



Πολυτεχνική Σχολή
Τμήμα Μηχανικών Η/Υ & Πληροφορικής

Διπλωματική Εργασία

Γεωγραφική ταυτοποίηση ελαίων με τεχνικές μηχανικής μάθησης

Παπαγεωργίου Δημήτριος

A.M. 1064280

Επιβλέπων

Μακρής Χρήστος

Αναπληρωτής Καθηγητής

Συνεπιβλέπων

Ηλίας Αριστείδης

ΕΔΙΠ

Πάτρα, 2024

© Copyright συγγραφής Παπαγεωργίου Δημήτριος, 2024

© Copyright θέματος Αριστείδης Ηλίας

Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Η έγκριση της διπλωματικής εργασίας από το Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών & Πληροφορικής του Πανεπιστημίου Πατρών δεν υποδηλώνει απαραίτητως και αποδοχή των απόψεων του συγγραφέα εκ μέρους του Τμήματος.

Στους γονείς μου

Περίληψη

Το ελαιόλαδο αποτελεί ένα προϊόν μεγάλης αξίας, ιδιαίτερα για τις χώρες με μεσογειακό κλίμα. Η ποιότητα και η προέλευση του καθορίζουν την εμπιστοσύνη των καταναλωτών και τις τιμές στην αγορά. Ωστόσο, ζητήματα νοθείας και ανακριβών δηλώσεων σχετικά με τη γεωγραφική προέλευση πλήττουν τη φήμη του κλάδου, με την ανάμειξη ελαιόλαδου με κατώτερης ποιότητας λάδια να είναι συχνά φαινόμενο. Οι παραδοσιακές μέθοδοι επαλήθευσης είναι χρονοβόρες και απαιτούν εξειδικευμένη γνώση, γεγονός που καθιστά αναγκαία την ανάπτυξη καινοτόμων προσεγγίσεων.

Η παρούσα διπλωματική εργασία στοχεύει στην ανάπτυξη και αξιολόγηση ενός συστήματος που προβλέπει με ακρίβεια τη γεωγραφική προέλευση του ελαιόλαδου μέσω ανάλυσης φασματοσκοπικών δεδομένων. Για τον σκοπό αυτό ελαιόλαδα διαφορετικών γεωγραφικών προελεύσεων της Πελοποννήσου, αναλύθηκαν μέσω φασματοσκοπίας FTIR (Fourier Transform Infrared). Τα δεδομένα αυτά προεπεξεργάστηκαν και αναλύθηκαν με διάφορους αλγόριθμους μηχανικής μάθησης, τόσο ταξινόμησης όσο και ομαδοποίησης.

Η εργασία φιλοδοξεί να προσφέρει ένα καινοτόμο εργαλείο για την εξασφάλιση ποιότητας και αυθεντικότητας του ελαιόλαδου. Το μοντέλο πρόβλεψης που αναπτύχθηκε θα μπορεί να συμβάλλει μελλοντικά στη βελτίωση της διαφάνειας στον αγροδιατροφικό τομέα.

Abstract

Olive oil is a product of great value, especially for countries with a Mediterranean climate. Its quality and origin determine consumer confidence and market prices. However, issues of adulteration and inaccurate declarations of geographical origin are damaging the reputation of the sector, with the mixing of olive oil with inferior oils being a frequent occurrence. Traditional verification methods are time-consuming and require specialized knowledge, which necessitates the development of innovative approaches

This thesis aims to develop and evaluate a system that accurately predicts the geographical origin of olive oil through spectroscopic data analysis. For this purpose, olive oils of different geographical origins of the Peloponnese were analyzed by Fourier Transform Infrared (FTIR) spectroscopy. These data were pre-processed and analyzed using various machine learning algorithms, both classification and clustering.

The work aspires to provide an innovative tool to ensure the quality and authenticity of olive oil. The developed prediction model will be able to contribute in the future to improve transparency in the agri-food sector.

Περιεχόμενα

<i>Περίληψη</i>	<i>ix</i>
<i>Abstract</i>	<i>x</i>
<i>Περιεχόμενα</i>	<i>xi</i>
1. Εισαγωγή	14
1.1 Ερευνητικό πρόβλημα	14
1.2 Σκοπός της εργασίας	14
1.3 Συνεισφορά	15
1.4 Διάρθρωση	15
2. Το ελαιόλαδο	18
2.1 Η χημική σύσταση του ελαιόλαδου	18
2.2 Παράγοντες που επηρεάζουν την ποιότητα του ελαιόλαδου	18
2.2.1 Το μικροκλίμα	20
2.3 Νοθεία	22
3. Φασματοσκοπία	26
4. Επεξεργασία δεδομένων FTIR	29
4.1 Καθαρισμός δεδομένων	29
4.2 Principal Component Analysis	29
4.3 Binning	30
5. Μηχανική μάθηση	33
Εποπτευόμενη μάθηση	34
5.1 Ταξινόμηση	35
5.2 Μέθοδοι επικύρωσης	36
5.2.1 Μέθοδος διαχωρισμού	36
5.2.2 Μέθοδος διασταυρούμενης επικύρωσης	37
5.3 Αλγόριθμοι ταξινόμησης	38
5.3.1 Δέντρα αποφάσεων	39
5.3.2 Navie Bayes	42
5.3.3 Support Vector Machine	45
5.3.4 Random Forest	46
5.3.5 Linear Discriminant Analysis	47
Μάθηση χωρίς επίβλεψη	49
5.4 Ομαδοποίηση	50

5.5 Αλγόριθμοι ομαδοποίησης.....	52
5.5.1 K-means.....	52
5.5.2 Agglomerative Method	54
5.5.3 DBSCAN.....	56
5. Λογισμικά που χρησιμοποιήθηκαν	60
5.1 Google Colab.....	60
6. Υλοποίηση	62
6.1 Συλλογή και Προεπεξεργασία Δεδομένων.....	62
6.2 Ταξινόμηση με τη μέθοδο διαχωρισμού.....	69
6.2.1 Linear Discriminant Analysis.....	69
6.2.2 Decision Tree Classifier.....	70
6.2.3 Logistic Regression.....	71
6.2.4 Gaussian Navie Bayes.....	72
6.2.5 Support Vector machine.....	73
6.2.6 Random Forest	74
6.3 Ταξινόμηση με τη μέθοδο διασταυρούμενη επικύρωσης	75
6.3.1 Linear Discriminant Analysis.....	75
6.3.2 Decision Tree Classifier.....	77
6.3.3 Logistic Regression.....	79
6.3.4 Gaussian Navie Bayes.....	80
6.3.5 Support Vector machine.....	82
6.3.6 Random Forest	84
6.4 Ομαδοποίηση	86
6.4.1 K-Means	86
6.4.2 DBSCAN.....	86
6.4.3 Agglomerative Method	87
7. Συζήτηση	92
7.1 Συμπεράσματα	92
7.2 Μελλοντική Εργασία	94
8. Βιβλιογραφία	96

1. Εισαγωγή

1.1 Ερευνητικό πρόβλημα

Η κύρια πρόκληση στη βιομηχανία ελαιόλαδου είναι η διασφάλιση της αυθεντικότητας και της γεωγραφικής προέλευσης των προϊόντων. Το ελαιόλαδο είναι ένα προϊόν υψηλής αξίας και τόσο η ποιότητα όσο και η προέλευσή του επηρεάζουν σημαντικά την εμπιστοσύνη των καταναλωτών και τις τιμές της αγοράς. Ωστόσο, ο κλάδος μαστίζεται από τα ζητήματα της νοθείας και της λανθασμένης επισήμανσης. Η νοθεία περιλαμβάνει την ανάμειξη ελαιόλαδου με λάδια χαμηλότερης ποιότητας, ενώ η λανθασμένη επισήμανση αφορά την ψευδή δήλωση της γεωγραφικής προέλευσης του λαδιού. Και οι δύο πρακτικές εξαπατούν τους καταναλωτές και θέτουν άδικα σε μειονεκτική θέση τους έντιμους παραγωγούς.

Οι παραδοσιακές μέθοδοι για την επαλήθευση της γεωγραφικής προέλευσης του ελαιόλαδου, όπως η αισθητηριακή αξιολόγηση και η χημική ανάλυση, είναι χρονοβόρες και απαιτούν εξειδικευμένη τεχνογνωσία. Αυτές οι μέθοδοι μπορεί επίσης να είναι ασυνεπείς, καθώς βασίζονται στην υποκειμενική ανθρώπινη κρίση και μπορεί να μην εντοπίζουν πάντα περίπλοκες τεχνικές απάτης. Επιπλέον, ενώ η χημική ανάλυση παρέχει κάποιες γνώσεις, μπορεί να μην καταγράψει την πλήρη πολυπλοκότητα της σύνθεσης του ελαιόλαδου που επηρεάζεται από τη γεωγραφική του προέλευση (Ballin, 2010; Lerma-García et al., 2010).

1.2 Σκοπός της εργασίας

Δεδομένων αυτών των προκλήσεων, υπάρχει επείγουσα ανάγκη για μια πιο αξιόπιστη, αποτελεσματική και αυτοματοποιημένη προσέγγιση για τον έλεγχο της ταυτότητας της γεωγραφικής προέλευσης του ελαιόλαδου. Αυτή η έρευνα αντιμετωπίζει αυτό το πρόβλημα αξιοποιώντας δεδομένα φασματοσκοπίας, τα οποία παρέχουν ένα ολοκληρωμένο χημικό προφίλ του ελαιόλαδου, και τεχνικές μηχανικής μάθησης, οι οποίες μπορούν να αναλύσουν πολύπλοκα σύνολα δεδομένων για τον εντοπισμό προτύπων και την πραγματοποίηση ακριβών προβλέψεων. Με την ανάπτυξη προγνωστικών μοντέλων που βασίζονται σε δεδομένα φασματοσκοπίας, αυτή η μελέτη

στοχεύει στη δημιουργία ενός ισχυρού συστήματος ικανού να προσδιορίζει τη γεωγραφική προέλευση των δειγμάτων ελαιόλαδου με υψηλή ακρίβεια.

1.3 Συνεισφορά

Η ικανότητα ακριβούς προσδιορισμού της προέλευσης του ελαιόλαδου, όχι μόνο προστατεύει τους καταναλωτές και διατηρεί την ακεραιότητα της αγοράς, αλλά υποστηρίζει επίσης τις ρυθμιστικές προσπάθειες για την επιβολή νόμων και προτύπων για την επισήμανση. Επιπλέον, ενδυναμώνει τους παραγωγούς που τηρούν τα πρότυπα ποιότητας προστατεύοντας τα προϊόντα τους από τον αθέμιτο ανταγωνισμό. Τελικά, αυτή η έρευνα επιδιώκει να δώσει μια επιστημονικά ορθή και τεχνολογικά προηγμένη λύση σε ένα μακροχρόνιο πρόβλημα στη βιομηχανία ελαιόλαδου.

1.4 Διάρθρωση

Η παρούσα διπλωματική εργασία χωρίζεται σε οκτώ κεφάλαια, το καθένα από τα οποία περιλαμβάνει συγκεκριμένες πληροφορίες για τη μεθοδολογία και τα αποτελέσματα της έρευνας.

Κεφάλαιο 1: Εισαγωγή

Στο κεφάλαιο αυτό παρουσιάζεται το ερευνητικό πρόβλημα, η σημασία του θέματος και ο σκοπός της εργασίας.

Κεφάλαιο 2: Το ελαιόλαδο

Αυτό το κεφάλαιο εστιάζει στη χημική σύσταση του ελαιόλαδου και τους παράγοντες που επηρεάζουν την ποιότητά του, όπως το μικρόκλιμα. Επιπλέον, εξετάζονται ζητήματα νοθείας και αναλύεται η σημαντικότητα της αυθεντικότητας του ελαιόλαδου στην αγορά.

Κεφάλαιο 3: Φασματοσκοπία

Παρουσιάζονται οι βασικές αρχές της φασματοσκοπίας, με έμφαση στη χρήση της μεθόδου FTIR για την ανάλυση του ελαιόλαδου.

Κεφάλαιο 4: Επεξεργασία δεδομένων FTIR

Στο κεφάλαιο αυτό περιγράφονται οι διαδικασίες προεπεξεργασίας των δεδομένων φασματοσκοπίας, όπως ο καθαρισμός δεδομένων με τη Ανάλυση Κύριων Συνιστωσών (PCA) και με τη χρήση της τεχνικής binning, για τη μείωση της διάστασης των δεδομένων.

Κεφάλαιο 5: Μηχανική Μάθηση

Εδώ παρουσιάζονται οι μέθοδοι μηχανικής μάθησης που εφαρμόστηκαν για την ταξινόμηση και ομαδοποίηση των δεδομένων. Αναλύονται οι αλγόριθμοι ταξινόμησης και ομαδοποίησης, καθώς και οι μέθοδοι επικύρωσης που χρησιμοποιήθηκαν.

Κεφάλαιο 6: Υλοποίηση

Το κεφάλαιο αυτό περιγράφει την εφαρμογή των μεθόδων που αναλύθηκαν προηγουμένως. Συγκεκριμένα, παρουσιάζεται η συλλογή και επεξεργασία των δεδομένων και αναλύονται τα αποτελέσματα από την εφαρμογή των αλγορίθμων ταξινόμησης και ομαδοποίησης.

Κεφάλαιο 7: Συζήτηση

Περιγράφονται τα βασικά συμπεράσματα που προκύπτουν από την έρευνα, ενώ γίνεται αναφορά στις μελλοντικές προοπτικές και πιθανές επεκτάσεις της έρευνας.

Κεφάλαιο 8: Βιβλιογραφία

Περιλαμβάνεται η πλήρης λίστα των πηγών που χρησιμοποιήθηκαν στην εργασία.

2. Το ελαιόλαδο

2.1 Η χημική σύσταση του ελαιόλαδου

Το ελαιόλαδο προέρχεται φυσικά από τον καρπό της ελιάς και είναι μια σύνθετη χημική μήτρα. Το σαπωνοποιήσιμο κλάσμα (saponifiable fraction) αποτελείται κυρίως από τριακυλογλυκερόλες, διακυλογλυκερόλες, μονοακυλογλυκερόλες, ελεύθερα λιπαρά οξέα και φωσφολιπίδια, τα οποία αντιπροσωπεύουν σχεδόν το 98% της σύνθεσης του ελαιόλαδου. Το ασαπωνοποιήσιμο κλάσμα (unsaponifiable fraction), το οποίο δεν σχηματίζεται από λιπαρές ενώσεις, αποτελεί μόνο το 2% του βάρους του ελαιόλαδου. Το μη πολικό κλάσμα περιλαμβάνει υδρογονάνθρακες, σκουαλένιο, στερόλες, τοκοφερόλες, πτητικές ενώσεις, καροτενοειδή και χρωστικές ουσίες. Αντίστοιχα, το πολικό κλάσμα περιέχει πολλές δευτερεύουσες ενώσεις, όπως πολυφαινόλες και απλές φαινόλες. Αυτά τα μόρια, γνωστά ως τριγλυκερίδια, αποτελούνται από τρία λιπαρά οξέα που συνδέονται με τη γλυκερίνη. Η σύνδεση αυτή είναι ευαίσθητη και μπορεί να διασπαστεί εύκολα, απελευθερώνοντας ελεύθερα λιπαρά οξέα, κάτι που επιταχύνει τη σταδιακή αποικοδόμηση και αλλοίωση της ποιότητας του ελαίου (Μπόσκου, 2015). Η ποσότητα των ελεύθερων λιπαρών οξέων που δεν είναι πλέον συνδεδεμένα με τα τριγλυκερίδια εκφράζεται ως ελεύθερη οξύτητα, η οποία μετρά το ποσοστό του ελαϊκού οξέος (το κύριο λιπαρό οξύ στο ελαιόλαδο). Η οξύτητα αποτελεί έναν από τους πιο παλαιούς και κοινούς δείκτες ποιότητας του ελαιόλαδου και εξαρτάται σε μεγάλο βαθμό από την ποιότητα και τη φρεσκάδα των ελιών που χρησιμοποιούνται (Gazeli et al., 2020).

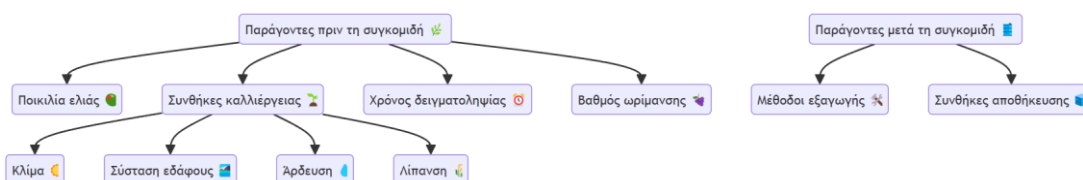
Τα δευτερεύοντα συστατικά συμβάλλουν παράλληλα στις σημαντικές βιολογικές ιδιότητες που κατέχει το ελαιόλαδο. Ιδιαίτερα, οι φαινολικές ενώσεις έχουν αποδοθεί με σημαντικά οφέλη για την υγεία, όπως αναφέρεται στον Κανονισμό (ΕΕ) 432/2012 της Επιτροπής, ο οποίος θέσπισε ισχυρισμό υγείας για τις πολυφαινόλες του ελαιόλαδου, οι οποίες συμβάλλουν στην προστασία των λιπιδίων του αίματος από το οξειδωτικό στρες (Kalogiouri et al., 2020).

2.2 Παράγοντες που επηρεάζουν την ποιότητα του ελαιόλαδου

Η παραγωγή ελαιόλαδου είναι μια πολύπλοκη διαδικασία που απαιτεί προσεκτικό σχεδιασμό και εκτέλεση για να διασφαλιστεί η υψηλή ποιότητα του τελικού προϊόντος. Η διαδικασία ξεκινά με τη συλλογή των ελιών, η οποία πρέπει να γίνεται στην

κατάλληλη στιγμή για να διατηρηθεί η ποιότητα του ελαιόλαδου. Στη συνέχεια, οι ελιές καθαρίζονται από φύλλα και κλαδιά και πλένονται για την απομάκρυνση των υπολειμμάτων. Ακολουθεί η σύνθλιψη των ελιών για την απελευθέρωση του φυτικού ελαίου που περιέχουν. Η εξαγωγή του ελαίου μπορεί να γίνει μηχανικά, με πίεση ή κενού, ή χημικά. Το ελαιόλαδο που παράγεται ενδέχεται να χρειάζεται περαιτέρω καθαρισμό για την απομάκρυνση τυχόν ακαθαρσιών, και στη συνέχεια αποθηκεύεται υπό κατάλληλες συνθήκες για να διατηρήσει τη φρεσκάδα και τις θρεπτικές του ιδιότητες.

Σε πολλές περιπτώσεις, το ελαιόλαδο υποβάλλεται σε διαδικασίες τυποποίησης για να πληροί συγκεκριμένα πρότυπα ποιότητας και γεύσης. Η διατήρηση της ποιότητας, της ασφάλειας και των αισθητηριακών χαρακτηριστικών, όπως το χρώμα, η γεύση και το άρωμα, είναι απαραίτητα για την εμπορική αξία του ελαιόλαδου. Αυτά τα χαρακτηριστικά καθορίζονται από τη χημική σύνθεση του ελαίου, η οποία επηρεάζεται από παράγοντες πριν και μετά τη συγκομιδή. Οι μέθοδοι εξαγωγής και οι συνθήκες αποθήκευσης αποτελούν τους σημαντικότερους παράγοντες μετά τη συγκομιδή που επηρεάζουν την ποιότητα του ελαιόλαδου (Ben-Hassine et al., 2013). Ανάμεσα στους παράγοντες πριν τη συγκομιδή, η ποικιλία της ελιάς αποτελεί κύρια πηγή διαφοροποίησης στη σύνθεση και στα αισθητηριακά χαρακτηριστικά (Aragicio & Harwood, 2013). Παράλληλα, η ποιότητα επηρεάζεται από τις συνθήκες καλλιέργειας, όπως οι περιβαλλοντικοί παράγοντες (κλίμα, σύσταση εδάφους) και οι αγρονομικές πρακτικές (άρδευση, λίπανση), καθώς και από τον χρόνο δειγματοληψίας και τον βαθμό ωρίμανσης των ελιών (Sayago et al., 2018).



Εικόνα 1. Σχηματική αναπαράσταση των παραγόντων πριν και μετά την συγκομιδή που επηρεάζουν την ποιότητα του ελαιόλαδου

2.2.1 Το μικροκλίμα

Πιο αναλυτικά, τα ποιοτικά χαρακτηριστικά του ελαιόλαδου που σχετίζονται με παράγοντες πριν την συγκομιδή, όπως η οξύτητα και η περιεκτικότητα σε πολυφαινόλες, ενδέχεται να επηρεάζονται σημαντικά από το μικροκλίμα της κάθε περιοχής.

Το μικροκλίμα αναφέρεται στο κλίμα μιας μικρής και συγκεκριμένης περιοχής που μπορεί να διαφέρει σημαντικά από το γενικό κλίμα μιας ευρύτερης περιοχής. Αυτή η διακύμανση μπορεί να συμβεί σε σχετικά μικρές αποστάσεις και επηρεάζεται από παράγοντες όπως το υψόμετρο, ο προσανατολισμός, η βλάστηση, τα υδάτινα σώματα και οι ανθρωπογενείς δομές. Αυτά τα στοιχεία μπορούν να αλλάξουν τη θερμοκρασία, την υγρασία, τα μοτίβα του ανέμου και τη βροχόπτωση στην περιοχή, δημιουργώντας ένα μοναδικό τοπικό κλίμα.

Η κατανόηση του μικροκλίματος είναι απαραίτητη για δραστηριότητες όπως η γεωργία, καθώς η ανάπτυξη των φυτών μπορεί να επηρεαστεί σημαντικά από τα τοπικά καιρικά πρότυπα και τις συνθήκες διαβίωσης. Συγκεκριμένα, η ίδια ποικιλία ελιάς μπορεί να παράγει ελαιόλαδα με διαφορετικά χαρακτηριστικά λόγω των διακυμάνσεων στο μικροκλίμα. Παρακάτω παρατίθενται όλοι οι τρόποι με τους οποίους το μικροκλίμα επιδρά στην ποιότητα του παραγόμενου ελαιόλαδου:

α) Θερμοκρασιακές διακυμάνσεις: Οι υψηλότερες θερμοκρασίες γενικά επιταχύνουν τη διαδικασία ωρίμανσης του καρπού, η οποία μπορεί να αυξήσει την περιεκτικότητα του ελαίου, αλλά μπορεί να μειώσει ορισμένες αρωματικές ενώσεις. Αντίθετα, οι ψυχρότερες συνθήκες τείνουν να επιβραδύνουν την ωρίμανση, οδηγώντας ενδεχομένως σε ένα πιο πυκνό, πιο παχύρρευστο έλαιο με πιο πλούσιο γευστικό προφίλ.

β) Έκθεση στο ηλιακό φως: Οι ελιές που δέχονται περισσότερο ηλιακό φως κατά την ανάπτυξή τους έχουν υψηλότερα επίπεδα πολυφαινόλων, οι οποίες όχι μόνο συμβάλλουν σε μια πιο στιβαρή και πολύπλοκη γεύση αλλά ενισχύουν επίσης την οξειδωτική σταθερότητα του ελαίου και τα οφέλη για την υγεία. Επιπλέον, η αυξημένη ηλιακή ακτινοβολία βελτιώνει την αποτελεσματικότητα της φωτοσύνθεσης, η οποία μπορεί να επηρεάσει θετικά τη θρεπτική ποιότητα του ελαίου. Ως αποτέλεσμα, τα ελαιόλαδα από πιο ηλιόλουστες περιοχές ή ελαιώνες μπορούν να παρουσιάζουν

ξεχωριστές γεύσεις και ανώτερη ποιότητα σε σύγκριση με εκείνα που καλλιεργούνται σε λιγότερο ηλιόλουστες συνθήκες.

γ) **Επίπεδα βροχόπτωσης και υγρασίας:** Οι υπερβολικές βροχοπτώσεις και η υψηλή υγρασία κοντά στην εποχή της συγκομιδής μπορεί να αυξήσουν τον κίνδυνο μυκητιακών ασθενειών, αυξάνοντας ενδεχομένως την οξύτητα του ελαίου και αλλάζοντας τη γεύση του. Αντίθετα, οι μέτριες βροχοπτώσεις συμβάλλουν στη διατήρηση μιας υγιούς ισορροπίας ανάπτυξης, ενισχύοντας τη γεύση και την ποιότητα του ελαίου. Οι περιοχές με χαμηλότερη υγρασία και ελεγχόμενη διαθεσιμότητα νερού βλέπουν συχνά ελιές με συμπυκνωμένη γεύση λόγω των καταπονημένων συνθηκών, που οδηγεί σε έλαιο με πιο πλούσια αρωματικά και φαινολικά προφίλ.

δ) **Άνεμος:** Σε περιοχές με ανέμους, η ροή του αέρα μπορεί να βοηθήσει στη μείωση της υγρασίας στους καρπούς και τα φύλλα, μειώνοντας την πιθανότητα μυκητιασικών ασθενειών και καταλήγοντας σε πιο υγιεινές ελιές και λάδι υψηλότερης ποιότητας. Επιπλέον, ο άνεμος δρα ως φυσικός αποτρεπτικός παράγοντας για τα παράσιτα, μειώνοντας την ανάγκη για χημικές επεξεργασίες και προάγοντας μια πιο βιολογική παραγωγή λαδιού. Ωστόσο, οι ισχυροί άνεμοι μπορούν επίσης να προκαλέσουν σωματικές βλάβες στα δέντρα και τις ελιές, μειώνοντας ενδεχομένως τη συνολική απόδοση και επηρεάζοντας την ποιότητα του ελαιόλαδου.

ε) **Τύπος του εδάφους:** Διαφορετικά εδάφη προσφέρουν διαφορετικά επίπεδα θρεπτικών συστατικών, τα οποία επηρεάζουν άμεσα την ανάπτυξη και την υγεία των ελαιόδεντρων. Τα καλά στραγγιζόμενα εδάφη, συνήθως αμμώδη ή πετρώδη, είναι ιδανικά για ελαιόδεντρα, καθώς προάγουν την υγιή ανάπτυξη των ριζών και αποτρέπουν την υπερχειλίση, η οποία μπορεί να οδηγήσει σε ασθένειες των ριζών. Αυτές οι συνθήκες επιτρέπουν τη βέλτιστη πρόσληψη θρεπτικών συστατικών, ζωτικής σημασίας για την παραγωγή ελιών υψηλής ποιότητας και, κατά συνέπεια, ανώτερου ελαιόλαδου. Επιπλέον, η σύνθεση ορυκτών του εδάφους μπορεί να επηρεάσει τις λεπτές γευστικές αποχρώσεις του λαδιού, δίνοντας στο λάδι κάθε περιοχής τον μοναδικό του χαρακτήρα. Έτσι, το είδος του εδάφους δεν επηρεάζει μόνο την υγεία και την παραγωγικότητα των ελαιόδεντρων αλλά και τη γεύση και την ποιότητα του παραγόμενου ελαιόλαδου.

στ) **Υψόμετρο:** Τα μεγαλύτερα υψόμετρα χαρακτηρίζονται συνήθως από χαμηλότερες θερμοκρασίες, οι οποίες μπορούν να επιβραδύνουν τη διαδικασία ωρίμανσης των ελιών, επιτρέποντας μεγαλύτερη περίοδο ανάπτυξης γεύσης. Αυτό συχνά οδηγεί σε ελαιόλαδο με πιο περίπλοκο και πιο διαφοροποιημένο γευστικό προφίλ. Επιπλέον, η αυξημένη έκθεση στην υπεριώδη ακτινοβολία σε υψηλότερα υψόμετρα μπορεί να ενισχύσει τη συγκέντρωση φαινολικών ενώσεων στις ελιές, ενισχύοντας τόσο τη γεύση όσο και τις αντιοξειδωτικές ιδιότητες του λαδιού.

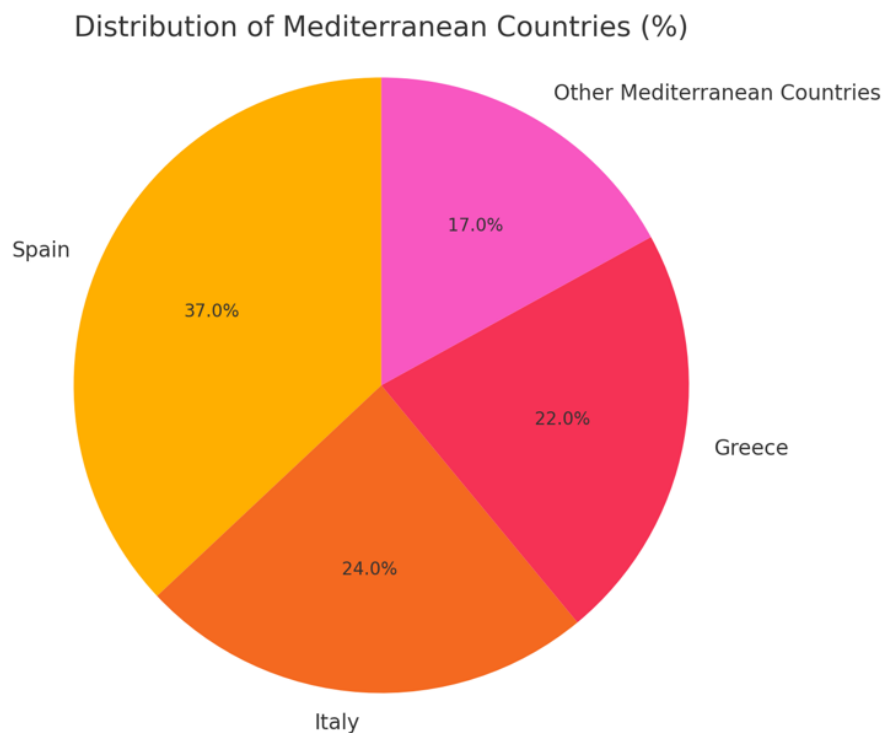
Αυτές οι μικροκλιματικές διαφορές οδηγούν σε διακυμάνσεις στη φαινολική σύνθεση, την οξύτητα, το χρώμα και το άρωμα του ελαίου. Ακόμη και όταν η ποικιλία της ελιάς είναι ίδια, αυτοί οι περιβαλλοντικοί παράγοντες μπορούν να αλλάξουν σημαντικά τις αισθητικές και χημικές ιδιότητες του παραγόμενου ελαιόλαδου. Αυτός είναι ο λόγος για τον οποίο το ελαιόλαδο από διαφορετικά κτήματα ή περιοχές, το καθένα με το δικό του μοναδικό μικροκλίμα, μπορεί να έχει εντυπωσιακά διαφορετική γεύση, παρά το γεγονός ότι είναι φτιαγμένο από τον ίδιο τύπο ελιάς. (Orlandi et al., 2020; Rapa & Ciano, 2022)

2.3 Νοθεία

Το ελαιόλαδο αποτελεί θεμελιώδες συστατικό της διατροφής σε όλες τις μεσογειακές χώρες από την προϊστορική εποχή (Bendini et al., 2007). Σήμερα, περίπου το 90% της παγκόσμιας παραγωγής ελαιόλαδου προέρχεται από την περιοχή της Μεσογείου, με τις ευρωπαϊκές χώρες να αντιπροσωπεύουν περίπου το 82% της παγκόσμιας παραγωγής. Η Ισπανία αποτελεί τον μεγαλύτερο παραγωγό, με περίπου το 37% της συνολικής παραγωγής παγκοσμίως, ακολουθούμενη από την Ιταλία με 24% και την Ελλάδα με 22%, (Gazeli et al., 2020)

Πρόκειται για προϊόν υψηλής αξίας, το οποίο επιτρέπεται να κυκλοφορεί με Προστατευόμενη Ονομασία Προέλευσης (ΠΟΠ) και Προστατευόμενη Γεωγραφική Ένδειξη (ΠΓΕ) (ΕΕ, Αρ. 1151/2012). Αυτές οι ετικέτες δόθηκαν για να ενθαρρύνουν τη διαφορετική γεωργική παραγωγή και να προστατεύσουν τα ονόματα των προϊόντων από απομίμηση ή κακή χρήση, καθώς και για να βοηθήσουν τους καταναλωτές να προσδιορίσουν καλύτερα τα ειδικά χαρακτηριστικά του προϊόντος. Τα χαρακτηριστικά συνδέονται άμεσα με την επικράτεια. Ο συνδυασμός περιβαλλοντικών παραγόντων και η ανθρώπινη παρέμβαση κατά την παραγωγή και την επεξεργασία καθιστούν το προϊόν

μοναδικό. Η αυθεντικότητα των τροφίμων περιλαμβάνει την επικύρωση ότι μια δηλωμένη προδιαγραφή ενός τρόφιμου είναι ακριβής. Ο έλεγχος της γνησιότητας των προϊόντων διατροφής μπορεί να αποτρέψει την εσφαλμένη περιγραφή, τη νοθεία ή την ψευδή σήμανση προέλευσης (Kalogiouri et al., 2020). Η επιβολή της νομοθεσίας για την επισήμανση διασφαλίζει ότι τα τρόφιμα περιγράφονται με ακρίβεια, προστατεύοντας τους καταναλωτές από την αγορά προϊόντων κατώτερης ποιότητας και τιμής με εσφαλμένη περιγραφή και προασπίζοντας τους έντιμους εμπόρους από τον αθέμιτο ανταγωνισμό. Έτσι, η ανάπτυξη αναλυτικών μεθόδων που εγγυώνται την αυθεντικότητα των τροφίμων είναι θεμελιώδης για τη λειτουργία της σύγχρονης κοινωνίας στη διεθνή αγορά τροφίμων.



