ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΩΝ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

“ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ & ΥΠΗΡΕΣΙΕΣ”

Τίτλος ΜΔΕ: "Χρήση μετα-χαρακτηριστικών για την βελτίωση της συσταδοποίησης σε Εφαρμογές Μηχανικής Μάθησης"

MSc Thesis Title:"On the Use of Metafeatures for Improving Clustering in Machine Learning Applications"

Γεώργιος Κουλός

Υποβάλλεται

για την εκπλήρωση των προϋποθέσεων λήψης

Μεταπτυχιακού Διπλώματος

στην ειδίκευση «ΜΔΑ/ΠΠΣ/ΠΔ»

του ΠΜΣ “Πληροφοριακά Συστήματα & Υπηρεσίες”

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Ιούνιος 2024

Επιβλέπων: Χρήστος Δουλκερίδης

Ακαδημαϊκή Θέση: Αναπληρωτής Καθηγητής

Συγγραφέας / Author . . . . . .

ΣΕΛΙΔΑ ΕΓΚΥΡΟΤΗΤΑΣ

Ονοματεπώνυμο Φοιτητή: Γεώργιος Κουλός
Τίτλος Μεταπτυχιακής Διπλωματικής Εργασίας: "Χρήση Μετα-χαρακτηριστικών για την Βελτίωση της Συσταδοποίησης σε Εφαρμογές Μηχανικής Μάθησης"

Η παρούσα Μεταπτυχιακή Διπλωματική Εργασία υποβάλλεται ως μερική εκπλήρωση των απαιτήσεων του Προγράμματος Μεταπτυχιακών Σπουδών "Πληροφοριακά Συστήματα & Υπηρεσίες" του Τμήματος Ψηφιακών Συστημάτων του Πανεπιστημίου Πειραιώς και εγκρίθηκε στις 13/6/2024 [ημερομηνία έγκρισης] από τα μέλη της Εξεταστικής Επιτροπής.

Εξεταστική Επιτροπή

Επιβλέπων/ουσα (Τμήμα Ψηφιακών Συστημάτων, Πανεπιστήμιο Πειραιώς): Δουλκερίδης Χρήστος, Αναπληρωτής Καθηγητής
Μέλος Εξεταστικής Επιτροπής: Χαλκίδη Μαρία, Αναπληρώτρια Καθηγήτρια
Μέλος Εξεταστικής Επιτροπής: Τελέλης Ορέστης, Επίκουρος Καθηγητής

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΑΥΘΕΝΤΙΚΟΤΗΤΑΣ

Ο Γεώργιος Κουλός, γνωρίζοντας τις συνέπειες της λογοκλοπής, δηλώνω υπεύθυνα ότι η παρούσα εργασία με τίτλο «Χρήση μετα-χαρακτηριστικών για την Βελτίωση της Συσταδοποίησης σε Εφαρμογές Μηχανικής Μάθησης», αποτελεί προϊόν αυστηρά προσωπικής εργασίας και όλες οι πηγές που έχω χρησιμοποιήσει, έχουν δηλωθεί κατάλληλα στις βιβλιογραφικές παραπομπές και αναφορές. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή.

Επιπλέον *δηλώνω υπεύθυνα ότι η συγκεκριμένη Μεταπτυχιακή Διπλωματική Εργασία έχει συγγραφεί από εμένα προσωπικά και δεν έχει υποβληθεί ούτε έχει αξιολογηθεί στο πλαίσιο κάποιου άλλου μεταπτυχιακού ή προπτυχιακού τίτλου σπουδών, στην Ελλάδα ή στο εξωτερικό.*

Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου. Σε κάθε περίπτωση, αναληθούς ή ανακριβούς δηλώσεως, υπόκειμαι στις συνέπειες που προβλέπονται τις διατάξεις που προβλέπει η Ελληνική και Κοινοτική Νομοθεσία περί πνευματικής ιδιοκτησίας.

Ο ΔΗΛΩΝ

Ονοματεπώνυμο: Γεώργιος Κουλός

Αριθμός Μητρώου: Μ2138

Υπογραφή:

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω τον καθηγητή μου Χρήστο Δουλκερίδη για την άψογη συνεργασία, καθώς και για την κατανόηση, υπομονή και καθοδήγηση που μου έδειξε, σε μια περίοδο ριζικών αλλαγών στην ζωή μου. Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου και την σύζυγο μου Χαρά, για την αμέριστη στήριξη που μου παρείχε για εκπόνηση αυτής της διατριβής και να την αφιερώσω στην κόρη μου Σαββίνα.

ΠΕΡΙΕΧΟΜΕΝΑ

Πίνακας περιεχομένων

ΠΕΡΙΛΗΨΗ

Η μηχανική μάθηση (ML), ένα βασικό στοιχείο της τεχνητής νοημοσύνης (AI), επιτρέπει στα συστήματα να μαθαίνουν από δεδομένα και να λαμβάνουν έξυπνες αποφάσεις χωρίς ρητό προγραμματισμό. Το πεδίο έχει εξελιχθεί σημαντικά από την πρόταση του Alan Turing για μια «μηχανή μάθησης» στη δεκαετία του 1950, που τώρα περιλαμβάνει τεχνικές όπως η εποπτευόμενη μάθηση, η μη εποπτευόμενη μάθηση και η ενισχυτική μάθηση. Μεταξύ αυτών, η ομαδοποίηση - ένας τύπος μάθησης χωρίς επίβλεψη - διαδραματίζει κρίσιμο ρόλο στην ανακάλυψη μοτίβων και δομών μέσα σε σύνολα δεδομένων.

Παρά τις δυνατότητές της, οι παραδοσιακές ροές εργασίας ML είναι πολύπλοκες και χρονοβόρες, περιλαμβάνοντας βήματα όπως η προεπεξεργασία δεδομένων, η επιλογή χαρακτηριστικών, η επιλογή μοντέλου και η ρύθμιση υπερπαραμέτρων. Αυτές οι διαδικασίες απαιτούν σημαντική εμπειρογνωμοσύνη και πόρους, θέτοντας εμπόδιο στην ευρύτερη υιοθέτηση της μηχανικής μάθησης. Το AutoML (Automated Machine Learning) αντιμετωπίζει αυτές τις προκλήσεις αυτοματοποιώντας τη ροή εργασίας ML. Χρησιμοποιεί προηγμένους αλγόριθμους για τον εξορθολογισμό των εργασιών, καθιστώντας την ML προσβάσιμη σε ένα ευρύτερο κοινό. Το AutoML όχι μόνο επιταχύνει την ανάπτυξη μοντέλων, αλλά και εκδημοκρατίζει το ML, προωθώντας την καινοτομία σε διάφορους τομείς. Ωστόσο, καθώς το AutoML εξελίσσεται, είναι ζωτικής σημασίας να αντιμετωπιστούν ζητήματα που σχετίζονται με την ερμηνευσιμότητα, τη δικαιοσύνη και το απόρρητο για να διασφαλιστεί η ηθική και αποτελεσματική εφαρμογή ML

Ο πυρήνας εστίασης αυτής της διατριβής είναι στη χρήση μεταχαρακτηριστικών για την ενίσχυση της απόδοσης των αλγορίθμων ομαδοποίησης σε εφαρμογές ML. Τα μεταχαρακτηριστικά είναι περιγραφικά χαρακτηριστικά που εξάγονται από σύνολα δεδομένων που παρέχουν πολύτιμες πληροφορίες σχετικά με τη δομή των δεδομένων, καθοδηγώντας την επιλογή και τη βελτιστοποίηση των αλγορίθμων ομαδοποίησης. Αξιοποιώντας τα μεταχαρακτηριστικά, είναι δυνατό να βελτιωθεί η ακρίβεια, η αποτελεσματικότητα και η ευρωστία της ομαδοποίησης. Οι παραδοσιακές ροές εργασίας ML για ομαδοποίηση περιλαμβάνουν πολύπλοκες, χρονοβόρες διαδικασίες που απαιτούν σημαντική εξειδίκευση στον τομέα. Αυτές οι διαδικασίες περιλαμβάνουν προεπεξεργασία δεδομένων, μηχανική χαρακτηριστικών, επιλογή αλγορίθμων και συντονισμό υπερπαραμέτρων. Η εισαγωγή των μεταχαρακτηριστικών εξορθολογίζει αυτές τις ροές εργασίας παρέχοντας μια μετα-επίπεδη κατανόηση των δεδομένων, η οποία με τη σειρά της διευκολύνει την αυτόματη επιλογή και διαμόρφωση αλγορίθμων ομαδοποίησης μέσω συστημάτων όπως το AutoML (Automated Machine Learning).

Τα συστήματα AutoML χρησιμοποιούν μεταχαρακτηριστικά για την αυτοματοποίηση διαφόρων σταδίων του αγωγού ML, μειώνοντας σημαντικά την ανάγκη για χειροκίνητη παρέμβαση και τεχνογνωσία. Τα μεταχαρακτηριστικά ενημερώνουν το σύστημα σχετικά με τις καταλληλότερες τεχνικές ομαδοποίησης και ρυθμίσεις παραμέτρων, ενισχύοντας τη συνολική απόδοση και αξιοπιστία των μοντέλων. Αυτός ο αυτοματισμός εκδημοκρατίζει το ML, καθιστώντας τις προηγμένες τεχνικές ομαδοποίησης προσβάσιμες σε ένα ευρύτερο κοινό και προωθώντας την καινοτομία σε διάφορους τομείς, όπως η επεξεργασία φυσικής γλώσσας, η υπολογιστική όραση και η υγειονομική περίθαλψη.Η διατριβή διερευνά την ανάπτυξη και εφαρμογή μεταχαρακτηριστικών σε εργασίες ομαδοποίησης, αποδεικνύοντας τον αντίκτυπό τους στη βελτίωση των αποτελεσμάτων ομαδοποίησης. Αναλύοντας συστηματικά διαφορετικούς τύπους μεταχαρακτηριστικών και την επίδρασή τους στην απόδοση ομαδοποίησης, η έρευνα παρέχει ένα ολοκληρωμένο πλαίσιο για την ενσωμάτωση μεταχαρακτηριστικών σε ροές εργασίας ML.

Η ανάλυση υπογραμμίζει την πολυπλοκότητα που συνεπάγεται η ομαδοποίηση συνόλων δεδομένων και τη σημασία της εξέτασης τόσο των μεταχαρακτηριστικών υψηλού επιπέδου όσο και των επιπτώσεων της προεπεξεργασίας δεδομένων. Τα αποτελέσματα δείχνουν ότι τα μεταχαρακτηριστικά ομαδοποιούν σταθερά σύνολα δεδομένων με παρόμοιες στατιστικές ιδιότητες, όπως τα σύνολα δεδομένων αποτυχίας ρομπότ, υποδεικνύοντας κοινές κλίμακες ή κατανομές. Αυτή η ευθυγράμμιση υποδηλώνει ότι τα μεταχαρακτηριστικά συχνά ενσωματώνουν στοιχεία όπως η διακύμανση, η ασυμμετρία ή η κύρτωση, ενδεικτικά της γενικής δομής των συνόλων δεδομένων. Οι μελλοντικές εργασίες θα μπορούσαν να περιλαμβάνουν λεπτομερέστερη εξέταση των σταδίων προεπεξεργασίας για την κατανόηση των συγκεκριμένων επιπτώσεών τους στα αποτελέσματα της ομαδοποίησης. Επιπλέον, η διερεύνηση εναλλακτικών τεχνικών ομαδοποίησης και η ενσωμάτωση της γνώσης θα μπορούσαν να αποσαφηνίσουν περαιτέρω τις διακρίσεις μεταξύ συνόλων δεδομένων που παρατηρούνται στην προεπεξεργασμένη ομαδοποίηση.

Συμπερασματικά, αυτή η διατριβή υπογραμμίζει το δυναμικό μετασχηματιστικό των μεταχαρακτηριστικών στην προώθηση μεθοδολογιών ομαδοποίησης στο πλαίσιο της μηχανικής μάθησης. Καθώς η ML συνεχίζει να διαπερνά διάφορους τομείς, η στρατηγική χρήση των μεταχαρακτηριστικών στα συστήματα AutoML μπορεί να οδηγήσει σε πιο ακριβείς, αποτελεσματικές και προσβάσιμες λύσεις ομαδοποίησης. Οι μελλοντικές εργασίες θα πρέπει να αντιμετωπίσουν τις προκλήσεις που σχετίζονται με την ερμηνευτικότητα, τη δικαιοσύνη και το απόρρητο του μοντέλου, ώστε να διασφαλιστεί η ηθική και αποτελεσματική ανάπτυξη των τεχνολογιών μηχανικής μάθησης.

Abstract

Machine learning (ML), a pivotal component of artificial intelligence (AI), enables systems to learn from data and make intelligent decisions without explicit programming. The field has evolved significantly since Alan Turing's proposal of a "learning machine" in the 1950s, now encompassing techniques such as supervised learning, unsupervised learning, and reinforcement learning. Among these, clustering—a type of unsupervised learning—plays a crucial role in discovering patterns and structures within datasets.

Despite its potential, traditional ML workflows are complex and time-consuming, involving steps such as data preprocessing, feature selection, model selection, and hyperparameter tuning. These processes require substantial expertise and resources, posing a barrier to broader ML adoption. AutoML (Automated Machine Learning) addresses these challenges by automating the ML workflow. It uses advanced algorithms to streamline tasks, making ML accessible to a wider audience. AutoML not only speeds up model development but also democratizes ML, fostering innovation across various fields. However, as AutoML progresses, it is crucial to address issues related to interpretability, fairness, and privacy to ensure ethical and effective ML applications

The core focus of this thesis is on the use of metafeatures to enhance the performance of clustering algorithms in ML applications. Metafeatures are descriptive characteristics extracted from datasets that provide valuable insights into the data’s structure, guiding the selection and optimization of clustering algorithms. By leveraging metafeatures, it is possible to improve clustering accuracy, efficiency, and robustness. Traditional ML workflows for clustering involve complex, time-consuming processes that require significant domain expertise. These processes include data preprocessing, feature engineering, algorithm selection, and hyperparameter tuning. The introduction of metafeatures streamlines these workflows by providing a meta-level understanding of the data, which

in turn facilitates the automatic selection and configuration of clustering algorithms through systems like AutoML (Automated Machine Learning).

AutoML systems utilize metafeatures to automate various stages of the ML pipeline, significantly reducing the need for manual intervention and expertise. Metafeatures inform the system about the most suitable clustering techniques and parameter settings, enhancing the overall performance and reliability of the models. This automation democratizes ML, making advanced clustering techniques accessible to a broader audience and fostering innovation across various domains, such as natural language processing, computer vision, and healthcare. The thesis explores the development and application of metafeatures in clustering tasks, demonstrating their impact on improving clustering outcomes. By systematically analyzing different types of metafeatures and their influence on clustering performance, the research provides a comprehensive framework for integrating metafeatures into ML workflows.

The analysis highlights the complexities involved in clustering datasets and the importance of considering both high-level metafeatures and the impacts of data preprocessing. The results show that metafeatures consistently group datasets with similar statistical properties, such as the robot failure datasets, indicating common scales or distributions. This alignment suggests that metafeatures often incorporate elements like variance, skewness, or kurtosis, indicative of the general structure of datasets. Future work could involve a more detailed examination of preprocessing steps to understand their specific impacts on clustering outcomes. Additionally, exploring alternative clustering techniques and integrating domain knowledge could further clarify the distinctions between datasets observed in preprocessed clustering.

In conclusion, this thesis underscores the transformative potential of metafeatures in advancing clustering methodologies within machine learning. As ML continues to permeate various sectors, the strategic use of metafeatures in AutoML systems can lead to more accurate, efficient, and accessible clustering solutions. Future work will need to address challenges related to model interpretability, fairness, and privacy to ensure the ethical and effective deployment of Machine Learning technologies.

1.ΕΙΣΑΓΩΓΗ

Η μηχανική μάθηση (ML) έχει γίνει ακρογωνιαίος λίθος στον τομέα της τεχνητής νοημοσύνης (AI), προσφέροντας τη δυνατότητα αυτόματης μάθησης και βελτίωσης μέσω της εμπειρίας που έχει αποκτηθεί, χωρίς να είναι σαφώς προγραμματισμένο. Αυτό το πεδίο επικεντρώνεται στην ανάπτυξη αλγορίθμων που μπορούν να επεξεργαστούν, να αναλύσουν και να μάθουν από την επεξεργασία δεδομένων, και μετέπειτα μπορούν να πάρουν έξυπνες αποφάσεις με βάση τα μαθησιακά μοτίβα και τις πληροφορίες. Η ουσία της μηχανικής μάθησης έγκειται στην ικανότητά της να προσαρμόζεται σε νέα δεδομένα ανεξαρτήτως από τις οδηγίες και τους κανόνες ενός προγράμματος, επιτρέποντας της να κάνει προβλέψεις ή να παίρνει αποφάσεις βάσει των δεδομένων που είναι διαθέσιμα.

Η προέλευση της μηχανικής μάθησης έχει βαθιές ρίζες μέσα στην ιστορία της τεχνητής νοημοσύνης, με τις πρώιμες έννοιες των πρώτων υπολογιστών που θα μαθαίνουν σαν άνθρωποι, να κάνουν ήδη την εμφάνιση τους από τη δεκαετία του 1950. Ο Alan Turing, μια πρωτοποριακή φιγούρα στην πληροφορική, πρότεινε την ιδέα μιας «μηχανής μάθησης» στη σημαντική εργασία του «Computing Machinery and Intelligence» [1]. Από τότε, το πεδίο της μηχανικής μάθησης έχει εξελιχθεί, από απλή αναγνώριση προτύπων σε πολύπλοκους αλγόριθμους, ικανούς να μαθαίνουν μόνοι τους και να κάνουν εξαιρετικά ακριβείς προβλέψεις. Οι τρεις κύριοι τύποι μηχανικής μάθησης είναι η εποπτευόμενη μάθηση(supervised learning), όπου το μοντέλο εκπαιδεύεται σε επισημασμένα δεδομένα, μάθηση χωρίς επίβλεψη(unsupervised learning), όπου το μοντέλο μαθαίνει από δεδομένα χωρίς σημάνσεις για να βρει κρυμμένα μοτίβα και τέλος την ενισχυτική μάθηση(reinforcement learning), όπου ένας agent μαθαίνει να λαμβάνει αποφάσεις, κάνοντας ενέργειες σε ένα περιβάλλον για την επίτευξη στόχων.

Οι εφαρμογές της μηχανικής μάθησης είναι τεράστιες και ποικίλες, καλύπτοντας διάφορους τομείς, όπως η επεξεργασία φυσικής γλώσσας, η computer vision (ένα πεδίο της επιστήμης των υπολογιστών που εστιάζει στη δυνατότητα των υπολογιστών να αναγνωρίζουν και να κατανοούν αντικείμενα και ανθρώπους σε εικόνες και βίντεο), η υγειονομική περίθαλψη, οι χρηματοπιστωτικές υπηρεσίες και πολλά άλλα. Για παράδειγμα, οι αλγόριθμοι ML τροφοδοτούν τις δυνατότητες αναγνώρισης φωνής σε εικονικούς βοηθούς όπως η Siri και η Alexa, ενεργοποιούν τα συστήματα συστάσεων (recommendation systems) του Netflix και της Amazon και βοηθούν στη διάγνωση ασθενειών αναλύοντας ιατρικές εικόνες.

Παρά τις τεράστιες δυνατότητες και προοπτικές στο χώρο, οι παραδοσιακές ροές εργασίας της μηχανικής μάθησης περιλαμβάνουν πολύπλοκα, χρονοβόρα βήματα, συμπεριλαμβανομένης της προεπεξεργασίας δεδομένων (data preprocessing), της επιλογής χαρακτηριστικών (feature selection), της επιλογής μοντέλων (model selection) και του συντονισμού υπερπαραμέτρων (hyperparameter tuning). Καθένα από αυτά τα βήματα απαιτεί σημαντική γνώση και εξειδίκευση στον τομέα της μηχανικής μάθησης, καθιστώντας την απρόσιτη σε μη ειδικούς και δύσκολη στη χρήση, ακόμη και για έμπειρους επαγγελματίες.

Η αυτοματοποιημένη μηχανική μάθηση (AutoML) αναδύεται ως μια μετασχηματιστική λύση σε αυτές τις προκλήσεις, με στόχο την αυτοματοποίηση της διαδικασίας εφαρμογής της μηχανικής μάθησης από άκρο σε άκρο σε πραγματικά προβλήματα. Το AutoML επιδιώκει να κάνει τη μηχανική εκμάθηση πιο προσιτή, αποτελεσματική και επεκτάσιμη αυτοματοποιώντας εργασίες, όπως η μηχανική χαρακτηριστικών (feature engineering), η επιλογή μοντέλων (model selection) και η βελτιστοποίηση υπερπαραμέτρων (hyperparameter optimization). Ο στόχος είναι να δοθεί η δυνατότητα στους ειδικούς του τομέα χωρίς σημαντικό υπόβαθρο στη μηχανική μάθηση να χρησιμοποιούν αποτελεσματικά τις τεχνικές ML και να επιτρέπουν στους επαγγελματίες της ML να επιτυγχάνουν καλύτερα αποτελέσματα γρηγορότερα.

Τα εργαλεία AutoML, όπως το Cloud AutoML της Google, το AutoML του H2O και το Auto-sklearn, παρέχουν φιλικές προς το χρήστη διεπαφές και αυτοματοποιημένες ροές εργασίας που αφαιρούν μεγάλο μέρος της πολυπλοκότητας που εμπλέκεται στην ανάπτυξη μοντέλων μηχανικής μάθησης. Αυτά τα εργαλεία χρησιμοποιούν εξελιγμένους αλγόριθμους και τεχνικές μετα-μάθησης για την αναζήτηση πιθανών μοντέλων και διαμορφώσεων, προσδιορίζοντας τις πιο κατάλληλες προσεγγίσεις για ένα συγκεκριμένο σύνολο δεδομένων και εργασιών.

Η σημασία του AutoML δεν είναι μόνο στην αυτοματοποίηση της ροής εργασίας ML, αλλά και στην δυνατότητα χρήσης της μηχανικής μάθησης από ένα ευρύτερο κοινό, κάνοντας την πιο προσιτή σε μεγαλύτερο αριθμό χρηστών. Μειώνοντας το εμπόδιο εισόδου, το χρόνο και την τεχνογνωσία που απαιτούνται για την ανάπτυξη μοντέλων ML, το AutoML έχει τη δυνατότητα να επιταχύνει την καινοτομία και την εφαρμογή της μηχανικής μάθησης σε διάφορους τομείς.

Συμπερασματικά, καθώς η μηχανική μάθηση συνεχίζει να εξελίσσεται και να ενσωματώνεται σε διάφορες πτυχές της τεχνολογίας και της καθημερινής ζωής, το AutoML ξεχωρίζει ως μια καθοριστική εξέλιξη που θα μπορούσε να διευρύνει τη χρήση της ML και να προωθήσει μια νέα εποχή λύσεων που βασίζονται στην τεχνητή νοημοσύνη. Καθώς ο τομέας εξελίσσεται, θα είναι σημαντικό να αντιμετωπιστούν προκλήσεις, όπως η ερμηνευσιμότητα, η δικαιοσύνη και η ιδιωτικότητα στα μοντέλα που δημιουργούνται από το AutoML, διασφαλίζοντας ότι όχι μόνο αποδίδουν καλά αλλά και ευθυγραμμίζονται με τα δεοντολογικά πρότυπα και τις κοινωνικές αξίες.

1.1 Δήλωση προβλήματος

Παρά τις σημαντικές εξελίξεις και τις ευρέως διαδεδομένες εφαρμογές της μηχανικής μάθησης (ML) σε διάφορους τομείς, η ανάπτυξη επιτυχημένων μοντέλων ML σε σενάρια πραγματικού κόσμου παραμένει γεμάτη προκλήσεις. Ένα από τα μεγαλύτερα εμπόδια είναι η προϋπόθεση ύπαρξης εκτεταμένης εξειδίκευσης και γνώσης στον τομέα της επιστήμης των δεδομένων για την πλοήγηση στις περίπλοκες διαδικασίες προεπεξεργασίας δεδομένων, μηχανικής χαρακτηριστικών, επιλογής μοντέλων και βελτιστοποίησης υπερπαραμέτρων. Αυτά τα βήματα είναι ζωτικής σημασίας για την κατασκευή αποτελεσματικών μοντέλων ML, αλλά απαιτούν υψηλό επίπεδο τεχνικής επάρκειας και καταναλώνουν σημαντικό χρόνο και πόρους. Ως αποτέλεσμα, η πολυπλοκότητα και η τεχνογνωσία που απαιτούνται στις παραδοσιακές ροές εργασίας μηχανικής μάθησης μπορούν να λειτουργήσουν ως εμπόδια εισόδου για πολλούς οργανισμούς και άτομα, περιορίζοντας την ευρύτερη υιοθέτηση και εφαρμογή τεχνολογιών ML. Επιπλέον, η επαναληπτική φύση της ανάπτυξης μοντέλων ML, σε συνδυασμό με το συνεχώς αυξανόμενο μέγεθος και την πολυπλοκότητα των συνόλων δεδομένων, επιδεινώνει περαιτέρω αυτές τις προκλήσεις, καθιστώντας δύσκολη την αποτελεσματική κλιμάκωση των λύσεων ML και την παρακολούθηση της ταχείας εξέλιξης των πληροφοριών που βασίζονται σε δεδομένα.

Η αυτοματοποιημένη μηχανική μάθηση (AutoML) αναδεικνύεται ως μια πολλά υποσχόμενη λύση για την αντιμετώπιση αυτών των προκλήσεων με τον εξορθολογισμό της ροής εργασίας ML και τον δυνατότητα πρόσβασης σε τεχνολογίες ML, σε ένα ευρύτερο κοινό. Το AutoML στοχεύει στην αυτοματοποίηση των δύσκολων, χρονοβόρων και πολύπλοκων πτυχών της διαδικασίας ML, επιτρέποντας στους χρήστες χωρίς βαθιά τεχνική εξειδίκευση στην επιστήμη των δεδομένων να αναπτύξουν αποτελεσματικά μοντέλα ML. Ωστόσο, η υιοθέτηση του ίδιου του AutoML παρουσιάζει ένα νέο σύνολο προκλήσεων, συμπεριλαμβανομένης της διασφάλισης της ερμηνευσιμότητας και της διαφάνειας των μοντέλων που δημιουργούνται αυτόματα, της διατήρησης του απορρήτου και της

ασφάλειας των δεδομένων και της επίτευξης ισορροπίας μεταξύ αυτοματισμού και ανάγκης για προσαρμοσμένες λύσεις σε διαφορετικά ή ειδικά περιβάλλοντα. Ο πρωταρχικός στόχος του AutoML να καταστήσει το ML πιο προσιτό και αποτελεσματικό αντιπαραβάλλεται με την ανάγκη ανάπτυξης συστημάτων AutoML που είναι προσαρμόσιμα, αξιόπιστα και ευθυγραμμισμένα με ηθικά πρότυπα. Έτσι, η δήλωση προβλήματος περιστρέφεται γύρω από την ενίσχυση των δυνατοτήτων και την υιοθέτηση του AutoML ώστε να ξεπεραστούν τα εμπόδια που είναι εγγενή στις παραδοσιακές ροές εργασίας ML, ενώ παράλληλα αντιμετωπίζει τις αναδυόμενες προκλήσεις που σχετίζονται με την αυτοματοποίηση στην ανάπτυξη μοντέλων, για να αξιοποιήσει πλήρως τις δυνατότητες της ML στην προώθηση της καινοτομίας σε διάφορους τομείς.

Στον τομέα της μηχανικής μάθησης (ML) και ειδικότερα στο πλαίσιο της αυτοματοποιημένης μηχανικής μάθησης (AutoML), η επιλογή των κατάλληλων αλγορίθμων και η επακόλουθη ρύθμιση των υπερπαραμέτρων τους αποτελούν κομβικά στάδια που επηρεάζουν σημαντικά την απόδοση και την αποτελεσματικότητα των μοντέλων ML. Η επιλογή αλγορίθμων περιλαμβάνει τον προσδιορισμό της καταλληλότερης μεθόδου ML (π.χ. δέντρα αποφάσεων, νευρωνικά δίκτυα, μηχανές διανυσμάτων υποστήριξης) για ένα σύνολο δεδομένων και τύπο προβλήματος, το οποίο δεν είναι μια ασήμαντη εργασία λόγω του θεωρήματος "No Free Lunch". Αυτό το θεώρημα υποθέτει ότι κανένας μεμονωμένος αλγόριθμος δεν ξεπερνά καθολικά όλους τους άλλους σε όλα τα πιθανά προβλήματα, υπονοώντας την ανάγκη για μια προσεκτική και συγκεκριμένη διαδικασία επιλογής [5].

Ένα σημαντικό κομμάτι της διαδικασίας AutoML είναι η χρήση μεταχαρακτηριστικών (meta-features). Τα μεταχαρακτηριστικά είναι περιγραφικά στατιστικά και ιδιότητες των συνόλων δεδομένων που μπορούν να χρησιμοποιηθούν για την κατανόηση των δεδομένων και την καθοδήγηση της επιλογής αλγορίθμων. Παραδείγματα μεταχαρακτηριστικών περιλαμβάνουν τον αριθμό των δειγμάτων και των χαρακτηριστικών, την κατανομή των τιμών και την παρουσία κενών τιμών. Με τη χρήση αυτών των πληροφοριών, τα συστήματα AutoML μπορούν να βελτιώσουν την απόδοση και την αποτελεσματικότητα των μοντέλων ML. Τα μεταχαρακτηριστικά επιτρέπουν στα συστήματα ML να μαθαίνουν από προηγούμενες εμπειρίες και να εφαρμόζουν αυτήν τη γνώση σε νέα σύνολα δεδομένων. Με τη συλλογή και ανάλυση μεγάλου αριθμού μεταχαρακτηριστικών από διάφορα προβλήματα ML, τα συστήματα μπορούν να προβλέψουν ποιοι αλγόριθμοι και υπερπαράμετροι είναι πιθανό να αποδώσουν καλύτερα σε νέες εργασίες, μειώνοντας έτσι τον χρόνο και την προσπάθεια που απαιτούνται για την ανάπτυξη επιτυχημένων μοντέλων ML.

Ο συντονισμός υπερπαραμέτρων περιπλέκει περαιτέρω τη διαδικασία ανάπτυξης μοντέλων, καθώς οι περισσότεροι αλγόριθμοι ML έρχονται με υπερπαραμέτρους που πρέπει να βελτιστοποιηθούν. Αυτές οι υπερπαράμετροι, οι οποίες δε μαθαίνονται άμεσα από τα δεδομένα, μπορούν να επηρεάσουν σημαντικά τη διαδικασία μάθησης του μοντέλου και την τελική του απόδοση. Ο χειροκίνητος συντονισμός είναι συχνά κουραστικός και μη πρακτικός, ειδικά με την αυξανόμενη πολυπλοκότητα των αλγορίθμων και την απεραντοσύνη στον χώρο των υπερπαραμέτρων. Ως εκ τούτου, αυτοματοποιημένες τεχνικές βελτιστοποίησης υπερπαραμέτρων, όπως η αναζήτηση πλέγματος (grid search), η τυχαία αναζήτηση (random search), η μπεϋζιανή βελτιστοποίηση (Bayesian optimization) και πιο πρόσφατα, οι μεταευρετικοί αλγόριθμοι (metaheuristic algorithms), όπως οι γενετικοί αλγόριθμοι, έχουν γίνει αναπόσπαστα συστατικά των συστημάτων AutoML, με στόχο την αποτελεσματική αναζήτηση του χώρου υπερπαραμέτρων για βέλτιστες διαμορφώσεις [6].

Τα πλαίσια AutoML, όπως το Auto-sklearn, το TPOT και το H2O AutoML, ενσωματώνουν αυτά τα κρίσιμα βήματα, αυτοματοποιώντας την επιλογή αλγορίθμων και τη ρύθμιση των υπερπαραμέτρων τους για να κάνουν το ML πιο προσβάσιμο και να επιταχύνουν την ανάπτυξη μοντέλων υψηλής απόδοσης. Αυτά τα πλαίσια χρησιμοποιούν διάφορες στρατηγικές, συμπεριλαμβανομένων των μεθόδων μετα-μάθησης και μεθόδους συνόλων, για την αποτελεσματική πλοήγηση στις διαδικασίες επιλογής αλγορίθμων και συντονισμού υπερπαραμέτρων, δίνοντας έτσι

την δυνατότητα στην ML για μια μεγαλύτερη χρήση από περισσότερους χρήστες και επιτρέποντας ευρύτερη υιοθέτηση σε διάφορους τομείς [7],[8].

Η ενσωμάτωση της επιλογής αλγορίθμων και της ρύθμισης υπερπαραμέτρων στο AutoML όχι μόνο αντιμετωπίζει τα εμπόδια πολυπλοκότητας και τεχνογνωσίας που σχετίζονται με τις παραδοσιακές ροές εργασίας ML, αλλά παρουσιάζει επίσης νέες προκλήσεις. Αυτές περιλαμβάνουν τη διασφάλιση της επεκτασιμότητας των διαδικασιών αναζήτησης, τη διατήρηση της ερμηνευσιμότητας των επιλεγμένων μοντέλων και τη διασφάλιση της γενίκευσης των αυτοματοποιημένων αποφάσεων που λαμβάνονται από τα συστήματα AutoML. Καθώς ο τομέας εξελίσσεται, οι συνεχιζόμενες προσπάθειες έρευνας και ανάπτυξης επικεντρώνονται στην ενίσχυση της αποδοτικότητας, της αξιοπιστίας και της φιλικότητας προς το χρήστη των εργαλείων AutoML, υποσχόμενες έτσι να γεφυρώσουν περαιτέρω το χάσμα μεταξύ του αναπτυσσόμενου τοπίου δεδομένων και της αποτελεσματικής ανάπτυξης λύσεων ML.

1.2 Clustering

Η εποπτευόμενη(supervised learning) και η μη εποπτευόμενη μάθηση(unsupervised learning) αντιπροσωπεύουν τις δύο κύριες κατηγορίες μηχανικής μάθησης, καθεμία με ξεχωριστές μεθοδολογίες και εφαρμογές. Η εποπτευόμενη μάθηση χαρακτηρίζεται από τη χρήση επισημασμένων συνόλων δεδομένων για την εκπαίδευση αλγορίθμων, όπου κάθε σημείο δεδομένων εισόδου σχετίζεται με μια αντίστοιχη ετικέτα εξόδου. Αυτή η προσέγγιση επιτρέπει στο μοντέλο να μάθει μια συνάρτηση που χαρτογραφεί τις εισόδους στις επιθυμητές εξόδους, επιτρέποντας προβλέψεις ή ταξινομήσεις σε νέα δεδομένα. Η εποπτευόμενη μάθηση περιλαμβάνει ένα ευρύ φάσμα αλγορίθμων, συμπεριλαμβανομένης της γραμμικής παλινδρόμησης για συνεχή πρόβλεψη αποτελεσμάτων και λογιστική παλινδρόμηση, δέντρα αποφάσεων και νευρωνικά δίκτυα για εργασίες ταξινόμησης. Η αποτελεσματικότητα των εποπτευόμενων μοντέλων μάθησης αξιολογείται συχνά μέσω μετρήσεων, όπως η ακρίβεια, η αποτελεσματικότητα, η ανάκληση για εργασίες ταξινόμησης και το μέσο τετραγωνικό σφάλμα για εργασίες παλινδρόμησης. Η πρωταρχική πρόκληση στην εποπτευόμενη μάθηση έγκειται στην απόκτηση ενός επαρκώς μεγάλου και ολοκληρωμένου συνόλου δεδομένων με σήμανση, το οποίο μπορεί να είναι χρονοβόρο και δαπανηρό, αλλά είναι ζωτικής σημασίας για την εκπαίδευση ακριβών και γενικευμένων μοντέλων [9].

Η μάθηση χωρίς επίβλεψη, από την άλλη πλευρά, ασχολείται με δεδομένα που στερούνται ρητών ετικετών ή κατηγοριών εξόδου. Ο στόχος εδώ είναι να αποκαλυφθούν η υποκείμενη δομή ή μοτίβα μέσα στα δεδομένα. Η ανάλυση ομαδοποίησης, μια βασική τεχνική στη μάθηση χωρίς επίβλεψη, περιλαμβάνει την ομαδοποίηση σημείων δεδομένων σε ομάδες έτσι ώστε τα στοιχεία μέσα στο ίδιο σύμπλεγμα να μοιάζουν περισσότερο μεταξύ τους από εκείνα σε άλλα συμπλέγματα. Αυτό επιτυγχάνεται χωρίς προηγούμενη γνώση σχετικά με την ιδιότητα μέλους της ομάδας ή του συμπλέγματος των δεδομένων. Η ομαδοποίηση έχει ένα ευρύ φάσμα εφαρμογών, από την κατηγοριοποίηση πελατών στο μάρκετινγκ έως την ανάλυση γονιδιακής έκφρασης στη βιοπληροφορική. Αλγόριθμοι, όπως ο K-Means, ο hierarchical clustering και ο DBSCAN είναι από τους πιο ευρέως χρησιμοποιούμενους στην ανάλυση ομαδοποίησης, ο καθένας με τους δικούς του μηχανισμούς για τον ορισμό και τη βελτιστοποίηση συστάδων με βάση διάφορα μέτρα ομοιότητας ή πυκνότητας. Σε αντίθεση με την εποπτευόμενη μάθηση, όπου η απόδοση του μοντέλου μπορεί να αξιολογηθεί έναντι γνωστών labels, η μάθηση χωρίς επίβλεψη, ιδιαίτερα η ομαδοποίηση, συχνά βασίζεται σε εγγενείς μετρήσεις, όπως η βαθμολογία σιλουέτας (silhouette score) ή ο δείκτης Davies-Bouldin για τη μέτρηση της ποιότητας της ομαδοποίησης [10],[11]. Οι προσεγγίσεις ομαδοποίησης μπορούν πράγματι να κατηγοριοποιηθούν σε σκληρή διαμέριση (hard partitioning) και ασαφή διαμέριση (fuzzy partitioning) με βάση τον τρόπο με τον οποίο εκχωρούν σημεία δεδομένων σε συμπλέγματα.

1.3 Hard Partitioning

Στη σκληρή διαμέριση (hard partitioning), γνωστή και ως καθαρή διαμέριση, κάθε σημείο δεδομένων εκχωρείται ακριβώς σε ένα σύμπλεγμα, χωρίς επικάλυψη μεταξύ συμπλεγμάτων. Η εκχώρηση είναι οριστική, πράγμα που σημαίνει ότι ένα σημείο δεδομένων ανήκει εξ ολοκλήρου σε ένα σύμπλεγμα και καθόλου σε οποιοδήποτε άλλο. Ο K-Means είναι ένα χαρακτηριστικό παράδειγμα συμπλέγματος Hard Partitioning, όπου κάθε σημείο εκχωρείται στο πλησιέστερο κέντρο συμπλέγματος και τα μέλη ορίζονται χωρίς αμφιβολία. Η μαθηματική αναπαράσταση της σκληρής διαμέρισης μπορεί να εκφραστεί χρησιμοποιώντας μια συνάρτηση I_{ij} για ένα σημείο δεδομένων x_i και ένα cluster C_j :

$I_{ij} = 1$ έαν το x_i εκχωρείται στο σύμπλεγμα C_j. Αυτό σημαίνει ότι το σημείο δεδομένων x_i ανήκει στο σύμπλεγμα C_j , και κανένα άλλο.

$I_{ij} = 0$ έαν το x_i δεν εκχωρείται στο σύμπλεγμα C_j. . Αυτό σημαίνει ότι το σημείο δεδομένων x_i δεν ανήκει στο σύμπλεγμα C_j.

Αυτή η δυαδική αναπαράσταση (0 ή 1) χρησιμοποιείται για να υποδείξει την ιδιότητα μέλους σημείων δεδομένων σε συμπλέγματα για μεθόδους που περιλαμβάνουν σκληρή διαμέριση, όπου κάθε σημείο δεδομένων βρίσκεται αυστηρά στο ένα ή στο άλλο σύμπλεγμα, χωρίς επικάλυψη ή κοινή συμμετοχή.

1.4 Fuzzy Partitioning

Η ασαφής διαμέριση (fuzzy partitioning), από την άλλη πλευρά, επιτρέπει στα σημεία δεδομένων να ανήκουν σε πολλαπλά συμπλέγματα με ποικίλους βαθμούς συμμετοχής, εισάγοντας την έννοια της αβεβαιότητας στην κατηγοριοποίηση συμπλέγματος. Η προσέγγιση αυτή είναι ιδιαίτερα χρήσιμη σε σενάρια όπου τα όρια μεταξύ των συμπλεγμάτων δεν είναι σαφώς καθορισμένα. Ο αλγόριθμος Fuzzy C-Means (FCM) είναι μια δημοφιλής μέθοδος ασαφούς ομαδοποίησης όπου κάθε σημείο δεδομένων έχει έναν συντελεστή συμμετοχής για κάθε σύμπλεγμα, υποδεικνύοντας τη δύναμη της συσχέτισης ή της συμμετοχής σε αυτό το σύμπλεγμα. Το άθροισμα των συντελεστών συμμετοχής ενός σημείου σε όλα τα συμπλέγματα ισούται με 1. Ο συντελεστής coefficient ενός σημείου δεδομένων x_i σε ένα σύμπλεγμα C_j συμβολίζεται με u_{ij}, όπου u_{ij} κυμαίνεται από 0 έως 1, αντιπροσωπεύοντας τον βαθμό συμμετοχής:

$u_{ij} \in [0,1]$ and $\Sigma_{j=1}^{k} u_{ij}=1$ όπου k είναι ο αριθμός των συστάδων, και u_{ij} είναι ο βαθμός ιδιότητας μέλους του σημείου δεδομένων x_i στο cluster C_j.

Η διάκριση μεταξύ σκληρής και ασαφούς διαμέρισης υπογραμμίζει την προσαρμοστικότητα των τεχνικών ομαδοποίησης σε διαφορετικούς τύπους δεδομένων και προβληματικών τομέων. Ενώ η σκληρή διαμέριση παρέχει σαφείς, διακριτές ομαδοποιήσεις, η ασαφής διαμέριση καταγράφει τις αποχρώσεις και τις επικαλύψεις στα δεδομένα, προσφέροντας μια πιο λεπτομερή κατανόηση των σχέσεων σημείων δεδομένων και των συνθέσεων συμπλεγμάτων.

1.5 Algorithm Selection

Το πρόβλημα της επιλογής αλγορίθμων στη μηχανική μάθηση ενσωματώνει μια θεμελιώδη πρόκληση που εκτείνεται τόσο στις θεωρητικές όσο και στις πρακτικές διαστάσεις του πεδίου. Δεδομένης μιας συγκεκριμένης εργασίας ή συνόλου δεδομένων, ο καθορισμός του καταλληλότερου

αλγορίθμου μάθησης μπορεί να επηρεάσει σημαντικά την απόδοση και την αποτελεσματικότητα του μοντέλου. Αυτή η απόφαση δεν είναι τετριμμένη λόγω του θεωρήματος "No Free Lunch", το οποίο υποθέτει ότι κανένας αλγόριθμος δεν μπορεί να ξεπεράσει όλους τους άλλους σε όλα τα πιθανά προβλήματα [5]. Κατά συνέπεια, η επιλογή του αλγορίθμου εξαρτάται από τα χαρακτηριστικά των δεδομένων, τη φύση της εργασίας και τις επιθυμητές μετρήσεις απόδοσης, καθιστώντας την επιλογή αλγορίθμου ένα κεντρικό βήμα στο pipeline της μηχανικής μάθησης.

Στην πράξη, η διαδικασία επιλογής ενός κατάλληλου αλγορίθμου συχνά καθοδηγείται από εμπειρικό πειραματισμό και γνώσης του συγκεκριμένου τομέα. Οι επαγγελματίες μπορούν να βασίζονται στην εμπειρία τους ή σε καθιερωμένα σημεία αναφοράς για να επιλέξουν ένα αρχικό σύνολο αλγορίθμων για εξερεύνηση. Ωστόσο, αυτή η προσέγγιση μπορεί να είναι χρονοβόρα και μπορεί να μην εγγυάται τον προσδιορισμό του βέλτιστου αλγορίθμου, ειδικά σε νέους ή σύνθετους προβληματικούς τομείς. Επιπλέον, η ταχεία πρόοδος της έρευνας μηχανικής μάθησης εισάγει συνεχώς νέους αλγορίθμους και παραλλαγές, επεκτείνοντας τη δεξαμενή επιλογών και περιπλέκοντας τη διαδικασία λήψης αποφάσεων. Αυτή η αυξανόμενη πολυπλοκότητα υπογραμμίζει την ανάγκη για συστηματικές και αυτοματοποιημένες προσεγγίσεις στην επιλογή αλγορίθμων, οι οποίες μπορούν να καθοδηγήσουν μέσα σε ένα εκτεταμένο αλγοριθμικό τοπίο πιο αποτελεσματικά και με μεγαλύτερη ακρίβεια [12].

Η αυτοματοποιημένη μηχανική μάθηση (AutoML) έχει αναδειχθεί ως ένα πολλά υποσχόμενο παράδειγμα για την αντιμετώπιση της πρόκλησης της επιλογής αλγορίθμων. Τα πλαίσια AutoML στοχεύουν στην αυτοματοποίηση της διαδικασίας επιλογής και διαμόρφωσης αλγορίθμων μηχανικής μάθησης για τη βελτιστοποίηση της απόδοσης σε μια δεδομένη εργασία. Αυτά τα συστήματα χρησιμοποιούν συχνά μετα-μάθηση, η οποία αξιοποιεί ιστορικά δεδομένα απόδοσης σε ένα ευρύ φάσμα εργασιών και συνόλων δεδομένων για να ενημερώσει τη διαδικασία επιλογής, καθώς και της Bayesian βελτιστοποίησης, η οποία παρέχει μια προσέγγιση με κανόνες για τη διερεύνηση του χώρου των αλγορίθμων και των υπερπαραμέτρων τους [7],[3]. Με την ενσωμάτωση αυτών των τεχνικών, τα εργαλεία AutoML προσπαθούν να κάνουν τη μηχανική μάθηση έτοιμη για χρήση σε ευρύ κοινό πιο εύκολα, καθιστώντας τα μοντέλα υψηλής απόδοσης, προσβάσιμα σε μη ειδικούς και εξορθολογίζοντας τη διαδικασία ανάπτυξης για έμπειρους επαγγελματίες. Παρά τις προόδους στο AutoML, η επιλογή αλγορίθμων παραμένει ένας τομέας ενεργούς έρευνας, με συνεχείς προσπάθειες για την ενίσχυση της προσαρμοστικότητας, της αποτελεσματικότητας και της διαφάνειας αυτών των αυτοματοποιημένων συστημάτων.

1.6 Hyperparameter optimization HPO

Η βελτιστοποίηση υπερπαραμέτρων (HPO) είναι μια βασική διαδικασία στην ανάπτυξη μοντέλων μηχανικής μάθησης που περιλαμβάνει την εύρεση του βέλτιστου συνόλου υπερπαραμέτρων που αποδίδει την καλύτερη απόδοση. Αυτή η βελτιστοποίηση είναι ζωτικής σημασίας επειδή η επιλογή των υπερπαραμέτρων μπορεί να επηρεάσει σημαντικά την ικανότητα του αλγορίθμου μάθησης να μοντελοποιεί αποτελεσματικά το σύνολο των δεδομένων που μας ενδιαφέρει. Ο απώτερος στόχος του HPO είναι να αναζητήσει και να εντοπίσει με αποτελεσματικό τρόπο, μέσω του χώρου υπερπαραμέτρων, το συνδυασμό που μεγιστοποιεί την απόδοση του μοντέλου, που συχνά μετράται με μετρικές, όπως ακρίβεια, αποτελεσματικότητα, ανάκληση ή F1 score για περιπτώσεις ταξινόμησης και μέσο τετραγωνικό σφάλμα ή μέσο απόλυτο σφάλμα για περιπτώσεις παλινδρόμησης.

Οι υπερπαράμετροι είναι κρίσιμες για την αρχιτεκτονική των μοντέλων μηχανικής μάθησης. Είναι οι ρυθμίσεις για διαμόρφωση που χρησιμοποιούνται για τη δομή της μαθησιακής διαδικασίας. Σε αντίθεση με τις παραμέτρους του μοντέλου που μαθαίνονται απευθείας από τα δεδομένα εκπαίδευσης κατά τη διάρκεια της φάσης εκπαίδευσης του μοντέλου, οι υπερπαράμετροι ορίζονται

πριν από τη διαδικασία εκπαίδευσης και παραμένουν σταθερές κατά τη διάρκεια αυτής. Παραδείγματα υπερπαραμέτρων περιλαμβάνουν το ρυθμό μάθησης στην κάθοδο κλίσης, το βάθος ενός δέντρου αποφάσεων, τον αριθμό των κρυμμένων στρωμάτων και νευρώνων σε ένα νευρωνικό δίκτυο και τους όρους κανονικοποίησης στα μοντέλα παλινδρόμησης.

Η επιλογή των κατάλληλων υπερπαραμέτρων μπορεί να επηρεάσει σημαντικά την απόδοση ενός μοντέλου μηχανικής μάθησης. Μια πολύ μεγάλη τιμή για το ρυθμό μάθησης σε έναν αλγόριθμο βελτιστοποίησης μπορεί να προκαλέσει το μοντέλο να υπερβεί το ελάχιστο της συνάρτησης απώλειας, ενώ μια πολύ μικρή τιμή μπορεί να οδηγήσει σε μια οδυνηρά αργή σύγκλιση. Ομοίως, ένα υπερβολικά βαθύ δέντρο αποφάσεων μπορεί να υπερκαλύπτει τα δεδομένα εκπαίδευσης, συλλαμβάνοντας θόρυβο και όχι την υποκείμενη κατανομή δεδομένων, ενώ ένα πολύ ρηχό δέντρο μπορεί να μην ταιριάζει, αποτυγχάνοντας να συλλάβει σημαντικά μοτίβα. Έτσι, ο συντονισμός υπερπαραμέτρων, η διαδικασία επιλογής του συνόλου των βέλτιστων υπερπαραμέτρων για έναν αλγόριθμο μάθησης, γίνεται ένα κρίσιμο βήμα για την οικοδόμηση αποτελεσματικών μοντέλων μηχανικής μάθησης.

Ο συντονισμός υπερπαραμέτρων (hyperparameter tuning) παραδοσιακά περιλαμβάνει μεθόδους, όπως η grid search, όπου αξιολογείται συστηματικά ένα προκαθορισμένο εύρος τιμών για κάθε υπερπαράμετρο, η random search η οποία δειγματίζει τυχαία τον χώρο υπερπαραμέτρων και είναι συχνά πιο αποτελεσματική από την αναζήτηση πλέγματος και χειροκίνητο συντονισμό (manual tuning), βασιζόμενοι στη διαίσθηση και την εμπειρία του ασκούμενου. Ωστόσο, αυτές οι μέθοδοι μπορεί να είναι χρονοβόρες και μπορεί να μην εγγυώνται την εύρεση του βέλτιστου συνόλου υπερπαραμέτρων, ειδικά καθώς αυξάνεται η διάσταση του χώρου υπερπαραμέτρων.

