Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Χημικών Μηχανικών
Τομέας ΙΙ: Ανάλυσης, Σχεδιασμού και Ανάπτυξης Διεργασιών και Συστημάτων
Μονάδα Αυτόματης Ρύθμισης και Πληροφορικής

Διπλωματική Εργασία

Ανάπτυξη καινοτόμων μεθόδων νανοπληροφορικής για την αυτοματοποιημένη συσταδοποίηση υλικών και την πρόβλεψη ανεπιθύμητων ιδιοτήτων τους

Καλούτση Μαριάννα

Επιβλέπων: Καθηγητής Χαράλαμπος Σαρίμβεης

Αθήνα,
Ιούνιος 2021

National Technical University of Athens
School of Chemical Engineering
Department II: Department of Process Analysis and Plant Design
Unit of Process Control and Informatics

Diploma Thesis

Development of novel nano-informatics methods for the automated clustering of materials and the prediction of undesired properties

Kaloutsi Marianna

Supervisor: Professor Haralambos Sarimveis

Athens,
June 2021

Περιεχόμενα

Περίληψη v
Abstract vii
Ευχαριστίες ix
Εισαγωγή 1
Νανοϋλικά και νανοσωματίδια 3
1.1. Ιδιότητες νανοσωματιδίων 3
1.2. Κατηγοριοποίηση νανοσωματιδίων 4
1.2.1. Νανοσωματίδια μετάλλων 4
1.2.2. Νανοσωματίδια μεταλλικών οξειδίων 4
1.3. Εφαρμογές σε βιολογικά συστήματα 4
1.4. Τοξικότητα νανοσωματιδίων 5
1.5. Ανάγκη πρόβλεψης τοξικότητας 9
Ανάπτυξη προβλεπτικών μοντέλων 10
2.1. Μεθοδολογίες πρόβλεψης 10
2.1.1. Μεθοδολογία QSAR 10
2.1.2. Read-across 11
2.2. Προεπεξεργασία δεδομένων 12
2.2.1. Κανονικοποίηση δεδομένων 12
2.2.2. Εξαγωγή χαρακτηριστικών 13
2.2.3. Ανάλυση κύριων συνιστωσών 14
Αξιολόγηση παραγόμενων μοντέλων 16
3.1. Εσωτερική αξιολόγηση 16
3.1.1. K-Fold Cross Validation 16
3.1.2. Leave-one-out Cross Validation 17
3.2. Εξωτερική αξιολόγηση 18
Αλγόριθμος Kennard Stone 19
3.3. Στατιστικά εργαλεία 19
3.3.1. Πίνακας σύγχυσης 19
3.3.2. Ακρίβεια και βαθμός σφάλματος 20
3.3.3. Ειδικότητα και ευαισθησία 21
3.3.4. Συντελεστής συσχέτισης Matthews 21
3.4. Έλεγχος τυχαίας επιλογής 22
3.5. Διαίρεση σε training, validation και test set 23
3.6. Πεδίο εφαρμογής 24

Ανάπτυξη μεθοδολογίας 27
4.1. Μεθοδολογία συσταδοποίησης 27
4.2. Μεθοδολογίες Κατηγοριοποίησης νέων δειγμάτων 39
4.2.1. Μέθοδος Κατηγοριοποίησης 1 40
4.2.2. Μέθοδος Κατηγοριοποίησης 2 43
4.2.3. Μέθοδος Κατηγοριοποίησης 3 44
Μελέτες περιπτώσεων 45
5.1. Σύνολο νανοσωματιδίων υδρ(οξειδείων) μετάλλων 45
5.2. Σύνολο υπερ-παραμαγνητικών νανοσωματιδίων οξειδίου του σιδήρου 46
5.3. Σύνολο μεταλλικών οξειδίων Α 47
5.4. Σύνολο μεταλλικών οξειδίων Β 49
Παρουσίαση αποτελεσμάτων 50
6.1. Πρόβλεψη τοξικότητας χρησιμοποιώντας όλες τις διαθέσιμες μεταβλητές 50
6.1.1. Σύνολο δεδομένων MeHydOx 51
6.1.2. Σύνολο δεδομένων SPIONs 59
6.1.3. Σύνολο δεδομένων MeOx 65
6.1.4. Σύνολο δεδομένων Cytotox 72
6.2. Πρόβλεψη τοξικότητας εφαρμόζοντας επιλογή μεταβλητών 76
6.2.1. Σύνολο δεδομένων Cytotox 77
6.3. Πρόβλεψη τοξικότητας εφαρμόζοντας ανάλυση κύριων συνιστωσών 80
6.3.1. Σύνολο δεδομένων MeHydOx 81
6.3.2. Σύνολο δεδομένων SPIONs 86
6.3.3. Σύνολο δεδομένων MeOx 92
6.3.4. Σύνολο δεδομένων Cytotox 97
6.4. Συνολική αποτίμηση του μοντέλου και των μεθοδολογιών προεπεξεργασίας .. 101
6.4.1. Σύγκριση μεταξύ των Μεθόδων Κατηγοριοποίησης και των συνόλων δεδομένων 101
6.4.2. Σύγκριση επιδόσεων με βιβλιογραφικές δοκιμές μοντέλων 105
Συμπεράσματα και προτάσεις για περαιτέρω έρευνα 107
7.1. Ανάλυση αποτελεσμάτων και συμπεράσματα 108
7.2. Προτάσεις για μελλοντική έρευνα 110
Διάθεση του κώδικα στην επιστημονική κοινότητα 111
Παράρτημα 115
Γλώσσα προγραμματισμού Python 115
Βιβλιοθήκες 115
Βιβλιογραφία 118

Περίληψη

Στο πλαίσιο της παρούσας Διπλωματικής Εργασίας αναπτύχθηκε μια προβλεπτική μεθοδολογία συσταδοποίησης δεδομένων η οποία αποσκοπεί στην δημιουργία ομάδων νανοϋλικών με κοινή τοξική απόκριση. Η ομαδοποίηση αυτή, βασίζεται στις ιδιότητες ενός συνόλου δεδομένων νανοϋλικών με γνωστή τοξικότητα. Στη συνέχεια, αν οι ιδιότητες αυτές είναι γνωστές για νανοϋλικά για τα οποία δεν έχει μελετηθεί η τοξική τους συμπεριφορά, καθίσταται δυνατή η πρόβλεψή της, μέσω της αντιστοίχισης των δειγμάτων στις διαμορφωμένες ομάδες. Η μεθοδολογία παρουσιάζεται εκτενώς σε δεδομένα τοξικότητας νανοϋλικών σε βιολογικά συστήματα, χωρίς αυτό να σημαίνει ότι οι εφαρμογές της περιορίζονται μόνο σε παρόμοια δεδομένα. Η προτεινόμενη μεθοδολογία έχει καθολικό χαρακτήρα και δύναται να εφαρμοστεί και σε δεδομένα από άλλους επιστημονικούς τομείς (π.χ. υλικά εν γένει, φάρμακα κ.λπ.).

Η προτεινόμενη μεθοδολογία εναρμονίζεται με τις διεθνείς τάσεις προς αποφυγή των *in vivo* πειραματικών τεχνικών και την ανάπτυξη *in silico* προβλεπτικών μεθόδων οι οποίες βασίζονται στην Επιστήμη των Δεδομένων και τον προγραμματισμό, αξιοποιώντας ήδη διαθέσιμα δεδομένα και συσχετίσεις ιδιοτήτων και τοξικότητας γνωστών υλικών. Υπό αυτό το πρίσμα, ο σκοπός της προτεινόμενης μεθοδολογίας έγκειται στον ορθό και αυτοματοποιημένο χαρακτηρισμό νανοσωματιδίων ως τοξικά ή μη, λαμβάνοντας υπόψιν τη γνωστή συμπεριφορά συγγενών/παρόμοιων δειγμάτων. Πρόκειται δηλαδή για μια κατεξοχήν μεθοδολογία που εντάσσεται στο πλαίσιο των μεθοδολογιών grouping/read-across όπως αυτές περιγράφονται από τους κανονισμούς της Ευρωπαϊκής Ένωσης.

Ως πρώτο βήμα, η μεθοδολογία δημιουργεί σφαιρικές συστάδες (clusters) νανοσωματιδίων με καθορισμένο κέντρο και σύνορα, μέσω τεσσάρων διαδοχικών σταδίων βασισμένων σε λογικούς κανόνες και ενός τελικού σταδίου που βασίζεται στην επίλυση ενός προβλήματος γραμμικού προγραμματισμού. Η διαμόρφωση των συστάδων πραγματοποιείται τοποθετώντας κάθε κατηγοριοποιημένο νανοσωματίδιο στον πολυδιάστατο χώρο βάσει του συνόλου των ιδιοτήτων του, οι οποίες λειτουργούν ως συντεταγμένες. Στη συνέχεια, το μοντέλο προβλέπει την τοξικότητα μη κατηγοριοποιημένων νανοσωματιδίων, κατατάσσοντας τα σε μια εκ των συστάδων μέσω σύνθετων κριτήριων αξιολόγησης των σχετικών αποστάσεών τους από τα κέντρα και τα σύνορα των συστάδων.

Για την αξιολόγηση της μεθοδολογίας και την εκτίμηση της αποτελεσματικότητάς της, χρησιμοποιήθηκαν τέσσερα διαφορετικά σύνολα δεδομένων (datasets) που προέρχονται από τις δημοσιεύσεις των Forest *et al.* (2019), των Kotzabasaki *et al.* (2020), των Liu et *al.* (2013) καθώς και των Papadiamantis *et al.* (2020). Κάθε

σύνολο δεδομένων χαρακτηρίζεται από μια ομάδα ιδιοτήτων, οι οποίες τοποθετούν εκάστοτε δείγμα στο χώρο, καθώς και έναν κατηγορικό χαρακτηρισμό τοξικότητας, ο οποίος αποτελεί και τη μεταβλητή-στόχο. Στο πλαίσιο προεπεξεργασίας του συνόλου δεδομένων και αποσκοπώντας στη βελτίωση της ευρωστίας και τη μείωση του υπολογιστικού χρόνου του μοντέλου, δύναται μεθοδική επιλογή μερικών εκ των ιδιοτήτων προς διαμόρφωση των συστάδων ή η εφαρμογή της Ανάλυσης Κύριων Συνιστωσών (PCA). Η εφαρμογή τους ή μη αξιολογείται βάσει του μεγέθους του συνόλου καθώς και της επίδοσης του μοντέλου με ή χωρίς αυτή.

Μετά από διαδοχικές και ποικίλες μεθόδους επικύρωσης σε τέσσερα σύνολα δεδομένων, η προβλεπτική ικανότητα της προτεινόμενης μεθοδολογίας κρίνεται αρκετά ικανοποιητική. Τα σύνολα των Forest *et al.* (2019), Kotzabasaki *et al.* (2020) και Liu et *al.* (2013) δοκιμάστηκαν ως έχουν αλλά και με εφαρμογή Ανάλυσης Κύριων Συνιστωσών κατά την προεπεξεργασία τους, ενώ το σύνολο των Papadiamantis *et al* (2020) δοκιμάστηκε και με πρότερη επιλογή μεταβλητών. Η αξιολόγηση των προβλέψεων πραγματοποιήθηκε χρήσει των δεικτών ακρίβειας, ευαισθησίας, ειδικότητας και του συντελεστή συσχέτισης Matthews, οι οποίοι υπολογίζονται βάσει του πίνακα σύγχυσης των προβλέψεων. Χαρακτηριστικά αναφέρεται πως σε δοκιμές εξωτερικής αξιολόγησης και χωρίς μείωση των διαστάσεων των συνόλων επετεύχθη ακρίβεια για τα τέσσερα σύνολα δεδομένων $acc_1 = 0.86,\ acc_2 = 1,\ acc_3 = 0.86$ *και* $acc_4 = 0.85$ αντίστοιχα.

Η μέθοδος αναπτύχθηκε και αξιολογήθηκε σε περιβάλλον γλώσσας προγραμματισμού Python αξιοποιώντας τις διαθέσιμες βιβλιοθήκες NumPy, Pandas και scikit-learn. Συμπληρωματικά, για την επίλυση του προβλήματος γραμμικού προγραμματισμού στο πλαίσιο διαμόρφωσης των κατηγοριοποιημένων συστάδων, χρησιμοποιείται το υπολογιστικό πακέτο MIP της Python, δίνοντας πρόσβαση σε ποικιλία επιλυτών. Ως επέκταση της παρούσας διπλωματικής εργασίας, θεωρήθηκε χρήσιμη η διάθεση του μοντέλου προς την επιστημονική κοινότητα και κάθε ενδιαφερόμενο σε μορφή κώδικα Python. Μέσω της δημιουργίας αποθετηρίου στο GitHub, προσφέρεται ελεύθερη πρόσβαση στον κώδικα με σκοπό τη χρήση για πρόβλεψη τοξικότητας ή τη συμπλήρωσή και τη βελτίωση του.

Λέξεις κλειδιά

Νανοπληροφορική, νανοϋλικά, τοξικότητα, προβλεπτικό μοντέλο, μεθοδολογίες SAR και read-across, συσταδοποίηση, κατηγοριοποίηση, γραμμικός προγραμματισμός.

Abstract

In this Thesis, a predictive data clustering methodology has been developed which aims to create groups of nanomaterials with a common toxic response. Grouping is based on the properties of a dataset of nanomaterials with known toxic response. Therefore, if these properties are known for nanomaterials whose toxic behavior has not been tested, their toxicity can be predicted by assigning the samples to the formed clusters. The suggested method is extensively applicable to nanomaterial toxicity data, although its use is not limited to such data only. On the contrary, the proposed method is universal in nature and can be applied to data of different scientific fields, i.e., material science, Medicine etc.

The proposed methodology is in line with international trends to avoid *in vivo* experimental techniques and replace them with *in silico* predictive methods based on Data Science, Machine Learning, and programming, utilizing already available data and correlations between properties and toxicity of known materials. In this light, the purpose of the proposed methodology lies in the correct and automatic characterization of nanoparticles as toxic or non-toxic, considering the known behavior of related or similar samples. Thus, the method itself belongs to the grouping/read-across methodologies described by the regulations of the European Union.

Firstly, the methodology creates spherical clusters of nanoparticles with a defined center and boundaries, through four successive stages based on logical rules and a final stage based on the solution of a linear programming problem. The configuration of the clusters is carried out by placing each categorized nanoparticle in the multidimensional space based on its properties, which serve as coordinates. Then, the model predicts the untested nanoparticles by classifying them into one of the clusters through complex evaluation criteria of their relative distances from the centers and the boundaries of the clusters.

Four different datasets are used to evaluate the methodology and its effectiveness: Forest *et al.* (2019), Kotzabasaki *et al.* (2020), Liu *et al.* (2013) and Papadiamantis *et al.* (2020). Each dataset is characterized by a set of properties, which place the samples in the multidimensional space, as well as a categorical toxicity characterization, which is the target variable. In order to improve the robustness of the model and reduce the computational time needed, variable selection or Principal Component Analysis (PCA) can be applied on data before modeling. The criteria regarding their application are evaluated based on the size of the set as well as the performance of the model if they are applied.

After applying successive and varied validation methods in the four data sets, the predictive power of the proposed methodology is considered quite satisfactory. The sets of Forest *et al.* (2019), Kotzabasaki *et al.* (2020) and Liu *et al.* (2013) were tested as they are, as well as with application of PCA during data pretreatment. The set of Papadiamantis *et al.* (2020) was also tested with variable selection. The

predictive ability is evaluated using indicators like accuracy, sensitivity, specificity, and the Matthews correlation coefficient, which are calculated based on the confusion matrix. During external evaluation tests and without reducing the dimensions of the sets, the accuracy achieved in the data sets is $acc_1 = 0.86$, $acc_2 = 1$, $acc_3 = 0.86$, $acc_4 = 0.85$, respectively.