Εικόνα 2 Διαγραμματική απεικόνιση (pie-chart) των χωρών προέλευσης του ελαιόλαδου

Η περιεκτικότητα του ελαιόλαδου σε αντιοξειδωτικά, ωμέγα-3 λιπαρά οξέα και άλλες θρεπτικές ουσίες έχει αποδειχθεί να συμβάλλουν στην πρόληψη ασθενειών, όπως οι καρδιαγγειακές παθήσεις και ορισμένες μορφές καρκίνου.

Η αυξανόμενη ευαισθητοποίηση του κοινού για τα οφέλη της μεσογειακής διατροφής στην υγεία, της οποίας αναπόσπαστο μέρος αποτελεί το ελαιόλαδο, το έχει καταστήσει ένα από τα προϊόντα με τη μεγαλύτερη κατανάλωση παγκοσμίως, γεγονός που επηρεάζει και την τιμή του στην αγορά. Συγκεκριμένα, τα τελευταία είκοσι χρόνια,

χάρη στις προωθητικές προσπάθειες του Διεθνούς Συμβουλίου Ελαιόλαδου (ΔΟΕ) και της Ευρωπαϊκής Ένωσης (ΕΕ), το εμπορικό ενδιαφέρον έχει αυξηθεί κατακόρυφα. Ωστόσο, το συνεχώς αυξανόμενο εμπορικό ενδιαφέρον και οι σημαντικές διακυμάνσεις στις τιμές του ελαιόλαδου οδηγούν σε νοθεία, ιδίως όταν αφορά το έξτρα παρθένο ελαιόλαδο (Gazeli et al., 2020).

3. Φασματοσκοπία

Η φασματοσκοπία, μια ισχυρή αναλυτική τεχνική, παρέχει μια ολοκληρωμένη και λεπτομερή διερεύνηση της σύνθεσης και των ιδιοτήτων της ύλης, εξετάζοντας την αλληλεπίδραση μεταξύ του φωτός και ουσιών. Η φασματοσκοπία μπορεί να μας παρέχει πληροφορίες για τις μοριακές και ατομικές δομές διαφόρων υλικών. Με αυτή τη τεχνική καθίσταται δυνατή η αναγνώριση χημικών ενώσεων, η αξιολόγηση της καθαρότητας τους και η κατανόηση της σύνθετης μοριακής δυναμικής. Μέσω των ποικίλων εφαρμογών της, συμπεριλαμβανομένης της παρακολούθησης του περιβάλλοντος, της φαρμακευτικής ανάπτυξης και των αστρονομικών μελετών, προωθεί τις εξελίξεις σε πολλούς κλάδους, συμβάλλοντας στην επιστημονική πρόοδο και ενισχύοντας την κατανόησή μας για τον φυσικό κόσμο (Prakash et al., 2021).

Η φασματοσκοπία επεξεργάζεται δεδομένα και παράγει αποτελέσματα μέσω μιας σειράς συστηματικών βημάτων:

1. Προετοιμασία δείγματος: Το προς ανάλυση δείγμα παρασκευάζεται σε κατάλληλη μορφή, είτε είναι αέριο, υγρό ή στερεό. Η σωστή προετοιμασία διασφαλίζει την ακριβή αλληλεπίδραση μεταξύ του δείγματος και της πηγής φωτός.

2. Αλληλεπίδραση φωτός: Το δείγμα εκτίθεται σε συγκεκριμένο εύρος ηλεκτρομαγνητικής ακτινοβολίας (φως). Αυτό μπορεί να είναι με τη μορφή ορατού φωτός, υπεριώδους φωτός, υπέρυθρου φωτός, ακτινών X ή άλλων μηκών κύματος του ηλεκτρομαγνητικού φάσματος.

3. Απορρόφηση, εκπομπή ή σκέδαση: Καθώς το φως αλληλοεπιδρά με το δείγμα, τα μόρια ή τα άτομα μέσα στο δείγμα απορροφούν, εκπέμπουν ή σκεδάζουν το φως σε συγκεκριμένα μήκη κύματος. Αυτή η αλληλεπίδραση εξαρτάται από τα ενεργειακά επίπεδα των μορίων ή των ατόμων.

4. Ανίχνευση: Το φασματοσκοπικό όργανο, εξοπλισμένο με ανιχνευτές όπως φωτοδίοδοι, συσκευές συζευγμένου φορτίου (CCD) ή σωλήνες φωτοπολλαπλασιαστή, μετρά την ένταση του φωτός σε διαφορετικά μήκη κύματος. Τα δεδομένα που προκύπτουν είναι ένα φάσμα, που δείχνει πόσο φως απορροφάται, εκπέμπεται ή διασκορπίζεται σε κάθε μήκος κύματος.

5. Επεξεργασία Δεδομένων: Τα ακατέργαστα φασματικά δεδομένα επεξεργάζονται χρησιμοποιώντας αλγόριθμους λογισμικού. Αυτό περιλαμβάνει

βήματα όπως διόρθωση γραμμής βάσης, μείωση θορύβου και κανονικοποίηση για τη βελτίωση της ακρίβειας και της σαφήνειας του φάσματος.

6. Ανάλυση και Ερμηνεία: Το επεξεργασμένο φάσμα αναλύεται για τον εντοπισμό χαρακτηριστικών κορυφών ή μοτίβων. Αυτές οι κορυφές αντιστοιχούν σε συγκεκριμένες ενεργειακές μεταβάσεις στα μόρια ή τα άτομα του δείγματος, παρέχοντας πληροφορίες σχετικά με τη σύνθεση, τη δομή και τη συγκέντρωση του δείγματος.

7. Συγκριτική Ανάλυση: Τα λαμβανόμενα φασματικά δεδομένα συγκρίνονται συχνά με φάσματα αναφοράς από γνωστές ουσίες ή βάσεις δεδομένων. Αυτή η σύγκριση βοηθά στον εντοπισμό των χημικών ενώσεων που υπάρχουν στο δείγμα.

8. Ποσοτικά και Ποιοτικά Αποτελέσματα: Με βάση την ανάλυση, προκύπτουν ποσοτικά δεδομένα (όπως επίπεδα συγκέντρωσης) και ποιοτικά δεδομένα (όπως μοριακή δομή και λειτουργικές ομάδες). Αυτά τα αποτελέσματα ερμηνεύονται για να εξαχθούν σημαντικά συμπεράσματα σχετικά με το δείγμα.

Η φασματοσκοπία μετατρέπει την αλληλεπίδραση μεταξύ φωτός και ύλης σε λεπτομερείς, ενεργές πληροφορίες, επιτρέποντας στους επιστήμονες να κατανοήσουν και να χαρακτηρίσουν με ακρίβεια τις ιδιότητες διαφόρων ουσιών (Dans et al., 2021; Gautam et al., 2015; Mokari et al., 2023).

4. Επεξεργασία δεδομένων FTIR

4.1 Καθαρισμός δεδομένων

Πριν από την ταξινόμηση, τα δεδομένα Φασματοφωτομετρίας Υπερύθρου με Μετασχηματισμό Fourier (Fourier Transform InfraRed, FTIR) απαιτούν κατάλληλη προεπεξεργασία δεδομένων ώστε να καταστούν ερμηνεύσιμα και αξιοποιήσιμα για περαιτέρω ανάλυση. Αρχικά, χρησιμοποιήθηκε η μέθοδος Principal Component Analysis (PCA) χρησιμοποιείται για τη μείωση της διάστασης των δεδομένων, εξάγοντας τις πιο σημαντικές συνιστώσες που περιέχουν το μεγαλύτερο ποσοστό της πληροφορίας. Ακόμα εφαρμόστηκε και η μέθοδος Binning, η οποία οργανώνει τα συνεχόμενα ή διακριτά δεδομένα σε συγκεκριμένα διαστήματα, μειώνοντας τον αριθμό των μεταβλητών και διευκολύνοντας την ανάλυση. Με αυτές τις τεχνικές, τα δεδομένα καθίστανται πιο διαχειρίσιμα, επιτρέποντας τη σαφέστερη ερμηνεία τους και την περαιτέρω ανάλυση. (Lerma-García et al., 2010; Subramanian et al., 2011; Sun, 2009)

4.2 Principal Component Analysis

Η Principal Component Analysis (PCA) είναι μια ευρέως προτιμώμενη τεχνική στη μηχανική μάθηση για πολλούς επιτακτικούς λόγους. Η PCA χρησιμοποιείται για τη μείωση της διάστασης μεγάλων συνόλων δεδομένων, γεγονός που απλοποιεί την πολυπλοκότητα χωρίς να χάνει την ουσία των δεδομένων. Μετατρέποντας τις αρχικές μεταβλητές σε ένα νέο σύνολο ασυσχέτιστων μεταβλητών που ονομάζονται κύρια στοιχεία, η PCA διατηρεί τις πιο κρίσιμες πληροφορίες, ενώ απορρίπτει τον θόρυβο και τον πλεονασμό.

Ένα από τα κύρια πλεονεκτήματα της PCA είναι ότι ενισχύει την υπολογιστική απόδοση. Οι αλγόριθμοι μηχανικής μάθησης συχνά αποδίδουν καλύτερα και ταχύτερα με μειωμένες διαστάσεις εισόδου. Αυτή η μείωση όχι μόνο επιταχύνει τη διαδικασία εκπαίδευσης, αλλά βοηθά επίσης στον μετριασμό του κινδύνου υπερβολικής προσαρμογής αφαιρώντας λιγότερο σημαντικά χαρακτηριστικά που διαφορετικά θα μπορούσαν να οδηγήσουν σε πολυπλοκότητα του μοντέλου χωρίς να συμβάλλουν στην απόδοση.

Επιπλέον, η PCA διευκολύνει την καλύτερη οπτικοποίηση των δεδομένων. Τα δεδομένα υψηλών διαστάσεων μπορεί να είναι δύσκολο να ερμηνευτούν και να απεικονιστούν, αλλά προβάλλοντας αυτά τα δεδομένα σε κύρια στοιχεία, γίνεται

ευκολότερο να εξερευνηθούν και να κατανοηθούν τα υποκείμενα μοτίβα και οι σχέσεις.

Συνεπώς, η PCA ευνοείται στη μηχανική μάθηση επειδή βελτιστοποιεί την επεξεργασία δεδομένων, βελτιώνει την απόδοση του αλγορίθμου, μειώνει τον κίνδυνο υπερπροσαρμογής και ενισχύει την ερμηνευτικότητα των δεδομένων, καθιστώντας το ένα ισχυρό εργαλείο για αποτελεσματική ανάλυση δεδομένων και ανάπτυξη μοντέλων (Shlens, n.d.; Subramanian et al., 2011; Sun, 2009; Tang & Allen, 2021).

4.3 Binning

Το Binning, η διαδικασία ομαδοποίησης συνεχών ή διακριτών δεδομένων σε διαστήματα ή "bins", είναι μια ευρέως χρησιμοποιούμενη τεχνική προεπεξεργασίας δεδομένων στη μηχανική εκμάθηση. Αυτός ο μετασχηματισμός όχι μόνο απλοποιεί την πολυπλοκότητα των δεδομένων, αλλά παρέχει επίσης πολλαπλά οφέλη που συμβάλλουν στην σωστή λειτουργία των μοντέλων μηχανικής μάθησης. Ένα από τα πιο σημαντικά πλεονεκτήματα του binning έγκειται στην ικανότητά του να βελτιώνει την απόδοση συγκεκριμένων μοντέλων μηχανικής μάθησης. Ορισμένοι αλγόριθμοι, όπως τα δέντρα απόφασης (Decision trees), και τα τυχαία δάση (Random Forests) χειρίζονται τις κατηγορικές μεταβλητές πιο αποτελεσματικά από τις συνεχείς. Αυτοί οι αλγόριθμοι συχνά δυσκολεύονται να καθορίσουν τα βέλτιστα σημεία διαχωρισμού για συνεχείς μεταβλητές, οδηγώντας σε μη βέλτιστες δομές δέντρων. Το Binning βοηθά στην αντιμετώπιση αυτού του ζητήματος μετατρέποντας συνεχή δεδομένα σε ένα σύνολο προκαθορισμένων κατηγοριών, επιτρέποντας σε αυτά τα μοντέλα να επικεντρωθούν σε πιο ουσιαστικές διαιρέσεις εντός αυτών των διαστημάτων. Αυτή η διαδικασία μπορεί να μειώσει την υπερπροσαρμογή και να βελτιώσει τη γενίκευση, οδηγώντας σε πιο ακριβείς προβλέψεις. Οι ακραίες τιμές είναι μια κοινή πρόκληση στη μηχανική μάθηση, συχνά παραμορφώνουν τις προβλέψεις μοντέλων και οδηγούν σε μειωμένη απόδοση. Οι συνεχείς μεταβλητές, ειδικότερα, είναι ευαίσθητες σε ακραίες τιμές, οι οποίες μπορούν να παραμορφώσουν τις σχέσεις μεταξύ των χαρακτηριστικών και της μεταβλητής στόχου. Η μέθοδος αυτή προσφέρει μια αποτελεσματική λύση ομαδοποιώντας τις ακραίες τιμές στην ίδια κατηγορία με τις πιο τυπικές παρατηρήσεις, μειώνοντας έτσι την επιρροή τους στο μοντέλο. Οι ακραίες τιμές που εμπίπτουν στον ίδιο κάδο με τις κανονικές τιμές "εξάγονται κατά μέσο όρο", επιτρέποντας στο μοντέλο να επικεντρωθεί στις ευρύτερες τάσεις εντός των δεδομένων αντί να παραπλανηθεί από

άτυπες παρατηρήσεις. Αυτή η προσέγγιση κάνει το binning ιδιαίτερα χρήσιμο, όπου ο στόχος είναι να ελαχιστοποιηθεί η επίδραση του θορύβου ή των ανωμαλιών στα δεδομένα. Επιπλέον, το binning μπορεί να βελτιώσει την απόδοση αλγορίθμων που περιλαμβάνουν υπολογισμούς με βάση την απόσταση, όπως K-means. Σε αυτές τις περιπτώσεις, η μείωση του αριθμού των μοναδικών τιμών απλοποιεί τους υπολογισμούς απόστασης, οδηγώντας σε ταχύτερους χρόνους εκπαίδευσης και πρόβλεψης μοντέλων. Η ικανότητά του να εξομαλύνει δεδομένα, να μειώνει το θόρυβο και να βελτιώνει την απόδοση του μοντέλου το καθιστά μια προτιμώμενη επιλογή για επαγγελματίες που στοχεύουν στη δημιουργία ερμηνεύσιμων, αξιόπιστων και υψηλής απόδοσης μοντέλων μηχανικής εκμάθησης (Binning as a pretext task: improving self-supervised learning in tabular domains, n.d.; Dehuri et al., 2022).

5. Μηχανική μάθηση

Κατά την πάροδο των χρόνων, η δημιουργικότητα του ανθρώπινου εγκεφάλου οδήγησε στην εφεύρεση διαφόρων ειδών εργαλείων, τα οποία έκαναν την ανθρώπινη ζωή πιο εύκολη. Η μηχανική μάθηση (machine learning), είναι ένα από αυτά και βασίζεται σε διαφορετικούς αλγόριθμους για την επίλυση προβλημάτων με δεδομένα. Το είδος του αλγορίθμου που χρησιμοποιείται, για την επίλυση του προβλήματος, εξαρτάται από το είδος του προβλήματος, τον αριθμό των μεταβλητών, το είδος του μοντέλου που το περιγράφει καλύτερα και ούτω καθεξής (Mahesh, 2020).

Η μηχανική μάθηση υπερέχει σε αποτελεσματικότητα σε κλίμακα αυτοματοποιώντας την επεξεργασία και την ανάλυση τεράστιων συνόλων δεδομένων. Οι αλγόριθμοι μηχανικής μάθησης έχουν σχεδιαστεί για να βελτιώνονται αυτόνομα μέσω της εμπειρίας, δίνοντάς τους τη δυνατότητα να χειρίζονται σύνθετες εργασίες πιο γρήγορα και με ακρίβεια, καθώς εκτίθενται σε περισσότερα δεδομένα. Αυτή η δυνατότητα είναι ιδιαίτερα ωφέλιμη σε περιβάλλοντα με μεγάλο όγκο πληροφοριών, όπως οι χρηματοπιστωτικές αγορές ή η ανάλυση μεγάλων δεδομένων, όπου η ταχύτητα και η ακρίβεια είναι κρίσιμες. Κατά συνέπεια, η μηχανική μάθηση, όχι μόνο ενισχύει τη λειτουργική απόδοση, αλλά οδηγεί επίσης σε σημαντικές προόδους σε τομείς όπου η επεκτασιμότητα είναι ζωτικής σημασίας.

Η μηχανική μάθηση εκτιμάται ιδιαίτερα για την προγνωστική της δύναμη, η οποία της επιτρέπει να προβλέπει αποτελέσματα με βάση δεδομένα που έχει ξανά χρησιμοποιήσει. Αυτή η δυνατότητα είναι απαραίτητη για εφαρμογές που απαιτούν ακριβείς προβλέψεις, όπως εκτίμηση κινδύνου, πρόβλεψη ζήτησης και εξατομικευμένες συστάσεις. Αναλύοντας μοτίβα από δεδομένα του παρελθόντος, τα μοντέλα μηχανικής μάθησης μπορούν να κάνουν τεκμηριωμένες προβλέψεις που βοηθούν τις επιχειρήσεις και τους οργανισμούς, να λαμβάνουν προληπτικές αποφάσεις που βασίζονται σε δεδομένα.

Η μηχανική μάθηση ενισχύει σημαντικά την αυτοματοποίηση επιτρέποντας στα συστήματα να λαμβάνουν αποφάσεις και να εκτελούν εργασίες με ελάχιστη ανθρώπινη παρέμβαση. Αυτή η ικανότητα μετασχηματίζει σε διάφορους κλάδους, βελτιώνοντας τις λειτουργίες και μειώνοντας το κόστος. Για παράδειγμα, στην κατασκευή, οι αλγόριθμοι μηχανικής μάθησης μπορούν να προβλέψουν τις βλάβες του εξοπλισμού και να προγραμματίσουν τη συντήρηση, ελαχιστοποιώντας έτσι το χρόνο διακοπής της

λειτουργίας. Στην εξυπηρέτηση πελατών, τα εναλλακτικά μέσα εξυπηρέτησης (chatbots) που τροφοδοτούνται από μηχανική μάθηση μπορούν να χειριστούν αποτελεσματικά τα ερωτήματα ρουτίνας, επιτρέποντας στους ανθρώπινους παράγοντες να επικεντρωθούν σε πιο περίπλοκα ζητήματα. Ο αυτοματισμός που καθοδηγείται από τη μηχανική μάθηση, όχι μόνο ενισχύει την παραγωγικότητα, αλλά εξασφαλίζει επίσης σταθερή ποιότητα και λειτουργική αξιοπιστία.

Η μηχανική μάθηση είναι ικανή να χειρίζεται προβλήματα που είναι πολύ περίπλοκα για τις παραδοσιακές προσεγγίσεις. Εντοπίζοντας μοτίβα και σχέσεις σε τεράστια και πολύπλευρα σύνολα δεδομένων, οι αλγόριθμοι μηχανικής μάθησης μπορούν να αποκρυπτογραφήσουν πολυπλοκότητες που ο άνθρωπος δεν μπορούσε. Η ικανότητα της μηχανικής μάθησης να διαχειρίζεται και να κατανοεί αυτήν την πολυπλοκότητα όχι μόνο οδηγεί σε βαθύτερες γνώσεις και πιο ακριβή μοντέλα, αλλά επιτρέπει επίσης καινοτόμες λύσεις σε μερικά από τα πιο προκλητικά προβλήματα που αντιμετωπίζουν οι βιομηχανίες σήμερα. Υπερέχει στην προώθηση της συνεχούς βελτίωσης καθώς τα μοντέλα της μαθαίνουν και προσαρμόζονται δυναμικά από νέα δεδομένα. Αυτή η συνεχής διαδικασία μάθησης επιτρέπει σε αυτά τα συστήματα να βελτιώσουν τους αλγόριθμους τους και να βελτιώσουν την απόδοσή τους με την πάροδο του χρόνου, διασφαλίζοντας ότι παραμένουν αποτελεσματικά ακόμη και όταν αλλάζουν οι συνθήκες. Αυτή η δυνατότητα είναι ιδιαίτερα πολύτιμη όπου τα μοτίβα μπορούν να αλλάζουν συχνά.

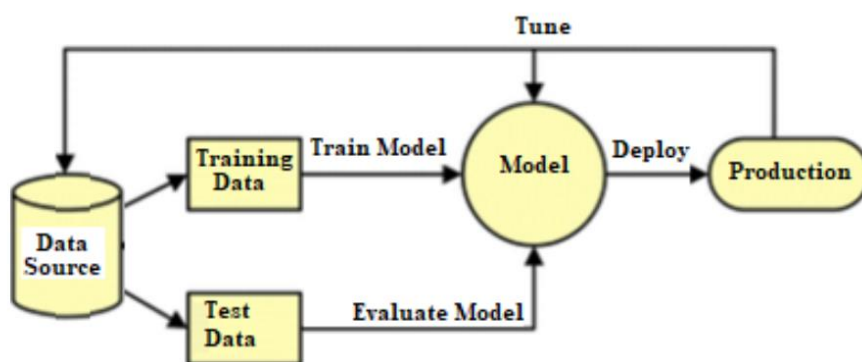
Ουσιαστικά, τα δυνατά σημεία της μηχανικής μάθησης έγκεινται στην επεκτασιμότητα, την προσαρμοστικότητα και την αποτελεσματικότητά της, γεγονός που την καθιστά απαραίτητο εργαλείο στο σύγχρονο τεχνολογικό τοπίο όπου τα δεδομένα είναι άφθονα και η ζήτηση για ταχύτερες και ακριβέστερες διαδικασίες αυξάνεται συνεχώς (Bishop, 2009).

Εποπτευόμενη μάθηση

Η εποπτευόμενη μάθηση (Supervised Learning) είναι ένας υπότυπος της μηχανικής μάθησης, όπου τα δεδομένα που χρησιμοποιούνται για την εκπαίδευση του μοντέλου είναι επισημασμένα. Χρησιμοποιείται στην ανάπτυξη προγνωστικών μοντέλων, καθώς αξιοποιεί επισημασμένα σύνολα δεδομένων για την εκπαίδευση αλγορίθμων, επιτρέποντάς τους έτσι να κάνουν ακριβείς προβλέψεις ή ταξινομήσεις. Μέσω της διαδικασίας τροφοδοσίας του αλγορίθμου με ζεύγη εισόδου-εξόδου, το μοντέλο

μαθαίνει να αντιστοιχίζει τις εισόδους στις επιθυμητές εξόδους, συλλαμβάνοντας αποτελεσματικά τα μοτίβα και τις σχέσεις που έχουν μεταξύ τους τα δεδομένα.

Η διαδικασία της εποπτευόμενης μάθησης περιλαμβάνει πολλά βασικά βήματα. Συνοπτικά, συλλέγεται ένα μεγάλο σύνολο δεδομένων με κάθε κατηγορία να διαθέτει μία ετικέτα, διασφαλίζοντας ότι αντιπροσωπεύει με ακρίβεια τον τομέα του προβλήματος. Στη συνέχεια, το σύνολο δεδομένων χωρίζεται σε υποσύνολα εκπαίδευσης (train) και επικύρωσης (test). Κατά τη διάρκεια της εκπαίδευσης, ο αλγόριθμος προσαρμόζει επαναληπτικά τις εσωτερικές του παραμέτρους για να ελαχιστοποιήσει τη διαφορά μεταξύ των προβλέψεών του και των πραγματικών αποτελεσμάτων. Αυτό επιτυγχάνεται μέσω τεχνικών όπως η κατάβαση κλίσης (Gradient Descent) και η οπισθοδιάδοση (Backpropagation). (Huynh, 2023; Sarker, 2021; Verdonck et al., 2024; Xiaoming et al., n.d.).



Εικόνα 3 = Διαγραμματική απεικόνιση της διαδικασίας εκπαίδευσης μοντέλου μηχανικής μάθησης (Mahesh, 2020)

5.1 Ταξινόμηση

Η ταξινόμηση (Classification), αποτελεί μία από τις βασικές τεχνικές στην επιβλεπόμενη μάθηση, όπου ο αλγόριθμος μαθαίνει από ένα σετ επισημασμένων δεδομένων για να μπορέσει να προβλέψει την κατηγορία ή την κλάση νέων δεδομένων και έτσι να τα ταξινομήσει σε προκαθορισμένες κατηγορίες με βάση τα χαρακτηριστικά τους. Στο πλαίσιο της μηχανικής μάθησης, η ταξινόμηση απαντά στο ερώτημα «σε ποια κλάση ανήκει αυτό το δείγμα;» και εφαρμόζεται όταν οι εξοδοί (outputs) έχουν πεπερασμένο και διακριτό αριθμό τιμών. Η επεξεργασία των δεδομένων για να γίνουν κατάλληλα για εκπαίδευση, θα περιλαμβάνει τον καθαρισμό, τη μετατροπή και το διαχωρισμό των δεδομένων σε σετ εκπαίδευσης και δοκιμής. Το μοντέλο μαθαίνει από το σετ δεδομένων εκπαίδευσης (train data) όπου χρησιμοποιείται

έναν αλγόριθμο ταξινόμησης (όπως η λογιστική παλινδρόμηση, τα δέντρα αποφάσεων ή τα νευρωνικά δίκτυα) για να δημιουργήσει ένα μοντέλο που προβλέπει τις κλάσεις των νέων δεδομένων. Το μοντέλο εκτιμάται στο σετ δεδομένων επικύρωσης (test data) για να δει πόσο καλά λειτουργεί σε δεδομένα που δεν έχει δει προηγουμένως. Οι προβλέψεις του μοντέλου ελέγχονται συγκριτικά με τις πραγματικές κλάσεις των δεδομένων δοκιμής. Μετά την επιτυχή εκτίμηση, το μοντέλο μπορεί να εφαρμοστεί σε νέα δεδομένα για να προβλέψει τις κλάσεις τους με εμπιστοσύνη (Domingos, 2012; Kathole et al., 2019; Wagner et al., 2003).

5.2 Μέθοδοι επικύρωσης

Η αξιολόγηση των μοντέλων μηχανικής μάθησης είναι ζωτικής σημασίας για την αποφυγή υπερπροσαρμογής και για τη βελτιστοποίηση της απόδοσής τους σε νέα δεδομένα. Δύο βασικές μέθοδοι επικύρωσης (validation) που χρησιμοποιούνται είναι η μέθοδος διαχωρισμού (split) και η διασταυρούμενη επικύρωση (cross-validation), καθεμία με τα δικά της πλεονεκτήματα και εφαρμογές.

5.2.1 Μέθοδος διαχωρισμού

Στη μηχανική μάθηση και στην ανάπτυξη μοντέλων, η μέθοδος διαχωρισμού (split) είναι μια κρίσιμη διαδικασία που χρησιμοποιείται για τη διαίρεση ενός συνόλου δεδομένων σε ξεχωριστά υποσύνολα. Διαιρώντας το σύνολο δεδομένων σε διακριτά υποσύνολα, συνήθως σε σετ εκπαίδευσης (train) και σετ δοκιμής (test), το μοντέλο μπορεί να μάθει και να εκπαιδευτεί από την κατηγορία δεδομένων train ενώ η απόδοσή του αξιολογείται σε ένα ξεχωριστό, αόρατο τμήμα δεδομένων test. Αυτή η διαδικασία βοηθά στην αποφυγή της υπερβολικής προσαρμογής, όπου ένα μοντέλο μπορεί να έχει τέλεια απόδοση σε δεδομένα εκπαίδευσης αλλά να μην μπορεί να προβλέψει με ακρίβεια σε νέα, αόρατα δεδομένα. Η αξιολόγηση του μοντέλου σε ένα δοκιμαστικό σύνολο επιτρέπει στους προγραμματιστές να μετρήσουν την ικανότητά του να κάνει προβλέψεις σε σενάρια πραγματικού κόσμου, διασφαλίζοντας ότι το μοντέλο δεν απομνημονεύει απλώς τα δεδομένα εκπαίδευσης αλλά μαθαίνει πραγματικά τα μοτίβα των δεδομένων. Επιπλέον, μια στρατηγική διαχωρισμού, όπως η διαίρεση 80-20 ή 70-30 για εκπαίδευση και επικύρωση αντίστοιχα, παρέχει μια ισορροπία μεταξύ επαρκών δεδομένων για εκμάθηση και επαρκών δεδομένων για τη δοκιμή της απόδοσης,

οδηγώντας σε πιο ισχυρά και αξιόπιστα μοντέλα μηχανικής εκμάθησης. (Tan et al., 2021)

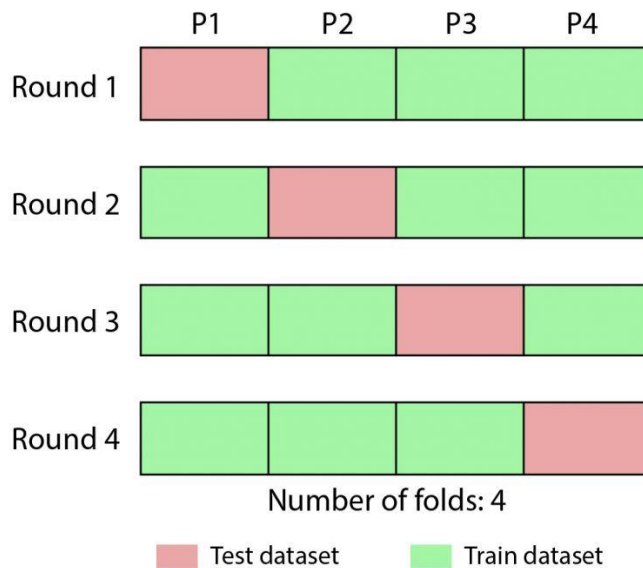
5.2.2 Μέθοδος διασταυρούμενης επικύρωσης

Η διασταυρούμενη επικύρωση (cross validation) είναι μια πιο εκτεταμένη μέθοδος επικύρωσης που προσφέρει μια πιο ακριβή και ισχυρή εκτίμηση της απόδοσης ενός μοντέλου. Σε αντίθεση με τη μέθοδο διαχωρισμού, όπου το σύνολο δεδομένων χωρίζεται μια φορά, η διασταυρούμενη επικύρωση χωρίζει τα δεδομένα σε πολλαπλά υποσύνολα (συνήθως K τμήματα) και η διαδικασία εκπαίδευσης και δοκιμής επαναλαμβάνεται πολλές φορές.