Η αυτοματοποιημένη μηχανική μάθηση (AutoML) αναδεικνύεται ως λύση στις προκλήσεις της ρύθμισης υπερπαραμέτρων, με στόχο την αυτοματοποίηση της διαδικασίας εφαρμογής της μηχανικής μάθησης. Τα εργαλεία AutoML χρησιμοποιούν πιο εξελιγμένες και αποτελεσματικές μεθόδους βελτιστοποίησης υπερπαραμέτρων, όπως βελτιστοποίηση Bayesian, βελτιστοποίηση gradient-based και εξελικτικούς αλγόριθμους. Η μπεϋζιανή βελτιστοποίηση, για παράδειγμα, δημιουργεί ένα μοντέλο στατιστικών πιθανοτήτων της συνάρτησης που αντιστοιχίζει τις υπερπαραμέτρους, στην μετρική αξιολόγησης στόχου και το χρησιμοποιεί για να επιλέξει τις υπερπαραμέτρους με τις περισσότερες πιθανότητες επιτυχίας, για αξιολόγηση στην πραγματική αντικειμενική συνάρτηση. Αυτή η προσέγγιση είναι πιο αποτελεσματική από την αναζήτηση πλέγματος (grid search) ή τυχαίας αναζήτησης (random search), καθώς αξιοποιεί τα αποτελέσματα προηγούμενων αξιολογήσεων για τη βελτίωση της διαδικασίας αναζήτησης [13].

Η ενσωμάτωση του AutoML στη ρύθμιση υπερπαραμέτρων όχι μόνο επιταχύνει τον κύκλο ανάπτυξης του μοντέλου, αλλά την κάνει και προσβάσιμη σε επαγγελματίες χωρίς βαθιά εμπειρία στον τομέα. Οι πλατφόρμες AutoML, όπως το Auto-sklearn, το H2O AutoML και το Cloud AutoML της Google παρέχουν φιλικές προς το χρήστη διεπαφές που αφαιρούν την πολυπλοκότητα του συντονισμού υπερπαραμέτρων, επιτρέποντας στους χρήστες να επικεντρωθούν περισσότερο στην επίλυση προβλημάτων και λιγότερο στις περίπλοκες ρυθμίσεις των αλγορίθμων [2],[14].

1.7 Στρατηγικές για Hyperparameter Optimization:

Grid Search:

Η grid search είναι μία από τις απλούστερες μορφές HPO, όπου ορίζεται ένα πλέγμα τιμών υπερπαραμέτρων και πραγματοποιείται εξαντλητική αναζήτηση σε αυτό το πλέγμα. Για κάθε συνδυασμό υπερπαραμέτρων, το μοντέλο εκπαιδεύεται και αξιολογείται η απόδοσή του. Ο κύριος

περιορισμός της αναζήτησης πλέγματος είναι η υπολογιστική της αναποτελεσματικότητας, ειδικά καθώς αυξάνεται ο αριθμός των υπερπαραμέτρων και το μέγεθος του χώρου αναζήτησης.

Random Search:

Η random search αντιμετωπίζει ορισμένες από τις ανεπάρκειες της grid search επιλέγοντας τυχαίους συνδυασμούς υπερπαραμέτρων για αξιολόγηση. Ενώ μπορεί να φαίνεται λιγότερο συστηματική, η τυχαία αναζήτηση μπορεί συχνά να φτάσει σε σχεδόν βέλτιστες διαμορφώσεις γρηγορότερα από την αναζήτηση πλέγματος, ειδικά σε χώρους υψηλών διαστάσεων, καθώς δεν είναι όλες οι υπερπαράμετροι εξίσου σημαντικές για την απόδοση του μοντέλου [6].

Bayesian Optimization:

Η bayesian optimization είναι μια πιο εξελιγμένη προσέγγιση που δημιουργεί ένα μοντέλο πιθανοτήτων της αντικειμενικής συνάρτησης και το χρησιμοποιεί για να επιλέξει τις πιο ελπιδοφόρες υπερπαραμέτρους προς αξιολόγηση. Αυτή η μέθοδος είναι ιδιαίτερα αποτελεσματική καθώς ενσωματώνει προηγούμενη γνώση για τη λήψη εξυπνότερων αποφάσεων σχετικά με τα σημεία στον υπερπαραμετρικό χώρο που θα εξερευνήσουμε στη συνέχεια, εξισορροπώντας την εξερεύνηση και την εκμετάλλευση [13].

Gradient-Based Optimization:

Για ορισμένα μοντέλα, ειδικά μοντέλα βαθιάς μάθησης, είναι δυνατή η χρήση μεθόδων που βασίζονται σε διαβαθμίσεις για τη βελτιστοποίηση των υπερπαραμέτρων. Αυτή η προσέγγιση απαιτεί η αντικειμενική συνάρτηση να είναι διαφοροποιήσιμη σε σχέση με τις υπερπαραμέτρους, κάτι που δεν είναι πάντα εφικτό για όλους τους τύπους μοντέλων.

Evolutionary Algorithms:

Οι εξελικτικοί αλγόριθμοι, εμπνευσμένοι από τη διαδικασία της φυσικής επιλογής, χρησιμοποιούν μηχανισμούς, όπως η μετάλλαξη, η διασταύρωση και η επιλογή για να εξελίξουν ένα σύνολο υπερπαραμέτρων προς καλύτερη απόδοση. Οι γενετικοί αλγόριθμοι είναι ένα δημοφιλές παράδειγμα αυτής της προσέγγισης, όπου ένας πληθυσμός συνόλων υπερπαραμέτρων εξελίσσεται σε αρκετές γενιές, με τα σύνολα με τις καλύτερες επιδόσεις να τροποποιούνται και να συνδυάζονται για να παράγουν την επόμενη γενιά [15].

Meta-Learning:

Η μετα-μάθηση περιλαμβάνει τη μάθηση από προηγούμενες εμπειρίες κατάρτισης μοντέλων για την πρόβλεψη των καλύτερων υπερπαραμέτρων για ένα νέο σύνολο δεδομένων. Αξιοποιώντας ιστορικά δεδομένα απόδοσης, η μετα-μάθηση μπορεί να αποτελέσει ένα καλό σημείο εκκίνησης για τη βελτιστοποίηση υπερπαραμέτρων, μειώνοντας τον χώρο αναζήτησης και το υπολογιστικό κόστος [16].

1.8 Προκλήσεις και μελλοντικές κατευθύνσεις

Ενώ το HPO μπορεί να βελτιώσει σημαντικά την απόδοση του μοντέλου, εισάγει επίσης προκλήσεις, συμπεριλαμβανομένου του υπολογιστικού κόστους, ειδικά για πολύπλοκα μοντέλα και μεγάλα σύνολα δεδομένων. Επιπλέον, η στοχαστική φύση πολλών αλγορίθμων μάθησης προσθέτει ένα επίπεδο αβεβαιότητας στη διαδικασία βελτιστοποίησης. Οι μελλοντικές κατευθύνσεις στο HPO μπορεί να περιλαμβάνουν πιο προηγμένες προσεγγίσεις μετα-μάθησης, καλύτερη ενσωμάτωση με αγωγούς AutoML και την ανάπτυξη πιο υπολογιστικά αποτελεσματικών αλγορίθμων βελτιστοποίησης που μπορούν να χειριστούν την πολυπλοκότητα των σύγχρονων εργασιών μηχανικής μάθησης. Παρά τις εξελίξεις στο AutoML, ο συντονισμός υπερπαραμέτρων παραμένει ένας τομέας ενεργού έρευνας. Προκλήσεις, όπως η αποτελεσματική εξερεύνηση χώρων υψηλών διαστάσεων, η αντιμετώπιση της στοχαστικής φύσης πολλών αλγορίθμων μάθησης και η διασφάλιση της αναπαραγωγιμότητας και της ερμηνευσιμότητας της διαδικασίας συντονισμού αποτελούν αντικείμενο συνεχιζόμενης έρευνας. Καθώς η μηχανική μάθηση συνεχίζει να εξελίσσεται, η ανάπτυξη πιο αποτελεσματικών, ισχυρών και διαισθητικών μεθόδων συντονισμού υπερπαραμέτρων θα είναι ζωτικής σημασίας για την αξιοποίηση του πλήρους δυναμικού των μοντέλων μηχανικής μάθησης σε διάφορους τομείς.

1.9 Joint problem

Το κοινό πρόβλημα της συνδυασμένης επιλογής αλγορίθμων και βελτιστοποίησης υπερπαραμέτρων, κοινώς συντομογραφία CASH, αποτελεί κεντρική πρόκληση στον τομέα της αυτοματοποιημένης μηχανικής μάθησης (AutoML). Εκτείνεται πέρα από τις μεμονωμένες εργασίες επιλογής του καλύτερου αλγορίθμου μηχανικής μάθησης για ένα δεδομένο σύνολο δεδομένων και συντονισμού των υπερπαραμέτρων του. Αντ 'αυτού, το CASH αντιμετωπίζει αυτά τα δύο κρίσιμα ζητήματα ταυτόχρονα, αναγνωρίζοντας ότι η βέλτιστη επιλογή αλγορίθμου είναι εγγενώς συνυφασμένη με τη συγκεκριμένη διαμόρφωση των υπερπαραμέτρων του. Αυτό το πρόβλημα διατυπώνεται ως εύρεση του καλύτερου αλγορίθμου A από ένα σύνολο υποψήφιων αλγορίθμων A και της βέλτιστης διαμόρφωσης υπερπαραμέτρων λ από τον αντίστοιχο χώρο υπερπαραμέτρων Λ_A για κάθε αλγόριθμο, έτσι ώστε η απόδοση του αλγορίθμου A με υπερπαραμέτρους λ σε ένα σύνολο δεδομένων D μεγιστοποιείται [17].

Η αποτελεσματική αντιμετώπιση του προβλήματος του CASH είναι ζωτικής σημασίας για την αυτοματοποίηση του pipeline της μηχανικής μάθησης και για να καταστούν τα μοντέλα υψηλής απόδοσης προσβάσιμα τόσο σε εμπειρογνώμονες όσο και σε μη ειδικούς. Η πολυπλοκότητα αυτού του προβλήματος προκύπτει από την απεραντοσύνη και την ετερογένεια του συνδυασμένου χώρου των αλγορίθμων και των υπερπαραμέτρων τους, γεγονός που μπορεί να οδηγήσει σε συνδυαστική έκρηξη και να καταστήσει εξαντλητικές στρατηγικές αναζήτησης οι οποίες να καθίστανται υπολογιστικά ανέφικτες. Επιπλέον, το τοπίο απόδοσης στο πρόβλημα του CASH, είναι συχνά μη κυρτό και γεμάτο με λύσεις που είναι βέλτιστες σε ένα συγκεκριμένο περιορισμένο πεδίο (local optima), αυξάνοντας την πρόκληση. Έχουν αναπτυχθεί διάφορες στρατηγικές για την αντιμετώπιση του προβλήματος CASH, συμπεριλαμβανομένης της μετα-μάθησης, η οποία αξιοποιεί ιστορικά δεδομένα απόδοσης για να καθοδηγήσει την αναζήτηση του καλύτερου αλγορίθμου και διαμόρφωσης υπερπαραμέτρων, και της Bayesian βελτιστοποίησης, η οποία πλοηγείται αποτελεσματικά στον χώρο αναζήτησης εξισορροπώντας την εξερεύνηση νέων διαμορφώσεων με την εκμετάλλευση γνωστών πολλά υποσχόμενων περιοχών [7].

Οι συνέπειες της επίλυσης του προβλήματος CASH εκτείνονται πολύ πέρα από την απλή υπολογιστική αποτελεσματικότητα. Αυτοματοποιώντας την επιλογή των αλγορίθμων και τη διαμόρφωσή τους, τα συστήματα AutoML στοχεύουν στην χρήση της μηχανικής μάθησης από

ευρύτερο κοινό, επιτρέποντας σε επαγγελματίες χωρίς βαθιά εμπειρία μηχανικής μάθησης να αναπτύξουν αποτελεσματικά μοντέλα. Αυτό έχει τη δυνατότητα να επιταχύνει την καινοτομία και την εφαρμογή της μηχανικής μάθησης σε διάφορους τομείς, όπως τους τομείς της υγειονομικής περίθαλψης και της χρηματοδότησης, έως την ενέργεια και τις μεταφορές. Ωστόσο, το πρόβλημα των CASH θέτει επίσης προκλήσεις όσον αφορά στην ερμηνευσιμότητα και στη διαφάνεια των αυτοματοποιημένων αποφάσεων, καθώς το σκεπτικό πίσω από την επιλογή ενός συγκεκριμένου αλγορίθμου και συνόλου υπερπαραμέτρων μπορεί να μην είναι άμεσα εμφανές. Οι μελλοντικές ερευνητικές κατευθύνσεις στο AutoML και το πρόβλημα CASH περιλαμβάνουν την ανάπτυξη πιο αποτελεσματικών στρατηγικών αναζήτησης, τη βελτίωση της ευρωστίας και της γενίκευσης των λύσεων και την ενίσχυση της ερμηνευσιμότητας των συστημάτων AutoML για την ενίσχυση της εμπιστοσύνης και της υιοθέτησης μεταξύ των χρηστών [3].

2.Meta-Learning: Γεφύρωση γνώσεων μεταξύ εργασιών και μετα-χαρακτηριστικά

2.1 Εισαγωγή

Η μετα-μάθηση, ή «μαθαίνω για να μάθω», βρίσκεται στην πρώτη γραμμή των ερευνητικών συνόρων της μηχανικής μάθησης, προσφέροντας ένα παράδειγμα που επιτρέπει στους αλγόριθμους να αξιοποιήσουν προηγούμενες μαθησιακές εμπειρίες για να ενισχύσουν τις μελλοντικές μαθησιακές εργασίες. Η ουσία της μετα-μάθησης έγκειται στην ικανότητά της να γενικεύει τη γνώση που αποκτάται από μια ποικιλία μαθησιακών εργασιών και να εφαρμόζει αυτή τη συγκεντρωτική σοφία για την αντιμετώπιση νέων προβλημάτων που δεν έχουν εμφανιστεί μέχρι τότε, πιο αποδοτικά και αποτελεσματικά. Αυτή η προσέγγιση έρχεται σε αντίθεση με τις παραδοσιακές μεθόδους μηχανικής μάθησης, οι οποίες συνήθως ξεκινούν από το μηδέν για κάθε νέα εργασία, χωρίς να επωφελούνται από προηγούμενες μαθησιακές εμπειρίες [18].

Ένα βασικό συστατικό της μετα-μάθησης είναι η έννοια των μετα-χαρακτηριστικών, τα οποία είναι χαρακτηριστικά υψηλού επιπέδου που εξάγονται από σύνολα δεδομένων για να αναπαραστήσουν τις ιδιότητές τους, όπως η πολυπλοκότητα, η ποικιλομορφία ή η παρουσία θορύβου. Τα μετα-χαρακτηριστικά μπορούν να περιλαμβάνουν, μεταξύ άλλων, απλά περιγραφικά στατιστικά χαρακτηριστικά, θεωρητικά-πληροφοριακά μέτρα, χαρακτηριστικά βασισμένα σε μοντέλα και αποτελέσματα ορόσημων. Αναλύοντας αυτά τα μετα-χαρακτηριστικά, ένα σύστημα μετα-μάθησης μπορεί να εντοπίσει ομοιότητες μεταξύ νέων εργασιών και εργασιών που εκτελέστηκαν προηγουμένως, καθοδηγώντας την επιλογή αλγορίθμων ή υπερπαραμέτρων που είναι πιθανό να αποδώσουν καλά στη νέα εργασία με βάση τις προηγούμενες επιδόσεις σε παρόμοιες εργασίες [19].

2.2 Ο ρόλος των μετα-χαρακτηριστικών στη μετα-μάθηση

Τα μετα-χαρακτηριστικά διαδραματίζουν κεντρικό ρόλο στη μετα-μάθηση παρέχοντας μια περιγραφική περίληψη των εγγενών χαρακτηριστικών των συνόλων δεδομένων. Για παράδειγμα, απλά στατιστικά μετα-χαρακτηριστικά μπορούν να περιγράψουν τον αριθμό των στιγμιότυπων και των χαρακτηριστικών, την ασυμμετρία και την κύρτωση των χαρακτηριστικών των κατανομών ή την ισορροπία των κλάσεων-στόχων. Τα πληροφοριακά-θεωρητικά μέτρα, όπως η εντροπία και η αμοιβαία πληροφόρηση, προσφέρουν πληροφορίες σχετικά με την πολυπλοκότητα και τον πλεονασμό εντός των δεδομένων. Τα χαρακτηριστικά που βασίζονται σε μοντέλα μπορεί να

περιλαμβάνουν την απόδοση απλών μοντέλων, όπως δέντρα αποφάσεων ή τους πλησιέστερους γείτονες στο σύνολο δεδομένων, χρησιμεύοντας ως επιδόσεις που υποδηλώνουν την ελκυστικότητα του συνόλου δεδομένων. Τα αποτελέσματα που θα αποτελούν ορόσημο, συγκεκριμένα, είναι μετα-χαρακτηριστικά που βασίζονται στην απόδοση απλών και γρήγορων αλγορίθμων, παρέχοντας μια χονδρική βάση για το τι μπορεί να επιτευχθεί σε ένα σύνολο δεδομένων και ενημερώνοντας έτσι τη διαδικασία μετα-μάθησης σχετικά με δυνητικά κατάλληλα και πολύπλοκα μοντέλα [20].

Η εξαγωγή και η ανάλυση των μετα-χαρακτηριστικών επιτρέπει στο σύστημα μετα-μάθησης να κατασκευάσει μια βάση μετα-γνώσης, ένα αποθετήριο γνώσεων που συλλέγονται από διάφορες μαθησιακές εργασίες. Αυτή η μετα-γνωσιακή βάση υποστηρίζει την ικανότητα του συστήματος να λαμβάνει τεκμηριωμένες αποφάσεις σχετικά με την επιλογή αλγορίθμων, τη ρύθμιση υπερπαραμέτρων, ακόμη και τα βήματα προεπεξεργασίας δεδομένων για νέες εργασίες. Με τη χαρτογράφηση των μετα-χαρακτηριστικών ενός νέου συνόλου δεδομένων σε αυτή τη βάση γνώσεων, το σύστημα μπορεί να εντοπίσει ιστορικές εργασίες που είναι παρόμοιας φύσης και να χρησιμοποιήσει τις αντίστοιχες επιτυχημένες στρατηγικές ως σημείο εκκίνησης, υπερπηδώντας έτσι την κατά τα άλλα μακρά και υπολογιστικά δαπανηρή διαδικασία μάθησης από το μηδέν.

2.3 Προκλήσεις και μελλοντικές κατευθύνσεις

Ενώ η μετα-μάθηση και η χρήση μετα-χαρακτηριστικών υπόσχονται πολλά για την προώθηση της αποτελεσματικότητας και της προσαρμοστικότητας των συστημάτων μηχανικής μάθησης, θέτουν επίσης σημαντικές προκλήσεις. Η επιλογή και η μηχανική των πληροφοριακών μετα-χαρακτηριστικών που συλλαμβάνουν την ουσία των μαθησιακών εργασιών, είναι μια μη τετριμμένη προσπάθεια που απαιτεί προσεκτική εξέταση και γνώση του τομέα. Επιπλέον, η ίδια η διαδικασία μετα-μάθησης πρέπει να σχεδιαστεί έτσι ώστε να αξιοποιεί αποδοτικά και αποτελεσματικά τη βάση μετα-γνώσης, απαιτώντας εξελιγμένους αλγόριθμους ικανούς να διακρίνουν διαφοροποιημένα πρότυπα και σχέσεις εντός των μεταδεδομένων.

Η μελλοντική έρευνα στη μετα-μάθηση και τα μετα-χαρακτηριστικά είναι πιθανό να διερευνήσει πιο προηγμένες μεθόδους για την εξαγωγή και επιλογή μετα-χαρακτηριστικών, αξιοποιώντας τεχνικές από τη βαθιά μάθηση και την εκμάθηση μη εποπτευόμενων χαρακτηριστικών για τον αυτόματο εντοπισμό των πιο σημαντικών χαρακτηριστικών των συνόλων δεδομένων. Επιπλέον, η ενσωμάτωση της μετα-μάθησης με άλλα αναδυόμενα παραδείγματα, όπως η μάθηση μεταφοράς (transfer learning) και η μάθηση λίγων βολών (few-shot learning), παρουσιάζει μια συναρπαστική οδό για την ανάπτυξη πιο ισχυρών και ευέλικτων συστημάτων μάθησης που μπορούν να προσαρμοστούν γρήγορα σε ένα ευρύ φάσμα εργασιών με ελάχιστα δεδομένα. Καθώς το πεδίο εξελίσσεται, ο απώτερος στόχος θα είναι η δημιουργία συστημάτων μετα-μάθησης που όχι μόνο υπερέχουν στην επιλογή αλγορίθμων και τη βελτιστοποίηση υπερπαραμέτρων, αλλά και ενσωματώνουν το ευρύτερο όραμα αυτόνομων παραγόντων μάθησης ικανών για συνεχή μάθηση και προσαρμογή σε δυναμικά περιβάλλοντα.

Οι περίπλοκες καταστάσεις της μετα-μάθησης επεκτείνονται στη σφαίρα της αλγοριθμικής προσαρμοστικότητας, όπου το σύστημα όχι μόνο προσδιορίζει τα καταλληλότερα μοντέλα μηχανικής μάθησης για μια δεδομένη εργασία, αλλά προσαρμόζει επίσης αυτά τα μοντέλα ως απάντηση στα εξελισσόμενα τοπία δεδομένων. Αυτή η προσαρμοστικότητα είναι ζωτικής σημασίας σε δυναμικά περιβάλλοντα όπου οι κατανομές δεδομένων μπορεί να μετατοπιστούν με την πάροδο του χρόνου, ένα φαινόμενο γνωστό ως μετατόπιση εννοιών (concept drift). Σε τέτοια σενάρια, τα συστήματα μετα-μάθησης που είναι εξοπλισμένα με την ικανότητα να αναγνωρίζουν αυτές τις αλλαγές μέσω αλλαγών στα μετα-χαρακτηριστικά μπορούν να προσαρμόσουν προληπτικά τους αλγόριθμους μάθησης ή τις παραμέτρους τους για να διατηρήσουν τη βέλτιστη απόδοση. Αυτό το επίπεδο ανταπόκρισης σηματοδοτεί ένα άλμα προς πραγματικά ευφυή συστήματα που μιμούνται την ανθρώπινη ευελιξία

μάθησης, βελτιώνοντας συνεχώς τις γνώσεις και τις στρατηγικές τους με βάση νέες εμπειρίες και ιδέες.

Επιπλέον, η συγχώνευση της μετα-μάθησης με άλλα εξελιγμένα παραδείγματα μηχανικής μάθησης, όπως η ενισχυτική μάθηση (reinforcement learning) και η αναζήτηση νευρωνικής αρχιτεκτονικής (neural architecture), προαναγγέλλει μια νέα εποχή συστημάτων AutoML. Αυτά τα προηγμένα συστήματα θα μπορούσαν αυτόνομα να διερευνήσουν και να καινοτομήσουν νέους αλγόριθμους μάθησης ή αρχιτεκτονικές νευρωνικών δικτύων προσαρμοσμένες σε συγκεκριμένες εργασίες, καθοδηγούμενες από τις αρχές της μετα-μάθησης. Με αυτόν τον τρόπο, θα μπορούσαν να ανακαλύψουν νέες λύσεις που ξεπερνούν τα υπάρχοντα μοντέλα, ωθώντας τα όρια του τι είναι εφικτό στη μηχανική μάθηση. Αυτή η οραματική προσέγγιση όχι μόνο ενισχύει τις δυνατότητες των μεμονωμένων μοντέλων, αλλά συμβάλλει και στη συλλογική πρόοδο του πεδίου, καθώς αυτές οι καινοτόμες λύσεις εμπλουτίζουν τη βάση μετα-γνώσης, δημιουργώντας έναν ενάρετο κύκλο μάθησης και ανακάλυψης.

Η εξέλιξη της μετα-μάθησης και η εφαρμογή της στην επιλογή αλγορίθμων και τη βελτιστοποίηση υπερπαραμέτρων συμπυκνώνει τη φιλοδοξία της κοινότητας μηχανικής μάθησης να αναπτύξει συστήματα που όχι μόνο μαθαίνουν από δεδομένα αλλά και μαθαίνουν πώς να μαθαίνουν πιο αποτελεσματικά. Καθώς η έρευνα σε αυτόν τον τομέα εξελίσσεται, η συνέργεια μεταξύ μετα-μάθησης, μετα-χαρακτηριστικών και άλλων μαθησιακών παραδειγμάτων υπόσχεται να καταλύσει σημαντικές καινοτομίες, καθιστώντας τη μηχανική μάθηση πιο προσιτή, αποτελεσματική και ισχυρή. Το ταξίδι προς την επίτευξη αυτού του οράματος θα σηματοδοτηθεί αναμφίβολα από προκλήσεις, αλλά οι πιθανές ανταμοιβές - όσον αφορά τόσο στις πρακτικές εφαρμογές όσο και στις θεωρητικές γνώσεις - το καθιστούν ένα βαθιά συναρπαστικό σύνορο στην προσπάθεια να ξεκλειδώσουμε όλες τις δυνατότητες της τεχνητής νοημοσύνης.

2.4 Πίνακας meta-features

Ο πίνακας 1 που επισυνάπτεται προσφέρει μια πολύτιμη συλλογή μετα-χαρακτηριστικών που χρησιμοποιούνται πιο συχνά, στην ανάλυση και κατανόηση συνόλων δεδομένων για εφαρμογές μηχανικής μάθησης. Τα μετα-χαρακτηριστικά που αναφέρονται, όπως οι «παρουσίες Nr» και τα «χαρακτηριστικά Nr», χρησιμεύουν για τη μετάδοση βασικών διαστάσεων των δεδομένων, όπως το μέγεθος και η πολυπλοκότητα, οι οποίες είναι κρίσιμες στα αρχικά στάδια της αξιολόγησης του συνόλου δεδομένων. Η ασυμμετρία και η κύρτωση είναι ζωτικής σημασίας για την κατανόηση της υποκείμενης κατανομής των χαρακτηριστικών μέσα στα δεδομένα, επηρεάζοντας τις αποφάσεις προεπεξεργασίας και την επιλογή των κατάλληλων αλγορίθμων. Οι μετρήσεις συσχέτισης παρέχουν πληροφορίες σχετικά με την αλληλεξάρτηση χαρακτηριστικών, οι οποίες μπορούν να ενημερώσουν τις τεχνικές επιλογής χαρακτηριστικών και μείωσης διαστάσεων, καθοδηγώντας τον επιστήμονα δεδομένων προς πιο ενημερωμένες στρατηγικές μοντελοποίησης.

Name	Formula	Rationale	Variants
Nr instances	n	Speed, Scalability (Michie et al., 1994)	p/n, $log(n)$, log(n/p)
Nr features	p	Curse of dimensionality (Michie et al., 1994)	$log(p)$, % categorical
Nr classes	c	Complexity, imbalance (Michie et al., 1994)	ratio min/maj class
Nr missing values	m	Imputation effects (Kalousis, 2002)	% missing
Nr outliers	o	Data noisiness (Rousseeuw and Hubert, 2011)	o/n
Skewness	$\frac{E(X-\mu_X)^3}{\sigma_X^3}$	Feature normality (Michie et al., 1994)	min,max,μ,σ,q_1,q_3
Kurtosis	$\frac{E(X-\mu_X)^4}{\sigma_X^4}$	Feature normality (Michie et al., 1994)	min,max,μ,σ,q_1,q_3
Correlation	$\rho_{X_1X_2}$	Feature interdependence (Michie et al., 1994)	min,max,μ,σ,ρ_{XY}
Covariance	$cov_{X_1X_2}$	Feature interdependence (Michie et al., 1994)	min,max,μ,σ,cov_{XY}
Concentration	$\tau_{X_1X_2}$	Feature interdependence (Kalousis and Hilario, 2001)	min,max,μ,σ,τ_{XY}
Sparsity	sparsity(X)	Degree of discreteness (Salama et al., 2013)	min,max,μ,σ
Gravity	gravity(X)	Inter-class dispersion (Ali and Smith-Miles, 2006a)	
ANOVA p-value	$p_{val_{\mathbf{X}_1X_2}}$	Feature redundancy (Kalousis, 2002)	$p_{val_{XY}}$ (Soares et al., 2004)
Coeff. of variation	$\frac{\sigma_Y}{\mu_Y}$	Variation in target (Soares et al., 2004)	
PCA ρ_{λ_1}	$\sqrt{\frac{\lambda_1}{1+\lambda_1}}$	Variance in first PC (Michie et al., 1994)	$\frac{\lambda_1}{\sum_i \lambda_i}$ (Michie et al., 1994)
PCA skewness		Skewness of first PC (Feurer et al., 2014)	PCA kurtosis
PCA 95%	$\frac{dim_{pca_{var}}}{p}$	Intrinsic dimensionality (Bardenet et al., 2013)	
Class probability	$P(\mathbf{C})$	Class distribution (Michie et al., 1994)	min,max,μ,σ
Class entropy	$H(\mathbf{C})$	Class imbalance (Michie et al., 1994)	
Norm. entropy	$\frac{H(\mathbf{X})}{log_2 n}$	Feature informativeness (Castiello et al., 2005)	min,max,μ,σ
Mutual inform.	$MI(\mathbf{C},\mathbf{X})$	Feature importance (Michie et al., 1994)	min,max,μ,σ
Uncertainty coeff.	$\frac{MI(\mathbf{C},\mathbf{X})}{H(\mathbf{C})}$	Feature importance (Agresti, 2002)	min,max,μ,σ
Equiv. nr. feats	$\frac{H(C)}{MI(C,X)}$	Intrinsic dimensionality (Michie et al., 1994)	
Noise-signal ratio	$\frac{H(X)-MI(C,X)}{MI(C,X)}$	Noisiness of data (Michie et al., 1994)	
Fisher's discrimin.	$\frac{(\mu_{c1}-\mu_{c2})^2}{\sigma_{c1}^2-\sigma_{c2}^2}$	Separability classes c_1, c_2 (Ho and Basu, 2002)	See Ho:2002
Volume of overlap		Class distribution overlap (Ho and Basu, 2002)	See Ho and Basu (2002)
Concept variation		Task complexity (Vilalta and Drissi, 2002)	See Vilalta (1999)
Data consistency		Data quality (Köpf and Iglezakis, 2002)	See Köpf and Iglezakis (2002)
Nr nodes, leaves	$\lvert\eta\rvert, \lvert\psi\rvert$	Concept complexity (Peng et al., 2002)	Tree depth
Branch length		Concept complexity (Peng et al., 2002)	min,max,μ,σ
Nodes per feature	$\lvert\eta_X\rvert$	Feature importance (Peng et al., 2002)	min,max,μ,σ
Leaves per class	$\frac{\lvert\psi\rvert}{\lvert\psi\rvert}$	Class complexity (Filchenkov and Pendryak, 2015)	min,max,μ,σ
Leaves agreement	$\frac{n_{\psi_i}}{n}$	Class separability (Bensusan et al., 2000)	min,max,μ,σ
Information gain		Feature importance (Bensusan et al., 2000)	min,max,μ,σ, gini
Landmarker(1NN)	$P(\theta_{1NN}, t_j)$	Data sparsity (Pfahringer et al., 2000)	See Pfahringer et al. (2000)
Landmarker(Tree)	$P(\theta_{Tree}, t_j)$	Data separability (Pfahringer et al., 2000)	Stump,RandomTree
Landmarker(Lin)	$P(\theta_{Lin}, t_j)$	Linear separability (Pfahringer et al., 2000)	Lin.Disciminant
Landmarker(NB)	$P(\theta_{NB}, t_j)$	Feature independence (Pfahringer et al., 2000)	See Ler et al. (2005)
Relative LM	$P_{a,j} - P_{b,j}$	Probing performance (Fürnkranz and Petrak, 2001)	
Subsample LM	$P(\theta_i, t_j, s_t)$	Probing performance (Soares et al., 2001)	

Πίνακας 1:Επισκόπηση των μετα-χαρακτηριστικών που χρησιμοποιούνται συνήθως. Ομάδες από πάνω προς τα κάτω: απλές, στατιστικές, θεωρητικές πληροφορίες, πολυπλοκότητα, βασισμένες σε μοντέλα και ορόσημα.

Κατά την ανάπτυξη του αναλυτικού πλαισίου μας, προέκυψε ένα ολοκληρωμένο σύνολο μετα-χαρακτηριστικών για να ενθυλακωθούν διάφορες διαστάσεις του υπό διερεύνηση συνόλου δεδομένων. Αυτά τα μετα-χαρακτηριστικά κατηγοριοποιούνται σχολαστικά σε τρεις διακριτές ομάδες, καθεμία από τις οποίες αντικατοπτρίζει μια διαφορετική πτυχή των εγγενών χαρακτηριστικών των δεδομένων. Απλές μετα-δυνατότητες, όπως ο αριθμός των παρουσιών και των δυνατοτήτων, παρέχουν μια θεμελιώδη κατανόηση της κλίμακας και της πολυπλοκότητας του συνόλου δεδομένων. Τα στατιστικά μετα-χαρακτηριστικά, συμπεριλαμβανομένων των μέτρων της κεντρικής τάσης, της μεταβλητότητας και των ιδιοτήτων κατανομής, όπως η ασυμμετρία και η κύρτωση, προσφέρουν βαθύτερες γνώσεις σχετικά με τη στατιστική φύση των δεδομένων. Τα πληροφοριακά-θεωρητικά μετα-χαρακτηριστικά, δηλαδή η εντροπία συνόλου δεδομένων και η μέγιστη εντροπία χαρακτηριστικών, ποσοτικοποιούν τον βαθμό αβεβαιότητας και το πληροφοριακό περιεχόμενο που ενσωματώνεται στο σύνολο δεδομένων. Αυτή η πολύπλευρη προσέγγιση στην εξαγωγή μετα-χαρακτηριστικών επιτρέπει μια λεπτή ανάλυση του συνόλου δεδομένων, διευκολύνοντας τον

εντοπισμό των υποκείμενων μοτίβων και ενημερώνοντας τις επακόλουθες αποφάσεις επεξεργασίας δεδομένων και μοντελοποίησης. Τα συνοπτικά μετα-χαρακτηριστικά και οι αντίστοιχες υπολογιστικές μεθοδολογίες τους περιγράφονται στον Πίνακα 2, χρησιμεύοντας ως κεντρικό σημείο αναφοράς για την επακόλουθη ανάλυση δεδομένων.

.

	Meta-feature	**Τύπος**	**Περιγραφή**
Simple Meta-features	num_instances	len(numeric_dataset)	Ο συνολικός αριθμός παρουσιών (γραμμών) στο σύνολο δεδομένων.
	num_features	numeric_dataset.shape[1]	Ο συνολικός αριθμός αριθμητικών χαρακτηριστικών (στηλών) στο σύνολο δεδομένων.
Statistical Meta-features	mean_feature_correlation	corr_matrix.mean().mean()	Ο μέσος όρος των μέσων συσχετίσεων μεταξύ όλων των ζευγών χαρακτηριστικών.
	std_dev_feature_correlation	corr_matrix.stack().std()	Η τυπική απόκλιση όλων των συσχετίσεων χαρακτηριστικών, παρέχοντας πληροφορίες σχετικά με την εξάπλωση συσχέτισης.
	max_feature_correlation	corr_matrix.stack().max()	Η μέγιστη τιμή συσχέτισης μεταξύ όλων των ζευγών χαρακτηριστικών.
	mean_feature_std_dev	numeric_dataset.std().mean()	Ο μέσος όρος των τυπικών αποκλίσεων κάθε χαρακτηριστικού, με ένδειξη της συνολικής μεταβλητότητας.
	median_feature_skewness	Numeric_dataset. apply(skew, nan_policy='omit').median()	Η διάμεση ασυμμετρία σε όλα τα χαρακτηριστικά, υποδεικνύοντας συμμετρία της κατανομής δεδομένων.
	dataset_skewness	numeric_dataset. apply(skew, nan_policy='omit').mean()	Η μέση ασυμμετρία σε όλα τα χαρακτηριστικά, που δείχνει την ασυμμετρία του συνόλου δεδομένων στο σύνολό του.
	dataset_kurtosis	numeric_dataset. apply(kurtosis, nan_policy='omit').mean()	Η μέση κύρτωση σε όλα τα χαρακτηριστικά, υποδεικνύοντας την «ουρά» της κατανομής δεδομένων.
	total_mean_of_column_means	numeric_dataset.mean().mean()	Ο μέσος όρος των μέσων κάθε χαρακτηριστικού, παρέχοντας μια συνολική κεντρική τάση.
	std_dev_of_column_std_devs	numeric_dataset.std().std()	Η τυπική απόκλιση των τυπικών αποκλίσεων κάθε χαρακτηριστικού, που δείχνει την εξάπλωση της μεταβλητότητας.
Information-Theoretic Meta-features	dataset_entropy	numeric_dataset. apply(calculate_entropy).mean()	Η μέση εντροπία σε όλα τα χαρακτηριστικά, που δείχνει το μέσο επίπεδο διαταραχής ή τυχαιότητας.
	numeric_dataset. apply(calculate_entropy).mean()	numeric_dataset. apply(calculate_entropy).max()	Η μέγιστη εντροπία μεταξύ όλων των χαρακτηριστικών, υποδεικνύοντας το χαρακτηριστικό με την υψηλότερη διαταραχή.

Πίνακας 2: Επισκόπηση των μετα-χαρακτηριστικών που χρησιμοποιήθηκαν στην παρούσα διατριβή

3.Εισαγωγή στην Ανάλυση Ομαδοποίησης

3.1 Εισαγωγή

Η ανάλυση ομαδοποίησης (clustering analysis) αποτελεί ακρογωνιαίο λίθο της μη εποπτευόμενης μάθησης στον τομέα της επιστήμης των δεδομένων και της μηχανικής μάθησης, προσφέροντας ένα ισχυρό μέσο για την αποκάλυψη εγγενών δομών εντός συνόλων δεδομένων χωρίς προκαθορισμένες ετικέτες ή κατηγορίες. Αυτή η μεθοδολογία είναι ζωτικής σημασίας σε μια πληθώρα εφαρμογών, που κυμαίνονται από την τμηματοποίηση πελατών στο μάρκετινγκ έως την ανάλυση γονιδιακής έκφρασης στη γονιδιωματική, λόγω της ικανότητάς της να οργανώνει τεράστιες ποσότητες δεδομένων σε σημαντικές ομάδες με βάση μετρήσεις ομοιότητας ή απόστασης.

Η ουσία της ανάλυσης ομαδοποίησης έγκειται στην ικανότητά της να κοσκινίζει μη δομημένα και σύνθετα σύνολα δεδομένων για να εντοπίζει μοτίβα, σχέσεις και ομαδοποιήσεις που δεν είναι άμεσα εμφανείς. Στο πλαίσιο αυτής της διατριβής, η ανάλυση ομαδοποίησης είναι απαραίτητη λόγω της δυνατότητάς της να αποκαλύψει κρυφές δομές μέσα στα δεδομένα, παρέχοντας πληροφορίες που είναι κρίσιμες για το συγκεκριμένο ερευνητικό ερώτημα. Είτε η εστίαση είναι στην κατανόηση της συμπεριφοράς των καταναλωτών, στην ανίχνευση δόλιων δραστηριοτήτων είτε στην αποκάλυψη βιολογικών γνώσεων από γονιδιωματικά δεδομένα, η ομαδοποίηση χρησιμεύει ως διερευνητικό εργαλείο που μπορεί να καθοδηγήσει τη διατύπωση υποθέσεων, την επιλογή χαρακτηριστικών και τις επακόλουθες αναλυτικές διαδικασίες.

3.2 Η ανάγκη για ανάλυση ομαδοποίησης

Η ανάλυση ομαδοποίησης είναι ιδιαίτερα πολύτιμη σε σενάρια όπου τα δεδομένα δεν φέρουν ετικέτα και οι σχέσεις μεταξύ των σημείων δεδομένων είναι άγνωστες. Αυτό συμβαίνει συχνά σε διερευνητικές ερευνητικές φάσεις όπου ο στόχος είναι η κατανόηση της υποκείμενης δομής των δεδομένων. Για παράδειγμα, στην έρευνα αγοράς, η ομαδοποίηση μπορεί να αποκαλύψει ξεχωριστές ομάδες πελατών με βάση τις αγοραστικές συνήθειες, τα δημογραφικά στοιχεία και τις προτιμήσεις, επιτρέποντας στοχευμένες στρατηγικές μάρκετινγκ και εξατομικευμένες εμπειρίες πελατών.

Στον επιστημονικό τομέα, όπως στη γονιδιωματική ή την πρωτεϊνωματική, η ανάλυση ομαδοποίησης χρησιμοποιείται για τον εντοπισμό ομάδων γονιδίων ή πρωτεϊνών που παρουσιάζουν παρόμοια πρότυπα έκφρασης, υποδηλώνοντας μια κοινή λειτουργία ή συμμετοχή σε παρόμοιες βιολογικές διεργασίες. Αυτό μπορεί να βοηθήσει σημαντικά στον σχολιασμό των γονιδίων, στην κατανόηση των μηχανισμών της νόσου και στην ανακάλυψη πιθανών θεραπευτικών στόχων.

3.3 Σημασία της ομαδοποίησης στην έρευνα

Η σημασία της ομαδοποίησης στην έρευνα δεν μπορεί να υπερεκτιμηθεί. Όχι μόνο διευκολύνει την ανακάλυψη εγγενών ομαδοποιήσεων μέσα στα δεδομένα, αλλά βοηθά επίσης στη συμπίεση και περίληψη δεδομένων, στην εξαγωγή χαρακτηριστικών και στην ανίχνευση ανωμαλιών. Μειώνοντας την πολυπλοκότητα των δεδομένων και επισημαίνοντας σημαντικά μοτίβα, η ανάλυση ομαδοποίησης επιτρέπει στους ερευνητές να επικεντρωθούν στις πιο σχετικές πτυχές των δεδομένων τους, καθιστώντας τις επακόλουθες αναλύσεις πιο εύχρηστες και ερμηνεύσιμες.

Επιπλέον, οι μέθοδοι ομαδοποίησης είναι απίστευτα ευέλικτες, με ένα ευρύ φάσμα αλγορίθμων διαθέσιμων για την κάλυψη διαφορετικών τύπων δεδομένων, κλιμάκων και υποθέσεων

διανομής. Από μεθόδους διαμέρισης, όπως ο K-Means, ιεραρχικές τεχνικές ομαδοποίησης, μεθόδους που βασίζονται στην πυκνότητα, όπως το DBSCAN, έως ομαδοποίηση βάσει μοντέλων, όπως μοντέλα Gaussian Mix, η επιλογή του αλγορίθμου μπορεί να προσαρμοστεί στα συγκεκριμένα χαρακτηριστικά και απαιτήσεις του εν λόγω συνόλου δεδομένων.

3.4 Clustering στο πλαίσιο της παρούσας διπλωματικής εργασίας

Σε αυτή τη διατριβή, η ανάλυση ομαδοποίησης χρησιμοποιείται ως θεμελιώδες βήμα για την πλοήγηση μέσα από την πολυπλοκότητα του συνόλου δεδομένων, με στόχο την αποκάλυψη μοτίβων που μπορούν να ενημερώσουν και να βελτιώσουν την κατεύθυνση της έρευνας. Η εφαρμογή της ομαδοποίησης θα πλαισιωθεί στο συγκεκριμένο ερευνητικό ερώτημα, αναφέροντας λεπτομερώς πώς η ταυτοποίηση των συστάδων μπορεί να οδηγήσει σε βαθύτερη κατανόηση των δεδομένων και να συμβάλει στην απάντηση του βασικού ερευνητικού ερωτήματος.

Η ενότητα της μεθοδολογίας θα επεξεργαστεί περαιτέρω την επιλογή του βέλτιστου αλγορίθμου ομαδοποίησης, το σκεπτικό πίσω από αυτήν την επιλογή και τις μετρήσεις που χρησιμοποιούνται για την αξιολόγηση της ποιότητας και της συνάφειας των προσδιορισμένων συνεργατικών σχηματισμών. Μέσα από μια σχολαστική εξέταση των αποτελεσμάτων ομαδοποίησης, αυτή η διατριβή θα δείξει πώς η ανάλυση ομαδοποίησης μπορεί να οδηγήσει σε νέες ιδέες, να υποστηρίξει τη λήψη αποφάσεων και ενδεχομένως να αποκαλύψει νέους δρόμους για έρευνα.

Η ανάλυση ομαδοποίησης, με την ικανότητά της να αποκαλύπτει κρυφές δομές και να απλοποιεί πολύπλοκα σύνολα δεδομένων, είναι ένα ανεκτίμητο εργαλείο στο οπλοστάσιο των επιστημόνων δεδομένων και των ερευνητών. Η εφαρμογή του σε αυτή τη διατριβή στοχεύει να αξιοποιήσει αυτά τα οφέλη για να ρίξει φως στα υποκείμενα πρότυπα μέσα στα δεδομένα, παρέχοντας μια σταθερή βάση για περαιτέρω ανάλυση και συμβάλλοντας στους πρωταρχικούς στόχους της έρευνας.

3.4 Αναλυτικές Μεθοδολογίες Αλγορίθμων Ομαδοποίησης

Αυτή η ενότητα εμβαθύνει στις περίπλοκες μεθοδολογίες και τις υποκείμενες αρχές μηχανικής των επιλεγμένων αλγορίθμων ομαδοποίησης: K-Means, Agglomerative Clustering, Spectral Clustering και Gaussian Mix Models (GMM). Η επιλογή κάθε αλγορίθμου βασίστηκε στη μοναδική ικανότητά τους να αποσαφηνίζουν τη δομή του συνόλου δεδομένων, προσφέροντας διαφορετικές προοπτικές και ιδέες για τις εγγενείς ομαδοποιήσεις.

3.4.1 Ομαδοποίηση K-Means

Μεθοδολογία: Η ομαδοποίηση K-Means λειτουργεί με μια απλή αλλά αποτελεσματική επαναληπτική προσέγγιση βελτίωσης. Η διαδικασία ξεκινά με την τυχαία αρχικοποίηση των k centroids, όπου k είναι μια παράμετρος καθορισμένη από το χρήστη που καθορίζει τον αριθμό των συστάδων. Κάθε σημείο δεδομένων στο σύνολο δεδομένων αντιστοιχίζεται στη συνέχεια στο πλησιέστερο κεντροειδές με βάση την Ευκλείδεια απόσταση, σχηματίζοντας k συστάδες. Τα κεντροειδή υπολογίζονται εκ νέου ως ο μέσος όρος όλων των σημείων που αποδίδονται στο σμήνος τους και οι αντιστοιχίσεις ενημερώνονται με βάση αυτά τα νέα κεντροειδή. Αυτή η διαδικασία

επαναλαμβάνεται μέχρι τη σύγκλιση, που συνήθως ορίζεται από μια ελάχιστη αλλαγή στις κεντροειδείς θέσεις μεταξύ των επαναλήψεων [21].

Αρχές μηχανικής: Η απλότητα του K-Means έγκειται στη γραμμική πολυμορφότητά του O(n), καθιστώντας το εξαιρετικά επεκτάσιμο για μεγάλα σύνολα δεδομένων. Ωστόσο, η απόδοση του αλγορίθμου είναι ευαίσθητη στην αρχική τοποθέτηση των κεντροειδών. Τεχνικές όπως η μέθοδος αρχικοποίησης K-Means++ μετριάζουν αυτό εξασφαλίζοντας μια πιο διασκορπισμένη αρχική κεντροειδή κατανομή, ενισχύοντας τη σύγκλιση και την ποιότητα του συμπλέγματος [21].

3.4.2 Agglomerative Clustering

Μεθοδολογία: Η συσσωμάτωση (agglomerative clustering) είναι μια ιεραρχική τεχνική ομαδοποίησης από κάτω προς τα πάνω. Αρχικά, κάθε σημείο δεδομένων θεωρείται ως μεμονωμένο σύμπλεγμα. Ζεύγη συστάδων συγχωνεύονται καθώς κάποιος ανεβαίνει στην ιεραρχία, με βάση ένα κριτήριο σύνδεσης που μετρά την ανομοιότητα μεταξύ συνόλων παρατηρήσεων. Αυτή η διαδικασία συνεχίζεται μέχρι να παραμείνει ένα μόνο σύμπλεγμα ή να επιτευχθεί ένας προκαθορισμένος αριθμός συμπλέγματος, με το δενδρόγραμμα να απεικονίζει αυτή την ιεραρχική διαδικασία συγχώνευσης [22]. Τα κοινά κριτήρια σύνδεσης περιλαμβάνουν:

1. Μονή σύνδεση: Η ελάχιστη απόσταση μεταξύ σημείων σε οποιαδήποτε δύο συστάδες.
2. Πλήρης σύνδεση: Η μέγιστη απόσταση μεταξύ σημείων σε οποιαδήποτε δύο σμήνη.
3. Μέση σύνδεση: Η μέση απόσταση μεταξύ όλων των ζευγών σημείων σε οποιαδήποτε δύο συστάδες.
4. Μέθοδος Ward: Η αύξηση του αθροίσματος των τετραγωνισμένων αποστάσεων εντός των συστάδων μετά τη συγχώνευση.

Αρχές μηχανικής: Η συσσωμάτωση δεν απαιτεί τον εκ των προτέρων προσδιορισμό του αριθμού των συμπλεγμάτων. Το δενδρόγραμμα, ένα δενδροειδές διάγραμμα που λαμβάνεται από ιεραρχική ομαδοποίηση, επιτρέπει στους αναλυτές να επιλέξουν τον αριθμό των συστάδων κόβοντας το δέντρο στο επιθυμητό επίπεδο. Η πολυπλοκότητα του αλγορίθμου είναι παραδοσιακά O(n^3) αλλά μπορεί να μειωθεί σε O($n^2 \log n$) με αποτελεσματικές δομές δεδομένων.

3.4.3 Spectral Clustering

Μεθοδολογία: Η φασματική ομαδοποίηση (spectral clustering) μετατρέπει το πρόβλημα ομαδοποίησης σε πρόβλημα διαμέρισης γραφήματος. Τα κύρια βήματα περιλαμβάνουν την κατασκευή ενός γραφήματος ομοιότητας από τα δεδομένα, όπου οι κόμβοι αντιπροσωπεύουν σημεία δεδομένων και οι ακμές αντικατοπτρίζουν την ομοιότητα μεταξύ των σημείων. Στη συνέχεια, υπολογίζεται το γράφημα Laplacian και οι ιδιοτιμές και τα ιδιοδιανύσματά του χρησιμοποιούνται για τη μείωση των διαστάσεων του συνόλου δεδομένων. Η ομαδοποίηση, ακολούθως, εκτελείται σε αυτόν τον μειωμένο χώρο, συχνά χρησιμοποιώντας K-Means, για τον εντοπισμό μη κυρτών συστάδων [23].