The method was developed and evaluated in a Python programming language environment using the available NumPy, Pandas and scikit-learn libraries. Additionally, to solve the linear programming problem during the configuration of categorized clusters, the Python MIP package was used, giving access to a variety of solvers. As an extension of this Thesis, it was considered useful to make the model available to the scientific community and anyone interested, in Python code format. By creating a repository on GitHub, free access to the code is offered to be used for predicting toxicity or supplementing and improving it.

Key Words

Nanoinformatics, nanomaterials, toxicity, predictive model, SAR and read-across methods, clustering, categorization, linear programming.

Ευχαριστίες

Η παρούσα Διπλωματική Εργασία με τίτλο «Ανάπτυξη καινοτόμων μεθόδων νανοπληροφορικής για την αυτοματοποιημένη συσταδοποίηση υλικών και την πρόβλεψη ανεπιθύμητων ιδιοτήτων τους» εκπονήθηκε στη Μονάδα Αυτόματης Ρύθμισης και Πληροφορικής του Εθνικού Μετσόβιου Πολυτεχνείου, υπό την επίβλεψη του Καθηγητή ΕΜΠ Χαράλαμπου Σαρίμβει κατά το Ακαδημαϊκό Έτος 2020-21.

Με την ολοκλήρωση αυτής, ολοκληρώνονται και οι σπουδές μου στην Σχολή Χημικών Μηχανικών του ΕΜΠ, υπό πρωτόγνωρες, για όλους, συνθήκες. Παρά τη δυσκολία αυτών, η στήριξη από το ακαδημαϊκό και προσωπικό περιβάλλον ήταν άπλετη και καθοριστική για την πορεία της Εργασίας αυτής και κατ' επέκταση, των σπουδών μου. Για το λόγο αυτό, θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντά μου, καθηγητή Χ. Σαρίμβει για την ανάθεση του συγκεκριμένου θέματος και τις πολύτιμες συμβουλές του καθ' όλη τη διάρκεια της, καθώς και την Δήμητρα Δανάη Βάρσου, υποψήφια διδάκτορα της Μονάδας Αυτόματης Ρύθμισης και Πληροφορικής, χωρίς την καθημερινή καθοδήγηση, προθυμία και υπομονή της οποίας, η παρούσα Εργασία δεν θα ήταν η ίδια. Επίσης, ευχαριστώ τα μέλη της τριμελούς εξεταστικής επιτροπής για το χρόνο που αφιέρωσαν στην μελέτη και αξιολόγησή της. Τέλος, οφείλω ένα μεγάλο ευχαριστώ στην οικογένεια και τους φίλους, που προσέφεραν άπλετα την αγάπη τους και τη στήριξή τους, όλα αυτά τα χρόνια.

Εισαγωγή

Η Επιστήμη της Νανοτεχνολογίας, η οποία βασίζεται στις διαρκώς εξελισσόμενες δυνατότητες των νανουλικών, αναπτύσσεται με ταχείς ρυθμούς και εισέρχεται σε ολοένα και περισσότερα επιστημονικά πεδία, ξεκλειδώνοντας νέες αχαρτογράφητες περιοχές και εφαρμογές. Η δομή των νανοϋλικών τους προσδίδει τεράστια ποικιλομορφία ως προς τη συμπεριφορά και τις ιδιότητές τους, χάρη στα ιδιαίτερα φαινόμενα που λαμβάνουν χώρα στην επιφάνειά τους, καθιστώντας τα ιδανική πρώτη ύλη για πληθώρα εφαρμογών σε τομείς όπως η ιατρική, η βιολογία και η ηλεκτρονική. Νέες μελέτες πραγματοποιούνται διαρκώς με σκοπό την αναζήτηση τρόπων εκμετάλλευσης και τροποποίησης των εγγενών χαρακτηριστικών τους.

Το μεγάλο επιστημονικό ενδιαφέρον για τα νανοϋλικά και τις δυνατότητές τους, έχει επιφέρει και σοβαρούς προβληματισμούς, για τους οποίους η επιστημονική έρευνα είναι ελλιπής και δύσκολη στη διεκπεραίωση, ως προς τις πιθανές αρνητικές επιπτώσεις που μπορεί να έχει η χρήση τους σε ζώντες οργανισμούς και βιολογικά συστήματα. Υπάρχουν σημαντικές ενδείξεις πως τα νανοσωματίδια, όταν έρθουν σε επαφή με τον οργανισμό τροποποιούν τη δομή τους και επηρεάζουν τις φυσικοχημικές συσχετίσεις μεταξύ αυτών και του βιολογικού περιβάλλοντος στο οποίο βρίσκονται, οδηγώντας συχνά σε καταστροφή κυττάρων ή πρόκληση οξειδωτικού στρες. Η συμπεριφορά αυτή, αποκαλούμενη νανοτοξικότητα (nanotoxicity) δημιουργεί άμεση ανάγκη εις βάθος μελέτης των νανουλικών, των ιδιοτήτων τους και των διακυμάνσεων σε αυτές ώστε να μπορέσει να εκτιμηθεί η ασφάλεια έκθεσης του ανθρώπινου οργανισμού και του περιβάλλοντος με τα νανοσωματίδια, χωρίς επιπτώσεις στην υγεία και τη βιωσιμότητα.

Η διύλιση περισσότερων δεδομένων για τη συμπεριφορά των νανοσωματιδίων καθίσταται δυνατή μέσω πειραμάτων και εργαστηριακών δοκιμών. Ωστόσο, καθώς η επιστημονική κοινότητα τείνει να απομακρυνθεί από τεχνικές χρήσης πειραματόζωων θέτοντας σοβαρούς δεοντολογικούς προβληματισμούς, εναλλακτικές λύσεις έχουν βρεθεί μέσω *in vitro* και *in silico* δοκιμών. Οι τελευταίες, μάλιστα, δύναται να αποτελέσουν μια ιδιαίτερα ανταγωνιστική λύση, καθώς μειώνουν σημαντικά το κόστος και το χρόνο διεξαγωγής των δοκιμών. Σε αυτό το πλαίσιο λειτουργεί και ο Ευρωπαϊκός Οργανισμός Χημικών Προϊόντων (European Chemicals Agency, ECHA), ο οποίος, μέσω του κανονισμού REACH (Registration, Evaluation, Authorization and Restriction of Chemicals) διευκολύνει την πρόσβαση σε δεδομένα τοξικότητας προς οποιοδήποτε ενδιαφερόμενο, επιβάλλοντας τη δημιουργία ψηφιακών βιβλιοθηκών στις βιομηχανίες που δραστηριοποιούνται στο χώρο παραγωγής χημικών.

Το κενό πληροφορίας των ιδιοτήτων των νανοϋλικών δύναται να καλύψει ο ταχεία αναπτυσσόμενος τομέας της Νανοπληροφορικής, στο πλαίσιο της οποίας προσφέρονται υπολογιστικές τεχνικές πρόβλεψης της συμπεριφοράς αυτών, όπως τα μοντέλα [Q]SAR (Qualitative/Quantitative Structure-Activity Relationships). Ωστόσο παρά την ικανοποιητική τους επίδοση, απαιτούν μεγάλα σύνολα δεδομένων για να εκπαιδευτούν, γεγονός που καθιστά δύσκολη την ευρεία χρήση τους στον τομέα της Νανοπληροφορικής, όπου τα διαθέσιμα δεδομένα είναι περιορισμένα. Λόγω της πολυπλοκότητας της δομής των νανοσωματιδίων και της ποικιλίας των ρυθμίσιμων ιδιοτήτων τους, αλλά και του όγκου των νανοϋλικών που βρίσκονται ήδη στην αγορά ή σε διαδικασία παραγωγής είναι αδύνατο σε περιορισμένο χρόνο να μετρηθούν όλες οι παράμετροι κάθε δείγματος ώστε να δημιουργηθούν επαρκή δεδομένα. Έναντι των [Q]SAR μοντέλων χρησιμοποιείται συχνά η μεθοδολογία read-across, η οποία παράγει αξιόπιστες προβλέψεις τοξικότητας ακόμα και με μικρό όγκο δεδομένων, βασιζόμενη σε νανοσωματίδια με ήδη γνωστές ιδιότητες και υψηλή φυσικοχημική συγγένεια με τα ζητούμενα δείγματα.

Υπό το πρίσμα των παραπάνω, στην παρούσα Διπλωματική Εργασία αναπτύχθηκε ένα κατηγορικό μοντέλο μαθηματικού προγραμματισμού που αποσκοπεί στην πρόβλεψη της τοξικότητας νανοσωματιδίων. Το μοντέλο αξιοποιεί νανοσωματίδια με γνωστές ιδιότητες και τοξικότητα και τα ομαδοποιεί στον πολυδιάστατο χώρο βάσει της εγγύτητας και της τοξικότητάς τους μέσω μιας διαδικασίας διαδοχικών σταδίων που αποτελούνται από λογικούς κανόνες. Στη συνέχεια, επιστρατεύονται τρία διαφορετικά κριτήρια κατηγοριοποίησης, για την πρόβλεψη της τοξικότητας δειγμάτων με άγνωστο τοξικό χαρακτήρα.

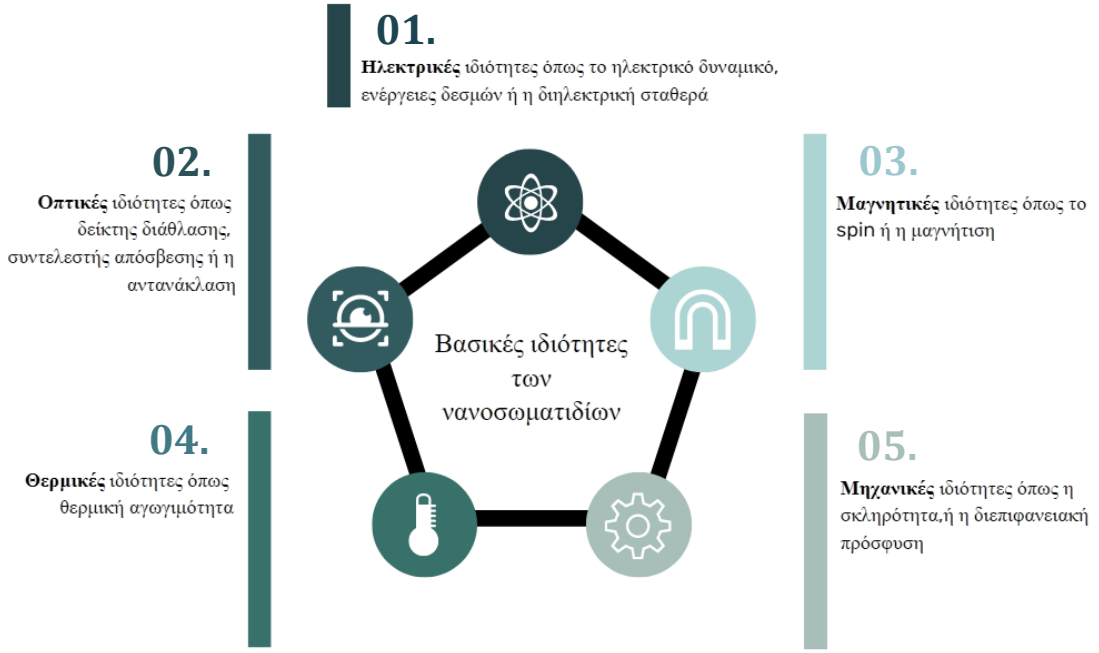
Κεφάλαιο 1
Νανοϋλικά και νανοσωματίδια

Καθώς οι δυνατότητες της τεχνολογίας ολοένα και αυξάνονται και οι επεμβάσεις στις ιδιότητες των συμβατικών υλικών προσεγγίζουν ανεξερεύνητες περιοχές, το πεδίο ανάπτυξης των νανοϋλικών και των νανοσωματιδίων διευρύνεται διαρκώς, προσφέροντας νέες εφαρμογές και λύσεις σε, έως τώρα, άλυτα προβλήματα στο χώρο της μηχανικής, της ιατρικής, της επιστήμης ή της βιομηχανίας.

Τα νανοϋλικά, τα οποία διαθέτουν μια τουλάχιστον διάσταση μικρότερη των 100 nm, λόγω της ιδιαιτερότητάς τους να τροποποιούν τις ιδιότητές τους ανάλογα το σχήμα και το μέγεθός τους, αποτελούν πλέον αντικείμενο πληθώρας μελετών και πειραμάτων [1]. Μέσω αυτών, αναζητούνται τρόποι εκμετάλλευσης των εγγενών χαρακτηριστικών τους αλλά και τροποποίησης αυτών, ώστε να διευρυνθούν οι πρακτικές εφαρμογές τους.

1.1. Ιδιότητες νανοσωματιδίων

Η διαφορά στη συμπεριφορά μεταξύ των υλικών και των αντίστοιχων στη νάνο-κλίμακα οφείλεται σε δύο κύριους παράγοντες: τα επιφανειακά και κβαντικά φαινόμενα. Τα δυο αυτά στοιχεία διαδραματίζουν καθοριστικό ρόλο στην διαμόρφωση της χημικής δραστικότητας (chemical reactivity) και των ξεχωριστών μηχανικών, οπτικών, ηλεκτρικών και μαγνητικών ιδιοτήτων τους [2] οι οποίες βασίζονται στη σύσταση, το μέγεθος και σχήμα, τη σταθερότητα, την



Σχήμα 1.1.1: Σχεδιάγραμμα βασικών ιδιοτήτων των νανοσωματιδίων

τραχύτητα επιφάνειας και την τοπογραφία της [3]. Οι βασικές εξ αυτών παρουσιάζονται στο σχήμα 1.1.1.

1.2. Κατηγοριοποίηση νανοσωματιδίων

Τα νανουλικά μπορούν να κατηγοριοποιηθούν βάσει πληθώρας κριτηρίων με βασικότερο, αυτό της σύστασης. Συγκεκριμένα, εντοπίζονται κατηγορίες νανοσωματιδίων μετάλλου, άνθρακα, ζεόλιθου, κεραμικά, ημιαγωγοί ή πολυμερικά, μεταξύ άλλων, κάθε μια από τις οποίες εμφανίζει ιδιαίτερες ιδιότητες και συμπεριφορές [3].

1.2.1. Νανοσωματίδια μετάλλων

Τα νανοσωματίδια που βασίζονται σε μέταλλα, όπως ο χρυσός, ο λευκόχρυσος, το τιτάνιο, ο σίδηρος ή ο άργυρος, αποκτούν ξεχωριστές οπτικοηλεκτρικές ιδιότητες εξαιτίας της συμπεριφοράς τους ως προς τον εντοπισμό επιφανειακού συντονισμού πλάσματος (localized surface plasmon resonance, LSPR) , αναπτύσσοντας επίσης ευρεία όρια απορρόφησης στην ορατή ζώνη του ηλεκτρομαγνητικού φάσματος [4]. Τα χαρακτηριστικά αυτά, τα καθιστούν ιδανικά στο πεδίο της έρευνας, ιδιαίτερα της ιατρικής και της βιολογίας.

