ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΜΗΧΑΝΟΛΟΓΩΝ ΚΑΙ ΑΕΡΟΝΑΥΠΗΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΤΟΜΕΑΣ ΟΡΓΑΝΩΣΗΣ ΚΑΙ ΔΙΟΙΚΗΣΗΣ

ΣΠΟΥΔΑΣΤΙΚΗ ΕΡΓΑΣΙΑ

Αλγόριθμοι ταξινόμησης στην Μηχανική Μάθηση: Μέθοδοι και τεχνικές υλοποίησης

Σταμάτιος Πασπαλιάρης
ΑΜ: 1048418

Επιβλέπων: Νικόλαος Καρακαπιλίδης, Καθηγητής

Πάτρα, 2023

Περίληψη

Στόχος της παρούσας εργασίας είναι η παράθεση του θεωρητικού υποβάθρου για την υλοποίηση ευφυών υπολογιστικών συστημάτων, που υπάγονται στο πλαίσιο της Τεχνητής Νοημοσύνης. Το συγκεκριμένα συστήματα, βασίζονται στην δια λειτουργικότητα επιμέρους συστημάτων Μηχανικής Μάθησης. Με βάση αυτή την παραδοχή, η εργασία εστιάζεται στην παρουσίαση των τεχνικών υλοποίησης, των κρίσιμων παραμέτρων και των πεδίων εφαρμογής τριών βασικών αλγορίθμων επιτηρούμενης Μηχανικής Μάθησης. Συγκεκριμένα, οι αλγόριθμοι που επιλέγονται είναι τα δέντρα αποφάσεων, ο ταξινομητής κατά Naïve Bayes και ο αλγόριθμος των κοντινότερων γειτόνων (k-NN). Για κάθε έναν από τους παραπάνω αλγορίθμους παρατίθεται και ένα εκτενές παράδειγμα, με στόχο την πλήρη αποσαφήνιση του τρόπου υλοποίησής τους. Επιπλέον, έμφαση δίνεται και στο στάδιο προ – επεξεργασίας των δεδομένων, το οποίο θεωρείται ιδιαίτερα σημαντικό για την αποτελεσματική υλοποίηση των αλγορίθμων. Τέλος, δίνεται το σύνολο των μετρικών ελέγχου της αποτελεσματικότητας των αλγορίθμων ταξινόμησης.

Λέξεις κλειδιά: Τεχνητή Νοημοσύνη, αλγόριθμοι Μηχανικής Μάθησης, επιτηρούμενη μάθηση, Δέντρα Αποφάσεων, Naïve Bayes, k-NN.

Abstract

This work aims to present the theoretical background for the implementation of intelligent computing systems, which fall within the framework of Artificial Intelligence. The specific systems are based on the inter-functionality of individual Machine Learning algorithms. Based on this assumption, this work focuses on presenting the implementation techniques, critical issues and application fields of three basic supervised Machine Learning algorithms. Specifically, the algorithms chosen are Decision Trees, the Naïve Bayes classifier and the k-NN nearest neighbors algorithm. For each of the above algorithms, an extensive example is listed, with the aim of fully clarifying how they are implemented. Furthermore, emphasis is placed on the data pre-processing stage, which is considered important for the effective implementation of the algorithms. Finally, the evaluation metrics for the classification algorithms, are given.

Key words: Artificial Intelligence, Machine Learning algorithms, Supervised learning, Decision Trees, Naïve Bayes, k-NN.

Περιεχόμενα

Λίστα Πινάκων

Λίστα Εικόνων

Λίστα Σχημάτων

Εισαγωγή

Γενικά στοιχεία

Τα τελευταία χρόνια, οι προσπάθειες για την δημιουργία μοντέλων ικανών να προσομοιώνουν την ανθρώπινη σκέψη και τον τρόπο λήψης αποφάσεων, έχουν ενταθεί ιδιαίτερα. Τα συστήματα μέσω των οποίων υλοποιούνται οι εν λόγω προσομοιώσεις, υπάγονται στο γενικότερο πλαίσιο της Τεχνητής Νοημοσύνης. Ανατρέχοντας στην διεθνή βιβλιογραφία, μπορεί να παρατηρήσει κανείς, μεγάλο όγκο τέτοιων εφαρμογών καθώς και μελετών περίπτωσης, σε διαφορετικούς κλάδους. Ενδεικτικά, σημειώνεται ότι μεγάλη πληθώρα εφαρμογών Τεχνητής Νοημοσύνης, παρατηρείται στον κλάδο της μηχανικής, της ιατρικής, των νευροεπιστημών καθώς και στην εφοδιαστική αλυσίδα.

Στην πραγματικότητα, η υλοποίηση συστημάτων Τεχνητής Νοημοσύνης, αποτελεί το ανώτερο ιεραρχικά επίπεδο των συστημάτων εξόρυξης γνώσης. Για να καταστεί εφικτή η ευφυία των υπολογιστών, θα πρέπει να λειτουργούν ταυτόχρονα και να επικοινωνούν διάφορα άλλα συστήματα, με κυριότερα από αυτά να θεωρούνται τα συστήματα Μηχανικής Μάθησης καθώς και τα συστήματα «βαθιάς» γνώσης (deep learning). Σύμφωνα με την συγκεκριμένη παραδοχή, γίνεται αντιληπτό ότι το πλέον ουσιαστικό μέρος για την ανάπτυξη εφαρμογών Τεχνητής Νοημοσύνης, είναι ο στοχευμένος σχεδιασμός συστημάτων Μηχανικής Μάθησης.

Αναφορικά με τους αλγορίθμους Μηχανικής Μάθησης, σημειώνεται ότι έχουν αναπτυχθεί πολλές προσεγγίσεις, με στόχο την επίλυση διαφορετικών προβλημάτων. Ωστόσο όλες οι προσεγγίσεις, υπάγονται σε συγκεκριμένες μεθοδολογίες υλοποίησης και ως εκ τούτου μπορούν να ταξινομηθούν σε τρεις επιμέρους κατηγορίες: τους αλγόριθμους επιτηρούμενης μάθησης, τους αλγόριθμους μη επιτηρούμενης μάθησης και τους αλγόριθμους ενίσχυσης γνώσης. Μια επιπλέον κατηγορία, είναι οι αλγόριθμοι ημί-επιτηρούμενης μάθησης, οι οποίοι μεθοδολογικά βρίσκονται στο μεταίχμιο μεταξύ των δύο πρώτων κατηγοριών.

Δομή εργασίας

Η παρούσα εργασία, εστιάζεται στην παρουσίαση αλγορίθμων επιτηρούμενης Μηχανικής Μάθησης, σχετικά με την επίλυση του προβλήματος της ταξινόμησης. Συγκεκριμένα, επιλέγονται τρεις διαδεδομένοι αλγόριθμοι ταξινόμησης: Δέντρα αποφάσεων, Naïve Bayes και k-NN. Για κάθε έναν από τους συγκεκριμένους αλγορίθμους γίνεται τεχνική παρουσίαση ως προς την υλοποίησή τους, δίνοντας έμφαση στα σημεία ιδιαίτερου ενδιαφέροντος και τις

κρίσιμες παραμέτρους για την αποτελεσματική υλοποίηση. Επιπλέον, το θεωρητικό υπόβαθρο εμπλουτίζεται μέσω παράθεσης αναλυτικών παραδειγμάτων υλοποίησης, τα οποία στοχεύουν στην διασαφήνιση των μεθόδων.

Η εργασία δομείται σε τέσσερα (4) επιμέρους κεφάλαια. Το πρώτο κεφάλαιο αφορά την παρουσίαση ορισμένων θεωρητικών στοιχείων, σχετικά με τον τρόπο ταξινόμησης των αλγορίθμων Μηχανικής Μάθησης, ανάλογα με τον τρόπο υλοποίησής τους και ανάλογα με το πρόβλημα που μπορούν να επιλύσουν. Επιπρόσθετα, παρατίθενται και ορισμένες εφαρμογές στο πλαίσιο της Μηχανικής Μάθησης, προκειμένου να διασαφηνιστεί η χρηστικότητα και η αποτελεσματικότητα των μεθόδων, σε σύνθετα προβλήματα.

Στο δεύτερο κεφάλαιο, γίνεται αναφορά στα απαραίτητα προκάτοχα στάδια των αλγορίθμων (προ – επεξεργασία δεδομένων), τα οποία συμβάλουν σημαντικά στην βελτίωση της αποτελεσματικότητας τους. Ενώ δίνεται και το σύνολο των μετρικών ελέγχου της αποτελεσματικότητας των αλγορίθμων, το οποίο είναι κοινό για οποιονδήποτε αλγόριθμο χρησιμοποιηθεί και ως εκ τούτου αποτελεί στοιχείο συγκριτικής θεώρησης και συμπερασμοτολογίας σχετικά με την καταλληλόλητα του εκάστοτε αλγόριθμου για ένα συγκεκριμένο πρόβλημα.

Το τρίτο κεφάλαιο, αποτελεί στην πραγματικότητα τον βασικό κορμό της εργασίας. Το συγκεκριμένο κεφάλαιο, δομείται σε τρεις επιμέρους υποενότητες, κάθε μία από τις οποίες αναφέρεται σε έναν συγκεκριμένο αλγόριθμο ταξινόμησης, από αυτούς που αναφέρονται παραπάνω. Για κάθε έναν αλγόριθμο, δίνεται η αρχή λειτουργίας, τα σημαντικότερα τεχνικά στοιχεία καθώς και ένα αναλυτικό παράδειγμα, με εφαρμογή των όσων αναφέρονται στο θεωρητικό μέρος. Στην τελευταία ενότητα της εργασίας, γίνεται μια σύνοψη των όσων αναφέρθηκαν στην εργασία και προτείνονται μελλοντικές προεκτάσεις της.

Κεφάλαιο 1: Βασικές έννοιες και ορισμοί

Στο πρώτο κεφάλαιο της εργασίας, αποδίδονται ορισμένες βασικές έννοιες και ορισμοί, σχετικά με την Τεχνητή Νοημοσύνη (Artificial Intelligence) και την Μηχανική Μάθηση (Machine Learning). Πιο συγκεκριμένα, στο πρώτο μέρος, αποτυπώνεται ο ορισμός της Τεχνητής Νοημοσύνης, γίνεται μια σύντομη ιστορική αναδρομή για την εξέλιξη της έννοιας και παρατίθενται οι εννοιολογικές και τεχνικές προσεγγίσεις υλοποίησης έξυπνων συστημάτων, καθώς επίσης και τα δυνητικά επίπεδα αυτονομίας και αποτελεσματικότητας κάθε κατηγορίας τέτοιων συστημάτων. Έπειτα, παρουσιάζεται ο τρόπος δόμησης ευφυών συστημάτων, μέσα από την χρήση αλγορίθμων Μηχανικής Μάθησης και καταγράφονται ορισμένες πολύ βασικές εφαρμογές που έχουν υλοποιηθεί. Στο δεύτερο μέρος του κεφαλαίου, δίνεται ο ορισμός της Μηχανικής Μάθησης, γίνεται ταξινόμηση των αλγορίθμων σε επιμέρους κατηγορίες ανάλογα με την μέθοδο σχεδιασμού και λειτουργίας (επιβλεπόμενη, μη επιβλεπόμενη μάθηση κ.λπ.) και τέλος, παρατίθενται οι κύριες κατηγορίες προβλημάτων που αντιμετωπίζονται με αλγορίθμους Μηχανικής Μάθησης, μαζί με κάποιους ευρέως διαδεδομένους αλγορίθμους ανά κατηγορία.

1.1 Τεχνητή Νοημοσύνη

Η Τεχνητή Νοημοσύνη αποτελεί έναν από τους πιο επίκαιρους κλάδους της επιστήμης των υπολογιστών, ο οποίος αποκτά όλο και περισσότερο έδαφος μέσω εφαρμογών σε πολλούς και διαφορετικούς επιστημονικούς κλάδους, όπως η κλασσική μηχανική, η ιατρική, η βιολογία και άλλοι. Στην πραγματικότητα, πρόκειται για τον κλάδο, ο οποίος προσπαθεί να δημιουργήσει υπολογιστικά μοντέλα, τα οποία να μιμούνται σε μεγάλο βαθμό τις δυνατότητες του ανθρώπινου εγκεφάλου. Ο συνδυασμός, της πολυδιάστατης ανθρώπινης λογικής με τις τεράστιες δυνατότητες αποθήκευσης και επεξεργασίας δεδομένων, που παρέχουν οι υπολογιστές, μπορούν να οδηγήσουν στην ταχεία επίλυση σύνθετων προβλημάτων.

Ενώ η Τεχνητή Νοημοσύνη, θεωρείται ως μια πολύ νέα προσέγγιση στην επιστήμη των υπολογιστών, στην πραγματικότητα αυτή η θεώρηση δεν είναι ιδιαίτερα βάσιμη. Οι πρώτες προσπάθειες για την δημιουργία ευφυών μοντέλων με χρήση υπολογιστών, χρονολογούνται στην δεκαετία του 1940, σε πανεπιστήμια της Νέας Υόρκης και του Σικάγο, τα οποία στηρίζονταν στην δημιουργία τεχνητών νευρώνων. Η ιδέα και η λογική του συγκεκριμένου σχεδιασμού συστημάτων, θεωρήθηκε ιδιαίτερα πρωτοπόρα για την εποχή, γεγονός που οδήγησε στην διοργάνωση πολλών σχετικών ημερίδων και συνεδρίων καθώς και σε

γενικότερη ένταση της έρευνας γύρω από τον κλάδο. Ως αποκορύφωμα των ερευνών, θεωρείται η δημιουργία του πρώτου υπολογιστή τεχνητού νευρωνικού δικτύου, γύρω στο 1950 καθώς και η καθιέρωση μιας νέας γλώσσας προγραμματισμού, εν ονόματει List, για την εξυπηρέτηση σκοπών ανάπτυξης συστημάτων Τεχνητής Νοημοσύνης. Ωστόσο, τα τότε διαθέσιμα υπολογιστικά συστήματα, δεν είχαν την απαραίτητη ισχύ για την επεξεργασία του όγκου δεδομένων, που απαιτούνταν για την βελτίωση της αποτελεσματικότητα των αλγορίθμων. Γεγονός ου οδήγησε στην απομάκρυνση από την ιδέα της Τεχνητής Νοημοσύνης, για περισσότερα από 30 χρόνια, μέχρι να φτάσουμε στην δεκαετία του 2000, όπου η ανάπτυξη της ισχύος των υπολογιστών αλλά και η καθιέρωση του διαδικτύου, επανάφεραν πάλι την ιδέα στο προσκήνιο. Φτάνοντας στο σήμερα, ορισμένες εταιρείες κολοσσοί, όπως η Google, η Amazon και άλλες, επενδύουν τεράστια χρηματικά ποσά για την δημιουργία ευφυών συστημάτων, τα οποία πλέον θεωρούνται ιδιαιτέρως αποτελεσματικά, αφού δίνουν δυνατότητες γνώσης, οι οποίες είναι θεωρητικά αδύνατον να εξερευνηθούν από τον ανθρώπινο εγκέφαλο, χωρίς υπολογιστική υποστήριξη.

Ανατρέχοντας κανείς στην βιβλιογραφία περί Τεχνητής Νοημοσύνης, θα διαπιστώσει ότι υπάρχουν διαφορετικοί ορισμοί οι οποίοι προσπαθούν να αποδώσουν διαφορετικές διαστάσεις λειτουργίας των συγκεκριμένων συστημάτων (Κάρλος, 2020). Ενδεικτικά αναφέρεται οι κάτωθι:

- Συστήματα που σκέπτονται σαν τον άνθρωπο
- Συστήματα που λειτουργούν και αποφασίζουν ορθολογικά
- Συστήματα που μιμούνται τον άνθρωπο
- Συστήμα που μιμούνται την λογική του ανθρώπου

Αν συνδυάσουμε τις παραπάνω διαστάσεις, συμπεραίνουμε ότι ο ορισμός που έχει δοθεί από τον Παγκόσμιο Οργανισμό Οικονομικής Συνεργασίας και Ανάπτυξης (OECD) είναι πιθανότατα ο πιο πλήρης, σύμφωνα με τον οποίο «Τεχνητή Νοημοσύνη είναι ένα υπολογιστικό σύστημα, το οποίο χρησιμοποιώντας ένα σύνολο δεδομένων, προσδιορισμένο από έναν άνθρωπο ή ένα άλλο σύστημα, μπορεί να λειτουργεί ορθολογικά και να λαμβάνει αποφάσεις, οι οποίες επηρεάζουν διάφορα άλλα συστήματα εικονικά ή πραγματικά».

Τα επίπεδα αυτονομίας και αποτελεσματικότητας των αποφάσεων, έχουν να κάνουν με το σχεδιασμό των συστημάτων καθώς και με τον όγκο των προς επεξεργασία δεδομένων. Ειδικότερα, μπορούμε να κατηγοριοποιήσουμε τα συστήματα Τεχνητής Νοημοσύνης, ανάλογα με τον βαθμό αυτονομίας σε τρεις μεγάλες κατηγορίες. Στην πρώτη κατηγορία υπάγονται τα απλά ή γειτονικά συστήματα Τεχνητής Νοημοσύνης (Artificial Narrow

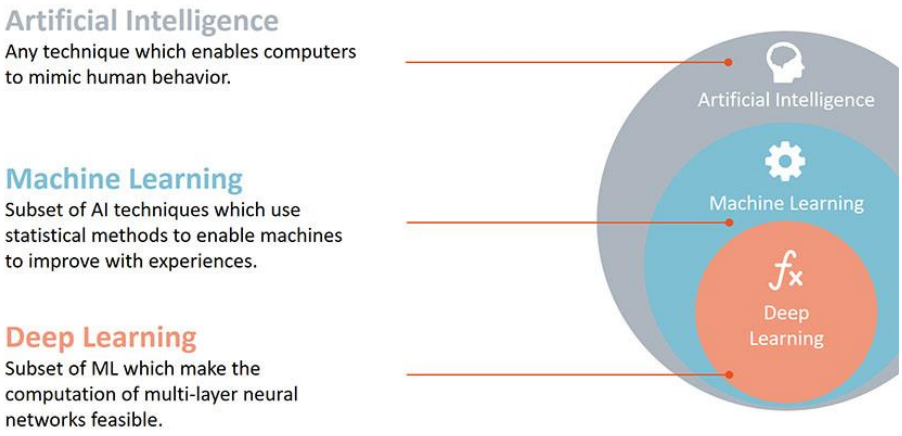
Intelligence – ANI), των οποίων η αποτελεσματικότητα είναι υψηλή όταν εφαρμόζονται σε περιορισμένο εύρος δεδομένων και χαρακτηρίζονται εν γένει από υψηλό βαθμό αυτονομίας. Βασικό χαρακτηριστικό είναι ότι επιτελούν μια και μόνο μία διαδικασία πολύ καλύτερα από τον άνθρωπο. Στην συνέχεια, τα συστήματα γενικής Τεχνητής Νοημοσύνης (Artificial General Intelligence – AGI), των οποίων η αποτελεσματικότητα κυμαίνεται σε ικανοποιητικά – αποδεκτά επίπεδα, αλλά για ένα μεγάλο εύρος εφαρμογών και η αυτονομία τους είναι σχετικά περιορισμένη, διότι απαιτούν ερμηνεία για την λήψη αποφάσεων. Κύριο στοιχείο σχεδιασμού αυτών των συστημάτων, είναι η προσομοίωση των ανθρώπινων ενεργειών και αποφάσεων σε πολλά επίπεδα. Η τελευταία κατηγορία συστημάτων, αποτελεί την μετεξέλιξη της γενικής Τεχνητής Νοημοσύνης και ονομάζεται εξαιρετικά ευφυή συστήματα Τεχνητής Νοημοσύνης (Artificial Super Intelligence - ASI). Πρόκειται για μια πλήρως αυτόνομη και αποτελεσματική κατηγορία, η οποία λειτουργεί σε όλα τα επίπεδα καλύτερα και ταχύτερα από τον ανθρώπινο νου. Στην πραγματικότητα, αυτή η κατηγορία αποτελεί την στόχευση των επιστημόνων για το άμεσο μέλλον, ωστόσο δεν μπορούμε να θεωρήσουμε ότι βρισκόμαστε επί του παρόντος στο συγκεκριμένο επίπεδο.

Από τεχνικής σκοπιάς, δηλαδή ως προς τον τρόπο υλοποίησης, θα μπορούσαμε να διακρίνουμε τους αλγορίθμους Τεχνητής Νοημοσύνης σε δύο επιμέρους κατηγορίες. Στους αλγορίθμους που υλοποιούνται με συμβολικό τρόπο και σε αυτούς που υλοποιούνται με μη συμβολικό ή υπολογιστικό τρόπο. Στην πρώτη κατηγορία, οι αλγόριθμοι δημιουργούνται με τρόπο ώστε να χρησιμοποιούν τον τρόπο σκέψης το ανθρώπου, προκειμένου να λαμβάνουν αποφάσεις. Αντίθετα, στην δεύτερη κατηγορία, οι προσπάθειες επικεντρώνονται στην αποτύπωση ορισμένων βιολογικών και όχι νοητικών ικανοτήτων και χαρακτηριστικών του ανθρώπου, και στην προσπάθεια εξέλιξης αυτών. Χαρακτηριστική περίπτωση αποτελούν, τα τεχνητά δίκτυα νευρώνων, τα οποία είναι εμπνευσμένα από το βιολογικό νευρωνικό σύστημα του ανθρώπινου εγκεφάλου.

Ολοκληρώνοντας, σημειώνουμε ότι οι διαστάσεις της ευφυίας και των έξυπνων αποφάσεων που λαμβάνουν τα συστήματα Τεχνητής Νοημοσύνης, στηρίζονται στον συνδυασμό γνώσεων οι οποίες έχουν προκύψει από την χρήση αλγορίθμων Μηχανικής Μάθησης (Machine Learning) ή Βαθιάς Μάθησης (Deep Learning). Στην Εικόνα 1, παρουσιάζονται τα επίπεδα μάθησης και δημιουργίας γνώσης, μέσα από την χρήση δεδομένων, μέχρι να φτάσουμε στο επίπεδο της ευφυίας που ταυτίζεται με την Τεχνητή Νοημοσύνη και είναι το ανώτερο.

Όπως μπορούμε να παρατηρήσουμε στην Εικόνα 1, η δημιουργία ενός ευφυούς συστήματος, επιτυγχάνεται σε τρίτο στάδιο και έχει απαραίτητα προ απαιτούμενα, την

λειτουργία ενός αλγορίθμου Μηχανικής Μάθησης, για εξαγωγή γνώσης από τα δεδομένα (επίπεδο 1) και εν συνεχεία την χρήση μεθόδων για κριτική αξιολόγηση της γνώσης που έχει εξαχθεί (επίπεδο 2). Η ταυτόχρονη λειτουργία και ανατροφοδότηση μεταξύ των δύο αυτών επιπέδων, αποτελεί στην πραγματικότητα ένα σύστημα Τεχνητής Νοημοσύνης.



Εικόνα 1 Επίπεδα μάθησης και δημιουργίας γνώσης μέσα από ανάλυση δεδομένων

(**Πηγή**: https://rapidminer.com/)

1.2 Σύγχρονες εφαρμογές της Τεχνητής Νοημοσύνης με χρήση αλγορίθμων Μηχανικής Μάθησης

Στην προηγούμενη ενότητα, σημειώθηκε ότι οι εφαρμογές της Τεχνητής Νοημοσύνης, αποκτούν όλους και περισσότερο έδαφος σε πολλά και διαφορετικά πεδία. Επίσης, αναφέρθηκε ότι η δημιουργία ευφυών συστημάτων, απαιτεί σε μεγάλο βαθμό την χρήση αλγορίθμων Μηχανικής Μάθησης και την κριτική αξιολόγηση της αποτελεσματικότητας αυτών. Στην παρούσα ενότητα, παρουσιάζονται ορισμένα από τα πιο διάσημα συστήματα Τεχνητής Νοημοσύνης, τα οποία βασίζονται σε αλγορίθμους Μηχανικής Μάθησης, σε ένα πιο τεχνικό επίπεδο. Προφανώς, οι εφαρμογές και οι μελέτες περίπτωσης που μπορεί να βρει κανείς ανατρέχοντας στο διαδίκτυο, είναι πάρα πολλές, ωστόσο επιλέχθηκαν να καταγραφούν μόνο ορισμένες από αυτές, κυρίως διότι θεωρούνται αρκετά εξελιγμένες και επειδή έφεραν σε μεγάλο βαθμό μια πρωτοποριακή τάση, στους κλάδους που εφαρμόστηκαν.

Επεξεργασία και κατανόησης φυσικής γλώσσας (Natural Language Processing)

Αποτελεί μία από τις πιο καινοτόμες εφαρμογές της Τεχνητής Νοημοσύνης, κατά την οποία σχεδιάζονται και εκπαιδεύονται αλγόριθμοι, οι οποίοι να μπορούν να έρθουν σε απευθείας επικοινωνία με τους ανθρώπους, χωρίς να είναι απαραίτητη κάποια διαδικασία κωδικοποίησης

των λέξεων. Για να καταστεί εφικτή η επικοινωνία ανθρώπου και υπολογιστή, χρησιμοποιούνται αλγόριθμοι, οι οποίοι λαμβάνοντας ως πηγές κείμενα και άλλα γραπτά εγχειρίδια, προσπαθούν να κατανοήσουν την φυσική διάλεκτο. Στην πραγματικότητα, οι προσπάθειες επικεντρώνονται στην δημιουργία συστάδων λέξεων, οι οποίες μέσω ενός κοινού χαρακτηριστικού, ερμηνεύονται ως προς την φυσική τους σημασία. Στην συνέχεια, συνδυάζοντας το σύνολο των λέξεων μιας πρότασης ή φράσης και αντιστοιχίζοντας με την αντίστοιχη ερμηνεία την κάθε λέξη, γίνεται προσπάθεια διερμηνείας της φυσικής γλώσσας από τον αλγόριθμο. Όπως γίνεται αντιληπτό, αυτή η διαδικασία γίνεται εξαιρετικά σύνθετη, αν ληφθούν υπόψιν οι διάφοροι ιδιωματισμοί που παρατηρούνται στις εκάστοτε φυσικές γλώσσες. Για τον λόγο αυτό οι αλγόριθμοι χρειάζονται πάρα πολλά δεδομένα, για να λειτουργούν σωστά. Τα τελευταία χρόνια, με την ανάπτυξη του διαδικτύου και των έγγραφων πηγών (άρθρα, σχόλια σε δημοσιεύσεις κ.λπ.) που δημοσιεύονται, οι αλγόριθμοι έχουν διαθέσιμο τεράστιο όγκο δεδομένων προς εκπαίδευση, γεγονός που έχει οδηγήσει στην βελτίωση της αποτελεσματικότητάς τους. Μπορούμε να διατυπώσουμε διάφορα παραδείγματα κειμενογράφων και μεταφραστών, με μία από τις σημαντικότερες και ευρέως γνωστές εφαρμογές να είναι ο αλγόριθμος μετάφρασης της Google (Google Translate algorithm).