Αρχές μηχανικής: Η φασματική ομαδοποίηση υπερέχει στον εντοπισμό μη κυρτών σμηνών και μπορεί να αποκαλύψει πολύπλοκες δομές που οι παραδοσιακές μέθοδοι, όπως τα K-Means, δεν μπορούν. Η ευελιξία του στην επιλογή της μέτρησης ομοιότητας (π.χ. Gaussian συνάρτηση ακτινικής βάσης) του επιτρέπει να προσαρμόζεται σε διάφορους τύπους δεδομένων και κατανομές. Ωστόσο, το υπολογιστικό κόστος της αποσύνθεσης ιδιοτιμών, το οποίο είναι O(n^3) , μπορεί να είναι ένας περιοριστικός παράγοντας για μεγάλα σύνολα δεδομένων.

3.4.4 Μοντέλα Gaussian Mixture (GMM)

Μεθοδολογία: Το GMM είναι ένα πιθανοτικό μοντέλο που υποθέτει ότι τα σημεία δεδομένων παράγονται από ένα μείγμα πεπερασμένου αριθμού κατανομών Gauss με άγνωστες παραμέτρους. Το GMM χρησιμοποιεί τον αλγόριθμο Expectation-Maximization (EM) για να εκτιμήσει τις παραμέτρους των κατανομών Gauss. Ο αλγόριθμος EM εκτελεί επαναληπτικά δύο βήματα: το βήμα προσδοκίας (E-step), το οποίο εκτιμά τις πιθανότητες συμμετοχής κάθε σημείου δεδομένων σε κάθε σύμπλεγμα (κατανομή) και το βήμα μεγιστοποίησης (M-step), το οποίο ενημερώνει τις παραμέτρους των κατανομών Gauss με βάση αυτές τις πιθανότητες [24].

Αρχές μηχανικής: Το GMM παρέχει μια προσέγγιση ήπιας ομαδοποίησης, προσφέροντας όχι μόνο αναθέσεις συστάδων αλλά και την πιθανότητα συμμετοχής σε κάθε σύμπλεγμα, καταγράφοντας τον βαθμό αβεβαιότητας στις ταξινομήσεις. Αυτό το πιθανοτικό πλαίσιο επιτρέπει στο GMM να μοντελοποιεί συστάδες διαφορετικών μεγεθών και σχημάτων, καθιστώντας το ευέλικτο σε ποικίλα σύνολα δεδομένων. Η πολυπλοκότητα του αλγορίθμου επηρεάζεται κυρίως από τις επαναλήψεις EM και τον αριθμό των συνιστωσών, με γενικό υπολογιστικό κόστος $O(n \cdot k \cdot d^2)$ για n σημεία δεδομένων, *k* Gaussian συνιστώσες και *d* διαστάσεις.

3.5 Μετρήσεις αξιολόγησης στη διαδικασία ομαδοποίησης

Η αποτελεσματικότητα των αλγορίθμων ομαδοποίησης εξαρτάται από ισχυρές μετρήσεις αξιολόγησης που παρέχουν πληροφορίες σχετικά με την ποιότητα των παραγόμενων συμπλεγμάτων. Σε αυτή τη μελέτη, τρεις βασικές μετρήσεις - Silhouette Score, Calinski-Harabasz Index και Davies-Bouldin Index - χρησιμοποιήθηκαν για την αξιολόγηση και τη σύγκριση της απόδοσης των διαμορφώσεων ομαδοποίησης σε διαφορετικούς αλγόριθμους. Αυτή η ενότητα εμβαθύνει στη μεθοδολογία και τη σημασία αυτών των μετρήσεων, απεικονίζοντας την εφαρμογή τους στη διαδικασία ομαδοποίησης.

3.5.1 Βαθμολογία σιλουέτας (Silhouette Score)

Το Silhouette Score είναι μια δημοφιλής μέτρηση που χρησιμοποιείται για τη μέτρηση της ποιότητας της ομαδοποίησης. Μετρά πόσο παρόμοιο είναι ένα αντικείμενο με το δικό του σύμπλεγμα (συνοχή) σε σύγκριση με άλλα συμπλέγματα (διαχωρισμός). Η βαθμολογία σιλουέτας για ένα μόνο δείγμα υπολογίζεται ως $\frac{b-a}{\max(a,b)}$, όπου a είναι η μέση απόσταση εντός του συμπλέγματος (η μέση απόσταση μεταξύ του δείγματος και όλων των άλλων σημείων της ίδιας ομάδας) και b είναι η μέση πλησιέστερη απόσταση συμπλέγματος (η μέση απόσταση μεταξύ του δείγματος και όλων των σημείων της πλησιέστερης ομάδας στην οποία το δείγμα δεν αποτελεί μέρος). Η βαθμολογία της σιλουέτας κυμαίνεται από -1 έως 1, όπου μια υψηλή τιμή δείχνει ότι το αντικείμενο ταιριάζει καλά με το δικό του σύμπλεγμα και ελάχιστα ταιριάζει με τα γειτονικά σμήνη [25].

Στο πλαίσιο αυτής της μελέτης, υπολογίστηκε η βαθμολογία σιλουέτας για κάθε διαμόρφωση ομαδοποίησης για να προσδιοριστεί ο βαθμός προσαρμογής των σημείων δεδομένων εντός των συστάδων, καθοδηγώντας έτσι την επιλογή του βέλτιστου αριθμού συστάδων και άλλων υπερπαραμέτρων.

3.5.2 Δείκτης Calinski-Harabasz

Ο δείκτης Calinski-Harabasz, γνωστός και ως κριτήριο αναλογίας διακύμανσης, είναι μια μέτρηση που αξιολογεί την ποιότητα του συμπλέγματος συγκρίνοντας τη διασπορά εντός του συμπλέγματος, με τη διασπορά μεταξύ των συμπλεγμάτων . Ορίζεται ως $\frac{\mathrm{Tr}(B_k)}{\mathrm{Tr}(W_k)} \times \frac{N-k}{k-1}$, όπου $Tr(B_k)$ είναι το ίχνος της μήτρας διασποράς μεταξύ ομάδων, $Tr(W_k)$ είναι το ίχνος της μήτρας διασποράς εντός συμπλέγματος, N είναι ο αριθμός των σημείων και k είναι ο αριθμός των συστάδων. Μια υψηλότερη βαθμολογία Calinski-Harabasz δείχνει ότι τα σμήνη είναι πυκνά και καλά διαχωρισμένα, κάτι που είναι επιθυμητό σε μια καλή διαμόρφωση ομαδοποίησης [26]. Αυτός ο δείκτης χρησιμοποιήθηκε για να συμπληρώσει τη βαθμολογία της σιλουέτας, παρέχοντας μια πρόσθετη προοπτική για τη δομή ομαδοποίησης, δίνοντας έμφαση στα χαρακτηριστικά διασποράς των συστάδων.

3.5.3 Δείκτης Davies-Bouldin

Ο δείκτης Davies-Bouldin είναι ένα σύστημα εσωτερικής αξιολόγησης όπου η επικύρωση του πόσο καλά έχει γίνει η ομαδοποίηση γίνεται χρησιμοποιώντας ποσότητες και χαρακτηριστικά εγγενή στο σύνολο δεδομένων. Ο δείκτης Davies-Bouldin ορίζεται ως

$\frac{1}{n}\Sigma_{i=1}^{n} max_{j\neq i}(\frac{\sigma_i+\sigma_j}{d(c_i,c_j)})$, όπου σ είναι η μέση απόσταση όλων των στοιχείων σε ένα σμήνος από το κεντροειδές του σμήνους, $c_i \; and \; c_j$ είναι τα κεντροειδή των συστάδων i και j και $d(c_i, c_j)$ είναι η απόσταση μεταξύ κεντροειδών c_i και c_j. . Ο δείκτης στοχεύει σε χαμηλότερες τιμές, με χαμηλότερο δείκτη Davies-Bouldin που σχετίζεται με ένα μοντέλο με καλύτερο διαχωρισμό μεταξύ των συστάδων [27].

Αυτός ο δείκτης είναι ιδιαίτερα χρήσιμος για τον εντοπισμό σεναρίων όπου τα συμπλέγματα δεν ήταν καλά διαχωρισμένα ή όταν υπήρχε σημαντική διακύμανση εντός των συμπλεγμάτων, βοηθώντας στη βελτίωση των διαμορφώσεων ομαδοποίησης.

3.6 Υπολογιστική Υλοποίηση

Η διαδικασία αξιολόγησης ενθυλακώθηκε σε μια συνάρτηση Python με την ονομασία ”calc_results”, σχεδιασμένη να υπολογίζει τις προαναφερθείσες μετρήσεις για μια δεδομένη έξοδο ομαδοποίησης και να προσθέτει τα αποτελέσματα σε ένα ολοκληρωμένο λεξικό αποτελεσμάτων. Αυτή η αυτοματοποιημένη προσέγγιση διευκόλυνε την αποτελεσματική σύγκριση διαφορετικών διαμορφώσεων ομαδοποίησης, επιτρέποντας τη συστηματική αξιολόγηση της απόδοσης των αλγορίθμων σε ποικίλες ρυθμίσεις παραμέτρων. Η ενσωμάτωση του Silhouette Score, του Calinski-Harabasz Index και του Davies-Bouldin Index στο πλαίσιο αξιολόγησης παρείχε μια πολύπλευρη εικόνα των αποτελεσμάτων ομαδοποίησης, περιλαμβάνοντας πτυχές συνοχής, διαχωρισμού και διασποράς. Αυτή η αυστηρή μεθοδολογία αξιολόγησης διασφάλισε ότι οι επιλεγμένες διαμορφώσεις ομαδοποίησης όχι μόνο βελτιστοποιήθηκαν για συγκεκριμένες παραμέτρους αλγορίθμων, αλλά και ευθυγραμμίστηκαν με την εγγενή δομή του συνόλου δεδομένων, ενισχύοντας έτσι την αξιοπιστία και την ερμηνευσιμότητα των αποτελεσμάτων ομαδοποίησης.

3.7 Συμπέρασμα

Η επιλογή των K-Means, Agglomerative Clustering, Spectral Clustering και GMM για αυτή τη διατριβή βασίζεται στις ποικίλες μεθοδολογίες τους και στις ξεχωριστές προοπτικές που προσφέρουν στη δομή του συνόλου δεδομένων. Οι αρχές μηχανικής και οι υπολογιστικές αποχρώσεις κάθε αλγορίθμου εξετάστηκαν προσεκτικά για να διασφαλιστεί ότι ευθυγραμμίζονται με τα χαρακτηριστικά του συνόλου δεδομένων και τους ερευνητικούς στόχους. Η επόμενη φάση συντονισμού υπερπαραμέτρων στοχεύει στην περαιτέρω βελτίωση αυτών των αλγορίθμων, βελτιστοποιώντας την απόδοσή τους για να αποφέρει τα πιο συνεκτικά και διορατικά αποτελέσματα ομαδοποίησης.

3.8 Διαδικασία ρύθμισης υπερπαραμέτρων για αλγόριθμους ομαδοποίησης

Η διαδικασία ρύθμισης υπερπαραμέτρων είναι ένα θεμελιώδες βήμα στη βελτιστοποίηση των αλγορίθμων ομαδοποίησης, που περιλαμβάνει μια συστηματική εξερεύνηση του χώρου παραμέτρων για τον εντοπισμό διαμορφώσεων που μεγιστοποιούν την απόδοση του αλγορίθμου. Αυτή η ενότητα παρέχει μια λεπτομερή επισκόπηση της προσέγγισης συντονισμού υπερπαραμέτρων που υιοθετήθηκε για τα K-Means, Agglomerative Clustering, Spectral Clustering και Gaussian Mix Models (GMM), δίνοντας έμφαση στις συγκεκριμένες παραμέτρους που διερευνώνται μέσω της αναζήτησης πλέγματος (grid search) και των τεχνικών τυχαίας αναζήτησης (random search).

3.8.1 K-Means

Grid Search παράμετροι:

1. n_clusters: Ο αριθμός των συστάδων κυμαινόταν από 2 έως 10, αντανακλώντας ένα ευρύ αλλά υπολογιστικά εφικτό εύρος για τη διερεύνηση πιθανών ομαδοποιήσεων εντός των δεδομένων.
2. init: Η μέθοδος αρχικοποίησης εναλλάσσονταν μεταξύ του προεπιλεγμένου "k-means++" για έξυπνη τοποθέτηση κεντροειδούς και του "random" για συγκρίσεις γραμμής βάσης.
3. n_init: Ο αριθμός των αρχικών κεντροειδών σπόρων δοκιμάστηκε στα 10 και 20 για να εκτιμηθεί η επίδραση των πολλαπλών αρχικοποιήσεων στη σταθερότητα του τελικού διαλύματος.
4. tol: Η ανοχή για σύγκλιση διερευνήθηκε στα σημεία 1e-4 και 1e-3 για τον προσδιορισμό της ευαισθησίας του αλγορίθμου στα κριτήρια σύγκλισης.

Random Search παράμετροι:

Επεκτάθηκε το εύρος n_clusters έως και 20 για να διερευνήσει πιο λεπτομερείς δυνατότητες ομαδοποίησης με στοχαστικό τρόπο.

Διατηρήθηκαν άλλες παράμετροι όπως στην αναζήτηση πλέγματος, αλλά επετράπηκε ένας ευρύτερος χώρος εξερεύνησης λόγω της τυχαιοποιημένης φύσης της αναζήτησης.

3.8.2 Agglomerative Clustering

Grid Search Parameters:

1. n_clusters: Παρόμοια με το K-Means, ο αριθμός των συστάδων ποικίλει από 2 έως 10 για τον προσδιορισμό της βέλτιστης λεπτομέρειας των ιεραρχικών ομαδοποιήσεων.
2. ‘linkage’: Το κριτήριο linkage διέφερε μεταξύ "ward", "complete" και "average" για την αξιολόγηση διαφορετικών στρατηγικών συγχώνευσης συστάδων με βάση τις αποστάσεις μεταξύ συστάδων.

Random Search Parameters:

1. Επεκτάθηκε n_clusters έως 20 για τη διερεύνηση λεπτομερέστερων ιεραρχικών δομών.
2. Εισάχθηκε η "ενιαία" σύνδεση με την τυχαία αναζήτηση για να συμπεριλάβει ένα πιο ποικίλο σύνολο δυναμικών ομαδοποίησης, όπου η εστίαση είναι στα πλησιέστερα σημεία μεταξύ των συμπλεγμάτων.

3.8.3 Spectral Clustering

Grid Search Parameters:

1. n_clusters: Κυμαίνεται από 2 έως 10, με στόχο την αποτύπωση σημαντικών ομαδοποιήσεων στο γράφημα ομοιότητας.
2. assign_labels: Εναλλάσσονται μεταξύ "kmeans" και "discretize" για την κατανόηση της επίδρασης διαφορετικών στρατηγικών ανάθεσης συστάδων μετά την ιδιοσύνθεση.
3. n_init: Ορίστηκαν οι επιλογές 10 και 20 για να αξιολογήσουμε τη συνέπεια των αποτελεσμάτων ομαδοποίησης με πολλαπλές αρχικοποιήσεις.
4. n_neighbors: Δοκιμάστηκε με 5 και 10 για να εκτιμηθεί πώς ο αριθμός των γειτόνων στον πίνακα συνάφειας επηρεάζει τη δομή του γραφήματος και την επακόλουθη ομαδοποίηση.

Random Search Parameters:

Χρησιμοποιήθηκαν εκτεταμένα εύρη n_clusters και n_neighbors σε 20 και 15, αντίστοιχα, επιτρέποντας μια ευρύτερη και πιο τυχαία εξερεύνηση συνδυασμών παραμέτρων που θα μπορούσαν να αποκαλύψουν μη προφανή μοτίβα ομαδοποίησης.

3.8.4 Gaussian Mixture Models (GMM)

Grid Search Parameters:

1. n_components: Ποικίλλει από 2 έως 10 για τον προσδιορισμό του βέλτιστου αριθμού κατανομών Gauss που ταιριάζουν καλύτερα στα δεδομένα.
2. covariance_type: Περιλαμβάνονται οι λέξεις "πλήρης", "δεμένη", "διαγώνια" και "σφαιρική" για να κατανοήσουμε πώς διαφορετικές υποθέσεις σχετικά με τη δομή της συνδιακύμανσης επηρεάζουν την ομαδοποίηση.
3. init_params: Εναλλάσσονται μεταξύ "kmeans" και "random" για να εξεταστεί η επίδραση της αρχικοποίησης στη σύγκλιση και την ακρίβεια του μοντέλου μείγματος.
4. n_init: Ορίστηκαν οι τιμές 1 και 2 για να αξιολογήσουμε την επίδραση πολλαπλών προσπαθειών αρχικοποίησης στη σταθερότητα του μοντέλου.

Random Search Parameters:

Επεκτάθηκε το n_components έως 20 και n_init έως 5, παρέχοντας έναν ευρύτερο στοχαστικό χώρο αναζήτησης για να αποκαλύψει ενδεχομένως πιο σύνθετα μοντέλα μείγματος που θα μπορούσαν να συλλάβουν καλύτερα την κατανομή των δεδομένων.

3.8.5 Μέτρηση αξιολόγησης: Silhouette Score

Σε όλους τους αλγορίθμους, η βαθμολογία σιλουέτας χρησιμοποιήθηκε ως κύρια μέτρηση αξιολόγησης για τη μέτρηση της ποιότητας των διαμορφώσεων ομαδοποίησης. Η ικανότητα αυτής της μέτρησης να εξισορροπεί τη συνοχή και τον διαχωρισμό των συμπλεγμάτων την κατέστησε ιδανική επιλογή για τη σύγκριση διαφορετικών ρυθμίσεων υπερπαραμέτρων, καθοδηγώντας την επιλογή της πιο αποτελεσματικής λύσης ομαδοποίησης [25].

3.9 Υπολογιστικές εκτιμήσεις

Οι υπολογιστικές απαιτήσεις του συντονισμού υπερπαραμέτρων ήταν σημαντικές, ιδιαίτερα για αλγορίθμους με υψηλότερη πολυπλοκότητα και εκτεταμένα πλέγματα παραμέτρων. Χρησιμοποιήθηκαν τεχνικές παράλληλης επεξεργασίας για την επιτάχυνση των διαδικασιών πλέγματος και τυχαίας αναζήτησης, επιτρέποντας την ταυτόχρονη αξιολόγηση πολλαπλών συνόλων παραμέτρων. Επιπλέον, εφαρμόστηκαν μηχανισμοί έγκαιρης διακοπής όπου ήταν εφικτό για τον τερματισμό της αναζήτησης όταν οι βελτιώσεις σταθεροποιήθηκαν, διατηρώντας υπολογιστικούς πόρους.

3.10 Συμπέρασμα

Η διαδικασία συντονισμού υπερπαραμέτρων (hyperparameter tuning), που περιγράφεται λεπτομερώς εδώ, υπογραμμίζει τη μεθοδική προσέγγιση που ακολουθείται για τη βελτίωση της απόδοσης των βασικών αλγορίθμων ομαδοποίησης. Με την προσεκτική πλοήγηση στο χώρο παραμέτρων μέσω στρατηγικών πλέγματος και τυχαίας αναζήτησης και αξιοποιώντας τη βαθμολογία σιλουέτας ως ολοκληρωμένη μέτρηση απόδοσης, αυτή η διαδικασία διασφαλίζει την εξαγωγή βελτιστοποιημένων μοντέλων ομαδοποίησης προσαρμοσμένων στα μοναδικά χαρακτηριστικά του συνόλου δεδομένων.

4. Οι βιβλιοθήκες και ο ρόλος τους στην ανάλυση δεδομένων

4.1 Εισαγωγή

Στον τομέα της ανάλυσης δεδομένων και της μηχανικής μάθησης, αρκετές βιβλιοθήκες Python ξεχωρίζουν για την ολοκληρωμένη λειτουργικότητα και την ευκολία χρήσης τους. Αυτή η ενότητα εμβαθύνει στις βασικές βιβλιοθήκες που χρησιμοποιούνται στην παρούσα διατριβή, διευκρινίζοντας τη σημασία και την εφαρμογή τους σε διάφορα στάδια της ανάλυσης, από την προεπεξεργασία δεδομένων έως την αξιολόγηση μοντέλων.

4.2 Pandas

Το Pandas, μια βιβλιοθήκη ανοιχτού κώδικα, έχει καθιερωθεί ως απαραίτητο εργαλείο για χειρισμό και ανάλυση δεδομένων στην Python. Προσφέρει DataFrame και Series ως κύριες δομές δεδομένων, οι οποίες επιτρέπουν την αποτελεσματική αποθήκευση και το χειρισμό δεδομένων σε μορφή πίνακα και χρονοσειρών. Αυτές οι δομές δεδομένων είναι εξοπλισμένες με ένα ολοκληρωμένο σύνολο λειτουργιών για την επεξεργασία, τη συγκέντρωση και το μετασχηματισμό δεδομένων, διευκολύνοντας πολύπλοκες εργασίες ανάλυσης δεδομένων με ευκολία. Η δυνατότητα αβίαστης εισαγωγής, καθαρισμού και χειρισμού συνόλων δεδομένων καθιστά το Pandas απαραίτητο εργαλείο στην εργαλειοθήκη της επιστήμης των δεδομένων [28].

Ένα από τα ξεχωριστά χαρακτηριστικά του Pandas είναι οι ισχυρές δυνατότητες εισόδου/εξόδου, υποστηρίζοντας ένα ευρύ φάσμα μορφών αρχείων, όπως CSV, Excel, JSON, HTML και HDF5. Αυτή η ευελιξία καθιστά το Pandas μια λύση για επιστήμονες δεδομένων που ασχολούνται με διαφορετικές πηγές δεδομένων. Επιπλέον, η απρόσκοπτη ενσωμάτωσή του με βάσεις δεδομένων, επιτρέποντας την άμεση φόρτωση δεδομένων σε DataFrames, εξορθολογίζει τη φάση προετοιμασίας δεδομένων των έργων ανάλυσης.

Η Pandas υπερέχει στο χειρισμό δεδομένων που λείπουν (missing values), μια κοινή πρόκληση σε σύνολα δεδομένων πραγματικού κόσμου. Παρέχει ευέλικτα εργαλεία για τη συμπλήρωση, απόθεση ή παρεμβολή τιμών που λείπουν, επιτρέποντας ισχυρές διαδικασίες καθαρισμού δεδομένων. Επιπλέον, η εξελιγμένη λειτουργικότητα χρονοσειρών, συμπεριλαμβανομένης της δημιουργίας εύρους ημερομηνιών, της μετατροπής συχνότητας και των στατιστικών κινούμενων παραθύρων, το καθιστά ιδιαίτερα ισχυρό για ανάλυση χρονοσειρών.

Οι δυνατότητες της βιβλιοθήκης για ομαδοποίηση (grouping) και pivoting, ενσωματωμένες σε λειτουργίες, όπως groupby και pivot_table, προσφέρουν διαισθητικούς και αποτελεσματικούς τρόπους συγκέντρωσης και αναδιαμόρφωσης δεδομένων. Αυτό διευκολύνει την εξερεύνηση των μοτίβων και των σχέσεων εντός του συνόλου δεδομένων, ένα κρίσιμο βήμα στην ανάλυση δεδομένων.

Η δύναμη του Panda έγκειται, επίσης, στην ικανότητά του να εκτελεί συγχωνεύσεις και ενώσεις συνόλων δεδομένων υψηλής απόδοσης, παρόμοιες με τις λειτουργίες SQL, επιτρέποντας τον συνδυασμό διαφορετικών πηγών δεδομένων σε ένα συνεκτικό σύνολο δεδομένων έτοιμο για ανάλυση.

4.3 NumPy

Το NumPy, συντομογραφία του Numerical Python, είναι μια θεμελιώδης βιβλιοθήκη για αριθμητικούς υπολογισμούς στην Python. Εισάγει ένα ισχυρό n-διαστάσεων αντικείμενο συστοιχίας, το οποίο χρησιμεύει ως δομικό στοιχείο για ένα ευρύ φάσμα επιστημονικών και μαθηματικών βιβλιοθηκών που βασίζονται σε Python, συμπεριλαμβανομένων των Pandas, SciPy και Scikit-learn. Το NumPy είναι θεμελιώδες για την επιστημονική πληροφορική στην Python. Εισάγει ισχυρά n-διαστάσεων αντικείμενα συστοιχίας και ένα ευρύ φάσμα συναρτήσεων για γραμμική άλγεβρα, μετασχηματισμούς Fourier και δυνατότητες τυχαίων αριθμών. Η αποτελεσματικότητα και η ταχύτητα του NumPy, που προέρχονται από τη βάση κώδικα C και Fortran, το καθιστούν μια βασική βιβλιοθήκη για την εκτέλεση μαθηματικών συναρτήσεων υψηλού επιπέδου σε μεγάλους, πολυδιάστατους πίνακες και μήτρες [29]. Αυτές οι πράξεις περιλαμβάνουν ένα ευρύ φάσμα, από τη βασική αριθμητική έως τις πιο σύνθετες μαθηματικές συναρτήσεις, όπως οι πράξεις γραμμικής άλγεβρας, οι μετασχηματισμοί Fourier και η δημιουργία τυχαίων αριθμών.

Η δυνατότητα μετάδοσης συστοιχιών του NumPy είναι ιδιαίτερα αξιοσημείωτη, επιτρέποντας λειτουργίες από άποψη στοιχείων σε συστοιχίες διαφορετικών σχημάτων, αποφεύγοντας έτσι την ανάγκη για ρητούς βρόχους. Αυτό όχι μόνο οδηγεί σε πιο συνοπτικό και ευανάγνωστο κώδικα, αλλά επίσης βελτιώνει σημαντικά την υπολογιστική απόδοση. Οι δυνατότητες της βιβλιοθήκης για τεμαχισμό (slicing) και ευρετηρίαση (indexing) παρέχουν ένα ευέλικτο και διαισθητικό περιβάλλον εργασίας για πρόσβαση και χειρισμό δεδομένων πίνακα. Οι προηγμένες δυνατότητες ευρετηρίου, όπως η δυαδική ευρετηρίαση (boolean indexing) και η fancy indexing, προσφέρουν ισχυρές μεθόδους επιλογής και φιλτραρίσματος δεδομένων.

Το NumPy διαδραματίζει επίσης κρίσιμο ρόλο στη διαλειτουργικότητα μεταξύ των βιβλιοθηκών ανάλυσης δεδομένων. Η δομή του πίνακα χρησιμεύει ως η τυπική μορφή αριθμητικών δεδομένων, διευκολύνοντας την απρόσκοπτη ανταλλαγή και χειρισμό δεδομένων σε διαφορετικές βιβλιοθήκες και εργαλεία στο οικοσύστημα της Python.

4.4 Scikit-learn

Το Scikit-learn είναι μια κορυφαία βιβλιοθήκη στο οικοσύστημα της Python για μηχανική μάθηση. Προσφέρει ένα ευρύ φάσμα αλγορίθμων για ταξινόμηση, παλινδρόμηση, ομαδοποίηση, μείωση διαστάσεων και επιλογής μοντέλου, καθιστώντας το ένα ευέλικτο εργαλείο τόσο για αρχάριους επιστήμονες δεδομένων όσο και για έμπειρους επαγγελματίες. Η φιλοσοφία σχεδιασμού της βιβλιοθήκης δίνει έμφαση στην ευκολία χρήσης, την απόδοση και την ευελιξία, διασφαλίζοντας ότι οι σύνθετες εργασίες μηχανικής μάθησης είναι προσβάσιμες και αποτελεσματικές. Το συνεπές API και η ολοκληρωμένη τεκμηρίωσή (documentation) του το καθιστούν προσβάσιμο τόσο για αρχάριους όσο και για ειδικούς. Αυτή η διατριβή αξιοποιεί τους αλγόριθμους ομαδοποίησης του Scikit-learn,

όπως οι KMeans, AgglomerativeClustering, SpectralClustering και GaussianMix, καθώς και τα εργαλεία προεπεξεργασίας και αξιολόγησης μοντέλων όπως το MinMaxScaler, το StandardScaler και διάφορες μετρήσεις ομαδοποίησης [30].

Βασικά χαρακτηριστικά και αλγόριθμοι:

1. Αλγόριθμοι εποπτευόμενης μάθησης (Supervised Learning): Το Scikit-learn παρέχει μια εκτεταμένη σειρά αλγορίθμων για εποπτευόμενες μαθησιακές εργασίες, συμπεριλαμβανομένων δημοφιλών μεθόδων, όπως Support Vector Machines, Random Forests, Gradient Boosting και k-Nearest Neighbors. Αυτοί οι αλγόριθμοι είναι εξοπλισμένοι για να χειριστούν ένα ευρύ φάσμα εφαρμογών, από την ανίχνευση ανεπιθύμητων μηνυμάτων έως την πρόβλεψη της συμπεριφοράς των καταναλωτών.
2. Αλγόριθμοι μάθησης χωρίς επίβλεψη (Unsupervised Learning): Στη σφαίρα της μη εποπτευόμενης μάθησης, το Scikit-learn προσφέρει αλγόριθμους ομαδοποίησης (όπως K-Means, DBSCAN και ιεραρχική ομαδοποίηση), μαζί με τεχνικές μείωσης διαστάσεων, όπως η Ανάλυση Κύριων Συνιστωσών (PCA) και η t-Distributed Stochastic Neighbor Embedding (t-SNE). Αυτά τα εργαλεία είναι ανεκτίμητα για την ανακάλυψη υποκείμενων μοτίβων και δομών σε δεδομένα χωρίς προκαθορισμένες ετικέτες.
3. Αξιολόγηση και επιλογή μοντέλου (Model Evaluation and Selection): Το Scikit-learn παρέχει ολοκληρωμένα εργαλεία για την αξιολόγηση μοντέλων και τη ρύθμιση υπερπαραμέτρων, συμπεριλαμβανομένης της διασταυρούμενης επικύρωσης (cross-validation), της αναζήτησης πλέγματος (grid search) και των μετρικών για την αξιολόγηση της απόδοσης του μοντέλου. Αυτές οι λειτουργίες είναι ζωτικής σημασίας για την ανάπτυξη ισχυρών μοντέλων που γενικεύονται καλά σε νέα δεδομένα.
4. Pipeline and Feature Extraction: Η λειτουργία του pipeline της βιβλιοθήκης επιτρέπει τη δημιουργία σύνθετων αγωγών επεξεργασίας και μοντελοποίησης δεδομένων, εξορθολογίζοντας τη ροή εργασίας από την προεπεξεργασία δεδομένων έως την αξιολόγηση μοντέλων. Επιπλέον, οι ενότητες εξαγωγής χαρακτηριστικών διευκολύνουν τη μετατροπή δεδομένων κειμένου και εικόνας σε μορφή κατάλληλη για αλγόριθμους μηχανικής μάθησης.

Κοινότητα και οικοσύστημα

Το Scikit-learn επωφελείται από μια ζωντανή κοινότητα μελών που συνεισφέρουν και ένα τεράστιο οικοσύστημα συμπληρωματικών βιβλιοθηκών. Η συμβατότητά του με άλλες βιβλιοθήκες, όπως το Pandas, το NumPy και το Matplotlib επιτρέπει τον απρόσκοπτο χειρισμό δεδομένων, τους αριθμητικούς υπολογισμούς και την οπτικοποίηση, δημιουργώντας ένα συνεκτικό περιβάλλον για έργα μηχανικής μάθησης από άκρο σε άκρο.

4.5 Matplotlib

Το Matplotlib είναι μια ολοκληρωμένη βιβλιοθήκη για τη δημιουργία στατικών, κινούμενων και διαδραστικών απεικονίσεων στην Python. Η ευελιξία του επιτρέπει τη δημιουργία σύνθετων plots (γραφημάτων) με σχετικά απλό κώδικα. Το Matplotlib αποτελεί μια θεμελιώδη βιβλιοθήκη για οπτικοποίηση δεδομένων στην Python, προσφέροντας μια ολοκληρωμένη σουίτα λειτουργιών σχεδίασης για τη δημιουργία στατικών, κινούμενων και διαδραστικών απεικονίσεων. Η ευελιξία και η προσαρμοστικότητά του το καθιστούν απαραίτητο εργαλείο για την εξερεύνηση, ανάλυση και παρουσίαση δεδομένων. Η ενσωμάτωση της βιβλιοθήκης με το Pandas και το NumPy διευκολύνει την οπτικοποίηση των συνόλων δεδομένων και την ερμηνεία των αναλυτικών αποτελεσμάτων, καθιστώντας την ένα ανεκτίμητο εργαλείο για την εξερεύνηση και παρουσίαση δεδομένων [31].

Δυνατότητες σχεδίασης

1. Βασικά έως σύνθετα γραφήματα: Από απλές γραφικές παραστάσεις γραμμών και ιστογράμματα έως πολύπλοκα γραφήματα διασποράς και απεικονίσεις 3D, το Matplotlib παρέχει ένα ευρύ φάσμα λειτουργιών σχεδίασης. Αυτή η ευελιξία υποστηρίζει ένα ευρύ φάσμα αναγκών οπτικοποίησης, από την προκαταρκτική εξερεύνηση δεδομένων έως τη δημοσίευση ερευνητικών ευρημάτων.
2. Προσαρμογή και τελειοποίηση(Fine-tuning): Το Matplotlib προσφέρει εκτεταμένες επιλογές προσαρμογής, επιτρέποντας την τελειοποίηση της αισθητικής των γραφημάτων, συμπεριλαμβανομένων χρωμάτων, legend, ετικετών (label) και σχολιασμών. Αυτό το επίπεδο ελέγχου διασφαλίζει ότι οι απεικονίσεις επικοινωνούν αποτελεσματικά τις υποκείμενες πληροφορίες δεδομένων.
3. Ενσωμάτωση (Integration) με Pandas: Το Matplotlib ενσωματώνεται άψογα με το Pandas, επιτρέποντας την άμεση σχεδίαση από αντικείμενα DataFrame και Series. Αυτή η ενοποίηση βελτιστοποιεί τη διαδικασία απεικόνισης στις ροές εργασίας της ανάλυσης δεδομένων.

Εφαρμογές και αντίκτυπος

Η ικανότητα του Matplotlib να παράγει πολύ αξιόλογα και ολοκληρωμένα γραφήματα το έχει καταστήσει βασικό στοιχείο σε ακαδημαϊκά και επαγγελματικά περιβάλλοντα. Χρησιμοποιείται ευρέως στην επιστημονική πληροφορική, τη χρηματοδότηση και διάφορους τομείς που απαιτούν ολοκληρωμένη ανάλυση δεδομένων και οπτικοποίηση.

4.6 SciPy

Η SciPy, μια βιβλιοθήκη Python ανοιχτού κώδικα, αποτελεί αναπόσπαστο μέρος της επιστημονικής και τεχνικής πληροφορικής. Βασιζόμενη στις θεμελιώδεις δυνατότητες του NumPy, το SciPy επεκτείνει τη λειτουργικότητά του με μια συλλογή μαθηματικών αλγορίθμων και συναρτήσεων ευκολίας (convenience functions). Είναι οργανωμένη σε υποπακέτα που καλύπτουν διαφορετικούς επιστημονικούς τομείς πληροφορικής, καθιστώντας την ένα ευέλικτο εργαλείο για ένα ευρύ φάσμα εφαρμογών στη μηχανική, την επιστήμη και τα μαθηματικά. Στην παρούσα διπλωματική εργασία, λειτουργίες από το SciPy, όπως η σύνδεση(linkage), το δενδρόγραμμα, το fcluster και διάφορα στατιστικά εργαλεία, χρησιμοποιούνται για ιεραρχική ομαδοποίηση και στατιστική ανάλυση [32].

Βασικά στοιχεία και λειτουργίες:

1. Βελτιστοποίηση και ελαχιστοποίηση (Optimization and Minimization): Η SciPy παρέχει ισχυρά εργαλεία βελτιστοποίησης, συμπεριλαμβανομένων λειτουργιών για την ελαχιστοποίηση ή τη μεγιστοποίηση στόχων, την εύρεση ρίζας (root finding) και την προσαρμογή καμπύλης (curve fitting.). Αυτά τα εργαλεία είναι απαραίτητα σε διάφορους τομείς, όπως η φυσική, η χημεία και η οικονομία, όπου η προσαρμογή μοντέλων και η βελτιστοποίηση του συστήματος είναι ζωτικής σημασίας.
2. Στατιστική ανάλυση: Η ενότητα scipy.stats προσφέρει ένα ευρύ φάσμα στατιστικών συναρτήσεων και κατανομών, συμπεριλαμβανομένων συνοπτικών στατιστικών, κατανομών πιθανότητας, δοκιμών υποθέσεων και συναρτήσεων συσχέτισης. Αυτή η

ολοκληρωμένη σουίτα στατιστικών εργαλείων υποστηρίζει εξελιγμένη ανάλυση και ερμηνεία δεδομένων.

3. Γραμμική Άλγεβρα: Με βάση τις δυνατότητες του πίνακα και του πίνακα του NumPy, η ενότητα γραμμικής άλγεβρας της SciPy, scipy.linalg, περιλαμβάνει προηγμένα χαρακτηριστικά, όπως αποσυνθέσεις πινάκων, προβλήματα ιδιοτιμών και συναρτήσεις πινάκων. Αυτές οι λειτουργίες είναι ζωτικής σημασίας για την επίλυση σύνθετων μαθηματικών προβλημάτων στη μηχανική και τις φυσικές επιστήμες.
4. Ολοκλήρωση και Διαφορικές Εξισώσεις (Integration and Differential Equations): Το SciPy παρέχει λειτουργίες για αριθμητική ολοκλήρωση και επίλυση συνηθισμένων και μερικών διαφορικών εξισώσεων. Αυτές οι δυνατότητες είναι ιδιαίτερα χρήσιμες σε τομείς, όπως η αστροφυσική, η κβαντική μηχανική και τα βιομαθηματικά, όπου η μοντελοποίηση συνεχών συστημάτων είναι κοινή.
5. Επεξεργασία σήματος και εικόνας: Με λειτουργίες φιλτραρίσματος, συνέλιξης και μετασχηματισμών Fourier, το module επεξεργασίας σήματος της SciPy υποστηρίζει την ανάλυση και το χειρισμό σημάτων και εικόνων. Αυτό χρησιμοποιείται ευρέως σε τομείς, όπως η ιατρική απεικόνιση, η επεξεργασία ήχου και οι τηλεπικοινωνίες.

Κοινότητα και Ανάπτυξη

Η ανάπτυξη της SciPy καθοδηγείται από μια μεγάλη κοινότητα συνεισφερόντων, εξασφαλίζοντας συνεχή βελτίωση και την προσθήκη αλγορίθμων αιχμής. Η εκτενής τεκμηρίωση και τα σεμινάριά του διευκολύνουν την υιοθέτησή του τόσο από ερευνητές όσο και από επαγγελματίες.

4.7 Seaborn

Το Seaborn είναι μια βιβλιοθήκη οπτικοποίησης Python βασισμένη στο Matplotlib που παρέχει μια διεπαφή υψηλού επιπέδου για τη σχεδίαση ελκυστικών και ενημερωτικών στατιστικών γραφικών. Το Seaborn βασίζεται στις θεμελιώδεις δυνατότητες σχεδίασης του Matplotlib, προσφέροντας μια διεπαφή υψηλότερου επιπέδου για τη δημιουργία στατιστικά ενημερωμένων απεικονίσεων. Έχει σχεδιαστεί για να λειτουργεί απρόσκοπτα με τα Pandas DataFrames, επιτρέποντας τη σύνθετη οπτικοποίηση δεδομένων με συνοπτικό κώδικα. Είναι ιδιαίτερα κατάλληλο για την οπτικοποίηση σύνθετων συνόλων δεδομένων, προσφέροντας μια ποικιλία μοτίβων οπτικοποίησης και συνδυασμών χρωμάτων για την αποκάλυψη πληροφοριών εντός των δεδομένων. Η χρήση του Seaborn σε αυτή τη διατριβή βοηθά στη δημιουργία πιο περίτεχνων γραφημάτων για εξερεύνηση δεδομένων και παρουσίαση αποτελεσμάτων [33].

Χαρακτηριστικά απεικόνισης

1. Στατιστικά γραφήματα: Η Seaborn υπερέχει στη δημιουργία γραφημάτων που αναδεικνύουν τις στατιστικές ιδιότητες των δεδομένων, όπως γραφήματα κατανομής, γραφήματα παλινδρόμησης και κατηγορικά γραφήματα. Αυτές οι απεικονίσεις ενισχύονται με την εκτίμηση της πυκνότητας του πυρήνα, επιτρέποντας μια βαθύτερη κατανόηση των κατανομών δεδομένων και των σχέσεων.
2. Αισθητική προσαρμογή(Aesthetic Customization): Το Seaborn διαθέτει πολλά built-in θέματα και μια ευέλικτη διεπαφή για την προσαρμογή της εμφάνισης των γραφημάτων. Αυτό διασφαλίζει ότι οι απεικονίσεις δεν είναι μόνο ενημερωτικές αλλά και αισθητικά ευχάριστες, διευκολύνοντας τη σαφή και αποτελεσματική επικοινωνία των πληροφοριών δεδομένων.
3. Πλέγματα όψεων και γραφήματα ζευγών(Facet Grids and Pair Plots): Για ανάλυση πολλαπλών μεταβλητών, το Seaborn παρέχει λειτουργίες, όπως το FacetGrid και το pairplot για τη δημιουργία πλεγμάτων γραφημάτων με βάση διαφορετικά υποσύνολα

δεδομένων ή συνδυασμούς μεταβλητών. Αυτή η δυνατότητα είναι ανεκτίμητη για την εξερεύνηση σύνθετων συνόλων δεδομένων και τον εντοπισμό μοτίβων σε πολλαπλές διαστάσεις.

Εφαρμογές στην Επιστήμη Δεδομένων

Η ικανότητα του Seaborn να δημιουργεί εξελιγμένες απεικονίσεις με ελάχιστο κώδικα το καθιστά αγαπημένο μεταξύ των επιστημόνων δεδομένων για διερευνητική ανάλυση δεδομένων. Η εστίαση της βιβλιοθήκης στα στατιστικά γραφήματα βοηθά στον έλεγχο υποθέσεων και στη λήψη αποφάσεων, ενισχύοντας τη ροή εργασίας ανάλυσης δεδομένων.

4.8 Συμπέρασμα

Η συγχώνευση αυτών των βιβλιοθηκών αποτελεί τη ραχοκοκαλιά της υπολογιστικής ανάλυσης σε αυτή τη διατριβή. Τα Pandas και NumPy μαζί αποτελούν τις πιο σημαντικές βιβλιοθήκες του χειρισμού δεδομένων και των αριθμητικών υπολογιστικών εργασιών στην Python. Το Pandas, με τις πλούσιες δομές δεδομένων και λειτουργίες του, απλοποιεί τις εργασίες καθαρισμού, μετασχηματισμού και ανάλυσης δεδομένων, καθιστώντας το απαραίτητο εργαλείο για τους επιστήμονες δεδομένων. Το NumPy, από την άλλη πλευρά, παρέχει τα θεμελιώδη δομικά στοιχεία για αποτελεσματικούς αριθμητικούς υπολογισμούς και χρησιμεύει ως βάση για ένα τεράστιο οικοσύστημα επιστημονικών υπολογιστικών βιβλιοθηκών. Η συνέργεια μεταξύ αυτών των δύο βιβλιοθηκών ενισχύει σημαντικά την παραγωγικότητα και την αποτελεσματικότητα στις ροές εργασίας ανάλυσης δεδομένων. Το Scikit-learn και το Matplotlib, το καθένα στους αντίστοιχους τομείς της μηχανικής μάθησης και της οπτικοποίησης δεδομένων, παρέχουν ισχυρά και αποτελεσματικά εργαλεία που είναι θεμελιώδη για τη ροή εργασίας της επιστήμης των δεδομένων. Το Scikit-learn απλοποιεί την εφαρμογή αλγορίθμων μηχανικής μάθησης, ενώ το Matplotlib προσφέρει ένα πλούσιο σύνολο λειτουργιών για την οπτικοποίηση σύνθετων συνόλων δεδομένων. Μαζί, αποτελούν ουσιαστικό μέρος του οικοσυστήματος επιστήμης δεδομένων Python, διευκολύνοντας την ανάπτυξη διορατικών και προγνωστικών μοντέλων και την αποτελεσματική επικοινωνία ευρημάτων που βασίζονται σε δεδομένα. Η SciPy και η Seaborn, η καθεμία στις αντίστοιχες θέσεις της επιστημονικής πληροφορικής και της στατιστικής οπτικοποίησης δεδομένων, συμβάλλουν σημαντικά στην αποδοτικότητα και την αποτελεσματικότητα των διαδικασιών ανάλυσης δεδομένων. Η SciPy προσφέρει ένα ολοκληρωμένο σύνολο μαθηματικών και επιστημονικών εργαλείων, ενώ η Seaborn απλοποιεί τη δημιουργία σύνθετων, ενημερωτικών απεικονίσεων. Μαζί, αυτές οι βιβλιοθήκες δημιουργούν μια ισχυρή εργαλειοθήκη που υποστηρίζει τις περίπλοκες διαδικασίες ανάλυσης δεδομένων, από την προεπεξεργασία έως την αξιολόγηση και την οπτικοποίηση μοντέλων.

5.Ροή εργασιών προεπεξεργασίας δεδομένων για ανάλυση μηχανικής μάθησης

5.1 Εισαγωγή

Το στάδιο προεπεξεργασίας δεδομένων είναι ένα κρίσιμο στοιχείο στο pipeline της μηχανικής μάθησης, διασφαλίζοντας ότι τα σύνολα δεδομένων είναι καθαρά, συνεπή και έτοιμα για ανάλυση. Αυτή η ενότητα περιγράφει τη συστηματική προσέγγιση που ακολουθείται για την προεπεξεργασία

πολλαπλών συνόλων δεδομένων, διασφαλίζοντας ότι βελτιστοποιούνται για επόμενες εργασίες μηχανικής μάθησης.

5.2 Αρχική αξιολόγηση συνόλου δεδομένων

Κάθε σύνολο δεδομένων, που προσδιορίζεται με το μοναδικό του όνομα, υποβάλλεται σε μια αρχική αξιολόγηση για την κατανόηση της δομής, του περιεχομένου και της ποιότητάς του. Αυτό το βήμα περιλαμβάνει την καταγραφή του ονόματος του συνόλου δεδομένων και την καταγραφή του αρχικού σχήματός του, παρέχοντας μια γραμμή βάσης για την αξιολόγηση των αποτελεσμάτων των βημάτων προεπεξεργασίας.

5.2.1 Καθαρισμός δεδομένων και Standardization:

- Επιλογή αριθμητικών στηλών: Για να διατηρηθεί η συνέπεια και η υπολογιστική σκοπιμότητα, διατηρούνται μόνο αριθμητικές στήλες για ανάλυση. Από αυτό το βήμα εξαιρούνται οι μη αριθμητικοί τύποι δεδομένων, οι οποίοι ενδέχεται να απαιτούν διαφορετικές τεχνικές προεπεξεργασίας.
- Τυποποίηση ονομάτων στηλών: Τα ονόματα των στηλών μετατρέπονται σε πεζά για την τυποποίηση της ονομασίας σε όλο το σύνολο των δεδομένων. Αυτή η πρακτική ελαχιστοποιεί πιθανά σφάλματα στο χειρισμό και την ανάλυση δεδομένων λόγω της ευαισθησίας πεζών-κεφαλαίων
- Κατάργηση διπλότυπων γραμμών: Οι διπλότυπες καταχωρίσεις καταργούνται για να διασφαλιστεί η μοναδικότητα των δεδομένων, αποτρέποντας τα στρεβλά αποτελέσματα ανάλυσης.
- Εξαίρεση στηλών μοναδικού αναγνωριστικού: Οι στήλες που περιέχουν μοναδικά αναγνωριστικά (π.χ. πρωτεύοντα κλειδιά) ή έχουν «αναγνωριστικό» στα ονόματά τους καταργούνται, καθώς δεν συμβάλλουν στην προγνωστική μοντελοποίηση ή ανάλυση.
- Εξάλειψη σταθερών στηλών: Οι στήλες με πανομοιότυπα δεδομένα σε όλες τις σειρές καταργούνται, καθώς δεν προσφέρουν διακύμανση και, επομένως, καμία προγνωστική τιμή.
- Διαγραφή στηλών με υψηλή έλλειψη: Οι στήλες με τιμές που λείπουν και υπερβαίνουν ένα προκαθορισμένο όριο (30% σε αυτήν την περίπτωση) εξαιρούνται από το σύνολο δεδομένων για λόγους διατήρησης της ποιότητας και της ακεραιότητας των δεδομένων.
- Φιλτράρισμα βάσει Z-Score: Οι ακραίες τιμές προσδιορίζονται χρησιμοποιώντας τα Z-scores, ένα μέτρο για το πόσο μακριά είναι μια παρατήρηση από το μέσο όρο, όσον αφορά στις τυπικές αποκλίσεις. Οι σειρές με οποιαδήποτε στήλη που έχει βαθμολογία Z μεγαλύτερη από 3 θεωρούνται ακραίες τιμές και αφαιρούνται, καθώς μπορούν να στρεβλώσουν σημαντικά τα αποτελέσματα της επακόλουθης ανάλυσης.
- MinMax Scaling: Το MinMaxScaler εφαρμόζεται για την κλιμάκωση αριθμητικών χαρακτηριστικών σε ένα καθορισμένο εύρος (προεπιλογή 0 έως 1). Αυτή η μέθοδος κλιμάκωσης είναι ιδιαίτερα χρήσιμη όταν οι αλγόριθμοι είναι ευαίσθητοι στο μέγεθος των χαρακτηριστικών.
- Καταλογισμός KNN: Ο αλγόριθμος K-Nearest Neighbors (KNN) χρησιμοποιείται για τον καταλογισμό τιμών που λείπουν. Αυτή η μέθοδος προβλέπει τις τιμές που λείπουν με βάση τις ομοιότητες μεταξύ των χαρακτηριστικών, παρέχοντας μια πιο λεπτή προσέγγιση από τον απλό μέσο (simple mean) ή διάμεσο καταλογισμό (median imputation).

- Κάθε σύνολο δεδομένων, τώρα καθαρισμένο, τυποποιημένο και τεκμαρτό, επαναφέρεται όσον αφορά στο ευρετήριό του για να διασφαλιστεί η συνέχεια και η ευκολία πρόσβασης. Στη συνέχεια, τα σύνολα δεδομένων συγκεντρώνονται σε μια συλλογή, σηματοδοτώντας την ολοκλήρωση της φάσης προεπεξεργασίας.

5.2.2 Καταγραφή και χειρισμός σφαλμάτων

Καθ' όλη τη διάρκεια της ροής εργασίας προεπεξεργασίας, διατηρούνται λεπτομερή αρχεία καταγραφής για την παρακολούθηση της προόδου και την τεκμηρίωση τυχόν προβλημάτων που προκύπτουν. Αυτή η πρακτική διευκολύνει τον εντοπισμό σφαλμάτων και διασφαλίζει τη διαφάνεια στη διαδικασία προετοιμασίας δεδομένων.

