τα συμπεράσματα που προέκυψαν από την εργασία καθώς και οι ενέργειες που προτείνεται να εκτελεστούν μελλοντικά.

Στη συνέχεια παρατίθεται η βιβλιογραφία και οι αναφορές της εργασίας.

2. Μηχανική Μάθηση

2.1 Εισαγωγή

Η Τεχνητή Νοημοσύνη (Artificial Intelligence), σύμφωνα με τους Βλαχάβα et al 2006, είναι ο τομέας της επιστήμης των υπολογιστών που ασχολείται με τη σχεδίαση και την υλοποίηση προγραμμάτων τα οποία είναι ικανά να μιμηθούν τις ανθρώπινες γνωστικές ικανότητες, εμφανίζοντας έτσι χαρακτηριστικά που συνήθως αποδίδονται σε ανθρώπινη συμπεριφορά, όπως η επίλυση προβλημάτων, η αντίληψη μέσω της όρασης, η μάθηση, η εξαγωγή συμπερασμάτων, η κατανόηση φυσικής γλώσσας, κ.λπ.

Ο κλάδος στον οποίο τα συλλεχθέντα δεδομένα, επεξεργάζονται και αναλύονται με πληθώρα διαφορετικών τεχνικών, με τελικό στόχο την αυτοματοποιημένη δημιουργία μοντέλων, ονομάζεται μηχανική μάθηση, που αποτελεί κομμάτι της Τεχνητής Νοημοσύνης.

Για τη μηχανική μάθηση έχουν δοθεί διάφοροι ορισμοί που μας βοηθούν στην καλύτερη κατανόησή της:

- «Πεδίο μελέτης που δίνει στους υπολογιστές την ικανότητα να μαθαίνουν, χωρίς να έχουν ρητά προγραμματιστεί» (Samuel 1959)
- «Η μελέτη υπολογιστικών μεθόδων για την απόκτηση νέας γνώσης, νέων δεξιοτήτων και νέων τρόπων οργάνωσης της υπάρχουσας γνώσης» (Carbonell 1987)
- «Ένα πρόγραμμα υπολογιστή λέμε ότι μαθαίνει από την εμπειρία E ως προς κάποια κλάση εργασιών T και μέτρο απόδοσης P, αν η απόδοσή του σε εργασίες από το T, όπως μετριέται από το P, βελτιώνεται μέσω της εμπειρίας E» (Mitchell 1997)
- «Κάτι μαθαίνει όταν αλλάζει τη συμπεριφορά του κατά τέτοιο τρόπο ώστε να αποδίδει καλύτερα στο μέλλον» (Witten & Frank 2000)

Η μηχανική μάθηση ασχολείται με τη δημιουργία αλγορίθμων οι οποίοι αλληλοεπιδρούν με τα δεδομένα, «μαθαίνουν» από τα δεδομένα και δημιουργούν μοντέλα. Στόχος των παραγόμενων μοντέλων είναι οι όσο το δυνατόν ακριβέστερες και επιτυχείς προβλέψεις και η λήψη των βέλτιστων δυνατών αποφάσεων.

Οι αλγόριθμοι της μηχανικής μάθησης αντιμετωπίζουν και επιλύουν σύνθετα προβλήματα τα οποία δεν έχουν λυθεί ή είναι πολύ δύσκολο να αντιμετωπιστούν με τις παραδοσιακές προγραμματιστικές τεχνικές.

Ο συνεχώς αυξανόμενος όγκος των δεδομένων καθώς και οι ποικίλες μορφές τους, η αλματώδης αύξηση της υπολογιστικής ισχύος και η προσιτή αποθήκευση δεδομένων έχουν συντελέσει στην αύξηση του ενδιαφέροντος για τη μηχανική μάθηση.

2.2 Ιστορική αναδρομή

Καταγράφοντας, εν συντομία, τους βασικούς ιστορικούς σταθμούς της μηχανικής μάθησης, η αφετηρία βρίσκεται στο 1950, όπου ο Alan Turing δημοσίευσε μια εργασία στην οποία έθεσε το ερώτημα: «Μπορούν οι μηχανές να σκεφτούν;». Ακόμα, δημιούργησε το «Test Turing» προκειμένου να ελέγξει αν μια μηχανή είναι «έξυπνη», ορίζοντας ότι, η μηχανή πρέπει να ξεγελάσει έναν άνθρωπο και να τον κάνει να πιστεύσει ότι είναι ένας άλλος άνθρωπος και όχι ένας υπολογιστής.

Το 1952 ο Arthur Samuel δημοσίευσε λογισμικό που έπαιζε «ντάμα» (checkers) και βελτιωνόταν σε κάθε παρτίδα. Ο ίδιος, το 1959 ήταν ο πρώτος που χρησιμοποίησε τον όρο «Machine Learning».

Ανάμεσα στα έτη 1980 και 1987 δημιουργήθηκαν τα πρώτα «Expert Systems» που ήταν βασισμένα σε κανόνες και έτσι αυξήθηκε το ενδιαφέρον για τη μηχανική μάθηση. Το 1997 ο Deep Blue της IBM, αναμετρήθηκε εκ νέου και νίκησε τον παγκόσμιο πρωταθλητή Gary Kasparov στο σκάκι, αποτελώντας τον πρώτο υπολογιστή που κατάφερε να νικήσει έναν ειδικό.

Τέλος, το 2017 η εταιρία Waymo κυκλοφόρησε τα πρώτα αυτόνομα αυτοκίνητα που κινούνταν χωρίς ανθρώπινη παρέμβαση και παρουσία.

2.3 Είδη μηχανικής μάθησης

Τα είδη μηχανικής μάθησης είναι:

- μάθηση με επίβλεψη (supervised learning)
- μάθηση χωρίς επίβλεψη (unsupervised learning)
- μάθηση με ενίσχυση (reinforcement learning)

και παρουσιάζονται στις επόμενες σελίδες. Κάθε είδος περιλαμβάνει τεχνικές και μεθόδους που εφαρμόζονται στα προς επίλυση προβλήματα.

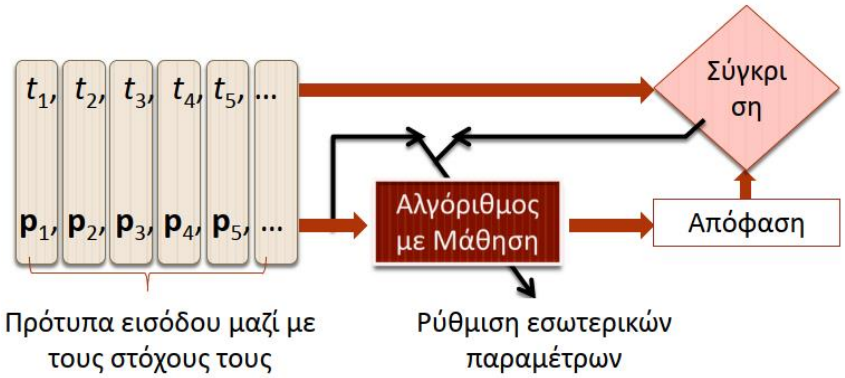
3. Μάθηση με επίβλεψη

3.1 Εισαγωγή

Στη μάθηση με επίβλεψη στόχος είναι η εκπαίδευση του συστήματος ώστε να «μάθει» μια συνάρτηση. Η συνάρτηση αυτή αποκαλείται συνάρτηση στόχος και θα χρησιμοποιηθεί για να προβλεφθούν οι άγνωστες τιμές νέων δεδομένων οι οποίες θα εισαχθούν στο σύστημα. Οι τιμές αυτές είναι και το τελικό ζητούμενο της διαδικασίας.

Η μάθηση της συνάρτησης στόχου από μια μέθοδο της μάθησης με επίβλεψη, γίνεται με χρήση των δεδομένων εκπαίδευσης. Τα δεδομένα εκπαίδευσης είναι συνήθως ζεύγη εκ των προτέρων γνωστών τιμών, δηλαδή γνωστής εισόδου αλλά και γνωστής εξόδου, που είναι η τιμή της συνάρτησης στόχου. Η ονομασία μάθηση με επίβλεψη προέρχεται καθώς παρέχεται η σωστή τιμή της συνάρτησης εξόδου ανάλογα με τα δεδομένα που δίνονται ως είσοδο.

Πιο συγκεκριμένα, χρησιμοποιούνται ως είσοδο τα πρότυπα εισόδου p_1, p_2, p_3, … μαζί με τους αντίστοιχους γνωστούς στόχους t_1, t_2, t_3,… (εικόνα 2). Τα πρότυπα είναι συνήθως διανύσματα κάποιου n-διάστατου χώρου.



Εικόνα 2: Μάθηση με επίβλεψη

Πηγή: Διαμαντάρας & Μπότσης 2019

Σύμφωνα με τους Βλαχάβα et al 2006, η μάθηση με επίβλεψη βασίζεται στην υπόθεση της επαγωγικής μάθησης:

«Κάθε υπόθεση που έχει βρεθεί να προσεγγίζει καλά τη συνάρτηση στόχο για ένα αρκετά μεγάλο σύνολο παραδειγμάτων, θα προσεγγίζει το ίδιο καλά τη συνάρτηση στόχο και για περιπτώσεις που δεν έχει εξετάσει».

Έτσι, βάσει επαγωγής, αφού η συνάρτηση στόχος έχει εκπαιδευτεί και προσεγγίζει καλά τα γνωστά δεδομένα, θα προσεγγίζει καλά και τα άγνωστα δεδομένα.

Το σύνολο δεδομένων εισόδου, των οποίων οι τιμές είναι γνωστές, χωρίζεται σε δύο υποσύνολα. Το πρώτο υποσύνολο αποτελεί το σύνολο εκπαίδευσης (training set) και

χρησιμοποιείται ώστε το σύστημα να «μάθει» τη συνάρτηση στόχο και να κατασκευαστεί το μοντέλο. Το δεύτερο υποσύνολο ονομάζεται σύνολο ελέγχου (test set) και χρησιμοποιείται για τον έλεγχο και την επικύρωσή του μοντέλου.

Τα προβλήματα μάθησης με επίβλεψη είναι τα πιο δημοφιλή ανάμεσα στα προβλήματα μηχανικής μάθησης. Διακρίνονται σε δύο κατηγορίες ανάλογα με την έξοδο της συνάρτησης που κατασκευάζεται:

- προβλήματα κατηγοριοποίησης ή ταξινόμησης (classification): Οι στόχοι είναι διακριτές τιμές, δηλαδή $t_i \in \{0,1\}$ ή $t_i \in \{0,1,...,C\}$, και αντιστοιχούν σε κλάσεις αντικειμένων.
- προβλήματα παλινδρόμησης (regression): Οι στόχοι είναι είτε συνεχείς τιμές είτε διακριτές τιμές των οποίων όμως το πλήθος είναι απεριόριστο, δηλαδή $t_i \in R$, και αντιστοιχούν σε τιμές κάποιων ποσοτήτων.

Και στις δύο κατηγορίες προβλημάτων της μάθησης με επίβλεψη, ο στόχος είναι κοινός: η δημιουργία ενός μοντέλου πρόβλεψης, το οποίο ονομάζεται κατηγοριοποιητής ή ταξινομητής (classifier). Με τη βοήθεια του κατηγοριοποιητή γίνεται πρόβλεψη της τιμής μιας μεταβλητής που δεν είναι γνωστή, με χρήση των ήδη γνωστών τιμών άλλων μεταβλητών. Οι μεταβλητές των οποίων γνωρίζουμε τις τιμές τους ονομάζονται ανεξάρτητες μεταβλητές ενώ εξαρτημένη μεταβλητή ονομάζεται η μεταβλητή για τις τιμές της οποίας γίνεται η πρόβλεψη.

Οι βασικότερες τεχνικές της μηχανικής μάθησης με επίβλεψη είναι:

- Γραμμική Παλινδρόμηση (Linear Regression)
- Δέντρα Απόφασης / Ταξινόμησης (Decision / Classification Trees)
- Κατηγοριοποίηση με βάση τους κοντινότερους γείτονες (k-Nearest Neighbors)
- Κατηγοριοποιητής Κανόνων (Rule-Based Classifier)
- Κατηγοριοποιητές Bayes
- Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines)
- Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks)

3.2 Γραμμική Παλινδρόμηση

Παλινδρόμηση (regression) είναι η διαδικασία με την οποία προσδιορίζεται η σχέση της εξαρτημένης μεταβλητής y που ονομάζεται και έξοδος, με τις ανεξάρτητες μεταβλητές x_1, x_2, …, x_m που αποκαλούνται και είσοδοι. Σκοπός της παλινδρόμησης είναι η πρόβλεψη της τιμής της μεταβλητής y όταν γνωρίζουμε τις τιμές των ανεξάρτητων μεταβλητών. Σύμφωνα με τους Tan et al 2006, στόχος της παλινδρόμησης είναι να βρει μια συνάρτηση η οποία να μπορεί να προσαρμοστεί στα δεδομένα εισόδου με ένα ελάχιστο σφάλμα. Η ανάλυση μέσω παλινδρόμησης πραγματοποιείται για να προσδιοριστούν οι συσχετίσεις ανάμεσα σε δύο ή περισσότερες μεταβλητές οι οποίες έχουν τη σχέση αιτίου-αποτελέσματος και για να γίνονται προβλέψεις για το θέμα που αναλύεται, χρησιμοποιώντας τις συσχετίσεις αυτές.

Στην παλινδρόμηση εφαρμόζονται τεχνικές για τη μοντελοποίηση της σχέσης της εξαρτημένης και των ανεξάρτητων μεταβλητών. Η εξαρτημένη μεταβλητή μοντελοποιείται ως συνάρτηση των ανεξάρτητων μεταβλητών, των συντελεστών παλινδρόμησης και ενός τυχαίου όρου σφάλματος (Xin & Xiao 2009).

Υπάρχουν διάφορα μοντέλα παλινδρόμησης. Στη γραμμική παλινδρόμηση (linear regression) η αναμενόμενη τιμή εξόδου μοντελοποιείται με μια γραμμική συνάρτηση των μεταβλητών εισόδου:

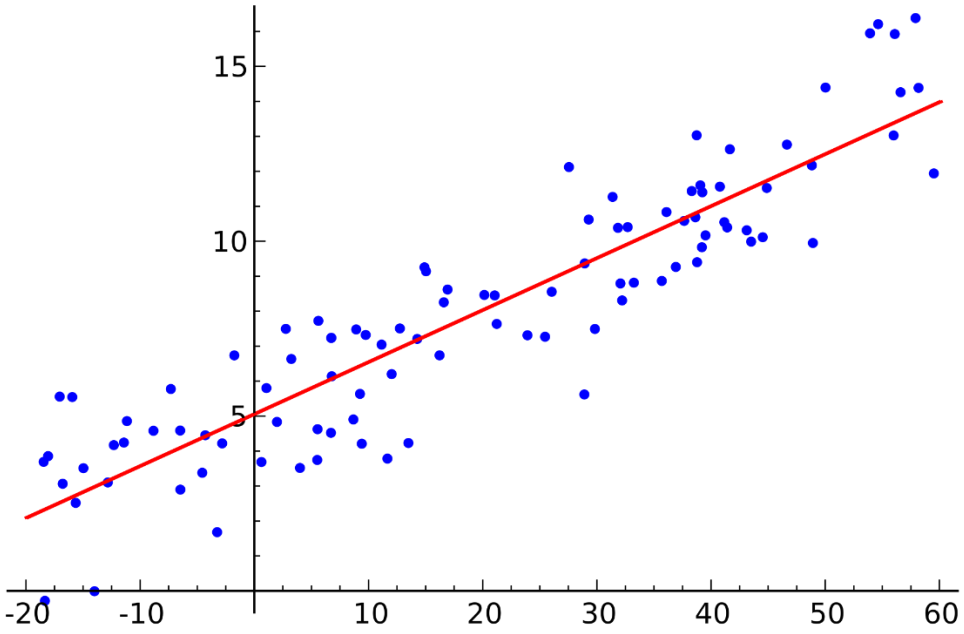
$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + ... + \beta_n x_{nj}, j = 1, 2, ..., m$$

με m τον αριθμό δεδομένων εκπαίδευσης. Ζητείται ο υπολογισμός των συντελεστών β_i.

Η μέθοδος των ελαχίστων τετραγώνων (least squares method) είναι μια από τις πλέον γνωστές και ευρέως χρησιμοποιούμενες μεθόδους επίλυσης. Με τη μέθοδο αυτή, ελαχιστοποιείται το σφάλμα που προκύπτει ανάμεσα στην εκτιμώμενη τιμή και στα πραγματικά δεδομένα.

Παράδειγμα απλής γραμμικής παλινδρόμησης

Με δεδομένα τα n σημεία (x, y) και ανεξάρτητη μεταβλητή τη x, στόχος της γραμμικής παλινδρόμησης είναι η εύρεση εξίσωσης ευθείας, της οποίας η γραφική παράσταση πλησιάζει «αρκετά κοντά» τα σημεία αυτά. Η κόκκινη ευθεία της εικόνας 3, αποτελεί την ευθεία της παλινδρόμησης της Y πάνω στη X, και αποτελεί τη βέλτιστη εξίσωση που μοντελοποιεί τα σημεία.

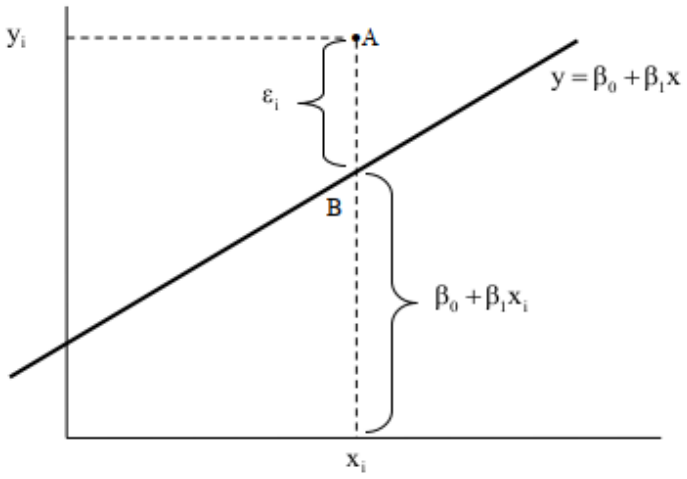


Εικόνα 3: Παράδειγμα απλής γραμμικής παλινδρόμησης

Πηγή: https://el.wikipedia.org/wiki/Απλή_γραμμική_παλινδρόμηση

Η εξίσωση της ευθείας δίνεται από τον τύπο $y = \beta_0 + \beta_1 x$. Πρέπει να υπολογιστούν οι τιμές των συντελεστών β_0 και β_1 ώστε η παραγόμενη ευθεία να προσεγγίζει τα σημεία, περιγράφοντας έτσι τη σχέση μεταξύ των X και Y.

Για τον υπολογισμό των συντελεστών β_0 και β_1 χρησιμοποιούνται μέθοδοι που ελαχιστοποιούν τις κατακόρυφες αποστάσεις $\varepsilon_i = y_i - \beta_0 - \beta_1 x$ των σημείων (x_i, y_i) από την ευθεία $y = \beta_0 + \beta_1 x$. Τα σφάλματα στην απλή γραμμική παλινδρόμηση αναπαριστώνται από τις κατακόρυφες αποστάσεις ε_i (εικόνα 4).



Εικόνα 4: Σφάλμα στην απλή γραμμική παλινδρόμηση

Το μοντέλο παλινδρόμησης συχνά βασίζεται σε μεγάλο βαθμό στην ικανοποίηση των υποκείμενων υποθέσεων. Η ανάλυση παλινδρόμησης έχει επικριθεί ως χρησιμοποιούμενη καταχρηστικά για το σκοπό αυτό, σε πολλές περιπτώσεις όπου δεν είναι δυνατή η επαλήθευση των κατάλληλων υποθέσεων. Ένας σημαντικός παράγοντας για την κριτική αυτή οφείλεται στο γεγονός ότι ένα μοντέλο παλινδρόμησης είναι πιο εύκολο να επικριθεί παρά να βρει μια μέθοδο για την προσαρμογή του μοντέλου παλινδρόμησης (Cook & Weisberg 1982). Ωστόσο, ο έλεγχος των υποθέσεων του μοντέλου δεν θα πρέπει ποτέ να παραβλέπεται στην ανάλυση παλινδρόμησης (Xin & Xiao 2009).

3.3 Μέθοδοι κατηγοριοποίησης

3.3.1 Δέντρα απόφασης / ταξινόμησης

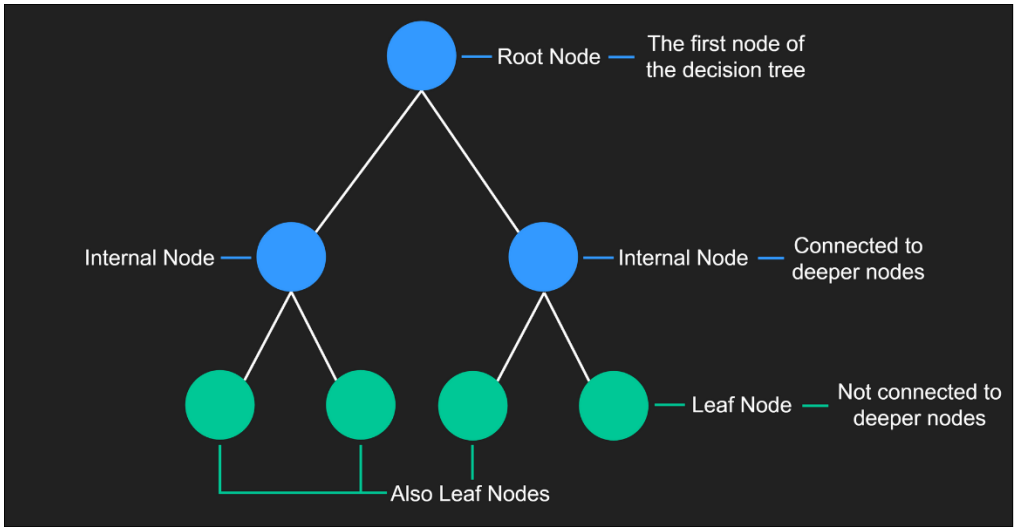
Τα δέντρα απόφασης ή ταξινόμησης (Decision trees ή Classification trees) αποτελούν μια μέθοδο για την εύρεση συναρτήσεων στόχου που εφαρμόζεται σε προβλήματα στα οποία η έξοδος έχει διακριτές τιμές (προβλήματα κατηγοριοποίησης). Για την εύρεση της συνάρτησης γίνεται χρήση της τεχνικής «διαίρει και βασίλευε»: το σύνολο των παρατηρήσεων των δεδομένων του προβλήματος διαιρείται σε υποσύνολα παρατηρήσεων, με κριτήριο την τιμή ως προς κάποιο χαρακτηριστικό. Η διαδικασία αυτή επαναλαμβάνεται αναδρομικά. Έτσι, το παραγόμενο μοντέλο αναπαρίσταται ως ανεστραμμένη δενδρική δομή.

Ως τεχνική είναι πολύ δημοφιλής λόγω της απλότητας αλλά και της διαφάνειάς της. Το γραφικό αποτέλεσμα που αναπαριστά μια ιεραρχική δομή, καθιστά ευκολότερη την ερμηνεία των αποτελεσμάτων σε σύγκριση με άλλες τεχνικές.

Η δενδροειδής δομή που εξάγεται, αποτελείται από κλαδιά και κόμβους και περιγράφει τα δεδομένα. Ισχύουν οι επόμενοι ορισμοί:

- Ρίζα (root) ονομάζεται ο πρώτος κόμβος, δηλαδή ο ανώτερος κόμβος του δέντρου και είναι ο μοναδικός κόμβος χωρίς εισερχόμενα κλαδιά.
- Εσωτερικός κόμβος (ή κόμβος εξέτασης) (internal node) ονομάζεται κάθε κόμβος ο οποίος έχει ακριβώς ένα εισερχόμενο κλαδί και έχει και εξερχόμενα κλαδιά.
- Φύλλο ή τερματικός κόμβος ή κόμβος απόφασης (leaf node) ονομάζεται κάθε κόμβος ο οποίος έχει ακριβώς ένα εισερχόμενο κλαδί και κανένα εξερχόμενο.

Ένα δέντρο ταξινόμησης φαίνεται στην εικόνα 5:



Εικόνα 5: Γενική αναπαράσταση δέντρου ταξινόμησης

Πηγή: https://mlfromscratch.com/decision-tree-classification/#/

Η ρίζα και κάθε εσωτερικός κόμβος περιέχουν συνθήκες ελέγχου της τιμής ενός χαρακτηριστικού των παραδειγμάτων. Κάθε κλαδί δηλώνει μια διαφορετική διακριτή τιμή του χαρακτηριστικού. Σε κάθε φύλλο εκχωρείται μια ετικέτα κατηγορίας και άρα τα φύλλα είναι η απόφαση κατηγοριοποίησης σε μια κατηγορία (κλάση).

Κατά συνέπεια, τα φύλλα αποτελούν κλάσεις (classes) και έτσι ένα δέντρο αποφάσεων είναι ένα σύνολο κανόνων ταξινόμησης (classification rules) (Κοτταρά 2019).

Με μαθηματικούς όρους, ένα δέντρο απόφασης είναι μια ιεραρχημένη συλλογή σύνθετων διαζευκτικών προτάσεων οι οποίες αποτελούνται από ένα σύνολο λογικών συζεύξεων που αναφέρονται σε τιμές χαρακτηριστικών συγκεκριμένων παραδειγμάτων (Φλώρου 2017).

Η κατασκευή των δέντρων γίνεται με διάφορους τρόπους. Η διαφορά τους έγκειται στους αλγορίθμους που χρησιμοποιούνται για την επιλογή του κριτηρίου διαχωρισμού του συνόλου των δεδομένων.

Για τη σωστή δημιουργία ενός αλγορίθμου κατασκευής δέντρων αποφάσεων, πρέπει τα δεδομένα πρέπει να έχουν κατηγοριοποιηθεί στις αντίστοιχες κλάσεις. Ο αλγόριθμος δημιουργεί τα δέντρα από το σύνολο δεδομένων και η ποιότητα των δέντρων είναι άμεσα συνυφασμένη με την ακρίβεια της κατηγοριοποίησης και με το μέγεθος του δένδρου (Φλώρου 2017).

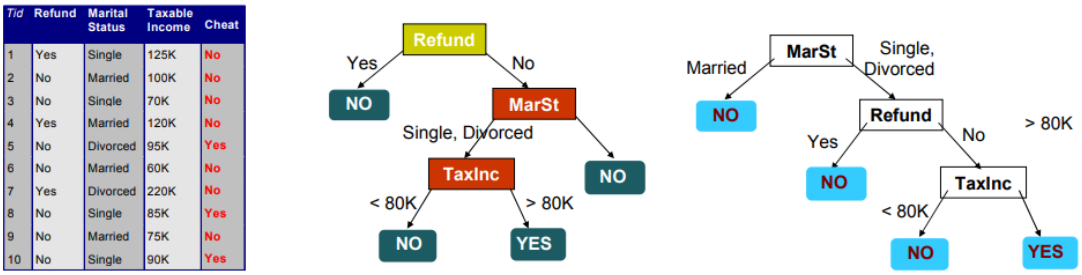
Η κατασκευή του δέντρου απόφασης γίνεται συνήθως σε δύο φάσεις:

Στην πρώτη φάση χτίζεται το αρχικό, μεγάλο, δέντρο με τον επαναλαμβανόμενο διαχωρισμό του συνόλου δεδομένων σύμφωνα με τις τιμές των ανεξάρτητων μεταβλητών. Μπορούν να χρησιμοποιηθούν διάφοροι αλγόριθμοι για τον καθορισμό της σειράς επιλογής της ανεξάρτητης μεταβλητής, αλλά κάθε φορά επιλέγεται εκείνη η ανεξάρτητη μεταβλητή που διαχωρίζει καλύτερα τα δεδομένα. Η πρώτη φάση του αλγορίθμου ολοκληρώνεται όταν

καταλήξει σε κόμβο από τον οποίο δεν μπορεί να αρχίσει νέος διαχωρισμός του συνόλου των δεδομένων. Αυτός ο κόμβος δεν έχει παιδιά, οπότε, συμπερασματικά, είναι ένα φύλλο του δέντρου.

Στη δεύτερη φάση χρησιμοποιούνται διάφορες τεχνικές που υλοποιούν το κλάδεμα του δέντρου. Με τον τρόπο αυτό καθορίζεται το τελικό μέγεθος του δέντρου. Ως κλάδεμα νοούνται οι τεχνικές που αφαιρούν εκείνα τα τμήματα του δέντρου που σχετίζονται με ένα μη σημαντικό χαρακτηριστικό. Στόχος των τεχνικών αποτελεί το τελικό δέντρο να μην είναι υπερβολικά εξειδικευμένο στην κατηγοριοποίηση νέων παραδειγμάτων, αλλά ταυτόχρονα να έχει τη μεγαλύτερη δυνατή ακρίβεια. Έτσι ελαχιστοποιείται και ο μικρότερος αριθμός παραγόμενων κανόνων. Είναι δυνατόν να παραχθούν διαφορετικά δέντρα απόφασης που να έχουν διαφορετική απόδοση στην κατηγοριοποίηση του ίδιου συνόλου εκπαίδευσης. Αυτό οφείλεται στις επιλογές των χαρακτηριστικών που επιλέγονται ως ρίζα και ως κόμβοι – γονείς.

Στην εικόνα 6 διακρίνονται δύο διαφορετικά δέντρα για το ίδιο σύνολο δεδομένων εκπαίδευσης.



Εικόνα 6: Διαφορετικά δέντρα προερχόμενα από το ίδιο σύνολο δεδομένων

Πηγή: Tan et al 2006

Βάσει των Breiman et al 1984, η πολυπλοκότητα ενός δέντρου επιδρά σημαντικά στην ακρίβειά του. Ο υπολογισμός της πολυπλοκότητας του δέντρου μπορεί να γίνει με χρήση διάφορων μετρικών όπως ο συνολικός αριθμός των κόμβων, ο συνολικός αριθμός των φύλλων, ο αριθμός των χαρακτηριστικών που χρησιμοποιούνται και το βάθος του δέντρου.

Η κατηγοριοποίηση ενός νέου παραδείγματος γίνεται με βάση τη διαδρομή από τη ρίζα του δέντρου μέχρι ένα φύλλο, σύμφωνα με τα αποτελέσματα των ενδιάμεσων της διαδρομής ελέγχων. Πιο αναλυτικά, αρχικά στον κόμβο της ρίζας, γίνεται ο πρώτος έλεγχος και έτσι καθορίζεται το κλαδί στο οποίο αντιστοιχεί η τιμή του επιλεγμένου χαρακτηριστικού. Ακολούθως, θεωρούμε τον κόμβο στον οποίο οδηγεί το επιλεχθέν κλαδί. Η διαδικασία αυτή επαναλαμβάνεται έως οδηγήσει σε κάποιο φύλλο του δέντρου. Όλα τα παραδείγματα που καταλήγουν στο ίδιο φύλλο, κατηγοριοποιούνται με τον ίδιο τρόπο, οπότε υπάρχει μοναδικό

μονοπάτι, με αφετηρία τη ρίζα και οδηγεί σε κάθε φύλλο. Το μονοπάτι αυτό εκφράζει τον κανόνα που χρησιμοποιήθηκε για την κατηγοριοποίηση του παραδείγματος.

Τα δέντρα απόφασης παρουσιάζουν αρκετά πλεονεκτήματα. Το βασικότερο δεν είναι παρά η ευκολία και η αποτελεσματικότητα χρήσης τους. Ακόμα, από ένα δέντρο απόφασης μπορούν να εξαχθούν κανόνες If – Then εύκολα κατανοητοί και ερμηνεύσιμοι. Για να γίνει αυτό αρκεί η συνένωση των ελέγχων που γίνονται κατά διαδρομή από τη ρίζα σε κάθε φύλλο, λαμβάνοντας ως τιμή για την κατηγοριοποίηση την πρόβλεψη της κατηγοριοποίησης σε κάθε φύλλο. Σύμφωνα με τον Quinlan 1987, το επαγόμενο σύνολο κανόνων μπορεί ακολούθως να απλουστευθεί, ώστε να βελτιωθεί η δυνατότητα κατανόησής του από ένα χρήστη και ίσως και η ακρίβειά του. Ένα ακόμα πλεονέκτημα των δέντρων απόφασης είναι η επιτυχής εφαρμογή τους σε μεγάλες βάσεις δεδομένων, καθώς το μέγεθος της βάσης δεδομένων είναι ανεξάρτητο από το μέγεθος του παραγόμενου δέντρου.

Ως μειονεκτήματα των δέντρων αποφάσεων πρέπει να αναφερθούν η αδυναμία χειρισμού των συνεχών δεδομένων καθώς οι τιμές των χαρακτηριστικών πρέπει να χωριστούν σε διαστήματα. Η δυσκολία χειρισμού των ελλιπών δεδομένων αποτελεί επίσης πρόβλημα καθώς δεν είναι δυνατόν να ανακαλυφθούν οι σωστές διακλαδώσεις που πρέπει να ακολουθηθούν. Ακόμα, στα δέντρα απόφασης δε λαμβάνονται υπόψη τυχόν υπάρχουσες συσχετίσεις ανάμεσα στα χαρακτηριστικά.

