Πτυχιακή Εργασία

Η δύναμη της μηχανικής μάθησης και εφαρμογή της σε προβλήματα classification στην start-up VivaDrive

Στέφανος Κυπριτίδης

Επιβλέπων:
Καθηγητής Πάνος Λουρίδας

Βαρσοβία 2022

Περίληψη

Η ολοένα και μεγαλύτερη χρήση της τεχνολογίας και συγκεκριμένα της μηχανικής μάθησης για την αντιμετώπιση ποικίλων προβλημάτων αφορά όλους τους κλάδους. Όχι μόνο διευκολύνει τις ζωές όλων αλλά μπορεί να συμβάλλει στην λήψη σημαντικών αποφάσεων. Η παρούσα εργασία ξεκινά με μια εισαγωγή στη μηχανική μάθηση, στα διάφορα μοντέλα και όρους της. Επίσης, περιλαμβάνει μια εκτεταμένη ανάλυση δεδομένων και μηχανικής μάθησης. Συγκεκριμένα, αναλύει και αντιμετωπίζει τρία ξεχωριστά προβλήματα classification μιας πραγματικής start-up εταιρίας σχετικά με την αντικατάσταση συμβατικών οχημάτων με τα κατάλληλα ηλεκτρικά. Για κάθε πρόβλημα δοκιμάζονται πληθώρα διαφορετικών αλγορίθμων και τεχνικών μηχανικής μάθησης προκειμένου να βρεθούν οι συνδυασμοί εκείνων που αντιμετωπίζουν το κάθε πρόβλημα αποτελεσματικότερα. Αξίζει να σημειωθεί πως από την ανάλυση φαίνονται λίγες μόνο από τις δυνατότητες της μηχανικής μάθησης. Ωστόσο, υπαινίσσονται οι αμέτρητες δυνατότητες εφαρμογής της στο παρόν και στο μέλλον.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Στη σημερινή εποχή η τεχνολογία αποτελεί αναπόσπαστο κομμάτι της καθημερινής μας ζωής. Το διαδίκτυο, οι υπολογιστές, τα smart-phones και γενικά τα smart-devices συνιστούν αντικείμενα καθημερινής χρήσης για τους περισσότερους από εμάς. Παράλληλα με την χρήση αυτών όμως παράγεται ένας τεράστιος όγκος δεδομένων, τα οποία πρέπει να αποθηκευτούν ενώ μπορούν να οδηγήσουν στην απόκτηση πολύτιμης γνώσης. Συγκεκριμένα, η μηχανική μάθηση είναι ένας από τους τρόπους απόκτησης ευφυΐας μέσα από τα δεδομένα. Σύμφωνα με τον [1], η μηχανική μάθηση δεν αφορά μόνο την αποθήκευση μεγάλων ποσοτήτων δεδομένων, αλλά ταυτόχρονα αποτελεί μέρος της τεχνητής νοημοσύνης. Αναλυτικά, μέσω της τεχνητής νοημοσύνης επιτυγχάνεται η βελτίωση των προγραμμάτων υπολογιστή προκειμένου να εκτελούν εργασίες που απαιτούν την συμμετοχή ανθρώπινου παράγοντα, όπως η λήψη αποφάσεων.

Η μηχανική μάθηση είναι ένα τεράστιο διεπιστημονικό πεδίο που βασίζεται πάνω σε έννοιες από τον επιστήμη των υπολογιστών, τη στατιστική, γνωσιακή επιστήμη, μηχανική, θεωρία βελτιστοποίησης και πολλούς άλλους κλάδους από τα μαθηματικά και τις επιστήμες [2]. Επίσης, η μηχανική μάθηση μπορεί να εφαρμοστεί σε διάφορους τομείς της ζωής μας όπως για παράδειγμα την υγειονομική περίθαλψη, την παραγωγή, την εκπαίδευση, την χρηματοοικονομική μοντελοποίηση, την αστυνόμευση και το μάρκετινγκ [3]. Ένα χαρακτηριστικό παράδειγμα εφαρμογής μηχανικής μάθησης συνιστά η διαφοροποίηση των έγκυρων email από τα ανεπιθύμητα email [1].

Σκοπός της παρούσας εργασίας είναι να αναλύσει την επίδραση που ασκεί η τεχνολογία και συγκεκριμένα η μηχανική μάθηση για την αντιμετώπιση προβλημάτων κατηγοριοποίησης (classification) και την λήψη αποφάσεων σχετικά με αυτά. Για να πραγματοποιηθεί αυτό υλοποιήθηκε μια ανάλυση περίπτωσης διάφορων πραγματικών προβλημάτων classification στα πλαίσια της τεχνολογικής εταιρίας VivaDrive. Η συγκεκριμένη start-up αποτελεί μια τεχνολογική εταιρία που ασχολείται με την διαχείριση στόλων οχημάτων και την σταδιακή αντικατάσταση συμβατικών οχημάτων με ηλεκτρικά. Συγκεκριμένα, η ανάλυση περίπτωσης συνιστά μια ανάλυση δεδομένων και μηχανικής μάθησης αφού εξερευνά τα δεδομένα και στη συνέχεια υλοποιεί μοντέλα μηχανικής μάθησης είδους classification για την πρόβλεψη οχημάτων που είναι κατάλ-

ληλα για αντικατάσταση με ηλεκτρικά οχήματα.

Η ανάλυση περίπτωσης ασχολείται με την επίλυση τριών ξεχωριστών και διαφορετικών προβλημάτων classification. Το πρώτο πρόβλημα αφορά την δημιουργία μοντέλων μηχανικής μάθησης με στόχο την πρόβλεψη συμβατικών οχημάτων ως κατάλληλα ή μη κατάλληλα για αντικατάσταση από ηλεκτρικά. Επιπροσθέτως, το δεύτερο πρόβλημα σχετίζεται με την δημιουργία μοντέλων που είναι ικανά να προβλέψουν πόσο κατάλληλα μπορεί ένα ηλεκτρικό όχημα να αντικαταστήσει το συγκεκριμένο συμβατικό. Τέλος, το τρίτο πρόβλημα αφορά την κατασκευή μοντέλων με στόχο την πρόβλεψη όλων των μαρκών ηλεκτρικών οχημάτων που μπορούν να αντικαταστήσουν το εξετασθέν συμβατικό.

Κεφάλαιο 2

Βιβλιογραφική Ανασκόπηση

Η παρούσα ενότητα περιλαμβάνει μια βιβλιογραφική ανασκόπηση σχετικά με την μηχανική μάθηση. Αναλυτικά, θα παρουσιαστούν οι υπάρχουσες γνώσεις, θα περιγραφούν τα διάφορα είδη μηχανικής μάθησης και θα αναλυθούν οι όροι και τα μοντέλα που θα χρησιμοποιήσουμε στην εργασία. Έτσι θα γίνουν κατανοητές οι ακριβείς σημασίες των όρων και η χρήση των μοντέλων, με ακόλουθο την ομαλή μετάβαση από ενότητα σε ενότητα και τελικά την κατανόηση της σημαντικότητας και της δύναμης της μηχανικής μάθησης. Συγκεκριμένα, στη πρώτη υποενότητα θα εξεταστεί ο διαχωρισμός της μηχανικής μάθησης σε επιβλεπόμενη (supervised), μερικώς επιβλεπόμενη (semi-supervised) και μη επιβλεπόμενη (unsupervised) μάθηση και θα αναλυθεί το κάθε είδος. Στην δεύτερη υποενότητα θα ακολουθήσουν τα διαφορετικά είδη classification, εξηγώντας το καθένα από αυτά. Στη συνέχεια, θα εξεταστούν οι διαφορετικοί αλγόριθμοι που χρησιμοποιούνται για classification, επισημαίνοντας τα πλεονεκτήματα και τα μειονεκτήματα του καθενός. Τέλος, θα αναλυθούν διάφοροι όροι σχετικά με την μηχανική μάθηση που θα χρησιμοποιηθούν στο κομμάτι της ανάλυσης περίπτωσης της Vivadrive.

2.1 Είδη Μηχανικής Μάθησης

2.1.1 Επιβλεπόμενη

Σύμφωνα με τους [4], τα μοντέλα επιβλεπόμενης μηχανικής μάθησης προσπαθούν να ανακαλύψουν τη σχέση μεταξύ των χαρακτηριστικών εισόδου (ανεξάρτητες μεταβλητές) και ενός συγκεκριμένου χαρακτηριστικό στόχου (εξαρτημένη μεταβλητή). Για την ακρίβεια, τα μοντέλα αυτά μαθαίνουν από ένα σύνολο προεπισημασμένων δεδομένων, δηλαδή έχοντας για κάθε δεδομένο την ετικέτα (στόχο) που του αντιστοιχεί. Έτσι, μέσα από τα δεδομένα και τις αντίστοιχες ετικέτες τους, το μοντέλο κατασκευάζει τελικά το δικό του σύστημα πιθανολογικής χαρτογράφησης (από δεδομένο σε ετικέτα) για να χρησιμοποιηθεί για νέες εισόδους δεδομένων [5].

Κάποιες εφαρμογές της επιβλεπόμενης μάθησης είναι τα διαγνωστικά, η πρόβλεψη προσωπικότητας [6], η πρόβλεψη επίδοσης μαθητών [7] και η αναγνώριση εικόνων [8]. Μάλιστα, σύμφωνα με τους [2], η επιβλεπόμενη μηχανική μάθηση διαχωρίζεται περαιτέρω σε δύο κατηγορίες, την κατηγοριοποίηση (classification) και την παλινδρόμηση (regression). Πιο συγκεκριμένα, στην παλινδρόμηση το χαρακτηριστικό στόχου παίρνει συνεχείς τιμές ενώ στην κατηγοριοποίηση το χαρακτηριστικό στόχου παίρνει ετικέτες που αντιστοιχούν σε διαφορετικές κατηγορίες. Τέλος, τα δέντρα αποφάσεων, Naïve Bayes, support vector machine και random forest συνιστούν παραδείγματα αλγορίθμων κατηγοριοποίησης, ενώ η γραμμική παλινδρόμηση αποτελεί παράδειγμα παλινδρόμησης [6].

2.1.2 Μη-Επιβλεπόμενη

Στο συγκεκριμένο είδος μηχανικής μάθησης, δεν υπάρχουν δεδομένα με ετικέτες και σαφείς οδηγίες για την προεκπαίδευση ενός συγκεκριμένου μοντέλου. Έτσι, σε αυτά τα μοντέλα η εκπαίδευση απουσιάζει. Συγκεκριμένα, η ανάλυση πραγματοποιείται με βάση τα υπάρχοντα δεδομένα και ταυτόχρονα εστιάζει στα κοινά χαρακτηριστικά και δομές σε μια ομάδα [9]. Παρόμοια, σύμφωνα με τους [6], στα μοντέλα μη επιβλεπόμενης μάθησης δίνεται τεράστιος όγκος δεδομένων μαζί με συγκεκριμένες οδηγίες για να βρεθεί κάποιο μοτίβο μεταξύ των δεδομένων.

Όσον αφορά την έξοδο των μοντέλων αυτών, τα δεδομένα μπορούν να οργανωθούν σε διαφορετικούς τύπους, όπως συσταδοποίηση (clustering), ανίχνευση ανωμαλιών, συσχέτιση και αυτόματη κωδικοποίηση [10]. Επιπρόσθετα, διάφορες δημοφιλείς εφαρμογές μη επιβλεπόμενης μηχανικής μάθησης συνιστούν η τμηματοποίηση πελατών, το στοχευμένο μάρκετινγκ και τα recommendations όπως για παράδειγμα στην αρχική σελίδα του youtube [6]. Τέλος, οι κύριοι αλγόριθμοι μη επιβλεπόμενης μάθησης είναι η K-means συσταδοποίηση, K-Nearest neighbors (K-NN), principal component analysis (PCA) και οι κανόνες συσχέτισης [11].

2.1.3 Μερικώς-Επιβλεπόμενη

Σε αυτή την προσέγγιση το σύστημα εκπαιδεύεται όχι μόνο με δεδομένα με ετικέτα αλλά και χωρίς ετικέτα στη φάση του ελέγχου (testing) ενώ ταυτόχρονα γίνεται χρήση και των δύο παραπάνω ειδών μηχανικής μάθησης. Ο κύριος στόχος αυτής της προσέγγισης αποτελεί η επίτευξη καλύτερης ακρίβειας (accuracy αλλά και precision) σε σύγκριση με τα παραδοσιακά είδη επιβλεπόμενης και μη επιβλεπόμενης μάθησης [10]. Αντίστοιχα, σύμφωνα με τους [3], η μερικώς επιβλεπόμενη μάθηση χρησιμοποιεί δεδομένα χωρίς ετικέτα για να αυξήσει τα δεδομένα με ετικέτα στο πλαίσιο ενός μοντέλου επιβλεπόμενης μάθησης και μάλιστα μπορεί να χρησιμοποιεί αρχιτεκτονικές μη επιβλεπόμενης μάθησης σε συνδυασμό με διαδικασίες βελτιστοποίησης που κάνουν χρήση ετικετών. Επιπρόσθετα, η μερικώς επιβλεπόμενη συσταδοποίηση και η μερικώς επιβλεπόμενη κατηγοριοποίηση αποτελούν τους δύο διαφορετικούς τύπους εξόδου που υπάρχουν [10]. Τέλος, μερικά παραδείγματα αλγορίθμων αυτής της προσέγγισης συνιστούν το self-training, active learning και το expectation maximisation [12].

2.2 Είδη κατηγοριοποίησης

Σύμφωνα με τον [13], η κατηγοριοποίηση μπορεί να διακριθεί σε binary, multi-class και multi-label classification.

2.2.1 Binary classification

Σύμφωνα με τους [14], το binary classification παίζει σημαντικό ρόλο στη διαδικασία της μηχανικής μάθησης. Αναλυτικά, το binary classification χρησιμοποιείται για δυαδικά προβλήματα δηλαδή καταστάσεις όπου η πρόβλεψη συνιστά μια απόφαση της μορφής ναι ή όχι [15]. Έτσι, το συγκεκριμένο είδος αναφέρεται για κατηγοριοποιήσεις όπου υπάρχουν μόνο δύο ετικέτες-κατηγορίες [13]. Επιπρόσθετα, σύμφωνα με τους [16], όταν υπάρχουν μόνο δύο κατηγορίες και ταυτόχρονα υπάρχει ένα λογικό επίπεδο ισορροπίας μεταξύ των δεδομένων των δύο κατηγοριών συνιστάται απόλυτα η χρήση binary classification.

Όπως αναφέρθηκε ήδη, το binary classification περιλαμβάνει μια κατηγορία κανονικής κατάστασης και μια άλλη κατηγορία ασυνήθιστης κατάστασης. Μερικά παραδείγματα binary classification συνιστούν η ανίχνευση ανεπιθύμητης αλληλογραφίας, η πρόβλεψη μετατροπής (αγορά ή όχι), η ανίχνευσης καρκίνου [13] και η ανίχνευση ψεύτικου προφίλ [14]. Τέλος, δημοφιλείς αλγόριθμοι που χρησιμοποιούνται για προβλήματα binary classification αποτελούν τα δέντρα αποφάσεων, το logistic regression, Naïve Baynes και το support vector machine [13].

2.2.2 Multi-class classification

Σύμφωνα με τους [17], καθώς δίνεται μεγαλύτερη έμφαση για τη συγκέντρωση δεδομένων από τα ιδρύματα, τα προβλήματα multi-class θα γίνονται ολοένα και πιο σημαντικά. Συγκεκριμένα, το multi-class classification αναφέρεται σε κατηγοριοποιήσεις όπου υπάρχουν περισσότερες από δύο ετικέτες-κατηγορίες. Μάλιστα, απουσιάζει η έννοια του κανονικού και του ασυνηθίστου αφού τα δεδομένα αντιστοιχίζονται σε μια από τις διάφορες γνωστές κατηγορίες [13].