1.2.2. Νανοσωματίδια μεταλλικών οξειδίων

Συνδυάζοντας μόρια μετάλλων με οξυγόνο, σχηματίζονται νανοσωματίδια μεταλλικών οξειδίων όπως TiO_2, Fe_2O_3, Al_2O_3 τα οποία διαφοροποιούνται ως προς τις επιφανειακές ιδιότητές τους, άρα και στην ενέργεια κενού ζώνης. Το χαρακτηριστικό αυτό διευκολύνει τη χρήση τους σε καταλύτες, αισθητήρες ή ημιαγωγούς. Παράλληλα, η αύξηση της ενεργής επιφάνειας αυξάνει σημαντικά τη συμβατότητα με βιολογικούς οργανισμούς [3].

1.3. Εφαρμογές σε βιολογικά συστήματα

Με βάση την πληθώρα των ξεχωριστών χαρακτηριστικών και ιδιοτήτων τους, οι εφαρμογές των νανουλικών διευρύνονται διαρκώς. Σε αυτές περιλαμβάνονται πολλοί διαφορετικοί τομείς, συμπεριλαμβανομένης και της ιατρικής. Στην ιατρική κοινότητα, τα νανοϋλικά χρήζουν ευρείας αποδοχής καθώς, λόγω του μοναδικού τους μεγέθους καθίσταται ιδανικά για χορήγηση φαρμάκων με μεγάλη ακρίβεια ως προς την ποσότητα και το σημείο χορήγησης στον οργανισμό, οδηγώντας σε νέες έρευνες για βιοδιασπώμενα νανοσωματίδια μεταφοράς φαρμακευτικών ουσιών.

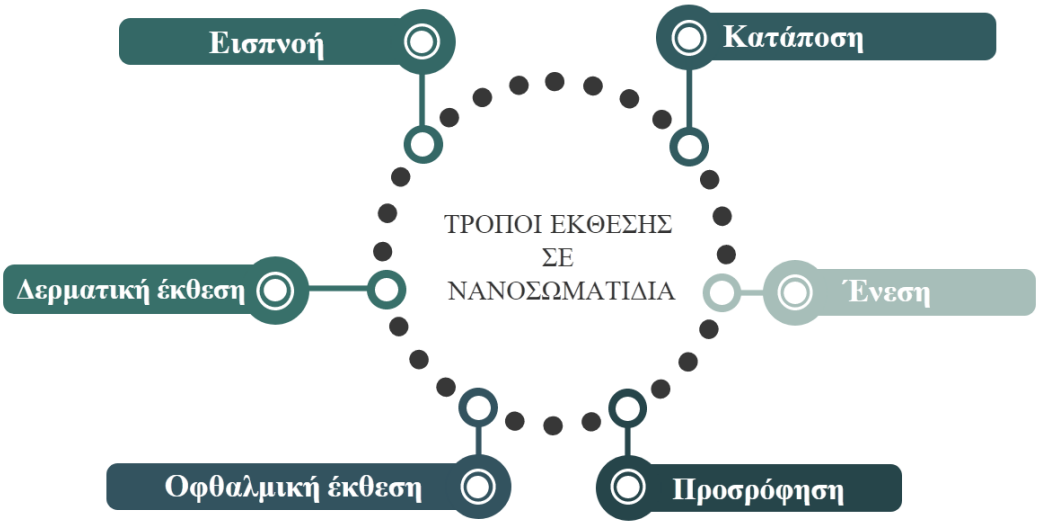
Ταυτόχρονα, οι ημιαγωγοί και τα μεταλλικά νανοσωματίδια αποδεικνύονται χρήσιμα και στην ανίχνευση και θεραπεία του καρκίνου, αξιοποιώντας τις μοναδικές επιφανειακές ιδιότητές τους [4], ενώ τα βιολογικά νανοϋλικά που βασίζονται στο γραφένιο, μελετώνται εις βάθος για τις ικανότητές τους στην

κατασκευή κυτταρικών ιστών. Εκτός της μηχανικής τους δύναμης και σκληρότητας, τα νανοσωματίδια γραφενίου εμφανίζουν ιδιότητες επιτάχυνσης του πολλαπλασιασμού και της διαφοροποίησης μεσεγχυματικών βλαστικών κυττάρων, διατηρώντας τη κυτταρική συμβατότητά τους. Το σύνολο αυτών των χαρακτηριστικών, συνδυαστικά με την υψηλή ηλεκτρική αγωγιμότητα που εμφανίζουν, τα καθιστά ιδανικά για εφαρμογές βιολογικών αισθητήρων που απαιτούν ηλεκτρική διέγερση [5].

1.4. Τοξικότητα νανοσωματιδίων

Εκτός των πλεονεκτημάτων εφαρμογής τους, τα νανοσωματίδια έχουν αρνητικές και συχνά τοξικές συνέπειες στον οργανισμό, σε μεγαλύτερο βαθμό από την ίδια ουσία όταν αυτή δεν διαθέτει διαστάσεις στη μικροκλίμακα [1]. Μάλιστα, θεωρείται πως η τοξικότητα που μπορούν τα νανοσωματίδια να προκαλέσουν στον οργανισμό αυξάνεται αντιστρόφως ανάλογα με το μέγεθος των μορίων που το απαρτίζουν, καθιστώντας συχνά τα νανοσωματίδια ιδιαίτερα επικίνδυνα για τους ανθρώπους και το περιβάλλον. Παράλληλα, η απρόβλεπτη συμπεριφορά τους σε βιολογικά συστήματα δημιουργεί άμεση ανάγκη μελέτης της συμπεριφοράς τους και δημιουργίας μοντέλων πρόβλεψης αυτής, ώστε η χρήση τους να καταστεί ασφαλέστερη.

Συγκεκριμένα, νανοσωματίδια διαμέτρου μικρότερης των 10 nm, εμφανίζουν συμπεριφορά που προσεγγίζει αυτή ενός αερίου, οδηγώντας σε εύκολη είσοδο σε ανθρώπινους ιστούς και δημιουργώντας διαταραχές στο βιολογικό περιβάλλον τους. Κατά την εισπνοή, διαχέονται μέσω της αναπνευστικής οδού και του



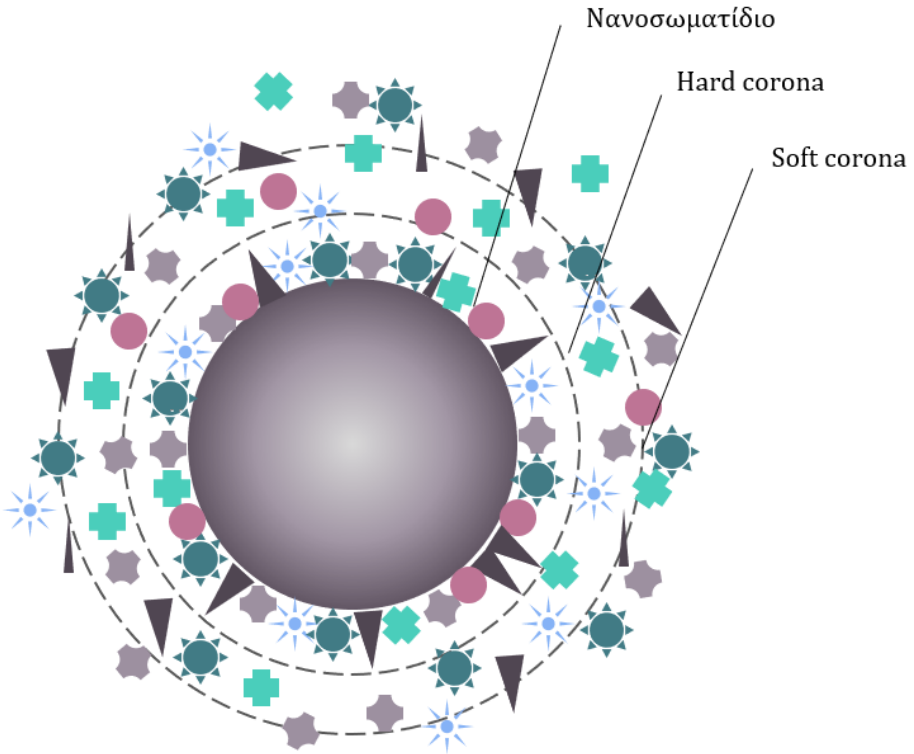
Σχήμα 1.4.1: Σχεδιάγραμμα μεθόδων εισόδου των νανοσωματιδίων στον οργανισμό

οισοφάγου στο συκώτι, την καρδιά, τη σπλήνα και τον εγκέφαλο ενώ συχνά, συσσωρεύονται στο συκώτι [1]. Οι διάφοροι τρόποι με τους οποίους τα σωματίδια μπορούν να εισέλθουν στον ανθρώπινο οργανισμό παρουσιάζονται

στο σχήμα 1.4.1. Κατά την παραμονή τους στον οργανισμό, ανάλογα το σχήμα και τη σύστασή τους, μπορούν να προκαλέσουν σημαντική ζημιά στα κύτταρα μέσω οξειδωτικού στρες ή τραυματισμού οργανιδίων. Για την πρόβλεψη της πιθανής τοξικότητάς τους, έντονες μελέτες πραγματοποιούνται με σκοπό να εντοπιστούν οι φυσικοχημικοί και διαστατικοί παράγοντες που επηρεάζουν την εκδήλωση τοξικής συμπεριφοράς σε βιολογικά περιβάλλοντα.

Παρά την μακροχρόνια θεώρηση πως οι κύριες παράμετροι αντιστοιχούν στο ύψος της δόσης, τη διάσταση και την αντοχή των σωματιδίων, πρόσφατες έρευνες αναθεωρούν και συμπεριλαμβάνουν πληθώρα παραγόντων όπως η μάζα, η συσσώρευση ή οι ιδιότητες της επιφάνειας τους [2]. Σημαντικό ρόλο στην τοξικότητα φαίνεται πως κατέχει το μέγεθος του σωματιδίου. Συγκεκριμένα, μικρά σωματίδια με μεγαλύτερη ειδική επιφάνεια (Specific Surface Area, SSA), ευνοούνται στην αλληλεπίδραση με κυτταρικά στοιχεία όπως πρωτεΐνες, νουκλεϊκά οξέα, λιπίδια ή υδατάνθρακες και εισέρχονται με μεγαλύτερη ευκολία στα κύτταρα.

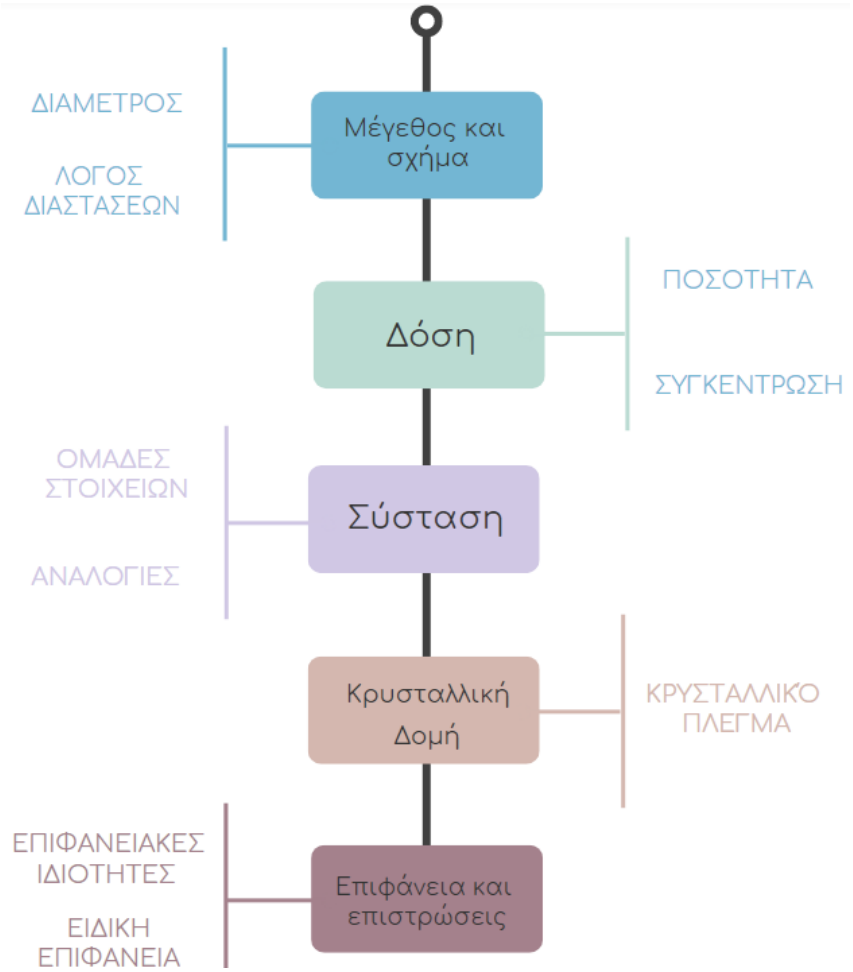
Μάλιστα, η εις βάθος μελέτη του μηχανισμού με τον οποίο τα νανοσωματίδια και οι πρωτεΐνες αλληλοεπιδρούν συμβάλει ουσιαστικά στην πρόβλεψη της τοξικότητας τους. Κατά τη διάρκεια αυτής της αλληλεπίδρασης είναι πιθανή η αλλαγή φάσης, η συσσωμάτωση σωματιδίων ή η αναδόμηση της επιφάνειάς τους, καθορίζοντας την δραστικότητα και την βιοδιαθεσιμότητά τους. Στην επιφάνεια του νανοσωματιδίου και γύρω από αυτό, όπως φαίνεται και στο σχήμα 1.4.2, μπορεί να σχηματιστεί πρωτεϊνικό στέμμα (protein corona ή PC το οποίο



Σχήμα 1.4.2: Απεικόνιση του συσσωματώματος νανοσωματιδίου- πρωτεϊνών και της δημιουργίας πρωτεϊνικού στέμματος

αναμενόμενα αλλάζει την βιολογική ταυτότητα των βίο-μορίων και διαμορφώνει την αντίδραση του ανοσοποιητικού συστήματος στο ξένο σώμα, τοξική ή μη [6]. Ο σχηματισμός του στέμματος χαρακτηρίζεται ως μια δυναμική διαδικασία ανταγωνισμού μεταξύ των πρωτεϊνών για την πρόσδεσή τους στην επιφάνεια των νανοσωματιδίων και την δημιουργία διεπιφάνειας και εξαρτάται τόσο από τα σχετικά μεγέθη των μορίων όσο και από το βιολογικό περιβάλλον στο οποίο αναπτύσσεται (αίμα, κυτταρόπλασμα) [7]. Το πρωτεϊνικό στέμμα περιλαμβάνει δυο στρώματα, το 'μαλακό στέμμα' ή Soft Corona και το 'σκληρό στέμμα' ή Hard Corona. Το στρώμα Soft Corona σχηματίζεται πρώτο και περιλαμβάνει πρωτεΐνες με μεγαλύτερη κινητικότητα οι οποίες είναι αδύναμα προσδεμένες στην επιφάνεια του νανοσωματιδίου. Καθώς η ισορροπία προσρόφησης σταθεροποιείται, προσκολλώνται στην επιφάνεια πρωτεΐνες με μεγαλύτερη χημική συγγένεια, σχηματίζοντας το στρώμα Hard Corona [6].