Συστήματα συστάσεων

Τα συστήματα συστάσεων αποτελούν την κύρια κατηγορία αλγορίθμων με την οποία αλληλοεπιδρά ένας χρήστης κατά την περιήγηση του στο διαδίκτυο. Μπορούμε να παρατηρήσουμε τέτοιους αλγορίθμους στις προτάσεις που γίνονται σε διάφορα μέσα κοινωνικής δικτύωσης, όπως το YouTube, το Instagram, το Tweeter και άλλα, ακόμα και στις μηχανές αναζήτησης πληροφοριών της Google.

Αποτελούν μια κατηγορία συστημάτων, τα οποία προσπαθούν να συλλέξουν πληθώρα και ποικιλία δεδομένων για κάθε χρήστη, βάσει των οποίων καθορίζουν τις προτιμήσεις του και συνθέτουν το δυναμικό του προφίλ. Εν συνεχεία του προτείνουν πράγματα που να είναι αρκετά συναφή με το προφίλ του, βάσει προτιμήσεων και κριτικών που έχουν διαπιστωθεί από άλλους χρήστες με παρεμφερές προφίλ. Η ανάπτυξη του διαδικτύου συνέδραμε καθοριστικά στην εξέλιξη και εδραίωση τους. Όπως αναφέρθηκε, ένα τέτοιο σύστημα χρειάζεται πληθώρα ασυσχέτιστων, κυρίως, δεδομένων για να λειτουργεί αποτελεσματικά. Με την χρήση του διαδικτύου στην καθημερινότητα, κάτι τέτοιο είναι εφικτό για τον κάθε απλό χρήστη, ο οποίο κάνει σχόλια, likes ή απλές αναζητήσεις στις μηχανές της Google.

Τα τελευταία χρόνια, η χρήση των συστημάτων συστάσεων, θεωρείται ως μια από τις τεχνολογίες αιχμής, γεγονός που έχει εντείνει την έρευνα γύρω από την ανάπτυξη τους. Αποτέλεσμα, είναι η δημιουργία πολλών και διαφορετικών προσεγγίσεων για την υλοποίηση των συστημάτων, όπως για παράδειγμα τα συστήματα συνεργατικού φιλτραρίσματος, τα συστήματα βασισμένα στην γνώση ή στο περιεχόμενο και διάφορες άλλες πιο σύνθετε υβριδικές προσεγγίσεις.

Τεχνολογία Blockchain

Πρόκειται για την πιο κλασική εφαρμογή της Τεχνητής Νοημοσύνης σε βιομηχανικό επίπεδο. Αποτελείται από συνεργαζόμενους αλγορίθμους, οι οποίοι βελτιστοποιούν την χρήση των διαθέσιμων πόρων (υλικών και άυλων) των οργανισμών και συντονίζουν την ροή των υλικών σε όλο το μήκος της εφοδιαστικής αλυσίδας, από τον αρχικό παραγωγό ως τον τελικό καταναλωτή. Τα δεδομένα μεταφέρονται σε πραγματικό χρόνο, σε όλα τα μπλοκ της αλυσίδας, δίνοντας έτσι την ευκαιρία του καλύτερου μετασχηματισμού των διαδικασιών.

Η εφαρμογή του Blockchain, θεωρείται ότι συνεισφέρει καθοριστικά, στον σχεδιασμό ευέλικτων παραγωγικών διαδικασιών καθώς και στην τεκμηριωμένη λήψη αποφάσεων μεταξύ όλων των εμπλεκομένων. Επίσης, θεωρείται ότι μέσω της μεθόδου οι οργανισμοί μπορούν να ολοκληρώσουν τις διαδικασίες τους και να μειώσουν σημαντικά τα κόστη τους. Για τον λόγο αυτό, αρκετές πάρα πολύ μεγάλες εταιρείες έχουν επενδύσει τεράστια χρηματικά ποσά, για την ανάπτυξη εφαρμογών blockchain.

Στα προσεχή χρόνια, με την περαιτέρω ανάπτυξη του διαδικτύου των πραγμάτων (IoT), οι εφαρμογές αυτές αναμένεται να λαμβάνουν αποφάσεις σε πρώτο χρόνο, σχετικά με τον προγραμματισμό των παραγωγών και των δρομολογίων, επιλύοντας έτσι ένα από τα κυριότερα ζητήματα αποφάσεων, στον βιομηχανικό τομέα. Κλασσικά συστήματα, που υπάγονται στην βιομηχανική λειτουργία, πρόκειται να ολοκληρωθούν, βελτιώνοντας σημαντικά την αποδοτικότητα των οργανισμών.

Ανάλυση φυσικής εικόνας και ανάπτυξη ηλεκτρονικών παιχνιδιών

Η ανάλυση της φυσικής εικόνας, αποτελεί μία από τις πρώτες απόπειρες εφαρμογής της Τεχνητής Νοημοσύνης. Οι αλγόριθμοι αυτής της κατηγορίας, προσπαθούν να αναγνωρίσουν γεωγραφικά τοπία, πρόσωπα και αντικείμενα μέσα από ένα στιγμιότυπο (φωτογραφία) ή από ένα σύνολο στιγμιότυπων (βίντεο). Για την αναγνώριση, αναλύεται κάθε ένα από τα pixel μιας εικόνας, και αποθηκεύεται προσωρινά. Στην συνέχεια, γίνεται προσπάθεια ταυτοποίησης των

αποθηκευμένων pixel, με άλλα pixel που είναι καταχωρημένα και τακτοποιημένα σε μια βάση δεδομένων. Όσο περισσότερα pixel αντιστοιχιστούν με τα αποθηκευμένα, τόσο πιο εύκολα μπορεί να γίνει η ταυτοποίηση μεταξύ των εικονιζόμενων στοιχείων. Χαρακτηριστικό παράδειγμα τέτοιας εφαρμογής, παρατηρείται σε όλες τις συλλογές εικόνων των σύγχρονων κινητών τηλεφώνων (smartphones), εκεί όπου για κάθε νέα λήψη, αποδίδεται μια ονομασία στην εικόνα, με βάση τα στοιχεία του απεικονιζόμενου προσώπου ή γεωγραφικού τοπίου.

Βασισμένη, σε εντελώς παρόμοια λογική είναι και η δημιουργία των ηλεκτρονικών παιχνιδιών. Μόνο που στην περίπτωση αυτή, σχεδιάζονται αλγόριθμοι με στόχο να δημιουργήσουν pixels, τα οποία να ταυτοποιούνται σε μεγάλο βαθμό με την διαθέσιμη πληροφορία που υπάρχει. Οι τελευταίες εκδόσεις διάσημων ηλεκτρονικών παιχνιδιών, φαίνεται να έχουν προσεγγίσει σε μεγάλο βαθμό την φυσικότητα των προσώπων και των τοπίων (βλ. Εικόνα 2). Επίσης, φαίνεται η τεχνολογία στα ηλεκτρονικά παιχνίδια να προχωράει σε ακόμα υψηλότερα επίπεδα, συνδυάζοντας στοιχεία όπως η ψυχολογική διακύμανση των παικτών, οι αντιδράσεις τους σε παρόμοιες παρελθοντικές καταστάσεις και άλλα. Στην πραγματικότητα, αυτές τα δεδομένα λαμβάνονται από τους αλγορίθμους, μέσω της αλληλεπίδρασης των χρηστών με τις συσκευές εισόδου των παιχνιδιών (πληκτρολόγιο, τηλεχειριστήριο κ.λπ.) και τον τρόπο χειρισμού τους.

Εικόνα 2 Στιγμιότυπο από την τελευταία έκδοση διάσημου ηλεκτρονικού παιχνιδιού

1.3 Μηχανική Μάθηση

Η μάθηση αποτελεί μία από τις κυριότερες νοητικές λειτουργίες του ανθρώπου, η οποία σχετίζεται με τους τρόπους και τα μέσα με τα οποία ο ανθρώπινος εγκέφαλος επεξεργάζεται τις πληροφορίες, τα ερεθίσματα και τα βιώματα που λαμβάνει και στην συνέχεια τα χρησιμοποιεί αυτόματα προκειμένου να βελτιώσει ορισμένες από τις δραστηριότητες που εκτελεί με υψηλή συχνότητα στην καθημερινότητά του. Αυτή η ερμηνεία της μάθησης επαφίεται κατά κύριο λόγο στην ψυχολογική διάσταση, ωστόσο αποτελεί τον θεμέλιο λίθο για την δημιουργία των μοντέλων μηχανικής μάθησης.

Είναι γενικά αποδεκτό, ότι η ισχυρή πλειοψηφία των συστημάτων μηχανικής, που έχουν δημιουργηθεί από τον άνθρωπο, είναι κατά μεγάλο βαθμό εμπνευσμένα από παρατηρήσεις λειτουργίας αντίστοιχων φυσικών συστημάτων. Έτσι, και στην περίπτωση της μηχανικής

μάθησης, έχουν δημιουργηθεί αλγόριθμοι, οι οποίοι προσπαθούν να προσομοιώσουν την λειτουργία του εγκεφάλου κατά την διαδικασία της μάθησης, μέσω της χρήσης υπολογιστή, θεωρώντας ότι κάτι τέτοιο θα ήταν αποδοτικό και λειτουργικό και στην περίπτωση των τεχνητών συστημάτων.

Πιο συγκεκριμένα, οι αλγόριθμοι μηχανικής μάθησης, αποσκοπούν στην ανάλυση και επεξεργασία ενός συνόλου δεδομένων, με έναν προκαθορισμένο τρόπο, με απώτερο σκοπό την εξαγωγή προτύπων, μοτίβων και συσχετίσεων που διέπουν τα δεδομένα και μέσω της κατάλληλη ερμηνείας αυτών, να κάνουν κάποιες προβλέψεις σχετικά με τυχόν τάσεις που υπάρχουν. Με τον τρόπο αυτό, γίνεται μια προσπάθεια ερμηνείας των μελλοντικών συνθηκών, κάτι που γενικά είναι ιδιαίτερα επιθυμητό για την λήψη αποφάσεων σε διάφορους τομείς των επιχειρήσεων και της επιστήμης εν γένει. Είναι λογικό, ότι σε αρκετές περιπτώσεις οι προβλέψεις είτε αποτυγχάνουν είτε δεν παρέχουν την επιθυμητή ακρίβεια. Για τον λόγο αυτό, μια από τις σημαντικότερες λειτουργίες των αλγορίθμων είναι η ικανότητά τους να αποτυπώνουν το ποσοστό σφάλματος, με βάση το οποίο επαναπροσδιορίζεται η λειτουργία τους, με τρόπο που να εγγυάται την προοδευτική μείωση του σφάλματος. Όπως γίνεται αντιληπτό, για να είναι αποδοτική μια τέτοια διαδικασία, απαιτείται μεγάλος όγκος δεδομένων, για την εκπαίδευση του μοντέλου. Συνεπώς, θεωρείται ότι οι αλγόριθμοι, είναι χρήσιμοι κι αποτελεσματικοί όταν εφαρμόζονται σε μεγάλες βάσεις δεδομένων, οι οποίες μάλιστα να έχουν και μεγάλο εύρος μεταβλητών.

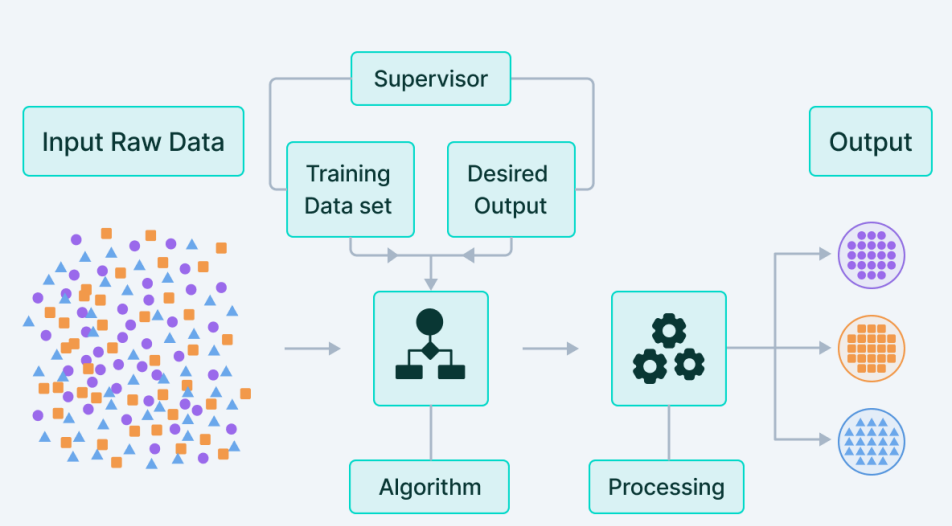
Οι τεχνικές μηχανικής μάθησης, αποτελούν στην πραγματικότητα, την εξέλιξη και την επέκταση των τεχνικών ανάλυσης και εξόρυξης δεδομένων. Τα τελευταία χρόνια, αποτελούν ένα εξαιρετικά ενδιαφέρον ερευνητικό πεδίο, με ολοένα και περισσότερες εφαρμογές και μελέτες περίπτωσης. Αυτό αιτιολογείται υπό την έννοια ότι, με την είσοδο των ηλεκτρονικών υπολογιστών στην συντριπτική πλειοψηφία των καθημερινών λειτουργιών ανθρώπων και οργανισμών αλλά και με την εδραίωση του διαδικτύου, δημιουργείται καθημερινά τεράστιος όγκος δεδομένων, ο οποίος σαφέστατα μπορεί να αποδώσει χρήσιμες πληροφορίες, όταν ακολουθείται ένας αποτελεσματικός τρόπος ανάλυσης. Το τελευταίο διάστημα, έχει δημιουργηθεί πληθώρα συστημάτων, βασισμένα σε αλγορίθμους μηχανικής μάθησης, σε πολλούς και διαφορετικούς τομείς, όπως για παράδειγμα στο εμπόριο, στις εφοδιαστικές αλυσίδες, στη βιομηχανία, στην ιατρική και σε άλλες.

1.4 Ταξινόμηση αλγορίθμων Μηχανικής Μάθησης

Λόγω της χρηστικότητας και της αποτελεσματικότητας τους, τα τελευταία χρόνια έχουν αναπτυχθεί πολλοί και διαφορετικοί αλγόριθμοι, οι οποίοι χρησιμοποιούνται στην επίλυση διαφορετικών προβλημάτων και είναι αποτελεσματικοί σε βάσεις δεδομένων με διαφορετικά χαρακτηριστικά. Είναι λοιπόν αρκετά σημαντική η ταξινόμηση των αλγορίθμων καθώς και η αποτύπωση των προβλημάτων στα οποία είναι αποτελεσματικοί, προκειμένου να μπορεί κανείς να επιλέξει τον κατάλληλο, για να οδηγηθεί σε χρήσιμα συμπεράσματα. Κατά γενική ομολογία, θα μπορούσαν να δημιουργηθούν πολλές και διαφορετικές προσεγγίσεις ως προς τα κριτήρια ταξινόμησης των αλγορίθμων, ωστόσο το κριτήριο που επικρατεί στην πλειοψηφία των βιβλιογραφικών αναφορών, είναι η φύση του προς επίλυση προβλήματος.

Με βάση αυτό το χαρακτηριστικό, μπορούμε να ταξινομήσουμε τους αλγορίθμους σε τρεις μεγάλες κατηγορίες:

- **Αλγόριθμοι επιτηρούμενης μάθησης (supervised learning algorithms):** στην συγκεκριμένη κατηγορία, υπάγονται οι αλγόριθμοι, των οποίων η υλοποίηση βασίζεται στην δημιουργία μιας συνάρτησης εκπαίδευσης με την χρήση διαθέσιμων δεδομένων, τα οποία χρησιμοποιούνται μερικώς σαν είσοδοι (περίπου το 70%) και μερικώς για τον έλεγχο της αποδοτικότητας της συνάρτησης (περίπου το 30%). Όσο μεγαλύτερος ο αριθμός των δεδομένων, τόσο καλύτερα θεωρείται ότι εκπαιδεύεται η συνάρτηση και κατ’ επέκταση λειτουργεί αποτελεσματικά και μπορεί να επεκταθεί και να γενικευθεί. Τα δεδομένα εισόδου, ονομάζονται δεδομένα εκπαίδευσης και έχουν μια γνωστή ετικέτα ή αποτέλεσμα (label). Το μοντέλο προετοιμάζεται μέσω μιας εκπαιδευτικής διαδικασίας στην οποία ο χρήστης παρέχει στον αλγόριθμο εκπαίδευσης τα δεδομένα εισόδου και έτσι ο αλγόριθμος κάνει προβλέψεις, ελέγχει την σύγκλιση των προβλέψεων με τα πραγματικά δεδομένα και προσπαθεί να μειώσει τα σφάλματα και να βελτιώσει τις προβλέψεις, όταν οι αποκλίσεις ξεπεράσουν ένα προκαθορισμένο όριο. Αυτή η διαδικασία εκπαίδευσης συνεχίζεται μέχρι το μοντέλο επιτύχει ένα επιθυμητό επίπεδο ακρίβειας από τα δεδομένα εισόδου. Μόλις τα σφάλματα προσεγγίσουν το επιθυμητό επίπεδο ακριβείας, τότε οι αλγόριθμοι μπορούν να χρησιμοποιηθούν για προβλέψεις με άγνωστα δεδομένα εξόδου. Σύμφωνα με την βιβλιογραφία, οι αλγόριθμοι της επιτηρούμενης μάθησης, χρησιμοποιούνται σε προβλήματα ταξινόμησης (classification), πρόγνωσης (prediction) και διερμηνείας (interpretation).



Εικόνα 3 Διαγραμματική απεικόνιση λειτουργίας αλγορίθμων επιτηρούμενης μάθησης

(**Πηγή**: https://www.v7labs.com/)

Παρόμοιας λογικής με τους αλγορίθμους επιτηρούμενης μάθησης, είναι οι αλγόριθμοι ημι-επιτηρούμενη μάθησης. Η μοναδική διαφορά έγκειται στο ότι χρησιμοποιούνται σε βάσεις δεδομένων που αποτελούνται από δεδομένα με ετικέτα (labeled) και δεδομένα χωρίς ετικέτα (unlabeled). Ως προς την υλοποίηση των αλγορίθμων, αυτό σημαίνει ότι τα επισημασμένα δεδομένα περιέχουν σημαντικές πληροφορίες για το προς επίλυση πρόβλημα και τα χαρακτηριστικά του, , ενώ τα δεδομένα χωρίς ετικέτα(μη επισημασμένα) δεν έχουν αυτές τις πληροφορίες. Με την χρήση αυτού του συνδυασμού οι αλγόριθμοι μαθαίνουν να επισημαίνουν μη επισημασμένα δεδομένα, κάνοντας τα πιο χρήσιμα δεδομένα για την εκπαίδευση του μοντέλου. Έτσι υπάρχει ένα επιθυμητό πρόβλημα πρόβλεψης, αλλά το μοντέλο πρέπει να μάθει τις δομές για να οργανώνει τα δεδομένα καθώς και να κάνει προβλέψεις.

- **Αλγόριθμοι μη επιτηρούμενης μάθησης (unsupervised learning algorithms):**
  Η λογική σχεδιασμού αυτών των αλγορίθμων, διαφέρει σημαντικά από την προηγούμενη κατηγορία. Πιο συγκεκριμένα, στην μη-επιτηρούμενη μάθηση ο αλγόριθμος σχεδιάζεται ώστε να διατρέχει τα δεδομένα και να προσπαθεί να εξάγει μοτίβα σύνδεσης ή συσχετίσεις μεταξύ τους. Δεν υπάρχει κάποια οδηγία ή κάποιος άνθρωπος - χείριστης για την εκπαίδευση. Τα δεδομένα εισόδου δεν φέρουν ετικέτα και δεν έχουν γνωστό αποτέλεσμα. Αντ' αυτού, η μηχανή καθορίζει τις συσχετίσεις και τις σχέσεις αναλύοντας διαθέσιμα δεδομένα. Στη διαδικασία μη επιτηρούμενης εκμάθησης, είναι απαραίτητη η προσπέλαση μεγάλου όγκου δεδομένων,

προκειμένου να δημιουργηθούν μοτίβα και γενικά συμπεράσματα. Όσον αφορά την τεχνική πλευρά της εξαγωγής μοτίβων, οι προσπάθειες επικεντρώνονται στην οργάνωση των δεδομένων με οποιοδήποτε τρόπο που να περιγράψει τη δομή τους. Αυτό μπορεί να σημαίνει είτε ομαδοποίηση των δεδομένων σε συστοιχίες είτε τοποθέτηση με τρόπο τέτοιο που να είναι πιο οργανωμένα.Όσο περισσότερα δεδομένα αξιολογεί ο αλγόριθμος τόσο βελτιώνεται η ικανότητα του να λαμβάνει αποφάσεις σχετικά με τα δεδομένα. Οι συγκεκριμένοι αλγόριθμοι, χρησιμοποιούνται σε προβλήματα, όπως η ομαδοποίηση (clustering), η μείωση διαστάσεων (dimensionality reduction) και η εκμάθηση κανόνων συσχέτισης(association rule learning).

- **Αλγόριθμοι ενισχυτικής μάθησης (reinforced learning algorithms):** Η ενισχυτική μάθηση (reinforcement learning) επικεντρώνεται σε μεθόδους διδασκαλίας, οπού ο αλγόριθμος μηχανικής μάθησης παρέχει ένα σύνολο ενεργειών, παραμέτρων και τελικών τιμών. Καθορίζοντας τους κανόνες, ο αλγόριθμος προσπαθεί να εξερευνήσει διαφορετικές επιλογές και δυνατότητες, παρακολουθώντας και αξιολογώντας κάθε αποτέλεσμα έως ότου προσδιορίσει πιο είναι το βέλτιστο. Η ενισχυτική μάθηση ακολουθεί τη λογική του trial & error. Μαθαίνει από προηγούμενες εμπειρίες και προσαρμόζει την απάντηση στην καλύτερη δυνατή. Χαρακτηριστικό παράδειγμα χρήσης τέτοιων αλγορίθμων, αποτελούν οι ρομποτικοί μηχανισμοί.

Συνοψίζοντας, τα όσα αναφέρθηκαν στην συγκεκριμένη ενότητα, παρατίθεται ο Πίνακας 1, με στόχο την διαγραμματική παρουσίαση της ταξινόμησης των αλγορίθμων μηχανικής μάθησης.

Ταξινόμηση αλγορίθμων Μηχανικής Μάθησης		**Μορφή δεδομένων**	
		Δομημένα - με ετικέτες	**Αδόμητα - χωρίς ετικέτες**
Μεταβλητές βάσης	**Διακριτές**	Επιτηρούμενη μάθηση - classification	Μη επιτηρούμενη μάθηση - clustering
	Συνεχείς	Επιτηρούμενη μάθηση - regression	Μη επιτηρούμενη μάθηση - dimensionality reduction

Πίνακας 1 Ταξινόμηση αλγορίθμων Μηχανικής Μάθησης σύμφωνα με την μορφή των δεδομένων

1.5 Κατηγορίες προβλημάτων Μηχανικής Μάθησης

Όπως προκύπτει από την ανάλυση που προηγήθηκε στην προηγούμενη ενότητα, οι αλγόριθμοι Μηχανικής Μάθησης που έχουν αναπτυχθεί, μπορούν να επιλύσουν συγκεκριμένες κατηγορίες προβλημάτων, ανάλογα με τον τύπο και την δομή των δεδομένων που χρησιμοποιούνται. Οι βασικότερες και ευρέως διαδεδομένες κατηγορίες προβλημάτων, είναι η ταξινόμηση (classification), οι προβλέψεις παλινδρόμησης με βάση συγκεκριμένα χαρακτηριστικά (regression), η ομαδοποίηση (clustering), η μείωση διαστάσεων των μεταβλητών της βάσης (dimensionality reduction), καθώς και η ανάλυση συσχετίσεων (association analysis). Η ταξινόμηση και η παλινδρόμηση, αντιμετωπίζονται μέσω αλγορίθμων επιβλεπόμενης μάθησης ενώ οι υπόλοιπες κατηγορίες προβλημάτων, αντιμετωπίζονται με αλγορίθμους μη επιβλεπόμενης μάθησης (βλ. Εικόνα 3).

Στην παρούσα ενότητα, αναλύεται κάθε μια από τις επιμέρους κατηγορίες ξεχωριστά, ενώ καταγράφονται και ορισμένα από τους πιο γνωστούς αλγορίθμους, για την επιτέλεση του κάθε σκοπού. Στο σημείο αυτό αναφέρουμε, ότι ανάλογα με την εμπειρία του χρήστη σε ένα πρόβλημα και το επίπεδο ικανοτήτων προγραμματισμού, είναι δυνατό να αναπτυχθούν εντελώς νέοι αλγόριθμοι ή παραλλαγές των υφισταμένων, οι οποίοι να λειτουργούν πιο αποτελεσματικά από τις κλασσικές προσεγγίσεις.

