ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

Πολυτεχνική Σχολή
Τμήμα Μηχανικών Η/Υ & Πληροφορικής

Διπλωματική Εργασία

Τεχνικές μηχανικής μάθησης για ταυτοποίηση γεωγραφικής προέλευσης προϊόντων

Ιωάννης Μακαντάσης
Α.Μ. 1054299

Επιβλέπων
Καθηγητής Χρήστος Ι. Μπούρας

Συνεπιβλέπων
Στυλιανός Κουρής, Καθηγητής

Μέλος Επιτροπής Αξιολόγησης
Εύη Παπαϊωάννου, Επίκουρη Καθηγήτρια

Πάτρα, 2023

Πρόλογος

Η διπλωματική εργασία με τίτλο "Τεχνικές Μηχανικής Μάθησης για ταυτοποίηση Γεωγραφικής Προέλευσης Προϊόντων" ανοίγει τον δρόμο για την εξερεύνηση ενός σημαντικού πεδίου της επιστήμης της μηχανικής μάθησης. Η ταυτοποίηση της γεωγραφικής προέλευσης προϊόντων αποτελεί ένα σημαντικό θέμα στον τομέα του εμπορίου και της αλυσίδας εφοδιασμού, καθώς επιτρέπει στους καταναλωτές να γνωρίζουν περισσότερα για την προέλευση και την ποιότητα των προϊόντων που αγοράζουν.

Η τεχνητή νοημοσύνη και η μηχανική μάθηση έχουν αναδείξει νέες δυνατότητες για την αντιμετώπιση αυτής της πρόκλησης. Χρησιμοποιώντας τεχνικές μηχανικής μάθησης, μπορούμε να αναπτύξουμε μοντέλα πρόβλεψης που βασίζονται σε δεδομένα, όπως πληροφορίες περί γεωγραφικής προέλευσης, χαρακτηριστικά προϊόντων, και ποιοτικές παραμέτρους. Τέτοια μοντέλα μπορούν να χρησιμοποιηθούν για την αυξημένη διαφάνεια της αλυσίδας εφοδιασμού, την εξάλειψη της απάτης, και τη βελτίωση της ποιότητας των προϊόντων.

Στην πορεία αυτής της διπλωματικής εργασίας, θα εξετάσουμε τις διάφορες τεχνικές μηχανικής μάθησης που μπορούν να εφαρμοστούν για την ταυτοποίηση της γεωγραφικής προέλευσης προϊόντων. Θα διερευνήσουμε τη χρήση αλγορίθμων μηχανικής μάθησης, όπως τα νευρωνικά δίκτυα, τα δέντρα αποφάσεων, και οι μέθοδοι μάθησης με επίβλεψη και χωρίς επίβλεψη.

Επίσης, θα εξετάσουμε τα πλεονεκτήματα και τις προκλήσεις που προκύπτουν κατά την εφαρμογή αυτών των τεχνικών στον πραγματικό κόσμο. Θα διερευνήσουμε πώς οι εταιρίες και οι καταναλωτές μπορούν να επωφεληθούν από αυτές τις τεχνικές, ενισχύοντας την εμπιστοσύνη στα προϊόντα και την προστασία των καταναλωτών.

Αυτή η διπλωματική εργασία ανοίγει τον δρόμο για περαιτέρω έρευνα και ανάπτυξη στον τομέα της ταυτοποίησης γεωγραφικής προέλευσης προϊόντων μέσω της μηχανικής μάθησης. Αναμένουμε ότι αυτή η εργασία θα προσφέρει πολύτιμες πληροφορίες και θα ενθαρρύνει την κοινότητα της επιστήμης των δεδομένων και της μηχανικής μάθησης να συνεχίσει την έρευνα σε αυτόν τον σημαντικό τομέα.

Περίληψη

Η παρούσα διπλωματική εργασία εξετάζει τον ρόλο της μηχανικής μάθησης στην ανάπτυξη μοντέλων πρόβλεψης για τη γεωγραφική προέλευση προϊόντων. Αυτή η έρευνα αποκαλύπτει τον τρόπο με τον οποίο η μηχανική μάθηση μπορεί να χρησιμοποιηθεί για την αύξηση της διαφάνειας στην αλυσίδα εφοδιασμού, την ανίχνευση απάτης και τη βελτίωση της ποιότητας των προϊόντων. Η εργασία εξετάζει διάφορες τεχνικές μηχανικής μάθησης, όπως νευρωνικά δίκτυα και δέντρα αποφάσεων, καθώς και προκλήσεις και πλεονεκτήματα που προκύπτουν από την εφαρμογή αυτών των τεχνικών. Η διπλωματική εργασία ανοίγει τον δρόμο για περαιτέρω έρευνα στον τομέα της ταυτοποίησης γεωγραφικής προέλευσης προϊόντων με χρήση της μηχανικής μάθησης.

Abstract

This thesis examines the role of machine learning in the development of predictive models for the geographic origin of products. This research reveals how machine learning can be used to increase transparency in the supply chain, detect fraud and improve product quality. The paper examines various machine learning techniques, such as neural networks and decision trees, as well as challenges and advantages arising from the application of these techniques. The thesis paves the way for further research in the area of identifying the geographical origin of products using machine learning.

Περιεχόμενα

Λίστα Εικόνων

Λίστα Πινάκων

1

Εισαγωγή

1.1 Σημασία του προβλήματος

Η σύγχρονη αγορά τροφίμων έχει αναπτυχθεί σε ένα παγκόσμιο, πολύπλοκο σύστημα με πολλών ειδών προϊόντα που διακινούνται με αυξανόμενο ρυθμό. Αυτή η ανάπτυξη της βιομηχανίας τροφίμων υπογραμμίζει την ανάγκη για ακριβή αναγνώριση και πιστοποίηση των προϊόντων βάσει της γεωγραφικής τους προέλευσης, του είδους, της ποικιλίας και των ιδιαίτερων χαρακτηριστικών τους. Ένα από τα πιο διαδεδομένα προϊόντα στην αγορά τροφίμων το οποίο μάλιστα, είναι υψίστης σημασίας για την οικονομία των μεσογειακών χωρών, αποτελεί το αγνό παρθένο ελαιόλαδο. Αναλόγως την χώρα παραγωγής, το ελαιόλαδο εμφανίζει διαφορετικές ποιοτικές ιδιαιτερότητες, γευστικές, αρωματικές και ποιοτικές (π.χ. συγκέντρωση σε β-καροτένια, αντιοξειδωτικά, λιπαρά οξέα κλπ.), τις οποίες ο καταναλωτής μπορεί να διακρίνει ώστε να επιλέξει το προϊόν που επιθυμεί να αγοράσει. Έτσι, οι καταναλωτές ενδιαφέρονται ολοένα και περισσότερο για τη γεωγραφική προέλευση των τροφίμων τους, θεωρώντας την ως ένδειξη ποιότητας και αυθεντικότητας. Επομένως, η δυνατότητα αξιόπιστης και γρήγορης επαλήθευσης της γεωγραφικής προέλευσης προϊόντων όπως το ελαιόλαδο έχει σημαντικά οφέλη τόσο για τους κατασκευαστές, τους εμπόρους όσο και τους καταναλωτές.

Η γεωγραφική προέλευση ενός προϊόντος διατροφής όπως το ελαιόλαδο μπορεί να επηρεάσει δραστικά τις αισθητηριακές και θρεπτικές του ιδιότητες, οι οποίες επηρεάζονται κατά κύριο λόγο από τη φύση του εδάφους, τις κλιματικές συνθήκες και τις συγκεκριμένες ποικιλίες ελιάς. Έτσι, η γεωγραφική προέλευση μεταφράζεται σε μια μοναδική χημική σύνθεση και αισθητηριακό προφίλ που μπορεί να χρησιμεύσει ως ένα είδος «βιολογικού διαβατηρίου». Αυτή η εγγενής σχέση μεταξύ της καταγωγής και της ποιότητας των προϊόντων αποτελεί τη βάση για τις ονομασίες προστατευόμενης προέλευσης και τις γεωγραφικές ενδείξεις, οι οποίες είναι νομικοί μηχανισμοί που έχουν θεσπιστεί για την πιστοποίηση της προέλευσης των γεωργικών προϊόντων.

Ωστόσο, η αύξηση της ζήτησης για προϊόντα με πιστοποιημένη προέλευση συνοδεύτηκε από αύξηση των περιπτώσεων απάτης και νοθείας. Το οικονομικό κίνητρο για νόθευση και απάτη είναι μεγάλο αφού ένα προϊόν υψηλής ποιότητας προέλευσης μπορεί να έχει σημαντικά υψηλότερη τιμή από ένα κοινό ή νοθευμένο ισοδύναμο. Ως εκ τούτου, η διασφάλιση της γνησιότητας αυτών των προϊόντων δεν είναι μόνο θέμα προστασίας των καταναλωτών αλλά και ζήτημα θεμιτού ανταγωνισμού και οικονομικής δικαιοσύνης.

Σε αυτό το πλαίσιο, νέες μεθοδολογίες για ακριβή και γρήγορο έλεγχο της γνησιότητας και της προέλευσης των τροφίμων έχουν μεγάλη ζήτηση. Η φασματοσκοπία διάσπασης που επάγεται από λέιζερ (Laser-induced breakdown spectroscopy- LIBS) ξεχωρίζει ως ένα πολλά υποσχόμενο εργαλείο για την αντιμετώπιση αυτών των αναγκών. Η φασματοσκοπία διάσπασης που προκαλείται από λέιζερ (LIBS) είναι μια ισχυρή αναλυτική τεχνική που χρησιμοποιείται για τον προσδιορισμό της χημικής σύνθεσης ενός ευρέος φάσματος υλικών. Η τεχνική βασίζεται στη χρήση ενός λέιζερ υψηλής ισχύος για την εξάτμιση ενός μικρού τμήματος ενός δείγματος, δημιουργώντας ένα πλάσμα. Στη συνέχεια, το φως που εκπέμπεται από το πλάσμα αναλύεται για να προσδιοριστεί η σύνθεση του δείγματος. Το LIBS είναι μια γρήγορη και εξαιρετικά ευαίσθητη τεχνική που είναι σε θέση να παρέχει σε πραγματικό χρόνο, επιτόπια ανάλυση ενός ευρέος φάσματος υλικών, καθιστώντας το πολύτιμο εργαλείο για μια ποικιλία εφαρμογών.

Αυτό είναι όπου η μηχανική μάθηση μπαίνει στην εικόνα. Οι αλγόριθμοι μηχανικής μάθησης έχουν τη δυνατότητα να ξεδιαλύνουν μοτίβα και συσχετισμούς εντός των φασμάτων LIBS που είναι πέρα από την εμβέλεια των συμβατικών στατιστικών προσεγγίσεων. Διαφορετικοί τύποι αλγορίθμων μηχανικής μάθησης, που κυμαίνονται από μηχανές διανυσμάτων υποστήριξης έως k-πλησιέστερους γείτονες (k-NN) και νευρωνικά δίκτυα, έχουν χρησιμοποιηθεί για αυτήν την διπλωματική εργασία, παρουσιάζοντας πολλά υποσχόμενα αποτελέσματα.

Η υπόθεση αυτής της διπλωματικής εργασίας βρίσκεται στο σημείο τομής αυτών των τεχνολογικών προόδων στη φασματοσκοπία και τη μηχανική μάθηση. Στόχος μας είναι να εξερευνήσουμε, να επιλέξουμε και να εφαρμόσουμε κατάλληλους αλγόριθμους μηχανικής μάθησης για την αξιοποίηση των φασματοσκοπικών δεδομένων που παρέχει η LIBS, διασφαλίζοντας έτσι τον ακριβή και αποτελεσματικό προσδιορισμό της γεωγραφικής προέλευσης των προϊόντων διατροφής. Αυτή η εφαρμογή της μηχανικής εκμάθησης στα δεδομένα LIBS για έλεγχο ταυτότητας τροφίμων υπόσχεται την παροχή ενός επιπλέον επιπέδου διασφάλισης στους καταναλωτές, τις ρυθμιστικές αρχές και τη βιομηχανία, την ενίσχυση των εμπορικών σημάτων και την ενίσχυση της αξίας του προϊόντος.

1.2 Στόχοι της Εργασίας

Ο πρωταρχικός στόχος αυτής της διπλωματικής εργασίας είναι να αξιοποιήσει τη δύναμη της μηχανικής μάθησης για να ενισχύσει τις δυνατότητες της φασματοσκοπίας διάσπασης που προκαλείται από λέιζερ (LIBS) στον προσδιορισμό της γεωγραφικής προέλευσης των προϊόντων διατροφής, με ιδιαίτερη έμφαση στο ελαιόλαδο. Στόχος της μελέτης είναι να δημιουργηθεί μια μεθοδολογία που χρησιμοποιεί τεχνικές μηχανικής εκμάθησης για την αξιόπιστη, γρήγορη και ακριβή επαλήθευση της προέλευσης του προϊόντος σε πραγματικό χρόνο.

Η εργασία χωρίζεται σε δύο βασικά μέρη. Το πρώτο μέρος αφορά την εύρεση και την εφαρμογή κατάλληλων αλγορίθμων μηχανικής μάθησης για την ανάλυση των φασματοσκοπικών δεδομένων που λαμβάνονται μέσω της τεχνικής LIBS. Αυτό συνεπάγεται μια ολοκληρωμένη μελέτη τόσο μέσω των εποπτευόμενων όσο και των μη εποπτευόμενων αλγορίθμων εκμάθησης, συμπεριλαμβανομένων, ενδεικτικά, των μηχανών υποστήριξης διανυσμάτων (Support Vector Machines-SVMs), των αλγορίθμων k-Nearest Neighbors (k-NN) και των μεθόδων που βασίζονται σε νευρωνικά δίκτυα (Neural Networks-NNs). Ένα βασικό μέρος αυτής της διαδικασίας περιλαμβάνει την μελέτη της αποτελεσματικότητας των αλγορίθμων αυτών σε σχέση με την φύση των δεδομένων που θα αναλυθούν στην παρούσα εργασία. Αυτή η διαδικασία θα περιλαμβάνει επίσης την εξέταση άλλων στρατηγικών μηχανικής μάθησης όπως οι μέθοδοι προεπεξεργασίας των δεδομένων (preprocessing methods), οι οποίες περιλαμβάνουν την ομαδοποίηση και την μείωση των διαστάσεων των δεδομένων με στόχο την βελτίωση της απόδοσης των επιλεγμένων αλγορίθμων.

Το δεύτερο μέρος, εφόσον έχουν επιλεγεί κατάλληλα οι αλγόριθμοι μηχανικής μάθησης, στοχεύει στην εκπαίδευση των μοντέλων πρόβλεψης και την βελτιστοποίησή τους, εστιάζοντας στην επίτευξη όσο το δυνατόν υψηλότερων ποσοστών επιτυχούς πρόβλεψης της γεωγραφικής προέλευσης. Η εκπαίδευση και η βελτιστοποίηση του μοντέλου περιλαμβάνουν την βελτιστοποίηση των υπερπαραμέτρων (hyperparameters tuning), την ελαχιστοποίηση της υπερπροσαρμογής των μοντέλων (overfitting) και την διερεύνηση διαφόρων μεθόδων για τη βελτίωση της ακρίβειας (αλλαγές στα kernels, degrees, estimators, max_depths) έτσι ώστε το μοντέλο πρόβλεψης να μπορεί να γενικευτεί στα πειραματικά δεδομένα. Η απόδοση αυτών των εκπαιδευμένων μοντέλων θα αξιολογηθεί και θα συγκριθεί με τις υπάρχουσες μεθοδολογίες για να επικυρωθεί η αποτελεσματικότητά τους και να αναδειχθούν τα πιθανά οφέλη της προσέγγισής μας.

1.3 Συνεισφορά

Στην παρούσα διατριβή, επιδιώκεται η παρουσίαση μιας ολοκληρωμένης κατανόησης του συνεργατικού δυναμικού του LIBS και της μηχανικής μάθησης για την ενίσχυση των διαδικασιών επαλήθευσης της γνησιότητας των τροφίμων. Πραγματοποιείται προσπάθεια συνεισφοράς πολύτιμων γνώσεων σε αυτόν τον ταχέως εξελισσόμενο τομέα και προτείνονται πρακτικές λύσεις για την ενίσχυση του ελέγχου της γνησιότητας των προϊόντων και της πιστοποίησης γεωγραφικής προέλευσης.

1.4 Διάρθρωση της Διπλωματικής Εργασίας

Στο **Κεφάλαιο 2** πραγματοποιείται ανασκόπηση βιβλιογραφίας που παρουσιάζεται μια εις βάθος ανασκόπηση της υπάρχουσας βιβλιογραφίας σχετικά με την αυθεντικότητα των τροφίμων και τη γεωγραφική προέλευση, τη φασματοσκοπία διάσπασης που προκαλείται από λέιζερ (LIBS) και την εφαρμογή της μηχανικής μάθησης στη φασματοσκοπική ανάλυση. Οι τρέχουσες προκλήσεις και τα κενά στον τομέα που στοχεύει να αντιμετωπίσει η μελέτη μας παρουσιάζονται επίσης σε αυτό το κεφάλαιο.

Στην συνέχεια, στο **Κεφάλαιο 3** περιγράφεται το μεθοδολογικό πλαίσιο της έρευνας. Αρχικά, αναλύονται οι διαφορετικοί τύποι μηχανικής εκμάθησης συνοδευόμενοι από την περιγραφή των μοντέλων που αξιοποιούνται στην παρούσα διπλωματική εργασία. Επιπρόσθετα, πραγματοποιείται επεξήγηση της διαδικασίας αξιολόγησης των αλγορίθμων.

Στο **Κεφάλαιο 4** περιγράφεται λεπτομερώς η τεχνική LIBS. Σε αυτό το κεφάλαιο παρουσιάζεται η λειτουργία καθώς και οι βασικές αρχές της συγκεκριμένης τεχνικής, καθώς και ορισμένες προκλήσεις στην ανάλυση δεδομένων LIBS. Επίσης, αναλύεται η διαδικασία ενοποίησης του LIBS με την μηχανική εκμάθηση ενώ περιγράφονται περιπτώσεις μελετών της τεχνικής LIBS.

Στο **Κεφάλαιο 5**, πραγματοποιείται επισκόπηση του συνόλου των δεδομένων καθώς και η αντιστοίχιση του ονόματος γεωγραφικής προέλευσης του λαδιού σε αριθμητική τιμή. Στην συνέχεια, παρουσιάζονται τα μοντέλα που αξιοποιήθηκαν για την ανάλυση καθώς και οι μέθοδοι αξιολόγησής τους. Τέλος, πραγματοποιείται η επεξήγηση των αποτελεσμάτων συνοδευόμενη από λεπτομερή συζήτηση των αποτελεσμάτων.

Το **Κεφάλαιο 6** περιλαμβάνει σημαντικά συμπεράσματα και προτάσεις για μελλοντική έρευνα. Στο παρόν κεφάλαιο ολοκληρώνεται η μελέτη με περιληπτική ανάλυση των ευρημάτων, τα αποτελέσματα της μελέτης και συστάσεις για μελλοντική έρευνα στο πεδίο. Σε αυτό το κεφάλαιο εξετάζονται επίσης

οι περιορισμούς της μελέτης και προτείνονται πιθανές μελλοντικές κατευθύνσεις που θα μπορούσαν να βασιστούν σε αυτήν την έρευνα.

2

Ανασκόπηση βιβλιογραφίας

2.1 Αυθεντικότητα τροφίμων και γεωγραφική προέλευση

Η αυθεντικότητα των τροφίμων και η πιστοποίηση γεωγραφικής προέλευσης έχουν γίνει σημαντικά θέματα στην επιστήμη των τροφίμων λόγω της αυξανόμενης ζήτησης για διαφάνεια στην παραγωγή τροφίμων (Smith, 2016). Οι καταναλωτές σήμερα δεν ενδιαφέρονται μόνο για το θρεπτικό περιεχόμενο των τροφίμων τους αλλά και για τη γεωγραφική τους προέλευση, θεωρώντας το ως ένδειξη ποιότητας και γνησιότητας (Charlebois et al., 2016).

Η γεωγραφική προέλευση ενός τρόφιμου μπορεί να επηρεάσει δραστικά τις αισθητηριακές και θρεπτικές του ιδιότητες, οι οποίες καθορίζονται από διάφορους παράγοντες, συμπεριλαμβανομένων των εδαφολογικών συνθηκών, του κλίματος και συγκεκριμένων καλλιεργητικών πρακτικών (Bellassen & Giraud, 2013). Αυτά τα χαρακτηριστικά σχηματίζουν ένα μοναδικό χημικό και αισθητηριακό προφίλ για κάθε προϊόν διατροφής, που συχνά αναφέρεται ως «terroir» του προϊόντος (Teixeira et al., 2018).

Η συσχέτιση μεταξύ γεωγραφικής προέλευσης και ποιότητας των προϊόντων διατροφής οδήγησε στην καθιέρωση ετικετών όπως η Προστατευόμενη Ονομασία Προέλευσης (ΠΟΠ) και η Προστατευόμενη Γεωγραφική Ένδειξη (ΠΓΕ). Αυτές οι ετικέτες πιστοποιούν την προέλευση των γεωργικών προϊόντων και συμβάλλουν στη διατήρηση της ποιότητας, της φήμης και της αυθεντικότητας αυτών των προϊόντων (Barham, 2003).

Ωστόσο, η αυξανόμενη ζήτηση για προϊόντα με πιστοποίηση προέλευσης έχει οδηγήσει σε αυξημένες περιπτώσεις απάτης στα τρόφιμα (Spink & Moyer, 2011). Η νοθεία και η εσφαλμένη επισήμανση των προϊόντων διατροφής για οικονομικό όφελος θέτουν σοβαρές προκλήσεις για την αυθεντικότητα και την ασφάλεια των τροφίμων (Herman & Conti, 2013).

Έχουν χρησιμοποιηθεί διάφορες τεχνικές για τον έλεγχο της γνησιότητας των τροφίμων και την πιστοποίηση γεωγραφικής προέλευσης, συμπεριλαμβανομένης της ανάλυσης σταθερού λόγου

ισοτόπων (Ruth & Brunton, 2007), της γραμμικής κωδικοποίησης DNA (Galimberti et al., 2013) και της τεχνολογίας αισθητήρων (Zhang et al, 2014).

Μεταξύ αυτών, το Laser-Induced Breakdown Spectroscopy (LIBS) έχει αναδειχθεί ως ένα πολλά υποσχόμενο εργαλείο για ταχεία και ακριβή δοκιμή γνησιότητας τροφίμων (Cremers & Radziemski, 2006). Το LIBS χρησιμοποιεί έναν παλμό λέιζερ υψηλής ενέργειας για να αφαιρέσει την επιφάνεια ενός δείγματος και να δημιουργήσει ένα πλάσμα. Το προκύπτον φάσμα ατομικών εκπομπών μπορεί να χρησιμοποιηθεί για τον προσδιορισμό της στοιχειακής σύνθεσης του δείγματος και έτσι να παρέχει μια ένδειξη της γεωγραφικής προέλευσής του (Gunduz et al., 2017).

Η χρήση αλγορίθμων μηχανικής μάθησης σε συνδυασμό με το LIBS είναι μια πολλά υποσχόμενη στρατηγική προσέγγιση για αξιόπιστο, γρήγορο και ασφαλή έλεγχο της γεωγραφικής προέλευσης των προϊόντων διατροφής (El-Abassy et al., 2019).

Συμπερασματικά, η διατήρηση της γνησιότητας των τροφίμων και η επαλήθευση της γεωγραφικής προέλευσης των προϊόντων διατροφής είναι απαραίτητα στις τρέχουσες οικονομικές συνθήκες. Η πρόοδος σε τεχνολογίες όπως το LIBS και η μηχανική μάθηση παρέχουν ισχυρά εργαλεία για την αντιμετώπιση αυτών των ζητημάτων (Santos et al., 2020).

2.2 Φασματοσκοπία διάσπασης που προκαλείται από λέιζερ (LIBS)

Το Laser-Induced Breakdown Spectroscopy (LIBS) είναι μια τεχνική φασματοσκοπίας ατομικής εκπομπής που χρησιμοποιείται όλο και περισσότερο για ανάλυση υλικών σε διάφορους τομείς, συμπεριλαμβανομένης της γεωλογίας, της μεταλλουργίας και πιο πρόσφατα, της επιστήμης των τροφίμων (Cremers & Radziemski, 2006). Η τεχνική είναι γνωστή για την ταχύτητα, την ευκολία, την ευελιξία και την ικανότητά της να αναλύει κάθε είδους δείγμα, στερεό, υγρό ή αέριο, με ελάχιστη προετοιμασία (Hahn & Omenetto, 2010).

Το LIBS λειτουργεί με βάση την αρχή της αφαίρεσης με λέιζερ. Ένας παλμός λέιζερ υψηλής ενέργειας εστιάζεται στην επιφάνεια του δείγματος, οδηγώντας σε τοπική θέρμανση και εξάτμιση μιας μικρής περιοχής του δείγματος. Αυτή η εξάτμιση δημιουργεί ένα νέφος πλάσματος που αποτελείται από άτομα, ιόντα και ηλεκτρόνια. Καθώς το πλάσμα ψύχεται, τα άτομα και τα ιόντα στο πλάσμα εκπέμπουν φωτόνια, παράγοντας ένα φάσμα εκπομπής φωτός που είναι χαρακτηριστικό των στοιχείων που υπάρχουν στο δείγμα (Schieber, et al., 2019).

Ένα από τα πιο σημαντικά πλεονεκτήματα του LIBS είναι η δυνατότητα του να διεξάγει επιτόπια ανάλυση σε πραγματικό χρόνο χωρίς να απαιτείται προετοιμασία δείγματος. Αυτό είναι ιδιαίτερα χρήσιμο στη βιομηχανία τροφίμων όπου οι γρήγορες και μη καταστροφικές δοκιμές είναι ζωτικής

σημασίας (Gondal et al., 2006). Επιπλέον, το LIBS μπορεί να ανιχνεύσει όλα τα στοιχεία, συμπεριλαμβανομένων των ελαφρών στοιχείων όπως ο άνθρακας, το άζωτο και το οξυγόνο, τα οποία αποτελούν πρόκληση για άλλες φασματοσκοπικές τεχνικές μέτρησης (Anzano, 2006).

Ωστόσο, παρά τα πολλά πλεονεκτήματά του, το LIBS έρχεται με προκλήσεις. Τα φάσματα LIBS είναι συχνά πολύπλοκα και δύσκολα ερμηνεύσιμα λόγω του πλήθους των φασματικών γραμμών και της παρουσίας μοριακών και ατομικών ειδών, μαζί με τις επιδράσεις των παραμέτρων του πλάσματος στα σχήματα και τις εντάσεις των φασματικών γραμμών (Bilge et al., 2017). Αυτή η πολυπλοκότητα απαιτεί τη χρήση προηγμένων αναλυτικών τεχνικών για ακριβή ερμηνεία και ποσοτικοποίηση των στοιχειακών συγκεντρώσεων (Motlagh et al., 2017).

Μια πολλά υποσχόμενη προσέγγιση για την αντιμετώπιση αυτών των προκλήσεων είναι η χρήση αλγορίθμων μηχανικής μάθησης. Η μηχανική μάθηση, με την δυνατότητά της να χειρίζεται πολυμεταβλητά δεδομένα και την ικανότητά της για αναγνώριση προτύπων, είναι κατάλληλη για την ερμηνεία των φασμάτων LIBS. Διαφορετικοί τύποι αλγορίθμων μηχανικής μάθησης, που κυμαίνονται από μηχανές υποστήριξης διανυσμάτων έως νευρωνικά δίκτυα, έχουν δείξει πολλά υποσχόμενα για τη βελτίωση της ακρίβειας και της ακρίβειας της ανάλυσης LIBS (Chen et al., 2020).

Η ενοποίηση του LIBS με τη μηχανική μάθηση ανοίγει νέους δρόμους για ταχεία, επιτόπια ανάλυση και σε πραγματικό χρόνο ανάλυση της αυθεντικότητας των τροφίμων και του προσδιορισμού της γεωγραφικής προέλευσης. Για παράδειγμα, μια μελέτη των El-Abassy et al. (2019) απέδειξε την επιτυχημένη εφαρμογή του LIBS σε συνδυασμό με αλγόριθμους μηχανικής μάθησης για τον γρήγορο εντοπισμό και αναγνώριση πλαστού ελαιόλαδου.

Συμπερασματικά, το LIBS, με τα μοναδικά του πλεονεκτήματα και τις δυνατότητες ενσωμάτωσης με προηγμένες τεχνικές ανάλυσης δεδομένων, όπως η μηχανική μάθηση, μπορεί να αποτελέσει ένα ισχυρό εργαλείο για την επαλήθευση της γνησιότητας των τροφίμων και της γεωγραφικής προέλευσης. Περαιτέρω εξελίξεις σε αυτόν τον τομέα μπορούν να φέρουν επανάσταση στους ελέγχους γνησιότητας των τροφίμων, παρέχοντας αξιόπιστες, γρήγορες και μη καταστροφικές μεθόδους για τη διασφάλιση της ακεραιότητας των προϊόντων διατροφής και την προώθηση πρακτικών δίκαιου εμπορίου.