3.3.1.1 Ο αλγόριθμος ID3

Ο ID3 (Iterative Dichotomiser 3) αποτελεί έναν από τις πιο απλούς και διαδεδομένους αλγορίθμους για την κατασκευή δένδρων απόφασης. Προτάθηκε το 1986 από τον Quinlan. Είναι ένας αναδρομικός αλγόριθμος που χρησιμοποιεί στρατηγική από πάνω-προς-τα-κάτω (top-down) για να κατασκευάσει το δέντρο απόφασης, στο οποίο αναπαρίστανται οι συσχετίσεις στα δεδομένα εκπαίδευσης. Στόχος της στρατηγικής του είναι στο παραχθέν δέντρο να απαιτείται ο ελάχιστος αριθμός συγκρίσεων κατά την κατηγοριοποίηση ενός νέου αντικειμένου. Ο ID3 για να εφαρμοστεί, πρέπει οι τιμές των μεταβλητών να είναι διακριτές, οπότε τυχόν συνεχείς αριθμητικές τιμές πρέπει να μετατραπούν και να ενταχθούν στις κατηγορίες (κλάσεις) που ορίζονται.

Με τον τρόπο αυτό όμως εισέρχεται η έννοια της υποκειμενικότητας την επιλογής των κατηγοριών καθώς αυτή μπορεί να γίνει με πολλούς τρόπους. Έτσι η επιτυχία στην τελική δημιουργία του δενδροειδούς μοντέλου εξαρτάται από το αρχικό μέρος της διαδικασίας, την προετοιμασία των δεδομένων, παρά από την επιτυχή εκτέλεση του αλγορίθμου και της ερμηνείας των αποτελεσμάτων.

Ο αλγόριθμος μπορεί γενικά να περιγραφεί ως εξής:

- Αρχικά επιλέγεται το πιο κατάλληλο χαρακτηριστικό για έλεγχο στη ρίζα.

- Στη συνέχεια, για κάθε μία από τις δυνατές τιμές αυτού του χαρακτηριστικού δημιουργούνται οι αντίστοιχοι απόγονοι (παιδιά) της ρίζας. Τα δεδομένα μοιράζονται στους νέους κόμβους (απόγονοι) με βάση με την τιμή τους, για το χαρακτηριστικό το οποίο ορίστηκε (και άρα λειτουργεί ως έλεγχος) ως ρίζα του δέντρου.
- Η παραπάνω διαδικασία εκτελείται επαναληπτικά για κάθε νέο κόμβο. Τα χαρακτηριστικά επιλέγονται βάσει των δεδομένων που ανήκουν στον κάθε κόμβο.
- Ένας κόμβος μετατρέπεται σε φύλλο όταν τα δεδομένα που ανήκουν στον κόμβο αυτό ανήκουν όλα στην ίδια κατηγορία. Η κατηγορία αυτή αποτελεί την τιμή του φύλλου.
- Αν σε κάποιο βάθος τελειώσουν τα χαρακτηριστικά προς έλεγχο, τότε ο κόμβος γίνεται τερματικός. Η τιμή του κόμβου θα είναι εκείνη που πλειοψηφεί με βάση τα δεδομένα του κόμβου αυτού.

Ο διαχωρισμός των παρατηρήσεων στον ID3 γίνεται με κριτήριο το Κέρδος Πληροφορίας (Information Gain). Για κάθε χαρακτηριστικό υπολογίζεται το Κέρδος Πληροφορίας και ο αλγόριθμος επιλέγει εκείνο με τη μεγαλύτερη τιμή. Σύμφωνα με τον Κύρκο 2015 έτσι δημιουργούνται υποσύνολα μεγαλύτερης ομοιογένειας.

Το Κέρδος Πληροφορίας G(S,A) εκφράζει τη μείωση της εντροπίας που θα προκύψει, αν το σύνολο παρατηρήσεων S διαχωριστεί σε υποσύνολα με βάση τις τιμές του χαρακτηριστικού A (Κύρκος 2015). Η εντροπία είναι ένα μέτρο της ανομοιογένειας του συνόλου S. Η εντροπία παίρνει τιμές στο διάστημα [0, 1] και αν το μέτρο της εντροπίας ισούται με μηδέν το σύνολο είναι τέλεια ταξινομημένο. Αντίθετα, αν το μέτρο της εντροπίας ισούται με 1, η σύνθεση του συνόλου είναι εντελώς τυχαία.

Αν το σύνολο S έχει s παρατηρήσεις και αν το χαρακτηριστικό της κλάσης μπορεί να πάρει c διαφορετικές τιμές και το πλήθος των παρατηρήσεων με τιμή κλάσης i είναι s_i, τότε η εντροπία του S δίνεται από την εξίσωση:

$$E(S) = \sum_{i=1}^{s} (-p_i \log_2(p_i))$$

με p_i το ποσοστό των παρατηρήσεων που ανήκουν στην κλάση i $\left(p_i = \frac{s_i}{s} \right)$.

Το Κέρδος Πληροφορίας αναπαριστά τη μείωση της εντροπίας του συνόλου εκπαίδευσης S αν ο διαχωρισμός γίνει με βάση τη μεταβλητή A. Δίνεται από τον τύπο:

$$G(S, A) = E\left(S\right) - \sum_{i=1}^{s} \frac{S_i}{s} \cdot E\left(S_i\right)$$

Όπου E(S) είναι η εντροπία του συνόλου S και ο δεύτερος όρος της διαφοράς είναι η εντροπία του διαχωρισμού του S ανάλογα με τις τιμές του χαρακτηριστικού A.

Ο διαχωρισμός των δεδομένων γίνεται με βάση τις τιμές του χαρακτηριστικού που έχει το μεγαλύτερο Κέρδος Πληροφορίας. Η γενική ιδέα του αλγορίθμου ID3, προκειμένου να επιτευχθεί ο διαχωρισμός ενός μη ομοιογενούς συνόλου δειγμάτων σε έναν κόμβο του δέντρου απόφασης που ανήκει σε συγκεκριμένο κλαδί, σύμφωνα με τη Γεωργούλη 2015,

- Για όλα τα χαρακτηριστικά που δεν έχουν χρησιμοποιηθεί στο συγκεκριμένο κλαδί, υπολογίζεται η εντροπία σε σχέση με τα δείγματα.
- Επιλέγεται το χαρακτηριστικό που έχει το μέγιστο κέρδος πληροφορίας.
- Κατασκευάζεται ο κόμβος του χαρακτηριστικού αυτού.

Στον αλγόριθμο ID3 η συνάρτηση στόχος έχει συνήθως δύο διακριτές τιμές. Ο χειρισμός των μεταβλητών που παίρνουν πραγματικές τιμές και έχουν έξοδο με περισσότερες από δύο τιμές, μπορεί να υλοποιηθεί αλλά απαιτεί επέκταση του βασικού αλγορίθμου. Ο Quinlan 1986, προτείνει επίσης και ένα επαναληπτικό πλαίσιο για την επιτάχυνση της διαδικασίας παραγωγής του δέντρου όταν το σύνολο των δεδομένων εκπαίδευσης είναι πολύ μεγάλο. Στα πλεονεκτήματα του ID3 πρέπει να καταγραφεί η λειτουργία του αλγορίθμου παρά την πιθανή απουσία τιμών ορισμένων χαρακτηριστικών.

3.3.1.2 Οι αλγόριθμοι C4.5 και C5.0

Ο αλγόριθμος C4.5 είναι μια δημοφιλής τεχνική δημιουργίας δέντρων κατηγοριοποίησης. Έχει προταθεί και αυτός από τον Quinlan το 1993 και αποτελεί την πλέον διαδεδομένη βελτίωση του ID3. Ο πηγαίος κώδικας είναι γραμμένος σε γλώσσα C και αριθμεί περί τις 9.000 γραμμές κώδικα – σημαντική αναβάθμιση από τις περίπου 600 γραμμές κώδικα γλώσσας Pascal του αλγορίθμου ID3. Οι βελτιώσεις αφορούν στα παρακάτω σημεία:

- Διάσπαση: Όπως σημειώνει ο Quinlan το Κέρδος Πληροφορίας τείνει να ευνοεί τα γνωρίσματα στα οποία το πλήθος των τιμών είναι μεγάλο, με αποτέλεσμα τη δημιουργία μεγάλου αριθμού μικρών και πολύ ομοιογενών συνόλων. Με την τροποποίηση του κριτηρίου διαχωρισμού και τη χρήση του επονομαζόμενου κριτηρίου «Λόγος Κέρδους» (Gain Ratio) αντιμετωπίζεται η παραπάνω δυσκολία. Ο Λόγος Κέρδους δίνεται από τον τύπο:

$$\text{GainRatio}(S, A) = \frac{G(S, A)}{E(S, A)}$$

  Με το Λόγο Κέρδους επιτυγχάνεται η κανονικοποίηση του Κέρδους Πληροφορίας ως προς την Εντροπία και με τη χρήση του βελτιώνεται η ακρίβεια και μειώνεται η πολυπλοκότητα των δέντρων (Κύρκος 2015).
- Ελλιπή δεδομένα: Αν μια τιμή λείπει από ένα χαρακτηριστικό γίνεται πρόβλεψη της τιμής αυτής, με βάση τις υπόλοιπες τιμές του χαρακτηριστικού. Με αυτόν τον τρόπο

το συγκεκριμένο χαρακτηριστικό λαμβάνεται υπόψη στη δημιουργία του δέντρου, ειδάλλως θα αγνοούνταν καθώς θα ήταν ελλιπές.

- Αριθμητικά δεδομένα: Είναι δυνατή η χρήση χαρακτηριστικών που περιλαμβάνουν αριθμητικές τιμές. Για να επιτευχθεί αυτό, ταξινομούνται οι τιμές κάθε αριθμητικού πεδίου σε αύξουσα σειρά και καθορίζεται μια τιμή που λειτουργεί ως σύνορο: οι παρατηρήσεις χωρίζονται σε αυτές που οι τιμές τους είναι μικρότερες ή ίσες του συνόρου και σε αυτές που οι τιμές τους είναι μεγαλύτερες του συνόρου. Έτσι προκύπτουν δύο διακριτές τιμές, που λαμβάνουν τις τιμές βάσει των καθορισμένων περιοχών συνεχών τιμών.
- Κλάδεμα: Εφαρμόζονται δύο στρατηγικές κλαδέματος στον C4.5: η αντικατάσταση του υποδένδρου και η ανύψωση του υποδένδρου. Στην αντικατάσταση του υποδένδρου ο αλγόριθμος ξεκινάει από τα φύλλα και προχωρά προς τη ρίζα. Σε κάθε κόμβο του δέντρου ελέγχεται αν το υποδένδρο που ξεκινά αυτόν θα παραμείνει ή θα αντικατασταθεί από φύλλο. Η αντικατάσταση πραγματοποιείται όταν το σφάλμα που προκύπτει είναι κοντά στο σφάλμα με το αρχικό υποδένδρο. Στην ανύψωση του υποδένδρου σε κάθε κόμβο του δένδρου γίνεται έλεγχος αν το υποδένδρο με ρίζα αυτόν τον κόμβο θα παραμείνει ως έχει ή αν θα αντικατασταθεί από το συχνότερα χρησιμοποιούμενο υποδένδρο του, με βάση υπολογισμούς αύξησης της συχνότητας λαθών.

Γενικά, τα βήματα του αλγορίθμου μπορούν να περιγραφούν ως εξής:

- Επιλογή του χαρακτηριστικού με το οποίο επιτυγχάνεται ο καλύτερος διαχωρισμός με χρήση του μέτρου Λόγου Κέρδους.
- Διαχωρισμός των δεδομένων σε υποσύνολα βάσει των τιμών του χαρακτηριστικού αυτού.
- Επανάληψη της διαδικασίας για κάθε υποσύνολο που περιέχει περισσότερες από μία κατηγορίες.
- Τερματισμός του αλγορίθμου όταν υπάρχουν υποσύνολα που περιέχουν περισσότερες από μία κατηγορίες ή όταν έχουν χρησιμοποιηθεί όλα τα χαρακτηριστικά.

Ο αλγόριθμος C5.0 μπορεί να χαρακτηριστεί ως εμπορική έκδοση του C4.5, υπό την έννοια ότι συναντάται και χρησιμοποιείται ευρύτατα στα πακέτα λογισμικού εξόρυξης γνώσης, καθώς προσανατολίζεται στη χρήση μεγάλων συνόλων δεδομένων. Η σχετική βιβλιογραφία είναι περιορισμένη και συμπεράσματα προκύπτουν κυρίως από την αξιολόγηση του πηγαίου κώδικα του αλγορίθμου που δημοσιεύτηκε το 2011. Κατά την εκτέλεση του η χρήση της μνήμης του συστήματος είναι βελτιωμένη σε ποσοστό της τάξης του 90%, η ταχύτητά του είναι αυξημένη και η ακρίβεια των παραγόμενων κανόνων είναι μεγαλύτερη. Ο αλγόριθμος υλοποιεί μια εργασία τελικού κλαδέματος, κατά την οποία απομακρύνονται τα υποδένδρα μέχρις ότου η

τιμή σφάλματος ξεπεράσει τη βασική τιμή του τυπικού σφάλματος, με αποτέλεσμα τη δημιουργία απλούστερων δένδρων.

3.3.2 Κατηγοριοποίηση με βάση τους κοντινότερους γείτονες

Οι Κατηγοριοποιητές Βασισμένοι σε Παραδείγματα (Instance Based Classifiers) είναι μια οικογένεια κατηγοριοποιητών στους οποίους η μάθηση βασίζεται στην αναλογία (Κύρκος 2015). Δεν υπάρχει στάδιο εκπαίδευσης, ούτε δημιουργείται ένα μοντέλο, μέχρι ένα ζητηθεί η κατηγοριοποίηση ενός νέου παραδείγματος. Από την ιδιότητα αυτή, οι κατηγοριοποιητές ανήκουν στην κατηγορία αλγορίθμων «οκνηρής μάθησης» και καλούνται «οκνηροί» (lazy classifiers). Στη φάση της μάθησης αποθηκεύουν τα δεδομένα του συνόλου εκπαίδευσης στη μνήμη, ενώ η κατηγοριοποίηση του νέου παραδείγματος γίνεται με τη σύγκρισή του με τα γνωστά παραδείγματα του συνόλου δεδομένων που είναι αποθηκευμένα στη μνήμη.

Μια από τις δημοφιλέστερες τεχνικές κατηγοριοποίησης της κατηγορίας αυτής, είναι η κατηγοριοποίηση των k-κοντινότερων γειτόνων (k-Nearest Neighbors – k-NN). Η κατηγοριοποίηση του νέου παραδείγματος γίνεται με βάση τις αντίστοιχες κατηγοριοποιήσεις των k κοντινότερων παραδειγμάτων, τα οποία ονομάζονται «γείτονες» του.

Κάθε νέο παράδειγμα τοποθετείται στο χώρο σαν καινούριο σημείο και η κατηγορία του προκύπτει από τις κατηγορίες των k κοντινότερων γειτονικών του σημείων. Οι κοντινότεροι γείτονες ενός παραδείγματος υπολογίζονται χρησιμοποιώντας κάποια από τις γνωστές μετρικές. Οι συνηθέστερες είναι η απόσταση της Ευκλείδειας Γεωμετρίας, που χρησιμοποιείται όταν τα χαρακτηριστικά λαμβάνουν αριθμητικές τιμές, και η απόσταση Manhattan που προτιμάται όταν τα δεδομένα λαμβάνουν ποιοτικές τιμές.

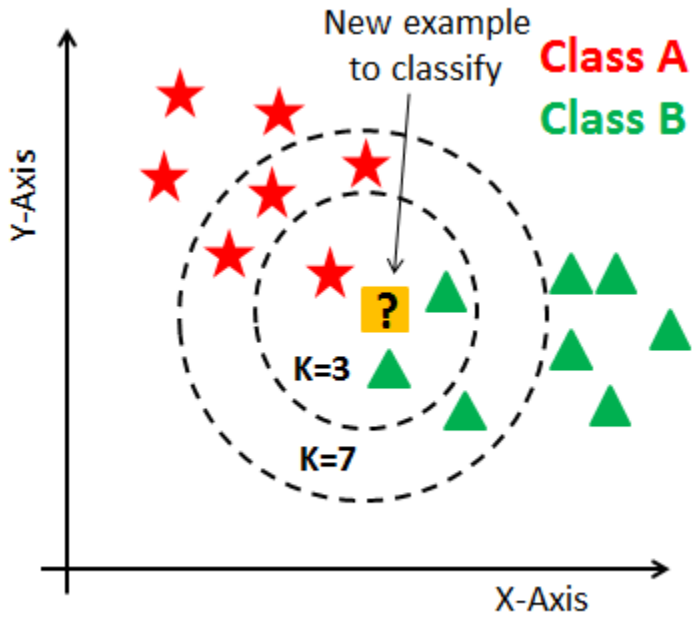
Ο τύπος της Ευκλείδειας απόστασης μεταξύ δύο παραδειγμάτων X, Y διαστάσεων n, δίνεται από τον τύπο:

$$d(X,Y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

Η απόσταση Manhattan δίνεται από τον τύπο:

$$d(X,Y)=\sum_{i=1}^{n}(x_i-y_i),\ \text{με}\ \ x_i-y_i=\begin{cases}0,\ \text{αν}\ x_i=y_i\\1,\ \text{αν}\ x_i\neq y_i\end{cases}$$

Στον αλγόριθμο k-NN η τιμή της παραμέτρου k ορίζεται από το χρήστη. Στο σχήμα της εικόνας 7, η κατηγοριοποίηση του νέου άγνωστου παραδείγματος διαφέρει ανάλογα με την επιλεχθείσα τιμή του k. Αν k=3, τότε το νέο παράδειγμα κατηγοριοποιείται ως τρίγωνο, ενώ αν k=7 τότε κατηγοριοποιείται ως αστέρι.



Εικόνα 7: Προσδιορισμός κατηγορίας με βάση τους 3 και 7 κοντινότερους γείτονες

Πηγή: https://www.kdnuggets.com/2020/11/most-popular-distance-metrics-knn.html

Υπάρχουν παραλλαγές του αλγορίθμου k-NN. Μια από αυτές ορίζει να μη λαμβάνεται η απόφαση για την κατηγοριοποίηση με ισοτιμία μεταξύ των γειτόνων, αλλά η απόφαση κατηγοριοποίησης να επηρεάζεται περισσότερο από τα σημεία που βρίσκονται πλησιέστερα στο προς κατηγοριοποίηση παράδειγμα.

Η μέθοδος των k-κοντινότερων γειτόνων χαρακτηρίζεται ως πολύ αποτελεσματική μέθοδος. Στα πλεονεκτήματά της συγκαταλέγεται το γεγονός πως μπορεί να χρησιμοποιηθεί για αριθμητικά και για ποιοτικά δεδομένα, ακόμα και στην περίπτωση που ο αριθμός δεδομένων του συνόλου εκπαίδευσης είναι μικρός.

Από την άλλη μεριά, έχει υψηλό υπολογιστικό κόστος κατά την κατηγοριοποίηση νέων παραδειγμάτων, εξαιτίας του αριθμού των συγκρίσεων που απαιτούνται για το νέο παράδειγμα, καθώς αφορούν κάθε στοιχείο του συνόλου εκπαίδευσης δεδομένων. Ακόμα, δυσκολία υπάρχει και στο ζήτημα της επιλογής της βέλτιστης τιμής της παραμέτρου k, ώστε να ληφθεί το αποτέλεσμα με την καλύτερη δυνατή ακρίβεια.

3.3.3 Κατηγοριοποιητής Κανόνων

Ένας κατηγοριοποιητής κανόνων (Rule-Based Classifier) είναι μια τεχνική για την κατηγοριοποίηση εγγραφών χρησιμοποιώντας μια συλλογή από κανόνες If – then. Οι κανόνες If – then είναι της μορφής if (συνθήκη) then (κανόνας). Το αριστερό μέρος του κανόνα ονομάζεται προηγούμενος κανόνας (rule antecedent) ή συνθήκη εισόδου (precondition). Αν η συνθήκη είναι Αληθής τότε ικανοποιείται η συνθήκη και ενεργοποιείται ο κανόνας. Το δεξί μέρος του κανόνα ονομάζεται επακόλουθος κανόνας (rule consequent) και περιέχει την προβλεπόμενη κατηγορία.

Για τη δημιουργία ενός κατηγοριοποιητή κανόνων πρέπει να εντοπιστεί ένα σύνολο από κανόνες οι οποίοι προσδιορίζουν τις σχέσεις μεταξύ των χαρακτηριστικών ενός συνόλου δεδομένων και κάθε κατηγορίας. Υπάρχουν δύο γενικές κατηγορίες μεθόδων για την δημιουργία κανόνων κατηγοριοποίησης: οι άμεσες μέθοδοι που εξάγουν τους κανόνες κατηγοριοποίησης κατευθείαν από τα δεδομένα και οι έμμεσες μέθοδοι που εξάγουν τους κανόνες κατηγοριοποίησης με χρήση άλλων μεθόδων κατηγοριοποίησης, όπως π.χ. τα δέντρα απόφασης και τα νευρωνικά δίκτυα.

Οι άμεσες μέθοδοι διαχωρίζουν το σύνολο των παρατηρήσεων σε μικρότερα σύνολα ώστε κάθε παράδειγμα που ανήκει σε ένα υποσύνολο να μπορεί να κατηγοριοποιηθεί με χρήση ενός απλού κανόνα κατηγοριοποίησης. Οι μέθοδοι αυτοί ονομάζονται «τεχνικές κάλυψης» καθώς στόχος τους είναι η δημιουργία κανόνων που να καλύπτει μια κατηγορία. Σύμφωνα με τους Tan et al 2006, ο αλγόριθμος σειριακής κάλυψης (sequential covering) χρησιμοποιείται συχνά για να εξαχθούν κανόνες άμεσα από τα δεδομένα. Ο αλγόριθμος εξάγει τους κανόνες χρησιμοποιώντας μια κατηγορία κάθε φορά για σύνολα δεδομένων που περιέχουν περισσότερες από δύο κατηγορίες. Το κριτήριο βάσει του οποίου αποφασίζεται ποια κατηγορία θα εξαχθεί πρώτη εξαρτάται από διάφορους παράγοντες (π.χ. το ποσοστό των παραδειγμάτων που ανήκουν σε κάθε κατηγορία).

Οι έμμεσες μέθοδοι χρησιμοποιούν τους κανόνες κατηγοριοποίησης για να παρέχουν μια περιληπτική περιγραφή πιο πολύπλοκων μοντέλων κατηγοριοποίησης. Στην περίπτωση που οι κανόνες εξάγονται από ένα δέντρο απόφασης, δεν υπάρχει πάντα ισοδυναμία ανάμεσα στους κανόνες και το δέντρο. Βασική διαφορά είναι ότι στο δέντρο απόφασης υπάρχει σειρά με την οποία γίνεται ο διαχωρισμός, σε αντίθεση με τους κανόνες κατηγοριοποίησης.

Για την παραγωγή κανόνων από ένα δέντρο απόφασης, κάθε διαδρομή από τη ρίζα του δέντρου προς το φύλλο, μπορεί να εκφραστεί ως ένας κανόνας κατηγοριοποίησης. Οι συνθήκες ελέγχου που υπάρχουν στη διαδρομή αυτή, δημιουργούν συζεύξεις καθώς συνδέονται μεταξύ τους με χρήση του λογικού τελεστή AND και δημιουργούν το αριστερό μέρος του κανόνα. Το δεξί μέρος του κανόνα, η κατηγορία, προκύπτει από το φύλλο του δέντρου. Το πλήθος των παραγόμενων κανόνων είναι ίσο με το πλήθος των φύλλων του δέντρου απόφασης.

Αν οι παραγόμενοι κανόνες είναι πολύπλοκοι, τότε, μπορεί να εφαρμοστεί μια τεχνική κλαδέματος για την αφαίρεση των μη σημαντικών συνθηκών, οι οποίες καθορίζονται συγκρίνοντας το ρυθμό σφάλματος του νέου, πιο απλού, κανόνα με το ρυθμό σφάλματος του αρχικού, πλήρη, κανόνα.

Για την παραγωγή κανόνων από ένα νευρωνικό δίκτυο χρησιμοποιείται ο γράφος του δικτύου στον οποίο πρακτικά εμπεριέχονται οι κανόνες αυτοί. Ο μεγάλος αριθμός εισόδων του νευρωνικού δικτύου είναι ένα πρόβλημα για την εξαγωγή των κανόνων, καθώς αν ένας κόμβος του νευρωνικού δικτύου έχει n εισόδους με την κάθε είσοδο να λαμβάνει τρεις διαφορετικές τιμές, τότε προκύπτουν μόνο για το συγκεκριμένο κόμβο, 3n συνδυασμοί εισόδων. Για την αντιμετώπιση του προβλήματος αυτού, οι τιμές εξόδου των κόμβων των κρυμμένων στρωμάτων καθώς και των κόμβων των στρωμάτων εξόδου του νευρωνικού δικτύου, διακριτοποιούνται με χρήση κατάλληλων αλγορίθμων. Ακολουθεί το κλάδεμα του δέντρου αφαιρώντας τους περιττούς κόμβους εισόδου και τις συνδέσεις τους με τους κόμβους του κρυφού επιπέδου και τις συνδέσεις τους με τους κόμβους εξόδου. Έτσι, απομένουν μόνο οι σημαντικοί κόμβοι και οι συνδέσεις και εξάγονται οι κανόνες απεικόνισης των σχέσεων εισόδου και εξόδου, μεταξύ των κόμβων του νευρωνικού δικτύου.

Το πλεονέκτημα των κατηγοριοποιητών κανόνων είναι πως είναι εύκολα αντιληπτοί από τον άνθρωπο. Αν ένας κατηγοριοποιητής κανόνων επιτρέπει την ενεργοποίηση πολλών κανόνων για ένα παράδειγμα, τότε είναι πιθανό να δημιουργηθεί ένα πολύπλοκο όριο απόφασης. Η απόδοσή τους είναι συγκρίσιμη με τους κατηγοριοποιητές δέντρων απόφασης.

3.3.4 Κατηγοριοποιητές Bayes

Οι κατηγοριοποιητές Bayes βασίζονται στη Στατιστική και χρησιμοποιούνται σε προβλήματα όπου ζητείται η πρόβλεψη πιθανοτήτων εμφάνισης κατηγοριών και η εκτίμηση της πιθανότητας μια παρατήρηση να ανήκει σε μια εκ των προκαθορισμένων κατηγοριών. Θεμέλιος λίθος είναι το αντίστοιχο θεώρημα του Bayes το οποίο υπολογίζει την υπό συνθήκη πιθανότητα $P(H \mid X)$, δηλαδή την πιθανότητα επαλήθευσης της υπόθεσης H δεδομένου ότι ισχύει το γεγονός X, και δίνεται από τον τύπο:

$$P(H \mid X) = \frac{P(H) \cdot P(X \mid H)}{P(X)}$$

όπου:

- $P(H \mid X)$ είναι η δεσμευμένη πιθανότητα να ισχύει η υπόθεση H έχοντας ως δεδομένη την υπόθεση X. Είναι γνωστή ως «εκ των υστέρων πιθανότητα του H».

- $P(H)$ είναι η πιθανότητα πραγματοποίησης της υπόθεσης Η, δηλαδή να ισχύει η υπόθεση Η, και είναι γνωστή ως «εκ των προτέρων πιθανότητα του Η».
- $P(X \mid H)$ είναι η δεσμευμένη πιθανότητα να συμβεί το γεγονός Χ δεδομένης της υπόθεσης Η, η οποία μπορεί να υπολογιστεί από τις πληροφορίες του υπό εξέταση προβλήματος.
- $P(X)$ είναι η πιθανότητα να πραγματοποιηθεί το γεγονός Χ, το οποίο αντιστοιχεί στην ανεξάρτητη μεταβλητή.

Η απόδοση της κατηγοριοποίησης κατά Bayes είναι αρκετά υψηλή και επιπλέον χαρακτηρίζεται από μεγάλη ταχύτητα σε μεγάλες Βάσεις Δεδομένων. Στα μειονεκτήματα καταγράφεται το υψηλό υπολογιστικό κόστος όλων των πιθανοτήτων των κλάσεων.

Ιδιαίτερα χρήσιμη και πρακτική μέθοδος μάθησης είναι ο «Αφελής/Απλός ταξινομητής Bayes (simple/naive Bayes classifier)» με τη βοήθεια του οποίου δημιουργούνται πιθανοτικά μοντέλα πρόβλεψης που αφορούν είτε ποιοτικά είτε ποσοτικά χαρακτηριστικά.

3.3.4.1 Ο αφελής κατηγοριοποιητής Bayes

Ο «αφελής κατηγοριοποιητής Bayes» (Naïve Bayes Classifier) είναι μία μέθοδος κατηγοριοποίησης που βασίζεται στον κανόνα του Bayes. Είναι μια απλή μέθοδος στην οποία γίνεται η υπόθεση ότι η επίδραση ενός χαρακτηριστικού σε μια δεδομένη κλάση είναι ανεξάρτητη από τις τιμές των υπολοίπων χαρακτηριστικών (Ζουμπουλίδης 2012). Με την υπόθεση αυτή, που είναι γνωστή ως «υπό συνθήκη ανεξαρτησία» (conditional independence) και η οποία όμως δεν ισχύει πάντα, οι υπολογισμοί απλουστεύονται.

Έστω D το σύνολο των δεδομένων εκπαίδευσης, Χ ένα παράδειγμα του συνόλου D και $C_1, C_2, ..., C_m$ οι m το πλήθος κλάσεις του προβλήματος. Ο «αφελής κατηγοριοποιητής Bayes» θα κατηγοριοποιήσει το παράδειγμα Χ στην κλάση C_i, αν η εκ των υστέρων πιθανότητα της κλάσης i είναι μεγαλύτερη από τις εκ των υστέρων πιθανότητες των λοιπών κλάσεων. Δηλαδή, αν:

$$P(C_i \mid X) > P(C_j \mid X), \text{ για κάθε } j \text{ με } 1 \le j \le m, i \ne j.$$

Για τον υπολογισμό των πιθανοτήτων ισχύουν τα ακόλουθα:

Καταρχάς, έστω Χ (διάνυσμα $(x_1, x_2, ..., x_n)$) που αντιστοιχεί σε μια παρατήρηση από το σύνολο δεδομένων και έστω Η η υπόθεση ότι η παρατήρηση αυτή ανήκει στην κλάση C_i, θεωρώντας ότι υπάρχουν m κλάσεις. Χρησιμοποιώντας το θεώρημα του Bayes, υπολογίζεται η πιθανότητα η παρατήρηση Χ να ανήκει στην κλάση C_i από τον τύπο:

$$P(C_i \mid X) = \frac{P(C_i) \cdot P(X \mid C_i)}{P(X)}$$

Προκειμένου μια παρατήρηση να ενταχθεί σε μια κλάση, υπολογίζονται οι πιθανότητες για κάθε κλάση, και η παρατήρηση κατηγοριοποιείται στην κλάση που η τιμή της πιθανότητας είναι μεγαλύτερη.

Καθώς η τιμή της πιθανότητας P(X) είναι σταθερή για όλες τις κλάσεις και η $P(C_i)$ υπολογίζεται εύκολα από τη συχνότητα εμφάνισης κάθε κατηγορίας στα δεδομένα εκπαίδευσης (από το λόγο πλήθους παρατηρήσεων που ανήκουν στην κλάση C_i, προς το συνολικό πλήθος παρατηρήσεων), απομένει ο υπολογισμός του $P(X \mid C_i)$. Με την παραδοχή της ανεξαρτησίας μεταξύ των μεταβλητών εισόδου, ο υπολογισμός του $P(X \mid C_i)$ γίνεται από τον τύπο:

$$P(X \mid C_i) = \prod_{\kappa=1}^{n} P(x_\kappa \mid C_i)$$

με x_κ την αντίστοιχη τιμή της διάστασης κ στο διάνυσμα X.

Στα πλεονεκτήματα του αφελούς κατηγοριοποιητή Bayes συγκαταλέγονται η ευκολία χρήσης του και η δυνατότητα χειρισμού ελλιπών δεδομένων με την παράλειψη των αντίστοιχων πιθανοτήτων.

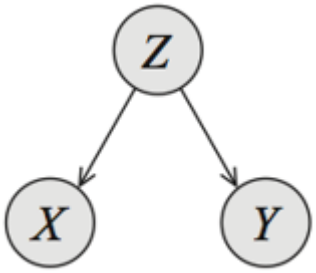
Θεωρητικά, ο αφελής κατηγοριοποιητής Bayes έχει το ελάχιστο ποσοστό σφάλματος εν συγκρίσει με τους άλλους κατηγοριοποιητές. Στην πράξη, ωστόσο, αυτό δεν επιβεβαιώνεται πάντα εξαιτίας ανακριβειών στις υποθέσεις των δεδομένων και την υποθετική ανεξαρτησία των κλάσεων. Η ανεξαρτησία των χαρακτηριστικών είναι πιο σπάνια ως περίπτωση και πολλές φορές αντιμετωπίζεται με το να αγνοούνται τα χαρακτηριστικά τα οποία εξαρτώνται από άλλα. Στα μειονεκτήματα πρέπει να αναφερθεί ακόμα η έλλειψη χειρισμού συνεχών δεδομένων, αν και η περίπτωση αυτή μπορεί να λυθεί με το χωρισμό των τιμών των συνεχών χαρακτηριστικών σε διαστήματα – όμως ο τρόπος χωρισμού επηρεάζει τα αποτελέσματα.