Πέρα από το ίδιο το μέγεθος, οι σχέση μεταξύ των διαστάσεων των σωματιδίων μπορεί να διαμορφώσει εξίσου την τοξικότητα. Συγκεκριμένα οι ίνες νανουλικών, οι οποίες χαρακτηρίζονται από μεγάλο λόγο διαστάσεων (aspect ratio), άμεσα σχετιζόμενες με την βίο-αντοχή τους (biodurability), υποδεικνύουν ισχυρή τοξική συμπεριφορά [2], [8].

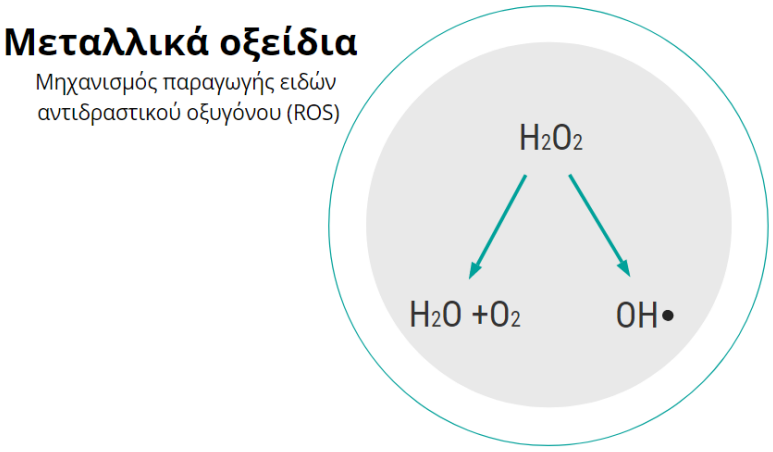


Σχήμα 1.4.3: Σχεδιάγραμμα βασικών παραμέτρων που επηρεάζουν την τοξικότητα των νανοσωματιδίων

Το επιφανειακό φορτίο των σωματιδίων, καθορίζοντας τον τρόπο με τον οποίο έρχονται σε επαφή με βιομόρια και οργανίδια, επηρεάζει σημαντικά την κυτταρική τοξικότητα (cytotoxicity) [8]. Παρά τις εκτεταμένες έρευνες, δεν μπορεί να εξακριβωθεί η ακριβής σχέση ποσότητας – τοξικότητας. Μάλιστα, υπάρχουν ενδείξεις πως, ενάντια στην ενστικτώδη θεώρηση, μικρότερη δόση είναι ικανή να προκαλέσει μεγαλύτερη ζημιά στα κύτταρα και πως ουσιαστικότερο ρόλο έχει η επιφάνεια και όχι η μάζα των νανοσωματιδίων.

Όσον αφορά στην επίδραση που έχει η συγκέντρωση της δόσης των σωματιδίων στην τοξικότητα, αποδεικνύεται πως υψηλότερη συγκέντρωση οδηγεί σε μεγαλύτερη συσσωμάτωση αυτών, δυσκολεύοντας την κίνηση και είσοδό τους στα κύτταρα και την καταστροφή τους. Παράλληλα, συνδυαστικά με τη χημική σύσταση, η κρυσταλλική δομή καθορίζει σε μεγάλο βαθμό τις βασικές φυσικοχημικές ιδιότητες των ατόμων και των μορίων, άρα και την τοξική τους συμπεριφορά. Τέλος, η συμπεριφορά των σωματιδίων μπορεί να τροποποιηθεί από την ύπαρξη η όχι επίστρωσης (coating), η οποία έχει άμεση επίδραση στις φυσικοχημικές ιδιότητές τους. Σε αυτό συμβάλλει και η απρόβλεπτη αλληλεπίδραση που μπορεί να έχει το ίδιο το νανοσωματίδιο με την επίστρωσή του [2]. Συνοπτικά, οι παράγοντες που φαίνεται να επηρεάζουν την τοξικότητα των νανοσωματιδίων παρουσιάζονται στο σχήμα 1.4.3.

Εκτός των φυσικοχημικών και διαστατικών ιδιοτήτων των νανοσωματιδίων που διαμορφώνουν την τοξικότητά τους, κάθε είδος νανοσωματιδίων εμφανίζει ξεχωριστή τοξική συμπεριφορά. Συγκεκριμένα, ενδιαφέρουσα απόκριση επιδεικνύουν τα νανοσωματίδια μεταλλικών οξειδίων, τα οποία είναι ικανά να ενισχύσουν το οξειδωτικό στρες που προκαλούν μέσω της παραγωγής δραστικών μορφών οξυγόνου (reactive oxygen species, ROS), όπως φαίνεται στο σχήμα 1.4.4. Αυτές, «επιτίθενται» στα βιομόρια του οργανισμού και τροποποιούν



Σχήμα 1.4.4: Απεικόνιση της αντιδραστικής συμπεριφοράς των νανοσωματιδίων μεταλλικών οξειδίων προς παραγωγή τοξικών μορίων

τη δομή των μιτοχονδρίων και την αλυσίδα μεταφοράς ηλεκτρονίων. Άμεσα συνέπεια είναι η διαταραχή του κυτταρικού κύκλου και η καταστολή του

πολλαπλασιασμού τους [8]. Καθώς η σχέση αίτιου και αποτελέσματος μεταξύ ιδιοτήτων και τοξικότητας των σωματιδίων γίνεται ολοένα και πιο πολύπλοκη, ικανοποιητικές λύσεις φαίνεται να προσφέρει η θεώρηση της ποσοτικής σχέσης δομής και συμπεριφοράς ή [Q]SAR.

1.5. Ανάγκη πρόβλεψης τοξικότητας

Οι διαρκώς αυξανόμενες εφαρμογές των νανουλικών στην ιατρική, την επιστήμη και την τεχνολογία ενισχύουν την ανάγκη μελέτης της τοξικότητάς τους. Μέχρι πολύ πρόσφατα, οι μελέτες αυτές πραγματοποιούνταν με δοκιμές σε ζώα, *in vivo*. Σταδιακά, οι *in vitro* δοκιμές κατέστησαν προσιτές. Πλέον, δύναται οι τοξικολογικές δοκιμές να πραγματοποιηθούν και *in silico* μέσω υπολογιστικών μοντέλων που αξιοποιούν δεδομένα, μεθοδολογίες, λογισμικά και αλγορίθμους για την οργάνωση, ανάλυση, προσομοίωση και πρόβλεψης της τοξικότητας των υλικών. Οι υπολογιστικές μέθοδοι μπορούν να δράσουν συμπληρωματικά στις *in vivo* και *in vitro* δοκιμές ώστε να ελαχιστοποιηθούν οι δοκιμές σε ζώα, ο πειραματικός χρόνος και το κόστος ή να βελτιωθεί η ακρίβεια των προβλέψεων της τοξικότητας. Παράλληλα, η απουσία πειράματος επιτρέπει την πρόβλεψη τοξικής συμπεριφοράς στο στάδιο του σχεδιασμού (safety-by-design), πριν το χημικό είδος προς εξέταση να συντεθεί, διευκολύνοντας τις επιστημονικές μελέτες [9]. Γίνεται δηλαδή εφικτό να σχεδιάζονται νανοϋλικά εκ των προτέρων ασφαλή που θα χρειάζονται ελάχιστες ή και καθόλου δοκιμές πριν βρουν εφαρμογή στο εμπόριο.

Κεφάλαιο 2
Ανάπτυξη προβλεπτικών μοντέλων

2.1. Μεθοδολογίες πρόβλεψης

Η νανοπληροφορική ως ‘επιστήμη και πρακτική ασχολείται με τον καθορισμό της πληροφορίας που αφορά στη νάνο-κλίμακα καθώς και την ανάπτυξη και υλοποίηση μηχανισμών συλλογής επικύρωσης, αποθήκευσης, διαμοιρασμού, ανάλυσης, μοντελοποίησης και εφαρμογής αυτής της πληροφορίας’ [10]. Με τα εργαλεία που ως επιστήμη έχει αναπτύξει, είναι δυνατή μια υπολογιστική προσέγγιση εκτίμησης της ασφάλειας και της τοξικότητας των νανοσωματιδίων η οποία προσμετρά την πολυπλοκότητά και την ποικιλία τους στη διαμόρφωση της τελικής πρόβλεψης. Στο πεδίο πρόβλεψης της τοξικότητας, η υπολογιστική τοξικολογία (computational toxicology) συνδυάζει την τοξικολογία, τη βιο-στατιστική και την επιστήμη των υπολογιστών για την μοντελοποίηση βιολογικών συστημάτων και την ανάπτυξη προβλεπτικής ικανότητας ως προς τους κινδύνους έκθεσης των οργανισμών σε άγνωστες ουσίες, αναδεικνύοντας τις κυριότερες παραμέτρους οι οποίες ρυθμίζουν την βιολογική δραστηριότητα που οδηγεί σε τοξικότητα [11].

2.1.1. Μεθοδολογία QSAR

Βασική μεθοδολογία στην πρόβλεψη συμπεριφοράς και τοξικότητας των υλικών αποτελεί η QSAR μεθοδολογία, η οποία περιλαμβάνει ένα σύνολο κανόνων και βημάτων πρόβλεψης, βάσει της θεώρησης πως υλικά και σωματίδια με παρόμοια δομή και χαρακτηριστικά θα συμπεριφέρονται με παρόμοια τρόπο ως προς την τοξικότητά τους σε κάποιο βιολογικό περιβάλλον. Αν αυτή η υπόθεση ισχύει, τότε ο κίνδυνος που μπορεί να προκαλεί ένα υλικό μπορεί να προβλεφθεί από ήδη υπάρχοντα τοξικολογικά δεδομένα γνωστών υλικών [12]. Αν η τοξική συμπεριφορά αποτελεί συνάρτηση των ιδιοτήτων των υλικών, η σχέση αυτή μπορεί να εκφραστεί μαθηματικά:

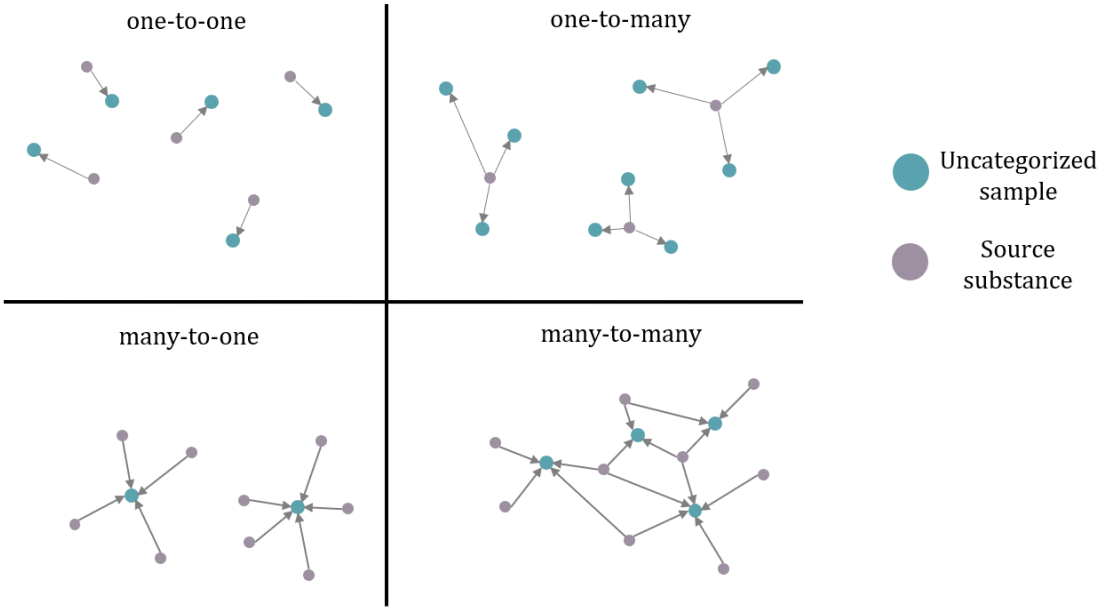
$$toxicity = f(properties)$$

Το σύνολο των ιδιοτήτων που χαρακτηρίζουν τα υλικά και διαμορφώνουν το μοντέλο καλούνται ‘περιγραφείς’ ή descriptors, ενώ η εξαρτημένη μεταβλητή πρόβλεψης, σε αυτή την περίπτωση η τοξικότητα, καλείται endpoint. Εκτός της μοντελοποίησης, βασικό βήμα στις μεθοδολογίες QSAR αποτελεί ο υπολογισμός των περιγραφέων που θα χρησιμοποιηθούν σε αυτή. Σε περίπτωση που ο αριθμός τους είναι δυσανάλογα μεγάλος συγκριτικά με τον αριθμό των δειγμάτων, συχνά ενθαρρύνεται η επιλογή μεταβλητών (feature selection) ώστε να μειωθούν οι διαστάσεις του συνόλου δεδομένων, να μειωθεί ο υπολογιστικός χρόνος και να βελτιωθεί η ευρωστία του μοντέλου [9]. Τα μοντέλα QSAR χρησιμοποιούνται

ευρέως καθώς είναι ικανά να μοντελοποιήσουν όχι μόνο μέσω παλινδρόμησης και ποσοτικοποίησης των σχέσεων ιδιοτήτων και τοξικότητας, αλλά και μέσω κατηγοριοποίησης των υλικών σε τοξικά και μη τοξικά, προσομοιώνοντας την ποιοτική αλληλεπίδραση περιγραφέων και μεταβλητή πρόβλεψης. Στην περίπτωση απουσίας ποσοτικών σχέσεων, το μοντέλο καλείται SAR [9].

2.1.2. Read-across

Η τεχνική read-across αφορά στη διαδικασία κάλυψης κενών απούσας πληροφορίας και μπορεί να εφαρμοστεί τόσο στην προσέγγιση ανάλογων όσο και στη προσέγγιση κατηγοριών. Βάσει του επίσημου ορισμού που αποδίδει ο REACH, η τεχνική χρησιμοποιείται για την πρόβλεψη της μεταβλητής πρόβλεψης μιας ουσίας, αξιοποιώντας πληροφορίες της ίδιας από άλλες ουσίες (source



Σχήμα 2.1.0.1: Απεικόνιση των διαφορετικών πρακτικών read-across που αφορούν στην πρόβλεψη ιδιοτήτων χημικών ουσιών

substances), ακολουθώντας την ίδια θεώρηση με την QSAR μεθοδολογία πως υλικά με παρόμοια δομικά χαρακτηριστικά θα επιδεικνύουν παρόμοια φυσικοχημική και τοξικολογική συμπεριφορά. Πριν την πρόβλεψη του endpoint μιας χημικής ουσίας, ένα σύνολο ουσιών με γνωστά endpoints δημιουργεί ομάδες (groups), βάσει των οποίων η νέα ουσία κατηγοριοποιείται [13].

Διακρίνονται δύο διαφορετικές προσεγγίσεις ως προς τον τρόπο πρόβλεψης των endpoints κατά την εφαρμογή της στρατηγικής read-across. Η προσέγγιση των ανάλογων (analogue approach) αφορά στις χημικές ουσίες 'των οποίων οι έμφυτες φυσικοχημικές, περιβαλλοντικές ή τοξικολογικές ιδιότητες πιθανολογούνται να είναι παρόμοιες με αυτές άλλων ουσιών βάσει κοινών δομικών και φυσικοχημικών ιδιοτήτων'. Η προσέγγιση εφαρμόζεται όταν ο αριθμός των ουσιών είναι περιορισμένος και δεν υπάρχουν εμφανείς ομοιότητες

στις ιδιότητες τους [14]. Αντίθετα, η προσέγγιση των κατηγοριών (category approach), ορίζει ως κατηγορία ένα 'σύνολο χημικών ουσιών των οποίων οι φυσικοχημικές ή περιβαλλοντικές ιδιότητες πιθανολογούνται να είναι παρόμοιες μεταξύ τους βάσει κοινών δομικών ιδιοτήτων' και αφορά σύνολο ουσιών με κοινά χαρακτηριστικά [14].