1. Ταξινόμηση

Η ταξινόμηση είναι μια διαδικασία κατηγοριοποίησης ενός συνόλου δεδομένων σε κλάσεις, η οποία μπορεί να εφαρμοστεί τόσο σε δομημένα όσο και σε μη δομημένα δεδομένα. Θεωρείται μία από τις πιο απλές κατηγορίες προβλημάτων και υπάρχει πληθώρα εφαρμογών στη βιβλιογραφία. Οι αλγόριθμοι σχεδιάζονται έτσι ώστε να χρησιμοποιούν ένα διαθέσιμο σύνολο δεδομένων, για την δημιουργία των κλάσεων και στην συνέχεια μέσω μιας συνάρτησης κατηγοριοποιούν τα νέα – «άγνωστα» δεδομένα σε κάθε μία από τις επιμέρους κλάσεις. Οι αλγόριθμοι, υπάγονται στην κατηγορία της επιβλεπόμενης μάθησης και χρειάζονται δύο διαφορετικές βάσεις για να λειτουργήσουν αποδοτικά. Η πρώτη βάση χρησιμοποιείται για την εκπαίδευση των μοντέλων και ονομάζεται βάση εκπαίδευσης (training data set).Τα δεδομένα (data points) της δεύτερης βάσης, ταξινομούνται στις επιμέρους κλάσεις που έχουν δημιουργηθεί κατά το στάδιο της εκπαίδευσης. Οι κλάσεις αναφέρονται συχνά ως στόχοι (targets), ετικέτες (labels) ή κατηγορίες (categories). Σαφώς, η εφαρμογή των μοντέλων μπορεί να γίνει και σε μία βάση δεδομένων, αρκεί αυτή να διατιμηθεί σε δύο μέρη. Στην περίπτωση της διάτμησης, σύμφωνα με την διαθέσιμη βιβλιογραφία, προτείνονται ποσόστωσης της τάξης 80-20 ή 70-30, με τα περισσότερα δεδομένα να χρησιμοποιούνται για

την εκπαίδευση. Στις περισσότερες εφαρμογές, όπου οι βάσεις είναι δυναμικές, αυτό συνεπάγεται ότι ο όγκος δεδομένων θα αυξάνεται, οπότε και το μοντέλο θα εκπαιδεύεται καλύτερα, γεγονός που σημαίνει μείωση των λαθών ταξινόμησης και αύξηση αποδοτικότητας.

Βασική προϋπόθεση, για την αποδοτικότητα των μοντέλων ταξινόμησης, είναι η ύπαρξη ενός τουλάχιστον διακριτού χαρακτηριστικού, για να καταστεί εφικτή η δημιουργία των κλάσεων. Ενώ οι μεταβλητές των εγγραφών, μπορούν να είναι είτε διακριτές είτε συνεχείς (Karacapilidis et al., 2013). Στην περίπτωση που το διακριτό χαρακτηριστικό των κλάσεων είναι συνεχές (π.χ. ηλικιακή ομάδα, δείκτης μάζας σώματος κ.λπ.) τότε οι αλγόριθμοι υπάγονται στην κατηγορία της παλινδρόμησης (regression), ενώ στην αντίθετη περίπτωση υπάγονται στην απλή ταξινόμηση (classification).

Επιπρόσθετα, σημειώνεται ότι τα μοντέλα της ταξινόμησης είναι πιο χρήσιμα κυρίως ως επεξηγηματικά εργαλεία, τα οποία εξυπηρετούν περιγραφικούς σκοπούς, ως προς τα δεδομένα. Για παράδειγμα, είναι προτιμότερο και πιο αποδοτικό να χρησιμοποιήσουμε ένα μοντέλο ταξινόμησης για να τοποθετήσουμε την πιστοληπτική ικανότητα ενός δυνητικού δανειολήπτη σε μία από τις κατηγορίες ισχυρή, μέση και χαμηλή, παρά να χρησιμοποιήσουμε ένα τέτοιο μοντέλο για τον υπολογισμό των επιτοκίων ανά περίπτωση.

Ορισμένοι ευρέως διαδεδομένοι αλγόριθμοι ταξινόμησης είναι οι:

- k – Nearest Neighbors (k-NN)
- Logistic Regression
- Decision Trees
- Random Forests
- Super Vector Machines (SVM)
- Naïve - Bayes
- Artificial Neural Networks

2. Παλινδρόμηση

Στην παλινδρόμηση (regression) το μοντέλο σχεδιάζεται με σκοπό να μπορεί να εκτιμά και να κατανοεί τις σχέσεις μεταξύ των δεδομένων. Η παλινδρόμηση ασχολείται με τη διατύπωση μαθηματικών περιορισμών και σχέσεων που υφίστανται μεταξύ των μεταβλητών, οι οποίες βελτιώνονται επαναληπτικά, χρησιμοποιώντας ένα μέτρο σφάλματος στις προβλέψεις που γίνονται από το μοντέλο. Οι μέθοδοι παλινδρόμησης αποτελούνται από ένα σύνολο στατιστικών μεθόδων και εργαλείο, για τον λόγο αυτό σε πολλές βιβλιογραφικές αναφορές συμπεριλαμβάνονται στην στατιστική μηχανική μάθηση.

Στην πραγματικότητα, βασίζονται στην λογική ότι οι περισσότερες από τις μεταβλητές της βάσης έχουν έναν βαθμό εξάρτησης μεταξύ τους, καθώς και ότι επηρεάζουν σε κάποιο βαθμό μια ανεξάρτητη με αυτές μεταβλητή. Η αποτύπωση της μαθηματικής σχέσης μεταξύ των εξαρτημένων και ανεξάρτητων μεταβλητών, χρησιμοποιείται στην συνέχεια για την δημιουργία προβλέψεων ως προς τις ανεξάρτητες. Σε επόμενο στάδιο, χρησιμοποιούνται στατιστικά εργαλεία και μετρικές για την ποσοτικοποίηση της απόδοσης, μέσω της οποίας προκύπτουν αναπροσαρμογές του μαθηματικού μοντέλου. Η διαδικασία εκτελείται κυκλικά (σε loop), έως ότου ο βαθμός ακρίβειας των προβλέψεων να φτάσει στο επιθυμητό επίπεδο.

Κλασικοί αλγόριθμοι προβλέψεων μέσω παλινδρόμησης, θεωρούνται οι:

- Linear Regression
- Polynomial Regression
- Neural Network Regression

3. Ομαδοποίηση - Συσταδοποίηση

Ο βασικός στόχος των αλγορίθμων αυτής της κατηγορίας είναι η ομαδοποίηση ενός συνόλου δεδομένων, με ανομοιογένειες μεταξύ τους, σε συγκεκριμένες συστάδες με συμπαγές και συναφές περιεχόμενο δεδομένων. Με πιο απλά λόγια, από ένα γενικό σύνολο επιδιώκεται η δημιουργία ειδικών ομάδων δεδομένων.

Η συσταδοποίηση είναι εντελώς διαφορετική διεργασία από την ταξινόμηση, καθώς δεν βασίζει την λειτουργία της στην κατάταξη νέων δεδομένων σε ήδη κατασκευασμένες ομάδες δεδομένων, αλλά αποτελεί την κύρια διαδικασία για την δημιουργία των συστάδων δεδομένων. Υπό αυτή την έννοια, θα μπορούσαμε να θεωρήσουμε ότι η συσταδοποίηση είναι μια απαραίτητη διαδικασία για την ορθή διεξαγωγή της ταξινόμησης, στην συνέχεια. Μια επιπλέον ουσιαστική διαφορά, έγκειται στο γεγονός ότι οι αλγόριθμοι που χρησιμοποιούνται για ομαδοποίηση, παρουσιάζουν συχνά πολύ χαμηλή ικανότητα προβλέψεων.

Χαρακτηριστικά παραδείγματα ομαδοποίησης, θεωρούνται η γλωσσολογική και υφολογική ανάλυση κειμένων, ανάλογα με την νοηματική ερμηνεία των λέξεων, η ανάλυση του φυσικού λόγου (Natural Language Processing – NLP), η ηχητική ανάλυση σε μαγνητοσκοπήσεις και άλλα. Ορισμένοι αλγόριθμοι επίλυση αυτών των προβλημάτων είναι:

- K – Means Clustering
- Mean – Shift Clustering
- Gaussian Mixture
- DBSCAN Clustering

4. Μείωση διαστάσεων μεταβλητών

Ορισμένες από τις κατηγορίες προβλημάτων, που αναφέρθηκαν παραπάνω, απαιτούν την χρήση μεγάλου αριθμού δεδομένων εντελώς ασυσχέτιστων μεταξύ τους, προκειμένου να λειτουργήσουν αποτελεσματικά. Ωστόσο, σε άλλες περιπτώσεις, δεν είναι όλα τα χαρακτηριστικά στα σύνολα δεδομένων που δημιουργούνται σημαντικά για την εκπαίδευση των αλγορίθμων μηχανικής εκμάθησης. Ορισμένα χαρακτηριστικά μπορεί να είναι άσχετα και μερικά μπορεί να μην επηρεάζουν το αποτέλεσμα της πρόβλεψης (Reddy et al., 2020). Η ύπαρξη άσχετων μεταβλητών, όταν δεν είναι επιθυμητή, δημιουργία αρκετά μεγάλη πολυπλοκότητα στα μοντέλα, η οποία οδηγεί εν τέλει σε χαμηλά ποσοστά αποτελεσματικότητας στις προβλέψεις. Η παράβλεψη ή η αφαίρεση αυτών των άσχετων ή λιγότερο σημαντικών χαρακτηριστικών μειώνει την επιβάρυνση των αλγορίθμων μηχανικής μάθησης. Για την εύρεση και αφαίρεση των "περιττών" μεταβλητών, έχουν αναπτυχθεί ορισμένοι βοηθητικοί αλγόριθμοι Μηχανικής Μάθησης. Κάποιοι εκ των βασικότερων είναι οι:

- Feature selection and extraction
- Principal Component Analysis (PCA)
- Non-negative matrix factorization (NMF)
- Linear discriminant analysis (LDA)
- Generalized discriminant analysis (GDA)

5. Ανάλυση συσχετίσεων

Σε αυτή την κατηγορία προβλημάτων, το ζητούμενο είναι να ανακαλυφθούν μοτίβα και συσχετίσεις μεταξύ των δεδομένων, τα οποία είναι αδύνατον να εντοπίσει ο ανθρώπινος νους, με την χρήση λογικών συμπερασμάτων. Η επιδίωξη αυτή, στηρίζεται σε στατιστικές θεωρήσεις, σύμφωνα με τις οποίες υπάρχουν συνδέσεις μεταξύ μεταβλητών, μέσω των οποίων η τιμή της μίας επηρεάζει σε κάποιο βαθμό την τιμή κάποιας άλλης ή κάποιων άλλων. Στην πραγματικότητα αυτό ισχύει, ωστόσο μπορεί οι συσχετίσεις να προκύπτουν μέσα από εντελώς διαφορετικά και ανομοιογενή χαρακτηριστικά μεταξύ των δεδομένων. Αυτές, επιδιώκονται να βρεθούν, κάνοντας χρήση τεράστιου όγκου δεδομένων και ανομοιογενών μεταβλητών. Στην αντίπερα όχθη, πολλές φορές συναντώνται και οι σχέσεις μεταξύ συναφών, αλλά όχι όμοιων, ομάδων δεδομένων. Αντίθετα, με τις μη προφανείς περιπτώσεις, τα μοτίβα μεταξύ των συναφών κατηγοριών δεν ενδιαφέρουν τόσο καθώς δεν παρέχουν κάποια ιδιαίτερη γνώση για τα μοτίβα συσχέτισης, παρά μόνο την απολύτως βασική. (Tan, 2006).

Μια από τις πιο γνωστές εφαρμογές, της συγκεκριμένης κατηγορίας, η οποία θεωρητικά ανέδειξε την ισχύ των στατιστικών ισχυρισμών, είναι αυτή της ανάλυσης του καλαθιού του super market (market basket analysis). Σύμφωνα με την συγκεκριμένη ανάλυση επιδιώκεται η συσχέτιση των επιλογών που γίνονται κατά την αγορά προϊόντων, από διάφορους αγοραστές. Ενώ σε δεύτερο στάδιο γίνεται προσπάθεια για την δημιουργία μοτίβων και για γενίκευση αυτών.

Κεφάλαιο 2: Μεθοδολογικό πλαίσιο υλοποίησης αλγορίθμων Μηχανικής Μάθησης

Η υλοποίηση ενός ολοκληρωμένου συστήματος μηχανικής μάθησης, απαιτεί την οργάνωση και την υλοποίηση ενός συνόλου από προκάτοχα στάδια καθώς και στάδια αξιολόγησης της αποτελεσματικότητας των αλγορίθμων. Τα προκάτοχα στάδια των αλγορίθμων, ονομάζονται συχνά και στάδια προ - επεξεργασίας των δεδομένων (data preprocessing), ενώ η αποτελεσματικότητα υπολογίζεται με βάση συγκεκριμένης μετρικές (evaluation metrics). Και τα δύο συγκεκριμένα στάδια, είναι πολύ σημαντικά, διότι εξασφαλίζουν την ερμηνεία των τελικών αποτελεσμάτων και δίνουν την δυνατότητα παρακολούθησης της εξελικτικής πορείας των αποφάσεων των αλγορίθμων. Στο παρόν κεφάλαιο, αναλύονται τόσα τα απαραίτητα προκάτοχα στάδια του αλγορίθμου, όσο και τα απαραίτητα στάδια ελέγχου της αποτελεσματικότητας, για την υλοποίηση ενός ολοκληρωμένου συστήματος. Το μεθοδολογικό πλαίσιο που περιγράφεται στο παρόν κεφάλαιο, έχει εφαρμογή κυρίως σε αλγορίθμους επιτηρούμενης μάθησης, και πιο συγκεκριμένα σε αλγορίθμους ταξινόμησης, με τους οποίους θα ασχοληθούμε και στην συνέχεια της εργασίας.

2.1 Στάδια προ-επεξεργασίας δεδομένων

Όπως αναφέρθηκε και στο εισαγωγικό μέρος του κεφαλαίου, η αποτελεσματική χρήση αλγορίθμων μηχανικής μάθησης, απαιτεί την υλοποίηση μεθόδων για «καθαρισμό» των δεδομένων προκειμένου αυτά να έρθουν στην κατάλληλη μορφή προς επεξεργασία. Σε διάφορες βιβλιογραφικές πηγές, γίνεται αναφορά για την μεθοδολογία που ακολουθείται σχετικά με την προ-επεξεργασία των δεδομένων. Στην παρούσα εργασία, γίνεται εστίαση σε τρεις από τις κυριότερες πηγές προβλημάτων που πηγάζουν από τα ίδια τα δεδομένα και σε επόμενο στάδιο καταγράφονται ορισμένες μεθοδολογίες επίλυσης των προβλημάτων αυτών. Οι κυριότερες πηγές προβλημάτων σχετίζονται με τις χαμένες τιμές (missing values), με τον θόρυβο στα δεδομένα και τέλος με την βέλτιστη επιλογή των μεταβλητών για την επίλυση ενός προβλήματος (Osama et al., 2023). Κάθε μία από τις έννοιες αυτές αναλύονται κάτωθι:

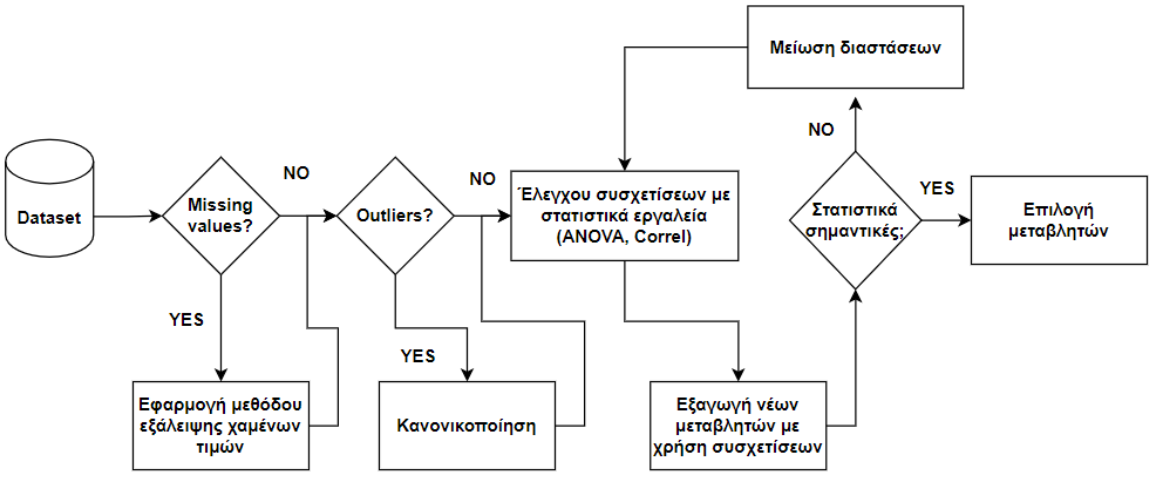
- **Χαμένες τιμές:** Είναι πολύ συχνό φαινόμενο η ύπαρξη ορισμένων «κενών» τιμών, ιδίως σε βάσεις δεδομένων οι οποίες σχετίζονται με πραγματικά προβλήματα, όπως για παράδειγμα από τον χώρο της βιομηχανίας. Όπως γίνεται αντιληπτό πολλά είναι τα αίτια που οδηγούν στην ύπαρξη κενών τιμών, ενδεικτικά μπορούμε να αναφέρουμε ότι κάποιες φορές δεν είναι διαθέσιμα όλα τα απαραίτητα δεδομένα κι έτσι στην βάση συγκρατούνται μόνο τα γνωστά στοιχεία, επιπρόσθετα το ανθρώπινο λάθος ως προς

την καταχώρηση των δεδομένων μπορεί να παίξει σημαντικό ρόλο στην έλλειψη επαρκούς στοιχείων, καθώς επίσης και οι λανθασμένες αποφάσεις σχετικά με το ποια δεδομένα είναι σημαντικά και ποια όχι, μπορούν να οδηγήσουν στην απώλεια σημαντικής πληροφορίας, η οποία αν δεν υπάρχει πιθανότατα εμποδίζει την αποτελεσματική υλοποίηση των αλγορίθμων μηχανικής μάθησης.

- **Θόρυβος:** Ο θόρυβος αποτελεί ίσως την πιο σημαντική προβληματική διάσταση στο πεδίο της ανάλυσης των δεδομένων. Στην βιβλιογραφία ο όρος θόρυβος, χρησιμοποιείται για να περιγράψει δύο συνθήκες που παρατηρούνται στις βάσεις δεδομένων. Η πρώτη συνθήκη σχετίζεται με την ύπαρξη λανθασμένων καταχωρίσεων μέσα στην βάση, οι οποίες μπορεί να σχετίζονται είτε με τον ανθρώπινο παράγοντα είτε με καθαρά τεχνικούς παράγοντες, όπως για παράδειγμα ο τρόπος λειτουργίας κάποιων μηχανικών διατάξεων και του εξοπλισμού γενικότερα, ή τέλος μπορεί να σχετίζονται με τον τρόπο ερμηνείας των δεδομένων που συλλέγονται, ο οποίος επηρεάζει άμεσα και τον τρόπο με τον οποίο αποθηκεύονται και ερμηνεύονται τα δεδομένα, όπως για παράδειγμα ο τρόπος αποκωδικοποίησης κάποιων σημάτων και ενδείξεων. Τα λάθη που προκαλούν τέτοιου είδους θόρυβο, είναι συνήθως πάρα πολύ δύσκολο να εντοπιστούν, διότι είναι άμεσα συνδεδεμένα με την λειτουργία και χαρακτηρίζονται ως λογικά λάθη εκχώρησης. Η δεύτερη συνθήκη για την δημιουργία θορύβου στα δεδομένα, σχετίζεται με την ύπαρξη ακραίων τιμών για κάποιες μεταβλητές μέσα στην βάση (outliers). Αυτή η κατηγορία θορύβου, θα λέγαμε ότι επιδρά πάρα πολύ σημαντικά στην υλοποίηση των αλγορίθμων μηχανικής μάθησης, ιδίως σε αυτούς που ανήκουν στην κατηγορία των επιτηρούμενων αλγορίθμων. Αυτό αιτιολογείται διότι οι αλγόριθμοι κατά την εκπαίδευση τους προσπαθούν να λάβουν την περισσότερη δυνατή πληροφορία σχετικά με τα δεδομένα, και στην προσπάθεια τους αυτή διατρέχουν όλα τα δεδομένα και έτσι «μαθαίνουν» πληροφορία η οποία σχετίζεται και με τα outliers. Ωστόσο, κάτι τέτοιο δεν είναι σωστό, γιατί τα outliers αποτελούν συνήθως μεμονωμένα περιστατικά και δεν σχετίζονται άμεσα με τις γενικές συνθήκες που επικρατούν στο πρόβλημα. Στην πραγματικότητα, αν τα outliers δεν εξομαλυνθούν, οι αλγόριθμοι που προκύπτουν είναι υπέρ-εστιασμένοι και μη αποτελεσματικοί για χρήση σε άλλες βάσεις δεδομένων που περιγράφουν το ίδιο πρόβλημα ή ακόμα και σε μεταγενέστερα δεδομένα της ίδιας βάσης.
- **Βέλτιστη επιλογή μεταβλητών:** Μόλις ολοκληρωθεί η διαδικασία επίλυσης των προαναφερθέντων προβληματικών διαστάσεων, στο επόμενο στάδιο ακολουθεί η

επιλογή του βέλτιστου αριθμού μεταβλητών για την επίλυση του προβλήματος και την εξαγωγή χρήσιμων και ουσιαστικών συμπερασμάτων. Στην πραγματικότητα, οι βάσεις δεδομένων, ιδίως αυτές που αντλούνται από επιχειρησιακούς χώρους, περιέχουν πολύ μεγάλο αριθμό μεταβλητών, ο οποίος μπορεί ως ένα βαθμό να σχετίζεται με το πρόβλημα που επιλύεται αλλά μπορεί να μην σχετίζεται και σχεδόν καθόλου, αλλά να φορά πληροφορία που είναι χρήσιμη για τον οργανισμό σχετικά με άλλες αποφάσεις που λαμβάνει. Η επιλογή των μεταβλητών για την επίλυση ενός προβλήματος μηχανικής μάθησης, θα πρέπει να γίνεται με γνώμονα δύο κριτήρια. Αφενός θα πρέπει ο αριθμός των μεταβλητών να είναι τέτοιος που να περιγράφει όλες τις διαστάσεις του προβλήματος και αφετέρου όλες οι ανεξάρτητες μεταβλητές θα πρέπει να έχουν κάποια άμεση και στατιστικά σημαντική συσχέτιση με τις εξαρτημένες, έτσι ώστε να υποστηρίζεται η εξαγωγή αιτιωδών σχέσεων μεταξύ τους. Ολοκληρώνοντας, σημειώνεται ότι στο πλαίσιο της αποτελεσματικής λειτουργίας των αλγορίθμων μηχανικής μάθησης, θα πρέπει να αποφεύγονται οι αλληλοεπικαλύψεις μεταξύ των μεταβλητών (overlap). Με άλλα λόγια αν παρατηρηθεί συσχέτιση μεταξύ ορισμένων ανεξάρτητων μεταβλητών μέσω μιας συγκεκριμένης μαθηματικής σχέσης, θα πρέπει να προτιμάται η δημιουργία μιας νέας μεταβλητής που θα ενώνει τις δύο επιμέρους μεταβλητές, μέσω της μαθηματικής αυτής σχέσης, έναντι της χρήσης και των δύο μεταβλητών ως ξεχωριστές οντότητες.

Συνοψίζοντας τα όσα αναφέρθηκαν στην παρούσα ενότητα σχετικά με την αρχική επεξεργασία των δεδομένων, παρατίθεται η Εικόνα 4, η οποία αποδίδει με γραφικό τρόπο την προτεινόμενη αλγοριθμικής διαδικασία του σταδίου της προ-επεξεργασίας.



Εικόνα 4 Αλγοριθμική διαδικασία προ-επεξεργασίας δεδομένων

2.2 Μέθοδοι εξάλειψης χαμένων τιμών

Όπως σημειώθηκε και παραπάνω, το πρόβλημα των χαμένων τιμών είναι αρκετά σημαντικό διότι συνεπάγεται την απώλεια χρήσιμης πληροφορίας. Ανατρέχοντας στην βιβλιογραφία, μπορούμε να διαπιστώσουμε ότι έχουν κατά καιρούς προταθεί διαφορετικές μέθοδοι για την εξάλειψη των χαμένων τιμών. Ορισμένες από αυτές είναι αρκετά ολιστικές και θα λέγαμε ότι δεν επιλύουν επί της ουσίας το πρόβλημα μάθησης των αλγορίθμων (Κύρκος, 2015). Τέτοιες προσεγγίσεις μπορεί να προβλέπουν την **διαγραφή ολόκληρων γραμμών**, στις οποίες συμπεριλαμβάνονται κενές τιμές σε κάποιες στήλες, ή να προβλέπουν την συμπλήρωση των κενών κελιών με την χρήση κάποιας **σταθερής τιμής**, έτσι ώστε να είναι συμπληρωμένη πλήρως όλη η βάση δεδομένων και ταυτόχρονα να υπάρχει μια ένδειξη σχετικά με τα άγνωστα δεδομένα. Όπως γίνεται αντιληπτό, αμφότερες οι προσεγγίσεις αυτές αντενδείκνυνται διότι στην πρώτη περίπτωση χάνεται περαιτέρω πληροφορία από αυτή που είχε ήδη χαθεί και στην δεύτερη μεταβιβάζεται μια λανθασμένη πληροφορία, για την οποία οι αλγόριθμοι θα εκπαιδευτούν κανονικά, μιας και είναι αδύνατο να κατανοήσουν ότι αυτά τα δεδομένα δεν αφορούν πραγματικά στοιχεία του προβλήματος, αλλά παίζουν τον ρόλο της ετικέτας επί της ουσίας. Οπότε και στις δύο περιπτώσεις, οδηγούμαστε σε σημαντική μείωση της αποτελεσματικότητας, η οποία μάλιστα σχετίζεται με το στάδιο της εκπαίδευσης.