3.3.4.2 Δίκτυα Bayes

Τα δίκτυα Bayes (Bayes Net) είναι κατηγοριοποιητές βάσει πιθανοτήτων και αποτελούν επέκταση του «Αφελούς Κατηγοριοποιητή Bayes» επιτρέποντας την ανεξαρτησία υποσυνόλων των μεταβλητών εισόδου. Οι εξαρτήσεις των μεταβλητών αναπαρίστανται γραφικά με ένα Κατευθυνόμενο Ακυκλικό Γράφο του οποίου οι κόμβοι αναπαριστούν μεταβλητές και τα βέλη ορίζουν τις σχέσεις εξάρτησης (Κύρκος 2015). Πιο συγκεκριμένα, η κατεύθυνση του βέλους από τη μεταβλητή X προς τη μεταβλητή Y, σημαίνει ότι η μεταβλητή Y (τέκνο της X)

εξαρτάται από τη X (γονέας της Y). Τα βέλη αναπαριστούν τις πιθανοτικές εξαρτήσεις μεταξύ των μεταβλητών. Οι εξαρτήσεις αυτές αναπαριστώνται με τις υπό συνθήκη πιθανότητες οι οποίες δημιουργούν έναν πίνακα πιθανοτήτων P, που ονομάζεται Πίνακας Υπό Συνθήκη Πιθανοτήτων (Conditional Probability Table).

Αν X, Y,Z τρεις μεταβλητές ενός Δικτύου Bayes, τότε οι μεταβλητές X, Y ονομάζονται υπό συνθήκη ανεξάρτητες (εικόνα 8), αν οι τιμές της μεταβλητής X, έχοντας ως δεδομένες τις τιμές των μεταβλητών Y και Z, εξαρτώνται μόνο από τις τιμές της μεταβλητής Z. Αν, δηλαδή, ισχύει η ισότητα $P(X \mid Y, Z) = P(X \mid Z)$.



Εικόνα 8: Υπό συνθήκη ανεξάρτητες μεταβλητές

Πηγή: https://ermongroup.github.io/cs228-notes/representation/directed/

Στα Δίκτυα Bayes ισχύει η τοπική ιδιότητα Markov: κάθε μεταβλητή είναι υπό συνθήκη ανεξάρτητη από τους μη απογόνους της όταν είναι δεδομένοι οι γονείς της (Κύρκος 2015).

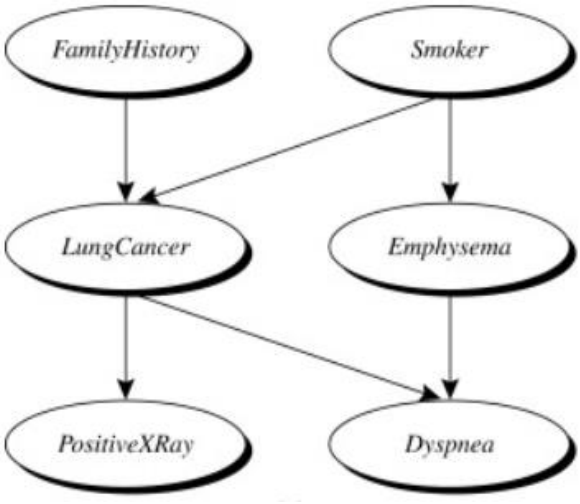
Στον Πίνακα Υπό Συνθήκη Πιθανοτήτων απεικονίζεται για κάθε μεταβλητή X η πιθανότητα P(X|Γονέας(X)) της μεταβλητής X. Αν έχουμε n το πλήθος μεταβλητές $X_1, X_2, \ldots, X_n$ τότε η πιθανότητα εμφάνισης της παρατήρησης $x_1, x_2, \ldots, x_n$ υπολογίζεται από τον τύπο

$$P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid \text{Γονέας}(x_i))$$

Ο Γονέας(X) είναι ένας κόμβος που βρίσκεται στο ακριβώς παραπάνω επίπεδο του γράφου, από αυτό της μεταβλητής X και συνδέεται απευθείας με τη μεταβλητή X. Για να δημιουργηθεί ένα μοντέλο Δικτύου Bayes πρέπει να κατασκευαστεί ο γράφος του Δικτύου και να υπολογιστεί ο Πίνακας Υπό Συνθήκη Πιθανοτήτων.

Ένας Δίκτυο Bayes χρησιμοποιείται για προβλέψεις ως κατηγοριοποιητής με έναν από τους κόμβους του Δικτύου να αντιπροσωπεύει τη μεταβλητή της κλάσης. Υπολογίζονται οι τιμές των πιθανοτήτων για μια παρατήρηση που αφορούν όλες τις τιμές της κλάσης και η παρατήρηση κατηγοριοποιείται στην κλάση με τη μεγαλύτερη πιθανότητα. Ο υπολογισμός των τιμών του Πίνακα Υπό Συνθήκη Πιθανοτήτων δεν παρουσιάζει συνήθως ιδιαίτερες δυσκολίες, όμως, η κατασκευή του γράφου είναι δύσκολη εργασία και είτε σχεδιάζεται από ειδικούς είτε αυτόματα με διάφορες μεθόδους που έχουν προταθεί από ερευνητές. Πρέπει να σημειωθεί πως

είναι δυνατή η τροποποίηση του αυτοματοποιημένου γράφου από τους ειδικούς, αξιοποιώντας με τον τρόπο αυτό προηγούμενη γνώση. Ένα Δίκτυο Bayes παρουσιάζεται στην εικόνα 9.



	FH, S	FH, ~S	~FH, S	~FH, ~S
LC	0.8	0.5	0.7	0.1
~LC	0.2	0.5	0.3	0.9

Εικόνα 9: Δίκτυο Bayes με τον αντίστοιχο Πίνακα Υπό Συνθήκη Πιθανοτήτων

Πηγή: Russell et al 1997

Στα πλεονεκτήματα των Δικτύων Bayes είναι η δυνατότητα χειρισμού αριθμητικών και ονομαστικών μεταβλητών, καθώς οι υψηλές αποδόσεις που επιτυγχάνουν. Ακόμα, τα δίκτυα Bayes είναι κατανοητά με ευκολία από τους ανθρώπους καθώς με το γράφο του Δικτύου οπτικοποιούνται οι σχέσεις μεταξύ της κλάσης και των μεταβλητών εισόδου.
Στα μειονεκτήματα των δικτύων Bayes πρέπει να αναφερθούν τα επόμενα: η δημιουργία του γράφου από τα δεδομένα δεν γίνεται από μια διαδικασία γενικά αποδεκτή και το υψηλό υπολογιστικό κόστος υπολογισμού των πιθανοτήτων.

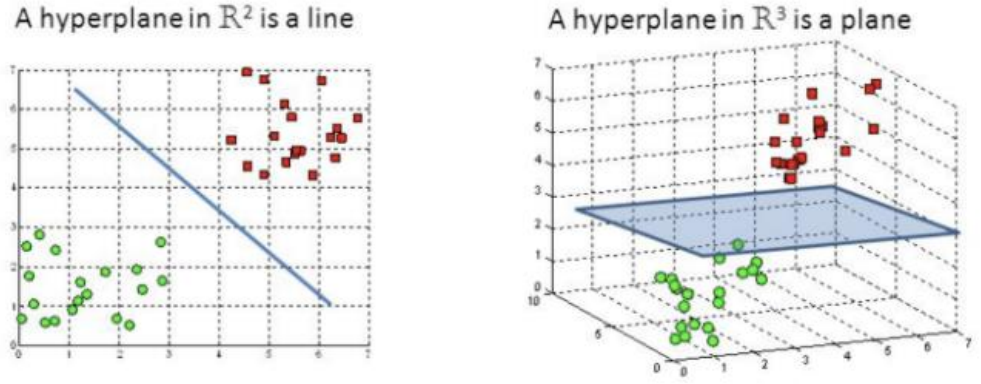
3.3.5 Μηχανές διανυσμάτων υποστήριξης

Οι Μηχανές Διανυσμάτων Υποστήριξης (ΜΔΥ) (Support vector machines), αποτελούν μια από τις πιο δημοφιλείς μεθόδους παρεμβολής και κατηγοριοποίησης. Τις πρότεινε το 1995, ο Vapnik και τα τελευταία χρόνια παρατηρείται αυξανόμενη χρήση σε εφαρμογές όπως αναγνώριση χειρόγραφων ψηφίων, εντοπισμό προσώπων σε εικόνες, κατηγοριοποίηση κειμένου, αναγνώριση αντικειμένων κ.ά.
Στην περίπτωση της κατηγοριοποίησης, οι ΜΔΥ προσπαθούν να βρουν ένα υπερεπίπεδο (ή υπερεπιφάνεια) (hyperplane/hypersurface). Στόχος του υπερεπιπέδου είναι ο διαχωρισμός, στο χώρο παραδειγμάτων, των αρνητικών από τα θετικά παραδείγματα. Καθώς υπάρχουν περισσότερα από ένα υπερεπίπεδα τα οποία είναι κατάλληλα για το διαχωρισμό, το υπερεπίπεδο επιλέγεται έτσι ώστε απόστασή του από τα κοντινότερα θετικά αλλά και αρνητικά

παραδείγματα να είναι η μεγαλύτερη δυνατή και ονομάζεται υπερεπίπεδο μεγίστου περιθωρίου. Οι παρατηρήσεις διαχωρίζονται σε κλάσεις και οι νέες παρατηρήσεις κατηγοριοποιούνται βάσει της θέσης τους ως προς το υπερεπίπεδο, δηλαδή, τα υπερεπίπεδα αποτελούν τα όρια απόφασης για την κατηγοριοποίηση των στοιχείων σε κλάσεις.

Το περιθώριο (margin) d υπολογίζεται από τις αποστάσεις του υπερεπιπέδου από το πλησιέστερο θετικό και το πλησιέστερο αρνητικό παράδειγμα και ορίζεται ως η ελάχιστη απόσταση μιας παρατήρησης από το υπερεπίπεδο.

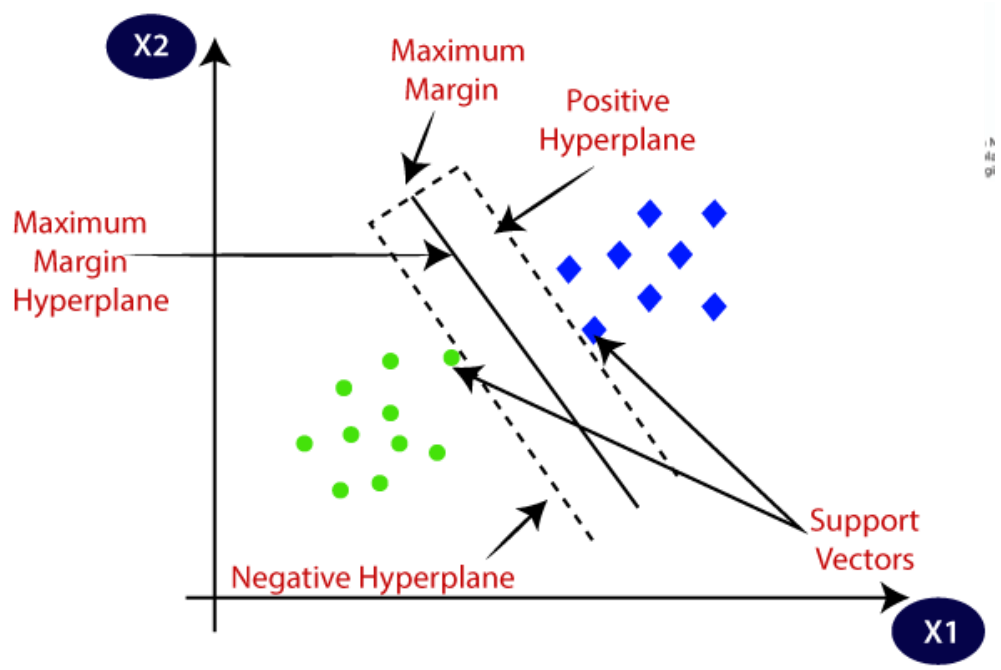
Στον 2-διάστατο χώρο $\mathbb{R}^2$ το υπερεπίπεδο είναι ευθεία, ενώ στον 3-διάστατο χώρο $\mathbb{R}^3$ είναι ένα επίπεδο (εικόνα 10), με την γενικότερη χρήση του όρου υπερεπίπεδο να αναφέρεται σε εκείνη την επιφάνεια που επιτελεί τον διαχωρισμό ανάλογα με τις διαστάσεις των δεδομένων εισόδου.



Εικόνα 10: Υπερεπίπεδα στον $\mathbb{R}^2$ και στον $\mathbb{R}^3$

Πηγή: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Τα σημεία που βρίσκονται στο όριο του περιθωρίου καλούνται διανύσματα υποστήριξης (support vectors) (εικόνα 11).



Εικόνα 11: Βέλτιστο υπερεπίπεδο και διανύσματα υποστήριξης

Πηγή: https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

Η διαδικασία προσδιορισμού του υπερεπιπέδου μέγιστου περιθωρίου, σύμφωνα με τον Κύρκο 2015, έχει ως εξής:

Το υπερεπίπεδο διαχωρισμού έχει εξίσωση $y = wx + b$ με w, b όπου το w διάνυσμα ίδιων διαστάσεων με το χαρακτηριστικό διάνυσμα x και κάθετο στο επίπεδο και b πραγματικός αριθμός.

Έτσι, τα πλησιέστερα θετικά και αρνητικά παραδείγματα βρίσκονται στα υπερεπίπεδα με εξισώσεις $wx + b = +1$ και $wx + b = -1$ αντίστοιχα. Έστω x_1 το πλησιέστερο θετικό παράδειγμα και x_2 το πλησιέστερο αρνητικό παράδειγμα. Έχουμε:

$$wx_1 + b = +1$$

$$wx_2 + b = -1$$

Από όπου προκύπτει ότι:

$$w\left(x_1 - x_2\right) = 2 \Leftrightarrow \|w\| \cdot d = 2 \Leftrightarrow d = \frac{2}{\|w\|}$$

Ακόμα, ισχύει ότι:

$$wx_i + b \geq +1 \Leftrightarrow y_i = +1$$

$$wx_i + b \leq -1 \Leftrightarrow y_i = -1$$

άρα,

$$y_i\left(wx_i + b\right) \geq 1 \, \forall i \in \{1, 2, \ldots, n\}$$

Για να επιτευχθεί το μέγιστο περιθώριο d, πρέπει να λυθεί το πρόβλημα βελτιστοποίησης

$$\min_{w,b} \frac{1}{2}\|w\|^2$$

$$y_i\left(wx_i + b\right) \geq 1 \forall i \in \{1, 2, ..., n\}$$

κάτι που επιτυγχάνεται με χρήση λογισμικού.

Τελικώς, η κατηγοριοποίηση μιας εγγραφής x γίνεται με υπολογισμό της τιμής της συνάρτησης

$$f\left(x\right) = wx + b$$

όπου αν $f\left(x\right) > 0$ είναι $y\left(x\right) = +1$ αλλιώς είναι $y\left(x\right) = -1$.

Οι μηχανές διανυσμάτων υποστήριξης έχουν καλή απόδοση και χαμηλό υπολογιστικό κόστος καθώς κατά την κατηγοριοποίηση νέων παραδειγμάτων, χρησιμοποιούν ένα μικρό τμήμα του συνόλου εκπαίδευσης δεδομένων, το οποίο πρέπει να καταχωρηθεί στη μνήμη. Ο διαχωρισμός των κλάσεων μπορεί να γίνει με γραμμικό και μη-γραμμικό τρόπο. Στην περίπτωση δυαδικών κλάσεων επιτυγχάνουν υψηλές επιδόσεις ενώ η απόδοσή τους παραμένει υψηλή όταν τα δεδομένα έχουν πολλές στήλες αλλά σχετικά λίγες γραμμές, ενώ παρουσιάζουν ανεκτικότητα όταν το πλήθος των παραδειγμάτων ανά κλάση διαφέρει. Ακόμα λειτουργούν πολύ καλά με δεδομένα πολλών διαστάσεων. Μειονεκτούν στο ότι είναι χρονοβόρα διαδικασία και έχουν έλλειψη εκφραστικής ισχύος.

3.3.6 Τεχνητά Νευρωνικά δίκτυα

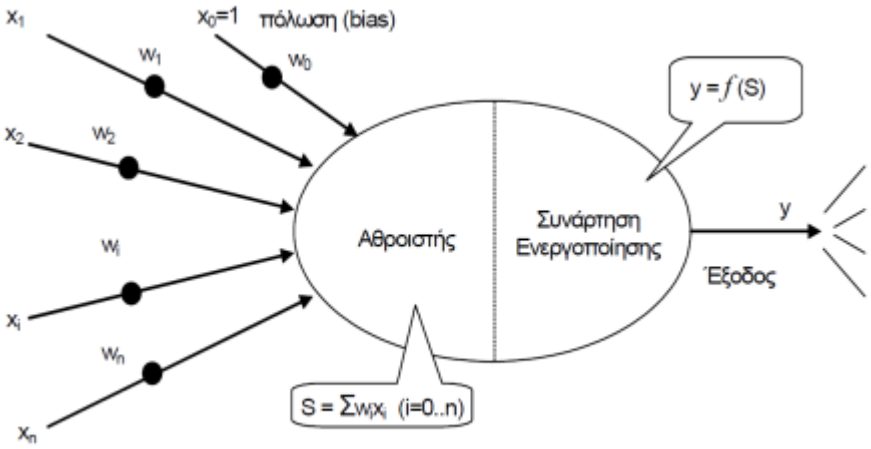
Τα Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks) είναι εμπνευσμένα, και η αρχιτεκτονική τους βασίζεται, στα Βιολογικά Νευρωνικά Δίκτυα και πιο συγκεκριμένα στον ανθρώπινο εγκέφαλο. Είναι μια ιδιαίτερα δημοφιλής τεχνική και είναι αποτέλεσμα των εργασιών προσομοίωσης του τρόπου λειτουργίας του ανθρώπινου εγκεφάλου.

Ο ανθρώπινος εγκέφαλος περιέχει νευρικά κύτταρα που ονομάζονται νευρώνες (neurons) και συνδέονται με νευρίτες (νηματοειδείς ίνες). Αντίστοιχα, στα νευρωνικά δίκτυα, ο νευρώνας αποτελεί τη βασική δομική μονάδα και δέχεται τιμές εισόδου βάσει των οποίων υπολογίζει την τιμή εξόδου. Οι νευρώνες συνδέονται μεταξύ τους με κατευθυνόμενα βέλη ή συνδέσεις τροφοδοτώντας με τον τρόπο αυτό, ως εισόδους, άλλους νευρώνες. Στις συνδέσεις αποδίδονται και αριθμητικές τιμές που ονομάζονται βάρη w (weights).

Η επεξεργασία που διενεργεί ένας νευρώνας ολοκληρώνεται σε δύο στάδια (εικόνα 12):

- Αρχικά, οι τιμές εξόδου των συνδεδεμένων νευρώνων πολλαπλασιάζονται με τα αντίστοιχα βάρη των συνδέσεων, δίνοντας έτσι τις τιμές εισόδου του νευρώνα. Στο πρώτο στάδιο, οι τιμές εισόδου αθροίζονται.

- Στο δεύτερο στάδιο, η συνάρτηση ενεργοποίησης (activation function) του νευρώνα, λαμβάνει το άθροισμα των τιμών εισόδου και δίνει μια νέα τιμή, που αποτελεί την έξοδο του νευρώνα. Έτσι, η συνάρτηση ενεργοποίησης ονομάζεται και συνάρτηση μετασχηματισμού.

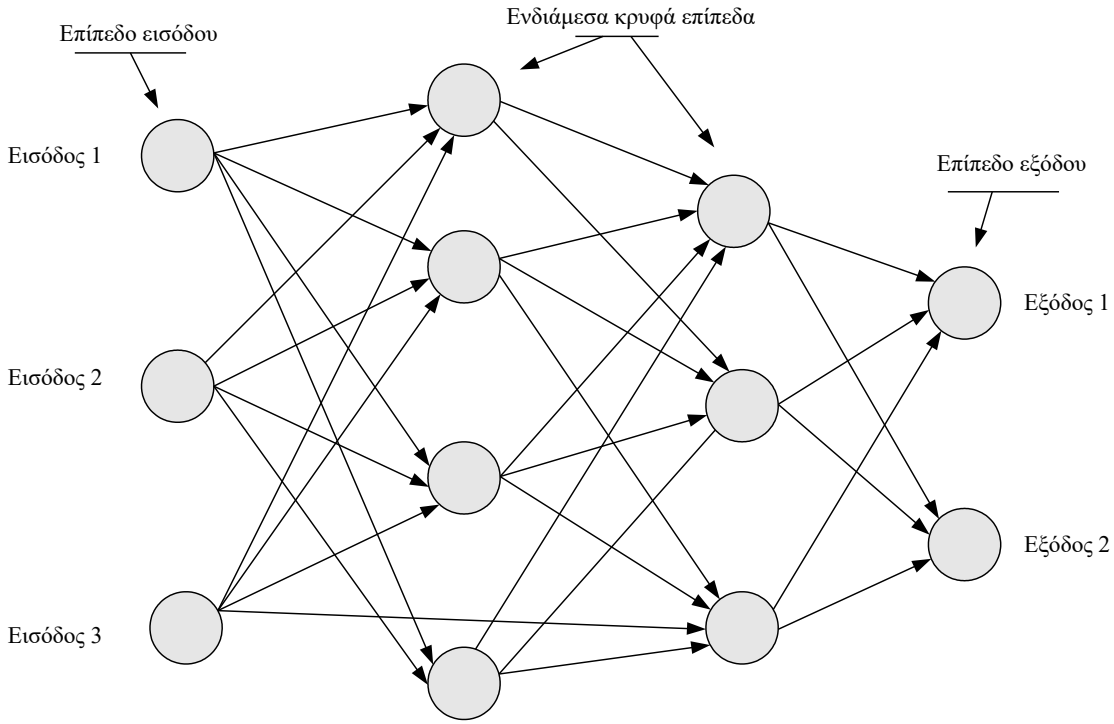


Εικόνα 12: Μοντέλο Τεχνητού Νευρώνα

Πηγή: Μυλωνάς 2022

Η βασική απαίτηση για τη συνάρτηση ενεργοποίησης είναι να είναι μη γραμμική ώστε να μοντελοποιεί μη γραμμικά φαινόμενα. Η πλέον συνήθης συνάρτηση ενεργοποίησης είναι η σιγμοειδής συνάρτηση $f(x) = \frac{1}{1+e^{-x}}$ ενώ άλλες συναρτήσεις ενεργοποίησης είναι η βηματική, η συνάρτηση προσήμου και η λογιστική.

Ένα νευρωνικό δίκτυο αναπαρίσταται ως ένας κατευθυνόμενος γράφος (εικόνα 13). Κάθε κόμβος αυτού του γράφου είναι ένας νευρώνας και οι ακμές του είναι οι συνδέσεις των νευρώνων.



Εικόνα 13: Γράφος νευρωνικού δικτύου

Πηγή: Μυλωνάς 2022

Ένα τυπικό νευρωνικό δίκτυο αποτελείται από επίπεδα (layers), και πιο συγκεκριμένα:

- Το πρώτο επίπεδο ονομάζεται επίπεδο εισόδου (input layer). Περιλαμβάνει κόμβους που δέχονται τις τιμές από τις αντίστοιχες ανεξάρτητες μεταβλητές.
- Τα ενδιάμεσα κρυφά επίπεδα (hidden layers), οι νευρώνες των οποίων δέχονται τις τιμές από τους νευρώνες του πρώτου επιπέδου πολλαπλασιασμένες με τα αντίστοιχα βάρη συνδέσεων. Το άθροισμα των τιμών αυτών το λαμβάνει ως είσοδο η συνάρτηση ενεργοποίησης. Η τιμή εξόδου της συνάρτησης είναι η τιμή εξόδου του κρυφού νευρώνα και πολλαπλασιάζεται με τα βάρη των αντίστοιχων συνδέσεων, ώστε να μεταβιβαστεί στον αντίστοιχο νευρώνα του επιπέδου εξόδου. Είναι δυνατόν να υπάρχουν διαφορετικές συναρτήσεις μετασχηματισμού σε νευρώνες διαφορετικών επιπέδων (Κύρκος 2015). Το πλήθος των κρυφών επιπέδων καθορίζεται από το χρήστη του νευρωνικού δικτύου.
- Το επίπεδο εξόδου, το πλήθος κόμβων του οποίου καθορίζεται από το πλήθος των κατηγοριών, όταν αντιμετωπίζεται ένα πρόβλημα κατηγοριοποίησης.

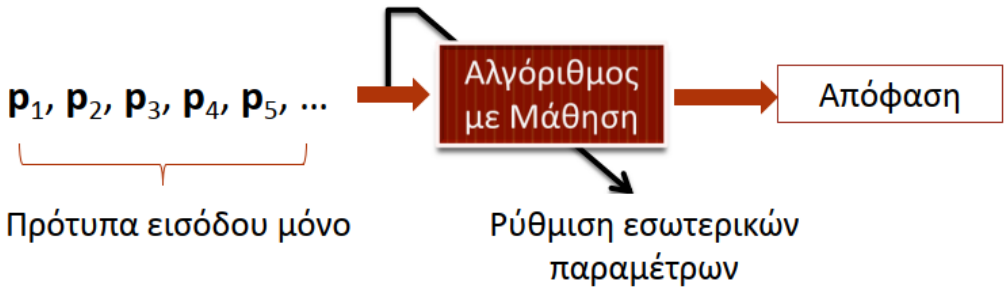
Τα πλεονεκτήματα των νευρωνικών δικτύων είναι ότι η απόδοσή τους είναι καλή σε πολύπλοκα προβλήματα εκεί όπου η χρήση άλλων τεχνικών δεν κρίνεται επαρκής. Παραμετροποιούνται εύκολα και βελτιώνονται από τη μάθηση. Όμως η κατανόησή τους από το χρήστη είναι δύσκολα σε σχέση με άλλες τεχνικές, π.χ. δέντρα απόφασης. Η δημιουργία

κανόνων από τα νευρωνικά δίκτυα δεν είναι διαδικασία που μπορεί να χαρακτηριστεί εύκολη. Χρησιμοποιούνται για αριθμητικές τιμές δεδομένων.

4. Μάθηση χωρίς επίβλεψη

4.1 Εισαγωγή

Η μηχανική μάθηση χωρίς επίβλεψη είναι η κατηγορία μηχανικής μάθησης που χρησιμοποιείται για εκμάθηση από εμπειρία καθώς και για να εκπαιδεύσει ένα μοντέλο χωρίς να χρειάζεται επίβλεψη από το χειριστή, ή με άλλα λόγια στόχος είναι η αυτοματοποιημένη παραγωγή γνώσης. Στη μάθηση χωρίς επίβλεψη, σύμφωνα με τους Βλαχάβα et al 2006, το σύστημα πρέπει μόνο του να ανακαλύψει συσχετίσεις ή ομάδες σε ένα σύνολο δεδομένων, δημιουργώντας πρότυπα, χωρίς να είναι γνωστό αν υπάρχουν, πόσα και ποια είναι.
Το μοντέλο παίρνει ως είσοδο δεδομένα και επαναλαμβάνει τη διαδικασία εκπαίδευσης μέχρι να φτάσει σε μία ικανοποιητική επίδοση. Το μοντέλο χρησιμοποιεί τα πρότυπα εισόδου p_1, p_2, p_3, … χωρίς όμως τους αντίστοιχους στόχους, καθώς δεν είναι γνωστοί. Το διάγραμμα μάθησης απεικονίζεται στην εικόνα 14:



Εικόνα 14: Μάθηση χωρίς επίβλεψη

Πηγή: Διαμαντάρας & Μπότσης 2019

Το μοντέλο μπορεί να εφαρμόσει τη γνώση που έχει και να την εκμεταλλευτεί για να κάνει προβλέψεις ή να λάβει αποφάσεις χωρίς να χρειάζεται επίβλεψη. Στόχος είναι η ανακάλυψη συσχετίσεων ανάμεσα στα δεδομένα, που ονομάζεται «ανακάλυψη κανόνων συσχέτισης» (association rule mining) ή ο χωρισμός των δεδομένων σε ομάδες ώστε τα δεδομένα κάθε ομάδας να μοιάζουν το μέγιστο δυνατό μεταξύ τους και να διαφέρουν το μέγιστο δυνατό από τα δεδομένα των άλλων ομάδων, διαδικασία που ονομάζεται «ομαδοποίηση» (clustering).

4.2 Ανακάλυψη Κανόνων Συσχέτισης

Η ανακάλυψη κανόνων συσχέτισης έχει ως στόχο την ανακάλυψη συσχετίσεων ανάμεσα στα δεδομένα ενός συνόλου. Παρουσιάστηκε το 1993 από τους Agrawal et al σαν μια τεχνική ανάλυσης καλαθιού αγορών. Ένα καλάθι αγορών περιέχει τα αντικείμενα που αγοράζουν οι πελάτες ενός καταστήματος. Η ανακάλυψη των κανόνων συσχέτισης μελετά και αναλύει τις αγορές που κάνουν οι πελάτες ενός καταστήματος και έχει ως στόχο είναι τον εντοπισμό εκείνων των αντικειμένων που αγοράζονται μαζί από τους πελάτες. Με τον τρόπο αυτό αναλύονται οι συνήθειες και οι ανάγκες αγορών των πελατών.

Ένας κανόνας συσχέτισης είναι μια σχέση της μορφής:

$$X \to Y$$

Με X, Y υποσύνολα των εμπορευμάτων του καταστήματος που δεν έχουν κοινά στοιχεία, δηλαδή ισχύει $X \cap Y = \varnothing$.

Για παράδειγμα ο κανόνας

$$\{X_1, X_2\} \to \{Y_1, Y_2\}$$

σημαίνει ότι όταν ένας πελάτης αγοράζει τα προϊόντα X_1 και X_2 (δηλαδή τα προϊόντα αυτά υπάρχουν μαζί στο καλάθι αγορών) τότε αγοράζει και τα προϊόντα Y_1 και Y_2.

Για τη μέτρηση της ισχύος ενός κανόνα, ορίζονται δύο βασικές μετρικές: η υποστήριξη και η εμπιστοσύνη.

- Η υποστήριξη (support) του κανόνα $X \to Y$ είναι το ποσοστό των συναλλαγών που περιέχουν και το X και το Y, δηλαδή

$$\text{supp}(X \to Y) = P(X \cup Y)$$

- Η εμπιστοσύνη (confidence) του κανόνα $X \to Y$ είναι η δεσμευμένη πιθανότητα εμφάνισης του Y όταν εμφανίζεται το X, δηλαδή

$$\text{conf}(X \to Y) = P(Y \mid X) \text{ ή } \text{conf}(X \to Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)}$$

Για την ανακάλυψη κανόνων συσχέτισης πρέπει να οριστούν από το χρήστη οι κατώτατες δεκτές τιμές (όρια) για την υποστήριξη και την εμπιστοσύνη. Από όλους τους κανόνες που προκύπτουν απορρίπτονται όσοι έχουν υποστήριξη και εμπιστοσύνη κάτω από το όριο και οι απομένουν οι υπόλοιποι κανόνες, που θεωρούνται ισχυροί.

Η υποστήριξη και η εμπιστοσύνη απεικονίζουν τη συχνότητα εφαρμογής και την αξιοπιστία του κανόνα αντίστοιχα. Για παράδειγμα ένας κανόνας που έχει υποστήριξη 0,5%, σημαίνει ότι το 0,5% όλων των πελατών αγοράζει τα προϊόντα αυτά μαζί και ως χαμηλό ποσοστό δεν έχει πρακτικό ενδιαφέρον, άρα υψηλή υποστήριξη συνεπάγεται κανόνα που έχει μεγαλύτερο ενδιαφέρον.

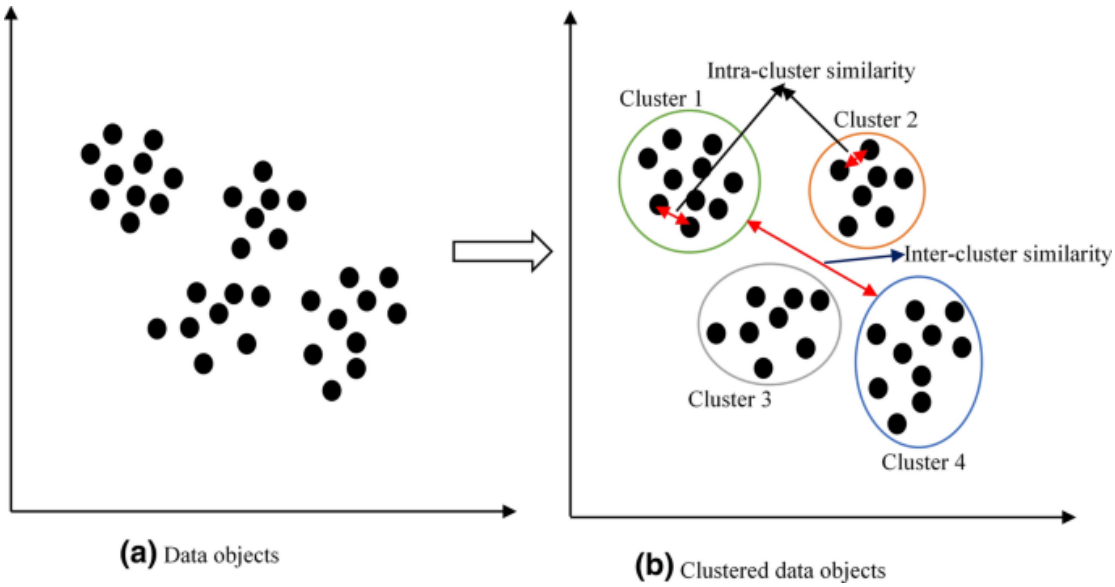
Εμπιστοσύνη 70% σημαίνει ότι το 70% των πελατών που αγοράζει τα προϊόντα του συνόλου X, αγοράζει και τα προϊόντα του συνόλου Y.