Η μεθοδολογία read-across μπορεί να εφαρμοστεί μέσω 4 διαφορετικών τεχνικών: ένα προς ένα (one to one), ένα προς πολλά (one to many), πολλά προς ένα (many to one) και πολλά προς πολλά (many to many), όπως παρουσιάζεται και στο σχήμα 2.1.1. Αν η μεταβλητή πρόβλεψης μιας άλλης ουσίας αποτελεί το μοναδικό τρόπο για την πραγματοποίηση πρόβλεψης, τότε εμφανίζονται οι περιπτώσεις ένα προς ένα ή ένα προς πολλά. Αντίθετα, όταν υπάρχει επαρκής όγκος πληροφορίας, περισσότερες άλλες ουσίες μπορούν να αξιοποιηθούν και εφαρμόζεται μια εκ των πολλά προς ένα και πολλά προς πολλά τεχνικών [15].

2.2. Προεπεξεργασία δεδομένων

Πριν την διαμόρφωση οποιουδήποτε προβλεπτικού μοντέλου, συνήθως απαιτείται επεξεργασία των δειγμάτων που αποσκοπεί σε ένα πιο εύκολα διαχειρίσιμο σύνολο δεδομένων. Δείγματα με ακραίες (outliers) ή λιγότερες από τις απαιτούμενες τιμές, σύνολα δεδομένων με προβληματικά δείγματα ή δείγματα που εμφανίζονται πολλαπλές φορές είναι πιθανό να αποπροσανατολίζουν τη μεθοδολογία διαμόρφωσης του προβλεπτικού μοντέλου και να δημιουργούν δυσκολίες στην εκπαίδευση του και την αποκωδικοποίηση του συνόλου δεδομένων [16].

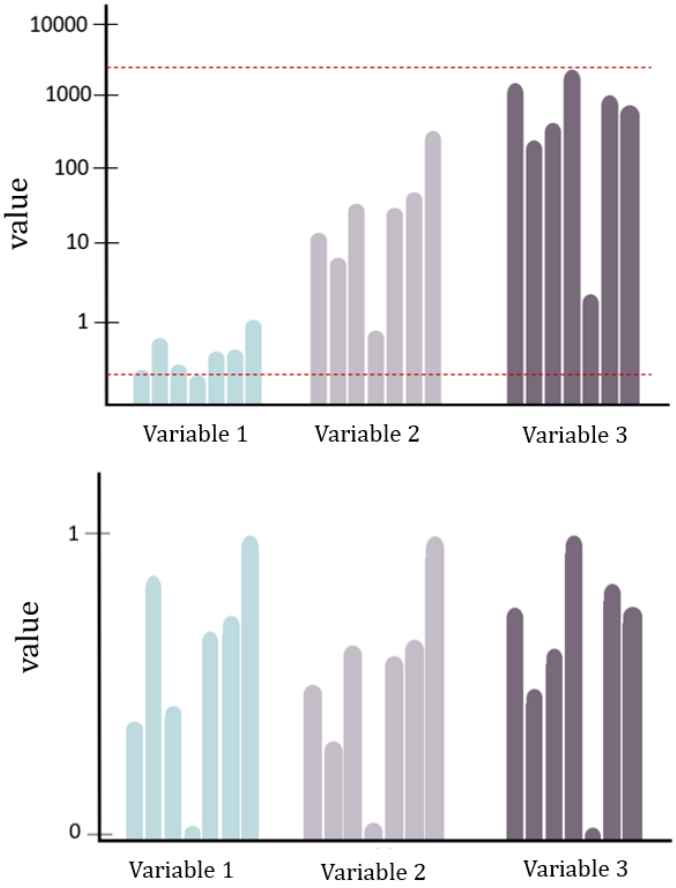
Πρώτο στάδιο κατά την προεπεξεργασία αποτελεί η απομάκρυνση μη αποδεκτών δειγμάτων που δεν πληρούν τα παραπάνω κριτήρια, ώστε να μειωθεί ο πιθανός θόρυβος (noise) ή οι αναξιόπιστες τιμές. Στην συνέχεια, ακολουθεί η κανονικοποίηση των τιμών (normalization) και η εξαγωγή των χαρακτηριστικών των συνόλων (feature extraction), όπως περιγράφονται παρακάτω.

2.2.1. Κανονικοποίηση δεδομένων

Οι στατιστικές μέθοδοι στις οποίες συμμετέχουν πολλές μεταβλητές (multivariate) λειτουργούν αποδοτικότερα όταν όλες οι μεταβλητές χρησιμοποιούν την ίδια κλίμακα μεγεθών ή είναι κανονικοποιημένες. Η απαίτηση αυτή αποσκοπεί στην μετρίαση του σχετικού βάρους (weight) κάθε μεταβλητής ώστε όλες να συμμετέχουν εξίσου στη διαμόρφωση του τελικού μοντέλου. Για την κανονικοποίηση, στο πλαίσιο αυτής της μελέτης χρησιμοποιήθηκε η μετατροπή:

$$f_{i,scaled} = \frac{f_i - f_{min}}{f_{max} - f_{min}} \qquad (2.2.1)$$

όπου $f_{i,scaled}, f_i$ οι κανονικοποιημένες και μη κανονικοποιημένες τιμές, αντίστοιχα, της μεταβλητής f του δείγματος i και f_{min}, f_{max} οι ελάχιστες και μέγιστες τιμές αυτής. Με αυτή την αριθμητική τροποποίηση, οι σχετικές διαφορές μεταξύ των μεταβλητών παραμένουν σταθερές, χωρίς να αλλοιώνονται οι αλληλεπιδράσεις τους, οι συμμετρίες και οι κατανομές τους [17], αν και το εύρος των τιμών περιορίζεται σε τιμές μεταξύ 0 και 1. Διαγραμματική απεικόνιση της κανονικοποίησης παρουσιάζεται στο σχήμα 2.2.1.



Σχήμα 2.2.1: Απεικόνιση της κανονικοποίησης συνόλου δεδομένων τριών μεταβλητών και επτά δειγμάτων

2.2.2. Εξαγωγή χαρακτηριστικών

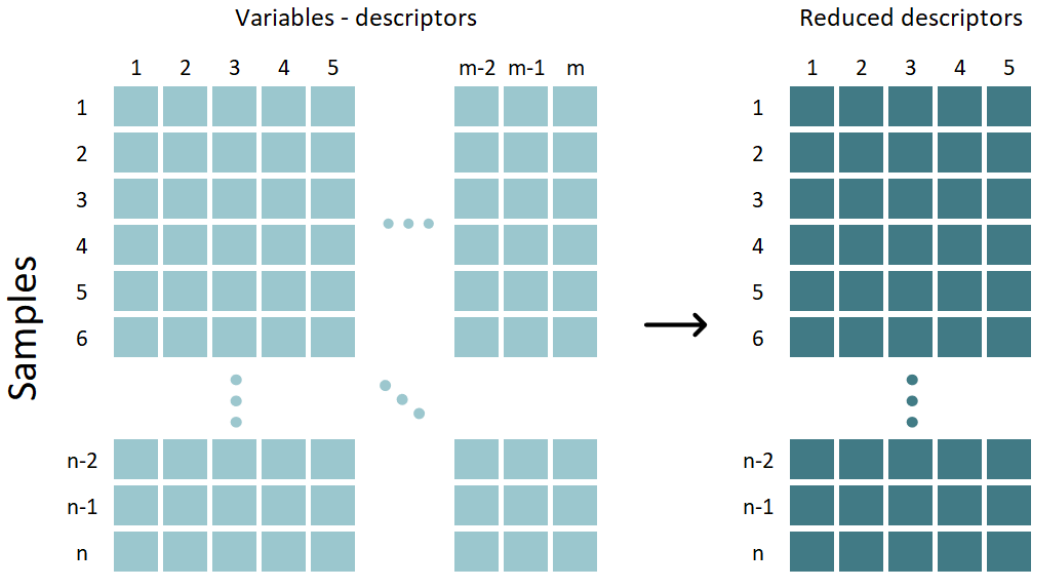
Σε μερικά σύνολα δεδομένων, είναι πιθανό πολλές μεταβλητές να περισσεύουν, μη προσφέροντας χρήσιμη πληροφορία για την πρόβλεψη της τάξης του εκάστοτε δείγματος ή να σχετίζονται έντονα μεταξύ τους, καθιστώντας κάποιες από αυτές πλεονάζουσες. Η μείωση των μεταβλητών που χρησιμοποιούνται βάσει του συνόλου των διαθέσιμων βασίζεται στον εντοπισμό τέτοιων μοτίβων και συμπεριφορών, απομακρύνοντας τη μη σχετική πληροφορία, αυξάνοντας την απόδοση του μοντέλου και μειώνοντας το υπολογιστικό κόστος. Μάλιστα, παρά την απομάκρυνση δεδομένων από το μοντέλο, η αναδιαμόρφωση του συνόλου δεδομένων μπορεί να οδηγήσει σε μεγαλύτερη ακρίβεια [18].

Στην παρούσα μελέτη, πέραν της απλής απομάκρυνσης κάποιων μεταβλητών, εξετάζεται η σύνθεση νέων και η αντικατάσταση των παλαιών μέσω μαθηματικών και λογικών συνδυασμών ώστε να εξασφαλιστεί η βέλτιστη χρήση της διαθέσιμης πληροφορίας.

2.2.3. Ανάλυση κύριων συνιστωσών

Η αναδιαμόρφωση του συνόλου δεδομένων πραγματοποιήθηκε βάσει της μεθόδου ανάλυσης των κύριων συνιστωσών (Principal Component Analysis, PCA). Η PCA βασίζεται στον συνδυασμό των αρχικών ιδιοτήτων και την κατάλληλη τροποποίηση τους ώστε να περιλαμβάνει όσο το δυνατόν περισσότερη αρχική πληροφορία σε μικρότερο σύνολο δεδομένων. Οι νέες διαμορφωμένες μεταβλητές καλούνται κύριες συνιστώσες (Principal Components, PCs) και θα μπορούσαν να περιγραφούν ως περίληψη των αρχικών μεταβλητών [19].

Συγκεκριμένα, οι κύριες συνιστώσες είναι γραμμικός συνδυασμός των αρχικών μεταβλητών, ορθογωνικές και κανονικοποιημένες. Η μεθοδολογία χτίζει τις PCs με τέτοιο τρόπο ώστε η πρώτη συνιστώσα να αντιστοιχεί στη μέγιστη ολική διακύμανση στον πίνακα των μεταβλητών. Οι επόμενες συνιστώσες αντιστοιχίζονται με τις υπόλοιπες διακυμάνσεις, κατά φθίνουσα σειρά, διατηρώντας την ανεξαρτησία τους από τις προηγούμενες.



Σχήμα 2.2.2: Απεικόνιση ης διαδικασίας μείωσης μεταβλητών σε σύνολο δεδομένων με m αρχικές μεταβλητές και n δείγματα

Στο σύνολό τους οι κύριες συνιστώσες PCs θα είναι σε αριθμό λιγότερες ή ίσες με τις αρχικές, όπως φαίνεται στο σχήμα 2.2.2. Σε αυτή την περίπτωση, η PCA μπορεί να χαρακτηριστεί και ως μείωση διαστάσεων (dimensionality reduction). Η μείωση αποσκοπεί στην ευκολότερη ανάλυση των δεδομένων από το

προτεινόμενο μοντέλο, χωρίς να χαθεί πολύτιμη πληροφορία. Η ορθή επιλογή του αριθμού των τελικών PCs αποτελεί σημαντική διαδικασία, λαμβάνοντας υπόψιν πως όσο το k πλησιάζει το 1, τόσο απλούστερο το σύνολο δεδομένων, ενώ όσο πλησιάζει το m, τόσο περισσότερη πληροφορία διατηρείται.

Η μεθοδολογία χρησιμοποιείται συχνά σε σύνολα δεδομένων με μεγάλο αριθμό μεταβλητών συγκριτικά με τα δείγματα, ή σε περιπτώσεις δυσνόητων σχέσεων μεταξύ των ανεξάρτητων μεταβλητών, για την άμβλυνση της πολυπλοκότητας των προβλημάτων. Για την εφαρμογή της, έγινε χρήση της βιβλιοθήκης της Python 'sklearn.decomposition' ή οποία περιλαμβάνει τη μεθοδολογία PCA στα εργαλεία της, αφού προηγουμένως είχε προηγηθεί κανονικοποίηση των δεδομένων ώστε να εξασφαλιστεί παρόμοια βαρύτητα στο τελικό σύνολο μεταβλητών για κάθε μια εξ αυτών.

Κεφάλαιο 3
Αξιολόγηση παραγόμενων μοντέλων

Η αξιολόγηση ενός μοντέλου ως προς την ακρίβειά του είναι σημαντική τόσο για την επίγνωση της ικανότητάς του να κατηγοριοποιεί ορθά μελλοντικά δείγματα όσο και για την σύγκριση μεταξύ διαθέσιμων προβλεπτικών μοντέλων και λήψη αποφάσεων για το βέλτιστο εξ αυτών [20].

Ιδανικά, μια μέθοδος αξιολόγησης πρέπει να χαρακτηρίζεται από χαμηλή μεροληψία και διακύμανση (bias και variance, αντίστοιχα). Ωστόσο, συχνά οι δύο ιδιότητες δρουν ανταγωνιστικά μεταξύ τους καθώς, όσο περισσότερο το μοντέλο προσαρμόζεται στα δεδομένα βάσει των οποίων διαμορφώνεται, τόσο μειώνεται η ικανότητά του να επεξεργάζεται ελαφρώς διαφοροποιημένα δείγματα οδηγώντας σε υπερβολική προσαρμογή (overfitting).

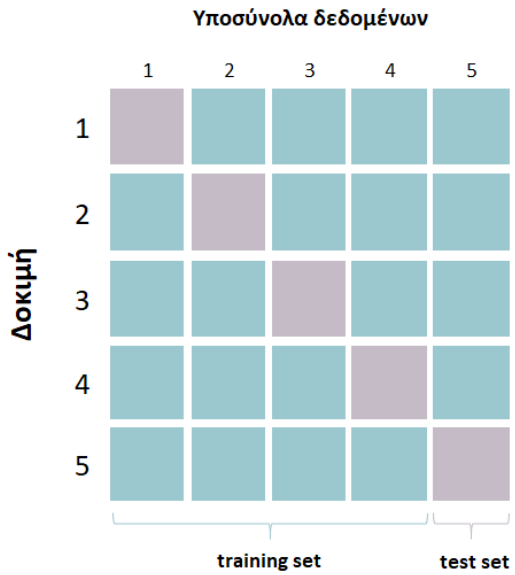
Για την εξασφάλιση της ορθής αξιολόγησης της μεθοδολογίας που αναπτύχθηκε, αξιοποιήθηκαν διαφορετικές μέθοδοι επικύρωσης, όπως αυτές παρουσιάζονται παρακάτω. Βασική αρχή όλων αποτελεί ο διαχωρισμός του σετ των δεδομένων σε δύο πλήρως ανεξάρτητα σύνολα: το σύνολο εκπαίδευσης (training set) και το σύνολο δοκιμών (test set). Το πρώτο εξ αυτών αξιοποιείται από το προβλεπτικό μοντέλο για να διαμορφώσει τις παραμέτρους του, ενώ το δεύτερο αξιοποιείται για την αξιολόγηση τα προβλεπτικής ικανότητάς. Καθώς αυτή η διαδικασία, περιορίζει τα διαθέσιμα δείγματα που μπορούν να χρησιμοποιηθούν σε κάθε στάδιο, ιδιαίτερα σε μικρά σύνολα δεδομένων, έχουν αναπτυχθεί πιο σύνθετες διαδικασίες επικύρωσης, κάποιες εκ των οποίων αξιοποιήθηκαν και στην παρούσα μελέτη [21].