Η σχολαστική προεπεξεργασία των συνόλων δεδομένων θέτει τα θεμέλια για ισχυρά και αξιόπιστα μοντέλα μηχανικής μάθησης. Με την αντιμετώπιση ζητημάτων, όπως οι ελλιπείς τιμές, οι ακραίες τιμές και τα άσχετα χαρακτηριστικά, τα δεδομένα μετατρέπονται σε μια καθαρή, συνεπή μορφή που ευνοεί την ανάλυση. Αυτή η ροή προεπεξεργασίας όχι μόνο βελτιώνει την απόδοση του μοντέλου, αλλά συμβάλλει και σε πιο ακριβή και ερμηνεύσιμα αποτελέσματα, υπογραμμίζοντας τη σημασία της διεξοδικής προετοιμασίας δεδομένων στη διοχέτευση μηχανικής μάθησης.

5.3 Ανάλυση της διαδικασίας επιλογής και υπολογισμού μετα-χαρακτηριστικών

Η διαδικασία υπολογισμού των μετα-χαρακτηριστικών διαδραματίζει κρίσιμο ρόλο στην κατανόηση των χαρακτηριστικών των συνόλων δεδομένων, ειδικά στο πλαίσιο της μηχανικής μάθησης και της εξόρυξης δεδομένων. Τα μετα-χαρακτηριστικά παρέχουν πληροφορίες υψηλού επιπέδου σχετικά με το σύνολο δεδομένων, οι οποίες μπορεί να είναι καθοριστικές σε εργασίες, όπως ο χαρακτηρισμός συνόλου δεδομένων, η ανάλυση πολυπλοκότητας και η επιλογή αλγορίθμων. Η συνάρτηση “calculate_meta_features” έχει σχεδιαστεί για να εξάγει ένα ολοκληρωμένο σύνολο μετα-χαρακτηριστικών από ένα συγκεκριμένο σύνολο δεδομένων, εστιάζοντας κυρίως στα αριθμητικά χαρακτηριστικά του. Αυτή η ενότητα εμβαθύνει στο σκεπτικό πίσω από την επιλογή συγκεκριμένων μετα-χαρακτηριστικών και τη μεθοδολογία που χρησιμοποιείται στον υπολογισμό τους.

5.3.1 Επιλογή αριθμητικών δεδομένων

Η συνάρτηση ξεκινά φιλτράροντας το σύνολο δεδομένων ώστε να περιλαμβάνει μόνο αριθμητικούς τύπους δεδομένων. Αυτή η εστίαση στα αριθμητικά δεδομένα οφείλεται στην ποσοτική φύση των μετα-χαρακτηριστικών που υπολογίζονται, τα οποία απαιτούν αριθμητικές πράξεις, όπως μέσοι, τυπικές αποκλίσεις και συσχετίσεις. Αυτό το βήμα διασφαλίζει ότι οι επόμενοι υπολογισμοί είναι ουσιαστικοί και εφαρμόσιμοι.

5.3.2 Χειρισμός τιμών που λείπουν για υπολογισμό συσχέτισης

Οι τιμές που λείπουν μπορούν να επηρεάσουν σημαντικά τον υπολογισμό των συσχετίσεων, οδηγώντας σε μεροληπτικά ή απροσδιόριστα αποτελέσματα. Για να μετριαστεί αυτό, η συνάρτηση συμπληρώνει τις τιμές που λείπουν με τον μέσο όρο των αντίστοιχων στηλών τους. Αυτή η μέθοδος υπολογισμού διατηρεί τη συνολική κατανομή των δεδομένων, επιτρέποντας παράλληλα τον υπολογισμό πινάκων συσχέτισης.

5.3.3 Απλά μετα-χαρακτηριστικά(Simple Meta-features)

- Αριθμός παρουσιών (num_instances): Αντιπροσωπεύει το συνολικό αριθμό παρατηρήσεων στο σύνολο δεδομένων, παρέχοντας πληροφορίες σχετικά με το μέγεθός του και ενδεχομένως την πολυπλοκότητά του.
- Αριθμός χαρακτηριστικών (num_features): Υποδεικνύει τη διάσταση του συνόλου δεδομένων, η οποία σχετίζεται άμεσα με την «κατάρα της διαστασιακότητας» (curse of dimensionality) και μπορεί να επηρεάσει την επιλογή αλγορίθμων μηχανικής μάθησης.

5.3.4 Στατιστικά μετα-χαρακτηριστικά (Statistical Meta-features)

- Συσχετίσεις χαρακτηριστικών (Feature Correlations): Η μέση τιμή, η τυπική απόκλιση και η μέγιστη τιμή των συσχετίσεων χαρακτηριστικών (mean_feature_correlation, std_dev_feature_correlation, max_feature_correlation) υπολογίζονται για την κατανόηση των αλληλεξαρτήσεων μεταξύ των χαρακτηριστικών. Οι υψηλοί συσχετισμοί μπορεί να υποδηλώνουν πλεονασμό στα δεδομένα, ενώ ένα ευρύ φάσμα συσχετίσεων μπορεί να υποδηλώνει ποικίλες σχέσεις μεταξύ των χαρακτηριστικών.
- Τυπική απόκλιση και ασυμμετρία (Standard Deviation and Skewness): Η μέση τυπική απόκλιση (mean_feature_std_dev) και η διάμεση ασυμμετρία (median_feature_skewness) των χαρακτηριστικών παρέχουν πληροφορίες σχετικά με τη μεταβλητότητα και την ασυμμετρία της κατανομής δεδομένων. Επιπλέον, η ασυμμετρία και η κύρτωση σε επίπεδο συνόλου δεδομένων (dataset_skewness, dataset_kurtosis) προσφέρουν μια προβολή υψηλού επιπέδου των χαρακτηριστικών κατανομής των δεδομένων.

5.3.5 Πληροφοριακά-Θεωρητικά Μετα-χαρακτηριστικά (Information-Theoretic Meta-features)

Εντροπία: Η εντροπία κάθε χαρακτηριστικού υπολογίζεται για να εκτιμηθεί η ποσότητα πληροφοριών ή αβεβαιότητας μέσα στα δεδομένα. Οι μέσες και μέγιστες τιμές εντροπίας (dataset_entropy, max_feature_entropy) σε όλα τα χαρακτηριστικά παρέχουν ένα μέτρο της συνολικής μη προβλεψιμότητας και της παρουσίας τυχαιότητας στο σύνολο δεδομένων.

5.3.6 Μεθοδολογικές εκτιμήσεις

Η συνάρτηση χρησιμοποιεί διανυσματικές λειτουργίες και μεθόδους DataFrame για τον αποτελεσματικό υπολογισμό των μετα-χαρακτηριστικών, αξιοποιώντας τη βελτιστοποιημένη απόδοση των Pandas και NumPy. Ο χειρισμός εξαιρέσεων ενσωματώνεται για να διασφαλιστεί η ευρωστία, με την καταγραφή να εφαρμόζεται για την παρακολούθηση της διαδικασίας και την καταγραφή τυχόν σφαλμάτων.

5.3.7 Συμπέρασμα

Η συνάρτηση “calculate_meta_features” ενσωματώνει μια προσεκτική προσέγγιση για την εξαγωγή σημαντικών χαρακτηριστικών από σύνολα δεδομένων. Εστιάζοντας σε αριθμητικά δεδομένα και χρησιμοποιώντας μια σειρά στατιστικών και πληροφοριακών θεωρητικών μέτρων, παρέχει ένα ολοκληρωμένο προφίλ της δομής και των ιδιοτήτων του συνόλου δεδομένων. Αυτά τα μετα-χαρακτηριστικά μπορούν να διευκολύνουν διάφορες μετα-αναλυτικές εργασίες, συμπεριλαμβανομένης της σύγκρισης συνόλων δεδομένων, της αξιολόγησης πολυπλοκότητας και της καθοδήγησης της επιλογής κατάλληλων τεχνικών επεξεργασίας δεδομένων και μηχανικής μάθησης.

5.4 Μετα-χαρακτηριστικά Ανάλυση ροής εργασίας για μείωση διαστάσεων

Η συνάρτηση “meta_features_workflow” ενσωματώνει μια δομημένη προσέγγιση για τη σύνθεση και τη μείωση της διάστασης των μετα-χαρακτηριστικών που εξάγονται από τη συλλογή των προεπεξεργασμένων συνόλων δεδομένων. Αυτή η διαδικασία είναι ζωτικής σημασίας για την κατανόηση των γενικών χαρακτηριστικών των συνόλων δεδομένων και τη διευκόλυνση της σύγκρισης ή της ομαδοποίησής τους με βάση τα μετα-χαρακτηριστικά. Η ροή εργασίας περιλαμβάνει διάφορα βασικά βήματα, καθένα από τα οποία συμβάλλει στον τελικό στόχο της μείωσης των διαστάσεων και της ανάλυσης μετα-χαρακτηριστικών.

5.5 Εξαγωγή μετα-χαρακτηριστικών

Η ροή εργασίας ξεκινά από μια συνεχή εφαρμογή πάνω σε κάθε σύνολο δεδομένων από τη συλλογή των προεπεξεργασμένων δεδομένων. Για κάθε σύνολο δεδομένων, ενεργοποιείται η συνάρτηση “calculate_meta_features” για την εξαγωγή ενός ολοκληρωμένου συνόλου μετα-χαρακτηριστικών, που ενθυλακώνουν διάφορες στατιστικές και πληροφοριακές-θεωρητικές ιδιότητες του συνόλου δεδομένων. Αυτό το βήμα είναι κρίσιμο για την καταγραφή των χαρακτηριστικών υψηλού επιπέδου κάθε συνόλου δεδομένων, τα οποία είναι καθοριστικής σημασίας στις επόμενες φάσεις ανάλυσης.

5.6 Συλλογή μετα-χαρακτηριστικών

Τα μετα-χαρακτηριστικά που εξάγονται για κάθε σύνολο δεδομένων συλλέγονται σε έναν πίνακα, ο οποίος στη συνέχεια μετατρέπεται σε DataFrame (metafeatures_df). Αυτό το βήμα συλλογής

διευκολύνει τον χειρισμό και την ανάλυση των μετα-χαρακτηριστικών σε όλα τα σύνολα δεδομένων, παρέχοντας μια ενοποιημένη προβολή των εξαγόμενων χαρακτηριστικών.

5.7 Κανονικοποίηση μετα-χαρακτηριστικών

Δεδομένου του ποικίλου εύρους και της κλίμακας των μετα-χαρακτηριστικών, η κανονικοποίηση χρησιμοποιείται για την τυποποίηση των κλιμάκων τους. Το StandardScaler εφαρμόζεται στο DataFrame των μετα-χαρακτηριστικών, διασφαλίζοντας ότι κάθε μετα-χαρακτηριστικό έχει μέσο όρο 0 και τυπική απόκλιση 1. Η διαδικασία περιλαμβάνει την αφαίρεση του μέσου όρου κάθε χαρακτηριστικού από τα σημεία δεδομένων και στη συνέχεια τη διαίρεση με την τυπική απόκλιση κάθε χαρακτηριστικού:

$$z = \frac{(x - \mu)}{\sigma}$$

όπου:

z είναι η τυποποιημένη τιμή,

x είναι η αρχική τιμή,

μ είναι ο μέσος όρος του χαρακτηριστικού, και

σ είναι η τυπική απόκλιση του χαρακτηριστικού.

Μετά από αυτόν τον μετασχηματισμό, η προκύπτουσα κατανομή κάθε χαρακτηριστικού επικεντρώνεται γύρω από το 0 (μέσος όρος = 0) και έχει διακύμανση μονάδας (τυπική απόκλιση = 1). Αυτό μπορεί να είναι ιδιαίτερα χρήσιμο σε αλγόριθμους που υποθέτουν ότι τα δεδομένα διανέμονται κανονικά ή σε αλγόριθμους που είναι ευαίσθητοι στην κλίμακα των δεδομένων, όπως οι μηχανές διανυσμάτων υποστήριξης (Support Vector Machines) ή οι k-πλησιέστεροι γείτονες. Αυτή η κανονικοποίηση είναι ζωτικής σημασίας για την πρόληψη των αποκλίσεων κλίμακας που μπορούν να επηρεάσουν τα αποτελέσματα αλγορίθμων μείωσης διαστάσεων και ομαδοποίησης που ακολουθούν.

5.8 Μείωση διαστάσεων με χρήση PCA

Η Ανάλυση Κύριων Συνιστωσών (PCA) χρησιμοποιείται για τη μείωση των διαστάσεων των κανονικοποιημένων μετα-χαρακτηριστικών. Επιλέγοντας έναν αριθμό-στόχο κύριων συνιστωσών (σε αυτή την περίπτωση επιλέχθηκαν 2), το PCA μετατρέπει τα μετα-χαρακτηριστικά σε ένα νέο χώρο όπου οι άξονες (κύρια συστατικά) συλλαμβάνουν τη μέγιστη διακύμανση εντός των δεδομένων. Αυτό το βήμα είναι θεμελιώδες για την απόσταξη της ουσίας των μετα-χαρακτηριστικών σε μια πιο διαχειρίσιμη και ερμηνεύσιμη μορφή, διευκολύνοντας την οπτική αναπαράσταση και την περαιτέρω ανάλυση.

5.9 Αποτέλεσμα και ερμηνεία

Η συνάρτηση εξάγει τα μετασχηματισμένα μετα-χαρακτηριστικά (metafeatures_pca) μαζί με την εξηγούμενη αναλογία διακύμανσης των κύριων συνιστωσών. Ο εξηγούμενος λόγος διακύμανσης

παρέχει πληροφορίες σχετικά με το ποσοστό της διακύμανσης του συνόλου δεδομένων που συλλαμβάνεται από κάθε κύρια συνιστώσα, χρησιμεύοντας ως μέτρο της αποτελεσματικότητας της μείωσης των διαστάσεων.

5.10 Χειρισμός σφαλμάτων και καταγραφή

Καθ' όλη τη διάρκεια της ροής εργασίας, υπάρχουν ισχυροί μηχανισμοί χειρισμού σφαλμάτων και καταγραφής για τη διασφάλιση της ομαλής εκτέλεσης και την παροχή ενημερωτικών σχολίων σχετικά με τη διαδικασία. Τυχόν εξαιρέσεις που συναντώνται κατά τον υπολογισμό μετα-χαρακτηριστικών ή το PCA καταγράφονται, βοηθώντας στην αντιμετώπιση προβλημάτων και τη βελτίωση της διαδικασίας.

5.11 Συμπέρασμα

Η συνάρτηση "meta_features_workflow" ενσωματώνει μια ολοκληρωμένη προσέγγιση για την εξαγωγή μετα-χαρακτηριστικών, την κανονικοποίηση και τη μείωση των διαστάσεων. Αξιοποιώντας τη στατιστική τυποποίηση και το PCA, συμπυκνώνει αποτελεσματικά τον υψηλής διάστασης χώρο των μετα-χαρακτηριστικών σε μια πιο ερμηνεύσιμη και αναλύσιμη μορφή. Αυτή η ροή εργασίας όχι μόνο ενισχύει την κατανόηση των χαρακτηριστικών του συνόλου των δεδομένων σε μετα-επίπεδο, αλλά ανοίγει και το δρόμο για προηγμένες αναλύσεις, όπως η μετα-μάθηση και η ομαδοποίηση συνόλων δεδομένων με βάση εγγενείς ιδιότητες.

6.Ανάλυση ομαδοποίησης μετα-χαρακτηριστικών μειωμένων διαστάσεων

6.1 Εισαγωγή

Η ανάλυση ομαδοποίησης των μετα-χαρακτηριστικών μειωμένων διαστάσεων είναι ένα σημαντικό βήμα για την κατανόηση των εγγενών ομαδοποιήσεων εντός του συνόλου των δεδομένων με βάση τα χαρακτηριστικά υψηλού επιπέδου τους. Αυτή η διαδικασία περιλαμβάνει την εφαρμογή διαφόρων αλγορίθμων ομαδοποίησης στα μετα-χαρακτηριστικά μετασχηματισμού PCA, με στόχο την αποκάλυψη μοτίβων και ομοιοτήτων που μπορεί να μην είναι άμεσα εμφανείς. Η συνάρτηση "clustering_and_plot" ενσωματώνει αυτή την ανάλυση, χρησιμοποιώντας μια σειρά τεχνικών ομαδοποίησης και οπτικοποιώντας τα αποτελέσματά τους.

6.2 Επιλογή αλγορίθμων ομαδοποίησης

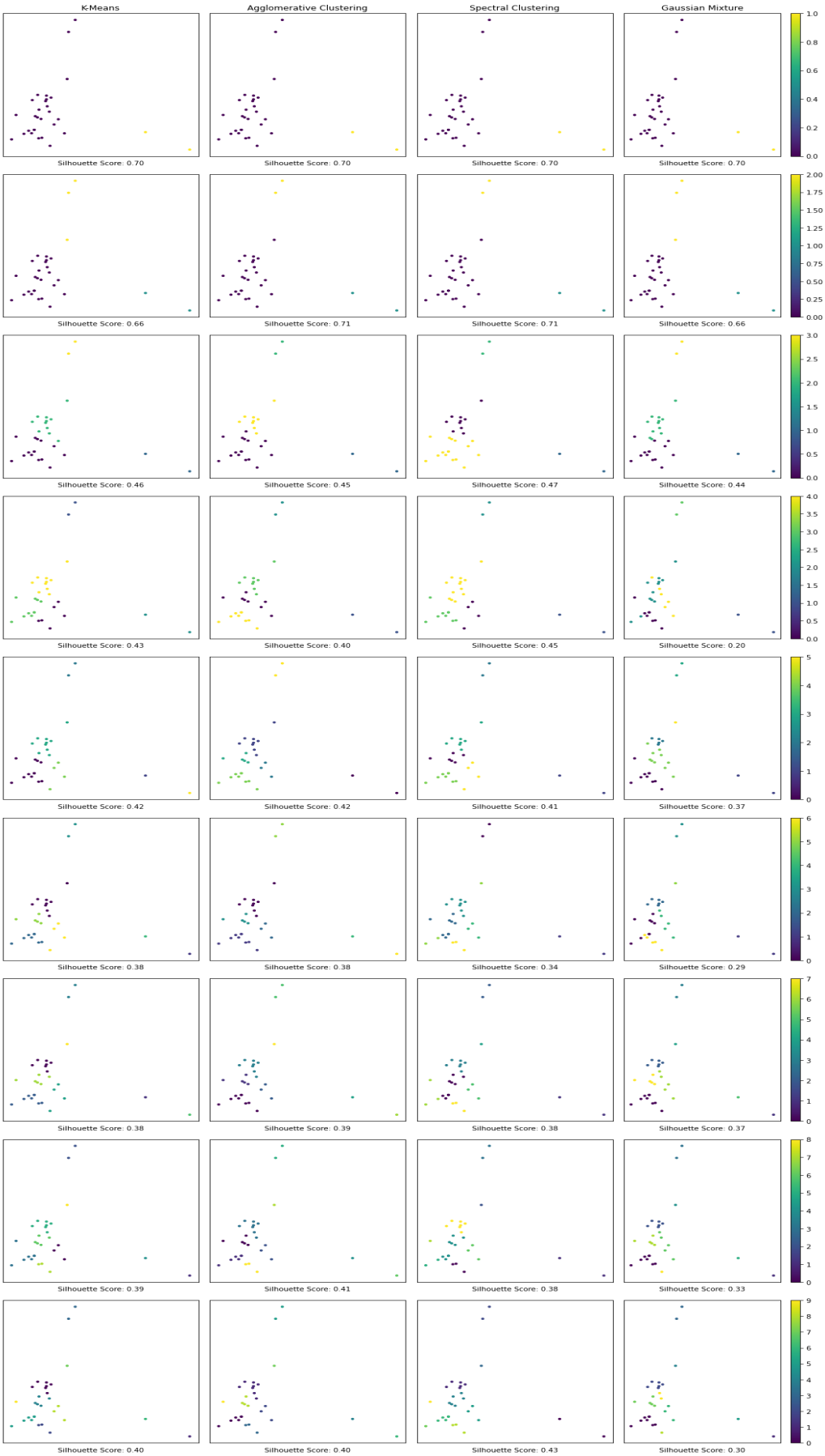
Τέσσερις αλγόριθμοι ομαδοποίησης επιλέγονται για αυτήν την ανάλυση, καθένας από τους οποίους φέρνει μια μοναδική προσέγγιση στην ομαδοποίηση δεδομένων:

- K-Means: Μια ευρέως χρησιμοποιούμενη μέθοδος διαμέρισης που διαιρεί τα δεδομένα σε k clusters ελαχιστοποιώντας τις διακυμάνσεις εντός του συμπλέγματος.
- Agglomerative Clustering: Μια ιεραρχική τεχνική ομαδοποίησης που δημιουργεί ένθετα συμπλέγματα με προοδευτική συγχώνευση ή διαίρεση υπαρχόντων συμπλεγμάτων βάσει μετρήσεων απόστασης.
- Spectral Clustering: Χρησιμοποιεί το φάσμα (ιδιοτιμές) του πίνακα ομοιότητας των δεδομένων για να εκτελέσει μείωση διαστάσεων πριν από την ομαδοποίηση, ιδιαίτερα αποτελεσματική για μη κυρτά σμήνη.
- Gaussian Mix: Ένα πιθανοτικό μοντέλο που υποθέτει ότι τα σημεία δεδομένων παράγονται από ένα μείγμα διαφόρων κατανομών Gauss, παρέχοντας μια προσέγγιση μαλακής ομαδοποίησης.

6.3 Διαδικασία ομαδοποίησης και οπτικοποίηση

Για κάθε αλγόριθμο ομαδοποίησης, η συνάρτηση επαναλαμβάνεται σε ένα προκαθορισμένο εύρος αριθμών συμπλέγματος (από 2 έως 10) για να εξερευνήσει διαφορετικές διαμορφώσεις ομαδοποίησης. Αυτό το εύρος επιλέγεται για να παρέχει μια ευρεία εικόνα των πιθανών ομαδοποιήσεων χωρίς μεγάλη υπολογιστική πολυπλοκότητα.

Για κάθε ρύθμιση παραμέτρων, ο αλγόριθμος εφαρμόζεται στα μετα-χαρακτηριστικά με μειωμένη PCA (metafeatures_pca) και οι ετικέτες που προκύπτουν χρησιμοποιούνται για τη σχεδίαση των σημείων δεδομένων σε μια γραφική παράσταση διασποράς. Κάθε δευτερεύουσα γραφική παράσταση αντιπροσωπεύει ένα διαφορετικό αποτέλεσμα ομαδοποίησης, με τα σημεία δεδομένων χρωματισμένα με βάση το εκχωρημένο σύμπλεγμά τους. Η βαθμολογία σιλουέτας, ένα μέτρο της συνοχής και του διαχωρισμού των συστάδων, υπολογίζεται και εμφανίζεται για κάθε διαμόρφωση, προσφέροντας μια ποσοτική αξιολόγηση της ποιότητας της ομαδοποίησης.



Εικόνα 1 Εμφάνιση clusters από 1-9 συσταδοποίησεις για τους 4 αλγόριθμους της εργασίας

6.4 Εκτιμήσεις και προσαρμογές

Η χρήση της παραμέτρου random_state εξασφαλίζει αναπαραγωγιμότητα σε αλγορίθμους που περιλαμβάνουν στοχαστικές διαδικασίες, όπως το K-Means και το Gaussian Mix. Η συνάρτηση χρησιμοποιεί μια διάταξη πλέγματος (grid) για τα subplots, με κάθε σειρά να αντιστοιχεί σε διαφορετικό αριθμό συστάδων και κάθε στήλη να αντιπροσωπεύει διαφορετικό αλγόριθμο ομαδοποίησης. Αυτή η διάταξη διευκολύνει τη σύγκριση των αποτελεσμάτων ομαδοποίησης μεταξύ αλγορίθμων και αριθμών συμπλέγματος.

6.5 Χειρισμός σφαλμάτων και καταγραφή

Οι ισχυροί μηχανισμοί χειρισμού σφαλμάτων και καταγραφής είναι ενσωματωμένοι για την καταγραφή και την αναφορά τυχόν προβλημάτων που προέκυψαν κατά τη διαδικασία ομαδοποίησης. Αυτή η προληπτική προσέγγιση βοηθά στην αντιμετώπιση προβλημάτων και διασφαλίζει την ακεραιότητα της ανάλυσης.

6.6 Συμπέρασμα

Η συνάρτηση “clustering_and_plot” προσφέρει ένα ολοκληρωμένο πλαίσιο για την ανάλυση ομαδοποίησης των μετα-χαρακτηριστικών μειωμένων διαστάσεων, αξιοποιώντας ένα ποικίλο σύνολο αλγορίθμων για την αποκάλυψη εγγενών ομαδοποιήσεων δεδομένων. Μέσω της επαναληπτικής εξερεύνησης και της συστηματικής οπτικοποίησης, παρέχει πολύτιμες πληροφορίες σχετικά με τις σχέσεις και τις ομοιότητες μεταξύ των συνόλων δεδομένων σε μετα-επίπεδο. Αυτή η ανάλυση όχι μόνο ενισχύει την κατανόηση των χαρακτηριστικών του συνόλου δεδομένων, αλλά ενημερώνει επίσης τις στρατηγικές διαχείρισης δεδομένων και την επιλογή αλγορίθμων για επακόλουθες εργασίες μηχανικής μάθησης. Αυτή η λεπτομερής εξέταση της διαδικασίας ανάλυσης ομαδοποίησης υπογραμμίζει την αξία της στην αποκάλυψη σύνθετων μοτίβων μέσα σε σύνολα δεδομένων, εμπλουτίζοντας έτσι τις φάσεις προεπεξεργασίας και εξερεύνησης δεδομένων των εργασιών της μηχανικής μάθησης.

6.7 Αξιολόγηση της απόδοσης ομαδοποίησης χρησιμοποιώντας μετρικές

Η συνάρτηση “print_metrics” και η ενσωμάτωσή της στη ροή εργασιών ομαδοποίησης διαδραματίζουν κεντρικό ρόλο στην αξιολόγηση της αποτελεσματικότητας διαφορετικών αλγορίθμων ομαδοποίησης σε μετα-χαρακτηριστικά μειωμένων διαστάσεων. Η αξιολόγηση αυτή είναι ζωτικής σημασίας για την κατανόηση των εγγενών ομαδοποιήσεων εντός των δεδομένων και την επιλογή της καταλληλότερης προσέγγισης ομαδοποίησης. Η διαδικασία περιλαμβάνει τον υπολογισμό και την ανάλυση τριών βασικών μετρήσεων:

- Silhouette Score
- Calinski-Harabasz Index
- Davies-Bouldin Index.

Number of Clusters: 2
2024-03-25 13:45:23,432 - INFO -
KMeans Clustering:
2024-03-25 13:45:23,440 - INFO - Silhouette Score: 0.7016926707639081
2024-03-25 13:45:23,441 - INFO - Calinski-Harabasz Index: 24.696761721757714
2024-03-25 13:45:23,442 - INFO - Davies-Bouldin Index: 0.39337481902034815
2024-03-25 13:45:23,444 - INFO -
Agglomerative Clustering:
2024-03-25 13:45:23,452 - INFO - Silhouette Score: 0.7016926707639081
2024-03-25 13:45:23,454 - INFO - Calinski-Harabasz Index: 24.696761721757714
2024-03-25 13:45:23,455 - INFO - Davies-Bouldin Index: 0.39337481902034815
2024-03-25 13:45:23,466 - INFO -
GMM Clustering:
2024-03-25 13:45:23,472 - INFO - Silhouette Score: 0.5516640508674385
2024-03-25 13:45:23,473 - INFO - Calinski-Harabasz Index: 12.628214418480601
2024-03-25 13:45:23,474 - INFO - Davies-Bouldin Index: 0.6262753006953257
2024-03-25 13:45:23,520 - INFO -
Spectral Clustering:
2024-03-25 13:45:23,527 - INFO - Silhouette Score: 0.7016926707639081
2024-03-25 13:45:23,528 - INFO - Calinski-Harabasz Index: 24.696761721757714
2024-03-25 13:45:23,529 - INFO - Davies-Bouldin Index: 0.39337481902034815
2024-03-25 13:45:23,531 - INFO -
Number of Clusters: 3
2024-03-25 13:45:23,575 - INFO -
KMeans Clustering:
2024-03-25 13:45:23,587 - INFO - Silhouette Score: 0.6584116067787859
2024-03-25 13:45:23,588 - INFO - Calinski-Harabasz Index: 49.239047348725336
2024-03-25 13:45:23,589 - INFO - Davies-Bouldin Index: 0.4555746174103638
2024-03-25 13:45:23,592 - INFO -
Agglomerative Clustering:
2024-03-25 13:45:23,600 - INFO - Silhouette Score: 0.7130112952714299
2024-03-25 13:45:23,601 - INFO - Calinski-Harabasz Index: 47.87354326650443
2024-03-25 13:45:23,602 - INFO - Davies-Bouldin Index: 0.3152841741857307
2024-03-25 13:45:23,616 - INFO -
GMM Clustering:
2024-03-25 13:45:23,623 - INFO - Silhouette Score: 0.6584116067787859
2024-03-25 13:45:23,624 - INFO - Calinski-Harabasz Index: 49.239047348725336
2024-03-25 13:45:23,626 - INFO - Davies-Bouldin Index: 0.4555746174103638
2024-03-25 13:45:23,683 - INFO -
Spectral Clustering:
2024-03-25 13:45:23,701 - INFO - Silhouette Score: 0.7130112952714299
2024-03-25 13:45:23,704 - INFO - Calinski-Harabasz Index: 47.87354326650443
2024-03-25 13:45:23,709 - INFO - Davies-Bouldin Index: 0.3152841741857307
2024-03-25 13:45:23,724 - INFO -
Number of Clusters: 4
2024-03-25 13:45:23,880 - INFO -
KMeans Clustering:
2024-03-25 13:45:23,889 - INFO - Silhouette Score: 0.46105913487948047
2024-03-25 13:45:23,893 - INFO - Calinski-Harabasz Index: 66.21038100107786
2024-03-25 13:45:23,894 - INFO - Davies-Bouldin Index: 0.5191247149151964
2024-03-25 13:45:23,898 - INFO -
Agglomerative Clustering:
2024-03-25 13:45:23,908 - INFO - Silhouette Score: 0.4520529248317854
2024-03-25 13:45:23,909 - INFO - Calinski-Harabasz Index: 63.8650792743607
2024-03-25 13:45:23,911 - INFO - Davies-Bouldin Index: 0.5073387869765577
2024-03-25 13:45:23,942 - INFO -
GMM Clustering:
2024-03-25 13:45:23,950 - INFO - Silhouette Score: 0.4446459705893575
2024-03-25 13:45:23,951 - INFO - Calinski-Harabasz Index: 61.458711527156936
2024-03-25 13:45:23,953 - INFO - Davies-Bouldin Index: 0.5766412987708252
2024-03-25 13:45:24,041 - INFO -

Spectral Clustering:
2024-03-25 13:45:24,050 - INFO - Silhouette Score: 0.46105913487948047
2024-03-25 13:45:24,052 - INFO - Calinski-Harabasz Index: 66.21038100107785
2024-03-25 13:45:24,053 - INFO - Davies-Bouldin Index: 0.5191247149151964
2024-03-25 13:45:24,054 - INFO -
Number of Clusters: 5
2024-03-25 13:45:24,112 - INFO -
KMeans Clustering:
2024-03-25 13:45:24,122 - INFO - Silhouette Score: 0.4539785625772311
2024-03-25 13:45:24,124 - INFO - Calinski-Harabasz Index: 66.30742877645282
2024-03-25 13:45:24,125 - INFO - Davies-Bouldin Index: 0.5895247871470216
2024-03-25 13:45:24,128 - INFO -
Agglomerative Clustering:
2024-03-25 13:45:24,140 - INFO - Silhouette Score: 0.39576934382664336
2024-03-25 13:45:24,141 - INFO - Calinski-Harabasz Index: 61.38986649637344
2024-03-25 13:45:24,143 - INFO - Davies-Bouldin Index: 0.7647966418872073
2024-03-25 13:45:24,179 - INFO -
GMM Clustering:
2024-03-25 13:45:24,187 - INFO - Silhouette Score: 0.40364344341159125
2024-03-25 13:45:24,188 - INFO - Calinski-Harabasz Index: 52.71423379351341
2024-03-25 13:45:24,189 - INFO - Davies-Bouldin Index: 0.44736210785671837
2024-03-25 13:45:24,249 - INFO -
Spectral Clustering:
2024-03-25 13:45:24,259 - INFO - Silhouette Score: 0.45047495106504937
2024-03-25 13:45:24,260 - INFO - Calinski-Harabasz Index: 62.584917493996635
2024-03-25 13:45:24,261 - INFO - Davies-Bouldin Index: 0.5644776924475338
2024-03-25 13:45:24,262 - INFO -
Number of Clusters: 6
2024-03-25 13:45:24,318 - INFO -
KMeans Clustering:
2024-03-25 13:45:24,328 - INFO - Silhouette Score: 0.394592424005015
2024-03-25 13:45:24,329 - INFO - Calinski-Harabasz Index: 65.51341256781068
2024-03-25 13:45:24,330 - INFO - Davies-Bouldin Index: 0.5346296482561802
2024-03-25 13:45:24,333 - INFO -
Agglomerative Clustering:
2024-03-25 13:45:24,347 - INFO - Silhouette Score: 0.4195383568914647
2024-03-25 13:45:24,348 - INFO - Calinski-Harabasz Index: 63.03935483067662
2024-03-25 13:45:24,349 - INFO - Davies-Bouldin Index: 0.5879616110339275
2024-03-25 13:45:24,411 - INFO -
GMM Clustering:
2024-03-25 13:45:24,421 - INFO - Silhouette Score: 0.30111529577061275
2024-03-25 13:45:24,422 - INFO - Calinski-Harabasz Index: 54.725652614114395
2024-03-25 13:45:24,423 - INFO - Davies-Bouldin Index: 0.6348797102327728
2024-03-25 13:45:24,487 - INFO -
Spectral Clustering:
2024-03-25 13:45:24,497 - INFO - Silhouette Score: 0.40259725910196204
2024-03-25 13:45:24,498 - INFO - Calinski-Harabasz Index: 64.36509727566234
2024-03-25 13:45:24,500 - INFO - Davies-Bouldin Index: 0.6340599909485568
2024-03-25 13:45:24,501 - INFO -
Number of Clusters: 7
2024-03-25 13:45:24,562 - INFO -
KMeans Clustering:
2024-03-25 13:45:24,573 - INFO - Silhouette Score: 0.3755943759071448
2024-03-25 13:45:24,574 - INFO - Calinski-Harabasz Index: 70.081937696509
2024-03-25 13:45:24,576 - INFO - Davies-Bouldin Index: 0.5075058491763162
2024-03-25 13:45:24,578 - INFO -
Agglomerative Clustering:
2024-03-25 13:45:24,592 - INFO - Silhouette Score: 0.3833494067001565
2024-03-25 13:45:24,594 - INFO - Calinski-Harabasz Index: 68.27102236600261
2024-03-25 13:45:24,595 - INFO - Davies-Bouldin Index: 0.48580292165359124
2024-03-25 13:45:24,637 - INFO -
GMM Clustering:
2024-03-25 13:45:24,653 - INFO - Silhouette Score: 0.32296408943115323

2024-03-25 13:45:24,654 - INFO - Calinski-Harabasz Index: 59.48541766735094
2024-03-25 13:45:24,655 - INFO - Davies-Bouldin Index: 0.6049191196888861
2024-03-25 13:45:24,741 - INFO -
Spectral Clustering:
2024-03-25 13:45:24,752 - INFO - Silhouette Score: 0.38548688023809163
2024-03-25 13:45:24,753 - INFO - Calinski-Harabasz Index: 53.23936066731837
2024-03-25 13:45:24,754 - INFO - Davies-Bouldin Index: 0.6037675802299125
2024-03-25 13:45:24,755 - INFO -
Number of Clusters: 8
2024-03-25 13:45:24,822 - INFO -
KMeans Clustering:
2024-03-25 13:45:24,835 - INFO - Silhouette Score: 0.37158027944870187
2024-03-25 13:45:24,836 - INFO - Calinski-Harabasz Index: 79.36937832979085
2024-03-25 13:45:24,837 - INFO - Davies-Bouldin Index: 0.4747524623743483
2024-03-25 13:45:24,844 - INFO -
Agglomerative Clustering:
2024-03-25 13:45:24,855 - INFO - Silhouette Score: 0.39211681754918426
2024-03-25 13:45:24,857 - INFO - Calinski-Harabasz Index: 78.45809326514444
2024-03-25 13:45:24,863 - INFO - Davies-Bouldin Index: 0.44795504907714717
2024-03-25 13:45:24,993 - INFO -
GMM Clustering:
2024-03-25 13:45:25,005 - INFO - Silhouette Score: 0.35387702787478226
2024-03-25 13:45:25,007 - INFO - Calinski-Harabasz Index: 73.65578492835792
2024-03-25 13:45:25,008 - INFO - Davies-Bouldin Index: 0.4588956767178946
2024-03-25 13:45:25,094 - INFO -
Spectral Clustering:
2024-03-25 13:45:25,105 - INFO - Silhouette Score: 0.3794028025472168
2024-03-25 13:45:25,106 - INFO - Calinski-Harabasz Index: 56.81250865610377
2024-03-25 13:45:25,107 - INFO - Davies-Bouldin Index: 0.6191398333740401
2024-03-25 13:45:25,108 - INFO -
Number of Clusters: 9
2024-03-25 13:45:25,187 - INFO -
KMeans Clustering:
2024-03-25 13:45:25,199 - INFO - Silhouette Score: 0.3867864973043854
2024-03-25 13:45:25,200 - INFO - Calinski-Harabasz Index: 91.34306394077433
2024-03-25 13:45:25,201 - INFO - Davies-Bouldin Index: 0.44541483202482646
2024-03-25 13:45:25,205 - INFO -
Agglomerative Clustering:
2024-03-25 13:45:25,218 - INFO - Silhouette Score: 0.40551529472198067
2024-03-25 13:45:25,219 - INFO - Calinski-Harabasz Index: 89.88943438693596
2024-03-25 13:45:25,221 - INFO - Davies-Bouldin Index: 0.4643391163903593
2024-03-25 13:45:25,312 - INFO -
GMM Clustering:
2024-03-25 13:45:25,326 - INFO - Silhouette Score: 0.2874671667432968
2024-03-25 13:45:25,327 - INFO - Calinski-Harabasz Index: 67.69982327945533
2024-03-25 13:45:25,328 - INFO - Davies-Bouldin Index: 0.5673251297792251
2024-03-25 13:45:25,415 - INFO -
Spectral Clustering:
2024-03-25 13:45:25,427 - INFO - Silhouette Score: 0.3596120388103604
2024-03-25 13:45:25,429 - INFO - Calinski-Harabasz Index: 52.24288801515447
2024-03-25 13:45:25,431 - INFO - Davies-Bouldin Index: 0.629242685511702

6.8 Βαθμολογία σιλουέτας (Silhouette Score)

Ορισμός: Η βαθμολογία σιλουέτας μετρά πόσο παρόμοιο είναι ένα αντικείμενο με το δικό του σύμπλεγμα σε σύγκριση με άλλα σμήνη. Η τιμή κυμαίνεται από -1 έως 1, όπου μια υψηλή τιμή υποδεικνύει ότι το αντικείμενο ταιριάζει καλά με το δικό του σύμπλεγμα και δεν ταιριάζει καλά με γειτονικά συμπλέγματα.

Χρήση: Σε αυτή τη ροή εργασίας, η βαθμολογία σιλουέτας υπολογίζεται για κάθε αποτέλεσμα ομαδοποίησης, παρέχοντας πληροφορίες σχετικά με τη συνοχή και τον διαχωρισμό των σχηματισμένων συμπλεγμάτων. Μια υψηλότερη μέση βαθμολογία υποδηλώνει καλύτερα καθορισμένες ομάδες.

6.9 Δείκτης Calinski-Harabasz

Ορισμός: Γνωστός και ως κριτήριο λόγου διακύμανσης, ο δείκτης αυτός είναι ο λόγος του αθροίσματος της διασποράς μεταξύ συστάδων και της διασποράς εντός συστάδων για όλες τις συστάδες. Οι υψηλότερες τιμές γενικά υποδεικνύουν ότι οι συστάδες είναι πυκνές και καλά διαχωρισμένες, κάτι που είναι επιθυμητό.

Χρήση: Ο δείκτης Calinski-Harabasz συμπληρώνει τη βαθμολογία σιλουέτας προσφέροντας ένα εναλλακτικό μέτρο της ποιότητας των συστάδων, ιδιαίτερα χρήσιμο για την αξιολόγηση του πόσο συμπαγής είναι μια συστάδα καθώς και για τον διαχωρισμό των συστάδων.

6.10 Δείκτης Davies-Bouldin

Ορισμός: Αυτός ο δείκτης δηλώνει τη μέση «ομοιότητα» μεταξύ των συστάδων, όπου η ομοιότητα είναι ένα μέτρο που συγκρίνει την απόσταση μεταξύ των συστάδων με το μέγεθος των ίδιων των συστάδων. Ένας χαμηλότερος δείκτης Davies-Bouldin σχετίζεται με μια καλύτερη διαμόρφωση ομαδοποίησης.

Χρήση: Αξιολογώντας τον δείκτη Davies-Bouldin για κάθε σύνολο συμπλεγμάτων, η ροή εργασίας παρέχει μια πρόσθετη προοπτική για την απόδοση ομαδοποίησης, εστιάζοντας στη διασπορά και τον διαχωρισμό συμπλεγμάτων.

6.11 Επαναληπτική ομαδοποίηση και αξιολόγηση των μετρικών

Η ροή εργασίας εφαρμόζει επαναληπτικά τέσσερις διακριτούς αλγόριθμους ομαδοποίησης (K-Means, Agglomerative Clustering, Gaussian Mix Models και Spectral Clustering) σε μια σειρά αριθμητικών συμπλεγμάτων. Για κάθε ρύθμιση παραμέτρων, η ομαδοποίηση εκτελείται στα μετα-χαρακτηριστικά που έχουν μειωθεί μέσω PCA και υπολογίζονται οι καθορισμένες μετρικές για την αξιολόγηση της ποιότητας του συμπλέγματος. Αυτή η επαναληπτική προσέγγιση επιτρέπει μια ολοκληρωμένη σύγκριση αλγορίθμων και των ρυθμίσεων τους, βοηθώντας στην επιλογή της πιο αποτελεσματικής στρατηγικής ομαδοποίησης βάσει ποσοτικών μετρικών.

6.11.1 Καταγραφή και χειρισμός σφαλμάτων

Καθ’ όλη τη διάρκεια της διαδικασίας, η λεπτομερής καταγραφή παρέχει ανατροφοδότηση σε πραγματικό χρόνο σχετικά με την απόδοση του συμπλέγματος, συμπεριλαμβανομένων των υπολογισμένων μετρικών τιμών για κάθε αλγόριθμο και του αριθμού συμπλέγματος, όπως έχει

ρυθμιστεί. Οι μηχανισμοί χειρισμού σφαλμάτων διασφαλίζουν την ευρωστία της ροής εργασίας, καταγράφοντας τυχόν εξαιρέσεις κατά τον υπολογισμό των μετρικών ή την ομαδοποίηση, διευκολύνοντας έτσι την αντιμετώπιση προβλημάτων και διασφαλίζοντας τη συνέχεια της ανάλυσης.

6.11.2 Συμπέρασμα

Η αξιολόγηση της απόδοσης ομαδοποίησης χρησιμοποιώντας τη συνάρτηση “print_metrics” και η επακόλουθη επαναληπτική διαδικασία ομαδοποίησης παρέχει ένα συστηματικό και ποσοτικό πλαίσιο για την ανάλυση των τάσεων ομαδοποίησης εντός του συνόλου των δεδομένων. Αξιοποιώντας τις βασικές μετρικές, όπως ο Silhouette Score, ο Calinski-Harabasz Index και ο Davies-Bouldin Index, η ροή εργασίας προσφέρει μια λεπτή κατανόηση της ποιότητας ομαδοποίησης, καθοδηγώντας την επιλογή του καταλληλότερου αλγορίθμου ομαδοποίησης και διαμόρφωσης για τα δεδομένα που επεξεργαζόμαστε. Αυτή η αυστηρή προσέγγιση στην αξιολόγηση ομαδοποίησης είναι απαραίτητη για την αποκάλυψη ουσιαστικών προτύπων στα δεδομένα και τη λήψη τεκμηριωμένων αποφάσεων στο ευρύτερο πλαίσιο των έργων μηχανικής μάθησης και ανάλυσης δεδομένων. Η λεπτομερής ανάλυση της διαδικασίας αξιολόγησης υπογραμμίζει τον κρίσιμο ρόλο της στη διασφάλιση της αποτελεσματικότητας και της αξιοπιστίας των αποτελεσμάτων ομαδοποίησης, εμπλουτίζοντας έτσι τις φάσεις εξερεύνησης και ανάλυσης δεδομένων των έργων μηχανικής μάθησης.

6.12 Συστηματική Αξιολόγηση και Βελτιστοποίηση Αλγορίθμων Ομαδοποίησης

Η συστηματική αξιολόγηση και βελτιστοποίηση των αλγορίθμων ομαδοποίησης μέσω της συνάρτησης “calc_results” και η επακόλουθη ανάλυση αντιπροσωπεύουν μια σχολαστική προσέγγιση για τον εντοπισμό των πιο αποτελεσματικών διαμορφώσεων ομαδοποίησης. Αυτή η διαδικασία είναι καθοριστική για τη βελτιστοποίηση της ομαδοποίησης των συνόλων δεδομένων με βάση τα μετα-χαρακτηριστικά τους, διευκολύνοντας τελικά μια βαθύτερη κατανόηση των χαρακτηριστικών και των σχέσεων στο σύνολο των δεδομένων.

6.12.1 Υπολογισμός και καταγραφή μετρήσεων ομαδοποίησης

Για κάθε αποτέλεσμα ομαδοποίησης, η συνάρτηση υπολογίζει τρεις κρίσιμες μετρικές: Silhouette Score, Calinski-Harabasz Index και Davies-Bouldin Index. Αυτές οι μετρικές προσφέρουν συλλογικά μια ολοκληρωμένη αξιολόγηση της ποιότητας των συμπλεγμάτων, λαμβάνοντας υπόψη πτυχές, όπως η συνοχή, ο διαχωρισμός και η συμπαγής.

- Το Silhouette Score αξιολογεί πόσο κοντά είναι κάθε σημείο σε ένα σύμπλεγμα με σημεία στα γειτονικά σμήνη, παρέχοντας έτσι πληροφορίες σχετικά με την καταλληλότητα των εκχωρήσεων συμπλέγματος.
- Ο δείκτης Calinski-Harabasz αξιολογεί την εγκυρότητα του συμπλέγματος με βάση την αναλογία του αθροίσματος μεταξύ των συστάδων προς το άθροισμα των τετραγώνων εντός του συμπλέγματος, επισημαίνοντας την πυκνότητα και τον καλό διαχωρισμό των συστάδων.

- Ο δείκτης Davies-Bouldin μετρά τη μέση ομοιότητα μεταξύ κάθε συμπλέγματος και του πιο παρόμοιου, όπου οι χαμηλότερες τιμές υποδεικνύουν καλύτερη ομαδοποίηση.

Κάθε σύνολο μετρήσεων προσαρτάται σε μια λίστα αποτελεσμάτων, δημιουργώντας ένα ολοκληρωμένο αρχείο απόδοσης ομαδοποίησης σε διάφορους αλγόριθμους και αριθμούς συμπλεγμάτων.

6.12.2 Επαναληπτική ομαδοποίηση σε αλγόριθμους και μετρήσεις συμπλεγμάτων

Η επαναληπτική διαδικασία περιλαμβάνει πολλαπλούς αλγόριθμους ομαδοποίησης (K-Means, Agglomerative Clustering, Gaussian Mix Models και Spectral Clustering) σε ένα εύρος αριθμών συστάδων (από 2 έως 9). Αυτό το εύρος εξασφαλίζει μια διεξοδική εξερεύνηση πιθανών διαμορφώσεων ομαδοποίησης, επιτρέποντας τον προσδιορισμό των βέλτιστων ρυθμίσεων με βάση τις υπολογισμένες μετρήσεις.

6.12.3 Ανάλυση και προσδιορισμός βέλτιστων διαμορφώσεων ομαδοποίησης

Μετά τον υπολογισμό των μετρήσεων για κάθε σενάριο ομαδοποίησης, τα αποτελέσματα που συλλέγονται μετατρέπονται σε ένα DataFrame για ανάλυση. Αυτή η δομημένη μορφή διευκολύνει τον εντοπισμό των καλύτερων συνδυασμών ομαδοποίησης με βάση τις βαθμολογίες των μετρικών. Η ανάλυση περιλαμβάνει:

- Καλύτερος συνδυασμός ομαδοποίησης: Προσδιορισμός της μοναδικής και καλύτερης διαμόρφωσης ομαδοποίησης σε όλους τους αλγόριθμους και τον αριθμό συμπλεγμάτων, λαμβάνοντας υπόψη τα αποτελέσματα για την ισορροπία μεταξύ της βαθμολογίας Silhouette score, του δείκτη Calinski-Harabasz και του δείκτη Davies-Bouldin.
- Καλύτερος συνδυασμός ανά αλγόριθμο: Προσδιορισμός της βέλτιστης διαμόρφωσης ομαδοποίησης για κάθε αλγόριθμο, επιτρέποντας πληροφορίες για συγκεκριμένους αλγόριθμους σχετικά με την απόδοση ομαδοποίησης.
- Συνολικά καλύτερος συνδυασμός: Επισήμανση της πιο αποτελεσματικής ρύθμισης ομαδοποίησης σε όλα τα εξεταζόμενα σενάρια, παρέχοντας μια σαφή σύσταση για την καλύτερη προσέγγιση ομαδοποίησης με βάση τις αξιολογημένες μετρικές.

Best Clustering Combination				
Algorithm	**Num_Clusters**	**Silhouette_Score**	**Calinski_Harabasz**	**Davies_Bouldin**
Agglomerative	3	0.713011	47.873543	0.315284

Πίνακας 3 Αλγόριθμος με το καλύτερο συνδυασμό αποτελεσμάτων δεικτών

Best Combination Per Algorithm				
Algorithm	**Num_Clusters**	**Silhouette_Score**	**Calinski_Harabasz**	**Davies_Bouldin**
Agglomerative	3	0.713011	47.873543	0.315284
GMM	3	0.658412	49.239047	0.455575
KMeans	2	0.701693	24.696762	0.393375
Spectral	3	0.713011	47.873543	0.315284

Πίνακας 4 Καλύτερος συνδυασμός αποτελεσμάτων δεικτών ανά αλγόριθμο

Overall Best Combination				
Algorithm	**Num_Clusters**	**Silhouette_Score**	**Calinski_Harabasz**	**Davies_Bouldin**
Agglomerative	3	0.713011	47.873543	0.315284

Πίνακας 5 Συναολικά καλύτερος συνδυασμός αποτελεσμάτων δεικτών σε αλγόριθμο

6.12.4 Χειρισμός σφαλμάτων και ολοκληρωμένη καταγραφή

Ο ισχυρός χειρισμός σφαλμάτων διασφαλίζει την ανθεκτικότητα της διαδικασίας αξιολόγησης, με κατάλληλη καταγραφή για κάθε βήμα για την παροχή διαφάνειας και τη διευκόλυνση του εντοπισμού σφαλμάτων. Αυτή η προσοχή στη λεπτομέρεια εξασφαλίζει την αξιοπιστία των αποτελεσμάτων και την ομαλή εκτέλεση της ανάλυσης.