Ωστόσο, η ερμηνεία των φασμάτων LIBS είναι πολύπλοκη λόγω της πολυμεταβλητής φύσης των δεδομένων, που απαιτούν προηγμένες τεχνικές ανάλυσης δεδομένων (De Giacomo et al., 2016). Εδώ μπαίνει η μηχανική μάθηση. Η μηχανική μάθηση, ένα υποσύνολο της τεχνητής νοημοσύνης, έχει τη δυνατότητα να εξάγει σημαντικά μοτίβα και συσχετισμούς από πολύπλοκα σύνολα δεδομένων (Bishop, 2006).

Αλγόριθμοι μηχανικής μάθησης όπως μηχανές υποστήριξης διανυσμάτων, k-Nearest Neighbors και νευρωνικά δίκτυα έχουν εφαρμοστεί με επιτυχία σε διάφορους τομείς, συμπεριλαμβανομένου του

ελέγχου αυθεντικότητας τροφίμων (Borràs et al., 2016)[15]. Αυτοί οι αλγόριθμοι μπορούν να χειριστούν τη σύνθετη και πολυμεταβλητή φύση των δεδομένων LIBS, παρέχοντας ισχυρά μοντέλα για δοκιμές γνησιότητας τροφίμων (Chen et al., 2020).

2.3 Μηχανική Μάθηση στη Φασματοσκοπική Ανάλυση

Η μηχανική μάθηση (ML), ένα υποσύνολο της τεχνητής νοημοσύνης (AI), είναι ένα ισχυρό υπολογιστικό εργαλείο γνωστό για τις δυνατότητες αναγνώρισης προτύπων και την δυνατότητα του να χειρίζεται και να ερμηνεύει πολύπλοκα πολυμεταβλητά δεδομένα (Bishop, 2006). Τα τελευταία χρόνια, η μηχανική μάθηση χρησιμοποιείται όλο και περισσότερο για την ανάλυση φασματοσκοπικών δεδομένων, συμπεριλαμβανομένης της φασματοσκοπίας διάσπασης που προκαλείται από λέιζερ (LIBS) (Chen et al., 2020).

Η ενσωμάτωση της μηχανικής μάθησης με τη φασματοσκοπία επιτρέπει την εξαγωγή πολύτιμων πληροφοριών από πολύπλοκα και θορυβώδη φάσματα, βελτιώνοντας έτσι την ακρίβεια και την ακρίβεια της ανάλυσης (Goodacre, 2004). Αυτή η συνεργατική προσέγγιση είναι ιδιαίτερα σημαντική στο πλαίσιο της επαλήθευσης της γνησιότητας των τροφίμων και της γεωγραφικής προέλευσης, όπου η ερμηνεία των φασμάτων LIBS απαιτεί προηγμένες αναλυτικές τεχνικές (El-Abassy et al., 2019).

Έχουν διερευνηθεί διάφοροι αλγόριθμοι μηχανικής μάθησης για φασματοσκοπική ανάλυση δεδομένων, ο καθένας με τα δυνατά του σημεία και τους περιορισμούς του. Για παράδειγμα, οι μηχανές διανυσμάτων υποστήριξης (SVM) έχουν δείξει πολλά υποσχόμενες εργασίες ταξινόμησης, διακρίνοντας μεταξύ διαφορετικών κατηγοριών που βασίζονται σε ένα υπερεπίπεδο σε ένα χώρο χαρακτηριστικών υψηλών διαστάσεων (Boser et al., 1992). Ομοίως, ο αλγόριθμος k-Nearest Neighbors (k-NN), ο οποίος ταξινομεί ένα δείγμα με βάση την πλειοψηφία των k πλησιέστερων γειτόνων του στο χώρο χαρακτηριστικών, έχει επίσης εφαρμοστεί στη φασματοσκοπική ανάλυση (Cover & Hart, 1967).

Τα νευρωνικά δίκτυα, ειδικά τα μοντέλα βαθιάς μάθησης, έχουν επιδείξει τις δυνατότητές τους στο χειρισμό πολύπλοκων, μη γραμμικών δεδομένων (Schmidhuber, 2015). Τα συνελικτικά νευρωνικά δίκτυα (CNN), ένας τύπος μοντέλου βαθιάς μάθησης που έχει σχεδιαστεί για να μαθαίνει αυτόματα και προσαρμοστικά χωρικές ιεραρχίες χαρακτηριστικών, έχουν δείξει πολλά υποσχόμενα αποτελέσματα στην ανάλυση φασματικών δεδομένων (Krizhevsky et al., 2012).

Εκτός από αυτά, τεχνικές μείωσης διαστάσεων, όπως η ανάλυση κύριου συστατικού (PCA-Principal component analysis) και η ενσωμάτωση στοχαστικού γείτονα (t-SNE) έχουν χρησιμοποιηθεί για την απλοποίηση σύνθετων φασματικών δεδομένων, καθιστώντας ευκολότερη την οπτική απεικόνιση και την ερμηνεία (Van Der Maaten & Hinton, 2008).

Η εφαρμογή της μηχανικής μάθησης στη φασματοσκοπική ανάλυση δεν είναι χωρίς προκλήσεις. Η υπερπροσαρμογή, όπου το μοντέλο αποδίδει καλά στα δεδομένα εκπαίδευσης αλλά κακώς σε μη ορατά δεδομένα, είναι ένα σύνηθες πρόβλημα στη μηχανική μάθηση (Hawkins et al., 2004). Οι τεχνικές τακτοποίησης και οι στρατηγικές διασταυρούμενης επικύρωσης (cross-validation) χρησιμοποιούνται συχνά για τον μετριασμό της υπερπροσαρμογής και την ενίσχυση της ικανότητας γενίκευσης του μοντέλου (Rifkin & Klautau, 2004).

Παρά αυτές τις προκλήσεις, η δυνατότητα της μηχανικής μάθησης στη φασματοσκοπική ανάλυση είναι αναμφισβήτητη. Όπως φαίνεται από διάφορες μελέτες, η μηχανική μάθηση, όταν ενσωματώνεται με το LIBS, μπορεί να παρέχει ένα ισχυρό και αποτελεσματικό εργαλείο για την επαλήθευση της γνησιότητας των τροφίμων και της γεωγραφικής προέλευσης (Borràs et al., 2016). Αυτή η ολοκληρωμένη προσέγγιση υπόσχεται να φέρει επανάσταση στις δοκιμές τροφίμων, προσφέροντας μια γρήγορη, μη καταστροφική και αξιόπιστη λύση για τη διασφάλιση της ακεραιότητας της αλυσίδας εφοδιασμού τροφίμων μας.

2.4 Τρέχουσες προκλήσεις και κενά στο πεδίο

Παρά τις σημαντικές προόδους στη χρήση της φασματοσκοπίας διάσπασης που προκαλείται από λέιζερ (LIBS) σε συνδυασμό με τη μηχανική εκμάθηση για την αυθεντικότητα των τροφίμων και την επαλήθευση γεωγραφικής προέλευσης, υπάρχουν αρκετές προκλήσεις και κενά στο πεδίο.

Πρώτον, μια κρίσιμη πρόκληση έγκειται στην εγγενή πολυπλοκότητα των φασμάτων LIBS. Το πλήθος των φασματικών γραμμών, μαζί με την παρουσία τόσο μοριακών όσο και ατομικών ειδών, μπορεί να περιπλέξει την ερμηνεία των δεδομένων (Bilge et al., 2017). Επιπλέον, η μεταβλητότητα στα φασματικά σήματα λόγω αλλαγών στις συνθήκες του πλάσματος, όπως η θερμοκρασία και η πίεση, μπορεί επίσης να επηρεάσει τα σχήματα και τις εντάσεις των φασματικών γραμμών, θέτοντας πρόσθετες προκλήσεις για ανάλυση (Harilal et al., 2019).

Για να ξεπεραστεί αυτό, υπάρχει μια αυξανόμενη ανάγκη για ισχυρούς αλγόριθμους μηχανικής μάθησης ικανούς να χειρίζονται αποτελεσματικά την πολυπλοκότητα και τη μεταβλητότητα των φασμάτων LIBS. Ενώ έχουν εφαρμοστεί αρκετοί αλγόριθμοι μηχανικής μάθησης σε αυτό το πλαίσιο, υπάρχει ακόμα περιθώριο βελτίωσης όσον αφορά την ακρίβεια της ανάλυσης (Chen et al., 2020).

Μια άλλη πρόκληση έγκειται στην έλλειψη μεγάλων, υψηλής ποιότητας συνόλων δεδομένων για την εκπαίδευση μοντέλων μηχανικής εκμάθησης. Για να λειτουργήσουν αποτελεσματικά οι αλγόριθμοι μηχανικής μάθησης, απαιτούν σημαντικές ποσότητες δεδομένων για εκπαίδευση, επικύρωση και δοκιμή (Bishop, 2006). Ωστόσο, η συγκέντρωση αυτών των μεγάλων συνόλων δεδομένων για

προϊόντα τροφίμων είναι συχνά δύσκολη λόγω υλικοτεχνικών θεμάτων, ρυθμιστικών περιορισμών και κόστους.

Επιπλέον, το ζήτημα της υπερπροσαρμογής είναι μια κοινή πρόκληση στη μηχανική μάθηση, ειδικά όταν έχουμε να κάνουμε με περιορισμένα ή θορυβώδη δεδομένα (Hawkins et al., 2004). Η υπερπροσαρμογή συμβαίνει όταν ένα μοντέλο μαθαίνει τον θόρυβο στα δεδομένα εκπαίδευσης και όχι το υποκείμενο μοτίβο, οδηγώντας σε κακή απόδοση σε αόρατα δεδομένα. Οι τεχνικές τακτοποίησης και οι στρατηγικές διασταυρούμενης επικύρωσης (cross-validation) μπορούν να βοηθήσουν στον μετριασμό της υπερβολικής προσαρμογής, αλλά δεν είναι αλάνθαστες (Rifkin & Klautau, 2004).

Ένα άλλο κενό στον τομέα είναι η έλλειψη τυποποίησης στην εφαρμογή του LIBS και της μηχανικής μάθησης στην ανάλυση τροφίμων. Οι μέθοδοι που χρησιμοποιούνται για την απόκτηση δεδομένων, την προ επεξεργασία και την ανάλυση ποικίλλουν συχνά στη σχετική βιβλιογραφία, καθιστώντας δύσκολη τη σύγκριση των αποτελεσμάτων και την επικύρωση μοντέλων (Markley et al., 2017). Η καθιέρωση τυποποιημένων διαδικασιών για την ανάλυση LIBS, σε συνδυασμό με τη μηχανική μάθηση, θα ενίσχυε την αξιοπιστία και τη συγκρισιμότητα των αποτελεσμάτων σε διάφορες μελέτες και εργαστήρια.

Τέλος, υπάρχει ανάγκη να γεφυρωθεί το χάσμα μεταξύ της έρευνας και της εφαρμογής αυτών των τεχνολογιών. Ενώ το LIBS και η μηχανική μάθηση έχουν δείξει πολλά υποσχόμενα σε εργαστηριακές περιπτώσεις, η πρακτική εφαρμογή τους σε βιομηχανικές ρυθμίσεις για την αυθεντικότητα των τροφίμων σε πραγματικό χρόνο και την επαλήθευση γεωγραφικής προέλευσης είναι ακόμη εκκολαπτόμενη. Απαιτείται περισσότερη δουλειά για τη μεταφορά αυτών των τεχνολογιών από το εργαστήριο στην πράξη, συμπεριλαμβανομένης της ανάπτυξης συστημάτων και διεπαφών φιλικών προς τον χρήστη και αντιμετώπισης κανονιστικών και υλικοτεχνικών προκλήσεων (Santos et al., 2020).

Η αντιμετώπιση αυτών των προκλήσεων και κενών είναι σημαντική για την αξιοποίηση του πλήρους δυναμικού του LIBS και της μηχανικής μάθησης για τη διασφάλιση της αυθεντικότητας των τροφίμων και τη διαφύλαξη της ακεραιότητας της αλυσίδας εφοδιασμού τροφίμων.

3

Θεωρητικό υπόβαθρο της Μηχανικής Μάθησης

3.1 Εισαγωγή στη Μηχανική Μάθηση

Η μηχανική μάθηση, ένα υποσύνολο της τεχνητής νοημοσύνης (AI), είναι ένα καινοτόμο πεδίο που στοχεύει στη δημιουργία αλγορίθμων και μοντέλων που επιτρέπουν στους υπολογιστές να μαθαίνουν από δεδομένα και να λαμβάνουν αποφάσεις ή προβλέψεις χωρίς ρητό προγραμματισμό (Russell and Norvig, 2016). Η αυξανόμενη διαθεσιμότητα δεδομένων υψηλών διαστάσεων και υπολογιστικών πόρων τα τελευταία χρόνια έχει οδηγήσει σε αύξηση της χρήσης μεθόδων μηχανικής μάθησης σε διάφορους επιστημονικούς κλάδους, συμπεριλαμβανομένης της φασματοσκοπίας.

Στην ουσία, οι αλγόριθμοι μηχανικής μάθησης μαθαίνουν από δεδομένα αναγνωρίζοντας μοτίβα και δημιουργώντας σχέσεις μεταξύ μεταβλητών εισόδου και εξόδου. Αυτοί οι αλγόριθμοι έχουν σχεδιαστεί για να βελτιώνουν την απόδοσή τους με την πάροδο του χρόνου καθώς εκτίθενται σε περισσότερα δεδομένα. Αυτή η δυνατότητα εκμάθησης διαφοροποιεί τη μηχανική μάθηση από την παραδοσιακή πληροφορική όπου οι αλγόριθμοι είναι ρητά προγραμματισμένοι για να εκτελούν συγκεκριμένες εργασίες (Bishop, 2006).

Η μηχανική μάθηση μπορεί γενικά να κατηγοριοποιηθεί σε τρεις τύπους: μάθηση με επίβλεψη, μάθηση χωρίς επίβλεψη και ενισχυτική μάθηση. Η εποπτευόμενη μάθηση περιλαμβάνει εκμάθηση από δεδομένα με ετικέτα για την πρόβλεψη αποτελεσμάτων για μη ορατά δεδομένα (Visible Data). Συνήθως χρησιμοποιείται για εργασίες όπως η παλινδρόμηση και η ταξινόμηση όπου η μεταβλητή εξόδου ή στόχος είναι γνωστή κατά τη διάρκεια της εκπαίδευσης (James et al., 2013).

Από την άλλη πλευρά, η μάθηση χωρίς επίβλεψη περιλαμβάνει τη μάθηση από δεδομένα χωρίς ετικέτα για την εύρεση εγγενών δομών ή προτύπων μέσα στα δεδομένα. Αυτός ο τύπος μάθησης

χρησιμοποιείται συνήθως για εργασίες όπως η ομαδοποίηση και η μείωση διαστάσεων όπου η απόδοση δεν είναι γνωστή κατά τη διάρκεια της εκπαίδευσης (Hinton & Sejnowski, 1999).

Η ενισχυτική μάθηση είναι ένας τύπος μάθησης όπου ένας πράκτορας μαθαίνει να λαμβάνει αποφάσεις αλληλοεπιδρώντας με ένα περιβάλλον και λαμβάνοντας την **απάντηση** σωστό ή λάθος. Αυτός ο τύπος μάθησης χρησιμοποιείται σε τομείς όπου η λήψη αποφάσεων είναι διαδοχική και το αποτέλεσμα εξαρτάται από μια σειρά ενεργειών (Sutton & Barto, 2018).

Στη φασματοσκοπία, η μηχανική μάθηση προσφέρει μια πολλά υποσχόμενη προσέγγιση για την αντιμετώπιση πολύπλοκων φασματικών δεδομένων. Μέσω των δυνατοτήτων αναγνώρισης προτύπων και της δυνατότητάς της να χειρίζεται δεδομένα με πολλές μεταβλητές, η μηχανική μάθηση μπορεί να βοηθήσει στην εξαγωγή χρήσιμων πληροφοριών από φασματοσκοπικά δεδομένα, οδηγώντας σε βελτιωμένη απόδοση ως προς την ανάλυση (Goodacre, 2004). Η ενσωμάτωση της μηχανικής μάθησης με φασματοσκοπικές τεχνικές, όπως η φασματοσκοπία διάσπασης με λέιζερ (LIBS), επιτρέπει την ανάπτυξη ισχυρών, αποτελεσματικών και δυνητικά σε πραγματικό χρόνο αναλυτικών εργαλείων για διάφορες εφαρμογές, συμπεριλαμβανομένης της γνησιότητας των τροφίμων και της επαλήθευσης γεωγραφικής προέλευσης.

3.2 Τύποι αλγορίθμων μηχανικής μάθησης

Η μηχανική μάθηση προσφέρει μια σειρά αλγορίθμων που μπορούν να ταξινομηθούν ευρέως σε τρεις τύπους: Εποπτευόμενη μάθηση, Μη εποπτευόμενη μάθηση και Ενισχυτική μάθηση. Κάθε τύπος βρίσκει εφαρμογή σε διαφορετικές καταστάσεις ανάλογα με το είδος των δεδομένων και την εργασία στο χέρι.

3.2.1 Εποπτευόμενη μάθηση

Η εποπτευόμενη μάθηση (supervised learning) είναι ένας τύπος μηχανικής μάθησης όπου το μοντέλο εκπαιδεύεται να μάθει μια συσχέτιση μεταξύ εισόδου (input) και εξόδου (output) βάσει ετικετών (labels) που παρέχονται για τα δεδομένα εκπαίδευσης. Κατά τη διάρκεια της εκπαίδευσης, το μοντέλο προσπαθεί να εκτιμήσει μια συνάρτηση που αντιστοιχεί την είσοδο στην επιθυμητή έξοδο.

Ο αλγόριθμος της εποπτευόμενης μάθησης εκπαιδεύεται με ένα σύνολο δεδομένων εκπαίδευσης, το οποίο περιλαμβάνει ζεύγη εισόδου-εξόδου. Το μοντέλο προσαρμόζει τις εσωτερικές παραμέτρους του μέσω επαναλαμβανόμενων διαδικασιών, με στόχο να μειώσει το σφάλμα μεταξύ της πραγματικής εξόδου και της εκτιμώμενης εξόδου από το μοντέλο.

Δύο βασικές επιμέρους εργασίες στην εποπτευόμενη μάθηση είναι η ταξινόμηση και η παλινδρόμηση. Η ταξινόμηση αφορά την πρόβλεψη μιας ετικέτας και η παλινδρόμηση αφορά την πρόβλεψη μιας ποσότητας. Οι δημοφιλείς αλγόριθμοι για την εποπτευόμενη μάθηση περιλαμβάνουν γραμμική παλινδρόμηση για εργασίες παλινδρόμησης και λογιστική παλινδρόμηση, μηχανές διανυσμάτων υποστήριξης (SVM), δέντρα αποφάσεων, τυχαία δάση και νευρωνικά δίκτυα για εργασίες ταξινόμησης (James et al., 2013).

Η επιτυχία της εποπτευόμενης μάθησης στη φασματοσκοπική ανάλυση δεδομένων μπορεί να αποδοθεί στην δυνατότητά της να χειρίζεται δεδομένα υψηλών διαστάσεων και στην ικανότητά της να προβλέπει αποτελέσματα με βάση προηγούμενα πρότυπα. Στο πλαίσιο του LIBS, η εποπτευόμενη μάθηση μπορεί να χρησιμοποιηθεί για την ταξινόμηση δειγμάτων με βάση τα φάσματα τους ή την πρόβλεψη της συγκέντρωσης ενός συγκεκριμένου στοιχείου σε ένα δείγμα (Chen et al., 2020).

3.2.2 Μάθηση χωρίς επίβλεψη

Η μάθηση χωρίς επίβλεψη περιλαμβάνει μάθηση από δεδομένα χωρίς ετικέτα. Ο στόχος της μάθησης χωρίς επίβλεψη είναι να βρει εγγενή δομή ή πρότυπα στα δεδομένα, όπως η ομαδοποίηση, ή να βρει αναπαραστάσεις των δεδομένων που περιέχουν περισσότερη πληροφορία ή χρήση (Goodfellow et al., 2016).

Οι συνήθεις εργασίες μάθησης χωρίς επίβλεψη περιλαμβάνουν την ομαδοποίηση (clustering) , που είναι η διαδικασία της ανάθεσης των δεδομένων σε ομάδες (clusters) με βάση τις ομοιότητες τους. Οι αλγόριθμοι ομαδοποίησης αναζητούν μοτίβα και δομές στα δεδομένα, προσπαθώντας να δημιουργήσουν ομάδες παραδειγμάτων που είναι παρόμοια μεταξύ τους. Περιλαμβάνουν ,επίσης, τη διάσπαση (Dimensionality reduction), όπου είναι η διαδικασία μείωσης της διάστασης των δεδομένων ενώ διατηρείται η σημασιολογική πληροφορία. Σκοπός είναι να μειωθεί η πολυπλοκότητα των δεδομένων και να απομονωθούν οι σημαντικότερες χαρακτηριστικές που επηρεάζουν την δομή των δεδομένων. Άλλη μία συνήθης εργασία μη εποπτευομένης μάθησης είναι η αναγωγή διάστασης (feature extraction). Η αναγωγή διάστασης είναι η διαδικασία μετατροπής των αρχικών χαρακτηριστικών ενός δείγματος σε ένα νέο χαρακτηριστικό χώρο με λιγότερες διαστάσεις. Αυτό μπορεί να γίνει μέσω μεθόδων όπως η ανάλυση κύριων συνιστωσών (Principal Component Analysis) ή η αναγωγή διάστασης βασισμένη σε μοντέλα (Model-based Dimensionality Reduction). Μέσω αυτών των τύπων και τεχνικών, η μη εποπτευόμενη μάθηση επιτρέπει την ανακάλυψη κρυμμένων πληροφοριών και την ανάλυση μη ετικετοποιημένων δεδομένων. Οι δημοφιλείς αλγόριθμοι μάθησης χωρίς επίβλεψη περιλαμβάνουν τον αλγόριθμο k-means για ομαδοποίηση και ανάλυση κύριου συστατικού (PCA) για μείωση διαστάσεων (Hinton & Sejnowski, 1999).

Η μάθηση χωρίς επίβλεψη έχει εφαρμοστεί ευρέως στη φασματοσκοπική ανάλυση δεδομένων για την αναγνώριση προτύπων και την εξαγωγή χαρακτηριστικών από δεδομένα υψηλών διαστάσεων. Αυτό επέτρεψε βελτιωμένη ερμηνεία και ταξινόμηση σύνθετων φασματοσκοπικών δεδομένων, όπως τα φάσματα LIBS (Borràs et al., 2016).

3.2.3 Ενισχυτική Μάθηση

Η ενισχυτική μάθηση (reinforcement learning) είναι ένας τύπος μηχανικής μάθησης που ασχολείται με τον τρόπο που ένας αλγόριθμος αλληλοεπιδρά με ένα περιβάλλον και αποκτά γνώση μέσω δοκιμών και σφαλμάτων. Σε αντίθεση με άλλους τύπους μηχανικής μάθησης που βασίζονται σε ετικετοποιημένα δεδομένα, η ενισχυτική μάθηση εστιάζει στην εκπαίδευση ενός αλγορίθμου να προσαρμόζει τη συμπεριφορά του βάσει των ανταμοιβών που λαμβάνει από το περιβάλλον.

Ο αλγόριθμος ενισχυτικής μάθησης αποκτά γνώση μέσω της διαδικασίας δοκιμής και σφάλματος. Αρχικά, ο αλγόριθμος αλληλοεπιδρά με το περιβάλλον και παράγει μια δράση. Το περιβάλλον ανταποκρίνεται στη δράση παρέχοντας μια ανταμοιβή ή ένα σήμα αξιολόγησης της επίδοσης της δράσης. Ο αλγόριθμος συγκεντρώνει αυτήν την ανταμοιβή και προσπαθεί να μάθει τις πιο επικερδείς δράσεις με βάση τις αλληλεπιδράσεις που έχει ήδη έχει εκτελέσει.

Η ενισχυτική μάθηση χρησιμοποιεί ένα μοντέλο γνωστό ως Markov Decision Process (MDP) για να περιγράψει την αλληλεπίδραση μεταξύ του αλγορίθμου και του περιβάλλοντος. Το MDP αναπαριστά την κατάσταση του περιβάλλοντος, τις διαθέσιμες δράσεις και τις ανταμοιβές που συνδέονται με κάθε δράση. Ο αλγόριθμος ενισχυτικής μάθησης στοχεύει στην εύρεση της βέλτιστης πολιτικής λήψης αποφάσεων, η οποία προσδιορίζει ποια δράση θα πρέπει να εκτελείται σε κάθε κατάσταση, με σκοπό τη μεγιστοποίηση των συνολικών ανταμοιβών που λαμβάνονται από το περιβάλλον.

Οι αλγόριθμοι ενισχυτικής μάθησης χρησιμοποιούνται σε πολλές εφαρμογές, όπως στη ρομποτική, στις αυτόνομες οχήματα, στις συστάσεις στο διαδίκτυο και στη διαχείριση πόρων.

Αν και η ενισχυτική μάθηση δεν εφαρμόζεται συνήθως στη φασματοσκοπική ανάλυση δεδομένων, έχει δείξει δυνατότητες σε τομείς όπου η λήψη αποφάσεων είναι διαδοχική και το αποτέλεσμα εξαρτάται από μια σειρά ενεργειών. Με περαιτέρω έρευνα και ανάπτυξη, η ενισχυτική μάθηση θα μπορούσε να προσφέρει καινοτόμες λύσεις για τη βελτιστοποίηση της απόκτησης και ανάλυσης φασματικών δεδομένων στο LIBS.

Συμπερασματικά, κάθε τύπος μηχανικής μάθησης προσφέρει μοναδικές δυνατότητες που μπορούν να αξιοποιηθούν για φασματοσκοπική ανάλυση δεδομένων. Η επιλογή του αλγορίθμου εξαρτάται από τη φύση των δεδομένων και την εργασία. Με τη σωστή εφαρμογή, η μηχανική εκμάθηση μπορεί να βελτιώσει σημαντικά την αναλυτική απόδοση φασματοσκοπικών τεχνικών όπως το LIBS.

3.3 Αλγόριθμοι Μηχανικής Μάθησης για Φασματοσκοπική Ανάλυση

3.3.1 Υποστήριξη διανυσματικών μηχανών (SVM)

Οι Μηχανές Διανυσμάτων Υποστήριξης (SVM) είναι ισχυρά και ευέλικτα μοντέλα, ικανά να εκτελούν γραμμική ή μη γραμμική ταξινόμηση, παλινδρόμηση και ακόμη πιο ακραία ανίχνευση (Cortes & Vapnik, 1995). Τα SVM είναι ιδιαίτερα κατάλληλα για ταξινόμηση σύνθετων αλλά μικρού ή μεσαίου μεγέθους συνόλων δεδομένων. Τα SVM μπορούν να χειριστούν δεδομένα υψηλών διαστάσεων, γεγονός που τα καθιστά ιδανικά για φασματοσκοπική ανάλυση δεδομένων.

Στο πλαίσιο της φασματοσκοπίας, τα SVM έχουν χρησιμοποιηθεί για μια ποικιλία εργασιών. Για παράδειγμα, έχουν εφαρμοστεί SVM για την ταξινόμηση του ελαιόλαδου σύμφωνα με τη γεωγραφική του προέλευση χρησιμοποιώντας φάσματα LIBS, επιτυγχάνοντας υψηλή ακρίβεια (Santos Jr et al., 2015). Τα SVM λειτουργούν καλά επίσης στην αντιμετώπιση της «κατάρας της διάστασης» στα φασματοσκοπικά δεδομένα, διαχειριζόμενοι αποτελεσματικά πολλά χαρακτηριστικά εισόδου.

Στην ουσία, το SVM είναι ένας αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται για την ταξινόμηση και την παλινδρόμηση. Η βασική ιδέα πίσω από το SVM είναι η εύρεση μιας υπερεπιφάνειας απόφασης (decision boundary) που διαχωρίζει τα δεδομένα εισόδου σε δύο κατηγορίες, με τον καλύτερο δυνατό τρόπο.

Για να κατανοήσουμε την μαθηματική υπόσταση του SVM, πρέπει να αναφερθούμε σε έννοιες όπως οι ανά επιπέδων (linearly separable) και μη γραμμικά διαχωρίσιμα (non-linearly separable) δείγματα, η βαθμίδα (margin), τα ακραία δείγματα (support vectors) και οι συναρτήσεις με πολλές μεταβλητές (multivariate functions).

Ας πάρουμε το πιο απλό παράδειγμα ταξινόμησης με δύο διαστάσεις. Έστω ότι έχουμε δύο κατηγορίες δειγμάτων που παριστάνονται από τα σύμβολα "+1" και "-1". Το SVM προσπαθεί να βρει την υπερεπιφάνεια απόφασης που διαχωρίζει αυτά τα δείγματα με το μεγαλύτερο δυνατό περιθώριο (margin).

Η υπερεπιφάνεια απόφασης είναι μια γραμμή στον διΦδιάστατο χώρο (ή υπερεπίπεδο σε περισσότερες διαστάσεις) και παριστάνεται από την εξίσωση:

$w^T * x + b = 0$

όπου w είναι το διάνυσμα κατεύθυνσης (weight vector) που καθορίζει την κατεύθυνση της υπερεπιφάνειας απόφασης, x είναι το διάνυσμα εισόδου και b είναι η σταθερά προμήθειας (bias term).