Σύμφωνα με τον [13], ο αριθμός των ετικετών-κατηγοριών διαφέρει ανάλογα με το είδος του προβλήματος. Για παράδειγμα, ένα μοντέλο αναγνώρισης προσώπου μπορεί να προβλέψει ότι μια φωτογραφία ανήκει σε ένα άτομο μεταξύ των χιλιάδων ή δεκάδων χιλιάδων ατόμων που υπάρχουν στο σύστημα. Αντίθετα, στο παράδειγμα κατηγοριοποίησης καρκίνου του εγκεφάλου υπάρχουν πέντε διαφορετικές ετικέτες-κατηγορίες εκ των οποίων η μια αντιστοιχεί σε φυσιολογικό δείγμα [18]. Τέλος, τα δέντρα αποφάσεων, το random forest, το Gradient Boosting και ο Naïve Baynes είναι από τους πιο διαδεδομένους αλγόριθμους για τα προβλήματα multi-class classification [13].

2.2.3 Multi-label classification

Το multi-label classification διαφέρει και από το binary αλλά και από το multi-class classification αφού κάθε περιστατικό-παράδειγμα μπορεί να συσχετίζεται με πολλές ετικέτες-κατηγορίες [19]. Μάλιστα, το multi-label classification απαιτείται όλο και περισσότερο στις εφαρμογές όπως την κατηγοριοποίηση πρωτεϊνικής λειτουργίας, κατηγοριοποίηση μουσικής [20], κατηγοριοποίηση κειμένου, κατηγοριοποίηση βίντεο και στη βιοπληροφορική [19]. Τέλος, ένα απλό παράδειγμα multi-label classification είναι η κατηγοριοποίηση ενός τραγουδιού το οποίο μπορεί να ανήκει σε διάφορα μουσικά είδη όπως rock and ballad [20].

Τα προβλήματα multi-label αντιμετωπίζονται είτε με την μέθοδο του μετασχηματισμού προβλήματος είτε με την μέθοδο της προσαρμογής αλγορίθμου [21]. Πιο συγκεκριμένα, οι προσαρμογές των παραδοσιακών αλγορίθμων classification που χρησιμοποιούνται για την επίλυση προβλημάτων multi-label, ονομάζονται multi-label εκδοχές. Μερικά παραδείγματα αποτελούν τα multi-label δέντρα αποφάσεων, τα multi-label Random Forests και το multi-label Gradient Boosting [13]. Αντίθετα, με τον μετασχηματισμό προβλήματος ένα multi-label πρόβλημα μετατρέπεται σε ένα ή περισσότερα binary ή multi-class προβλήματα [19].

Ειδικότερα, παρουσιάζονται όλες διαφορετικές προσεγγίσεις για την επίλυση προβλημάτων τύπου multi-label:

- Binary Relevance είναι μια προσέγγιση μετασχηματισμού προβλήματος η οποία αντιμετωπίζει κάθε ετικέτα ως ένα ξεχωριστό πρόβλημα τύπου binary classification [22].
- Classifier Chains είναι μια άλλη προσέγγιση μετασχηματισμού προβλήματος όπου ο πρώτος classifier εκπαιδεύεται μόνο στα δεδομένα εισόδου και στη συνέχεια κάθε επόμενος classifier εκπαιδεύεται στα δεδομένα εισόδου αλλά και σε όλους τους προηγούμενους classifiers της αλυσίδας [22].
- Το Label Powerset είναι επίσης μια μέθοδος μετασχηματισμού προβλήματος όπου το πρόβλημα μετατρέπεται από multi-label σε multi-class με έναν classifier να εκπαιδεύεται για όλους τους μοναδικούς συνδυασμούς ετικετών που βρίσκονται στα δεδομένα εκπαίδευσης [22].
- Οι προσαρμοσμένοι αλγόριθμοι προσαρμόζουν τον αλγόριθμο για να εκτελεί κατευθείαν κατηγοριοποίηση multi-label, αντί να μετατρέπει το πρόβλημα σε διαφορετικά υποσύνολα προβλημάτων. Ένα παράδειγμα τέτοιου αλγορίθμου αποτελεί ο MLkNN που είναι η multi-label έκδοση του k-NearestNeighbors [22].
- Οι ensemble προσεγγίσεις μπορούν επίσης να επιλύσουν προβλήματα multi-label παράγοντας συνήθως καλύτερα αποτελέσματα [22].
- Τα νευρωνικά δίκτυα είναι ένα άλλο παράδειγμα αλγορίθμου που υποστηρίζει τα προβλήματα multi-label classification και μάλιστα μπορούν να ρυθμιστούν για την αποτελεσματική επίλυση τους [23].

2.3 Αλγόριθμοι κατηγοριοποίησης

Τα προβλήματα κατηγοριοποίησης μπορούν να αντιμετωπιστούν από διάφορες παραλλαγές αλγορίθμων μηχανικής μάθησης με ακριβή και αποδοτικά αποτελέσματα [24]. Μάλιστα, η παρουσία της βιβλιοθήκης ανοιχτού κώδικα Scikit-learn διευκολύνει σε μεγάλο βαθμό τους ερευνητές να λύσουν προβλήματα κατηγοριοποίησης χρησιμοποιώντας αλγορίθμους μηχανικής μάθησης και τις παραλλαγές τους [25]. Σε αυτό το σημείο θα εξεταστούν διάφοροι αλγόριθμοι μηχανικής μάθησης και παραλλαγές μαζί με τα πλεονεκτήματα και τα μειονεκτήματα τους.

2.3.1 Decision Trees

Τα δέντρα αποφάσεων ή αλλιώς decision trees είναι ένας από τους αλγορίθμους που χρησιμοποιείται σε τεράστιο βαθμό στη στατιστική και στη μηχανική μάθηση [1]. Βέβαια, είναι ένας από τους πιο απλούς αλγορίθμους μηχανικής μάθησης αφού δημιουργεί κανόνες συσχέτισης για να βρει και να προβλέψει τις ετικέτες-στόχους [24]. Συγκεκριμένα, αποτελεί έναν ιεραρχικό σχεδιασμό που εφαρμόζει την προσέγγιση διαίρει και βασίλευε, ενώ ταυτόχρονα συνιστά μια μη παραμετρική τεχνική [1]. Επιπρόσθετα, στα δέντρα αυτά τα δεδομένα μπορούν να μοντελοποιηθούν σε ιεραρχικές δομές χρησιμοποιώντας μια σειρά συγκρίσεων της μορφής if-else [26].

Τα δέντρα αποφάσεων χρησιμοποιούνται όχι μόνο σε μεγάλα αλλά και σε μικρά dataset, ενώ μπορούν να διαχειριστούν και αριθμητικά αλλά και κατηγορικά δεδομένα [27]. Επιπλέον, τα δέντρα αποφάσεων συνήθως κατασκευάζονται σε δύο φάσεις, την ανάπτυξη του δέντρου και το κλάδεμα του δέντρου [2]. Επεξηγηματικά, η ανάπτυξη του δέντρου αφορά τον επαναλαμβανόμενο διαχωρισμό των δεδομένων προπόνησης με βάση τα βέλτιστα κριτήρια μέχρις ότου το μεγαλύτερο μέρος των εγγραφών να ανήκει σε μια ετικέτα-κατηγορία [28]. Αντίθετα, το κλάδεμα του δέντρου αφορά την μείωση του μεγέθους του δέντρου προκειμένου να γίνει ευκολότερο στην κατανόηση [29]. Επιπρόσθετα, σύμφωνα με τους [30] το κλάδεμα του δέντρου μπορεί να χρησιμοποιηθεί και για αντιμετώπιση του φαινομένου overfitting όπου ο αλγόριθμος μαθαίνει να κατηγοριοποιεί τέλεια όλα τα δεδομένα προπόνησης.

Σύμφωνα με τους [4], τα δέντρα αποφάσεων παρουσιάζουν αρκετά πλεονεκτήματα ως αλγόριθμος κατηγοριοποίησης. Αρχικά, τα δέντρα αποφάσεων έχουν αυτονόητη λογική και είναι εύκολα στο να τα ακολουθήσει κανείς όταν συμπιέζονται. Μάλιστα, όταν έχουν ένα λογικό μέγεθός γίνεται εύκολη και η κατανόηση τους από μη επαγγελματίες χρήστες. Αυτό επιτυγχάνεται επίσης από το γεγονός ότι τα δέντρα αποφάσεων μπορούν να μετατραπούν εύκολα σε ένα σύνολο από κανόνες [4]. Επιπρόσθετα, τα δέντρα αποφάσεων μπορούν να διαχειριστούν εύκολα όχι μόνο ακραίες τιμές-outliers [31] αλλά και δεδομένα που απουσιάζουν-missing values [4]. Τέλος, αφού τα δέντρα συνιστούν μη παραμετρική τεχνική, δεν απαιτείται κάποιος προσδιορισμός παραμέτρων λειτουργικής μορφής [31].

Παρόλα αυτά, σύμφωνα με τον [1], τα δέντρα αποφάσεων εμφανίζουν και περιορισμούς. Αρχικά, μερικές φορές τα δέντρα αποφάσεων μπορεί να είναι υπολογιστικά ακριβά. Ακόμα, τα δέντρα αποφάσεων εύκολα καταλήγουν στο φαινόμενο overfitting όταν δεν χρησιμοποιούν-

ται τρόποι αποφυγής του. Εξάλλου, σύμφωνα με τον [32], κάποιες φορές δημιουργούνται πολύ σύνθετα δέντρα που δεν γενικεύουν καλά. Παράλληλα, τα δέντρα αποφάσεων μπορεί να είναι ασταθή αφού μικρές παραλλαγές στα δεδομένα μπορεί να οδηγήσουν στη δημιουργία ενός εντελώς διαφορετικού δέντρου. Τέλος, πρακτικά τα δέντρα αποφάσεων χρησιμοποιούνται ευρέως για κατηγοριοποίηση, ενώ είναι λιγότερο κατάλληλα για προβλήματα παλινδρόμησης [1].

2.3.2 Support Vector Machine

Το Support Vector Machine (SVM) συνιστά μια από τις πιο σημαντικές και βολικές τεχνικές για την επίλυση προβλημάτων που σχετίζονται με κατηγοριοποίηση δεδομένων [2]. Παράλληλα, το SVM είναι και μια καθιερωμένη και συχνά χρησιμοποιούμενη τεχνική κατηγοριοποίησης [10]. Σύμφωνα με τους [33], ανεξάρτητα που νέες τεχνικές classification προτάσσονται, το SVM παραμένει ένας από τους πιο δημοφιλείς και ευρέως χρησιμοποιούμενους αλγορίθμους κατηγοριοποίησης. Αν και σε πρώτο στάδιο είχε αναπτυχθεί για προβλήματα binary classification, επεκτάσεις multi-class classification έχουν υλοποιηθεί.

Το SVM μπορεί να χρησιμοποιηθεί για οποιονδήποτε αριθμό διανυσματικών διαστάσεων των δεδομένων [10]. Αναλυτικά, στην κατηγοριοποίηση το SVM καθορίζει ένα βέλτιστο διαχωριστικό υπερεπίπεδο χρησιμοποιώντας την έννοια του περιθωρίου (margin). Το περιθώριο είναι ουσιαστικά η απόσταση μεταξύ του υπερεπίπεδου και των πλησιέστερων σημείων σε αυτό εκατέρωθεν. Στόχος αποτελεί η μεγιστοποίηση του περιθωρίου για καλύτερη γενίκευση των δεδομένων-σημείων [1]. Πιο συγκεκριμένα, στις δύο διαστάσεις το υπερεπίπεδο είναι μια γραμμή [10].

Η χρήση SVM σε classification προβλήματα προσφέρει διάφορα πλεονεκτήματα. Αρχικά, σύμφωνα με τον [32], το SVM είναι αποτελεσματικό σε χώρους υψηλών διαστάσεων. Επιπλέον, είναι και αποδοτικό στην χρήση μνήμης αφού χρησιμοποιεί μόνο ένα υποσύνολο σημείων εκπαίδευσης για την εύρεση του υπερεπίπεδου. Επιπρόσθετα, ένα ακόμα πλεονέκτημα συνιστά η ικανότητά του SVM να αντιμετωπίσει ποικίλα προβλήματα κατηγοριοποίησης όπως προβλήματα υψηλών διαστάσεων και μη γραμμικά διαχωρίσιμα προβλήματα [2]. Εξάλλου, η εκπαίδευση είναι σχετικά εύκολη, το μοντέλο μπορεί να χρησιμοποιηθεί τόσο με συνεχή όσο και με κατηγορικά δεδομένα και μπορεί να αντιμετωπίσει δεδομένα με σφάλματα. Ακόμα, το trade-off μεταξύ πολυπλοκότητας του μοντέλου και σφάλματος ελέγχεται εύκολα. Τέλος, η ακρίβεια των προβλέψεων είναι πολύ υψηλή, γίνονται καλές γενικεύσεις και παράγεται μια μοναδική βέλτιστη λύση [1].

Παράλληλα, η χρήση SVM συνοδεύεται και από μερικά μειονεκτήματα. Αρχικά, ο αλγόριθμος δεν παρέχει άμεσα εκτιμήσεις πιθανοτήτων, με αποτέλεσμα αυτές να υπολογίζονται με χρήση του ακριβού five-fold cross-validation [32], το οποίο θα αναλυθεί στην συνέχεια. Ακόμη, η επίτευξη άριστων αποτελεσμάτων προϋποθέτει την σωστή ρύθμιση κάποιων βασικών παραμέτρων [2]. Εκτός από αυτό, το SVM είναι πολύ δύσκολο να ερμηνευτεί εκτός κι αν τα χαρακτηριστικά των δεδομένων είναι ερμηνεύσιμα. Τέλος, κάποιες φορές η χρήση SVM μπορεί να είναι υπολογιστικά ακριβή [1].

2.3.3 K-Nearest-Neighbours

Ο αλγόριθμος K-Nearest-Neighbours (KNN) βασίζεται στην αρχή ότι μέσα σε ένα σύνολο δεδομένων οι παρατηρήσεις δεδομένων θα βρίσκονται σε κοντινή απόσταση σχετικά με άλλες παρατηρήσεις που έχουν παρόμοια χαρακτηριστικά [30]. Ειδικότερα, οι παρατηρήσεις δεδομένων παρουσιάζονται σε έναν ν-διάστατο χώρο, όπου ν είναι ο αριθμός των χαρακτηριστικών των δεδομένων. Έτσι, ένα νέο σημείο κατηγοριοποιείται ανάλογα με την ομοιότητά του με τα υπόλοιπα σημεία δεδομένων που είναι ήδη αποθηκευμένα γνωστά ως δεδομένα εκπαίδευσης [1]. Συγκεκριμένα, το Κ προσδιορίζει τον αριθμό των πλησιέστερων γειτόνων που πρέπει να εξεταστούν προκειμένου να κατηγοριοποιηθεί η εξετάζουσα παρατήρηση [2]. Τέλος, για Κ≥3 γίνεται ψηφοφορία για την κατηγοριοποίηση της νέας παρατήρησης με βάση την πιο κοινή κατηγορία μεταξύ των Κ πλησιέστερων γειτόνων [1].

Η επιλογή του Κ επηρεάζει την απόδοση του KNN αλγόριθμου. Επεξηγηματικά, όταν υπάρχει θόρυβος στα δεδομένα, ένα μικρό Κ μπορεί να έχει ως επακόλουθο οι θορυβώδεις περιπτώσεις να κερδίσουν την πλειοψηφία [30]. Αντίθετα, εάν το Κ είναι πολύ μεγάλο, ο αλγόριθμος μπορεί να κατηγοριοποιήσει εσφαλμένα τη νέα παρατήρηση επειδή οι πλειοψηφία των πλησιέστερων γειτόνων μπορεί να βρίσκονται πολύ μακριά από τη συγκεκριμένη παρατήρηση [1].

Ο αλγόριθμος KNN εμφανίζει τα δικά του πλεονεκτήματα. Αρχικά, ο αλγόριθμος είναι αποτελεσματικός για μεγάλα σύνολα δεδομένα και ταυτόχρονα κατάλληλος για θορυβώδη δεδομένα εκπαίδευσης. Επιπροσθέτως, ο KNN χαρακτηρίζεται από απλότητα και διαφάνεια και έτσι μπορεί να υλοποιηθεί και να κατανοηθεί με ευκολία [2]. Τέλος, η διαδικασία εκπαίδευσης είναι ταχύτατη και έχει μηδενικό κόστος [1].