6.12.5 Συμπέρασμα

Η λεπτομερής διαδικασία αξιολόγησης και βελτιστοποίησης που ενσωματώνεται στη συνάρτηση “calc_results” και η επακόλουθη ανάλυση αντιπροσωπεύουν μια αυστηρή προσέγγιση στην ανάλυση ομαδοποίησης. Αξιοποιώντας βασικές μετρικές και μια συστηματική εξερεύνηση των διαμορφώσεων για την ομαδοποίηση, αυτή η μεθοδολογία παρέχει ανεκτίμητες πληροφορίες σχετικά με την εγγενή ομαδοποίηση των συνόλων δεδομένων με βάση τα μετα-χαρακτηριστικά. Τέτοιες πληροφορίες είναι ζωτικής σημασίας για τη λήψη αποφάσεων βάσει δεδομένων σε έργα μηχανικής μάθησης, ενισχύοντας την κατανόηση των χαρακτηριστικών των συνόλων δεδομένων και ενημερώνοντας την επιλογή αναλυτικών στρατηγικών. Αυτό το ολοκληρωμένο πλαίσιο για την αξιολόγηση και βελτιστοποίηση της ομαδοποίησης υπογραμμίζει τη σημασία μιας δομημένης και μετρικής προσέγγισης στην ομαδοποίηση, διευκολύνοντας την ανακάλυψη σημαντικών μοτίβων και ομαδοποιήσεων μέσα σε σύνθετα σύνολα δεδομένων.

7. Αξιοποίηση βέλτιστων διαμορφώσεων συμπλέγματος για διορατική οπτικοποίηση δεδομένων

7.1 Εισαγωγή

Η διαδικασία προσδιορισμού των βέλτιστων διαμορφώσεων ομαδοποίησης και οπτικοποίησης των συμπλεγμάτων που παράγονται, είναι ένα σημαντικό βήμα προς την κατανόηση της εγγενούς δομής εντός των συνόλων δεδομένων. Αυτή η ενότητα εμβαθύνει στη μεθοδολογία που χρησιμοποιήθηκε για να επιτευχθεί αυτό, αξιοποιώντας το λεξικό “best_configs” και τη συνάρτηση “plot_clusters_names”.

7.2 Προσδιορισμός βέλτιστων παραμέτρων ομαδοποίησης

Η προσέγγιση ξεκινά με τη συνάθροιση βέλτιστων παραμέτρων ομαδοποίησης για κάθε αλγόριθμο, με βάση προηγούμενες αναλύσεις που προσδιορίζουν τις διαμορφώσεις που μεγιστοποιούν την καλύτερη απόδοση σύμφωνα με επιλεγμένες μετρικές. Αυτές οι πληροφορίες

αποθηκεύονται συστηματικά στο λεξικό “best_configs”, το οποίο αντιστοιχίζει κάθε αλγόριθμο στο βέλτιστο σύνολο παραμέτρων, όπως ο αριθμός των συστάδων.

Εξαγωγή παραμέτρων: Για κάθε αλγόριθμο, ο βέλτιστος αριθμός συμπλεγμάτων (n_clusters) εξάγεται από το best_per_algorithm DataFrame, το οποίο είναι μια περίληψη των καλύτερων διαμορφώσεων ανά αλγόριθμο.

Ενοποίηση παραμέτρων: Οι εξαγόμενες παράμετροι ενοποιούνται στο λεξικό best_configs, διευκολύνοντας την εύκολη πρόσβαση και εφαρμογή σε επόμενα βήματα ομαδοποίησης και οπτικοποίησης.

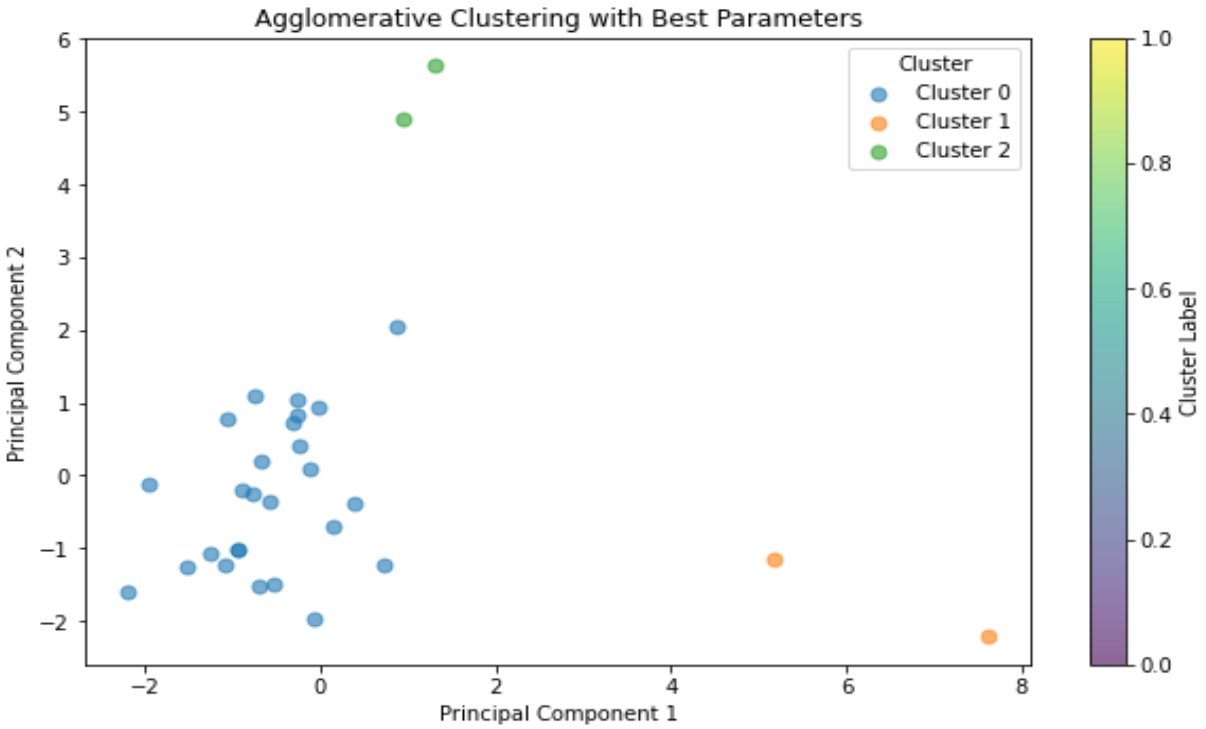
Updated best_configs for Agglomerative: {'n_clusters': 3}
Updated best_configs for GMM: {'n_clusters': 3}
Updated best_configs for KMeans: {'n_clusters': 2}
Updated best_configs for Spectral: {'n_clusters': 3}

7.3 Οπτικοποίηση των συμπλεγμάτων

Η συνάρτηση “plot_clusters_names” διαδραματίζει κεντρικό ρόλο στην οπτικοποίηση του τρόπου ομαδοποίησης των συνόλων δεδομένων κάτω από τη βέλτιστη διαμόρφωση κάθε αλγορίθμου. Αυτή η οπτικοποίηση δεν είναι μόνο ζωτικής σημασίας για την ποιοτική αξιολόγηση, αλλά και για την ανταλλαγή πληροφοριών με ένα ευρύτερο κοινό.

Οπτικοποίηση συμπλέγματος: Κάθε σύνολο δεδομένων σχεδιάζεται σύμφωνα με τα χαρακτηριστικά μείωσης PCA, με διαφορετικά χρώματα που αντιπροσωπεύουν ξεχωριστά συμπλέγματα. Αυτή η οπτική αναπαράσταση επιτρέπει μια διαισθητική κατανόηση του τρόπου ομαδοποίησης των συνόλων δεδομένων με βάση τα μετα-χαρακτηριστικά τους.

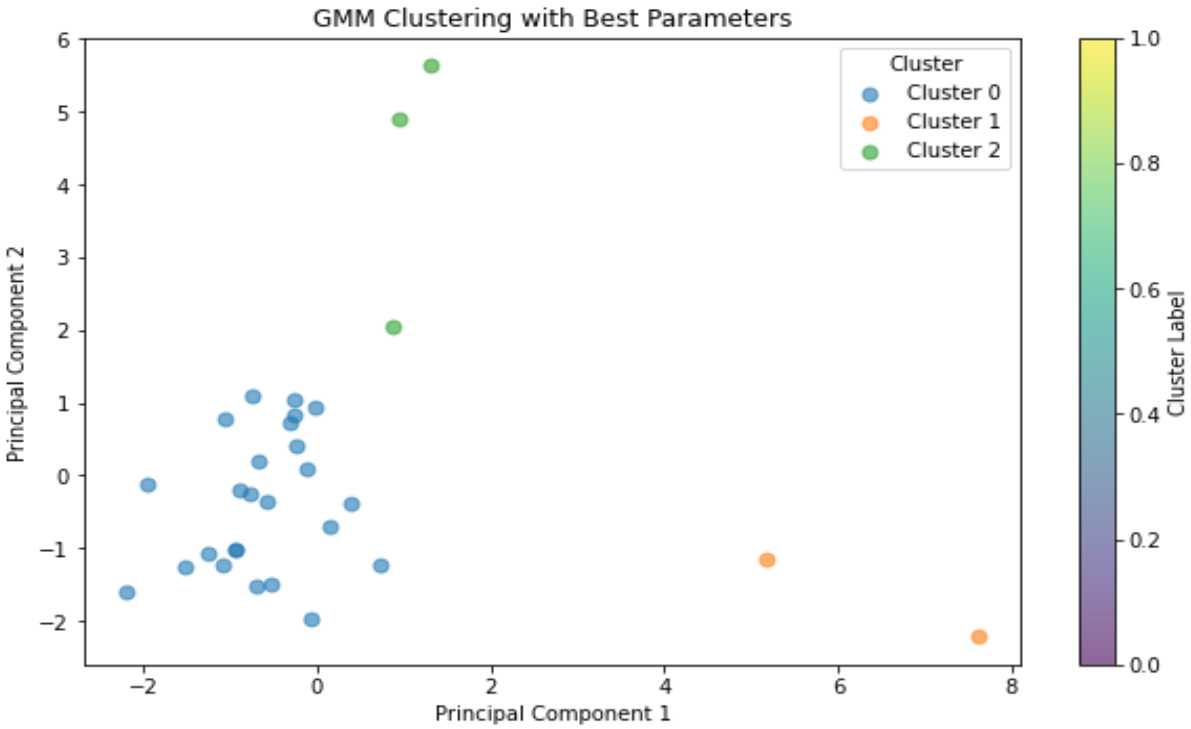
Επισήμανση συμπλέγματος και συνόλου δεδομένων: Κάθε σύμπλεγμα σχολιάζεται με μια ετικέτα και παρέχεται ένα υπόμνημα για λόγους σαφήνειας. Επιπλέον, με την χρήση της συνάρτησης ομαδοποιούνται τα δεδομένα ονομαστικά, βάσει του συμπλέγματος στο οποίο έχουν κατανεμηθεί, προσφέροντας μια σαφή αναπαράσταση του αποτελέσματος ομαδοποίησης.



Εικόνα 2 Agglomerative CLustering με τις καλύτερες παραμέτρους

Dataset names per cluster for Agglomerative	
Cluster 0	pima, planning_relax, pm10, pollution, prnn_fglass, quake, rabe_266, residential_building, seeds, sonar, stock, synthetic_control, tecator, urban_land_cover, vehicle, vertebral_column_2classes, vertebral_column_3classes, vinnie, visualizing_environmental, visualizing_galaxy, volcanoes_a2, wifi_localization, wine, winequality-red, winequality-white, yeast
Cluster 1	robot_failures_lp4, robot_failures_lp5
Cluster 2	triazines_cl, triazines_reg

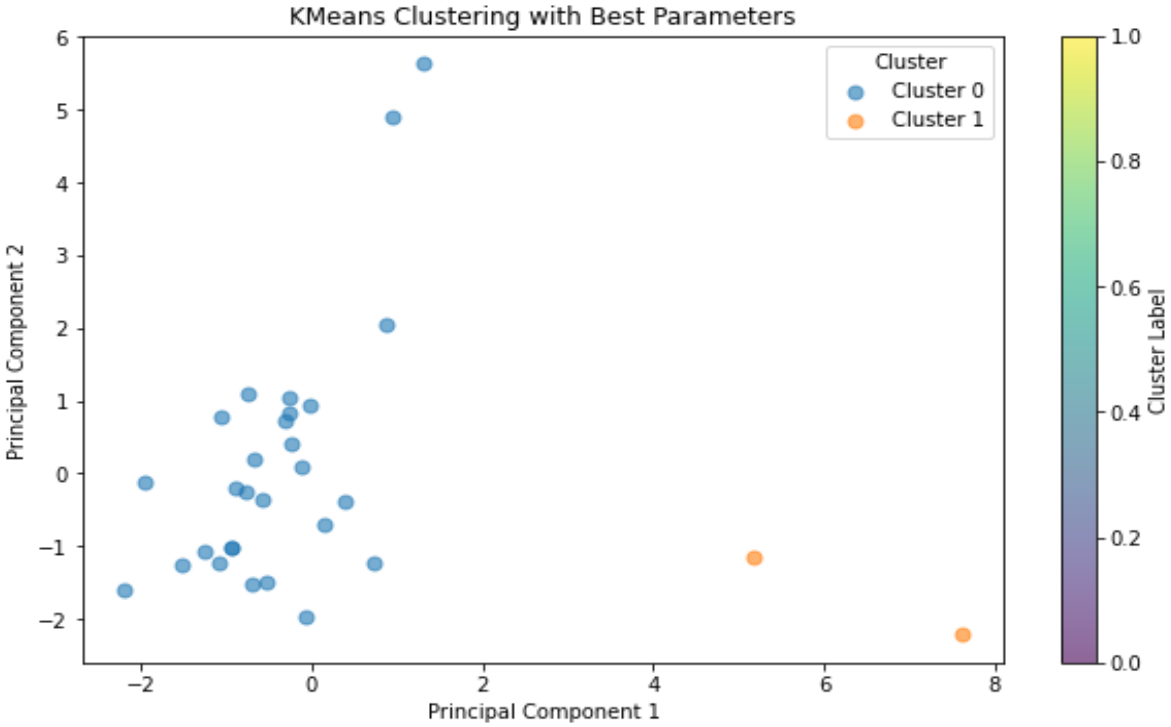
Πίνακας 6 Ονόματα datasets ανά cluster με Agglomerative



Εικόνα 3 GMM CLustering με τις καλύτερες παραμέτρους

Dataset names per cluster for GMM	
Cluster 0	pima, planning_relax, pm10, pollution, prnn_fglass, quake, rabe_266, residential_building, seeds, sonar, stock, synthetic_control, tecator, urban_land_cover, vehicle, vertebral_column_2classes, vertebral_column_3classes, vinnie, visualizing_environmental, visualizing_galaxy, volcanoes_a2, wifi_localization, wine, winequality-red, winequality-white
Cluster 1	robot_failures_lp4, robot_failures_lp5
Cluster 2	triazines_cl, triazines_reg, yeast

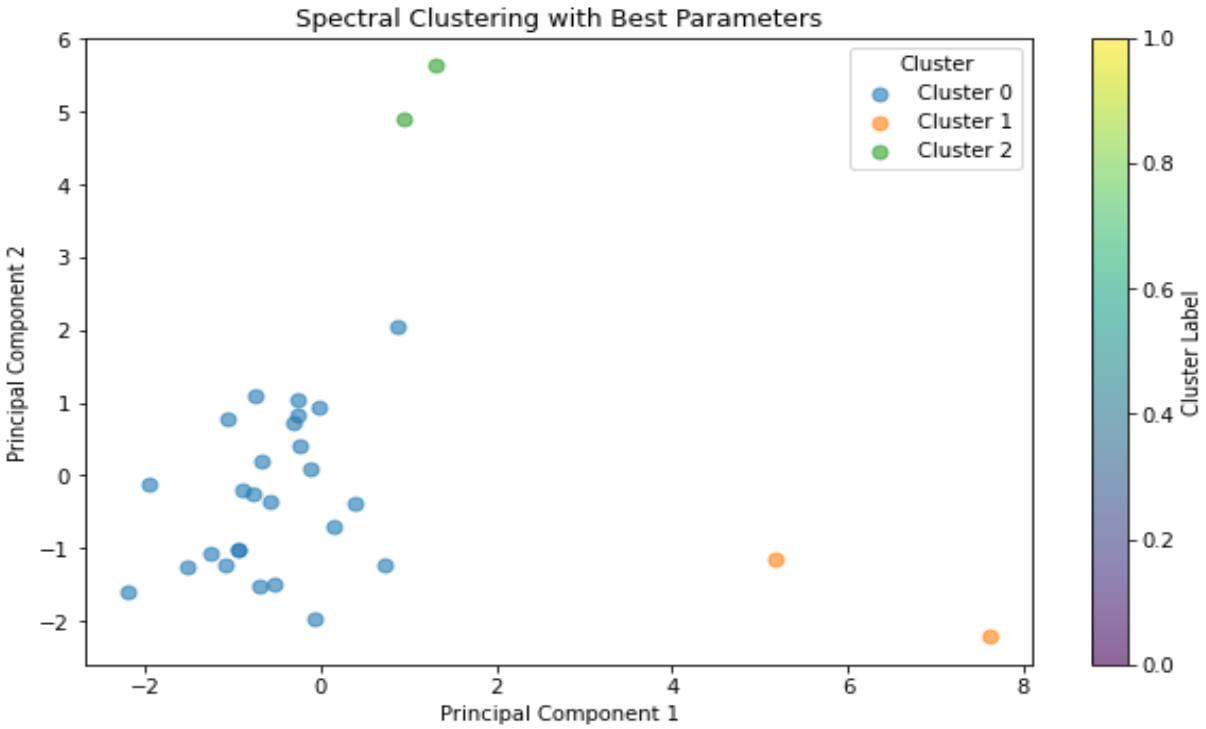
Πίνακας 7 Ονόματα datasets ανά cluster με GMM



Εικόνα 4 KMeans CLustering με τις καλύτερες παραμέτρους

Dataset names per cluster for KMeans	
Cluster 0	pima, planning_relax, pm10, pollution, prnn_fglass, quake, rabe_266, residential_building, seeds, sonar, stock, synthetic_control, tecator, triazines_cl, triazines_reg, urban_land_cover, vehicle, vertebral_column_2classes, vertebral_column_3classes, vinnie, visualizing_environmental, visualizing_galaxy, volcanoes_a2, wifi_localization, wine, winequality-red, winequality-white, yeast
Cluster 1	robot_failures_lp4, robot_failures_lp5

Πίνακας 8 Ονόματα datasets ανά cluster με KMeans



Εικόνα 5 Spectral CLustering με τις καλύτερες παραμέτρους

Dataset names per cluster for Spectral	
Cluster 0	pima, planning_relax, pm10, pollution, prnn_fglass, quake, rabe_266, residential_building, seeds, sonar, stock, synthetic_control, tecator, urban_land_cover, vehicle, vertebral_column_2classes, vertebral_column_3classes, vinnie, visualizing_environmental, visualizing_galaxy, volcanoes_a2, wifi_localization, wine, winequality-red, winequality-white, yeast
Cluster 1	robot_failures_lp4, robot_failures_lp5
Cluster 2	triazines_cl, triazines_reg

Πίνακας 9 Ονόματα datasets ανά cluster με Spectral

7.4 Ομαδοποίηση και ερμηνεία συνόλου δεδομένων

Το τελευταίο βήμα περιλαμβάνει την ομαδοποίηση του συνόλου των δεδομένων βάσει των ονομάτων τους στις αναθέσεις συμπλέγματός τους για την παρουσίαση αυτών των πληροφοριών με οργανωμένο τρόπο. Αυτό το βήμα γεφυρώνει το χάσμα μεταξύ των αριθμητικών εκχωρήσεων συμπλέγματος και των πραγματικών συνόλων δεδομένων, επιτρέποντας μια πιο ερμηνευτική ανάλυση των αποτελεσμάτων ομαδοποίησης.

- Cluster-Dataset Mapping: Κάθε σύνολο δεδομένων αντιστοιχίζεται στο αντίστοιχο σύμπλεγμα, διευκολύνοντας την κατανόηση των συνόλων δεδομένων που μοιράζονται παρόμοια χαρακτηριστικά σύμφωνα με τον αλγόριθμο ομαδοποίησης και τις επιλεγμένες παραμέτρους.
- Δημιουργία πληροφοριών: Εξετάζοντας ποια σύνολα δεδομένων ομαδοποιούνται, οι ερευνητές και οι αναλυτές μπορούν να δημιουργήσουν πληροφορίες σχετικά με τις

ομοιότητες και τις διαφορές μεταξύ των συνόλων δεδομένων, ενδεχομένως καθοδηγώντας περαιτέρω ανάλυση ή δημιουργία υποθέσεων.

7.5 Συμπέρασμα

Ο μεθοδικός προσδιορισμός των βέλτιστων διαμορφώσεων ομαδοποίησης και η επακόλουθη οπτικοποίηση και ανάλυση των συμπλεγμάτων παρέχουν ένα ισχυρό πλαίσιο για την κατανόηση της υποκείμενης δομής εντός των συνόλων δεδομένων. Αυτή η διαδικασία όχι μόνο βοηθά στην επικύρωση των αποτελεσμάτων ομαδοποίησης, αλλά χρησιμεύει και ως ένα ισχυρό εργαλείο για τη διερευνητική ανάλυση δεδομένων, προσφέροντας ένα γραφικό και ερμηνευτικό φακό μέσω του οποίου μπορούμε δούμε πολύπλοκες σχέσεις των συνόλων δεδομένων.

8.Hyperparameter Grid Search: Βελτίωση της αποτελεσματικότητας και της ακρίβειας των ομαδοποιήσεων

8.1 Εισαγωγή

Η μεθοδολογία για την ενίσχυση της αποτελεσματικότητας και της ακρίβειας ομαδοποίησης μέσω της αναζήτησης πλέγματος (grid search) υπερπαραμέτρων περιλαμβάνει μια σχολαστική εξερεύνηση του χώρου παραμέτρων για κάθε επιλεγμένο αλγόριθμο ομαδοποίησης. Αυτή η ενότητα παρέχει μια περιεκτική ανάλυση αυτής της διαδικασίας, η οποία είναι ζωτικής σημασίας για τη βελτιστοποίηση των αποτελεσμάτων ομαδοποίησης και τη διασφάλιση ισχυρής τμηματοποίησης δεδομένων.

8.2 Ορισμός πλέγματος υπερπαραμέτρων

Κάθε αλγόριθμος ομαδοποίησης διαθέτει μοναδικές παραμέτρους που επηρεάζουν σημαντικά την απόδοσή του και τα αποτελέσματα ομαδοποίησης. Το λεξικό param_grids ενσωματώνει αυτές τις παραμέτρους για τέσσερις βασικούς αλγορίθμους: K-Means, Agglomerative Clustering, Spectral Clustering και Gaussian Mix.

- Παράμετροι K-Means: Ο αριθμός των συστάδων (n_clusters), η μέθοδος αρχικοποίησης (init), ο αριθμός αρχικοποιήσεων (n_init) και η ανοχή για σύγκλιση (tol) ποικίλλουν για να διερευνηθεί η συμπεριφορά του αλγορίθμου υπό διαφορετικές συνθήκες.
- Παράμετροι Agglomerative Clustering: Παράμετροι, όπως ο αριθμός των συστάδων (n_clusters) και τα κριτήρια σύνδεσης (linkage) προσαρμόζονται για να εξεταστεί πώς διαφορετικές ιεραρχικές στρατηγικές ομαδοποίησης επηρεάζουν την τελική ομαδοποίηση.
- Παράμετροι Spectral Clustering: Η εξερεύνηση αυτού του αλγορίθμου περιλαμβάνει τον αριθμό των συστάδων (n_clusters), τη μέθοδο εκχώρησης ετικετών (assign_labels), τον

αριθμό αρχικοποιήσεων (n_init) και τον αριθμό των γειτόνων(neighbors) στο πλησιέστερο γράφημα (n_neighbors).
- Παράμετροι Gaussian Mixture: Η διακύμανση του αριθμού των συστατικών (n_components), του τύπου συνδιακύμανσης (covariance_type), της μεθόδου αρχικοποίησης (init_params) και του αριθμού αρχικοποιήσεων (n_init) βοηθά στην κατανόηση της επίδρασης των υποθέσεων κατανομής στην ομαδοποίηση.

8.3 Εκτέλεση Grid Search

Η συνάρτηση “perform_grid_search” διενεργείται μέσα στον καθορισμένο χώρο υπερπαραμέτρων για κάθε αλγόριθμο, χρησιμοποιώντας τη βαθμολογία σιλουέτας ως μέτρηση για την αξιολόγηση της απόδοσης ομαδοποίησης. Αυτή η βαθμολογία, μετρώντας τη μέση απόσταση εντός του συμπλέγματος σε σχέση με την πλησιέστερη απόσταση συμπλέγματος, παρέχει ένα σαφές κριτήριο για τη βελτιστοποίηση και την επιλογή των καλύτερων παραμέτρων αλγορίθμου.

- Επαναληπτικός έλεγχος παραμέτρων: Για κάθε συνδυασμό παραμέτρων που ορίζονται στο param_grids, ο αλγόριθμος εφαρμόζεται στο σύνολο δεδομένων μετα-χαρακτηριστικών με μειωμένη PCA (metafeatures_pca) και αξιολογείται η προκύπτουσα ομαδοποίηση.
- Βέλτιστη αναγνώριση παραμέτρων: Ο συνδυασμός που δίνει την υψηλότερη βαθμολογία σιλουέτας θεωρείται βέλτιστος για τον αλγόριθμο, ενθυλακώνοντας την πιο αποτελεσματική διαμόρφωση ομαδοποίησης των χαρακτηριστικών των δεδομένων.

Best configurations for each algorithm Grid Search		
Algorithm	**Best Params**	**Best Score**
K-Means	'init': 'k-means++', 'n_clusters': 2, 'n_init': 10, 'tol': 0.0001	0.7016926707639081
Agglomerative Clustering	'linkage': 'ward', 'n_clusters': 3	0.7130112952714299
Spectral Clustering	'assign_labels': 'kmeans', 'n_clusters': 3, 'n_init': 10, 'n_neighbors': 5	0.7130112952714299
Gaussian Mixture	'covariance_type': 'tied', 'init_params': 'kmeans', 'n_components': 3, 'n_init': 1	0.7130112952714299

Πίνακας 10 Καλύτεροι παράμετροι ανά αλγόριθμο με Grid Search

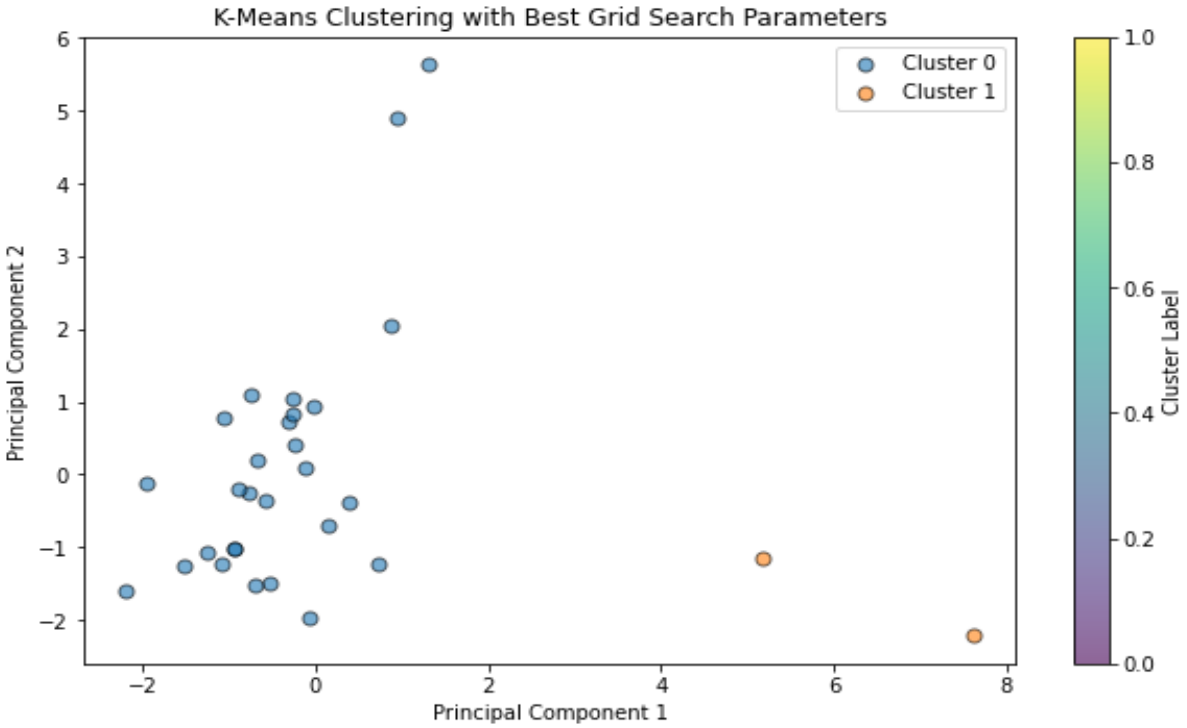
8.4 Οπτικοποίηση βέλτιστης ομαδοποίησης

Μετά τον προσδιορισμό των καλύτερων παραμέτρων στο προηγούμενο βήμα, η συνάρτηση “plot_best_configuration” χρησιμοποιείται για την οπτική αναπαράσταση του αποτελέσματος ομαδοποίησης υπό τις βέλτιστες ρυθμίσεις. Αυτό το βήμα είναι ζωτικής σημασίας για την επικύρωση της ομαδοποίησης και τη διασφάλιση ότι η αλγοριθμική ομαδοποίηση ευθυγραμμίζεται με την αναμενόμενη τμηματοποίηση δεδομένων.

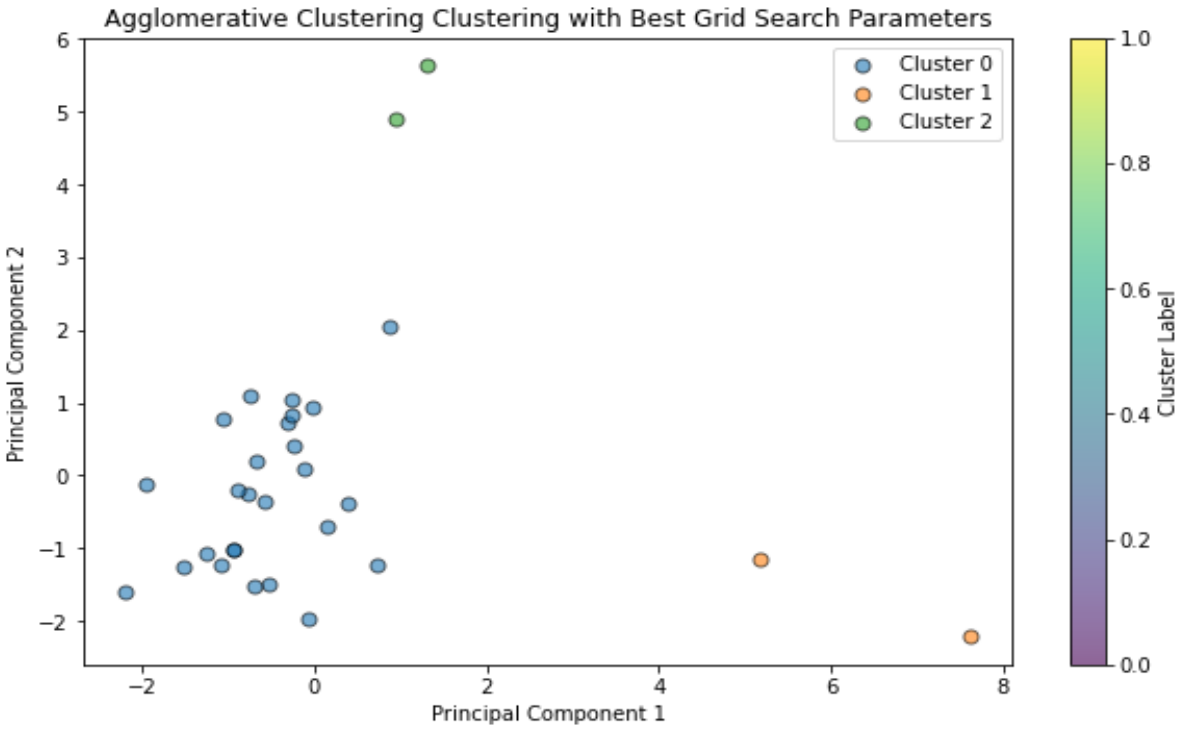
- Αναπαράσταση συμπλέγματος: Κάθε σημείο δεδομένων, που αντιπροσωπεύει ένα σύνολο δεδομένων στον χώρο μειωμένου PCA, σχεδιάζεται με ένα μοναδικό χρώμα που αντιστοιχεί

στην διαμόρφωση του συμπλέγματος, διευκολύνοντας την άμεση οπτική αξιολόγηση της ποιότητας και της συνοχής του συμπλέγματος.

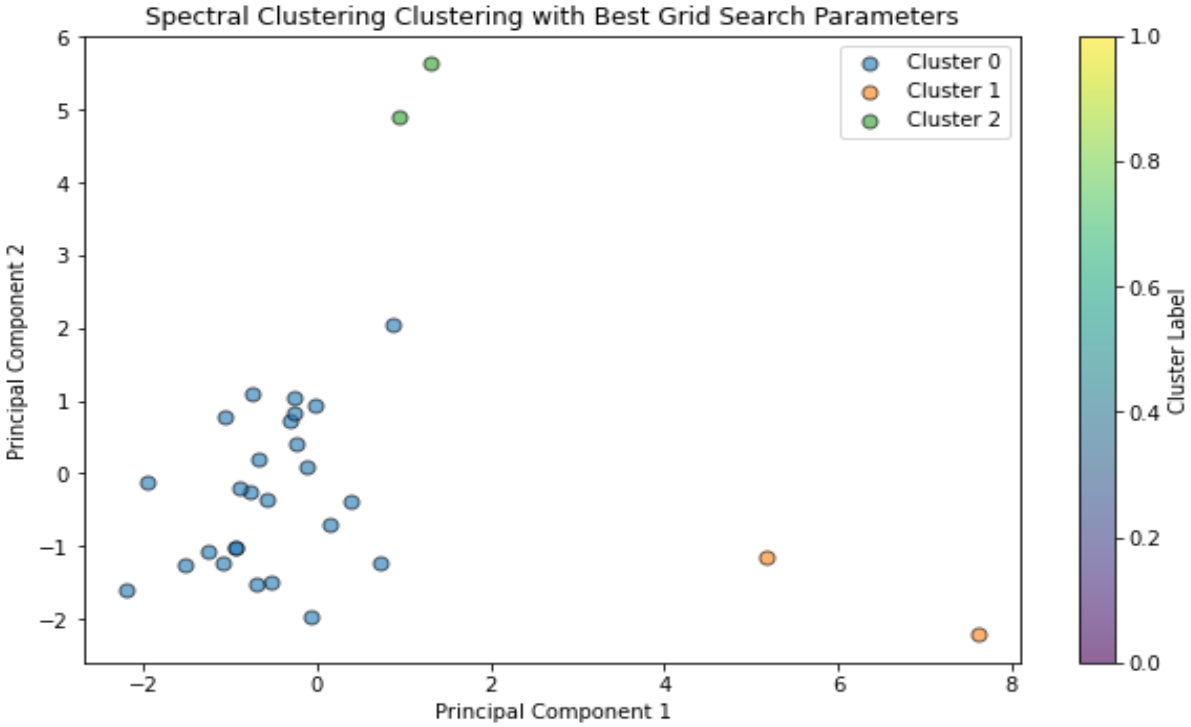
- Απεικονίσεις: Τα γραφήματα που δημιουργούνται παρέχουν μια γραφική ερμηνεία του τρόπου ομαδοποίησης των συνόλων δεδομένων κάτω από την καλύτερη διαμόρφωση κάθε αλγορίθμου, χρησιμεύοντας ως βάση για περαιτέρω αναλυτικές γνώσεις και συζητήσεις.



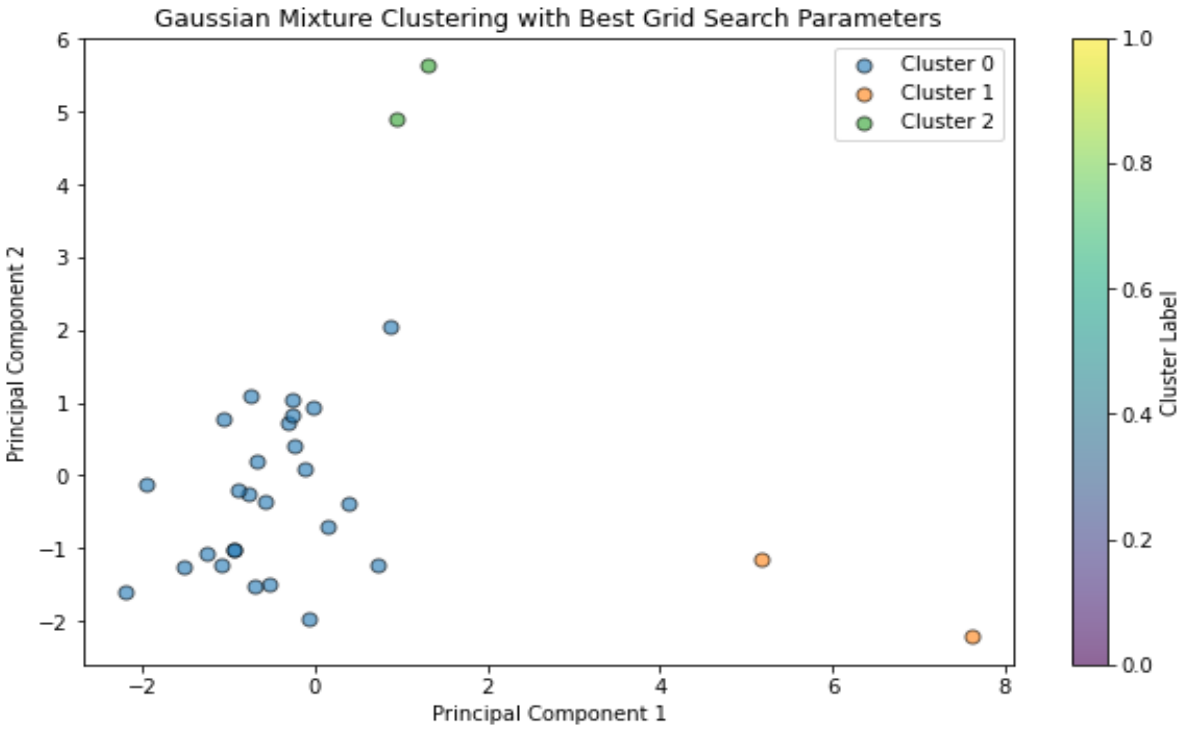
Εικόνα 6 KMeans Clustering με τις καλύτερες παραμέτρους σε Grid Search



Εικόνα 7 Agglomerative Clustering με τις καλύτερες παραμέτρους σε Grid Search



Εικόνα 8 Spectral Clustering με τις καλύτερες παραμέτρους σε Grid Search



Εικόνα 9 Gaussian Mixture Clustering με τις καλύτερες παραμέτρους σε Grid Search

8.5 Συμπέρασμα

Η μεθοδολογία grid search των υπερπαραμέτρων αποτελεί ακρογωνιαίο λίθο στην επιδίωξη εκλεπτυσμένων και ακριβών αποτελεσμάτων ομαδοποίησης. Με τη μεθοδική διερεύνηση του χώρου παραμέτρων και την αξιοποίηση της βαθμολογίας σιλουέτας για την αξιολόγηση της απόδοσης, αυτή η προσέγγιση εξασφαλίζει τον εντοπισμό βέλτιστων διαμορφώσεων ομαδοποίησης. Οι προκύπτουσες απεικονίσεις βοηθούν περαιτέρω στην ερμηνευτική κατανόηση της δομής των δεδομένων, προωθώντας μια βαθύτερη κατανόηση των υποκείμενων μοτίβων μέσα στα σύνολα δεδομένων. Η εκτέλεση αυτής της διαδικασίας συντονισμού υπερπαραμέτρων (hyperparameter tuning) όχι μόνο ενισχύει την απόδοση των μοντέλων ομαδοποίησης, αλλά συμβάλλει σημαντικά και στην ακρίβεια της ανάλυσης της μελέτης, διασφαλίζοντας ότι τα αποτελέσματα ομαδοποίησης είναι στατιστικά ισχυρά και διαισθητικά ερμηνεύσιμα.

9.Βελτίωση ομαδοποίησης μέσω συντονισμού υπερπαραμέτρων τυχαίας αναζήτησης (Random Search Hyperparameter Tuning)

9.1 Εισαγωγή

Ο συντονισμός υπερπαραμέτρων τυχαίας αναζήτησης αντιπροσωπεύει μια κεντρική προσέγγιση στη μηχανική μάθηση για τη βελτιστοποίηση της απόδοσης του αλγορίθμου. Σε αντίθεση με την αναζήτηση πλέγματος, η οποία διερευνά μεθοδικά έναν καθορισμένο χώρο παραμέτρων, η τυχαία αναζήτηση λαμβάνει δείγματα τιμών των παραμέτρων από μια καθορισμένη κατανομή, σε έναν σταθερό αριθμό επαναλήψεων. Αυτή η μέθοδος μπορεί συχνά να είναι πιο αποδοτική και αποτελεσματική, ειδικά όταν πρόκειται για χώρους υψηλών διαστάσεων ή όταν η βέλτιστη περιοχή παραμέτρων είναι μικρή.

9.2 Ορισμός κατανομών παραμέτρων

Για κάθε αλγόριθμο ομαδοποίησης - K-Means, Agglomerative Clustering, Spectral Clustering και Gaussian Mix - ορίζονται διακριτές κατανομές παραμέτρων:

- K-Means: Διερευνά τις διακυμάνσεις στον αριθμό των συμπλεγμάτων, τις μεθόδους αρχικοποίησης, τον αριθμό των αρχικοποιήσεων και την ανοχή για σύγκλιση.
- Agglomerative Clustering: Διερευνά διαφορετικούς αριθμούς συστάδων και κριτήρια σύνδεσης, δίνοντας έμφαση στην ιεραρχική φύση του αλγορίθμου.
- Spectral Clustering: Προσαρμόζει τον αριθμό των συμπλεγμάτων, τις στρατηγικές εκχώρησης ετικετών, τον αριθμό των αρχικοποιήσεων και τον αριθμό των γειτόνων για το πλησιέστερο γράφημα.
- Gaussian Mix: Μεταβάλλει τον αριθμό των συστατικών, τον τύπο συνδιακύμανσης, τον αριθμό των αρχικοποιήσεων και τη μέθοδο αρχικοποίησης για να κατανοήσει την επίδραση στις πιθανοτικές εκχωρήσεις συμπλεγμάτων.

9.3 Εκτέλεση τυχαίας αναζήτησης (Random Search)

Η συνάρτηση “random_search_clustering” έχει σχεδιαστεί για να επαναλαμβάνεται μέσω δειγμάτων παραμέτρων για δεδομένο αριθμό επαναλήψεων (n_iter). Κάθε συνδυασμός παραμέτρων εφαρμόζεται στο σύνολο δεδομένων και το αποτέλεσμα ομαδοποίησης αξιολογείται χρησιμοποιώντας τη βαθμολογία σιλουέτας, μια μετρική που αξιολογεί τη συνοχή και τον διαχωρισμό των συστάδων που προκύπτουν.

- Δειγματοληψία παραμέτρων: Λαμβάνονται δείγματα παραμέτρων από τις καθορισμένες κατανομές, επιτρέποντας μια ευρεία και ποικίλη εξερεύνηση του χώρου παραμέτρων.
- Αξιολόγηση απόδοσης: Η βαθμολογία σιλουέτας χρησιμεύει ως κριτήριο για την αξιολόγηση της ποιότητας της ομαδοποίησης, καθοδηγώντας την επιλογή του καλύτερου συνόλου παραμέτρων.

9.4 Προσδιορισμός βέλτιστων διαμορφώσεων

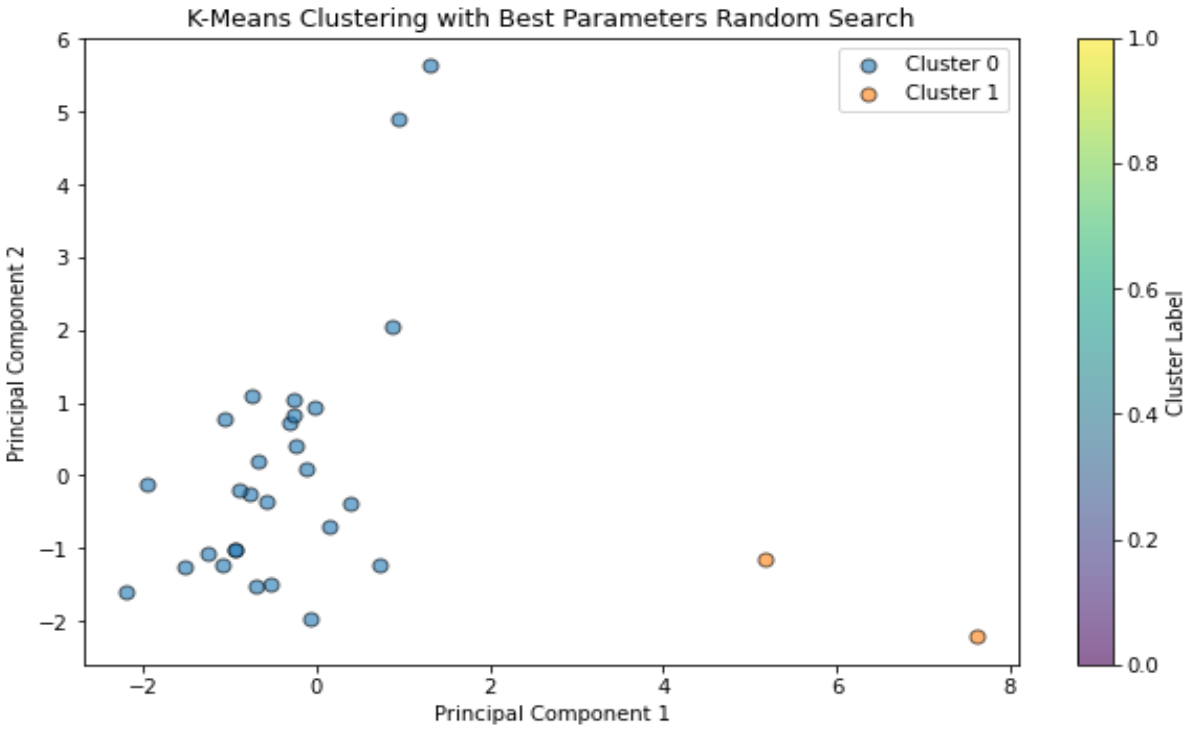
Μετά την τυχαία αναζήτηση, καθορίζονται οι καλύτερες διαμορφώσεις παραμέτρων για κάθε αλγόριθμο με βάση τις υψηλότερες βαθμολογίες σιλουέτας που λαμβάνονται. Αυτές οι βέλτιστες διαμορφώσεις αντιπροσωπεύουν τις πιο αποτελεσματικές ρυθμίσεις για κάθε αλγόριθμο ομαδοποίησης, δεδομένων των χαρακτηριστικών των δεδομένων, διασφαλίζοντας βελτιωμένη ποιότητα ομαδοποίησης και ερμηνευσιμότητα.