Οι κανόνες συσχέτισης παρουσιάζουν πολλά πλεονεκτήματα. Αφενός η αναπαράστασή τους είναι εύκολα κατανοητή από τους ανθρώπους, αφετέρου παρουσιάζουν πολύ καλή απόδοση όταν υπάρχει μεγάλος όγκος διαθέσιμων δεδομένων. Το βασικό μειονέκτημα τους είναι η επιλογή των ουσιαστικών κανόνων από σύνολο των κανόνων που παράγονται.

4.3 Ομαδοποίηση

Ομαδοποίηση (clustering) είναι ο χωρισμός ενός συνόλου δεδομένων σε ομάδες (clusters) με βάση κάποιο μέτρο ομοιότητας. Τα δεδομένα χωρίζονται με βάση ορισθέντα κριτήρια και ζητούμενο είναι όλα τα στοιχεία μιας ομάδας να μοιάζουν μεταξύ τους το περισσότερο δυνατό και να διαφέρουν το περισσότερο δυνατό από τα στοιχεία των άλλων ομάδων.

Σύμφωνα με τους Tan et al 2006, η ομαδοποίηση ομαδοποιεί τα αντικείμενα δεδομένων με βάση μόνο τις πληροφορίες που βρίσκονται στα δεδομένα και που περιγράφουν τα αντικείμενα και τις σχέσεις τους (εικόνα 15). Όσο πιο μεγάλη είναι η ομοιότητα εντός μιας ομάδας και όσο πιο μεγάλη είναι η διαφορά μεταξύ ομάδων, τόσο πιο καλή ή διακριτή είναι η ομαδοποίηση.



Εικόνα 15:Ομαδοποίηση δεδομένων

Πηγή: Ezugwu et al 2020

Στην περίπτωση που τα δεδομένα λαμβάνουν αριθμητικές τιμές, ως κριτήριο ομοιότητας δύο δεδομένων μπορεί να χρησιμοποιηθεί η Ευκλείδεια απόστασή τους. Για δύο δεδομένα x, y καθένα m χαρακτηριστικών, η Ευκλείδεια απόστασή τους είναι

$$d(x, y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$$

Ενώ μπορεί να χρησιμοποιηθεί και η απόσταση Manhattan

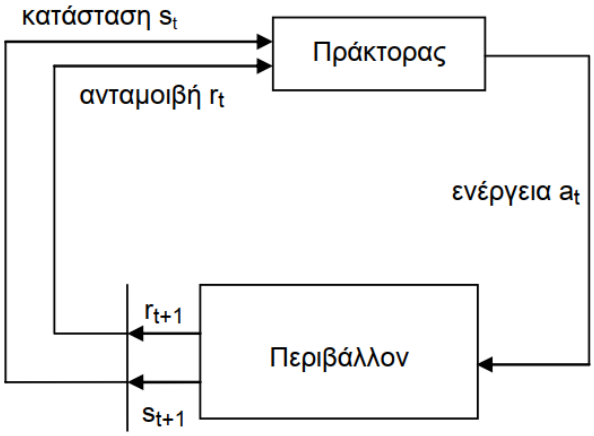
$$d(x, y) = \sum_{i=1}^{m} |x_i - y_i|$$

Για τα χαρακτηριστικά που είναι διακριτά, αν η τιμή τους είναι ίδια η απόστασή τους είναι 0, αλλιώς είναι 1.

Οι πιο γνωστές τεχνικές ομαδοποίησης είναι οι αλγόριθμοι K-μέσων (K-means) και DBSCAN.

5. Μάθηση με ενίσχυση

Η ενισχυτική μάθηση ασχολείται με τη λύση του προβλήματος της μάθησης μιας βέλτιστης συμπεριφοράς ενός πράκτορα που ενεργεί σε ένα περιβάλλον. Ο πράκτορας δρα σε ένα περιβάλλον με στόχο να λάβει τη μέγιστη ανταμοιβή. Μέσω της συνεχούς αλληλεπίδρασής του με το περιβάλλον λαμβάνει μια αναπαράσταση της τρέχουσας κατάστασης και στη συνέχεια προβαίνει σε ενέργειες βάσει κάποιας πολιτικής που ακολουθεί. Μετά την εκτέλεση ενεργειών λαμβάνουν χώρα αριθμητικές ανταμοιβές από το περιβάλλον. Ακολούθως ο πράκτορας βελτιώνει τις ενέργειες και τις εκτελεί, διαδικασία που επαναλαμβάνεται και μέσω της οποίας επιτυγχάνεται η μάθηση. Ο τύπος μάθησης με ενίσχυση αποτυπώνεται στην εικόνα 16:



Εικόνα 16: Μάθηση με ενίσχυση

Πηγή: https://www.kdnuggets.com/2018/03/5-things-reinforcement-learning.html

Η μάθηση με ενίσχυση διαφέρει από τη μάθηση με επίβλεψη, καθώς εδώ, δεν είναι γνωστή εκ των προτέρων η επιθυμητή έξοδος και επιπλέον οι βέλτιστες δυνατές ενέργειες δεν έχουν ρητά οριστεί, και διαφέρει από τη μάθηση χωρίς επίβλεψη, καθώς, ο πράκτορας δέχεται πληροφορίες από το περιβάλλον στο οποίο ενεργεί.

Σύμφωνα με τον Παρτάλα 2009, τα προβλήματα ενισχυτικής μάθησης μοντελοποιούνται συνήθως ως Μαρκοβιανές Διαδικασίες Απόφασης (Markov Decision Processes). Μια Μαρκοβιανή Διαδικασία Απόφασης περιγράφεται από μια τετράδα $< S, A, T, R >$ όπου S είναι το πεπερασμένο σύνολο των δυνατών καταστάσεων, A το πεπερασμένο σύνολο των ενεργειών, $T: S \times A \rightarrow S$ είναι η συνάρτηση μετάβασης (transition function) η οποία δεδομένης μιας κατάστασης και μιας ενέργειας επιστρέφει την επόμενη κατάσταση του περιβάλλοντος. Τέλος, η συνάρτηση ανταμοιβής, $R: S \rightarrow R$, απεικονίζει κάθε κατάσταση του περιβάλλοντος σε μια πραγματική αριθμητική τιμή (ανταμοιβή).

Υπάρχουν διάφοροι αλγόριθμοι για τη μάθηση με ενίσχυση, με δημοφιλέστερους τους Monte Carlo, Temporal Difference (TD) (χρονικών διαφορών), SARSA και Q-Learning.

6. Εφαρμογές της μηχανικής μάθησης

Υπάρχει μεγάλη πληθώρα εφαρμογών της μηχανικής μάθησης σε διάφορα πεδία, τα οποία συνεχώς αυξάνονται.

Ξεκινώντας από την εκπαίδευση, χρησιμοποιείται για να βελτιώσει την απόδοση των συστημάτων εκπαίδευσης και να παρέχει πληροφορίες σχετικά με την απόδοση των μαθητών. Είναι δυνατή η προσαρμογή του περιεχομένου του μαθήματος στο επίπεδο κατανόησης του εκάστοτε μαθητή, η παρακολούθηση της προόδου του μαθητή και η ανάλογη ανανέωση του περιεχομένου, ο εντοπισμός προβλημάτων στην απόκτηση γνώσης και η πρόταση παρεμβάσεων καθώς και η παροχή συμβουλευτικής υποστήριξης στους μαθητές και η καθοδήγησή τους σε επιπλέον πηγές γνώσης. Ακόμα, είναι δυνατή η πρόταση επιλογών κατάλληλων προγραμμάτων μαθημάτων για κάθε μαθητή ενώ μπορεί να χρησιμοποιηθεί για να δημιουργήσει προσαρμοσμένα προγράμματα εκπαίδευσης και για να προβλέψει την απόδοση των μαθητών και την πρόταση εξατομικευμένων και εκπαιδευτικά στοχευμένων ενεργειών.

Στον τομέα της ασφάλειας, η μηχανική μάθηση χρησιμοποιείται για τον εντοπισμό του κακόβουλου λογισμικού, αναγνωρίζοντας συγκεκριμένα χαρακτηριστικά σε επίπεδο κώδικα ή σε επίπεδο προσπέλασης δεδομένων. Ακόμα χρησιμοποιείται για την αναγνώριση των ανεπιθύμητων και κακόβουλων ηλεκτρονικών μηνυμάτων και ιστοσελίδων. Επιπρόσθετα, αποτελεί βασικό εργαλείο εντοπισμού ανεπιθύμητων κριτικών σε υπηρεσίες π.χ. ξενοδοχείων, αναγνωρίζοντας συνήθη χαρακτηριστικά του περιεχομένου τους, με πολύ υψηλά ποσοστά επιτυχίας. Όλες οι παραπάνω ενέργειες αποτελούν σοβαρό πρόβλημα για την αποτελεσματική χρήση και την επικοινωνία στο Διαδίκτυο. Πολύ βασική είναι η δυνατότητα προσαρμογής των μεθόδων της μηχανικής μάθησης στις τροποποιήσεις των μεθόδων των δημιουργών των κακόβουλων ενεργειών, δυνατότητα που δεν είναι εύκολη με τη χρήση των παραδοσιακών κανόνων εντοπισμού που βασίζονται σε κανόνες.

Στον χρηματοοικονομικό τομέα, η μηχανική μάθηση χρησιμοποιείται επιτυχώς για την πρόβλεψη της τιμής των μετοχών, για την ισοτιμία νομισμάτων, τον εντοπισμό επενδυτικών ευκαιριών κ.λπ. Αξιοποιεί τον υψηλό διαθέσιμο όγκο δεδομένων, όπως το ιστορικό των τιμών των μετοχών, τις παρελθούσες ισοτιμίες, τις τιμές του χρυσού, τα επιτόκια κ.ά.

Στην βιοπληροφορική, στον κλάδο της βιολογίας που μελετά την επίλυση προβλημάτων μέσω υπολογιστή, η μηχανική μάθηση χρησιμοποιείται στην πρόβλεψη καρκίνου και άλλων ασθενειών βάσει της γονιδιακής έκφρασης, στην ανίχνευση γονιδίων σε αλυσίδες DNA κ.λπ.

Στο εμπόριο και στη διαφήμιση, η μηχανική μάθηση χρησιμοποιείται ως προσωποποιημένο σύστημα συστάσεων προϊόντων και διαφημίσεων στις επιχειρήσεις παροχής προϊόντων και υπηρεσιών μέσω Διαδικτύου. Αξιοποιεί δεδομένα όπως οι προηγούμενες αγορές του πελάτη, το ιστορικό περιήγησης, τα σχόλια και οι αξιολογήσεις του σε προϊόντα, αλλά και πιθανά

διαθέσιμα δημογραφικά δεδομένα όπως η ηλικία του, το φύλο του, η χώρα, η διεύθυνση κατοικίας κ.λπ.

Στην αυτόνομη οδήγηση με ήδη γνωστά τα αυτοκίνητα των Google και Tesla, το πεδίο εφαρμογής είναι ιδιαίτερα απαιτητικό καθώς υπάρχει μεγάλο πλήθος παραμέτρων που πρέπει να λαμβάνεται υπόψη, όπως φωτεινοί σηματοδότες, πεζοί, λοιπά οχήματα, προτεραιότητες κ.ά. Μπορεί να φανεί ιδιαίτερα χρήσιμο σε περιπτώσεις ατόμων με ειδικές ανάγκες των οποίων η καθημερινότητα θα διευκολυνθεί και θα εμπλουτιστεί.

Στον ιατρικό τομέα παρέχεται η δυνατότητα ιατρικών διαγνώσεων μέσω παρακολούθησης ασθενών που νοσηλεύονται σε νοσοκομεία. Ακόμα μπορεί να χρησιμοποιηθεί για την υποβοήθηση της διάγνωσης μέσω υπολογιστή, όπου γίνεται πρόβλεψη της ασθένειας με χρήση μοντέλων και ιατρικών δεδομένων. Έτσι οι διαγνώσεις μπορούν να είναι ταχύτερες, εγκυρότερες αλλά και πιο εξεζητημένες, όπως για παράδειγμα σε περιπτώσεις σπάνιων ασθενειών όπου οι διαγνώσεις είναι δυσκολότερες και προϋποθέτουν επιπλέον γνώσεις αλλά και εξειδίκευση από τον θεράποντα ιατρό. Χρησιμοποιείται ακόμα για την αναγνώριση της συναισθηματικής του κατάστασης του ασθενούς, για την αναγνώριση κινήσεων και ενεργειών ασθενών προειδοποιώντας για μη ομαλές καταστάσεις κ.λπ.

Καθώς ο κατάλογος εφαρμογών της μηχανικής μάθησης είναι μακροσκελής, ακολουθεί ενδεικτική αναφορά λοιπών τομέων εφαρμογής που περιλαμβάνει την οπτική αναγνώριση χαρακτήρων, την αναγνώριση φυσικής ομιλίας (π.χ. αυτόματα συστήματα κράτησης θέσεων, Siri, Alexa, Cortana), την αναγνώριση εικόνας, τις μηχανές αναζήτησης, την υπολογιστική όραση, την πρόβλεψη καιρού, την πρόβλεψη και σύσταση προϊόντων προς αγορά, την πρόβλεψη κίνησης κυκλοφορίας, την αυτόματη μετάφραση κ.ά.

7. Βιβλιογραφική επισκόπηση

Το πλήθος των σχετιζόμενων με την απόδοση των μαθητών παραγόντων είναι μεγάλο. Οι παράγοντες αυτοί ενδέχεται να είναι σχετικοί με την εν γένει εκπαιδευτική διαδικασία ή να είναι προσωπικοί, οικογενειακοί, κοινωνικοί ή άλλοι παράγοντες.

Διάφορες μελέτες έχουν πραγματοποιηθεί που προσπαθούν να αναδείξουν τους παράγοντες αυτούς και αναλύσουν τη μεταξύ τους σχέση. Βαθύτερος στόχος όλων των μελετών είναι η εύρεση εκείνων των παραγόντων που επηρεάζουν στο μεγαλύτερο βαθμό την μαθησιακή πορεία και ο εντοπισμός των μαθητών των οποίων η πορεία απαιτεί παρεμβάσεις ώστε να επιτευχθεί το καλύτερο δυνατό αποτέλεσμα.

Μέθοδοι κατηγοριοποίησης έχουν χρησιμοποιηθεί στην εκπαίδευση με διάφορους στόχους. Ακολουθεί ενδεικτική αναφορά και μελέτες που έλαβαν χώρα το 2006:

- Για την ταξινόμηση των μαθητών σε hint-driven (προτιμούν τις υποδείξεις), failure-driven (προτιμούν την αποτυχία) ή καμία σημαντική προτίμηση ανάμεσα στις δύο αυτές κατηγορίες, καθώς και την ανακάλυψη των κοινών παρανοήσεων των μαθητών στη χρήση του SlideTutor (Yudelson at al 2006).
- Για την αναγνώριση των μαθητών που έχουν λιγότερα μαθησιακά κίνητρα και την εύρεση διορθωτικών ενεργειών για τη μείωση του αριθμού εγκατάλειψης των σπουδών (Cocea & Weibelzahl 2006).
- Για την πρόβλεψη της επιτυχίας μαθημάτων (Hamalainen & Vinni 2006).

Εστιάζοντας με μεγαλύτερη λεπτομέρεια σε σχετικές εργασίες, ήδη από το 2003 χρησιμοποιήθηκαν δεδομένα για 227 μαθητές που καταγράφηκαν από το Διαδικτυακό Σύστημα Διαχείρισης Μάθησης LON-CAPA. Στόχος της μελέτης ήταν η πρόβλεψη του τελικού βαθμού και η επιτυχία ή αποτυχία στο μάθημα «Physics for Scientists and Engineers I» του πανεπιστημίου Michigan. Ο αριθμός των κλάσεων ήταν είτε δύο (pass, fail), είτε τρεις (low, middle, high) είτε εννέα σύμφωνα με τους βαθμούς. Ανάλογα με τον επιθυμητό αριθμό των κλάσεων ξεχώρισαν οι αλγόριθμοι k-Nearest Neighbor (k-NN) (82,3%) στην περίπτωση των δύο κλάσεων και CART (Classification And Regression Tree) (59,9% και 33,1%) (Minaei-Bidgoli & Punch 2003).

Επίσης το 2003, στο Ελληνικό Ανοικτό Πανεπιστήμιο, χρησιμοποίησαν δεδομένα από 354 φοιτητές και τις επιδόσεις τους στην ενότητα «Εισαγωγή στην Πληροφορική». Τα δεδομένα ήταν προσωπικού χαρακτήρα όπως φύλο, ηλικία, οικογενειακή κατάσταση, αριθμός παιδιών, απασχόληση (καθόλου, μερική, ολική, υπερωρίες), εργασία σχετική με ηλεκτρονικούς υπολογιστές κ.ά. Τα ακαδημαϊκά δεδομένα περιλάμβαναν την παρουσία ή μη στις προγραμματισμένες συναντήσεις καθώς και τους βαθμούς σε εργασίες. Πιο αποδοτικός αποδείχθηκε ο αλγόριθμος Naïve Bayes με τον αλγόριθμο BP (Back Propagation) των νευρωνικών δικτύων να ακολουθεί. Πρέπει όμως να σημειωθεί πως οι διαφορές μεταξύ όλων

των αλγορίθμων ήταν μικρές. Τα ποσοστά ακρίβειας άγγιζαν το 63% όταν χρησιμοποιούνταν μόνο τα προσωπικά δεδομένα που ήταν διαθέσιμα από την έναρξη του ακαδημαϊκού έτους και ξεπερνούσαν το 83% πριν το μέσο του εξαμήνου (Kotsiantis et al 2003).

Σε μελέτη του 2011 χρησιμοποίησαν δεδομένα από τις επιδόσεις των φοιτητών στις εξετάσεις σε τέσσερα πανεπιστημιακά μαθήματα (Advanced Computer Architecture, Management Information System, Advanced Database System, Object Oriented Modeling & Design). Με τη χρήση του Weka, αναφέρουν ότι ο αλγόριθμος ZeroR είχε ποσοστό ακρίβειας 84,06% με την κλάση πρόβλεψης να αριθμεί πέντε δυνατές τιμές (distinction, first, second, pass, ATKT) (Aher & Lobo 2011).

Σε μελέτη του 2013 στόχος ήταν η πρόβλεψη της συνολικής απόδοσης των φοιτητών στο πανεπιστήμιο (excellent, very good, good, average, bad). Η μελέτη υλοποιήθηκε με χρήση του Weka. Το σύνολο δεδομένων αριθμούσε 10330 παραδείγματα με 14 χαρακτηριστικά προσωπικού και ακαδημαϊκού χαρακτήρα. Αν και η ακρίβεια των αλγορίθμων ήταν κάτω του 70% (κυμάνθηκε από 52% ως 67%), πρέπει να σημειωθεί πως ο αλγόριθμος J48 είχε την καλύτερη απόδοση με τα υψηλότερα ποσοστά ακρίβειας. Ακολούθησε ο k-NN τους Naïve Bayes και Bayes Net να είναι τελευταίοι (Kabakchieva 2013), σε αντίθεση με τους Mueen et al 2016 όπου ο αλγόριθμος Naïve Bayes πρώτευσε με ακρίβεια της τάξης του 86%. Στην τελευταία εργασία, στο Weka χρησιμοποιήθηκαν δεδομένα προσωπικού χαρακτήρα, ακαδημαϊκά αλλά και σχετικά με το Σύστημα Διαχείρισης Μάθησης. Ορισμένα από τα προσωπικού χαρακτήρα είναι ηλικία, πόλη γέννησης, καθημερινές ώρες μελέτης, επίπεδο μόρφωσης πατέρα και μητέρας, απόσταση από το πανεπιστήμιο, ακαδημαϊκά είναι οι βαθμοί, οι εργασίες, βαθμοί σε τεστ εργαστηρίων κ.ά. Χαρακτηριστικά σχετικά με τη χρήση του συστήματος από τους προπτυχιακούς φοιτητές στα μαθήματα (Programming Fundamental και Advanced Operating System) είναι για παράδειγμα οι απαντήσεις στο forum, ο συνολικός αριθμός λέξεων που έγραψε, η συνολική ώρα που διατέθηκε online κ.ά. Ακολούθησαν ο Multilayer Perception και ο J48 με ακρίβεια 82,7% και 79,2% αντίστοιχα. Ο Multilayer Perception ήταν αυτός που αναδείχθηκε με ακρίβεια 74,8% σε μελέτη του 2014 όπου το σύνολο δεδομένων αριθμούσε 165 εγγραφές, για άλλη μια φορά με χρήση του Weka. Τα δεδομένα περιείχαν χαρακτηριστικά όπως βαθμοί στη θεωρία, βαθμοί εργαστηρίων, οικογενειακό εισόδημα, μόρφωση γονέων κ.ά. (Ruby & David 2014).

Το 2017 χρησιμοποίησαν δεδομένα από το Sardar Patel Institute of Technology College MCA Department. Εφάρμοσαν τους αλγορίθμους Naïve Bayes, J48, ZeroR and Random Tree στο Weka. Στόχος ήταν ο διαχωρισμός των μαθητών σε fast, average και slow learners. Ο αλγόριθμος με το μεγαλύτερο ποσοστό ακρίβειας ήταν ο RandomTree με ποσοστό ακρίβειας 95,45%. Ακολούθησε ο J48 με ακρίβεια 72,72% ενώ οι Naïve Bayes και ZeroR είχαν ακρίβεια 68,18% και 36,36% αντίστοιχα (Mhetre & Nagar 2017).

Επίσης το 2017, χρησιμοποίησαν δεδομένα που αποτελούνταν από 300 εγγραφές των 24 χαρακτηριστικών, τα οποία συνέλεξαν από τρία πανεπιστήμια της Ινδίας. Τα δεδομένα περιείχαν προσωπικά, οικογενειακά αλλά και ακαδημαϊκά χαρακτηριστικά. Στόχος ήταν η κατηγοριοποιημένη πρόβλεψη της τελικής απόδοσης των μαθητών (best, very good, good, pass, fail). Στο συγκεκριμένο σετ δεδομένων ο αλγόριθμος Random Forest είχε ακρίβεια 99% και ακολούθησαν οι αλγόριθμοι PART, J48, Bayes Net με ποσοστά ακρίβειας 74,33%, 73% και 65,33% αντίστοιχα (Hussain et al 2017).

Σε μελέτη που πραγματοποιήθηκε το 2018 το σύνολο δεδομένων περιείχε 507 εγγραφές 18 χαρακτηριστικών από τις επιδόσεις των φοιτητών το προηγούμενο εξάμηνο. Τα δεδομένα ήταν προσωπικού χαρακτήρα (φύλο, τόπος γέννησης, σχολείο φοίτησης κ.ά.) αλλά και ακαδημαϊκού (βαθμοί μαθημάτων). Ζητούμενο ήταν η πρόβλεψη βαθμού (fail, good, excellent, outstanding). Διαπίστωσαν, με χρήση του Weka πως ο αλγόριθμος J48 είχε ποσοστό ακρίβειας 97,43% (Dey et al 2018).

Σε πιο πρόσφατη μελέτη, το 2021, τα αποτελέσματα αναδεικνύουν τους αλγορίθμους Naïve Bayes και Random Forest να ξεχωρίζουν με ακρίβεια 63,33% και 63% αντίστοιχα, για την πρόβλεψη του τελικού βαθμού με τα διαθέσιμα δεδομένα μετά το τρίτο εξάμηνο. Ο J48 σημείωσε ακρίβεια 55,67% στο δείγμα 300 εγγραφών που χρησιμοποιήθηκε. Η κατάταξη δεν άλλαξε για την πρόβλεψη μετά το τέταρτο εξάμηνο, απλά αυξήθηκε η ακρίβεια (69,67%, 67,6% και 61,67% αντίστοιχα). Τα δεδομένα προήλθαν από τρία τμήματα πανεπιστημίου Computer and Information Science της Σαουδικής Αραβίας και τα αποτελέσματα προέκυψαν με χρήση του Weka (Alturki & Alturki 2021).

Από τη βιβλιογραφία δεν απουσιάζουν οι συγκριτικές μελέτες. Οι Ashraf et al 2018 παρουσιάζουν έρευνες που διεξήχθησαν μεταξύ 2011 και 2017. Τις διαχωρίζουν ανάλογα τις μεθόδους κατηγοριοποίησης που χρησιμοποίησαν, τις μεταβλητές που περιλαμβάνονταν στα δεδομένα, την κλάση πρόβλεψης κ.λπ. Γίνεται καταγραφή των αποτελεσμάτων, της ακρίβειας των μεθόδων και άλλων μετρικών.

Συμπερασματικά, στις μελέτες το πρόγραμμα που χρησιμοποιήθηκε ήταν κατά βάση το Weka. Εφαρμόστηκαν οι γνωστοί αλγόριθμοι κατηγοριοποίησης με τα αποτελέσματα να διαφέρουν ανάλογα με το διαθέσιμο σετ δεδομένων. Οι διαφορές που παρατηρούνται αφορούν τους αλγορίθμους κατηγοριοποίησης που σημείωσαν τα μεγαλύτερα ποσοστά ακρίβειας σε κάθε μελέτη, αλλά και τα ποσοστά αυτά, καθώς σε κάποιες μελέτες ήταν ιδιαίτερα υψηλά ενώ σε άλλες χαρακτηρίζονται χαμηλά.

8. WEKA

8.1 Εισαγωγή

Το Weka (Waikato Environment for Knowledge Analysis) είναι ένα πρόγραμμα εφαρμογών μηχανικής μάθησης και εξόρυξης γνώσης από δεδομένα. Δημιουργήθηκε στο τμήμα Επιστήμης Υπολογιστών του Πανεπιστημίου του Waikato της Νέας Ζηλανδίας. Χρησιμοποιείται ευρέως για ερευνητικούς και εκπαιδευτικούς σκοπούς.

Είναι ένα πακέτο λογισμικού ανοιχτού κώδικα με GNU-General Public License (περισσότερες λεπτομέρειες για την Γενική Άδεια Δημόσιας Χρήσης είναι διαθέσιμες στον ιστότοπο https://www.gnu.org/licenses/licenses.en.html). Το Weka είναι υλοποιημένο στην γλώσσα προγραμματισμού Java. Η Java είναι μια δημοφιλής αντικειμενοστραφής γλώσσα προγραμματισμού διαθέσιμη για όλες τις πλατφόρμες υπολογιστών.

Το λογισμικό του Weka είναι διαθέσιμο για τα λειτουργικά συστήματα Windows, Linux, Macintosh. Η πρώτη έκδοση κυκλοφόρησε το 1997, ενώ πιο πρόσφατη έκδοση για τα Windows είναι η 3.8.6, η οποία χρησιμοποιήθηκε στην παρούσα εργασία. Στην επίσημη ιστοσελίδα του πανεπιστημίου του Waikato της Νέας Ζηλανδίας (https://waikato.github.io/weka-wiki/downloading_weka/) είναι διαθέσιμη για λήψη η πιο πρόσφατη έκδοση του λογισμικού. Υπάρχει επίσης και η developer version που απευθύνεται σε προγραμματιστές.

Στην ιστοσελίδα του πανεπιστημίου υπάρχουν επιπλέον και σύνδεσμοι που οδηγούν σε μαθήματα σχετικά με τη χρήση του Weka για μηχανική μάθηση και εξόρυξη δεδομένων καθώς και διάφορες άλλες σχετικές πληροφορίες (εικόνα 17).

Εικόνα 17: Η σελίδα του Weka στο πανεπιστήμιο Waikato

https://www.cs.waikato.ac.nz/ml/weka/index.html

Το όνομα και το λογότυπο (logo) του προγράμματος είναι εμπνευσμένα από ένα ενδημικό στη Νέα Ζηλανδία, χωρίς φτερά, πουλί που φέρει το όνομα weka (εικόνα 18).

Εικόνα 18: Το logo του προγράμματος και το πουλί weka
Πηγή: https://nzbirdsonline.org.nz/species/weka

8.2 Αλγόριθμοι και δυνατότητες του Weka.

Το Weka περιλαμβάνει μια μεγάλη γκάμα αλγορίθμων μηχανικής μάθησης. Περιλαμβάνονται αλγόριθμοι κατηγοριοποίησης, παλινδρόμησης, ομαδοποίησης, εύρεσης κανόνων συσχέτισης κ.ά. Οι αλγόριθμοι αυτοί μπορούν να χρησιμοποιηθούν είτε μέσω του γραφικού περιβάλλον του προγράμματος, είτε μέσω γραμμής εντολών (command line).
Το Weka παρέχει ακόμα εργαλεία για την προεπεξεργασία (preprocessing) των δεδομένων. Η διαδικασία της προεπεξεργασίας περιλαμβάνει τον καθαρισμό, την επεξεργασία και το μετασχηματισμό των δεδομένων, όπως π.χ. διακριτοποίηση (discretization), συγχώνευση ονομαστικών τιμών (merge nominal values) κ.ά. Παρέχει επίσης οπτικές αναπαραστάσεις των δεδομένων. Ακόμα, παρέχει δυνατότητες επεξεργασίας των αποτελεσμάτων (post-processing), αξιολόγησή τους, αλλά και την οπτικοποίησή (visualization) τους.

8.3 Η μορφή δεδομένων του WEKA.

Ο υποστηριζόμενος τύπος αρχείων του Weka είναι τα αρχεία με επέκταση .ARFF (Attribute-Relation File Format). Πρόκειται για ένα αρχείο κειμένου ASCII το οποίο περιλαμβάνει μια λίστα δεδομένων που δεν είναι παρά γραμμές με τιμές χαρακτηριστικών. Ένα αρχείο . ARFF έχει την δομή που φαίνεται στην εικόνα 19 όπου απεικονίζεται τμήμα του αρχείου των δεδομένων που χρησιμοποιήθηκε στην παρούσα εργασία:

```
@RELATION exams

@ATTRIBUTE gender {male, female}
@ATTRIBUTE raceethnicity {A, B, C, D, E}
@ATTRIBUTE parentaleducation {bachelor, master, somecollege, associate, highschool
@ATTRIBUTE lunch {standard, freereduced}
@ATTRIBUTE preparationcourse {none, completed}
@ATTRIBUTE math numeric
@ATTRIBUTE reading numeric
@ATTRIBUTE writing numeric

@DATA
female,B,bachelor,standard,none,72,72,74
female,C,somecollege,standard,completed,69,90,88
female,B,master,standard,none,90,95,93
male,A,associate,freereduced,none,47,57,44
male,C,somecollege,standard,none,76,78,75
female,B,associate,standard,none,71,83,78
female,B,somecollege,standard,completed,88,95,92
male,B,somecollege,freereduced,none,40,43,39
male,D,highschool,freereduced,completed,64,64,67
female,B,highschool,freereduced,none,38,60,50
male,C,associate,standard,none,58,54,52
male,D,associate,standard,none,40,52,43
female,B,highschool,standard,none,65,81,73
male,A,somecollege,standard,completed,78,72,70
female,A,master,standard,none,50,53,58
female,C,highschool,standard,none,69,75,78
male,C,highschool,standard,none,88,89,86
female,B,highschool,freereduced,none,18,32,28
male,C,master,freereduced,completed,46,42,46
female,C,associate,freereduced,none,54,58,61
male,D,highschool,standard,none,66,69,63
female,B,somecollege,freereduced,completed,65,75,70
male,D,somecollege,standard,none,44,54,53
female,C,highschool,standard,none,69,73,73
```

Εικόνα 19: Δομή αρχείου .arff

Διακρίνονται δύο βασικά ξεχωριστά τμήματα. Το τμήμα Κεφαλίδας (header) και το τμήμα Δεδομένων (data).

Στο τμήμα Κεφαλίδας, αρχικά υπάρχει η γραμμή που ξεκινά με το @relation. Στη γραμμή αυτή υπάρχει το όνομα της σχέσης (relation) που περιγράφει τα δεδομένα και δεν μπορεί να παραληφθεί.

Ακολουθεί η γραμμή @attribute στην οποία δηλώνονται όλα τα χαρακτηριστικά (ονόματα πεδίων) που περιλαμβάνονται στο σύνολο δεδομένων που θα εισαχθεί στο πρόγραμμα και πρέπει να είναι μοναδικά. Η δήλωση των χαρακτηριστικών πρέπει να ακολουθεί τη σύνταξη: @attribute <attribute-name> <datatype>.

Στο <attribute-name> δηλώνεται το όνομα του χαρακτηριστικού και στο <datatype> ο τύπος των δεδομένων του.

Οι τύποι των δεδομένων που υποστηρίζει το Weka είναι οι εξής τέσσερις:

- Αριθμητικός (numeric)
- Ονομαστικός (nominal)
- Αλφαριθμητικός (string)
- Ημερομηνία (date)

Ο αριθμητικός (numeric) τύπος μπορεί να είναι πραγματικός ή ακέραιος. Στον ονομαστικό (nominal) προσδιορίζονται εντός αγκυλών οι τιμές που μπορεί να λάβει το εν λόγω χαρακτηριστικό, ως εξής:

@attribute gender {male; female}

όπου το χαρακτηριστικό gender μπορεί να λάβει μόνο τις δύο τιμές που περικλείονται στο ζεύγος αγκυλών.

Ο αλφαριθμητικός (string) τύπος επιτρέπει στα χαρακτηριστικά να λάβουν αυθαίρετες (μη ορισμένες) τιμές.

Ο τύπος δεδομένων ημερομηνία (date) επιτρέπει τη χρήση ημερομηνιών συγκεκριμένης μορφής.