Στον αντίποδα, σημειώνεται ότι έχουν αναπτυχθεί και ορισμένες άλλες προσεγγίσεις, οι οποίες είναι περισσότερο **ορθολογικές** και ακόμα και αν δεν μπορούν να ανασύρουν πλήρως τη χαμένη πληροφορία, τουλάχιστον μετριάζουν την απώλεια και κυρίως δεν εμποδίζουν την διαδικασία αποτελεσματικής μάθησης των αλγορίθμων παρεμβάλλοντας μη έγκυρη πληροφορία. Σύμφωνα με τον Κύρκο (2015), οι κυριότερες ορθολογικές μέθοδοι επίλυσης του προβλήματος χαμένων τιμών είναι:

- **Αντικατάσταση των χαμένων τιμών με την μέση τιμή της στήλης** που ανήκουν, όταν πρόκειται για συνεχείς μεταβλητές **ή αντικατάσταση με την συνηθέστερη ετικέτα** (label) όταν πρόκειται για διακριτές μεταβλητές, οι οποίες σχετίζονται κατά βάση με αλφαριθμητικά χαρακτηριστικά. Αυτή η μέθοδος βασίζεται στην λογική ότι οι δειγματικοί μέσοι για κάθε πληθυσμό, έχουν πάντοτε την τάση να παρουσιάζουν μια μορφή τυπικής κανονικής κατανομής, το οποίο με άλλα λόγια σημαίνει ότι η ισχυρή πλειοψηφία των παρατηρήσεων συγκεντρώνεται γύρω από την μέση τιμή, συνεπώς αν τεθεί η μέση τιμή σαν αντικατάσταση των χαμένων τιμών, τότε μετριάζεται σημαντικά το ποσοστό της χαμένης πληροφορίας.

- **Δημιουργία κάθε πιθανού συνδυασμού μεταξύ των στηλών, για τον προσδιορισμό των χαμένων τιμών.** Αυτή η μέθοδος αν και είναι αποτελεσματική σε σχετικά μικρές βάσεις δεδομένων, είναι πρακτικά αδύνατη προς υλοποίηση σε μεγάλες βάσεις με πολυδιάστατη πληροφορία, καθώς αυξάνει έντονα τις υπολογιστικές απαιτήσεις για την δημιουργία των συνδυασμών και επιπλέον μπορεί να οδηγήσει σε υπολογισμό τιμών μη εύκολα ερμηνεύσιμων. Στην πραγματικότητα, θα πρέπει να υπάρχει κάποιο είδος συσχέτισης μεταξύ των μεταβλητών προκειμένου να λειτουργήσει αποτελεσματικά η μέθοδος.
- **Χρήση αλγορίθμων τεχνητής νοημοσύνης για την αυτόματη συμπλήρωση των χαμένων τιμών.** Αν η αναλογία των χαμένων τιμών προς τις υπαρκτές είναι σχετικά μικρή (μικρότερη από 20%), τότε οι διαθέσιμοι αλγόριθμοι μπορούν με αρκετά υψηλά επίπεδα να συμπληρώσουν αυτόματα τις χαμένες τιμές μέσω προβλέψεων. Στην ουσία, σε αυτήν την περίπτωση παρεμβάλλεται ένα επιπλέον στάδιο υλοποίησης των αλγορίθμων, κατά το οποίο η βάση εκπαίδευσης του μοντέλου, διαιρείται σε μια μικρότερη βάση εκπαίδευσης, η οποία περιέχει τα γνωστά δεδομένα και σε μια μικρότερη βάση ελέγχου, η οποία περιέχει τις μεταβλητές με τις απολεσθείσες τιμές. Έτσι, χρησιμοποιώντας κάποιον αλγόριθμο επιτηρούμενης μάθησης για πρόβλεψη (π.χ. κάποια μορφή παλινδρόμησης) μπορούν να γίνουν ακριβείς προβλέψεις σχετικά με τις χαμένες τιμές, ιδίως αν η αρχική βάση δεδομένων είναι αρκετά μεγάλη σε έκταση. Μόλις συμπληρωθούν οι κενές τιμές, τότε ολόκληρη η βάση χρησιμοποιείται για επανεκπαίδευση του μοντέλου, με στόχο πλέον την πρόβλεψη εντελώς άγνωστων μεταβλητών.

Ολοκληρώνοντας, σημειώνεται ότι υπάρχει καταγεγραμμένη και μια τελευταία μέθοδος εξάλειψης του φαινομένου, η οποία υπό τις κατάλληλες συνθήκες αποτελεί την βέλτιστη δυνατή επιλογή συγκριτικά με όλες όσες αναφέρθηκαν ανωτέρω. Η μέθοδος αυτή σχετίζεται με την εύρεση των πραγματικών τιμών για τα απολεσθέντα δεδομένα. Η διαδικασία αυτή μπορεί να υλοποιηθεί μέσω κάποιου είδους έρευνας στον οργανισμό από τον οποίον έχουν αντληθεί τα δεδομένα ή με την χρήση τεχνικών κατιδεασμού (brainstorming). Ωστόσο, αν κι η μέθοδος αυτή θα ήταν η ιδανική, στην πραγματικότητα οι αναλύσεις γίνονται ασύγχρονα σχετικά με την πραγματική ημερομηνία των δεδομένων και κατά κύριο λόγω οι αναλύσεις λαμβάνουν χώρα σε δεδομένα παρελθόντος χρόνου, στοχεύοντας στην εξερεύνηση κάποιων μοτίβων και πλαισίων τα οποία θα είναι χρήσιμα για την πρόβλεψη μελλοντικών συνθηκών.

Κατά συνέπεια, η παραπάνω μέθοδος θα λέγαμε ότι είναι **ουτοπική**, για τις περισσότερες κατηγορίες προβλημάτων.

2.3 Μέθοδοι εξάλειψης θορύβου από τα δεδομένα

Όπως αναλύθηκε και στην πρώτη ενότητα του παρόντος κεφαλαίου, οι δύο βασικότερες πηγές θορύβου στα δεδομένα θεωρούνται η ύπαρξη λανθασμένων εγγραφών, οι οποίες οφείλονται είτε στον ανθρώπινο παράγοντα είτε σε κάποια δυσλειτουργία τεχνική καθώς επίσης και η ύπαρξη ακραίων τιμών, οι οποίες εμποδίζουν τους αλγορίθμους να εστιάσουν στο «ουσιαστικό» τμήμα της βάσης και να εκπαιδευτούν αποτελεσματικά. Στην πραγματικότητα, η πρώτη κατηγορία θορύβου αφορά λογικά λάθη τα οποία είναι πολύ δύσκολο να εντοπιστούν και ακόμα πιο σύνθετες είναι οι απαιτούμενες παρεμβάσεις για την εξάλειψη και διόρθωσή τους. Πιο συγκεκριμένα, μπορούμε να θεωρήσουμε ότι απαιτούν μια σειρά από υποθετικά σενάρια, τα οποία επιφορτίζουν την διαδικασία της εκπαίδευσης με επιπρόσθετη αβεβαιότητα. Στην παρούσα εργασία, δεν θα γίνει αναφορά σε βάθος για τέτοιου είδους τεχνικές, αλλά η εστίαση θα επικεντρωθεί στην εξάλειψη των ακραίων – ασυνήθιστων τιμών (outliers).

Η εξάλειψη του θορύβου από τα δεδομένα, λόγω της ύπαρξης των ακραίων τιμών, μπορεί να γίνει χρησιμοποιώντας κάποια μορφή κανονικοποίησης. Αρχικά, θα πρέπει να σημειωθεί ότι οι ακραίες τιμές είναι φαινόμενο που παρατηρείται σε συνεχείς μεταβλητές, ενώ δεν έχει νόημα να οριστεί το πρόβλημα των ακραίων τιμών σε διακριτές μεταβλητές. Στην περίπτωση των διακριτών μεταβλητών, δεν υπεισέρχεται θόρυβος στα δεδομένα λόγω αυτής της συνθήκης. Αναφορικά με την κανονικοποίηση, αυτή αποτελεί μια διαδικασία που μπορεί να υλοποιηθεί με χρήση διάφορων τεχνικών. Στο πλαίσιο της εργασίας, θα αναφέρουμε δύο από τις πλέον κλασικές τεχνικές, οι οποίες είναι η **κανονικοποίηση με βάση την τιμή z** (z-score) και η **κανονικοποίηση με υποβιβασμό της τάξης μεγέθους** (βλ. Πίνακα 2)**.**

Μέθοδος κανονικοποίησης	**Τύπος υπολογισμού**
Με βάση την τιμή z	$z = \frac{x - \mu}{\sigma}$ Όπου: μ: μέση τιμή μεταβλητής και σ: τυπική απόκλιση τιμών μεταβλητής

Με υποβιβασμό τάξης μεγέθους	$\begin{cases} x_{τελικό} = \frac{x_{αρχικό}}{10^k} \\ k = \frac{\ln\left(x_{max_{αρχικό}}\right)}{\ln(10)} \end{cases}$
	Όλες οι νέες τιμές: μεταξύ 0 και 1

Πίνακας 2 Τύπου κανονικοποίησης ακραίων τιμών δεδομένων

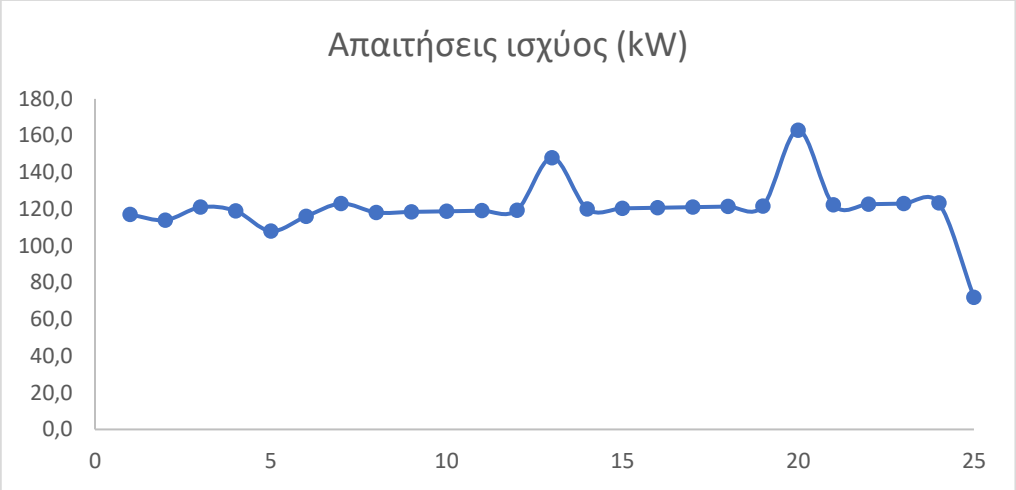
(Πηγή: Κύρκος, 2015)

Επιπρόσθετα, σημειώνεται ότι πέραν των παραπάνω τεχνικών, υπάρχουν και άλλες τεχνικές κανονικοποίησης όπως είναι τεχνική με τις ελάχιστες και μέγιστες τιμές και άλλες. Επιπλέον, σε ορισμένα προβλήματα και εφ' όσον κρίνεται από την ομάδα ανάλυσης ότι η διαδικασία κανονικοποίησης δεν έχει μεγάλη σημασία, είναι δυνατόν ακόμη και να απαλειφθούν εντελώς οι ακραίες τιμές. Αυτή η μέθοδος μπορεί να είναι χρήσιμη είτε στην περίπτωση που οι ακραίες τιμές είναι ελάχιστες έναντι των υπολοίπων είτε στην περίπτωση που είναι άμεσα συσχετισμένες με ειδικές συνθήκες κάποιων ημερών. Για παράδειγμα μια ακραία ένδειξη στους αισθητήρες μιας μηχανής συσκευασίας σε ένα εργοστάσιο, μπορεί να είναι άμεσα συσχετισμένη με κάποια ειδική θερμοκρασιακή ή λειτουργική συνθήκη (διαρροή ρελέ διαφυγής) που επικράτησε σε κάποια συγκεκριμένη μέρα λειτουργίας. Αν αυτή η συνθήκη δεν αποτελεί την καθημερινή πραγματικότητα, τότε τα αντίστοιχα δεδομένα είναι προτιμότερο να απαλειφθούν εντελώς προκειμένου να μην αλλοιώσουν την δυνατότητα γενίκευσης των μοντέλων.

2.3.1 Εφαρμογή κανονικοποίησης ακραίων τιμών

Με στόχο την αποσαφήνιση των όσων αναφέρθηκαν ανωτέρω σχετικά με την κανονικοποίηση των τιμών, μέσω συγκεκριμένων τεχνικών, παρατίθεται μια σύντομη μελέτη περίπτωσης, η οποία αφορά στις ενεργειακές απαιτήσεις μιας παραγωγικής μονάδας σε καθημερινή βάση, μετρημένη σε kW για διάστημα 25 ημερών, με χρήση κατάλληλα τοποθετημένων αισθητήρων. Όλη η πληροφορία σχετικά με τις αρχικά μετρήσεις παρουσιάζεται στο Σχήμα 1. Όπως προκύπτει από το Σχήμα 1, υπάρχουν τρεις ακραίες τιμές, οι οποίες δεν υπακούν στο γενικότερο μοτίβο περιγραφής των ενεργειακών απαιτήσεων της μονάδας. Οι δύο εξ' αυτών των τιμών ξεφεύγουν σε υψηλότερα επίπεδα από την μέση κατανάλωση και μία είναι σε αρκετά χαμηλότερο επίπεδο, από αυτή. Με γνώμονα ότι στην μηχανική μάθηση το κύριο ζητούμενο είναι η εξαγωγή των μοτίβων, συμπεραίνουμε ότι οι τρεις αυτές τιμές πρόκειται να μειώσουν την αποτελεσματικότητα της διαδικασίας μάθησης και έτσι θεωρούνται outliers και κατ' επέκταση θα πρέπει είτε να απαλειφθούν εντελώς είτε

να εξομαλυνθούν. Εν προκειμένω επιλέγεται η εξομάλυνση των τιμών αυτών, χρησιμοποιώντας τον τύπο υποβιβασμό τάξης μεγέθους που αναφέρθηκε παραπάνω, καθώς φαίνεται να είναι ο πιο κατάλληλο για αυτή την εφαρμογή. Οι εξομαλυμένες τιμές αποδίδονται στον Πίνακα 3.



Σχήμα 1 Εντοπισμός ακραίων τιμών – Εφαρμογή

Για να καταστεί εφικτός ο υπολογισμός των εξομαλυμένων τιμών, χρειάζεται σε αρχικό στάδιο να υπολογιστεί ο συντελεστής k. Με δεδομένο ότι η μέγιστη τιμή απαίτησης σε ισχύ είναι τα 163 kW, όπως προκύπτει από το Σχήμα 1, και εφαρμόζοντας τον τύπο υπολογισμού του k, που δίνεται στον Πίνακα 2, τελικά προκύπτει: k = 2,121.

Ημέρα	**Απαιτήσεις ισχύος (kW)**	**Υποβιβασμός τάξης μεγέθους**
1	117,0	0,718
2	114,0	0,699
3	121,0	0,742
4	119,0	0,730
5	108,0	0,663
6	116,0	0,712
7	123,0	0,755
8	118,1	0,725
9	118,5	0,727
10	118,8	0,729
11	119,1	0,731
12	119,4	0,733
13	148,0	0,908

14	120,1	0,737
15	120,4	0,739
16	120,7	0,741
17	121,0	0,743
18	121,4	0,745
19	121,7	0,746
20	163,0	1,000
21	122,3	0,750
22	122,6	0,752
23	123,0	0,754
24	123,3	0,756
25	72,0	0,442

Πίνακας 3 Κανονικοποίηση ακραίων τιμών - Εφαρμογή

2.4 Τεχνικές αξιολόγησης αλγορίθμων Μηχανικής Μάθησης

Στις προηγούμενες ενότητες του παρόντος κεφαλαίου αναλύθηκαν οι τεχνικές που υλοποιούνται κατά το στάδιο της προ-επεξεργασίας των δεδομένων, το οποίο αποσκοπεί στην δημιουργία της κατάλληλης μορφής και δομής των δεδομένων, έτσι ώστε να μπορούν να υλοποιηθούν οι αλγόριθμοι μηχανικής μάθησης αποτελεσματικά. Στην παρούσα ενότητα, παρουσιάζονται ορισμένες τεχνικές αξιολόγησης της λειτουργίας των αλγορίθμων. Όπως γίνεται κατανοητό, το παρόν στάδιο είναι ακόλουθο της εφαρμογής των αλγορίθμων. Αναφέρεται, επίσης, ότι κατά κύριο λόγο οι μέθοδοι αξιολόγησης θα επικεντρωθούν ως προς τον έλεγχο των αλγορίθμων ταξινόμησης, διότι σε αυτούς επικεντρώνεται και το υπόλοιπο μέρος της εργασίας.

Το κυριότερο στοιχείο, από το οποίο προκύπτουν οι κυριότερες μετρικές αξιολόγησης της αποτελεσματικότητας των αλγορίθμων ταξινόμησης, είναι η μήτρα σύγχυσης (confusion matrix). Η συγκεκριμένη μήτρα έχει δύο διαστάσεις, η πρώτη διάσταση σχετίζεται με τις προβλέψεις των αλγορίθμων σχετικά με την κατηγοριοποίηση στις επιμέρους κλάσεις και τοποθετείται στον κατακόρυφο άξονα της μήτρας. Αντίστοιχα, υπάρχει και η διάσταση των πραγματικών τιμών και κλάσεων, η οποία τοποθετείται στον οριζόντιο άξονα της μήτρας. Στη συνήθη μορφή της, μια μήτρα σύγχυσης περιέχει τέσσερις επιμέρους τιμές στο κύριο μέρος της. Οι τιμές αυτές συμβολίζονται ως TP, TN, FP, και FN και ερμηνεύονται ως ορθά θετικές, ορθά αρνητικές, εσφαλμένα θετικές και εσφαλμένα αρνητικές προβλέψεις. Μια τυπική μήτρα

σύγχυσης παρουσιάζεται στον Πίνακα 3, στο κάτωθι κεφάλαιο και συγκεκριμένα στην ανάλυση του αλγορίθμου των δέντρων αποφάσεων.

Οι (Kaur et al., 2023) ανατρέχοντας διεθνείς δημοσιεύσεις της τελευταίας 10-ετίας, οι οποίες αφορούσαν εφαρμογές της μηχανικής μάθησης σε διάφορες βάσεις δεδομένων, και πιο συγκεκριμένα τεχνικές ανάλυσης φυσικού λόγου, οι οποίες υπάγονται στο γενικότερο πλαίσιο της ταξινόμησης – κατηγοριοποίησης, συγκέντρωσαν και κατέγραψαν τις κυριότερες μετρικές που είχαν χρησιμοποιηθεί στις περισσότερες μελέτες. Στην συνέχεια παρατίθενται ορισμένες από τις μετρικές αυτές, μέσω του Πίνακα 2.

Ονομασία μετρικής ελέγχου	**Συνάρτηση υπολογισμού**	**Επεξήγηση**
Precision (P)	$Precision = {TP}/{(TP + FP)}$	Ο λόγος των σωστών θετικών προβλέψεων προς τις συνολικές θετικές προβλέψεις
Recall (R)	$Recall = {TP}/{(TP + FN)}$	Ο λόγος των σωστών θετικών προβλέψεων προς το σύνολο των σωστών θετικών και λανθασμένα αρνητικών προβλέψεων. Είναι επίσης γνωστή ως ευαισθησία
F - Score	$F - Score = {2 * Precision * Recall}/{Precision + Recall}$	Σταθμισμένη μετρική που προκύπτει από το precision και το recall
Ακρίβεια - Accuracy	$Accuracy = {(TP + TN)}/{(TP + FP + TN + FN)}$	Λόγος σωστών προβλέψεων προς το συνολικό αριθμό των μεταβλητών

Εξειδίκευση - Specificity	$Specificity = {TN}/{(FP + TN)}$	Υπολογισμός αρνητικών απαντήσεων που προβλέφθηκαν ορθά
Περιοχή κάτω από την καμπύλη – Area under Curve - AUC	$AUC: FPR = 1 - Specifity = {FP}/{(FP + TN)}$ $FPR: False\ Positive\ Ratio$	Προσδιορισμός ικανότητας αλγορίθμου να διαχωρίζει τις κλάσεις μεταξύ τους

Πίνακας 4 Μετρικές αξιολόγησης αποτελεσμάτων ταξινόμησης στους αλγορίθμους μηχανικής μάθησης

(**Πηγή:** Kaur, et al., 2023)

Καταλήγοντας, σημειώνεται ότι οι συναρτήσεις υπολογισμού των μετρικών που παρατίθενται στον παραπάνω πίνακα είναι εκπεφρασμένες ως προς δύο βασικές κλάσεις. Ωστόσο, αυτές οι σχέσεις μπορούν, με σχετική ευκολία, να επεκταθούν έτσι ώστε να καλύπτουν τον έλεγχο μοντέλων που κατηγοριοποιούν δεδομένα σε περισσότερες από μία κλάσεις.

Κεφάλαιο 3: Αλγόριθμοι Ταξινόμησης

3.1 Δέντρα αποφάσεων (Decision trees)

Τα δέντρα αποφάσεων αποτελούν έναν από τους πιο διάσημους και ευρέως χρησιμοποιούμενους αλγορίθμους μηχανικής μάθησης και υπάγονται στην κατηγορία των αλγορίθμων επιτηρούμενης μάθησης. Οι κλασσικές προσεγγίσεις στην υλοποίηση των συγκεκριμένων αλγορίθμων, χρησιμοποιούν κυρίως διακριτές ή κατηγορικές μεταβλητές δεδομένων. Ωστόσο, η αποτελεσματικότητα των δέντρων αποφάσεων, έχει οδηγήσει στην ένταση της έρευνας σχετικά με την επέκταση της χρήσης των αλγορίθμων και σε βάσεις δεδομένων με συνεχείς μεταβλητές (Quinlan, 1986). Ορισμένες από τις πιο αποτελεσματικές εφαρμογές των δέντρων αποφάσεων, έχουν αναπτυχθεί για να επιλύσουν προβλήματα σχετικά με εξόρυξη δεδομένων από ηλεκτρονικές πηγές, αναγνώριση προτύπων και μοτίβων, κατηγοριοποίηση δεδομένων και χαρακτηριστικών και άλλα (Γκίκας, 2022).

Το κυριότερο πεδίο εφαρμογών για τα δέντρα αποφάσεων, είναι η κατηγοριοποίηση των δεδομένων σε διαφορετικές κλάσεις ή συστάδες με συγκεκριμένα χαρακτηριστικά. Για την επίτευξη αυτού του σκοπού, η υλοποίηση βασίζεται σε μια λογική ακολουθία της μορφής: if – then if – else if - else. Συμπεραίνουμε λοιπόν ότι τα εξαιρετικά επίπεδα αποτελεσματικότητας των αλγορίθμων, προκύπτουν μέσα από απλές, σχετικά, προγραμματιστικές προσεγγίσεις. Ωστόσο, στα δέντρα αποφάσεων υπάρχουν ορισμένες άλλες παράμετροι που αυξάνουν την πολυπλοκότητα της θεμελιώδους λογικής σχεδιασμού και υλοποίησης. Πιο συγκεκριμένα, θα πρέπει να εξετάζονται παράμετροι σχετικά με το μέγεθος των δέντρων, έτσι ώστε να εξασφαλίζεται το μέγιστο εφικτό κέρδος πληροφορίας και ταυτόχρονα η μείωση της εντροπίας του δέντρου. Παράλληλα, θα ο σχεδιασμός θα πρέπει να προβλέπει και να αποτρέπει φαινόμενα υπερεστίασης και υποεστιάσης των δέντρων, καθώς αυτό αποτελεί την πιο σημαντική κατηγορία πηγών αποτυχίας αυτών των αλγορίθμων (Giannopoulos et al., 2021). Στην συνέχεια, γίνεται αναλυτική περιγραφή της δομής και του τρόπου λειτουργίας ενός τυπικού δέντρου απόφασης, και παράλληλα γίνεται σχολιασμός σχετικά με το κέρδος πληροφορίας στα δέντρα και τις μεθόδους αποφυγής της υπερεστίασης και υποεστίασης.

Όσον αφορά τους λόγους για τους οποίους οι συγκεκριμένοι αλγόριθμοι θεωρούνται από τους πλέον σημαντικούς για τη μηχανική μάθηση, αυτοί σχετίζονται με τον υψηλό βαθμό διερμηνείας των αποτελεσμάτων που επιτρέπουν (interpretable results), καθώς και με το γεγονός ότι είναι εύκολη η διαγραμματική απεικόνιση της συνολικής διαδικασίας (Rokach and Maimon, 2008). Όπως γίνεται εύκολα αντιληπτό, όσο πιο κατανοητά είναι τα αποτελέσματα

ενός αλγορίθμου, τόσο πιο εύκολα μπορούν να τεκμηριωθούν οι αποφάσεις που προτείνονται καθώς επίσης και να γίνουν βελτιωτικές ενέργειες για την αναβάθμιση της αποτελεσματικότητας, ως προς την ορθότητα των προβλέψεων, και της αποδοτικότητα, ως προς τον χρόνο που απαιτείται για την υλοποίηση.