Best configurations for each algorithm random search		
Algorithm	**Best Params**	**Best Score**
K-Means	'tol': 0.0001, 'n_init': 20, 'n_clusters': 2, 'init': 'random'	0.7016926707639081
Agglomerative Clustering	'linkage': 'ward', 'n_clusters': 3	0.7130112952714299
Spectral Clustering	'n_neighbors': 15, 'n_init': 10, 'n_clusters': 3, 'assign_labels': 'discretize'	0.7130112952714299
Gaussian Mixture	'n_init': 2, 'n_components': 3, 'init_params': 'kmeans', 'covariance_type': 'tied'	0.7130112952714299

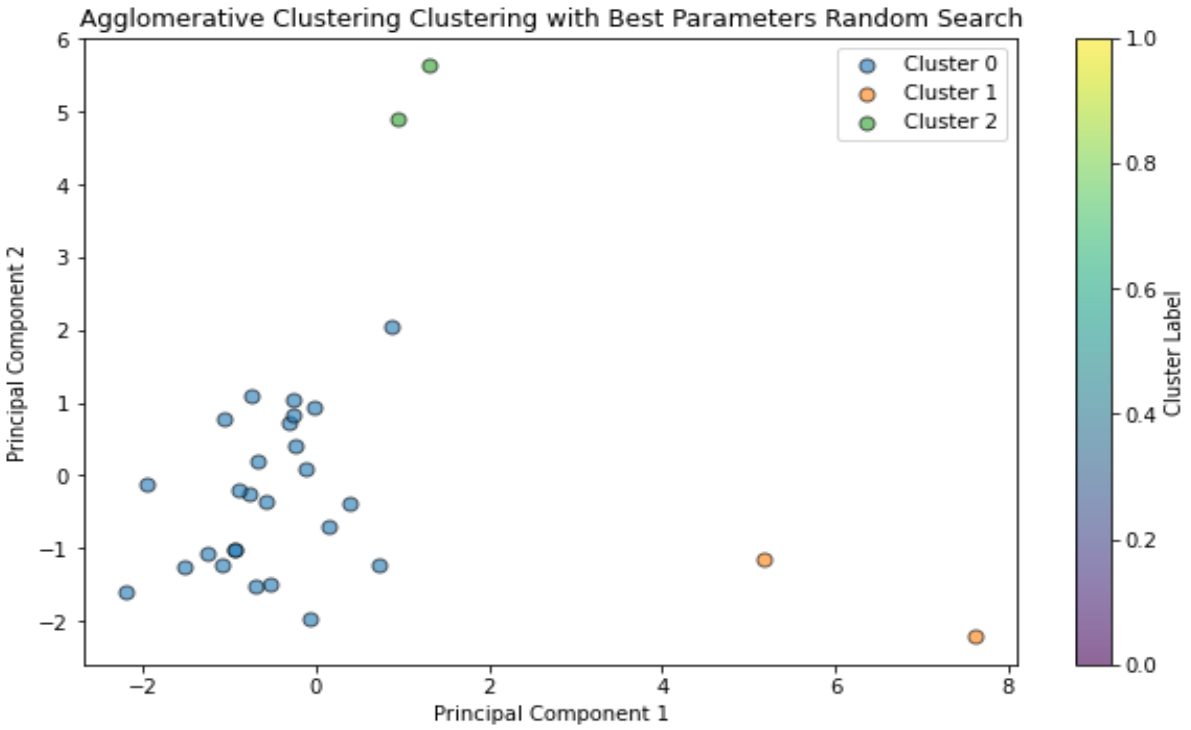
Πίνακας 11 Καλύτεροι παράμεροι ανά αλγόριθμο με Random Search

9.5 Οπτικοποίηση Αποτελεσμάτων Ομαδοποίησης

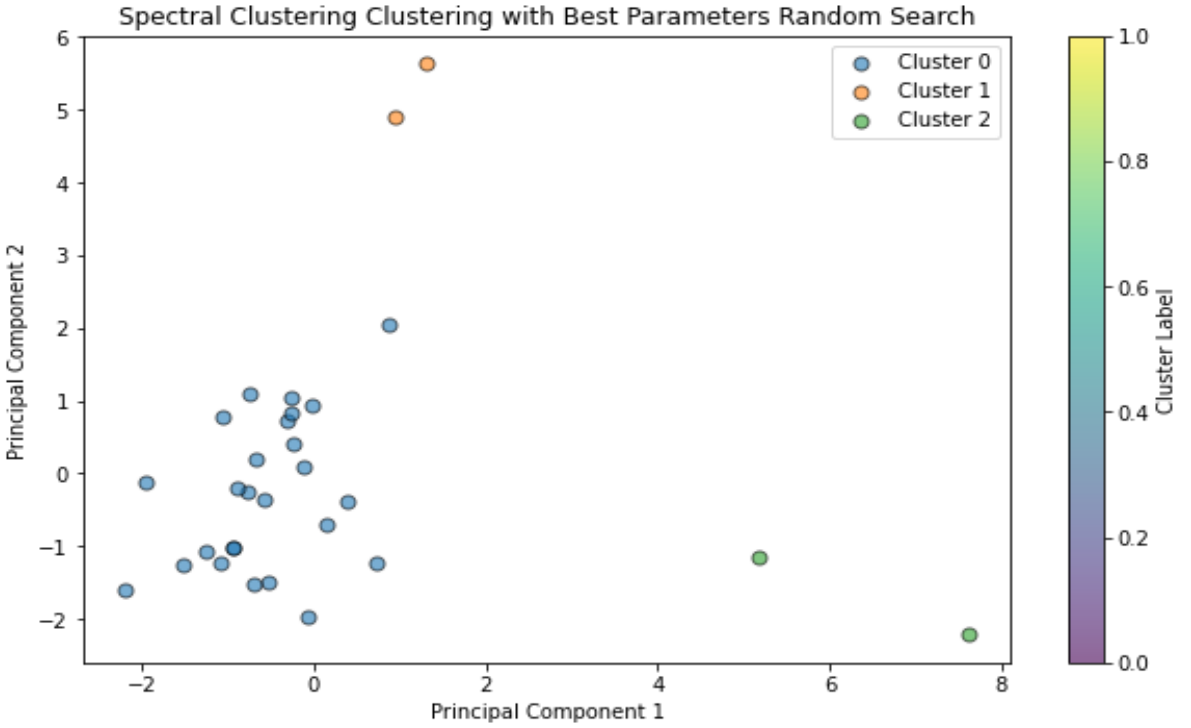
Η συνάρτηση “plot_clusters” απεικονίζει τα σημεία δεδομένων στον μειωμένο χώρο PCA, χρωματισμένα σύμφωνα με τις αντιστοιχίσεις του συμπλέγματος στις βέλτιστες ρυθμίσεις παραμέτρων. Αυτή η οπτικοποίηση όχι μόνο επιβεβαιώνει την αποτελεσματικότητα των επιλεγμένων παραμέτρων, αλλά παρέχει και πληροφορίες σχετικά με τη δομή των δεδομένων, διευκολύνοντας μια βαθύτερη κατανόηση των υποκείμενων μοτίβων.



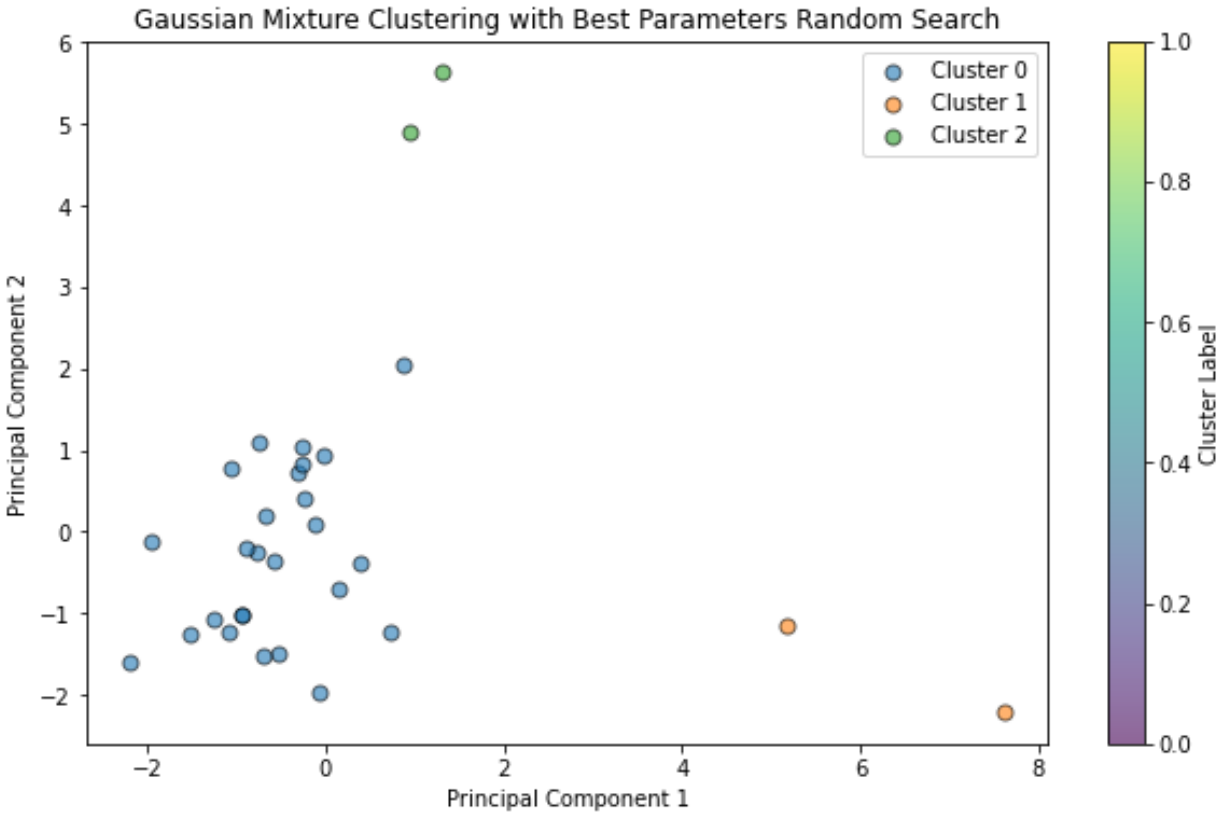
Εικόνα 10 KMeans Clustering με τις καλύτερες παραμέτρους σε Random Search



Εικόνα 11 Agglomerative Clustering με τις καλύτερες παραμέτρους σε Random Search



Εικόνα 12 Spectral Clustering με τις καλύτερες παραμέτρους σε Random Search



Εικόνα 13 Gaussian Mixture Clustering με τις καλύτερες παραμέτρους σε Random Search

9.6 Συμπέρασμα

Ο συντονισμός υπερπαραμέτρων τυχαίας αναζήτησης αποτελεί μια ισχυρή μέθοδο για τη βελτίωση της απόδοσης των αλγορίθμων ομαδοποίησης. Εξερευνώντας αποτελεσματικά έναν ευρύ χώρο παραμέτρων και εστιάζοντας σε διαμορφώσεις που αποδίδουν τις υψηλότερες βαθμολογίες σιλουέτας, αυτή η προσέγγιση εξασφαλίζει την επιλογή βέλτιστων ρυθμίσεων για κάθε αλγόριθμο. Οι επακόλουθες απεικονίσεις των αποτελεσμάτων ομαδοποίησης υπό αυτές τις βέλτιστες συνθήκες επικυρώνουν περαιτέρω τη διαδικασία συντονισμού, προσφέροντας μια γραφική αναπαράσταση της εγγενούς δομής των δεδομένων. Η χρήση της τυχαίας αναζήτησης σε αυτό το πλαίσιο όχι μόνο εξορθολογίζει τη διαδικασία συντονισμού υπερπαραμέτρων, αλλά συμβάλλει σημαντικά και στη μεθοδολογική ακρίβεια της μελέτης, διασφαλίζοντας ότι τα αποτελέσματα ομαδοποίησης είναι στατιστικά βελτιστοποιημένα και ουσιαστικά ερμηνεύσιμα.

Οι διαφορές στις καλύτερες διαμορφώσεις και βαθμολογίες μεταξύ τυχαίας αναζήτησης και αναζήτησης πλέγματος μπορούν να αποδοθούν στις ξεχωριστές στρατηγικές αναζήτησης. Η τυχαία αναζήτηση διερευνά τον χώρο παραμέτρων ευρύτερα και μερικές φορές μπορεί να βρει καλύτερες ή ισοδύναμες διαμορφώσεις πιο αποτελεσματικά από την αναζήτηση πλέγματος, η οποία αξιολογεί συστηματικά όλους τους πιθανούς συνδυασμούς εντός του καθορισμένου πλέγματος παραμέτρων. Η ομοιότητα στις καλύτερες βαθμολογίες δείχνει ότι και οι δύο μέθοδοι μπορούν να βελτιστοποιήσουν αποτελεσματικά τις υπερπαραμέτρους, αλλά οι συγκεκριμένες καλύτερες παράμετροι ενδέχεται να διαφέρουν λόγω της εγγενούς τυχαιότητας στη μέθοδο τυχαίας αναζήτησης και της εξαντλητικής φύσης της αναζήτησης πλέγματος. Οι διαφορές στις καλύτερες διαμορφώσεις από την τυχαία αναζήτηση και την αναζήτηση πλέγματος μπορεί να υποδηλώνουν τη διερευνητική φύση της τυχαίας αναζήτησης, η οποία μερικές φορές μπορεί να σκοντάψει τυχαία σε αποτελεσματικές διαμορφώσεις,

σε αντίθεση με τη συστηματική προσέγγιση της αναζήτησης πλέγματος. Οι παραλλαγές στα «n_init», «init_params» και «n_clusters»/«n_components» υπογραμμίζουν τις μοναδικές λύσεις που μπορεί να βρει κάθε μέθοδος στον ίδιο αλγόριθμο. Οι συνεπείς καλύτερες βαθμολογίες σε όλες τις μεθόδους υποδηλώνουν ότι και οι δύο είναι σε θέση να εντοπίσουν μοντέλα υψηλής απόδοσης, αλλά η διαδρομή που ακολουθούν για να φτάσουν εκεί μπορεί να διαφέρει σημαντικά.

10. Μέθοδος αγκώνα για βέλτιστη επιλογή συμπλέγματος και οπτικοποίηση μετά το Hyperparameter tuning

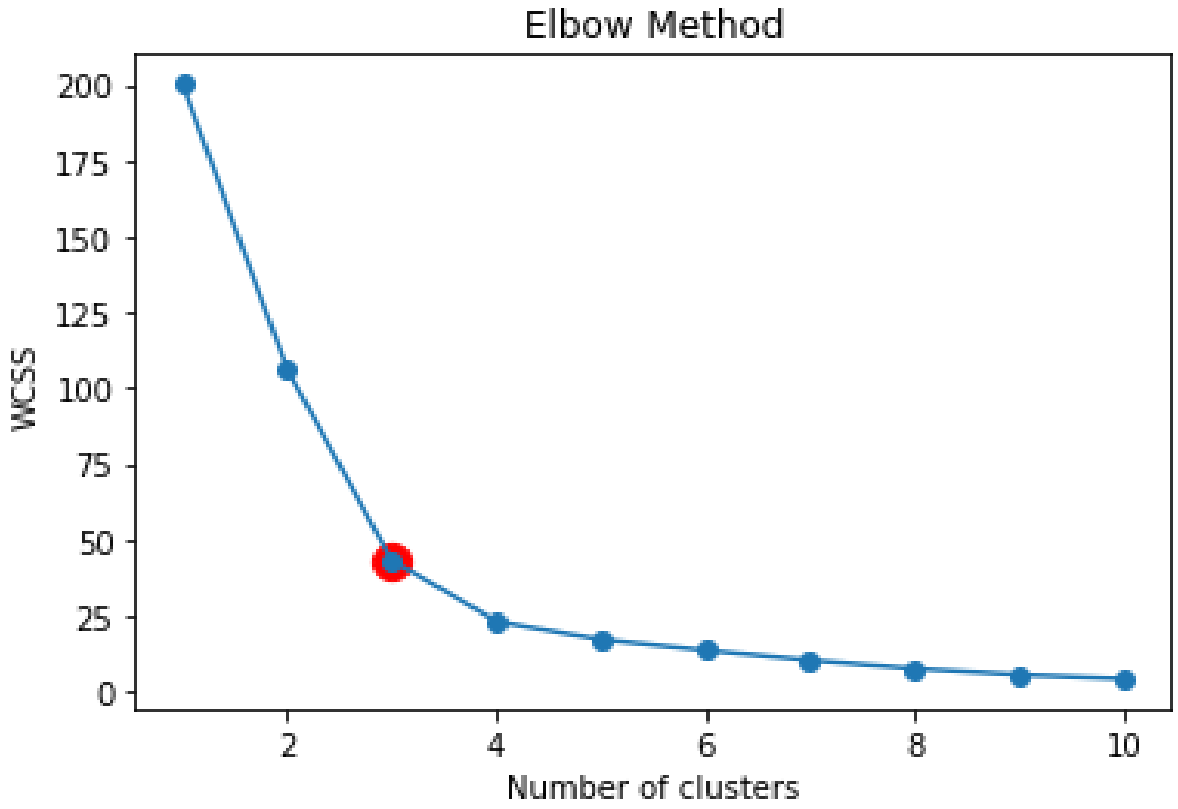
10.1 Εισαγωγή

Η μέθοδος αγκώνα αποτελεί θεμελιώδη τεχνική στον τομέα της ανάλυσης συστάδων, που χρησιμοποιείται κυρίως για τον προσδιορισμό του βέλτιστου αριθμού συστάδων. Σχεδιάζοντας το άθροισμα των τετραγώνων εντός του συμπλέγματος (WCSS) έναντι του αριθμού των συστάδων, αυτή η μέθοδος διευκολύνει τον προσδιορισμό ενός σημείου "κάμψης" ή "αγκώνα", πέρα από το οποίο η μείωση του WCSS γίνεται οριακή. Αυτό το σημείο θεωρείται συνήθως ως ο ιδανικός αριθμός συμπλεγμάτων, εξισορροπώντας μεταξύ της ελαχιστοποίησης του WCSS και της αποφυγής υπερβολικής τμηματοποίησης.

10.2 Εφαρμογή της μεθόδου αγκώνα

Η συνάρτηση “calculate_wcss” υπολογίζει το WCSS για μια περιοχή αριθμών συμπλέγματος, παρέχοντας μια ολοκληρωμένη εικόνα του τρόπου με τον οποίο μεταβάλλονται ανάλογα με τον αριθμό των συμπλεγμάτων. Αρχικοποιώντας τα K-Means με παραμέτρους, όπως 'k-means++' για κεντροειδή αρχικοποίηση και καθορισμένες επαναλήψεις, αυτή η λειτουργία εξασφαλίζει ισχυρή και αποτελεσματική ομαδοποίηση.

Μετά τον υπολογισμό, η συνάρτηση “plot_elbow” οπτικοποιεί τη σχέση μεταξύ του αριθμού των συμπλεγμάτων και του WCSS, επιτρέποντας μια οπτική επιθεώρηση για το σημείο αγκώνα. Αυτή η γραφική αναπαράσταση όχι μόνο βοηθά στη διαισθητική επιλογή του βέλτιστου αριθμού συμπλέγματος, αλλά ενισχύει και την ερμηνευσιμότητα της προσέγγισης ομαδοποίησης.



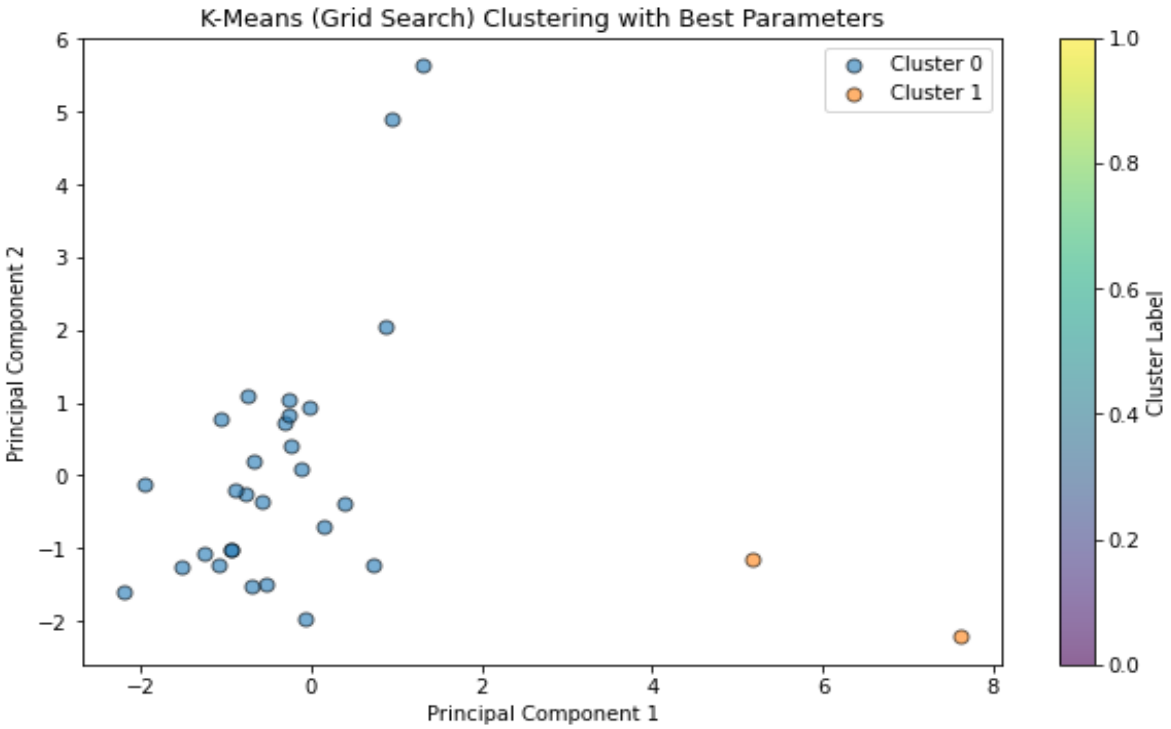
Εικόνα 14 Μέθοδος του αγκώνα για αριθμό cluster

10.3 Πρακτική εφαρμογή και ιδέες

Οι υπολογισμένες τιμές WCSS και η επακόλουθη γραφική παράσταση αγκώνα προσφέρουν κρίσιμες πληροφορίες σχετικά με την εγγενή ομαδοποίηση εντός του συνόλου δεδομένων. Η αναγνώριση του σημείου αγκώνα χρησιμεύει ως κατευθυντήρια γραμμή για την επιλογή ενός δικαιολογημένου αριθμού συστάδων, διασφαλίζοντας έτσι ότι το μοντέλο ομαδοποίησης συλλαμβάνει σημαντικές δομές χωρίς να περιπλέκει υπερβολικά το μοντέλο.

10.4 Οπτικοποίηση συμπλέγματος μετά το Hyperparameter Tuning

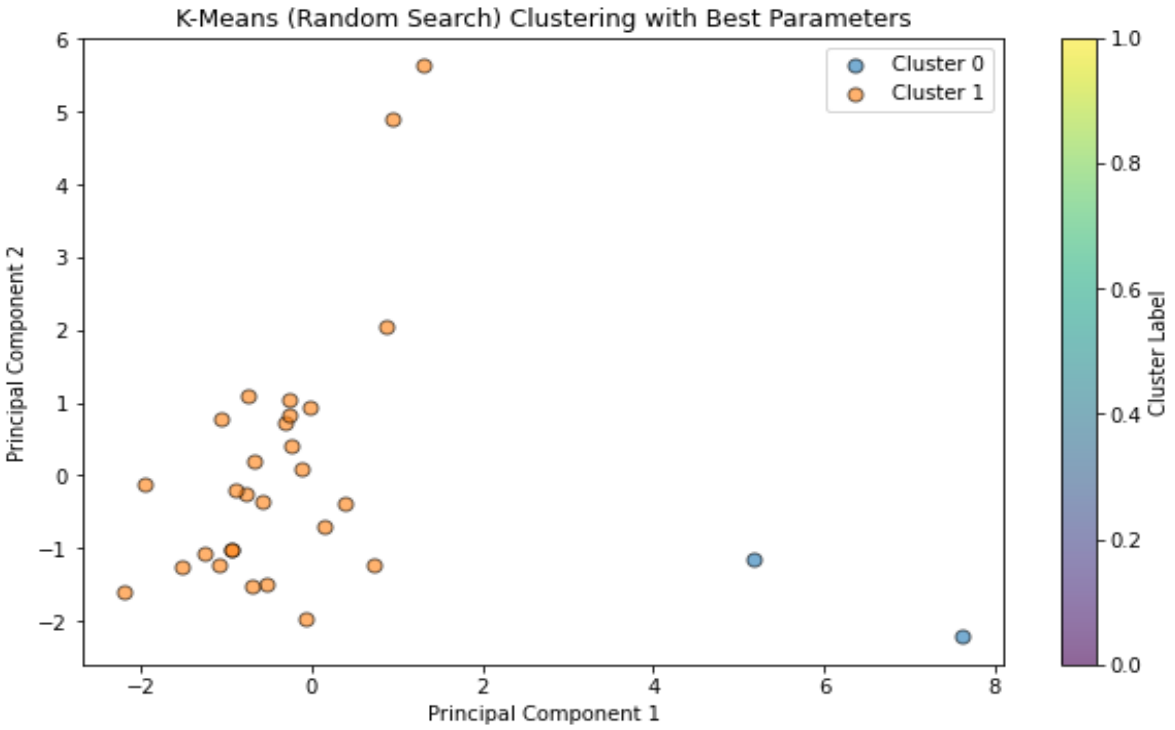
Μετά τον προσδιορισμό των βέλτιστων υπερπαραμέτρων μέσω μεθόδων, όπως το πλέγμα και η τυχαία αναζήτηση, η συνάρτηση “plot_clusters_and_print_names” παρέχει μια οπτική και αναλυτική αναπαράσταση των αποτελεσμάτων ομαδοποίησης. Με τη χαρτογράφηση των σημείων των δεδομένων στις αντίστοιχες ομάδες τους σε ένα μειωμένο διαστατικό χώρο (συνήθως μέσω PCA), αυτή η απεικόνιση υπογραμμίζει την αποτελεσματικότητα των επιλεγμένων παραμέτρων και τη φυσική κατάτμηση εντός των δεδομένων.



Εικόνα 15 KMeans Clustering με τις καλύτερες παραμέτρους σε Grid Search

Dataset names per cluster for K-Means (Grid Search)	
Cluster 0	pima, planning_relax, pm10, pollution, prnn_fglass, quake, rabe_266, residential_building, seeds, sonar, stock, synthetic_control, tecator, triazines_cl, triazines_reg, urban_land_cover, vehicle, vertebral_column_2classes, vertebral_column_3classes, vinnie, visualizing_environmental, visualizing_galaxy, volcanoes_a2, wifi_localization, wine, winequality-red, winequality-white, yeast
Cluster 1	robot_failures_lp4, robot_failures_lp5

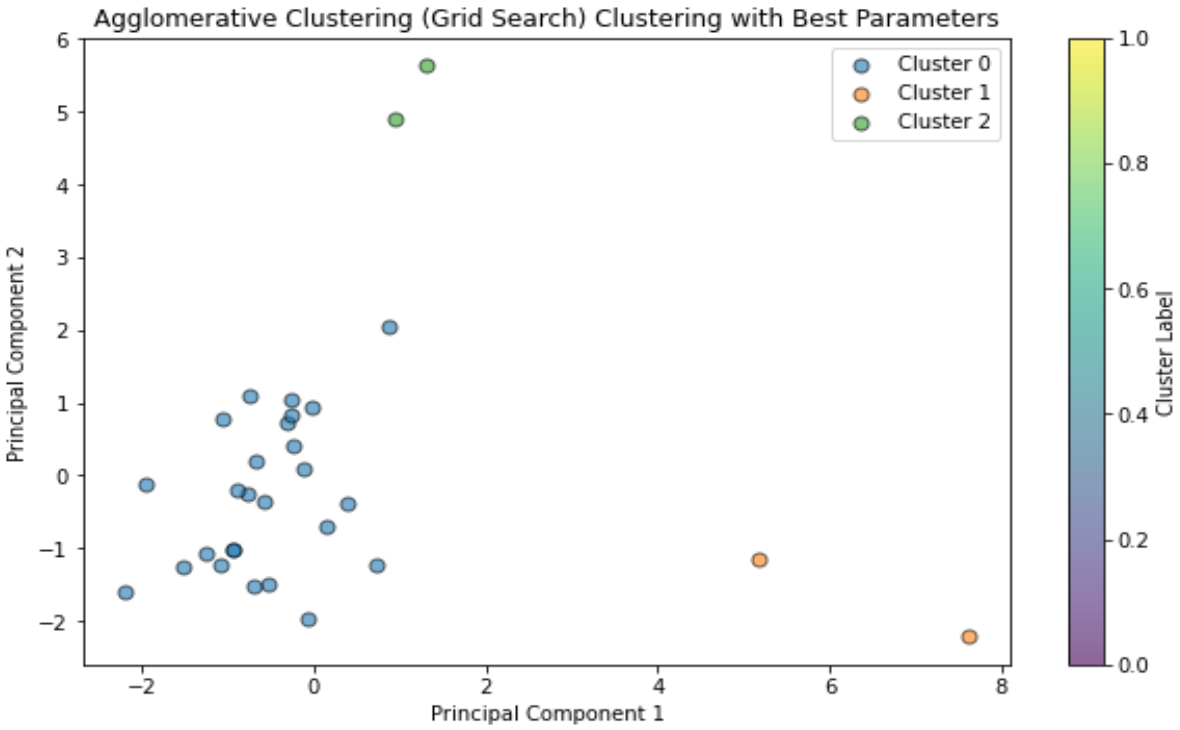
Πίνακας 12 Ονόματα datasets ανά cluster με Kmeans σε Grid Search



Εικόνα 16 KMeans Clustering με τις καλύτερες παραμέτρους σε Random Search

Dataset names per cluster for K-Means (Random Search)	
Cluster 0	pima, planning_relax, pm10, pollution, prnn_fglass, quake, rabe_266, residential_building, seeds, sonar, stock, synthetic_control, tecator, triazines_cl, triazines_reg, urban_land_cover, vehicle, vertebral_column_2classes, vertebral_column_3classes, vinnie, visualizing_environmental, visualizing_galaxy, volcanoes_a2, wifi_localization, wine, winequality-red, winequality-white, yeast
Cluster 1	robot_failures_lp4, robot_failures_lp5

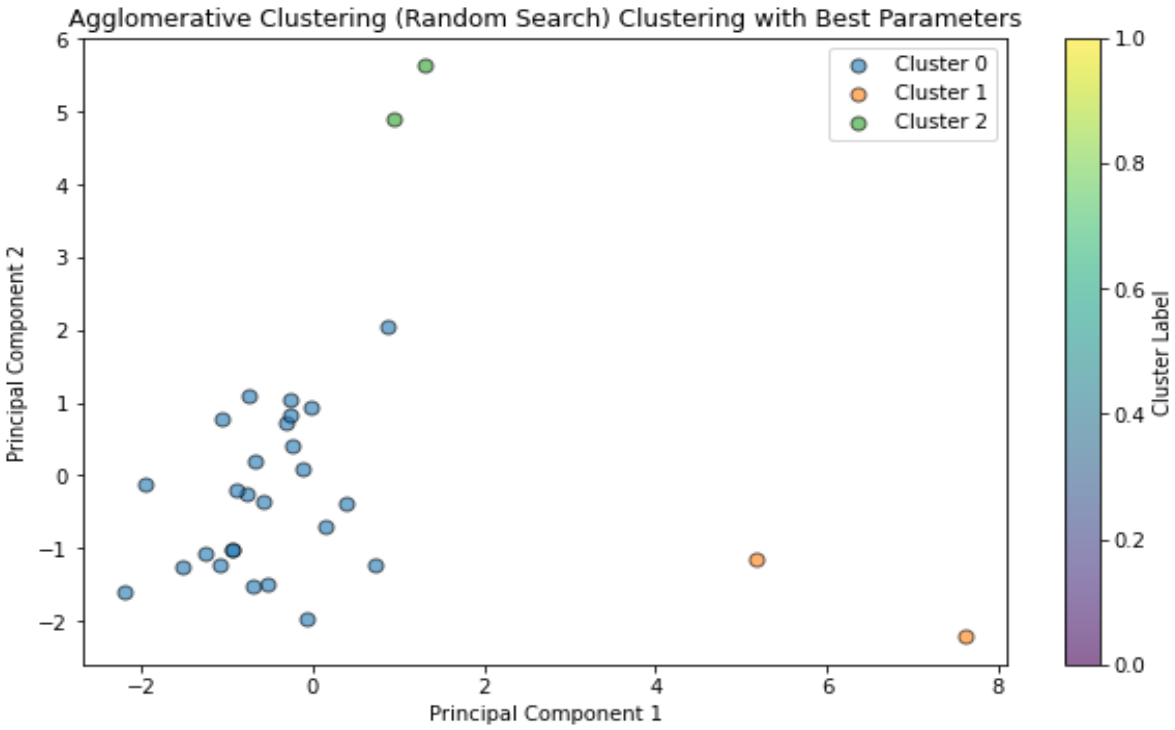
Πίνακας 13 Ονόματα datasets ανά cluster με Kmeans σε Random Search



Εικόνα 17 Agglomerative Clustering με τις καλύτερες παραμέτρους σε Grid Search

Dataset names per cluster for Agglomerative Clustering (Grid Search)	
Cluster 0	pima, planning_relax, pm10, pollution, prnn_fglass, quake, rabe_266, residential_building, seeds, sonar, stock, synthetic_control, tecator, triazines_cl, triazines_reg, urban_land_cover, vehicle, vertebral_column_2classes, vertebral_column_3classes, vinnie, visualizing_environmental, visualizing_galaxy, volcanoes_a2, wifi_localization, wine, winequality-red, winequality-white, yeast
Cluster 1	robot_failures_lp4, robot_failures_lp5
Cluster 2	triazines_cl, triazines_reg

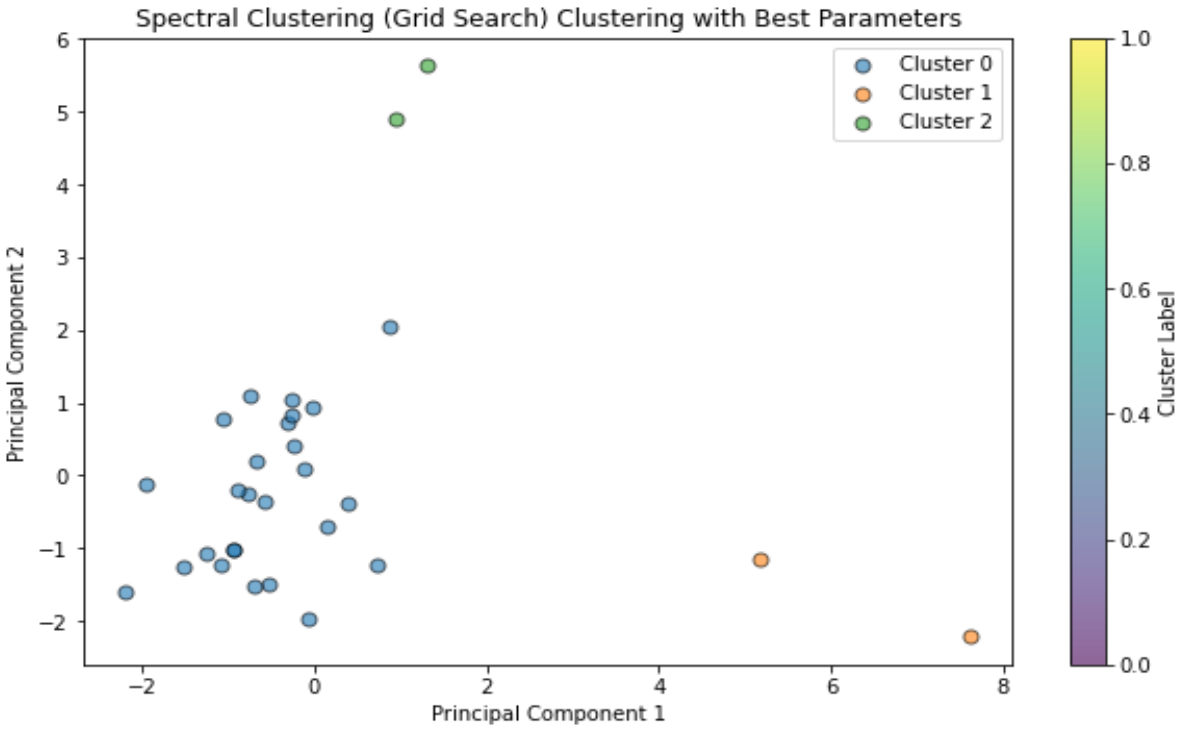
Πίνακας 14 Ονόματα datasets ανά cluster με Agglomerative σε Grid Search



Εικόνα 18 Agglomerative Clustering με τις καλύτερες παραμέτρους σε Random Search

Dataset names per cluster for Agglomerative Clustering (Random Search)	
Cluster 0	pima, planning_relax, pm10, pollution, prnn_fglass, quake, rabe_266, residential_building, seeds, sonar, stock, synthetic_control, tecator, triazines_cl, triazines_reg, urban_land_cover, vehicle, vertebral_column_2classes, vertebral_column_3classes, vinnie, visualizing_environmental, visualizing_galaxy, volcanoes_a2, wifi_localization, wine, winequality-red, winequality-white, yeast
Cluster 1	robot_failures_lp4, robot_failures_lp5
Cluster 2	triazines_cl, triazines_reg

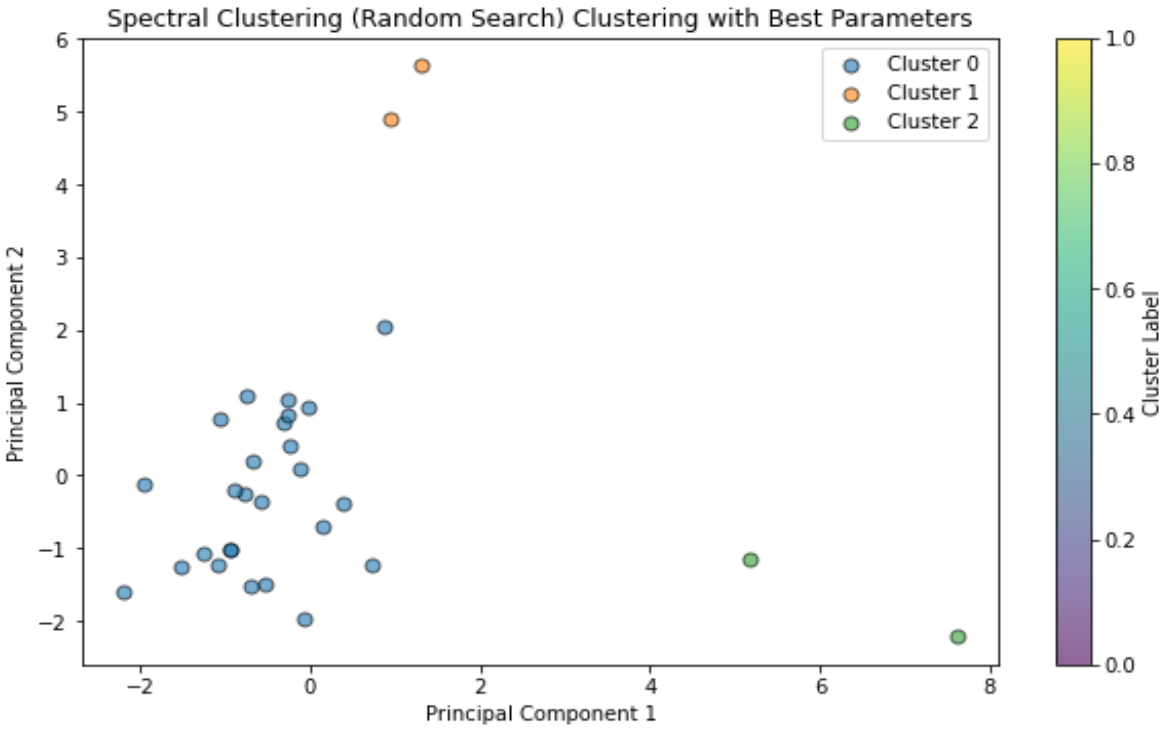
Πίνακας 15 Ονόματα datasets ανά cluster με Agglomerative σε Random Search



Εικόνα 19 Spectral Clustering με τις καλύτερες παραμέτρους σε Grid Search

Dataset names per cluster for Spectral Clustering (Grid Search)	
Cluster 0	pima, planning_relax, pm10, pollution, prnn_fglass, quake, rabe_266, residential_building, seeds, sonar, stock, synthetic_control, tecator, triazines_cl, triazines_reg, urban_land_cover, vehicle, vertebral_column_2classes, vertebral_column_3classes, vinnie, visualizing_environmental, visualizing_galaxy, volcanoes_a2, wifi_localization, wine, winequality-red, winequality-white, yeast
Cluster 1	robot_failures_lp4, robot_failures_lp5
Cluster 2	triazines_cl, triazines_reg

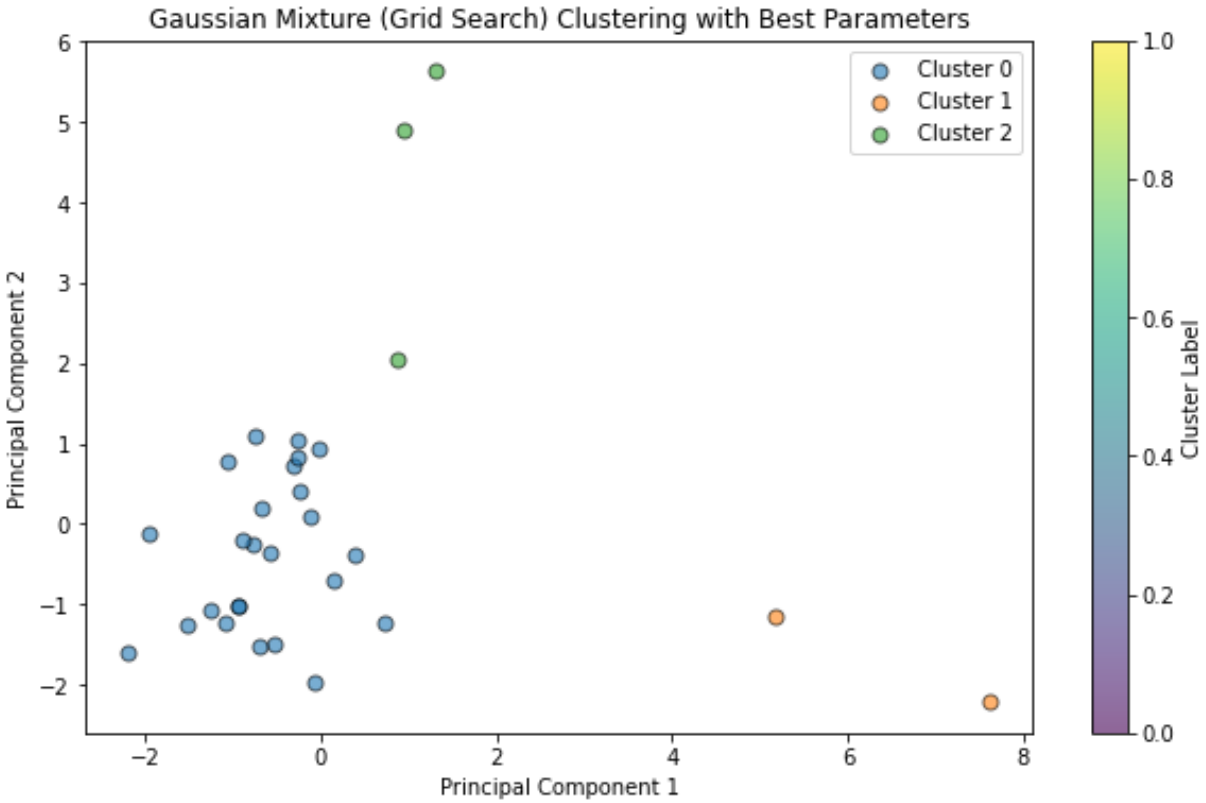
Πίνακας 16 Ονόματα datasets ανά cluster με Spectral σε Grid Search



Εικόνα 20 Spectral Clustering με τις καλύτερες παραμέτρους σε Random Search

Dataset names per cluster for Spectral Clustering (Random Search)	
Cluster 0	pima, planning_relax, pm10, pollution, prnn_fglass, quake, rabe_266, residential_building, seeds, sonar, stock, synthetic_control, tecator, triazines_cl, triazines_reg, urban_land_cover, vehicle, vertebral_column_2classes, vertebral_column_3classes, vinnie, visualizing_environmental, visualizing_galaxy, volcanoes_a2, wifi_localization, wine, winequality-red, winequality-white, yeast
Cluster 1	robot_failures_lp4, robot_failures_lp5
Cluster 2	triazines_cl, triazines_reg

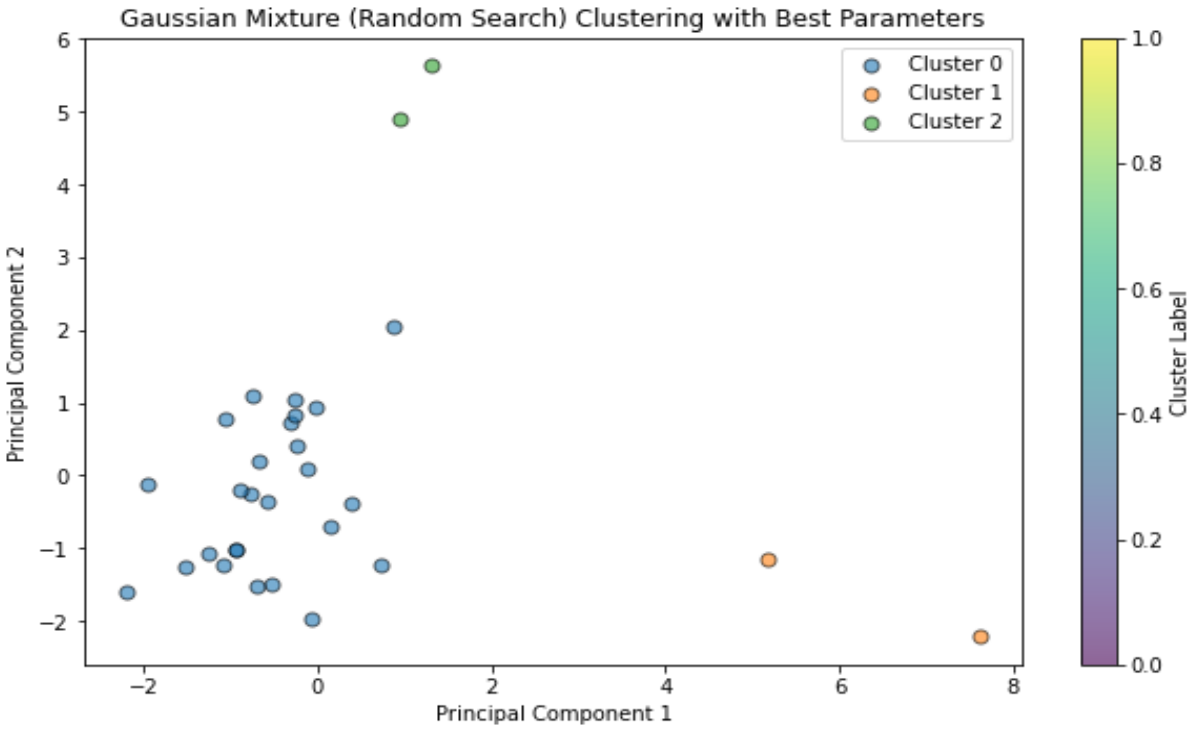
Πίνακας 17 Ονόματα datasets ανά cluster με Spectral σε Random Search



Εικόνα 21 Gaussian Mixture Clustering με τις καλύτερες παραμέτρους σε Grid Search

Dataset names per cluster for Gaussian Mixture (Grid Search)	
Cluster 0	pima, planning_relax, pm10, pollution, prnn_fglass, quake, rabe_266, residential_building, seeds, sonar, stock, synthetic_control, tecator, triazines_cl, triazines_reg, urban_land_cover, vehicle, vertebral_column_2classes, vertebral_column_3classes, vinnie, visualizing_environmental, visualizing_galaxy, volcanoes_a2, wifi_localization, wine, winequality-red, winequality-white
Cluster 1	robot_failures_lp4, robot_failures_lp5
Cluster 2	triazines_cl, triazines_reg, , yeast

Πίνακας 18 Ονόματα datasets ανά cluster με Gaussian σε Grid Search.



Εικόνα 22 Gaussian Mixture Clustering με τις καλύτερες παραμέτρους σε Random Search

Dataset names per cluster for Gaussian Mixture (Random Search)	
Cluster 0	pima, planning_relax, pm10, pollution, prnn_fglass, quake, rabe_266, residential_building, seeds, sonar, stock, synthetic_control, tecator, urban_land_cover, vehicle, vertebral_column_2classes, vertebral_column_3classes, vinnie, visualizing_environmental, visualizing_galaxy, volcanoes_a2, wifi_localization, wine, winequality-red, winequality-white, yeast
Cluster 1	robot_failures_lp4, robot_failures_lp5
Cluster 2	triazines_cl, triazines_reg

Πίνακας 19 Ονόματα datasets ανά cluster με Gaussian σε Random Search.

10.5 Πληροφορίες από την απεικόνιση συμπλέγματος

Τα γραφήματα ομαδοποίησης εξυπηρετούν πολλαπλούς σκοπούς: επικυρώνουν τη συνοχή και τον διαχωρισμό που επιτυγχάνεται από τη διαδικασία ομαδοποίησης, προσφέρουν ένα απτό μέσο για την αξιολόγηση των επιπτώσεων του συντονισμού υπερπαραμέτρων και ενδεχομένως αποκαλύπτουν υποκείμενα μοτίβα δεδομένων που θα μπορούσαν να ενημερώσουν περαιτέρω ανάλυση ή εφαρμογές πραγματικού κόσμου.

10.6 Ανάλυση και επεξήγηση των αποτελεσμάτων της ομαδοποίησης

Αυτή η ενότητα παρουσιάζει την ανάλυση των αποτελεσμάτων ομαδοποίησης που λαμβάνονται από πολλαπλά σύνολα δεδομένων χρησιμοποιώντας διάφορους αλγόριθμους ομαδοποίησης. Η ρύθμιση υπερπαραμέτρων για αυτούς τους αλγορίθμους πραγματοποιήθηκε χρησιμοποιώντας μεθόδους αναζήτησης πλέγματος και τυχαίας αναζήτησης. Το πρωταρχικό κριτήριο αξιολόγησης για την ποιότητα ομαδοποίησης ήταν η βαθμολογία σιλουέτας, η οποία μετρά πόσο

παρόμοιο είναι ένα αντικείμενο με το δικό του σύμπλεγμα σε σύγκριση με άλλα σμήνη. Οι αλγόριθμοι ομαδοποίησης που εφαρμόζονται στο αποθετήριο συνόλων δεδομένων μας έχουν αποκαλύψει ενδιαφέρουσες ομαδοποιήσεις που μπορεί να αντικατοπτρίζουν υποκείμενα μοτίβα και σχέσεις μεταξύ των συνόλων δεδομένων. Τα αποτελέσματα από τα μοντέλα K-Means, Agglomerative Clustering, Spectral Clustering και Gaussian Mix παρέχουν μια πολύπλευρη εικόνα του τρόπου με τον οποίο αυτά τα σύνολα δεδομένων ομαδοποιούνται στο πλαίσιο των εγγενών χαρακτηριστικών τους.

10.6.1 Επισκόπηση αποτελεσμάτων

Τα αποτελέσματα ομαδοποίησης υποδεικνύουν ένα συνεπές μοτίβο σε διαφορετικούς αλγορίθμους, με μικρές παραλλαγές στις εκχωρήσεις συμπλεγμάτων ανάλογα με τον αλγόριθμο και τη μέθοδο ρύθμισης υπερπαραμέτρων που χρησιμοποιείται. Συγκεκριμένα, τα σύνολα δεδομένων «robot_failures_lp4» και «robot_failures_lp5» συχνά ομαδοποιήθηκαν, υποδηλώνοντας υψηλό βαθμό ομοιότητας μεταξύ αυτών των δύο συνόλων δεδομένων, όσον αφορά στα μεταχαρακτηριστικά τους και ο οποίος αναγνωρίστηκε σε διαφορετικούς αλγόριθμους ομαδοποίησης και μεθόδους συντονισμού υπερπαραμέτρων.

10.6.2 Ομαδοποίηση K-Means

Στην περίπτωση του K-Means, παρατηρούμε ένα μεγάλο σύμπλεγμα (Cluster 0) το οποίο φαίνεται να ομαδοποιεί την πλειοψηφία των συνόλων δεδομένων μαζί. Αυτό υποδηλώνει ότι για τις επιλεγμένες παραμέτρους και τον αριθμό των συστάδων, ο αλγόριθμος K-Means προσδιορίζει μια ευρεία ομοιότητα σε αυτά τα σύνολα δεδομένων. Η ομοιότητα αυτή θα μπορούσε να βασίζεται σε διάφορους παράγοντες, όπως η διαστασιολόγηση των δεδομένων, η εξάπλωση των σημείων δεδομένων ή άλλες στατιστικές ιδιότητες που καταγράφονται στον χώρο δυνατοτήτων. Είναι ενδιαφέρον ότι ένα πολύ μικρό σύνολο δεδομένων (στο Cluster 1) φαίνεται να είναι αρκετά διαφορετικό από τα υπόλοιπα, υποδεικνύοντας ενδεχομένως μοναδικά χαρακτηριστικά που δεν μοιράζονται με τη μεγαλύτερη ομάδα. Η ομαδοποίηση K-Means, τόσο στην αναζήτηση πλέγματος όσο και στην τυχαία αναζήτηση, έδειξε σαφή διαχωρισμό μεταξύ των περισσότερων υπόλοιπων συνόλων δεδομένων και των συνόλων δεδομένων «robot_failures», υποδεικνύοντας μια σαφή διαφορά στα χαρακτηριστικά δεδομένων αυτών των συνόλων δεδομένων σε σύγκριση με άλλα. Η πλειοψηφία των συνόλων δεδομένων ομαδοποιήθηκε σε ένα μεγάλο σύμπλεγμα, με τα «robot_failures_lp4» και «robot_failures_lp5» να σχηματίζουν ένα μικρότερο, ξεχωριστό σύμπλεγμα.