Για παράδειγμα, σε μια διασταυρούμενη επικύρωση 5 πτυχών (5-fold cross-validation), το σύνολο δεδομένων χωρίζεται σε 5 ίσα τμήματα. Σε κάθε επανάληψη, το μοντέλο εκπαιδεύεται σε 4 από τα τμήματα και δοκιμάζεται στο πέμπτο. Αυτή η διαδικασία επαναλαμβάνεται έτσι ώστε κάθε τμήμα να χρησιμοποιηθεί ως σύνολο δοκιμής μία φορά. Στο τέλος, η απόδοση του μοντέλου υπολογίζεται ως ο μέσος όρος των μετρήσεων από όλες τις επαναλήψεις.

Η διασταυρούμενη επικύρωση είναι ιδιαίτερα χρήσιμη όταν τα δεδομένα είναι περιορισμένα, καθώς επιτρέπει τη μέγιστη αξιοποίηση του συνόλου δεδομένων. Κάθε δείγμα συμμετέχει τόσο στην εκπαίδευση όσο και στη δοκιμή, διασφαλίζοντας ότι κανένα τμήμα των δεδομένων δεν αποκλείεται από τη διαδικασία μάθησης. Επίσης, παρέχει μια πιο αξιόπιστη εκτίμηση της απόδοσης του μοντέλου, καθώς μειώνει τον κίνδυνο υπερπροσαρμογής σε συγκεκριμένα δεδομένα εκπαίδευσης.

Επιπλέον, η διασταυρούμενη επικύρωση επιτρέπει την προσαρμογή των παραμέτρων του μοντέλου με μεγαλύτερη ακρίβεια, ενώ προσφέρει πιο σταθερές και αξιόπιστες μετρήσεις απόδοσης, ιδιαίτερα σε σύνολα δεδομένων μικρού ή μεσαίου μεγέθους (Kohavi, n.d.; Tibshirani & Tibshirani, 2009).



Εικόνα 4. Διαδικασία διασταυρωμένης επικύρωσης με τέσσερις επαναλήψεις(cross-validation)
<https://blog.quantinsti.com/cross-validation-machine-learning-trading-models/>

5.3 Αλγόριθμοι ταξινόμησης

Οι αλγόριθμοι ταξινόμησης (classification algorithms) υπερέχουν στον χειρισμό κατηγορικών δεδομένων, κάτι που αποτελεί σημαντικό πλεονέκτημα σε πολλές εφαρμογές μηχανικής εκμάθησης. Αυτοί οι αλγόριθμοι μπορούν να επεξεργάζονται κατηγορικές μεταβλητές απευθείας, εξαλείφοντας την ανάγκη για εκτεταμένα βήματα προ επεξεργασίας, όπως η κωδικοποίησή τους σε αριθμητικές μορφές. Αυτή η άμεση προσέγγιση όχι μόνο απλοποιεί τη διαδικασία προετοιμασίας δεδομένων αλλά διατηρεί επίσης την ακεραιότητα της αρχικής δομής των δεδομένων. Ως αποτέλεσμα, τα μοντέλα ταξινόμησης μπορούν να αξιοποιήσουν πιο αποτελεσματικά τα εγγενή μοτίβα εντός κατηγορικών δεδομένων, οδηγώντας σε πιο ακριβή και αποτελεσματικά αποτελέσματα. Είναι κατάλληλοι για το χειρισμό μεγάλων συνόλων δεδομένων και πολύπλοκων προβλημάτων με πολλά χαρακτηριστικά. Αυτή η επεκτασιμότητα διασφαλίζει ότι τα μοντέλα ταξινόμησης μπορούν να εκπαιδευτούν αποτελεσματικά σε τεράστιες ποσότητες δεδομένων, ένα τυπικό σενάριο σε πολλά σύγχρονα περιβάλλοντα δεδομένων. Με την αποτελεσματική κλιμάκωση, αυτοί οι αλγόριθμοι διατηρούν ισχυρή απόδοση ακόμη και όταν το σύνολο δεδομένων μεγαλώνει, επιτρέποντάς τους να παρέχουν συνεχώς διορατικές και αξιόπιστες προβλέψεις. Αυτό το χαρακτηριστικό είναι ιδιαίτερα χρήσιμο σε τομείς όπως το ηλεκτρονικό εμπόριο, τα μέσα κοινωνικής δικτύωσης, όπου ο όγκος των δεδομένων αυξάνεται συνεχώς, στη δημιουργία

προβλέψεων για νέα, αόρατα δεδομένα, από το φιλτράρισμα ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου έως και την ιατρική διάγνωση (Domingos, 2012; Kathole et al., 2019; Wagner et al., 2003).

5.3.1 Δέντρα αποφάσεων

Τα δέντρα αποφάσεων (decision trees) είναι ένας τύπος αλγορίθμου επιβλεπόμενης μάθησης που χρησιμοποιείται τόσο για καθήκοντα ταξινόμησης όσο και για παλινδρόμηση. Είναι εύκολα στην απεικόνιση, κάνοντάς τα δημοφιλή επιλογή μεταξύ των επιστημονικών δεδομένων για προγνωστικά μοντέλα. Τα δέντρα αποφάσεων λειτουργούν επαναλαμβάνοντας τη διαίρεση των δεδομένων σε μικρότερα υποσύνολα με αποτέλεσμα να αναπτυχθεί ένα δενδροειδές σε όλη τη διάρκεια αυτής της διαδικασίας. Κάθε κόμβος στο δέντρο αντιπροσωπεύει ένα σημείο "απόφασης" και διαχωρίζει τα δεδομένα με βάση το καλύτερο δυνατό χαρακτηριστικό για να μεγιστοποιήσει το κέρδος πληροφορίας. Βασικά στοιχεία των δέντρων αποφάσεων είναι ο ριζικός κόμβος (root node), οι κόμβοι απόφασης (decision nodes) και οι κόμβοι φύλλων (leaf nodes).

Ριζικός κόμβος

Ο ριζικός κόμβος (root node) σε ένα δέντρο αποφάσεων αποτελεί το κεντρικό σημείο από το οποίο ξεκινάει η διαδικασία ταξινόμησης ή πρόβλεψης. Ο ριζικός κόμβος επιλέγει την καλύτερη μεταβλητή πρόβλεψης που διαιρεί το σύνολο δεδομένων σε υποσύνολα που είναι όσο το δυνατόν πιο ομοιογενή όσον αφορά τη μεταβλητή στόχο. Αυτή η επιλογή μεταβλητής βασίζεται σε κριτήρια όπως το “Information Gain” ή το “Gini Impurity”, τα οποία βοηθούν στον προσδιορισμό της μεταβλητής που διαχωρίζει καλύτερα τα δεδομένα σε ομάδες που είναι καθαρές ως προς τη μεταβλητή αποτελέσματος. Χρησιμοποιώντας την επιλεγμένη μεταβλητή πρόβλεψης, ο ριζικός κόμβος χωρίζει τα δεδομένα σε δύο ή περισσότερα υποσύνολα. Αυτά τα υποσύνολα είναι πιο ομοιογενή ή πιο καθαρά σε σχέση με τη μεταβλητή στόχο, που σημαίνει ότι τα δεδομένα σε κάθε υποσύνολο έχουν παρόμοιες ή ίδιες τιμές για τη μεταβλητή αποτελέσματος. Χρησιμοποιώντας την επιλεγμένη μεταβλητή πρόβλεψης, ο ριζικός κόμβος χωρίζει τα δεδομένα σε δύο ή περισσότερα υποσύνολα. Κάθε υποσύνολο σχηματίζει έναν κλάδο που οδηγεί σε πρόσθετους κόμβους απόφασης ή κόμβους φύλλων. Η απόφαση του ριζικού κόμβου υπαγορεύει τη διαδρομή που θα ακολουθήσουν τα δεδομένα διαδρομής στο δέντρο. Τα σημεία δεδομένων που πληρούν

τη συνθήκη που δοκιμάστηκε στη ρίζα μετακινούνται σε έναν κλάδο, ενώ αυτά που δεν το κάνουν μετακινούνται στους άλλους κλάδους. Αυτή η διακλάδωση συνεχίζεται κάτω από το δέντρο, αυξάνοντας την ομοιογένεια των υποσυνόλων με κάθε διάσπαση. Επιλέγοντας πρώτα την πιο περιγραφική μεταβλητή, ο ριζικός κόμβος διασφαλίζει ότι οι διαχωρισμοί σε κάθε επόμενο κόμβο πρέπει να επεξεργάζονται συνολικά λιγότερα δεδομένα, γεγονός που βελτιστοποιεί την ικανότητα του δέντρου να κάνει ακριβείς προβλέψεις γρήγορα. Η επιλογή της μεταβλητής στη ρίζα επηρεάζει ολόκληρη τη δομή του δέντρου αποφάσεων. Μια σωστά επιλεγμένη ρίζα θα δημιουργήσει ένα καλά ισορροπημένο δέντρο που απαιτεί λιγότερες διασπάσεις για τον καθαρισμό των δεδομένων, μειώνοντας την πολυπλοκότητα του δέντρου.

Κόμβοι απόφασης

Οι κόμβοι απόφασης (decision nodes) είναι τα θεμελιώδη στοιχεία σε ένα δέντρο αποφάσεων που διχάζουν με σύνεση το σύνολο δεδομένων σε πιο ομοιογενή υποσύνολα, μια διαδικασία καθοριστική για τη βελτίωση της σαφήνειας και της καθαρότητας των προβλέψεων που γίνονται. Αυτοί οι κόμβοι αξιολογούν σχολαστικά τα συγκεκριμένα χαρακτηριστικά, επιλέγοντας αυτό που κατανέμει βέλτιστα τα δεδομένα, διευκολύνοντας έτσι μια δομημένη και διορατική ανατομή του συνόλου δεδομένων σε υποσύνολα που είναι σημαντικά πιο ομοιόμορφα ως προς τη μεταβλητή στόχο. Κάθε κόμβος απόφασης ενσωματώνει ένα κρίσιμο σημείο απόφασης όπου τα δεδομένα χωρίζονται κατά μήκος των γραμμών που υπαγορεύονται από τα εγγενή χαρακτηριστικά του χαρακτηριστικού, με κάθε κλάδο που προέρχεται από τον κόμβο να αντιπροσωπεύει ένα πιθανό αποτέλεσμα της δοκιμής. Αυτός ο μεθοδικός διαχωρισμός όχι μόνο βελτιώνει τη δομή των δεδομένων, αλλά βοηθά επίσης στη σταδιακή αποσαφήνιση των υποκείμενων προτύπων, ενισχύοντας τελικά έναν μηχανισμό λήψης αποφάσεων που είναι ταυτόχρονα ισχυρός και ερμηνεύσιμος. Η στρατηγική τοποθέτηση και λειτουργία αυτών των κόμβων απόφασης υποστηρίζει καθοριστικά την ικανότητα του Δέντρου Αποφάσεων να απομυθοποιεί πολύπλοκα σύνολα δεδομένων, καθιστώντας τα ανεκτίμητα στην προγνωστική μοντελοποίηση.

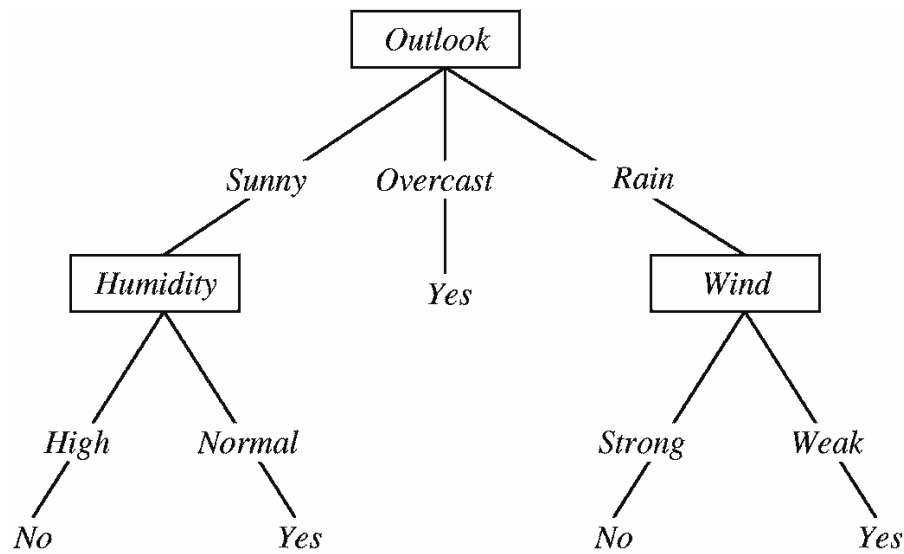
Κόμβοι φύλλων

Οι κόμβοι φύλλων (leaf nodes) αντιπροσωπεύουν το αποκορύφωμα της διαδικασίας λήψης αποφάσεων μέσα σε ένα Δέντρο Αποφάσεων, που χρησιμεύουν ως οριστικοί ταξινομητές ή προγνωστικοί παράγοντες μετά τη διέλευση των δεδομένων μέσα από

το σχολαστικό πλαίσιο των Κόμβων Αποφάσεων. Κάθε κόμβος φύλλων ενσωματώνει μια τελική απόφαση, μια έξοδο που έχει τελειοποιηθεί μέσω διαδοχικών επιπέδων τμηματοποίησης δεδομένων που ξεκινούν από τον κόμβο ρίζας και πραγματοποιούνται από τους Κόμβους Απόφασης. Αυτοί οι κόμβοι στέκονται ως οι τελικές αποθήκες κατηγορικών ή συνεχών αποτελεσμάτων, ανάλογα με τη φύση της εργασίας στο χέρι - είτε πρόκειται για ταξινόμηση είτε παλινδρόμηση. Οι διαδρομές που οδηγούν σε αυτούς τους κόμβους φύλλων δημιουργούνται μέσω μιας σειράς στρατηγικών διαχωρισμών, καθεμία σχεδιασμένη για να μεγιστοποιήσει την ομοιογένεια και να ελαχιστοποιήσει την εντροπία των υποσυνόλων που προκύπτουν. Κατά συνέπεια, οι κόμβοι φύλλων όχι μόνο σηματοδοτούν την επίλυση της υπολογιστικής διαδικασίας αλλά επίσης ενσωματώνουν την αποσταγμένη ουσία των δεδομένων εισόδου, παρέχοντας σαφείς, εφαρμόσιμες ιδέες. Ο ρόλος τους είναι κρίσιμος στη μετατροπή των πολύπλοκων μονοπατιών και χαρακτηριστικών δεδομένων σε απλές, κατανοητές προβλέψεις, υπογραμμίζοντας έτσι την απαραίτητη χρησιμότητά τους στην αρχιτεκτονική του Δέντρου Αποφάσεων.

Τα οφέλη από τη χρήση των δέντρων αποφάσεων είναι ότι τα δέντρα απόφασης κατανοούνται και ερμηνεύονται εύκολα. Η οπτική αναπαράσταση των κόμβων απόφασης, των κλάδων και των κόμβων φύλλων διευκολύνει την οπτικοποίηση του τρόπου λήψης των αποφάσεων, κάτι που είναι ιδιαίτερα πλεονεκτικό για την παρουσίαση και την κατανόηση της διαδικασίας λήψης αποφάσεων. Μπορούν να διαχειρίζονται τόσο κατηγορικά όσο και συνεχή δεδομένα, καθιστώντας τα ευέλικτα για διάφορους τύπους δεδομένων εισόδου. Τα δέντρα απόφασης εκτελούν εγγενώς την επιλογή χαρακτηριστικών. Οι κορυφαίοι κόμβοι στα δέντρα αποφάσεων είναι οι πιο σημαντικές μεταβλητές στο σύνολο δεδομένων και παρέχουν άμεσες πληροφορίες για τα πιο προγνωστικά χαρακτηριστικά. Είναι ικανά να καταγράφουν μη γραμμικές σχέσεις μεταξύ χαρακτηριστικών, κάτι που πολλά γραμμικά μοντέλα δεν μπορούν να κάνουν, χωρίς να απαιτούν μετασχηματισμούς στα δεδομένα. Τα δέντρα αποφάσεων χρησιμεύουν ως ισχυρό εργαλείο για προγνωστικές αναλύσεις, βοηθώντας στην απλοποίηση σύνθετων συνόλων δεδομένων χαρτογραφώντας σαφείς διαδρομές αποφάσεων και είναι ανεκτίμητης αξίας σε τομείς όπου η σαφής ερμηνεία και η ευκολία επικοινωνίας είναι ζωτικής σημασίας. Αυτό τα καθιστά ελκυστική επιλογή για αρχικές έρευνες στο σύνολο δεδομένων και παρέχει ένα καλό σημείο αναφοράς για την

προγνωστική απόδοση πιο περίπλοκων αλγορίθμων. (James et al., 2017; Quinlan, 1986)



Εικόνα 5. Διαγραμματική απεικόνιση δέντρου απόφασης (<https://towardsdatascience.com/decision-tree-in-machine-learning-e380942a4c96>)

5.3.2 Navie Bayes

Ο Naive Bayes (NB) είναι μια τεχνική ταξινόμησης που βασίζεται στο θεώρημα Bayes με μια υπόθεση ανεξαρτησίας μεταξύ των προγνωστικών (Mahesh, 2020). Ο NB είναι ένας δημοφιλής αλγόριθμος που χρησιμοποιείται κυρίως για εργασίες ταξινόμησης. Η αποτελεσματικότητά του προκύπτει από την υπόθεση ότι όλα τα χαρακτηριστικά είναι ανεξάρτητα το ένα από το άλλο δεδομένης της κλάσης, παρόλο που αυτό συχνά δεν συμβαίνει στα δεδομένα του πραγματικού κόσμου.

Ο αλγόριθμος Naive Bayes είναι μια επέκταση του Θεωρήματος του Bayes που υποστηρίζει σύνολα δεδομένων υψηλών διαστάσεων. Απλοποιεί τους υπολογισμούς υποθέτοντας ανεξαρτησία χαρακτηριστικών δεδομένης της κατηγορίας εξόδου. Αυτή η υπόθεση επιτρέπει την αποτελεσματική ταξινόμηση των δεδομένων, παρά τις πολυπλοκότητες που εισάγει η διάσταση. Ο Naive Bayes είναι γνωστός για την υπολογιστική του αποτελεσματικότητα, καθιστώντας τον ιδανικό για σύνολα δεδομένων υψηλών διαστάσεων που συναντώνται συχνά στην ταξινόμηση κειμένου και σε εφαρμογές σε πραγματικό χρόνο, όπως η ανίχνευση ανεπιθύμητων μηνυμάτων.

Αυτή η αποτελεσματικότητα πηγάζει από την υπόθεση ότι τα χαρακτηριστικά είναι ανεξάρτητα, απλοποιώντας τον υπολογισμό των πιθανοτήτων υπό όρους σε απλούς πολλαπλασιασμούς και αναζητήσεις πιθανοτήτων. Κατά συνέπεια, ο Naive Bayes μπορεί γρήγορα να εκπαιδευτεί σε μεγάλα σύνολα δεδομένων και να κάνει γρήγορες

προβλέψεις, κάτι που είναι ζωτικής σημασίας σε περιβάλλοντα που απαιτούν επεξεργασία και αποφάσεις σε πραγματικό χρόνο. Η γραμμική επεκτασιμότητά του σε σχέση με τον αριθμό των προβλέψεων και των σημείων δεδομένων ενισχύει περαιτέρω την καταλληλότητά του για εφαρμογές μεγάλης κλίμακας, διασφαλίζοντας σταθερή απόδοση ακόμη και όταν αυξάνεται ο όγκος δεδομένων.

Αυτός ο συνδυασμός ταχύτητας, απλότητας και επεκτασιμότητας καθιστά τον Naive Bayes έναν προτιμώμενο αλγόριθμο, ειδικά σε ρυθμίσεις περιορισμένων πόρων. Είναι ιδιαίτερα ικανός στη διαχείριση δεδομένων που λείπουν σε ένα σύνολο δεδομένων. Αυτή η ικανότητα οφείλεται στην πιθανολογική βάση του, η οποία εγγενώς εξυπηρετεί την απουσία δεδομένων παραλείποντας απλώς τιμές που λείπουν κατά τη φάση υπολογισμού πιθανοτήτων. Σε αντίθεση με τα μοντέλα που απαιτούν πλήρεις περιπτώσεις ή καταλογισμό καταχωρήσεων που λείπουν, ο Naive Bayes αντιμετωπίζει κάθε χαρακτηριστικό ανεξάρτητα, επιτρέποντάς του να κάνει προβλέψεις χρησιμοποιώντας τα διαθέσιμα δεδομένα χωρίς να εισάγει ή να μαντέψει τις τιμές που λείπουν.

Ο Naive Bayes επιδεικνύει αξιοσημείωτη ευελιξία σε ένα ευρύ φάσμα εφαρμογών, από την ταξινόμηση κειμένου και τον εντοπισμό ανεπιθύμητων μηνυμάτων έως την ιατρική διάγνωση και την ανάλυση συναισθημάτων. Η υποκείμενη πιθανολογική του προσέγγιση του επιτρέπει να διαχειρίζεται αποτελεσματικά προβλήματα ταξινόμησης τόσο δυαδικών όσο και πολλαπλών κλάσεων. Επιπλέον, ο Naive Bayes μπορεί να προσαρμοστεί για να χειρίζεται συνεχή δεδομένα, υποθέτοντας διαφορετικές στατιστικές κατανομές, όπως την Gaussian για εισόδους πραγματικών τιμών.

Αυτή η ευελιξία το καθιστά ιδιαίτερα χρήσιμο σε διάφορους τομείς, επιτρέποντάς του να παρέχει ισχυρή απόδοση ανεξάρτητα από τη φύση των δεδομένων ή την πολυπλοκότητα της εργασίας. Η ικανότητά του να παρέχει πιθανολογικές πληροφορίες σχετικά με τις συμμετοχές στην τάξη ενισχύει τη χρησιμότητά του, καθιστώντας τον ένα αξιόπιστο και διορατικό εργαλείο για πολλές προκλήσεις πρόβλεψης μοντελοποίησης.

Αυτός ο αλγόριθμος όχι μόνο ταξινομεί τα δεδομένα αλλά και ποσοτικοποιεί τη βεβαιότητα μέσω πιθανολογικών αποτελεσμάτων, προσφέροντας σαφείς πληροφορίες για την πιθανότητα κάθε κατηγορίας. Αυτό το πιθανοτικό αποτέλεσμα είναι εξαιρετικά ωφέλιμο για διαδικασίες λήψης αποφάσεων, όπου η κατανόηση του επιπέδου εμπιστοσύνης των προβλέψεων είναι ζωτικής σημασίας. Για παράδειγμα, στην ιατρική

διάγνωση, η γνώση των πιθανοτήτων διαφόρων ασθενειών βοηθά στην πιο αποτελεσματική αξιολόγηση του κινδύνου.

Ο Naive Bayes είναι εγγενώς ανθεκτικός σε άσχετα χαρακτηριστικά λόγω του αγνωστικισμού των χαρακτηριστικών του, που σημαίνει ότι αντιμετωπίζει κάθε χαρακτηριστικό ανεξάρτητα κατά τον υπολογισμό των πιθανοτήτων. Αυτό το χαρακτηριστικό του επιτρέπει να διατηρεί την απόδοση ακόμη και όταν το σύνολο δεδομένων περιλαμβάνει μη πληροφοριακούς προγνωστικούς παράγοντες, οι οποίοι ενδέχεται να υποβαθμίσουν την απόδοση πιο περίπλοκων μοντέλων.

Παρά τις απλοποιήσεις του, ο Naive Bayes μπορεί να είναι εξαιρετικά αποτελεσματικός, ειδικά υπό τις κατάλληλες συνθήκες, όπως όταν οι προγνωστικοί παράγοντες είναι ανεξάρτητοι ή οι συσχετισμοί είναι ελάχιστοι. Η ικανότητά του να χειρίζεται μεγάλους όγκους δεδομένων με υψηλή ταχύτητα και καλές μετρήσεις απόδοσης τον καθιστά ένα ανεκτίμητο εργαλείο για αρχική διερευνητική ανάλυση επιστημονικών δεδομένων. Συχνά χρησιμεύει ως σημείο αναφοράς για τη σύγκριση της απόδοσης πιο εξελιγμένων αλγορίθμων. Ο Naive Bayes παραμένει μια δημοφιλής επιλογή λόγω της απλότητας, της αποτελεσματικότητάς του και των διαισθητικών πιθανοτήτων που παρέχει.

Το Θεώρημα του Bayes, στο οποίο βασίζεται ο αλγόριθμος Naive Bayes, υιοθετεί μια σημαντική απλοποιημένη μορφή την ανεξαρτησία των προβλεπτών. Αυτή η "αφελής" υπόθεση ανεξαρτησίας προϋποθέτει ότι κάθε χαρακτηριστικό συνεισφέρει ανεξάρτητα στην πιθανότητα της κλάσης που θα καταλήξει ένα δεδομένο, χωρίς να επηρεάζεται από την παρουσία ή τις τιμές άλλων χαρακτηριστικών (John, n.d.).

$$P(c | x) = \frac{P(x | c)P(c)}{P(x)}$$

The diagram shows the formula for Bayes' Theorem with four labels and arrows pointing to the components: 'Likelihood' points to $P(x | c)$, 'Class Prior Probability' points to $P(c)$, 'Posterior Probability' points to $P(c | x)$, and 'Predictor Prior Probability' points to $P(x)$.

$$P(c | X) = P(x_1 | c) \times P(x_2 | c) \times \dots \times P(x_n | c) \times P(c)$$

Εικόνα 6 Το θεώρημα του Naive Bayes https://www.saedsayad.com/naive_bayesian.htm

5.3.3 Support Vector Machine

Οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines - SVM) αποτελούν έναν ισχυρό αλγόριθμο μηχανικής μάθησης, βασισμένο στη θεωρία Vapnik-Chervonenkis. Ένας από τους βασικούς στόχους του SVM είναι η μεγιστοποίηση του περιθωρίου μεταξύ των κλάσεων δεδομένων, δηλαδή η απόσταση μεταξύ ενός υπερεπιπέδου και των πλησιέστερων σημείων δεδομένων, τα οποία ονομάζονται διανύσματα υποστήριξης. Αυτό επιτυγχάνει τη δημιουργία ενός σαφούς και βέλτιστου ορίου απόφασης, μειώνοντας τον κίνδυνο υπερπροσαρμογής και διασφαλίζοντας ότι το μοντέλο είναι ανθεκτικό σε παραλλαγές και θόρυβο.

Η μέθοδος που χρησιμοποιούν τα SVM για τη δημιουργία αυτού του ορίου είναι αποτελεσματική, καθώς το μοντέλο εστιάζει στα κρίσιμα δεδομένα που βρίσκονται πιο κοντά στο όριο απόφασης. Αυτό μειώνει την υπολογιστική πολυπλοκότητα, διότι το μοντέλο δεν επεξεργάζεται όλο το σύνολο δεδομένων αλλά μόνο τα διανύσματα υποστήριξης. Αυτή η ιδιότητα καθιστά τα SVM ιδανικά για την ανίχνευση περιπτώσεων με ακραίες τιμές ή θόρυβο, καθώς εστιάζουν στις πιο σημαντικές πληροφορίες για τη βελτιστοποίηση του μοντέλου.

Ένα άλλο κρίσιμο στοιχείο του SVM είναι η χρήση του Kernel Trick. Αυτή η τεχνική επιτρέπει στα SVM να χειρίζονται μη γραμμικά διαχωρισμένα δεδομένα με την εφαρμογή ενός μετασχηματισμού σε υψηλότερη διάσταση, όπου τα δεδομένα είναι γραμμικά διαχωρίσιμα. Χωρίς να χρειάζονται επεξεργασία υψηλής υπολογιστικής πολυπλοκότητας, τα SVM χρησιμοποιούν συναρτήσεις πυρήνα όπως την ακτινική συνάρτηση βάσης (RBF) ή την πολυωνυμική συνάρτηση για να διευκολύνουν την ταξινόμηση πολύπλοκων δεδομένων.

Ένα από τα μεγαλύτερα πλεονεκτήματα των SVM είναι η ικανότητά τους να διαχειρίζονται προβλήματα υψηλών διαστάσεων, όπου ο αριθμός των χαρακτηριστικών υπερβαίνει τον αριθμό των παραδειγμάτων. Αυτή η ιδιότητα τα καθιστά ιδιαίτερα κατάλληλα για εφαρμογές όπως η ταξινόμηση κειμένου ή γονιδιώματος. Επιπλέον, η ανθεκτικότητα των SVM σε θόρυβο και ακραίες τιμές, χάρη στη στρατηγική της μεγιστοποίησης του περιθωρίου, τους δίνει τη δυνατότητα να αποφεύγουν την υπερπροσαρμογή και να γενικεύουν καλά σε νέα δεδομένα.

Συνοπτικά, τα SVM είναι ένας αλγόριθμος υψηλής αποδοτικότητας για προβλήματα μηχανικής μάθησης, με ιδιαίτερη ικανότητα στη διαχείριση δεδομένων υψηλών

διαστάσεων και την προσαρμογή μέσω διαφορετικών συναρτήσεων πυρήνα. Η ισχυρή τους θεωρητική βάση και η ικανότητά τους να διαχωρίζουν με ακρίβεια τις κλάσεις σε σύνθετα προβλήματα καθιστούν τα SVM κορυφαία επιλογή σε πολλές εφαρμογές ταξινόμησης. (Burges, 1998; Cortes et al., 1995)

5.3.4 Random Forest

Τα τυχαία δάση (Random Forests) αποτελούν έναν από τους πιο διαδεδομένους αλγόριθμους μηχανικής μάθησης, καθώς μπορούν να χρησιμοποιηθούν τόσο για ταξινομήσεις όσο και για παλινδρομήσεις. Ο βασικός τους μηχανισμός στηρίζεται στη δημιουργία πολλαπλών δέντρων αποφάσεων κατά τη φάση εκπαίδευσης, όπου κάθε δέντρο είναι κατασκευασμένο από ένα τυχαίο υποσύνολο δεδομένων και χαρακτηριστικών. Η συνδυασμένη πρόβλεψη από τα πολλά δέντρα μειώνει τον κίνδυνο υπερπροσαρμογής και αυξάνει την ακρίβεια των προβλέψεων.