3.1.1 Δομή των Δέντρων Αποφάσεων

Ένα τυπικό δέντρο αποφάσεων αποτελείται από κόμβους και από κλαδιά, τα οποία οδηγούν σε επόμενους κόμβους μέχρις ότου να καταλήξει στην τελική κατηγοριοποίηση των δεδομένων. Σύμφωνα με αυτή την περιγραφή, συμπεραίνεται η top – down λογική των δέντρων ή με άλλα λόγια η προσπάθεια ανάλυσης του αρχικού προβλήματος σε επιμέρους στάδια για την διευκόλυνση της τελικής κατάταξης. Οι κόμβοι που συναντώνται, μπορούν να διαχωριστούν σε κύριους και δευτερεύοντες κόμβους ή αλλιώς κόμβους κλαδιών, ανάλογα με το ιεραρχικό επίπεδο που τοποθετούνται στο δέντρο. Όσο περισσότερη ανάλυση απαιτείται για την εκπαίδευση του μοντέλου, τόσο περισσότεροι κόμβοι χαμηλότερης ιεραρχίας δημιουργούνται. Στη γενική μορφή των δέντρων, παρατηρείται συνήθως ένας μόνο κύριος κόμβος, ο οποίος τοποθετείται στο ανώτατο επίπεδο, από τον οποίο δημιουργούνται διάφορα επιμέρους κλαδιά που είτε καταλήγουν σε κάποια κλάση είτε σε κάποιον άλλον κόμβο χαμηλότερου επιπέδου.

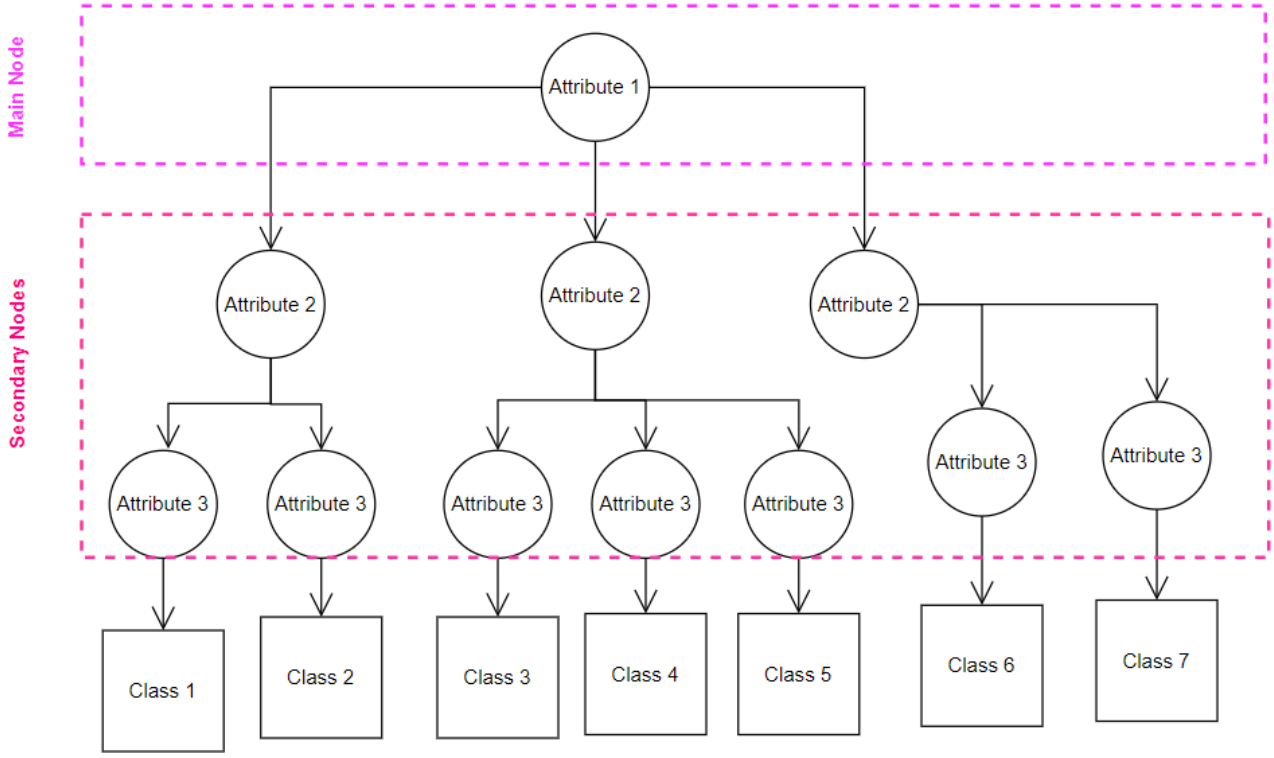
Όσο περισσότεροι δευτερεύοντες κόμβοι υπάρχουν στην δομή του δέντρου τόσο καλύτερα εκπαιδεύεται το μοντέλο στα χαρακτηριστικά μιας βάσης δεδομένων. Ωστόσο, η εκπαίδευση του μοντέλου θα πρέπει να είναι σχεδιασμένη με τρόπο που να επιτρέπει την γενίκευση με αποτελεσματικό τρόπο και σε άλλες βάσεις, δηλαδή με τρόπο που να αποτρέπει την υπερστίαση. Σαφώς, όσο αυξάνει η πολυπλοκότητα του προβλήματος, τόσο διαφοροποιούνται οι μορφές των δέντρων, προκειμένου να ερμηνεύσουν με ακρίβεια τα δεδομένα.

Όσον αφορά την αντιστοίχιση των περιεχομένων μιας βάσης δεδομένων με την παραπάνω λογική δημιουργίας των δέντρων, θα μπορούσαμε να αναφέρουμε τα κάτωθι:

- Οι κόμβοι αναπαριστούν τις μεταβλητές ή τα χαρακτηριστικά της βάσης δεδομένων, είτε στο σύνολό τους είτε μόνο ορισμένες από αυτές, ανάλογα με την φύση του προβλήματος. Πιο συγκεκριμένα, ο κύριος κόμβος συμπεριλαμβάνει το σύνολο των χαρακτηριστικών και για το λόγο αυτό συνδέεται με όλους τους επιμέρους κόμβους, ενώ οι δευτερεύοντες κόμβοι περιλαμβάνουν τουλάχιστον ένα χαρακτηριστικό από το σύνολο.
- Τα κλαδιά των δέντρων αναπαριστούν τις πιθανές τιμές που μπορούν να ληφθούν για μια συγκεκριμένη μεταβλητή της βάσης, δηλαδή για έναν κόμβο. Για παράδειγμα, αν

εξεταζόταν μια μεταβλητή, που λαμβάνει τρεις διαφορετικές τιμές στο σύνολο της βάσης, τότε η μεταβλητή θα αναπαρίσταντο με έναν κόμβο, από τον οποίο θα εξέρχονταν τρία διαφορετικά κλαδιά.

Στην Εικόνα 5, παρουσιάζεται η γραφική αναπαράσταση ενός δέντρου αποφάσεων, για την κατηγοριοποίηση περιγραφικών δεδομένων. Το δέντρο της Εικόνας 5, σύμφωνα με τα όσα αναφέρθηκαν παραπάνω, προκύπτει από μια βάση δεδομένων με τρεις κύριες μεταβλητές και επτά επιμέρους κλάσεις – κατηγορίες δεδομένων.



Εικόνα 5 Σχηματική αναπαράσταση ενός δέντρου αποφάσεων

Πριν την υλοποίηση ενός δέντρου αποφάσεων, είναι πολύ σημαντικό να έχει προηγηθεί το φιλτράρισμα της βάσης δεδομένων και έχουν επιλυθεί προβλήματα σχετικά με την διαχείριση ή κανονικοποίηση των ακραίων τιμών, την συμπλήρωση ή απομόνωση των κενών τιμών, την μείωση του αριθμού των χαρακτηριστικών, έτσι ώστε να λαμβάνονται υπόψιν μόνο τα ουσιαστικά χαρακτηριστικά και το μοντέλο να είναι εύκολα ερμηνεύσιμο και λογικό. Μετά την υλοποίηση των παραπάνω προκάτοχων σταδίων, ακολουθεί ο διαχωρισμός του συνόλου των δεδομένων, σε τρεις επιμέρους βάσεις, κάθε μία από τις οποίες εξυπηρετεί διαφορετικούς σκοπούς. Οι απαραίτητες επιμέρους βάσεις είναι:

[1]**Training data set**: Το μεγαλύτερο τμήμα της βάσης δεδομένων, χρησιμοποιείται προκειμένου να εκπαιδευτεί το μοντέλο. Δεν υπάρχει ξεκάθαρος τρόπος για τον υπολογισμό του τμήματος της βάσης που θα πρέπει να χρησιμοποιηθεί για να

εκπαιδευτεί αποτελεσματικά το μοντέλο. Ωστόσο, στην βιβλιογραφία έχουν επικρατήσει ορισμένοι εμπειρικοί κανόνες, σύμφωνα με τους οποίους μια ποσόστωση 70-30 ή 80-20, οδηγεί σε αποτελεσματική εκπαίδευση και παράλληλα δεν δημιουργεί άμεσο κίνδυνο για υπερστίαση του μοντέλου. Βέβαια, η επιλογή ενός ικανού συνόλου δεδομένων και κατ' επέκταση η ποσόστωση μεταξύ των βάσεων, εξαρτάται σε μεγάλο βαθμό και από τα ιδιαίτερα χαρακτηριστικά του εκάστοτε προβλήματος.

[2]**Validation data set**: Μόλις ολοκληρωθεί η εκπαίδευση του μοντέλου, τότε δοκιμάζεται η αποτελεσματικότητά του ως προς την κατηγοριοποίηση των υπόλοιπων δεδομένων, δηλαδή του συνόλου που δεν χρησιμοποιήθηκε για την εκπαίδευση. Για τον έλεγχο της αποτελεσματικότητας έχουν αναπτυχθεί διάφοροι δείκτες. Μια πολύ γνωστή προσέγγιση, η οποία βρίσκει εφαρμογή σε όλους σχεδόν τους αλγορίθμους κατηγοριοποίησης, είναι η δημιουργία της μήτρας αποτελεσμάτων ή σύγχυσης (confusion matrix). Στην συγκεκριμένη μήτρα καταγράφονται τέσσερις δείκτες αποτελεσμάτων: το ποσοστό των απαντήσεων που χαρακτηρίστηκαν ως θετικές και ήταν πράγματι θετικές(True Positive - TP), το ποσοστό των απαντήσεων που ορθώς χαρακτηρίστηκαν ως αρνητικές (True Negative - TN), και αντίστοιχα οι λανθασμένοι χαρακτηρισμοί θετικών και αρνητικών (False Positive – FP και False Negative – FN). Όπως είναι προφανές, η συγκεκριμένη μήτρα, βρίσκει εφαρμογή κυρίως σε προβλήματα κατηγοριοποίησης μεταξύ δύο κλάσεων. Ωστόσο, μπορεί εύκολα να επεκταθεί και για περισσότερες κλάσεις, βασισμένη στον ίδιο συλλογισμό. Στον Πίνακα 5, παρουσιάζεται η τυπική δομή μιας μήτρας σύγχυσης, δύο κλάσεων.

		ΠΡΟΒΛΕΨΕΙΣ	
ΠΡΑΓΜΑΤΙΚΕΣ ΤΙΜΕΣ		Κλάση 1 (*θετική*)	Κλάση 2 (*αρνητική*)
	Κλάση 1 (*θετική*)	True Positive	False Negative
	Κλάση 2 (*αρνητική*)	False Positive	True Negative

Πίνακας 5 Τυπική μήτρα αποτελεσμάτων - κατηγοριοποίηση δύο κλάσεων

Η παραπάνω μήτρα είναι πολύ σημαντική, διότι εκτός του ότι παρέχει μια ξεκάθαρη απεικόνιση για την ικανότητα προβλέψεων του μοντέλου, δίνει επιπλέον την δυνατότητα υπολογισμού διαφόρων δεικτών αποτελεσματικότητας, όπως του δείκτη ακρίβειας (βλ. παρακάτω σχέση) καθώς και πιο σύνθετων στατιστικών δεικτών.

$$\boldsymbol{Ακρίβεια} = \frac{True\ Positive + True\ Negative}{Πλήθος\ των\ δεδομένων\ για\ κατηγοριοποίηση}$$

Ο υπολογισμός των δεικτών είναι καθοριστικός για την αξιολόγηση της εκπαίδευσης και της πληρότητας του μοντέλου. Πιο συγκεκριμένα, πριν την έναρξη της υλοποίησης θα πρέπει να έχουν προκαθοριστεί τα επιθυμητά επίπεδα ακρίβειας. Αν το δέντρο αποφάσεων, που έχει επιλεγεί δεν μπορεί να φτάσει τα επιθυμητά επίπεδα ακρίβειας, τότε η διαδικασία εκπαίδευσης θα πρέπει να επιθεωρηθεί και να βελτιωθεί ή ακόμα και να επιλεγεί διαφορετική προσέγγιση υλοποίησης. Ακολουθείται, λοιπόν, μια κυκλική διαδικασία, η οποία τερματίζεται όταν το μοντέλο φτάσει στην επιθυμητή ακρίβεια.

[3]**Test data set**: Μετά την ολοκλήρωση του ελέγχου του μοντέλου, θα πρέπει να διερευνηθεί η αποτελεσματικότητά του και σε άλλες βάσεις δεδομένων, οι οποίες δεν σχετίζονται κάπως με την βάση εκπαίδευσης. Αυτό μπορεί να γίνει είτε χρησιμοποιώντας εντελώς διαφορετική βάση, αλλά με ίδια χαρακτηριστικά, είτε διαιρώντας εξ' αρχής την βάση σε τρία τμήματα, έτσι ώστε το μεγαλύτερο τμήμα να χρησιμοποιηθεί για εκπαίδευση, το μικρότερο για έγκριση των αποτελεσμάτων λειτουργίας και το τελευταίο για προβλέψεις, με την υπόθεση ότι είναι εντελώς «άγνωστο» για το μοντέλο.

Ολοκληρώνοντας την παρούσα παράγραφο, σημειώνεται ότι ο τρόπος διαχωρισμού της βάσης, δεν θα πρέπει να γίνεται με αυθαίρετο τρόπο, δηλαδή «χειροκίνητα», αλλά θα πρέπει να εφαρμόζονται στατιστικοί μέθοδοι δειγματοληψίας, διότι με τον τρόπο αυτό επιτυγχάνεται η βέλτιστη δυνατή εκπαίδευση του μοντέλου στις διαστάσεις του προβλήματος.

3.1.2 Η υπερ - εστίαση (overfitting) στα δέντρα αποφάσεων

Ένα από τα κύρια ζητήματα που θα πρέπει να αντιμετωπιστούν κατά την υλοποίηση ενός δέντρου αποφάσεων, είναι η υπερστίαση, η οποία παρατηρείται συχνά σε αυτά τα μοντέλα και αποτελεί τον βασικό λόγο μείωσης της αποτελεσματικότητάς τους. Ως υπερστίαση ορίζεται το φαινόμενο κατά το οποίο το μοντέλο μαθαίνει σε πολύ μεγάλη λεπτομέρεια τα χαρακτηριστικά και τις ιδιομορφίες της βάσης εκπαίδευσης, γεγονός που κατ' επέκταση επιφέρει αρνητικά αποτελέσματα ως προς την ικανότητα του δέντρου να προβλέπει αποτελεσματικά σε διαφορετικές βάσεις δεδομένων.

Το πρόβλημα της υπερεστίασης, στις περισσότερες βιβλιογραφικές αναφορές, είναι συσχετισμένο με τον σχεδιασμό των κατάλληλων συνθηκών τερματισμού των μοντέλων. Πιο συγκεκριμένα, αν οι συνθήκες τερματισμού, κατά το στάδιο της εκπαίδευσης των μοντέλων, δεν είναι κατάλληλα σχεδιασμένη, τότε το δέντρο διατρέχει λεπτομερώς ολόκληρη την βάση

και κατηγοριοποιεί όλα τα διαθέσιμα δεδομένα σε αντίστοιχες κλάσεις. Μόλις ολοκληρωθεί η προσπέλαση και δεν υπάρχουν άλλα δεδομένα, τότε το μοντέλο επιδιώκει την εξαγωγή συσχετίσεων μεταξύ των διάφορων χαρακτηριστικών, χρησιμοποιώντας πιθανότητες και σενάρια. Είναι προφανές ότι οι συσχετίσεις που πιθανότατα υπάρχουν σε μια συγκεκριμένη βάση, δεν αποτελούν κοινό παράγοντα για όλες τις υπόλοιπες βάσεις, με παρεμφερές περιεχόμενο. Κατά συνέπεια, η δημιουργία τέτοιου είδους συσχετίσεων, μειώνει σημαντικά τον βαθμό αποτελεσματικότητας και επεκτασιμότητας των μοντέλων.

Ένα επιπλέον χαρακτηριστικό που οδηγεί στην υπερεστίαση, είναι η ύπαρξη κενών κελιών και ακραίων τιμών στα δεδομένα εκπαίδευσης. Αμφότερες οι δύο περιπτώσεις δημιουργούν θόρυβο, ο οποίος είναι δύσκολο να ερμηνευθεί από το μοντέλο, και άρα θα πρέπει να απαλείφεται εξ' αρχής, κατά τα προκαταρκτικά στάδια υλοποίησης. Για παράδειγμα, η ύπαρξη κενών τιμών, θεωρείται από το μοντέλο ως μια ξεχωριστή κλάση, για την οποία θα πρέπει να εκπαιδευτεί. Στη συνέχεια, μόλις γίνει χρήση του μοντέλου σε «άγνωστα» δεδομένα, το μοντέλο θα επιδιώκει την εύρεση δεδομένων προς κατηγοριοποίηση στην κλάση των κενών στοιχείων. Όπως γίνεται αντιληπτό, η προσέγγιση αυτή δεν είναι η πιο αποδοτική, καθώς γίνεται προσπάθεια εύρεσης στοιχείων, που γενικά δεν υπάρχουν, κι έτσι ο χρόνος υλοποίησης αυξάνεται κατά πολύ.

Για την αντιμετώπιση του προβλήματος της υπερεστίασης και κατ' επέκταση για την μείωση της έκτασης των δέντρων, έχουν αναπτυχθεί πολλές προσεγγίσεις, οι οποίες διαφοροποιούνται ανάλογα με τον αλγόριθμο που χρησιμοποιείται για την υλοποίηση του δέντρου. Ορισμένες από τις πιο διαδεδομένες προσεγγίσεις υλοποίησης, είναι ο αλγόριθμος CART, ο C4.5, ο ID3 και άλλοι. Όλες οι ανωτέρω προσεγγίσεις, στοχεύουν στην πραγματικότητα σε μια μορφή «κλαδέματος» (pruning) των αρχικών δέντρων που δημιουργούνται, προκειμένου να μειωθούν οι διαστάσεις στον απολύτως απαραίτητο βαθμό, ο οποίος εξασφαλίζει το μέγιστο δυνατό κέρδος πληροφορίας. Αναφορικά με το «κλάδεμα» των δέντρων, θα μπορούσαμε να αναφέρουμε ότι οι δύο κυρίαρχες προσεγγίσεις είναι το «προ – κλάδεμα» και το «μετά – κλάδεμα» (Han, et al., 2012).

Οι προσεγγίσεις υλοποίησης του «προ - κλαδέματος» επικεντρώνονται στην δημιουργία κανόνων και συνθηκών τερματισμού της εκπαίδευσης των μοντέλων, μετά από ένα ορισμένο στάδιο. Πρόκειται, στην πραγματικότητα, για προγνωστικές προσεγγίσεις (proactive approaches). Βασίζονται, κατά κύριο λόγο, στην λογική ότι από ένα στάδιο ανάλυσης και έπειτα, η δομή του δέντρου θα πρέπει να μεγαλώσει υπερβολικά πολύ προκειμένου να επιτευχθεί σημαντικά βελτιωμένη μάθηση. Κατά συνέπεια, εξετάζεται ο βαθμός κέρδους

πληροφορίας ανάλογα με το μέγεθος και γίνονται ορισμένες παραδοχές και trade – offs, έτσι ώστε να τερματιστεί η διαδικασία ανάπτυξης του δέντρου.

Αντίθετα, στις προσεγγίσεις υλοποίησης «μετά – κλαδέματος» το μοντέλο δεν τερματίζεται κατά την διαδικασία της μάθησης και δεν δημιουργούνται συνθήκες τερματισμού από την αρχή της διαδικασίας. Τα δέντρα αναπτύσσονται πλήρως και στη συνέχεια ελέγχεται ο βαθμός αποτελεσματικότητάς τους και ο χρόνος υλοποίησης τους, μέσω εφαρμογής σε διαφορετικές βάσεις δεδομένων, μέσω ενός συνόλου από ανεξάρτητες μεταβλητές. Έπειτα, ακολουθεί μια κυκλική διαδικασία αφαίρεσης κόμβων και μετατροπής αυτών σε απλά κλαδιά, δηλαδή κλάσεις, έως ότου να βρεθεί ο αριθμός των κόμβων που παρέχει την μέγιστη δυνατή πληροφορία. Γενικά, σε αυτή την περίπτωση, για την επιλογή του βέλτιστου αριθμού κόμβων, εξετάζεται μια σειρά από παραμέτρους, όπως χρόνος υλοποίησης, βαθμός ακρίβειας ανάλογα με τους κόμβους και διάφοροι άλλοι δείκτες στατιστικής σημαντικότητας των αποτελεσμάτων.

Ανατρέχοντας στην βιβλιογραφία, μπορεί να διαπιστώσει κανείς ότι έχουν καταγραφεί αρκετά πλεονεκτήματα και μειονεκτήματα για κάθε μία από τις παραπάνω μεθόδους «κλαδέματος» των δέντρων. Ενώ αρκετή συζήτηση γίνεται γύρω από το τρόπο επιλογής των κριτηρίων αξιολόγησης της εκπαίδευσης. Αναμφίβολα, το προς επίλυση πρόβλημα παίζει σημαντικό ρόλο στην επιλογή των κατάλληλων μεθόδων. Ωστόσο, θα θεωρούσαμε ότι η μέθοδος του «μετά – κλαδέματος» είναι περισσότερο δια δραστική με τον χρηστή και επιτρέπει την καλύτερη εποπτεία κατά την εξέλιξη του μοντέλου, οπότε ίσως είναι πιο αποτελεσματική σε περιπτώσεις που δεν υπάρχει τεράστια εμπειρία σχετικά με το πρόβλημα ή όταν οι βάσεις δεδομένων που χρησιμοποιούνται δεν είναι εξαιρετικά μεγάλες (Witten, et al., 2011).

3.1.3 Διαδεδομένοι αλγόριθμοι υλοποίησης δέντρων αποφάσεων

Αλγόριθμος ID3

Αποτελεί έναν από τους πιο κλασσικούς αλγορίθμους, για την ανάπτυξη ενός δέντρου αποφάσεων, με πληθώρα εφαρμογών σε πολλά και διαφορετικά προβλήματα. Στη συγκεκριμένη προσέγγιση, η κατηγοριοποίηση – ταξινόμηση των δεδομένων στις επιμέρους κλάσεις, γίνεται με βάση δύο κριτήρια: το κέρδος πληροφορίας και την εντροπία της πληροφορίας (Shang et al., 2013).