Από την άλλη πλευρά, ο KNN χαρακτηρίζεται από αρκετά μειονεκτήματα. Καταρχάς, το υπολογιστικό κόστος είναι υψηλό καθώς απαιτείται ο υπολογισμός της απόστασης κάθε εξετάζουσας παρατήρησης από όλα τα δείγματα εκπαίδευσης. Ακόμη, απαιτείται ο προσδιορισμός του αριθμού Κ των πλησιέστερων γειτόνων [32] που μπορεί να κάνει μεγάλη διαφορά όπως αναφέρθηκε προηγουμένως. Εκτός αυτού, χαρακτηριστικά των δεδομένων που δεν έχουν μεγάλη σημασία μπορεί να προκαλέσουν προβλήματα στα αποτελέσματα του KNN [2]. Τέλος, σύμφωνα με τον [1], ο αλγόριθμος έχει ανάγκη από τεράστια μνήμη για να αποθηκεύσει όλα τα δεδομένα εκπαίδευσης.

2.3.4 Naïve Bayes

Ο classifier Naïve Bayes (NB) αποτελεί έναν από τους πιο συχνά χρησιμοποιούμενους αλγορίθμους εποπτευόμενης μηχανικής μάθησης [10]. Επίσης, ο NB συχνά χρησιμοποιείται για κατηγοριοποίηση κειμένου όπως κατηγοριοποίηση αρχείων και ανίχνευση ανεπιθύμητης αλληλογραφίας. Παράλληλα, ο NB υπολογίζει ένα σύνολο πιθανοτήτων από συνδυασμούς τιμών του συνόλου δεδομένων [33].

Αναλυτικότερα, η λειτουργία του NB στηρίζεται πάνω στον κανόνα του Bayes. Ο συγκε-

κριμένος κανόνας προσπαθεί να υπολογίσει την πιθανότητα εμφάνισης ενός γεγονότος με βάση σχετικές προηγούμενες γνώσεις και συνθήκες [10]. Αποτελεσματικά, ο NB κάνει μια υπόθεση ανεξαρτησίας μεταξύ των διαφορετικών χαρακτηριστικών των δεδομένων. Μάλιστα, με την βοήθεια στατιστικών μεθόδων υπολογίζονται οι πιθανότητες μιας εισόδου-παρατήρησης να σχετίζεται με τις διάφορες κατηγορίες. Τέλος, η έξοδος του αλγορίθμου είναι η κατηγορία με την υψηλότερη πιθανότητα [12].

Διάφορα πλεονεκτήματα συνοδεύονται από την χρήση του NB. Αρχικά, το κύριο πλεονέκτημα του NB είναι ο σύντομος υπολογιστικός χρόνος για την εκπαίδευση [30]. Επίσης, η διαδικασία λήψης αποφάσεων στον NB είναι ταχύτατη σχετικά με άλλους classifiers και λειτουργεί με καλά αποτελέσματα ακόμα και με μικρό αριθμό δεδομένων εκπαίδευσης [33]. Άλλωστε, ο αλγόριθμος του NB είναι εξαιρετικά επεκτάσιμος, γρήγορος και εύκολος στην εφαρμογή σε ένα σύστημα [10]. Ακόμα, σύμφωνα με τους [34], ο NB όχι μόνο απαιτεί λίγο αποθηκευτικό χώρο τόσο κατά την εκπαίδευση όσο και κατά την κατηγοριοποίηση αλλά και περιλαμβάνει λίγες παραμέτρους σχετικά με άλλους classifiers όπως SVM. Τέλος, λαμβάνονται υπόψη στοιχεία από πολλά χαρακτηριστικά για να γίνει η τελική πρόβλεψη και ταυτόχρονα ο NB μπορεί να διαχειριστεί missing values και θορυβώδη δεδομένα.

Παρόλα αυτά, η χρήση NB έρχεται και με διάφορα προβλήματα. Καταρχάς, το κύριο μειονέκτημα συνιστά ότι μπορεί να χρησιμοποιηθεί μόνο εάν τα χαρακτηριστικά των δεδομένων είναι εντελώς ανεξάρτητα το ένα από το άλλο. Πρακτικά, αυτό δεν είναι πάντα εφικτό [10]. Ωστόσο, σύμφωνα με την έρευνα των [34], ο NB μπορεί να λειτουργήσει ικανοποιητικά σε ορισμένες περιπτώσεις όπου τα χαρακτηριστικά δεν είναι πλήρως ανεξάρτητα αλλά με λιγότερο ακριβείς εκτιμήσεις και μειωμένη απόδοση. Αυτό οφείλεται στο γεγονός ότι ο αλγόριθμος μπορεί να αυξήσει την επιρροή από τα δύο χαρακτηριστικά που εξαρτώνται μεταξύ τους και να μειώσει την επιρροή από τα άλλα, οδηγώντας σε μεροληψία-bias στην κατηγοριοποίηση. Τέλος, ο NB μπορεί εμφανίσει προβλήματα λόγω υπερευαισθησίας σε περιττά ή άσχετα χαρακτηριστικά.

2.3.5 Logistic regression

Η λογιστική παλινδρόμηση αποτελεί ένας από τους πιο απλούς αλγορίθμους μηχανικής μάθησης. Πιο συγκεκριμένα, χρησιμοποιείται σε διάφορα προβλήματα κατηγοριοποίησης όπως ανάλυση κειμένου, εξόρυξη δεδομένων και ανάκτηση πληροφοριών [35]. Μάλιστα, η λογιστική παλινδρόμηση εξετάζει τη σχέση μεταξύ μιας δυαδικής (εξαρτώμενης) μεταβλητής όπως η παρουσία ή η απουσία ασθένειας, και διάφορων προγνωστικών (επεξηγηματικών ή ανεξάρτητων) μεταβλητών όπως δημογραφικά στοιχεία ασθενών [36]. Παρόλα αυτά, η λογιστική παλινδρόμηση μπορεί να είναι πολυωνυμικής φύσεως, δηλαδή να έχει τρεις ή περισσότερες κατηγορίες-ετικέτες. Τέλος, ο αλγόριθμος είναι γνωστός για την ικανότητά του να προβλέπει την πιθανότητα της μεταβλητής στόχου [35] χρησιμοποιώντας την βασική logistic function [33].

Διάφορα οφέλη και μειονεκτήματα συμπεριλαμβάνονται στην χρήση της λογιστικής παλινδρόμησης. Αρχικά, η λογιστική παλινδρόμηση είναι ιδανική για προβλήματα binary classification αφού έχει σχεδιαστεί για αυτά [33]. Επιπλέον, είναι πολύ χρήσιμη για την κατανόηση της επιρροής πολλών ανεξάρτητων μεταβλητών σε μια μεμονωμένη μεταβλητή στόχου. Από την άλλη

πλευρά, γίνεται η υπόθεση ότι όλες οι προγνωστικές μεταβλητές είναι ανεξάρτητες μεταξύ τους [32]. Έτσι, προτού εκτελεστεί ο αλγόριθμος πρέπει να εντοπιστούν όλες οι σημαντικές ανεξάρτητες μεταβλητές και να αφαιρεθούν οι μη σχετικές και εξαρτημένες [33]. Τέλος, η λογιστική παλινδρόμηση απαιτεί να μην υπάρχουν δεδομένα που λείπουν – missing values [32].

2.3.6 Stochastic Gradient Descent

Ο αλγόριθμος Stochastic Gradient Descent (SGD) συνιστά μια απλή και πολύ αποτελεσματική προσέγγιση για την προσαρμογή γραμμικών μοντέλων [32]. Αναλυτικότερα, βασίζεται στην λειτουργία των συναρτήσεων κυρτής απώλειας (loss function) του SVM και της λογιστικής παλινδρόμησης. Μάλιστα, αποδεικνύεται ότι είναι ένας ισχυρός classifier για multi-class προβλήματα αφού συνδυάζει πολλαπλούς binary classifiers με την μέθοδο one-vs-all [24]. Τέλος, στον αλγόριθμο εξετάζονται επανειλημμένα τα δεδομένα εκπαίδευσης και κάθε φορά χρησιμοποιείται ένα παράδειγμα εκπαίδευσης προκειμένου να τροποποιηθούν οι παράμετροι ανάλογα με την κλίση του σφάλματος σε σχέση με το μεμονωμένο παράδειγμα εκπαίδευσης [37].

Ο SGD σημειώνει τα δικά του πλεονεκτήματα και μειονεκτήματα. Αρχικά, ο SGD είναι αλγόριθμος που κατανοείται σχετικά εύκολα. Ακόμη, μπορεί να χρησιμοποιηθεί αποτελεσματικά σε μεγάλα dataset αφού χρησιμοποιεί μόνο ένα παράδειγμα εκπαίδευσης (batch size) ανά επανάληψη [24]. Επιπλέον, σύμφωνα με τον [32], ο SGD χαρακτηρίζεται από αποδοτικότητα και ευκολία στην υλοποίηση και υποστηρίζει διάφορες λειτουργίας απώλειας (loss function) και ποινές στην κατηγοριοποίηση. Ωστόσο, για την επίτευξη καλών αποτελεσμάτων πρέπει οι υπερπαράμετροι του SGD να έχουν ρυθμιστεί σωστά. Τέλος, ο αλγόριθμος μπορεί να έχει προβλήματα με θορυβώδη δεδομένα καθώς τα παραδείγματα που επιλέγονται επαναληπτικά είναι τυχαία, και ταυτόχρονα εμφανίζει ευαισθησία στην κλιμάκωση χαρακτηριστικών (feature scaling) [24].

2.3.7 Random Forest

Ο αλγόριθμος random forest (RF) είναι τύπου ensemble learning που χρησιμοποιεί πλήθος δέντρων απόφασης κατά την διάρκεια της εκπαίδευσης [38]. Ειδικότερα, ο RF επιλέγει τυχαία υποσύνολα χαρακτηριστικών από το συνολικό σετ εκπαίδευσης για τη δημιουργία πολλαπλών δέντρων απόφασης και έπειτα επιστρέφει την μέση πρόβλεψη όλων των δέντρων [35]. Τέλος, ο RF γενικά προπονεί καλύτερα μοντέλα από τα τυπικά μοντέλα μηχανικής μάθησης [4].

Η χρήση του RF συνοδεύεται με τα δικά του οφέλη και περιορισμούς. Αρχικά, ο RF λύνει το πρόβλημα του overfit κατά τη διάρκεια της προπόνησης που είχε αναφερθεί ως μειονέκτημα των δέντρων απόφασης [35]. Εκτός από αυτό, ο αλγόριθμος RF έχει πιο ακριβή αποτελέσματα από τα δέντρα απόφασης στις περισσότερες περιπτώσεις [32]. Από την άλλη πλευρά, με την χρήση του RF δεν υπάρχει πια κάποιο δέντρο που μπορεί να ερμηνευτεί δυσκολεύοντας πολύ την ερμηνευσιμότητα που αποτελεί πλεονέκτημα των δέντρων απόφασης [4]. Τέλος, ο RF είναι πολύπλοκος αλγόριθμος που εφαρμόζεται δύσκολα και με αργές προβλέψεις σε πραγματικό χρόνο [32].

2.3.8 Extra Trees Classifier

Ο αλγόριθμος extra trees ή αλλιώς extremely randomized forest είναι κι αυτός τύπου ensemble learning. Αναλυτικά, δημιουργείται ένα σύνολο από μη κλαδεμένα δέντρα απόφασης όπου και τα χαρακτηριστικά και ο διαχωρισμός επιλέγονται πολύ τυχαία [39]. Μάλιστα, η λειτουργία του αλγορίθμου μοιάζει σχετικά με του random forest αλλά διαφέρει ως προς τον τρόπο δημιουργίας δέντρων σε ένα δάσος δέντρων απόφασης. Συγκεκριμένα, τυχαία δείγματα των Κ βέλτιστων χαρακτηριστικών χρησιμοποιούνται για την απόφαση και το κριτήριο Gini χρησιμοποιείται για την επιλογή του καλύτερου χαρακτηριστικού για τον διαχωρισμό δεδομένων στο δέντρο. Έτσι, με αυτή την προσέγγιση κατασκευάζονται αποσυσχετισμένα (decorrelated) δέντρα [24].

Τα extra-trees έχουν τα δικά τους θετικά και αρνητικά. Αρχικά, ο αλγόριθμος είναι γνωστός για την υψηλή ακρίβεια και υπολογιστική απόδοση [40]. Επιπρόσθετα, ο αλγόριθμος μπορεί να αντιμετωπίσει ακραίες τιμές, να προσδιορίσει τα σημαντικά χαρακτηριστικά μεταξύ των άσχετων και να χρησιμοποιηθεί σε εφαρμογές εξόρυξης μεγάλης κλίμακας [39]. Αντίθετα, τα extra trees εμφανίζουν μεγαλύτερη πολυπλοκότητα αφού είναι κατά μέσο όρο περίπου δύο επίπεδα ή λιγότερο βαθύτερα συγκριτικά με το Random Forest και το Tree Bagging. Τέλος, ο αλγόριθμος του extra-trees έχει διάφορες παραμέτρους που πρέπει να ρυθμιστούν, οι οποίες παίζουν σημαντικό ρόλο στην παραγωγή καλών αποτελεσμάτων [41].

2.3.9 Νευρωνικά δίκτυα

Τα νευρωνικά δίκτυα συνιστούν ένα πολύ γνωστό εργαλείο και ανήκουν στον τομέα της τεχνητής νοημοσύνης. Επιπροσθέτως, υπάρχουν διάφορες παραλλαγές νευρωνικών δικτύων, ενώ οι εφαρμογές τους καλύπτουν τομείς όπως αναγνώριση προτύπων, εντοπισμός και classification προβλήματα [35]. Μάλιστα, τα νευρωνικά δίκτυα είναι εμπνευσμένα από τους νευρώνες του ανθρώπινου εγκεφάλου όπου οι κόμβοι διασυνδέονται με τέτοιο τρόπο ώστε η έξοδος κάθε κόμβου μετατρέπεται σε είσοδος άλλου κόμβου. Ναι μεν κάθε κόμβος αποκτά περισσότερες από μία εισόδους, αλλά η παραγόμενη έξοδος αποτελείται από μία μόνο τιμή [6].

Διάφορα πλεονεκτήματα χαρακτηρίζουν την χρήση νευρωνικών δικτύων. Αρχικά, τα νευρωνικά δίκτυα τείνουν να παράγουν καλύτερα αποτελέσματα όταν υπάρχουν πολλαπλές διαστάσεις και συνεχή χαρακτηριστικά [30]. Ακόμα, μπορούν να χρησιμοποιηθούν για την επίλυση προβλημάτων γραμμικού και μη γραμμικού προγραμματισμού, ενώ έχουν επιτυχία στην επίλυση διαφορετικών ειδών προβλημάτων όπως κατηγοριοποίηση, συσταδοποίηση και παλινδρόμηση. Τέλος, χαρακτηρίζονται ως ισχυρά και ευέλικτα αφού μαθαίνουν από τα δεδομένα εκπαίδευσης χωρίς να απαιτείται γνώση της διαδικασίας παραγωγής τους [1].

Από την άλλη πλευρά, τα νευρωνικά δίκτυα έχουν και τα μειονεκτήματα τους. Αρχικά, η παρουσία μη σχετικών χαρακτηριστικών μπορεί να οδηγήσει σε αναποτελεσματική εκπαίδευση του νευρωνικού δικτύου. Ακόμη, απαιτείται μεγάλη ποσότητα δεδομένων προκειμένου να επιτευχθεί η μέγιστη ακρίβεια προβλέψεων [30]. Τέλος, η επιλογή της βέλτιστης αρχιτεκτονικής νευρωνικού δικτύου συνήθως δεν μπορεί να είναι γνωστή εκ των προτέρων αφού τα νευρωνικά

δίκτυα περιλαμβάνουν τη διαδικασία δοκιμής και σφάλματος [1].

2.4 Όροι μηχανικής μάθησης

Όπως αναφέρθηκε ήδη στην υποενότητα supervised learning, αρχικά χρησιμοποιείται ένα σύνολο προεπισημασμένων δεδομένων (training data) για να εκπαιδεύσουν το μοντέλο κατάλληλα. Έπειτα αφού ολοκληρωθεί η εκπαίδευση και κατασκευαστεί το σύστημα πιθανολογικής χαρτογράφησης από δεδομένα σε ετικέτες-κατηγορίες, μπορεί να δοκιμαστεί σε νέες εισόδους δεδομένων γνωστά ως testing data [5].