Η κατηγοριοποίηση ενός δείγματος x γίνεται με βάση τον πολλαπλασιασμό του με το διάνυσμα κατεύθυνσης w και την προσθήκη της σταθεράς b. Αν το αποτέλεσμα είναι θετικό, το δείγμα ανήκει στην κατηγορία "+1", ενώ αν είναι αρνητικό, ανήκει στην κατηγορία "-1". Για να βρούμε τον καλύτερο δυνατό διαχωριστικό περιθώριο (margin), πρέπει να ελαχιστοποιήσουμε το μήκος του διανύσματος κατεύθυνσης w, δηλαδή το $\|w\|$, με τον περιορισμό ότι τα δείγματα είναι σωστά κατηγοριοποιημένα, δηλαδή:

$y_i * (w^T * x_i + b) \geq 1$

όπου y_i είναι η ετικέτα κατηγορίας του δείγματος x_i.

Οι διανύσματα x_i που είναι κοντινά στο περιθώριο καλούνται ακραία δείγματα (support vectors), καθώς καθορίζουν τη θέση και το πλάτος του περιθωρίου.

Για μη γραμμικά διαχωρίσιμα δείγματα, μπορούμε να χρησιμοποιήσουμε μια μη γραμμική συνάρτηση μετασχηματισμού (kernel function) για να μεταφέρουμε τα δείγματα σε έναν χώρο υψηλότερης διάστασης, όπου είναι πιθανό να γίνεται ο γραμμικός διαχωρισμός τους. Έτσι, ο αλγόριθμος SVM μπορεί να εφαρμοστεί και σε αυτές τις περιπτώσεις.

Οι παραπάνω έννοιες αποτελούν μια εισαγωγή στο μαθηματικό υπόβαθρο του SVM. Ο αλγόριθμος περιλαμβάνει πολλές λεπτομέρειες, όπως ο υπολογισμός των πολλαπλασιαστών Lagrange και η επίλυση του προβλήματος βελτιστοποίησης, που υπερβαίνουν τα πλαίσια αυτής της απάντησης. Ωστόσο, αυτή η επισκόπηση θα πρέπει να σας δώσει μια καλή ιδέα για τον τρόπο λειτουργίας και το μαθηματικό υπόβαθρο του SVM.

3.3.2 Τεχνικές ensemble και αλγοριθμοι με δέντρα αποφάσεων

Τα δέντρα αποφάσεων και ο αλγόριθμος Random Forest βασίζονται κατά κύριο λόγο σε τεχνικές ensemble με bagging ή με boosting.

Το boosting ακολουθεί μια προσέγγιση "σειριακού" συνδυασμού μοντέλων. Οι αλγόριθμοι boosting, όπως ο AdaBoost και ο Gradient Boosting, εκπαιδεύουν μοντέλα σειριακά, δίνοντας μεγαλύτερη έμφαση στα δείγματα που δυσκολεύουν το προηγούμενο μοντέλο να προβλέψει σωστά. Κάθε νέο μοντέλο προσπαθεί να "επικεντρωθεί" στα δείγματα που προηγούμενα προβλέφθηκαν λανθασμένα, εστιάζοντας στα λάθη του προηγούμενου μοντέλου. Τα μοντέλα εκπαιδεύονται σειριακά και οι προβλέψεις τους συνδυάζονται, συνήθως με βάρη, για να παραχθεί η τελική πρόβλεψη του ensemble.

Ο στόχος του boosting είναι να δημιουργήσει ένα ισχυρό μοντέλο εστιάζοντας στα δείγματα που είναι πιο δύσκολα για το μοντέλο να τα προβλέψει σωστά.

Το bagging (Bootstrap Aggregating) είναι μια τεχνική συνόδου (ensemble) μάθησης που στοχεύει στη βελτίωση της απόδοσης των μοντέλων μηχανικής μάθησης. Η ιδέα του bagging είναι να δημιουργηθούν πολλά ανεξάρτητα μοντέλα και να συνδυαστούν οι προβλέψεις τους για την τελική πρόβλεψη. Ο όρος "bootstrap" αναφέρεται στην μέθοδο εκπαίδευσης των ανεξάρτητων μοντέλων. Κατά τη διάρκεια της εκπαίδευσης, δημιουργούνται τυχαία υποσύνολα δεδομένων με επανατοποθέτηση (resampling) από το αρχικό σύνολο δεδομένων. Αυτό σημαίνει ότι ένα δείγμα μπορεί να επιλέγεται πολλές φορές ή να μην επιλέγεται καθόλου για τη δημιουργία ενός υποσυνόλου. Η χρήση της επανατοποθέτησης επιτρέπει την παραγωγή πολλών διαφορετικών υποσυνόλων δεδομένων που χρησιμοποιούνται για την εκπαίδευση των μοντέλων. Κάθε μοντέλο εκπαιδεύεται σε ένα από τα υποσύνολα δεδομένων και παράγει μια πρόβλεψη για κάθε δείγμα εισόδου. Οι προβλέψεις αυτές συνδυάζονται, συνήθως με μέσο όρο ή πλειοψηφία, για να παραχθεί η τελική πρόβλεψη του συνόλου των μοντέλων. Η συνδυασμένη πρόβλεψη είναι συχνά πιο ακριβής και σταθερή από τις προβλέψεις των μεμονωμένων μοντέλων.

Το bagging μπορεί να χρησιμοποιηθεί με οποιοδήποτε αλγόριθμο μηχανικής μάθησης, αλλά συνήθως συνδυάζεται με δυνατά μοντέλα όπως δέντρα απόφασης , k-NN (k-Nearest Neighbors), SVM (Support Vector Machines) κ.α.

3.3.2.1 Δέντρα αποφάσεων

Τα δέντρα αποφάσεων είναι ένας αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται για την ανάλυση και τη λήψη αποφάσεων σε προβλήματα ταξινόμησης και παλινδρόμησης. Λειτουργούν με την αρχή της διαίρεσης των δεδομένων σε διαφορετικές κατηγορίες με βάση την τιμή κάποιων χαρακτηριστικών.

Η διαδικασία του δέντρου αποφάσεων ξεκινάει με έναν κόμβο που αναπαριστά ολόκληρο το σύνολο δεδομένων. Στη συνέχεια, επιλέγεται ένα χαρακτηριστικό (ή μια μεταβλητή) από τα δεδομένα που θα χρησιμοποιηθεί για τον διαχωρισμό των δεδομένων σε διάφορες κατηγορίες. Αυτή η επιλογή γίνεται με βάση κάποιον αλγόριθμο επιλογής χαρακτηριστικού, όπως ο δείκτης πληροφορίας (index of information) ή η κατακερματισμένη παρακολούθηση (fragmented monitoring). Μετά την επιλογή του χαρακτηριστικού, δημιουργείται ένας κόμβος αποφάσεων που συνδέεται με το επιλεγμένο χαρακτηριστικό. Τα δεδομένα διαιρούνται σε υποσύνολα με βάση τις διάφορες τιμές του χαρακτηριστικού. Αυτή η διαδικασία επαναλαμβάνεται για κάθε υποσύνολο δεδομένων που δημιουργήθηκε, μέχρι να φτάσουμε σε τερματικούς κόμβους. Οι τερματικοί κόμβοι είναι οι κόμβοι που δεν χρειάζονται περαιτέρω διαχωρισμό. Σε ένα πρόβλημα ταξινόμησης, ο τερματικός κόμβος

μπορεί να αντιστοιχεί σε μια κατηγορία ή μια τελική απόφαση, ενώ σε ένα πρόβλημα παλινδρόμησης μπορεί να αναπαριστά μια τιμή. Η διαδικασία της δημιουργίας του δέντρου αποφάσεων συνεχίζεται μέχρι να δημιουργηθούν τερματικοί κόμβοι για όλα τα υποσύνολα δεδομένων ή μέχρι να επιτευχθεί κάποια συνθήκη τερματισμού, όπως η επίτευξη ενός μέγιστου βάθους του δέντρου ή η απόδοση του μοντέλου. Όταν ένα νέο δείγμα εισάγεται στο δέντρο αποφάσεων, ακολουθείται το μονοπάτι που ξεκινά από τη ρίζα και καταλήγει σε έναν τερματικό κόμβο. Η απόφαση που λαμβάνεται στον τερματικό κόμβο χρησιμοποιείται για την πρόβλεψη ή την κατηγοριοποίηση του δείγματος.

Τα δέντρα αποφάσεων είναι αποτελεσματικά μοντέλα μηχανικής μάθησης που προσφέρουν ερμηνευσιμότητα και αποτελούν μια δημοφιλή επιλογή για πολλά προβλήματα αναλυτικής λήψης αποφάσεων. Μπορούν επίσης να χρησιμοποιηθούν για την εξαγωγή χαρακτηριστικών, καθώς οι σημαντικότερες μεταβλητές εμφανίζονται στην κορυφή του δέντρου.

3.3.2.2 Random Forest

Το Random Forest είναι ένας αλγόριθμος μηχανικής μάθησης που βασίζεται στην ιδέα της συνδυαστικής ταξινόμησης πολλαπλών δέντρων αποφάσεων. Ανήκει στην κατηγορία των ensemble μοντέλων, όπου συνδυάζονται πολλά μοντέλα για να πετύχουν καλύτερη απόδοση. Η ιδέα πίσω από το Random Forest είναι ότι αντί να βασιστούμε σε ένα μόνο δέντρο αποφάσεων για την ταξινόμηση ή την πρόβλεψη, δημιουργούμε ένα δέντρων τυχαίων αποφάσεων και συνδυάζουμε τις προβλέψεις τους για να πάρουμε το τελικό αποτέλεσμα. Κάθε δέντρο στο τυχαίο δάσος εκπαιδεύεται με διαφορετικό υποσύνολο των δεδομένων εκπαίδευσης και των χαρακτηριστικών. Η διαδικασία εκπαίδευσης του Random Forest ξεκινά με την επιλογή τυχαίων υποσυνόλων των δεδομένων εκπαίδευσης για κάθε δέντρο που θα δημιουργηθεί στο τυχαίο δάσος. Η επιλογή γίνεται με αντικατάσταση, πράγμα που επιτρέπει στα δεδομένα να εμφανίζονται σε περισσότερα από ένα υποσύνολα. Έπειτα, γίνεται η εκπαίδευση των δέντρων αποφάσεων. Για κάθε υποσύνολο δεδομένων, δημιουργείται ένα δέντρο αποφάσεων χρησιμοποιώντας την ίδια διαδικασία με ένα κανονικό δέντρο αποφάσεων. Κάθε κόμβος διαχωρίζεται χρησιμοποιώντας ένα χαρακτηριστικό που επιλέγεται τυχαία από ένα υποσύνολο των διαθέσιμων χαρακτηριστικών. Τέλος, γίνεται ένας συνδυασμός προβλέψεων όταν ένα νέο δείγμα πρέπει να ταξινομηθεί ή να προβλεφθεί. Κάθε δέντρο στο τυχαίο δάσος προβλέπει την κατηγορία ή την τιμή του δείγματος. Το τελικό αποτέλεσμα προκύπτει από τον συνδυασμό των προβλέψεων όλων των δέντρων, είτε μέσω πλειοψηφίας είτε μέσω κάποιας μέσης τιμής, ανάλογα με το είδος του προβλήματος (ταξινόμηση ή παλινδρόμηση).

Το Random Forest έχει πολλαπλά πλεονεκτήματα, όπως η ικανότητα αντιμετώπισης υψηλής διαστασιμότητας, η μείωση του overfitting και η ευκολία στη χρήση.

3.3.3 Νευρωνικά Δίκτυα

Τα νευρωνικά δίκτυα είναι μια κατηγορία αλγορίθμων μηχανικής μάθησης που εμπνέονται από τον τρόπο λειτουργίας του ανθρώπινου εγκεφάλου. Αυτά τα δίκτυα αποτελούνται από μια συλλογή απλών μονάδων επεξεργασίας, γνωστών ως νευρώνες, που συνεργάζονται για να επιτύχουν εκμάθηση από δεδομένα.

Τα νευρωνικά δίκτυα αποτελούνται από διάφορα επίπεδα. Το πρώτο επίπεδο είναι η είσοδος, όπου τα δεδομένα εισάγονται στο δίκτυο. Τα ενδιάμεσα επίπεδα, γνωστά ως κρυφά επίπεδα, αποτελούνται από νευρώνες που εφαρμόζουν μη γραμμικές μετασχηματιστικές λειτουργίες στα δεδομένα εισόδου. Το τελευταίο επίπεδο είναι το επίπεδο εξόδου, όπου παράγονται οι εκτιμήσεις ή οι προβλέψεις του δικτύου.

Η δημιουργία ενός νευρωνικού δικτύου αποτελείται από αρκετά βήματα. Αρχικά, πρέπει να οριστεί η αρχιτεκτονική του δικτύου, που περιλαμβάνει τον αριθμό των επιπέδων και των νευρώνων που θα χρησιμοποιηθούν. Ένα γνωστό παράδειγμα είναι το πολυεπίπεδο πλήρως συνδεδεμένο δίκτυο (multi-layer perceptron), το οποίο χρησιμοποιήθηκε στο πείραμα αυτό. Στη συνέχεια, γίνεται η αρχικοποίηση των παραμέτρων του δικτύου, όπως τα βάρη (weights) και οι μεταβλητές ομαλοποίησης (bias). Αυτές οι παράμετροι αρχικοποιούνται με τυχαίες τιμές ή με κάποιες προκαθορισμένες τιμές, ανάλογα με την επιλογή μας. Έπειτα, πραγματοποιείται η εκπαίδευση του δικτύου. Κατά τη διάρκεια αυτής της διαδικασίας, τα δεδομένα εισόδου περνούν μέσα από το δίκτυο και υπολογίζονται οι προβλέψεις του. Γίνεται σύγκριση αυτών των προβλέψεων με τις πραγματικές τιμές και υπολογίζεται ένα μέτρο σφάλματος, γνωστό ως χαμένο (loss). Ο στόχος είναι να ελαχιστοποιηθεί αυτός ο χαμένος, χρησιμοποιώντας τεχνικές βελτιστοποίησης όπως η μέθοδος της ανάπτυξης διάδοσης προς τα πίσω (back propagation). Η διαδικασία της εκπαίδευσης επαναλαμβάνεται πολλές φορές, με τη χρήση διαφορετικών παρτίδων δεδομένων (batching), με σκοπό τη βελτίωση της απόδοσης του δικτύου. Αφού ολοκληρωθεί η εκπαίδευση, το νευρωνικό δίκτυο αξιολογείται σε ένα σύνολο δεδομένων ελέγχου για να μετρηθεί η απόδοσή του. Αυτό μπορεί να γίνει με την υπολογιστική ακρίβεια, την ακρίβεια ταξινόμησης ή άλλες μετρικές, ανάλογα με τον τύπο του προβλήματος.

Τα νευρωνικά δίκτυα έχουν αναπτυχθεί σε μια πληθώρα αρχιτεκτονικών και παραλλαγών που προσαρμόζονται σε διάφορες εφαρμογές. Εδώ είναι μερικές από τις πιο γνωστές αρχιτεκτονικές νευρωνικών δικτύων:

1. Τα Προσαρμοστικά Νευρωνικά Δίκτυα (Artificial Neural Networks - ANNs): Αυτή είναι η βασική μορφή νευρωνικών δικτύων, όπου οι νευρώνες οργανώνονται σε στρώματα και κάθε

νευρώνας συνδέεται με τους νευρώνες του επόμενου στρώματος. Τα ANNs είναι κατάλληλα για προβλήματα όπως η ταξινόμηση εικόνων και η πρόβλεψη.

2. Τα Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks - CNNs): Αυτή η αρχιτεκτονική έχει σχεδιαστεί ειδικά για επεξεργασία εικόνων και αναγνώριση προτύπων. Χρησιμοποιεί συνελίξεις για την ανίχνευση χαρακτηριστικών σε διαφορετικά μέρη της εικόνας και τονίζει την ιεραρχική δομή των δεδομένων.
3. Τα Αναδρομικά Νευρωνικά Δίκτυα (Recurrent Neural Networks - RNNs): Αυτή η αρχιτεκτονική έχει σχεδιαστεί για την επεξεργασία ακολουθιών δεδομένων, όπως φυσική γλώσσα, ομιλία και χρονοσειρές. Οι RNNs έχουν μια αναδρομική δομή που επιτρέπει την αποθήκευση και την αξιοποίηση πληροφοριών από προηγούμενα στοιχεία της ακολουθίας.
4. Τα Γεννητικά Νευρωνικά Δίκτυα (Generative Neural Networks - GANs): Αυτή η αρχιτεκτονική χρησιμοποιείται για τη δημιουργία νέων δεδομένων με βάση ένα υπάρχον σύνολο δεδομένων. Τα GANs αποτελούνται από δύο ανταγωνιστικά νευρωνικά δίκτυα, έναν γεννήτορα (generator) που παράγει νέα δεδομένα και έναν διακριτή (discriminator) που προσπαθεί να αξιολογήσει την πραγματικότητα των δεδομένων.

Στον τομέα της φασματοσκοπίας, τα νευρωνικά δίκτυα έχουν χρησιμοποιηθεί για την ανάλυση πολύπλοκων φασματικών δεδομένων και την επίτευξη υψηλών επιπέδων ακρίβειας σε εργασίες ταξινόμησης και πρόβλεψης. Μια μελέτη εφάρμοσε ένα συνελικτικό νευρωνικό δίκτυο (CNN) σε δεδομένα LIBS για ταξινόμηση εδάφους και πέτυχε ποσοστό ακρίβειας σημαντικά υψηλότερο από τους παραδοσιακούς αλγόριθμους μηχανικής μάθησης (Mehdipour et al., 2019).

3.3.4 Τεχνικές Ομαδοποίησης και Μείωσης Διαστάσεων

Οι τεχνικές ομαδοποίησης, όπως το k-means και οι τεχνικές μείωσης διαστάσεων, όπως το Principal Component Analysis (PCA) είναι αλγόριθμοι μάθησης χωρίς επίβλεψη που μπορούν να βοηθήσουν στην αποκάλυψη της εγγενούς δομής των δεδομένων. Το K-means χρησιμοποιείται για την κατανομή των δεδομένων εισόδου σε k διακριτές ομάδες, ενώ το PCA χρησιμοποιείται για τη μείωση της διάστασης των δεδομένων διατηρώντας παράλληλα το μεγαλύτερο μέρος της διακύμανσης (Jolliffe και Cadima, 2016).

Η ανάλυση κυρίαρχων συστατικών (Principal Component Analysis ή PCA) είναι μια τεχνική μείωσης διαστάσεων που χρησιμοποιείται στην επεξεργασία και ανάλυση δεδομένων. Σκοπός της είναι να μετασχηματίσει ένα σύνολο πολλαπλών χαρακτηριστικών σε ένα νέο χώρο χαρακτηριστικών, λαμβάνοντας υπόψη τις κύριες κατευθύνσεις της μεγαλύτερης διακύμανσης στα δεδομένα. Με αυτόν

τον τρόπο, μπορούμε να μειώσουμε τον αριθμό των χαρακτηριστικών, διατηρώντας τα σημαντικότερα χαρακτηριστικά που εξηγούν την πλειονότητα της διακύμανσης των δεδομένων.

Ο τρόπος λειτουργίας του PCA είναι ο εξής. Πρώτον, αφαιρείται ο μέσος όρος από κάθε χαρακτηριστικό έτσι ώστε να κεντραριστούν τα δεδομένα. Στη συνέχεια, γίνεται ο υπολογισμός του πίνακα συνδιακύμανσης (covariance matrix) για τα κεντραρισμένα δεδομένα ο οποίος καταγράφει την κοινή μεταβλητότητα μεταξύ των διαφορετικών χαρακτηριστικών. Έπειτα, γίνεται υπολογισμός των ιδιοδιανύσματων και των ιδιοτιμών του πίνακα συνδιακύμανσης. Τα ιδιοδιανύσματα αναπαριστούν τις κύριες κατευθύνσεις της μεγαλύτερης διακύμανσης, ενώ οι ιδιοτιμές αντιστοιχούν στη διακύμανση κατά μήκος αυτών των κατευθύνσεων. Επιπλέον, γίνεται η επιλογή των κυρίαρχων συνιστωσών, κρατώντας τα πρώτα N ιδιοδιανύσματα με τις μεγαλύτερες ιδιοτιμές για να διατηρηθεί το σημαντικότερο μέρος της διακύμανσης των δεδομένων. Εδώ, το N είναι ο αριθμός των κύριων συνιστωσών που πρέπει να κρατηθούν, ο οποίος μπορεί να είναι μικρότερος από τον αριθμό των αρχικών χαρακτηριστικών. Τέλος, χρησιμοποιούνται τα επιλεγμένα ιδιοδιανύσματα για να μετασχηματιστούν τα αρχικά δεδομένα στο νέο χώρο κυρίαρχων συνιστωσών, μειώνοντας τις διαστάσεις των δεδομένων.

Η PCA είναι ιδιαίτερα χρήσιμη για μείωση των διαστάσεων όταν αντιμετωπίζουμε δεδομένα με πολλά χαρακτηριστικά και θέλουμε να διατηρήσουμε τις κύριες πτυχές της διακύμανσης. Χρησιμοποιείται επίσης για οπτικοποίηση και εξερεύνηση δεδομένων υψηλών διαστάσεων.

Στη φασματοσκοπία, αυτές οι τεχνικές έχουν χρησιμοποιηθεί για την ταξινόμηση φασμάτων και την ανίχνευση ακραίων τιμών. Για παράδειγμα, η PCA έχει εφαρμοστεί σε δεδομένα LIBS για αρχική διερευνητική ανάλυση, μείωση θορύβου και ανίχνευση κρυφών μοτίβων στα δεδομένα (Sancey et al., 2014).

Κάθε αλγόριθμος μηχανικής μάθησης έχει τα δυνατά και τα αδύνατα σημεία του και η επιλογή του αλγορίθμου εξαρτάται από τη συγκεκριμένη εργασία, τη φύση των δεδομένων και τους διαθέσιμους υπολογιστικούς πόρους. Όταν επιλεγούν και βελτιστοποιηθούν κατάλληλα, αυτοί οι αλγόριθμοι μπορούν να παρέχουν ισχυρά και αποτελεσματικά εργαλεία για φασματοσκοπική ανάλυση δεδομένων.

3.4 Αξιολόγηση μοντέλων μηχανικής μάθησης

Η δημιουργία ενός μοντέλου μηχανικής μάθησης είναι μισή δουλειά που χρειάζεται να πραγματοποιηθεί, το άλλο μισό διασφαλίζει ότι αποδίδει καλά και γενικεύει σε αόρατα δεδομένα. Αυτό απαιτεί αυστηρές στρατηγικές και μετρήσεις αξιολόγησης, συμπεριλαμβανομένης της αλληλοεπικύρωσης, της τακτοποίησης και των κατάλληλων μετρήσεων απόδοσης.

3.4.1 Διασταυρούμενη επικύρωση

Η διασταυρούμενη επικύρωση (cross validation) είναι ένα ισχυρό προληπτικό μέτρο κατά της υπερβολικής προσαρμογής, της τάσης ενός μοντέλου να μαθαίνει τον θόρυβο στα δεδομένα εκπαίδευσης αντί για το υποκείμενο μοτίβο. Παρέχει μια ισχυρή μέθοδο για την εκτίμηση της απόδοσης ενός μοντέλου σε αόρατα δεδομένα (Kohavi, 1995).

Στη διασταυρούμενη επικύρωση k-fold, τα δεδομένα χωρίζονται σε k υποσύνολα ίσου μεγέθους. Στη συνέχεια, το μοντέλο εκπαιδεύεται k φορές, κάθε φορά χρησιμοποιώντας k-1 υποσύνολα ως δεδομένα εκπαίδευσης και το υπόλοιπο υποσύνολο ως δεδομένα επικύρωσης. Στη συνέχεια, η απόδοση του μοντέλου υπολογίζεται κατά μέσο όρο στους k γύρους που έτρεξε ο αλγόριθμος. Αυτή η διαδικασία διασφαλίζει ότι όλα τα σημεία δεδομένων χρησιμοποιούνται τόσο για εκπαίδευση όσο και για επικύρωση, οδηγώντας σε μια πιο ισχυρή εκτίμηση της απόδοσης του μοντέλου (James et al., 2013).

3.4.2 Τακτοποίηση

Η τακτοποίηση είναι μια τεχνική που χρησιμοποιείται για την αποφυγή της υπερπροσαρμογής προσθέτοντας έναν όρο ποινής στη συνάρτηση απώλειας που το μοντέλο στοχεύει να ελαχιστοποιήσει. Αυτή η ποινή αποθαρρύνει το μοντέλο να αποδίδει υπερβολική σημασία σε οποιοδήποτε μεμονωμένο χαρακτηριστικό, μειώνοντας έτσι την πολυπλοκότητα του μοντέλου και βελτιώνοντας τη γενίκευσή του (Ng, 2004).

Δύο κοινές μορφές τακτοποίησης είναι η L1 και η L2. Η τακτοποίηση L1 μπορεί να οδηγήσει σε αραιές λύσεις όπου ορισμένα βάρη χαρακτηριστικών είναι μηδενικά, εκτελώντας αποτελεσματικά την επιλογή χαρακτηριστικών. Η τακτοποίηση L2, από την άλλη πλευρά, συρρικνώνει τους συντελεστές αλλά δεν τους μηδενίζει απαραίτητα (Tibshirani, 1996).

3.4.3 Μετρήσεις απόδοσης

Η επιλογή των μετρήσεων απόδοσης εξαρτάται από τη συγκεκριμένη εργασία – ταξινόμηση, παλινδρόμηση ή ομαδοποίηση. Για εργασίες ταξινόμησης, οι κοινές μετρήσεις περιλαμβάνουν την ακρίβεια, την ανάκληση, τη βαθμολογία F1 και την περιοχή κάτω από την καμπύλη ROC (AUC-ROC). Για εργασίες παλινδρόμησης, οι κοινές μετρήσεις περιλαμβάνουν το μέσο απόλυτο σφάλμα (MAE), το μέσο τετράγωνο σφάλμα (MSE) και το R-squared (James et al., 2013).

Είναι σημαντικό να επιλέγονται οι σωστές μετρήσεις με βάση την εργασία και το συγκεκριμένο πλαίσιο. Για παράδειγμα, σε μια εργασία ταξινόμησης όπου οι κλάσεις είναι μη ισορροπημένες, η

ακρίβεια μπορεί να είναι παραπλανητική και μετρήσεις όπως η ακρίβεια, η ανάκληση ή η βαθμολογία F1 μπορεί να είναι πιο κατάλληλες (Davis and Goadrich, 2006).

Συμπερασματικά, η αξιολόγηση της απόδοσης ενός μοντέλου μηχανικής μάθησης είναι εξίσου σημαντική με την κατασκευή του ίδιου του μοντέλου. Είναι σημαντικό να χρησιμοποιηθούν κατάλληλες στρατηγικές αξιολόγησης και μετρήσεις για να διασφαλιστεί ότι το μοντέλο όχι μόνο ταιριάζει καλά στα δεδομένα αλλά και γενικεύεται σε μη ορατά δεδομένα.

4

Θεωρητικό Πλαίσιο LIBS και Ενσωμάτωση Μηχανικής Μάθησης

4.1 Εισαγωγή στην τεχνική LIBS

Η τεχνική φασματοσκοπίας πλάσματος επαγόμενου από λέιζερ, Laser-Induced Breakdown Spectroscopy-LIBS είναι μια τεχνική φασματοσκοπίας εκπομπής που χρησιμοποιείται σε διάφορες επιστημονικές και βιομηχανικές εφαρμογές. Η τεχνική στηρίζεται στην εστίαση της δέσμης ενός ισχυρού λέιζερ στην επιφάνεια ενός δείγματος, δημιουργώντας έτσι ένα πλάσμα που εκπέμπει φως. Το εκπεμπόμενο φως περιέχει τις φασματικές υπογραφές των στοιχείων που υπάρχουν στο δείγμα (Cremers & Radziemski, 2006).

Τα βασικά πλεονεκτήματα της τεχνικής LIBS περιλαμβάνουν την ταχεία ανάλυση, την ελάχιστη προετοιμασία του δείγματος και την δυνατότητα ανάλυσης οποιασδήποτε φάσης ύλης - στερεάς, υγρής ή αέριας. Έχει βρει εφαρμογές σε διάφορους τομείς όπως η γεωλογία, η αρχαιολογία, οι περιβαλλοντικές επιστήμες, και σε πιο σχετικές με αυτή τη διατριβή, την αυθεντικότητα των τροφίμων και την επαλήθευση της γεωγραφικής τους προέλευσης (Miziolek et al., 2006).

Παρόλο που το LIBS παρέχει ολοκληρωμένα φασματικά δεδομένα, η πολυπλοκότητα και οι υψηλές ή διαστάσεις αυτών των δεδομένων απαιτούν πολύπλοκες τεχνικές ανάλυσης. Αυτό είναι όπου η μηχανική μάθηση μπαίνει στην εικόνα, προσφέροντας μια σειρά μεθόδων ικανών να αναλύουν και να ερμηνεύουν αποτελεσματικά τα δεδομένα LIBS.