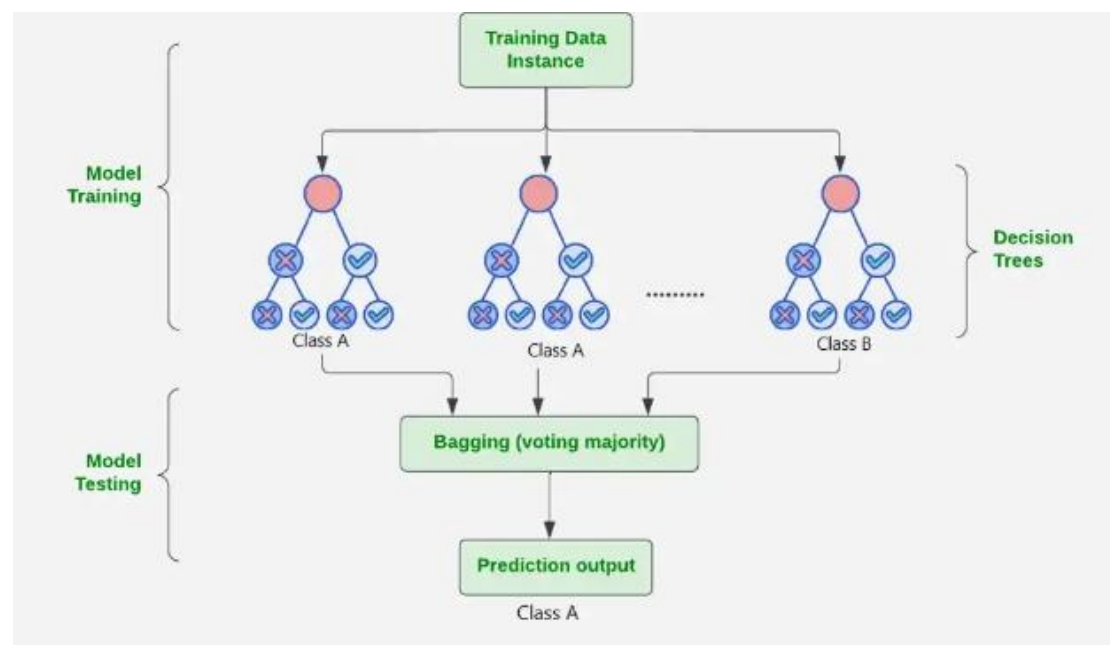
Η βασική λειτουργία των τυχαίων δασών στηρίζεται στην αρχή του συνόλου (ensemble learning), όπου πολλές μονάδες (τα δέντρα αποφάσεων) συνεργάζονται για να παραχθεί ένα τελικό αποτέλεσμα. Αυτό επιτρέπει στον αλγόριθμο να είναι πιο ανθεκτικός σε δεδομένα με θόρυβο ή ακραίες τιμές, καθώς κάθε δέντρο έχει κατασκευαστεί από διαφορετικό τμήμα των δεδομένων και έτσι δεν επηρεάζεται το σύνολο από τυχόν ακραίες παρατηρήσεις.

Ένα από τα κύρια πλεονεκτήματα των τυχαίων δασών είναι η ικανότητά τους να προσδιορίζουν τη σχετικότητα των χαρακτηριστικών. Κατά τη διάρκεια της εκπαίδευσης, ο αλγόριθμος μπορεί να εντοπίσει ποια χαρακτηριστικά συμβάλλουν περισσότερο στη βελτίωση της ακρίβειας του μοντέλου, παρέχοντας πολύτιμη πληροφόρηση για την επιλογή των πιο κρίσιμων μεταβλητών. Αυτό επιτυγχάνεται μέσω μετρικών όπως η ανομοιογένεια Gini (impurity) και η εντροπία για ταξινομήσεις, ή η διακύμανση για παλινδρομήσεις.

Επιπλέον, τα τυχαία δάση έχουν την ικανότητα να διαχειρίζονται σύνολα δεδομένων που περιλαμβάνουν τόσο αριθμητικά όσο και κατηγορικά χαρακτηριστικά, χωρίς να απαιτείται κλιμάκωση των δεδομένων. Αυτή η ιδιότητα καθιστά τον αλγόριθμο εύχρηστο, καθώς απαιτεί ελάχιστες ρυθμίσεις παραμέτρων για τη βελτιστοποίηση. Οι κύριες παράμετροι που χρειάζεται να οριστούν είναι ο αριθμός των δέντρων και ο αριθμός των χαρακτηριστικών που θα δειγματοληφθούν για τη δημιουργία κάθε δέντρου, καθιστώντας τον αλγόριθμο φιλικό προς τον χρήστη.

Η μέθοδος αυτή είναι λιγότερο ευαίσθητη στο θόρυβο των δεδομένων, καθώς ο τυχαίος χαρακτήρας της δειγματοληψίας μειώνει την πιθανότητα το μοντέλο να προσαρμοστεί σε ακραίες τιμές ή θορυβώδη δεδομένα. Επιπλέον, η προσέγγιση του αλγορίθμου με τα πολλαπλά δέντρα συμβάλλει στη βελτίωση της ακρίβειας χωρίς να αυξάνεται η μεροληψία του μοντέλου.

Δεδομένων αυτών των ιδιοτήτων, τα τυχαία δάση είναι ιδανικά για σύνθετα προβλήματα πρόβλεψης που περιλαμβάνουν πολλούς τύπους δεδομένων και θόρυβο. Παρέχουν επίσης καλή γενίκευση σε νέα δεδομένα και αξιοπιστία στις προβλέψεις τους, καθιστώντας τα έναν ισχυρό αλγόριθμο για ένα ευρύ φάσμα εφαρμογών της μηχανικής μάθησης, από την ανάλυση δεδομένων μέχρι την ανάπτυξη προγνωστικών μοντέλων (Breiman, 2001; Louppe, 2014).



Εικόνα 7 Απεικόνιση μεθόδου bagging με δέντρα απόφασης <https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/>

5.3.5 Linear Discriminant Analysis

Η Γραμμική Διακριτική Ανάλυση (Linear Discriminant Analysis - LDA) αποτελεί έναν σημαντικό αλγόριθμο μηχανικής μάθησης, ιδιαίτερος χρήσιμο για την εποπτευόμενη ταξινόμηση. Διακρίνεται για την ικανότητά της να διαχωρίζει κλάσεις σε καταστάσεις όπου τα όρια μεταξύ των κατηγοριών είναι σαφώς διακριτά. Η LDA επιδιώκει να μεγιστοποιήσει τη διαχωριστικότητα μεταξύ των κλάσεων, ελαχιστοποιώντας ταυτόχρονα τη διακύμανση εντός της κάθε κλάσης, κάτι που την

καθιστά ιδιαίτερα αποτελεσματική στη δημιουργία σαφών ορίων απόφασης. Αυτή η προσέγγιση μειώνει την υπολογιστική πολυπλοκότητα και ενισχύει την ερμηνευσιμότητα του μοντέλου, αποφεύγοντας την υπερπροσαρμογή που μπορεί να εμφανιστεί σε πιο σύνθετα μοντέλα, όπως τα νευρωνικά δίκτυα, ειδικά όταν τα δεδομένα είναι περιορισμένα.

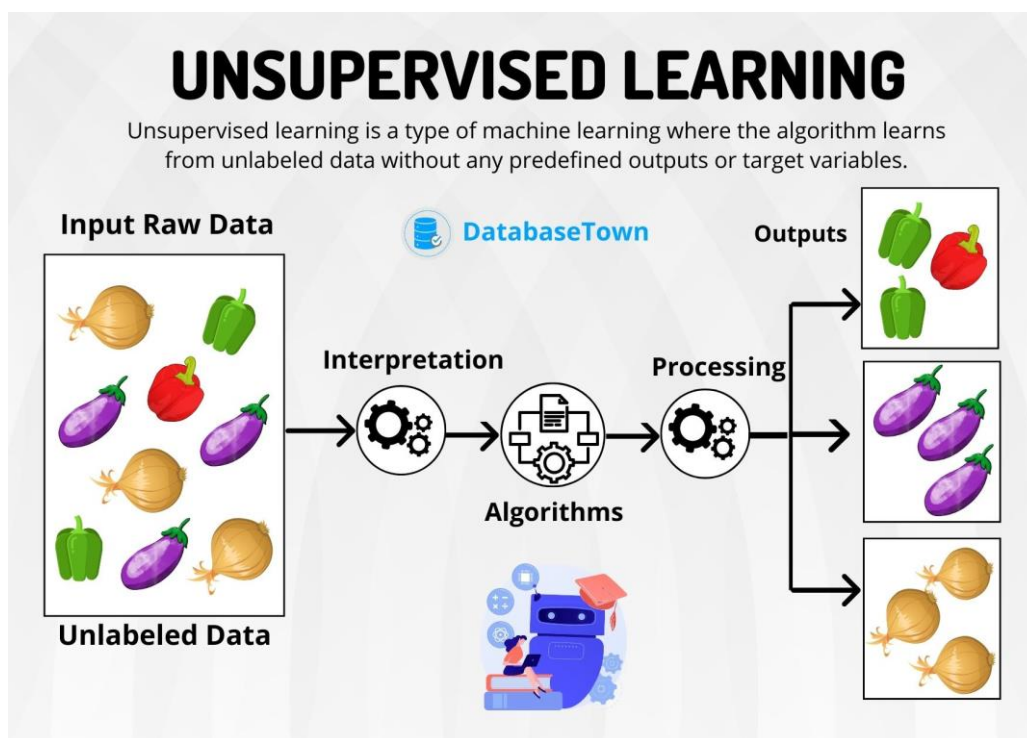
Σε καταστάσεις με περιορισμένα δεδομένα, η LDA έχει το πλεονέκτημα της καλής γενίκευσης, κάτι που την καθιστά ιδανική όταν η συλλογή μεγάλων συνόλων δεδομένων είναι δύσκολη ή δαπανηρή. Ενώ πιο περίπλοκα μοντέλα, όπως τα νευρωνικά δίκτυα, απαιτούν συχνά μεγάλα σύνολα δεδομένων για να αποδώσουν αποτελεσματικά, η LDA μπορεί να αποδώσει ικανοποιητικά με μικρότερες ποσότητες δεδομένων, διατηρώντας την απλότητα και την ευρωστία των αποτελεσμάτων της.

Η LDA βασίζεται στην υπόθεση ότι τα δεδομένα ακολουθούν κανονική κατανομή (Gaussian) και ότι οι κλάσεις έχουν ίσες συνδιακυμάνσεις. Όταν αυτές οι υποθέσεις ισχύουν, η LDA μπορεί να αποδειχθεί εξαιρετικά αποτελεσματική, ακόμα και συγκριτικά με πιο εύελictους αλγόριθμους, αφού αξιοποιεί πλήρως τη δομή των δεδομένων, αποφεύγοντας την προσαρμογή σε θόρυβο ή ακραίες τιμές. Αντίθετα, αλγόριθμοι όπως η λογιστική παλινδρόμηση μπορεί να απαιτούν πρόσθετες τεχνικές για να διαχειριστούν τέτοιες προκλήσεις.

Συνολικά, η LDA αποτελεί μια ισχυρή επιλογή, ειδικά όταν η διαφάνεια, η αποδοτικότητα και η καλή γενίκευση σε μικρά σύνολα δεδομένων είναι ζωτικής σημασίας. Προσφέρει μια καλή ισορροπία μεταξύ απλότητας και απόδοσης, καθιστώντας την ιδανική επιλογή σε προβλήματα που απαιτούν σαφή και αξιόπιστα μοντέλα ταξινόμησης. (Ghojogh & Crowley, 2019; Leung et al., n.d.; Tharwat et al., 2017)

Μάθηση χωρίς επίβλεψη

Η μάθηση χωρίς επίβλεψη (Unsupervised Learning) παίζει καθοριστικό ρόλο στη μηχανική μάθηση, επιτρέποντας στους αλγόριθμους να αναγνωρίζουν μοτίβα και δομές μέσα σε δεδομένα χωρίς προκαθορισμένη επισήμανση. Με την ανάλυση των δεδομένων εισόδου, διευκολύνει την ανακάλυψη σχέσεων μεταξύ των δεδομένων και τα ομαδοποιεί με βάση την ομοιότητά τους. Αυτή η προσέγγιση είναι ιδιαίτερα επωφελής στην ανάπτυξη μοντέλων, καθώς αποκαλύπτει κρυφά μοτίβα και δομές χωρίς να απαιτεί εκτεταμένη ανθρώπινη συμβολή, καθιστώντας την οικονομικά αποδοτική και επεκτάσιμη. Τελικά, η μάθηση χωρίς επίβλεψη συμβάλλει στη βαθύτερη κατανόηση σύνθετων συνόλων δεδομένων, ενισχύοντας την καινοτομία και τη λήψη τεκμηριωμένων αποφάσεων σε διάφορους τομείς (Figueiredo & Jain, n.d.; Jain, 2010; Rokach, 2009).



Εικόνα 8 Απεικόνιση μη επιβλεπόμενης μάθησης <https://databasetown.com/unsupervised-learning-types-applications/>

5.4 Ομαδοποίηση

Η ομαδοποίηση (clustering) είναι μια θεμελιώδης τεχνική στη μηχανική μάθηση χωρίς επίβλεψη καθώς διαδραματίζει κρίσιμο ρόλο στην αποκάλυψη πολύπλοκων προτύπων μέσα στα δεδομένα. Οργανώνοντας τα δεδομένα σε ομάδες με βάση τις ομοιότητές τους, η ομαδοποίηση επιτρέπει την οργάνωση μη δομημένων συνόλων δεδομένων σε ουσιαστικά υποσύνολα χωρίς προκαθορισμένες επισημάνσεις, προσφέροντας μία πιο διορατική ανάλυση. Αυτό όχι μόνο διευκολύνει τη βαθύτερη κατανόηση των πολύπλοκων συνόλων δεδομένων, αλλά είναι εξαιρετικά πολύτιμο σε εφαρμογές όπως η τμηματοποίηση πελατών, η ανάλυση αγοράς ή η αναγνώριση εικόνας, όπου η διάκριση παρόμοιων ομάδων είναι το κλειδί για την κατανόηση της συμπεριφοράς ή την πρόβλεψη αποτελεσμάτων. Μέσω αυτών των δυνατοτήτων, η ομαδοποίηση συμβάλλει σημαντικά στην ανακάλυψη πολύτιμων πληροφοριών, οδηγώντας τόσο στην καινοτομία όσο και στην ακρίβεια στη λήψη αποφάσεων βάσει των υπάρχοντων δεδομένων.

Εκτός από τη βασική της λειτουργία, η ομαδοποίηση μπορεί να χρησιμεύσει ως ένα ισχυρό βήμα προεπεξεργασίας σε αγωγούς μηχανικής εκμάθησης. Με τη χρήση της για την οργάνωση δεδομένων σε συμπλέγματα, επιτυγχάνονται πολλαπλά οφέλη. Πρώτον, συμβάλλει στη μείωση των διαστάσεων των δεδομένων, κάνοντάς τα πιο διαχειρίσιμα, ενώ ταυτόχρονα τονίζει τα πιο κρίσιμα χαρακτηριστικά που καθορίζουν τις σχέσεις μεταξύ των σημείων δεδομένων. Αυτό βελτιώνει την υπολογιστική αποδοτικότητα και διευκολύνει την κατανόηση των υποκείμενων μοτίβων.

Στη διαδικασία ανίχνευσης ανωμαλιών, η ομαδοποίηση βοηθά στον εντοπισμό ακραίων τιμών, διαχωρίζοντας τα κανονικά μοτίβα από τα δεδομένα που δεν ακολουθούν το σύννηθες μοτίβο, κάτι που επιτρέπει τη δημιουργία πιο ισχυρών μοντέλων ανίχνευσης. Επιπλέον, η ανάλυση των δεδομένων σε συμπλέγματα παρέχει πολύτιμες πληροφορίες σχετικά με την πολυπλοκότητα και τη διανομή των δεδομένων, βοηθώντας τους ερευνητές να επιλέξουν τους κατάλληλους αλγόριθμους μηχανικής μάθησης.

Η διαδικασία της ομαδοποίησης γενικά περιλαμβάνει τα εξής βήματα:

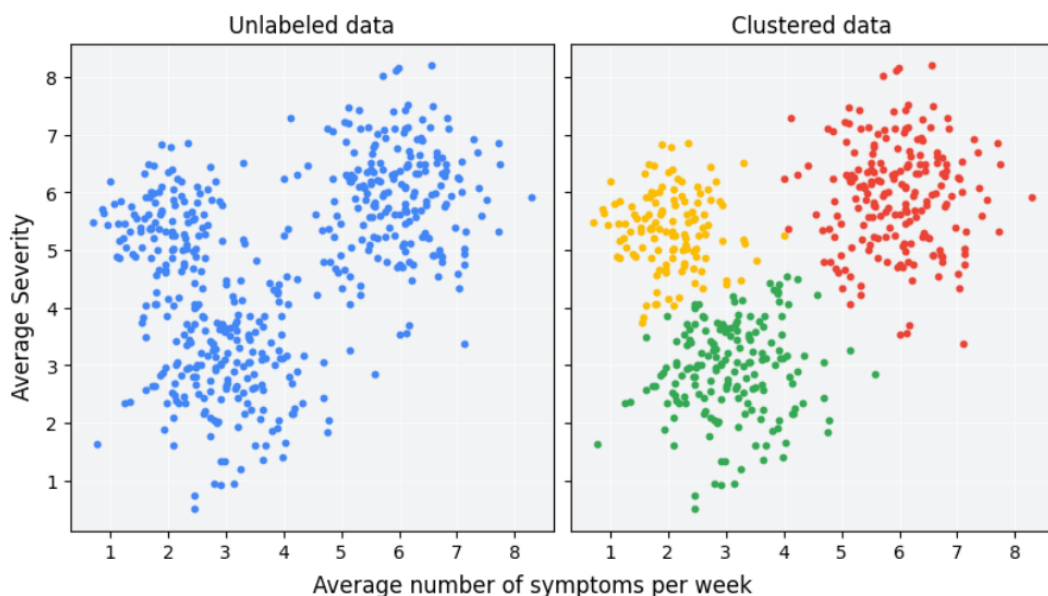
α) Αναπαράσταση δεδομένων (Data Representation): Κάθε σημείο δεδομένων αναπαρίσταται ως διάνυσμα χαρακτηριστικών σε έναν πολυδιάστατο χώρο. Οι διαστάσεις αυτές αντιστοιχούν στα χαρακτηριστικά των δεδομένων, όπως το ύψος, το βάρος ή το χρώμα, ανάλογα με το είδος των δεδομένων που εξετάζονται.

β) Μετρική απόστασης (Distance Metric): Για να μετρηθεί η ομοιότητα μεταξύ των σημείων δεδομένων, χρησιμοποιείται μια μετρική απόστασης, όπως η Ευκλείδεια απόσταση για αριθμητικά δεδομένα ή άλλες εξειδικευμένες συναρτήσεις απόστασης. Όσο πιο κοντά βρίσκονται τα σημεία, τόσο πιο όμοια θεωρούνται.

γ) Επιλογή αλγορίθμου ομαδοποίησης (Clustering Algorithm): Διάφοροι αλγόριθμοι μπορούν να εφαρμοστούν για την ομαδοποίηση των δεδομένων.

δ) Επανάληψη και σύγκλιση (Iteration and Convergence): Ο αλγόριθμος συνεχίζει να προσαρμόζει τα συμπλέγματα μέχρι να επιτευχθεί σύγκλιση, δηλαδή μέχρι τα συμπλέγματα να σταθεροποιηθούν και να μην αλλάζουν περαιτέρω.

ε) Αξιολόγηση (Evaluation): Δεδομένου ότι η ομαδοποίηση είναι χωρίς επιβλέψη, δεν υπάρχουν ετικέτες για να συγκριθούν τα αποτελέσματα. Αντίθετα, η αξιολόγηση γίνεται με εσωτερικά μέτρα όπως η "απόσταση εντός του συμπλέγματος" (πόσο κοντά είναι τα σημεία μεταξύ τους) και η "απόσταση μεταξύ των συστάδων" (πόσο διακριτά είναι τα συμπλέγματα μεταξύ τους). (Rodriguez et al., 2019; Serra-Burriel & Ames, 2022; Yin et al., 2024)



Εικόνα 9 Απεικόνιση ομαδοποίησης δεδομένων (clustering) <https://developers.google.com/machine-learning/clustering/overview>

5.5 Αλγόριθμοι ομαδοποίησης

5.5.1 K-means

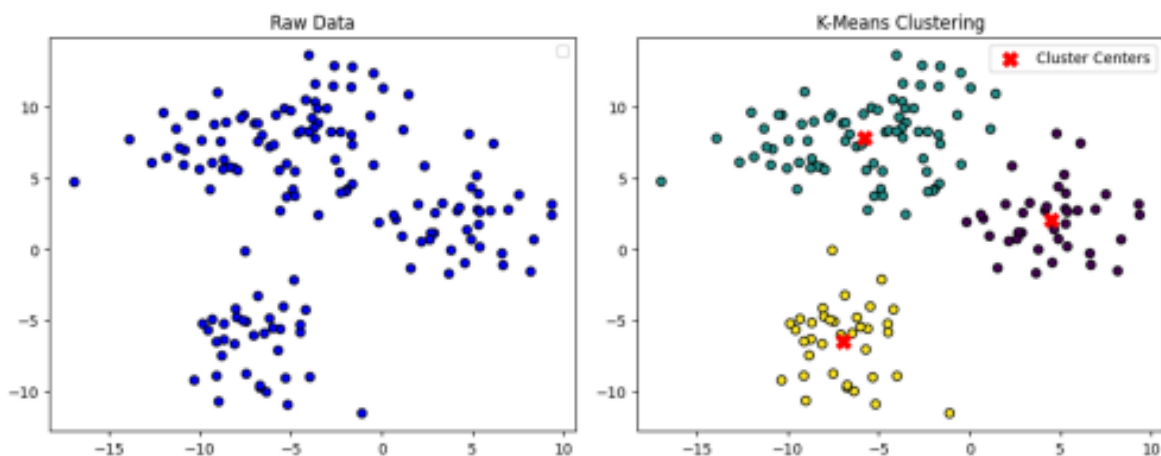
Ο K-means είναι ένας ευρέως προτιμώμενος αλγόριθμος στη μηχανική μάθηση λόγω της απλότητας, της υπολογιστικής απόδοσης και της ευελιξίας του στην αντιμετώπιση διαφόρων εργασιών ομαδοποίησης. Είναι ιδιαίτερα κατάλληλος για τη γρήγορη τμηματοποίηση μεγάλων συνόλων δεδομένων σε ουσιαστικά συμπλέγματα, καθώς ο αλγόριθμος κλιμακώνεται αποτελεσματικά με την αύξηση του μεγέθους των δεδομένων. Ο K-means λειτουργεί διαμερίζοντας επαναληπτικά ένα σύνολο δεδομένων σε K διακριτά συμπλέγματα με βάση την ομοιότητα των σημείων των δεδομένων. Η λειτουργία του αλγόριθμου βήμα προς βήμα:

- α) **Αρχικοποίηση (Initialization):** Ο αλγόριθμος ξεκινά επιλέγοντας K αρχικά κεντροειδή συμπλέγματα (αυτά μπορούν να επιλεγούν τυχαία ή να βασιστούν σε κάποια μέθοδο προετοιμασίας). Αυτά τα κεντροειδή αντιπροσωπεύουν το κέντρο κάθε συστάδας.
- β) **Εκχώρηση σημείων δεδομένων (Assignment of Data Points):** Κάθε σημείο στο σύνολο δεδομένων εκχωρείται στη συνέχεια στη πλησιέστερη συστάδα, τυπικά με βάση την Ευκλείδεια απόσταση μεταξύ του σημείου και του κέντρου. Αυτό το βήμα ομαδοποιεί τα σημεία δεδομένων σε K συμπλέγματα.
- γ) **Ενημέρωση κεντροειδών (Updating Centroids):** Αφού έχουν εκχωρηθεί όλα τα σημεία σε συμπλέγματα, τα κεντροειδή υπολογίζονται εκ νέου. Το νέο κέντρο κάθε συστάδας είναι η κεντρική θέση από όλων των σημείων που έχουν εκχωρηθεί σε αυτό το σύμπλεγμα. Αυτό το βήμα διασφαλίζει ότι τα κεντροειδή βρίσκονται ακριβώς στο κέντρο των σημείων που αντιπροσωπεύουν.
- δ) **Επανάθεση και Επανάληψη (Reassignment and Iteration):** Η διαδικασία εκχώρησης σημείων στο πλησιέστερο κέντρο και επανυπολογισμού των κεντροειδών επαναλαμβάνεται πολλές φορές. Κάθε επανάληψη βελτιώνει την τοποθέτηση των κεντροειδών και τις ομαδοποιήσεις των σημείων, μειώνοντας τη συνολική διακύμανση εντός των συστάδων.
- ε) **Σύγκλιση (Convergence):** Ο αλγόριθμος συνεχίζει να επαναλαμβάνεται έως ότου τα κεντροειδή σταματήσουν να αλλάζουν σημαντικά τις θέσεις τους,

υποδεικνύοντας ότι τα συμπλέγματα έχουν σταθεροποιηθεί. Σε αυτό το σημείο, ο αλγόριθμος συγκλίνει και σχηματίζονται οι τελικές συστάδες.

στ) **Αποτελέσματα (Result):** Μόλις ο αλγόριθμος αρχίσει να συγκλίνει, βγάζει τα τελικά K συμπλέγματα, με κάθε σύμπλεγμα να περιέχει σημεία που μοιάζουν περισσότερο μεταξύ τους παρά με άλλα σημεία που βρίσκονται σε διαφορετικά συμπλέγματα.

Ο K-means εντοπίζει αποτελεσματικά τα μοτίβα, καθιστώντας το ιδανικό για περιπτώσεις χρήσης όπως τμηματοποίηση πελατών, συμπίεση εικόνας και ανίχνευση ανωμαλιών. Ένα από τα βασικά πλεονεκτήματα του αλγορίθμου είναι η απλή εφαρμογή και η ευκολία κατανόησής του, καθιστώντας το προσβάσιμο τόσο σε αρχάριους όσο και σε έμπειρους επαγγελματίες. Είναι ιδιαίτερα αποτελεσματικό όταν εφαρμόζεται σε σύνολα δεδομένων με μέτριο αριθμό διαστάσεων, παρέχοντας αξιόπιστα αποτελέσματα σε σύντομο χρονικό διάστημα. Αν και υποθέτει ότι τα συμπλέγματα είναι σφαιρικά και παρόμοιων μεγεθών, η απόδοσή του μπορεί να βελτιστοποιηθεί με μεθόδους όπως η τεχνική ‘elbow’, η οποία βοηθά στον προσδιορισμό του βέλτιστου αριθμού συστάδων και προηγμένες στρατηγικές αρχικοποίησης κεντροειδών. Αυτές οι βελτιώσεις συμβάλλουν στη δημοτικότητα του ως λύση ομαδοποίησης σε μια ποικιλία πρακτικών εφαρμογών (Monath et al., 2021; Müllner, 2011; Ran et al., 2023; Tokuda et al., 2022).



Εικόνα 10 Απεικόνιση ομαδοποίησης K-Means (22) *A Rapid Review of Cluster*

5.5.2 Agglomerative Method

Η συγκεντρωτική μέθοδος (Agglomerative Method), που χρησιμοποιείται συχνά στην ιεραρχική ομαδοποίηση, είναι μια προσέγγιση από κάτω προς τα πάνω για την ομαδοποίηση παρόμοιων σημείων δεδομένων στη μηχανική μάθηση. Ξεκινά αντιμετωπίζοντας κάθε σημείο δεδομένων ως το δικό του σύμπλεγμα και συγχωνεύει επαναληπτικά ζεύγη συστάδων με βάση την ομοιότητά τους μέχρι να συνδυαστούν όλα τα σημεία σε ένα ενιαίο σύμπλεγμα ή να επιτευχθεί ένας επιθυμητός αριθμός συστάδων. Αυτή η τεχνική είναι εξαιρετικά αποτελεσματική στον εντοπισμό ένθετων συστάδων και να αποκαλύπτει την ιεραρχική δομή των δεδομένων. Στην ανάπτυξη μοντέλων, η συγκεντρωτική μέθοδος επιτρέπει την εξερεύνηση των σχέσεων εντός των δεδομένων, ενισχύοντας τις γνώσεις και τη λήψη αποφάσεων σε τομείς όπως η κατάταξη πελατών, η ταξινόμηση εικόνων και η ανάλυση γονιδιακής έκφρασης. Η ευελιξία του στον χειρισμό διαφορετικών κριτηρίων σύνδεσης και μέτρων απόστασης το καθιστά πολύτιμο εργαλείο, αν και η υπολογιστική του πολυπλοκότητα μπορεί να δημιουργήσει προκλήσεις για μεγάλα σύνολα δεδομένων.

Η συγκεντρωτική μέθοδος προτιμάται σε ορισμένες εφαρμογές μηχανικής εκμάθησης σε αντίθεση με τις μεθόδους επίπεδης ομαδοποίησης (όπως το k-means), δεν απαιτεί τον προκαθορισμό του αριθμού των συστάδων, γεγονός που επιτρέπει στον αλγόριθμο να βρίσκει δυναμικά τις βέλτιστες ομαδοποιήσεις καθώς εξερευνώνται τα δεδομένα. Αυτή η μέθοδος είναι ιδιαίτερα χρήσιμη όταν η φυσική δομή των δεδομένων είναι άγνωστη ή όταν θέλετε να απεικονίσετε πώς τα συμπλέγματα είναι ένθετα το ένα μέσα στο άλλο. Ένα από τα βασικά πλεονεκτήματα της συγκεντρωτικής μεθόδου είναι η ευελιξία της στην προσαρμογή σε διαφορετικούς τύπους δεδομένων και ανάγκες ομαδοποίησης. Αυτό προέρχεται από την ικανότητά της να χρησιμοποιεί διάφορα κριτήρια σύνδεσης και μετρικές απόστασης, που επιτρέπουν στους χρήστες να προσαρμόζουν τον αλγόριθμο με βάση τα χαρακτηριστικά των δεδομένων τους ή τους στόχους της ανάλυσης. Τα κριτήρια σύνδεσης καθορίζουν πώς υπολογίζεται η απόσταση μεταξύ των συστάδων κατά τη συγχώνευσή τους:

- **Η μοναδική σύνδεση (Single linkage)** θεωρεί τη μικρότερη απόσταση μεταξύ οποιωνδήποτε δύο σημείων από δύο διαφορετικά συμπλέγματα. Αυτή η προσέγγιση τείνει να σχηματίζει επιμήκεις συστάδες που μοιάζουν με αλυσίδα.