2.4.1 Μετρικές

Για την αξιολόγηση της επίδοσης κάθε αλγορίθμου όσον αφορά τα αποτελέσματα των προβλέψεων, υπάρχουν διάφορες μετρικές [42, p. 214]:

- Η μετρική accuracy ορίζεται ως

$$\frac{TN + TP}{FN + FP + TN + TP}$$

όπου TN ο αριθμός των true negative, TP ο αριθμός των true positive, FN ο αριθμός των false negative και ανάλογα. Αναλυτικά, η μετρική accuracy υπολογίζει πόσο συχνά προβλέπει σωστά ο συγκεκριμένος classifier. Η συγκεκριμένη μετρική είναι πολύ χρήσιμη όταν υπάρχει ισορροπία στις ετικέτες-κατηγορίες που προβλέπονται [43].

- Η μετρική precision ορίζεται ως

$$\frac{TP}{FP + TP}$$

Η μετρική precision εξηγεί πόσες από τις προβλεπόμενες περιπτώσεις ως positive αποδείχθηκαν όντως positive. Το precision είναι χρήσιμο στις περιπτώσεις όπου τα false positive έχει χειρότερες επιπτώσεις από τα false negative [43].

- Η μετρική recall ορίζεται ως

$$\frac{TP}{FN + TP}$$

Το recall εξηγεί πόσες από τις πραγματικές positive περιπτώσεις προβλέφθηκαν σωστά από το μοντέλο. Επίσης, είναι χρήσιμη μέτρηση σε περιπτώσεις όπου το false negative έχει χειρότερες επιπτώσεις από το false positive, όπως συχνά σε ιατρικές περιπτώσεις [43].

- Η μετρική F-measure ορίζεται ως

$$\frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

Η συγκεκριμένη μετρική αποτελεί τον αρμονικό μέσο όρο των precision και recall. Μάλιστα, μεγιστοποιείται όταν το precision είναι ίσο με το recall. Εκτός από αυτό, η μετρική F-measure τιμωρεί περισσότερο τις ακραίες τιμές. Τέλος, είναι χρήσιμη όταν τα false positive και τα false negative κοστίζουν το ίδιο και όταν ο αριθμός των true negative είναι υψηλός [43].

Η μετρική hamming loss χαρακτηρίζεται ως η πιο κοινή μετρική στη βιβλιογραφία που σχετίζεται με multi-label classification [44]. Συγκεκριμένα, μετρά τον αριθμό των ετικετών για τις οποίες η πρόβλεψή ήταν λάθος και έπειτα τον κανονικοποιεί [45]. Έτσι, όσο χαμηλότερη η τιμή της μετρικής hamming loss, τόσο καλύτερα συνιστούν τα αποτελέσματα. Αυτό σημαίνει ότι η απόδοση του μοντέλου είναι τέλεια όταν η τιμή της μετρικής ισούται με το μηδέν [44].

2.4.2 Εύρεση υπερπαραμέτρων

Ένας άλλος σημαντικός όρος που θα χρησιμοποιηθεί στην συνέχεια είναι το grid search. Το grid search συνιστά μια παραδοσιακή μέθοδος βελτιστοποίησης υπερπαραμέτρων [46]. Ουσιαστικά δοκιμάζει όλους τους συνδυασμούς ενός ορισμένου συνόλου υπερπαραμέτρων [47] προκειμένου να ανιχνεύσει τις καλύτερες κάνοντας ταυτόχρονα χρήση cross-validation για να μετρήσει την επίδοση των αποτελεσμάτων [48]. Παρόλα αυτά, ένα μειονέκτημα του grid search είναι ότι τα μοντέλα που δοκιμάζονται αυξάνονται εκθετικά συγκριτικά τον αριθμό των υπερπαραμέτρων [47].

Μια άλλη τεχνική βελτιστοποίησης υπερπαραμέτρων αποτελεί η Bayesian βελτιστοποίηση με Gaussian διαδικασίες. Η συγκεκριμένη τεχνική θεωρεί ότι η συνάρτηση αξιολόγησης του μοντέλου μηχανικής μάθησης ακολουθεί μια πολυμεταβλητή Gaussian κατανομή. Έτσι, πραγματοποιούνται εξυπνότερες επιλογές όσον αφορά την επιλογή της επόμενης υπερπαραμέτρου προς αξιολόγηση συγκριτικά με την τεχνική grid-search [49].

2.4.3 Cross validation

Σύμφωνα με τους [50], ο όρος cross-validation περιγράφει μια μέθοδο αξιολόγησης και σύγκρισης αλγορίθμων μάθησης. Συγκεκριμένα, η μέθοδος αυτή χωρίζει τα δεδομένα σε ένα τμήμα που χρησιμοποιείται για εκπαίδευση του μοντέλου και σε ένα άλλο τμήμα που χρησιμοποιείται για επικύρωση (έλεγχο) του μοντέλου. Η βασική μορφή cross-validation είναι το k-fold cross-validation όπου τα δεδομένα χωρίζονται σε k ίσα (ή περίπου ίσα) τμήματα γνωστά ως folds. Έτσι, ακολουθούν k επαναλήψεις του μοντέλου όπου σε κάθε επανάληψη υπάρχει ένα διαφορετικό fold για επικύρωση, ενώ ταυτόχρονα τα υπόλοιπα k-1 fold χρησιμοποιούνται για εκπαίδευση.

Μια συγκεκριμένη περίπτωση k-fold cross-validation είναι η leave-one-out cross-validation. Ειδικότερα, σε αυτή την περίπτωση το k ισούται με τον συνολικό αριθμό παρατηρήσεων. Απο-

τελεσματικά, σε κάθε επανάληψη χρησιμοποιείται μια μεμονωμένη παρατήρηση ως fold για επικύρωση και όλες οι υπόλοιπες παρατηρήσεις χρησιμοποιούνται για εκπαίδευση. Με την συγκεκριμένη μέθοδο το σφάλμα ελέγχου (testing) είναι περίπου μια αμερόληπτη εκτίμηση του πραγματικού σφάλματος πρόβλεψης. Ωστόσο, το υπολογιστικό κόστος της μεθόδου μπορεί να είναι ιδιαίτερα υψηλό για μεγάλα σύνολα δεδομένων [51].

2.4.4 Τεχνικές oversampling

Οι τεχνικές oversampling επιλύουν το ζήτημα της ανισορροπίας των δεδομένων. Αυτό επιτυγχάνεται με τη δημιουργία πρόσθετων δειγμάτων εκπαίδευσης για τη κατηγορία με την μειοψηφία των δειγμάτων και μάλιστα εστιάζει στην βελτίωση της απόδοσης του classifier [52]. Αναλυτικά, στην τεχνική random-oversampler τα δείγματα αντιγράφονται για τυχαίες φορές και συνδυάζονται με τον πληθυσμό δειγμάτων της πλειοψηφικής κατηγορίας. Από την άλλη πλευρά, η τεχνική smote παράγει νέα τεχνητά δείγματα χρησιμοποιώντας τον χώρο των χαρακτηριστικών αντί να επαναλαμβάνει τα υπάρχοντα δείγματα [53].

2.4.5 PCA

Η Principal Component Analysis (PCA) είναι από τις πιο δημοφιλείς πολυμεταβλητές στατιστικές τεχνικές και χρησιμοποιείται στους περισσότερους επιστημονικούς κλάδους [54]. Αναλυτικά, το PCA συνιστά έναν μαθηματικό αλγόριθμο που μειώνει τις διαστάσεις των δεδομένων, διατηρώντας το μεγαλύτερο μέρος της διακύμανσης στο dataset. Αυτό επιτυγχάνεται με τον εντοπισμό των principal components που είναι κατευθύνσεις κατά μήκος των οποίων η διακύμανση στα δεδομένα είναι μέγιστη [55].

Κεφάλαιο 3

Ανάλυση Περίπτωσης

3.1 Dataset

Η συγκεκριμένη έρευνα έγινε στα πλαίσια της πολωνικής start-up εταιρίας VivaDrive. Τα δεδομένα που χρησιμοποιήθηκαν ανήκουν στη συγκεκριμένη εταιρία και προέρχονται από την χρήση τεχνολογιών IoT σε οχήματα στόλου άλλης εταιρίας-πελάτη. Αφού συγκεντρώνονται τα απαραίτητα δεδομένα για το κατάλληλο χρονικό διάστημα, μέσω του αλγορίθμου της εταιρίας υπολογίζονται διάφορες μετρικές και βελτιστοποιήσεις και τελικά προτείνονται συγκεκριμένα ηλεκτρικά οχήματα ανάλογα με την καταλληλότητα τους. Στόχος της εταιρίας συνιστά η αντικατάσταση όσων περισσότερων συμβατικών οχημάτων με ηλεκτρικά οχήματα.

Τα αρχικά δεδομένα αφορούν δύο dataset. Αρχικά, το πρώτο dataset περιέχει όλα τα προφίλ οχημάτων και αποτελείται από 981 γραμμές και 20 στήλες. Ειδικότερα, ως προφίλ οχήματος θεωρείται ένα μοναδικό όχημα που παρακολουθήθηκε για μια συγκεκριμένη χρονική περίοδο. Το ίδιο όχημα μπορεί να αντιστοιχεί σε πολλά προφίλ οχημάτων όταν υπάρχουν πληροφορίες για αυτό για διαφορετικές χρονικές περιόδους. Έτσι, το dataset προφίλ οχημάτων περιέχει πληροφορίες για την οδήγηση του οχήματος για την χρονική περίοδο που αφορά. Επιπρόσθετα, το δεύτερο dataset αφορά τα ίδια τα οχήματα. Συγκεκριμένα, περιλαμβάνει 72 γραμμές και 53 στήλες. Αυτό σημαίνει ότι υπάρχουν δεδομένα για 72 μοναδικά οχήματα. Τέλος, το δεύτερο dataset έχει πληροφορίες για τα χαρακτηριστικά του οχήματος και γενικές πληροφορίες για την οδήγηση του.

Τα οχήματα του dataset βρίσκονται στην Πολωνία αφού και η start-up και η εταιρία κάτοχος των οχημάτων είναι πολωνικές. Επίσης, τα προφίλ οχημάτων έχουν δεδομένα για χρονική διάρκεια είτε ενός μηνός είτε τριών μηνών. Συγκεντρωτικά, υπάρχουν δεδομένα για τα προφίλ οχημάτων από τον Ιούνιο του 2020 έως και τον Ιούλιο του 2021.

3.2 Data pre-processing

Προτού γίνει οποιαδήποτε ανάλυση και μοντέλο μηχανικής μάθησης, τα δεδομένα πρέπει να συγκεντρωθούν, καθαριστούν και να μετατραπούν στην κατάλληλη μορφή. Το συγκεκριμένο βήμα χρειάστηκε αρκετή προσπάθεια και χρόνο, μιας και τα αρχικά dataset περιλάμβαναν πολλές άχρηστες πληροφορίες, στήλες σε λάθος μορφές και πληθώρα δεδομένων που πρέπει να επεξεργαστούν σε συγκεντρωτικά χαρακτηριστικά.

Αρχίζοντας με τα χαρακτηριστικά με άχρηστες πληροφορίες, διαγράφτηκαν οι στήλες με τον κωδικό στόλου, κωδικό οχήματος και την τελευταία ενημέρωση. Επιπλέον, δημιουργήθηκαν καινούργιες στήλες για τον πιο συχνό οδηγό του οχήματος και τον μήνα αρχής παρακολούθησης του προφίλ. Παράλληλα, δημιουργήθηκαν διάφορες καινούργιες στήλες για στατιστικά χαρακτηριστικά του οχήματος που ήταν αποθηκευμένα σε μια στήλη σε μορφή dictionary. Ακόμα, από τα δεδομένα των διαδρομών κατασκευάστηκαν μετρικές μέσου όρου, ποσοστών, ελαχίστου και μεγίστου όπως μέση διαδρομή και ποσοστό διαδρομών με φορτιστή ηλεκτροκίνητου οχήματος σε κοντινή απόσταση. Τέλος, διαγράφτηκαν τα προφίλ οχημάτων με περιορισμένες πληροφορίες για τα όρια ταχυτήτων στους δρόμους των διαδρομών, με περιορισμένη δραστηριότητα κίνησης και με περιορισμένες καταγεγραμμένες ταχύτητες διαδρομών.

Το τελικό dataset περιέχει 934 γραμμές και 73 στήλες. Η τελευταία στήλη αφορά το χαρακτηριστικό στόχου δηλαδή την ετικέτα που θέλουμε να προβλέψουμε. Προφανώς, η στήλη του χαρακτηριστικού στόχου έχει διαφορετική μορφή και τιμές για κάθε ξεχωριστό είδος προβλήματος που θα επιλυθεί. Παρακάτω φαίνεται ο πίνακας που περιγράφει τις διάφορες στήλες.

Πίνακας 3.1: Περιγραφή χαρακτηριστικών

Χαρακτηριστικό	**Περιγραφή**
monitoring start month	πρώτος μήνας παρακολούθησης προφίλ
monitoring days	αριθμός ημερών παρακολούθησης του προφίλ
active days	αριθμός ημερών που το όχημα πραγματοποίησε τουλάχιστον μια διαδρομή
data quality score	μετρική 1-10 που περιγράφει κατά πόσο οι εξεταζόμενες διαδρομές είναι αρκετές και ποιοτικές για την εξαγωγή ακριβών συμπερασμάτων
reliability score	μετρική 1-10 της μεταβλητότητας του καθημερινού μοτίβου ενός οδηγού
car use score	μετρική 1-10 που περιγράφει την πτώση στην εμβέλεια του πιθανού ηλεκτρικού οχήματος λόγω άσχημων συνθηκών θερμοκρασίας και δρόμων που οδηγήθηκε το όχημα
driver behavior score	μετρική 1-10 που περιγράφει την καταλληλότητα του οδηγού για οδήγηση ηλεκτρικού οχήματος όσον αφορά τη συμπεριφορά του (τάσεις υπερβολικής ταχύτητας)
Συνέχεια στην επόμενη σελίδα	

Πίνακας 3.1 – συνέχεια από την προηγούμενη σελίδα

Χαρακτηριστικό	Περιγραφή
	και την ταχύτητα (υπέρβαση του σημείου αποκοπής των ηλεκτρικών οχημάτων των 90 Km/h)
number of trips	αριθμός καταγεγραμμένων διαδρομών
fuel costs	συνολικά έξοδα για καύσιμα
body type	τύπος αμαξώματος αυτοκινήτου
driver name	όνομα οδηγού οχήματος πλειοψηφίας διαδρομών
fuelcost per trip	μέσο κόστος καυσίμων ανά διαδρομή
distance per speed limit 0-30, 30-50 etc.	συνολική απόσταση που καλύφθηκε ανά κατηγορία ορίων ταχύτητας
speed limit distribution 0-30, 30-50 etc.	κατανομή αποστάσεων ανά κατηγορία ορίων ταχύτητας
distance per temp class 0-10, 10-20 etc.	συνολική απόσταση που καλύφθηκε ανά κατηγορία θερμοκρασίας
distance with temp ratio	αναλογία συνολικής απόστασης όπου ήταν γνωστή η θερμοκρασία
states availability	ποσοστό διαδρομών με καταγεγραμμένη ταχύτητα
range drop speed	ποσοστό εμβέλειας που θα χαθεί λόγω των συνθηκών των δρόμων που οδηγήθηκε το όχημα
range drop temp	ποσοστό εμβέλειας που θα χαθεί λόγω των συνθηκών θερμοκρασίας που οδηγήθηκε το όχημα
total distance	συνολική απόσταση διαδρομών
distance with limits ratio	αναλογία απόστασης με γνωστά όρια ταχυτήτων
tot overspeeding distance	συνολική απόσταση όπου ο οδηγός υπερέβη το όριο ταχύτητας
overspeeding ratio	αναλογία απόστασης όπου ο οδηγός υπερέβη το όριο ταχύτητας
overspeeding distance per speed limit 0-30, 30-50 etc.	απόσταση που ο οδηγός υπερέβη το όριο ταχύτητας ανά κατηγορία ορίου ταχύτητας
overspeeding ratio per speed limit 0-30, 30-50 etc.	αναλογία απόστασης που ο οδηγός υπερέβη το όριο ταχύτητας ανά κατηγορία ορίου ταχύτητας
total over 90 distance	συνολική απόσταση που διανύθηκε με ταχύτητα μεγαλύτερη από 90 Km/h
over 90 ratio	αναλογία απόστασης που διανύθηκε με ταχύτητα μεγαλύτερη από 90 Km/h
total segments	αριθμός οδικών τμημάτων που διασχίστηκαν στις διαδρομές
segments with limits	αριθμός οδικών τμημάτων με γνωστά όρια ταχυτήτων
speed limits availability	αναλογία οδικών τμημάτων με γνωστά όρια ταχυτήτων
active days score	ποσοστό ημερών στις οποίες το όχημα ήταν ενεργό
recorded distance ratio	αναλογία καταγεγραμμένης απόστασης
average-max-median trip distance	στατιστικά για την απόσταση διαδρομής
average-max-median pause time	στατιστικά για το διάλειμμα ανάμεσα σε διαδρομές
percentage start home	ποσοστό διαδρομών που ξεκινάνε από την τοποθεσία
Συνέχεια στην επόμενη σελίδα	