Η χρήση της μηχανικής μάθησης σε συνδυασμό με την τεχνική LIBS και είναι ένας πρόσφατος αλλά ταχέως αναπτυσσόμενος τομέας έρευνας. Η κύρια εστίαση έγκειται στην ανάπτυξη μοντέλων μηχανικής μάθησης που μπορούν να προβλέψουν χαρακτηριστικά ενός δείγματος, όπως η γεωγραφική του προέλευση, με βάση τα φασματικά δεδομένα LIBS (Anzano & Laserna, 2016).

Για λόγους πληρότητας της παρούσας εργασίας στην συνέχεια παρουσιάζονται σύντομα οι θεμελιώδεις αρχές του LIBS και οι διάφοροι τύποι αλγορίθμων μηχανικής μάθησης που ταιριάζουν καλύτερα για την ανάλυση δεδομένων LIBS και τις προκλήσεις και τις μελλοντικές προοπτικές σε αυτόν τον ερευνητικό τομέα.

4.2 Βασικές αρχές της τεχνικής LIBS

Η τεχνική LIBS Laser-Induced Breakdown Spectroscopy (LIBS) είναι μια ευέλικτη φασματοσκοπική τεχνική που βασίζεται στην αλληλεπίδραση ενός ισχυρού παλμικού λέιζερ με το υπό μελέτη υλικό, με αποτέλεσμα την δημιουργία πλάσματος το οποίο εκπέμπει φάσμα το οποίο είναι αντιπροσωπευτικό της στοιχειακής σύνθεσης του υλικού (Cremers & Radziemski, 2006).

Όταν ο παλμός λέιζερ αλληλοεπιδρά με την επιφάνεια του υλικού, αφαιρεί μια μικρή ποσότητα του υλικού, δημιουργώντας ένα πλάσμα υψηλής θερμοκρασίας που αποτελείται από διεγερμένα ιόντα, άτομα και μόρια των στοιχείων που υπάρχουν στο δείγμα. Αυτό το πλάσμα ψύχεται γρήγορα και αποδιεγείρεται, εκπέμποντας χαρακτηριστικές ατομικές και ιοντικές φασματικές γραμμές, δηλ. το φάσμα LIBS (Miziolek et al., 2006).

Ολόκληρη η διαδικασία, από τον παλμό λέιζερ έως την ανίχνευση εκπομπής, είναι αξιοσημείωτα γρήγορη, λαμβάνει χώρα σε διάστημα μικροδευτερόλεπτων, γεγονός που συμβάλλει στην ανάλυση σε πραγματικό χρόνο (Anzano & Laserna, 2016).

Το φάσμα LIBS περιλαμβάνει κορυφές που αντιστοιχούν στις γραμμές ατομικής εκπομπής των συστατικών στοιχείων. Οι φασματικές γραμμές κάθε στοιχείου εμφανίζονται σε μοναδικά μήκη κύματος, χρησιμεύοντας ως «δακτυλικά αποτυπώματα» που επιτρέπουν την αναγνώριση των στοιχείων στο δείγμα. Η σχετική ένταση αυτών των γραμμών μπορεί να παρέχει ποιοτικές ή και ποσοτικές πληροφορίες σχετικά με τη συγκέντρωση των στοιχείων (Noll, 2012).

Διάφοροι παράγοντες επηρεάζουν την ποιότητα των φασμάτων LIBS, συμπεριλαμβανομένων των παραμέτρων λέιζερ (π.χ. ενέργειας, διάρκειας παλμού), της φύσης του δείγματος και του περιβάλλοντος. Έτσι, η απόκτηση συνεπών και αξιόπιστων φασμάτων LIBS συχνά απαιτεί προσεκτική βελτιστοποίηση αυτών των παραμέτρων (Hahn & Omenetto, 2010).

Πέρα από τη ανάλυση των στοιχείων, τα φάσματα LIBS μπορεί να παρέχουν πληροφορίες για ισοτοπικές αναλογίες, μοριακά είδη, ακόμη και φυσική κατάσταση. Ωστόσο, η εξαγωγή τέτοιων πληροφοριών απαιτεί συνήθως πιο σύνθετες τεχνικές ανάλυσης δεδομένων, εξ ου και το ενδιαφέρον για μεθοδολογίες μηχανικής μάθησης για την αντιμετώπιση αυτών των προκλήσεων (Liu, et al., 2017).

Συνοπτικά, το LIBS, με την ικανότητά του να παρέχει σε πραγματικό χρόνο, ελάχιστα καταστροφική στοιχειακή ανάλυση, προσφέρει σημαντικές δυνατότητες σε μια σειρά εφαρμογών. Ωστόσο, η ερμηνεία και η χρήση των δεδομένων LIBS είναι πολύπλοκες εργασίες λόγω της υψηλής διάστασης και της εγγενούς μεταβλητότητας των φασμάτων. Ως εκ τούτου, οι αποτελεσματικές μέθοδοι ανάλυσης δεδομένων είναι απαραίτητες για τη μετάφραση των μετρήσεων LIBS σε ουσιαστικά αποτελέσματα - ένας ρόλος που ταιριάζει κατάλληλα στις τεχνικές μηχανικής μάθησης.

4.2.1 Η διαδικασία του ablation και η δημιουργία πλάσματος

Η έναρξη μιας ανάλυσης φασματοσκοπίας διάσπασης που προκαλείται από λέιζερ (LIBS) ξεκινά με τη δημιουργία ενός πλάσματος από την επιφάνεια του δείγματος μέσω μιας διαδικασίας που ονομάζεται αφαίρεση με λέιζερ. Η αφαίρεση με λέιζερ προκαλείται από την απορρόφηση του παλμού λέιζερ υψηλής ενέργειας που ακτινοβολεί την επιφάνεια του δείγματος. Η απορροφούμενη ενέργεια σπάει τους ατομικούς ή μοριακούς δεσμούς του υλικού, εξατμίζοντάς το και ιονίζοντάς το για να δημιουργήσει ένα λοφίο πλάσματος (Pareja et al., 2018).

Η διαδικασία σχηματισμού πλάσματος είναι γρήγορη, και συμβαίνει σε κλάσματα δευτερολέπτου. Στην αρχή, το εξατμισμένο υλικό υφίσταται μια ταχεία διαστολή λόγω της υψηλής θερμοκρασίας και πίεσης που έχει. Αυτή η διαστολή, μαζί με τις απώλειες ενέργειας λόγω ακτινοβολίας και κρούσεων, οδηγεί σε ψύξη του πλάσματος και σχηματισμό κρουστικού κύματος (Hahn & Omenetto, 2012).

Το πλάσμα είναι μια σύνθετη οντότητα που αποτελείται από ένα μείγμα διεγερμένων ατόμων, ιόντων και ελεύθερων ηλεκτρονίων. Τα χαρακτηριστικά του, όπως η θερμοκρασία και η πυκνότητα ηλεκτρονίων, μπορεί να ποικίλλουν ανάλογα με το χρόνο και τη χωρική θέση, γεγονός που επηρεάζει τη φασματική εκπομπή και συνεπώς την ερμηνεία των δεδομένων LIBS (Miziolek et al., 2006).

4.2.2 Φασματοσκοπία Εκπομπής και Στοιχειακή Ανάλυση

Η καρδιά της ανάλυσης LIBS βρίσκεται στη φασματοσκοπία εκπομπής του πλάσματος. Καθώς το πλάσμα ψύχεται και τα συστατικά αποδιέγονται, εκπέμπουν φως σε χαρακτηριστικά μήκη κύματος. Αυτό το εκπεμπόμενο φως συλλέγεται και αναλύεται για να παραχθεί το φάσμα LIBS (Noll, 2012).

Το φάσμα LIBS αποτελείται από διάφορες φασματικές γραμμές εκπομπής , καθεμία από τις οποίες αντιστοιχεί σε μια συγκεκριμένη ατομική ή ιοντική μετάβαση. Τα μήκη κύματος αυτών των γραμμών χρησιμεύουν ως δακτυλικά αποτυπώματα για την στοιχειακή αναγνώριση, καθώς κάθε στοιχείο έχει ένα μοναδικό σύνολο φασματικών γραμμών. Επιπλέον, οι εντάσεις αυτών των γραμμών συσχετίζονται με τη σχετική συγκέντρωση των αντίστοιχων στοιχείων στο δείγμα (Cremers & Radziemski, 2006).

Ωστόσο, η σχέση μεταξύ της έντασης της φασματικής γραμμής και της στοιχειακής συγκέντρωσης δεν είναι απλή. Πολλοί παράγοντες μπορούν να επηρεάσουν την ένταση της φασματικής γραμμής, συμπεριλαμβανομένης της φυσικής κατάστασης του δείγματος, των επιδράσεων της μήτρας, των συνθηκών του πλάσματος και των παραμέτρων του λέιζερ (Senesi et al., 2009).

Αυτές οι πολυπλοκότητες στη φασματοσκοπία εκπομπής LIBS θέτουν σημαντικές προκλήσεις για την ερμηνεία των δεδομένων. Κατά συνέπεια, προηγμένες τεχνικές ανάλυσης δεδομένων, όπως αλγόριθμοι μηχανικής μάθησης, είναι απαραίτητες για την εξαγωγή σημαντικών πληροφοριών από τα φάσματα LIBS και την αποτελεσματική χρήση τους για εργασίες όπως ο προσδιορισμός γεωγραφικής προέλευσης ή η επαλήθευση γνησιότητας των τροφίμων.

4.3 Προκλήσεις στην ανάλυση δεδομένων LIBS

Ενώ η φασματοσκοπία LIBS προσφέρει πολλά πλεονεκτήματα, όπως ταχεία, in situ ανάλυση με ελάχιστη προετοιμασία δείγματος, η φάση ανάλυσης δεδομένων μπορεί να είναι δύσκολη λόγω διαφόρων παραγόντων που επηρεάζουν τα φάσματα LIBS.

Μία από τις κύριες προκλήσεις στην ανάλυση δεδομένων LIBS προκύπτει από την πολυπλοκότητα και την υψηλή διάσταση των φασμάτων. Κάθε φάσμα αποτελείται από πολλές φασματικές γραμμές, που αντιστοιχούν στις ποικίλες διαφορετικές ατομικές και ιοντικές μεταβάσεις των συστατικών στοιχείων του δείγματος. Οι κορυφές στα φάσματα μπορεί να επικαλύπτονται, καθιστώντας δύσκολη την αναγνώριση και τον ποσοτικό προσδιορισμό των στοιχείων (Cremers & Radziemski, 2006).

Η μεταβλητότητα στα φάσματα LIBS είναι μια άλλη σημαντική πρόκληση. Παραλλαγές μπορεί να προκύψουν λόγω διαφορών στη μήτρα του δείγματος, αλλαγές στις παραμέτρους λέιζερ και διακυμάνσεις στις συνθήκες περιβάλλοντος. Για παράδειγμα, το ίδιο υλικό μπορεί να αποδώσει διαφορετικά φάσματα LIBS ανάλογα με τη φυσική του κατάσταση (στερεό, υγρό, αέριο) ή εάν είναι μέρος μιας σύνθετης μήτρας (Miziolek et al., 2006).

Ένα σχετικό ζήτημα είναι το φαινόμενο matrix, ένα φαινόμενο όπου η παρουσία ορισμένων στοιχείων επηρεάζει το σήμα LIBS άλλων στοιχείων. Αυτά τα φαινόμενα μπορούν να παραμορφώσουν τη συσχέτιση μεταξύ της έντασης της φασματικής γραμμής και της συγκέντρωσης του στοιχείου, περιπλέκοντας την ποσοτική ανάλυση (Hahn & Omenetto, 2012).

Οι παραλλαγές από λήψη σε λήψη προσθέτουν άλλο ένα επίπεδο πολυπλοκότητας στην ανάλυση δεδομένων LIBS. Ακόμη και όταν αναλύεται το ίδιο σημείο σε ένα δείγμα, διαδοχικοί παλμοί λέιζερ μπορούν να παράγουν διαφορετικά φάσματα LIBS. Αυτές οι παραλλαγές προκύπτουν λόγω ελαφρών αλλαγών στη διαδικασία αφαίρεσης και στις συνθήκες πλάσματος με κάθε παλμό λέιζερ (Senesi et al., 2009).

Ενώ οι μέθοδοι βαθμονόμησης μπορούν να μετριάσουν ορισμένες από αυτές τις προκλήσεις, δεν είναι πάντα πρακτικές ή αποτελεσματικές, ιδιαίτερα σε εφαρμογές πεδίου όπου τα τυπικά δείγματα ενδέχεται να μην είναι άμεσα διαθέσιμα. Επιπλέον, οι μέθοδοι βαθμονόμησης συνήθως υποθέτουν μια γραμμική σχέση μεταξύ της έντασης της φασματικής γραμμής και της συγκέντρωσης στοιχείων, κάτι που δεν συμβαίνει πάντα (Noll, 2012).

Λόγω αυτών των πολυπλοκοτήτων, οι παραδοσιακές τεχνικές στατιστικής και φασματοσκοπικής ανάλυσης συχνά υστερούν στην αποτελεσματική ερμηνεία των δεδομένων LIBS. Έτσι, υπάρχει ένα αυξανόμενο ενδιαφέρον για τη χρήση αλγορίθμων μηχανικής μάθησης για την ανάλυση δεδομένων LIBS. Η μηχανική μάθηση προσφέρει ισχυρά εργαλεία ικανά να μοντελοποιούν πολύπλοκες, μη γραμμικές σχέσεις, να αντιμετωπίζουν δεδομένα υψηλών διαστάσεων και να μαθαίνουν από τη μεταβλητότητα των δεδομένων. Ως εκ τούτου, η μέθοδος υπόσχεται πολλά για την αντιμετώπιση των προκλήσεων στην ανάλυση δεδομένων LIBS και την αξιοποίηση του πλήρους δυναμικού του LIBS σε διάφορες εφαρμογές.

4.4 Ενοποίηση του LIBS με τη Μηχανική Μάθηση

Η ενσωμάτωση του LIBS με τη μηχανική μάθηση έχει γίνει μια ολοένα και πιο πολλά υποσχόμενη προσέγγιση για την αντιμετώπιση της πολυπλοκότητας και της υψηλής διάστασης των δεδομένων LIBS. Αξιοποιώντας την δυνατότητα της μηχανικής μάθησης να μοντελοποιεί πολύπλοκες, μη γραμμικές σχέσεις και να μαθαίνει από δεδομένα υψηλών διαστάσεων, μπορούν να πραγματοποιηθούν πιο διορατικές αναλύσεις. Αυτά περιλαμβάνουν βελτιωμένη φασματική αναγνώριση κορυφών, ανώτερη ποσοτικοποίηση των στοιχειακών συγκεντρώσεων και βελτιωμένη πρόβλεψη ιδιοτήτων όπως η γεωγραφική προέλευση ή η αυθεντικότητα των τροφίμων (Ferreira et al., 2018).

4.4.1 Προεπεξεργασία δεδομένων LIBS για Μηχανική Μάθηση

Προτού τα δεδομένα LIBS μπορούν να τροφοδοτηθούν σε ένα μοντέλο μηχανικής μάθησης, είναι συχνά απαραίτητα βήματα προεπεξεργασίας για τη βελτίωση της ποιότητας των δεδομένων και τη βελτίωση της αποτελεσματικότητας της επακόλουθης ανάλυσης.

Ένα κρίσιμο βήμα προεπεξεργασίας είναι η φασματική κανονικοποίηση, η οποία μετριάζει την επίδραση των διακυμάνσεων στη συνολική ένταση των φασμάτων, που συχνά προκαλούνται από αλλαγές στην ενέργεια ή την εστίαση του λέιζερ. Τεχνικές όπως η κανονικοποίηση ολικής έντασης,

όπου κάθε φάσμα διαιρείται με τη συνολική του ένταση, ή η κανονικοποίηση διανύσματος, όπου κάθε φάσμα διαιρείται με τον Ευκλείδειο κανόνα του, χρησιμοποιούνται συνήθως (Li et al., 2014).

Η διόρθωση γραμμής βάσης είναι ένα άλλο βασικό βήμα προεπεξεργασίας. Αφαιρεί το συνεχές φόντο ή τη «γραμμή βάσης» που μπορεί να υπάρχει στα φάσματα LIBS λόγω εκπομπής ευρείας ζώνης ή διάσπαρτου φωτός. Διάφορες μέθοδοι, όπως πολυωνυμική προσαρμογή, ασύμμετρα ελάχιστα τετράγωνα και μετασχηματισμός κυματιδίων, μπορούν να χρησιμοποιηθούν για τη διόρθωση γραμμής βάσης (Zhao et al., 2015).

Η μείωση του θορύβου είναι συχνά απαραίτητη για τη βελτίωση της αναλογίας σήματος προς θόρυβο των φασμάτων. Τεχνικές όπως τα φίλτρα εξομάλυνσης, ο μετασχηματισμός Fourier ή ο μετασχηματισμός κυματιδίων μπορούν να χρησιμοποιηθούν για τη μείωση του θορύβου στα φάσματα LIBS (Zhang et al., 2017).

Τέλος, η εξαγωγή χαρακτηριστικών είναι ένα κρίσιμο βήμα που μετατρέπει τα υψηλών διαστάσεων φάσματα LIBS σε χώρο χαμηλότερης διάστασης διατηρώντας παράλληλα τις πιο σημαντικές πληροφορίες. Τεχνικές όπως η Ανάλυση Κύριων Στοιχείων (PCA), η Γραμμική Διακριτική Ανάλυση (LDA) ή ο μετασχηματισμός κυματιδίων μπορούν να χρησιμοποιηθούν για την εξαγωγή χαρακτηριστικών (Gautam et al., 2015).

Η προεπεξεργασία των δεδομένων LIBS διαδραματίζει κρίσιμο ρόλο στην αποτελεσματικότητα της επακόλουθης ανάλυσης μηχανικής μάθησης. Βελτιώνοντας την ποιότητα των δεδομένων και μειώνοντας τις διαστάσεις τους, τα βήματα προεπεξεργασίας μπορούν να βελτιώσουν σημαντικά την απόδοση του μοντέλου μηχανικής εκμάθησης και να παρέχουν πιο ακριβή και ουσιαστικά αποτελέσματα.

4.4.2 Εφαρμογή της Μηχανικής Μάθησης στην Ανάλυση Δεδομένων LIBS

Η εφαρμογή της μηχανικής μάθησης στην ανάλυση δεδομένων LIBS γίνεται ολοένα και πιο διαδεδομένη λόγω της δυνατότητας αυτών των τεχνικών να αντιμετωπίσουν ορισμένες από τις εγγενείς προκλήσεις που σχετίζονται με τις παραδοσιακές μεθόδους ανάλυσης δεδομένων. Οι μέθοδοι μηχανικής μάθησης είναι πλεονεκτικές ως προς την δυνατότητά τους να μοντελοποιούν πολύπλοκες, μη γραμμικές σχέσεις, να χειρίζονται δεδομένα υψηλών διαστάσεων και να προσαρμόζονται στη μεταβλητότητα των δεδομένων.

Διάφοροι τύποι αλγορίθμων μηχανικής μάθησης έχουν χρησιμοποιηθεί για την ανάλυση δεδομένων LIBS, ο καθένας από τους οποίους προσφέρει μοναδικά οφέλη. Αυτά περιλαμβάνουν, αλλά δεν περιορίζονται σε, αλγόριθμους εποπτευόμενης μάθησης, αλγόριθμους μάθησης χωρίς επίβλεψη και μεθόδους βαθιάς μάθησης (Liu et al., 2017).

Οι εποπτευόμενοι αλγόριθμοι εκμάθησης, όπως οι Υποστήριξη Διανυσματικές Μηχανές (SVMs), το Τυχαίο Δάσος (RF) και τα Τεχνητά Νευρωνικά Δίκτυα (ANNs), έχουν εφαρμοστεί ευρέως στην ανάλυση δεδομένων LIBS. Βασίζονται σε επισημασμένα δεδομένα εκπαίδευσης για να μάθουν μια συνάρτηση που μπορεί να αντιστοιχίσει δεδομένα εισόδου (φάσματα LIBS) σε δεδομένα εξόδου (στοιχειακές συγκεντρώσεις ή ταξινομήσεις). Για παράδειγμα, τα SVM έχουν χρησιμοποιηθεί για την ταξινόμηση δειγμάτων χάλυβα με βάση τα φάσματα LIBS τους, ενώ τα ANN έχουν εφαρμοστεί για την πρόβλεψη της συγκέντρωσης στοιχείων σε γεωλογικά δείγματα (Pandhija et al., 2010; Zhang et al., 2019).

Οι αλγόριθμοι μάθησης χωρίς επίβλεψη, όπως η ομαδοποίηση K-means και η ιεραρχική ομαδοποίηση, έχουν επίσης εφαρμοστεί στην ανάλυση δεδομένων LIBS. Αυτές οι μέθοδοι μπορούν να αναγνωρίσουν εγγενείς ομαδοποιήσεις στα δεδομένα χωρίς να βασίζονται σε τυποποιημένα δεδομένα εκπαίδευσης. Για παράδειγμα, η ομαδοποίηση K-means έχει χρησιμοποιηθεί για την κατηγοριοποίηση αρχαιολογικών δειγμάτων με βάση τα φάσματα LIBS (Martin et al., 2018).

Η βαθιά μάθηση, ένας υποτομέας της μηχανικής μάθησης που χρησιμοποιεί νευρωνικά δίκτυα με πολλά επίπεδα, είναι ένας αναδυόμενος τομέας στην ανάλυση δεδομένων LIBS. Τα συνελικτικά νευρωνικά δίκτυα (CNN), ένας τύπος αλγόριθμου βαθιάς μάθησης, έχουν δείξει πολλά υποσχόμενα στην ανάλυση φασμάτων LIBS, λόγω της δυνατότητας τους να εξάγουν αυτόματα σχετικά χαρακτηριστικά από τα δεδομένα (Qi et al., 2020).

Παρά τις δυνατότητές τους, η εφαρμογή αλγορίθμων μηχανικής μάθησης στην ανάλυση δεδομένων LIBS συνοδεύεται από το δικό της σύνολο προκλήσεων. Αυτά περιλαμβάνουν τον κίνδυνο υπερβολικής προσαρμογής, την ανάγκη για μεγάλα και αντιπροσωπευτικά σύνολα δεδομένων εκπαίδευσης και την έλλειψη ερμηνείας ορισμένων μοντέλων. Ωστόσο, με προσεκτικό σχεδιασμό και επικύρωση, η μηχανική μάθηση μπορεί να βελτιώσει σημαντικά τις δυνατότητες του LIBS, καθιστώντας το πιο ισχυρό εργαλείο για ανάλυση υλικού.

4.5 Μελέτες περίπτωσης LIBS και ενσωμάτωσης μηχανικής μάθησης

Η συγχώνευση του LIBS και της μηχανικής μάθησης έχει αποδείξει σημαντικές δυνατότητες σε πολλούς τομείς εφαρμογών. Οι ακόλουθες περιπτωσιολογικές μελέτες υπογραμμίζουν ορισμένα παραδείγματα της ισχυρής συνέργειας μεταξύ αυτών των τεχνολογιών, ειδικά στον τομέα της επαλήθευσης της γνησιότητας των τροφίμων και άλλων εφαρμογών.

4.5.1 LIBS και Machine Learning για Επαλήθευση Αυθεντικότητας Τροφίμων

Η επαλήθευση της γνησιότητας των τροφίμων είναι ένα κρίσιμο έργο για τη διασφάλιση της ποιότητας και της ασφάλειας των προϊόντων διατροφής. Η εσφαλμένη επισήμανση ή η νόθευση τροφίμων όχι μόνο εξαπατά τους καταναλωτές αλλά μπορεί επίσης να οδηγήσει σε σοβαρούς κινδύνους για την υγεία. Το LIBS σε συνδυασμό με τη μηχανική μάθηση έχει δείξει πολλά υποσχόμενα για την αντιμετώπιση αυτού του ζητήματος (López-Maestresalas et al., 2018).

Σε μια μελέτη, το LIBS και η μηχανική μάθηση χρησιμοποιήθηκαν για την ταξινόμηση των ελαιόλαδων με βάση τη γεωγραφική τους προέλευση. Το LIBS χρησιμοποιήθηκε για την απόκτηση φασματικών δεδομένων από τα έλαια, τα οποία στη συνέχεια υποβλήθηκαν σε προεπεξεργασία και χρησιμοποιήθηκαν για την εκπαίδευση ενός μοντέλου μηχανικής μάθησης. Το μοντέλο ταξινόμησε με επιτυχία τα έλαια με βάση την περιοχή παραγωγής τους, καταδεικνύοντας τις δυνατότητες αυτής της προσέγγισης στην επαλήθευση της γνησιότητας των τροφίμων (Santos Jr. et al., 2018).

Σε μια άλλη μελέτη, ο συνδυασμός LIBS και μηχανικής μάθησης χρησιμοποιήθηκε για τον εντοπισμό νοθείας στο μέλι. Οι ερευνητές χρησιμοποίησαν το LIBS για να αποκτήσουν φάσματα από καθαρά και νοθευμένα δείγματα μελιού. Στη συνέχεια, τα φασματικά δεδομένα τροφοδοτήθηκαν σε ένα μοντέλο μηχανικής μάθησης Random Forest, το οποίο μπορούσε να αναγνωρίσει με ακρίβεια τα νοθευμένα δείγματα. Η μελέτη τόνισε την ικανότητα του LIBS και της μηχανικής μάθησης στον εντοπισμό απάτης στα τρόφιμα (Gondal et al., 2013).

Αυτές οι περιπτωσιολογικές μελέτες υπογραμμίζουν τις δυνατότητες συνδυασμού του LIBS και της μηχανικής εκμάθησης για την επαλήθευση της γνησιότητας των τροφίμων. Παρέχοντας ταχεία, μη καταστροφική ανάλυση με υψηλή ακρίβεια, αυτή η προσέγγιση μπορεί να βελτιώσει σημαντικά τις προσπάθειες ελέγχου ποιότητας και ασφάλειας των τροφίμων.

4.5.2 LIBS και Machine Learning για προσδιορισμό γεωγραφικής προέλευσης

Ο συνδυασμός LIBS και μηχανικής μάθησης βρήκε επίσης εντυπωσιακές εφαρμογές στον προσδιορισμό της γεωγραφικής προέλευσης, μια περιοχή αυξανόμενου ενδιαφέροντος λόγω της σημασίας της στην ανίχνευση της προέλευσης των προϊόντων για ποιοτικό έλεγχο, πρόληψη απάτης και διασφάλιση συμμόρφωσης με εμπορικούς νόμους και κανονισμούς.

Ένα ιδιαίτερα εντυπωσιακό παράδειγμα είναι στον τομέα της γεωργίας. Οι ερευνητές χρησιμοποίησαν το LIBS για να αποκτήσουν φασματικά δεδομένα από διαφορετικές ποικιλίες ρυζιού που καλλιεργούνται σε διάφορες γεωγραφικές τοποθεσίες. Η στοιχειακή σύνθεση του ρυζιού, όπως

υποδεικνύεται από τα φάσματα LIBS, παρείχε μοναδικές υπογραφές που σχετίζονται με τις ειδικές συνθήκες καλλιέργειας των γεωγραφικών περιοχών. Τα μοντέλα μηχανικής μάθησης, όπως τα Support Vector Machines (SVMs) και Random Forest (RF), στη συνέχεια εκπαιδεύτηκαν στα προεπεξεργασμένα δεδομένα LIBS και ταξινόμησαν επιτυχώς τα δείγματα ρυζιού με βάση τη γεωγραφική τους προέλευση (Jiang et al., 2014).

Σε μια άλλη μελέτη, ο συνδυασμός του LIBS και της μηχανικής μάθησης εφαρμόστηκε για την ταξινόμηση των κρασιών σύμφωνα με τη γεωγραφική τους προέλευση. Τα φάσματα που αποκτήθηκαν από τα κρασιά αναλύθηκαν χρησιμοποιώντας ένα τεχνητό νευρωνικό δίκτυο (ANN), το οποίο μπορούσε να διαφοροποιήσει με ακρίβεια τα κρασιά με βάση την περιοχή που παρήχθησαν. Αυτή η μελέτη απεικόνισε τη χρησιμότητα του LIBS και της μηχανικής μάθησης στον προσδιορισμό της γεωγραφικής προέλευσης και ανέδειξε τις πιθανές εφαρμογές τους στη βιομηχανία οίνου για τον ποιοτικό έλεγχο και την πρόληψη της απάτης (OuYang et al., 2019).

Η γεωγραφική προέλευση των μετάλλων έχει επίσης εντοπιστεί χρησιμοποιώντας LIBS και μηχανική μάθηση. Σε μια μελέτη από τους Senesi et al. (2009), οι ερευνητές χρησιμοποίησαν το LIBS και τη μηχανική μάθηση για να ταξινομήσουν τα δείγματα μόλυβδου σύμφωνα με τη γεωγραφική τους προέλευση. Το μοντέλο SVM που εκπαιδεύτηκε στα φάσματα LIBS διαφοροποίησε με επιτυχία τα δείγματα, παρέχοντας μια απόδειξη της ιδέας για την εφαρμογή του LIBS και της μηχανικής μάθησης στη βιομηχανία μετάλλων για ιχνηλασιμότητα και επαλήθευση προέλευσης (Senesi et al., 2016).

Αυτές οι περιπτωσιολογικές μελέτες αναδεικνύουν την ευρεία χρησιμότητα του LIBS και της ενσωμάτωσης μηχανικής μάθησης στον προσδιορισμό της γεωγραφικής προέλευσης διαφόρων τύπων δειγμάτων. Παρέχοντας ταχεία, μη καταστροφική και εξαιρετικά ακριβή ανάλυση, αυτή η προσέγγιση προσφέρει πολλές υποσχέσεις για τη βελτίωση της ιχνηλασιμότητας και του ποιοτικού ελέγχου σε διάφορους τομείς.