3.1. Εσωτερική αξιολόγηση

Οι μέθοδοι εσωτερικής αξιολόγησης (internal validation) εκτιμούν το μέτρο στο οποίο οι προβλέψεις του μοντέλου για τα δείγματα προς ανάλυση ή παρόμοια αυτών, αντιπροσωπεύουν τις πραγματικές τιμές τους [22].

3.1.1. K-Fold Cross Validation

Η μέθοδος της διασταυρούμενης επικύρωσης (cross validation) χρησιμοποιείται ως επαναληπτική διαδικασία αξιολόγησης ενός προτεινόμενου μοντέλου, ώστε να μελετηθεί εις βάθος η αξιοπιστία του. Το σύνολο δειγμάτων χωρίζεται σε k υποσύνολα (k-fold) ίδιου μεγέθους με τυχαίο τρόπο, όπως φαίνεται στο σχήμα 3.1.1. Ο αλγόριθμος επιλέγει όλα τα υποσύνολα, πλην ενός, και τα αξιοποιεί για την εκπαίδευση του, θεωρώντας το νέο σύνολο ως σύνολο εκπαίδευσης. Το



Σχήμα 3.1.1: Απεικόνιση της διαδικασίας χωρισμού του δειγματικού χώρου κατά την διασταυρούμενη επικύρωση με αριθμό χωρισμών n=5

υποσύνολο το οποίο δεν έχει αξιοποιηθεί για την διαμόρφωση του μοντέλου αποτελεί το σύνολο ελέγχου. Η διαδικασία επαναλαμβάνεται k φορές, με διαφορετικό σύνολο ελέγχου κάθε φορά. Από τα στατιστικά επιτυχίας και σφάλματος κάθε επανάληψης, δύναται να προκύψει ένα συνολικό σφάλμα για το συγκεκριμένο μοντέλο, μέσω εσωτερικής αξιολόγησης [21].

Σημειώνεται πως η γλώσσα προγραμματισμού Python, στην οποία αναπτύχθηκε η προτεινόμενη μεθοδολογία, εμπεριέχει, στην βιβλιοθήκη εργαλείων scikit-learn.model_selection, την αυτοματοποιημένη επαναληπτική διαδικασία K-Fold [23], η οποία και αξιοποιήθηκε στην μελέτη.

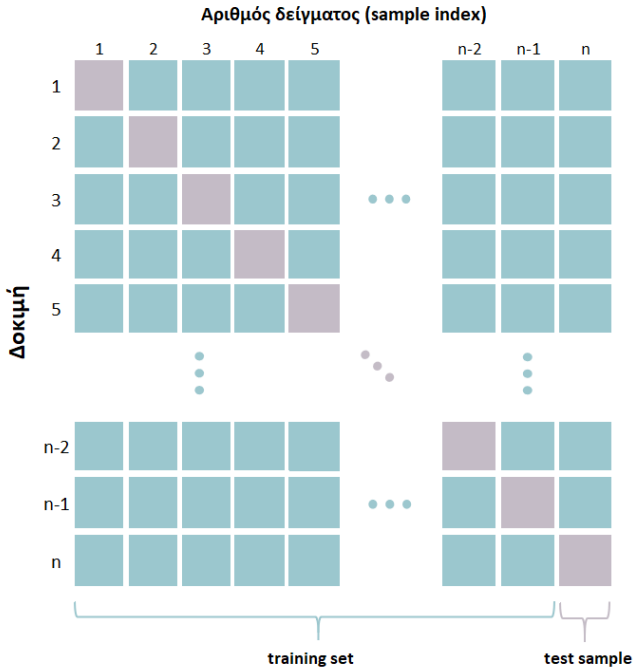
3.1.2. Leave-one-out Cross Validation

Η μέθοδος διασταυρούμενης επικύρωσης Leave-One-Out (LOO), η οποία μπορεί να θεωρηθεί υποπερίπτωση της k-fold Cross Validation, ακολουθεί την ίδια διαδικασία με αυτή, επαναλαμβάνοντας τον έλεγχο n φορές, όπου n ο αριθμός των δειγμάτων στο σύνολο δεδομένων. Μια ενδεικτική απεικόνιση της επικύρωσης LOO παρουσιάζεται στο σχήμα 3.1.2. Συγκεκριμένα, σε κάθε επανάληψη (n στο σύνολο) ο αλγόριθμος εκπαιδεύεται στο σύνολο των δειγμάτων πλην ενός, βάσει του ο οποίου πραγματοποιείται, στη συνέχεια, η πρόβλεψη και η αξιολόγηση του μοντέλου. Στο τέλος της διαδικασίας, προκύπτει το συνολικό σφάλμα βάσει των σωστών προβλέψεων εκ των συνολικών [21].

Η μεθοδολογία επικύρωσης LOO χρησιμοποιείται ιδιαίτερα σε μικρά σύνολα δεδομένων καθώς επιτρέπει την μέγιστη δυνατή εκμετάλλευση κάθε δείγματος επιτυγχάνοντας μεγάλη ακρίβεια. Φυσικά, μια τόσο εκτενής επαναληπτική διαδικασία θα είχε τεράστιες απαιτήσεις σε χρόνο και υπολογιστική ισχύ στην

περίπτωση ενός εκτενούς συνόλου δεδομένων, καθιστώντας την μη εφαρμόσιμη. Παράλληλα, και σε αντίθεση με την k-fold cross validation, η διαδικασία είναι πλήρως ντετερμινιστική και δεν εμπεριέχει στοιχεία τυχαιότητας, καθώς κάθε δυνατός συνδυασμός δειγμάτων έχει ληφθεί υπόψιν στην αξιολόγηση.

Στην ίδια βιβλιοθήκη scikit-learn.model_selection της Python, εμπεριέχεται και η μέθοδος Leave-One-Out, η οποία εφαρμόστηκε εξίσου, λόγω του μικρού μεγέθους των συνόλων δεδομένων.



Σχήμα 3.1.2: Απεικόνιση της διαδικασίας χωρισμού του δειγματικού χώρου κατά την διασταυρούμενη επικύρωση Leave-one-out

3.2. Εξωτερική αξιολόγηση

Ολοκληρώνοντας την εσωτερική αξιολόγηση, δύναται η δυνατότητα περαιτέρω επικύρωσης του μοντέλου μέσω εξωτερικής αξιολόγησης (external validation). Αυτή εκτιμά το βαθμό στον οποίο το διαμορφωμένο μοντέλο είναι εξίσου εφαρμόσιμο σε ένα διαφορετικό σύνολο δεδομένων [22] , το οποίο περιλαμβάνει ανεξάρτητα δείγματα ‘εύλογα συγγενούς’ πληθυσμού [24].

Ακολουθώντας την ίδια λογική διαχωρισμού με την εσωτερική επικύρωση, το σύνολο δεδομένων χωρίζεται σε δύο υποσύνολα, το σύνολο εκπαίδευσης και ελέγχουtraining και test set, αντίστοιχα. Ο χωρισμός μπορεί να γίνει με τυχαίο τρόπο (random sampling) ή ακολουθώντας μια λογική συσχέτισης του δειγματικού χώρου. Αν και η τυχαία επιλογή έχει ευκολότερη και ταχύτερη εφαρμογή, στο πεδίο των νευρωνικών δικτύων (Neural Networks) και της μηχανικής εκμάθησης (Machine Learning), προτιμάται η εφαρμογή μια δομημένης μεθοδολογίας διαχωρισμού του συνόλου δεδομένων. Μια τέτοια

μεθοδολογία μπορεί να εξασφαλίσει πως το σύνολο εκπαίδευσης θα περιλαμβάνει όλο το εύρος του δειγματικού χώρου, αποφεύγοντας προβλέψεις για δείγματα εκτός του χώρου στον οποίο το μοντέλο έχει εκπαιδευτεί [25].

Αλγόριθμος Kennard Stone

Μεταξύ των διαθέσιμων μεθόδων αυτής της λογικής, όπως ο γενετικός κώδικας (Genetic Algorithm, GA) ή ο αυτό-οργανωμένος χάρτης (Self Organizing Maps, SOMs) επιλέγεται προς εφαρμογή η μέθοδος Kennard Stone (KS), που αποτελεί την πλέον εφαρμοσμένη μεθοδολογία στο πεδίο [26].

Ο αλγόριθμος των Kennard και Stone βασίζεται στην επιλογή ενός αντιπροσωπευτικού τμήματος του δειγματικού χώρου ακολουθώντας βηματική διαδικασία διαχωρισμού. Συγκεκριμένα, ξεκινώντας από το ζεύγος δειγμάτων με την μεγαλύτερη ευκλείδεια απόσταση στο δειγματικό χώρο f διαστάσεων (ή μεταβλητών), επιλέγει με διαδοχικό τρόπο νέα δείγματα που απέχουν τις μέγιστες δυνατές αποστάσεις από τα ήδη επιλεγμένα δείγματα, έως ότου ο προκαθορισμένος αριθμός δειγμάτων του συνόλου εκπαίδευσης έχει επιλεχθεί. Η απόσταση στον f-διάστατο χώρο μεταξύ των δειγμάτων i και j υπολογίζεται ως:

$$d_{ij} = \sqrt{\sum_{v=1}^{f} \left(x_{iv} - x_{jv}\right)^2} \qquad (3.2.1)$$

όπου x το διάνυσμα των ιδιοτήτων-μεταβλητών κάθε δείγματος [27] .

3.3. Στατιστικά εργαλεία

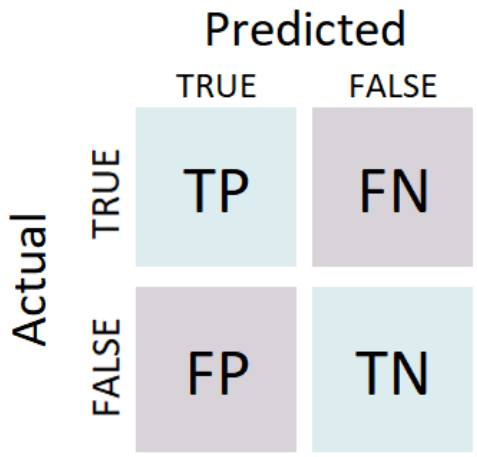
Για την καλύτερη αξιολόγηση του προτεινόμενου μοντέλου επιστρατεύονται μερικά ακόμη στατιστικά μεγέθη που ποσοτικοποιούν την αποτελεσματικότητά του, βάσει των μεθόδων επικύρωσης που αξιοποιούνται. Τα μεγέθη αυτά αποτελούν βασικά εργαλεία στην αξιολόγηση προβλεπτικών μοντέλων κατηγοριοποίησης.

3.3.1. Πίνακας σύγχυσης

Στην περίπτωση που η ζητούμενη πρόβλεψη σε ένα μοντέλο είναι δυαδικής φύσης, δηλαδή μπορεί να πάρει μόνο δύο τιμές, μαθηματικές ή λογικές, εμφανίζονται τέσσερις διαφορετικές πιθανότητες. Αν η μια τιμή-κλάση λάβει τη λογική έκφραση ορθή (True) και η δεύτερη λάβει την έκφραση ψευδής (False), οι σωστές προβλέψεις θα αντιστοιχούν στα μεγέθη True Positives (TP) και True Negatives (TN), αντίστοιχα. Αν σε ένα δείγμα λανθασμένα αποδοθεί η τιμή True, το δείγμα χαρακτηρίζεται ως False Positive, ενώ στην αντίθετη περίπτωση, το δείγμα χαρακτηρίζεται ως False Negative [21]. Οι αριθμοί των δειγμάτων από το training set που αντιστοιχούν στις τέσσερις διαφορετικές περιπτώσεις,

διαμορφώνουν τον πίνακα σύγχυσης (confusion matrix) του μοντέλου που αποτελεί βασικό εργαλείο οπτικοποίησης των επιδόσεών του (όπως στο σχήμα 3.3.1).

Αν και ο υπολογισμός των τεσσάρων μεγεθών κρίνεται σχετικά απλός, στο πλαίσιο της παρούσας μελέτης χρησιμοποιήθηκε η βιβλιοθήκη scikit-learn.metrics και το εργαλείο confusion matrix, το οποίο, δεδομένου των πραγματικών τιμών και των προβλέψεων, δημιουργεί αυτόματα τον πίνακα σύγχυσης στην επιθυμητή μορφή [28].



Σχήμα 3.3.1: Απεικόνιση πίνακα σύγχυσης και των στοιχείων του

3.3.2. Ακρίβεια και βαθμός σφάλματος

Οι πλέον ενδεικτικές τιμές αξιολόγησης της επίδοσης ενός προβλεπτικού μοντέλου είναι η ακρίβεια (accuracy) και ο βαθμός σφάλματος (error rate). Τα μεγέθη αντιστοιχούν στα κλάσματα των σωστών και εσφαλμένων προβλέψεων έναντι των συνολικών, αντίστοιχα. Αν μια πρόβλεψη είναι σωστή, συμβάλει στην αύξηση της ακρίβειας, ενώ αν είναι εσφαλμένη μεγεθύνει το σφάλμα [21].

Για τον υπολογισμό τους χρησιμοποιούνται οι τιμές των TP, TN, FP και FN:

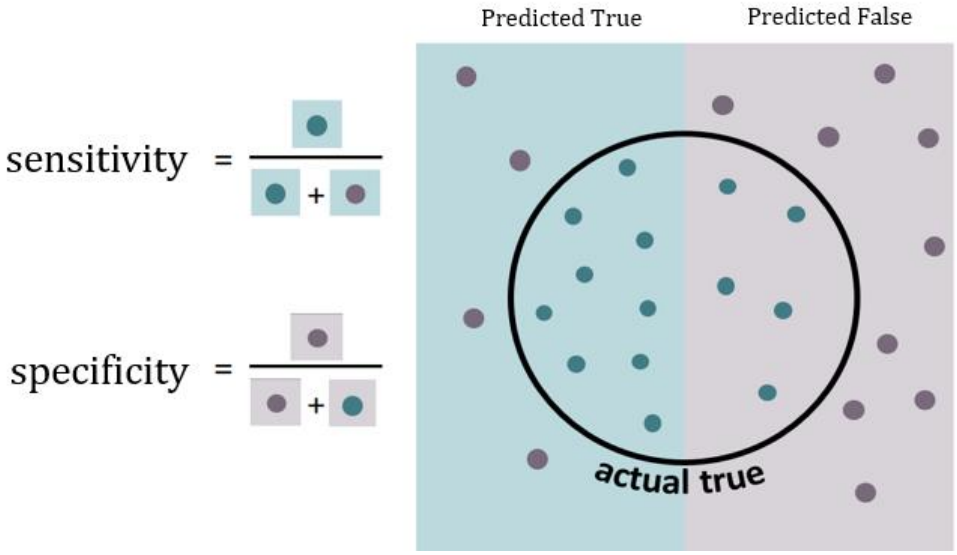
$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.3.1}$$

$$error\ rate = \frac{FP + FN}{TP + TN + FP + FN} = 1 - accuracy \tag{3.3.2}$$

Αν και αποτελούν βασικό χαρακτηριστικό ενός μοντέλου, οι δύο δείκτες εκτιμούν μόνο μερικώς την ικανότητα πρόβλεψης που κατέχει και δεν εμφανίζουν επαρκή ευαισθησία απέναντι σε δείγματα της μη κυρίαρχης ομάδας. Συμπληρωματικά, επιστρατεύονται μερικοί ακόμα στατιστικοί δείκτες για την διαμόρφωση μιας

ολοκληρωμένης εικόνας της επίδοσης του προτεινόμενου μοντέλου σε κάθε σύνολο δεδομένων [29].