Πίνακας 3.1 – συνέχεια από την προηγούμενη σελίδα

Χαρακτηριστικό	Περιγραφή
	σπιτιού του οδηγού
percentage start office	ποσοστό διαδρομών που ξεκινάνε από την τοποθεσία γραφείου του οδηγού
average trip duration sec	μέση διάρκεια διαδρομής σε δευτερόλεπτα
percentage charger nearby	ποσοστό διαδρομών με φορτιστή ηλεκτρικού οχήματος σε κοντινή απόσταση
percentage charger fast nearby	ποσοστό διαδρομών με γρήγορο φορτιστή ηλεκτρικού οχήματος σε κοντινή απόσταση
percentage charger 3F nearby	ποσοστό διαδρομών με 3F φορτιστή ηλεκτρικού οχήματος σε κοντινή απόσταση

3.3 Data visualization

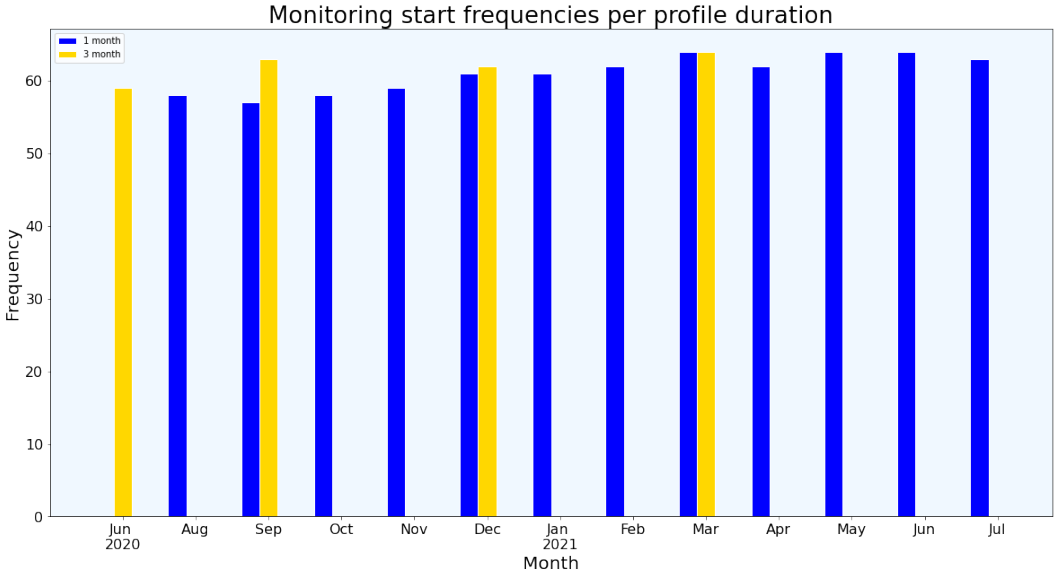
Σε αυτό το σημείο θα παρουσιαστούν μερικές οπτικοποιήσεις των χαρακτηριστικών για να γίνουν πιο κατανοητά τα δεδομένα, η μορφή τους και κάποιες σχέσεις μεταξύ χαρακτηριστικών. Προφανώς, με 73 διαφορετικά χαρακτηριστικά θα μπορούσαν να πραγματοποιηθούν αμέτρητα διαγράμματα όπου το καθένα θα είχε την δικιά του αξία. Ωστόσο, αποφασίστηκε να παρουσιαστούν μερικά από τα σημαντικά μόνο διαγράμματα αφού η συγκεκριμένη ανάλυση αφορά κυρίως την μηχανική μάθηση.

Το διάγραμμα 3.1 αφορά τις συχνότητες της αρχής παρακολούθησης των διαφορετικών προφίλ. Καταρχάς, κάθε προφίλ έχει δεδομένα παρακολούθησης του οχήματος είτε για τρεις μήνες είτε για έναν μήνα. Στην οπτικοποίηση φαίνεται ο αριθμός των προφίλ για κάθε μήνα όπου αρχίζει η παρακολούθηση του προφίλ. Μάλιστα, υπάρχουν ξεχωριστά χρώματα για τα προφίλ με διάρκεια ενός μήνα και για τα προφίλ με τρεις μήνες. Είναι κατανοητό ότι υπάρχουν δεδομένα για όλους τους διαφορετικούς μήνες και εποχές.

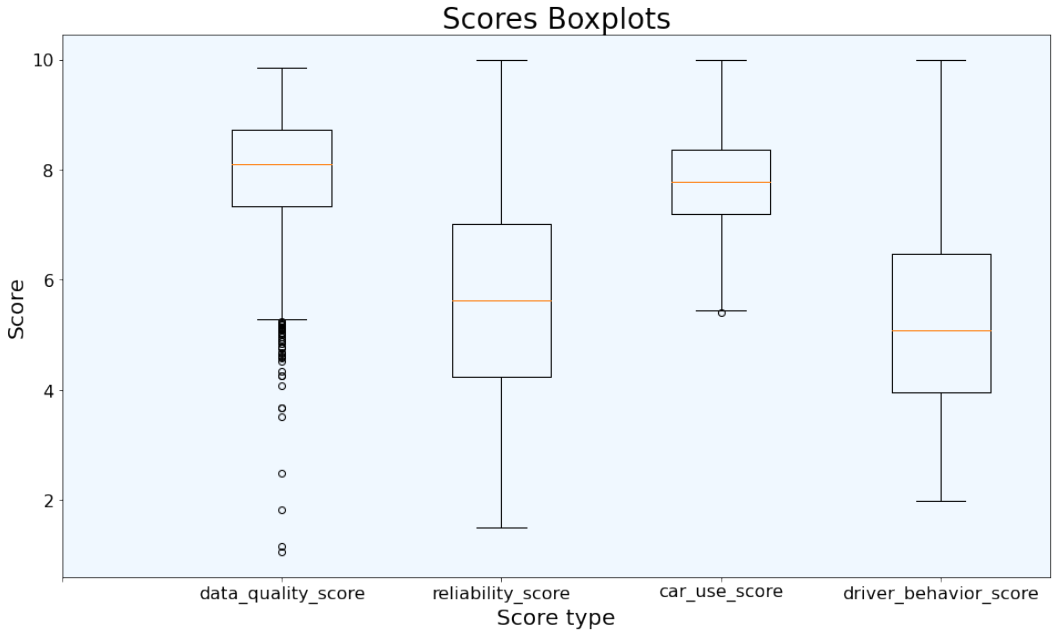
Η οπτικοποίηση 3.2 αφορά τα θηκογράμματα των διαφορετικών βαθμολογιών χαρακτηριστικών. Συγκεκριμένα, τα χαρακτηριστικά που περιγράφονται είναι το data quality score, reliability score, car use score και drivers behavior score. Επιπλέον, τα θηκογράμματα παρουσιάζουν την διάμεσο, τα τεταρτημόρια, το ελάχιστο και μέγιστο με βάση την κατανομή και τα outliers. Όλες οι βαθμολογίες φαίνεται ότι παίρνουν τιμές από το 1 έως το 10 ενώ κάθε βαθμολογία έχει ξεχωριστά στατιστικά χαρακτηριστικά.

Το διάγραμμα 3.3 περιγράφει την εξέλιξη του αριθμού των διαδρομών των οχημάτων με την πάροδο του χρόνου. Αναλυτικά, για κάθε μήνα φαίνονται οι συνολικές μοναδικές διαδρομές που οδηγήθηκαν από τα οχήματα του στόλου. Τον μήνα του Ιουνίου 2021 πραγματοποιήθηκαν οι περισσότερες διαδρομές, ενώ τον μήνα του Νοεμβρίου 2020 οι λιγότερες.

Σχήμα 3.1: Συχνότητες αρχής περιόδου παρακολούθησης



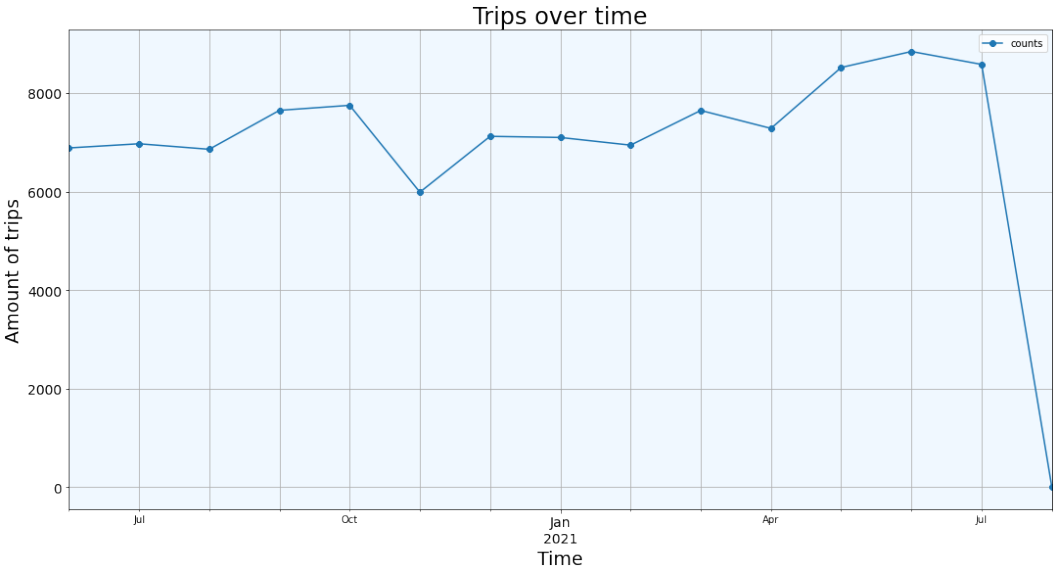
Σχήμα 3.2: Θηκογράμματα βαθμολογιών



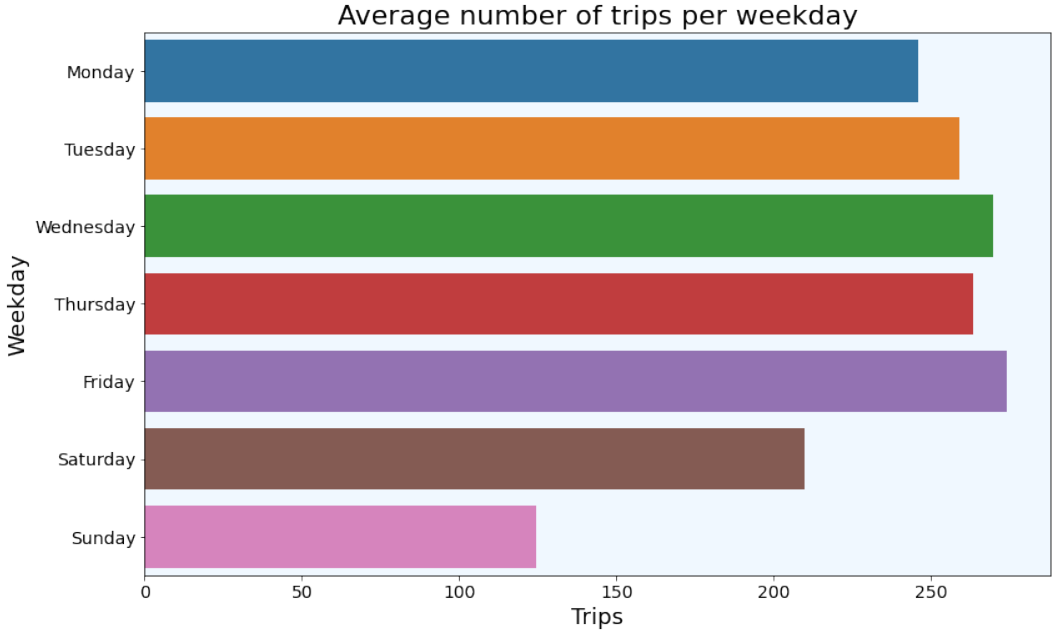
Το διάγραμμα 3.4 προβάλλει τον μέσο όρο διαδρομών ανά ημέρα εβδομάδας. Είναι ξεκάθαρο ότι κατά μέσο όρο τα σαββατοκύριακα καταγράφονται οι λιγότερες διαδρομές των οχημάτων του στόλου. Από την άλλη πλευρά, την Παρασκευή πραγματοποιούνται οι περισσότερες διαδρομές οχημάτων.

Η οπτικοποίηση 3.5 αφορά τις συχνότητες που τα διαφορετικά μοντέλα ηλεκτρικών οχημά-

Σχήμα 3.3: Αριθμός διαδρομών με την πάροδο του χρόνου

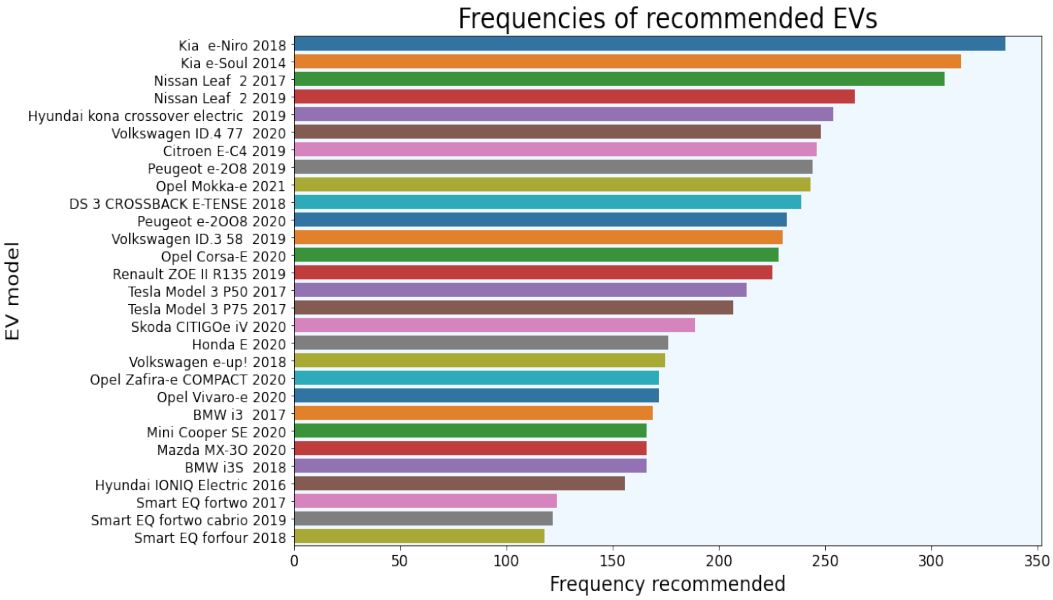


Σχήμα 3.4: Μέσος όρος διαδρομών ανά ημέρα εβδομάδας



των συστήθηκαν για αντικατάσταση ενός συμβατικού οχήματος. Συγκεκριμένα, τα μοντέλα Kia e-Niro 2018, Kia e-Soul 2014 και Nissan Leaf 2 2017 προτάθηκαν τις περισσότερες φορές, ενώ τα διάφορα μοντέλα Smart EQ τις λιγότερες φορές. Παρόλα αυτά, το συγκεκριμένο διάγραμμα, όπως και η ανάλυση στην συνέχεια, παρουσιάζει μόνο τα ηλεκτρικά οχήματα με λογικές τιμές. Με άλλα λόγια, δεν συμπεριλαμβάνονται τα πολυτελή και ακριβά ηλεκτρικά μοντέλα.