Η εντροπία πληροφορίας είναι ένα μέγεθος που δανείζονται οι αλγόριθμοι μηχανικής μάθησης από τη Θερμοδυναμική και χρησιμοποιείται για την δημιουργία ενός ποσοτικού μεγέθους μέτρησης της αβεβαιότητας ή της τυχαιότητας μεταξύ των δεδομένων, σε μία βάση (Ritchie, 1986). Όσο μεγαλύτερη είναι η τιμή της εντροπίας, τόσο μεγαλύτερος ο βαθμός

αβεβαιότητας των δεδομένων, και κατ' επέκταση τόσο μεγαλύτερος αριθμός κλάσεων απαιτείται προκειμένου να ταξινομηθούν με ακρίβεια όλα τα δεδομένα της βάσης. Όπως γίνεται αντιληπτό, μεγάλος αριθμός κλάσεων, με μικρό σχετικά αριθμό ταξινομήσεων σε αυτές, είναι πολύ πιθανό να οδηγήσει σε φαινόμενα υπερεστίασης των μοντέλων, οπότε απαιτείται ειδικός χειρισμός. Αντίθετα, σε περιπτώσεις όπου τα επίπεδα της εντροπίας παραμένουν χαμηλά, αυτό σημαίνει ότι η πλειοψηφία των δεδομένων μπορούν να ενταχθούν σε σημαντικά μικρότερο αριθμό κλάσεων. Με άλλα λόγια, πρόκειται για βάσεις δεδομένων με ξεκάθαρες συσχετίσεις μεταξύ των μεταβλητών και των χαρακτηριστικών εν γένει. Η λογική σχεδιασμού αυτού του μοντέλου στηρίζεται στην επιλογή ενός αριθμού κλάσεων, ο οποίος ταυτόχρονα να μειώνει την εντροπία πληροφορίας και να αυξάνει το κέρδος πληροφορίας, για κάθε κλάση. Για τον υπολογισμό της εντροπίας πληροφορίας, χρησιμοποιείται η κάτωθι σχέση:

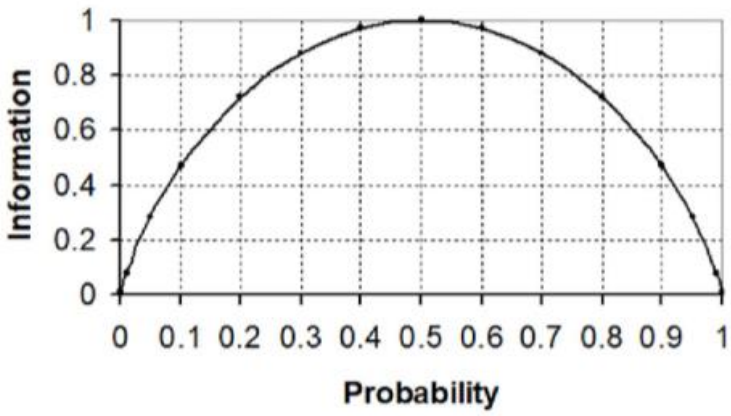
$$Εντροπία\ Πληροφορίας\ (Δ) = -\sum_{i=1}^{n} P(κ_i/Δ) * \log_2 P(κ_i/Δ)$$

$$Όπου: \begin{cases} Δ: το\ σύνολο\ των\ δεδομένων\ που\ εξετάζονται \\ n: το\ πλήθος\ των\ κλάσεων \\ P(κ_i/Δ): η\ πιθανότητα\ ύπαρξης\ μιας\ κλάσης\ i\ στο\ σύνολο\ δεδομένων\ Δ \end{cases}$$

Χρησιμοποιώντας την παραπάνω σχέση επαναληπτικά, ο αλγόριθμος μπορεί να επιλέξει το πλήθος των κλάσεων κατηγοριοποίησης, το οποίο επιφέρει την ελάχιστη τιμή της εντροπίας πληροφορίας. Στη συνέχεια, θα πρέπει να επαληθευτεί αν η επιλογή που έγινε, επιφέρει παράλληλα και την βελτιστοποίηση της ανακτημένης πληροφορίας. Αν και οι δύο αυτές συνθήκες ικανοποιούνται τότε προχωρά κανονικά η διαδικασία προβλέψεων ταξινόμησης, ειδάλλως γίνεται προσπάθεια εύρεσης του καλύτερου δυνατού συνδυασμού για μείωση εντροπίας και αύξηση κέρδους πληροφορίας. Ο υπολογισμός του κέρδους πληροφορίας, βασίζεται στο θεώρημα Shannon, η σχηματική αναπαράσταση του οποίου δίνεται στην Εικόνα 5. Σύμφωνα μ την συγκεκριμένη απεικόνιση, μπορούμε να συμπεράνουμε ότι οι πιθανότητες να λάβει κανείς την βέλτιστη δυνατή πληροφορία, μειώνονται σημαντικά όσο αυξάνει ο αριθμός των κλάσεων. Για την ταυτόχρονη μεγιστοποίηση της πιθανότητας και του όγκου πληροφορίας που εξάγεται, προκύπτει ότι ο αριθμός των κλάσεων θα πρέπει να είναι μικρότερος από τον μέγιστο δυνατό και μεγαλύτερος από το μέσο αριθμός κλάσεων. Η σχέση υπολογισμού του κέρδους πληροφορίας, δίνεται από τον τύπο:

$$Κέρδος\ Πληροφορίας\ (X_i, Δ) = Εντροπία\ (Δ) - \sum_{i,j} P\left(X_{i,j}\right) * E(X_i)$$

$$\text{Όπου: } \begin{cases} Xi: \tau\alpha\ \chi\alpha\rho\alpha\kappa\tau\eta\rho\iota\sigma\tau\iota\kappa\acute{\alpha}\ \tau\eta\varsigma\ \beta\acute{\alpha}\sigma\eta\varsigma \\ Xi,j: o\ \alpha\rho\iota\theta\mu\acute{o}\varsigma\ \tau\omega\nu\ \kappa\lambda\acute{\alpha}\sigma\epsilon\omega\nu\ \gamma\iota\alpha\ \tau\eta\nu\ \tau\alpha\xi\iota\nu\acute{o}\mu\eta\sigma\eta\ \tau\omega\nu\ \epsilon\pi\iota\mu\acute{\epsilon}\rho ous\ \chi\alpha\rho\alpha\kappa\tau\eta\rho\iota\sigma\tau\iota\kappa\acute{\omega}\nu \end{cases}$$



Εικόνα 6 Γραφική αναπαράσταση της συνάρτησης Shannon, περί κέρδους πληροφορίας

(Πηγή: Γκίκας, 2022)

Αλγόριθμος C4.5

Η βασική προβληματική διάσταση που έχει συνδεθεί με την λειτουργία του αλγορίθμου ID3, είναι ότι όταν εφαρμόζεται σε σχετικά μικρές βάσεις δεδομένων, οδηγεί στις περισσότερες περιπτώσεις σε υπερ-εστίαση του μοντέλου στη βάση εκπαίδευσης και ως εκ τούτου στην πολύ αποτελεσματική πρόβλεψη στα δεδομένα της ίδιας βάσης και συνάμα στην αδυναμία επέκτασης του, σε διαφορετικές βάσεις. Η προβληματική διάσταση συσχετίστηκε σε μεγάλο βαθμό με την αρχή λειτουργίας του αλγορίθμου, σύμφωνα με την οποία η ταξινόμηση ή η πρόβλεψη γίνεται με κύριο γνώμονα το κέρδος πληροφορίας. Πιο συγκεκριμένα, σε διάφορες μελέτες, διαπιστώθηκε ότι οι μεταβλητές που είχαν τις περισσότερες εγγραφές συγκριτικά με τις υπόλοιπες, θεωρούνται από τον αλγόριθμο, ότι περιέχουν την πιο «πλήρη» πληροφορία και έτσι τοποθετούνται στο υψηλότερο επίπεδο της ιεραρχίας του δέντρου, με την μορφή των κόμβων. Αυτή η προσέγγιση, σαφώς δεν είναι πάντοτε ορθολογική και ως εκ τούτου όταν εφαρμοστεί οδηγεί πιθανότατα σε αδυναμία των μοντέλων.

Για την επίλυση του συγκεκριμένου προβλήματος, αναπτύχθηκε ο αλγόριθμος C4.5, ο οποίος έχει παρόμοια φιλοσοφία ανάπτυξης του δέντρου, ωστόσο χρησιμοποιεί ως μετρική υλοποίησης και αξιολόγησης του δέντρου τον ρυθμό κέρδους πληροφορίας (Information Gain Ratio), ο οποίος υπολογίζεται από την σχέση:

$$Ρυθμός\ κέρδους\ πληροφορίας = \frac{Κέρδος\ Πληροφορίας\ (X_i, Δ)}{Διαμοιρασμός\ Πληροφορίας\ (X_i, Δ)}$$

Ο διαμοιρασμός της πληροφορίας (Split In Information), είναι το νέο μέγεθος που χρησιμοποιείται στην πραγματικότητα, καθώς το κέρδος πληροφορίας έχει ήδη υπολογιστεί

με βάση την προσέγγιση του αλγορίθμου ID3. Ο δείκτης αυτός χρησιμοποιείται για τον προσδιορισμό του βαθμού που μία μεταβλητή χαρακτηρίζεται από συγκεκριμένα δεδομένα ή τύπους δεδομένων που περιέχει. Επιπλέον, ο λόγος κέρδους πληροφορίας θεωρείται πολύ αποτελεσματικός ως προς την επιλογή του γνωρίσματος με τη μεγαλύτερη ικανότητα διαχωρισμού των παρατηρήσεων αποφεύγοντας γνωρίσματα που τείνουν να δημιουργήσουν δέντρα με πολλές διακλαδώσεις και κατ' επέκταση υπερ-εστιασμένα μοντέλα (Tan, 2006).

$$Διαμοιρασμός\ Πληροφορίας\ (X_i, Δ) = -\sum_{i,j} \frac{abs(X_{i,j})}{abs(X_i)} * log \frac{abs(X_{i,j})}{abs(X_i)}$$

3.1.4 Εφαρμογή δέντρου αποφάσεων

Με στόχο την διασαφήνιση των εννοιών που αναφέρθηκαν στις προηγούμενες ενότητες του κεφαλαίου, θεωρείται σκόπιμη η παράθεση ενός σύντομου και απλού παραδείγματος ανάπτυξης ενός δέντρου αποφάσεων. Στην συγκεκριμένη ενότητα, θα αναπτύξουμε ένα δέντρο αποφάσεων, το οποίο θα προβλέπει το άθλημα με το οποίο ασχολούνται οι έφηβοι στην Ελλάδα, ανάλογα με δύο μεταβλητές, το φύλλο και το ύψος τους.

Στις νεότερες ηλικίες τα πλέον διαδεδομένα αθλήματα θεωρούνται το ποδόσφαιρο, το μπάσκετ και το βόλεϊ. Στο συγκεκριμένο πρόβλημα θα χρησιμοποιηθούν μόνο αυτά τα τρία αθλήματα και θα παραληφθούν όλα τα υπόλοιπα, χάριν απλότητας. Μια άλλη παραδοχή απλούστευσης, σχετίζεται με το ύψος. Πιο συγκεκριμένα, το ύψος ενός ανθρώπου μετριέται είτε σε εκατοστά είτε σε μέτρα. Ωστόσο, για την διευκόλυνση της εφαρμογής, δεν θα χρησιμοποιηθούν συνεχείς τιμές στην μεταβλητή ύψος, αλλά τρεις διακριτές τιμές, οι οποίες είναι: κάτω από το μέσο ύψος, στο μέσο ύψος, πάνω από το μέσο ύψος, βασισμένες σε δημοσιευμένα στατιστικά στοιχεία. Για την εκπαίδευση του μοντέλου, θα χρησιμοποιήσουμε μια βάση που αποτελείται από 30 συνολικά δεδομένα. Γεγονός, που αν συνδυαστεί με τον εμπειρικό κανόνα διάτμησης των βάσεων εκπαίδευσης και ελέγχου (80-20 ή 70-30) που αναφέρθηκε παραπάνω, μας οδηγεί στο συμπέρασμα ότι ο μέγιστος αριθμός προβλέψεων που δυνητικά μπορεί να επιτύχει το μοντέλο με ακρίβεια, είναι 7 προβλέψεις. Στον Πίνακα 6, παρουσιάζονται συγκεντρωτικά στοιχεία σχετικά με τα δεδομένα εκπαίδευσης του μοντέλου.

Πίνακας 6 Δεδομένα εκπαίδευσης δέντρου αποφάσεων - Εφαρμογή

Συγκεντρωτικά στοιχεία		Άθλημα			
Φύλλο	Ύψος	Βόλλευ	Μπάσκετ	Ποδόσφαιρο	Γενικό Άθροισμα
Άνδρας		4	6	5	15
Άνδρας	Κάτω από το μέσο	1	2	1	4
Άνδρας	Μέσο	2	2	1	5
Άνδρας	Πάνω από το μέσο	1	2	3	6
Γυναίκα		4	5	6	15
Γυναίκα	Κάτω από το μέσο		1	1	2
Γυναίκα	Μέσο	2	3	5	10
Γυναίκα	Πάνω από το μέσο	2	1		3
Γενικό Άθροισμα		**8**	**11**	**11**	**30**

Σύμφωνα με τον Πίνακα 6, μπορούσε σε ένα πρώτο στάδιο να διαπιστώσουμε το ποσοστό ενασχόλησης με το κάθε άθλημα, το οποίο αναπαριστά την εκ των προτέρων πιθανότητα να προκύψει κατά την ταξινόμηση το κάθε άθλημα. Πιο συγκεκριμένα, παρατηρούμε ότι:

$$\begin{cases} P\,(ποδόσφαιρο) = \dfrac{παρατηρήσεις\ ποδοσφαίρου}{συνολικές\ παρατηρήσεις} = \dfrac{11}{30} = 36{,}7\% \\ P\,(μπάσκετ) = \dfrac{παρατηρήσεις\ μπάσκετ}{συνολικές\ παρατηρήσεις} = \dfrac{11}{30} = 36{,}7\% \\ P\,(βόλλευ) = \dfrac{παρατηρήσεις\ βόλλευ}{συνολικές\ παρατηρήσεις} = \dfrac{8}{30} = 26{,}7\% \end{cases}$$

Χρησιμοποιώντας τις εκ των προτέρων πιθανότητες, θα προχωρήσουμε στην δημιουργία του δέντρου αποφάσεων. Οι κόμβοι του δέντρου θα τοποθετηθούν με τέτοιο τρόπο, ώστε να εξασφαλίζεται το μεγαλύτερο κέρδος πληροφορίας (Information Gain – IG) από αυτό. Στην πραγματικότητα, οι κόμβοι θα προκύψουν μέσα από τις μεταβλητές φύλλο και ύψος, ως εκ τούτου προκύπτουν δύο διαφορετικοί συνδυασμοί: είτε να τοποθετηθεί ως κύριος κόμβος το φύλλο και ως δευτερεύοντας το ύψος (προσέγγιση i) είτε το αντίθετο (προσέγγιση ii). Στην συνέχεια ακολουθεί αναλυτική επίλυση και προσδιορισμός του κέρδους πληροφορίας σε κάθε μία προσέγγιση, ώστε να επιλεχθεί η τελική μορφή του δέντρου.

Προσέγγιση i: Κύριος κόμβος το φύλλο και δευτερεύοντας το ύψος

Ως προς το φύλλο μπορούμε να διακρίνουμε τα εξής:

$$\begin{cases} P\,(άνδρας) = \dfrac{άνδρες}{σύνολο\ δείγματος} = \dfrac{15}{30} = 50\% \\ P\,(γυναίκα) = \dfrac{γυναίκες}{σύνολο\ δείγματος} = \dfrac{15}{30} = 50\% \end{cases}$$

Εάν το φύλλο είναι άνδρας, τότε οι πιθανότητες ενασχόλησης με κάποιο από τα τρία αθλήματα, είναι οι κάτωθι:

$$\begin{cases} P(ποδόσφαιρο/άνδρας) = \dfrac{5}{15} = 33\% \\ P(μπάσκετ/άνδρας) = \dfrac{6}{15} = 40\% \\ P(βόλλευ/άνδρας) = \dfrac{4}{15} = 27\% \end{cases}$$

Στη συγκεκριμένη περίπτωση, η πληροφορία που λείπει, μπορεί να υπολογιστεί από την σχέση:

$$I_{άνδρας} = -0{,}33 * \log(0{,}33) - 0{,}40 * \log(0{,}40) - 0{,}27 * \log(0{,}27) = 0{,}4716$$

Αντίστοιχα αν το φύλλο είναι γυναίκα, τότε έχουμε ότι:

$$\begin{cases} P(ποδόσφαιρο/γυναίκα) = \dfrac{6}{15} = 40\% \\ P(μπάσκετ/γυναίκα) = \dfrac{5}{15} = 33\% \\ P(βόλλευ/γυναίκα) = \dfrac{4}{15} = 27\% \end{cases}$$

Κατά συνέπεια, εφαρμόζοντας την παραπάνω σχέση προκύπτει η πληροφορία που λείπει:

$$I_{γυναίκα} = I_{άνδρας} = 0{,}4716$$

Το κέρδος πληροφορίας, για την συγκεκριμένη δομή του δέντρου, προκύπτει από την αφαίρεση της συνολικής πληροφορίας που λείπει από τα δεδομένα με την πληροφορία που χάνεται με βάση την συγκεκριμένη δομή, δηλαδή:

$$\begin{aligned} IG_{προσέγγιση\ i} &= I_{total} - I_{sex} \\ &= \big[-p(ποδόσφαιρο) * \log\big(p(ποδόσφαιρο)\big) - p(μπάσκετ) \\ &* \log\big(p(μπάσκετ)\big) - p(βόλλευ) * \log\big(p(βόλλευ)\big)\big] \\ &- \big[p(άνδρας) * I_{άνδρας} + p(γυναίκα) * I_{γυναίκα}\big] \\ &= [-0{,}367 * \log(0{,}367) - 0{,}367 * \log(0{,}367) - 0{,}267 * \log(0{,}267)] \\ &- [0{,}5 * 0{,}4716 + 0{,}5 * 0{,}4716] = 0{,}4726 - 0{,}4716 = 0{,}001 \end{aligned}$$

Προσέγγιση ii: Κύριος κόμβος το ύψος και δευτερεύοντας το φύλλο

Ως προς το ύψος μπορούμε να διακρίνουμε τα εξής:

$$\begin{cases} P\ (χαμηλό\ ύψος) = \dfrac{6}{30} = 20\% \\ P(μέσο\ ύψος) = \dfrac{15}{30} = 50\% \\ P(μεγάλο\ ύψος) = \dfrac{9}{30} = 30\% \end{cases}$$

Ένα το ύψος είναι χαμηλό τότε η πιθανότητα ενασχόληση με κάποιο από τα τρία αθλήματα είναι:

$$\begin{cases} P(ποδόσφαιρο/χαμηλό\ ύψος) = \frac{2}{6} = 33{,}3\% \\ P(μπάσκετ/χαμηλό\ ύψος) = \frac{3}{6} = 50\% \\ P(βόλλευ/χαμηλό\ ύψος) = \frac{1}{6} = 16{,}7\% \end{cases}$$

Συνεπώς η μη ανακτηθείσα πληροφορία σε αυτή την περίπτωση είναι:

$$I_{χαμηλό\ ύψος} = -0{,}333 * \log(0{,}333) - 0{,}5 * \log(0{,}5) - 0{,}167 * \log(0{,}167) = 0{,}43935$$

Στην περίπτωση που το ύψος ισούται περίπου με το μέσο ύψος της συγκεκριμένης ηλικιακής ομάδας, η πιθανότητα ενασχόλησης είναι:

$$\begin{cases} P(ποδόσφαιρο/μέσο\ ύψος) = \frac{6}{15} = 40\% \\ P(μπάσκετ/μέσο\ ύψος) = \frac{5}{15} = 33{,}3\% \\ P(βόλλευ/μέσο\ ύψος) = \frac{4}{15} = 26{,}7\% \end{cases}$$

Συνεπώς η μη ανακτηθείσα πληροφορία σε αυτή την περίπτωση είναι:

$$I_{μέσο\ ύψος} = -0{,}4 * \log(0{,}4) - 0{,}333 * \log(0{,}333) - 0{,}267 * \log(0{,}267) = 0{,}47132$$

Τέλος, στην περίπτωση όπου το ύψος ξεπερνά τη μέση τιμή, μπορούμε να υπολογίσουμε τα κάτωθι:

$$\begin{cases} P(ποδόσφαιρο/μεγάλο\ ύψος) = \frac{3}{9} = 33{,}3\% \\ P(μπάσκετ/μεγάλο\ ύψος) = \frac{3}{9} = 33{,}3\% \\ P(βόλλευ/μεγάλο\ ύψος) = \frac{3}{9} = 33{,}3\% \end{cases}$$

Συνεπώς η μη ανακτηθείσα πληροφορία σε αυτή την περίπτωση είναι:

$$I_{μέσο\ ύψος} = -0{,}33 * \log(0{,}333) * 3 = 0{,}47278$$

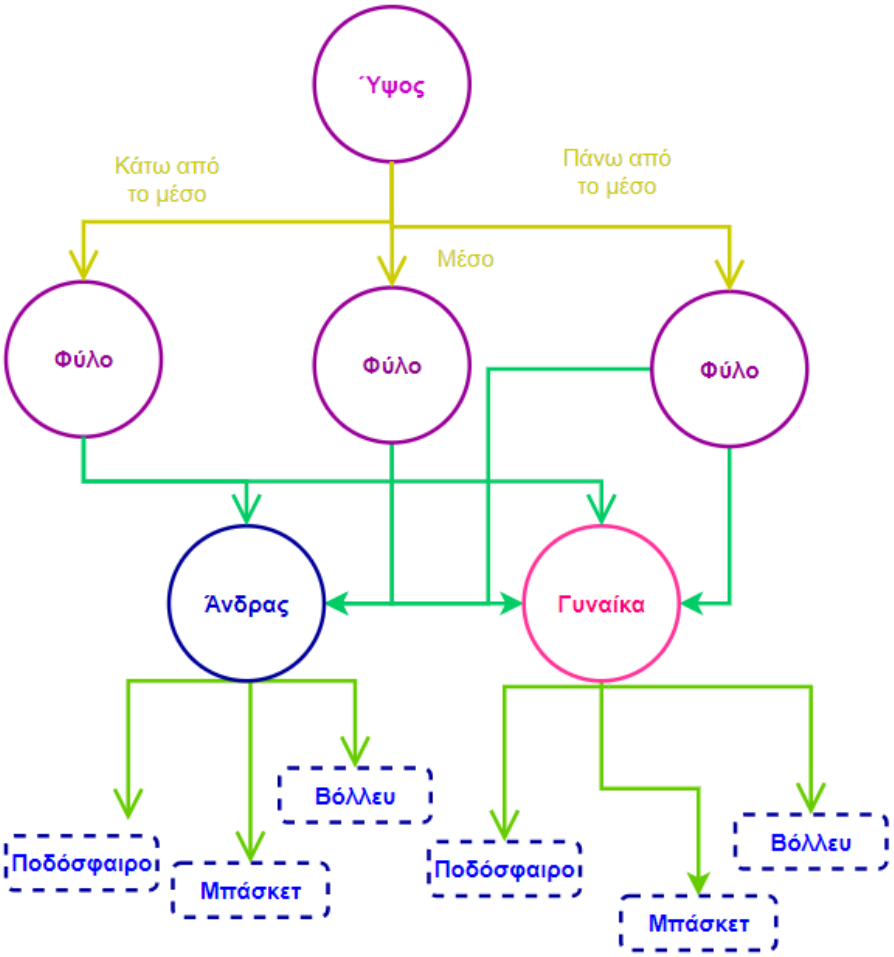
Με βάση όλους τους παραπάνω υπολογισμούς, η χαμένη πληροφορία στην περίπτωση επιλογής τους ύψους ως κύριου κόμβου είναι:

$$I_{ύψος} = p(χαμηλό\ ύψος) * I_{χαμηλό\ ύψος} + p(μέσο\ ύψος) * I_{μέσο\ ύψος} + p(μεγάλο\ ύψος) * I_{μεγάλο\ ύψος} = 0{,}2 * 0{,}43935 + 0{,}5 * 0{,}47132 + 0{,}3 * 0{,}47278 = 0{,}46536$$

Το κέρδος πληροφορίας στην δεύτερη προσέγγιση είναι:

$$IG_{προσέγγιση\ ii} = I_{total} - I_{ύψος} = 0{,}4726 - 0{,}46536 = 0{,}00724 > IG_{προσέγγιση\ i} = 0{,}001$$

Αφού το κέρδος πληροφορίας για την δεύτερη προσέγγιση προέκυψε μεγαλύτερο συγκριτικά με την πρώτη προσέγγιση, τότε σαφώς θα πρέπει να αναπτύξουμε το δέντρο αποφάσεων σύμφωνα με την δεύτερη προσέγγιση, άρα θεωρώντας ως κύριο κόμβο την ερώτηση περί ύψος και εν συνεχεία την ερώτηση περί φύλλου, ως δευτερεύοντα. Στην Εικόνα 7, παρουσιάζεται το δέντρο που προκύπτει σύμφωνα με την προσέγγιση ii.



Εικόνα 7 Υλοποίηση δέντρου αποφάσεων για την πρόβλεψη των αθλημάτων ενασχόλησης των εφήβων - Εφαρμογή

3.2 Ταξινόμηση κατά Naïve Bayes (Naïve Bayes classifier)

Πρόκειται για αλγορίθμους που χρησιμοποιούνται για την κατηγοριοποίηση – ταξινόμηση των δεδομένων σε επιμέρους κλάσεις και ανήκουν στην κατηγορία των αλγορίθμων επιτηρούμενης μάθησης. Θεωρούνται ότι έχουν πλεονέκτημα έναντι άλλων προσεγγίσεων ταξινόμησης, διότι είναι αρκετά απλοί ως προς την υλοποίηση και την αρχή λειτουργίας, ενώ ταυτόχρονα δίνουν την δυνατότητα εύκολης ερμηνείας των αποτελεσμάτων τους. Ανατρέχοντας στην βιβλιογραφία, διαπιστώνεται η ύπαρξη πολλών συστημάτων και εφαρμογών που έχουν αναπτυχθεί, με χρήση των αλγορίθμων αυτών, κυρίως σε προβλήματα αναγνώρισης ήχου, επεξεργασίας γραπτού ή φυσικού λόγου καθώς και επεξεργασίας εικόνων και βίντεο (Yao and Ye, 2020). Τα τελευταία χρόνια, παρατηρείται μια τάση για επέκταση του εύρους εφαρμογών των αλγορίθμων και σε διαφορετικά πεδία, όπως είναι η υπολογιστική νέφους (fog computing) (Onah, et al., 2021), τα ζητήματα ασφάλειας στον κυβερνοχώρο, η μοντελοποίηση της κατανάλωσης ενέργειας παραγωγικών και βιομηχανικών μονάδων (Venkata and Pandya, 2022) και άλλα πεδία.

Όσον αφορά την υλοποίηση και την αρχή λειτουργίας, πρόκειται για μεθόδους μάθησης που είναι βασισμένες στη στατιστική και πιο συγκεκριμένα στις κατανομές πιθανότητας. Οι Naïve Bayes είναι μία οικογένεια απλών πιθανολογικών ταξινομητών οι οποίοι βασίζονται στην εφαρμογή του θεωρήματος του Bayes και προϋποθέτουν την επεξεργασία δεδομένων με εντελώς ανεξάρτητες συσχετίσεις μεταξύ των χαρακτηριστικών. Πιο συγκεκριμένα, η λογική υλοποίησης των αλγορίθμων στηρίζεται στο γεγονός ότι οποιοδήποτε διαθέσιμο δεδομένο μπορεί να εμπίπτει σε μία ή περισσότερες από μια ομάδα κατηγοριών (ή όχι). Σε αρκετές αναφορές, οι αλγόριθμοι ονομάζονται και ως “naive = αφελής”, καθώς η αποτελεσματικοί λειτουργία τους στηρίζεται στην ανεξαρτησία μεταξύ των μεταβλητών, κάτι που στην πραγματικότητα είναι πρακτικά αδύνατο για τις περισσότερες από τις βάσεις δεδομένων που περιγράφουν πραγματικά δυναμικά προβλήματα.