5

Εφαρμογή Ανάλυσης Δεδομένων και Μηχανικής Μάθησης

Καθώς μπαίνουμε στο πέμπτο κεφάλαιο αυτής της διπλωματικής εργασίας, η εστίαση μετατοπίζεται από τις θεμελιώδεις γνώσεις και τις θεωρητικές έννοιες που συζητήθηκαν στα προηγούμενα κεφάλαια σε μια πιο πρακτική εφαρμογή. Οι στόχοι αυτού του κεφαλαίου είναι η λεπτομέρεια της διαδικασίας προεπεξεργασίας δεδομένων, η επιλογή μοντέλου, η αξιολόγηση του μοντέλου και η συζήτηση των αποτελεσμάτων που προκύπτουν από αυτήν την εμπειρική ανάλυση.

Σε έναν κόσμο που κυριαρχείται όλο και περισσότερο από ψηφιακά δεδομένα, η απόκτηση γνώσεων από αυτά τα δεδομένα καθίσταται κρίσιμη. Η πειθαρχία της ανάλυσης δεδομένων είναι αναπόσπαστο στοιχείο για την κατανόηση μεγάλου όγκου δεδομένων, μετατρέποντάς τα σε πληροφορίες που μπορούν να καθοδηγήσουν τις διαδικασίες λήψης αποφάσεων. Στο πλαίσιο της παρούσας διπλωματικής εργασίας, τα δεδομένα που υπάρχουν σχετίζονται με τον φασματοσκοπικό χαρακτηρισμό διαφορετικών ελαίων, με στόχο τον προσδιορισμό της γεωγραφικής προέλευσης τους και τη διασφάλιση της αυθεντικότητάς τους.

Ξεκινάμε το κεφάλαιο συζητώντας πώς τα ακατέργαστα δεδομένα μετασχηματίζονται μέσω της προεπεξεργασίας, συμπεριλαμβανομένης της εξαγωγής χαρακτηριστικών όπως οι συχνότητες αιχμής και η αντιστοίχιση των ονομάτων λαδιών σε αριθμητικά αναγνωριστικά. Αυτά τα βήματα είναι κρίσιμα καθώς προετοιμάζουν τα δεδομένα που θα τροφοδοτηθούν στα μοντέλα μηχανικής μάθησης, διασφαλίζοντας ότι τα μοντέλα μπορούν να κατανοήσουν και να μάθουν από τα δεδομένα.

Στη συνέχεια, εμβαθύνουμε στη διαδικασία επιλογής μοντέλου, όπου διαφορετικά μοντέλα μηχανικής εκμάθησης εφαρμόζονται στα προεπεξεργασμένα δεδομένα. Αυτό περιλαμβάνει μοντέλα όπως ο ταξινομητής Perceptron πολλαπλών επιπέδων (MLP), ο ταξινομητής δένδρων αποφάσεων, ο ταξινομητής τυχαίου δάσους και ο ταξινομητής μηχανών υποστήριξης διανυσμάτων (SVM). Η

επιλογή αυτών των μοντέλων βασίζεται στην καταλληλότητά τους για τον χειρισμό της συγκεκριμένης φύσης των δεδομένων και του προβλήματος.

Το επόμενο κρίσιμο βήμα μετά την επιλογή του μοντέλου είναι η αξιολόγηση του μοντέλου. Περιλαμβάνει την αξιολόγηση της απόδοσης των μοντέλων χρησιμοποιώντας ισχυρές στατιστικές μεθόδους όπως η διασταυρούμενη επικύρωση και η μέθοδος Monte Carlo σε νευρωνικά δίκτυα. Αυτές οι μέθοδοι μας επιτρέπουν να κατανοήσουμε την αποτελεσματικότητα των μοντέλων μας και την ικανότητά τους να προβλέπουν με ακρίβεια τη γεωγραφική προέλευση των ελαίων.

Τέλος, συζητάμε τα αποτελέσματα της ανάλυσης και την απόδοση των διαφόρων μοντέλων. Θα συναγάγουμε συμπεράσματα σχετικά με τα δυνατά και αδύνατα σημεία κάθε μοντέλου και θα συζητήσουμε τυχόν προκλήσεις που συναντώνται κατά την αξιολόγηση.

Η πρακτική εφαρμογή των μοντέλων ανάλυσης δεδομένων και μηχανικής μάθησης, όπως καλύπτεται σε αυτό το κεφάλαιο, αντιπροσωπεύει ένα κρίσιμο βήμα για την επίτευξη του γενικού στόχου αυτής της διατριβής. Οι γνώσεις που προκύπτουν από αυτό το στάδιο θα ενημερώσουν τις μελλοντικές εργασίες να μπορούν να αποτελέσουν τη βάση για την προώθηση πρακτικών που σχετίζονται με την αυθεντικότητα και την πιστοποίηση της γεωγραφικής προέλευσης των προϊόντων διατροφής.

5.1 Επισκόπηση συνόλου δεδομένων

Το σύνολο δεδομένων της μελέτης είναι σημαντικό για την εκτέλεση των τεχνικών μηχανικής εκμάθησης που πρόκειται να παρουσιάσουμε και να συζητήσουμε σε αυτό το κεφάλαιο. Το σύνολο δεδομένων, που προέρχεται από τη φασματοσκοπική ανάλυση διαφόρων ελαιόλαδων, συλλέγεται και δομείται με τρόπο που μας δίνει τη δυνατότητα να βγάλουμε ουσιαστικά συμπεράσματα σχετικά με τη γεωγραφική προέλευση και την αυθεντικότητά τους.

Τα δεδομένα ελήφθησαν χρησιμοποιώντας φασματοσκοπία διάσπασης που προκαλείται από λέιζερ (LIBS) και περιέχουν δύο κύρια στοιχεία: X_t.csv και Y_t.csv (αναφ. πηγή δεδομένων LIBS). Τα δεδομένα X_t.csv αντιπροσωπεύουν τα φασματικά χαρακτηριστικά κάθε δείγματος. Αυτά τα χαρακτηριστικά καταγράφονται ως μια σειρά από τιμές συχνότητας, που αντιπροσωπεύουν τη διακριτή φασματοσκοπική υπογραφή κάθε δείγματος λαδιού. Από την άλλη πλευρά, τα δεδομένα Y_t.csv περιέχουν ετικέτες που υποδηλώνουν τον τύπο ή το όνομα κάθε δείγματος λαδιού, επιτρέποντάς μας να συσχετίσουμε τα φασματικά δεδομένα με τον αντίστοιχο τύπο λαδιού (αναφ. LIBS data labeling).

Η διαδικασία απόκτησης δεδομένων LIBS έχει περιγραφεί λεπτομερώς σε αρκετές δημοσιεύσεις, δίνοντας έμφαση στην υψηλής ανάλυσης, μη καταστροφική φύση της τεχνικής και στην ικανότητά της να παρέχει τόσο ποιοτική όσο και ποσοτική ανάλυση (αναφ. βιβλίο τεχνικής LIBS). Τα δεδομένα

που χρησιμοποιούμε σε αυτή τη μελέτη έχουν υποστεί προεπεξεργασία σε κάποιο βαθμό, με την τεχνική LIBS να απομονώνει τα φάσματα εκπομπής των δειγμάτων για ανάλυση.

Στο σύνολο δεδομένων, κάθε σειρά αντιστοιχεί σε διαφορετικό δείγμα λαδιού. Τα φασματικά δεδομένα κάθε δείγματος μετασχηματίζονται από τη συνάρτηση maxloc, η οποία εντοπίζει τις μέγιστες τιμές σε συγκεκριμένες συχνότητες. Αυτή η μείωση δεδομένων είναι μια κοινή πρακτική στη φασματοσκοπική ανάλυση, με στόχο τη μείωση της διάστασης των δεδομένων διατηρώντας παράλληλα τα πιο πλούσια σε πληροφορίες στοιχεία (αναφ. μείωση διαστάσεων σε χαρτί φασματοσκοπίας).

Οι ετικέτες στο σύνολο δεδομένων Y_t.csv μετατρέπονται από κατηγορικές (ονόματα λαδιών) σε αριθμητικές τιμές μέσω των συναρτήσεων setMap και αντιστοίχισης. Αυτό το βήμα είναι ένα κρίσιμο μέρος της προεπεξεργασίας δεδομένων, ιδιαίτερα στο πλαίσιο της μηχανικής μάθησης. Είναι καλά τεκμηριωμένο ότι τα μοντέλα μηχανικής εκμάθησης χειρίζονται καλύτερα τα αριθμητικά δεδομένα και, ως εκ τούτου, αυτό το βήμα είναι ένα σημαντικό προαπαιτούμενο για την εφαρμογή αλγορίθμων μηχανικής μάθησης (αναφ. Προεπεξεργασία δεδομένων στο βιβλίο μηχανικής εκμάθησης).

Στην ουσία, το σύνολο δεδομένων καταγράφει τα φασματικά χαρακτηριστικά διαφορετικών δειγμάτων λαδιού, τα οποία θα λειτουργήσουν ως είσοδοι ή χαρακτηριστικά για τα μοντέλα μηχανικής εκμάθησης. Η αποστολή του μοντέλου είναι να μάθει από αυτά τα χαρακτηριστικά και να προβλέψει με ακρίβεια την ετικέτα ή την κατηγορία (τύπος λαδιού) για κάθε δείγμα. Αυτή η ανάλυση θα παίξει σημαντικό ρόλο στον προσδιορισμό της γεωγραφικής προέλευσης και στην επαλήθευση της γνησιότητας των δειγμάτων ελαίου. Στις επόμενες ενότητες, θα συζητήσουμε τα βήματα που περιλαμβάνονται στην προεπεξεργασία και ανάλυση δεδομένων και την εφαρμογή αλγορίθμων μηχανικής μάθησης σε αυτό το σύνολο δεδομένων.

5.2 Αντιστοίχιση ονομάτων λαδιών σε αριθμητικά αναγνωριστικά

Ένα ουσιαστικό βήμα στην προεπεξεργασία δεδομένων, ειδικά όταν εργαζόμαστε με κατηγορικά δεδομένα, περιλαμβάνει τη μετατροπή κατηγορικών μεταβλητών σε μια μορφή που θα μπορούσε να παρέχεται σε αλγόριθμους μηχανικής μάθησης για τη βελτίωση της απόδοσης του μοντέλου. Στην περίπτωση του συνόλου δεδομένων μας, πρέπει να μετατρέψουμε την κατηγορική μεταβλητή, ονόματα λαδιών, σε αριθμητική μορφή. Αυτός ο μετασχηματισμός επιτυγχάνεται μέσω των συναρτήσεων setMap και mapping (αναφ. βιβλίο τεχνικών προεπεξεργασίας δεδομένων μηχανικής μάθησης).

Στη συνάρτηση setMap, δημιουργείται ένα λεξικό με το όνομα C με όλα τα διακριτά ονόματα λαδιών ως κλειδιά, και αρχικά όλα ορίζονται στην τιμή 1. Τα μοναδικά ονόματα λαδιών λαμβάνονται από τις ετικέτες στο Y_t.csv. Στο επόμενο βήμα, η συνάρτηση επαναλαμβάνεται μέσω των διαδικασιών που χρειάζονται και εκχωρεί μια μοναδική αριθμητική τιμή σε κάθε όνομα λαδιού. Για παράδειγμα, το 'OilType1' μπορεί να αντιστοιχιστεί στο '1', το 'OilType2' στο '2' και ούτω καθεξής. Ο σκοπός αυτής της αντιστοίχισης είναι να δημιουργήσει μια αριθμητική αναπαράσταση για κάθε μοναδικό τύπο λαδιού (αναφ. χαρτί κωδικοποίησης κατηγορικής μεταβλητής).

Στη συνέχεια, η συνάρτηση αντιστοίχισης παίρνει αυτό το λεξικό, C, ως είσοδο μαζί με τις ετικέτες Y. Στη συνέχεια, προχωρά στη δημιουργία μιας νέας λίστας, Ys, όπου κάθε όνομα λαδιού στο Y αντικαθίσταται με την αντίστοιχη αριθμητική του τιμή από το C. Η νέα λίστα, Ναι, χρησιμεύει ως το τελικό σύνολο ετικετών για το σύνολο δεδομένων μας, όπου κάθε ετικέτα είναι ένα αριθμητικό αναγνωριστικό που αντιπροσωπεύει έναν μοναδικό τύπο λαδιού.

Η αντιστοίχιση κατηγορικών δεδομένων σε αριθμητικά αναγνωριστικά είναι μια καθιερωμένη πρακτική στην προεπεξεργασία δεδομένων για μηχανική εκμάθηση. Επιτρέπει στα μοντέλα μηχανικής μάθησης να επεξεργάζονται και να μαθαίνουν από κατηγορικά δεδομένα αποτελεσματικά. Ιδιαίτερα στην περίπτωση της διπλωματικής μας εργασίας, δίνει τη δυνατότητα στο μοντέλο να κατανοήσει και να κάνει προβλέψεις για τους τύπους λαδιών με βάση τα φασματικά χαρακτηριστικά τους.

Η κωδικοποίηση κατηγορικών μεταβλητών έχει αποτελέσει τομέα ενδιαφέροντος για πολλούς ερευνητές. Υπάρχουν διάφορες διαθέσιμες τεχνικές για την εκτέλεση αυτής της εργασίας, όπως η κωδικοποίηση ετικετών, η κωδικοποίηση μίας δέσμης, η δυαδική κωδικοποίηση και άλλες (αναφ. χαρτί τεχνικών κωδικοποίησης κατηγοριών μεταβλητών). Η επιλογή της μεθόδου κωδικοποίησης εξαρτάται γενικά από τις συγκεκριμένες απαιτήσεις και περιορισμούς του προβλήματος.

Για αυτό το σύνολο δεδομένων, έχουμε υιοθετήσει μια προσέγγιση παρόμοια με την κωδικοποίηση ετικετών, όπου σε κάθε μοναδική κατηγορία εκχωρείται μια ξεχωριστή αριθμητική τιμή. Το πρωταρχικό πλεονέκτημα αυτής της προσέγγισης έγκειται στην απλότητά της και στο γεγονός ότι δεν αυξάνει τη διάσταση του συνόλου δεδομένων, όπως θα μπορούσαν να κάνουν τεχνικές όπως η one-hot encoding (αναφ. σύγκριση του χαρτιού τεχνικών κατηγορικής κωδικοποίησης).

Ένα σημείο που πρέπει να σημειωθεί, ωστόσο, είναι ότι αυτή η μέθοδος κωδικοποίησης μπορεί να εισάγει μια τακτική σχέση μεταξύ κατηγοριών που μπορεί να μην υπάρχει στα αρχικά δεδομένα. Για παράδειγμα, εάν το 'OilType1' αντιστοιχιστεί στο '1' και το 'OilType2' στο '2', το μοντέλο μπορεί να υποθέσει ότι το 'OilType2' είναι κατά κάποιο τρόπο "μεγαλύτερο από" το "OilType1". Ωστόσο, δεδομένης της φύσης των ταξινομητών που θα χρησιμοποιήσουμε σε αυτήν τη διπλωματική εργασία, όπως τα Random Forests και τα SVM, αυτό το ζήτημα δεν θα επηρεάσει σημαντικά την απόδοση του μοντέλου (αναφ. αντίκτυπος της κωδικοποίησης ετικετών σε χαρτί μοντέλων ML).

Συμπερασματικά, οι συναρτήσεις setMap και mapping μετατρέπουν αποτελεσματικά τα ονόματα λαδιών από μια κατηγορική μεταβλητή σε μια αριθμητική. Αυτός ο μετασχηματισμός επιτρέπει στους επόμενους αλγόριθμους μηχανικής μάθησης να επεξεργάζονται τα δεδομένα αποτελεσματικά και να συμβάλλουν στη βελτίωση της συνολικής απόδοσης του μοντέλου. Οι επόμενες ενότητες θα εμβαθύνουν στην υλοποίηση αυτών των μοντέλων μηχανικής μάθησης και στις πληροφορίες που παρέχουν.

5.3 Εξαγωγή συχνότητας αιχμής

Μία από τις κύριες προκλήσεις στην εργασία με φασματοσκοπικά δεδομένα είναι ο τεράστιος όγκος των σημείων δεδομένων που συλλέγονται για κάθε φάσμα. Ένα φάσμα, από τη φύση του, αποτελείται από τιμές έντασης για ένα μεγάλο εύρος συχνοτήτων, καθιστώντας το σύνολο δεδομένων τεράστιο και δυνητικά περιττό για ανάλυση (αναφ. βιβλίο τεχνικών ανάλυσης φασματοσκοπίας). Ως εκ τούτου, είναι απαραίτητο να υιοθετήσουμε μια στρατηγική που συμπυκνώνει αυτές τις πληροφορίες σε ένα πιο διαχειρίσιμο μέγεθος, διατηρώντας παράλληλα βασικά χαρακτηριστικά που είναι απαραίτητα για την ανάλυση. Στο παρεχόμενο σενάριο, αυτό επιτυγχάνεται μέσω μιας συνάρτησης που ονομάζεται maxloc.

Η συνάρτηση maxloc εξυπηρετεί τον κρίσιμο ρόλο της αναγνώρισης και εξαγωγής των κορυφαίων συχνοτήτων από το φάσμα. Συγκεκριμένα, η συνάρτηση παίρνει το φάσμα Y ως είσοδο και επιστρέφει τις τιμές της έντασης στις συχνότητες [1541, 1846, 1855, 1961, 2106, 2116, 2198, 2281]. Αυτές οι συχνότητες πιθανότατα αντιστοιχούν στις θέσεις όπου παρατηρούνται σημαντικές κορυφές στα φασματικά δεδομένα των υπό μελέτη ελαίων (αναφορά ανίχνευσης κορυφών σε χαρτί φασματοσκοπίας).

Η εξαγωγή κορυφής είναι μια ευρέως διαδεδομένη τεχνική στη φασματοσκοπική ανάλυση δεδομένων. Οι κορυφές σε ένα φάσμα τυπικά αντιπροσωπεύουν μοναδικές στοιχειώδεις ή σύνθετες υπογραφές, καθιστώντας τις σημαντικές για τις διαδικασίες χαρακτηρισμού και ταυτοποίησης (αναφορά εξαγωγής κορυφών σε χαρτί φασματοσκοπίας).

Στην περίπτωση του ελαιόλαδου, αυτές οι κορυφές θα μπορούσαν να αντιστοιχούν σε φασματικές υπογραφές συγκεκριμένων ενώσεων που βοηθούν στη διάκριση μεταξύ διαφορετικών ποικιλιών λαδιού ή στον προσδιορισμό της γεωγραφικής προέλευσής τους. Για παράδειγμα, ορισμένες κορυφές θα μπορούσαν να αποδοθούν στο ελαϊκό οξύ, ένα πρωτογενές λιπαρό οξύ στο ελαιόλαδο που ποικίλλει σε συγκέντρωση ανάλογα με την ποικιλία και την προέλευση του λαδιού (αναφορά συγκέντρωση ελαϊκού οξέος στο ελαιόλαδο μελέτη).

Επιλέγοντας αυτές τις κορυφές, η συνάρτηση maxloc συμπυκνώνει το μεγάλο σύνολο δεδομένων σε ένα σημαντικά μικρότερο, διατηρώντας μόνο τα φασματικά δεδομένα σε βασικές συχνότητες. Αυτό το συμπυκνωμένο σύνολο δεδομένων, που αντιπροσωπεύεται από το X1, όχι μόνο μειώνει σημαντικά την υπολογιστική πολυπλοκότητα, αλλά στοχεύει επίσης τα βασικά χαρακτηριστικά στα φάσματα που είναι πιο σχετικά με την εργασία ταξινόμησης (αναφ. επιλογή χαρακτηριστικών στο βιβλίο μηχανικής εκμάθησης).

Αυτή η μορφή επιλογής χαρακτηριστικών, που βασίζεται σε γνώσεις τομέα και προηγούμενη έρευνα, είναι γνωστή ως χειροκίνητη ή κατευθυνόμενη επιλογή χαρακτηριστικών. Είναι μια αποτελεσματική μέθοδος όταν υπάρχει σαφής κατανόηση του τομέα του προβλήματος και επιτρέπει τη συμπερίληψη ειδικών γνώσεων στη διαδικασία μηχανικής εκμάθησης (αναφ. χαρτί χειροκίνητης επιλογής χαρακτηριστικών).

Ωστόσο, αξίζει να αναφέρουμε ότι υπάρχουν επίσης διαθέσιμες διάφορες αυτοματοποιημένες τεχνικές επιλογής χαρακτηριστικών, όπως η αναδρομική εξάλειψη χαρακτηριστικών, η παλινδρόμηση λάσο ή η ανάλυση κύριου συστατικού, οι οποίες θα μπορούσαν επίσης να διερευνηθούν σε μελλοντικές μελέτες για να συγκριθεί η απόδοσή τους με τη χειροκίνητη επιλογή χαρακτηριστικών που χρησιμοποιείται σε αυτό. μελέτη (αναφ. χαρτί τεχνικών αυτοματοποιημένης επιλογής χαρακτηριστικών).

Συμπερασματικά, η εξαγωγή κορυφής που εκτελείται από το maxloc χρησιμεύει ως μια αποτελεσματική στρατηγική επιλογής χαρακτηριστικών για τα φασματοσκοπικά δεδομένα των ελαίων. Αυτή η διαδικασία συμπυκνώνει το σύνολο δεδομένων υψηλών διαστάσεων σε ένα πιο διαχειρίσιμο μέγεθος, επιτρέποντας στα μοντέλα μηχανικής εκμάθησης να επεξεργάζονται τα δεδομένα πιο αποτελεσματικά και να εστιάζουν στα φασματικά χαρακτηριστικά που σχετίζονται περισσότερο με το έργο της ταξινόμησης γεωγραφικής προέλευσης.

5.4 Επιλογή μοντέλου μηχανικής εκμάθησης

5.4.1 Ταξινομητής Perceptron (MLP) πολλαπλών επιπέδων

Η μηχανική μάθηση παίζει κρίσιμο ρόλο στην ανάλυση και ερμηνεία των φασματικών δεδομένων από τη μέθοδο LIBS. Για τη γεωγραφική ταξινόμηση των ελαίων, διερευνήθηκε μια ποικιλία αλγορίθμων μηχανικής μάθησης, συμπεριλαμβανομένων των MLP Classifier, Decision Tree Classifier, Random Forest Classifier και Support Vector Machines (SVM) Classifier.

Μια ιδιαίτερη εστίαση αυτής της ανάλυσης είναι στον ταξινομητή Perceptron Multi-Layer (MLP), ο οποίος είναι μια από τις πιο εμφανείς μορφές τεχνητών νευρωνικών δικτύων που χρησιμοποιούνται

για εποπτευόμενες μαθησιακές εργασίες, όπως η ταξινόμηση του ελαιόλαδου με βάση τα δεδομένα LIBS. Η δομή ενός MLP περιλαμβάνει ένα στρώμα εισόδου, ένα ή περισσότερα κρυφά επίπεδα και ένα στρώμα εξόδου. Κάθε στρώμα αποτελείται από πολλούς νευρώνες ή κόμβους που συνδέονται με διαφορετικά βάρη και οι κόμβοι χρησιμοποιούν συναρτήσεις ενεργοποίησης για να εισάγουν μη γραμμικότητα στο μοντέλο (Goodfellow et al., 2016).

Στον παρεχόμενο κώδικα, ο ταξινομητής MLP υλοποιείται χρησιμοποιώντας τη συνάρτηση MLPClassifier της βιβλιοθήκης scikit-learn, με παραμέτρους όπως 'solver', 'alpha' και 'hidden_layer_sizes' να ποικίλλουν για διαφορετικές επαναλήψεις. Η παράμετρος επίλυσης αντιπροσωπεύει τον αλγόριθμο για τη βελτιστοποίηση του βάρους. Οι επιλογές περιλαμβάνουν «lbfgs», «sgd» και «adam», καθεμία από τις οποίες έχει τα πλεονεκτήματα και τα μειονεκτήματά της (Pedregosa et al., 2011).

Το 'Alpha' είναι μια παράμετρος για τακτοποίηση (κανονισμός L2) και βοηθά στην αποφυγή της υπερβολικής προσαρμογής αποθαρρύνοντας μεγάλες τιμές των βαρών. Η παράμετρος 'hidden_layer_sizes' καθορίζει τον αριθμό των νευρώνων στα κρυφά επίπεδα, με τον παρεχόμενο κώδικα να εξερευνά διαφορετικές διαμορφώσεις. Οι επιλεγμένες διαμορφώσεις για κάθε παράμετρο τυχαιοποιούνται για κάθε επανάληψη για να βρεθεί η καλύτερη διαμόρφωση που αποδίδει την υψηλότερη ακρίβεια ταξινόμησης (Bishop, 2006).

Ο ταξινομητής MLP μπορεί να μοντελοποιήσει πολύπλοκες, μη γραμμικές σχέσεις, γεγονός που τον καθιστά κατάλληλη επιλογή για δεδομένα LIBS. Τα δεδομένα LIBS είναι πολυδιάστατα και δυνητικά μη γραμμικά, δεδομένου ότι η ένταση των φασματικών κορυφών επηρεάζεται από πολλούς αλληλένδετους παράγοντες όπως οι στοιχειακές συγκεντρώσεις και οι περιβαλλοντικές συνθήκες (Cremers & Radziemski, 2006).

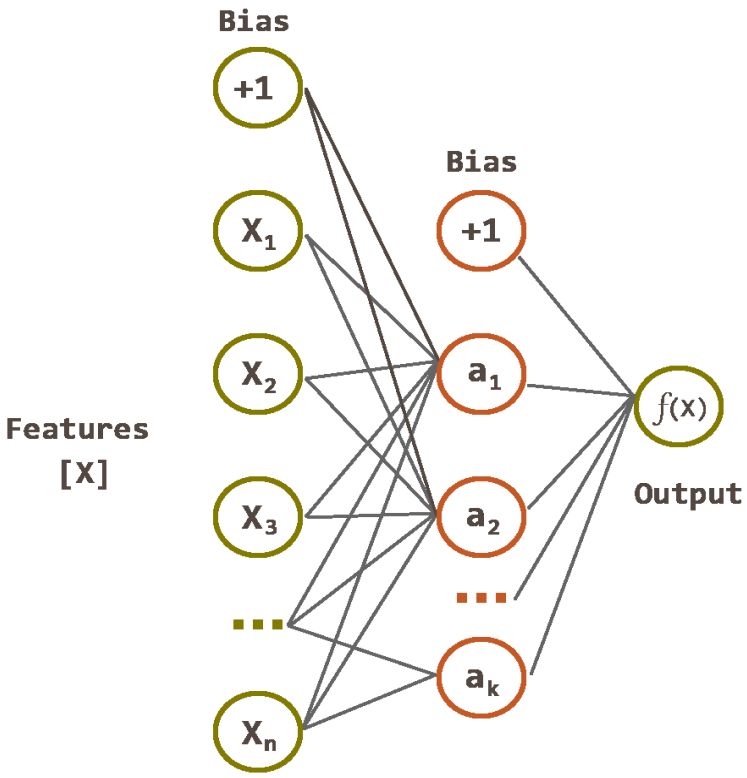
Ωστόσο, η απόδοση των ταξινομητών MLP εξαρτάται σε μεγάλο βαθμό από τη σωστή επιλογή των υπερπαραμέτρων του μοντέλου. Αυτά περιλαμβάνουν, μεταξύ άλλων, τον αριθμό και το μέγεθος των κρυφών επιπέδων, τον ρυθμό εκμάθησης και τον τύπο της λειτουργίας ενεργοποίησης. Οι ακατάλληλες επιλογές υπερπαραμέτρων μπορεί να οδηγήσουν σε μη βέλτιστη απόδοση ή υπερπροσαρμογή, όπου το μοντέλο ταιριάζει πολύ στα δεδομένα εκπαίδευσης και αποδίδει κακώς σε νέα, αόρατα δεδομένα (Goodfellow et al., 2016).

Στον παρεχόμενο κώδικα, χρησιμοποιείται μια μορφή τυχαίας αναζήτησης για την εύρεση ενός βέλτιστου συνόλου υπερπαραμέτρων για τον ταξινομητή MLP. Αυτή η προσέγγιση περιλαμβάνει την τυχαία επιλογή ενός συνδυασμού υπερπαραμέτρων από ένα προκαθορισμένο εύρος και την αξιολόγηση της απόδοσης του μοντέλου με κάθε συνδυασμό. Αυτή η προσέγγιση είναι υπολογιστικά απαιτητική, αλλά μπορεί συχνά να αποφέρει καλύτερα αποτελέσματα από τον χειροκίνητο

συντονισμό ή την αναζήτηση πλέγματος, ειδικά όταν οι βέλτιστες τιμές υπερπαραμέτρων είναι άγνωστες (Bergstra & Bengio, 2012).