Σχήμα 3.5: Συχνότητες προτεινόμενων ηλεκτρικών μοντέλων



3.4 Classification tasks

Η συγκεκριμένη ανάλυση αποτελείται από 3 διαφορετικά τμήματα. Αναλυτικά, αντιμετωπίζει 3 ξεχωριστά προβλήματα κατηγοριοποίησης χρησιμοποιώντας μοντέλα μηχανικής μάθησης και συγκρίνοντας τις επιδόσεις τους σε κάθε ξεχωριστό πρόβλημα με τις κατάλληλες μετρικές. Μάλιστα, το κάθε πρόβλημα είναι και ένα διαφορετικό είδος κατηγοριοποίησης όπως περιγράφηκε στο κεφάλαιο της βιβλιογραφικής ανασκόπησης.

Ένα χαρακτηριστικό που συνιστά ζωτικής σημασίας στις διάφορες κατηγοριοποιήσεις των προφίλ οχημάτων είναι τα προτεινόμενα ηλεκτρικά οχήματα του προφίλ. Φυσικά, κάθε προτεινόμενο ηλεκτρικό όχημα δεν είναι εξίσου κατάλληλο για ένα συγκεκριμένο προφίλ συμβατικού. Αυτός είναι ο λόγος για τον οποίο κάθε προτεινόμενο ηλεκτρικό όχημα σε ένα προφίλ έχει μια συγκεκριμένη βαθμολογία που ονομάζεται real score. Η μετρική real score είναι το εύρος-απόσταση που μπορεί να επιτευχθεί μέσω των υπαρχόντων σταθμών φόρτισης. Αναλυτικά, είναι η σταθμισμένη αναλογία της συνολικής απόστασης χιλιομέτρων που ανακτάται από τη φόρτιση του αυτοκινήτου με τους διαθέσιμους φορτιστές σε σχέση με αυτήν που ανακτάται ιδανικά ακολουθώντας την προτεινόμενη στρατηγική φόρτισης. Με απλά λόγια, η μετρική παίρνει τιμές από 1-10, όπου το 1 είναι το χειρότερο και το 10 είναι το καλύτερο. Έτσι, μια βαθμολογία 10 σημαίνει

ότι οι τρέχουσες υποδομές φόρτισης επιτυγχάνουν το βέλτιστο εύρος φόρτισης.

Με βάση τη συγκεκριμένη μετρική, ένα μοντέλο ηλεκτρικού οχήματος θεωρείται ως κατάλληλο ή ακατάλληλο να αντικαταστήσει το εξετασθέν προφίλ οχήματος. Πιο συγκεκριμένα, μετά από ανάλυση των κατανομών όλων των μετρικών real score και με βάση την εμπειρία του υπεύθυνου του data science team, η βαθμολογία real score του 5.2 τέθηκε ως το όριο που καθορίζει αν το συγκεκριμένο ηλεκτρικό όχημα μπορεί να αντικαταστήσει το συμβατικό. Με άλλα λόγια, ένα ηλεκτρικό όχημα που βαθμολογείται με real score 7 για ένα προφίλ οχήματος, θεωρείται ως κατάλληλο να αντικαταστήσει το συγκεκριμένο προφίλ συμβατικού οχήματος. Από την άλλη πλευρά, ένα ηλεκτρικό όχημα που βαθμολογείται με real score 5 για ένα προφίλ οχήματος, θεωρείται ως ακατάλληλο να αντικαταστήσει το συγκεκριμένο προφίλ. Τέλος, η συγκεκριμένη υπόθεση αφορά τα προβλήματα binary και multi-label classification, όπως θα αναλυθούν στην συνέχεια.

Ένα άλλο βασικό σημείο είναι ότι εξετάζονται μόνο τα ηλεκτρικά οχήματα με τιμή κάτω από 250,000 ζλότι Πολωνίας που αντιστοιχούν σε περίπου 55,000 ευρώ. Φυσικά, τα πιο ακριβά ηλεκτρικά οχήματα μπορούν να επιτύχουν καλύτερη εμβέλεια και απόδοση, αλλά ταυτόχρονα απαιτούν μεγαλύτερο κεφάλαιο για να αποκτηθούν. Αποτελεσματικά, λαμβάνονται υπόψη μόνο τα φθηνά και προσιτά για τους περισσότερους μοντέλα ηλεκτρικών οχημάτων σε όλη την έκταση της ανάλυσης. Τέλος, τα συγκεκριμένα μοντέλα ηλεκτρικών οχημάτων φαίνονται και στο διάγραμμα 3.5.

3.4.1 Binary classification

Η πρώτη εφαρμογή αφορά την πρόβλεψη ενός προφίλ οχήματος ως κατάλληλου ή ακατάλληλου για να αντικατασταθεί από ηλεκτρικό όχημα. Φυσικά, για κάθε προφίλ συμβατικού οχήματος υπάρχουν διάφορα προτεινόμενα ηλεκτρικά οχήματα, καθένα από τα οποία έχει μια ξεχωριστή βαθμολογία real score για το εξεταζόμενο προφίλ. Αναλυτικότερα, η συγκεκριμένη σχέση μπορεί να θεωρηθεί πολλά προς πολλά, αφού υπάρχουν πολλά προφίλ οχημάτων που αντιστοιχίζονται σε πολλά μοντέλα ηλεκτρικών οχημάτων. Για αυτό το λόγο, για κάθε προφίλ διατηρείται μόνο το μοντέλο ηλεκτρικού οχήματος με την υψηλότερη βαθμολογία real score.

Πίνακας 3.2: Συχνότητες κατηγοριών binary classification

Κατηγορία	Αριθμός προφίλ
Ακατάλληλο για αντικατάσταση από ηλεκτρικό όχημα	580
Κατάλληλο για αντικατάσταση από ηλεκτρικό όχημα	354

Ο πίνακας 3.2, περιγράφει πως κατηγοριοποιούνται τα 934 προφίλ οχημάτων. Ειδικότερα, ένα προφίλ οχήματος που κατηγοριοποιείται ως ακατάλληλο, σημαίνει ότι όλα τα μοντέλα ηλεκτρικών οχημάτων είχαν βαθμολογίες 5.2 ή λιγότερο στην μετρική real score. Από την άλλη πλευρά, ένα προφίλ οχήματος που κατηγοριοποιείται ως κατάλληλο, σημαίνει ότι τουλάχιστον ένα μοντέλο ηλεκτρικού οχήματος βαθμολογήθηκε με περισσότερο από 5.2 στη μετρική real

score. Τέλος, για την αξιολόγηση των μοντέλων της συγκεκριμένη εφαρμογής μηχανικής μάθησης χρησιμοποιήθηκε η μετρική accuracy.

Βήματα

Για κάθε αλγόριθμο ακολουθήθηκαν τα παρακάτω βήματα για τον υπολογισμό της μετρικής accuracy:

- Εύρεση των καλύτερων παραμέτρων για κάθε αλγόριθμο χρησιμοποιώντας grid-search και 10-fold cross validation. Με άλλα λόγια, κάθε δηλωμένος συνδυασμός παραμέτρων εκπαιδεύτηκε για 10 τυχαίους αλλά επαναλαμβανόμενους διαχωρισμούς δεδομένων σε εκπαίδευσης και ελέγχου. Ο συνδυασμός παραμέτρων με την υψηλότερη μέση τιμή της μετρικής accuracy για τους διαχωρισμούς δεδομένων σε εκπαίδευσης-ελέγχου δηλαδή ο συνδυασμός παραμέτρων που έκαναν τις περισσότερες σωστές προβλέψεις, αποθηκεύτηκε.
- Οι καλύτερες παράμετροι του προηγούμενου βήματος χρησιμοποιήθηκαν στη συνέχεια, για την εκπαίδευση του classifier και για τον υπολογισμό της μετρικής accuracy χρησιμοποιώντας leave-one-out cross-validation. Αυτό σημαίνει ότι για κάθε διαχωρισμό των δεδομένων σε εκπαίδευσης και ελέγχου, όπου τα δεδομένα ελέγχου περιέχουν κάθε φορά μόνο μια παρατήρηση, ο classifier εκπαιδεύτηκε, η παρατήρηση ελέγχου προβλέφθηκε και τελικά υπολογίστηκε η μέση τιμή της μετρικής accuracy.

Στη συνέχεια, ο classifier με την υψηλότερη μέση τιμή accuracy χρησιμοποιήθηκε με την τεχνική smote για να αυξήσει περαιτέρω τη μέση τιμή accuracy. Μάλιστα, χρησιμοποιήθηκαν διαφορετικές στρατηγικές επαναδειγματοληψίας μαζί με grid-search και 20-fold cross validation για την εύρεση της βέλτιστης στρατηγικής smote και παραμέτρων του classifier. Στη συνέχεια, η βέλτιστη στρατηγική επαναδειγματοληψίας και οι παράμετροι χρησιμοποιήθηκαν για την εκπαίδευση του classifier και τον υπολογισμό της μέσης τιμής accuracy χρησιμοποιώντας leave-one-out cross-validation.

Τέλος, χρησιμοποιήθηκε η Bayesian βελτιστοποίηση χρησιμοποιώντας Gaussian διαδικασίες προκειμένου να βρεθεί η βέλτιστη τιμή της παραμέτρου *n_estimators*. Η συγκεκριμένη παράμετρος είναι ένας φυσικός αριθμός που θα ήταν δύσκολο να βελτιστοποιηθεί με το grid-search, επειδή συνιστά μια εξαντλητική αναζήτηση παραμέτρων. Αφού βρέθηκε η βέλτιστη τιμή, χρησιμοποιήθηκε μαζί με την τεχνική smote και τον classifier για τον υπολογισμό πάλι της μετρικής accuracy κάνοντας χρήση leave-one-out cross-validation.

3.4.2 Multi-class classification

Η επόμενη εφαρμογή μηχανικής μάθησης ασχολείται με ένα πρόβλημα multi-class classification. Συγκεκριμένα, δημιουργήθηκαν διαφορετικές κατηγορίες που περιγράφουν πόσο κατάλληλα ένα προφίλ οχήματος μπορεί να αντικατασταθεί από ένα ηλεκτρικό όχημα. Όπως και στην εφαρμογή binary classification, για κάθε προφίλ οχήματος λαμβάνεται υπόψη μόνο το προτεινόμενο μοντέλο ηλεκτρικού οχήματος με την υψηλότερη βαθμολογία real score.

Πίνακας 3.3: Κατηγορίες multi-class classification

Κατηγορία	Περιγραφή	Διάστημα real score	Αριθμός προφίλ
0	Καμία σύσταση ηλεκτρικού οχήματος	-	90
1	Απογοητευτικό real score	$real\_score < 4$	375
2	Χαμηλό real score	$4 \leq real\_score < 6.5$	207
3	Καλό real score	$6.5 \leq real\_score < 10$	198
4	Τέλειο real score	$real\_score = 10$	64

Ο πίνακας 3.3 εξηγεί πως κατηγοριοποιούνται τα 934 προφίλ του τελικού dataset. Επίσης, περιγράφονται και οι διαφορετικές κατηγορίες και τα διαστήματα real score που αφορά η καθεμία. Φαίνεται ότι το dataset παρουσιάζει κάποια ανισορροπία όσον αφορά τις συχνότητες των προφίλ που ανήκουν στις διαφορετικές κατηγορίες. Τέλος, έγινε χρήση της μετρικής accuracy για την αξιολόγηση της επίδοσης των μοντέλων μηχανικής μάθησης.

Βήματα

Αρχικά, ακολουθήθηκαν τα παρακάτω βήματα για κάθε classifier για τον υπολογισμό της μέσης τιμής accuracy:

- Εύρεση των καλύτερων παραμέτρων για κάθε αλγόριθμο χρησιμοποιώντας grid-search και 10-fold cross validation. Έπειτα, αποθηκεύτηκαν οι παράμετροι με την υψηλότερη μέση τιμή της μετρικής accuracy για τους 10 διαχωρισμούς δεδομένων σε εκπαίδευσης-ελέγχου.
- Οι καλύτερες παράμετροι του προηγούμενου βήματος χρησιμοποιήθηκαν για την εκπαίδευση του classifier και για τον υπολογισμό της μετρικής accuracy χρησιμοποιώντας leave-one-out cross-validation.

Επιπρόσθετα, για τους classifiers SVC, random forest, Naive Baynes και Logistic regression πραγματοποιήθηκαν:

- Δημιουργία pipeline ενός standard scaler, PCA και του classifier. Οι καλύτερες παράμετροι βρέθηκαν για την τεχνική PCA ($n\_components$) και τον classifier χρησιμοποιώντας grid-search και 10-fold cross validation.

- Οι καλύτερες παράμετροι που βρέθηκαν, χρησιμοποιήθηκαν για την εκπαίδευση κάθε pipeline προκειμένου να υπολογιστεί η μέση τιμή accuracy με leave-one-out cross-validation.

Για τους classifiers με το υψηλότερο μέσο accuracy εφαρμόστηκαν τα επόμενα βήματα μαζί με τεχνικές random-oversampler και smote:

- Για τη χρήση smote και random-oversampler ορίστηκαν διαφορετικές στρατηγικές επαναδειγματοληψίας
- Το grid-search χρησιμοποιήθηκε με 250-fold cross validation προκειμένου να βρεθεί το pipeline που αποτελούνταν από την βέλτιστη τεχνική oversampling, την βέλτιστη στρατηγική επαναδειγματοληψίας και τις βέλτιστες παραμέτρους του classifier.
- Η βέλτιστη τεχνική oversampling, η βέλτιστη στρατηγική επαναδειγματοληψίας και οι βέλτιστες παράμετροι του classifier ορίστηκαν για την εκπαίδευση ενός pipeline για να υπολογιστεί η μέση τιμή accuracy χρησιμοποιώντας leave-one-out cross validation

Για την τεχνική oversampling και τον classifier που παρήγαγαν την υψηλότερη μέση τιμή accuracy, έγινε χρήση Bayesian βελτιστοποίηση με Gaussian διαδικασίες προκειμένου να βρεθεί η βέλτιστη τιμή της παραμέτρου *n_estimators* που παίρνει διακριτές τιμές. Αφού βρέθηκε η βέλτιστη τιμή, υπολογίστηκε πάλι η μέση τιμή accuracy για leave-one-out cross validation.

3.4.3 Multi-label classification

Η τρίτη εφαρμογή μηχανικής μάθησης επιλύει ένα πρόβλημα multi-label. Μάλιστα, η συγκεκριμένη εφαρμογή εστιάζει στην πρόβλεψη όλων των μαρκών ηλεκτρικών οχημάτων που είναι κατάλληλες να αντικαταστήσουν το εξετάζον προφίλ συμβατικού οχήματος. Όπως και στην εφαρμογή binary classification, προκειμένου ένα ηλεκτρικό όχημα να θεωρηθεί κατάλληλο να αντικαταστήσει ένα συμβατικό, πρέπει να έχει βαθμολογηθεί με περισσότερο από 5.2 στην μετρική real score. Έστω ότι ένα προφίλ μπορεί να αντικατασταθεί από τα μοντέλα ηλεκτρικών οχημάτων Kia e-Soul 2014, Kia e-Niro 2018 και BMW i3s 2018, τότε ο classifier θα πρέπει να προβλέψει τις μάρκες Kia και BMW ως κατάλληλες για το συγκεκριμένο προφίλ και όλες τις υπόλοιπες μάρκες ως ακατάλληλες για το προφίλ.