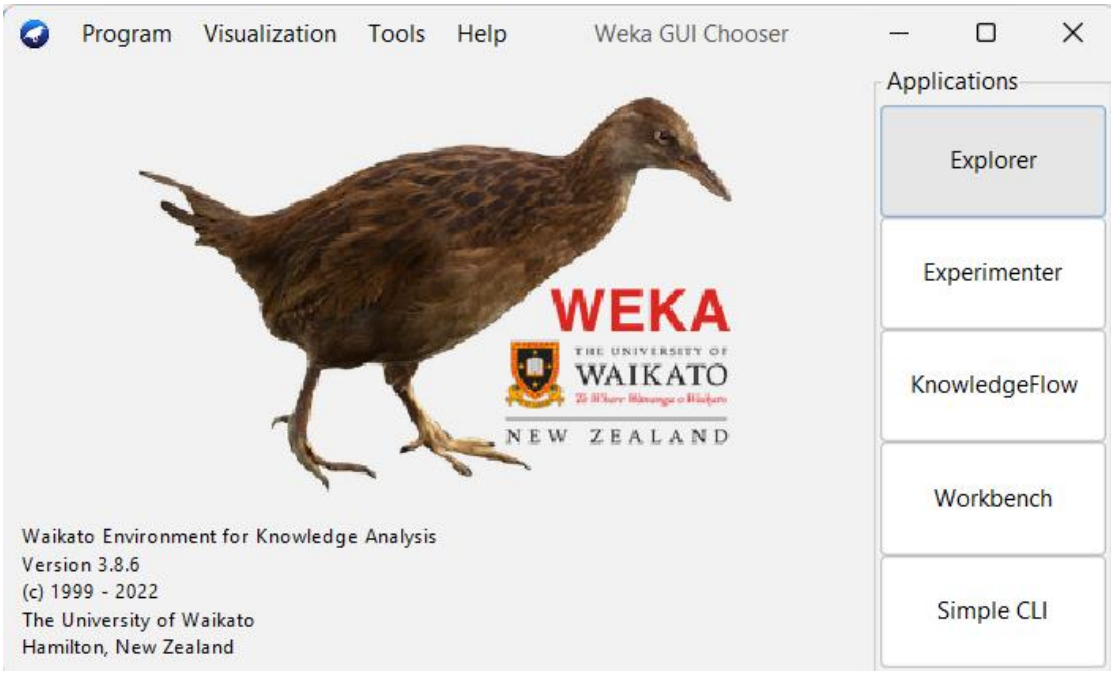
Στο τμήμα Δεδομένων, αρχικά υπάρχει η δήλωση @data που σηματοδοτεί την αρχή των δεδομένων. Τα δεδομένα είναι χωρισμένα σε γραμμές. Κάθε γραμμή περιλαμβάνει τις τιμές των χαρακτηριστικών που έχουν δηλωθεί στη γραμμή @attribute με την ίδια σειρά. Οι τιμές των χαρακτηριστικών χωρίζονται με κόμματα.

Ακόμα, μια γραμμή αποτελεί σχόλιο και αγνοείται κατά την ανάγνωση του αρχείου, όταν ξεκινά με το σύμβολο %.

Στο Weka ένα σύνολο δεδομένων (dataset) μπορεί να εισαχθεί ως αρχείο .ARFF, ως αρχείο CSV (Comma Separated Values), JSON, BSI (Binary Serialized Instances) κ.ά.. Σε αρκετές περιπτώσεις τα δεδομένα πρέπει να επεξεργαστούν ώστε να μπορούν να αναγνωριστούν από το Weka. Ειδικά για τα αρχεία .csv πρέπει να σημειωθεί ότι υπάρχει ένα μειονέκτημα σε σχέση με τον προτεινόμενο τύπο αρχείων.arff καθώς το σύνολο δεδομένων εκπαίδευσης και επαλήθευσης μπορεί να μην είναι συμβατά (https://waikato.github.io/weka-wiki/faqs/use_csv_files/).

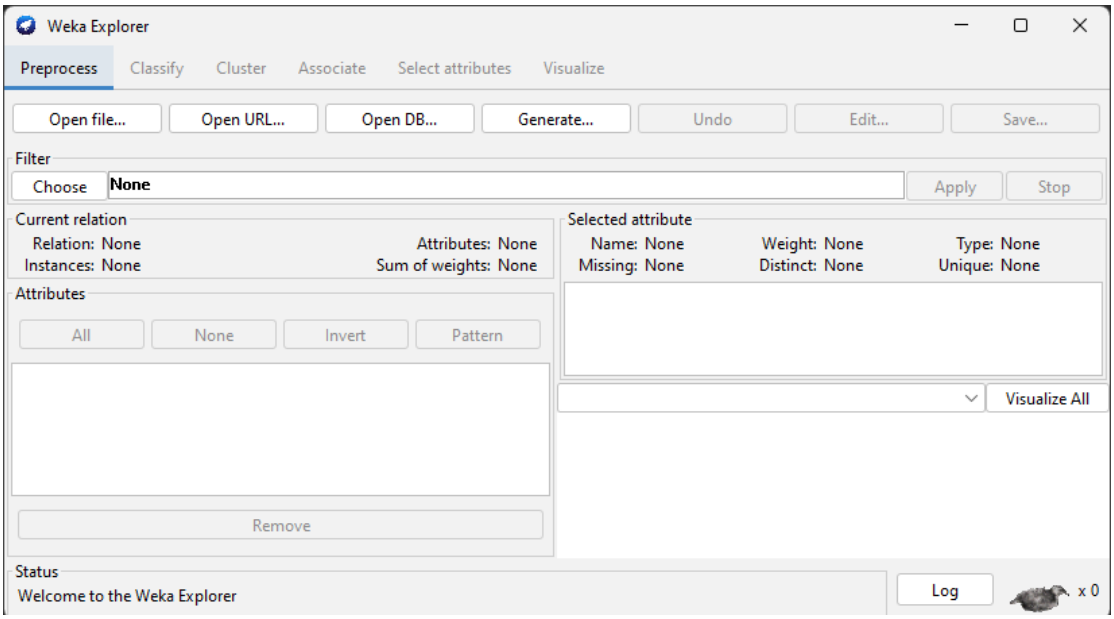
8.4 Το περιβάλλον διεπαφής (GUI) του Weka.

Η αρχική οθόνη του Weka φαίνεται στην εικόνα 20. Στο δεξί μέρος της οθόνης, στο πλαίσιο Applications, διακρίνονται οι επιλογές Explorer, Experimenter, KnowledgeFlow, Workbench, Simple CLI:



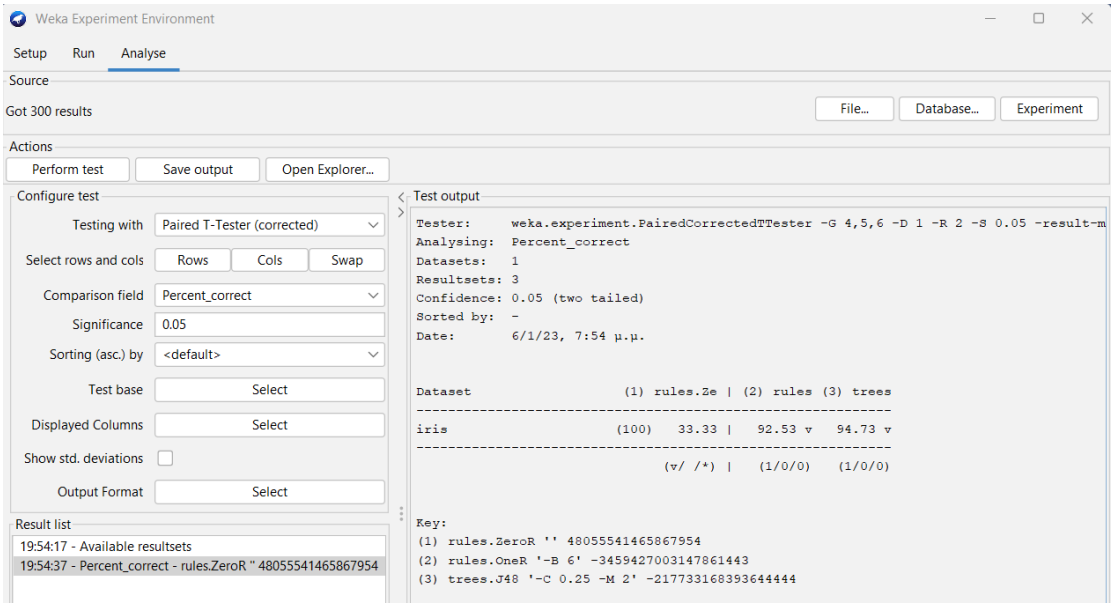
Εικόνα 20: Η κεντρική οθόνη του προγράμματος Weka

Explorer: Είναι το γραφικό περιβάλλον (διεπαφή) που απεικονίζεται στην εικόνα 21 και παρέχει πρόσβαση στις βασικές εργασίες του Weka. Αποτελεί τη δημοφιλέστερη επιλογή καθώς εκεί είναι διαθέσιμες οι επιλογές για το άνοιγμα του αρχείου δεδομένων, την εφαρμογή των φίλτρων, τους αλγορίθμους κατηγοριοποίησης, παλινδρόμησης, κ.λπ. Παρουσιάζεται αναλυτικά παρακάτω, καθώς αποτελεί το βασικό εργαλείο στο υλοποίησης της παρούσας εργασίας.



Εικόνα 21: Ο Weka Explorer

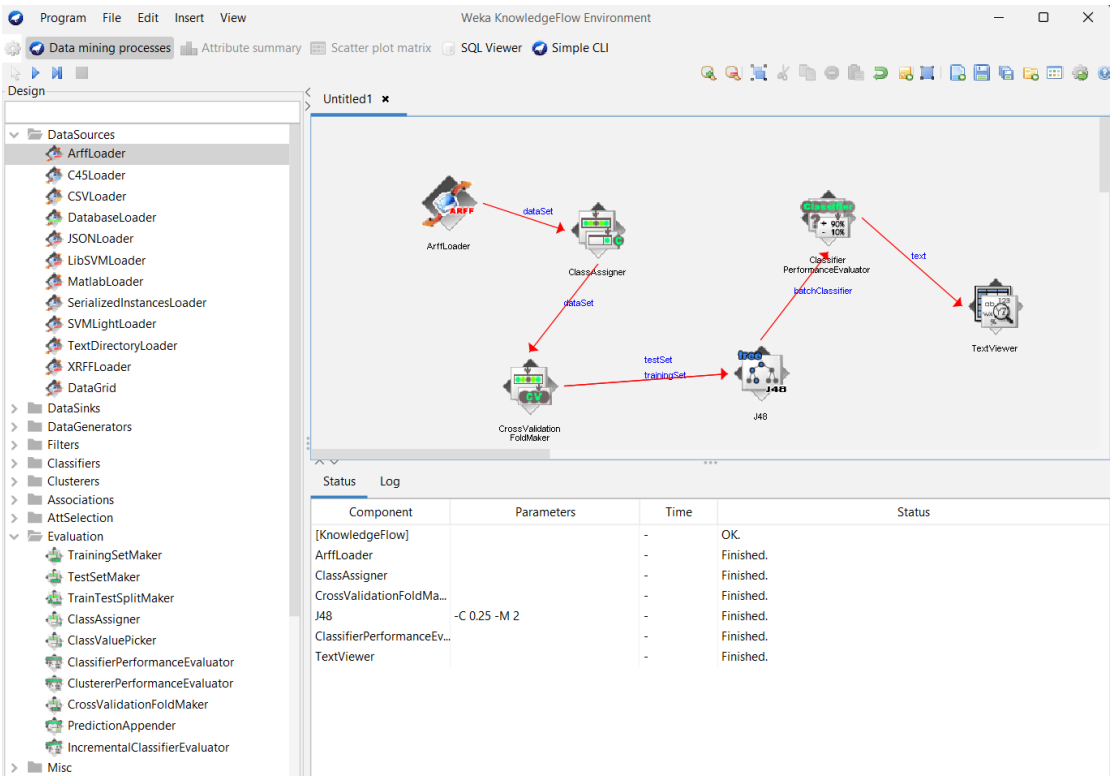
Experimenter: Μέσω του γραφικού περιβάλλοντος του Experimenter, έχουμε τη δυνατότητα δημιουργίας, εκτέλεσης και ανάλυσης πειραμάτων. Ένα πείραμα μπορεί να περιλαμβάνει την εκτέλεση περισσοτέρων του ενός αλγορίθμων σε ένα σύνολο δεδομένων και την ανάλυσή του για την εύρεση του στατιστικά καλύτερου αλγορίθμου. Γίνεται και παρουσιάζεται αξιολόγηση των αποτελεσμάτων των μεθόδων (εικόνα 22). Ουσιαστικά πρόκειται για το περιβάλλον μέσω του οποίου αναζητάται η απάντηση στην ερώτηση: «Ποια είναι η βέλτιστη μέθοδος και ποιες οι βέλτιστες παράμετροι για το συγκεκριμένο πρόβλημα;». Φυσικά, τα παραπάνω μπορούν να υλοποιηθούν και στον Explorer, αλλά μέσω του Experimenter η διαδικασία αυτοματοποιείται και είναι ευκολότερη η χρήση των φίλτρων και των κατηγοριοποιητών και συγκρίνονται ευκολότερα τα μοντέλα με τη συνοπτική παρουσίαση των αποτελεσμάτων τους. Το περιβάλλον του Experimenter μπορεί να χρησιμοποιηθεί για ερευνητικές εργασίες όπου εκτελούνται πειράματα διαφόρων μαθησιακών σχημάτων, με χρήση πολλών διαφορετικών σετ δεδομένων και πιθανόν διαφορετικές παραμέτρους.



Εικόνα 22: Η οθόνη αποτελεσμάτων του Experimenter

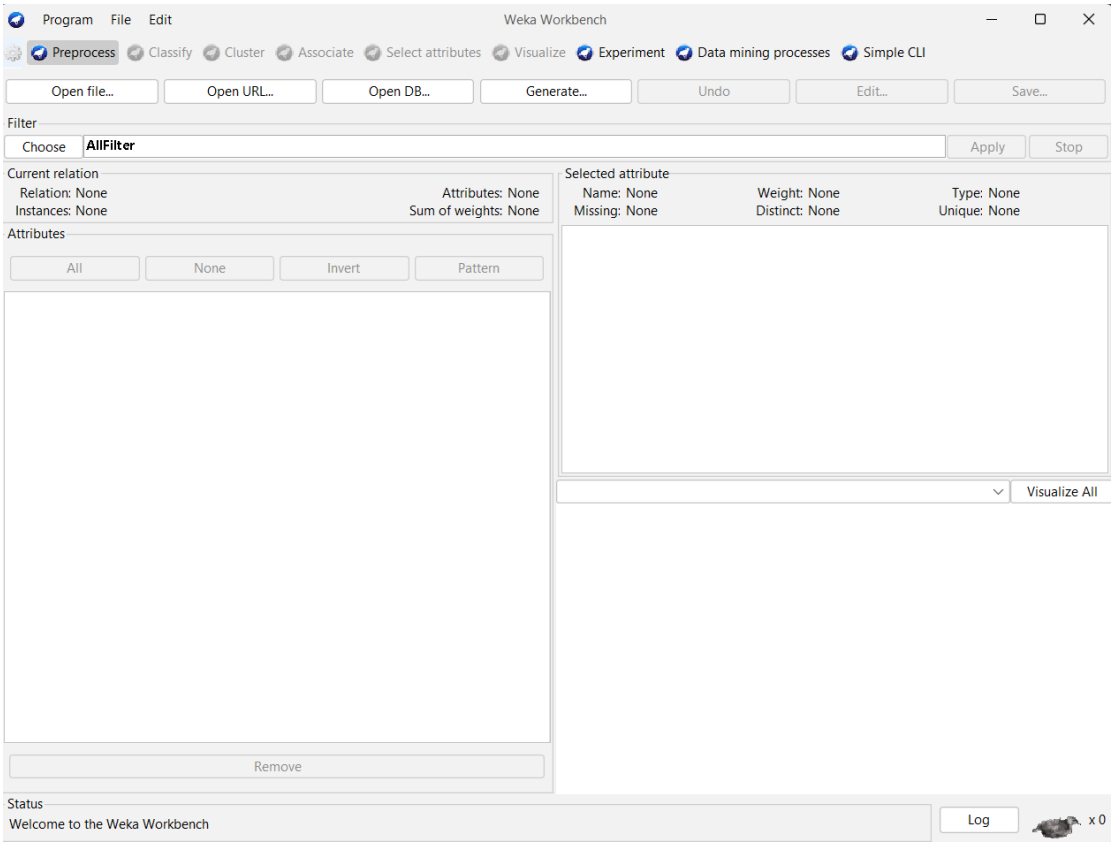
KnowledgeFlow: Στο περιβάλλον του KnowledgeFlow είναι δυνατή η εκτέλεση αλγορίθμων μέσω της σχεδίασης της ροής των δεδομένων. Πιο συγκεκριμένα, όπως φαίνεται και στην εικόνα 23, στο πλαίσιο Design παρέχονται όλα τα στοιχεία (components) που είναι διαθέσιμα προς χρήση. Τα στοιχεία αυτά είναι χωρισμένα σε κατηγορίες όπως DataSources, Filters, Classifiers, Evaluation, Visualization κ.ά. Ο χρήστης μπορεί να εισάγει τα στοιχεία αυτά (κάνοντας κλικ σε όποιο επιθυμεί και ακολούθως κλικάροντας στο πλαίσιο σχεδίασης) και να τα συνδέσει μεταξύ τους, δημιουργώντας με τον τρόπο αυτό την επιθυμητή ροή των

δεδομένων. Η σύνδεση γίνεται με δεξί κλικ στο στοιχείο και επιλογή της κατάλληλης εντολής, ανάλογα με το στοιχείο, και τέλος επιλογή του στοιχείου-στόχου, οπότε και δημιουργείται το κόκκινο βέλος που αναπαριστά τη σύνδεση των στοιχείων. Πρόκειται λοιπόν για ένα περιβάλλον που επιτρέπει την εκτέλεση εργασιών, όπως και στον Explorer, μέσω όμως διαφορετικού περιβάλλοντος. Παρουσιάζει ενδιαφέροντα χαρακτηριστικά που δεν υπάρχουν στον Explorer, όπως η δυνατότητα χειρισμού αυξητικά, δηλαδή όταν αλλάζουν οι τιμές ενός χαρακτηριστικού από την επεξεργασία κάποιου κατηγοριοποιητή – υπάρχουν διαθέσιμοι κατηγοριοποιητές ειδικά για το σκοπό αυτό. Για την αυξητική εκτέλεση πρέπει κάθε στοιχείο του συστήματος να έχει τη δυνατότητα να λειτουργήσει αυξητικά. Ακόμα είναι δυνατή η παράλληλη επεξεργασία ροών δεδομένων, η συγχώνευση φίλτρων για την αποδοτικότερη επεξεργασία των δεδομένων και όλα αυτά σε ένα ξεκάθαρο περιβάλλον. Το περιβάλλον KnowledgeFlow απευθύνεται σε πιο προχωρημένους χρήστες του Weka, που θέλουν να έχουν εικόνα της ροής των δεδομένων και των πληροφοριών μέσα στο σχεδιασμένο σύστημα.



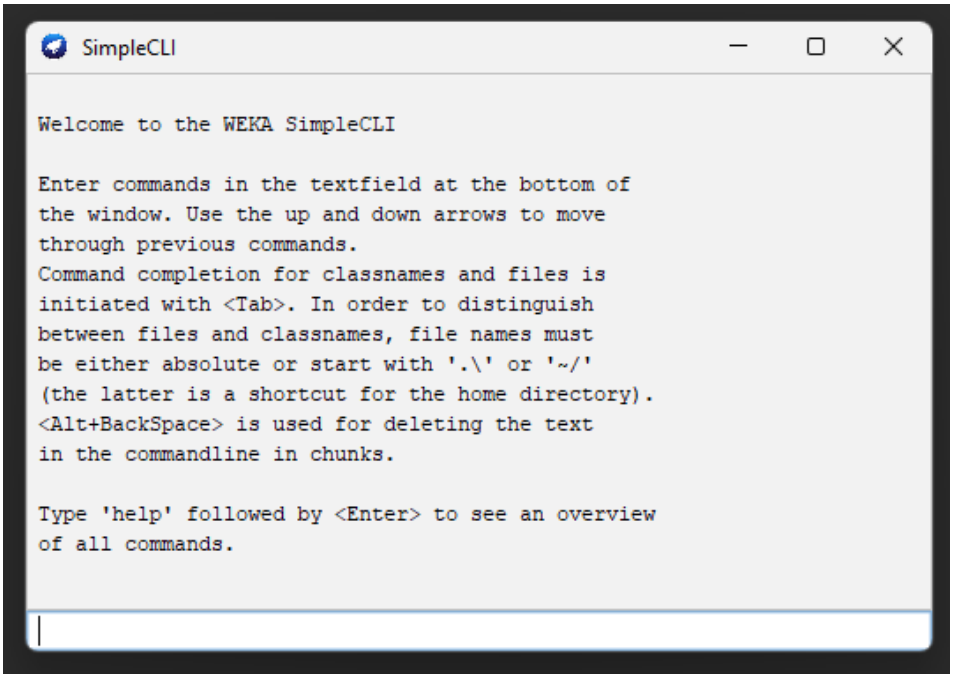
Εικόνα 23: Παράδειγμα χρήσης του KnowledgeFlow

Workbench: Είναι ένα περιβάλλον (εικόνα 24) το οποίο συνδυάζει τα παραπάνω γραφικά περιβάλλοντα σε μια ενιαία διάταξη, διευκολύνοντας την εύκολη και άμεση εναλλαγή μεταξύ τους. Είναι παραμετροποιήσιμος επιτρέποντας στο χρήστη να καθορίσει ποιες εφαρμογές και ποια πρόσθετα θα εμφανίζονται καθώς και τις σχετικές ρυθμίσεις.



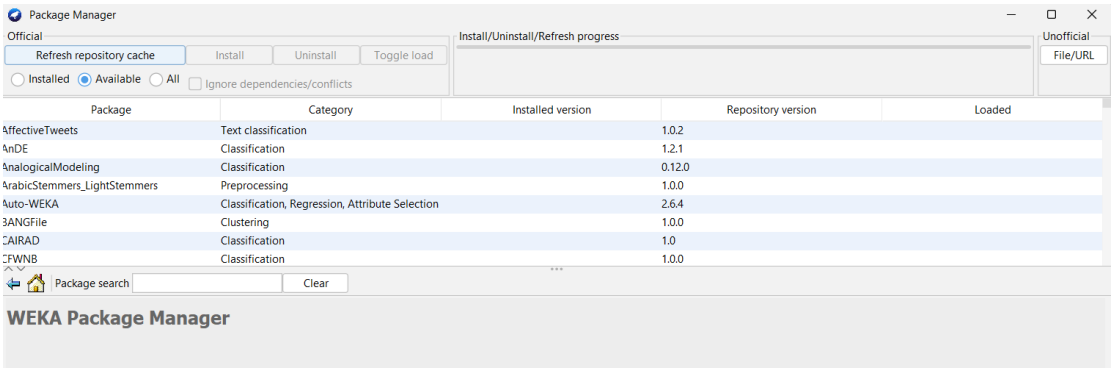
Εικόνα 24: Η κύρια οθόνη του περιβάλλοντος Workbench

SimpleCLI (Command Line Interface): Παρέχει πρόσβαση στο περιβάλλον γραμμής εντολών (εικόνα 25). Μέσω γραμμής εντολών μπορούν να εκτελεστούν όλες οι εντολές και οι αλγόριθμοι του προγράμματος. Η λειτουργικότητα του Simple CLI, δε διαφέρει από αυτή του Explorer και του Experimenter. Στο πλαίσιο διαλόγου που βρίσκεται στο κάτω μέρος του παραθύρου είναι δυνατή η πληκτρολόγηση εντολών για την άμεση εκτέλεσή τους. Απευθύνεται στους χρήστες που είναι εξοικειωμένοι με το Weka και τις εντολές του. Στην ιστοσελίδα https://docs.weka.io/getting-started-with-weka/manage-the-system-using-weka-cli υπάρχουν αναλυτικές πληροφορίες για τις διαθέσιμες εντολές και τη σύνταξή τους.



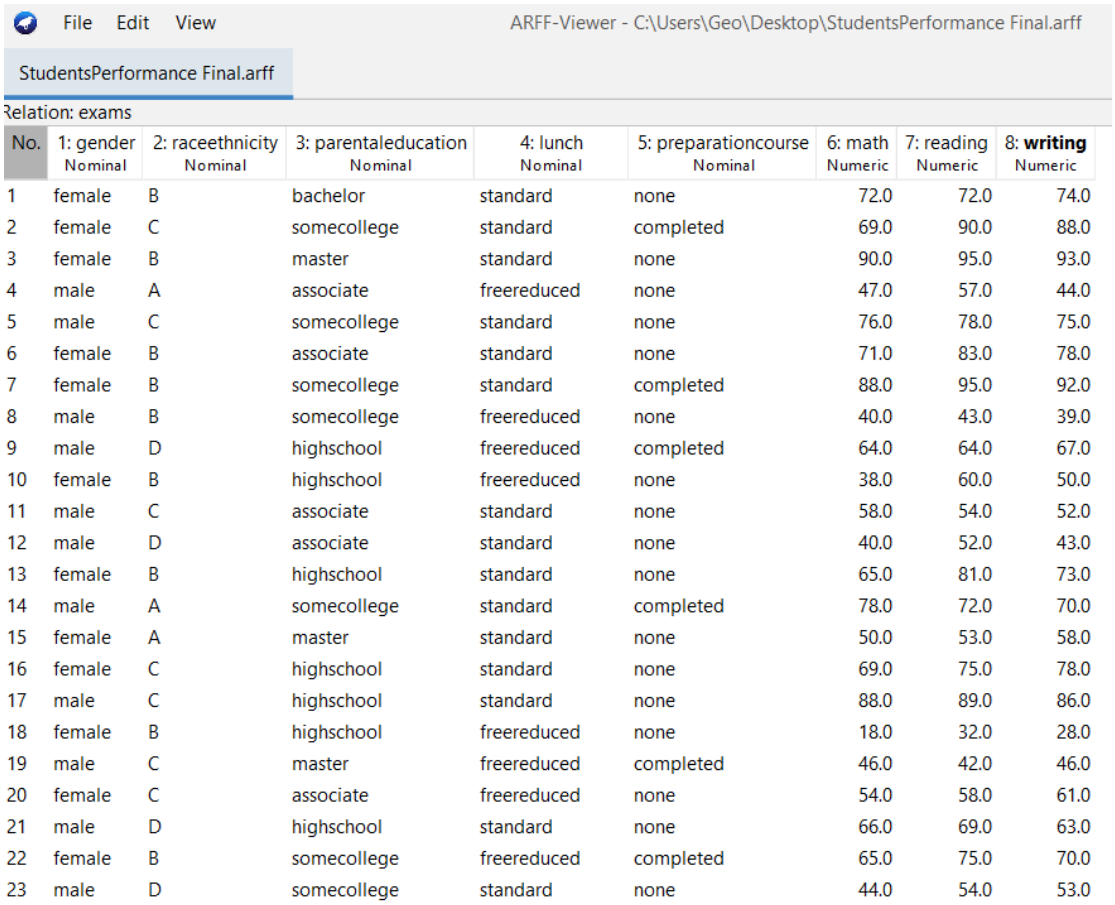
Εικόνα 25:Το περιβάλλον γραμμής εντολών (SimpleCLI) του Weka

Πρέπει ακόμα να σημειωθεί ότι στην κεντρική οθόνη, στο μενού Tools είναι διαθέσιμη η επιλογή Package manager μέσω της οποίας είναι δυνατή η εξερεύνηση και η εγκατάσταση πρόσθετων στο Weka, όπως για παράδειγμα νέων αλγορίθμων (εικόνα 26).



Εικόνα 26: Ο package manager του Weka

Τέλος, πάλι στο μενού Tools της κεντρικής οθόνης υπάρχει και η επιλογή ArffViewer με την οποία είναι δυνατό το άνοιγμα αρχείων διαφόρων μορφών (εικόνα 27), η επεξεργασία τους και η αποθήκευση σε μορφή .arff. Οι υποστηριζόμενες μορφές αρχείων διακρίνονται στην εικόνα 28.

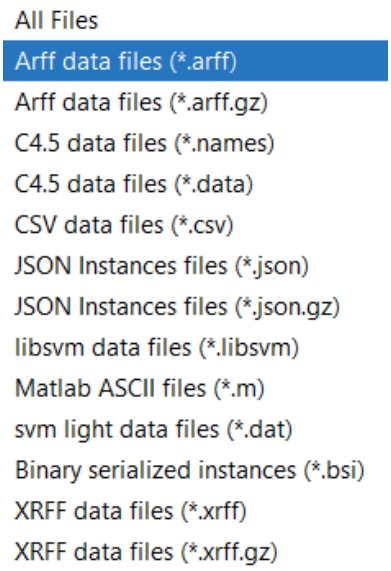

File Edit View ARFF-Viewer - C:\Users\Geo\Desktop\StudentsPerformance Final.arff

StudentsPerformance Final.arff

Relation: exams

No.	1: gender Nominal	2: raceethnicity Nominal	3: parentaleducation Nominal	4: lunch Nominal	5: preparationcourse Nominal	6: math Numeric	7: reading Numeric	8: **writing** Numeric
1	female	B	bachelor	standard	none	72.0	72.0	74.0
2	female	C	somecollege	standard	completed	69.0	90.0	88.0
3	female	B	master	standard	none	90.0	95.0	93.0
4	male	A	associate	freereduced	none	47.0	57.0	44.0
5	male	C	somecollege	standard	none	76.0	78.0	75.0
6	female	B	associate	standard	none	71.0	83.0	78.0
7	female	B	somecollege	standard	completed	88.0	95.0	92.0
8	male	B	somecollege	freereduced	none	40.0	43.0	39.0
9	male	D	highschool	freereduced	completed	64.0	64.0	67.0
10	female	B	highschool	freereduced	none	38.0	60.0	50.0
11	male	C	associate	standard	none	58.0	54.0	52.0
12	male	D	associate	standard	none	40.0	52.0	43.0
13	female	B	highschool	standard	none	65.0	81.0	73.0
14	male	A	somecollege	standard	completed	78.0	72.0	70.0
15	female	A	master	standard	none	50.0	53.0	58.0
16	female	C	highschool	standard	none	69.0	75.0	78.0
17	male	C	highschool	standard	none	88.0	89.0	86.0
18	female	B	highschool	freereduced	none	18.0	32.0	28.0
19	male	C	master	freereduced	completed	46.0	42.0	46.0
20	female	C	associate	freereduced	none	54.0	58.0	61.0
21	male	D	highschool	standard	none	66.0	69.0	63.0
22	female	B	somecollege	freereduced	completed	65.0	75.0	70.0
23	male	D	somecollege	standard	none	44.0	54.0	53.0

Εικόνα 27: Το σύνολο δεδομένων όπως εμφανίζεται στην εφαρμογή ArffViewer



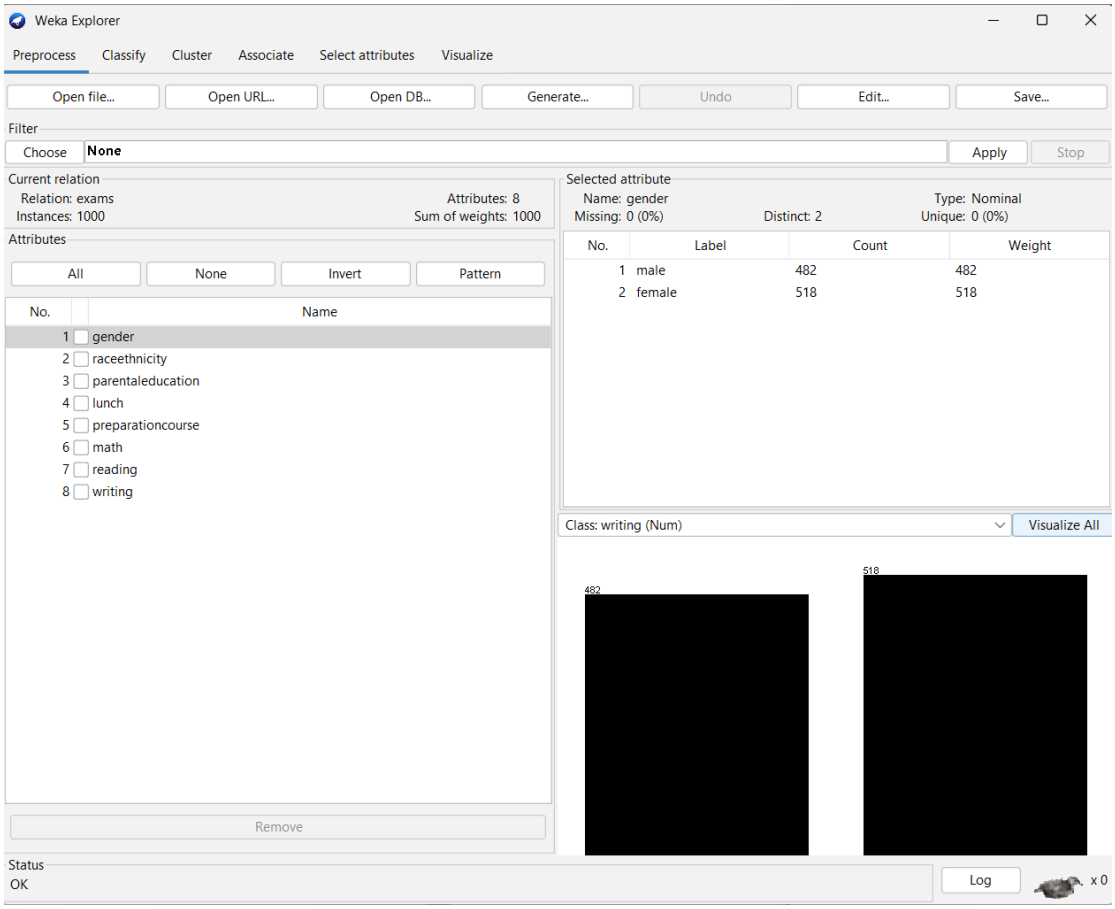
Εικόνα 28: Υποστηριζόμενοι τύποι αρχείων του ArffViewer

8.5 Ο Explorer του Weka.