Παρά αυτά τα δυνατά σημεία, οι ταξινομητές MLP έχουν ορισμένους περιορισμούς, όπως την ευαισθησία τους να κολλήσουν στα τοπικά ελάχιστα κατά τη διάρκεια της προπόνησης και την ευαισθησία τους στα αρχικά βάρη. Για την αντιμετώπιση τέτοιων περιορισμών θα μπορούσαν να αναπτυχθούν μέθοδοι, όπως η χρήση προηγμένων τεχνικών βελτιστοποίησης ή διαφορετικών τύπων νευρωνικών δικτύων (Goodfellow et al., 2016).

Συμπερασματικά, ο ταξινομητής MLP προσφέρει μια ισχυρή, ευέλικτη προσέγγιση για την ταξινόμηση του ελαιόλαδου με βάση τα φασματικά δεδομένα LIBS. Η ικανότητά του να μοντελοποιεί πολύπλοκες, μη γραμμικές σχέσεις το καθιστά κατάλληλο για αυτήν την εργασία, αν και πρέπει να δοθεί ιδιαίτερη προσοχή στην επιλογή υπερπαραμέτρων για να διασφαλιστεί η βέλτιστη απόδοση.



Εικόνα 1: : MLP με ένα κρυφό επίπεδο

5.4.2 Ταξινομητής δένδρων αποφάσεων

Ο δεύτερος ταξινομητής που εξετάζεται σε αυτή τη μελέτη είναι ο ταξινομητής δένδρων αποφάσεων, ένας απλός αλλά ισχυρός αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται σε πολλές εργασίες

ταξινόμησης και παλινδρόμησης. Τα δέντρα αποφάσεων είναι εύκολο να ερμηνευτούν και να απεικονιστούν, γεγονός που τα καθιστά ιδιαίτερα ελκυστικά για εφαρμογές όπου η κατανόηση της διαδικασίας λήψης αποφάσεων του μοντέλου είναι σημαντική, όπως η αυθεντικότητα των τροφίμων και ο προσδιορισμός της γεωγραφικής προέλευσης (Hastie et al., 2009).

Ένα δέντρο απόφασης δημιουργεί μοντέλα ταξινόμησης ή παλινδρόμησης με τη μορφή μιας δομής δέντρου. Αναλύει ένα σύνολο δεδομένων σε όλο και μικρότερα υποσύνολα, ενώ αναπτύσσει ένα σχετικό δέντρο αποφάσεων σταδιακά. Το τελικό αποτέλεσμα είναι ένα δέντρο με κόμβους απόφασης και κόμβους φύλλων, όπου κάθε εσωτερικός κόμβος αντιστοιχεί σε ένα χαρακτηριστικό, κάθε κόμβος φύλλου αντιστοιχεί σε μια ετικέτα κλάσης και κάθε κλάδος αντιπροσωπεύει έναν κανόνα (Quinlan, 1986).

Στον παρεχόμενο κώδικα Python, χρησιμοποιείται ο ταξινομητής δένδρων αποφάσεων από τη βιβλιοθήκη scikit-learn. Η παράμετρος 'max_depth' του ταξινομητή δένδρων αποφάσεων, η οποία αναφέρεται στον μέγιστο αριθμό επιπέδων στο δέντρο, ποικίλλει σε διαφορετικές επαναλήψεις. Η παράμετρος 'max_depth' παίζει κρίσιμο ρόλο στην αποτροπή του μοντέλου από την υπερβολική προσαρμογή των δεδομένων εκπαίδευσης περιορίζοντας την πολυπλοκότητα του μοντέλου. Εάν αυτή η παράμετρος ρυθμιστεί πολύ ψηλά, το δέντρο αποφάσεων μπορεί να γίνει υπερβολικά πολύπλοκο, οδηγώντας σε υπερβολική προσαρμογή όπου απομνημονεύει τα δεδομένα εκπαίδευσης και έχει κακή απόδοση σε αόρατα δεδομένα (Breiman et al., 1984).

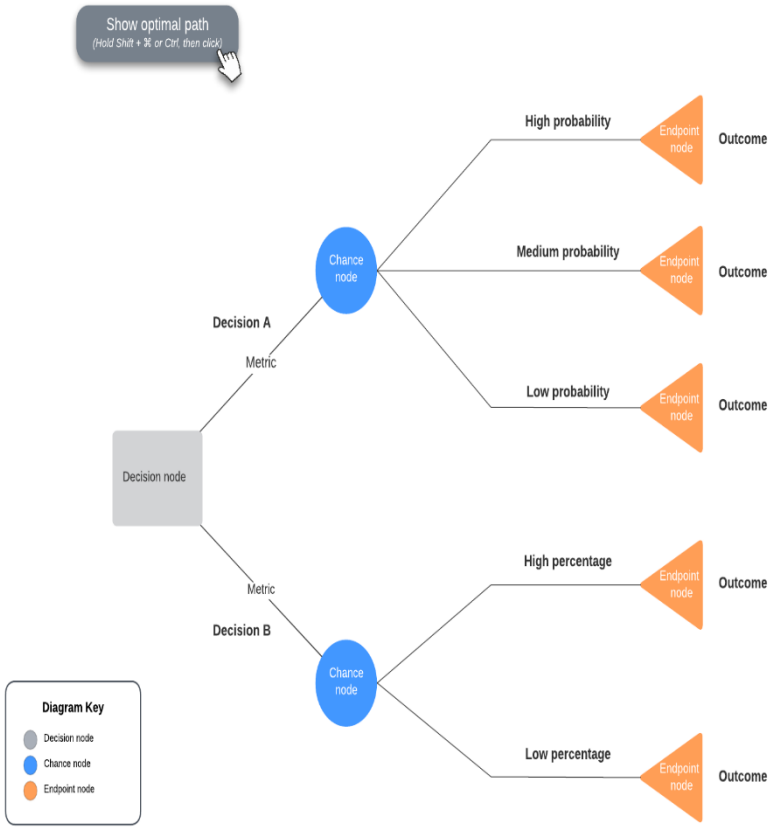
Ο ταξινομητής δέντρου αποφάσεων μπορεί να χειριστεί μη γραμμικές σχέσεις μεταξύ χαρακτηριστικών και μπορεί να είναι λιγότερο ευαίσθητος σε ακραίες τιμές στα δεδομένα σε σύγκριση με άλλους ταξινομητές, όπως το Support Vector Machines (Hastie et al., 2009). Τα δέντρα αποφάσεων είναι επίσης εξαιρετικά ερμηνεύσιμα, καθώς η δομή του δέντρου επιτρέπει μια σαφή απεικόνιση της διαδικασίας λήψης αποφάσεων. Αυτή η ερμηνεία είναι πολύτιμη στο πλαίσιο της επαλήθευσης της γνησιότητας των τροφίμων και της γεωγραφικής προέλευσης, όπου η κατανόηση της διαδικασίας λήψης αποφάσεων μπορεί να παρέχει πληροφορίες για τα πιο σημαντικά φασματικά χαρακτηριστικά για τη διάκριση διαφορετικών τύπων λαδιών (Quinlan, 1986).

Ωστόσο, τα δέντρα απόφασης έχουν επίσης τους περιορισμούς τους. Τα δέντρα απόφασης μπορούν εύκολα να προσαρμόσουν ή να υποχωρήσουν στο σύνολο δεδομένων εάν δεν είναι σωστά συντονισμένα και είναι επίσης γνωστό ότι έχουν υψηλή διακύμανση. Αυτό σημαίνει ότι αν αλλάξουμε ελαφρώς το σύνολο δεδομένων μας, το δέντρο αποφάσεων που προκύπτει μπορεί να φαίνεται πολύ διαφορετικό, οδηγώντας σε ασταθείς προβλέψεις (Hastie et al., 2009).

Σε αυτό το πλαίσιο, η προσέγγιση τυχαιοποίησης που χρησιμοποιείται στον κώδικα για την επιλογή της παραμέτρου 'max_depth' μπορεί να προσφέρει έναν αποτελεσματικό τρόπο για να βρεθεί μια ισορροπία μεταξύ υποπροσαρμογής και υπερπροσαρμογής. Επιπλέον, μέθοδοι συνόλου, όπως τα

τυχαία δάση, που κατασκευάζουν πολλαπλά δέντρα απόφασης και συγκεντρώνουν τα αποτελέσματά τους, μπορούν να χρησιμοποιηθούν για να ξεπεραστούν ορισμένοι από τους περιορισμούς των δέντρων μεμονωμένων αποφάσεων, όπως η υψηλή διακύμανσή τους (Breiman, 2001).

Συνοπτικά, ο ταξινομητής δένδρων αποφάσεων είναι μια εξαιρετικά ερμηνεύσιμη, μη παραμετρική μέθοδος κατάλληλη για την ταξινόμηση του ελαιόλαδου με βάση φασματικά δεδομένα LIBS. Αν και έχει ορισμένους περιορισμούς, αυτοί μπορούν να μετριαστούν μέσω προσεκτικής ρύθμισης παραμέτρων και χρήσης μεθόδων συνόλου.



Εικόνα 2: Διαδικασία ταξινόμησης Δένδρων Αποφάσεων

5.4.3 Ταξινομητής Random Forest

To Random Forest Classifier είναι ένας αλγόριθμος μηχανικής μάθησης συνόλου που είναι γνωστός για την στιβαρότητα, την ευελιξία και την εξαιρετική του απόδοση σε μια ποικιλία εφαρμογών, συμπεριλαμβανομένης της γνησιότητας των τροφίμων και της επαλήθευσης γεωγραφικής προέλευσης (Breiman, 2001). Ένα τυχαίο δάσος αποτελείται από ένα πλήθος δέντρων απόφασης και η παραγωγή του καθορίζεται από τις συνδυασμένες εξόδους αυτών των μεμονωμένων δέντρων.

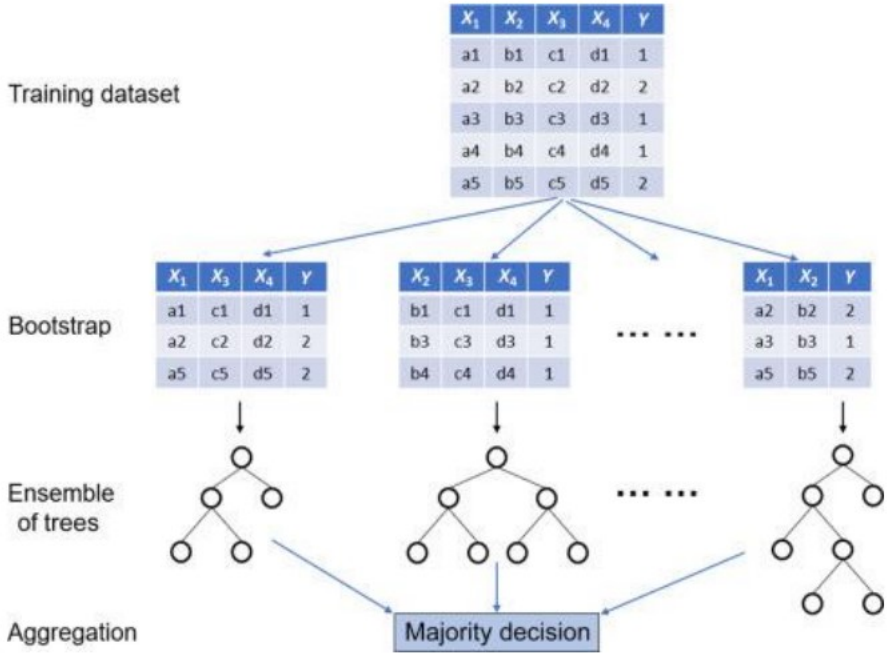
Στον κώδικα που παρέχεται, χρησιμοποιείται ο Random Forest Classifier από τη βιβλιοθήκη scikit-learn. Η κύρια παράμετρος που ποικίλλει σε διαφορετικές επαναλήψεις είναι «n_estimators», η οποία αναφέρεται στον αριθμό των δέντρων στο δάσος. Κάθε ένα από αυτά τα δέντρα είναι χτισμένο σε ένα δείγμα εκκίνησης από τα δεδομένα εκπαίδευσης και σε κάθε κόμβο, ένα τυχαίο υποσύνολο χαρακτηριστικών λαμβάνεται υπόψη για διαχωρισμό. Αυτή η διαδικασία εισάγει την τυχαιότητα στο μοντέλο, η οποία βοηθά στη μείωση της διακύμανσης και στη βελτίωση της γενίκευσης, μειώνοντας τον κίνδυνο υπερβολικής προσαρμογής (Cutler et al., 2007).

Τα Random Forests μπορούν να χειριστούν σύνολα δεδομένων υψηλών διαστάσεων και είναι λιγότερο επιρρεπή σε υπερπροσαρμογή σε σύγκριση με μεμονωμένα δέντρα αποφάσεων. Αυτό συμβαίνει επειδή κάθε δέντρο στο δάσος εκπαιδεύεται σε ένα διαφορετικό υποσύνολο δεδομένων, καθιστώντας το μοντέλο πιο ανθεκτικό σε θόρυβο και ακραίες τιμές (Díaz-Uriarte & Alvarez de Andrés, 2006). Αυτό είναι ιδιαίτερα σημαντικό στο πλαίσιο των φασματικών δεδομένων LIBS, τα οποία μπορεί να είναι υψηλών διαστάσεων και να περιέχουν ακραίες τιμές λόγω σφαλμάτων μέτρησης ή θορύβου.

Ένα άλλο πλεονέκτημα είναι ότι παρέχουν μέτρα σημασίας χαρακτηριστικών, τα οποία μπορούν να χρησιμοποιηθούν για να κατανοήσουμε ποια φασματικά χαρακτηριστικά είναι πιο σημαντικά για τη διάκριση διαφορετικών τύπων ελαίων (Strobl et al., 2008). Αυτή η ερμηνεία είναι ζωτικής σημασίας σε εφαρμογές επαλήθευσης γνησιότητας τροφίμων και γεωγραφικής προέλευσης, όπου η κατανόηση της διαδικασίας λήψης αποφάσεων μπορεί να οδηγήσει σε βελτιωμένες διαδικασίες ποιοτικού ελέγχου και ρυθμιστικά μέτρα.

Ωστόσο, παρά τα πλεονεκτήματά του, έχει τους περιορισμούς του. Ενώ το μοντέλο είναι λιγότερο επιρρεπές σε υπερπροσαρμογή από τα μεμονωμένα δέντρα απόφασης, μπορεί να προσαρμοστεί υπερβολικά εάν ο αριθμός των δέντρων είναι πολύ μεγάλος ή εάν τα δέντρα είναι πολύ περίπλοκα. Επιπλέον, τα Τυχαία Δάση μπορεί να είναι υπολογιστικά ακριβά για εκπαίδευση και ανάπτυξη, ιδιαίτερα σε μεγάλα σύνολα δεδομένων (Liaw et al., 2002).

Η τυχαιοποιημένη προσέγγιση που χρησιμοποιείται στον κώδικα για τη ρύθμιση της παραμέτρου 'n_estimators' επιτρέπει την αποτελεσματική εξερεύνηση του χώρου παραμέτρων του μοντέλου, η οποία μπορεί να βοηθήσει στον μετριασμό ορισμένων από αυτούς τους περιορισμούς. Παρά τις δυνατότητές του για υπερπροσαρμογή και υπολογιστικό κόστος, ο Random Forest Classifier παραμένει ένα ισχυρό εργαλείο για εργασίες ταξινόμησης που περιλαμβάνουν φασματικά δεδομένα υψηλών διαστάσεων, όπως τα φασματικά δεδομένα LIBS που χρησιμοποιούνται σε αυτή τη διπλωματική εργασία.



Εικόνα 3: Διαδικασία ταξινόμησης Τυχαίου Δάσους

5.4.4 Ταξινομητής Support Vector Machine (SVM)

Το Support Vector Machine (SVM) είναι ένας ισχυρός αλγόριθμος μηχανικής μάθησης που έχει εφαρμοστεί ευρέως σε πολλούς τομείς, όπως η αναγνώριση εικόνων, η κατηγοριοποίηση κειμένου και η βιοπληροφορική, μεταξύ άλλων (Cortes & Vapnik, 1995). Στο πλαίσιο της φασματοσκοπικής ανάλυσης δεδομένων, τα SVM μπορούν να χειριστούν αποτελεσματικά δεδομένα υψηλών διαστάσεων και να εκτελέσουν εργασίες ταξινόμησης δυαδικών καθώς και πολλαπλών κατηγοριών, καθιστώντας τα ελκυστική επιλογή για την ανάλυση δεδομένων LIBS για την αυθεντικότητα των τροφίμων και τον προσδιορισμό της γεωγραφικής προέλευσης.

Στον αλγόριθμο SVM, ο πρωταρχικός στόχος είναι να βρεθεί ένα υπερεπίπεδο που διαχωρίζει καλύτερα τα σημεία δεδομένων διαφορετικών κλάσεων στον χώρο χαρακτηριστικών. Το υπερεπίπεδο προσδιορίζεται με βάση τα σημεία δεδομένων (γνωστά ως διανύσματα υποστήριξης) που είναι πιο κοντά στο όριο απόφασης. Ο ταξινομητής SVM προσπαθεί να μεγιστοποιήσει το περιθώριο γύρω από το υπερεπίπεδο, οδηγώντας σε ένα ισχυρό μοντέλο που γενικεύει καλά σε αόρατα δεδομένα (Schölkopf & Smola, 2002).

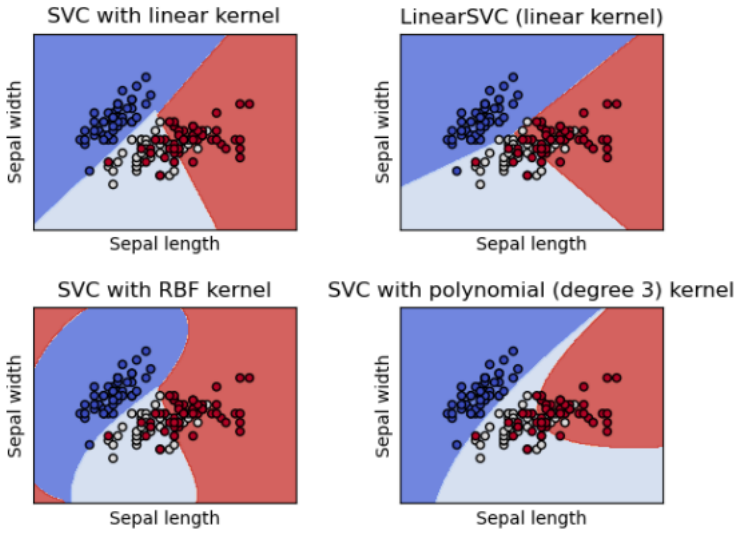
Στον παρεχόμενο κώδικα, χρησιμοποιείται ο ταξινομητής SVM από τη βιβλιοθήκη scikit-learn. Οι κρίσιμες παράμετροι που ποικίλλουν στις διάφορες επαναλήψεις περιλαμβάνουν «πυρήνα», «βαθμός» και «συντελεστής 0». Η παράμετρος «πυρήνας» καθορίζει τη συνάρτηση πυρήνα που χρησιμοποιείται για τη μετατροπή των δεδομένων σε χώρο υψηλότερης διάστασης για να διευκολυνθεί ο διαχωρισμός

των σημείων δεδομένων. Η παράμετρος 'degree' χρησιμοποιείται για πολυωνυμικό πυρήνα και υποδεικνύει το βαθμό του πολυωνύμου που χρησιμοποιείται για τον μετασχηματισμό. Η παράμετρος 'coef0' χρησιμοποιείται σε πολλές συναρτήσεις πυρήνα και ελέγχει πόσο επηρεάζεται το μοντέλο από πολυώνυμα υψηλού βαθμού στην περίπτωση του πολυωνύμου και του σιγμοειδούς πυρήνα.

Τα SVM, ιδιαίτερα με μη γραμμικές συναρτήσεις πυρήνα, μπορούν να συλλάβουν πολύπλοκες σχέσεις στα δεδομένα και να αποδίδουν καλά ακόμα και όταν τα δεδομένα δεν είναι γραμμικά διαχωρισμένα. Αυτό είναι ιδιαίτερα χρήσιμο στο πλαίσιο των φασματικών δεδομένων LIBS, τα οποία μπορούν να εμφανίσουν πολύπλοκα, μη γραμμικά μοτίβα (Mounier & Dubernet, 2004).

Ωστόσο, τα SVM έχουν ορισμένους περιορισμούς. Η επιλογή του πυρήνα και των σχετικών παραμέτρων μπορεί να επηρεάσει σημαντικά την απόδοση του μοντέλου και η επιλογή του κατάλληλου πυρήνα και παραμέτρων μπορεί να είναι δύσκολη. Τα SVM μπορούν επίσης να είναι υπολογιστικά εντατικά για μεγάλα σύνολα δεδομένων και η διαδικασία λήψης αποφάσεων του μοντέλου μπορεί να είναι δύσκολο να ερμηνευτεί, ιδιαίτερα με μη γραμμικούς πυρήνες (Hsu et al., 2003).

Η τυχαιοποιημένη προσέγγιση που χρησιμοποιείται στον κώδικα για τη ρύθμιση των παραμέτρων SVM επιτρέπει την αποτελεσματική εξερεύνηση του χώρου παραμέτρων του μοντέλου και μπορεί να βοηθήσει στην αντιμετώπιση αυτών των περιορισμών. Παρά τις πιθανές προκλήσεις, ο ταξινομητής SVM είναι ένα ισχυρό και ευέλικτο εργαλείο για την ανάλυση φασματικών δεδομένων LIBS για την αυθεντικότητα των τροφίμων και τον προσδιορισμό της γεωγραφικής προέλευσης.



Εικόνα 4: Αλγόριθμος SVC με την χρήση διαφορετικών μεθόδων πυρήνων

5.5 Αξιολόγηση Μοντέλου

5.5.1 Μέθοδος διασταυρούμενης επικύρωσης

Στη μηχανική μάθηση, η απόδοση ενός μοντέλου αξιολογείται συνήθως από την δυνατότητα του να προβλέπει με ακρίβεια νέα, αόρατα δεδομένα, τα οποία δεν εμπλέκονται στη διαδικασία εκπαίδευσης του μοντέλου. Μία από τις πιο συχνά χρησιμοποιούμενες μεθόδους για την εκτίμηση της απόδοσης ενός μοντέλου μηχανικής μάθησης σε αόρατα δεδομένα είναι η διασταυρούμενη επικύρωση (Kohavi, 1995).

Η διασταυρούμενη επικύρωση (cross validation), όπως υποδηλώνει το όνομά της, περιλαμβάνει την επικύρωση της απόδοσης ενός μοντέλου σε διαφορετικά υποσύνολα δεδομένων, δημιουργώντας ουσιαστικά πολλαπλούς διαχωρισμούς δοκιμής αμαξοστοιχίας. Είναι μια διαδικασία επαναδειγματοληψίας που χρησιμοποιείται για την αξιολόγηση μοντέλων μηχανικής μάθησης σε ένα περιορισμένο δείγμα δεδομένων. Η διαδικασία έχει μια μοναδική παράμετρο που ονομάζεται k που αναφέρεται στον αριθμό των ομάδων στις οποίες ένα δεδομένο δείγμα δεδομένων πρόκειται να χωριστεί (Kohavi, 1995).

Στο παρεχόμενο κώδικα Python, χρησιμοποιείται μια προσέγγιση διασταυρούμενης επικύρωσης για την αξιολόγηση της απόδοσης των μοντέλων μηχανικής εκμάθησης. Συγκεκριμένα, χρησιμοποιείται μια παραλλαγή διασταυρούμενης επικύρωσης γνωστή ως διασταυρούμενη επικύρωση k-fold. Σε αυτή τη μέθοδο, ολόκληρο το σύνολο δεδομένων χωρίζεται σε πτυχές ή υποσύνολα ίσου μεγέθους «k». Το μοντέλο μηχανικής εκμάθησης στη συνέχεια εκπαιδεύεται σε πτυχές «k-1» και το υπόλοιπο δίπλωμα χρησιμοποιείται ως το σετ δοκιμής για την αξιολόγηση της απόδοσης του μοντέλου. Αυτή η διαδικασία επαναλαμβάνεται 'k' φορές, με κάθε μία από τις πτυχές 'k' να χρησιμοποιείται ακριβώς μία φορά ως το σύνολο ελέγχου. Η τελική απόδοση του μοντέλου αναφέρεται συνήθως ως η μέση απόδοση στις επαναλήψεις «k».

Το κύριο πλεονέκτημα της διασταυρούμενης επικύρωσης k-fold είναι ότι επιτρέπει μια πιο ισχυρή εκτίμηση της απόδοσης του μοντέλου. Δεδομένου ότι κάθε σημείο δεδομένων πρέπει να βρίσκεται στο σύνολο δοκιμής ακριβώς μία φορά, η μέτρηση απόδοσης είναι λιγότερο πιθανό να επηρεαστεί από τυχόν ιδιαιτερότητες σε ένα συγκεκριμένο διαχωρισμό δοκιμής αμαξοστοιχίας (Hastie et al., 2009).

Η χρήση τετραπλής διασταυρούμενης επικύρωσης στο σενάριο υποδεικνύει ότι τα δεδομένα χωρίζονται σε τέσσερα ίσα μέρη και ο αλγόριθμος εκτελείται τέσσερις φορές, κάθε φορά με ένα διαφορετικό τέταρτο των δεδομένων που διατηρείται ως το σύνολο δοκιμής.

Παρά τα πολυάριθμα πλεονεκτήματά της, η διασταυρούμενη επικύρωση έχει τους περιορισμούς της. Πρώτον, υποθέτει ότι τα σημεία δεδομένων είναι ανεξάρτητα και πανομοιότυπα κατανεμημένα, κάτι που μπορεί να μην ισχύει πάντα, ειδικά στα δεδομένα χρονοσειρών (Varma & Simon, 2006). Δεύτερον, μπορεί να είναι υπολογιστικά εντατικό, ιδιαίτερα για μεγαλύτερα σύνολα δεδομένων και πιο πολύπλοκα μοντέλα. Τέλος, η επιλογή του «k» στη διασταυρούμενη επικύρωση k-fold μπορεί να επηρεάσει σημαντικά τα αποτελέσματα. Ένα μικρότερο «k» μπορεί να οδηγήσει σε υψηλή προκατάληψη, ενώ ένα μεγαλύτερο «k» μπορεί να οδηγήσει σε υψηλότερη διακύμανση (James et al., 2013).

Στο παρεχόμενο σενάριο Python, παρόλο που η 4-πλή διασταυρούμενη επικύρωση δεν απέδωσε ιδιαίτερα καλά αποτελέσματα, η συμπερίληψή της στην ανάλυση εξακολουθεί να είναι χρήσιμη καθώς παρέχει μια άλλη προοπτική για την απόδοση και την ευρωστία του μοντέλου.

Συμπερασματικά, η διασταυρούμενη επικύρωση είναι μια ουσιαστική τεχνική στη διαδικασία επιλογής και αξιολόγησης μοντέλων μηχανικής μάθησης. Παρέχει μια αξιόπιστη εκτίμηση της απόδοσης του μοντέλου και βοηθά στην επιλογή του βέλτιστου μοντέλου και των παραμέτρων του, κάτι που είναι κρίσιμο στο έργο της ανάλυσης δεδομένων LIBS για την αυθεντικότητα των τροφίμων και τον προσδιορισμό της γεωγραφικής προέλευσης.

5.5.2 Μέθοδος Monte Carlo στα νευρωνικά δίκτυα

Η μέθοδος Monte Carlo (MC), που πήρε το όνομά της από το καζίνο Monte Carlo στο Μονακό όπου οι εφευρέτες της μεθόδου πέρασαν το χρόνο τους, είναι μια στατιστική τεχνική που επιτρέπει αριθμητικές λύσεις σε πολύπλοκα προβλήματα. Η θεμελιώδης ιδέα πίσω από τη μέθοδο Monte Carlo είναι η χρήση της τυχαιότητας για την επίλυση προβλημάτων που μπορεί να είναι ντετερμινιστικά κατ' αρχήν. Η δυνατότητα εφαρμογής του, καλύπτει ένα ευρύ φάσμα πεδίων, από τις φυσικές επιστήμες έως τα οικονομικά και την επιστήμη των υπολογιστών (Metropolis & Ulam, 1949).

Στον τομέα της μηχανικής μάθησης και πιο συγκεκριμένα, στο πλαίσιο των νευρωνικών δικτύων, η μέθοδος Monte Carlo έχει σημαντικό ρόλο. Χρησιμοποιείται με διάφορους τρόπους, όπως συντονισμός υπερπαραμέτρων, επιλογή μοντέλου και για την εκτίμηση της αβεβαιότητας που σχετίζεται με τις προβλέψεις του μοντέλου (Neal, 2012).

Τα νευρωνικά δίκτυα, ειδικά τα πολυστρωματικά perceptrons (MLPs), έχουν μια εγγενή στοχαστική φύση λόγω της τυχαιοποιημένης αρχικοποίησης του βάρους τους και, σε ορισμένες περιπτώσεις, της τυχαίας εγκατάλειψης νευρώνων κατά τη διάρκεια της προπόνησης. Αυτή η τυχαιότητα σημαίνει ότι η εκπαίδευση της ίδιας αρχιτεκτονικής νευρωνικών δικτύων με τα ίδια δεδομένα θα μπορούσε να αποφέρει διαφορετικά αποτελέσματα κάθε φορά. Η μέθοδος Monte Carlo μπορεί να χρησιμοποιηθεί

για να διερευνήσει αυτά τα διαφορετικά αποτελέσματα και να αποκτήσει ένα πιο ισχυρό και γενικευμένο μοντέλο.