3.3.3. Ειδικότητα και ευαισθησία



Σχήμα 3.3.2: Απεικόνιση του δειγματικού χώρου ως προς τις πραγματικές και προβλεπόμενες τιμές και ορισμός της ειδικότητας και της ευαισθησίας

Η ευαισθησία (sensitivity) ενός μοντέλου διαμορφώνεται βάσει του αριθμού των δειγμάτων που διαγνώσθηκαν ορθώς ως True έναντι όλων όσων είναι True, ανεξαρτήτως της πρόβλεψής τους. Αντίστοιχα, η ειδικότητα (specificity) ορίζεται ως το πηλίκο του αριθμού των δειγμάτων που διαγνώσθηκαν ορθώς ως False έναντι όλων όσων είναι στην πραγματικότητα False (βλέπε Σχήμα 3.3.2) [21]. Χρησιμοποιώντας τις τιμές που ορίστηκαν στον πίνακα σύγχυσης, οι δύο τιμές υπολογίζονται ως:

$$sensitivity = \frac{TP}{TP + FN} \tag{3.3.3}$$

$$specificity = \frac{TN}{TN + FP} \tag{3.3.4}$$

3.3.4. Συντελεστής συσχέτισης Matthews

Για την καλύτερη δυνατή συσχέτιση μεταξύ μεταβλητών και πρόβλεψης, σε κάθε δοκιμή υπολογίστηκε και ο συντελεστής συσχέτισης Matthews (Matthews Correlation Coefficient), ο οποίος, βάσει των μεγεθών του πίνακα σύγχυσης ορίζεται ως:

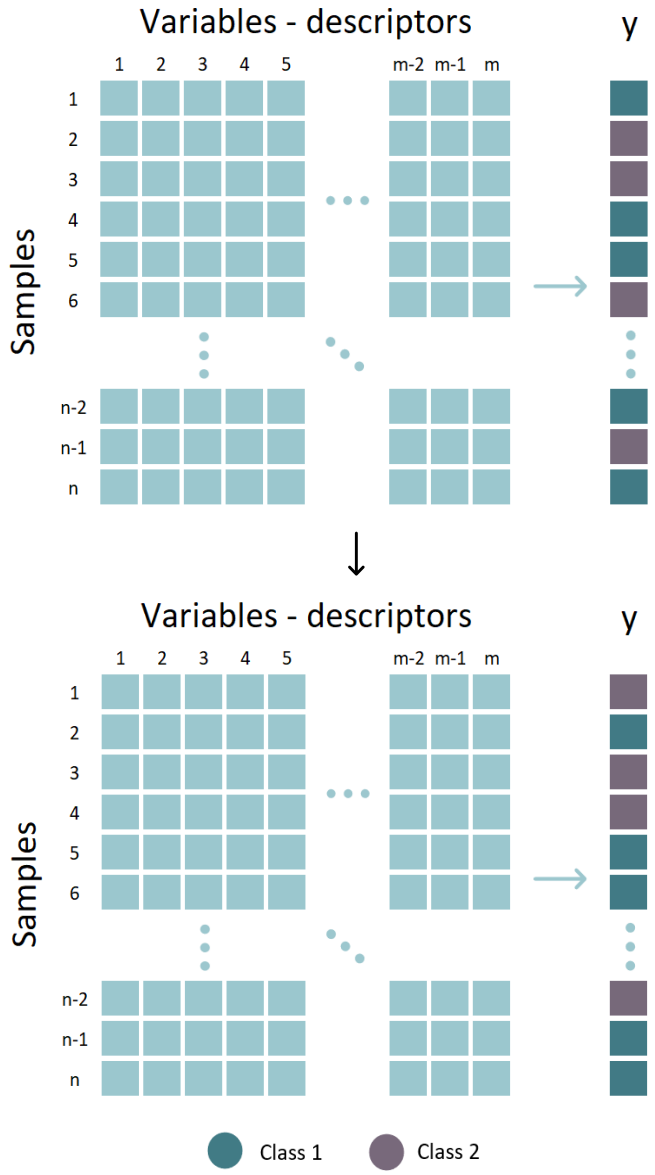
$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.3.5}$$

Ο συντελεστής Matthews εκτιμά άμεσα την επίδοση του μοντέλου σε σχέση με μια καθαρά τυχαία πρόβλεψη. Τιμές κοντά στο 1 υποδηλώνουν απόλυτα πετυχημένη μοντελοποίηση, ενώ, όσο η τιμή μικραίνει, η πρόβλεψη πλησιάζει την επίδοση ενός τυχαίου μοντέλου. Αντίστοιχα, τιμές πλησίον του -1, αντιστοιχούν σε πλήρως αποτυχημένο προβλεπτικό μοντέλο [30].

Στο πλαίσιο της παρούσας μελέτης, για τον υπολογισμό του συντελεστή αξιοποιήθηκε η βιβλιοθήκη scikit-learn.metrics και το εργαλείο matthews_corrcoef [31].

3.4. Έλεγχος τυχαίας επιλογής

Για την αξιολόγηση ενός μοντέλου, πέρα από την ικανοποιητική προβλεπτική ικανότητα σε άγνωστα δείγματα, απαιτείται και ο έλεγχος του ρόλου της τυχαίας



Σχήμα 3.4.1: Απεικόνιση της αναδιαμόρφωσης του διανύσματος y με τυχαίο τρόπο διατηρώντας ίδιο τον πίνακα των m descriptors για n αριθμό δειγμάτων , κατά τη διαδικασία του y scrambling

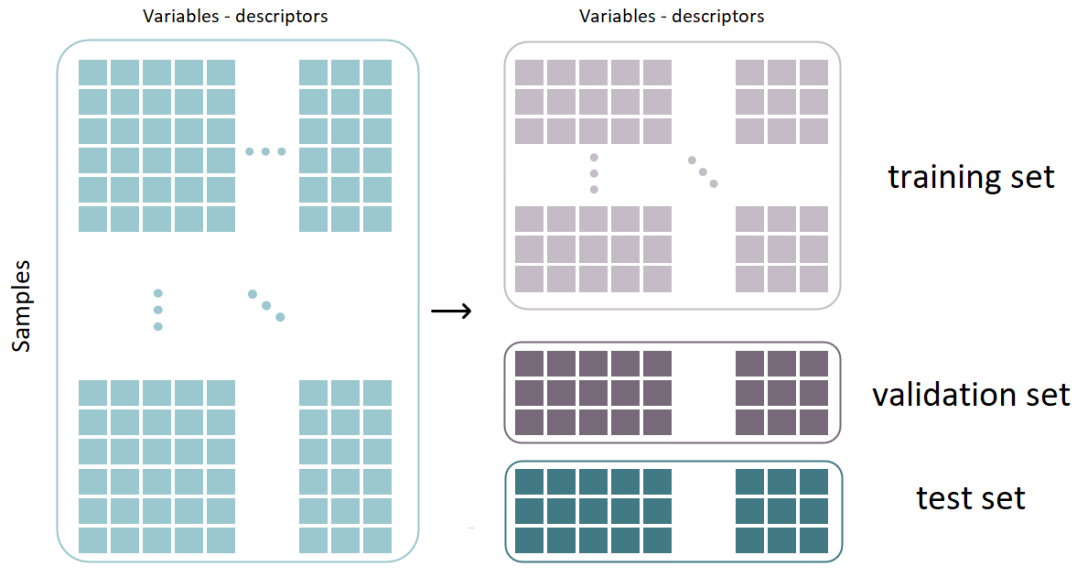
συσχέτισης των ανεξάρτητων μεταβλητών (descriptors x) με την έξοδο (y). Συγκεκριμένα, είναι πιθανό, κάποιες μεταβλητές x του μοντέλου να συσχετίζονται στατιστικά με το y, χωρίς να υπάρχει κάποια σύνδεση αίτιου-αποτελέσματος μεταξύ των δύο, οδηγώντας σε σωστές προβλέψεις, ανεξαρτήτως του μοντέλου που έχει διαμορφωθεί [32].

Για τον έλεγχο αυτής της πιθανότητας, πραγματοποιήθηκε διαδικασία τυχαιοποίησης του y (y randomization) (βλέπε Σχήμα 3.4.1). Σύμφωνα με αυτή, το προβλεπτικό μοντέλο διαμορφώνεται με την ίδια μεθοδολογία, βάσει του ορθού πίνακα των x μεταβλητών και μιας αναμεμειγμένης (scrambled) λίστας της εξόδου y. Η διαμόρφωση του μοντέλου επαναλαμβάνεται μερικές φορές, ανακατεύοντας κάθε φορά εκ νέου το διάνυσμα y μέσω συνάρτησης τυχαιότητας (random function). Αφού το μοντέλο 'χτιστεί' βασισμένο σε τυχαιοποιημένα δεδομένα, πραγματοποιήθηκε έλεγχος της συμπεριφοράς του σε άγνωστα δείγματα και παρατηρήθηκε η στατιστική επίδοση που πέτυχε [33].

Στη συγκεκριμένη δοκιμή, επιθυμητές είναι οι «φτωχές» προβλέψεις. Αν το μοντέλο με τα εσφαλμένα δεδομένα έχει αρνητική επίδοση, τότε μπορεί με ασφάλεια να εξαχθεί το συμπέρασμα πως η ικανοποιητική προβλεπτική συμπεριφορά στα σωστά δεδομένα δεν οφείλεται σε τυχαίες συσχετίσεις των μεταβλητών x και του y, αλλά είναι απόρροια αποτελεσματικής μοντελοποίησης.

3.5. Διαίρεση σε training, validation και test set

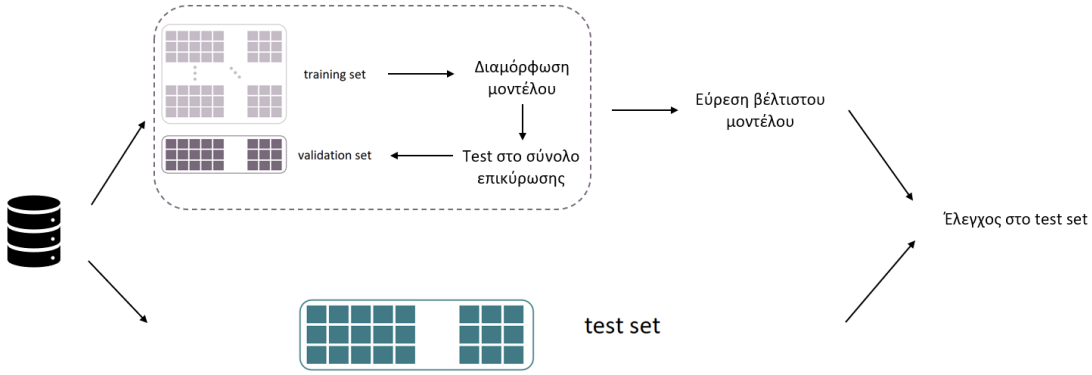
Σε περίπτωση που το μέγεθος του συνόλου δεδομένων το επιτρέπει, δύναται η διαίρεση του σε τρία υποσύνολα. Αρχικά, από το συνολικό dataset, το 20 − 30%, επονομαζόμενο ως σύνολο ελέγχου ,απομακρύνεται για την τελική αξιολόγηση



Σχήμα 3.5.1: Απεικόνιση του διαχωρισμού του συνόλου δεδομένων σε υποσύνολα εκπαίδευσης (training), επικύρωσης (validation) και ελέγχου (test) κατά την διαδικασία επιλογής μοντέλου (model selection)

του μοντέλου, όπως φαίνεται στο σχήμα 3.5.1. Στη συνέχεια, από το 70-80% που παραμένει, το αντίστοιχο 20-30% διαχωρίζεται ως σύνολο επικύρωσης (validation set). Τα υπόλοιπα δεδομένα θεωρούνται ως δεδομένα εκπαίδευσης και χρησιμοποιούνται για την εκπαίδευση του μοντέλου. Η διαδικασία παρουσιάζεται συνοπτικά στο σχήμα 3.5.2.

Το σύνολο εκπαίδευσης διαμορφώνει, ακολουθώντας τη μεθοδολογία τις παραμέτρους του μοντέλου το οποίο αξιολογείται από το σύνολο επικύρωσης. Καθώς το μοντέλο δεν έχει εκπαιδευτεί με δεδομένα του συνόλου επικύρωσης και η έξοδος y είναι άγνωστη σε αυτό, δύναται η εκτίμηση της ακρίβειας του μέσω αυτού. Μέσω ενός κύκλου ανατροφοδότησης, διαμορφώνονται οι βέλτιστες δυνατές παράμετροι που αποδίδουν το μικρότερο δυνατό σφάλμα στις προβλέψεις του συνόλου επικύρωσης. Η διαδικασία αυτή καλείται επιλογή μοντέλου ή model selection [34].



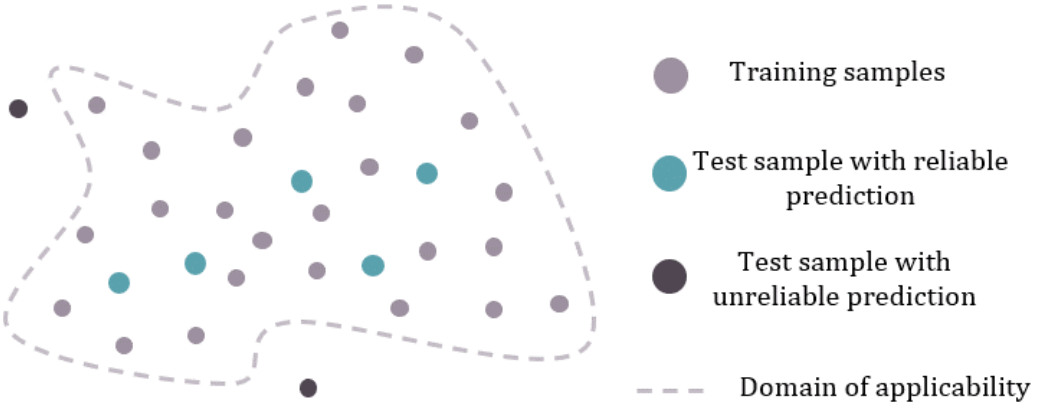
Σχήμα 3.5.2: Απεικόνιση της διαδικασίας επιλογής και ελέγχου του μοντέλου μέση επαναλαμβανόμενης διαδικασίας

Αφού το μοντέλο διαμορφωθεί πλήρως στο τέλος της επαναληπτικής διαδικασίας, το σύνολο ελέγχου χρησιμοποιείται για την τελική αξιολόγηση και υπολογισμό της στατιστικής επιτυχίας του βέλτιστου μοντέλου. Η αξία της παρουσίας ενός ανεξάρτητου συνόλου ελέγχου είναι μεγάλη, κυρίως για την εκτίμηση της ικανότητας γενίκευσης του μοντέλου σε άγνωστα προς το μοντέλο δεδομένα. Χωρίς τον έλεγχο μέσω του ανεξάρτητου συνόλου, η βελτιστοποίηση είναι πιθανό να οδηγήσει στην υπέρ-προσαρμογή του μοντέλου στα δεδομένα εκπαίδευσης και επικύρωσης, και την μοντελοποίηση του ίδιου του θορύβου των δεδομένων. Σε αυτή την περίπτωση το μοντέλο καθίσταται μη εύρωστο και μη αποτελεσματικό, ιδιαίτερα σε μεγάλα σύνολα δεδομένων [35].