Το περιβάλλον του Explorer φαίνεται στην εικόνα 21 και αποτελείται από έξι καρτέλες. Εδώ γίνεται η εισαγωγή των δεδομένων, εκτελούνται οι αλγόριθμοι και εμφανίζονται τα αποτελέσματα. Το περιβάλλον επιτρέπει την επεξεργασία των δεδομένων παρέχοντας ισχυρά εργαλεία στο χρήστη καθώς και την εκτέλεση πολλαπλών αλγορίθμων στην ίδια καρτέλα. Αρχικά είναι ενεργή μόνο η καρτέλα Preprocess ώστε να γίνει η εισαγωγή των δεδομένων, ενώ, μετά την εισαγωγή τους, ενεργοποιούνται οι υπόλοιπες καρτέλες. Τα δεδομένα που εισάγονται αποθηκεύονται από τον Explorer στη μνήμη. Έτσι δεν ενδείκνυται η χρήση του περιβάλλοντος του Explorer για προβλήματα στα οποία το σύνολο δεδομένων είναι μεγάλο.

8.5.1 Η καρτέλα Preprocess

Η καρτέλα Preprocess (εικόνα 29) αποτελείται από διάφορα πλαίσια πληροφοριών που είναι σχεδιασμένα ώστε να παρέχουν μια γενική εικόνα των δεδομένων. Επιπλέον, είναι διαθέσιμα διάφορα εργαλεία για την ανάλυση και επεξεργασία των δεδομένων.



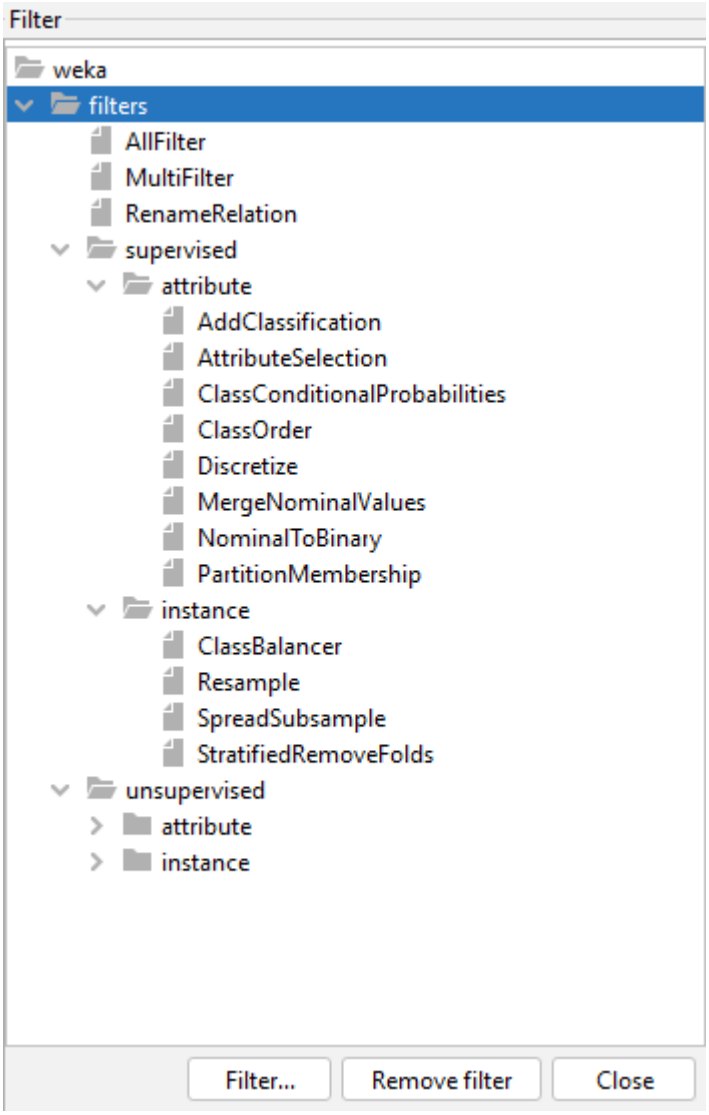
Εικόνα 29: Η καρτέλα Preprocess

Τα δεδομένα μπορούν να εισαχθούν μέσω τεσσάρων διαφορετικών επιλογών:

- Open file: Εισάγονται τα δεδομένα από αρχείο κατάλληλης μορφής.
- Open URL: Εισάγονται δεδομένα από μια διεύθυνση στο Διαδίκτυο.
- Open DB: Εισάγονται δεδομένα από μια βάση δεδομένων.
- Generate: Δημιουργούνται δεδομένα με χρήση αλγορίθμων που επιλέγει ο χρήστης.

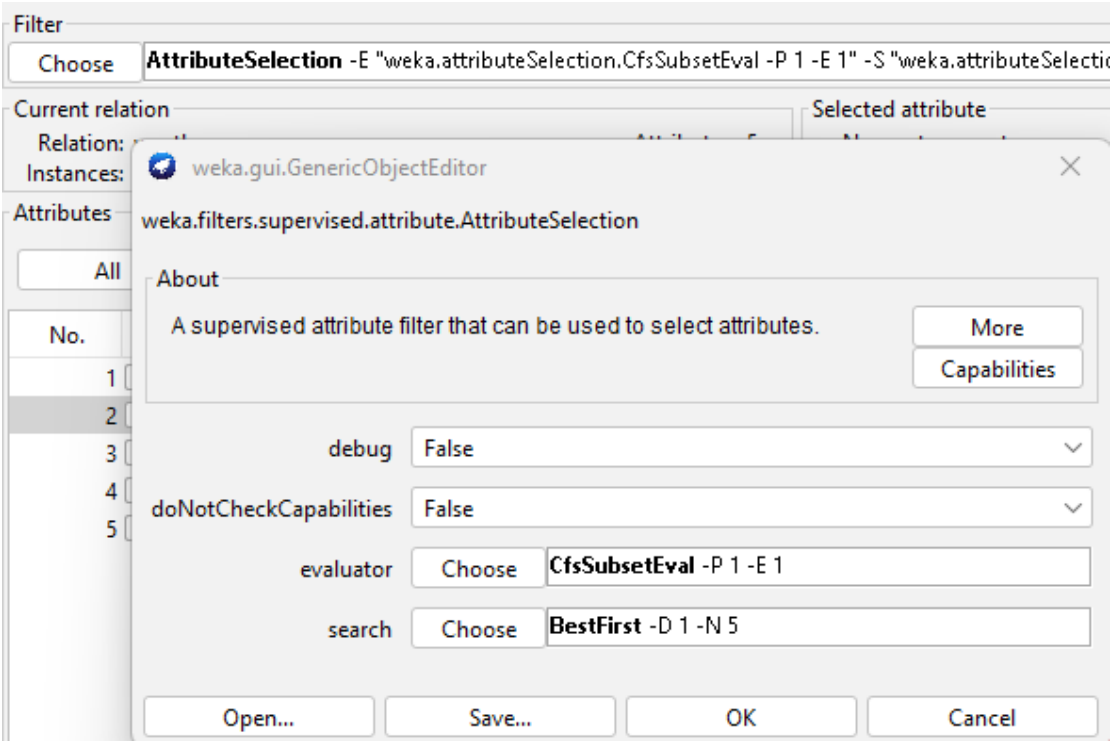
Μέσω της επιλογής Edit είναι δυνατή η επεξεργασία των δεδομένων. Εκεί μπορούν να τροποποιηθούν τιμές των χαρακτηριστικών, να προστεθούν ή να διαγραφούν στιγμιότυπα (instances). Το σετ δεδομένων μπορεί να αποθηκευτεί ως ξεχωριστό αρχείο μέσω της επιλογής Save.

Με τη χρήση φίλτρων τα δεδομένα μπορούν να τροποποιηθούν ώστε να έχουν μια περισσότερο κατανοητή και χρήσιμη μορφή. Φίλτρα ονομάζονται τα εργαλεία που παρέχει το Weka για την αυτοματοποιημένη προεπεξεργασία των δεδομένων και αφορούν δυνατότητες κανονικοποίησης αριθμητικών τιμών (normalization), διακριτοποίησης αριθμητικών τιμών (discretize), συγχώνευσης ονομαστικών πεδίων (merge) κ.ά. Τα διαθέσιμα φίλτρα του Weka είναι χωρισμένα σε κατηγορίες και φαίνονται στην εικόνα 30.



Εικόνα 30: Το μενού επιλογών των διαθέσιμων φίλτρων του Weka

Με τη χρήση του φίλτρου Attribute selection υλοποιείται η αυτόματη επιλογή των σημαντικών χαρακτηριστικών από τα δεδομένα ώστε να εφαρμοστούν σε αυτά οι αλγόριθμοι που θα επιλεγούν. Η επιλογή των σημαντικών χαρακτηριστικών μπορεί να γίνει με διάφορες μεθόδους. Η εξ' ορισμού μέθοδος του Weka είναι η CFS (Correlation-Based Feature Selection). Για την πρόσβαση στο παράθυρο επιλογών της μεθόδου αυτής, ο χρήστης πρέπει αφού επιλέξει το ομώνυμο φίλτρο, να κάνει κλικ στο πλαίσιο Attribute Selection, οπότε εμφανίζεται το παράθυρο του Generic Object Editor της εικόνας 31.



Εικόνα 31: Επιλογές του φίλτρου Attribute selection

Μετά τη χρήση του φίλτρου, ο αριθμός των πεδίων έχει μειωθεί και έχουν παραμείνει μόνο τα πεδία που χαρακτηρίζονται σημαντικά.

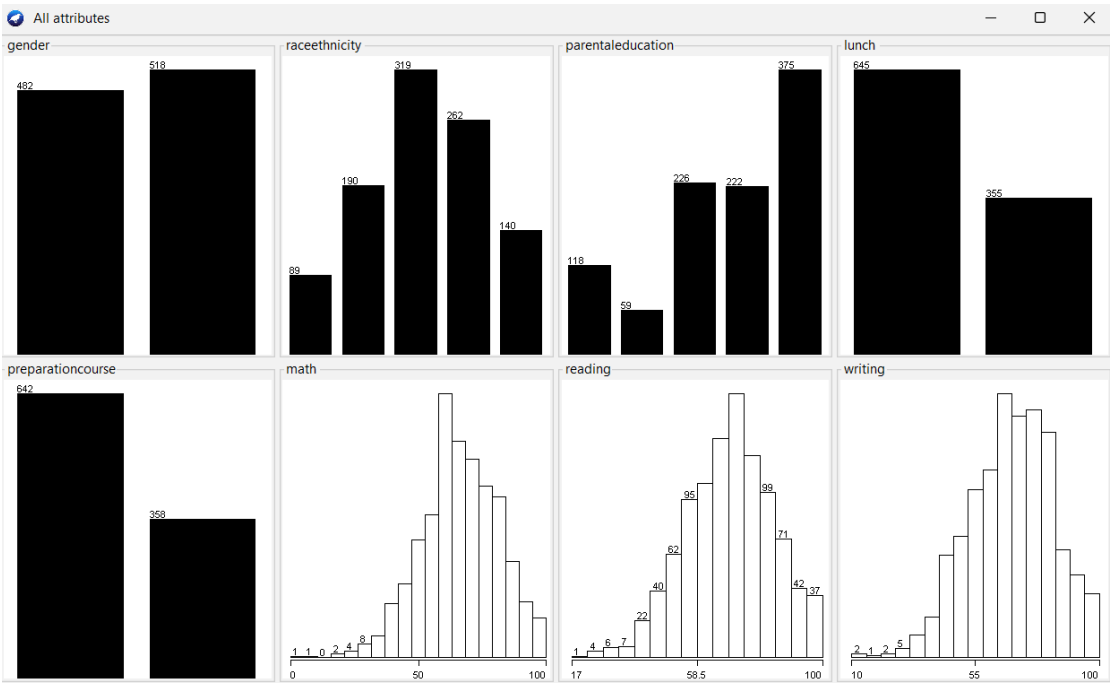
Μετά την φόρτωση του συνόλου δεδομένων εμφανίζονται βασικά στοιχεία των δεδομένων καθώς και σχετικά στατιστικά. Στο αριστερό τμήμα του παραθύρου, στο πλαίσιο Attributes, εμφανίζονται τα πεδία που απαρτίζουν το σύνολο δεδομένων. Είναι δυνατή η επιλογή όλων, η χειροκίνητη επιλογή ή η επιλογή τους βάσει κριτηρίων. Τα επιλεχθέντα πεδία μπορούν να αφαιρεθούν από τα δεδομένα με το πλήκτρο Remove.

Όταν επιλέγεται ένα χαρακτηριστικό, στο πλαίσιο Selected attribute, εμφανίζονται πληροφορίες σχετικές με αυτό:

- Name: το όνομά του,
- Type: ο τύπος του,
- Missing: το ποσοστό των στιγμιότυπων (instances) που απουσιάζουν τιμές δεδομένων,
- Distinct: το πλήθος των διαφορετικών τιμών που περιέχουν τα δεδομένα του επιλεγμένου χαρακτηριστικού,
- Unique: το ποσοστό των στιγμιότυπων των δεδομένων που έχουν μοναδική τιμή.

Ακόμα, εμφανίζονται οι τιμές που λαμβάνει και το πλήθος τους. Αν το πεδίο είναι αριθμητικού (numeric) τύπου εμφανίζονται επιπρόσθετα η ελάχιστη και η μέγιστη τιμή του, η μέση τιμή και η τυπική απόκλιση. Αν το πεδίο είναι ονομαστικού (nominal) τύπου εμφανίζονται οι τιμές του καθώς και το πλήθος των δεδομένων που λαμβάνει την τιμή αυτή.

Στο κάτω μέρος του δεξιού τμήματος του παραθύρου, ο χρήστης μπορεί να επιλέξει μία από τις κλάσεις (πεδία) των δεδομένων, ή μπορεί να επιλέξει no class. Το χαρακτηριστικό κλάσης (class) είναι εξ' ορισμού τελευταίο στη λίστα. Το Weka παρουσιάζει μέσω ραβδογραμμάτων, τις τιμές του πεδίου που έχει επιλεγεί στο πλαίσιο Attributes. Για κάθε τιμή, παρουσιάζεται το πλήθος των δεδομένων που τη λαμβάνουν, ως προς την επιλεχθείσα κλάση.
Με την επιλογή Visualize All, παρουσιάζεται η κατανομή όλων των τιμών των μεταβλητών ως προς την κλάση που έχει επιλεγεί στο drop down μενού Class. (εικόνα 32).

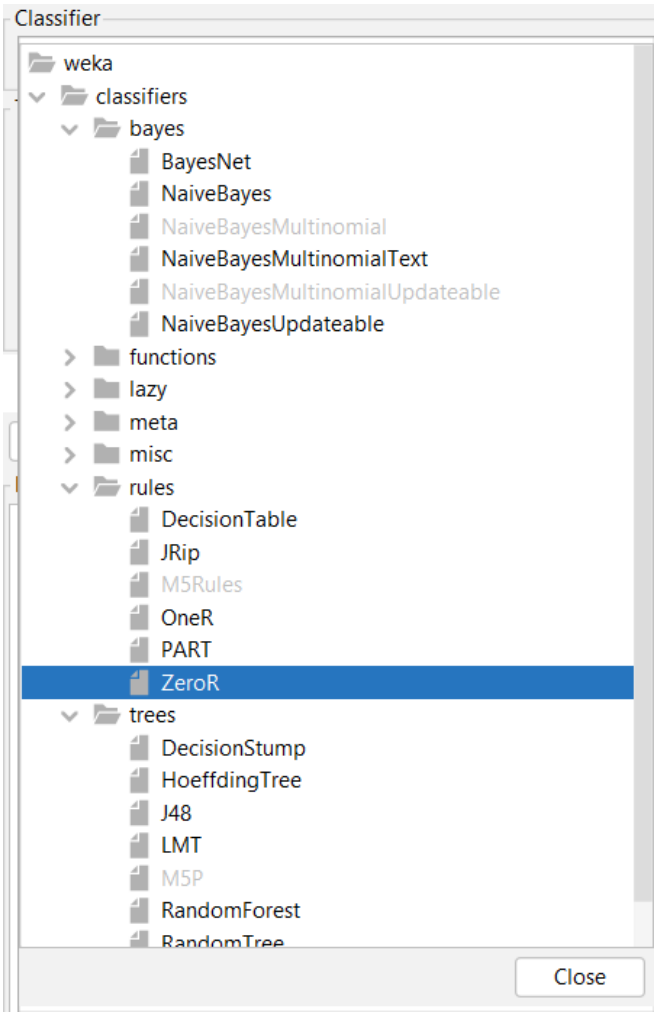


Εικόνα 32: Οπτικοποίηση του συνόλου δεδομένων

Με τη βοήθεια της καρτέλας Preprocess του Weka explorer, επιτελείται η διερευνητική ανάλυση δεδομένων (Exploratory Data Analysis / EDA). Αποτελεί ένα ουσιαστικό βήμα για την εξερεύνηση των δεδομένων, την κατανόηση της δομή τους, τη διαμόρφωση της αρχικής υπόθεσης και τον εντοπισμό ακραίων τιμών και ανωμαλιών (Thankachan 2017).

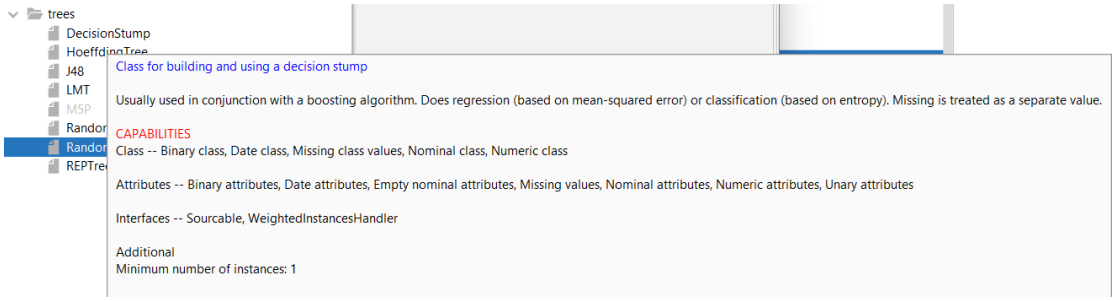
8.5.2 Η καρτέλα Classify

Η δεύτερη καρτέλα του Weka Explorer είναι η καρτέλα Classify. Με το πλήκτρο Choose, από το εμφανιζόμενο πλαίσιο διαλόγου, ο χρήστης επιλέγει τον αλγόριθμο κατηγοριοποίησης που θα εφαρμοστεί στα δεδομένα που έχουν ήδη φορτωθεί. Με τις μεθόδους κατηγοριοποίησης γίνεται πρόβλεψη τιμών μιας ποσότητας αριθμητικού ή ονομαστικού τύπου.
Οι διαθέσιμοι αλγόριθμοι είναι χωρισμένοι σε κατηγορίες και παρουσιάζονται σε δενδροειδή μορφή, όπως φαίνεται στην εικόνα 33.



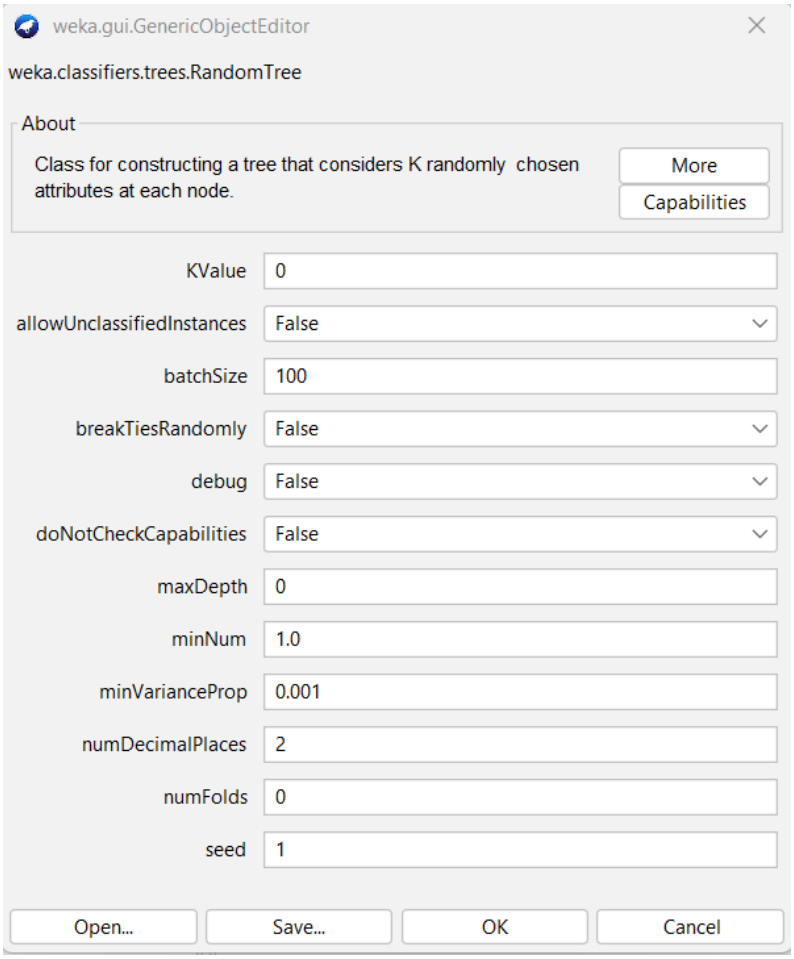
Εικόνα 33: Οι αλγόριθμοι κατηγοριοποίησης του Weka

Περιλαμβάνονται όλες οι γνωστές μέθοδοι κατηγοριοποίησης, όπως τα Δένδρα Αποφάσεων, τα Νευρωνικά Δίκτυα, τα Δίκτυα Bayes, οι Μηχανές Διανυσμάτων Υποστήριξης κ.ά. Αν ο χρήστης τοποθετήσει το δείκτη του ποντικιού πάνω από το όνομα ενός αλγορίθμου χωρίς να κάνει κλικ, εμφανίζεται ένα πλαίσιο με πληροφορίες σχετικές με το συγκεκριμένο αλγόριθμο, όπως φαίνεται για παράδειγμα την εικόνα 34.



Εικόνα 34: Πλαίσιο πληροφοριών για τον αλγόριθμο DecisionStump.

Αφού επιλεγεί ένας αλγόριθμος με το πλήκτρο Choose, ακριβώς δίπλα στο πλαίσιο εμφανίζεται το όνομά του. Με κλικ στο όνομα εμφανίζεται το παράθυρο της εικόνας 35 όπου μπορούν να οριστούν οι παράμετροι εκτέλεσης του αλγορίθμου.



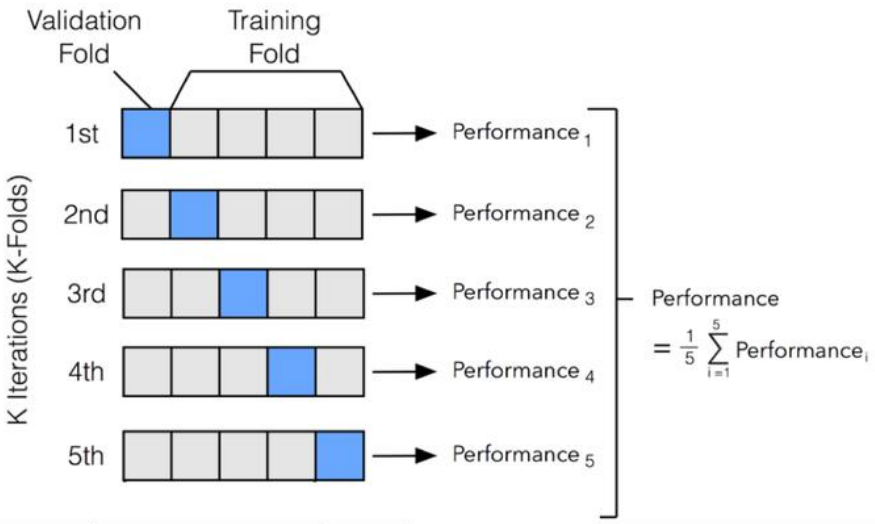
Εικόνα 35: Οι παράμετροι του αλγορίθμου RandromTree

Στη συνέχεια και πριν την εφαρμογή του αλγορίθμου, ο χρήστης ορίζει τη μέθοδο αξιολόγησης του κατηγοριοποιητή στο πλαίσιο Test Options, και στο αναδυόμενου μενού που βρίσκεται ακριβώς κάτω, επιλέγει το πεδίο της κλάσης. Η έννοια της αξιολόγησης του κατηγοριοποιητή σχετίζεται με την ικανότητα του παραγόμενου μοντέλου να κατηγοριοποιήσει, δηλαδή να προβλέψει, σωστά την κλάση άγνωστων παρατηρήσεων. Έχει μεγάλη σημασία ο καθορισμός της ακρίβειας ενός μοντέλου επιπρόσθετα διότι επιτρέπει τη σύγκριση διαφορετικών μοντέλων μεταξύ τους και την επιλογή του αποτελεσματικότερου.

Στο πλαίσιο Test options επιλέγεται η μέθοδος αξιολόγησης του κατηγοριοποιητή. Είναι διαθέσιμες οι παρακάτω τέσσερεις επιλογές, όπως διακρίνονται και στην εικόνα 37:

- Use training set: Η επίδοση του μοντέλου υπολογίζεται με χρήση του συνόλου εκπαίδευσης.

- Supplied test set: Η επικύρωση γίνεται με χρήση ενός διαφορετικού συνόλου δεδομένων.
- Cross-validation: Η επικύρωση γίνεται με την ομώνυμη μέθοδο της «Διασταυρωμένης Επικύρωσης», ορίζοντας το πλήθος των τμημάτων στο πλαίσιο Folds. Με τη μέθοδο αυτή το σύνολο δεδομένων χωρίζεται σε π.χ. 5 τυχαία υποσύνολα διαφορετικών παρατηρήσεων. Το ένα υποσύνολο θα χρησιμοποιηθεί ως το σύνολο επικύρωσης (validation set) ενώ τα υπόλοιπα τέσσερα σύνολα θα ενωθούν για να αποτελέσουν το σύνολο εκπαίδευσης (training set). Η διαδικασία αυτή επαναλαμβάνεται 5 φορές χρησιμοποιώντας κάθε φορά ένα διαφορετικό υποσύνολο ως σύνολο εκπαίδευσης και τα υπόλοιπα τέσσερα, αφού ενωθούν, ως σύνολο εκπαίδευσης. Η απόδοση του μοντέλου είναι ο μέσος όρος των παραπάνω επιδόσεων (εικόνα 36). Στο Weka, η προεπιλεγμένη τιμή είναι 10 folds.



Εικόνα 36: Cross-validation

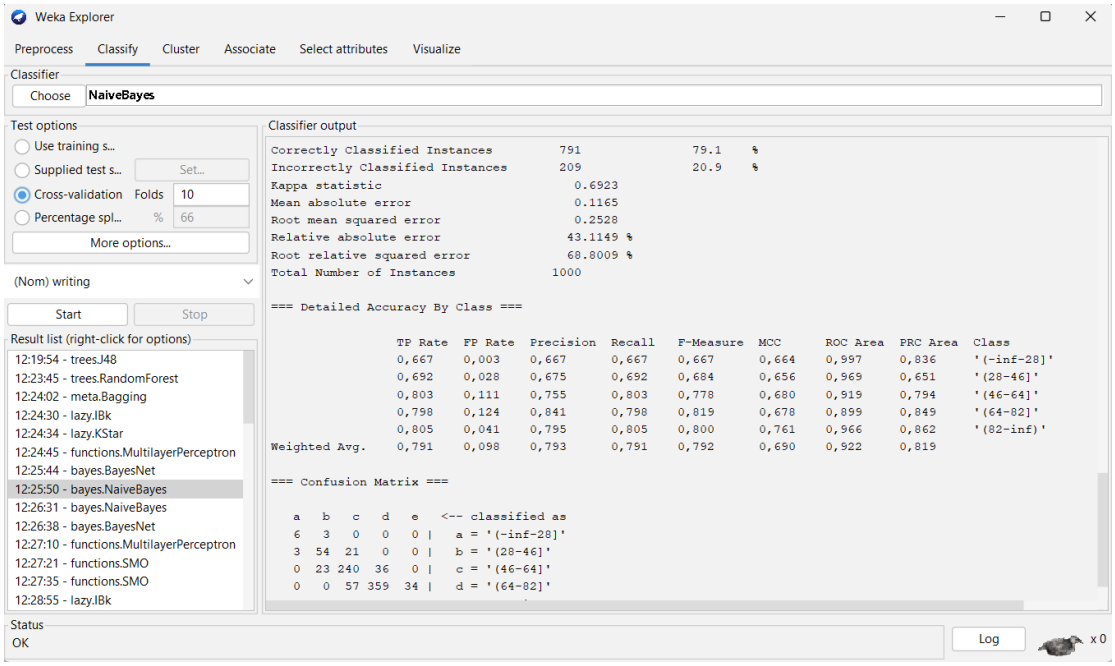
Πηγή:

https://zitaoshen.rbind.io/project/machine_learning/machine-learning-101-cross-vaildation/

- Percentage split: Θα εφαρμοστεί η μέθοδος holdout με την οποία θα χωριστεί το σύνολο των δεδομένων σε δύο υποσύνολα διαφορετικών παρατηρήσεων, με βάση τα ποσοστά που ορίζονται από το χρήστη. Το ένα σύνολο θα χρησιμοποιηθεί ως υποσύνολο εκπαίδευσης (training set) και το άλλο ως υποσύνολο επικύρωσης (validation set ή holdout set).

Η εφαρμογή του επιλεγμένου αλγορίθμου εκτελείται με το κουμπί Start. Ακολούθως στο πλαίσιο Result list εμφανίζεται το μοντέλο και στο πλαίσιο Classifier output εμφανίζονται τα αποτελέσματα του μοντέλου (εικόνα 37). Αυτά περιλαμβάνουν στοιχεία που είναι κοινά για

όλες τους κατηγοριοποιητές, όπως, μεταξύ άλλων, το πλήθος και το ποσοστό των ορθών και εσφαλμένων προβλέψεων, πληροφορίες σχετικά με την αναλυτική ακρίβεια ανά κλάση καθώς και ο Πίνακας Σύγχυσης. Ανάλογα με την επιλεχθείσα μέθοδο κατηγοριοποίησης εμφανίζονται επιπλέον πληροφορίες για στοιχεία και παραμέτρους που την αφορούν.



Εικόνα 37: Αποτελέσματα εκτέλεσης του Naïve Bayes

Τα αποτελέσματα στο πλαίσιο Classifier output είναι χωρισμένα σε τρία μέρη Run information, Classifier model (full training set) και Summary. Στο Run information εμφανίζονται οι πληροφορίες που αφορούν την εκτελεσθείσα κατηγοριοποίηση, όπως το όνομα του Scheme, Relation, Instances, Attributes που υπάρχουν στο αρχείο των δεδομένων και Test mode για την μέθοδο επικύρωσης που χρησιμοποιήθηκε. Στο Classifier model (full training set) αναπαριστάται το μοντέλο κατηγοριοποίησης. Στο Summary εμφανίζονται στατιστικά για την πρόβλεψη της κλάσης. Πολύ σημαντική πληροφορία είναι η Correctly Classified Instances, ως πλήθος και ως ποσοστό, καθώς πρόκειται για την απόδοση του αλγορίθμου. Στο Detailed Accuracy By Class, εμφανίζεται λεπτομερής αναφορά ανά κλάση για την ακρίβεια πρόβλεψης του κατηγοριοποιητή, ενώ ακολουθεί ο Πίνακας Σύγχυσης (Confusion Matrix).

Πρέπει ακόμα να σημειωθεί πως μπορούν να εφαρμοστούν περισσότερες από μία μέθοδοι, τα μοντέλα των οποίων εμφανίζονται στο πλαίσιο «Result list», απλοποιώντας την εναλλαγή μεταξύ των μοντέλων και την σύγκρισή τους. Όπως αναφέρει και ο τίτλος του πλαισίου («Result list – right-click for options»), κάνοντας δεξί κλικ στο όνομα ενός μοντέλου εμφανίζονται επιπλέον επιλογές και δυνατότητες που, και αυτές, διαφέρουν ανάλογα την

κατηγορία του μοντέλου. Χαρακτηριστικά αναφέρουμε πως μπορεί να γίνει οπτική αναπαράσταση του μοντέλου, όπως π.χ. του Δένδρου Αποφάσεων.