Σε αυτή τη μελέτη, όπως αναφέρεται στον παρεχόμενο κώδικα, η μέθοδος Monte Carlo εφαρμόζεται σε έναν ταξινομητή MLP. Αν και δεν δίνονται οι συγκεκριμένες λεπτομέρειες υλοποίησης, ένας πιθανός τρόπος εφαρμογής είναι μέσω του Monte Carlo Cross-Validation (MCCV). Στο MCCV, σε αντίθεση με τη διασταυρούμενη επικύρωση k-fold όπου το σύνολο δεδομένων διαιρείται συστηματικά σε πτυχές «k», η διαίρεση του συνόλου δεδομένων σε σύνολα εκπαίδευσης και επικύρωσης γίνεται τυχαία για έναν καθορισμένο αριθμό επαναλήψεων. Αυτή η διαδικασία επιτρέπει την καλύτερη εξερεύνηση της απόδοσης του μοντέλου σε διαφορετικά τυχαία τμήματα των δεδομένων, προσφέροντας έτσι μια πιο ολοκληρωμένη επισκόπηση της ευρωστίας και της γενίκευσης του μοντέλου (Xu & Liang, 2001).

Η μέθοδος Monte Carlo μπορεί επίσης να χρησιμοποιηθεί στον συντονισμό υπερπαραμέτρων σε νευρωνικά δίκτυα. Οι υπερπαράμετροι είναι παράμετροι των οποίων οι τιμές ορίζονται πριν ξεκινήσει η διαδικασία εκμάθησης. Αυτά περιλαμβάνουν τον αριθμό των κρυφών επιπέδων στο δίκτυο, τον αριθμό των νευρώνων σε κάθε κρυφό επίπεδο, τον ρυθμό εκμάθησης, κ.λπ. Μια πιθανή προσέγγιση είναι η χρήση μιας προσομοίωσης Monte Carlo για τυχαία δειγματοληψία τιμών για αυτές τις υπερπαραμέτρους από προκαθορισμένες περιοχές. το μοντέλο με αυτές τις τιμές και αξιολογήστε την απόδοσή του. Το καλύτερο σύνολο υπερπαραμέτρων θα ήταν τότε αυτό που έχει ως αποτέλεσμα την καλύτερη απόδοση μοντέλου (Bergstra & Bengio, 2012).

Παρά την στιβαρότητά της, η μέθοδος Monte Carlo έχει περιορισμούς. Μπορεί να είναι υπολογιστικά ακριβό, ειδικά για μεγάλα σύνολα δεδομένων και πολύπλοκα μοντέλα. Επιπλέον, η μέθοδος βασίζεται σε μεγάλο βαθμό στην τυχαιότητα, η οποία θα μπορούσε να οδηγήσει σε μεταβλητότητα στα αποτελέσματα (Robert & Casella, 2013).

Συνοψίζοντας, η μέθοδος Monte Carlo, όταν συνδυάζεται με μοντέλα νευρωνικών δικτύων όπως ο ταξινομητής MLP, παρέχει ένα ισχυρό εργαλείο για την αξιολόγηση και την επιλογή μοντέλων. Βοηθά στην απόκτηση μιας πιο ολοκληρωμένης κατανόησης της απόδοσης και της ευρωστίας του μοντέλου, ενισχύοντας έτσι τη συνολική αξιοπιστία των αποτελεσμάτων που λαμβάνονται.

5.6 Αποτελέσματα και συζήτηση

5.6.1 Επεξεργασία δεδομένων

Τα δεδομένα διαχωρίζονται: (1) στα δεδομένα εισόδου και (2) στις μεταβλητές εξόδου που είναι οι 3 γεωγραφικές προελεύσεις του ελαιόλαδου (Κρήτη, Πελοπόννησος και Λέσβος). Στην συνέχεια τα δύο

σύνολα διαιρέθηκαν στα δεδομένα εκπαίδευσης (training dataset) και τα δεδομένα εξέτασης (testing dataset) με ποσοστό 90% και 10% αντίστοιχα. Σύμφωνα με την περιγραφή της λειτουργίας των μοντέλων μηχανικής εκμάθησης στο κεφάλαιο 3, τα δεδομένα εκπαίδευσης αξιοποιούνται για την εκπαίδευση του μοντέλου στην αναγνώριση του επιπέδου ασφαλείας δημιουργώντας μοτίβα αναγνώρισης βάσει συγκεκριμένων χαρακτηριστικών. Για την διαδικασία της αξιολόγησης το μοντέλο επεξεργάζεται τα εισαγόμενα δεδομένα εξέτασης και τα ταξινομεί σε μία από τις τρείς γεωγραφικές περιοχές ώστε αυτά να συγκριθούν με τις πραγματικές.

Επίσης, στο πλαίσιο επεξεργασίας των δεδομένων εφαρμόστηκε η μέθοδος “maxloc” για την εξαγωγή σημαντικών πληροφοριών από τα δεδομένα που είναι κρίσιμες για την εκπαίδευση και την ελαχιστοποίηση του θορύβου των δεδομένων.

Για την εκπαίδευση των μοντέλων μηχανικής μάθησης, είναι απαραίτητο να μετατραπούν οι κατηγορίες/ετικέτες (labels) σε αριθμητικές τιμές. Στον συγκεκριμένο κώδικα, αξιοποιείται η συνάρτηση setMap για να δημιουργηθεί ένας χάρτης αντιστοίχισης (C) μεταξύ των μοναδικών ετικετών στη μεταβλητή στόχου. Αυτός ο χάρτης χρησιμοποιείται στη συνέχεια για να μετατραπούν οι ετικέτες σε αντίστοιχους αριθμούς που μπορούν να χρησιμοποιηθούν για την κατηγοριοποίηση.

5.6.2 Σύγκριση της απόδοσης των μοντέλων

Σε αυτήν την διπλωματική εργασία, χρησιμοποιήθηκαν πολλαπλά μοντέλα μηχανικής μάθησης, συμπεριλαμβανομένου του ταξινομητή Perceptron πολλαπλών επιπέδων (MLP), του ταξινομητή Decision Tree, του ταξινομητή Random Forest και του ταξινομητή Support Vector Machine (SVM), για την πρόβλεψη της αυθεντικότητας του λαδιού και της γεωγραφικής προέλευσης από τα δεδομένα LIBS. Η εφαρμογή διαφορετικών αλγορίθμων επέτρεψε μια ολοκληρωμένη εξερεύνηση των δυνατοτήτων των μοντέλων και της σχετικής απόδοσής τους.

Πίνακας 1: Ονοματολογία και συμβολισμός μοντέλων ταξινόμησης

Όνομα μοντέλου (ελληνικά)	Όνομα μοντέλου (αγγλικά)	Συμβολισμός μοντέλου
Μηχανές Διανυσμάτων Υποστήριξης	Support Vector Machines	SVM
Ταξινομητής Τυχαίων Δασών	Random Forests Classifier	RF
Ταξινομητής Δένδρων Αποφάσεων	Decision Trees	DT
Ταξινομητής Πολυεπίπεδου Perceptron	Multilayer Perceptron Classifier	MLP

Στο πλαίσιο της ανάπτυξης των μοντέλων, πραγματοποιείται εξέταση των παραμέτρων για τους τέσσερεις ταξινομητές (MLP, DT, SVM, RF). Συγκεκριμένα, πραγματοποιείται επαναληπτική εκπαίδευση με τυχαίους συνδυασμούς υπερπαραμέτρων με επανάληψη του βρόχου από 1 έως 20 με βήμα 1.

Για τον αλγόριθμο SVM οι συνδυασμοί αφορούν τις υπερπαραμέτρους:

- **kernel**: Το πυρήνας που χρησιμοποιείται για το SVM. Επιλέγεται τυχαία από τη λίστα kern, που περιλαμβάνει τους πυρήνες "linear," "poly," "rbf," και "sigmoid".
- **degree**: Η τάξη του πολυωνύμου πυρήνα (ισχύει μόνο για πυρήνα "poly"). Επιλέγεται τυχαία από το εύρος [2, 4].
- **coef0**: Η παράμετρος coef0 του πυρήνα (ίσχυεί για πυρήνες "poly" και "sigmoid"). Επιλέγεται τυχαία από το εύρος [0, 1].

Για τον αλγόριθμο RF οι συνδυασμοί αφορούν τις υπερπαραμέτρους:

- **max_depth**: Το μέγιστο βάθος των δέντρων που χρησιμοποιούνται στο Random Forest. Επιλέγεται τυχαία από το εύρος [2, 20].
- **n_estimators**: Ο αριθμός των δέντρων που θα συνθέσουν το Random Forest. Επιλέγεται τυχαία από το εύρος [10, 100].

Για τον αλγόριθμο DT οι συνδυασμοί αφορούν τις υπερπαραμέτρους:

- **max_depth**: Το μέγιστο βάθος του δέντρου αποφάσεων. Επιλέγεται τυχαία από το εύρος [2, 20].

Για τον αλγόριθμο MLP οι συνδυασμοί αφορούν τις υπερπαραμέτρους:

- **solver (αλγόριθμο επίλυσης**): Επιλέγεται τυχαία μεταξύ των τριών επιλογών 'lbfgs', 'sgd', και 'adam'.
- **alpha (λάμδα) για τον ρυθμιστή L2**: Επιλέγεται τυχαία μια τιμή από το διάστημα (0, 0.1).
- **hidden_layer_sizes (αριθμό των νευρώνων και των κρυφών στρωμάτων στο νευρωνικό δίκτυο**): Επιλέγεται τυχαία ένα ζευγάρι ακέραιων από το διάστημα (2, 100) για τον αριθμό των νευρώνων σε κρυφά επίπεδα

Αυτές οι παράμετροι δοκιμάζονται 20 φορές, και οι αποτελεσματικότητες και οι συνδυασμοί των παραμέτρων αποθηκεύονται. Στο τέλος εμφανίζεται ο καλύτερος συνδυασμός παραμέτρων με την υψηλότερη απόδοση.

5.6.2.1 Αλγόριθμος Μηχανών Διανυσμάτων Υποστήριξης (SVM)

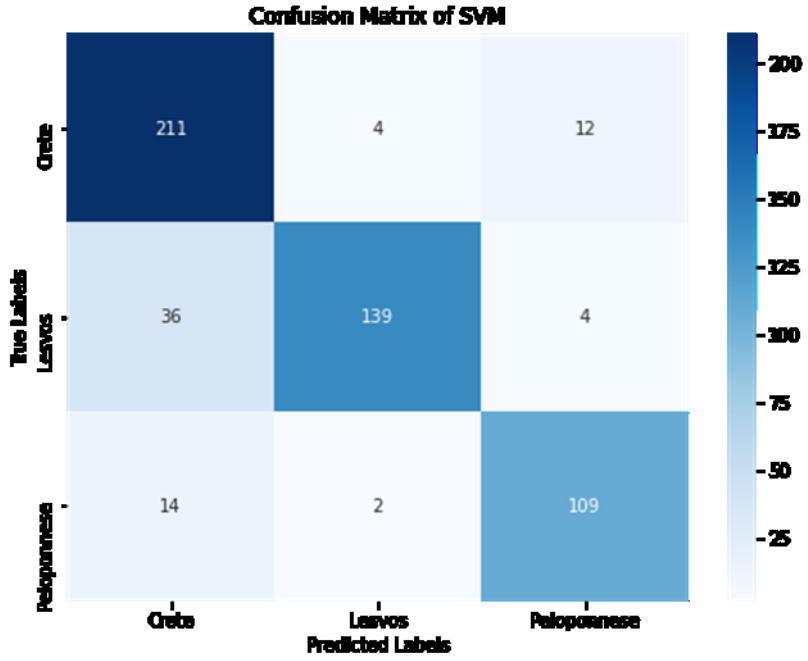
Ο αλγόριθμος SVM κατάφερε να ταξινομήσει ορθά μεγάλο ποσοστό των δειγμάτων για κάθε γεωγραφική περιοχή. Με βάση το Πίνακα 1, είναι εμφανές ότι το μοντέλο σημείωσε υψηλά ποσοστά ακρίβειας (precision) που σημαίνει ότι για τις τρείς γεωγραφικές περιοχές οι προβλέψεις σε υψηλά ποσοστά ήταν ορθές (0.88). Σχετικά με την ανάκληση (recall), και οι τρεις γεωγραφικές προελεύσεις σημείωσαν υψηλό ποσοστό ακρίβειας (0.86) το οποίο καταδεικνύει την υψηλή ικανότητα του μοντέλου να ανακαλύψει τις πραγματικές θετικές τιμές (δηλαδή τις πραγματικές γεωγραφικές προελεύσεις).

Με βάση τον Πίνακα 2, οι βέλτιστες παράμετροι που προέκυψαν με βάση την διαδικασία τυχαίας αναζήτησης που αναλύθηκε προηγουμένως ήταν:

kernel: ‘poly’

degree: 4

coef0: 0.8176590226704543



Εικόνα 5: Μήτρα σύγχυσης αλγόριθμου SVM

Πίνακας 2: Σύνοψη μοντέλου SVM

```
Best Parameters: ['poly', 4, 0.8176590226704543]
Confusion Matrix:
[[211   4  12]
 [ 36 139   4]
 [ 14   2 109]]
Classification Report:
              precision    recall  f1-score   support

       Crete       0.81      0.93      0.86       227
      Lesvos       0.96      0.78      0.86       179
 Peloponnese       0.87      0.87      0.87       125

    accuracy                           0.86       531
   macro avg       0.88      0.86      0.86       531
weighted avg       0.87      0.86      0.86       531
```

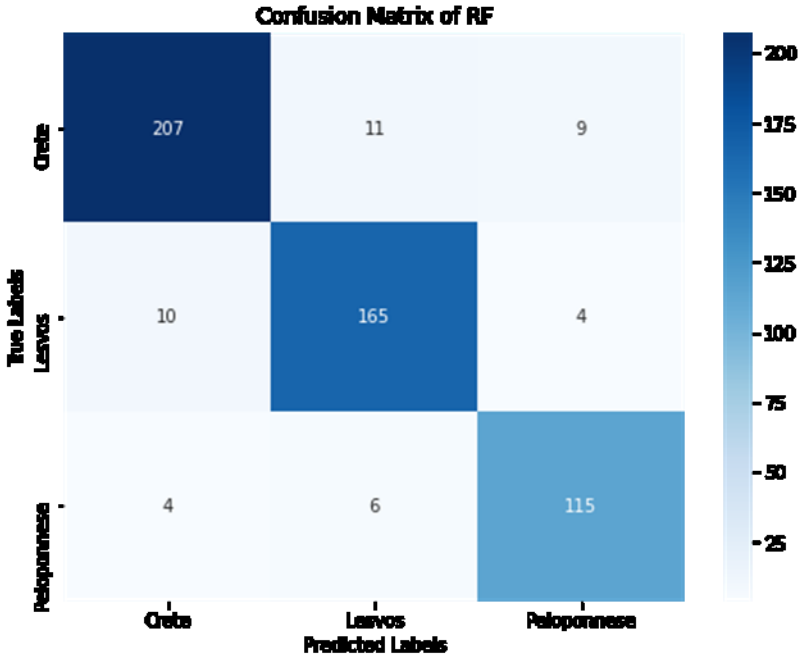
5.6.2.2 Αλγόριθμος Τυχαίων Δασών (RF)

Ο αλγόριθμος RF κατάφερε να ταξινομήσει ορθά πολύ μεγάλο ποσοστό των δειγμάτων για κάθε γεωγραφική περιοχή. Με βάση το Πίνακα 3 είναι εμφανές ότι το μοντέλο σημείωσε τα πιο υψηλά ποσοστά ακρίβειας (precision), συγκριτικά με τους άλλους αλγορίθμους που εξετάστηκαν, το οποίο αποδεικνύει ότι για τις τρείς γεωγραφικές περιοχές οι προβλέψεις σε υψηλά ποσοστά ήταν ορθές (0.91) . Σχετικά με την ανάκληση (recall), και οι τρεις γεωγραφικές προελεύσεις σημείωσαν πολύ υψηλό ποσοστό ακρίβειας (0.91) το οποίο καταδεικνύει την υψηλή ικανότητα του μοντέλου να ανακαλύψει τις πραγματικές θετικές τιμές (δηλαδή τις πραγματικές γεωγραφικές προελεύσεις).

Με βάση τον Πίνακα 3, οι βέλτιστες παράμετροι που προέκυψαν με βάση την διαδικασία τυχαίας αναζήτησης που αναλύθηκε προηγουμένως ήταν:

max_depth: 17

n_estimators: 31



Εικόνα 6: Μήτρα σύγχυσης αλγόριθμου RF

Πίνακας 3: Σύνοψη μοντέλου RF

```
Best Parameters: [17, 31]
Confusion Matrix:
[[206  12   9]
 [ 12 163   4]
 [  6   4 115]]
Classification Report:
              precision    recall  f1-score   support

       Crete       0.92      0.91      0.91       227
      Lesvos       0.91      0.91      0.91       179
 Peloponnese       0.90      0.92      0.91       125

    accuracy                           0.91       531
   macro avg       0.91      0.91      0.91       531
weighted avg       0.91      0.91      0.91       531
```

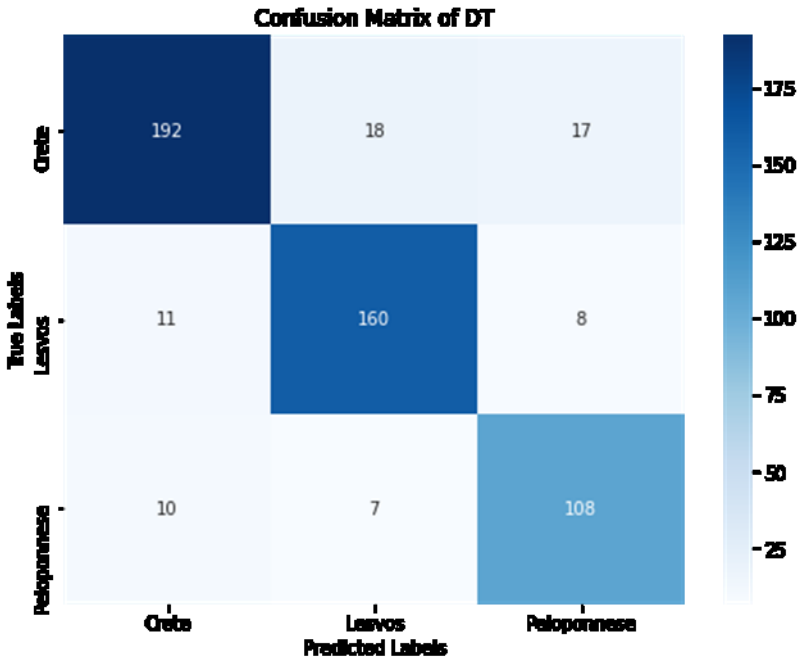
5.6.2.3 Αλγόριθμος Δένδρων Αποφάσεων (DT)

Ο αλγόριθμος DT κατάφερε να ταξινομήσει ορθά μεγάλο ποσοστό των δειγμάτων για κάθε γεωγραφική περιοχή. Με βάση το Πίνακα 4 είναι εμφανές ότι το μοντέλο σημείωσε υψηλά ποσοστά ακρίβειας (precision) που σημαίνει ότι για τις τρείς γεωγραφικές περιοχές οι προβλέψεις σε υψηλά ποσοστά ήταν ορθές (0.86) . Σχετικά με την ανάκληση (recall), και οι τρεις γεωγραφικές προελεύσεις σημείωσαν υψηλό ποσοστό ακρίβειας (0.86) το οποίο καταδεικνύει την υψηλή ικανότητα του

μοντέλου να ανακαλύψει τις πραγματικές θετικές τιμές (δηλαδή τις πραγματικές γεωγραφικές προελεύσεις).

Με βάση τον Πίνακα 4, οι βέλτιστες παράμετροι που προέκυψαν με βάση την διαδικασία τυχαίας αναζήτησης που αναλύθηκε προηγουμένως ήταν:

- max_depth: 13



Εικόνα 7: Μήτρα σύγχυσης αλγόριθμου DT

Πίνακας 4: Σύνοψη μοντέλου DT

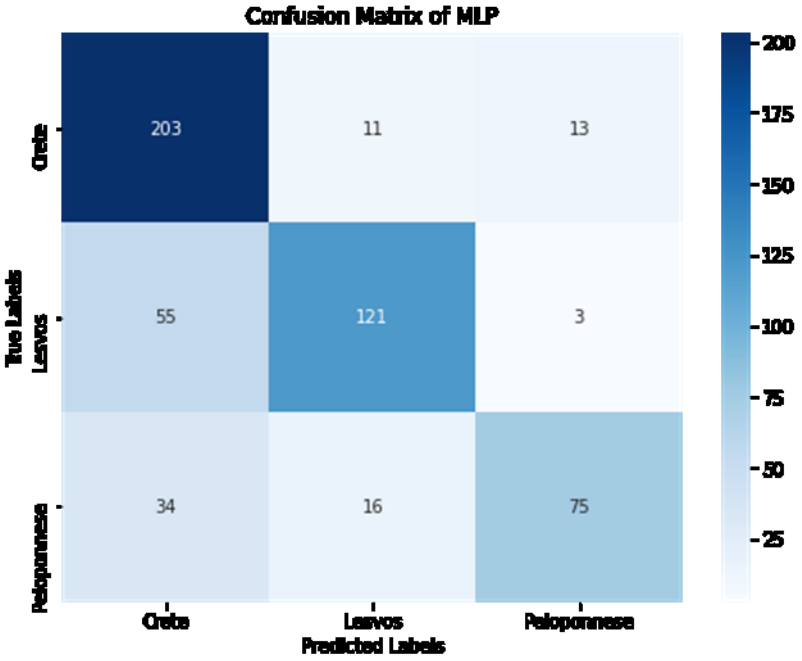
```
Best Parameters: [13, 22]
Confusion Matrix:
[[191  18  18]
 [ 11 162   6]
 [ 12   7 106]]
Classification Report:
              precision    recall  f1-score   support

       Crete       0.89      0.84      0.87       227
      Lesvos       0.87      0.91      0.89       179
 Peloponnese       0.82      0.85      0.83       125

    accuracy                           0.86       531
   macro avg       0.86      0.86      0.86       531
weighted avg       0.87      0.86      0.86       531
```

5.6.2.4 Αλγόριθμος Πολυεπίπεδου Perceptron (MLP)

Ο αλγόριθμος MLP κατάφερε να ταξινομήσει ορθά μεγάλο ποσοστό των δειγμάτων για κάθε γεωγραφική περιοχή. Με βάση το Πίνακα 5 είναι εμφανές ότι το μοντέλο σημείωσε υψηλά ποσοστά ακρίβειας (precision) που σημαίνει ότι για τις τρείς γεωγραφικές περιοχές οι προβλέψεις σε υψηλά ποσοστά ήταν ορθές. Σχετικά με την ανάκληση (recall), η Κρήτη σημείωσε υψηλό ποσοστό ενώ για την Πελοπόννησο και την Λέσβο το αντίστοιχο ποσοστό ήταν μειωμένο το οποίο καταδεικνύει την μειωμένη ικανότητα του μοντέλου να ανακαλύψει τις πραγματικές θετικές τιμές (δηλαδή τις πραγματικές γεωγραφικές προελεύσεις).

Με βάση τον Πίνακα 5, οι βέλτιστες παράμετροι που προέκυψαν με βάση την διαδικασία τυχαίας αναζήτησης που αναλύθηκε προηγουμένως ήταν:

- solver (αλγόριθμο επίλυσης): 'adam'.
- alpha (λάμδα) για τον ρυθμιστή L2: 0.038420144324223
- hidden_layer_sizes (αριθμό των νευρώνων και των κρυφών στρωμάτων στο νευρωνικό δίκτυο): (10, 9)



Εικόνα 8: Μήτρα σύγχυσης αλγόριθμου MLP

Πίνακας 5: Σύνοψη μοντέλου MLP

```
Best Parameters: ['adam', 0.03842019443242223, (10, 9)]
Confusion Matrix:
[[203  11  13]
 [ 55 121   3]
 [ 34  16  75]]
Classification Report:
              precision    recall  f1-score   support

       Crete       0.70      0.89      0.78       227
      Lesvos       0.82      0.68      0.74       179
 Peloponnese       0.82      0.60      0.69       125

    accuracy                           0.75       531
   macro avg       0.78      0.72      0.74       531
weighted avg       0.77      0.75      0.75       531
```

Ένας σημαντικός παράγοντας που πρέπει να ληφθεί υπόψη στη σύγκριση μοντέλων είναι η ακρίβεια των προβλέψεων. Η ακρίβεια είναι μια απλή μέτρηση που παρέχει την αναλογία των σωστών προβλέψεων προς τον συνολικό αριθμό των δειγμάτων εισόδου. Είναι το πιο διαισθητικό μέτρο απόδοσης και χρησιμοποιείται πιο συχνά σε προβλήματα ταξινόμησης. Ωστόσο, μπορεί να είναι παραπλανητικό εάν οι τάξεις είναι μη ισορροπημένες, καθώς δεν λαμβάνει υπόψη την κατανομή των τάξεων (Witten & Frank, 2005).

Μια άλλη βασική πτυχή που πρέπει να ληφθεί υπόψη κατά τη σύγκριση μοντέλων είναι η υπολογιστική τους απόδοση. Αυτός ο παράγοντας είναι κρίσιμος σε πρακτικές εφαρμογές, ειδικά όταν έχουμε να κάνουμε με μεγάλα σύνολα δεδομένων ή όταν ένα μοντέλο χρειάζεται να κάνει προβλέψεις σε πραγματικό χρόνο. Η αποτελεσματικότητα ενός μοντέλου καθορίζεται από την ταχύτητα εκπαίδευσης και πρόβλεψής του. Τα SVM, για παράδειγμα, μπορεί να είναι υπολογιστικά εντατικά, ειδικά για μεγάλα σύνολα δεδομένων ή όταν χρησιμοποιείται ένας μη γραμμικός πυρήνας (Burges, 1998).

Επιπλέον, η ερμηνεία του μοντέλου αποτελεί ουσιαστικό παράγοντα. Τα δέντρα απόφασης και τα τυχαία δάση τείνουν να είναι πιο ερμηνεύσιμα από τα MLP και τα SVM. Το πρώτο παρέχει ένα σαφές σύνολο κανόνων που οδηγούν σε μια πρόβλεψη, ενώ το δεύτερο, ειδικά με έναν μη γραμμικό πυρήνα ή πολλαπλά κρυφά επίπεδα, μπορεί να θεωρηθεί ως ένα «μαύρο κουτί» που παράγει προβλέψεις χωρίς μια σαφή, κατανοητή λογική (Breiman, 2001).

Εξετάζοντας τα αποτελέσματα, φαίνεται ότι ο ταξινομητής Random Forest ξεπερνά τα άλλα μοντέλα. Τα Random Forests είναι μοντέλα συνόλου που αξιοποιούν τη δύναμη πολλαπλών δέντρων

αποφάσεων για να κάνουν προβλέψεις. Είναι γνωστά για την στιβαρότητά τους έναντι της υπερβολικής προσαρμογής, την ικανότητά τους να χειρίζονται μεγάλο αριθμό χαρακτηριστικών και την υψηλή προγνωστική τους ακρίβεια (Breiman, 2001).

Παρόλο που ο ταξινομητής MLP δεν έδωσε τα καλύτερα αποτελέσματα σε αυτή τη συγκεκριμένη εργασία, αξίζει να σημειωθεί ότι τα MLP είναι ένα ισχυρό εργαλείο ικανό να μοντελοποιεί πολύπλοκες μη γραμμικές σχέσεις. Λειτουργούν ιδιαίτερα καλά όταν τα δεδομένα περιέχουν εισόδους υψηλών διαστάσεων και μη γραμμικά διαχωρίσιμες κλάσεις (Goodfellow et al., 2016).

Το SVM, παρά την υπολογιστική του πολυπλοκότητα, μπορεί να μοντελοποιήσει μη γραμμικά όρια απόφασης και είναι ανθεκτικό σε υπερπροσαρμογή, ειδικά σε χώρους υψηλών διαστάσεων. Ωστόσο, απαιτεί προσεκτική επιλογή του πυρήνα και των παραμέτρων του (Cortes & Vapnik, 1995).

Όσον αφορά τον ταξινομητή Decision Tree, είναι διαισθητικός και εύκολος στην ερμηνεία του, αλλά τείνει να είναι λιγότερο ακριβής σε σύγκριση με μεθόδους συνόλου ή πιο πολύπλοκα μοντέλα όπως τα MLP ή τα SVM (Quinlan, 1986).

Συμπερασματικά, ο ταξινομητής Random Forest φαίνεται να είναι το πιο αποτελεσματικό μοντέλο για τη συγκεκριμένη εργασία. Ωστόσο, είναι σημαντικό να σημειωθεί ότι το "καλύτερο" μοντέλο μπορεί να διαφέρει ανάλογα με την εργασία, τα δεδομένα που υπάρχουν και τις συγκεκριμένες απαιτήσεις της εφαρμογής. Επιπλέον, η χρήση της μεθόδου Monte Carlo και η διασταυρούμενη επικύρωση διασφαλίζουν ότι η αξιολόγηση αυτών των μοντέλων είναι ισχυρή και αξιόπιστη.