Συνολικά υπάρχουν 16 διαφορετικές μάρκες ηλεκτρικών οχημάτων στο dataset, η καθεμία από τις οποίες αποτελεί και μια ξεχωριστή ετικέτα που καλείται να προβλεφθεί ως αληθής ή ψευδής. Πιο συγκεκριμένα αυτές οι μάρκες είναι οι Nissan, Peugeot, Hyundai, Renault, Skoda, Honda, Mazda, Volkswagen, DS, BMW, Smart, Kia, Mini Cooper, Citroen, Tesla και Opel. Επιπρόσθετα, αφού το πρόβλημα είναι multi-label, η αξιολόγηση της επίδοσης των μοντέλων γίνεται με την μετρική hamming loss.

Βήματα

Για κάθε προσέγγιση και αλγόριθμο υλοποιήθηκαν τα παρακάτω βήματα για τον υπολογισμό της μέσης τιμής της μετρικής hamming loss:

- Εύρεση των καλύτερων παραμέτρων για κάθε προσέγγιση-αλγόριθμο χρησιμοποιώντας grid-search και 10-fold cross validation. Ο συνδυασμός παραμέτρων με τη χαμηλότερη μέση τιμή hamming loss για τους διαχωρισμούς δεδομένων εκπαίδευσης-ελέγχου αποθηκεύεται. Με άλλα λόγια, για κάθε προσέγγιση-αλγόριθμο κρατάμε τις παραμέτρους που επιτυγχάνουν την ακριβής πρόβλεψη των περισσότερων μεμονωμένων ετικετών.
- Οι καλύτερες παράμετροι του προηγούμενου βήματος χρησιμοποιήθηκαν για την εκπαίδευση της προσέγγισης-αλγορίθμου και τον υπολογισμό της μέσης τιμής της μετρικής hamming loss χρησιμοποιώντας leave-one-out cross-validation.

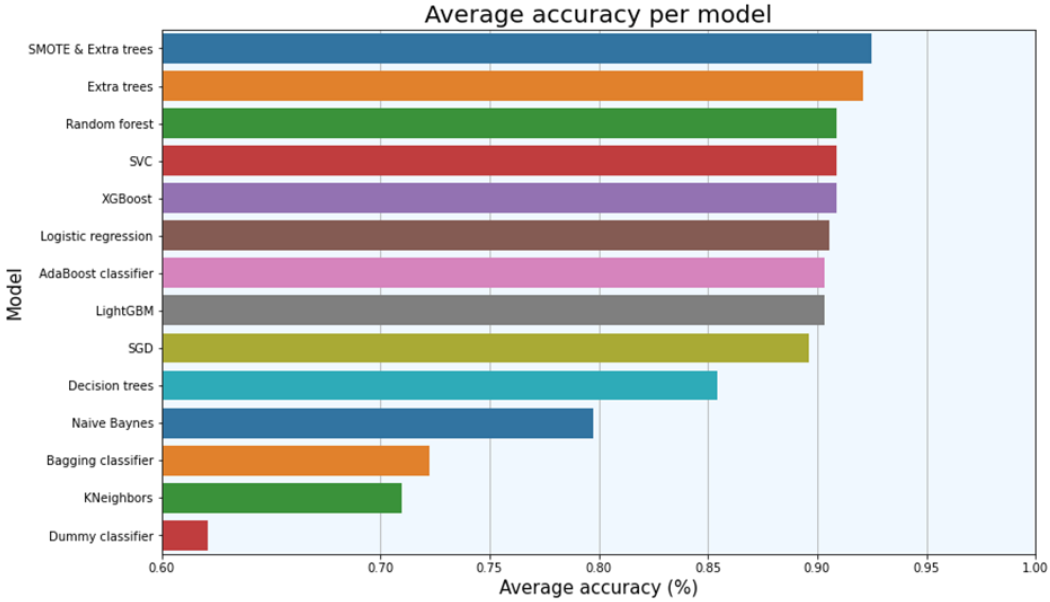
Κεφάλαιο 4

Αποτελέσματα και Ανάλυση

Σε αυτό το σημείο εξετάζονται οι επιδόσεις των αλγορίθμων μηχανικής μάθησης σε καθένα από τα προβλήματα classification. Για κάθε πρόβλημα συγκρίνονται οι διαφορετικοί αλγόριθμοι-pipeline αλγορίθμων και τελικά παρουσιάζεται εκείνο που επιτυγχάνει τα καλύτερα αποτελέσματα με βάση την μετρική του προβλήματος.

4.1 Binary classification

Σχήμα 4.1: Επιδόσεις αλγορίθμων στο binary classification



Στο πρόβλημα binary classification ο dummy classifier είχε το χαμηλότερο accuracy. Ανα-

λυτικά, ο dummy classifier προβλέπει πάντα την συχνότερη κατηγορία και χρησιμοποιείται μόνο για σύγκριση με τα άλλα έξυπνα μοντέλα. Οι classifiers KNeighbors, bagging και Naive Baynes ναι μεν έχουν υψηλότερο accuracy από τον dummy classifier, αλλά συγκριτικά με τα υπόλοιπα μοντέλα δεν φέρνουν καλά αποτελέσματα. Επίσης, τα δέντρα αποφάσεων ξεπερνούν το 85% accuracy και το SGD φτάνει στο 89.6% accuracy. Ωστόσο, οι καλύτεροι classifiers θεωρούνται τα extra trees, random forest, SVC, XGBoost, logistic regression, adaboost classifier και LightGBM αφού πετυχαίνουν πάνω από 90% average accuracy.

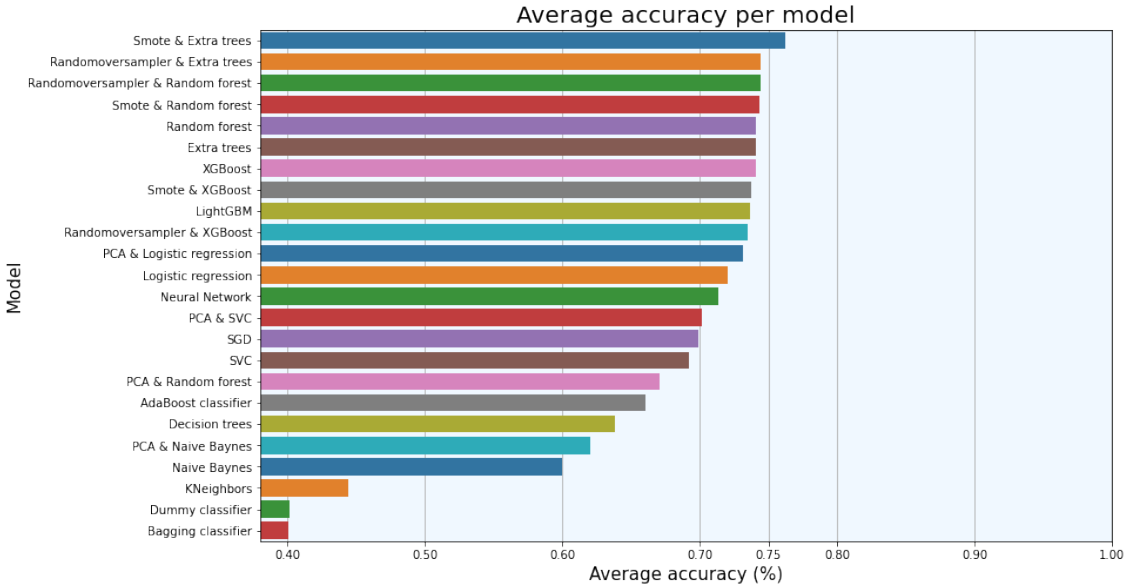
Τα extra trees αποτελούν τον αλγόριθμο που επιτυγχάνει το υψηλότερο accuracy με 92%. Μάλιστα, η χρήση της τεχνικής smote οδήγησε σε μια αύξηση 0.32 στη μετρική accuracy. Επιπλέον, η χρήση της Bayesian βελτιστοποίησης με Gaussian Processes για την εύρεση της βέλτιστης τιμής *n_estimators* είχε ως επακόλουθο την περεταίρω αύξηση της μετρικής accuracy κατά 0.107. Τέλος, το τελικό μοντέλο φτάνει το 92.5% accuracy και φαίνεται στο σχήμα 4.2

Σχήμα 4.2: Καλύτερο pipeline-αλγορίθμων binary classification

```
Pipeline(steps=[('smote',
                 SMOTE(random_state=42, sampling_strategy={0: 580, 1: 400})),
                ('extratreesclassifier',
                 ExtraTreesClassifier(class_weight='balanced', max_depth=10,
                                      max_features=None, n_estimators=65,
                                      random_state=7, warm_start=True))])
```

4.2 Multi-class classification

Σχήμα 4.3: Επιδόσεις αλγορίθμων στο multi-class classification



Στο πρόβλημα multi-class classification χρησιμοποιήθηκαν πληθώρα διαφορετικών αλγορίθμων και pipeline-αλγορίθμων. Οι χαμηλότερες επιδόσεις σημειώνονται από τον dummy, bagging

classifier και KNeighbors. Από την άλλη πλευρά, οι αλγόριθμοι και τα pipelines που περιέχουν τα extra trees, random forest, XGBoost και LightGBM καταγράφουν τα υψηλότερα accuracy με πάνω από 73%. Μια σημαντική παρατήρηση είναι ότι η χρήση της τεχνικής PCA κατάφερε να αυξήσει ικανοποιητικά το μέσο accuracy στα μοντέλα Naive Baynes, SVC και logistic regression αλλά όχι στο μοντέλο random forest. Ακόμα, η χρήση ovesampling τεχνικών στα μοντέλα XGBoost και LightGBM όχι μόνο δεν αύξησαν τη μετρική accuracy αλλά την μείωσαν κιόλας. Ωστόσο, η χρήση ovesampling τεχνικών στα μοντέλα random forest και extra trees είχε ως επακόλουθο την μικρή αύξηση του accuracy.

Και σε αυτό το πρόβλημα τα extra trees κατάφεραν να πετύχουν το υψηλότερο accuracy. Αναλυτικά, η χρήση extra trees σκόραρε 74% accuracy. Μάλιστα, όταν συνδυάστηκε με την χρήση της τεχνικής smote, το accuracy αυξήθηκε κατά 2 ολόκληρες μονάδες φτάνοντας στο 76%. Επίσης, η χρήση Bayesian βελτιστοποίησης με Gaussian Processes για την εύρεση της βέλτιστης τιμής *n_estimators* οδήγησε σε αύξηση του accuracy κατά 0.21. Τέλος, το pipeline smote-extra trees φτάνει το 76.23% accuracy και περιγράφεται στην εικόνα 4.4

Σχήμα 4.4: Καλύτερο pipeline-αλγορίθμων multi-class classification

```
Pipeline(steps=[('smote',
                 SMOTE(random_state=42,
                       sampling_strategy={0: 95, 1: 375, 2: 207, 3: 198,
                                          4: 67})),
                ('extratreesclassifier',
                 ExtraTreesClassifier(criterion='entropy', max_features=None,
                                      n_estimators=52, random_state=7,
                                      warm_start=True))])
```

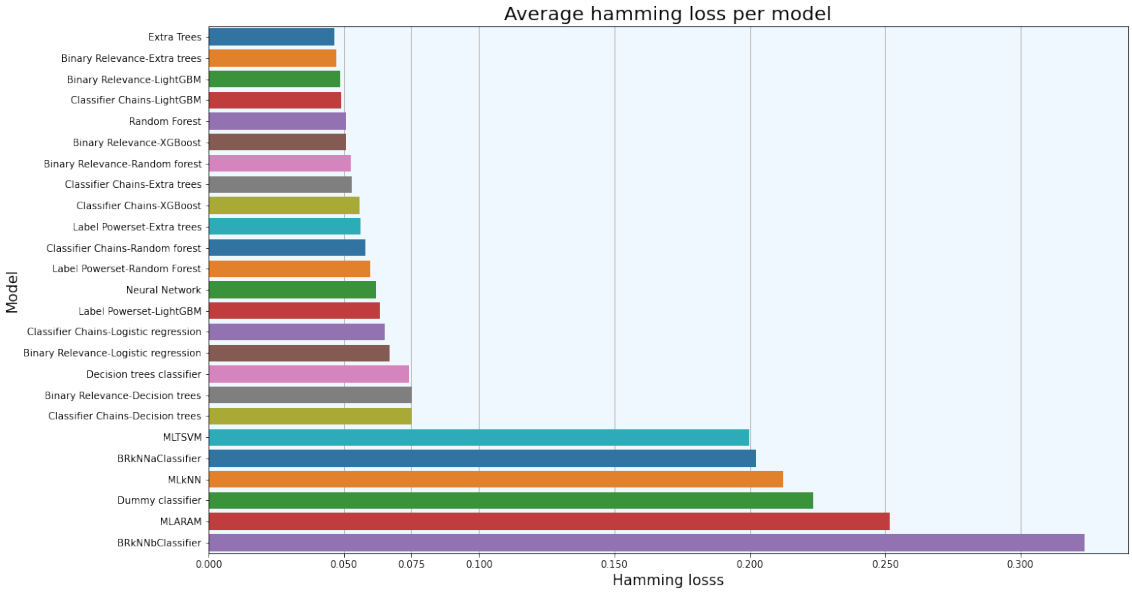
4.3 Multi-label classification

Τα μοντέλα του προβλήματος multi-label classification αξιολογούνται με την μετρική hamming loss. Έτσι, όσο χαμηλότερη η τιμή της μετρικής τόσο καλύτερη η επίδοση του μοντέλου. Αυτό οφείλεται στο γεγονός ότι η συγκεκριμένη μετρική περιγράφει απώλειες. Με βάση το διάγραμμα 4.5 οι χειρότερες επιδόσεις φαίνονται για τους προσαρμοσμένους αλγορίθμους και τον dummy classifier. Έπειτα, τα δέντρα απόφασης και logistic regression σε διάφορες προσεγγίσεις μετασχηματισμού προβλήματος απέδωσαν πολύ καλύτερα με hamming loss περίπου 0,07. Προχωρώντας, το νευρωνικό δίκτυο πέτυχε απώλεια hamming 0.062, ξεπερνώντας τα προαναφερθείσα μοντέλα. Επιπλέον, το random forest, τα extra trees, το XGBoost και το LightGBM που χρησιμοποιούνται σε διαφορετικές προσεγγίσεις μετασχηματισμού προβλημάτων παρήγαγαν τα καλύτερα αποτελέσματα.

Ο αλγόριθμος των extra trees έφτασε την ελάχιστη καταγεγραμμένη τιμή hamming loss με 0.0467. Μάλιστα, ο συγκεκριμένος αλγόριθμος ανήκει στις ensemble προσεγγίσεις. Το ακριβές μοντέλο extra trees περιγράφεται στο σχήμα 4.6

Τώρα θα εξεταστεί τι ακριβώς σημαίνει hamming loss τιμής 0.0467 στη συγκεκριμένη εφαρμογή. Αν έπρεπε να προβλεφθούν οι κατάλληλες μάρκες ηλεκτρικών οχημάτων για 4 δια-

Σχήμα 4.5: Επιδόσεις αλγορίθμων στο multi-label classification



φορετικά προφίλ οχημάτων (συνολικά $4 \cdot 16 = 64$ μάρκες), ο classifier extra trees θα προέβλεπε με ακρίβεια κατά μέσο όρο όλες τις ετικέτες μάρκας εκτός από 3 ετικέτες αφού $3/64 = 0.046875$ που είναι σχεδόν όσο και το hamming loss. Με άλλα λόγια, μόνο 3 ετικέτες μάρκας θα κατηγοριοποιηθούν εσφαλμένα (είτε μια κατάλληλη μάρκα ως ακατάλληλη είτε μια ακατάλληλη μάρκα ηλεκτρικού οχήματος ως κατάλληλη για ένα συγκεκριμένο προφίλ οχήματος) από τον συνολικό αριθμό των 64 ετικετών-μαρκών.

Σχήμα 4.6: Καλύτερος αλγόριθμος multi-class classification

```
ExtraTreesClassifier(criterion='entropy', max_features=None, n_estimators=125,
                     random_state=7, warm_start=True)
```

4.4 Εφαρμογή πιθανοτήτων

Σε όλα τα εξετασθέντα προβλήματα τα extra trees συμμετέχουν στην κατασκευή του καλύτερου pipeline-μοντέλου. Μια πολύ καλή χρήση του μοντέλου extra trees είναι ο υπολογισμός των πιθανοτήτων να ανήκει ένα προφίλ ελέγχου σε καθεμία από τις κατηγορίες του προβλήματος classification. Για να γίνει κατανοητό αυτό, θα παρουσιαστεί ένα παράδειγμα από το πρόβλημα multi-class classification που περιγράφτηκε ήδη.