8.5.2.1 Κριτήρια Εκτίμησης Αλγορίθμων

Όπως σημειώθηκε παραπάνω, στο πλαίσιο των αποτελεσμάτων της καρτέλας Classify εμφανίζονται πληροφορίες σχετικά με την αποτελεσματικότητα του μοντέλου, που αναλύονται ακολούθως:

- Πίνακας Σύγχυσης (Confusion Matrix)

Είναι ένας δισδιάστατος πίνακας, οι στήλες του οποίου αντιστοιχούν στις προβλέψεις και οι γραμμές στις πραγματικές τιμές κλάσης. Οι τιμές των κελιών είναι οι αληθινές θετικές, οι αληθινές αρνητικές, οι ψευδείς θετικές και οι ψευδείς αρνητικές προβλέψεις (Κύρκος 2015).
Τα δυνατά αποτελέσματα πρόβλεψης κατηγοριοποίησης των παραδειγμάτων (instances) είναι: True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN). Ο Πίνακας Σύγχυσης περιλαμβάνει τα αποτελέσματα αυτά (εικόνα 38):

	Πρόβλεψη κλάσης		
Πραγματική κλάση		Ναι	Όχι
	Ναι	TP	FN
	Όχι	FP	TN

Εικόνα 38: Πίνακας σύγχυσης

- Precision (Ακρίβεια)

Είναι το ποσοστό των παραδειγμάτων που είναι θετικά και έχουν κατηγοριοποιηθεί ως θετικά, δηλαδή το ποσοστό των αληθώς θετικών μεταξύ όλων των προβλεπόμενων θετικών. Όσο μεγαλύτερη είναι η ακρίβεια, τόσο μικρότερος είναι ο αριθμός των FP. Η ακρίβεια δίνεται από τον τύπο:

$$\text{Ακρίβεια} = \frac{TP}{TP + FP}$$

Ουσιαστικά η ακρίβεια είναι η πιθανότητα, αν επιλέξουμε τυχαία ένα αντικείμενο που έχει κατηγοριοποιηθεί σε μία κλάση, η κατηγοριοποίηση αυτή να είναι σωστή.

- Recall (Ανάκληση)

Το ποσοστό των παραδειγμάτων που κατηγοριοποιούνται σε μια κλάση διά το πραγματικό συνολικό της τάξης. Όσο μεγαλύτερη είναι η ανάκληση, τόσο ελαττώνεται ο αριθμός των θετικών παραδειγμάτων που έχουν δεν έχουν κατηγοριοποιηθεί σωστά. Η ανάκληση δίνεται από τον τύπο:

$$\text{Ανάκληση} = \frac{TP}{TP + FN}$$

Ουσιαστικά ανάκληση είναι η πιθανότητα ένα αντικείμενο της κλάσης, να ταξινομήθηκε όντως στην κλάση αυτή, δηλαδή πόσα από τα θετικά παραδείγματα βρήκε ο κατηγοριοποιητής. Η ανάκληση αυξάνεται, τότε η ακρίβεια μειώνεται και αντίστροφα.

- F-measure

Συνδυάζει τα μέτρα της Ακρίβειας (P) και της Ανάκλησης (R) με ίσα βάρη, όπως φαίνεται στον ακόλουθο τύπο:

$$F(R, P) = \frac{2 \cdot R \cdot P}{R + P}$$

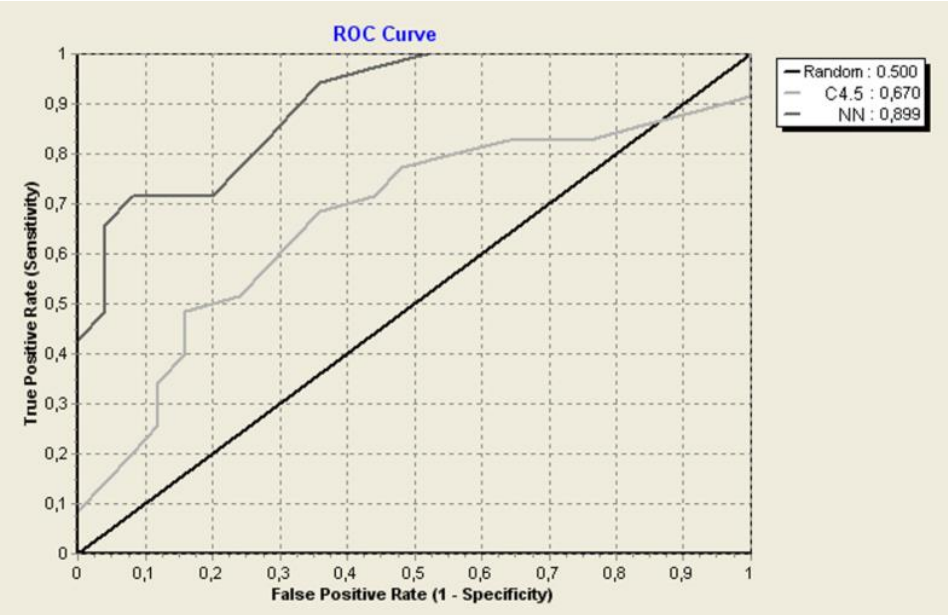
Η τιμή του F-measure έχει μεγάλη βαρύτητα στην αξιολόγηση των αποτελεσμάτων του αλγορίθμου.

- Weighted Avg (Σταθμισμένος Μέσος Όρος)

Υπολογίζεται με χρήση συντελεστών βαρύτητας, οπότε οι τιμές του συνόλου δεδομένων πολλαπλασιάζονται με διαφορετικούς προκαθορισμένους αριθμούς που ονομάζονται βάρη.

- Καμπύλες ROC (Receiver Operating Characteristics)

Ένα ισχυρό μέτρο για την εκτίμηση της ανά κλάση ακρίβειας του κατηγοριοποιητή είναι οι καμπύλες ROC (Κύρκος 2015). Στο διάγραμμα ROC περιγράφεται η σχέση του ποσοστού των σωστά ταξινομημένων θετικών παραδειγμάτων και του ποσοστού των ψευδώς ταξινομημένων θετικών παραδειγμάτων (Ζουμπουλίδης 2012). Είναι η απεικόνιση της απόδοση του κατηγοριοποιητή, ανεξάρτητα της κατανομής της τάξης ή του κόστους σφαλμάτων. Παράδειγμα καμπύλης ROC φαίνεται στην εικόνα 39.



Εικόνα 39: Παράδειγμα καμπύλης ROC

Πηγή: Κύρκος 2015

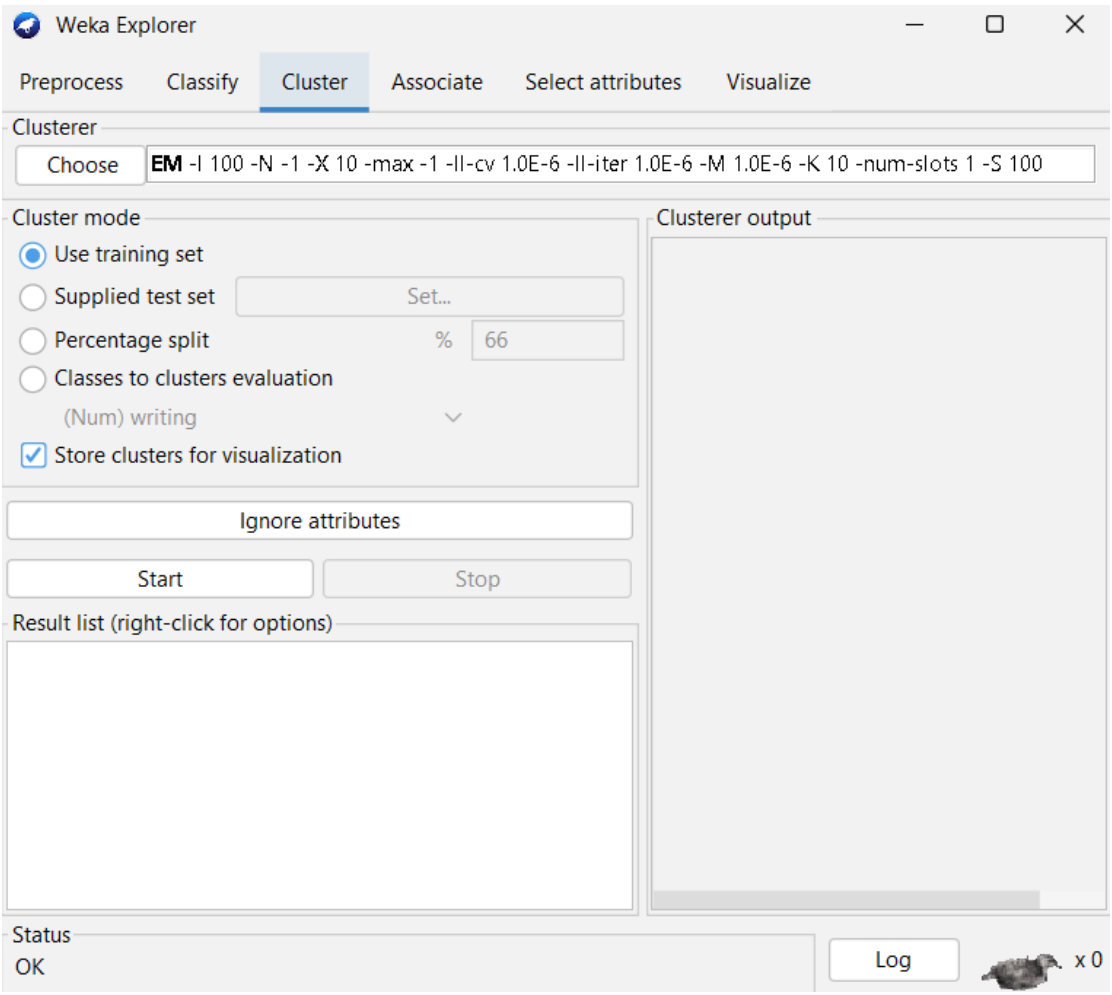
Ο οριζόντιος άξονας x΄x εκφράζει το FPR (False Positive Rate – ποσοστό των ψευδώς ταξινομημένων θετικών παραδειγμάτων) και ο κατακόρυφος άξονας y΄y εκφράζει το TPR (True Positive Rate – ποσοστό των σωστά ταξινομημένων θετικών παραδειγμάτων). Η απόδοση του κατηγοριοποιητή είναι καλύτερη όσο η γραφική παράσταση πλησιάζει τον άξονα y΄y.

Ένα ακόμα μέτρο αξιολόγησης των κατηγοριοποιητών που σχετίζεται με τις καμπύλες ROC, είναι το AUC (Area Under ROC Curve – Εμβαδόν κάτω από την καμπύλη ROC). Το AUC ερμηνεύεται ως το ποσοστό του χώρου ανάμεσα στην καμπύλη και τον οριζόντιο άξονα x΄x. Παίρνει τιμές ανάμεσα στο 0 και το 1, με αύξηση του εμβαδού να συνεπάγεται καλύτερη απόδοση του κατηγοριοποιητή. Τιμή UAC ίση με 0,5 σημαίνει ότι ο κατηγοριοποιητής ταξινομεί τυχαία τα δεδομένα.

Από το μενού επιλογών του δεξιού κλικ στο όνομα του μοντέλου στο πλαίσιο «Result list» ο χρήστης μπορεί να προβάλει και τις καμπύλες ROC του μοντέλου, μέσω της επιλογής «Visualize threshold curve».

8.5.3 Η καρτέλα Cluster

Η καρτέλα Cluster του Explorer απεικονίζεται στην εικόνα 40. Η δομή της είναι παρόμοια με αυτή της καρτέλας Classify που παρουσιάστηκε λεπτομερώς παραπάνω. Αρχικά, με το πλήκτρο Choose γίνεται η επιλογή του αλγορίθμου ομαδοποίησης που επιθυμεί ο χρήστης. Ακολούθως, με κλικ στο όνομα του επιλεγέντος αλγορίθμου γίνεται ρύθμιση των παραμέτρων εκτέλεσης του. Αφού επιλεγεί η μέθοδος επικύρωσης στο πλαίσιο Cluster mode, με το Start εκτελείται ο επιλεγμένος αλγόριθμος ομαδοποίησης και τα αποτελέσματα του εμφανίζονται στο πλαίσιο Clusterer output.



Εικόνα 40: Η καρτέλα Cluster

8.5.4 Η καρτέλα Associate

Η καρτέλα Associate είναι η πιο απλή από τις καρτέλες Classify και Cluster. Περιλαμβάνει αλγορίθμους για την εξόρυξη κανόνων συσχέτισης. Όπως και παραπάνω, επιλέγεται ένας

αλγόριθμος, καθορίζονται οι παράμετροι του, εκτελείται με το κουμπί Start και τα αποτελέσματα εμφανίζονται στο πλαίσιο Associator output.

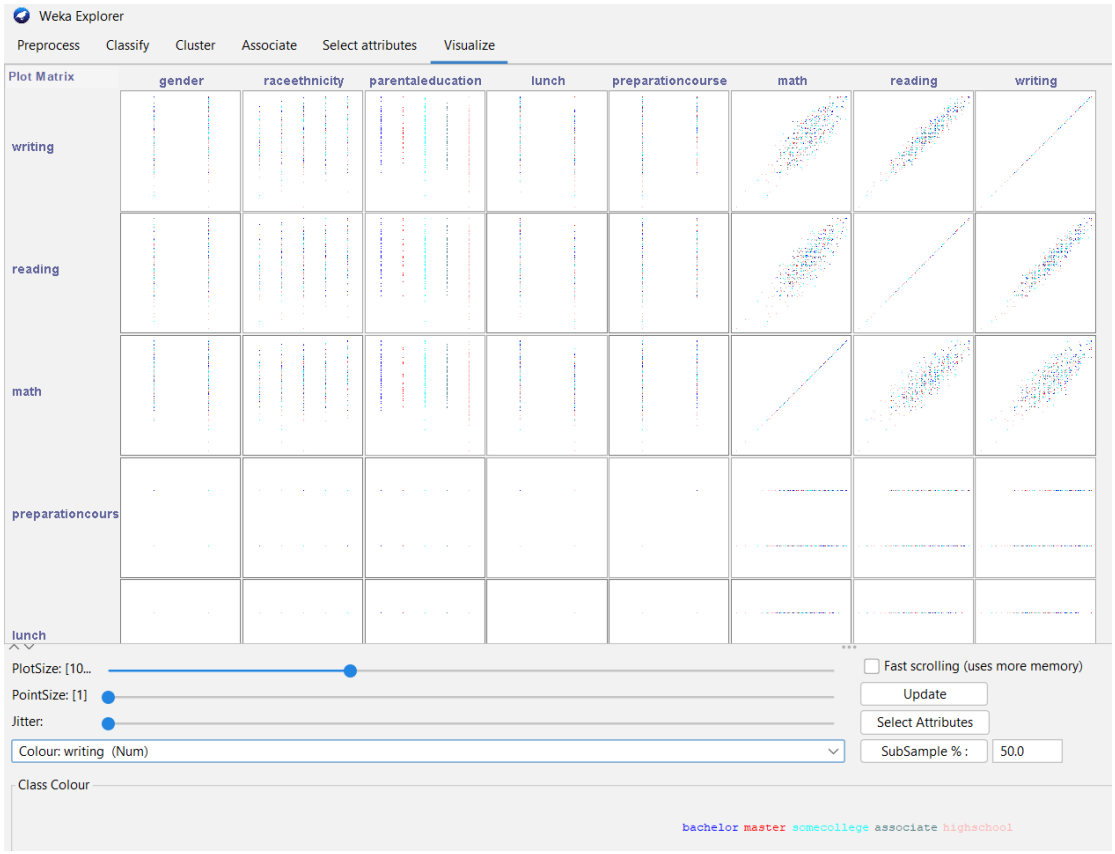
8.5.5 Η καρτέλα Select attributes

Στην καρτέλα Select Atttributes περιλαμβάνονται εργαλεία την εύρεση των σημαντικών χαρακτηριστικών του συνόλου δεδομένων. Ο χρήστης επιλέγει τη μέθοδο αξιολόγησης που επιθυμεί στο πλαίσιο Attribute Evaluator και τη μέθοδο αναζήτησης στο πλαίσιο Search Method. Κάνοντας κλικ σε κάποια από τις επιλεχθείσες μεθόδους εμφανίζονται οι σχετικές πληροφορίες και είναι δυνατή η παραμετροποίησή της. Μετά το κλικ στο κουμπί Start, στο πλαίσιο Attribute selection output εμφανίζονται τα αποτελέσματα της μεθόδου. Τα χαρακτηριστικά που δεν είναι σημαντικά μπορούν στη συνέχεια να διαγραφούν, από την καρτέλα Preprocess, αφού επιλεχθούν, μέσω του κουμπιού Remove. Η ίδια εργασία μπορεί να εκτελεστεί και με τη χρήση του φίλτρου Attribute Selection από την καρτέλα Preprocess (εικόνα 31). Ακόμα πρέπει να οριστεί το χαρακτηριστικό κλάσης. Οι διαθέσιμες μέθοδοι επικύρωσης είναι «Use full training set» και «cross-validation» που έχουν αναλυθεί προηγούμενα.

Ένα συχνά χρησιμοποιούμενο μέτρο αξιολόγησης είναι το μέτρο ReliefF. Αξιολογεί ένα χαρακτηριστικό μέσω επαναλαμβανόμενων δειγματοληψιών ενός παραδείγματος και λαμβάνοντας υπόψη τις τιμές του χαρακτηριστικού για το κοντινότερο παράδειγμα της ίδιας και της αντίθετης τάξης (Kononenko 1994). Σύμφωνα με τη Σερέτη 2020, στην ουσία εκτιμά την αξία ενός χαρακτηριστικού δειγματοληπτικά και λαμβάνοντας υπόψη την τιμή του δεδομένου χαρακτηριστικού για την πλησιέστερη παρουσία της ίδιας και διαφορετικής κλάσης. Συμπερασματικά, τα χαρακτηριστικά που είναι σχετικά με το χαρακτηριστικό κλάσης και έχουν τη μεγαλύτερη επιρροή σε αυτό, αξιολογούνται με υψηλές θετικές τιμές του μέτρου ReliefF, ενώ τα λιγότερο σχετικά λαμβάνουν βάρη με τιμές κοντά στο μηδέν.

8.5.6 Η καρτέλα Visualize

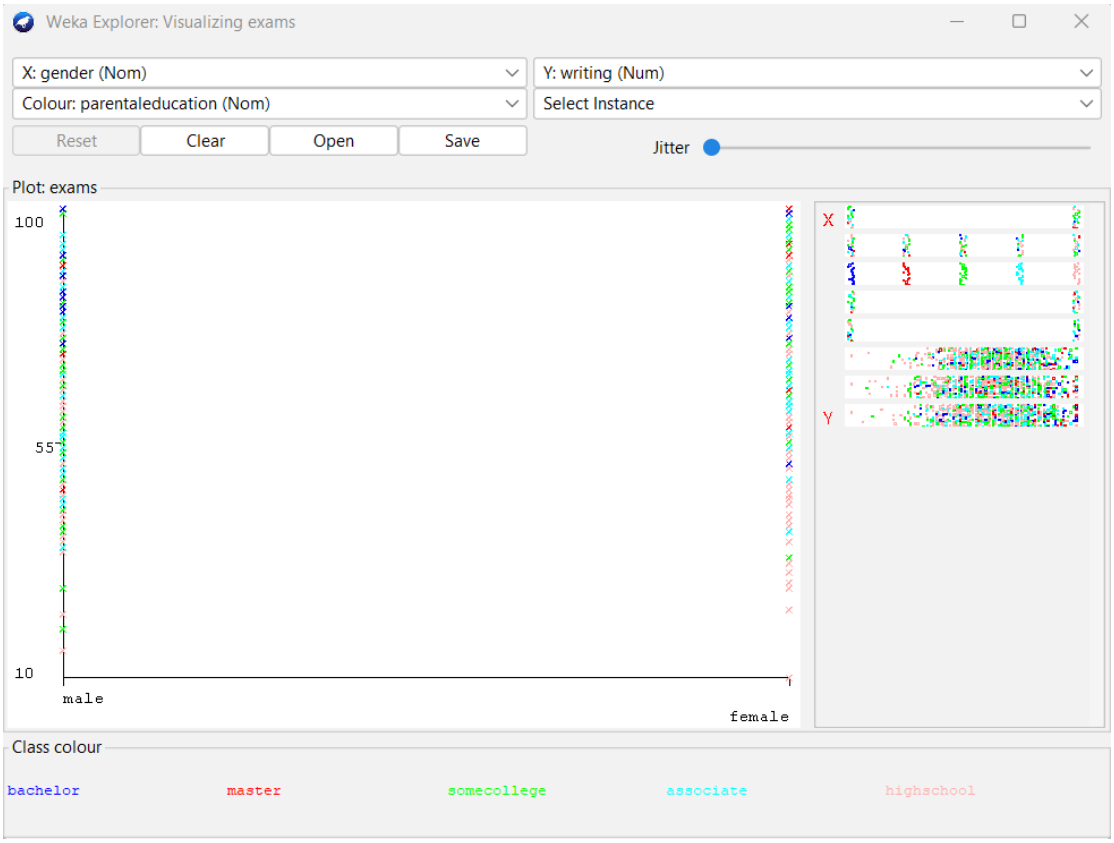
Η τελευταία καρτέλα του Weka Explorer είναι η καρτέλα Visualize. Σε αυτήν είναι διαθέσιμα εργαλεία που αφορούν την οπτικοποίηση των δεδομένων. Έτσι, ο χρήστης έχει μια διαφορετική εποπτεία των δεδομένων καθώς με την οπτικοποίηση παρουσιάζεται η διασπορά των παρατηρήσεων. Στην εικόνα 41 διακρίνονται τα διαγράμματα διασποράς για κάθε ζευγάρι χαρακτηριστικών που υπάρχει στα δεδομένα. Επιλέγοντας κάποιο διαφορετικό χαρακτηριστικό κλάσης και κάνοντας κλικ στο κουμπί Update, τα διαγράμματα χρωματίζονται ανάλογα.



Εικόνα 41: Η καρτέλα Visualize

Να σημειωθεί πως δίνεται δυνατότητα αλλαγή του μεγέθους του πίνακα διαγραμμάτων μέσω της ρύθμισης PlotSize, του μεγέθους των σημείων μέσω του PointSize. Ο πίνακας διαγραμμάτων μπορεί να σχεδιαστεί εκ νέου και να αφορά μόνο τις επιλογές που θα οριστούν αφού γίνει κλικ στο πλαίσιο Select Attributes (η επιλογή γίνεται με Shift+κλικ για συνεχόμενες και Ctrl+κλικ για μη συνεχόμενα χαρακτηριστικά της λίστας).

Με κλικ σε κάποιο από τα διαγράμματα εμφανίζεται μόνο του σε ξεχωριστό παράθυρο. Για παράδειγμα, στην εικόνα 42 υπάρχει το παράθυρο των μεταβλητών gender και writing. Είναι δυνατή η τροποποίηση των μεταβλητών των αξόνων επιλέγοντας διαφορετικές από τα drop down μενού, ενώ στο δεξί πλαίσιο του παραθύρου εμφανίζονται οι κατανομές των παρατηρήσεων για σταθερή μεταβλητή στον άξονα Y και διάφορες μεταβλητές στον άξονα X. Επίσης, είναι δυνατή η μεγέθυνση (zoom) κάποιου τμήματος του διαγράμματος, κάνοντας κλικ στο πλαίσιο Select Instance, επιλέγοντας π.χ. Rectangle, δημιουργώντας το πλαίσιο ενδιαφέροντος σύροντας το ποντίκι και, τέλος, κάνοντας κλικ στο κουμπί Submit.



Εικόνα 42: Διάγραμμα διασποράς δύο μεταβλητών

9. Εκτέλεση αλγορίθμων

9.1 Δεδομένα

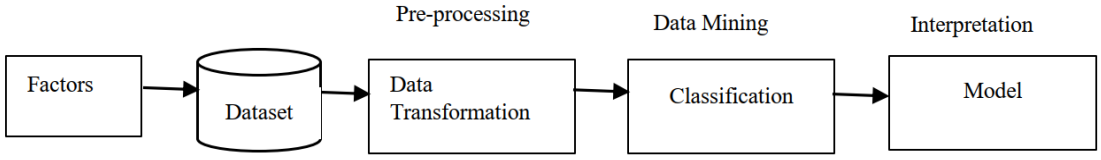
Για την έρευνα που διεξήχθη στην παρούσα εργασία έγινε χρήση του συνόλου δεδομένων που είναι διαθέσιμα στην ιστοσελίδα https://www.kaggle.com/code/spscientist/student-performance-in-exams/data. Τα δεδομένα αυτά αποτελούνται από 1000 παραδείγματα 8 χαρακτηριστικών και αφορούν την απόδοση των μαθητών λυκείου σε εξετάσεις στις Η.Π.Α. Πιο συγκεκριμένα τα χαρακτηριστικά είναι:

- Φύλο (gender), ονομαστικού τύπου, με τιμές male, female.
- Φυλή/Εθνικότητα (race/ethnicity), ονομαστικού τύπου, με τιμές group A, group B, group C, group D, group E.
- Μορφωτικό επίπεδο γονέων (parental education), ονομαστικού τύπου, με τιμές high school, associate, some college, bachelor, master.
- Μεσημεριανό (lunch), ονομαστικού τύπου, με τιμές free/reduced, standard.
- Μαθήματα προετοιμασίας (preparationcourse), ονομαστικού τύπου, με τιμές none, completed.
- Math Score, αριθμητικού τύπου, με τιμές από 0 έως 100.
- Reading Score, αριθμητικού τύπου, με τιμές από 0 έως 100.
- Writing Score, αριθμητικού τύπου, με τιμές από 0 έως 100.

Το αρχείο δεδομένων μετατράπηκε στη μορφή .arff.

9.2 Προεπεξεργασία των δεδομένων

Η προεπεξεργασία (preprocess) των δεδομένων είναι ένα από τα πιο σημαντικά και χρονοβόρα βήματα για την επιτυχή εξόρυξη δεδομένων. Είναι το στάδιο που προηγείται της εφαρμογής των αλγορίθμων κατηγοριοποίησης, καθώς στο στάδιο αυτό τα δεδομένα ελέγχονται και υφίστανται επεξεργασία ώστε είναι πιο ποιοτικά και τα παραγόμενα μοντέλα να είναι έγκυρα και ακριβή. Τα στάδια της διαδικασίας εξόρυξης δεδομένων απεικονίζονται στην εικόνα 43.



Εικόνα 43: Διαδικασία εξόρυξης δεδομένων

Πηγή: Mueen et al 2016

Τα αρχικά συλλεχθέντα δεδομένα δεν είναι χωρίς προβλήματα. Κάποια από τα συνήθη προβλήματα είναι:

- η ύπαρξη χαμένων τιμών, δηλαδή η απουσία τιμών σε κάποια χαρακτηριστικά. Αντιμετωπίζεται με διάφορους τρόπους, όπως π.χ. η διαγραφή όλου του στιγμιοτύπου (δεν προτιμάται καθώς μπορεί να μειώσει αρκετά το πλήθος των στιγμιοτύπων), η συμπλήρωση της τιμής που λείπει με μια σταθερή τιμή (μπορεί να δημιουργήσει θόρυβο) και η συμπλήρωση της τιμής που λείπει με το μέσο όρο των τιμών του χαρακτηριστικού για τα αριθμητικά χαρακτηριστικά ή με την πολυπληθέστερη τιμή για τα ονομαστικά χαρακτηριστικά, μέθοδος που συνήθως προτιμάται από τους ερευνητές.
- η ύπαρξη ακραίων τιμών (outliers) και εσφαλμένων τιμών στα δεδομένα, τα οποία στην περίπτωση αυτή ονομάζονται θορυβώδη. Σε περίπτωση ύπαρξης θορύβου είτε οι αριθμητικές τιμές των δεδομένων αντικαθίστανται από νέες, κατάλληλα προσαρμοσμένες τιμές (π.χ. μέσο όρο), είτε διαγράφονται τα στιγμιότυπα που τις περιέχουν.

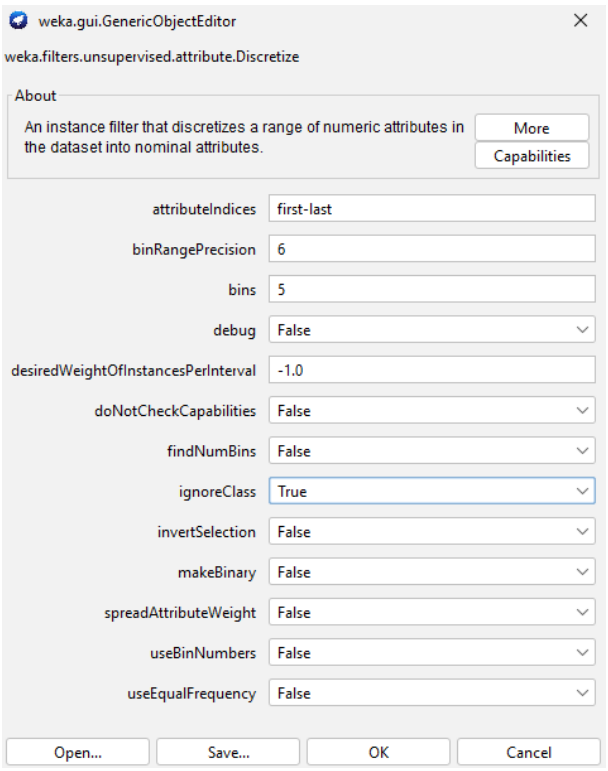
Οι διαδικασίες επίλυσης των παραπάνω προβλημάτων εντάσσονται στον «καθαρισμό δεδομένων» που είναι κομμάτι της προεπεξεργασίας των δεδομένων. Σύμφωνα με τους Fayyad et al 2003, το ποσοστό των δεδομένων που χρήζουν «καθαρισμό» μπορεί να φτάσει έως και 40%.

Ο μετασχηματισμός των δεδομένων είναι ένα ακόμα τμήμα της προεπεξεργασίας. Κατά το μετασχηματισμό τα δεδομένα προσαρμόζονται ώστε να πληρούν τις προϋποθέσεις των αλγορίθμων που θα εφαρμοστούν. Οι δύο συνήθεις διαδικασίες που εφαρμόζονται για το μετασχηματισμό των δεδομένων είναι η διακριτοποίηση (discretization) και η κανονικοποίηση (normalization). Με τη διακριτοποίηση γίνεται μετατροπή των αριθμητικών δεδομένων σε ονομαστικά. Χρησιμοποιείται όταν ο αλγόριθμος κατηγοριοποίησης που πρόκειται να εφαρμοστεί απαιτεί την ύπαρξη ονομαστικών δεδομένων. Με την κανονικοποίηση οι αριθμητικές τιμές μετατρέπονται σε άλλες αριθμητικές τιμές εντός επιθυμητού διαστήματος δοθέντος εύρους. Παραδείγματος χάριν, τα νευρωνικά δίκτυα αποδίδουν καλύτερα αποτελέσματα όταν οι τιμές των δεδομένων ανήκουν στο διάστημα [0,1], οπότε πριν την εφαρμογή του αλγορίθμου μετασχηματίζουμε τα δεδομένα ώστε οι τιμές να βρίσκονται στο διάστημα αυτό.

Ένα ακόμη τμήμα της προεπεξεργασίας των δεδομένων είναι η μείωση του όγκου των δεδομένων. Τα μεγάλου όγκου δεδομένα χρειάζονται περισσότερο χρόνο για την ανάλυσή τους και γενικότερα είναι δυσκολότερα στο χειρισμό τους. Μέσω κατάλληλων διαδικασιών, επιδιώκεται η μείωση του όγκου τους χωρίς να υπάρξει αλλοίωση των αποτελεσμάτων. Ένας τρόπος για την επίτευξη της μείωσης του όγκου των δεδομένων είναι μέσω της «επιλογής χαρακτηριστικών» (feature/attribute selection). Με τον τρόπο αυτό επιλέγονται και

παραμένουν στο σύνολο δεδομένων, μόνο εκείνα τα χαρακτηριστικά που είναι κατάλληλα για την εργασία εξόρυξης γνώσης που θα εκτελεστεί, ενώ τα άλλα εξαιρούνται και δε συμμετέχουν στο σύνολο δεδομένων κατά την εφαρμογή των αλγορίθμων. Τα χαρακτηριστικά αυτά είναι στατιστικά σημαντικά ως προς την μεταβλητή στόχο.

Στο Weka, οι λειτουργίες της προεπεξεργασίας βρίσκονται στην καρτέλα Preprocess του Explorer. Αφού φορτωθεί το αρχείο των δεδομένων, στην καρτέλα Preprocess εφαρμόζεται το φίλτρο Discretize (εικόνα 44). Όταν η τιμή της επιλογής ignoreClass είναι False, το φίλτρο δεν εφαρμόζεται στην τελευταία στήλη των δεδομένων, οπότε αλλάζουμε την τιμή σε True. Με τη χρήση του φίλτρου οι αριθμητικές τιμές μετατράπηκαν σε ονομαστικές και δημιουργήθηκαν 5 κλάσεις (-inf-28], (28-46], (46-64], (64-82], (82-inf]. Οι 5 κλάσεις μπορούν να αντιστοιχισθούν στις επιδόσεις: Πολύ χαμηλή, Χαμηλή, Μέτρια, Καλή, Πολύ καλή.



Εικόνα 44: Οι ρυθμίσεις του φίλτρου Discretize

9.3 Αποτελέσματα αλγορίθμων κατηγοριοποίησης σε δεδομένα 5 κλάσεων

Στη συνέχεια στην καρτέλα Classify, στο πλαίσιο Test Options επιλέχθηκε αρχικά Cross-validation με 10 folds (διασταυρωμένη επικύρωση με δέκα επαναλήψεις) και εφαρμόστηκαν οι αλγόριθμοι. Ακολούθως επιλέχθηκε Percentage split 66% και εφαρμόστηκαν εκ νέου.

Το χαρακτηριστικό κλάσης που επιλέχθηκε ήταν το Writing Score.