10.6.3 Ομαδοποίηση Agglomerative

Το Agglomerative Clustering εμφανίζει ένα παρόμοιο μεγάλο σύμπλεγμα με τα περισσότερα σύνολα δεδομένων ομαδοποιημένα (Cluster 0). Ωστόσο, διαφοροποιείται με την προσθήκη ενός τρίτου συμπλέγματος (Cluster 2), γεγονός που υποδηλώνει ότι η ιεραρχική φύση του Agglomerative Clustering είναι σε θέση να συλλάβει ένα λεπτότερο επίπεδο λεπτομέρειας στις σχέσεις δεδομένων. Η παρουσία ενός τρίτου συμπλέγματος θα μπορούσε να σημαίνει ότι αυτά τα σύνολα δεδομένων είναι κάπως ενδιάμεσα όσον αφορά στην ομοιότητά τους με τη μεγαλύτερη ομάδα και τη διακριτή ομάδα 1. Τα αποτελέσματα του Agglomerative Clustering ήταν παρόμοια με τα K-Means, με τα σύνολα δεδομένων «robot_failures» συχνά να διαχωρίζονται από τα υπόλοιπα. Μια ενδιαφέρουσα παρατήρηση είναι ο σχηματισμός ενός τρίτου συμπλέγματος στη μέθοδο αναζήτησης πλέγματος, που

περιέχει «triazines_cl» και «triazines_reg», γεγονός που υποδηλώνει ότι αυτά τα δύο σύνολα δεδομένων μοιράζονται κάποιες ομοιότητες που είναι διαφορετικές από τα άλλα σύνολα δεδομένων.

10.6.4 Ομαδοποίηση Spectral

Το Spectral Clustering αντικατοπτρίζει σε μεγάλο βαθμό τα σμήνη που βρέθηκαν από το Agglomerative Clustering, πιθανώς λόγω της ικανότητάς του να αναγνωρίζει την πολλαπλή δομή των δεδομένων. Η συνέπεια των αποτελεσμάτων ομαδοποίησης μεταξύ αυτών των δύο μεθόδων ενισχύει την έννοια ότι υπάρχει σημαντική ομοιότητα μεταξύ της πλειοψηφίας των συνόλων δεδομένων, με μερικές ακραίες τιμές που διαθέτουν μοναδικά χαρακτηριστικά. Το Spectral Clustering, γνωστό για την ικανότητά του να αναγνωρίζει μη γραμμικά όρια σμήνους, ομαδοποίησε επίσης ξεχωριστά τα σύνολα δεδομένων «robot_failures». Ο σχηματισμός των συνεργατικών σχηματισμών ήταν συνεπής με το K-Means και το Agglomerative Clustering, επικυρώνοντας περαιτέρω το διακριτικό χαρακτήρα των συνόλων δεδομένων «robot_failures».

10.6.5 Ομαδοποίηση Gaussian Mix Spectral

Η προσέγγιση Gaussian Mix Model (GMM), η οποία υποθέτει ότι τα σημεία δεδομένων παράγονται από ένα μείγμα διαφόρων κατανομών Gauss, παρείχε αποτελέσματα που ήταν σύμφωνα με τις άλλες μεθόδους ομαδοποίησης. Τα σύνολα δεδομένων «robot_failures» συχνά αναγνωρίζονταν ως διακριτό σύμπλεγμα, ενισχύοντας το μοτίβο που παρατηρήθηκε σε άλλους αλγορίθμους.

Οι μικρότερες ομάδες σε κάθε μέθοδο, ιδιαίτερα εκείνες που διαχωρίζονται σταθερά από την κύρια ομάδα σε διαφορετικές τεχνικές ομαδοποίησης, απαιτούν προσεκτικότερη διερεύνηση. Μπορεί να αντιπροσωπεύουν εξειδικευμένες περιοχές του χώρου του συνόλου δεδομένων με συγκεκριμένα χαρακτηριστικά που θα μπορούσαν να επηρεάσουν την επεξεργασία τους, όπως η απαίτηση διαφορετικών μοντέλων μηχανικής χαρακτηριστικών ή μηχανικής μάθησης. Η παρουσία αυτών των διακριτών ομάδων μπορεί να είναι ιδιαίτερα σημαντική για εργασίες που βασίζονται στη μετα-μάθηση, όπου η κατανόηση των αποχρώσεων διαφορετικών τύπων συνόλων δεδομένων είναι κρίσιμη για την επιλογή ή το σχεδιασμό αλγορίθμων προσαρμοσμένων σε κάθε τύπο.

Αυτές οι ομάδες δεν θα πρέπει να θεωρούνται απλώς ως διαχωρισμός των συνόλων δεδομένων, αλλά μάλλον ως χάρτης πορείας για περαιτέρω ανάλυση. Θα μπορούσαν να καθοδηγήσουν τους χρήστες προς την κατανόηση των συνόλων δεδομένων που ενδέχεται να απαιτούν εξειδικευμένη προεπεξεργασία ή ποια μπορεί να είναι κατάλληλα για μάθηση μεταφοράς λόγω της ομοιότητάς τους. Αυτή η ομαδοποίηση θα μπορούσε επίσης να έχει επιπτώσεις στην ανίχνευση ανωμαλιών εντός συνόλων δεδομένων, προσδιορίζοντας εκείνες που διαφέρουν σημαντικά από την πλειονότητα ως πιθανές ακραίες τιμές ή ειδικές περιπτώσεις.

Συνοπτικά, η ανάλυση ομαδοποίησης του αποθετηρίου συνόλων δεδομένων αποκαλύπτει μια κυρίαρχη ομάδα όπου τα σύνολα δεδομένων μοιράζονται ομοιότητες και μερικές μικρότερες ομάδες όπου τα σύνολα δεδομένων έχουν ξεχωριστές ιδιότητες. Αυτό αντικατοπτρίζει την ποικιλομορφία των δεδομένων που συναντάμε στη μηχανική μάθηση και την ανάγκη για λεπτή ανάλυση για να κατανοήσουμε τον καλύτερο τρόπο αξιοποίησης αυτής της ποικιλίας. Τέτοιες γνώσεις θα μπορούσαν να είναι ανεκτίμητες για αυτοματοποιημένους αγωγούς μηχανικής μάθησης και για την καθοδήγηση της εξερεύνησης και της προεπεξεργασίας νέων συνόλων δεδομένων σε ερευνητικά και εφαρμοσμένα περιβάλλοντα.

10.6.6 Ανάλυση ομαδοποίησης

Αυτή η ενότητα παρουσιάζει την ανάλυση των αποτελεσμάτων ομαδοποίησης που λαμβάνονται από πολλαπλά σύνολα δεδομένων χρησιμοποιώντας διάφορους αλγόριθμους ομαδοποίησης, δηλαδή K-Means, Agglomerative Clustering, Spectral Clustering και Gaussian Mix Models. Η ρύθμιση υπερπαραμέτρων για αυτούς τους αλγορίθμους πραγματοποιήθηκε χρησιμοποιώντας μεθόδους αναζήτησης πλέγματος και τυχαίας αναζήτησης. Το πρωταρχικό κριτήριο αξιολόγησης για την ποιότητα ομαδοποίησης ήταν η βαθμολογία σιλουέτας, η οποία μετρά πόσο παρόμοιο είναι ένα αντικείμενο με το δικό του σύμπλεγμα σε σύγκριση με άλλα σμήνη. Η συνεπής ομαδοποίηση των «robot_failures_lp4» και «robot_failures_lp5» σε διαφορετικούς αλγορίθμους υποδηλώνει ότι αυτά τα σύνολα δεδομένων έχουν μοναδικά χαρακτηριστικά που τα ξεχωρίζουν από τα υπόλοιπα. Αυτό μπορεί να οφείλεται σε συγκεκριμένα χαρακτηριστικά ή μοτίβα που υπάρχουν στα δεδομένα και διαφέρουν από εκείνα άλλων συνόλων δεδομένων.

Ο διαχωρισμός των «triazines_cl» και «triazines_reg» στη δική τους ομάδα σε ορισμένα από τα αποτελέσματα της συσσωμάτωσης θα μπορούσε να υποδηλώνει υψηλότερο επίπεδο ομοιότητας μεταξύ αυτών των δύο συνόλων δεδομένων. Αυτό μπορεί να αποδοθεί στα δεδομένα δομής που περιέχουν, τα οποία θα μπορούσαν να έχουν εγγενείς ομοιότητες που δεν υπάρχουν σε άλλα σύνολα δεδομένων.

Το μεγάλο σύμπλεγμα που περιέχει την πλειονότητα των συνόλων δεδομένων υποδηλώνει ότι, ενώ υπάρχουν διακριτές διαφορές μεταξύ ορισμένων συνόλων δεδομένων (όπως τα σύνολα δεδομένων «robot_failures» και «τριαζίνες»), πολλά σύνολα δεδομένων έχουν κοινά χαρακτηριστικά. Αυτό μπορεί να οφείλεται σε ομοιότητες στους τύπους δεδομένων, τις δυνατότητες ή τους τομείς από τους οποίους προέρχονται αυτά τα σύνολα δεδομένων.

Τα αποτελέσματα ομαδοποίησης υπογραμμίζουν την αποτελεσματικότητα διαφόρων αλγορίθμων ομαδοποίησης στον εντοπισμό διακριτών ομάδων μέσα σε μια συλλογή συνόλων δεδομένων. Τα συνεπή μοτίβα που παρατηρούνται σε διαφορετικούς αλγορίθμους και μεθόδους συντονισμού προσδίδουν εμπιστοσύνη στην ευρωστία αυτών των λύσεων ομαδοποίησης. Αυτή η ανάλυση παρέχει πολύτιμες πληροφορίες σχετικά με τις σχέσεις και τις ομοιότητες μεταξύ διαφορετικών συνόλων δεδομένων, οι οποίες μπορούν να είναι καθοριστικές για την κατανόηση των υποκείμενων δομών στα δεδομένα και την καθοδήγηση περαιτέρω εργασιών ανάλυσης δεδομένων ή μηχανικής μάθησης.

10.6.7 Απαρίθμηση ομαδοποιημένων συνόλων δεδομένων

Ένα αναπόσπαστο μέρος της ανάλυσης μετά την ομαδοποίηση περιλαμβάνει την καταχώρηση και καταγραφή των συνόλων δεδομένων σύμφωνα με τα cluster τους. Αυτή η κατηγοριοποίηση όχι μόνο διευκολύνει μια λεπτομερή κατανόηση των ομοιοτήτων των συνόλων δεδομένων, αλλά βοηθά και στη μετα-ανάλυση, όπου τα μοτίβα μεταξύ των ομάδων μπορούν να αποφέρουν μετα-γνώσεις σχετικά με τα χαρακτηριστικά των συνόλων δεδομένων και τις σχέσεις τους.

10.6.8 Συμπέρασμα

Η μέθοδος αγκώνα και οι επακόλουθες απεικονίσεις ομαδοποίησης μετά τον συντονισμό υπερπαραμέτρων αποτελούν βασικά συστατικά ενός ολοκληρωμένου πλαισίου ανάλυσης ομαδοποίησης. Με το σχολαστικό προσδιορισμό του βέλτιστου αριθμού ομάδων και την οπτική

αναπαράσταση των αποτελεσμάτων ομαδοποίησης, οι ερευνητές και οι επαγγελματίες μπορούν να διασφαλίσουν ότι τα μοντέλα ομαδοποίησής τους είναι στατιστικά ορθά και διαισθητικά ερμηνεύσιμα, ενισχύοντας έτσι τη συνολική ποιότητα και αξιοπιστία της ανάλυσης ομαδοποίησης. Αυτές οι μεθοδολογίες, βασισμένες σε ισχυρές στατιστικές αρχές και ενισχυμένες μέσω οπτικών αναλύσεων, ενδυναμώνουν την εξαγωγή σημαντικών μοτίβων και δομών από σύνθετα σύνολα δεδομένων, προωθώντας έτσι το πεδίο της ανάλυσης συστάδων.

11. Μετάβαση από το Hyperparameter Tuning στην εις βάθος ανάλυση συνόλου δεδομένων

11.1 Εισαγωγή

Μετά την εκτεταμένη ρύθμιση υπερπαραμέτρων που πραγματοποιήθηκε στα μεταχαρακτηριστικά από τη συλλογή των προεπεξεργασμένων συνόλων δεδομένων, αποκτήσαμε ένα ισχυρό σύνολο διαμορφώσεων που αντιπροσωπεύουν τις βέλτιστες ρυθμίσεις για διάφορα μοντέλα μηχανικής μάθησης. Αυτές οι διαμορφώσεις προέκυψαν τόσο μέσω αναζήτησης πλέγματος όσο και μέσω τυχαίων μεθόδων αναζήτησης και προσφέρουν μια βαθιά ανάλυση στις πιο αποτελεσματικές παραμέτρους για το συγκεκριμένο περιβάλλον της έρευνάς μας. Το επόμενο βήμα για την περαιτέρω ανάλυση σε αυτή την διατριβή, περιλαμβάνει μια στροφή από την μοντελοκεντρική εστίαση του hyperparameter tuning σε μια προοπτική επικεντρωμένη στα δεδομένα, ώστε να μπορέσουμε να προβούμε σε μια σύγκριση των αποτελεσμάτων με στόχο την συσχέτιση των μεταχαρακτηριστικών πάνω στα σύνολα δεδομένων στην μηχανική μάθηση.

Το σκεπτικό πίσω από αυτή την ανάλυση έγκειται στην υπόθεση ότι τα χαρακτηριστικά των προεπεξεργασμένων συνόλων δεδομένων - όπως συλλαμβάνονται από τα εξαγόμενα μεταχαρακτηριστικά - μπορεί να έχουν σημαντική επίδραση στις βέλτιστες διαμορφώσεις υπερπαραμέτρων που καθορίζονται μέσω συντονισμού. Για να επικυρώσουμε αυτή την υπόθεση, θα ξεκινήσουμε μια συγκριτική ανάλυση. Αυτή η ανάλυση όχι μόνο θα αναδείξει τις συνδέσεις μεταξύ των προεπεξεργασμένων συνόλων δεδομένων και των αποτελεσμάτων από τη ρύθμιση υπερπαραμέτρων, αλλά θα μας επιτρέψει επίσης να εξάγουμε συμπεράσματα σχετικά με την προγνωστική ισχύ των μεταχαρακτηριστικών σχετικά με την απόδοση του μοντέλου.

11.2 Συγκριτική Ανάλυση των δεδομένων Triazines και Robot Failures

Στην επικείμενη ανάλυση, θα δοθεί ιδιαίτερη προσοχή σε τέσσερα διακριτά σύνολα δεδομένων: δύο που σχετίζονται με τις τριαζίνες και δύο που σχετίζονται με το robot failure. Τα προκαταρκτικά αποτελέσματα από τις αυστηρές διαδικασίες συντονισμού υπερπαραμέτρων έχουν κατηγοριοποιήσει με συνέπεια αυτά τα σύνολα δεδομένων σε ξεχωριστά συμπλέγματα. Ένα τέτοιο μοτίβο είναι ενδεικτικό των εγγενών διαφορών στη δομή τους και ενδεχομένως στα υποκείμενα φαινόμενά τους. Αυτή η συνεπής ομαδοποίηση εγείρει ενδιαφέροντα ερωτήματα σχετικά με τη φύση αυτών των συνόλων δεδομένων και την επίδραση των χαρακτηριστικών τους στην απόδοση του μοντέλου. Εμβαθύνοντας σε μια ανάλυση αυτών των συνόλων δεδομένων, στοχεύουμε να επεξηγήσουμε τις τάσεις ομαδοποίησης τους και να διερευνήσουμε το βαθμό στον οποίο αυτές οι διακρίσεις επηρεάζουν την αποτελεσματικότητα της βελτιστοποίησης υπερπαραμέτρων.

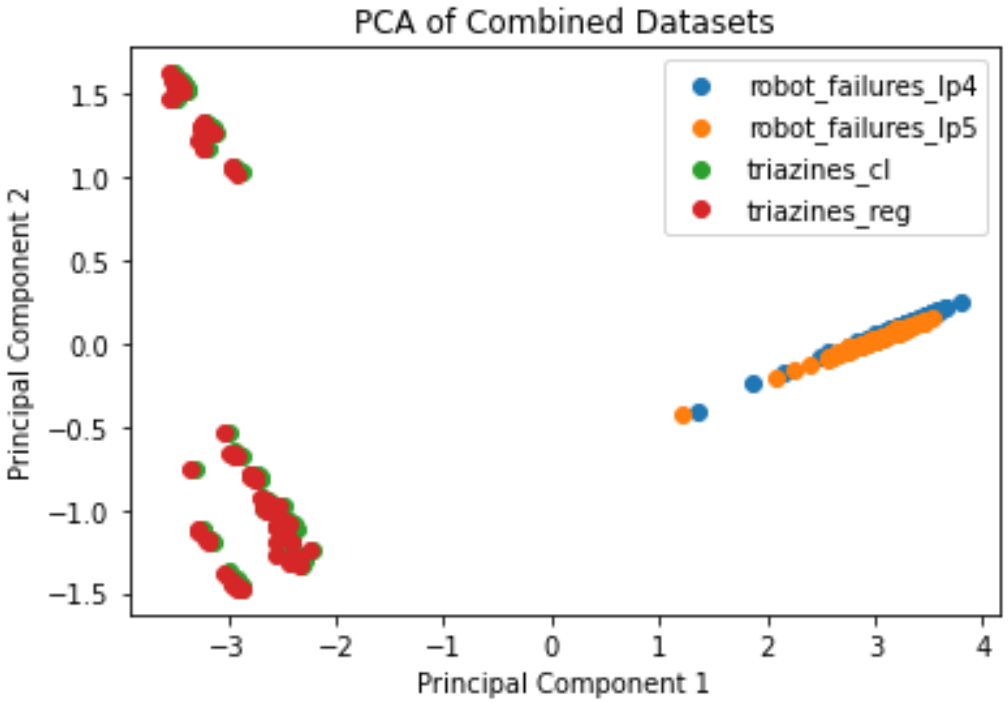
Στην ανάλυση θα χρησιμοποιηθεί μια λεπτομερή προσέγγιση, εξετάζοντας εξονυχιστικά κάθε σύνολο δεδομένων για να αποκαλύψουμε τα χαρακτηριστικά που συνέβαλαν στην ομαδοποίησή τους. Τα σύνολα δεδομένων triazines, που προέρχονται από δεδομένα χημικών ενώσεων, και τα σύνολα δεδομένων robot failure, που περιλαμβάνουν επιχειρησιακές μετρήσεις από ρομποτικά συστήματα, παρουσιάζουν μια μοναδική ευκαιρία να αντιπαραβληθούν και να συγκριθούν πώς οι διαφορετικοί τομείς δεδομένων ανταποκρίνονται σε μοντέλα μηχανικής μάθησης υπό ποικίλες ρυθμίσεις υπερπαραμέτρων. Αυτή η εξέταση όχι μόνο θα ενισχύσει την κατανόησή μας για τα σύνολα δεδομένων, αλλά θα αξιολογήσει και κριτικά την προγνωστική δύναμη των μεταχαρακτηριστικών στην πρόβλεψη των βέλτιστων υπερπαραμέτρων. Ο τελικός στόχος είναι να αντληθούν αξιοποιήσιμες πληροφορίες που μπορούν να βελτιώσουν τις μελλοντικές προσπάθειες συντονισμού υπερπαραμέτρων και να δημιουργήσουν μια πιο βαθιά σύνδεση μεταξύ των χαρακτηριστικών του συνόλου δεδομένων και της βελτιστοποίησης του μοντέλου μηχανικής μάθησης.

Dataset	Instances	Features
robot_failures_lp4	83	90
robot_failures_lp5	121	90
triazines_cl	107	58
triazines_reg	108	59

Πίνακας 20 Instances και features ανά dataset

11.3 PCA στα προεπεξεργασμένα δεδομένα

Στην ανάλυση κύριων συνιστωσών (PCA) στο σύνολο των δεδομένων robot failure και triazines, όπως παρουσιάζονται στο παρακάτω γράφημα, μπορούμε να παρατηρήσουμε πώς κατανέμονται τα σημεία δεδομένων από τα διαφορετικά σύνολα δεδομένων στο χώρο που ορίζεται από τις δύο πρώτες κύριες συνιστώσες.



Εικόνα 23 Διαχωρισμός συνόλων δεδομένων με 2 κύριες συνιστώσες PCA

Το διάγραμμα PCA απεικονίζει σαφώς το φυσικό διαχωρισμό των τεσσάρων συνόλων δεδομένων σε ομάδες με βάση τις δύο πρώτες κύριες συνιστώσες, οι οποίες αποτυπώνουν τις πιο

σημαντικές διακυμάνσεις εντός των συνδυασμένων δεδομένων. Στη φάση της διερευνητικής ανάλυσης δεδομένων, η Ανάλυση Κύριων Συνιστωσών (PCA) διεξήχθη σε τέσσερα βασικά σύνολα δεδομένων - δύο από τον τομέα των τριαζινών και δύο από το σύνολο δεδομένων του robot failure. Τα αποτελέσματα, όπως απεικονίζονται μέσω του παραπάνω γραφήματος scatter plot των δύο πρώτων κύριων συνιστωσών, διευκολύνουν την προκαταρκτική κατανόηση της δομής και διακύμανσης των δεδομένων. Για τα σύνολα δεδομένων triazines, το διάγραμμα PCA υποδηλώνει μια αξιοσημείωτη συμφωνία στην κατανομή δεδομένων. Τα σημεία δεδομένων που αντιπροσωπεύουν τόσο τα σύνολα δεδομένων triazines_cl όσο και triazines_reg παρατηρούνται να συνυπάρχουν στην ίδια περιοχή του γραφήματος. Αυτή η επικάλυψη δείχνει ότι ο μετασχηματισμός σε κύριο χώρο συστατικών αντικατοπτρίζει υψηλό βαθμό ομοιότητας στις ιδιότητες καταγραφής διακύμανσης αυτών των συνόλων δεδομένων. Το συμπέρασμα εδώ είναι ότι τυχόν διαφορές μεταξύ των συνόλων δεδομένων triazines δεν είναι οι κύριοι παράγοντες διακύμανσης και επομένως μπορεί να μην είναι σημαντικές όταν εξετάζεται η χρησιμότητα των συνόλων δεδομένων για εκπαίδευση μοντέλων εντός του χώρου χαρακτηριστικών με μειωμένο PCA.

Αντίθετα, τα σύνολα δεδομένων αποτυχίας ρομπότ, δηλαδή robot_failures_lp4 και robot_failures_lp5, αν και τα σημεία τους είναι πολύ κοντά, παρουσιάζουν μια πιο διάσπαρτη κατανομή σε όλη την πλοκή PCA. Αυτή η εξάπλωση είναι ενδεικτική μιας μεγαλύτερης μεταβλητότητας εντός του συνόλου δεδομένων σε σύγκριση με τις τριαζίνες. Κάθε σύμπλεγμα σημείων δεδομένων αποτυχίας ρομπότ φαίνεται να εκτείνεται κατά μήκος διαφορετικών αξόνων των κύριων συνιστωσών, υπονοώντας μοναδικά υποκείμενα χαρακτηριστικά ή λειτουργικές συνθήκες που επηρεάζουν το σύνολο δεδομένων. Παρ' όλα αυτά, η διασπορά είναι σχετικά μέτρια, γεγονός που υποδηλώνει ότι ενώ τα σύνολα δεδομένων ενσωματώνουν ξεχωριστά γεγονότα ή συνθήκες αποτυχίας, μοιράζονται ένα βαθμό ομοιότητας στον τρόπο με τον οποίο αυτά τα γεγονότα χαρακτηρίζονται από τα δεδομένα. Αυτή η λεπτή παραλλαγή εντός των συνόλων δεδομένων αποτυχίας ρομπότ θα μπορούσε να είναι ζωτικής σημασίας για την προγνωστική μοντελοποίηση, όπου η διαφοροποίηση μεταξύ τύπων αποτυχίας είναι συχνά ζωτικής σημασίας για τη διαγνωστική ακρίβεια.

Αυτή η ανάλυση υπογραμμίζει τη χρησιμότητα του PCA ως εργαλείου για την οπτική επιθεώρηση και υπόθεση σχετικά με τα χαρακτηριστικά του συνόλου δεδομένων πριν από την εφαρμογή μοντέλων μηχανικής μάθησης. Η γραφική αναπαράσταση βοηθά στη διάκριση της σχετικής ομοιογένειας εντός των δεδομένων των τριαζινών σε αντίθεση με την ελαφρώς ετερογενή φύση των δεδομένων αποτυχίας του ρομπότ. Περαιτέρω ομαδοποίηση ή ταξινόμηση θα μπορούσε να αποσαφηνίσει τον πιθανό αντίκτυπο αυτών των ευρημάτων στην επιλογή μοντέλων, τη μηχανική χαρακτηριστικών και, τελικά, την απόδοση πρόβλεψης. Ο πρωταρχικός στόχος είναι να αξιοποιηθούν αυτές οι πληροφορίες για την ενημέρωση στρατηγικών αποφάσεων στη διαδικασία ανάπτυξης μοντέλων, διασφαλίζοντας ότι οι επιλεγμένοι αλγόριθμοι είναι κατάλληλοι για να αξιοποιήσουν τα διακριτικά μοτίβα και τις δομές που αποκαλύπτονται στα δεδομένα.

11.4 Global stats

Στην αναλυτική διαδικασία, η αξιολόγηση των παγκόσμιων στατιστικών(global stats) είναι ένα θεμελιώδες βήμα για την απόκτηση μιας αρχικής κατανόησης των γενικών τάσεων του συνόλου δεδομένων. Για κάθε προεπεξεργασμένο σύνολο δεδομένων, υπολογίστηκαν συνολικές μετρήσεις, όπως ο μέσος όρος, η τυπική απόκλιση, η ελάχιστη, η μέγιστη και ο αριθμός των τιμών που λείπουν, παρέχοντας μια ποσοτική σύνοψη υψηλού επιπέδου. Ο μέσος όρος προσφέρει πληροφορίες σχετικά με την κεντρική τάση των χαρακτηριστικών, ενώ η τυπική απόκλιση αντικατοπτρίζει τη διασπορά ή τη μεταβλητότητα που υπάρχει στα δεδομένα. Το εύρος, που δίνεται από τις ελάχιστες και μέγιστες τιμές, συμπυκνώνει το εύρος των δεδομένων και μπορεί να ρίξει φως σε πιθανές ακραίες τιμές ή στην

ποικιλομορφία εντός των χαρακτηριστικών. Επιπλέον, ο αριθμός των ελλειπόντων τιμών, ο οποίος στην περίπτωση αυτή αναμένεται να είναι μηδενικός λόγω προηγούμενου υπολογισμού ή αφαίρεσης, επιβεβαιώνει την πληρότητα του συνόλου δεδομένων. Αυτά τα παγκόσμια στατιστικά στοιχεία όχι μόνο χρησιμεύουν ως σημείο ελέγχου για την ακεραιότητα των δεδομένων μετά την προεπεξεργασία, αλλά παρέχουν και ένα συγκριτικό καμβά για την κατανόηση των διαφορών και των ομοιοτήτων σε πολλαπλά σύνολα δεδομένων.

Αυτή η κατανόηση προωθείται με την κατασκευή ενός συγκριτικού πίνακα που καταγράφει τις αποστάσεις μεταξύ κάθε ζεύγους συνόλων δεδομένων όσον αφορά στις παγκόσμιες στατιστικές τους. Χρησιμοποιώντας έναν αριθμητικό πίνακα για τη συγκέντρωση αυτών των παγκόσμιων μετρήσεων και, στη συνέχεια, υπολογίζοντας έναν πίνακα απόστασης κατά ζεύγη, εμβαθύνουμε σε μια πιο λεπτή σύγκριση που υπερβαίνει τα μεμονωμένα σύνολα δεδομένων. Αυτός ο πίνακας απόστασης είναι ζωτικής σημασίας για την απεικόνιση της σχετικής θέσης κάθε συνόλου δεδομένων εντός του παγκόσμιου στατιστικού χώρου, επιτρέποντας τον εντοπισμό ομάδων ή ακραίων τιμών μεταξύ τους. Χρησιμεύει ως πρόδρομος για πιο σύνθετη μοντελοποίηση, επιτρέποντάς μας να προβλέψουμε πώς ορισμένα σύνολα δεδομένων μπορεί να συμπεριφέρονται σε σχέση με άλλα όταν υποβάλλονται σε αλγόριθμους μηχανικής μάθησης. Τελικά, αυτές οι παγκόσμιες στατιστικές και ο παραγόμενος πίνακας απόστασης θέτουν τις βάσεις για μια πιο λεπτή ανάλυση, προσφέροντας μια ποσοτική βάση από την οποία μπορούν να διακριθούν τα πρότυπα και να διατυπωθούν υποθέσεις σχετικά με τη συμπεριφορά του συνόλου δεδομένων.

Dataset	Global Mean	Global Std	Range	Total Missing
robot_failures_lp4	0.5960257969909475	0.22965668273560036	[0.0, 1.0]	0
robot_failures_lp5	0.5654426411924748	0.2161997652705985	[0.0, 1.0]	0
triazines_cl	0.16665066780668303	0.33296288920738	[0.0, 1.0000000000000002]	0
triazines_reg	0.17805734947975488	0.3395834048478034	[0.0, 1.0000000000000002]	0

Πίνακας 21 Χαρακτηριστικά συνόλων δεδομένων

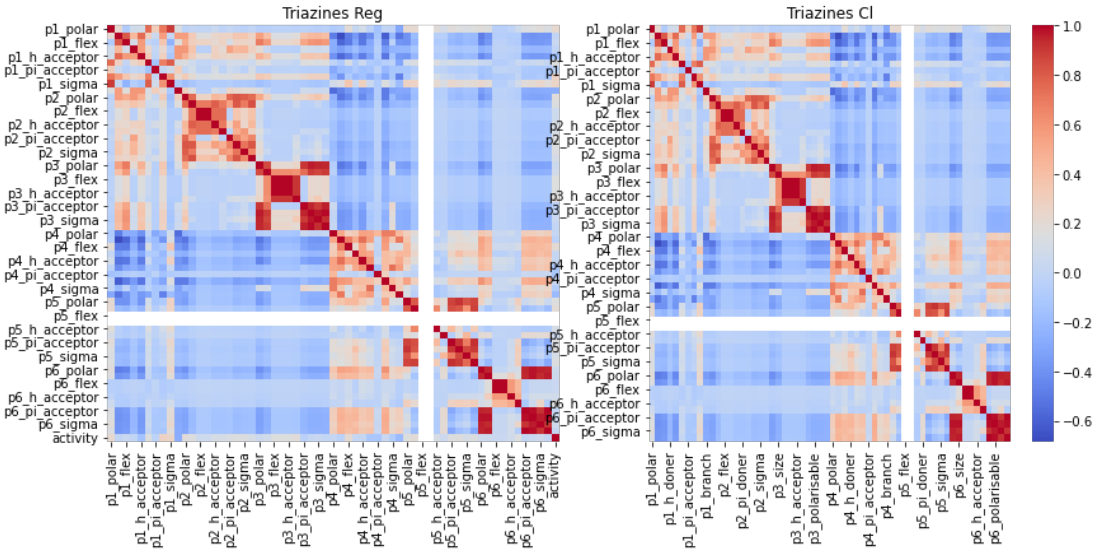
Οι μετρήσεις ομοιότητας μεταξύ των συνόλων δεδομένων παρέχουν ουσιαστικές πληροφορίες σχετικά με τη συγκριτική ανάλυσή τους. Τα σύνολα δεδομένων Triazines, triazines_cl και triazines_reg, παρουσιάζουν ιδιαίτερα υψηλό βαθμό ομοιότητας, με μια Ευκλείδεια απόσταση 0,01 να τα χωρίζει στον παγκόσμιο στατιστικό χώρο. Αυτή η εγγύτητα υπογραμμίζει τη συνέπεια της συνολικής μέσης και τυπικής απόκλισης, παρά τις πιθανές διαφορές στη δομή ή τις ιδιότητες που μπορεί να διαθέτουν τα σύνολα δεδομένων. Μια τέτοια στενή αντιστοίχιση υποδηλώνει ότι για εργασίες όπου τα παγκόσμια χαρακτηριστικά επηρεάζουν - όπως ορισμένες ταξινομήσεις ή παλινδρομήσεις - τα δύο σύνολα δεδομένων θα μπορούσαν ενδεχομένως να είναι εναλλάξιμα ή συνδυασμένα για να αυξήσουν τη διαδικασία εκπαίδευσης.

Ομοίως, παρατηρείται ισχυρή ομοιότητα μεταξύ των συνόλων δεδομένων robot failure, robot_failures_lp4 και robot_failures_lp5, που υποδεικνύεται από Ευκλείδεια απόσταση 0,03. Αυτό δείχνει ότι οι επιχειρησιακές συνθήκες ή τα συμβάντα αστοχίας που καταγράφονται σε αυτά τα σύνολα δεδομένων, αν και διακριτά, μοιράζονται κοινά στοιχεία στα στατιστικά προφίλ τους. Δεδομένου του παρόμοιου εύρους και της χαμηλής μεταβλητότητάς τους, τα μοντέλα που εκπαιδεύονται σε ένα από τα σύνολα δεδομένων αποτυχίας ρομπότ αναμένεται να έχουν παρόμοια απόδοση στο άλλο, υποθέτοντας ότι οι τρόποι αστοχίας που αντιπροσωπεύονται στα δεδομένα είναι ανάλογοι. Τα αποτελέσματα που προκύπτουν από αυτά τα μέτρα ομοιότητας είναι ελπιδοφόρα,

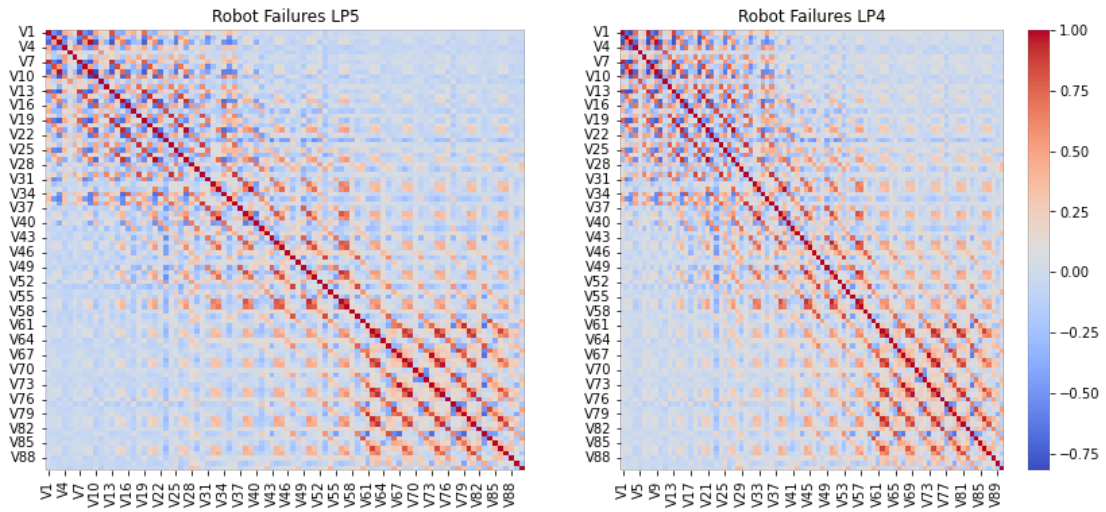
καθώς θα μπορούσαν ενδεχομένως να μειώσουν τον πλεονασμό στη συλλογή και την προεπεξεργασία δεδομένων σε τομείς όπου η απόκτηση δεδομένων είναι δύσκολη ή δαπανηρή.

11.5 Correlation heatmaps

Οι χάρτες θερμότητας συσχέτισης είναι μια οπτικά διαισθητική μέθοδος για την αναπαράσταση της ισχύος και της κατεύθυνσης των σχέσεων μεταξύ πολλαπλών μεταβλητών σε ένα σύνολο δεδομένων. Χρησιμοποιούν χρωματική κωδικοποίηση για να δηλώσουν τους συντελεστές συσχέτισης μεταξύ κάθε ζεύγους μεταβλητών. Συνήθως, τα ζεστά χρώματα υποδεικνύουν θετικές συσχετίσεις, ενώ τα ψυχρά χρώματα αντιπροσωπεύουν αρνητικές συσχετίσεις. Αυτά τα γραφήματα είναι ιδιαίτερα επωφελή για τον εντοπισμό μοτίβων και πιθανής εγγύτητας πριν από την επιλογή μοντέλων και τη μηχανική χαρακτηριστικών [37]. Μια μελέτη του Friendly (2002)[38] προτείνει ότι οι πίνακες συσχέτισης είναι απαραίτητοι για τη διερευνητική ανάλυση δεδομένων, παρέχοντας μια επισκόπηση του τρόπου με τον οποίο οι μεταβλητές σχετίζονται μεταξύ τους μέσα σε ένα πολυδιάστατο σύνολο δεδομένων. Οι χάρτες θερμότητας που παρουσιάζονται σε αυτή την έρευνα απεικονίζουν ισχυρά διαγώνια μοτίβα, χαρακτηριστικά θετικών συσχετίσεων μεταξύ μεταβλητών, και μικτές αποχρώσεις εκτός διαγωνίου, οι οποίες δείχνουν τόσο θετικές όσο και αρνητικές συσχετίσεις. Αυτές οι πληροφορίες που προέρχονται από τους χάρτες θερμότητας μπορούν να είναι καθοριστικές για την κατανόηση της υποκείμενης δομής των δεδομένων, καθώς και για την καθοδήγηση της προεπεξεργασίας δεδομένων και την επιλογή κατάλληλων αλγορίθμων για εργασίες μηχανικής μάθησης [39].



Εικόνα 24 Heatmaps των συνόλων δεδομένων triazines



Εικόνα 25 Heatmaps των συνόλων δεδομένων robot failure

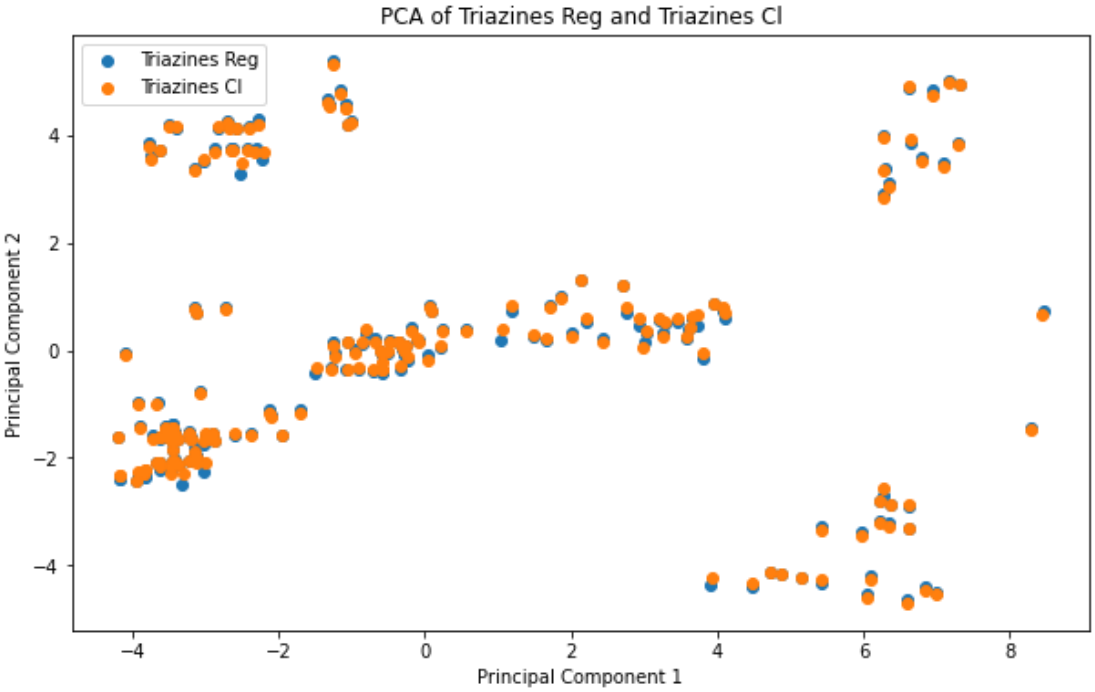
Με βάση τους παραπάνω χάρτες θερμότητας συσχέτισης για τα σύνολα δεδομένων «triazines_reg» και «triazines_cl», καθώς και τα σύνολα δεδομένων «robot_failures_lp5» και «robot_failures_lp4», μπορούμε να διακρίνουμε την εσωτερική δομή και τις σχέσεις μέσα σε κάθε σύνολο δεδομένων. Στην περίπτωση των συνόλων δεδομένων των τριαζινών, οι χάρτες θερμότητας δείχνουν μια έντονη διαγώνιο, υποδεικνύοντας ισχυρές θετικές συσχετίσεις μεταξύ ενός αριθμού μεταβλητών, κάτι που μας δείχνει ότι έχουμε χαρακτηριστικά που κινούνται παράλληλα, πιθανώς λόγω εγγενών χημικών ιδιοτήτων ή πειραματικών συνθηκών. Τα εκτός διαγωνίου μπλοκ έντονου χρώματος, τόσο σε ζεστούς όσο και σε ψυχρούς τόνους, υποδηλώνουν την παρουσία θετικών και αρνητικών συσχετίσεων μεταξύ διαφορετικών συνόλων χαρακτηριστικών, αντανακλώντας ενδεχομένως υποκείμενες βιοχημικές αλληλεπιδράσεις ή πειραματικές μεταβλητές που επηρεάζουν τα σύνολα δεδομένων με δομημένο τρόπο.

Όσον αφορά στα σύνολα δεδομένων αποτυχίας ρομπότ, οι χάρτες θερμότητας αποκαλύπτουν ένα διαφορετικό μοτίβο συσχέτισης. Ενώ υπάρχουν περιοχές ισχυρής θετικής συσχέτισης, όπως αναμένεται για λειτουργικά σχετιζόμενους ή πλεονάζοντες αισθητήρες, το συνολικό μοτίβο είναι λιγότερο συνεκτικό από αυτό των τριαζινών. Αυτό μπορεί να αντικατοπτρίζει μια πιο περίπλοκη αλληλεπίδραση χαρακτηριστικών εντός των δεδομένων αστοχίας του ρομπότ, όπου αλληλεπιδρούν διαφορετικοί τρόποι αστοχίας και λειτουργικές παράμετροι. Επιπλέον, οι χάρτες θερμότητας για αυτά τα σύνολα δεδομένων περιέχουν μεγαλύτερη διακύμανση από τη διαγώνιο, γεγονός που θα μπορούσε να είναι ενδεικτικό της ποικίλης φύσης των συμβάντων αστοχίας ή των επιχειρησιακών συνθηκών υπό τις οποίες συλλέχθηκαν τα δεδομένα. Τα ψυχρότερα χρώματα, που υποδηλώνουν αρνητικές συσχετίσεις, θα μπορούσαν να αντιστοιχούν σε ενδείξεις αισθητήρων που σχετίζονται αντιστρόφως μεταξύ τους υπό ορισμένες συνθήκες αστοχίας.

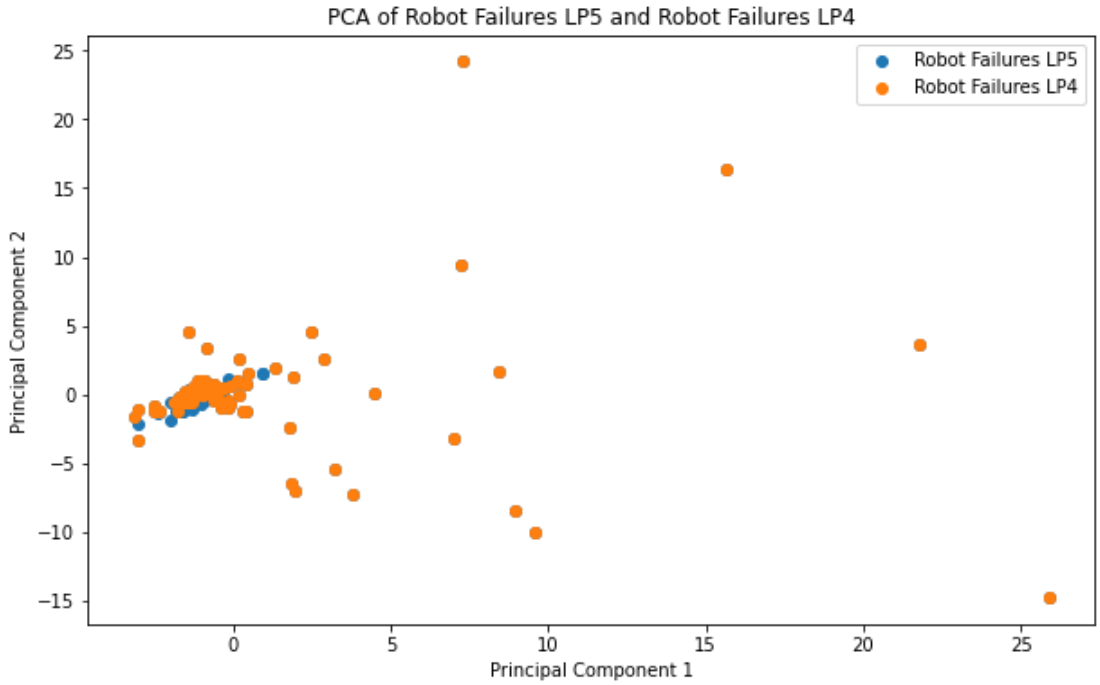
11.6 Scatter Plots

Τα γραφήματα scatter plots της ανάλυσης κύριων συνιστωσών (PCA) χρησιμοποιούνται ευρέως ως εργαλείο μείωσης διαστάσεων για την απλοποίηση της πολυπλοκότητας των δεδομένων υψηλών διαστάσεων, διατηρώντας παράλληλα τη διακύμανση που υπάρχει στο σύνολο δεδομένων στο μέγιστο δυνατό βαθμό. Αυτά τα γραφήματα σκέδασης διευκολύνουν την οπτικοποίηση των δεδομένων σε δύο ή τρεις διαστάσεις, γεγονός που επιτρέπει την παρατήρηση μοτίβων, συστάδων και ακραίων τιμών που μπορεί να μην είναι διακριτές στον αρχικό χώρο υψηλών διαστάσεων (Jolliffe,

2002). Το PCA το επιτυγχάνει αυτό μετατρέποντας τα δεδομένα σε ένα νέο σύστημα συντεταγμένων όπου οι μεγαλύτερες διακυμάνσεις από οποιαδήποτε προβολή των δεδομένων βρίσκονται στις πρώτες συντεταγμένες, που ονομάζονται κύριες συνιστώσες. Σχεδιάζοντας αυτά τα στοιχεία μεταξύ τους, οι ερευνητές μπορούν να προσδιορίσουν πώς ομαδοποιούνται μεμονωμένα σημεία δεδομένων, κάτι που μπορεί να είναι ιδιαίτερα διορατικό για τη διερευνητική ανάλυση δεδομένων, την ανίχνευση δομής, την αξιολόγηση της ποιότητας των δεδομένων και τη δημιουργία των βάσεων για περαιτέρω στατιστικές αναλύσεις ή εργασίες μηχανικής μάθησης [41]. Έτσι, τα γραφήματα σκέδασης PCA χρησιμεύουν ως απαραίτητο προκαταρκτικό βήμα στην ανάλυση δεδομένων, παρέχοντας ένα στιγμιότυπο της υποκείμενης δομής των δεδομένων σε μια σαφή και ερμηνεύσιμη οπτική μορφή.



Εικόνα 26 PCA Scatter plots των συνόλων δεδομένων triazines



Εικόνα 27 PCA Scatter plots των συνόλων δεδομένων robot failure

Τα PCA scatter plots που προκύπτουν από την ανάλυσή μας χρησιμεύουν ως οπτική απόδειξη της διακριτότητας και των πιθανών σχέσεων εντός των συνόλων δεδομένων Triazines και Robot Failures. Για τα σύνολα δεδομένων τριαζινών, η επικάλυψη σημείων από triazines_reg και triazines_cl δείχνει ότι μετά την προεπεξεργασία, την κλιμάκωση και τη μείωση των διαστάσεων, τα σύνολα δεδομένων διατηρούν ένα βαθμό ομοιότητας. Η διασπορά των σημείων κατά μήκος των κύριων συστατικών υποδηλώνει παραλλαγές στα σύνολα δεδομένων που ωστόσο καταγράφονται μέσα σε έναν κοινό χώρο χαρακτηριστικών, υπονοώντας πιθανές ομοιότητες στα χημικά χαρακτηριστικά τους ή στις μετρήσεις που καταγράφονται.

Αντίθετα, τα αποτελέσματα PCA για τα σύνολα δεδομένων αποτυχίας ρομπότ παρουσιάζουν έναν πιο διακριτό διαχωρισμό μεταξύ robot_failures_lp5 και robot_failures_lp4. Αυτή η διάκριση στο μετασχηματισμένο χώρο PCA θα μπορούσε να αντικατοπτρίζει διαφορετικές υποκείμενες συνθήκες λειτουργίας ή τρόπους αστοχίας που καταγράφονται από τα σύνολα δεδομένων. Ο διαχωρισμός θα μπορούσε να είναι κρίσιμος για διαγνωστικούς σκοπούς, όπου η διαφοροποίηση μεταξύ των τύπων αστοχιών είναι απαραίτητη.

Αυτά τα γραφήματα PCA υπογραμμίζουν την αξία του PCA ως εργαλείου για τη μείωση των διαστάσεων και την αποκάλυψη λανθανουσών δομών που μπορεί να μην είναι άμεσα εμφανείς στον χώρο υψηλότερων διαστάσεων. Προτείνουν ότι ενώ τα σύνολα δεδομένων τριαζινών μπορούν να χρησιμοποιηθούν εναλλακτικά ή να συνδυαστούν για ορισμένους τύπους ανάλυσης, τα σύνολα δεδομένων αποτυχίας ρομπότ ενδέχεται να απαιτούν πιο λεπτές προσεγγίσεις, με μοντέλα προσαρμοσμένα στα αντίστοιχα χαρακτηριστικά τους.