- **Η πλήρης σύνδεση** (Complete linkage) χρησιμοποιεί την πιο απομακρυσμένη απόσταση μεταξύ οποιωνδήποτε δύο σημείων σε διαφορετικά συμπλέγματα, γεγονός που οδηγεί σε πιο συμπαγή, σφαιρικά συμπλέγματα.
- **Η μέση σύνδεση** (Average linkage) υπολογίζει τη μέση απόσταση μεταξύ όλων των σημείων σε δύο συστάδες, προσφέροντας μια ισορροπία μεταξύ τους και μιας πλήρους σύνδεσης.

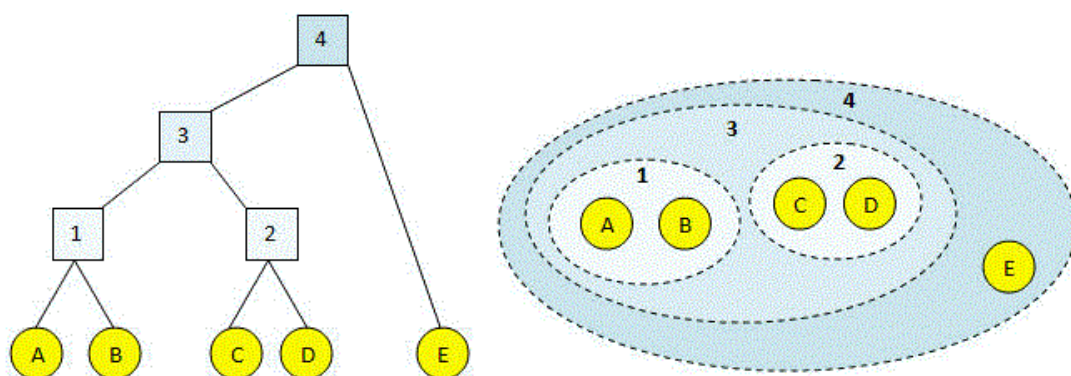
Οι μετρήσεις απόστασης καθορίζουν πώς μετράτε η ομοιότητα ή η ανομοιότητα μεταξύ των σημείων δεδομένων. Όπου οι μετρήσεις περιλαμβάνουν:

Ευκλείδεια απόσταση (Euclidean distance), η οποία μετρά την ευθεία απόσταση μεταξύ δύο σημείων.

Απόσταση Μανχάταν (Manhattan distance), που αθροίζει τις απόλυτες διαφορές μεταξύ των συντεταγμένων δύο σημείων, χρήσιμο για δομές δεδομένων που μοιάζουν με πλέγματα.

Ομοιότητα συνημιτόνου (Cosine similarity), που μετρά το συνημίτονο της γωνίας μεταξύ δύο διανυσμάτων, ιδανικό για κείμενο ή δεδομένα υψηλών διαστάσεων.

Στο πλαίσιο της συγκεντρωτικής μεθόδου, σημαίνει ότι ο αλγόριθμος θα παράγει σταθερά το ίδιο αποτέλεσμα κάθε φορά που εκτελείται στο ίδιο σύνολο δεδομένων με τις ίδιες παραμέτρους. Εφόσον αυτή η Μέθοδος είναι ντετερμινιστική, δεν θα υπάρχει τυχαιότητα στη διαδικασία ομαδοποίησης. Αυτό είναι ένα κρίσιμο χαρακτηριστικό γιατί εγγυάται ότι εάν κάποιος άλλος ερευνητής ή προγραμματιστής χρησιμοποιεί τα ίδια δεδομένα και ρυθμίσεις, θα καταλήξει σε πανομοιότυπα συμπλέγματα, καθιστώντας τα αποτελέσματα αξιόπιστα και συνεπή. Αν και είναι υπολογιστικά εντατικό για μεγάλα σύνολα δεδομένων, η ερμηνευτικότητά του είναι ένα σημαντικό πλεονέκτημα. Η ιεραρχική δομή μπορεί εύκολα να οπτικοποιηθεί μέσω δένδρογραμμάτων, παρέχοντας σαφείς πληροφορίες για το σχηματισμό συμπλέγματος και τις σχέσεις μεταξύ των σημείων δεδομένων. Αυτή η διαφάνεια καθιστά τη συγκεντρωτική μέθοδο ιδιαίτερα πολύτιμη για διερευνητική ανάλυση δεδομένων και εφαρμογές όπου η κατανόηση της διαδικασίας ομαδοποίησης είναι απαραίτητη. (Monath et al., 2021; Müllner, 2011; Ran et al., 2023; Tokuda et al., 2022)



Εικόνα 11 Απεικόνιση ιεραρχικής ομαδοποίησης <https://www.drive5.com/usearc>

5.5.3 DBSCAN

Ο DBSCAN (Density-Based Spatial Clustering of Applications with Noise) είναι ένας ισχυρός αλγόριθμος μηχανικής εκμάθησης χωρίς επίβλεψη, ιδιαίτερα κατάλληλος για εργασίες ομαδοποίησης όπου τα δεδομένα περιέχουν πολύπλοκα, μη γραμμικά μοτίβα ή θόρυβο. Σε αντίθεση με αλγόριθμους όπως το K-Means που βασίζονται σε προκαθορισμένους αριθμούς συστάδων και έχουν καλή απόδοση σε σφαιρικά συμπλέγματα, ο DBSCAN προσδιορίζει συστάδες με βάση την πυκνότητα των σημείων δεδομένων. Αυτό σημαίνει ότι μπορεί να ανακαλύψει συμπλέγματα αυθαίρετων σχημάτων και μεγεθών, χωρίς να απαιτείται προηγούμενη γνώση του πόσα συμπλέγματα υπάρχουν. Ο DBSCAN λειτουργεί κατηγοριοποιώντας τα σημεία σε core points, border points και outliers, με βάση το πόσο πυκνά είναι συσσωρευμένα σε μια συγκεκριμένη γειτονιά, που ορίζονται από δύο βασικές παραμέτρους: epsilon (ϵ), η μέγιστη απόσταση εντός των οποίων τα σημεία θεωρούνται γείτονες και MinPts, ο ελάχιστος αριθμός γειτονικών σημείων που απαιτούνται για να σχηματιστεί μια πυκνή περιοχή.

Κεντρικά σημεία

Τα κεντρικά σημεία (core points) είναι σημεία δεδομένων που βρίσκονται σε περιοχές υψηλής πυκνότητας. Ένα σημείο ταξινομείται ως κεντρικό σημείο εάν έχει τουλάχιστον MinPts γείτονες σε απόσταση ϵ . Αυτά τα βασικά σημεία είναι ουσιαστικά τα «κεντρικά» μέλη ενός συμπλέγματος, καθώς περιβάλλονται από αρκετά κοντινά σημεία για να σχηματίσουν μια πυκνή περιοχή συστάδων.

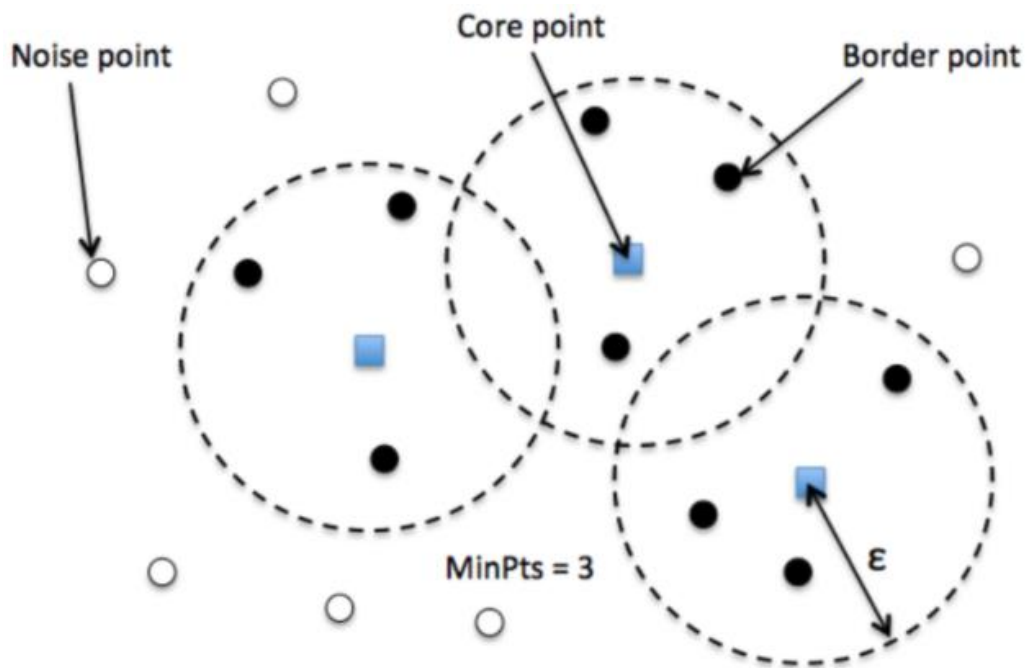
Οριακά σημεία

Τα οριακά σημεία (border points) είναι σημεία που δεν πληρούν την απαίτηση MinPts ώστε να είναι βασικά σημεία, αλλά βρίσκονται εντός της γειτονιάς ενός βασικού σημείου. Αν και δεν έχουν αρκετά γειτονικά σημεία για να σχηματίσουν το δικό τους σύμπλεγμα, είναι αρκετά κοντά στα σημεία του πυρήνα ώστε να αποτελούν μέρος του ορίου του συμπλέγματος. Τα συνοριακά σημεία βοηθούν στον καθορισμό του σχήματος ενός συμπλέγματος, αλλά είναι λιγότερο κεντρικά στη δομή του.

Ακραίες τιμές

Οι ακραίες τιμές (Outliers), ή τα σημεία θορύβου είναι σημεία που είναι πολύ μακριά από οποιαδήποτε βασικά σημεία και δεν πληρούν τις απαιτήσεις πυκνότητας για να ανήκουν σε οποιοδήποτε σύμπλεγμα. Αυτά τα σημεία είναι απομονωμένα, πέφτουν έξω από τις πυκνές περιοχές και συχνά αντιμετωπίζονται ως ανωμαλίες ή θόρυβος στα δεδομένα.

Αυτή η ταξινόμηση με βάση την πυκνότητα καθιστά τον DBSCAN αποτελεσματικό στον εντοπισμό συστάδων διαφορετικών σχημάτων και μεγεθών, ενώ χειρίζεται φυσικά τον θόρυβο. Σε αντίθεση με τους αλγόριθμους που εξαναγκάζουν κάθε σημείο σε ένα σύμπλεγμα, ο DBSCAN μπορεί να αγνοήσει τα ακραία σημεία, οδηγώντας σε πιο ουσιαστικά αποτελέσματα ομαδοποίησης σε σύνολα δεδομένων που περιέχουν ακανόνιστες δομές ή θόρυβο (Ester et al., 1996; Han et al., 2011; Küchenhoff et al., 2023).



Εικόνα 12 Απεικόνιση ομαδοποίησης DBSCAN. *Core Points*: ένα σημείο που έχει τουλάχιστον m σημεία σε απόσταση n από τον εαυτό του. *Border Points*: ένα σημείο που έχει τουλάχιστον ένα σημείο πυρήνα σε απόσταση n . *Outliers (Noise)*: ένα σημείο που δεν είναι ούτε πυρήνας ούτε σύνορο και έχει λιγότερα από m σημεία σε απόσταση n από τον εαυτό του. <https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html>

5. Λογισμικά που χρησιμοποιήθηκαν

5.1 Google Colab

Το Google Colab, είναι μια δωρεάν πλατφόρμα που βασίζεται σε cloud η οποία αναπτύχθηκε από την Google Research. Παρέχει στους χρήστες τη δυνατότητα να γράφουν και να εκτελούν κώδικα σε Python απευθείας από ένα πρόγραμμα περιήγησής, καθιστώντας τον ιδιαίτερα χρήσιμο για εργασίες όπως η μηχανική μάθηση, η ανάλυση δεδομένων και η βαθιά εκμάθηση (deep learning).

Ένα από τα πιο αξιοσημείωτα χαρακτηριστικά του Google Colab είναι ότι επιτρέπει στους χρήστες να αξιοποιούν ισχυρούς υπολογιστικούς πόρους χωρίς να χρειάζονται τοπικό υλικό. Το Colab προσφέρει πρόσβαση σε Μονάδες Επεξεργασίας Γραφικών (GPU) και Μονάδες Επεξεργασίας Τενσογράφου (Tensor Processing Units), καθιστώντας το πολύτιμο εργαλείο για έργα που απαιτούν υψηλή υπολογιστική ισχύ, όπως η εκπαίδευση σε βαθιά νευρωνικά δίκτυα (deep neural networks). Η πλατφόρμα είναι εφοδιασμένη με κοινές βιβλιοθήκες δεδομένων και μηχανικής μάθησης, όπως οι TensorFlow, PyTorch, Keras και SciKit-Learn, μειώνοντας τον χρόνο εγκατάστασης.

Το Google Colab είναι βασισμένο στα Jupyter Notebooks, μια ευρέως αποδεκτή διαδικτυακή εφαρμογή ανοιχτού κώδικα που επιτρέπει τη δημιουργία και την κοινή χρήση ανοιχτού κώδικα. Αυτό το καθιστά ιδιαίτερα χρήσιμο τόσο για πειραματισμό όσο και σε επαγγελματικό επίπεδο, καθώς ενσωματώνει την εκτέλεση κώδικα με λεπτομερείς επεξηγήσεις, όλα σε μια ενιαία διεπαφή σε έναν φορητού υπολογιστή.

Το Google Colab είναι ιδανικό για επιστήμονες, ερευνητές και εκπαιδευτικούς, καθώς επιτρέπει τη συλλογική εργασία σε ένα σημειωματάριο, τα οποία μπορεί εύκολα να κοινοποιηθεί και να τροποποιηθεί από πολλούς χρήστες. Επιπλέον, το Colab συνδέεται με το Google Drive, επιτρέποντας στους χρήστες να αποθηκεύουν και να έχουν πρόσβαση στην εργασία τους απευθείας από τον χώρο αποθήκευσης στο cloud. Η φιλική προς το χρήστη διεπαφή και οι προεγκατεστημένες βιβλιοθήκες μειώνουν περαιτέρω το εμπόδιο εισόδου για πολύπλοκες υπολογιστικές εργασίες.

Τα παραπάνω κατέστησαν το Google Colab εξαιρετικά αποτελεσματικό εργαλείο για την υλοποίηση της παρούσας διπλωματικής εργασίας.

6. Υλοποίηση

6.1 Συλλογή και Προεπεξεργασία Δεδομένων

Απόκτηση δεδομένων FTIR

Τα φάσματα υπέρυθρων μετασχηματισμού Fourier (FTIR) των δειγμάτων ελαιόλαδου ελήφθησαν χρησιμοποιώντας ένα φασματόμετρο FTIR. Κάθε δείγμα σαρώθηκε πολλές φορές για να εξασφαλιστεί η αξιοπιστία των δεδομένων και το μέσο φάσμα εξετάστηκε για περαιτέρω ανάλυση.

Οργάνωση δεδομένων

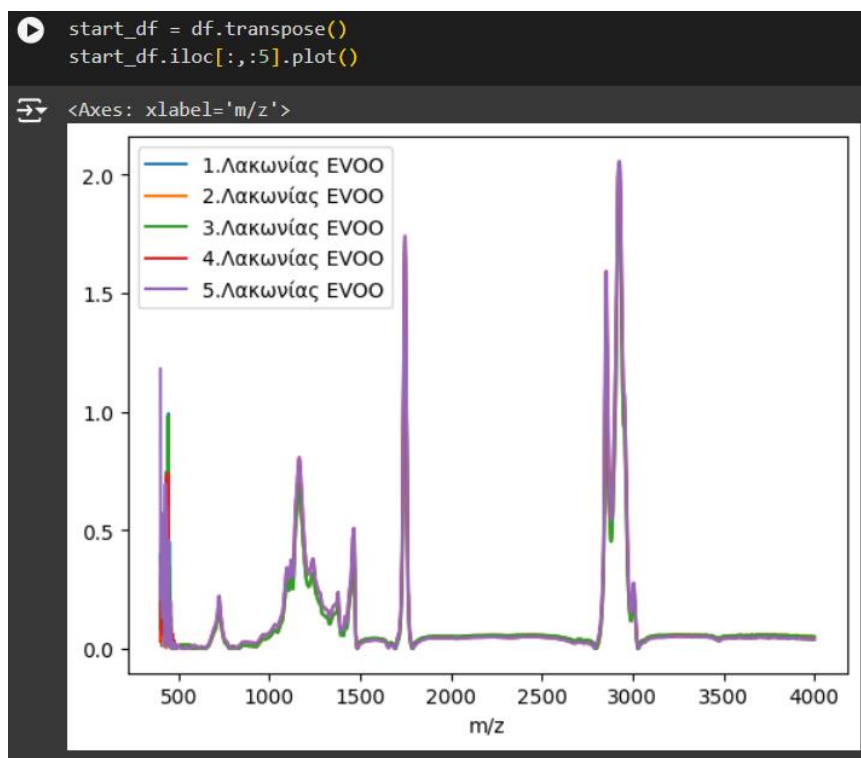
Τα συλλεχθέντα δεδομένα φασμάτων FTIR οργάνωθηκαν σε ένα pandas DataFrame. Οι σειρές είναι ευρετηριασμένες με κυματοαριθμούς (cm^{-1}), οι οποίοι αντιπροσωπεύουν τις συχνότητες των φασμάτων. Κάθε σειρά αντιστοιχεί σε έναν συγκεκριμένο αριθμό κύματος στον οποίο μετρήθηκε η απορρόφηση. Κάθε στήλη αντιπροσωπεύει ένα διαφορετικό δείγμα ελαιόλαδου από διάφορες περιοχές ("Λακωνίας EVOO", "Μεσσηνία EVOO"). Οι τιμές στις στήλες υποδεικνύουν τις ενδείξεις απορρόφησης στους αντίστοιχους κυματοαριθμούς. Αυτή η δομή επιτρέπει τη σύγκριση των φασμάτων απορρόφησης διαφορετικών δειγμάτων ελαιόλαδου.

```
init_df = pd.read_excel("FTIR_OLIVE_OILS.xlsx")
targets_df = pd.read_excel("target.xlsx")
#targets_df
init_df
```

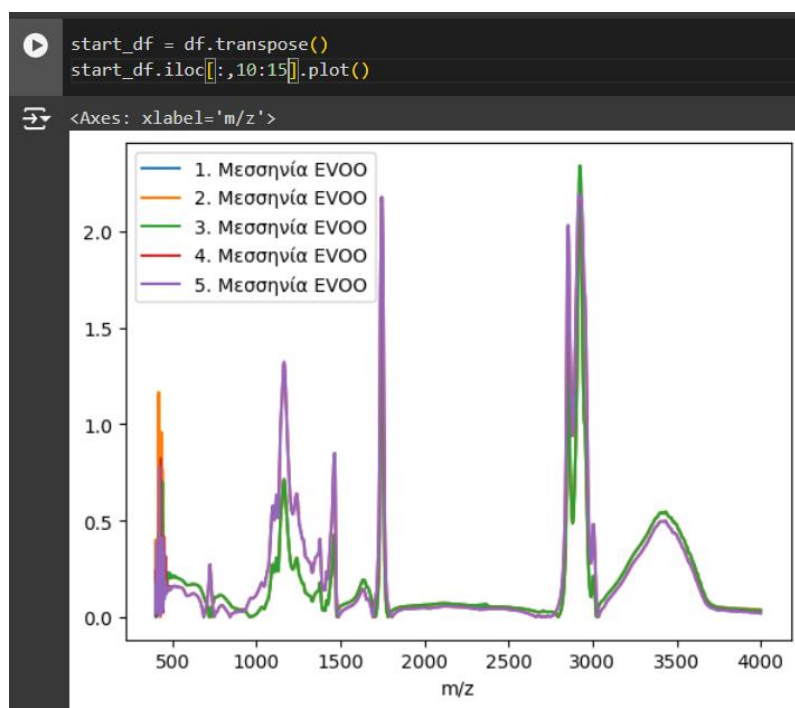
	m/z	1.Λακωνίας EVOO	2.Λακωνίας EVOO	3.Λακωνίας EVOO	4.Λακωνίας EVOO	5.Λακωνίας EVOO	6.Λακωνίας EVOO	7.Λακωνίας EVOO	8.Λακωνίας EVOO	9.Λακωνίας EVOO	...	1. Μεσσηνία EVOO	2. Μεσσηνία EVOO
0	4000	0.050010	0.051018	0.049442	0.037055	0.037733	0.036406	0.035452	0.036132	0.035118	...	0.036247	0.037292
1	3999	0.050219	0.051170	0.049293	0.037166	0.037898	0.036619	0.036033	0.036411	0.035402	...	0.036567	0.037548
2	3998	0.050329	0.051230	0.049105	0.037176	0.038042	0.036795	0.036552	0.036662	0.035631	...	0.036854	0.037778
3	3997	0.050320	0.051226	0.049064	0.036988	0.038089	0.036906	0.036733	0.036876	0.035710	...	0.036976	0.037806
4	3996	0.050269	0.051228	0.049107	0.036817	0.038165	0.036997	0.036709	0.037008	0.035725	...	0.037036	0.037805
...
3596	404	0.161851	0.117876	0.081993	0.269996	0.255273	0.139762	0.188326	0.106455	0.073786	...	0.038458	0.119610
3597	403	0.308020	0.237546	0.168276	0.297900	0.200698	0.162584	0.181503	0.077888	0.020203	...	0.014819	0.276288
3598	402	0.399361	0.331941	0.293851	0.262112	0.220719	0.173823	0.191225	0.465385	0.063983	...	0.000000	0.401701
3599	401	0.347775	0.351399	0.227431	0.109484	0.522756	0.144670	0.233046	0.710519	0.073186	...	0.029426	0.109485
3600	400	0.245388	0.327682	0.080096	0.029707	1.179704	0.096191	0.281287	0.554290	0.026014	...	0.143906	0.090661

3601 rows x 21 columns

Εικόνα 13. Αρχικός πίνακας δεδομένων FTIR ελαιόλαδου



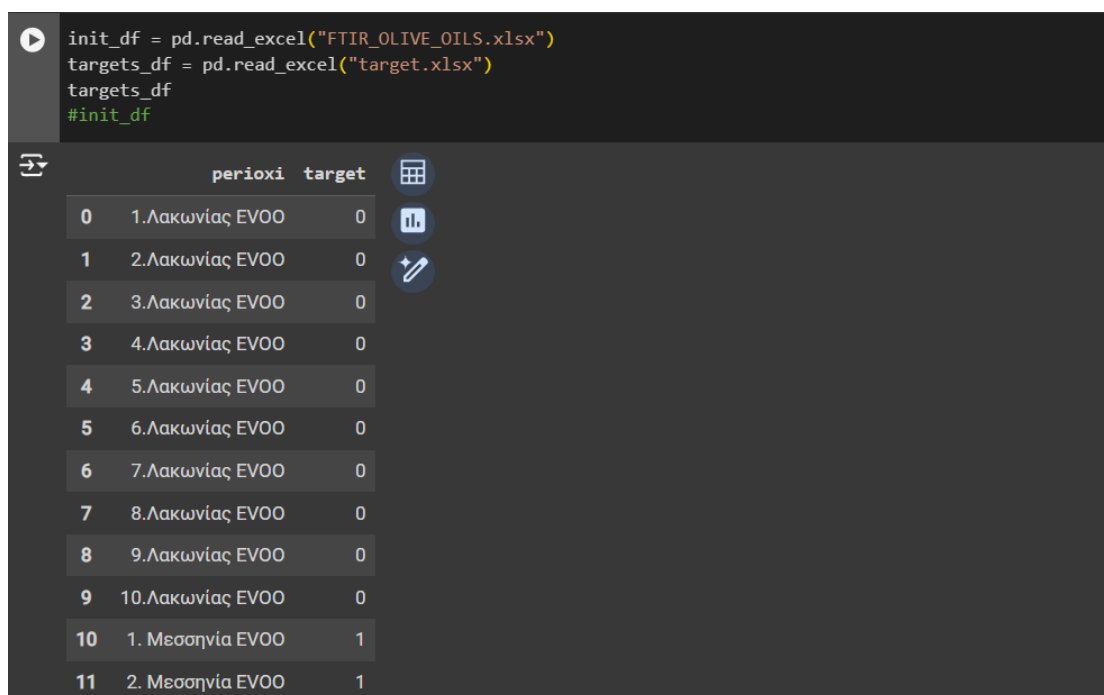
Εικόνα 14. Διάγραμμα φασμάτων FTIR για ελαιόλαδα Λακωνίας



Εικόνα 15. Διάγραμμα φασμάτων FTIR για ελαιόλαδα Μεσσηνίας

Στη συνέχεια δημιουργήθηκε ένας νέος πίνακας (targets_df) με τις τιμές-στόχους (0 για τη Λακωνία και 1 για τη Μεσσηνία), και τα δεδομένα προετοιμάστηκαν για εποπτευόμενη μηχανική εκμάθηση. Αυτός ο πίνακας χρησιμεύει ως ετικέτα για κάθε δείγμα ελαιόλαδου, υποδεικνύοντας την προέλευσή του. Κατά την εκπαίδευση του

μοντέλου, αυτές οι ετικέτες είναι απαραίτητες επειδή επιτρέπουν στο μοντέλο να μάθει την σχέση μεταξύ των δεδομένων FTIR (χαρακτηριστικά) και της προέλευσης (στόχος). Τελικά, αυτό βοηθά στη δημιουργία ενός μοντέλου πρόβλεψης που μπορεί να ταξινομήσει νέα δείγματα με βάση τα δεδομένα FTIR τους.



```
init_df = pd.read_excel("FTIR_OLIVE_OILS.xlsx")
targets_df = pd.read_excel("target.xlsx")
targets_df
#init_df
```

	perioxi	target
0	1.Λακωνίας ΕVOO	0
1	2.Λακωνίας ΕVOO	0
2	3.Λακωνίας ΕVOO	0
3	4.Λακωνίας ΕVOO	0
4	5.Λακωνίας ΕVOO	0
5	6.Λακωνίας ΕVOO	0
6	7.Λακωνίας ΕVOO	0
7	8.Λακωνίας ΕVOO	0
8	9.Λακωνίας ΕVOO	0
9	10.Λακωνίας ΕVOO	0
10	1. Μεσσηνία ΕVOO	1
11	2. Μεσσηνία ΕVOO	1

Εικόνα 16. Νέος πίνακας (targets_df) με τις τιμές-στόχους περιοχών για τα ελαιόλαδα.

Στη μηχανική μάθηση, συχνά γίνεται αντιστροφή στον πίνακα δεδομένων για να βρίσκεται σε μια πιο φιλική προς ανάλυση μορφή. Επομένως, με την αντιστροφή κάθε σειρά αντιπροσωπεύει ένα διαφορετικό δείγμα ελαιόλαδου, καθιστώντας ευκολότερη τη διαχείριση και την ανάλυση μεμονωμένων δειγμάτων, ενώ κάθε στήλη αντιπροσωπεύει ένα συγκεκριμένο χαρακτηριστικό ή μέτρηση, που ευθυγραμμίζεται με τον τρόπο με τον οποίο οι περισσότεροι αλγόριθμοι μηχανικής εκμάθησης αναμένουν δεδομένα εισόδου. Με αυτόν τον τρόπο εξασφαλίζεται συνέπεια στον τρόπο εκπαίδευσης και δοκιμής των μοντέλων, επιτρέποντας την απλή εφαρμογή αλγορίθμων και τον καλύτερο χειρισμό δεδομένων.

df = init_df.sort_values(by='m/z')

df = df.set_index('m/z')

df = df.transpose()

idx = df.index

df

m/z

400

401

402

403

404

405

406

407

408

409

...

3991

3992

1.Λακωνίας EVOO	0.245388	0.347775	0.399361	0.308020	0.161851	0.138374	0.133395	0.016431	0.103134	0.010366	...	0.050347	0.050329
2.Λακωνίας EVOO	0.327682	0.351399	0.331941	0.237546	0.117876	0.121794	0.133395	0.025780	0.145423	0.314348	...	0.051141	0.051228
3.Λακωνίας EVOO	0.080096	0.227431	0.293851	0.168276	0.081993	0.291221	0.417913	0.272547	0.130865	0.114710	...	0.049529	0.049452
4.Λακωνίας EVOO	0.029707	0.109484	0.262112	0.297900	0.269996	0.177952	0.117226	0.173889	0.270541	0.334053	...	0.037274	0.037296
5.Λακωνίας EVOO	1.179704	0.522756	0.220719	0.200698	0.255273	0.411121	0.570545	0.366348	0.215318	0.267823	...	0.038455	0.038594
6.Λακωνίας EVOO	0.096191	0.144670	0.173823	0.162584	0.139762	0.135366	0.124938	0.086220	0.030489	0.043206	...	0.037164	0.037160
7.Λακωνίας EVOO	0.281287	0.233046	0.191225	0.181503	0.188326	0.213624	0.244125	0.248644	0.270878	0.405536	...	0.036337	0.036162
8.Λακωνίας EVOO	0.554290	0.710519	0.465385	0.077888	0.106455	0.073992	0.031518	0.140177	0.255273	0.297213	...	0.036579	0.036559
9.Λακωνίας EVOO	0.026014	0.073186	0.063983	0.020203	0.073786	0.001270	0.136677	0.315231	0.485721	0.368094	...	0.035763	0.035749
10.Λακωνίας EVOO	0.099588	0.115432	0.094158	0.037357	0.027522	0.028439	0.030877	0.101269	0.220571	0.498239	...	0.037661	0.037764
1.Μεσσηνίας EVOO	0.143906	0.029426	0.000000	0.014819	0.038458	0.002693	0.053029	0.057508	0.046856	0.056421	...	0.037297	0.037142
2.Μεσσηνίας EVOO	0.090661	0.109485	0.401701	0.276288	0.119610	0.122582	0.154190	0.156412	0.166332	0.221379	...	0.038309	0.038287
3.Μεσσηνίας EVOO	0.015795	0.020748	0.123062	0.189729	0.212089	0.105898	0.053029	0.068653	0.052389	0.118781	...	0.035141	0.035119
4.Μεσσηνίας EVOO	0.026872	0.099505	0.245330	0.206340	0.093905	0.050233	0.131278	0.061089	0.047642	0.023961	...	0.021973	0.021854
5.Μεσσηνίας EVOO	0.166751	0.083968	0.018683	0.028789	0.138303	0.311144	0.385851	0.184587	0.263242	0.229983	...	0.020032	0.019861
6.Μεσσηνίας EVOO	0.166948	0.213028	0.237361	0.189251	0.093905	0.064607	0.155752	0.073716	0.005677	0.182671	...	0.017448	0.017035
7.Μεσσηνίας EVOO	0.096910	0.296708	0.424540	0.107940	0.329425	0.264855	0.109875	0.009427	0.005080	0.194142	...	0.047818	0.047835
8.Μεσσηνίας EVOO	0.134700	0.257683	0.370854	0.122365	0.044582	0.054244	0.000000	0.172465	0.378197	0.185845	...	0.046815	0.046743
9.Μεσσηνίας EVOO	0.096910	0.153868	0.642595	0.049683	0.187673	0.021761	0.404572	0.603035	0.443138	0.070740	...	0.045899	0.045852
10.Μεσσηνίας EVOO	0.276207	0.965352	0.753330	0.358872	0.613959	0.489691	0.176091	0.009438	0.039064	0.029293	...	0.026715	0.026426

Εικόνα 17. Πίνακας δεδομένων FTIR ελαιόλαδου με αντιστροφή (transpose)

Επεξεργασία δεδομένων

Όπως έχει αναφερθεί σε προηγούμενο κεφάλαιο, το binning είναι ένα βήμα προεπεξεργασίας στην ανάλυση δεδομένων και τη μηχανική μάθηση. Στο πλαίσιο της ανάλυσης φασμάτων FTIR για το ελαιόλαδο, η μεθοδολογία binning βοηθά στην εξομάλυνση των φασματικών δεδομένων.