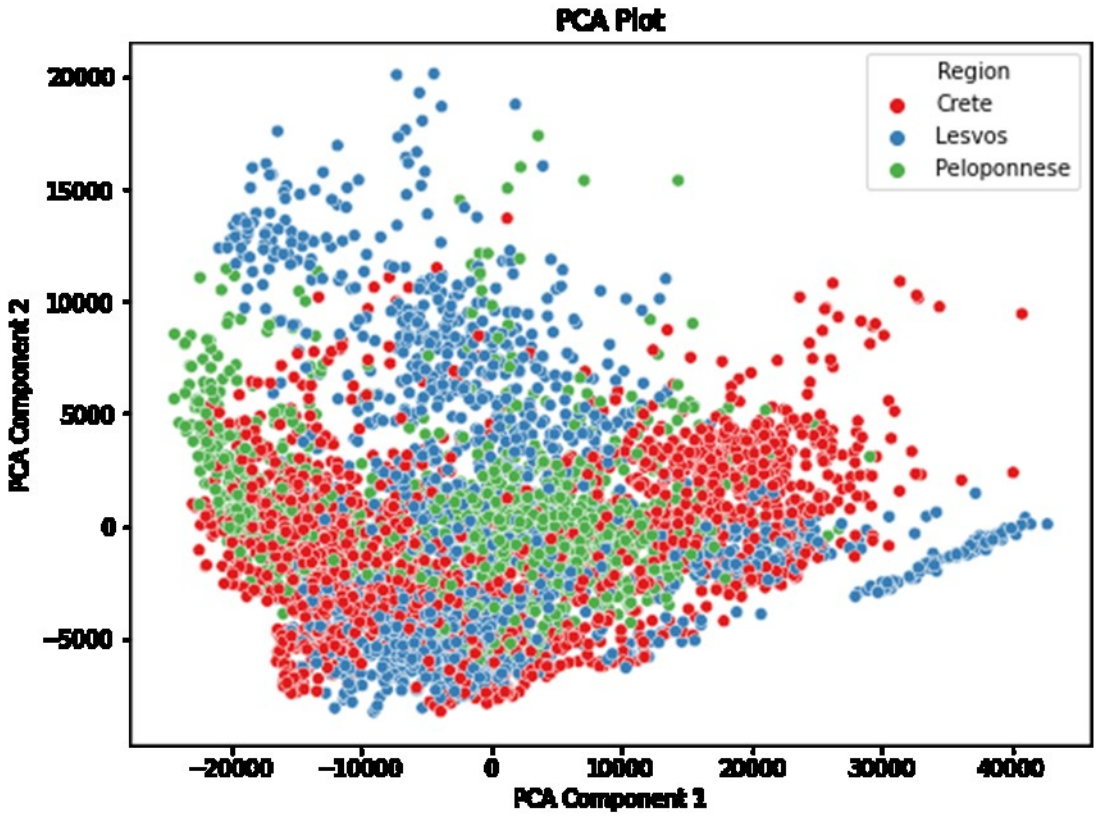
5.6.3 PCA Plot

Η Ανάλυση Κύριων Συνιστωσών (PCA) είναι μια αρκετά δημοφιλής στατιστική τεχνική που χρησιμοποιείται για τη μείωση των διαστάσεων των χαρακτηριστικών. Βασίζεται στην ανακατασκευή των χαρακτηριστικών, που μπορεί να συσχετίζονται μεταξύ τους, σε ένα σύνολο νέων χαρακτηριστικών, τα οποία είναι γραμμικά ανεξάρτητα μεταξύ τους και ονομάζονται Κύριες Συνιστώσες ή Principal Components. Αυτό γίνεται με σκοπό τη διατήρηση της σημαντικής πληροφορίας των αρχικών χαρακτηριστικών με λιγότερες διαστάσεις.

Η διαδικασία ξεκινά με την τυποποίηση των χαρακτηριστικών, ώστε να έχουν μέσο όρο 0 και τυπική απόκλιση 1. Στη συνέχεια, δημιουργούνται συνδυασμοί των χαρακτηριστικών ώστε να μην υπάρχει συσχέτιση μεταξύ τους. Οι διακυμάνσεις αναδιανέμονται έτσι ώστε η πρώτη Κύρια Συνιστώσα να περιέχει το μεγαλύτερο ποσοστό της συνολικής μεταβλητότητας. Έπειτα, οι υπόλοιπες Κύριες Συνιστώσες δημιουργούνται σε αύξουσα σειρά σημαντικότητας, λαμβάνοντας υπόψη τις διακυμάνσεις.

Κάθε Κύρια Συνιστώσα είναι ένας γραμμικός συνδυασμός των αρχικών χαρακτηριστικών, όπου οι συντελεστές στάθμισης υπολογίζονται έτσι ώστε να διασφαλίζεται η μέγιστη διακύμανση για κάθε Κύρια Συνιστώσα. Οι συσχετίσεις μεταξύ των αρχικών χαρακτηριστικών υπολογίζονται μέσω του πίνακα C, που περιέχει τις συνδιακυμάνσεις τους. Σημαντικό είναι να σημειώσουμε ότι αρχικά τα χαρακτηριστικά πρέπει να τυποποιηθούν, έτσι ώστε να έχουν μέση τιμή 0 και τυπική απόκλιση 1, προκειμένου να είναι αποτελεσματική η διαδικασία.

Τέλος, παρατηρείται ότι η επιτυχία της τεχνικής PCA εξαρτάται σε μεγάλο βαθμό από τον βαθμό συσχέτισης μεταξύ των αρχικών χαρακτηριστικών. Χαρακτηριστικά που έχουν πολύ υψηλή συσχέτιση ($r > \pm 0,99$) θεωρούνται πλεονάζοντα και δεν συμβάλλουν σημαντικά στη διαδικασία PCA, επομένως απορρίπτονται πριν την εφαρμογή της τεχνικής.



Εικόνα 9: Διάγραμμα PCA

5.6.4 Πληροφορίες και προκλήσεις με πολλαπλή επικύρωση

Η διασταυρούμενη επικύρωση είναι μια σημαντική στατιστική τεχνική που εφαρμόζεται στον τομέα της μηχανικής μάθησης για την αξιολόγηση της απόδοσης γενίκευσης των μοντέλων. Μας επιτρέπει να αποκτήσουμε μια πιο αμερόληπτη εκτίμηση της απόδοσης και της ευρωστίας του μοντέλου σε σύγκριση με την παραδοσιακή μέθοδο διαχωρισμού δοκιμής αμαξοστοιχίας. Επιπλέον, βοηθά στον

συντονισμό υπερπαραμέτρων βοηθώντας στην επιλογή του βέλτιστου συνόλου παραμέτρων που παράγουν το πιο ακριβές μοντέλο (Kohavi, 1995).

Σε αυτή την εργασία, εφαρμόστηκε η μέθοδος διασταυρούμενης επικύρωσης 4 φορές. Σύμφωνα με τα αποτελέσματα, οι βαθμολογίες διασταυρούμενης επικύρωσης διέφεραν σημαντικά μεταξύ διαφορετικών μοντέλων. Ο ταξινομητής Random Forest φάνηκε να έχει την υψηλότερη μέση βαθμολογία διασταυρούμενης επικύρωσης, υποδηλώνοντας ανώτερη ικανότητα γενίκευσης σε σύγκριση με τα άλλα μοντέλα.

Μια βασική ιδέα που προέκυψε από τη χρήση της διασταυρούμενης επικύρωσης σε αυτή τη μελέτη είναι η αποτελεσματικότητά της στην μετρίαση του κινδύνου υπερβολικής προσαρμογής. Χρησιμοποιώντας διαφορετικά υποσύνολα δεδομένων για εκπαίδευση και δοκιμή κατά τη διάρκεια κάθε πτυχής, παρέχει μια πιο αυστηρή αξιολόγηση της ικανότητας του μοντέλου να γενικεύει σε μη ορατά δεδομένα. Αυτό βοηθά στην αποφυγή του υπερβολικά πολύπλοκου χαρακτήρα των μοντέλων και της προσαρμογής τους πολύ κοντά στα δεδομένα εκπαίδευσης, κάτι που θα παρεμπόδιζε την απόδοσή τους σε νέα δεδομένα (Hastie et al., 2009).

Παρά τα πλεονεκτήματά της, η εφαρμογή διασταυρούμενης επικύρωσης εμφανίστηκε επίσης αρκετές προκλήσεις. Ένα από τα κύρια ζητήματα είναι το υπολογιστικό κόστος, ιδιαίτερα όταν έχουμε να κάνουμε με μεγάλα σύνολα δεδομένων και υπολογιστικά εντατικά μοντέλα όπως τα MLP και τα SVM. Η ανάγκη εκπαίδευσης και δοκιμής των χρόνων «k» του μοντέλου για διασταυρούμενη επικύρωση «k» μπορεί να αυξήσει δραματικά το υπολογιστικό κόστος, το οποίο μπορεί να μην είναι εφικτό σε ορισμένα σενάρια (Varma & Simon, 2006).

Επιπλέον, η επιλογή του «k» σε διασταυρούμενη επικύρωση k-fold μπορεί να επηρεάσει σημαντικά τα αποτελέσματα. Η διακύμανση της εκτίμησης που προκύπτει, μειώνεται καθώς το 'k' αυξάνεται, αλλά η προκατάληψη αυξάνεται. Έτσι, υπάρχει μια αντιστάθμιση μεταξύ της μεροληψίας και της διακύμανσης που πρέπει να αντιμετωπιστεί προσεκτικά (James et al., 2013).

Μια άλλη πρόκληση αφορά την υπόθεση της ανεξαρτησίας μεταξύ των παρατηρήσεων στις οποίες βασίζεται η διασταυρούμενη επικύρωση. Σε περιπτώσεις όπου οι παρατηρήσεις δεν είναι ανεξάρτητες, όπως δεδομένα χρονοσειρών ή δεδομένα με εγγενείς δομές ομάδας, η διασταυρούμενη επικύρωση θα μπορούσε να αποφέρει υπερβολικά αισιόδοξα αποτελέσματα που δεν αντιπροσωπεύουν με ακρίβεια την ικανότητα γενίκευσης του μοντέλου (Roberts et al., 2017).

Συμπερασματικά, η διασταυρούμενη επικύρωση, αν και ένα ισχυρό εργαλείο για την αξιολόγηση και τη βελτίωση της απόδοσης του μοντέλου Monte Carlo, παρουσιάζει επίσης το δικό της σύνολο προκλήσεων που πρέπει να εξεταστούν προσεκτικά. Για αυτήν την εργασία, τα οφέλη της στη βελτίωση της ευρωστίας των αξιολογήσεων απόδοσης των μοντέλων μηχανικής μάθησης υπερτερούσαν του υπολογιστικού κόστους και των άλλων προκλήσεων που αντιμετωπίζουν.

5.6.5 Προοπτικές της μεθόδου Monte Carlo στη νευρωνική αξιολόγηση

Η μέθοδος Monte Carlo (MC) έχει καθιερωθεί ως ένα ευέλικτο και ισχυρό εργαλείο στη σφαίρα της μηχανικής μάθησης, ιδιαίτερα στη νευρωνική αξιολόγηση, η οποία είναι η διαδικασία αξιολόγησης και συντονισμού μοντέλων νευρωνικών δικτύων.

Η υποκείμενη δύναμη της μεθόδου MC είναι η βάση της στη στοχαστική δειγματοληψία, επιτρέποντάς της να διερευνήσει την απόδοση ενός νευρωνικού δικτύου σε έναν τεράστιο χώρο παραμέτρων. Με την προσομοίωση διαφορετικών αποτελεσμάτων, μπορεί να προσφέρει αξιόπιστες εκτιμήσεις για το σφάλμα γενίκευσης ενός μοντέλου, να προσδιορίσει τις βέλτιστες υπερπαραμέτρους και να παρέχει πληροφορίες για την αβεβαιότητα του μοντέλου (Neal, 2012).

Συγκεκριμένα, σε αυτή τη μελέτη, η μέθοδος MC ήταν καθοριστική για την αξιολόγηση του ταξινομητή MLP. Βοήθησε στη μέτρηση της μεταβλητότητας στην απόδοση του μοντέλου λόγω της εγγενούς στοχαστικής φύσης του και στον προσδιορισμό της μέσης απόδοσής του σε διαφορετικές αρχικοποιήσεις και διαχωρισμούς δεδομένων.

Κοιτάζοντας το μέλλον, η μέθοδος MC έχει πολλά υποσχόμενες προοπτικές στο πλαίσιο της νευρωνικής αξιολόγησης. Καθώς τα νευρωνικά δίκτυα συνεχίζουν να αυξάνονται σε πολυπλοκότητα και διαστάσεις, η μέθοδος MC παρέχει έναν ισχυρό τρόπο πλοήγησης σε αυτόν τον υψηλών διαστάσεων χώρο και εντοπισμό βέλτιστων διαμορφώσεων. Μπορεί επίσης να χρησιμοποιηθεί σε μεθόδους συνόλου όπου εκπαιδεύονται πολλαπλά νευρωνικά δίκτυα και υπολογίζεται ο μέσος όρος των προβλέψεών τους, βελτιώνοντας τη συνολική απόδοση και την ευρωστία του μοντέλου (Goodfellow et al., 2016).

Ένας αναδυόμενος τομέας ενδιαφέροντος είναι τα Bayesian Neural Networks, όπου η μέθοδος MC, ιδιαίτερα η Markov Chain Monte Carlo (MCMC), μπορεί να χρησιμοποιηθεί για την εκτίμηση της μεταγενέστερης κατανομής των βαρών του δικτύου. Αυτή η προσέγγιση προσφέρει έναν τρόπο καταγραφής της αβεβαιότητας στις προβλέψεις του μοντέλου, που μπορεί να είναι κρίσιμης σημασίας σε ευαίσθητες εφαρμογές όπου μια εσφαλμένη πρόβλεψη μπορεί να έχει σοβαρές συνέπειες (Neal, 2012).

Ωστόσο, η μέθοδος MC δεν είναι χωρίς προκλήσεις. Μπορεί να είναι υπολογιστικά εντατικό, ειδικά για πολύπλοκα νευρωνικά δίκτυα και μεγάλα σύνολα δεδομένων. Επίσης, η εξάρτησή του από την τυχαιότητα σημαίνει ότι τα αποτελέσματα μπορεί να διαφέρουν μεταξύ διαφορετικών εκτελέσεων. Παρά αυτές τις προκλήσεις, τα οφέλη της μεθόδου MC στην προσφορά ισχυρών, ολοκληρωμένων αξιολογήσεων την καθιστούν ένα πολλά υποσχόμενο εργαλείο για νευρωνική αξιολόγηση.

5.7 Περίληψη

Σε αυτό το κεφάλαιο, εμβαθύναμε στη διαδικασία εφαρμογής τεχνικών μηχανικής εκμάθησης για την πρόβλεψη της αυθεντικότητας του λαδιού και της γεωγραφικής προέλευσης χρησιμοποιώντας δεδομένα LIBS. Εκπαιδεύτηκαν πολλά μοντέλα, συμπεριλαμβανομένων MLP, Decision Trees, Random Forests και SVMs και συγκρίθηκαν οι επιδόσεις τους.

Ο ταξινομητής Random Forest αναδείχθηκε ως το μοντέλο με τις καλύτερες επιδόσεις, επιδεικνύοντας τη δύναμη των μεθόδων συνόλου στον χειρισμό πολύπλοκων εργασιών ταξινόμησης. Ωστόσο, σημειώθηκε επίσης ότι το βέλτιστο μοντέλο μπορεί να ποικίλλει ανάλογα με τις ειδικές απαιτήσεις της εφαρμογής και τη φύση των δεδομένων.

Η διασταυρούμενη επικύρωση αναγνωρίστηκε ως ένα κρίσιμο εργαλείο για την αξιολόγηση της απόδοσης του μοντέλου, την μετρίαση της υπερπροσαρμογής και τη βελτίωση της ευρωστίας του μοντέλου. Παρά το υπολογιστικό του κόστος και την υπόθεση ανεξάρτητων παρατηρήσεων, τα οφέλη της διασταυρούμενης επικύρωσης στην προσφορά μιας αξιόπιστης εκτίμησης απόδοσης ήταν σαφή.

Ο ρόλος της μεθόδου Monte Carlo στη νευρωνική αξιολόγηση διερευνήθηκε επίσης. Επιτρέποντας τη στοχαστική δειγματοληψία σε διαφορετικά αποτελέσματα, η μέθοδος MC παρέχει αξιόπιστες εκτιμήσεις για την απόδοση ενός νευρωνικού δικτύου, βοηθά στον εντοπισμό βέλτιστων υπερπαραμέτρων και καταγράφει την αβεβαιότητα του μοντέλου. Παρά ορισμένες προκλήσεις, η μέθοδος MC έχει πολλά υποσχόμενες προοπτικές στη νευρωνική αξιολόγηση, ειδικά καθώς τα νευρωνικά δίκτυα συνεχίζουν να αυξάνονται σε πολυπλοκότητα.

Συμπερασματικά, η εφαρμογή τεχνικών μηχανικής μάθησης, που υποστηρίζονται από ισχυρές μεθόδους αξιολόγησης, όπως η διασταυρούμενη επικύρωση και η μέθοδος Monte Carlo, καταδεικνύει ενδιαφέρουσες δυνατότητες στον τομέα της γνησιότητας των τροφίμων και του προσδιορισμού της προέλευσης. Η μελλοντική εργασία μπορεί να επεκταθεί σε αυτές τις τεχνικές, να εξερευνήσει άλλα μοντέλα και να ενσωματώσει μεγαλύτερα, πιο διαφορετικά σύνολα δεδομένων για τη βελτίωση της ακρίβειας πρόβλεψης και της γενίκευσης.

6

Συμπεράσματα και Προοπτικές

6.1 Περίληψη ευρημάτων

Κατά τη διάρκεια αυτής της διπλωματικής εργασίας, η εστίαση επικεντρώθηκε κυρίως στον προσδιορισμό της γνησιότητας και της γεωγραφικής προέλευσης του ελαιόλαδου χρησιμοποιώντας δεδομένα φασματοσκοπίας διάσπασης που προκαλείται από λέιζερ (LIBS). Αξιοποιώντας μεθοδολογίες μηχανικής μάθησης, ιδιαίτερα Perceptrons Multi-Layer (MLPs), Decision Trees, Random Forests και Support Vector Machines (SVM), δημιουργήσαμε με επιτυχία μοντέλα πρόβλεψης και αναλύσαμε την απόδοσή τους.

Βρήκαμε ότι ο ταξινομητής Random Forest ξεπέρασε τα άλλα μοντέλα όσον αφορά την ακρίβεια πρόβλεψης και τη γενίκευση. Αυτό επιβεβαίωσε τη δύναμη των μεθόδων συνόλου στον χειρισμό πολύπλοκων προβλημάτων ταξινόμησης πολλαπλών τάξεων. Η διασταυρούμενη επικύρωση έπαιξε κρίσιμο ρόλο στη λήψη μιας πιο αμερόληπτης εκτίμησης απόδοσης, ενώ η μέθοδος Monte Carlo βοήθησε στην παροχή μιας ισχυρής κατανόησης της συμπεριφοράς του μοντέλου υπό διαφορετικές συνθήκες.

Μία από τις αξιοσημείωτες πτυχές ήταν η εφαρμογή του Monte Carlo Cross-Validation (MCCV) στην αξιολόγηση των νευρωνικών δικτύων, δείχνοντας πώς βοήθησε στην αποτύπωση της εγγενούς μεταβλητότητας και πολυπλοκότητας αυτών των μοντέλων. Σε αυτό το πλαίσιο, οι μέθοδοι Monte Carlo εμφανίστηκαν ως ένα πολλά υποσχόμενο εργαλείο για νευρωνική αξιολόγηση, ιδιαίτερα για την κατανόηση της συμπεριφοράς πολύπλοκων νευρωνικών δικτύων και για το χειρισμό του χώρου υψηλών διαστάσεων.

6.2 Αποτελέσματα της Μελέτης

Αυτή η έρευνα χρησιμεύει για να αναδείξει τις δυνατότητες των μεθοδολογιών μηχανικής μάθησης στο πλαίσιο της επαλήθευσης της γνησιότητας των τροφίμων και του προσδιορισμού της γεωγραφικής προέλευσης. Η χρήση της τεχνολογίας LIBS σε συνδυασμό με αλγόριθμους μηχανικής μάθησης θα μπορούσε να ανοίξει το δρόμο για ένα γρήγορο, μη καταστροφικό και αξιόπιστο σύστημα ανάλυσης τροφίμων, φέρνοντας δυνητικά επανάσταση στον τομέα της επιστήμης των τροφίμων και του ποιοτικού ελέγχου.

Οι επιπτώσεις αυτής της μελέτης εκτείνονται πέρα από το πεδίο της ανάλυσης του ελαιόλαδου. Οι τεχνικές και οι μεθοδολογίες που εφαρμόζονται εδώ θα μπορούσαν να προσαρμοστούν για άλλα προϊόντα διατροφής και, ευρύτερα, για οποιοδήποτε πρόβλημα ταξινόμησης όπου υπάρχουν διαθέσιμα φασματοσκοπικά δεδομένα. Ως εκ τούτου, αυτή η έρευνα έχει πιθανές επιπτώσεις σε πολλούς τομείς, από τη γεωργία και την περιβαλλοντική επιστήμη μέχρι την υγειονομική περίθαλψη και την επιστήμη των υλικών.

Η έρευνα υπογραμμίζει επίσης τη σημασία των ισχυρών τεχνικών αξιολόγησης μοντέλων σε εφαρμογές μηχανικής εκμάθησης. Η χρήση της διασταυρούμενης επικύρωσης και της μεθόδου Monte Carlo επέτρεψαν μια πιο αξιόπιστη εκτίμηση της απόδοσης του μοντέλου, υπογραμμίζοντας τη σημασία τους για τον μετριασμό της υπερπροσαρμογής και την καταγραφή της μεταβλητότητας του μοντέλου.

6.3 Συστάσεις για μελλοντική έρευνα

Ενώ αυτή η έρευνα έχει δείξει τις δυνατότητες των τεχνικών μηχανικής μάθησης στην ανάλυση ελαιόλαδου, υπάρχουν αρκετές κατευθύνσεις για μελλοντική εργασία. Πρώτον, θα μπορούσαν να διερευνηθούν άλλα μοντέλα μηχανικής μάθησης. Οι αλγόριθμοι βαθιάς μάθησης, όπως τα συνελικτικά νευρωνικά δίκτυα (CNN) και τα επαναλαμβανόμενα νευρωνικά δίκτυα (RNN), έχουν δείξει αξιοσημείωτη απόδοση στην αναγνώριση προτύπων και θα μπορούσαν ενδεχομένως να ενισχύσουν την προγνωστική ισχύ των μοντέλων.

Η χρήση τεχνικών επιλογής χαρακτηριστικών ή εξαγωγής θα μπορούσε επίσης να διερευνηθεί για τον εντοπισμό των πιο κατατοπιστικών φασματικών γραμμών στα δεδομένα LIBS. Τεχνικές όπως η ανάλυση κύριου στοιχείου (PCA) ή οι αλγόριθμοι επιλογής φασματικών χαρακτηριστικών θα μπορούσαν να βοηθήσουν στη μείωση της διάστασης των δεδομένων και ενδεχομένως να βελτιώσουν την απόδοση του μοντέλου που τρέξαμε.

Περαιτέρω έρευνα θα μπορούσε επίσης να εμβαθύνει στη χρήση των μεθόδων Monte Carlo για νευρωνική αξιολόγηση, ιδιαίτερα στο πλαίσιο των νευρωνικών δικτύων Bayes. Ο συνδυασμός τεχνικών Bayes και νευρωνικών δικτύων παρουσιάζει συναρπαστικές δυνατότητες, συμπεριλαμβανομένης της βελτιωμένης εκτίμησης αβεβαιότητας και της ερμηνευσιμότητας του μοντέλου.

Όσον αφορά τα δεδομένα, η μελλοντική έρευνα θα μπορούσε να επικεντρωθεί στην ενσωμάτωση μεγαλύτερων και πιο διαφορετικών συνόλων δεδομένων, συμπεριλαμβανομένων δειγμάτων από διαφορετικές γεωγραφικές περιοχές και διαφορετικές συνθήκες καλλιέργειας. Αυτό πιθανότατα θα βελτίωνε τη γενίκευση των μοντέλων και θα επέτρεπε πιο διαφοροποιημένες εργασίες ταξινόμησης.

Τέλος, η ανάπτυξη αυτών των μοντέλων στον πραγματικό κόσμο θα ήταν μια ενδιαφέρουσα κατεύθυνση για μελλοντική έρευνα. Αυτό θα μπορούσε να περιλαμβάνει την ανάπτυξη μιας φιλικής προς τον χρήστη διεπαφής για μη ειδικούς χρήστες ή την ενσωμάτωση των μοντέλων σε ένα αυτοματοποιημένο σύστημα ποιοτικού ελέγχου στην αλυσίδα παραγωγής ελαιόλαδου.

Βιβλιογραφία- Αναφορές

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, *13*(Feb), 281–305. http://www.jmlr.org/papers/v13/bergstra12a.html

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Cremers, D. A., & Radziemski, L. J. (2006). *Handbook of Laser-Induced Breakdown Spectroscopy*. Wiley. https://doi.org/10.1002/0470093013

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random Forests for Classification in Ecology. *Ecology*, *88*(11), 2783–2792. https://doi.org/10.1890/07-0539.1

Díaz-Uriarte, R., & Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, *7*(1), 3. https://doi.org/10.1186/1471-2105-7-3

Gondal, M. A., Hussain, T., Yamani, Z. H., & Baig, M. A. (2006). Detection of heavy metals in Arabian crude oil residue using laser induced breakdown spectroscopy. *Talanta*, *69*(5), 1072–1078. https://doi.org/10.1016/j.talanta.2005.11.023

Goodacre, R. (2004). Making sense of the metabolome using evolutionary computation: seeing the wood with the trees. *Journal of Experimental Botany*, *56*(410), 245–254. https://doi.org/10.1093/jxb/eri043

Hahn, D. W., & Omenetto, N. (2012). Laser-Induced Breakdown Spectroscopy (LIBS), Part II: Review of Instrumental and Methodological Approaches to Material Analysis and Applications to Different Fields. *Applied Spectroscopy*, *66*(4), 347–419. https://doi.org/10.1366/11-06574

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2). Springer.

Hinton, G., & Sejnowski, T. J. (1999). *Unsupervised learning: foundations of neural computation*. MIT press.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., & others. (2003). *A practical guide to support vector classification*. Taipei, Taiwan.

James, Gareth., Witten, Daniela., Hastie, Trevor., & Tibshirani, Robert. (2013). *An Introduction to Statistical Learning with Applications in R* (1st ed. 2013.). Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:2702042

Liaw, A., Wiener, M., & others. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.

Metropolis, N., & Ulam, S. (1949). The Monte Carlo Method. *Journal of the American Statistical Association*, *44*(247), 335–341. https://doi.org/10.1080/01621459.1949.10483310

Miziolek, A. W., Palleschi, V., & Schechter, I. (2006). *Laser induced breakdown spectroscopy*. Cambridge university press.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106. https://doi.org/10.1007/BF00116251

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Senesi, G. S., Dell'Aglio, M., Gaudiuso, R., De Giacomo, A., Zaccone, C., De Pascale, O., Miano, T. M., & Capitelli, M. (2009). Heavy metal concentrations in soils as determined by laser-induced breakdown spectroscopy (LIBS), with special emphasis on chromium. *Environmental Research*, *109*(4), 413–420. https://doi.org/10.1016/j.envres.2009.02.005

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(1), 307. https://doi.org/10.1186/1471-2105-9-307

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, *7*(1), 91. https://doi.org/10.1186/1471-2105-7-91

Xu, Q.-S., & Liang, Y.-Z. (2001). Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems*, *56*(1), 1–11. https://doi.org/10.1016/S0169-7439(00)00122-2

Σύντομο Βιογραφικό Συγγραφέα

Ιωάννης Μακαντάσης

Φοιτητής στο Τμήμα Μηχανικών Η/Υ & Πληροφορικής, Πανεπιστήμιο Πατρών

Γεννήθηκα στην Αθήνα το 1998 και επέλεξα να αφοσιωθώ στον κόσμο της τεχνολογίας. Παρακολουθώντας το πρόγραμμα σπουδών του Τμήματος Μηχανικών Η/Υ & Πληροφορικής στο Πανεπιστήμιο Πατρών, έχω αποκτήσει γνώσεις βαθιάς κατανόησης της επιστήμης της πληροφορικής.

Επιπλέον, εργάζομαι ως System Integration Engineer στην εταιρία SPACE HELLAS SA. Στο πλαίσιο αυτού του ρόλου, συμβάλλω στην ενοποίηση και συστηματοποίηση τεχνολογικών λύσεων, προσφέροντας αξιόπιστες λύσεις στον ψηφιακό χώρο.

Έχοντας ολοκληρώσει τις στρατιωτικές μου υποχρεώσεις στην Πολεμική Αεροπορία, ανέπτυξα πνεύμα συνεργασίας, ηγετικές ικανότητες και ανθεκτικότητα σε πιέσεις.

Με ενδιαφέρον για την εξέλιξη της τεχνολογίας, αναζητώ δυνατότητες να συνεχίσω να επεκτείνομαι επαγγελματικά και να συνεισφέρω στον τομέα της πληροφορικής και της τεχνολογίας