Οι συγκεκριμένοι αλγόριθμοι, επιλέγονται για την επίλυση διάφορων προβλημάτων ταξινόμησης, καθώς θεωρείται ότι έχουν δύο βασικά πλεονεκτήματα έναντι των υπολοίπων αλγορίθμων ταξινόμησης. Αρχικά, στηρίζονται σε ένα πολύ απλό μαθηματικό μοντέλο, το οποίο περιγράφεται παρακάτω, γεγονός που επιτρέπει την άμεση χρήση σε πολλές βάσεις δεδομένων και ταυτόχρονα δίνει τη δυνατότητα γρήγορων και εύκολων υπολογισμών, καθώς και εύκολα ερμηνεύσιμων μοντέλων χωρίς την απαίτηση ιδιαίτερα υψηλής υπολογιστικής ισχύος. Επιπρόσθετα, το κύριο πλεονέκτημα είναι ότι λειτουργεί με μεγάλη αξιοπιστία και αποτελεσματικότητα για μικρό όγκο δεδομένων, το οποίο χαρακτηριστικό είναι ιδιαίτερα

σημαντικό, ειδικά αν αναλογιστεί κανείς ότι τα περισσότερα μοντέλα μηχανικής μάθησης, χρειάζονται πάρα πολύ μεγάλο όγκο δεδομένων προκειμένου να είναι αποτελεσματικά. Γενικότερα, θα μπορούσαμε να θεωρήσουμε πως ισχύει ο εμπειρικός κανόνας, σύμφωνα με τον οποίο όσο πιο σύνθετες είναι οι προσεγγίσεις υλοποίησης ενός μοντέλου, τόσο μεγαλύτερος αριθμός δεδομένων απαιτείται προκειμένου να λειτουργούν αποτελεσματικά.

Από την άλλη πλευρά, η θεμελιώδης λογική της μεθόδου στηρίζεται στο γεγονός ότι όλες οι μεταβλητές της βάσης, θα πρέπει να είναι ανεξάρτητες μεταξύ τους, όπως σημειώθηκε και παραπάνω. Σε ορισμένες περιπτώσεις, η θεώρηση αυτή είναι εντελώς άτοπη και ως εκ τούτου τα μοντέλα αποτυγχάνουν ως προς την αποτελεσματική ταξινόμηση, καθώς δεν μπορούν να λάβουν υπόψιν τις υπάρχουσες προσεγγίσεις. Αυτός κατά βάση είναι ο κύριος λόγος χρησιμοποίησης των αλγορίθμων μόνο σε περιπτώσεις όπου οι μεταβλητές έχουν μεγάλο βαθμό ανεξαρτησίας μεταξύ τους, όπως για παράδειγμα η επεξεργασία εικόνων, καθώς και η μη χρήση σε ισχυρά συσχετισμένα δεδομένα, όπως για παράδειγμα τα ιατρικά δεδομένα ή τα δεδομένα παραγωγής στη βιομηχανία.

3.2.1 Στάδια υλοποίησης αλγορίθμων

Όπως αναφέρθηκε και παραπάνω, η αρχή λειτουργίας των αλγορίθμων στηρίζεται στο θεώρημα του Bayes. Σύμφωνα με το συγκεκριμένο θεώρημα, η πιθανότητα να πραγματοποιηθεί ένα ενδεχόμενο B με δεδομένο ότι έχει πραγματοποιηθεί το ανεξάρτητο ενδεχόμενο A, δίνεται από την παρακάτω σχέση:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Για να επεκταθεί η σχέση του Bayes, σε επίπεδο ταξινόμησης πολλών μεταβλητών, οι οποίες αντιπροσωπεύουν τα διάφορα χαρακτηριστικά της βάσης, θα πρέπει να ακολουθηθούν τα παρακάτω βήματα. Στο σημείο αυτό αναφέρεται ότι η διαδικασία μείωσης του αριθμού των μεταβλητών, που λαμβάνει χώρα πριν την υλοποίηση του αλγορίθμου, είναι πάρα πολύ σημαντική προκειμένου να διατηρηθεί η πολυπλοκότητα σε φυσιολογικά επίπεδα και το μοντέλο να εστιαστεί σε συγκεκριμένα χαρακτηριστικά.

Στάδιο 1: Δημιουργία διανύσματος μεταβλητών

Διατρέχοντας όλα τα χαρακτηριστικά της βάσης, ο αλγόριθμος δημιουργεί το διάνυσμα των μεταβλητών ή χαρακτηριστικών. Αν υποθέσουμε ότι το σύνολο των μεταβλητών (variables) είναι n στον αριθμό, τότε δημιουργείται ο κάτωθι μονοδιάστατος πίνακας, ο οποίος στην μαθηματική ανάλυση καλείται διάνυσμα.

$$V = [V_1, V_2, \dots, V_n]$$

Στο συγκεκριμένο στάδιο θα πρέπει να αποδίδεται ιδιαίτερη προσοχή, ως προς την επιλογή της βάσης εκπαίδευσης του μοντέλου, με γνώμονα αυτή να περιλαμβάνει όλα τα χαρακτηριστικά, δηλαδή τις μεταβλητές, για τα οποία θα γίνει στη συνέχεια η ταξινόμηση. Αν κάτι τέτοιο δεν συμβεί, δηλαδή αν υπάρξουν νέες μεταβλητές στο στάδιο του ελέγχου, για τις οποίες ο αλγόριθμος δεν έχει εκπαιδευθεί, τότε στην πραγματικότητα θα αγνοηθούν και θα θεωρηθούν ως μηδενικές κλάσεις, οδηγώντας σε αδυναμία ταξινόμησης σε αυτές.

Στάδιο 2: Κωδικοποίηση μεταβλητών εκπαίδευσης

Μόλις ολοκληρωθεί η διαδικασία δημιουργία του διανύσματος μεταβλητών, στη συνέχεια για τα δεδομένα κάθε μεταβλητής δημιουργείται μια ετικέτα, C_{vi}. Η ετικέτα χρησιμεύει στην ομαδοποίηση των δεδομένων και σε επόμενο στάδιο στον υπολογισμό της πιθανότητας, εμφάνισης κάποιας ομάδας θεωρώντας ως δεδομένο ότι διατρέχεται μια συγκεκριμένη μεταβλητή.

Στάδιο 3: Υπολογισμός πιθανότητας εμφάνισης μιας κλάσης σε κάθε μεταβλητή με εφαρμογή του θεωρήματος Bayes

Για τον υπολογισμός της πιθανότητας εύρεσης μιας κλάσης, στις διάφορες μεταβλητές του προβλήματος, ο τύπος του θεωρήματος Bayes, επεκτείνεται ως εξής:

$$P(C_i|V_1, V_2, \dots, V_n) = \frac{P(V_1, V_2, \dots, V_n|C_i) * P(V_i)}{P(V_1, V_2, \dots, V_n)} \; for \; 1 < i < k$$

Από πλευράς υπολογισμών, ο παρονομαστής του παραπάνω κλάσματος, δεν έχει κάποια ιδιαίτερη πολυπλοκότητας, καθώς υπολογίζεται απλώς κάνοντας χρήση των κανόνων της περιγραφικής στατιστικής και της αθροιστικής συνάρτησης πυκνότητας πιθανότητας. Αντίθετα, όσον αφορά τον αριθμητή, η σχέση υπολογισμού του μπορεί να απλοποιηθεί σημαντικά, κάνοντας εφαρμογή του αλυσιδωτού κανόνα υπολογισμού της από κοινού συνάρτησης πιθανότητας. Κατόπιν της εφαρμογής του συγκεκριμένου κανόνα και των αντίστοιχων απλοποιήσεων, μπορούμε τελικά να καταλήξουμε στην κάτωθι απλοποιημένη μορφή:

$$P(C_k|V_i) = \frac{P(C_k) * \prod(V_i, C_k)}{P(V_i)}$$

Μόλις προσδιοριστούν οι πιθανότητες της κάθε επιμέρους κλάσης σε κάθε μεταβλητή, στη συνέχεια ο αλγόριθμος αποδίδει στη μεταβλητή την «ετικέτα» με την μεγαλύτερη πιθανότητα εμφάνισης, θεωρώντας ότι κατά κύριο λόγο η μεταβλητή περιγράφετε από τη συγκεκριμένη ομάδα δεδομένων. Ολοκληρώνοντας, σημειώνεται ότι όσο περισσότερα δεδομένα καλείται να

επεξεργαστεί ο αλγόριθμος, τόσο αυξάνεται η δυσκολία και η διάρκεια των υπολογισμών πιθανότητας. Για την αντιμετώπιση αυτού του ζητήματος, θα πρέπει είτε να χρησιμοποιούνται εξ' αρχής σχετικά μικρές βάσεις δεδομένων, βασισμένες σε εμπειρικούς κανόνες για την εύρεση του κατάλληλου αριθμού εγγραφών ανά μεταβλητή, είτε να γίνεται προσπάθεια συρρίκνωσης ή ομαδοποίησης των εγγραφών, κατά το στάδιο της προ-επεξεργασίας. Στην συνέχεια της ενότητας, παρατίθεται αναλυτικό παράδειγμα υλοποίησης, προς διασαφήνιση των παραπάνω.

3.2.2 Εναλλακτικές προσεγγίσεις υλοποίησης

Τα στάδια υλοποίησης, που αναφέρθηκαν παραπάνω, αποτελούν το γενικό μεθοδολογικό πλαίσιο υλοποίησης των συγκεκριμένων αλγορίθμων. Στην πραγματικότητα, μπορεί να συναντήσει κανείς διάφορες παραλλαγές αλγορίθμων, οι οποίες βασίζονται στο παραπάνω πλαίσιο, αλλά διαφοροποιούνται ανάλογα με την ιδιομορφία του εκάστοτε προβλήματος, και πιο συγκεκριμένα ανάλογα με τον τύπο τον χρησιμοποιούμενων μεταβλητών. Οι εναλλακτικές προσεγγίσεις καταγράφονται παρακάτω, μαζί με τον τύπο των μεταβλητών που εξυπηρετούν. Στο σημείο αυτό αναφέρεται, ότι στη γενική περίπτωση μπορούν να χρησιμοποιηθούν και υβριδικές μορφές αλγορίθμων, δηλαδή συνδυασμός των παρακάτω μορφών, στην περίπτωση που υπάρχουν αρκετές μεταβλητές με διαφορετικό τύπο οι οποίες περιέχουν σημαντική πληροφορία και δεν μπορούν να απαλειφθούν ή στην περίπτωση που κριθεί πως το υβριδικό μοντέλο είναι απλώς πιο αποτελεσματικό, ως προς τις προβλέψεις ταξινόμησης, συγκριτικά με τα επιμέρους απλά μοντέλα.

Οι εναλλακτικές προσεγγίσεις ανάλογα με τον τύπο των μεταβλητών είναι:

- **Πολυωνυμικοί ταξινομητές κατά Bayes (Multinomial Naïve Bayes classifiers)**: Χρησιμοποιείται σε βάσεις δεδομένων, οι οποίες περιέχουν διακριτά χαρακτηριστικά, τα οποία μπορούν εύκολα να ταξινομηθούν με μια κύρια ετικέτα ανά χαρακτηριστικό. Ένα χαρακτηριστικό παράδειγμα, είναι η ταξινόμηση βιβλίων σε λογοτεχνικά, ιστορικά, επιστημονικά κ.λπ.
- **Ταξινομητές Bernoulli – Bayes:** Χρησιμοποιούνται επίσης, σε βάσεις με διακριτά χαρακτηριστικά, με κύριες ετικέτες περιγραφής. Η μοναδική διαφορά, έγκειται στο γεγονός ότι απαιτούν δυαδικές λογικές στα δεδομένα, τύπου 0 και 1, και κατ' επέκταση δημιουργούν και επεξεργάζονται μόνο δύο επιμέρους κλάσεις. Ένα παράδειγμα εφαρμογής, θα μπορούσε να ήταν ο χαρακτηρισμός εξαρτημάτων σε λειτουργικά ή ελαττωματικά, σε μια παραγωγική διαδικασία.

- **Gaussian Naïve Bayes algorithms:** Πρόκειται για αλγόριθμους που χρησιμοποιούνται σε βάσεις με συνεχείς μεταβλητές – χαρακτηριστικά. Στην ουσία, είναι η επέκταση της παραπάνω λογικής σε συνεχείς μεταβλητές, η οποία για να υλοποιηθεί, στηρίζεται στη λογική ότι οι συνεχείς μεταβλητές της βάσης, ακολουθούν την κατανομή Gauss. Αυτή η παραδοχή είναι συχνά ιδιαίτερα περιοριστική για τη γενίκευση της συγκεκριμένης μεθοδολογίας, ωστόσο αποτελεί έναν τρόπο υπολογισμού των πιθανοτήτων. Για την ισχύ της παραδοχής, θα πρέπει οι συνεχείς μεταβλητές να κανονικοποιηθούν ώστε να ακολουθούν την μορφή Gauss, χρησιμοποιώντας κατάλληλες σχέσεις. Στη συνέχεια θα πρέπει να ελέγχεται αν η κατανομή που δημιουργείται προσεγγίζει με επάρκεια την κατανομή Gauss και στη συνέχεια να γίνεται η υλοποίηση του μοντέλου. Σε κάθε περίπτωση, μετά την υλοποίηση, θα πρέπει να εξετάζεται η αποτελεσματικότητα των μοντέλων που προκύπτουν.

3.2.3 Εφαρμογή Naïve Bayes

Με στόχο την αποσαφήνιση των όσων καταγράφηκαν παραπάνω σχετικά με την λειτουργία των αλγορίθμων ταξινόμησης κατά Naïve Bayes, στο σημείο αυτό θα παραθέσουμε μια απλή εφαρμογή, στην οποία θα γίνουν αναλυτικά όλοι οι υπολογισμοί. Η εφαρμογή προέρχεται από ένα εκ των κυριότερων πεδίων εφαρμογής των αλγορίθμων, αυτό της εξαγωγής γνώσης από κείμενα (text mining).

Πιο συγκεκριμένα, στο πλαίσιο της εφαρμογής, επιθυμούμε να ταξινομήσουμε και να προβλέψουμε εάν κάποιο σχόλιο που έχει γίνει στα μέσα κοινωνικής δικτύωσης αφορά ένα συγκεκριμένο αθλητικό γεγονός, έτσι ώστε να αποκτήσουμε μια άμεση εικόνα, σχετικά με την εντύπωση που έχουν αποκομίσει άλλοι χρήστες. Για την εξυπηρέτηση του σκοπού αυτού, συλλέγονται προς ανάλυση τα παρακάτω δεδομένα και καταγράφονται στον Πίνακα 7. Τα δεδομένα αυτά, αποτελούν στην πραγματικότητα το τμήμα της βάσης που θα χρησιμοποιηθεί για την εκπαίδευση του μοντέλου, και στη συνέχεια, ένα άλλο string από άγνωστα δεδομένα, θα ταξινομηθεί σε κάποια από της δύο κατηγορίες.

Σχόλια	Κατηγορία
«Ένα καλό παιχνίδι»	Αθλητικά
«Οι εκλογές τελείωσαν»	Άλλο
«Πολύ ξεκάθαρο αποτέλεσμα»	Αθλητικά
«Ένα ξεκάθαρο και εύκολο παιχνίδι »	Αθλητικά
«Ένα οριακό εκλογικό αποτέλεσμα»	Άλλο

Πίνακας 7 Δεδομένα εκπαίδευσης αλγορίθμου Naïve Bayes, για εξόρυξη γνώσης από κείμενα - Εφαρμογή

Έχοντας ολοκληρώσει την εκπαίδευση του μοντέλου με τα δεδομένα του Πίνακα 7, στη συνέχεια θα πρέπει να ελεγχθεί ο βαθμός αποτελεσματικότητας του μοντέλου. Για τον έλεγχο λειτουργίας, μπορούν να χρησιμοποιηθούν δύο διαφορετικές προσεγγίσεις. Είτε να δοκιμαστεί η δυνατότητα πρόβλεψης σε μια αυτούσια φράση, για την οποία είναι γνωστή εκ των προτέρων η κατηγορία στην οποία ανήκει, δηλαδή Αθλητικά ή άλλο. Είτε να δοκιμαστεί μια εντελώς νέα φράση, για την οποία το μοντέλο δεν έχει εκπαιδευτεί και να δοκιμαστεί σε αυτή η αποτελεσματικότητά. Στο συγκεκριμένο παράδειγμα, επιλέγεται η δεύτερη περίπτωση, για δύο βασικούς λόγους. Αρχικά, διότι θέλουμε να ελέγξουμε και να διαπιστώσουμε τις δυνατότητες επέκτασης της εφαρμογής των μοντέλων, σε διαφορετικές βάσεις από αυτή που χρησιμοποιείται για την εκπαίδευση. Και αφετέρου, διότι είναι πρακτικά αδύνατό τα σχόλια που γίνονται στα μέσα κοινωνικής δικτύωσης θα είναι πάντοτε όμοια, με άλλα σχόλια που είχα γίνει για το ίδιο αντικείμενο στο παρελθόν.

Επιλέγοντας, να ελέγξουμε την αποτελεσματικότητα του μοντέλου σε άγνωστα δεδομένα (text string), θα πρέπει να σημειώσουμε ότι σε πρωταρχικό επίπεδο λειτουργίας, δεν θα βρεθεί καμία ομοιότητα. Έτσι, θα υλοποιηθεί ένα δεύτερο επίπεδο εύρεσης ομοιότητας, κατά το οποίο θα ελεγχθεί η κάθε φράση λέξη – λέξη ως προς την ομοιότητάς της με την φράση που επιθυμούμε να ταξινομήσουμε. Αυτό το επίπεδο λειτουργίας, ονομάζεται συχνά και ως «αφέλεια του αλγορίθμου», και είναι πολύ αποτελεσματικό σε τέτοιου είδους προβλήματα.

Ας θεωρήσουμε λοιπόν, ότι θέλουμε να ελέγξουμε αν η φράση: «ένα πολύ οριακό παιχνίδι». Σύμφωνα με τα όσα αναφέρθηκαν, θα πρέπει να ελέγξουμε λέξη-λέξη τη φράση και να την αντιστοιχίσουμε με κάποια από τις δύο κατηγορίες. Επομένως, χρησιμοποιούμε τις κάτωθι σχέσεις. Προτού προχωρήσουμε στον αναλυτικό υπολογισμό και την ταξινόμηση της φράσης, είναι σημαντικό να αναφερθεί ότι σύμφωνα με αυτή τη προσέγγιση, αν κάποια από τις λέξεις που αποτελούν την φράση ελέγχου δεν υπάρχουν καν στα δεδομένα εκπαίδευσης, τότε το γινόμενο υπολογισμού των πιθανοτήτων θα μηδενιστεί και έτσι το μοντέλο δεν θα μπορέσει να κάνει καμία πρόβλεψη σχετικά με την ταξινόμηση.

Αυτό αποτελεί ένα από τα κλασικά προβλήματα στην ανάλυση δεδομένων κειμένου, για την αποφυγή του οποίου έχουν αναπτυχθεί διάφορες βοηθητικές μέθοδοι. Μια από τις πιο απλές και αποτελεσματικές, είναι η μέθοδος εξομάλυνσης των δεδομένων κατά Laplace (Laplace smoothing method), σύμφωνα με την οποία σε κάθε επιμέρους διαφορετική λέξη προστίθεται μία μονάδα, προκειμένου καμία να μην μηδενιστεί η πιθανότητα που της αντιστοιχεί κατά τους υπολογισμούς. Στη συγκεκριμένη περίπτωση έχουμε 14 διαφορετικές λέξεις. Προς διασαφήνιση της μεθόδου παρατίθενται οι Πίνακες 8 και 9, στους οποίους καταγράφεται κάθε επιμέρους πιθανότητα που θα χρειαστεί για τον υπολογισμό των κάτωθι σχέσεων, πριν και μετά την εξομάλυνση Laplace.

Αρχικά για την κατηγορία Αθλητικά:

Λέξη	**p (λέξης/Αθλητικά)** ***πριν την εξομάλυνση***	**p(λέξης/Αθλητικά)** ***μετά την εξομάλυνση***
ένα	2/12 = **0,167**	(2+1)/(12+14) = **0,115**
πολύ	1/12 = **0,08**	(1+1)/(12+14) = **0,08**
οριακό	0/12 = **0**	(0+1)/(12+14) = **0,04**
παιχνίδι	2/12 = **0,167**	(2+1)/(12+14) = **0,115**

Πίνακας 8 Πιθανότητα εύρεσης μιας λέξης με ετικέτα Αθλητικά, πριν και μετά την εξομάλυνση Laplace - Εφαρμογή

Χρησιμοποιώντας τα δεδομένα του Πίνακα 5, τελικά υπολογίζουμε ότι:

$$p\ (ένα\ πολύ\ οριακό\ παιχνίδι|Αθλητικά)$$
$$= p(ένα|Αθλητικά) * p(πολύ|Αθλητικά) * p(οριακό|Αθλητικά)$$
$$* p(παιχνίδι|Αθλητικά) = 0{,}115 * 0{,}08 * 0{,}04 * 0{,}115 = 0{,}00004$$

Έπειτα, για την κατηγορία Άλλο:

Λέξη	**p (λέξης/Άλλο)** ***πριν την εξομάλυνση***	**p(λέξης/Άλλο)** ***μετά την εξομάλυνση***
ένα	1/7 = **0,143**	(1+1)/(7+14) = **0,095**
πολύ	0/7 = **0**	(0+1)/(7+14) = **0,048**
οριακό	1/7 = **0,143**	(1+1)/(7+14) = **0,095**
παιχνίδι	0/7 = **0**	(0+1)/(7+14) = **0,048**

Πίνακας 9 Πιθανότητα εύρεσης μιας λέξης με ετικέτα Άλλο, πριν και μετά την εξομάλυνση Laplace - Εφαρμογή

$$p\ (ένα\ πολύ\ οριακό\ παιχνίδι|Άλλο)$$
$$= p(ένα|Άλλο) * p(πολύ|Άλλο) * p(οριακό|Άλλο) * p(παιχνίδι|Άλλο)$$
$$= 0{,}095 * 0{,}048 * 0{,}095 * 0{,}048 = 0.00002$$

Συγκρίνοντας τις πιθανότητες που προέκυψαν μετά τους υπολογισμούς, διαπιστώνουμε ότι η πιθανότητα που αφορά την κατηγορία Αθλητικά είναι μεγαλύτερη από την πιθανότητα που αφορά την κατηγορία Άλλο. Κατά συνέπεια, ο αλγόριθμος θα ταξινομήσει την φράση: «ένα πολύ οριακό παιχνίδι» στην κατηγορία Αθλητικά, κάτι που είναι σωστό, σύμφωνα με την αρχική μας θεώρηση.

Με την χρήση του συγκεκριμένου παραδείγματος, διαπιστώνουμε πέραν των υπολοίπων, ότι οι συγκεκριμένοι αλγόριθμοι είναι πάρα πολύ αποτελεσματικοί σε μικρές βάσεις δεδομένων και σε εφαρμογές όπως αυτή της εξόρυξης γνώσης από κείμενα. Με τον τρόπο αυτό, αποδεικνύονται τα όσα καταγράφηκαν στο θεωρητικό πλαίσιο της παρούσας ενότητας.

3.3 Ο αλγόριθμος k-NN

3.3.1 Αρχή λειτουργίας αλγορίθμου

Πρόκειται για έναν επιτηρούμενο αλγόριθμο μηχανικής μάθησης, που έχει βρει μεγάλο πλήθος εφαρμογών, καθώς μπορεί να χρησιμοποιηθεί τόσο για την επίλυση προβλημάτων ταξινόμησης όσο και για πρόβλεψη μελλοντικών τιμών, εκτελώντας κάποιο είδος παλινδρόμησης (Κουρνέτας, 2017). Η πλήρης αγγλική ονομασία του αλγορίθμου είναι: k – Nearest Neighbors, ενώ η αντίστοιχη ελληνική k – πλησιέστεροι γείτονες.

Όπως μπορεί εύκολα να γίνει αντιληπτό, από την ονομασία του αλγορίθμου, η αρχή λειτουργίας του στηρίζεται στην εύρεση μοτίβων ομοιότητας μεταξύ k δεδομένων, τοποθετημένων σε μια χωρική απεικόνιση. Υπό αυτή την έννοια, μπορούμε σε ένα πρώτο επίπεδο ανάλυσης να σημειώσουμε ότι ένα πολύ κρίσιμο χαρακτηριστικό για την αποτελεσματική λειτουργία του αλγορίθμου, είναι ο προσδιορισμός της κατάλληλης τιμής k.

Η μη κατάλληλη επιλογή της σταθεράς k, εγκυμονεί κινδύνους μείωσης της αποτελεσματικότητας της διαδικασίας εκμάθησης, σε πρώτο στάδιο, και της ικανότητας προβλέψεων, σε επόμενο στάδιο. Χαρακτηριστικά αναφέρεται ότι, σε περίπτωση όπου επιλεγεί k σημαντικά μικρότερο από το ιδανικό, τότε στην πραγματικότητα μειώνεται σημαντικά η ικανότητα επεκτασιμότητας και προβλέψεων του αλγορίθμου, ενώ αντίθετα στην περίπτωση που επιλεγεί σημαντικά μεγαλύτερη τιμή για το k, έναντι της ιδανικής, τότε το μοντέλο αρχίζει να λαμβάνει υπόψιν του ομοιότητες μεταξύ αρκετά «μακρινών» δεδομένων, τα οποία θεωρεί ως «κοντινά» εσφαλμένα. Κατά συνέπεια, αυτό οδηγεί σε μείωση της αποτελεσματικότητας και της αξιοπιστίας των προβλέψεων ή των ταξινομήσεων στις εκάστοτε κλάσεις.

Υπάρχουν διάφορες προσεγγίσεις που έχουν αναπτυχθεί για τον υπολογισμό της τιμής k, με μία από τις ευρέως χρησιμοποιούμενες να είναι η χρήση του εμπειρικού κανόνα της τετραγωνικής ρίζας. Σύμφωνα με αυτόν τον κανόνα, η τιμή του k, προτείνεται να είναι ίση με την τετραγωνική ρίζα των παρατηρήσεων, στρογγυλοποιημένη στον πλησιέστερο ακέραιο, χωρίς κανένα δεκαδικό ψηφίο. Αν και ο συγκεκριμένο εμπειρικός κανόνας, φαίνεται να έχει αποδώσει ικανοποιητικά, στο συγκεκριμένο σημείο θα λέγαμε ότι είναι προτιμότερη μια κυκλική διαδικασία επαναλήψεων με ελαφρώς διαφοροποιημένες τιμές του k (try and error analysis), έτσι ώστε να βρεθεί εκείνη η τιμή που αποδίδει τα καλύτερα αποτελέσματα από πλευράς των μετρικών ελέγχου.