Τα συνεχή δεδομένα ομαδοποιήθηκαν σε διακριτά διαστήματα ή "κάδους". Αυτή η διαδικασία μειώνει την πολυπλοκότητα και τον θόρυβο στα δεδομένα, καθιστώντας ευκολότερη την ανάλυση και τη μοντελοποίηση. Αντί να εργάζεται με χιλιάδες μεμονωμένες μετρήσεις κυμάτων, το binning υπολογίζει τους μέσους όρους αυτών των μετρήσεων σε συγκεκριμένες περιοχές.

Τα βήματα που ακολουθήθηκαν είναι:

- **Συγκέντρωση:** Τα συνεχή σημεία των δεδομένων συγκεντρώνονται σε καθορισμένα εύρη, αντικαθιστώντας αυτές τις τιμές με μια μεμονωμένη τιμή με όπως ο μέσος όρος ή η διάμεσος.

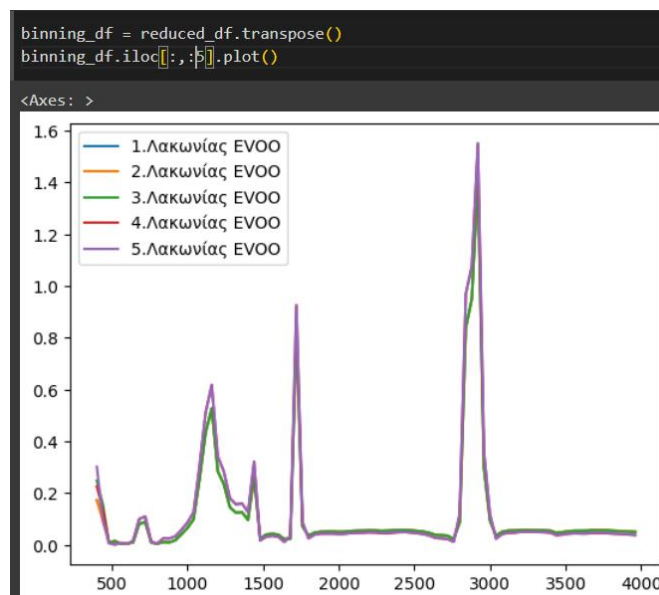
- **Μείωση διαστάσεων:** Με τη μείωση του αριθμού των σημείων, το binning απλοποιεί το σύνολο δεδομένων, το οποίο συμβάλλει στη βελτίωση της απόδοσης και της ερμηνείας των μοντέλων μηχανικής εκμάθησης.

Έτσι, διασφαλίζεται ότι τα δεδομένα που χρησιμοποιούνται για τη μηχανική εκμάθηση είναι πιο καθαρά, πιο διαχειρίσιμα και κατάλληλα για την παραγωγή αξιόπιστων και στιβαρών μοντέλων.

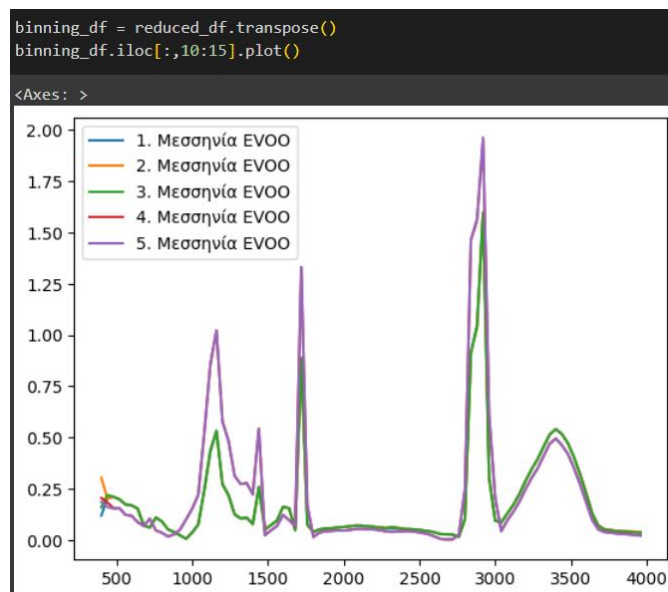
```
reduced_df = pd.DataFrame() #binning
b = 40
for i in range(400, 4000, b):
    ds = pd.Series(np.zeros(20), index=df.index)
    for j in range(b):
        ds = ds + df[i + j]
    reduced_df[i] = ds / b
reduced_df
```

	400	440	480	520	560	600	640	680	720	760	...	3600	3640	3680
1.Λακωνίας EVOO	0.171543	0.123592	0.008420	0.015683	0.005220	0.005195	0.010428	0.081935	0.089282	0.009088	...	0.055164	0.055531	0.057
2.Λακωνίας EVOO	0.174132	0.091530	0.010363	0.015648	0.005292	0.005332	0.010038	0.081404	0.088917	0.009068	...	0.055932	0.056058	0.058
3.Λακωνίας EVOO	0.248551	0.157272	0.009584	0.015440	0.005200	0.005216	0.009915	0.081198	0.088689	0.009051	...	0.054963	0.055121	0.057
4.Λακωνίας EVOO	0.226326	0.136374	0.008779	0.001622	0.008925	0.003736	0.016913	0.100625	0.110681	0.012563	...	0.044208	0.044218	0.046
5.Λακωνίας EVOO	0.302370	0.097273	0.006863	0.001693	0.008798	0.003384	0.016350	0.100001	0.110036	0.012373	...	0.045125	0.045090	0.047
6.Λακωνίας EVOO	0.176426	0.117240	0.012404	0.001789	0.008762	0.003488	0.016370	0.099961	0.109972	0.012317	...	0.044592	0.044436	0.047
7.Λακωνίας EVOO	0.368751	0.088404	0.009245	0.006233	0.024780	0.020305	0.037609	0.161932	0.176975	0.032210	...	0.044088	0.044043	0.047
8.Λακωνίας EVOO	0.348319	0.067717	0.010597	0.006572	0.024793	0.020265	0.036992	0.161347	0.176484	0.032039	...	0.044788	0.044639	0.047
9.Λακωνίας EVOO	0.223123	0.044175	0.010237	0.006166	0.024678	0.020146	0.036848	0.161107	0.176250	0.031998	...	0.044083	0.043898	0.047
10.Λακωνίας EVOO	0.175216	0.100736	0.006962	0.008421	0.006197	0.002509	0.012721	0.089806	0.098325	0.009750	...	0.046384	0.045900	0.048
1.Μεσσηνία EVOO	0.119526	0.203836	0.211400	0.198570	0.173351	0.169029	0.152715	0.069974	0.061162	0.109793	...	0.230847	0.132097	0.072
2.Μεσσηνία EVOO	0.304356	0.202677	0.210215	0.198646	0.173718	0.169469	0.154312	0.071220	0.061368	0.110209	...	0.231363	0.132333	0.072
3.Μεσσηνία EVOO	0.160189	0.220439	0.211397	0.199188	0.173850	0.169605	0.154634	0.071617	0.061448	0.110357	...	0.229579	0.130312	0.070

Εικόνα 18 Εφαρμογή binning στα δεδομένα FTIR ελαιολάδου



Εικόνα 19. Διάγραμμα φασμάτων FTIR ελαιόλαδου Λακωνίας μετά από εφαρμογή binning



Εικόνα 20 Διάγραμμα φασμάτων FTIR ελαιόλαδου Μεσσηνίας μετά από εφαρμογή binning

Αντίστοιχα, η **Principal Component Analysis (PCA)** τεχνική χρησιμοποιήθηκε για να μειωθεί η διάσταση του πίνακα. Τα αρχικά χαρακτηριστικά μετατράπηκαν σε ένα μικρότερο σύνολο μη συσχετισμένων στοιχείων, καταγράφοντας τη μεγαλύτερη απόκλιση στα δεδομένα.

```
pca = PCA(n_components=20) #συμπίεση ποσοτήτας δεδομένων με χρήση αλγορίθμου PCA
pca.fit(df)
pca_df = pd.DataFrame(pca.transform(df))
pca_df = pca_df.set_index(idx)
pca_df
```

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1.Λακωνίας EVOO	5.813614	-1.624776	-0.582810	0.405429	-0.254704	0.077716	-0.107428	-0.699066	0.019534	0.537032	0.009215	0.025760	-0.359992	-0.135292	0.300348
2.Λακωνίας EVOO	5.785569	-1.618963	-0.490950	0.168921	-0.480022	0.185703	-0.090342	0.032900	-0.179932	-0.366203	-0.221943	0.276345	-0.281746	-0.026966	-0.122527
3.Λακωνίας EVOO	5.868145	-1.635227	-0.603913	-0.332776	0.723741	-0.372969	0.065792	-0.071322	0.250864	0.564958	0.162719	0.466243	0.225859	0.039003	-0.204412
4.Λακωνίας EVOO	4.910128	0.249714	-0.267730	-0.250477	0.539635	-0.309784	-0.215746	-0.106239	0.280413	-0.255558	-0.296944	-0.443105	0.196881	0.395003	0.285213
5.Λακωνίας EVOO	4.926169	0.279646	-0.359776	-0.144073	0.556793	0.891096	-0.407155	0.815763	-0.240061	0.051477	-0.281267	-0.046465	-0.104401	-0.088326	0.014078
6.Λακωνίας EVOO	4.889717	0.248394	-0.393220	-0.590364	-0.165486	-0.218784	0.004164	0.091884	0.347754	-0.410904	0.507885	-0.103019	-0.018459	-0.020928	-0.271454
7.Λακωνίας EVOO	1.363696	6.114770	1.386541	0.747842	0.458715	-0.106803	-0.642316	-0.095990	-0.341491	-0.019523	0.511378	-0.092500	0.156041	-0.151302	0.024670
8.Λακωνίας EVOO	1.362625	6.039649	1.107317	0.105214	-0.066760	0.638737	1.029225	0.052504	0.055329	0.266513	0.060890	0.004421	-0.092908	0.294959	-0.055906
9.Λακωνίας EVOO	1.324505	6.081418	1.060073	0.036293	-0.749923	-0.571403	-0.202931	0.163931	0.540145	-0.063859	-0.475341	0.164075	-0.049156	-0.124440	0.009259
10.Λακωνίας EVOO	5.616398	-1.028602	-0.730414	0.295785	-0.571360	-0.099121	0.577296	-0.163426	-0.522515	-0.332162	0.045185	-0.257846	0.352800	-0.164272	-0.010035
1. Μεσσηνία EVOO	-1.825583	-5.400043	0.974673	-0.307825	-0.756270	-0.035521	-0.388843	0.331377	-0.086240	0.257780	0.182182	0.015553	-0.030000	0.006152	-0.033917

Εικόνα 21 Πίνακας δεδομένων FTIR ελαιόλαδου με εφαρμογή της PCA

Εκπαίδευση και η αξιολόγηση

Η εκπαίδευση και η αξιολόγηση μοντέλων είναι θεμελιώδη βήματα στη διαδικασία μηχανικής μάθησης. Η εκπαίδευση περιλαμβάνει τη διδασκαλία του μοντέλου να αναγνωρίζει μοτίβα στα δεδομένα, ενώ η αξιολόγηση αξιολογεί την απόδοση και τη γενίκευση του μοντέλου.

Η εκπαίδευση του μοντέλου έγινε, με τον διαχωρισμό των δεδομένων σε σύνολα εκπαίδευσης και δοκιμών. Συγκεκριμένα τα δεδομένα χωρίστηκαν σε 80% σετ εκπαίδευσης, που χρησιμοποιήθηκε για την εκμάθηση του μοντέλου και το υπόλοιπο 20% χρησιμοποιήθηκε για το σετ δοκιμών δηλαδή για την αξιολόγησή του. Κάθε φορά ορίζεται σαν X διαφορετικός τύπος δεδομένων όπως:

- Τα αρχικά δεδομένα που διαθέτουμε ($X = df$)
- Δεδομένα επεξεργασμένα με μέθοδο Binning ($X = reduced_df$)
- Δεδομένα επεξεργασμένα με Principal Component Analysis (PCA) ($X = pca_df$)

Σαν y ορίστηκε ο αρχικός πίνακας με όνομα targets ώστε να ολοκληρωθεί η μέθοδος Split.

```
# X = df # X = είναι με εισοδο τα αρχικα δεδομενα
X = reduced_df # X = είναι με εισοδο τα επεξεργασμενα δεδομενα μετα το binning
# X = pca_df # X = είναι με εισοδο τα επεξεργασμενα δεδομενα απο τον PCA
y = targets_df['target']

#split method
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
```

```
#apply all the algorithm
lda = LDA()
dt = DecisionTreeClassifier()
lr = LogisticRegression()
nb = GaussianNB()
svm = SVC()
rf = RandomForestClassifier()
```

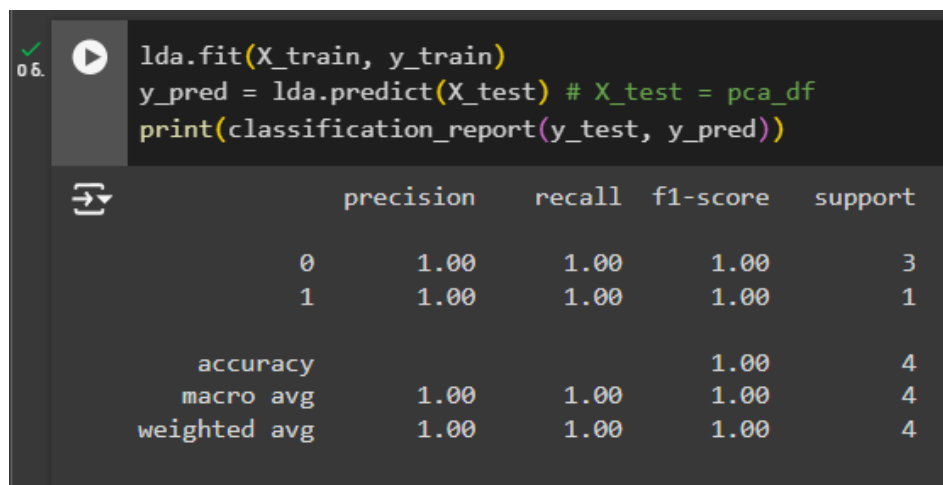
Εικόνα 22 Απεικόνιση εφαρμογής ταξινόμητων

6.2 Ταξινόμηση με τη μέθοδο διαχωρισμού

6.2.1 Linear Discriminant Analysis

Το LDA όπως έχει αναφερθεί παραπάνω, είναι μια πολύ γνωστή στατιστική μέθοδος που χρησιμοποιείται για τη μείωση και ταξινόμηση διαστάσεων. Υποθέτει ότι τα δεδομένα σε κάθε τάξη ακολουθούν μια κατανομή Gauss και στοχεύει να βρει έναν γραμμικό συνδυασμό χαρακτηριστικών που διαχωρίζει καλύτερα τις κλάσεις.

Μετά την εφαρμογή της τεχνικής διαχωρισμού (Split), όπου το σύνολο δεδομένων χωρίστηκε σε σύνολα εκπαίδευσης και δοκιμών, χρησιμοποιήθηκε η Γραμμική Διακριτική Ανάλυση (Linear Discriminant Analysis) για την αξιολόγηση της απόδοσης της ταξινόμησης. Συγκεκριμένα, η υλοποίηση του LDA εισήχθη από τη βιβλιοθήκη “from sklearn.discriminant_analysis import LinearDiscriminantAnalysis as LDA”. Κατά τη φάση της εκπαίδευσης, το μοντέλο προσαρμόστηκε στα δεδομένα εκπαίδευσης («X_train» και «y_train») και η απόδοσή του αξιολογήθηκε χρησιμοποιώντας το σύνολο δοκιμών («X_test») για αρχή στην εικόνα 1 δέχεται σαν είσοδο τα δεδομένα όπου έχουν υποστεί επεξεργασία Principal Component Analysis και στην εικόνα 2 δέχεται τα δεδομένα με την τεχνική Binning. Η αναφορά ταξινόμησης έδειξε ακρίβεια 100% και για τις δύο κατηγορίες, αντικατοπτρίζοντας τέλεια ταξινόμηση των δεδομένων δοκιμής. Αυτό το αποτέλεσμα καταδεικνύει την υψηλή αποτελεσματικότητα του LDA όταν εφαρμόζεται στο συγκεκριμένο σύνολο δεδομένων, καθώς και το πλεονέκτημα της χρήσης της τεχνικής split training για την επικύρωση αυτού του μοντέλου.



```
lda.fit(X_train, y_train)
y_pred = lda.predict(X_test) # X_test = pca_df
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	1.00	1.00	1.00	3
1	1.00	1.00	1.00	1
accuracy			1.00	4
macro avg	1.00	1.00	1.00	4
weighted avg	1.00	1.00	1.00	4

Εικόνα 23 Απεικόνιση LDA με δεδομένα εισόδου x = PCA

06.

```
lda.fit(X_train, y_train)
y_pred = lda.predict(X_test) # X_test = binning
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	1.00	1.00	1.00	2
1	1.00	1.00	1.00	2
accuracy			1.00	4
macro avg	1.00	1.00	1.00	4
weighted avg	1.00	1.00	1.00	4

Εικόνα 24 Απεικόνιση LDA με δεδομένα εισόδου x = binning

6.2.2 Decision Tree Classifier

Στο συγκεκριμένο στάδιο της ανάλυσης δεδομένων, έγινε χρήση του αλγορίθμου, ο οποίος εφαρμόστηκε μέσω της βιβλιοθήκης “sklearn.tree.DecisionTreeClassifier”. Ο αλγόριθμος αυτός ανήκει στην κατηγορία των αλγορίθμων επιβλεπόμενης μάθησης και χρησιμοποιεί μια δενδροειδή δομή για την κατάταξη των Decision Tree Classifier δεδομένων. Η διαδικασία της εκπαίδευσης πραγματοποιήθηκε στο σύνολο των δεδομένων (X_train και y_train), έχοντας χωριστεί με την μέθοδο split 80% - 20%. Η ακρίβεια του μοντέλου ανήλθε στο 100%, και στις δυο διαφορετικές εισόδους δεδομένων που τέθηκαν κάθε φορά. Αυτό υποδηλώνει την άριστη απόδοση του αλγορίθμου, τουλάχιστον για τα δεδομένα που χρησιμοποιήθηκαν για δοκιμή, με το μοντέλο να καταφέρνει να διαχωρίσει πλήρως τις κατηγορίες. Η χρήση του Decision Tree Classifier αποδείχθηκε ιδιαίτερα αποτελεσματική για την παρούσα ανάλυση, καθώς ο αλγόριθμος είναι γνωστός για την ευκολία ερμηνείας και την ικανότητά του να συλλαμβάνει μη γραμμικές σχέσεις μεταξύ των χαρακτηριστικών.

06.	▶	<pre>dt.fit(X_train, y_train) y_pred = dt.predict(X_test) # X_test = pca_df print(classification_report(y_test, y_pred))</pre>			
↔		precision	recall	f1-score	support
	0	1.00	1.00	1.00	3
	1	1.00	1.00	1.00	1
	accuracy			1.00	4
	macro avg	1.00	1.00	1.00	4
	weighted avg	1.00	1.00	1.00	4

Εικόνα 25. Απεικόνιση *Decision Tree Classifier* με δεδομένα εισόδου $x = \text{PCA}$

06.	▶	<pre>dt.fit(X_train, y_train) y_pred = dt.predict(X_test) # X_test = binning print(classification_report(y_test, y_pred))</pre>			
↔		precision	recall	f1-score	support
	0	1.00	1.00	1.00	2
	1	1.00	1.00	1.00	2
	accuracy			1.00	4
	macro avg	1.00	1.00	1.00	4
	weighted avg	1.00	1.00	1.00	4

Εικόνα 26. Απεικόνιση *Decision Tree Classifier* με δεδομένα εισόδου $x = \text{binning}$

6.2.3 Logistic Regression

Σε αυτό το στάδιο της ανάλυσης χρησιμοποιήθηκε ο αλγόριθμος Logistic Regression μέσω της βιβλιοθήκης “`sklearn.linear_model.LogisticRegression`”. Είναι ένας αλγόριθμος γραμμικής ταξινόμησης, ο οποίος χρησιμοποιείται κυρίως για δυαδική κατάταξη, αν και μπορεί να προσαρμοστεί και για περισσότερες κατηγορίες. Τα αποτελέσματα της ταξινόμησης και στις δυο περιπτώσεις έδειξαν ακρίβεια κατά 100% γεγονός που υποδεικνύει την τέλεια απόδοση του μοντέλου. Η μέθοδος αποδείχθηκε αποτελεσματική για την παρούσα ανάλυση, καθώς τα αποτελέσματα επιβεβαιώνουν τη σωστή γραμμική διάκριση των κατηγοριών.

06.

```
lr.fit(X_train, y_train)
y_pred = lr.predict(X_test) # X_test = pca_df
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1
1	1.00	1.00	1.00	3
accuracy			1.00	4
macro avg	1.00	1.00	1.00	4
weighted avg	1.00	1.00	1.00	4

Εικόνα 27 Απεικόνιση *Logistic Regression* με δεδομένα εισόδου $x = \text{PCA}$

06.


```
lr.fit(X_train, y_train)
y_pred = lr.predict(X_test) # X_test = binning
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	1.00	1.00	1.00	3
1	1.00	1.00	1.00	1
accuracy			1.00	4
macro avg	1.00	1.00	1.00	4
weighted avg	1.00	1.00	1.00	4

Εικόνα 28 Απεικόνιση *Logistic Regression* με δεδομένα εισόδου $x = \text{binning}$


6.2.4 Gaussian Navie Bayes

Στο συγκεκριμένο στάδιο της ανάλυσης, χρησιμοποιήθηκε ο αλγόριθμος Gaussian Naive Bayes μέσω της βιβλιοθήκης “`sklearn.naive_bayes.GaussianNB`”. Ο Gaussian Naive Bayes βασίζεται στην εφαρμογή του Θεωρήματος του Bayes και υποθέτει ότι τα χαρακτηριστικά των δεδομένων ακολουθούν την κανονική (Gaussian) κατανομή. Αυτός ο αλγόριθμος χρησιμοποιήθηκε κυρίως για γρήγορη και αποτελεσματική ταξινόμηση, ειδικά όταν τα χαρακτηριστικά είναι ανεξάρτητα μεταξύ τους. Ο αλγόριθμος έδειξε ποσοστό επιτυχίας 100%, κάτι που υποδεικνύει την άριστη απόδοση του μοντέλου.

06.  `nb.fit(X_train, y_train)`
`y_pred = nb.predict(X_test) # X_test = pca_df`
`print(classification_report(y_test, y_pred))`

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1
1	1.00	1.00	1.00	3
accuracy			1.00	4
macro avg	1.00	1.00	1.00	4
weighted avg	1.00	1.00	1.00	4

Εικόνα 29 Απεικόνιση *Gaussian Navie Bayes* με δεδομένα εισόδου $x = \text{PCA}$

06.  `nb.fit(X_train, y_train)`
`y_pred = nb.predict(X_test) # X_test = binning`
`print(classification_report(y_test, y_pred))`

	precision	recall	f1-score	support
0	1.00	1.00	1.00	2
1	1.00	1.00	1.00	2
accuracy			1.00	4
macro avg	1.00	1.00	1.00	4
weighted avg	1.00	1.00	1.00	4

Εικόνα 30 Απεικόνιση *Gaussian Navie Bayes* με δεδομένα εισόδου $x = \text{binning}$

6.2.5 Support Vector machine

Ο αλγόριθμος Support Vector Machine (SVM) είναι ένας ισχυρός αλγόριθμος επιβλεπόμενης μάθησης που χρησιμοποιείται για ταξινόμηση και παλινδρόμηση. Στη παρούσα διπλωματική εργασία εφαρμόστηκε μέσω της βιβλιοθήκης “`sklearn.svm import SVC`”. Δημιουργήθηκε ένα υπερεπίπεδο (hyperplane) που διαχωρίζει τα δεδομένα σε διαφορετικές κατηγορίες με τον βέλτιστο τρόπο. Η αναφορά ταξινόμησης έδειξε ότι το μοντέλο πέτυχε με ακρίβεια 100% την πρόβλεψη, για όλες τις κατηγορίες που του εισάχθηκαν. Η ικανότητα του SVM να διαχωρίζει γραμμικά και μη γραμμικά δεδομένα τον καθιστά ικανό ειδικά σε προβλήματα με υψηλή διάσταση χαρακτηριστικών.

06.

```
svm.fit(X_train, y_train)
y_pred = svm.predict(X_test) # X_test = pca_df
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1
1	1.00	1.00	1.00	3
accuracy			1.00	4
macro avg	1.00	1.00	1.00	4
weighted avg	1.00	1.00	1.00	4

Εικόνα 31 Απεικόνιση *Support Vector machine* με δεδομένα εισόδου $x = \text{PCA}$

06.


```
svm.fit(X_train, y_train)
y_pred = svm.predict(X_test) # X_test = binning
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	1.00	1.00	1.00	2
1	1.00	1.00	1.00	2
accuracy			1.00	4
macro avg	1.00	1.00	1.00	4
weighted avg	1.00	1.00	1.00	4

Εικόνα 32 Απεικόνιση *Support Vector machine* με δεδομένα εισόδου $x = \text{binning}$


6.2.6 Random Forest

Στην εφαρμογή του αλγόριθμου Random Forest Classifier μέσω της βιβλιοθήκης “`sklearn.ensemble.RandomForestClassifier`”. Είναι ένας αλγόριθμος συνόλου, ο οποίος συνδυάζει πολλαπλά δέντρα απόφασης για να δημιουργήσει ένα πιο ισχυρό και ακριβές μοντέλο ταξινόμησης. Το μοντέλο εκπαιδεύτηκε χρησιμοποιώντας τα δεδομένα εκπαίδευσης (X_{train} και y_{train}). Τα αποτελέσματα ήταν και εδώ 100% επιτυχής. Έτσι, ο Random Forest είναι ιδιαίτερα αποτελεσματικός αλγόριθμος, κυρίως λόγω της δυνατότητάς του να μειώνει την υπερπροσαρμογή μέσα από τη χρήση πολλών δέντρων.

06.  `rf.fit(X_train, y_train)`
`y_pred = rf.predict(X_test) # X_test = pca_df`
`print(classification_report(y_test, y_pred))`

	precision	recall	f1-score	support
0	1.00	1.00	1.00	3
1	1.00	1.00	1.00	1
accuracy			1.00	4
macro avg	1.00	1.00	1.00	4
weighted avg	1.00	1.00	1.00	4

Εικόνα 33 Απεικόνιση *Random Forest Classifier* με δεδομένα εισόδου $x = \text{PCA}$

06.  `rf.fit(X_train, y_train)`
`y_pred = rf.predict(X_test) # X_test = binning`
`print(classification_report(y_test, y_pred))`

	precision	recall	f1-score	support
0	1.00	1.00	1.00	3
1	1.00	1.00	1.00	1
accuracy			1.00	4
macro avg	1.00	1.00	1.00	4
weighted avg	1.00	1.00	1.00	4

Εικόνα 34 Απεικόνιση *Random Forest Classifier* με δεδομένα εισόδου $x = \text{binning}$

6.3 Ταξινόμηση με τη μέθοδο διασταυρούμενη επικύρωσης

6.3.1 Linear Discriminant Analysis

Χρησιμοποιώντας τη μέθοδο “cross_validate” πραγματοποιήθηκε λοιπόν, διασταυρούμενη επικύρωση με διάφορες μετρήσεις. Με βάση τον πίνακα της εικόνας, όλες οι μετρήσεις είναι σταθερά στο 100% στις 6 πτυχές διασταυρούμενης επικύρωσης στα δεδομένα τα οποία έχουν επεξεργαστεί με την μέθοδο Binning. Αυτό υποδηλώνει ότι το μοντέλο αποδίδει τέλεια στο συγκεκριμένο υπό αξιολόγηση σύνολο δεδομένων.

```
[32] k = 6
     cl_metrics = ('precision_weighted', 'recall_weighted', 'f1_weighted', 'accuracy')

[33] #start cross-validation method
     cv_results = cross_validate(lda, X, y, cv=k, scoring=cl_metrics) # X = binning
     results_df = pd.DataFrame(data={
         'precision_weighted': cv_results['test_precision_weighted'],
         'recall_weighted': cv_results['test_recall_weighted'],
         'f1-score_weighted': cv_results['test_f1_weighted'],
         'accuracy': cv_results['test_accuracy'],
     })
     results_df
```

	precision_weighted	recall_weighted	f1-score_weighted	accuracy
0	1.0	1.0	1.0	1.0
1	1.0	1.0	1.0	1.0
2	1.0	1.0	1.0	1.0
3	1.0	1.0	1.0	1.0
4	1.0	1.0	1.0	1.0
5	1.0	1.0	1.0	1.0

Εικόνα 35 Απεικόνιση *Linear Discriminant Analysis* με δεδομένα εισόδου $x = \text{binning}$

Μετά την εφαρμογή της τεχνικής Cross Validation, ο αλγόριθμος Linear Discriminant Analysis (LDA) χρησιμοποίησε τα δεδομένα που έχουν επεξεργαστεί με Principal Component Analysis (PCA). Για ακρίβεια, ανάκληση και βαθμολογία F1, το μοντέλο πέτυχε γενικά υψηλές τιμές, ειδικά σε ορισμένες πτυχές όπου όλες οι μετρήσεις έφτασαν το 1.000, υποδεικνύοντας τέλεια ταξινόμηση σε αυτές τις περιπτώσεις. Ωστόσο, σε ορισμένες πτυχές, το μοντέλο «δυσκολεύτηκε», όπως φαίνεται από τις χαμηλότερες τιμές αυτών των μετρήσεων, ιδιαίτερα για τη βαθμολογία και την ανάκληση F1, η οποία έπεσε σημαντικά, όπως στο 0,166667 και στο 0,333333 σε μία φορά. Συνολικά, η μέτρηση ακρίβειας, η οποία μετρά τη συνολική ορθότητα των προβλέψεων του μοντέλου, παρέμεινε σχετικά σταθερή, αλλά παρουσίασε επίσης κάποια διακύμανση μεταξύ 0,333333 και 0,750000 σε διαφορετικές πτυχές. Αυτό δείχνει ότι, ενώ το μοντέλο αποδίδει καλά σε ορισμένα υποσύνολα δεδομένων, υπάρχουν παραλλαγές στην απόδοσή του.

```
[146] k = 6
      cl_metrics = ('precision_weighted', 'recall_weighted', 'f1_weighted', 'accuracy')

#start cross-validation method
cv_results = cross_validate(lda, X, y, cv=k, scoring=cl_metrics) # X = PCA
results_df = pd.DataFrame(data={
    'precision_weighted': cv_results['test_precision_weighted'],
    'recall_weighted': cv_results['test_recall_weighted'],
    'f1-score_weighted': cv_results['test_f1_weighted'],
    'accuracy': cv_results['test_accuracy'],
})
results_df
```

/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1531: UndefinedWarning: The average of the results is not defined because the number of non-zero results is less than the number of results. (metric='precision_weighted', modifier='average', f1-score_weighted, recall_weighted, accuracy)

	precision_weighted	recall_weighted	f1-score_weighted	accuracy
0	0.833333	0.750000	0.733333	0.750000
1	0.833333	0.750000	0.733333	0.750000
2	1.000000	1.000000	1.000000	1.000000
3	0.111111	0.333333	0.166667	0.333333
4	1.000000	1.000000	1.000000	1.000000
5	1.000000	1.000000	1.000000	1.000000

Εικόνα 36 Απεικόνιση *Linear Discriminant Analysis* με δεδομένα εισόδου $x = \text{PCA}$

6.3.2 Decision Tree Classifier

Σε αυτήν την ανάλυση, ένας ταξινομητής Decision Tree (DT) αξιολογήθηκε χρησιμοποιώντας την τεχνική Cross Validation, συγκρίνοντας τα αποτελέσματα δύο μεθόδων προεπεξεργασίας. Όταν εφαρμόστηκαν τα δεδομένα με PCA, ο αλγόριθμος έδειξε μεταβλητότητα στην απόδοση. Ενώ ορισμένες επανάληψες πέτυχαν τέλεια αποτελέσματα άλλες παρουσίασαν πτώση. Αυτή η μεταβλητότητα υποδηλώνει ότι αν και ο PCA μείωσε αποτελεσματικά τη διάσταση του συνόλου δεδομένων, μπορεί να είχε ως αποτέλεσμα την απώλεια σημαντικών διακριτικών χαρακτηριστικών, επηρεάζοντας την απόδοση του ταξινομητή σε ορισμένες επαναλήψεις της επικύρωσης.