11.7 Separation Score

Στην συνάρτηση clustering_and_evaluate, ομαδοποιούμε και αξιολογούμε την ομοιότητα δύο συνόλων δεδομένων. Εκτελούμε διάφορα βασικά βήματα προεπεξεργασίας δεδομένων, όπως επιλογή αριθμητικών χαρακτηριστικών, καταλογισμό τιμών που λείπουν με τον μέσο όρο και κλιμάκωση για την τυποποίηση του χώρου δυνατοτήτων. Με αυτή την ομοιόμορφη προεπεξεργασία,

η συνάρτηση εφαρμόζει τον αλγόριθμο KMeans, μια δημοφιλή μη εποπτευόμενη μέθοδο μηχανικής μάθησης για τον εντοπισμό συστάδων εντός των συνδυασμένων δεδομένων και των δύο συνόλων δεδομένων. Η επιλογή δύο ομάδων (n_clusters=2) βασίζεται στην συνθήκη ότι η συνάρτηση συγκρίνει δύο διαφορετικά σύνολα δεδομένων.

Μετά την προσαρμογή των KMeans, η συνάρτηση δημιουργεί ετικέτες συμπλέγματος για κάθε σημείο δεδομένων, οι οποίες κατηγοριοποιούν κάθε σημείο δεδομένων σε ένα από τα δύο αναγνωρισμένα συμπλέγματα. Στη συνέχεια, υπολογίζει μια βαθμολογία διαχωρισμού συγκρίνοντας τις εκχωρήσεις συμπλέγματος εντός των υποσυνόλων των συνδυασμένων δεδομένων που αντιστοιχούν στα αρχικά σύνολα δεδομένων. Αυτή η βαθμολογία ποσοτικοποιεί το βαθμό στον οποίο τα δύο σύνολα δεδομένων μοιράζονται παρόμοια κατανομή δεδομένων στο χώρο δυνατοτήτων που δημιουργείται από τα βήματα προεπεξεργασίας. Μια βαθμολογία διαχωρισμού κοντά στο 1 θα έδειχνε ότι τα σύνολα δεδομένων έχουν χωριστεί σε διακριτά συμπλέγματα, ενώ μια βαθμολογία κοντά στο 0,5 θα υποδήλωνε ότι ο αλγόριθμος ομαδοποίησης δεν διαφοροποιείται καλά μεταξύ των συνόλων δεδομένων, πιθανώς λόγω πολύ παρόμοιων ή αλληλεπικαλυπτόμενων κατανομών δεδομένων. Αυτή η διαδικασία είναι πολύτιμη για την αξιολόγηση των σχετικών διαφορών ή ομοιοτήτων μεταξύ συνόλων δεδομένων σε ένα πλαίσιο μηχανικής μάθησης, η οποία μπορεί να είναι ιδιαίτερα χρήσιμη για τον προσδιορισμό της δυνατότητας μεταφοράς μοντέλων που εκπαιδεύονται σε ένα σύνολο δεδομένων σε άλλο ή της ανάγκης για ξεχωριστά μοντέλα για κάθε σύνολο δεδομένων.

Η διερεύνηση των συνόλων δεδομένων Triazines και Robot Failures αποκαλύπτει διορατικές διακρίσεις και ομοιότητες, οι οποίες είναι ιδιαίτερα εμφανείς στα πρότυπα ομαδοποίησης και στην ανάλυση κύριων συνιστωσών (PCA). Τα σύνολα δεδομένων Triazines, που περιλαμβάνουν Triazines Reg και Triazines Cl, έχουν επιδείξει αξιοσημείωτη ομοιότητα στη συμπεριφορά ομαδοποίησης, όπως αντικατοπτρίζεται από μια τέλεια βαθμολογία διαχωρισμού 1,00. Αυτό υποδηλώνει ότι, παρά τους επιδιωκόμενους σκοπούς των συνόλων δεδομένων για εργασίες παλινδρόμησης και ταξινόμησης, αντίστοιχα, μοιράζονται εγγενή χαρακτηριστικά δεδομένων που δε διακρίνονται στον εξεταζόμενο χώρο χαρακτηριστικών. Οι χάρτες θερμότητας των πινάκων συσχέτισης ενισχύουν περαιτέρω αυτή την παρατήρηση, απεικονίζοντας υψηλό βαθμό ομοιότητας στις αλληλεπιδράσεις μεταξύ των χαρακτηριστικών.

Αντίθετα, τα σύνολα δεδομένων Robot Failures, Robot Failures LP5 και Robot Failures LP4, απεικονίζουν μια έντονη διαφορά με βαθμολογία διαχωρισμού 0,00. Μια τέτοια βαθμολογία υποδηλώνει ότι κάθε σύνολο δεδομένων σχηματίζει ένα ξεχωριστό σύμπλεγμα χωρίς επικάλυψη, υπογραμμίζοντας θεμελιώδεις διαφορές στις συνθέσεις των χαρακτηριστικών τους. Τα γραφήματα διασποράς PCA απηχούν αυτή τη διχοτόμηση, τα σύνολα δεδομένων Triazines αναμειγνύονται και επικαλύπτονται εντός του χώρου χαρακτηριστικών με μειωμένο PCA, ενώ τα σύνολα δεδομένων Robot Failures διαχωρίζονται σε σαφώς διακριτές ομάδες. Τα σημεία πλοκής των πρώτων αναμειγνύονται, ενισχύοντας την ομοιότητά τους, ενώ ο διαχωρισμός των δεύτερων σε ανόμοιες περιοχές της πλοκής αναδεικνύει την ανομοιότητά τους.

Αυτά τα ευρήματα υπογραμμίζουν τη χρησιμότητα των τεχνικών μάθησης χωρίς επίβλεψη στη διάκριση των υποκείμενων δομών συνόλων δεδομένων. Τα παρατηρούμενα μοτίβα ομαδοποίησης μπορεί να προκύψουν από ποικίλες πειραματικές συνθήκες, πρωτόκολλα συλλογής δεδομένων ή εγγενείς πολυπλοκότητες εντός των δεδομένων. Ο υψηλός βαθμός συσχέτισης μεταξύ των συνόλων δεδομένων Triazines θα μπορούσε να υποδηλώνει ότι συλλαμβάνουν παρόμοια χημική υπογραφή, πιθανώς λόγω παρόμοιων χαρακτηριστικών ή ιδιοτήτων των ενώσεων που μετρήθηκαν. Από την άλλη, τα σύνολα δεδομένων Robot Failures θα μπορούσαν να αντικατοπτρίζουν διαφορετικούς τρόπους αστοχίας ή λειτουργικές συνθήκες, που εκδηλώνονται ως αποκλίνουσες υπογραφές δεδομένων που μπορούν εύκολα να διαχωριστούν μέσω ομαδοποίησης.

Περαιτέρω διερεύνηση των ειδικών χαρακτηριστικών που συμβάλλουν σε αυτά τα αποτελέσματα θα μπορούσε να παράσχει βαθύτερες γνώσεις. Για παράδειγμα, η ανάλυση των κεντροειδών των συστάδων KMeans θα μπορούσε να αποκαλύψει ποια χαρακτηριστικά έχουν τη

μεγαλύτερη επιρροή στον καθορισμό των συστάδων, ενώ οι βαθμολογίες σπουδαιότητας χαρακτηριστικών από μοντέλα μηχανικής μάθησης, όπως τα δέντρα αποφάσεων, θα μπορούσαν να προσδιορίσουν ποια χαρακτηριστικά προβλέπουν πιο έντονα την ιδιότητα μέλους του συνόλου δεδομένων. Τέτοιες αναλύσεις θα μπορούσαν να αποκαλύψουν τους παράγοντες που οδηγούν στις παρατηρούμενες ομοιότητες και διαφορές, προσφέροντας αξιοποιήσιμη νοημοσύνη για εφαρμογές συγκεκριμένου τομέα ή περαιτέρω προσπάθειες συλλογής δεδομένων.

Συμπερασματικά, η συγκριτική ανάλυση αποσαφηνίζει τις αποχρώσεις της ομοιότητας και της ασυμφωνίας των συνόλων δεδομένων, υπογραμμίζοντας τη δυνατότητα των τεχνικών μηχανικής μάθησης να παρέχουν ένα παράθυρο στην ουσία σύνθετων δεδομένων. Αυτή η διερευνητική ανάλυση δεδομένων όχι μόνο χρησιμεύει ως θεμέλιο για την προγνωστική μοντελοποίηση, αλλά συμβάλλει και στην καλύτερη κατανόηση της εγγενούς δομής των δεδομένων, με επιπτώσεις τόσο στη μεθοδολογία όσο και στην εφαρμογή στους τομείς της χημικής πληροφορικής και της ρομποτικής.

11.8 Εφαρμογή clustering στα προεπεξεργασμένα δεδομένα και ανάλυση αποτελεσμάτων

Σε αυτή την ενότητα της ανάλυσης μας, εμβαθύνουμε στη χρήση τεχνικών ομαδοποίησης σε προεπεξεργασμένα σύνολα δεδομένων. Ο πρωταρχικός στόχος είναι να ανακαλυφθούν ομάδες συνόλων δεδομένων που παρουσιάζουν συγγένεια με έναν συγκεκριμένο αλγόριθμο ομαδοποίησης, διακρίνοντας αποτελεσματικά τις ομοιότητες στις υποκείμενες δομές τους. Με την προεπεξεργασία των δεδομένων τυποποιούμε τα σύνολα δεδομένων για να αποκαλύψουμε τα πραγματικά χαρακτηριστικά τους χωρίς το θόρυβο άσχετων παραλλαγών. Η ομαδοποίηση τους μας επιτρέπει να παρατηρούμε όχι μόνο μεμονωμένες συμπεριφορές συνόλων δεδομένων, αλλά και να εντοπίζουμε μοτίβα μεταξύ ομάδων, προωθώντας μια βαθύτερη κατανόηση των εγγενών ομοιοτήτων τους. Αυτή η μεθοδική προσέγγιση μας επιτρέπει να κάνουμε παραλληλισμούς και αντιπαραβολές με προηγούμενα αποτελέσματα, οξύνοντας τη διορατικότητά μας για την αποτελεσματικότητα της αλγοριθμικής απόδοσης σε διαφορετικά σύνολα δεδομένων. Το τελικό αποτέλεσμα είναι μια πιο λεπτή εκτίμηση του ποιοι αλγόριθμοι ταιριάζουν καλύτερα σε ποιους τύπους δεδομένων, μια κατανόηση που είναι κρίσιμη για την ενίσχυση της ακρίβειας των προγνωστικών μοντέλων μας.

Best algorithm in preprocessed datasets	
Algorithm	**Datasets**
Spectral Clustering	pima, quake, residential_building, synthetic_control, vertebral_column_2classes, vertebral_column_3classes, visualizing_galaxy
K-Means	planning_relax, pm10, prnn_fglass, rabe_266, robot_failures_lp5, seeds, tecator, triazines_cl, triazines_reg, urban_land_cover, vehicle, visualizing_environmental, volcanoes_a2, wine, winequality-red, winequality-white
Agglomerative Clustering	pollution, robot_failures_lp4, stock, wifi_localization, yeast
Gaussian Mixture	sonar, vinnie

Πίνακας 22 Καλύτεροι αλγόριθμοι στα προεπεξεργασμένα σύνολα δεδομένων

Τα αποτελέσματα ομαδοποίησης ενισχύουν πράγματι την προηγούμενη ανάλυση. Το γεγονός ότι και τα δύο σύνολα δεδομένων triazines, triazines_cl και triazines_reg, μοιράζονται το K-Means ως τον καλύτερο αλγόριθμο ομαδοποίησης με βάση την υψηλότερη βαθμολογία σιλουέτας προσθέτει βάρος στο επιχείρημα ότι έχουν παρόμοιες δομές και μοτίβα. Η βαθμολογία σιλουέτας είναι ένα

μέτρο του πόσο παρόμοιο είναι ένα αντικείμενο μέσα στο δικό του σύμπλεγμα σε σύγκριση με αντικείμενα σε άλλα συμπλέγματα και μια υψηλότερη βαθμολογία σιλουέτας υποδεικνύει ένα μοντέλο με καλύτερα καθορισμένα συμπλέγματα. Δεδομένου ότι το K-Means καταγράφει ουσιαστικά τις εγγενείς συστάδες στα σύνολα δεδομένων triazines πιο αποτελεσματικά από άλλους αλγόριθμους, υποδηλώνει ότι αυτά τα σύνολα δεδομένων μπορεί να παρουσιάζουν ομοιογενή χαρακτηριστικά και κατανομές συστάδων, οι οποίες ευθυγραμμίζονται με τις προηγούμενες παρατηρήσεις PCA όπου τα σύνολα δεδομένων triazines εμφανίστηκαν στενά ομαδοποιημένα.

Αντίθετα, τα σύνολα δεδομένων αποτυχίας ρομπότ, robot_failures_lp4 και robot_failures_lp5, προσδιορίζονται ως τα πλέον κατάλληλα για διαφορετικούς αλγόριθμους ομαδοποίησης - Agglomerative Clustering και K-Means, αντίστοιχα. Αυτή η απόκλιση συνεπάγεται ότι τα υποκείμενα μοτίβα δεδομένων και οι δομές στα σύνολα δεδομένων αποτυχίας ρομπότ είναι πιο περίπλοκα και ενδέχεται να διαφέρουν σημαντικά, σύμφωνα με τα προηγούμενα αποτελέσματα PCA που υποδηλώνουν μεγαλύτερη μεταβλητότητα μεταξύ αυτών των δύο συνόλων δεδομένων. Η συσσωμάτωση και η ομαδοποίηση K-Means προσεγγίζουν την ομαδοποίηση με διαφορετικό τρόπο. Η πρώτη είναι μια ιεραρχική μέθοδος ομαδοποίησης που δημιουργεί μοντέλα με βάση την εγγύτητα των σημείων δεδομένων, ενώ η δεύτερη χωρίζει τα δεδομένα σε συστάδες με βάση το μέσο όρο των παρατηρήσεων μέσα σε κάθε σύμπλεγμα. Η προτίμηση για διαφορετικούς αλγορίθμους υπογραμμίζει την ανομοιότητα στα χαρακτηριστικά ομαδοποίησής τους και υποστηρίζει την ιδέα ότι αυτά τα σύνολα δεδομένων θα πρέπει να προσεγγίζονται με προσαρμοσμένες στρατηγικές ανάλυσης και μοντελοποίησης.

Επομένως, αυτά τα αποτελέσματα ομαδοποίησης προσφέρουν πολύτιμη επιβεβαίωση των αρχικών ευρημάτων και υπογραμμίζουν τη σημασία της κατανόησης των χαρακτηριστικών του συνόλου δεδομένων πριν από την επιλογή αλγορίθμων. Η συμφωνία στα αποτελέσματα ομαδοποίησης για τα σύνολα δεδομένων τριαζινών και η ασυμφωνία στα αποτελέσματα για τα σύνολα δεδομένων αποτυχίας ρομπότ είναι μια ενημερωτική ανακάλυψη που μπορεί να καθοδηγήσει μελλοντικές αποφάσεις προεπεξεργασίας δεδομένων, μηχανικής χαρακτηριστικών και κατάρτισης μοντέλων.

11.9 Ανάλυση αποτελεσμάτων

Στην ανάλυση που ακολουθεί, διερευνούμε την ασυμφωνία που παρατηρείται μεταξύ της ομαδοποίησης των μεταχαρακτηριστικών και των αποτελεσμάτων ομαδοποίησης που προέρχονται από τα προεπεξεργασμένα σύνολα δεδομένων, εστιάζοντας ειδικά στα σύνολα δεδομένων robot failure. Αυτή η απόκλιση εγείρει ενδιαφέροντα ερωτήματα σχετικά με τη φύση των συνόλων δεδομένων και τον αντίκτυπο της προεπεξεργασίας στα αποτελέσματα ομαδοποίησης.

11.10 Ομαδοποίηση βάσει μετα-χαρακτηριστικών έναντι προεπεξεργασμένης ομαδοποίησης δεδομένων

Κατά τη διάρκεια του αρχικού σταδίου της ανάλυσης, η ομαδοποίηση που βασίζεται σε μεταχαρακτηριστικά τοποθέτησε σταθερά τα σύνολα δεδομένων αποτυχίας ρομπότ μέσα στο ίδιο σύμπλεγμα. Τα μεταχαρακτηριστικά συνήθως ενσωματώνουν χαρακτηριστικά υψηλού επιπέδου των συνόλων δεδομένων, όπως διακύμανση χαρακτηριστικών, ασυμμετρία ή κύρτωση. Η ομοιότητα στα μεταχαρακτηριστικά υποδηλώνει ότι σε υψηλό επίπεδο, τα σύνολα δεδομένων αποτυχίας ρομπότ μοιράζονται γενικές στατιστικές ιδιότητες, πιθανώς λόγω παρόμοιων κλιμάκων χαρακτηριστικών ή κατανομών [42]. Αντίθετα, η ομαδοποίηση που εφαρμόστηκε απευθείας στα προεπεξεργασμένα σύνολα δεδομένων, η οποία λαμβάνει υπόψη τις μεμονωμένες τιμές χαρακτηριστικών και τις

αλληλεπιδράσεις τους, είχε ως αποτέλεσμα τον διαχωρισμό των συνόλων δεδομένων αποτυχίας ρομπότ σε διακριτά συμπλέγματα. Αυτό υποδηλώνει ότι όταν αναλύουμε τα δεδομένα σε πιο λεπτομερές επίπεδο, οι διαφορές γίνονται αρκετά έντονες ώστε να επηρεάζουν τα αποτελέσματα ομαδοποίησης.

11.11 Επιπτώσεις της προεπεξεργασίας στην ομαδοποίηση

Τα ίδια τα βήματα προεπεξεργασίας μπορούν να έχουν σημαντική επίδραση στα αποτελέσματα ομαδοποίησης. Η τυποποίηση, η κανονικοποίηση και η μείωση των διαστάσεων μπορούν να αλλάξουν το χώρο δεδομένων με τέτοιο τρόπο ώστε να αποκαλύπτουν ή να αποκρύπτουν μοτίβα μέσα στα δεδομένα [44]. Για παράδειγμα, η κλιμάκωση χαρακτηριστικών θα μπορούσε να μειώσει τον αντίκτυπο των ακραίων τιμών ή να ομαλοποιήσει την κλίμακα μεταξύ των χαρακτηριστικών, οδηγώντας έτσι σε διαφορετικά αποτελέσματα ομαδοποίησης. Ομοίως, εάν η ανάλυση κύριων συνιστωσών (PCA) χρησιμοποιήθηκε κατά τη διάρκεια της προεπεξεργασίας, θα μπορούσε να έχει μετασχηματίσει τα δεδομένα για να επισημάνει τις πιο σημαντικές διακυμάνσεις εις βάρος της παραμικρής λεπτομέρειας, η οποία θα μπορούσε να ήταν κρίσιμη στην αρχική ομαδοποίηση μεταχαρακτηριστικών [45].

11.12 Διαφοροποίηση στα αποτελέσματα ομαδοποίησης: Robot Failure Datasets

Η απόκλιση στην ομαδοποίηση μεταξύ της ανάλυσης μεταχαρακτηριστικών και των προεπεξεργασμένων συνόλων δεδομένων για τις περιπτώσεις αστοχίας ρομπότ μπορεί να αποδοθεί στην ιδιαιτερότητα των χαρακτηριστικών και στη φύση των δεδομένων που συλλέγονται. Είναι κατανοητό ότι τα μεταχαρακτηριστικά δεν κατέγραψαν ορισμένες λεπτές αποχρώσεις που έγιναν εμφανείς μετά την προεπεξεργασία. Αυτές οι λεπτές αποχρώσεις μπορεί να σχετίζονται με τους συγκεκριμένους τρόπους αστοχίας ή τις συνθήκες λειτουργίας που είναι μοναδικές για κάθε σύνολο δεδομένων αποτυχίας ρομπότ, οι οποίες γίνονται περισσότερο ή λιγότερο εμφανείς ανάλογα με τον τρόπο επεξεργασίας των δεδομένων πριν από την ομαδοποίηση [36]. Επιπλέον, η φύση των γεγονότων αποτυχίας, τα οποία μπορεί να είναι πολύ συγκεκριμένα και εξαρτώμενα από το πλαίσιο, θα μπορούσε να συμβάλει σε αυτή την ανισότητα. Τα βήματα προεπεξεργασίας—όπως η κανονικοποίηση, η κλιμάκωση ή ο χειρισμός τιμών που λείπουν—μεταβάλλουν τη δομή και την κατανομή των δεδομένων. Όταν τα μεταχαρακτηριστικά αντικατοπτρίζουν αυτές τις αλλαγές, ενδέχεται να αποτυπώνουν με μεγαλύτερη ακρίβεια τις λεπτές αποχρώσεις και τις ιδιαιτερότητες των τρόπων αστοχίας ή των συνθηκών λειτουργίας. Εάν εξακολουθούν να υπάρχουν αποκλίσεις, θα μπορούσαν να αποδοθούν στην εγγενή πολυπλοκότητα και ετερογένεια των συνόλων δεδομένων αποτυχίας, τα οποία ακόμη και η προηγμένη προεπεξεργασία και η εξαγωγή μεταχαρακτηριστικών ενδέχεται να μην ομογενοποιήσουν πλήρως [36],[46].

11.13 Συνέπεια και ομοιότητα των Triazines Datasets

Αντίθετα, τα σύνολα δεδομένων triazines παρέμειναν συνεπή τόσο στην ομαδοποίηση βάσει μεταχαρακτηριστικών όσο και στην προεπεξεργασμένη ομαδοποίηση δεδομένων. Αυτή η συνέπεια μπορεί να υποδηλώνει ότι οι ιδιότητες των χημικών ενώσεων και οι πειραματικές συνθήκες υπό τις οποίες συλλέχθηκαν τα δεδομένα είναι σταθερές και ανθεκτικές στις αλλαγές που εισάγονται από την προεπεξεργασία. Τα αποτελέσματα ομαδοποίησης, σε αυτή την περίπτωση, θα τονίσουν την ισχυρή φύση των δεδομένων των τριαζινών, υποδηλώνοντας ότι τα βασικά χαρακτηριστικά που οδηγούν στο σχηματισμό συστάδων διατηρούνται ακόμη και μετά την προεπεξεργασία [47].

11.12 Συμπεράσματα και μελλοντικές εργασίες

Αυτή η ανάλυση φωτίζει τις πολυπλοκότητες που εμπλέκονται στην ομαδοποίηση συνόλων δεδομένων και τη σημασία της εξέτασης τόσο των μεταχαρακτηριστικών υψηλού επιπέδου όσο και των επιπτώσεων της προεπεξεργασίας δεδομένων. Τονίζει την ανάγκη προσεκτικής εξέτασης των σταδίων προεπεξεργασίας και των πιθανών επιπτώσεών τους στα πρότυπα που αποκαλύπτονται μέσω της ομαδοποίησης. Οι μελλοντικές εργασίες θα μπορούσαν να περιλαμβάνουν μια πιο λεπτομερή εξέταση των βημάτων προεπεξεργασίας για την κατανόηση των συγκεκριμένων επιπτώσεών τους στα αποτελέσματα ομαδοποίησης. Τα μεταχαρακτηριστικά, τα οποία χρησιμεύουν ως περιγραφείς υψηλού επιπέδου συνόλων δεδομένων, έχουν ομαδοποιήσει με συνέπεια τα σύνολα δεδομένων robot failure, υποδηλώνοντας μια επιφανειακή ομοιότητα στις γενικές στατιστικές ιδιότητές τους. Αυτή η ευθυγράμμιση θα μπορούσε να υποδεικνύει παρόμοιες κλίμακες χαρακτηριστικών ή κατανομών σε αυτά τα σύνολα δεδομένων, καθώς τα μεταχαρακτηριστικά συχνά ενσωματώνουν χαρακτηριστικά, όπως η διακύμανση, η ασυμμετρία ή η κύρτωση, τα οποία είναι ενδεικτικά της γενικής δομής των συνόλων δεδομένων. Επιπλέον, η διερεύνηση εναλλακτικών τεχνικών ομαδοποίησης και η ενσωμάτωση γνώσεων τομέα θα μπορούσε να αποσαφηνίσει περαιτέρω τη διάκριση μεταξύ των συνόλων δεδομένων robot failure που παρατηρούνται στην προεπεξεργασμένη ομαδοποίηση.

Τα μεταχαρακτηριστικά διαδραματίζουν κεντρικό ρόλο στην κατανόηση και την αξιοποίηση συνόλων δεδομένων στο πλαίσιο της μηχανικής μάθησης. Πρόκειται ουσιαστικά για συνοπτικά στατιστικά στοιχεία ή χαρακτηριστικά που εξάγονται από σύνολα δεδομένων που καταγράφουν πληροφορίες υψηλού επιπέδου, όπως ο αριθμός των χαρακτηριστικών, η διακύμανση των χαρακτηριστικών, η ασυμμετρία, η παρουσία ακραίων τιμών και η συσχέτιση χαρακτηριστικών. Στη μηχανική μάθηση, τα μεταχαρακτηριστικά είναι ζωτικής σημασίας για διάφορους λόγους: βοηθούν στη σύγκριση συνόλων δεδομένων, καθοδηγούν τη διαδικασία επιλογής αλγορίθμων, ενημερώνουν τη μηχανική χαρακτηριστικών και παρέχουν ακόμη και πληροφορίες για την αναμενόμενη απόδοση του μοντέλου [18].

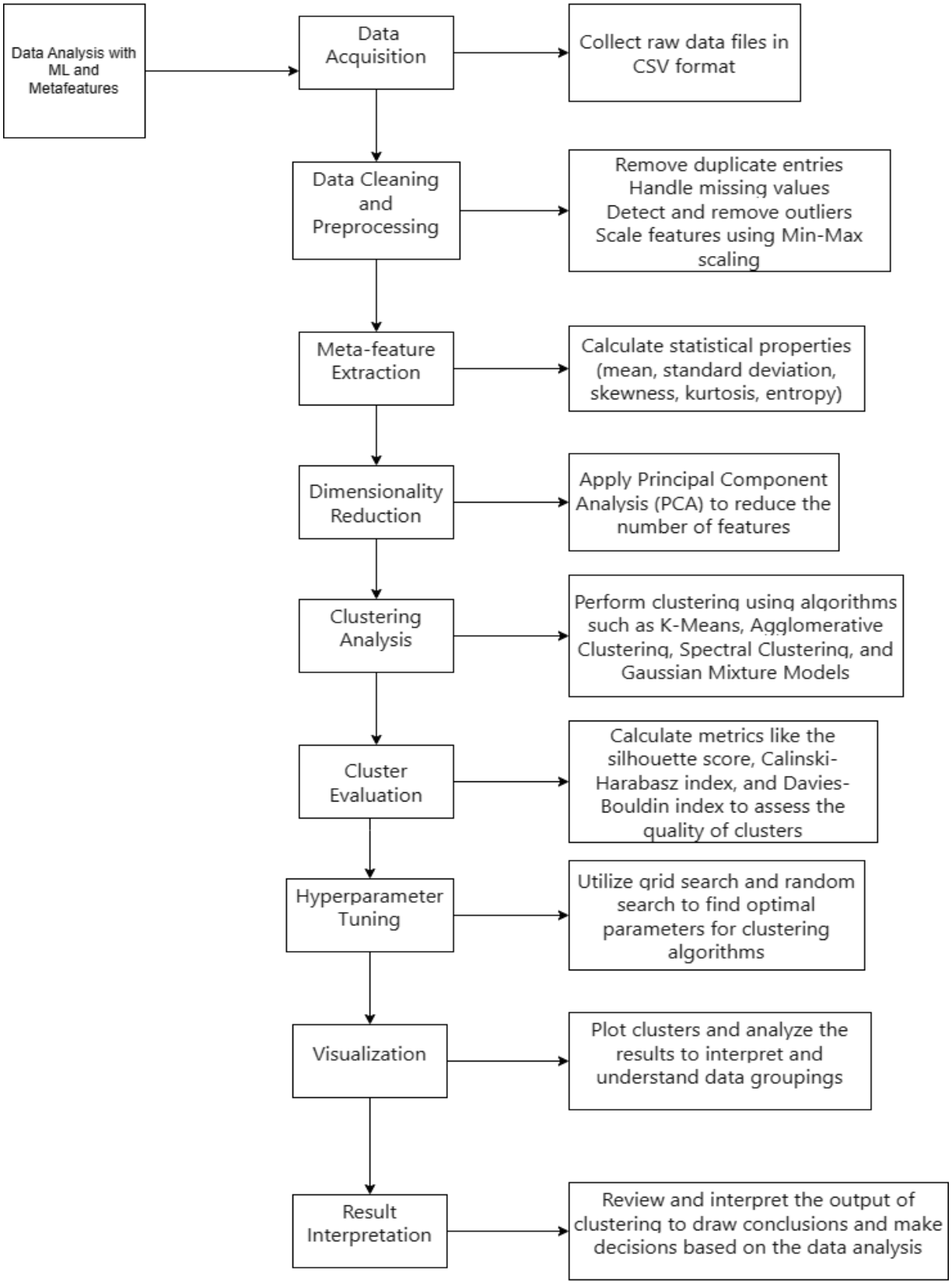
Στο πλαίσιο των παρατηρούμενων διαφορών στα αποτελέσματα ομαδοποίησης μεταξύ των συνόλων δεδομένων robot failure, τα μεταχαρακτηριστικά μπορεί να έχουν ομογενοποιήσει τα σύνολα δεδομένων συνοψίζοντας τα πολύπλοκα και δυνητικά μοναδικά χαρακτηριστικά τους σε ευρύτερα στατιστικά μέτρα. Μια τέτοια σύνοψη θα μπορούσε να παραβλέψει συγκεκριμένες συμπεριφορές χαρακτηριστικών ή αλληλεπιδράσεις που γίνονται σημαντικές μετά την προεπεξεργασία. Όταν η ομαδοποίηση πραγματοποιείται σε προεπεξεργασμένα σύνολα δεδομένων, τα δεδομένα εξετάζονται λεπτομερέστερα. Τα προεπεξεργασμένα δεδομένα μπορούν να επηρεάσουν τα αποτελέσματα ομαδοποίησης, γεγονός που θα μπορούσε να οδηγήσει στον παρατηρούμενο διαχωρισμό [48]. Η απόκλιση που παρατηρήθηκε στα αποτελέσματα ομαδοποίησης μετά την προεπεξεργασία - όπου τα σύνολα δεδομένων robot failure κατηγοριοποιήθηκαν διαφορετικά - υπογραμμίζει το μετασχηματιστικό αντίκτυπο της προεπεξεργασίας στα δεδομένα. Αυτή η διακύμανση υποδηλώνει την παρουσία αποχρώσεων, συγκεκριμένων χαρακτηριστικών εντός των συνόλων δεδομένων αποτυχίας ρομπότ που γίνονται πιο έντονα μέσω της προεπεξεργασίας, επηρεάζοντας την ομαδοποίηση των συνόλων δεδομένων [43]. Τέτοια ευρήματα τονίζουν τη σημασία της εξέτασης τόσο των μεταχαρακτηριστικών όσο και των λεπτομερών χαρακτηριστικών που προκύπτουν μετά την προεπεξεργασία, καθώς επηρεάζουν συλλογικά την επιλογή των μοντέλων μηχανικής μάθησης και την επακόλουθη απόδοσή τους.

Η σημασία των μεταχαρακτηριστικών στη μηχανική μάθηση δεν μπορεί να υπερεκτιμηθεί. Παρέχουν έναν προκαταρκτικό φακό μέσω του οποίου τα σύνολα δεδομένων μπορούν να κατανοηθούν και να συγκριθούν, βοηθώντας στην αρχική επιλογή αλγορίθμων και τεχνικών προεπεξεργασίας. Χρησιμεύουν ως πυξίδα για την πλοήγηση στον περίπλοκο χώρο της επιλογής αλγορίθμων, συχνά μέσω μετα-μάθησης, όπου η γνώση του τρόπου με τον οποίο ορισμένοι τύποι

δεδομένων ανταποκρίνονται σε διαφορετικούς αλγόριθμους μπορεί να χρησιμοποιηθεί για την πρόβλεψη των καλύτερων αλγορίθμων για νέα σύνολα δεδομένων [49]. Τα μεταχαρακτηριστικά συμβάλλουν, επίσης, στην κατανόηση του θεωρήματος «No Free Lunch» στη μηχανική μάθηση, το οποίο υποθέτει ότι κανένα μοντέλο δεν είναι καθολικά ανώτερο. Ως εκ τούτου, η επιλογή των μοντέλων θα πρέπει να ενημερώνεται από τα χαρακτηριστικά των διαθέσιμων δεδομένων [50]. . Μια ενημερωμένη προσέγγιση που αξιοποιεί τις μεταδυνατότητες για την επιλογή αλγορίθμων μπορεί να οδηγήσει σε πιο αποτελεσματικές και προσαρμοσμένες λύσεις μηχανικής μάθησης.

Στην περίπτωσή μας, η συνέπεια των αποτελεσμάτων ομαδοποίησης για τα σύνολα δεδομένων triazines επιβεβαιώνει την ευρωστία των μεταχαρακτηριστικών τους. Προτείνουν μια εγγενή ιδιότητα των δεδομένων που είναι ανθεκτική στην επίδραση της προεπεξεργασίας, ενώ τα σύνολα δεδομένων αποτυχίας ρομπότ, ίσως λόγω της πιο ευαίσθητης φύσης τους, παρουσιάζουν σημαντικές αλλαγές μετά την προεπεξεργασία, όπως φαίνεται στα ανόμοια αποτελέσματα ομαδοποίησης. Υπογραμμίζει τη σημασία της εξέτασης τόσο των μεταχαρακτηριστικών όσο και των αποτελεσμάτων της πραγματικής επεξεργασίας δεδομένων κατά τη λήψη αποφάσεων σε ροές εργασίας μηχανικής μάθησης. Οι μελλοντικές εργασίες θα πρέπει να εμβαθύνουν στην ανάπτυξη προηγμένων πλαισίων μετα-μάθησης που μπορούν να αξιοποιήσουν τα μεταχαρακτηριστικά όχι μόνο για να προβλέψουν τους καταλληλότερους αλγόριθμους, αλλά και να προτείνουν βέλτιστους αγωγούς προεπεξεργασίας για διαφορετικά σύνολα δεδομένων. Τέτοιες προσπάθειες θα μπορούσαν να αυτοματοποιήσουν και να βελτιώσουν τη διαδικασία επιλογής μοντέλου, οδηγώντας σε βελτιωμένη αποδοτικότητα και αποτελεσματικότητα στις ροές εργασίας μηχανικής μάθησης. Επιπλέον, η ενσωμάτωση πληροφοριών για συγκεκριμένους τομείς στην ανάλυση μεταχαρακτηριστικών θα μπορούσε να προσφέρει μια βαθύτερη κατανόηση των δεδομένων, ενισχύοντας περαιτέρω τη διαδικασία λήψης αποφάσεων σε εφαρμογές μηχανικής μάθησης.

Τα αντικρουόμενα αποτελέσματα στην ομαδοποίηση υπογραμμίζουν την ανάγκη για μια βαθύτερη εξέταση των μεταχαρακτηριστικών και του ρόλου τους στην προεπεξεργασία και την επιλογή μοντέλων. Αυτή η διατριβή θέτει τις βάσεις για τέτοιες εξερευνήσεις, αναδεικνύοντας τον κρίσιμο ρόλο των μεταχαρακτηριστικών, αναγνωρίζοντας παράλληλα τους περιορισμούς τους. Συνεχίζοντας να διερευνάται ο τρόπος με τον οποίο τα χαρακτηριστικά των δεδομένων και η προεπεξεργασία αλληλεπιδρούν με την απόδοση των αλγορίθμων, ο στόχος είναι να κινηθούμε σε πιο προσαρμοστικά και έξυπνα συστήματα μηχανικής μάθησης που κατανοούν καλύτερα και χρησιμοποιούν τα υποκείμενα μοτίβα μέσα στα δεδομένα. Καθώς προχωράμε στη μηχανική μάθηση και την εξόρυξη δεδομένων, η προσεκτική εξαγωγή και ανάλυση των μεταχαρακτηριστικών θα συνεχίσει να αποτελεί βασικό παράγοντα για την επιτυχή εφαρμογή μοντέλων σε πολύπλοκα προβλήματα του πραγματικού κόσμου.



Εικόνα 28 Ροή εργασιών κώδικα

Appendix A: Datasets

Κατά τη διάρκεια αυτής της διατριβής, χρησιμοποιήθηκε μια συλλογή 30 συνόλων δεδομένων για την επικύρωση των προτεινόμενων μοντέλων και μεθόδων, κυρίως αριθμητικής φύσης, για να εξασφαλίσει μια ισχυρή και ολοκληρωμένη αξιολόγηση. Αυτά τα σύνολα δεδομένων καλύπτουν διάφορους τομείς, αντικατοπτρίζοντας ένα ευρύ φάσμα πραγματικών εφαρμογών και προκλήσεων στην ανάλυση δεδομένων. Κάθε σύνολο δεδομένων επιλέχθηκε για τα μοναδικά χαρακτηριστικά του και τη συνάφειά του με συγκεκριμένα ερευνητικά ερωτήματα που εξετάστηκαν σε αυτή τη μελέτη. Αυτά τα σύνολα δεδομένων, που προέρχονται από διάφορα αξιόπιστα αποθετήρια, όπως το UCI Machine Learning Repository, το Kaggle και άλλα ακαδημαϊκά ιδρύματα, προσφέρουν μια ευρεία αναπαράσταση αριθμητικών δεδομένων σε πολλούς τομείς.

1. Pima Dataset: Προερχόμενο από ιατρική έρευνα, αυτό το σύνολο δεδομένων περιέχει διαγνωστικές μετρήσεις που σχετίζονται με παθήσεις διαβήτη μεταξύ του ινδικού πληθυσμού Pima.

2. Planning Relax Dataset: Αυτό το σύνολο δεδομένων αφορά σε προβλήματα προγραμματισμού και σχεδιασμού, όπου ο στόχος είναι η βελτιστοποίηση ακολουθιών εργασιών υπό ορισμένους περιορισμούς.

3. PM10 Dataset: Εστιάζοντας σε περιβαλλοντικά δεδομένα, παρακολουθεί τα επίπεδα αιωρούμενων σωματιδίων (PM10), παρέχοντας πληροφορίες σχετικά με την ποιότητα του αέρα και τις τάσεις ρύπανσης.

4. Pollution Dataset: Περιλαμβάνοντας διάφορους περιβαλλοντικούς δείκτες, αυτό το σύνολο δεδομένων είναι ζωτικής σημασίας για τη μελέτη των επιπτώσεων των βιομηχανικών και αστικών δραστηριοτήτων στην ποιότητα του αέρα.

5. PRNN FGlass Dataset: Μια συλλογή δεδομένων που χρησιμοποιούνται στην αναγνώριση προτύπων, ειδικά για την ταξινόμηση τύπων γυαλιού, ένα βασικό καθήκον στην εγκληματολογική ανάλυση.

6. Quake Dataset: Προερχόμενο από σεισμολογικές μελέτες, αυτό το σύνολο δεδομένων περιλαμβάνει σεισμικές κυματομορφές και συναφή χαρακτηριστικά για την ανίχνευση και ανάλυση σεισμών.

7. Rabe 266 Dataset: Συνήθως χρησιμοποιείται στη στατιστική μάθηση, περιλαμβάνει σημεία δεδομένων για συγκριτική αξιολόγηση αλγορίθμων παλινδρόμησης και ταξινόμησης.

8. Residential Building Dataset: Περιέχει οικονομικές και δομικές παραμέτρους κτιρίων κατοικιών, οι οποίες είναι ζωτικής σημασίας για την ανάλυση και τα μοντέλα πρόβλεψης της αγοράς ακινήτων.

9. Robot Failures LP4 and LP5 Datasets: Αυτά τα σύνολα δεδομένων προέρχονται από τον τομέα της ρομποτικής, περιέχουν ενδείξεις αισθητήρων και λειτουργικά δεδομένα που βοηθούν στη διάγνωση και την πρόβλεψη μηχανικών βλαβών σε ρομπότ.

10. Seeds Dataset: Χρησιμοποιείται στη γεωργική έρευνα, περιλαμβάνει χαρακτηριστικά διαφόρων τύπων σπόρων και χρησιμοποιείται για την κατηγοριοποίηση και διαφοροποίηση των γεωργικών προϊόντων.

11. Sonar Dataset: Χρησιμοποιείται για την αναγνώριση αντικειμένων κάτω από τη θάλασσα. Αυτό το σύνολο δεδομένων περιλαμβάνει σήματα σόναρ και τα χαρακτηριστικά τους για εργασίες πλοήγησης και ανίχνευσης υποβρυχίων.

12. Stock Dataset: Περιέχει δεδομένα χρηματιστηριακής αγοράς, συμπεριλαμβανομένων των διακυμάνσεων των τιμών και του όγκου συναλλαγών, απαραίτητα για τη χρηματοοικονομική μοντελοποίηση και την οικονομετρική ανάλυση.

13. Synthetic Control Dataset: Αποτελείται από τεχνητά παραγόμενα δεδομένα, βοηθά στη δοκιμή των μηχανισμών ελέγχου σε συνθετικά περιβάλλοντα για διάφορες μελέτες προσομοίωσης.

14. Tecator Dataset: Αυτό χρησιμοποιείται στην επιστήμη των τροφίμων για φασματογραφική ανάλυση, παρέχοντας δεδομένα σχετικά με τις ιδιότητες των τροφίμων και τις αξιολογήσεις ποιότητας.

15. Triazines CL και Reg Datasets: Περιέχει χημικά δεδομένα που σχετίζονται με την κατηγορία ζιζανιοκτόνων τριαζινών, που χρησιμοποιούνται για τη μελέτη χημικών ιδιοτήτων και βιολογικής δραστηριότητας.

16. Urban Land Cover Dataset: Περιλαμβάνει δεδομένα δορυφορικών εικόνων για την ταξινόμηση της αστικής κάλυψης γης, ζωτικής σημασίας για τον πολεοδομικό σχεδιασμό και τις αποφάσεις πολιτικής χρήσης γης.

17. Vehicle Dataset: Αυτό το σύνολο δεδομένων αποτελείται από ανάλυση σιλουέτας οχήματος για την αναγνώριση και ταξινόμηση διαφορετικών τύπων οχημάτων με βάση τα σχήματά τους.

18. Vertebral Column 2Classes και 3Classes Datasets: Περιέχουν εμβιομηχανικά χαρακτηριστικά ορθοπεδικών ασθενών, χρήσιμα για ιατρική διάγνωση που σχετίζεται με παθήσεις της σπονδυλικής στήλης.

19. Vinnie Dataset: Ένα εξειδικευμένο σύνολο δεδομένων που χρησιμοποιείται για εργασίες αναγνώρισης εικόνας, το οποίο περιέχει χαρακτηριστικά που εξάγονται από οπτικά δεδομένα.

20. Visualizing Environmental και Galaxy Datasets: Αυτά τα σύνολα δεδομένων είναι καθοριστικής σημασίας για εργασίες αστρονομικής και περιβαλλοντικής απεικόνισης, που περιέχουν σημεία δεδομένων που βοηθούν στη δημιουργία προσομοιώσεων και μοντέλων.

21. Volcanoes A2 Dataset: Περιλαμβάνει γεωλογικά δεδομένα που σχετίζονται με ηφαιστειακές δραστηριότητες, καθοριστικά για την ανάλυση και πρόβλεψη ηφαιστειακών κινδύνων.

22. WiFi Localization Dataset: Περιέχει δεδομένα ισχύος σήματος από δίκτυα Wi-Fi που χρησιμοποιούνται για υπηρεσίες εντοπισμού θέσης σε εσωτερικούς χώρους και βάσει τοποθεσίας.

23. Wine και Wine Quality (Red and White) Datasets: Εστίαση στα χαρακτηριστικά διαφορετικών δειγμάτων οίνων, απαραίτητα για την αξιολόγηση της ποιότητας και την ταξινόμηση στην αμπελουργία.

24. Yeast Dataset: Περιλαμβάνει χαρακτηριστικά σχετιζόμενα με τη βιολογία ζύμης, που χρησιμοποιούνται για διάφορες γενετικές και πρωτεομικές μελέτες.

Αυτά τα σύνολα δεδομένων υποβλήθηκαν σε προεπεξεργασία για να διασφαλιστεί η συμβατότητα με τις αναλυτικές τεχνικές που χρησιμοποιήθηκαν σε αυτήν την έρευνα. Τα σύνολα δεδομένων έχουν υποβληθεί σε αυστηρή προεπεξεργασία, συμπεριλαμβανομένης της κανονικοποίησης για την τυποποίηση του feature scale, την εύρεση και το χειρισμό ελλειπουσών τιμών καθώς και της ανίχνευσης ακραίων τιμών για τη διασφάλιση της ποιότητας των δεδομένων. Αυτή η αυστηρή προετοιμασία διασφαλίζει την αξιοπιστία και την εγκυρότητα των ερευνητικών ευρημάτων και συμπερασμάτων που εξάγονται από την ανάλυση. Αυτά τα βήματα προεπεξεργασίας είναι κρίσιμα για την εγκυρότητα των υπολογιστικών μοντέλων που αναπτύσσονται στην παρούσα διατριβή. Η αριθμητική φύση αυτών των συνόλων δεδομένων παρουσιάζει μια ευκαιρία για την εφαρμογή προηγμένων τεχνικών στατιστικής και μηχανικής μάθησης για την αποκάλυψη μοτίβων και την εξαγωγή ουσιαστικών συμπερασμάτων.

Βιβλιογραφία:

[1] Turing, A. M. (1950). Computing Machinery and Intelligence. Mind, LIX(236).

[2] Feurer, M., & Hutter, F. (2019). Hyperparameter Optimization. In Automated Machine Learning: Methods, Systems, Challenges. Springer, Cham.

[3] Hutter, F., Kotthoff, L., & Vanschoren, J. (Eds.). (2019). Automated Machine Learning: Methods, Systems, Challenges. Springer, Cham.

[4] He, Q., & Zhuang, F. (2021). AutoML: A Survey of the State-of-the-Art. Knowledge-Based Systems, 212, 106622.

[5] Wolpert, D. H., & Macready, W. G. (1997). No Free Lunch Theorems for Optimization. IEEE Transactions on Evolutionary Computation, 1(1).

[6] Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research.

[7] Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., & Hutter, F. (2015). Efficient and Robust Automated Machine Learning. In Advances in Neural Information Processing Systems.

[8] Olson, R. S., & Moore, J. H. (2019). TPOT: A Tree-Based Pipeline Optimization Tool for Automating Machine Learning. In Workshop on Automatic Machine Learning at ICML.

[9] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.

[10] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters.

[11] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd.

[12] Kotthoff, L. (2014). Algorithm selection for combinatorial search problems: A survey. AI Magazine.

[13] Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. In Advances in Neural Information Processing Systems.

[14] Gijsbers, P., LeDell, E., Thomas, J., Poirier, S., Bischl, B., & Vanschoren, J. (2019). An Open Source AutoML Benchmark. In arXiv preprint arXiv:1907.00909.

[15] Eiben, A. E., & Smith, J. E. (2015). Introduction to Evolutionary Computing. Springer.

[16] Vanschoren, J. (2018). Meta-learning: A survey. arXiv preprint arXiv:1810.03548.

[17] Thornton, C., Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2013). Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '13).

[18] Vilalta, R., & Drissi, Y. (2002). A perspective view and survey of meta-learning. Artificial Intelligence Review, 18(2).

[19] Brazdil, P., Giraud-Carrier, C., Soares, C., & Vilalta, R. (2008). Metalearning: Applications to Data Mining. Springer Science & Business Media.

[20] Pfahringer, B., Bensusan, H., & Giraud-Carrier, C. (2000, August). Meta-learning by landmarking various learning algorithms. In Proceedings of the Seventeenth International Conference on Machine Learning (ICML '00).

[21] Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics.

[22] Rokach, L., & Maimon, O. (2005). Clustering methods. In Data mining and knowledge discovery handbook. Springer, Boston, MA.

[23] Von Luxburg, U. (2007). A tutorial on spectral clustering. Statistics and computing.

[24] Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society. Series B (methodological).

[25] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics.

[26] Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics-theory and Methods.

[27] Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence.

[28] McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference.

[29] Van Der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy Array: A Structure for Efficient Numerical Computation. Computing in Science & Engineering.

[30] Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research.

[31] Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering.

[32] Jones, E., Oliphant, T., Peterson, P., et al. (2001). SciPy: Open Source Scientific Tools for Python.

[33] Waskom, M. (2021). Seaborn: Statistical Data Visualization. Journal of Open Source Software.

[34] Caliński, T., & Harabasz, J. (1974). "A dendrite method for cluster analysis." Communications in Statistics.

[35] Kaufman, L., & Rousseeuw, P. J. (2009). "Finding Groups in Data: An Introduction to Cluster Analysis". John Wiley & Sons.

[36] Pedregosa et al. (2011). "Scikit-learn: Machine Learning in Python." Journal of Machine Learning Research.

[37] Wilcox, R. R. (2005). Introduction to robust estimation and hypothesis testing. Academic Press.

[38] Friendly, M. (2002). Corrgrams: Exploratory displays for correlation matrices. The American Statistician.

[39] Michael, J. A., Kurz, A., & Holowka, D. (2013). Correlation heatmaps: a tool for understanding ecology and evolution. Trends in Ecology & Evolution.

[40] Jolliffe, I. T. (2002). Principal Component Analysis. Springer Series in Statistics. New York: Springer-Verlag.

[41] Ringnér, M. (2008). What is principal component analysis? Nature Biotechnology.

[42] Guyon, I., Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research.

[43] Hastie, T., Tibshirani, R., Friedman, J. (2009). The Elements of Statistical Learning. Springer Series in Statistics.

[44] García, S., Luengo, J., Herrera, F. (2015). Data Preprocessing in Data Mining. Springer.

[45] Jolliffe, I. T. (2002). Principal Component Analysis for Special Types of Data. Springer Series in Statistics.

[46] Liu, H., Motoda, H. (2007). Computational Methods of Feature Selection. Chapman and Hall/CRC.

[47] Handl, J., Knowles, J., Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. Bioinformatics.

[48] Brazdil, P., Soares, C., & Da Costa, J. P. (2003). Ranking learning algorithms: Using IBL and meta-learning on accuracy and time results. Machine Learning.

[49] Lemke, C., & Gabrys, B. (2010). Meta-learning for time series forecasting and forecast combination. Neurocomputing.

[50] Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. Neural computation.

[51] Poulakis, Y., Doulkeridis, C., & Kyriazis, D. (2020). AutoClust: A Framework for Automated Clustering based on Cluster Validity Indices. In *Proceedings of the 20th IEEE International Conference on Data Mining (ICDM)*. University of Piraeus, Greece.

[52] Vanschoren, J. (2018). Meta-Learning: A Survey.

[53] Poulakis, G. (2020). Unsupervised AutoML: A Study on Automated Machine Learning in the Context of Clustering [Master's thesis, University of Piraeus]. Department of Digital Systems.