Το επόμενο στοιχείο ενδιαφέροντος από τεχνικής σκοπιάς, είναι ο προσδιορισμός των πραγματικών αποστάσεων μεταξύ των σημείων, τα οποία προκύπτουν από την γραφική απεικόνιση των δεδομένων, σε ένα δισδιάστατο διάγραμμα. Στην πραγματικότητα, οι αποστάσεις μεταξύ των σημείων, αποτελούν μια επιπρόσθετη μετρική ελέγχου, ειδικά σχεδιασμένη για τον αλγόριθμο k-NN. Πιο συγκεκριμένα, έχοντας γνωστές τις αποστάσεις των σημείων, γίνεται αρκετά απλός ο έλεγχος της αποτελεσματικότητας του αλγορίθμου και κατ' επέκταση της σωστής επιλογής του k, για την δημιουργία της «γειτονιάς».

Μια από τις πιο κλασικές τεχνικές για τον υπολογισμό των αποστάσεων μεταξύ των σημείων, είναι η χρήση του τύπου της Ευκλείδειας Απόστασης, ο οποίος για δύο σημεία i και j, με συντεταγμένες (xi,yi) και (xj,yj), αντίστοιχα, ισούται με:

$$d\ (i,j) = \sqrt{\frac{(x^i - y^i)^2 + (x^j - y^j)^2}{2}}$$

Σαφώς, η παραπάνω σχέση μπορεί να επεκταθεί και για περισσότερα από δύο σημεία, μετατρέποντας το άθροισμα των δύο όρων σε άθροισμα όλων των επιμέρους όρων και αντικαθιστώντας τον αριθμό 2, με το σύνολο των εξεταζόμενων σημείων, έστω n. Επιπλέον, σημειώνεται ότι η Ευκλείδεια Απόσταση, δεν αποτελεί την μοναδική μετρική υπολογισμού των πραγματικών αποστάσεων, αλλά υπάρχουν κι άλλες μετρικές όπως για παράδειγμα η απόσταση κατά Chebychev η απόσταση κατά Minkowski και άλλες.

3.3.2 Εφαρμογή k-NN

Με στόχο την αναλυτική παρουσίαση της λειτουργίας του αλγορίθμου και την πλήρη αποσαφήνιση των όσων αναφέρθηκαν στην προηγούμενη ενότητα, στην παρούσα ενότητα παρατίθεται μια σύντομή εφαρμογή του αλγορίθμου k-NN, σε πρόβλημα ταξινόμησης – κατηγοριοποίησης.

Πιο συγκεκριμένα το πρόβλημα αφορά στην απόδοση ενός χαρακτηρισμού, για ένα τελικό προϊόν από μια παραγωγική διαδικασία, εκ των: ελλαττωματικό ή λειτουργικό ή χρειάζεται βελτίωση. Για την πλήρωση των αναγκών της εφαρμογής, θεωρούμε ότι τα δύο κύρια σημεία ελέγχου, για την απόδοση μιας εκ των παραπάνω ετικετών σε ένα προϊόν, είναι το βάρος και το μήκος. Πιο συγκεκριμένα αν το βάρος είναι κυμαίνεται από 10 έως 13 κιλά τότε το προϊόν είναι λειτουργικό, ενώ αν το βάρος ξεπερνά τα 13 κιλά ή είναι μικρότερο από 10 κιλά τότε το προϊόν είναι ελαττωματικό, τέλος στο βάρος δεν μπορεί να γίνει κάποια βελτιωτική παρέμβαση. Αντίστοιχα, αν το μήκος κυμαίνεται από 2 έως 3 μέτρα τότε το προϊόν είναι

λειτουργικό, ενώ αν κυμαίνεται από 2 έως 4 μέτρα τότε θέλει βελτίωση, και τέλος αν ξεπερνά τα 4 μέτρα ή είναι μικρότερο από 2 μέτρα τότε είναι ελαττωματικό.

Στα πλαίσια δειγματοληπτικού ελέγχου, το τμήμα ποιότητα συνέλεξε τα δεδομένα που παρουσιάζονται στον Πίνακα 10 και με βάση τα δεδομένα αυτά θέλει να αναπτύξει ένα σύστημα πρόβλεψης της κλάσης των μελλοντικών εξαρτημάτων, χρησιμοποιώντας τον αλγόριθμο k-NN.

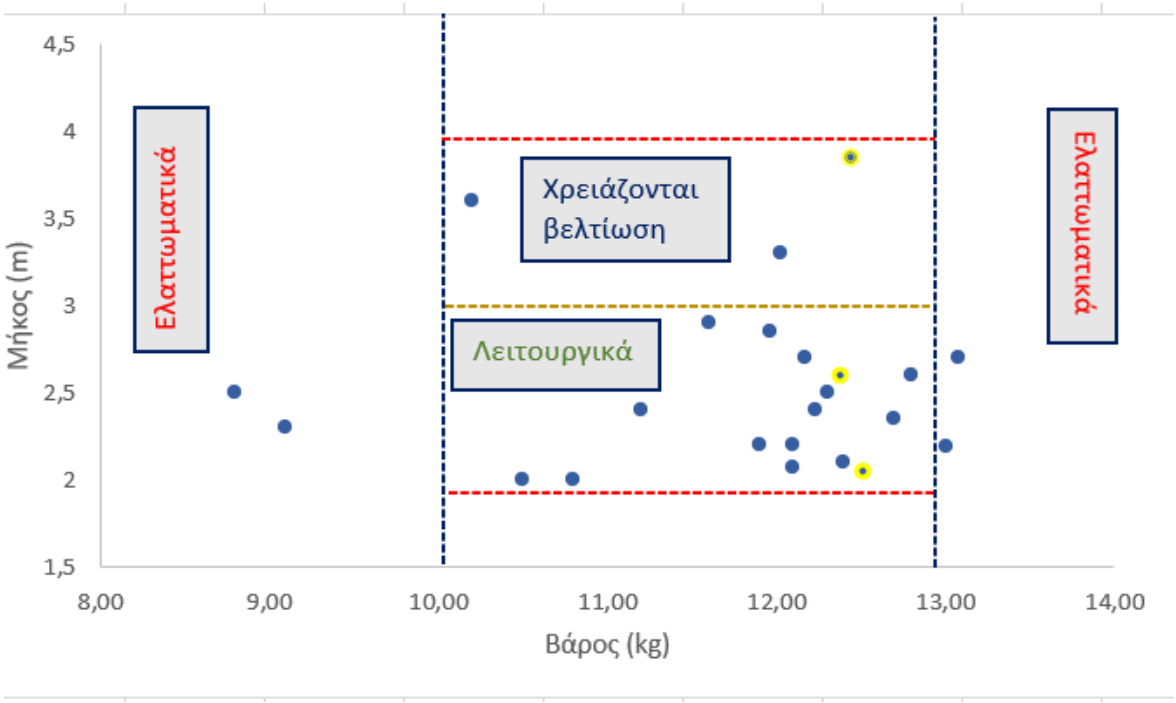
Βάρος	Μήκος	Χαρακτηρισμός
10,80	2	Λειτουργικό
11,20	2,4	Λειτουργικό
13,08	2,7	Ελαττωματικό
12,10	2,2	Λειτουργικό
12,40	2,1	Λειτουργικό
10,20	3,6	Χρειάζεται βελτίωση
10,50	2	Λειτουργικό
9,10	2,3	Ελαττωματικό
8,80	2,5	Ελαττωματικό
11,60	2,9	Λειτουργικό
11,90	1,8	Ελαττωματικό
12,70	2,35	Λειτουργικό
13,00	2,19	Λειτουργικό
12,80	2,6	Λειτουργικό
11,96	2,85	Λειτουργικό
12,03	3,3	Χρειάζεται βελτίωση
12,10	1,9	Ελαττωματικό
12,17	2,7	Λειτουργικό
12,23	2,4	Λειτουργικό
12,30	2,5	Λειτουργικό
12,37	2,6	Λειτουργικό
12,44	4	Χρειάζεται βελτίωση
12,51	2,05	Λειτουργικό

Πίνακας 10 Βάση δεδομένων για υλοποίηση αλγορίθμου k-NN - Εφαρμογή

Αναλύοντας τον Πίνακα 10, αρχικά μπορούμε να συμπεράνουμε ότι έχουμε συνολικά 23 μετρήσεις. Αυτή η παρατήρηση, μας οδηγεί σε δύο πορίσματα. Αρχικά, για την ανάπτυξη ενός συστήματος μηχανικής μάθησης, χρειάζεται η ισχυρή πλειοψηφία των δεδομένων να χρησιμοποιηθεί για την εκπαίδευση του συστήματος. Το ποσοστό των δεδομένων που θα χρησιμοποιηθεί για την εκπαίδευση προκύπτει μέσα από εμπειρικούς κανόνες, με επικρατέστερη προσέγγιση το «80-20», δηλαδή για εκπαίδευση το 80% των δεδομένων και για έλεγχο το υπολειπόμενο 20%. Ωστόσο, επειδή στο συγκεκριμένο παράδειγμα η έκταση της βάσης είναι πολύ περιορισμένη, επιλέγεται η προσέγγιση «90-10», με στόχο την καλύτερη δυνατή εκπαίδευση.

Σύμφωνα με το παραπάνω, πόρισμα τα προς εκπαίδευση δεδομένα θα προκύψουν από τις 20 πρώτες γραμμές και για έλεγχο θα χρησιμοποιηθούν οι τελευταίες 3. Αναφορικά με το k, χρησιμοποιώντας τον κανόνα της τετραγωνικής ρίζας, θα χρησιμοποιήσουμε την τιμή 5, η οποία προκύπτει από την στρογγυλοποίηση του αριθμού 4,79 που αποτελεί το ακριβές αποτέλεσμα της ρίζας του 23.

Τέλος, θα εφαρμόσουμε τον αλγόριθμο k-NN για την ταξινόμηση των τριών δεδομένων ελέγχου και θα υπολογίσουμε την αποτελεσματικότητα του μοντέλου. Για την εξυπηρέτηση του σκοπού αυτού, χρησιμοποιείται το Σχήμα 2.



Σχήμα 2 Έλεγχος αποτελεσματικότητας αλγορίθμου k-NN - Εφαρμογή

Τα «κιτρινισμένα» δεδομένα σχετίζονται με τις προβλέψεις του αλγορίθμου k-NN. Όπως μπορεί εύκολα να διαπιστωθεί, οι προβλέψεις συμπίπτουν σε μεγάλο βαθμό με την πραγματική θέση των δεδομένων και σε κάθε περίπτωση βρίσκονται στην σωστή κλάση, όπως αυτή προέκυψε από τις μετρήσεις. Κατά συνέπεια μπορούμε να συμπεράνουμε ότι αλγόριθμος είναι ιδιαίτερα αποτελεσματικός, ακόμα και αν το πλήθος των δεδομένων είναι αρκετά μικρό. Ένας άλλος τρόπος ελέγχου, θα ήταν η εφαρμογή των ευκλείδειων αποστάσεων και η ταξινόμηση με βάση αυτές, στις αντίστοιχες κλάσεις.

Κεφάλαιο 4: Επίλογος

4.1 Σύνοψη

Σκοπός της παρούσας εργασίας, ήταν η παρουσίαση μιας ολοκληρωμένης μεθοδολογίας για την επίλυση του προβλήματος της ταξινόμησης, με χρήση αλγορίθμων Μηχανικής Μάθησης. Η συγκεκριμένη κατηγορία προβλημάτων επιλέχθηκε διότι βρίσκει μεγάλο πεδίο εφαρμογών σε διαφορετικούς κλάδους. Οι αλγόριθμοι της ταξινόμησης, υπάγονται στο ευρύτερο πλαίσιο των αλγορίθμων επιτηρούμενης μάθησης, και ως εκ τούτου επιτρέπουν την μεγαλύτερη αλληλεπίδραση με τον χρήστη και παράλληλα αποδίδουν αποτελέσματα τα οποία είναι εύκολα ερμηνεύσιμα.

Στην πραγματικότητα, η επίτευξη υψηλών επιπέδων αποτελεσματικότητας κατά την λειτουργία των αλγορίθμων Μηχανικής Μάθησης, απαιτεί την διενέργεια ορισμένων προκάτοχων σταδίων καθώς επίσης και την θέσπιση συγκεκριμένων μετρικών για την αξιολόγησης της αποτελεσματικότητας. Τα στάδια ανάλυσης, πριν την εφαρμογή των αλγορίθμων, ονομάζονται στην βιβλιογραφία ως στάδια προ – επεξεργασίας των δεδομένων και αποσκοπούν στην μορφοποίηση των δεδομένων, έτσι ώστε αυτά να έρθουν στην κατάλληλη μορφή και δομή, για επεξεργασία.

Κατόπιν ανασκόπησης στην βιβλιογραφία διαπιστώθηκαν διάφορες τεχνικές για την επίλυση συγκεκριμένων ζητημάτων, κατά το στάδιο της προ – επεξεργασίας των δεδομένων. Τα ζητήματα που φαίνεται να επηρεάζουν πιο άμεσα την αποτελεσματικότητα των αλγορίθμων, είναι το ζήτημα των χαμένων τιμών, η ύπαρξη θορύβου στα δεδομένα καθώς και το πρόβλημα επιλογής του βέλτιστου αριθμού των μεταβλητών για την διενέργεια των αλγορίθμων. Για τα ανωτέρω ζητήματα, έγινε αναλυτική τεκμηρίωση των τεχνικών αντιμετώπισης ανάλογα με τα εκάστοτε χαρακτηριστικά του συνόλου δεδομένων. Εκτός από τα παραπάνω, σημαντικό μέρος του σταδίου προ – επεξεργασίας των δεδομένων, αποτελεί και ο διαχωρισμός της βάσης δεδομένων, σε βάση εκπαίδευσης και βάση ελέγχου των αποτελεσμάτων της εκπαίδευσης. Όπως διαπιστώθηκε, αυτή η διαδικασία γίνεται με κύριο γνώμονα τα εμπειρικά στοιχεία. Πιο συγκεκριμένα, παρατηρήθηκε ότι δεν υπάρχει συγκεκριμένος τρόπος διαμοιρασμού των βάσεων δεδομένων, ωστόσο έχουν επικρατήσει ορισμένοι εμπειρικοί κανόνες, όπως αυτός του 80% - 20% ή του 70% - 30%, με το μεγαλύτερο ποσοστό των διαθέσιμων δεδομένων να χρησιμοποιείται πάντοτε κατά την διαδικασία εκπαίδευσης των αλγορίθμων.

Αναφορικά με τις μετρικές αξιολόγησης των αποτελεσμάτων των αλγορίθμων, η ανασκόπηση έδειξε την ύπαρξη πολλών μετρικών, με τις περισσότερες από αυτές να βασίζονται στην μήτρα

σύγχυσης των αποτελεσμάτων. Συνεπώς, ιδιαίτερη βαρύτητα αποδόθηκε στην ερμηνεία των στοιχείων της εν λόγω μήτρας, ενώ παράλληλα καταγράφηκαν και ορισμένες διαδεδομένες μετρικές που προκύπτουν με βάση την συγκεκριμένη μήτρα.

Τέλος, στο επίκεντρο της εργασίας ήταν η χρήση και οι εφαρμογές των ίδιων των αλγορίθμων ταξινόμησης. Στην βιβλιογραφία, υπάρχει μεγάλη πληθώρα αλγορίθμων ταξινόμησης, καθώς και διάφορες παραλλαγές αυτών. Ωστόσο, στο πλαίσιο της εργασίας επιλέχθηκαν τρεις από τους πιο διαδεδομένους αλγορίθμους ταξινόμησης: τα δέντρα αποφάσεων, ο ταξινομητής Naïve Bayes και ο αλγόριθμος των k – πλησιέστερων γειτόνων (k-NN). Η επιλογή τριών αλγορίθμων, έγινε με γνώμονα την καλύτερη δυνατή εστίαση στις τεχνικές υλοποίησής τους καθώς και στα ιδιαίτερα σημεία ενδιαφέροντος. Επιπρόσθετα, οι συγκεκριμένοι αλγόριθμοι, βρίσκουν εφαρμογή σε ένα ευρύ πεδίο, από απλά προβλήματα ταξινόμησης σε σχεσιακές βάσεις δεδομένων μέχρι προβλήματα επεξεργασίας φυσικής γλώσσας (NLP). Συνεπώς, η αναλυτική καταγραφή των συγκεκριμένων αλγορίθμων, παρέχει την δυνατότητα αντιμετώπισης μεγάλης γκάμας προβλημάτων.

Ενδεικτικά προβλήματα από διαφορετικούς κλάδους, παρατέθηκαν κατόπιν της ανάλυσης των τεχνικών υλοποίησης των παραπάνω αλγορίθμων, με γνώμονα την εστίαση σε δυνητικές εφαρμογές που μπορούν να αναπτυχθούν. Η παράθεση διαφορετικών παραδειγμάτων σε κάθε αλγόριθμο, βασίστηκε στην καταλληλόλητα των αλγορίθμων για την επίλυση προβλημάτων με συγκεκριμένα χαρακτηριστικά.

4.2 Προτάσεις για επέκταση της εργασίας

Όπως σημειώθηκε και παραπάνω, στην συγκεκριμένη εργασία επιλέχθηκε η εστίαση σε τρεις αλγορίθμους, με γνώμονα την εστίαση στις τεχνικές τους. Ωστόσο, υπάρχουν πολλοί αλγόριθμοι ταξινόμησης που δεν αναφέρθηκαν, λόγω περιορισμών ως προς την έκταση. Συνεπώς, ως κύρια επέκταση της εργασίας, θεωρούμε την παράθεση επιπλέον αλγορίθμων ταξινόμησης (π.χ. Νευρωνικά Δίκτυα) και την αναλυτική περιγραφή των τεχνικών υλοποίησής τους, μέσω παράθεσης των αντίστοιχων μαθηματικών τύπων υπολογισμού.

Επιπλέον, αναφέρθηκε ότι η καταλληλόλητα των αλγορίθμων ποικίλει ανάλογα με το είδος των εφαρμογών. Με βάση την διαπίστωση αυτή, και λαμβάνοντας υπόψιν ότι η εργασία θα έχει εμπλουτιστεί και με επιπρόσθετους αλγορίθμους ταξινόμησης, θεωρείται ιδιαίτερα σημαντική η διενέργεια μιας συγκριτικής θεώρησης μεταξύ των αλγορίθμων, με βάση δύο διαστάσεις: α/ την πολυπλοκότητα των μαθηματικών προσεγγίσεων και β/ το πεδίο εφαρμογής, μέσω παράθεσης αντίστοιχων παραδειγμάτων χρήσης. Ακόμα, η εργασία θα

μπορούσε να εμπλουτιστεί μέσω της παράθεσης ολοκληρωμένων μελετών περίπτωσης τόσο από τον ειχειρηματικό τομέα όσο και από επιστημονικές – ερευντηικές εφαρμογές. Η παράθεση μελετών περίπτωσης, βοηθά σημαντικά στην διαπίστωση των οφελών που προκύτουν λόγω της εφαρμογής τέτοιων προσεγγίσεων, σε πραγματικές συνθήκες. Επομένως, με τον τρόπο αυτόν θα δημιουργούνταν ένα ολοκληρωμένο εγχειρίδιο πληροφόρησης, σχετικά με την ταξινόμηση των δεδομένων.

Καταλήγοντας, σημειώνεται ότι το σύνολο της εργασίας, αφορά το απαραίτητο θεωρητικό υπόβαθρο για την επίλυση ενός σημαντικού βιομηχανικού προβλήματος. Πιο συγκεκριμένα, το σύνολο των τεχνικών που αναλύεται παραπάνω, θα χρησιμοποιηθούν σε πραγματικά βιομηχανικά δεδομένα σχετικά με την παραγωγή, έτσι ώστε να αναπτυχθεί ένα σύστημα ταξινόμησης για την αυτόματη ανίχνευση ελαττωματικών προϊόντων από μια παραγωγική διαδικασία.

Βιβλιογραφία

[1] Giannopoulos, P., Kournetas, G. and Karacapilidis, N. (2021), "On the integration of Machine Learning algorithms and Operations Research techniques in the development of a hybrid Recommender System", *Intelligent Decision Technologies*, vol. 15, pp. 497 – 510.

[2] Han, J., Kamber, M. and Pei, J. (2012), Data Mining: Concepts and Techniques 3rd edition, Morgan Kaufmann Publishers

[3] Kanakaris, N. Karacapilidis, N. and Lazanas, A. (2019), "On the advancement of Project Management through a flexible integration of Machine Learning and Operations Research tools", In: Proceedings of the 8th International Conference on Operations Research and Enterprise Systems.

[4] Karacapilidis, N., Tzagarakis, M. and Christodoulou, S. (2013), "On a meaningful exploitation of machine and human reasoning to tackle data-intensive decision making", *Intelligent Decision Technologies Journal*, vol. 7, no 3, pp. 225-236.

[5] Kaur, R., Ginige, J.A., Obst, O., 2023. AI-based ICD coding and classification approaches using discharge summaries: A systematic literature review. Expert Syst. Appl. 213. https://doi.org/10.1016/j.eswa.2022.118997

[6] Onah, O. J., Abdulhamid, S. M., Abdullahi, M, Hassan, I. H. and Al-Ghusham, A. (2021), "Genetic Algorithm based feature selection and Naïve Bayes for anomaly detection in fog computing environment", *Machine Learning with Applications*, vol. 6, ISSN 2666-8270, https://doi.org/10.1016/j.mlwa.2021.100156.

[7] Osama, S., Shaban, H., Ali, A.A., 2023. Gene reduction and machine learning algorithms for cancer classification based on microarray gene expression data: A comprehensive review. Expert Syst. Appl. 213. https://doi.org/10.1016/j.eswa.2022.118946

[8] Quinlan, J. R. (1986), "Simplifying decision trees", *International Journal of Man-Machine Studies*, vol. 27, pp. 221-234.

[9] Reddy, T. G., Priya S. R. M., Parimala, M., Chowdhary, C. L, Reddy, P. K. M., Hakak, S. and Khan, W. Z. (2020), "A deep neural networks based model for uninterrupted marine environment monitoring", *Computer Communications*, vol. 157, pp. 64-75, ISSN 0140-3664, https://doi.org/10.1016/j.comcom.2020.04.004.

[10] Ritchie, D. (1986), "SHANNON AND WEAVER: Unravelling the Paradox of Information", *Communication Research*, vol. 13, issue 2, pp. 278–298.

[11] Rokach L. and Maimon O. (2008). "Data mining with decision trees. Theory and applications", *The World Scientific Journal*.

[12] Shang, C., Li, M., Feng, S., Jiang, Q., Fan, J. (2013), "Feature selection via maximizing global information gain for text classification", *Knowledge-Based Systems*, vol. 54, pp. 298-309, ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2013.09.019.

[13] Spedicato, G.A., Dutang, C., & Petrini, L. (2018), "Machine Learning Methods to Perform Pricing Optimization. A Comparison with Standard GLMs".

[14] Tan, P. N., Steinbach, M., Kumar, V. (2006), "Classification: Alternative Techniques" in: Introduction to Data Mining, Pearson Publisher.

[15] Venkata, P. and Pandya, V. (2022), "Data mining model and Gaussian Naive Bayes based fault diagnostic analysis of modern power system networks", *Materials Today: Proceedings*, vol. 62, part 13, pp. 7156-7161, ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2022.03.035.

[16] Witten, I. H., Frank, E. and Hall, M. A. (2011), Data mining: Practical Machine Learning tools and techniques, Amsterdam: Morgan Kaufmann, ISBN: 978-0-12-374856-0.

[17] Yao J. and Ye, Y. (2020), "The effect of image recognition traffic prediction method under deep learning and naive Bayes algorithm on freeway traffic safety", *Image and Vision Computing*, vol. 103, ISSN 0262-8856, https://doi.org/10.1016/j.imavis.2020.103971.

[18] Γκίκας, Δ. (2022), Διδακτορική Διατριβή: «Εξόρυξη δεδομένων για βελτιωμένη λήψη αποφάσεων στο μάρκετινγκ: εφαρμογές σε δεδομένα συμπεριφοράς καταναλωτών σε φυσικό και ψηφιακό περιβάλλον με την χρήση μοντέλου μηχανικής μάθησης», Τμήμα Διοίκησης Επιχειρήσεων Αγροτικών Προϊόντων και Τροφίμων, Πανεπιστήμιο Πατρών.

[19] Γκόλια, Γ. Ν. (2021), Διπλωματική Εργασία: «Διερεύνηση χρήσης τεχνικών μηχανικής μάθησης για ταξινόμηση (classification) βιογραφικών», Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Εθνικό Μετσόβιο Πολυτεχνείο.

[20] Κάρλος, Σ. (2020), Διδακτορική Διατριβή: «Ανάπτυξη μερικώς επιβλεπόμενων αλγορίθμων μηχανικής μάθησης», Τμήμα Μαθηματικών, Πανεπιστήμιο Πατρών.

[21] Κύρκος, Ε. (2015). Προεπεξεργασία Δεδομένων [Κεφάλαιο]. Στο Κύρκος, Ε. 2015. Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων [Προπτυχιακό εγχειρίδιο]. Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις.

[22] Κουρνέτας, Γ. (2017), Μεταπτυχιακή Διπλωματική Εργασία: «Παραγωγή βιοοικονομικών μοντέλων για την πρόβλεψη της ανάπτυξης του ψαριού στην ιχθυοκαλλιέργεια, με την χρήση μεθόδων της εξόρυξης δεδομένων», Τμήμα Οικονομικών Επιστημών, Πανεπιστήμιο Πατρών.