Αντίθετα, χρησιμοποιώντας τη μέθοδο binning, ο Decision Tree πέτυχε τέλεια αποτελέσματα σε όλες τις επανάληψες επικύρωσης, με όλες τις μετρήσεις να φτάνουν σταθερά στο 1,0. Αυτό δείχνει ότι το binning διατήρησε τις βασικές πληροφορίες που απαιτούνται για τη σωστή ταξινόμηση όλων των περιπτώσεων από το Decision Tree χωρίς λανθασμένες ταξινομήσεις. Η ομοιόμορφη απόδοση σε όλες τις πτυχές υποδηλώνει υψηλό επίπεδο γενίκευσης από το μοντέλο όταν εκπαιδεύεται με χαρακτηριστικά που μετασχηματίζονται μέσω binning.

```

cv_results = cross_validate(dt, X, y, cv=k, scoring=cl_metrics)
results_df = pd.DataFrame(data={
    'precision_weighted': cv_results['test_precision_weighted'], # X = binning
    'recall_weighted': cv_results['test_recall_weighted'],
    'f1-score_weighted': cv_results['test_f1_weighted'],
    'accuracy': cv_results['test_accuracy'],
})
results_df

```

	precision_weighted	recall_weighted	f1-score_weighted	accuracy
0	1.0	1.0	1.0	1.0
1	1.0	1.0	1.0	1.0
2	1.0	1.0	1.0	1.0
3	1.0	1.0	1.0	1.0
4	1.0	1.0	1.0	1.0
5	1.0	1.0	1.0	1.0

Εικόνα 37 Απεικόνιση *Decision Tree* με δεδομένα εισόδου $x = \text{binning}$

```

cv_results = cross_validate(dt, X, y, cv=k, scoring=cl_metrics)
results_df = pd.DataFrame(data={
    'precision_weighted': cv_results['test_precision_weighted'], # X = PCA
    'recall_weighted': cv_results['test_recall_weighted'],
    'f1-score_weighted': cv_results['test_f1_weighted'],
    'accuracy': cv_results['test_accuracy'],
})
results_df

```

	precision_weighted	recall_weighted	f1-score_weighted	accuracy
0	1.000000	1.000000	1.000000	1.000000
1	0.833333	0.750000	0.733333	0.750000
2	1.000000	1.000000	1.000000	1.000000
3	0.833333	0.666667	0.666667	0.666667
4	1.000000	1.000000	1.000000	1.000000
5	1.000000	1.000000	1.000000	1.000000

Εικόνα 38 Απεικόνιση *Decision Tree* με δεδομένα εισόδου $x = \text{PCA}$

6.3.3 Logistic Regression

Στη διαδικασία αξιολόγησης του μοντέλου με τη χρήση της τεχνικής Cross Validation, εφαρμόστηκαν δύο διαφορετικές μέθοδοι προεπεξεργασίας: binning και PCA. Στην πρώτη περίπτωση, όπου το binning χρησιμοποιήθηκε ως μέθοδος μετασχηματισμού χαρακτηριστικών, το μοντέλο πέτυχε βέλτιστα αποτελέσματα σε όλες τις μετρήσεις, με κάθε πτυχή της διασταυρούμενης επικύρωσης να επιστρέφει τέλεια βαθμολογία 100%. Αυτό υποδηλώνει ότι τα χαρακτηριστικά που εξήχθησαν μέσω του binning ήταν εξαιρετικά αποτελεσματικά στην καταγραφή της διακύμανσης εντός του συνόλου δεδομένων, οδηγώντας σε εξαιρετικά ακριβή ταξινόμηση χωρίς να παρατηρούνται λανθασμένες ταξινομήσεις. Η ομοιομορφία των αποτελεσμάτων σε πολλαπλές πτυχές υποδηλώνει επίσης ότι το μοντέλο γενικεύεται καλά στο σύνολο δεδομένων όταν περιορίζεται στα κύρια στοιχεία.

Αντίθετα, η δεύτερη περίπτωση, όπου το PCA χρησιμοποιήθηκε ως τεχνική μείωσης διαστάσεων, απέδωσε σημαντικά χαμηλότερη απόδοση. Η ακρίβεια, η ανάκληση, η βαθμολογία F1 και η ακρίβεια μειώθηκαν σημαντικά, με τιμές που κυμαίνονται μεταξύ 0,11 και 0,67 σε διαφορετικές πτυχές. Αυτά τα αποτελέσματα υποδηλώνουν ότι η διάκριση συνεχών χαρακτηριστικών μέσω του PCA είχε ως αποτέλεσμα την απώλεια σημαντικών πληροφοριών που είναι απαραίτητες για ακριβή ταξινόμηση. Οι κυμαινόμενες επιδόσεις σε διαφορετικές πτυχές διασταυρούμενης επικύρωσης υποδηλώνουν έλλειψη σταθερότητας και γενίκευσης του μοντέλου σε αυτήν τη μέθοδο προεπεξεργασίας.

```

cv_results = cross_validate(lr, X, y, cv=k, scoring=cl_metrics)
results_df = pd.DataFrame(data={
    'precision_weighted': cv_results['test_precision_weighted'], # X = binning
    'recall_weighted': cv_results['test_recall_weighted'],
    'f1-score_weighted': cv_results['test_f1_weighted'],
    'accuracy': cv_results['test_accuracy'],
})
results_df

```

	precision_weighted	recall_weighted	f1-score_weighted	accuracy
0	1.0	1.0	1.0	1.0
1	1.0	1.0	1.0	1.0
2	1.0	1.0	1.0	1.0
3	1.0	1.0	1.0	1.0
4	1.0	1.0	1.0	1.0
5	1.0	1.0	1.0	1.0

Εικόνα 39 Απεικόνιση *Logistic Regression* με δεδομένα εισόδου $x = \text{binning}$

```

cv_results = cross_validate(lr, X, y, cv=k, scoring=cl_metrics)
results_df = pd.DataFrame(data={
    'precision_weighted': cv_results['test_precision_weighted'], # X = PCA
    'recall_weighted': cv_results['test_recall_weighted'],
    'f1-score_weighted': cv_results['test_f1_weighted'],
    'accuracy': cv_results['test_accuracy'],
})
results_df

```

	precision_weighted	recall_weighted	f1-score_weighted	accuracy
0	0.250000	0.500000	0.333333	0.500000
1	0.250000	0.500000	0.333333	0.500000
2	0.111111	0.333333	0.166667	0.333333
3	0.111111	0.333333	0.166667	0.333333
4	0.444444	0.666667	0.533333	0.666667
5	0.444444	0.666667	0.533333	0.666667

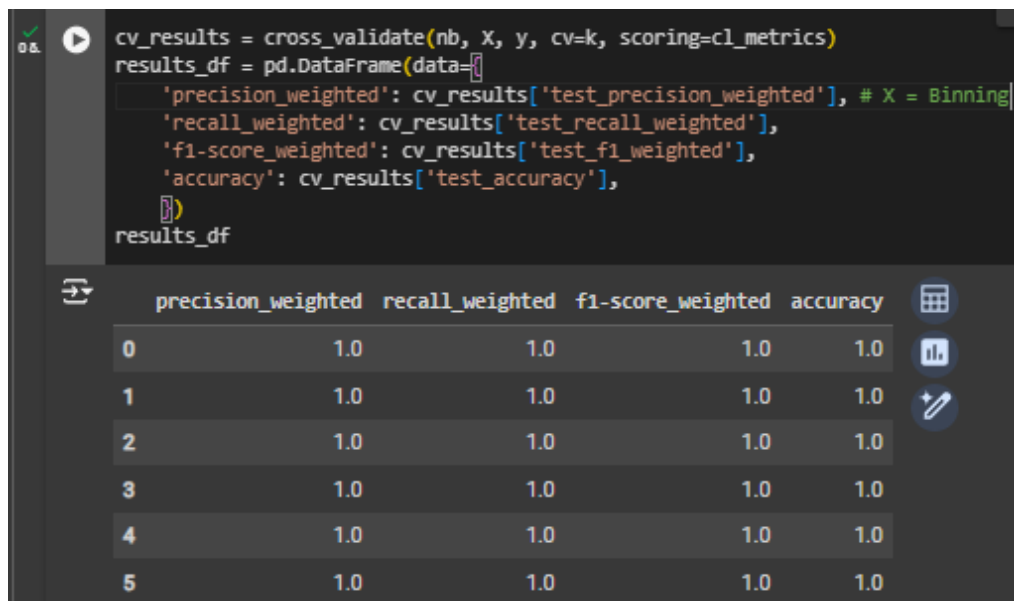
Εικόνα 40 Απεικόνιση *Logistic Regression* με δεδομένα εισόδου $x = \text{PCA}$

6.3.4 Gaussian Navie Bayes

Ενας ταξινομητής Naive Bayes (NB) αξιολογήθηκε χρησιμοποιώντας την τεχνική Cross Validation, με δύο διαφορετικές μεθόδους προεπεξεργασίας.

Για το πρώτο μοντέλο, όπου το binning εφαρμόστηκε ως τεχνική μετασχηματισμού χαρακτηριστικών, τα αποτελέσματα ήταν εξαιρετικά, με τις τιμές να φτάνουν σταθερά το 100% σε όλες τις πτυχές διασταυρούμενης επικύρωσης. Αυτό υποδηλώνει ότι το binning ήταν μια εξαιρετικά αποτελεσματική μέθοδος προεπεξεργασίας σε αυτή τη συγκεκριμένη περίπτωση, συλλαμβάνοντας επαρκείς πληροφορίες ώστε ο ταξινομητής Naive Bayes να προβλέψει τέλεια το αποτέλεσμα σε όλα τα σύνολα δοκιμών.

Στη δεύτερη περίπτωση, το PCA χρησιμοποιήθηκε για τη μείωση διαστάσεων και ενώ η απόδοση ήταν ακόμα ισχυρή, εμφάνισε κάποια μεταβλητότητα στην διασταυρούμενη επικύρωση. Ενώ η πλειονότητα των βαθμολογιών ήταν αρκετά καλές, ορισμένες πτυχές είδαν την ακρίβεια να πέφτει στο 0,25 και να ανακαλείται στο 0,5, με αποτέλεσμα τη συνολική βαθμολογία F1 να είναι 0,33. Η ακρίβεια σε αυτές τις πτυχές ήταν επίσης χαμηλότερη στο 50%. Αυτό υποδεικνύει ότι ενώ το PCA μείωσε τη διάσταση, ο μετασχηματισμός μπορεί να είχε οδηγήσει σε απώλεια κρίσιμων πληροφοριών για ορισμένα υποσύνολα δεδομένων, με αποτέλεσμα χαμηλότερη προγνωστική απόδοση για αυτές τις συγκεκριμένες πτυχές.



```
cv_results = cross_validate(nb, x, y, cv=k, scoring=cl_metrics)
results_df = pd.DataFrame(data={
    'precision_weighted': cv_results['test_precision_weighted'], # X = Binning
    'recall_weighted': cv_results['test_recall_weighted'],
    'f1-score_weighted': cv_results['test_f1_weighted'],
    'accuracy': cv_results['test_accuracy'],
})
results_df
```

	precision_weighted	recall_weighted	f1-score_weighted	accuracy
0	1.0	1.0	1.0	1.0
1	1.0	1.0	1.0	1.0
2	1.0	1.0	1.0	1.0
3	1.0	1.0	1.0	1.0
4	1.0	1.0	1.0	1.0
5	1.0	1.0	1.0	1.0

Εικόνα 41 Απεικόνιση Naive Bayes με δεδομένα εισόδου x = binning

```

cv_results = cross_validate(nb, X, y, cv=k, scoring=cl_metrics)
results_df = pd.DataFrame(data={
    'precision_weighted': cv_results['test_precision_weighted'], # X = PCA
    'recall_weighted': cv_results['test_recall_weighted'],
    'f1-score_weighted': cv_results['test_f1_weighted'],
    'accuracy': cv_results['test_accuracy'],
})
results_df

```

	precision_weighted	recall_weighted	f1-score_weighted	accuracy
0	0.250000	0.500000	0.333333	0.500000
1	1.000000	1.000000	1.000000	1.000000
2	1.000000	1.000000	1.000000	1.000000
3	1.000000	1.000000	1.000000	1.000000
4	0.444444	0.666667	0.533333	0.666667
5	0.444444	0.666667	0.533333	0.666667

Εικόνα 42 Απεικόνιση *Naive Bayes* με δεδομένα εισόδου $x = \text{PCA}$

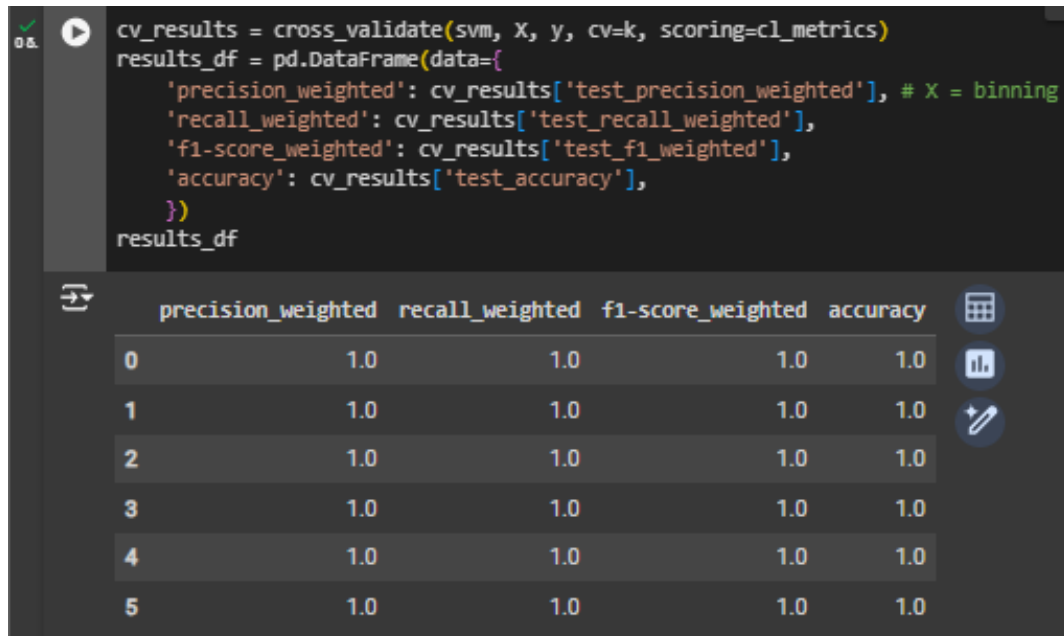
6.3.5 Support Vector machine

Ο Support Vector Machine (SVM) αξιολογήθηκε χρησιμοποιώντας την τεχνική Cross Validation με δύο μεθόδους προεπεξεργασίας: binning και Principal Component Analysis (PCA). Οι μετρήσεις απόδοσης, όπως η ακρίβεια, η ανάκληση, η βαθμολογία F1 και η ακρίβεια, αξιολογήθηκαν για να μετρηθεί η αποτελεσματικότητα αυτών των προσεγγίσεων προεπεξεργασίας.

Στο πρώτο σενάριο, όπου χρησιμοποιήθηκε το binning, το SVM πέτυχε άσογα αποτελέσματα, με όλες τις μετρήσεις απόδοσης να έχουν βαθμολογία 100% σε όλες τις πτυχές διασταυρούμενης επικύρωσης. Αυτό δείχνει ότι το binning διατήρησε αρκετές πληροφορίες ώστε το SVM να ταξινομεί τέλεια όλες τις παρουσίες, με αποτέλεσμα μηδενικά σφάλματα κατά τη δοκιμή.

Αντίθετα, στο δεύτερο μέρος, όπου εφαρμόστηκε PCA, είχε ως αποτέλεσμα σημαντική πτώση της απόδοσης. Ενώ μία από τις πτυχές διασταυρούμενης επικύρωσης διατήρησε τέλεια απόδοση (1,0 για όλες τις μετρήσεις), άλλες πτυχές επέδειξαν πολύ χαμηλότερη ακρίβεια, ανάκληση, βαθμολογία F1 και ακρίβεια. Για παράδειγμα, σε ορισμένες πτυχές, η ακρίβεια έπεσε στο 0,25, η ανάκληση στο 0,5 και η συνολική βαθμολογία F1 στο 0,33, με την ακρίβεια να ακολουθεί παρόμοια πτωτική τάση. Αυτά

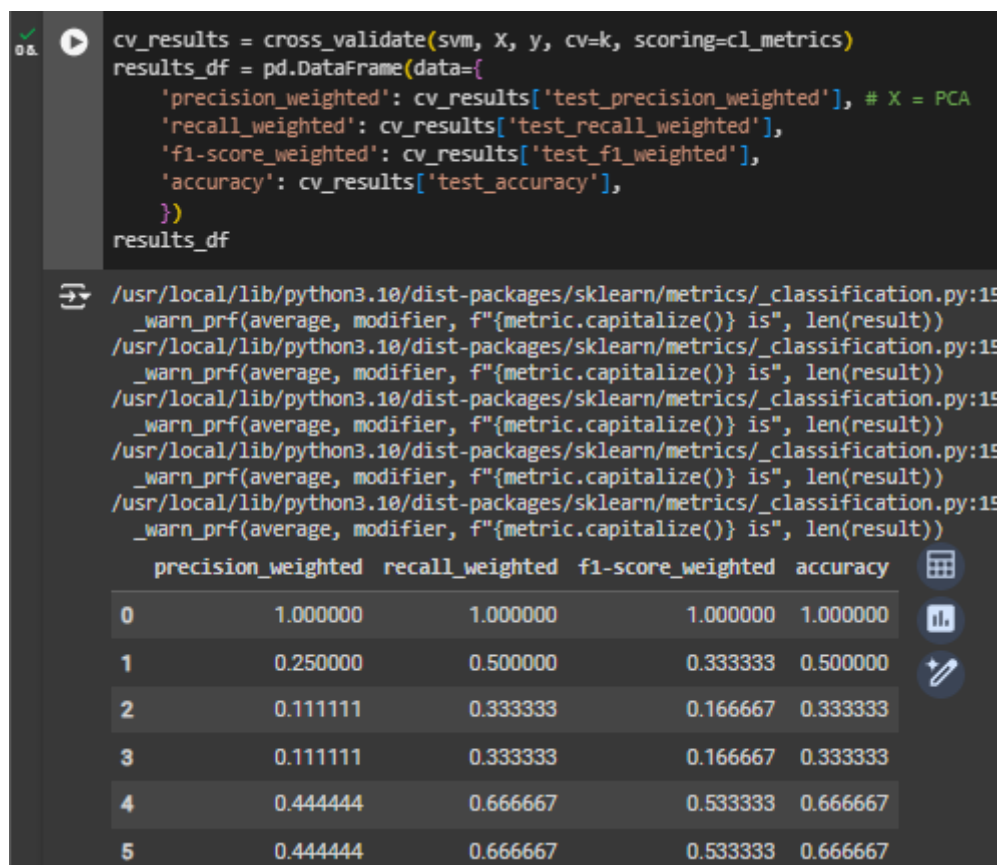
τα αποτελέσματα υποδηλώνουν ότι η μείωση της διάστασης μέσω του PCA μπορεί να έχει οδηγήσει σε απώλεια κρίσιμων πληροφοριών που είναι απαραίτητες για την ακριβή ταξινόμηση των δεδομένων από το SVM.



```
cv_results = cross_validate(svm, X, y, cv=k, scoring=cl_metrics)
results_df = pd.DataFrame(data={
    'precision_weighted': cv_results['test_precision_weighted'], # X = binning
    'recall_weighted': cv_results['test_recall_weighted'],
    'f1-score_weighted': cv_results['test_f1_weighted'],
    'accuracy': cv_results['test_accuracy'],
})
results_df
```

	precision_weighted	recall_weighted	f1-score_weighted	accuracy
0	1.0	1.0	1.0	1.0
1	1.0	1.0	1.0	1.0
2	1.0	1.0	1.0	1.0
3	1.0	1.0	1.0	1.0
4	1.0	1.0	1.0	1.0
5	1.0	1.0	1.0	1.0

Εικόνα 43 Απεικόνιση *Support Vector Machine* με δεδομένα εισόδου $x = \text{binning}$



```
cv_results = cross_validate(svm, X, y, cv=k, scoring=cl_metrics)
results_df = pd.DataFrame(data={
    'precision_weighted': cv_results['test_precision_weighted'], # X = PCA
    'recall_weighted': cv_results['test_recall_weighted'],
    'f1-score_weighted': cv_results['test_f1_weighted'],
    'accuracy': cv_results['test_accuracy'],
})
results_df
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:15
_warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:15
_warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:15
_warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:15
_warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:15
_warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
```

	precision_weighted	recall_weighted	f1-score_weighted	accuracy
0	1.000000	1.000000	1.000000	1.000000
1	0.250000	0.500000	0.333333	0.500000
2	0.111111	0.333333	0.166667	0.333333
3	0.111111	0.333333	0.166667	0.333333
4	0.444444	0.666667	0.533333	0.666667
5	0.444444	0.666667	0.533333	0.666667

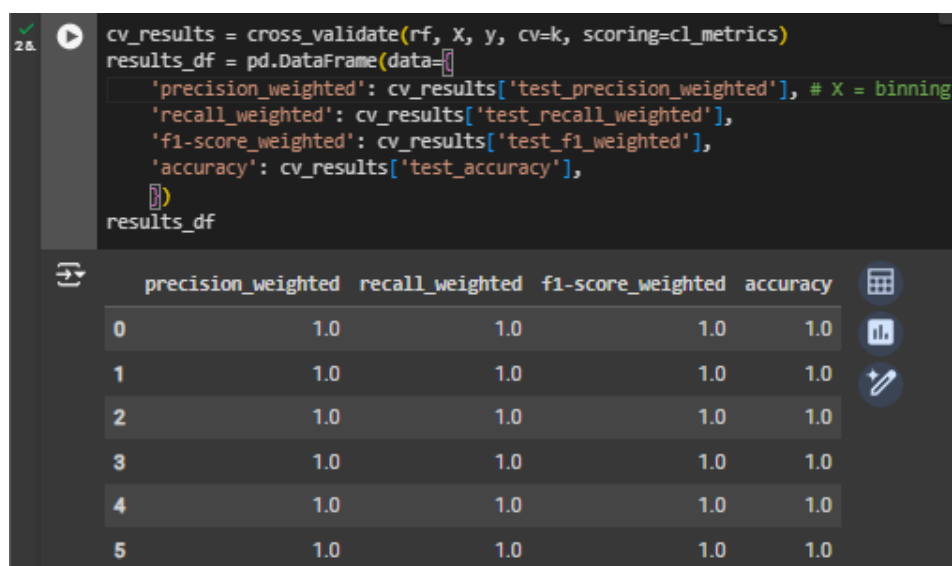
Εικόνα 44 Απεικόνιση *Support Vector Machine* με δεδομένα εισόδου $x = \text{PCA}$

6.3.6 Random Forest

Ενας ταξινομητής Random Forest (RF) αξιολογήθηκε χρησιμοποιώντας την τεχνική Cross Validation, με δύο διαφορετικές μεθόδους. Ο στόχος ήταν να αξιολογηθεί η επίδραση αυτών των τεχνικών προεπεξεργασίας στην απόδοση ταξινόμησης του μοντέλου μέσω κάποιων μετρήσεων.

Στην πρώτη περίπτωση, το binning εφαρμόστηκε ως τεχνική μετασχηματισμού χαρακτηριστικών και το μοντέλο πέτυχε τέλεια αποτελέσματα, βαθμολογώντας με ακρίβεια 100%. Αυτό υποδηλώνει ότι το binning διατήρησε αποτελεσματικά τις σχετικές πληροφορίες στο σύνολο δεδομένων.

Αντίθετα, η χρήση του PCA οδήγησε σε πτυχές που πέτυχαν υψηλή ακρίβεια, ανάκληση, βαθμολογίες F1 και ακρίβεια (1,0 σε ορισμένες περιπτώσεις), άλλες είδαν πτώση της απόδοσης, με μετρήσεις όπως η ακρίβεια και η ανάκληση να μειώνονται έως και 0,11 και 0,33, αντίστοιχα. Αυτή η μεταβλητότητα υποδεικνύει ότι η PCA, αν και επιτυχής στη μείωση της διαστάσεων, μπορεί να έχει οδηγήσει σε απώλεια κρίσιμων πληροφοριών που είναι απαραίτητες για ακριβή ταξινόμηση. Ως αποτέλεσμα, το μοντέλο δυσκολεύτηκε να διατηρήσει σταθερή απόδοση σε όλα τα σύνολα επικύρωσης, υπογραμμίζοντας την αντιστάθμιση μεταξύ μείωσης διαστάσεων και ακρίβειας ταξινόμησης.



```
cv_results = cross_validate(rf, X, y, cv=k, scoring=cl_metrics)
results_df = pd.DataFrame(data=[
    'precision_weighted': cv_results['test_precision_weighted'], # X = binning
    'recall_weighted': cv_results['test_recall_weighted'],
    'f1-score_weighted': cv_results['test_f1_weighted'],
    'accuracy': cv_results['test_accuracy'],
])
results_df
```

	precision_weighted	recall_weighted	f1-score_weighted	accuracy
0	1.0	1.0	1.0	1.0
1	1.0	1.0	1.0	1.0
2	1.0	1.0	1.0	1.0
3	1.0	1.0	1.0	1.0
4	1.0	1.0	1.0	1.0
5	1.0	1.0	1.0	1.0

Εικόνα 45 Απεικόνιση Random Forest με δεδομένα εισόδου x = binning

```
cv_results = cross_validate(rf, X, y, cv=k, scoring=cl_metrics)
results_df = pd.DataFrame(data={
    'precision_weighted': cv_results['test_precision_weighted'], # X = PCA
    'recall_weighted': cv_results['test_recall_weighted'],
    'f1-score_weighted': cv_results['test_f1_weighted'],
    'accuracy': cv_results['test_accuracy'],
})
results_df
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1
_warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1
_warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1
_warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
```

	precision_weighted	recall_weighted	f1-score_weighted	accuracy
0	0.833333	0.750000	0.733333	0.750000
1	1.000000	1.000000	1.000000	1.000000
2	0.833333	0.666667	0.666667	0.666667
3	0.111111	0.333333	0.166667	0.333333
4	0.444444	0.666667	0.533333	0.666667
5	0.444444	0.666667	0.533333	0.666667

Εικόνα 46 Απεικόνιση *Random Forest* με δεδομένα εισόδου $x = \text{PCA}$

Εν κατακλείδι, η σύγκριση των διαφορετικών αλγορίθμων (Linear Discriminant Analysis, Naive Bayes, Support Vector Machines, Random Forest, και Decision Tree) σε συνδυασμό με τις δύο μεθόδους προεπεξεργασίας (binning και Principal Component Analysis) ανέδειξε την επίδραση των τεχνικών αυτών στην απόδοση των μοντέλων. Συγκεκριμένα, η χρήση binning οδήγησε σε σταθερά άριστες αποδόσεις για όλα τα μοντέλα, με απόλυτες τιμές ακριβείας, ανάκλησης, F1-score και ακρίβειας στο 100%, γεγονός που υποδηλώνει ότι η μέθοδος αυτή διατηρεί επαρκώς τις κρίσιμες πληροφορίες για την ταξινόμηση. Αντίθετα, η PCA παρουσίασε αστάθεια σε ορισμένα μοντέλα, οδηγώντας σε πτώση της απόδοσης, κυρίως σε περιπτώσεις όπως οι SVM και LDA. Συνολικά, η επιλογή της κατάλληλης μεθόδου προεπεξεργασίας αποδεικνύεται καθοριστική για την ακρίβεια των αλγορίθμων και την ικανότητά τους να γενικεύουν τα αποτελέσματά τους σε διαφορετικά σύνολα δεδομένων.

6.4 Ομαδοποίηση

6.4.1 K-Means

Στην παρούσα ανάλυση χρησιμοποιήθηκε ο αλγόριθμος K-Means για τη δημιουργία δύο ομάδων (clusters) με βάση τα δεδομένα των δειγμάτων ελαιόλαδου. Ο αλγόριθμος εφαρμόστηκε με την παράμετρο $n_clusters=2$, η οποία αντιστοιχούσε στις δύο γνωστές κατηγορίες: δειγμάτων από Λακωνία ($target=0$) και δείγματα από Μεσσηνία ($target=1$).

Σύμφωνα με τα αποτελέσματα του ο K-Means κατάφερε να διαχωρίσει σωστά τα δείγματα ελαιόλαδου με βάση τις πραγματικές κατηγορίες τους, χωρίς σφάλματα στην ταξινόμηση. Αυτό δείχνει ότι ο συγκεκριμένος αλγόριθμος είναι ιδιαίτερα αποτελεσματικός για το συγκεκριμένο σετ δεδομένων.

Η καλή απόδοση του αλγορίθμου στην ομαδοποίηση των δειγμάτων δείχνει ότι οι δύο κατηγορίες (Λακωνίας και Μεσσηνίας) έχουν διακριτά χαρακτηριστικά, επιτρέποντας στον αλγόριθμο να τις διαχωρίσει με μεγάλη ακρίβεια. Ο **K-Means** είναι ένας αποδοτικός αλγόριθμος για ομαδοποιήσεις αυτού του είδους, όπου τα δεδομένα είναι σχετικά απλά και οι ομάδες είναι σαφώς οριοθετημένες.