Οι αλγόριθμοι που εκτελέστηκαν ήταν οι:

- J48 που αποτελεί την υλοποίηση του C4.5 στο Weka, από τα Δέντρα Απόφασης
- Naïve Bayes (NB) και Bayes Net (BN) από τους κατηγοριοποιητές Bayes
- SMO (Sequential Minimal Optimization) από Μηχανές Διανυσμάτων Υποστήριξης
- Multilayer perceptron (MP) από τα Τεχνητά Νευρωνικά Δίκτυα.
- IBk που αποτελεί την υλοποίηση του k-NN στο Weka, από τους αλγορίθμους Οκνηρής Μάθησης (με Ευκλείδεια απόσταση).

Τα αποτελέσματα που προέκυψαν καταγράφονται στον παρακάτω πίνακα 1:

Αλγόριθμος	**Test Options**	**Συνολική Ακρίβεια %**	**TP Rate**	**Πίνακας σύγχυσης**
J48	Cross validation	79,1	0,778 0,641 0,826 0,791 0,799	a b c d e <-- classified as 7 2 0 0 0 | a = '(-inf-28]' 3 50 25 0 0 | b = '(28-46]' 0 22 247 30 0 | c = '(46-64]' 0 0 61 356 33 | d = '(64-82]' 0 0 0 33 131 | e = '(82-inf)'
	Percentage split	78,8235	0,667 0,625 0,792 0,852 0,667	a b c d e <-- classified as 2 1 0 0 0 | a = '(-inf-28]' 0 10 6 0 0 | b = '(28-46]' 0 5 76 15 0 | c = '(46-64]' 0 0 15 138 9 | d = '(64-82]' 0 0 0 21 42 | e = '(82-inf)'
Naïve Bayes	Cross validation	79,1	0,667 0,692 0,803 0,798 0,805	a b c d e <-- classified as 6 3 0 0 0 | a = '(-inf-28]' 3 54 21 0 0 | b = '(28-46]' 0 23 240 36 0 | c = '(46-64]' 0 0 57 359 34 | d = '(64-82]' 0 0 0 32 132 | e = '(82-inf)'
	Percentage split	80,8824	0,667 0,688 0,823 0,802 0,841	a b c d e <-- classified as 2 1 0 0 0 | a = '(-inf-28]' 0 11 5 0 0 | b = '(28-46]' 0 3 79 14 0 | c = '(46-64]' 0 0 17 130 15 | d = '(64-82]' 0 0 0 10 53 | e = '(82-inf)'

Αλγόριθμος	Test Options	Συνολική Ακρίβεια %	TP Rate	Πίνακας σύγχυσης
Bayes Net	Cross validation	79,2	0,667 0,705 0,803 0,798 0,805	a b c d e <-- classified as 6 3 0 0 0 | a = '(-inf-28]' 2 55 21 0 0 | b = '(28-46]' 0 23 240 36 0 | c = '(46-64]' 0 0 57 359 34 | d = '(64-82]' 0 0 0 32 132 | e = '(82-inf)'
	Percentage split	81,1765	0,667 0,688 0,833 0,802 0,841	a b c d e <-- classified as 2 1 0 0 0 | a = '(-inf-28]' 0 11 5 0 0 | b = '(28-46]' 0 3 80 13 0 | c = '(46-64]' 0 0 17 130 15 | d = '(64-82]' 0 0 0 10 53 | e = '(82-inf)'
SMO	Cross validation	78,6	0,778 0,679 0,816 0,771 0,823	a b c d e <-- classified as 7 2 0 0 0 | a = '(-inf-28]' 3 53 22 0 0 | b = '(28-46]' 0 27 244 28 0 | c = '(46-64]' 0 0 64 347 39 | d = '(64-82]' 0 0 0 29 135 | e = '(82-inf)'
	Percentage split	81,1765	0,667 0,688 0,844 0,796 0,841	a b c d e <-- classified as 2 1 0 0 0 | a = '(-inf-28]' 0 11 5 0 0 | b = '(28-46]' 0 5 81 10 0 | c = '(46-64]' 0 0 18 129 15 | d = '(64-82]' 0 0 0 10 53 | e = '(82-inf)'
MP	Cross validation	74,5	0,667 0,628 0,726 0,789 0,720	a b c d e <-- classified as 6 2 1 0 0 | a = '(-inf-28]' 1 49 27 1 0 | b = '(28-46]' 0 28 217 53 1 | c = '(46-64]' 0 0 56 355 39 | d = '(64-82]' 0 0 0 46 118 | e = '(82-inf)'
	Percentage split	75	0,667 0,625 0,656 0,809 0,778	a b c d e <-- classified as 2 1 0 0 0 | a = '(-inf-28]' 0 10 6 0 0 | b = '(28-46]' 0 11 63 21 1 | c = '(46-64]' 0 2 17 131 12 | d = '(64-82]' 0 0 0 14 49 | e = '(82-inf)'

Αλγόριθμος	Test Options	Συνολική Ακρίβεια %	TP Rate	Πίνακας σύγχυσης
IBk	Cross validation	k=1 ⇨ 73,4 k=3 ⇨ 77,5 k=5 ⇨ 79,2	(k=5) 0,556 0,564 0,823 0,840 0,726	(k=5) a b c d e <-- classified as 5 2 1 1 0 | a = '(-inf-28]' 0 44 33 1 0 | b = '(28-46]' 0 14 246 39 0 | c = '(46-64]' 0 0 53 378 19 | d = '(64-82]' 0 0 0 45 119 | e = '(82-inf)'
	Percentage split	k =1 ⇨ 76,4706 k=3 ⇨ 80,5882 k =5 ⇨ 81,7647	(k=5) 0,333 0,563 0,813 0,883 0,746	(k=5) a b c d e <-- classified as 1 1 0 1 0 | a = '(-inf-28]' 0 9 7 0 0 | b = '(28-46]' 0 2 78 16 0 | c = '(46-64]' 0 0 13 143 6 | d = '(64-82]' 0 0 0 16 47 | e = '(82-inf)'

Πίνακας 1: Απόδοση αλγορίθμων στο σύνολο των δεδομένων

Διαπιστώνεται πως τη μεγαλύτερη απόδοση (81,7647%) είχε ο αλγόριθμος IBk (k-NN) με τιμή της παραμέτρου k=5 με percentage split 66%. Δεύτεροι, με ίδια απόδοση 81,1765%, είναι οι SMO και Bayes Net πάλι με percentage split 66%. Γενικά παρατηρείται πως όλοι οι αλγόριθμοι, πλην του J48, είχαν μεγαλύτερη απόδοση με percentage split 66% έναντι της επικύρωσης cross validation. Εξετάζοντας την απόδοση μόνο ως προς την επικύρωση cross validation, η μεγαλύτερη απόδοση είναι 79,2% των αλγορίθμων Bayes Net και IBk με k=5. Με 79,1% έπονται οι αλγόριθμοι J48 και Naïve Bayes.

Προκειμένου να επιτευχθούν όσο το δυνατόν πιο έγκυρα αποτελέσματα, στο πλαίσιο της προεπεξεργασίας των δεδομένων θα χρησιμοποιηθεί και η επιλογή χαρακτηριστικών (Select attributes) για τον εντοπισμό των χαρακτηριστικών που είναι οι στατιστικά σημαντικές ως προς το χαρακτηριστικό κλάσης. Με την επιλογή χαρακτηριστικών θα υπολογιστεί η συσχέτιση μεταξύ κάθε χαρακτηριστικού και του χαρακτηριστικού κλάσης (μεταβλητή εξόδου) ώστε τελικά να επιλεγούν τα χαρακτηριστικά που έχουν τη μεγαλύτερη συσχέτιση.
Μετά την εφαρμογή του φίλτρου Discretize και τη δημιουργία 5 διαστημάτων, στην καρτέλα Select attributes επιλέγεται ως Attribute Evaluator ο ReliefFAtributeEval, οπότε και ως Search method πρέπει να είναι η επιλεγμένη η μέθοδος Ranker.
Επιλέγοντας ως κλάση Writing και εκκινώντας τη διαδικασία εμφανίζονται τα εξής αποτελέσματα της εικόνας 45.

```
Ranked attributes:
 0.4585    7 reading
 0.3034    6 math
 0.1383    1 gender
 0.0579    5 preparationcourse
 0.0458    3 parentaleducation
 0.0328    4 lunch
 0.0155    2 raceethnicity

Selected attributes: 7,6,1,5,3,4,2 : 7
```

Εικόνα 45: Αποτελέσματα καρτέλας Select Attributes

Διαγράφοντας από τα δεδομένα τα 4 τελευταία χαρακτηριστικά που είναι τα λιγότερο σημαντικά καθώς έχουν χαμηλούς συντελεστές (η διαγραφή πραγματοποιείται εύκολα από την καρτέλα Preprocess, επιλέγοντας τα χαρακτηριστικά και κάνοντας κλικ στο Remove), γίνεται εκ νέου εκτέλεση των αλγορίθμων και τα αποτελέσματα παρουσιάζονται στον πίνακα 2:

Αλγόριθμος	**Test Options**	**Συνολική Ακρίβεια %**	**TP Rate**	**Πίνακας σύγχυσης**
J48	Cross validation	79,1	0,889 0,705 0,813 0,773 0,835	a b c d e <-- classified as 8 1 0 0 0 | a = '(-inf-28]' 3 55 20 0 0 | b = '(28-46]' 0 26 243 30 0 | c = '(46-64]' 0 0 63 348 39 | d = '(64-82]' 0 0 0 27 137 | e = '(82-inf)'
	Percentage split	81,4706	0,667 0,750 0,833 0,802 0,841	a b c d e <-- classified as 2 1 0 0 0 | a = '(-inf-28]' 0 12 4 0 0 | b = '(28-46]' 0 5 80 11 0 | c = '(46-64]' 0 0 17 130 15 | d = '(64-82]' 0 0 0 10 53 | e = '(82-inf)'
Naïve Bayes	Cross validation	79,4	0,889 0,705 0,816 0,778 0,835	a b c d e <-- classified as 8 1 0 0 0 | a = '(-inf-28]' 3 55 20 0 0 | b = '(28-46]' 0 23 244 32 0 | c = '(46-64]' 0 0 62 350 38 | d = '(64-82]' 0 0 0 27 137 | e = '(82-inf)'
	Percentage split	81,4706	0,667 0,688 0,844 0,802 0,841	a b c d e <-- classified as 2 1 0 0 0 | a = '(-inf-28]' 0 11 5 0 0 | b = '(28-46]' 0 3 81 12 0 | c = '(46-64]' 0 0 17 130 15 | d = '(64-82]' 0 0 0 10 53 | e = '(82-inf)'

Αλγόριθμος	Test Options	Συνολική Ακρίβεια %	TP Rate	Πίνακας σύγχυσης
Bayes Net	Cross validation	79,4	0,889 0,705 0,816 0,778 0,835	a b c d e <-- classified as 8 1 0 0 0 | a = '(-inf-28]' 3 55 20 0 0 | b = '(28-46]' 0 23 244 32 0 | c = '(46-64]' 0 0 62 350 38 | d = '(64-82]' 0 0 0 27 137 | e = '(82-inf)'
	Percentage split	81,4706	0,667 0,688 0,844 0,802 0,841	a b c d e <-- classified as 2 1 0 0 0 | a = '(-inf-28]' 0 11 5 0 0 | b = '(28-46]' 0 3 81 12 0 | c = '(46-64]' 0 0 17 130 15 | d = '(64-82]' 0 0 0 10 53 | e = '(82-inf)'
SMO	Cross validation	79,2	0,889 0,718 0,819 0,771 0,829	a b c d e <-- classified as 8 1 0 0 0 | a = '(-inf-28]' 3 56 19 0 0 | b = '(28-46]' 0 26 245 28 0 | c = '(46-64]' 0 0 65 347 38 | d = '(64-82]' 0 0 1 27 136 | e = '(82-inf)'
	Percentage split	81,4706	0,667 0,750 0,844 0,796 0,841	a b c d e <-- classified as 2 1 0 0 0 | a = '(-inf-28]' 0 12 4 0 0 | b = '(28-46]' 0 5 81 10 0 | c = '(46-64]' 0 0 18 129 15 | d = '(64-82]' 0 0 0 10 53 | e = '(82-inf)'
MP	Cross validation	79,1	0,889 0,692 0,809 0,796 0,787	a b c d e <-- classified as 8 1 0 0 0 | a = '(-inf-28]' 0 54 24 0 0 | b = '(28-46]' 0 21 242 36 0 | c = '(46-64]' 0 0 54 358 38 | d = '(64-82]' 0 0 0 35 129 | e = '(82-inf)'
	Percentage split	81,7647	0,667 0,750 0,823 0,815 0,841	a b c d e <-- classified as 2 1 0 0 0 | a = '(-inf-28]' 0 12 4 0 0 | b = '(28-46]' 0 3 79 14 0 | c = '(46-64]' 0 0 15 132 15 | d = '(64-82]' 0 0 0 10 53 | e = '(82-inf)'

<table>
<tr><th>Αλγόριθμος</th><th>Test Options</th><th>Συνολική Ακρίβεια %</th><th>TP Rate</th><th>Πίνακας σύγχυσης</th></tr>
<tr><td rowspan="2">IBk</td><td>Cross validation</td><td>k=1 ⇨ 79,5
k=3 ⇨ 79,3
k=5 ⇨ 78,7</td><td>(k=1)
0,889
0,705
0,809
0,787
0,829</td><td>(k=1)
a b c d e <-- classified as
8 1 0 0 0 | a = '(-inf-28]'
0 55 23 0 0 | b = '(28-46]'
0 22 242 35 0 | c = '(46-64]'
0 0 57 354 39 | d = '(64-82]'
0 0 0 28 136 | e = '(82-inf)'</td></tr>
<tr><td>Percentage split</td><td>k =1 ⇨
79,7059
k=3 ⇨
79,4118
k =5 ⇨
79,4118</td><td>(k=1)
0,667
0,750
0,823
0,864
0,603</td><td>(k=1)
a b c d e <-- classified as
2 1 0 0 0 | a = '(-inf-28]'
0 12 4 0 0 | b = '(28-46]'
0 3 79 14 0 | c = '(46-64]'
0 0 15 140 7 | d = '(64-82]'
0 0 0 25 38 | e = '(82-inf)'</td></tr>
</table>

Πίνακας 2: Απόδοση αλγορίθμων στο νέο σύνολο δεδομένων που περιλαμβάνει μόνο τα σημαντικά χαρακτηριστικά

Μεγαλύτερη απόδοση είχε ο Multilayer Perceptron (81,7647%) με percentage split 66%. Ακολουθούν με 81,4706% οι J48, Naïve Bayes, Bayes Net, SMO επίσης με percentage split 66%.

Εστιάζοντας στην επικύρωση validation cross πρωτεύει ο IBk με k=1 (79,5%) και ακολουθούν με 79,4% οι Naïve Bayes και Bayes Net.

9.4 Αποτελέσματα αλγορίθμων κατηγοριοποίησης σε δεδομένα 2 κλάσεων

Από εκπαιδευτικής άποψης είναι ιδιαίτερης σημασία ο εντοπισμός των μαθητών των οποίων η απόδοση κυμαίνεται σε χαμηλά επίπεδα, οπότε ελλοχεύει ο κίνδυνος είτε να χαθεί η χρονιά είτε να δημιουργηθούν στους μαθητές μαθησιακά κενά τα οποία θα αποτελέσουν τροχοπέδη για το υπόλοιπο των σπουδών τους. Οπότε η αναζήτηση των μαθητών που ο βαθμός τους θα είναι μικρότερος του 55 στα 100 αποτελεί ένα επιπλέον ζητούμενο για την παρούσα εργασία.
Με τη χρήση του φίλτρου Discretize δημιουργούνται οι κλάσεις (-inf, 55] και (55, inf).
Τα αποτελέσματα των αλγορίθμων απεικονίζονται στον πίνακα 3:

Αλγόριθμος	**Test Options**	**Συνολική Ακρίβεια %**	**TP Rate**	**Πίνακας σύγχυσης**
J48	Cross validation	93,8	0,862 0,959	a b <-- classified as 187 30 | a = '(-inf-55]' 32 751 | b = '(55-inf)'
	Percentage split	95	0,830 0,972	a b <-- classified as 44 9 | a = '(-inf-55]' 8 279 | b = '(55-inf)'
Naïve Bayes	Cross validation	93,1	0,866 0,949	a b <-- classified as 188 29 | a = '(-inf-55]' 40 743 | b = '(55-inf)'
	Percentage split	94,4118	0,830 0,965	a b <-- classified as 44 9 | a = '(-inf-55]' 10 277 | b = '(55-inf)'
Bayes Net	Cross validation	93	0,866 0,948	a b <-- classified as 188 29 | a = '(-inf-55]' 41 742 | b = '(55-inf)'
	Percentage split	94,4118	0,830 0,965	a b <-- classified as 44 9 | a = '(-inf-55]' 10 277 | b = '(55-inf)'

Αλγόριθμος	Test Options	Συνολική Ακρίβεια %	TP Rate	Πίνακας σύγχυσης
SMO	Cross validation	93,6	0,853 0,959	a b <-- classified as 185 32 | a = '(-inf-55]' 32 751 | b = '(55-inf)'
	Percentage split	95	0,830 0,972	a b <-- classified as 44 9 | a = '(-inf-55]' 8 279 | b = '(55-inf)'
MP	Cross validation	91,3	0,783 0,949	a b <-- classified as 170 47 | a = '(-inf-55]' 40 743 | b = '(55-inf)'
	Percentage split	92,6471	0,774 0,955	a b <-- classified as 41 12 | a = '(-inf-55]' 13 274 | b = '(55-inf)'
IBk	Cross validation	k=1 ⇨ 91,4 k=3 ⇨ 92,9 k=5 ⇨ 92,9	(k=3) 0,843 0,953	(k=3) a b <-- classified as 183 34 | a = '(-inf-55]' 37 746 | b = '(55-inf)'
	Percentage split	k=1 ⇨ 93,5294 k=3 ⇨ 95,2941 k=5 ⇨ 95,8824	(k=5) 0,830 0,983	(k=5) a b <-- classified as 44 9 | a = '(-inf-55]' 5 282 | b = '(55-inf)'

Πίνακας 3: Αποτελέσματα αλγορίθμων σε δεδομένα 2 κλάσεων

Υψηλότερη απόδοση είχε ο IBk με k=5 και k=3, 95,8824% kai 95,2941% αντίστοιχα. Ακολουθούν οι αλγόριθμοι J48 και SMO. Με percentage split 66% η απόδοση και των δύο ήταν 95% και με cross validation ήταν 93,8% και 93,6% αντίστοιχα.

Με την επιλογή χαρακτηριστικών της καρτέλας Select Attributes και επιλέγοντας ως Attribute Evaluator τον ReliefFAtributeEval όπως και παραπάνω, προκύπτουν τα αποτελέσματα της εικόνας 46:

```
Ranked attributes:
 0.4765    7 reading
 0.2215    1 gender
 0.1678    5 preparationcourse
 0.1368    3 parentaleducation
 0.1344    6 math
 0.1305    2 raceethnicity
 0.0131    4 lunch

Selected attributes: 7,1,5,3,6,2,4 : 7
```

Εικόνα 46: Αποτελέσματα καρτέλας Select Attributes

Διαγράφοντας το χαρακτηριστικό lunch και εκτελώντας εκ νέου τους αλγορίθμους, προκύπτουν τα αποτελέσματα του πίνακα 4:

Αλγόριθμος	**Test Options**	**Συνολική Ακρίβεια %**	**TP Rate**	**Πίνακας σύγχυσης**

J48	Cross validation	93,8	0,862
			0,959
				`a b <-- classified as`
				`187 30 | a = '(-inf-55]'`
				`32 751 | b = '(55-inf)'`

	Percentage split	95	0,830
			0,972
				`a b <-- classified as`
				`44 9 | a = '(-inf-55]'`
				`8 279 | b = '(55-inf)'`

Naïve Bayes	Cross validation	93,3	0,871
			0,950
				`a b <-- classified as`
				`189 28 | a = '(-inf-55]'`
				`39 744 | b = '(55-inf)'`

	Percentage split	95	0,849
			0,969
				`a b <-- classified as`
				`45 8 | a = '(-inf-55]'`
				`9 278 | b = '(55-inf)'`

Bayes Net	Cross validation	93,3	0,871
			0,950
				`a b <-- classified as`
				`189 28 | a = '(-inf-55]'`
				`39 744 | b = '(55-inf)'`

	Percentage split	95	0,849
			0,969
				`a b <-- classified as`
				`45 8 | a = '(-inf-55]'`
				`9 278 | b = '(55-inf)'`

Αλγόριθμος	Test Options	Συνολική Ακρίβεια %	TP Rate	Πίνακας σύγχυσης
SMO	Cross validation	93,6	0,853 0,959	a b <-- classified as 185 32 | a = '(-inf-55]' 32 751 | b = '(55-inf)'
	Percentage split	95	0,830 0,972	a b <-- classified as 44 9 | a = '(-inf-55]' 8 279 | b = '(55-inf)'
MP	Cross validation	91,9	0,779 0,958	a b <-- classified as 169 48 | a = '(-inf-55]' 33 750 | b = '(55-inf)'
	Percentage split	92,3529	0,717 0,962	a b <-- classified as 38 15 | a = '(-inf-55]' 11 276 | b = '(55-inf)'
IBk	Cross validation	k=1 ⇨ 92,3 k=3 ⇨ 93,6 k=5 ⇨ 93,7	(k=5) 0,848 0,962	(k=5) a b <-- classified as 184 33 | a = '(-inf-55]' 30 753 | b = '(55-inf)'
	Percentage split	k=1 ⇨ 93,5294 k=3 ⇨ 95,2941 k=5 ⇨ 95,2941	(k=3) 0,811 0,979	(k=3) a b <-- classified as 43 10 | a = '(-inf-55]' 6 281 | b = '(55-inf)'

Πίνακας 4: Αποτελέσματα μετά τη διαγραφή του χαρακτηριστικού lunch

Με percentage split 66% υψηλότερη απόδοση είχε ο IBk με k=3 και k=5 (95,2941%). Ακολούθως, πολύ μικρή διαφορά (απόδοση 95%) είχαν οι J48, Naïve Bayes, Bayes Net και SMO. Με cross validation ο J48, IBk (k=5) και οι SMO και IBk (k=3) είχαν 93,8%, 93,7% και 93,6% οι δύο τελευταίοι αντίστοιχα.

10. Συμπεράσματα

Η μηχανική μάθηση γνωρίζει εντυπωσιακή ανάπτυξη τα τελευταία χρόνια. Ένας από τους τομείς στους οποίους βρίσκει εφαρμογή είναι η εκπαίδευση με τη συμβολή της να μελετάται από τους ερευνητές, όπως παρουσιάστηκε και στη βιβλιογραφική επισκόπηση στο κεφάλαιο 7. Η χρήση των υπολογιστών στην εκπαίδευση, ιδιαίτερα την τριτοβάθμια, δημιουργεί μεγάλο όγκο δεδομένων προς ανάλυση. Η μηχανική μάθηση μπορεί να συμβάλει στην εκπαιδευτική διαδικασία μέσω της πρόβλεψης της απόδοσης των μαθητών, εκτός άλλων εφαρμογών. Γνωρίζοντας εγκαίρως βαθμούς και σκιαγραφώντας τη μαθησιακή πορεία μαθητών, είναι δυνατή η υλοποίηση δράσεων και παρεμβάσεων στην εκπαιδευτική διαδικασία, έχοντας ως τελικό στόχο την αύξηση του ενδιαφέροντος και τη βελτίωση της απόδοσης του μαθητή, κάτι που αποτελεί ζητούμενο της μαθητικής αλλά και της εκπαιδευτικής κοινότητας. Στην παρούσα εργασία έγινε χρήση εκπαιδευτικών δεδομένων σε αλγορίθμους κατηγοριοποίησης με στόχο την πρόβλεψη της απόδοσης των μαθητών, ερευνώντας, με τον τρόπο αυτό, πιθανή συμβολή στη διδασκαλία.

Τα δεδομένα χωρίστηκαν σε 5 κλάσεις (απόδοση πολύ χαμηλή, χαμηλή, μέτρια, καλή, πολύ καλή) αρχικά και σε 2 κλάσεις στη συνέχεια (απόδοση χαμηλή ή καλή). Εφαρμόστηκαν 6 αλγόριθμοι κατηγοριοποίησης στο αρχικό σύνολο δεδομένων και στο σύνολο δεδομένων που προέκυψε μετά την επιλογή χαρακτηριστικών. Ως μέθοδοι επικύρωσης χρησιμοποιήθηκαν cross-validation με 10 folds αλλά και percentage split 66%.

Δεν υπάρχει κάποιος συγκεκριμένος αλγόριθμος που να ξεχώρισε σημειώνοντας τις υψηλότερες επιδόσεις σε κάθε πείραμα. Οι αλγόριθμοι είχαν διαφορετικές αποδόσεις, κάτι που επιβεβαιώνει τα αποτελέσματα των λοιπών εργασιών, όπου οι αλγόριθμοι διέφεραν σε απόδοση σε κάθε εργασία, ανάλογα με τα δεδομένα. Η απόδοση των αλγορίθμων κρίνεται από ικανοποιητική (της τάξης του 80%) έως υψηλή καθώς φτάνει περίπου το 96%.

Η μείωση των χαρακτηριστικών δεν επιδρά ουσιαστικά στην απόδοση των αλγορίθμων, στο συγκεκριμένο σύνολο δεδομένων. Η απόδοση επηρεάζεται ελάχιστα, σε ποσοστό μικρότερο από 1%, πλην του αλγορίθμου MP στο αρχικό σύνολο δεδομένων και με τις δύο μεθόδους επικύρωσης που η διαφορά στην απόδοση αυξήθηκε κατά 5% περίπου.

Σε επόμενες μελλοντικές έρευνες μπορούν να εφαρμοστούν περισσότεροι αλγόριθμοι σε ξεχωριστά μαθήματα διαφορετικών κατευθύνσεων, στην προσπάθεια ανακάλυψης χαρακτηριστικών των μαθητών που δυσκολεύονται ή επιτυγχάνουν σε συγκεκριμένα μαθήματα ή ακόμα και μαθητών που έχουν λιγότερα μαθησιακά κίνητρα. Επιπρόσθετα, μπορούν να χρησιμοποιηθούν μεγαλύτερα σύνολα δεδομένων. Τα δεδομένα μπορούν να εμπλουτιστούν με πολλά χαρακτηριστικά που θα προσεγγίζουν περισσότερο την καθημερινότητα των μαθητών, η οποία περιλαμβάνει τη χρήση ηλεκτρονικών συσκευών, κοινωνικών δικτύων κ.λπ., αλλά και χαρακτηριστικά σχετικά με τη χρήση συστημάτων

διαχείρισης μάθησης όπου αυτά χρησιμοποιούνται. Το δείγμα μαθητών θα μπορούσε να είναι μεγαλύτερο και να περιλαμβάνει μαθητές που βρίσκονται σε κάθε γωνιά της χώρας (ή και του κόσμου), καθώς και να εφαρμοστεί σε μεγαλύτερο βάθος χρόνου ώστε να μελετηθούν τυχόν συνέπειες και παρεκκλίσεις στην πορεία και απόδοση των μαθητών από τις πρόσφατες συνθήκες που δημιουργήθηκαν λόγω της πανδημίας και των καραντινών (lockdown).

Βιβλιογραφία – Αναφορές

Ελληνική

Βλαχάβας Ι., Κεφαλάς Π., Βασιλειάδης Ν, Κόκκορας Φ., Σακελλαρίου Η., 2006, «Τεχνητή Νοημοσύνη», Γκιούρδας

Διαμαντάρας Κ., Μπότσης Δ., 2019, «Μηχανική Μάθηση», Κλειδάριθμος

Γεωργούλη Κ., 2015, «Τεχνητή Νοημοσύνη Μια εισαγωγική προσέγγιση», www.kallipos.gr

Ζουμπουλίδης Η., 2012, «Εφαρμογή μεθόδων εξόρυξης γνώσης στον εντοπισμό παραποιημένων λογιστικών καταστάσεων», Ε.Α.Π.

Κοτταρά Π., 2019, «Ανάπτυξη διαδικτυακής βιβλιοθήκης μοντέλων μηχανικής μάθησης για την πρόβλεψη ιδιοτήτων νανοϋλικών», Ε.Μ.Π.

Κύρκος Ε., 2015, «Επιχειρηματική ευφυία και εξόρυξη δεδομένων», www.kallipos.gr

Μυλωνάς Φ., 2022, «Εφαρμογές μηχανικής ευφυΐας και μάθησης», Σημειώσεις μαθήματος, Πανεπιστήμιο Δυτικής Αττικής

Παρτάλας Ι., 2009, «Μέθοδοι ενισχυτικής μάθησης σε συστήματα πρακτόρων», Α.Π.Θ.

Σερέτη Χ., 2020, «Διαχείριση δεδομένων βιβλιοθηκών στην πλατφόρμα WEKA», Πανεπιστήμιο Δυτικής Αττικής

Φλώρου Ε., 2017, «Αυτόματη διάκριση μεταφορικής και κυριολεκτικής σημασίας σε σώματα κειμένων της ελληνικής με υπολογιστικές μεθόδους Μηχανικής Μάθησης», Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Ξένη

Agrawal, R.; Imieliński, T.; Swami, A., 1993, “Mining association rules between sets of items in large databases”. Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93. p. 207

Aher S., Lobo L., 2011, “Data Mining in Educational System using WEKA”

Alturki S., Alturki N., 2021, “Using Educational Data Mining to predict students’ academic performance for applying early interventions”

Ashraf A., Anwe S., Khan M., 2018, “A Comparative Study of Predicting Student’s Performance by use of Data Mining Techniques”, American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS) Volume 44, No 1, pp 122-136

Baker, R.S., Yacef K., 2009, “The state of educational data mining in 2009: A review and future visions”. JEDM-Journal of Educational Data Mining. 1 (1): 2017.

Breiman L., Friedman J., Stone C., Olshen R., 1984, “Classification and Regression Trees Wadsworth statistics/probability series”, Taylor & Francis, ISBN-13: 978-0412048418

Cocea M., Weibelzahl S., 2006, “Can log files analysis estimate learners’ level of motivation?”, Proceedings of the workshop week Lernen—Wissensentdeckung—Adaptivität, Hildesheim, pp 32–35

Cook D.R., Weisberg S., 1982, “Residuals and influence in Regression”, Chapman and Hall, ISBN 0-412-24280-0

Dey A., Khasnabis A., Kumar A., 2018, “Prediction and Analysis of Student Performance by Data Mining in WEKA”

Ezugwu A., Shukla A., Agbaje M., Oyelade O., Jose-Garcia A., Agushaka J., 2020, “Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature”, Neural Computing & Applications 33, 6247–6306

Fayyad U. M., Piatetsky-Shapiro G., Uthurusamy R., 2003, Summary from the KDD-03 Panel – Data Mining: The Next 10 Years. ACM SIGKDD Exploration Newsletter, 5(2), 191-196

Hamalainen W., Vinni M., 2006, “Comparison of machine learning methods for intelligent tutoring systems, Proceedings of the eighth international conference in intelligent tutoring systems”, Taiwan, pp 525–534

Hussain S., Dahan N., Ba-Alwi F., Ribata N., 2017, “Educational Data Mining and Analysis of Students’ Academic Performance Using WEKA”

Kabakchieva D., 2013, “Predicting Student Performance by Using Data Mining Methods for Classification”

Kononenko I., 1994, Estimating attributes: Analysis and extensions of relief, Proceedings of the European Conference on Machine Learning, Springer

Kotsiantis S., Pierrakeas C., Pintelas P., 2003, “Preventing Student Dropout in Distance Learning Using Machine Learning Techniques”, LNAI 2774, pp. 267-274

Mhetre V., Nagar M., 2017, “Classification based data mining algorithms to predict slow, average and fast learners in educational system using Weka”

Minaei-Bidgoli B., Punch W., 2003, “Using Genetic Algorithms for Data Mining Optimization in an Educational Web-Based System”

Mueen A., Zafar B., Manzoor U., 2016, “Modeling and Predicting Students' Academic Performance Using Data Mining Techniques”, International Journal of Modern Education and Computer Science (IJMECS), Vol.8, No.11, pp.36-42, DOI: 10.5815/ijmecs.2016.11.05

Picciano, A. 2012, “The evolution of big data and learning analytics in American higher education”, Journal of Asynchronous Learning Networks, Volume 16, Issue 3.

Quinlan J. R., 1987, “Simplifying decision trees”, International Journal of Man-Machine Studies - Special Issue: Knowledge Acquisition for Knowledge-based Systems

Romero C., Ventura S., “Educational data mining: A survey from 1995 to 2005”, Expert Systems with Applications, 33, 135-146

Ruby J., David K., 2014, “Predicting the Performance of Students in Higher Education Using Data Mining Classification Algorithms - A Case Study”, International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Russell, Binder, Koller, and Kanazawa, 1997 “Adaptive Probabilistic Networks with Hidden Variables”, Springer

Thankachan, K., 2017, “Automating anomaly detection for exploratory data analytics”, International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, 711-715, DOI: 10.1109/ICICI.2017.8365228

Tan P.-N., Steinbach M., Kumar V., 2006, “Introduction to Data Mining”, Addison Wesley

Xin Y., Xiao G. S., 2009, “Linear Regression Analysis Theory and Computing”, World Scientific Publishing Co. Pte. Ltd, ISBN-13 978-981-283-410-2

Vapnik, V., “The Nature of Statistical Learning Theory”, 1995

Yudelson MV, Medvedeva O., Legowski E., Castine M., Jukic D., Rebecca C., 2006, “Mining student learning data to develop high level pedagogic strategy in a medical ITS”, Proceedings of AAAI workshop on educational data mining, Boston, pp 1–8