Στο πίνακα 4.1 περιγράφονται οι πιθανότητες των προφίλ ελέγχου να ανήκουν στις διάφορες κατηγορίες με βάση το καλύτερο μοντέλο extra-trees για το πρόβλημα multi-class classification που αναλύθηκε. Σε κάθε στήλη παρουσιάζονται οι πιθανότητες και με bold φαίνεται η αληθής πρόβλεψη. Αναλυτικά, το πρώτο προφίλ των δεδομένων-ελέγχου προβλέφθηκε ως καλό real score όσον αφορά την καταλληλότητα του να αντικατασταθεί από ηλεκτρικό όχημα. Μάλιστα,

Πίνακας 4.1: Εφαρμογή πιθανοτήτων σε testing data

Κατηγορία	Περιγραφή	Προφίλ 1	Προφίλ 2	Προφίλ 3
0	Καμία σύσταση	19	19	38.5
1	Απογοητευτικό real score	0	**71**	25
2	Χαμηλό real score	6	10	**36.5**
3	Καλό real score	**75**	0	0
4	Τέλειο real score	0	0	0

η συγκεκριμένη πρόβλεψη αντιστοιχεί σε πιθανότητα 75% και είναι αληθής αφού το προφίλ είχε στην πραγματικότητα καλό real score. Επιπρόσθετα, για το δεύτερο προφίλ έγινε η πρόβλεψη ότι ανήκει στην κατηγορία απογοητευτικό real score με πιθανότητα 71%. Πράγματι, η συγκεκριμένη πρόβλεψη είναι σωστή. Ακόμα, το τρίτο προφίλ προβλέφθηκε ως καμία σύσταση ηλεκτρικού οχήματος με πιθανότητα 38.5% που είναι και η υψηλότερη μεταξύ των άλλων κατηγοριών. Ωστόσο, το συγκεκριμένο προφίλ αντιστοιχεί στη κατηγορία χαμηλό real score με πιθανότητα 36.5%.

Η χρήση των πιθανοτήτων είναι αρκετά σημαντική στα προβλήματα classification. Καταρχάς, στην περίπτωση του multi-class classification μας, το pipeline extra trees προβλέπει σωστά 76.23% των περιπτώσεων το οποίο αποτελεί ένα ικανοποιητικό αποτέλεσμα. Βέβαια, η χρήση των πιθανοτήτων μπορεί να δώσει μια πιο ολοκληρωμένη εικόνα σε κάθε πρόβλεψη, αποκλείοντας κάποιες κατηγορίες και δίνοντας βαρύτητα σε άλλες. Αυτό φαίνεται στο παράδειγμα του τρίτου προφίλ ελέγχου. Ναι μεν η πρόβλεψη δεν ήταν ακριβής, αλλά είχε γίνει γνωστό το ενδεχόμενο το προφίλ να ανήκει στην κατηγορία χαμηλό real score. Αποτελεσματικά, σε περίπτωση που θα επιτρεπόταν και δεύτερη πρόβλεψη, αυτή θα ήταν σωστή.

Κεφάλαιο 5

Συμπεράσματα

Ανακεφαλαιώνοντας, η συγκεκριμένη εργασία ανέλυσε τρία διαφορετικά είδη προβλημάτων classification στα πλαίσια μια πραγματική εταιρίας που ασχολείται με την τεχνολογία και την μηχανική μάθηση. Άλλωστε, παρουσιάστηκαν και δοκιμάστηκαν πληθώρα τεχνικών και αλγορίθμων μηχανικής μάθησης και συνδυασμών τους προκείμενου να ανιχνευτούν εκείνοι που μπορούν να επιλύσουν κάθε ξεχωριστό πρόβλημα καλύτερα. Τα τελικά μοντέλα μηχανικής μάθησης πραγματοποιούν προβλέψεις σχετικά με την αντικατάσταση συμβατικών οχημάτων με ηλεκτρικά, την καταλληλότητα τους για αντικατάσταση από ηλεκτρικά και τις μάρκες ηλεκτρικών οχημάτων κατάλληλες για κάθε όχημα.

Τα περιγραφέντα μοντέλα μπορούν να διευκολύνουν την λήψη αποφάσεων των χρηστών του στόλου οχημάτων όσον αφορά την αντικατάσταση συμβατικών οχημάτων. Συγκεκριμένα, τα μοντέλα δίνουν γρήγορα μια εικόνα του οχήματος αφού προβλέπουν άμεσα σημαντική γνώση σχετικά με αυτό. Έτσι, μπορεί να εξοικονομηθεί πολύτιμος χρόνος, μιας και ο χρήστης δεν πρέπει να περιμένει να εκτελεστεί ολόκληρος ο αλγόριθμος electric-vehicle-recommendation. Ο συγκεκριμένος αλγόριθμος είναι πολύ δαπανηρός αφού πρέπει να υπολογίσει την βέλτιστη στρατηγική φόρτισης. Τέλος, σε μελλοντικό στάδιο τα μοντέλα μπορούν να εξελιχθούν και να συνδυαστούν με περισσότερα δεδομένα ώστε να κάνουν προσωποποιημένες προτάσεις μοντέλων ηλεκτρικών οχημάτων σε κάθε χρήστη μαζί με τα έξοδα που θα εξοικονομούνται.

Μέσα από την παρούσα ανάλυση περίπτωσης αντικατοπτρίζεται η δύναμη της μηχανικής μάθησης. Ειδικότερα, εξετάστηκε ένα μικρού μεγέθους πρόβλημα μιας εξειδικευμένης βιομηχανίας με αρκετή επιτυχία. Σίγουρα, η μηχανική μάθηση μπορεί να χρησιμοποιηθεί, αν δεν χρησιμοποιείται ήδη, για την αντιμετώπιση παρόμοιων προβλημάτων classification αλλά και διαφορετικών σε άλλες βιομηχανίες. Ήδη υπάρχει τεράστια ποικιλία τεχνικών και μοντέλων μηχανικής μάθησης που μπορούν να χρησιμοποιηθούν εύκολα για διάφορους σκοπούς και προβλήματα. Εν κατακλείδι, οι δυνατότητες εφαρμογής της μηχανικής μάθησης στην σημερινή εποχή είναι αμέτρητες και φαίνεται ότι στο μέλλον η χρήση της είναι μονόδρομος για όλους τους κλάδους.

Βιβλιογραφία

[1] A. E. Mohamed, "Comparative study of four supervised machine learning techniques for classification," *International Journal of Applied*, vol. 7, no. 2, 2017.

[2] A. A. Soofi and A. Awan, "Classification techniques in machine learning: applications and issues," *Journal of Basic and Applied Sciences*, vol. 13, pp. 459–465, 2017.

[3] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[4] O. Maimon and L. Rokach, "Data mining and knowledge discovery handbook," 2005.

[5] R. Sathya, A. Abraham *et al.*, "Comparison of supervised and unsupervised learning algorithms for pattern classification," *International Journal of Advanced Research in Artificial Intelligence*, vol. 2, no. 2, pp. 34–38, 2013.

[6] G. Sharma, G. Rani, V. S. Dhaka *et al.*, "A review on machine learning techniques for prediction of cardiovascular diseases," in *2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)*. IEEE, 2020, pp. 237–242.

[7] R. Katarya *et al.*, "A review: Predicting the performance of students using machine learning classification techniques," in *2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*. IEEE, 2019, pp. 36–41.

[8] P. Meshram, S. Ray *et al.*, "Field-level crop classification using an optimal dataset of multi-temporal sentinel-1 and polarimetric radarsat-2 sar data with machine learning algorithms," *Journal of the Indian Society of Remote Sensing*, vol. 49, no. 12, pp. 2945–2958, 2021.

[9] F. Qian, A. Pathak, Y. C. Hu, Z. M. Mao, and Y. Xie, "A case for unsupervised-learning-based spam filtering," *ACM SIGMETRICS performance evaluation review*, vol. 38, no. 1, pp. 367–368, 2010.

[10] R. Mansoor, N. D. Jayasinghe, and M. M. A. Muslam, "A comprehensive review on email spam classification using machine learning algorithms," in *2021 International Conference on Information Networking (ICOIN)*. IEEE, 2021, pp. 327–332.

[11] D. Johnson, "Unsupervised machine learning: What is, algorithms, example," Dec 2021, accessed: 2021-12-22. [Online]. Available: https://www.guru99.com/unsupervised-machine-learning.html#6

[12] M. Binkhonain and L. Zhao, "A review of machine learning algorithms for identification and classification of non-functional requirements," *Expert Systems with Applications: X*, vol. 1, p. 100001, 2019.

[13] J. Brownlee, "4 types of classification tasks in machine learning," Aug 2020, accessed: 2021-12-24. [Online]. Available: https://machinelearningmastery.com/types-of-classification-in-machine-learning/

[14] R. Kumari and S. K. Srivastava, "Machine learning: A review on binary classification," *International Journal of Computer Applications*, vol. 160, no. 7, 2017.

[15] P. Jeatrakul and K. Wong, "Comparing the performance of different neural networks for binary classification problems," in *2009 Eighth International Symposium on Natural Language Processing*. IEEE, 2009, pp. 111–115.

[16] C. Bellinger, S. Sharma, and N. Japkowicz, "One-class versus binary classification: Which and when?" in *2012 11th International Conference on Machine Learning and Applications*, vol. 2. IEEE, 2012, pp. 102–106.

[17] A. Chatterjee, M. Vallières, and J. Seuntjens, "Overlooked pitfalls in multi-class machine learning classification in radiation oncology and how to avoid them," *Physica Medica*, vol. 70, pp. 96–100, 2020.

[18] V. Panca and Z. Rustam, "Application of machine learning on brain cancer multiclass classification," in *AIP Conference Proceedings*, vol. 1862, no. 1. AIP Publishing LLC, 2017, p. 030133.

[19] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Machine learning*, vol. 85, no. 3, pp. 333–359, 2011.

[20] G. Tsoumakas and I. Katakis, "Multi-label classification," *International Journal of Data Warehousing and Mining*, vol. 3, no. 3, p. 1–13, 2007.

[21] K. Nooney, "Deep dive into multi-label classification..! (with detailed case study)," Jun 2018, accessed: 2021-12-24. [Online]. Available: https://towardsdatascience.com/journey-to-the-center-of-multi-label-classification-384c40229bff

[22] S. Jain, "Multi label classification: Solving multi label classification problems," Dec 2020, accessed: 2022-02-22. [Online]. Available: https://www.analyticsvidhya.com/blog/2017/08/introduction-to-multi-label-classification/

[23] J. Brownlee, "Multi-label classification with deep learning," Aug 2020, accessed: 2022-02-22. [Online]. Available: https://machinelearningmastery.com/multi-label-classification-with-deep-learning/

[24] A. Ishaq, M. Umer, M. F. Mushtaq, C. Medaglia, H. U. R. Siddiqui, A. Mehmood, and G. S. Choi, "Extensive hotel reviews classification using long short term memory," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 10, pp. 9375–9385, 2021.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[26] E. Alpaydin, *Introduction to machine learning*. MIT press, 2020.

[27] A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti, and M. Alazab, "A comprehensive survey for intelligent spam email detection," *IEEE Access*, vol. 7, pp. 168 261–168 295, 2019.

[28] L. Rutkowski, M. Jaworski, L. Pietruczuk, and P. Duda, "Decision trees for mining data streams based on the gaussian approximation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 108–119, 2013.

[29] D. D. Patil, V. Wadhai, and J. Gokhale, "Evaluation of decision tree pruning algorithms for complexity and classification accuracy," *International Journal of Computer Applications*, vol. 11, no. 2, pp. 23–30, 2010.

[30] S. B. Kotsiantis, I. Zaharakis, P. Pintelas *et al.*, "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering*, vol. 160, no. 1, pp. 3–24, 2007.

[31] R. Timofeev, "Classification and regression trees (cart) theory and applications," *Humboldt University, Berlin*, pp. 1–40, 2004.

[32] R. Garg, "7 types of classification algorithms," January 2018, accessed: 2021-12-25. [Online]. Available: https://analyticsindiamag.com/7-types-classification-algorithms/

[33] A. Sihombing and A. C. M. Fong, "Fake review detection on yelp dataset using classification techniques in machine learning," in *2019 International Conference on contemporary Computing and Informatics (IC3I)*. IEEE, 2019, pp. 64–68.

[34] K. M. Al-Aidaroos, A. A. Bakar, and Z. Othman, "Naive bayes variants in classification learning," in *2010 International Conference on Information Retrieval & Knowledge Management (CAMP)*. IEEE, 2010, pp. 276–281.

[35] F. E. Ayo, O. Folorunso, F. T. Ibharalu, and I. A. Osinuga, "Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions," *Computer Science Review*, vol. 38, p. 100311, 2020.

[36] T. Ayer, J. Chhatwal, O. Alagoz, C. E. Kahn Jr, R. W. Woods, and E. S. Burnside, "Comparison of logistic regression and artificial neural network models in breast cancer risk estimation," *Radiographics*, vol. 30, no. 1, pp. 13–22, 2010.

[37] A. Ng, "Cs229 lecture notes," *CS229 Lecture notes*, vol. 1, no. 1, pp. 1–3, 2000.

[38] L. Abhishek, "Optical character recognition using ensemble of svm, mlp and extra trees classifier," in *2020 International Conference for Emerging Technology (INCET)*. IEEE, 2020, pp. 1–4.

[39] V. Saxena and A. Aggarwal, "Comparative study of select non parametric and ensemble machine learning classification techniques," in *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*. IEEE, 2020, pp. 110–115.

[40] C. Désir, C. Petitjean, L. Heutte, M. Salaun, and L. Thiberville, "Classification of endomicroscopic images of the lung based on random subwindows and extra-trees," *IEEE transactions on biomedical engineering*, vol. 59, no. 9, pp. 2677–2683, 2012.

[41] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.

[42] Y. Li and Z. Chen, "Performance evaluation of machine learning methods for breast cancer prediction," *Appl Comput Math*, vol. 7, no. 4, pp. 212–216, 2018.

[43] S. K. Agrawal, "Evaluation metrics for classification model: Classification model metrics," Jul 2021, accessed: 2021-12-29. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/

[44] M. Pushpa and S. Karpagavalli, "Multi-label classification: problem transformation methods in tamil phoneme classification," *Procedia computer science*, vol. 115, pp. 572–579, 2017.

[45] S. Destercke, "Multilabel prediction with probability sets: the hamming loss case," in *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Springer, 2014, pp. 496–505.

[46] P. Liashchynskyi and P. Liashchynskyi, "Grid search, random search, genetic algorithm: a big comparison for nas," *arXiv preprint arXiv:1912.06059*, 2019.

[47] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization." *Journal of machine learning research*, vol. 13, no. 2, 2012.

[48] I. Syarif, A. Prugel-Bennett, and G. Wills, "Svm parameter optimization using grid search and genetic algorithm to improve classification performance," *Telkomnika*, vol. 14, no. 4, p. 1502, 2016.

[49] "Bayesian optimization using gaussian processes," accessed: 2022-02-23. [Online]. Available: https://scikit-optimize.github.io/stable/modules/generated/skopt.gp_minimize.html

[50] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation." *Encyclopedia of database systems*, vol. 5, pp. 532–538, 2009.

[51] D. Berrar, "Cross-validation." 2019.

[52] G. Kovács, "An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets," *Applied Soft Computing*, vol. 83, p. 105662, 2019.

[53] A. Amin, S. Anwar, A. Adnan, M. Nawaz, N. Howard, J. Qadir, A. Hawalah, and A. Hussain, “Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study,” *IEEE Access*, vol. 4, pp. 7940–7957, 2016.

[54] H. Abdi and L. J. Williams, “Principal component analysis,” *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[55] M. Ringnér, “What is principal component analysis?” *Nature biotechnology*, vol. 26, no. 3, pp. 303–304, 2008.