6.4.2 DBSCAN

Ο αλγόριθμος DBSCAN (Density-Based Spatial Clustering of Applications with Noise) για την ομαδοποίηση των δειγμάτων ελαιόλαδου. Οι παράμετροι που χρησιμοποιήθηκαν ήταν $min_samples=3$ και $eps=1.1$, με στόχο τον εντοπισμό πυκνωτικών περιοχών στα δεδομένα.

Ο DBSCAN κατάφερε να αναγνωρίσει τη δομή των δεδομένων και να τα ταξινομήσει σε δύο σαφείς ομάδες, χωρίς να δημιουργήσει θόρυβο ή να απορρίψει δεδομένα που βρίσκονται εκτός ομάδας (outliers). Αυτή η επιτυχής ομαδοποίηση δείχνει ότι η επιλογή των παραμέτρων ήταν κατάλληλη για τα συγκεκριμένα δεδομένα, επιτρέποντας στον αλγόριθμο να διακρίνει τις διαφορές στις πυκνότητες των δειγμάτων από τις διαφορετικές γεωγραφικές περιοχές.

```

clusters_df = targets_df

kmeans = KMeans(n_clusters=2)
clusters_df['kmeans'] = kmeans.fit_predict(X)

agglomerative = AgglomerativeClustering(n_clusters=2, linkage='average')
clusters_df['agglomerative'] = agglomerative.fit_predict(X)

dbscan = DBSCAN(min_samples=3, eps=1.1)
clusters_df['dbscan'] = dbscan.fit_predict(X)

print(clusters_df)

```

	perioxi	target	kmeans	agglomerative	dbscan
0	1. Λακωνίας	EV00	0	0	0
1	2. Λακωνίας	EV00	0	0	0
2	3. Λακωνίας	EV00	0	0	0
3	4. Λακωνίας	EV00	0	0	0
4	5. Λακωνίας	EV00	0	0	0
5	6. Λακωνίας	EV00	0	0	0
6	7. Λακωνίας	EV00	0	0	0
7	8. Λακωνίας	EV00	0	0	0
8	9. Λακωνίας	EV00	0	0	0
9	10. Λακωνίας	EV00	0	0	0
10	1. Μεσσηνία	EV00	1	1	1
11	2. Μεσσηνία	EV00	1	1	1
12	3. Μεσσηνία	EV00	1	1	1
13	4. Μεσσηνία	EV00	1	1	1
14	5. Μεσσηνία	EV00	1	1	1
15	6. Μεσσηνία	EV00	1	1	1
16	7. Μεσσηνία	EV00	1	1	1
17	8. Μεσσηνία	EV00	1	1	1
18	9. Μεσσηνία	EV00	1	1	1
19	10. Μεσσηνία	EV00	1	1	1

Εικόνα 47 Απεικόνιση αλγορίθμων ομαδοποίησης *K-Means*, *Agglomerative Method*, *DBSCAN*

6.4.3 Agglomerative Method

Στην παραπάνω εικόνα παρουσιάζονται και τα αποτελέσματα της μεθόδου Agglomerative Method για την ταξινόμηση των δειγμάτων. Ο αλγόριθμος, με τον οποίο έγινε η ομαδοποίηση σε δύο κύριες ομάδες (clusters), παρουσίασε παρόμοια αποτελέσματα με τον K-Means, με τα περισσότερα δείγματα να ταξινομούνται σωστά σε σχέση με την πραγματική ετικέτα τους.

- Όλα τα δείγματα από τη Λακωνία (target=0) ταξινομήθηκαν στο cluster 0.
- Όλα τα δείγματα από τη Μεσσηνία (target=1) ταξινομήθηκαν στο cluster 1.

Αυτό δείχνει ότι ο αλγόριθμος μπόρεσε να διαχωρίσει σωστά τα δεδομένα, δημιουργώντας σαφείς ομάδες με βάση τα χαρακτηριστικά των δειγμάτων.

Στα δεδομένα εφαρμόστηκε ξανά ο αλγόριθμος Agglomerative Method χρησιμοποιώντας τη μέθοδο σύνδεσης "average" για την ομαδοποίηση των δειγμάτων.

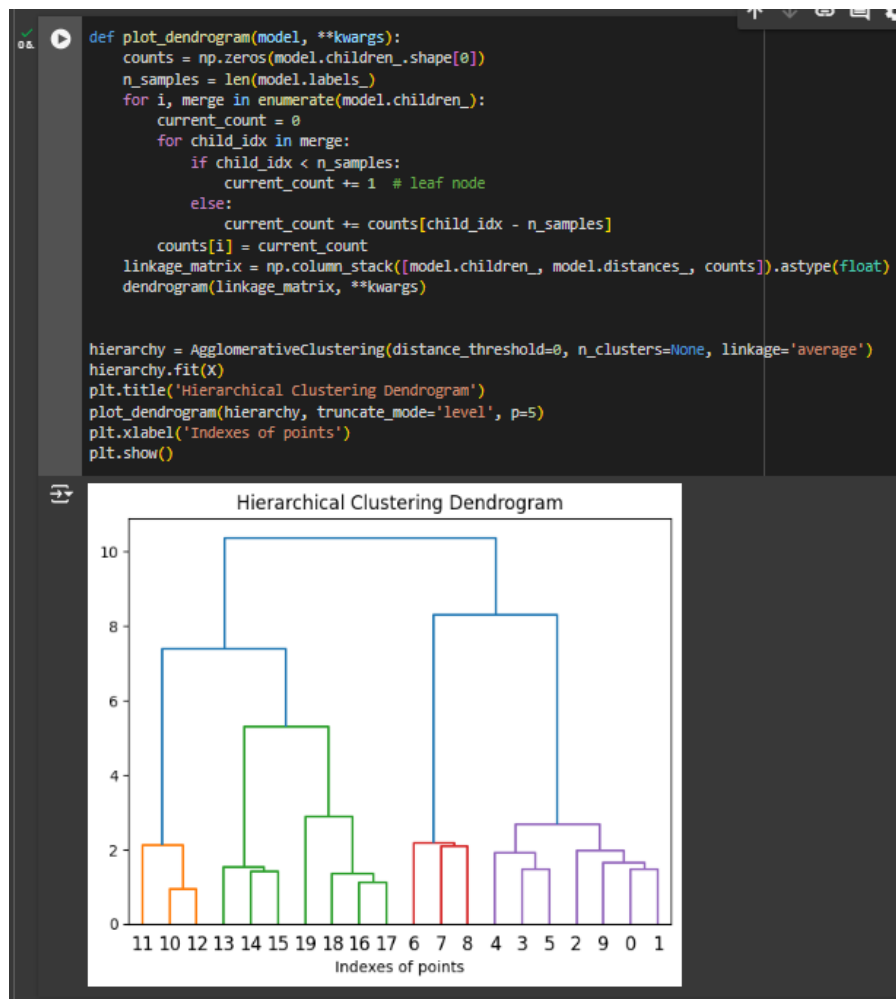
Στην συνέχεια απεικονίζεται η εικόνα με το δενδρόγραμμα που προέκυψε όπου

απεικονίζει τη διαδικασία συγχώνευσης των σημείων, δείχνοντας με ποιον τρόπο τα δείγματα συγχωνεύονται σε ομάδες (clusters) καθώς προχωρά η διαδικασία.

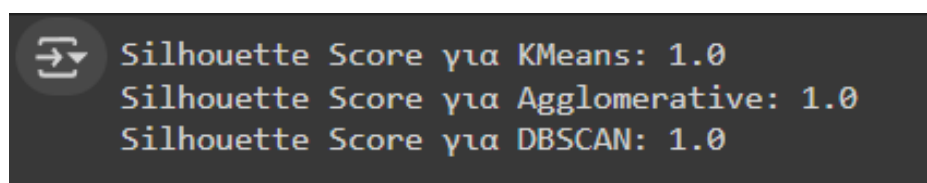
Σύμφωνα με το δενδρόγραμμα αρχικά παρατηρούμε δύο κύριες ομάδες (clusters) οι οποίες διαχωρίζονται στη κορυφή του δένδρου. Κάθε κύρια ομάδα περιλαμβάνει υπό-ομάδες (sub-clusters) συγκεκριμένα:

- Στην αριστερή ομάδα (πράσινες και πορτοκαλί γραμμές), παρατηρείται ότι τα δείγματα 11, 10, 12 ανήκουν στην υπό ομάδα πορτοκαλί και 13, 14, 15, 19, 18, 16 και 17 σχηματίζουν την πράσινη υπό ομάδα υποδηλώνοντας ότι τα συγκεκριμένα δείγματα ταιριάζουν περισσότερο μεταξύ τους.
- Στη δεξιά ομάδα (μοβ και κόκκινες γραμμές), τα δείγματα 6, 7, 8 ανήκουν στην κόκκινη υπό ομάδα ενώ τα δείγματα 4, 3, 5, 2, 9, 0, 1 σχηματίζουν την τελευταία και μοβ υπό ομάδα.

Ο αλγόριθμος Agglomerative Clustering όχι μόνο διαχωρίζει τα δεδομένα σε δύο βασικά clusters, αλλά ταυτόχρονα εντοπίζει την συσχέτιση των δειγμάτων εντός των clusters. Οι υπό ομάδες αυτές δείχνουν ότι ορισμένες περιοχές ή δείγματα έχουν πιο κοντινά χαρακτηριστικά και είναι πιο "σφιχτά" ομαδοποιημένα σε σχέση με άλλα.



Εικόνα 48 Απεικόνιση δενδρόγραμμα ομαδοποίησης *Agglomerative Method*,



Εικόνα 49 Απεικόνιση μετρικών αποδόσεων ομαδοποίησης

Ο Silhouette Score αποτελεί έναν κρίσιμο δείκτη αξιολόγησης της ποιότητας των αποτελεσμάτων ομαδοποίησης. Με τιμές από -1 έως 1, δείχνει πόσο καλά ομαδοποιούνται τα δεδομένα: ένα σκορ κοντά στο 1 σημαίνει ότι τα δεδομένα είναι πολύ κοντά στους στόχους της κάθε ομάδας, ενώ ένα αρνητικό σκορ δείχνει μια κακή ομαδοποίηση.

Ο αλγόριθμος K-Means πέτυχε Silhouette Score 1.0, που σημαίνει ότι ο διαχωρισμός των δεδομένων στις δύο ομάδες (Λακωνία και Μεσσηνία) ήταν απόλυτα σωστός. Παρά τα γνωστά μειονεκτήματα του K-Means, όπως η ευαισθησία στην επιλογή των αρχικών κέντρων, τα αποτελέσματα εδώ δείχνουν ότι η ομαδοποίηση ήταν ιδανική για αυτό το σύνολο δεδομένων. Ο Agglomerative πέτυχε επίσης Silhouette Score 1.0, υποδεικνύοντας ότι και η ιεραρχική μέθοδος ομαδοποίησης διαχώρισε τέλεια τα δεδομένα. Η μέθοδος αυτή αποδεικνύεται ιδιαίτερα σταθερή, ειδικά όταν τα δεδομένα έχουν σαφείς σχέσεις, όπως στην παρούσα περίπτωση. Το γεγονός ότι ο αλγόριθμος δεν απαιτεί προκαθορισμένα κέντρα ή αυθαίρετη κατανομή των δεδομένων του δίνει ένα πλεονέκτημα σε τέτοιες εφαρμογές. Ο αλγόριθμος DBSCAN, με Silhouette Score 1.0, επιβεβαιώνει την εξαιρετική του απόδοση. Η ικανότητά του να διαχωρίζει πυκνές περιοχές από αραιές και να ανιχνεύει τον θόρυβο, τον καθιστά ιδανικό για το συγκεκριμένο σύνολο δεδομένων. Επιπλέον, το γεγονός ότι δεν απαιτεί έναν καθορισμό κάποιου αριθμού clusters εκ των προτέρων του προσδίδει σημαντική ευελιξία.

7. Συζήτηση

7.1 Συμπεράσματα

Η νοθεία στο ελαιόλαδο και η ψευδής επισήμανση της γεωγραφικής του προέλευσης αποτελούν διαρκή πρόκληση για τη βιομηχανία, θέτοντας σε κίνδυνο την εμπιστοσύνη των καταναλωτών και τη βιωσιμότητα των έντιμων παραγωγών. Η αυθεντικότητα και η γεωγραφική προέλευση του ελαιόλαδου επηρεάζουν σημαντικά την ποιότητα του προϊόντος και την τιμή του στην αγορά, με αποτέλεσμα οι καταναλωτές να πληρώνουν αδίκως για λάδι χαμηλότερης ποιότητας ή για λάδι που δεν προέρχεται από τις δηλωμένες περιοχές. Οι παραδοσιακές μέθοδοι ελέγχου, όπως η αισθητηριακή αξιολόγηση και η χημική ανάλυση, αν και αποτελεσματικές, είναι χρονοβόρες, κοστοβόρες και απαιτούν εξειδικευμένη γνώση. Επιπλέον, αυτές οι μέθοδοι βασίζονται στην ανθρώπινη κρίση, η οποία ενδέχεται να μην εντοπίζει πάντα σύνθετες τεχνικές απάτης. Αυτές οι αδυναμίες καθιστούν αναγκαία την υιοθέτηση νέων, πιο αυτοματοποιημένων και ακριβών τεχνικών για την ανίχνευση της προέλευσης του ελαιόλαδου.

Για την επίλυση του προβλήματος της γεωγραφικής προέλευσης, χρησιμοποιήθηκαν δεδομένα φασματοσκοπίας FTIR από δείγματα ελαιόλαδου διαφορετικών περιοχών της Πελοποννήσου. Αρχικά, τα δεδομένα επεξεργάστηκαν με τις μεθόδους Principal Component Analysis (PCA) και binning, ώστε να μειωθεί η διάσταση των δεδομένων και να αναλυθούν πιο εύκολα από τους αλγορίθμους μηχανικής μάθησης. Στη συνέχεια, εφαρμόστηκαν αλγόριθμοι ταξινόμησης, όπως Decision Tree, Naive Bayes, Support Vector Machine, Random Forests και Linear Discriminant Analysis, με τεχνική εκμάθησης αρχικά Split Validation όπου όλοι οι αλγόριθμοι παρουσίασαν τέλεια απόδοση στα δεδομένα, ανεξάρτητα από το αν χρησιμοποιήθηκαν τα δεδομένα μετά από PCA ή binning. Αυτό πιθανόν οφείλεται στη μικρή ποσότητα δεδομένων και στο γεγονός ότι τα χαρακτηριστικά των δειγμάτων είναι αρκετά διαχωρίσιμα. Οι ίδιοι αλγόριθμοι εφαρμόστηκαν και με τεχνική εκμάθησης Cross Validation όπου όλοι οι αλγόριθμοι παρουσίασαν τέλεια απόδοση με τη μέθοδο binning, γεγονός που υποδεικνύει ότι αυτή η προσέγγιση διατηρεί τα απαραίτητα χαρακτηριστικά για την επιτυχή ταξινόμηση. Αντίθετα, η χρήση PCA προκάλεσε σημαντικές πτώσεις στην απόδοση των περισσότερων αλγορίθμων, ιδιαίτερα του Logistic Regression και του Naive Bayes. Αυτό δείχνει ότι η μέθοδος

PCA μπορεί να αφαιρέσει σημαντική πληροφορία κατά τη μείωση διαστάσεων, οδηγώντας σε συνδυασμό με την τεχνική Cross Validation λιγότερο αξιόπιστες προβλέψεις.

Τέλος χρησιμοποιήθηκαν και οι αλγόριθμοι ομαδοποίησης, όπως K-means, Agglomerative Clustering και DBSCAN. Όλοι οι αλγόριθμοι που εξετάστηκαν πέτυχαν άριστα αποτελέσματα με Silhouette Score 1.0. Αυτό δείχνει ότι τα δείγματα ελαιόλαδου από τις περιοχές της Λακωνίας και της Μεσσηνίας διαθέτουν σαφώς διαχωρίσιμα χαρακτηριστικά, επιτρέποντας στους αλγορίθμους να τα ομαδοποιήσουν τέλεια. Παρά τις διαφορές στις προσεγγίσεις τους, όλοι οι αλγόριθμοι ήταν εξίσου αποτελεσματικοί. Η επιτυχία του K-Means δείχνει ότι οι ομάδες είναι σχετικά ομοιογενείς και ισοκατανεμημένες, ενώ η σταθερότητα του Agglomerative Clustering υπογραμμίζει την ικανότητα του αλγορίθμου να αναγνωρίζει την ιεραρχική δομή των δεδομένων. Τέλος, η ακριβής απόδοση του DBSCAN υποδηλώνει ότι τα δεδομένα έχουν καθαρές πυκνότητες χωρίς θόρυβο, κάνοντας τον ιδανικό για τέτοιες περιπτώσεις. Τα αποτελέσματα δείχνουν ότι το σύνολο των δεδομένων ήταν ιδανικό για τη συγκεκριμένη ανάλυση και ότι οι αλγόριθμοι ομαδοποίησης μπορούν να χρησιμοποιηθούν επιτυχώς για την ανάλυση παρόμοιων δεδομένων.

Τα αποτελέσματα που προέκυψαν από την ανάλυση δείχνουν ότι οι αλγόριθμοι που χρησιμοποιήθηκαν επιτυγχάνουν μεγάλα ποσοστά ακρίβειας στις προβλέψεις τους, γεγονός που υποδεικνύει ότι το μοντέλο ταιριάζει υπερβολικά στα δεδομένα. Αυτό μπορεί να αποδοθεί στο μικρό μέγεθος του δείγματος, καθώς με περισσότερα δεδομένα η ακρίβεια θα μπορούσε να μειωθεί, αποτυπώνοντας πιο ρεαλιστικά την πρόβλεψη της γεωγραφικής προέλευσης. Η χρήση τεχνικών όπως το Cross Validation προσπάθησε να μετριάσει αυτή την τάση, αλλά η υπερβολική ακρίβεια υπογραμμίζει την ανάγκη για μεγαλύτερο σύνολο δεδομένων. Στο μέλλον, η επέκταση του δείγματος θα επιτρέψει την ανάπτυξη πιο γενικεύσιμων μοντέλων, τα οποία θα είναι ικανά να προβλέψουν με ακρίβεια την προέλευση σε πραγματικά περιβάλλοντα. Επιπλέον, η ενσωμάτωση επιπλέον χαρακτηριστικών και μεταβλητών, όπως οι κλιματικές συνθήκες, θα μπορούσε να βελτιώσει την ακρίβεια των μοντέλων.

7.2 Μελλοντική Εργασία

Σε μελλοντικές έρευνες, θα ήταν χρήσιμο να εξεταστούν διαφορετικές προσεγγίσεις για την αντιμετώπιση του περιορισμένου αριθμού δειγμάτων, όπως η χρήση τεχνικών αύξησης δεδομένων (data augmentation) ή η συγκέντρωση μεγαλύτερων συνόλων δεδομένων από διαφορετικές περιοχές και έτη παραγωγής. Επιπλέον, η ενσωμάτωση πολυμεταβλητών αναλύσεων, όπως η χρήση δεδομένων από άλλες μεθόδους φασματοσκοπίας ή η συνδυασμένη ανάλυση με χημικά χαρακτηριστικά του ελαιόλαδου, θα μπορούσε να οδηγήσει σε πιο ακριβείς και γενικεύσιμες προβλέψεις.

Άλλη κατεύθυνση για μελλοντική έρευνα είναι η ανάπτυξη υβριδικών μοντέλων μηχανικής μάθησης που συνδυάζουν ταξινόμηση και ομαδοποίηση, βελτιώνοντας την κατανόηση των μοτίβων στα δεδομένα. Επίσης, θα μπορούσε να εξεταστεί η χρήση αλγορίθμων βαθιάς μάθησης (deep learning), ειδικά σε μεγαλύτερα σύνολα δεδομένων, που θα επιτρέψουν την ανάλυση ακόμα πιο πολύπλοκων σχέσεων στα δεδομένα φασματοσκοπίας.

Τέλος, είναι σημαντικό να αναπτυχθούν εργαλεία που επιτρέπουν τη διαφάνεια και την ερμηνευσιμότητα των μοντέλων, ώστε να μπορούν οι παραγωγοί και οι καταναλωτές να εμπιστεύονται τα αποτελέσματα των αναλύσεων. Επενδύοντας σε μια διεπιστημονική προσέγγιση που συνδυάζει τις γνώσεις από τη φασματοσκοπία, τη χημεία και τη μηχανική μάθηση.

8. Βιβλιογραφία

- Ballin, N. Z. (2010). Authentication of meat and meat products. In *Meat Science* (Vol. 86, Issue 3, pp. 577–587). <https://doi.org/10.1016/j.meatsci.2010.06.001>
- BINNING AS A PRETEXT TASK: IMPROVING SELF-SUPERVISED LEARNING IN TABULAR DOMAINS*. (n.d.).
- Bishop, C. M. . (2009). *Pattern recognition and machine learning*. Springer Science + Business Media.
- Breiman, L. (2001). *Random Forests* (Vol. 45).
- Burges, C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. In *Data Mining and Knowledge Discovery* (Vol. 2).
- Cortes, C., Vapnik, V., & Saitta, L. (1995). Support-Vector Networks Editor. In *Machine Learning* (Vol. 20). Kluwer Academic Publishers.
- Dans, P. W., Foglia, S. D., & Nelson, A. J. (2021). Data processing in functional near-infrared spectroscopy (Fnirs) motor control research. In *Brain Sciences* (Vol. 11, Issue 5). MDPI AG. <https://doi.org/10.3390/brainsci11050606>
- Dehuri, S., Prasad Mishra, B. S., Mallick, P. K., & Cho, S.-B. (Eds.). (2022). *Biologically Inspired Techniques in Many Criteria Decision Making* (Vol. 271). Springer Nature Singapore. <https://doi.org/10.1007/978-981-16-8739-6>
- Domingos, P. (2012). A few useful things to know about machine learning. In *Communications of the ACM* (Vol. 55, Issue 10, pp. 78–87). <https://doi.org/10.1145/2347736.2347755>
- Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. www.aaai.org
- Figueiredo, M. A. T., & Jain, A. K. (n.d.). *Unsupervised Learning of Finite Mixture Models*.
- Gautam, R., Vanga, S., Ariese, F., & Umapathy, S. (2015). Review of multidimensional data processing approaches for Raman and infrared spectroscopy. *EPJ Techniques and Instrumentation*, 2(1). <https://doi.org/10.1140/epjti/s40485-015-0018-6>
- Gazeli, O., Bellou, E., Stefas, D., & Couris, S. (2020). Laser-based classification of olive oils assisted by machine learning. *Food Chemistry*, 302. <https://doi.org/10.1016/j.foodchem.2019.125329>

- Ghojogh, B., & Crowley, M. (2019). *Linear and Quadratic Discriminant Analysis: Tutorial*. <http://arxiv.org/abs/1906.02590>
- Han, J., Kamber, M., & Pei, J. (2011). *Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)*.
- Huynh, P. K. (2023). *Knowledge Integration in Domain-Informed Machine Learning and Multi-Scale Modeling of Nonlinear Dynamics in Complex Systems*. <https://doi.org/10.13140/RG.2.2.32234.29128>
- Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651–666. <https://doi.org/10.1016/j.patrec.2009.09.011>
- James, Gareth., Witten, Daniela., Hastie, Trevor., & Tibshirani, Robert. (2017). *An introduction to statistical learning : with applications in R*. Springer : Springer Science+Business Media.
- John, G. H. (n.d.). 338 *Estimating Continuous Distributions in Bayesian Classifiers*. <http://robotics.stanford.edu/~gjohn/>
- Kalogiouri, N. P., Aalizadeh, R., Dasenaki, M. E., & Thomaidis, N. S. (2020). Application of High Resolution Mass Spectrometric methods coupled with chemometric techniques in olive oil authenticity studies - A review. In *Analytica Chimica Acta* (Vol. 1134, pp. 150–173). Elsevier B.V. <https://doi.org/10.1016/j.aca.2020.07.029>
- Kathole, A. B., Halgaonkar, P. S., & Nikhade, A. A. (2019). Machine learning & its classification techniques. *International Journal of Innovative Technology and Exploring Engineering*, 8(9 Special Issue 3), 138–142. <https://doi.org/10.35940/ijitee.i3028.0789s319>
- Kohavi, R. (n.d.). *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*. [http://roboticsStanfordedu/"ronnyk](http://roboticsStanfordedu/)
- Küchenhoff, L., Lukas, P., Metz-Zumaran, C., Rothhaar, P., Ruggieri, A., Lohmann, V., Höfer, T., Stanifer, M. L., Boulant, S., Talemi, S. R., & Graw, F. (2023). Extended methods for spatial cell classification with DBSCAN-CellX. *Scientific Reports*, 13(1). <https://doi.org/10.1038/s41598-023-45190-4>
- Lerma-García, M. J., Ramis-Ramos, G., Herrero-Martínez, J. M., & Simó-Alfonso, E. F. (2010). Authentication of extra virgin olive oils by Fourier-transform infrared spectroscopy. *Food Chemistry*, 118(1), 78–83. <https://doi.org/10.1016/j.foodchem.2009.04.092>

- Leung, H., Zhihuiilai, H., & Xianyiizhang, H. (n.d.). *Information Fusion and Data Science Series Editor: Feature Learning and Understanding Algorithms and Applications*. <http://www.springer.com/series/15462>
- Louppe, G. (2014). *Understanding Random Forests: From Theory to Practice*. <http://arxiv.org/abs/1407.7502>
- Mahesh, B. (2020). Machine Learning Algorithms - A Review. *International Journal of Science and Research (IJSR)*, 9(1), 381–386. <https://doi.org/10.21275/art20203995>
- Mokari, A., Guo, S., & Bocklitz, T. (2023). Exploring the Steps of Infrared (IR) Spectral Analysis: Pre-Processing, (Classical) Data Modelling, and Deep Learning. In *Molecules* (Vol. 28, Issue 19). Multidisciplinary Digital Publishing Institute (MDPI). <https://doi.org/10.3390/molecules28196886>
- Monath, N., Dubey, K. A., Guruganesh, G., Zaheer, M., Ahmed, A., McCallum, A., Mergen, G., Najork, M., Terzihan, M., Tjanaka, B., Wang, Y., & Wu, Y. (2021). Scalable Hierarchical Agglomerative Clustering. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1245–1255. <https://doi.org/10.1145/3447548.3467404>
- Müllner, D. (2011). *Modern hierarchical, agglomerative clustering algorithms*. <http://arxiv.org/abs/1109.2378>
- Orlandi, F., Rojo, J., Picornell, A., Oteros, J., Pérez-Badia, R., & Fornaciari, M. (2020). Impact of climate change on olive crop production in Italy. *Atmosphere*, 11(6). <https://doi.org/10.3390/atmos11060595>
- Prakash, O., Pathak, A., Kumar, A., Juyal, V. K., Joshi, H. C., Gangola, S., Patni, K., Bhandari, G., Suyal, D. C., & Nand, V. (2021). Spectroscopy and Its Advancements for Environmental Sustainability. In *Bioremediation of Environmental Pollutants: Emerging Trends and Strategies* (pp. 317–338). Springer International Publishing. https://doi.org/10.1007/978-3-030-86169-8_14
- Quinlan, J. R. (1986). Induction of Decision Trees. In *Machine Learning* (Vol. 1).
- Ran, X., Xi, Y., Lu, Y., Wang, X., & Lu, Z. (2023). Comprehensive survey on hierarchical clustering algorithms and the recent developments. *Artificial Intelligence Review*, 56(8), 8219–8264. <https://doi.org/10.1007/s10462-022-10366-3>

- Rapa, M., & Ciano, S. (2022). A Review on Life Cycle Assessment of the Olive Oil Production. In *Sustainability (Switzerland)* (Vol. 14, Issue 2). MDPI. <https://doi.org/10.3390/su14020654>
- Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. da F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PLoS ONE*, 14(1). <https://doi.org/10.1371/journal.pone.0210236>
- Rokach, L. (2009). A survey of Clustering Algorithms. In *Data Mining and Knowledge Discovery Handbook* (pp. 269–298). Springer US. https://doi.org/10.1007/978-0-387-09823-4_14
- Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. In *SN Computer Science* (Vol. 2, Issue 3). Springer. <https://doi.org/10.1007/s42979-021-00592-x>
- Sayago, A., González-Domínguez, R., Beltrán, R., & Fernández-Recamiales, Á. (2018). Combination of complementary data mining methods for geographical characterization of extra virgin olive oils based on mineral composition. *Food Chemistry*, 261, 42–50. <https://doi.org/10.1016/j.foodchem.2018.04.019>
- Serra-Burriel, M., & Ames, C. (2022). Machine Learning-Based Clustering Analysis: Foundational Concepts, Methods, and Applications. In *Acta Neurochirurgica, Supplementum* (Vol. 134, pp. 91–100). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/978-3-030-85292-4_12
- Shlens, J. (n.d.). *A Tutorial on Principal Component Analysis*.
- Subramanian, A., Alvarez, V. B., Harper, W. J., & Rodriguez-Saona, L. E. (2011). Monitoring amino acids, organic acids, and ripening changes in Cheddar cheese using Fourier-transform infrared spectroscopy. *International Dairy Journal*, 21(6), 434–440. <https://doi.org/10.1016/j.idairyj.2010.12.012>
- Sun, D.-Wen. (2009). *Infrared spectroscopy for food quality analysis and control*. Academic Press/Elsevier.
- Tan, J., Yang, J., Wu, S., Chen, G., & Zhao, J. (2021). A critical look at the current train/test split in machine learning. <http://arxiv.org/abs/2106.04525>
- Tang, T. M., & Allen, G. I. (2021). Integrated Principal Components Analysis. In *Journal of Machine Learning Research* (Vol. 22). <http://jmlr.org/papers/v22/20-084.html>.

- Tharwat, A., Gaber, T., Ibrahim, A., & Hassanien, A. E. (2017). Linear discriminant analysis: A detailed tutorial. *AI Communications*, 30(2), 169–190. <https://doi.org/10.3233/AIC-170729>
- Tibshirani, R. J., & Tibshirani, R. (2009). A bias correction for the minimum error rate in cross-validation. *Annals of Applied Statistics*, 3(2), 822–829. <https://doi.org/10.1214/08-AOAS224>
- Tokuda, E. K., Comin, C. H., & Costa, L. da F. (2022). Revisiting agglomerative clustering. *Physica A: Statistical Mechanics and Its Applications*, 585. <https://doi.org/10.1016/j.physa.2021.126433>
- Verdonck, T., Baesens, B., Óskarsdóttir, M., & vanden Broucke, S. (2024). Special issue on feature engineering editorial. *Machine Learning*, 113(7), 3917–3928. <https://doi.org/10.1007/s10994-021-06042-2>
- Wagner, M., Naik, D., & Pothén, A. (2003). Protocols for disease classification from mass spectrometry data. *Proteomics*, 3(9), 1692–1698. <https://doi.org/10.1002/pmic.200300519>
- Xiaoming, D., Ying, C., Xiaofang, Z., Yu, G., & Campus, Z. (n.d.). Study on Feature Engineering and Ensemble Learning for Student Academic Performance Prediction. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 13, Issue 5). www.ijacsa.thesai.org
- Yin, H., Aryani, A., Petrie, S., Nambissan, A., Astudillo, A., & Cao, S. (2024). A Rapid Review of Clustering Algorithms. <http://arxiv.org/abs/2401.07389>