3.6. Πεδίο εφαρμογής

Στην προσπάθεια διαμόρφωσης ενός αποτελεσματικού μοντέλου που παράγει αξιόπιστες προβλέψεις, είναι σημαντικός ο εντοπισμός του πεδίου εφαρμογής του (domain of applicability). Σκοπός αποτελεί η προβλεπόμενη ιδιότητα να αντιστοιχεί σε δείγματα που βρίσκονται πλησίον των δειγμάτων από τα οποία

διαμορφώθηκε το μοντέλο, καθώς η αξιοπιστία του είναι άμεση συνάρτηση της εγγύτητας των δειγμάτων εκπαίδευσης και ελέγχου. Στο πλαίσιο αυτό, ο καθορισμός του πεδίου εφαρμογής αποτελεί βασική αρχή κατά την ανάπτυξη ενός QSAR μοντέλου, βάσει του Οργανισμού Οικονομικής Συνεργασίας και Ανάπτυξης (ΟΟΣΑ) (OECD, Organization for Economic Cooperation and Development) το οποίο μπορεί να οριστεί ως ο δομικός, βιολογικός ή φυσικοχημικός νοητός χώρος εντός τους οποίου διαμορφώθηκε το μοντέλο. Ως εκ τούτου, οι προβλέψεις πρέπει να αφορούν παρεμβολή και όχι προεκβολή του χώρου αυτού και μόνο τότε, μπορεί να θεωρηθεί η πρόβλεψη φερέγγυα (βλέπε Σχήμα 3.6.1) [36].



Σχήμα 3.6.1: Ποιοτική απεικόνιση στο δισδιάστατο δειγματικό χώρο του πεδίου εφαρμογής και των δειγμάτων ελέγχου με αξιόπιστες και μη αξιόπιστες προβλέψεις

Για τον εντοπισμό του πεδίου εφαρμογής, υπολογίζεται το αριθμητικό κατώφλι ($threshold$) του μοντέλου κάτω του οποίου η πρόβλεψη θεωρείται αξιόπιστη:

$$threshold = 3 \times \frac{p^1}{t} \qquad (3.6.1)$$

Όπου, p ο αριθμός των διαθέσιμων μεταβλητών που χρησιμοποιήθηκαν κατά τη μοντελοποίηση και t ο αριθμός των δειγμάτων εκπαίδευσης.

Για κάθε δείγμα i του συνόλου ελέγχου για το οποίο πραγματοποιείται πρόβλεψη, υπολογίζεται η τιμή μόχλευσης h_i ($leverage\ value$) η οποία αποτελεί ουσιαστικά την απόστασή του από το κέντρο βάρους ($centroid$) του δείγματος εκπαίδευσης:

$$h_i = x_i(X^T X)^{-1} x_i^T \qquad (3.6.2)$$

Όπου x_i το διάνυσμα που περιλαμβάνει το σύνολο των ιδιοτήτων του δείγματος προς εξέταση και X το σύνολο εκπαίδευσης. Αν η υπολογιζόμενη τιμή είναι

[1] Σε ορισμένες περιπτώσεις χρησιμοποιείται ως p, η τιμή των διαθέσιμων μεταβλητών + 1 [14]

μικρότερη από το κατώφλι που υπολογίστηκε, η πρόβλεψη του συγκεκριμένου δείγματος θεωρείται αξιόπιστη καθώς βρίσκεται εντός του πεδίου εφαρμογής. Ο υπολογισμός πραγματοποιείται για το σύνολο των δειγμάτων ελέγχου σε κάθε μοντέλο ώστε να επικυρωθεί η εγκυρότητά τους [37].

Κεφάλαιο 4
Ανάπτυξη μεθοδολογίας

4.1. Μεθοδολογία συσταδοποίησης

Στην προσπάθεια ανάπτυξης μεθοδολογιών *in silico* για την πρόβλεψη ανεπιθύμητων ιδιοτήτων των νανοσωματιδίων, ο Ευρωπαϊκός Οργανισμός Χημικών Προϊόντων πρότεινε την τεχνική read-across. Η τεχνική αυτή καθιστά δυνατή την πρόβλεψη ιδιοτήτων βασιζόμενη στην υπόθεση πως παρόμοια υλικά θα χαρακτηρίζονται και από παρόμοιες ιδιότητες. Η ροή εργασιών της κατηγοριοποίησης με μεθοδολογία read-across ακολουθεί τέσσερα στάδια: τον χαρακτηρισμό των νανοδομών, τη συλλογή δεδομένων σε μορφή πίνακα, την ανάπτυξη υπόθεσης ομαδοποίησης και την τελική αξιολόγηση αυτής της υπόθεσης. Στο πλαίσιο της προσπάθειας εύρεσης της καλύτερης υπόθεσης ομαδοποίησης, πραγματοποιούνται επαναλήψεις δοκιμής και σφάλματος, οι οποίες αυξάνουν το χρόνο και το κόστος των δοκιμών.

Υπό αυτό το πρίσμα, σκοπός αυτής της εργασίας αποτελεί η αυτοματοποίηση της διαδικασίας επιλογής της βέλτιστης υπόθεσης ομαδοποίησης, διαμορφώνοντας συστάδες νανοσωματιδίων[2] στον πολυδιάστατο χώρο. Η συγκεκριμένη μεθοδολογία χρησιμοποιείται για την κατηγοριοποίηση δειγμάτων μεταξύ δύο ομάδων – κλάσεων, ανάλογα τις ιδιότητές τους, οι οποίες μπορούν να είναι κατηγορικές ή συνεχείς και βασίζεται στη μέθοδο που προτείνεται από τους Ma et al. (2020) [51] στη δημοσίευση 'Spherical Classification of Data, a New Rule-Based Learning Method'.

Οι ιδιότητες του κάθε δείγματος αποτελούν τις συντεταγμένες του στον πολυδιάστατο χώρο και το βασικό κριτήριο υπολογισμού των αποστάσεων όπως αυτές περιγράφονται παρακάτω. Στο σύνολο των δειγμάτων (S), πέρα από ένα σύνολο «συντεταγμένων» έχει αποδοθεί και μια εκ των δύο κλάσεων k, k= 1 ή 2, σχηματίζοντας δύο σύνολα δειγμάτων S^1 και S^2 με $S = S^1 \cup S^2$ και $S^1 \cap S^2 = \emptyset$.

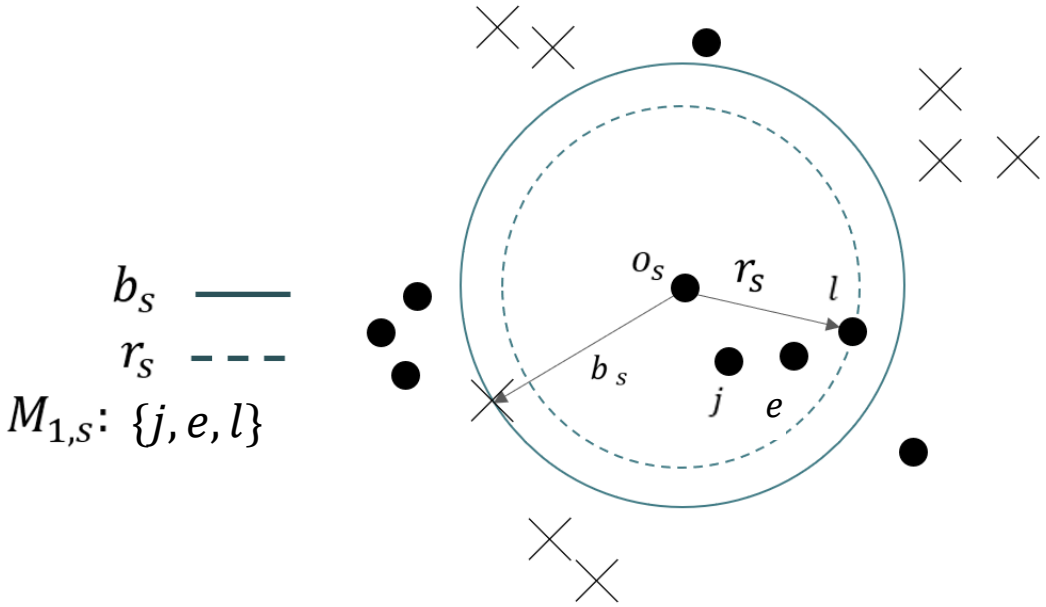
Η μεθοδολογία απαιτεί τη δημιουργία και επεξεργασία clusters C (συστάδες) με τις ιδιότητες:

$$C : \{o, b, r, h, M_1, M_2\}$$

όπου:

[2] Υπενθυμίζεται ότι η μεθοδολογία έχει καθολικό χαρακτήρα και στα πλαίσια της παρούσας Εργασίας παρουσιάζεται η εφαρμογή της σε δεδομένα τοξικότητας. Όπου 'νανοσωματίδια' υπονοείται κάθε εν δυνάμει δείγμα.

o : κέντρο του C με βάση το σύνολο συντεταγμένων των δειγμάτων που το απαρτίζουν
b : απόσταση μεταξύ του o και του πλησιέστερου δείγματος αντίθετης κλάσης εκτός του C
r : απόσταση μεταξύ του o και του πιο απομακρυσμένου δείγματος ίδιας κλάσης εντός του C
h : ομοιογένεια/καθαρότητα (purity) του C
M_1 : δείγματα ίδιας κλάσης που ανήκουν στο C
M_2 : δείγματα αντίθετης κλάσης που ανήκουν στο C



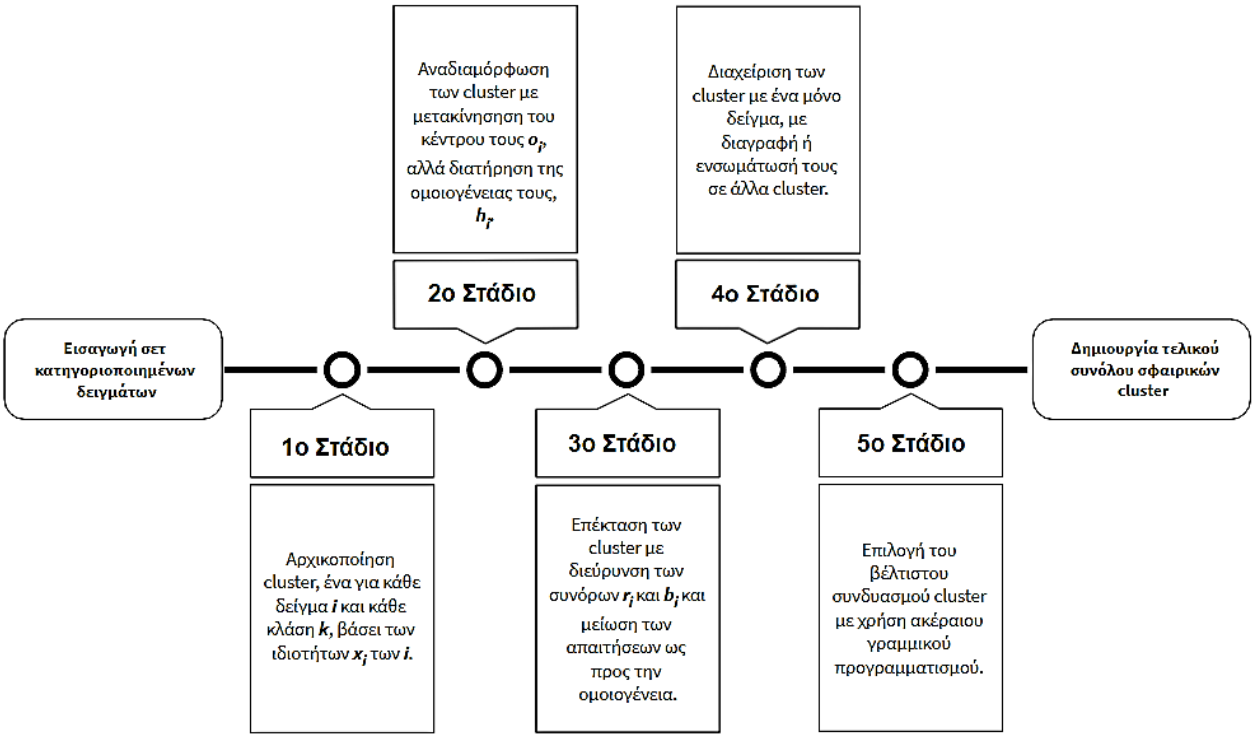
Σχήμα 4.1.1: Απεικόνιση ενδεικτικού δειγματικού χώρου και του cluster C_s που προκύπτει από το δείγμα s. Οι κουκίδες και τα 'x' αντιστουχούν στον συμβολισμό των δυο κλάσεων. Το C_s έχει διευρυμένη και συντηρητική ακτίνα τα b_s και r_s, αντίστοιχα, ενώ το σύνολο $M_{2,s}$ είναι κενό.

Τα μεγέθη αυτά ορίζουν ένα cluster το οποίο χαρακτηρίζεται από ένα κέντρο o, μια συντηρητική και μια διευρυμένη ακτίνα r και b, εντός των οποίων εντοπίζονται δείγματα και των δύο κλάσεων, και μιας καθαρότητας h (βλέπε Σχήμα 4.1.1). Η εκάστοτε συστάδα ανήκει σε γνωστή κλάση, 1 ή 2, ανάλογα την κλάση στην οποία ανήκει το δείγμα από το οποίο προέρχεται. Σημειώνεται ότι μια κλάση μπορεί να διαθέτει περισσότερες της μιας συστάδας.

Για την ανάπτυξη της μεθόδου που ακολουθεί, χρησιμοποιείται ο συμβολισμός:

S^k : το σύνολο των αριθμημένων δειγμάτων, ένα για κάθε κλάση k
x : οι συντεταγμένες του δείγματος, που αντιστοιχούν στο σύνολο των ιδιοτήτων/μεταβλητών του (features)
M : το σύνολο των δειγμάτων (και των δύο κλάσεων) τα οποία εντοπίζονται εντός του C
P : το σύνολο των αριθμημένων συστάδων, ένα για κάθε κλάση

J : το σύνολο των συστάδων που καλύπτουν εκάστοτε δείγμα και ανήκουν στην ίδια κλάση με αυτό



Σχήμα 4.1.2: Σκιαγράφηση των σταδίων της προτεινόμενης μεθοδολογίας

Η μέθοδος εφαρμόζεται σε πέντε στάδια, όπως αυτά παρουσιάζονται στο σχήμα 4.1.2 και, αφού ολοκληρωθεί, ακολουθεί η διαδικασία κατηγοριοποίησης νέων δειγμάτων.

Στάδιο 1ο: Δημιουργία των συστάδων

Στο πρώτο στάδιο, πραγματοποιείται η αρχικοποίηση των συστάδων, δημιουργώντας μια συστάδα για κάθε δείγμα κάθε κλάσης, όπως αυτό παρουσιάζεται ενδεικτικά στον τρισδιάστατο χώρο στο σχήμα 4.1.3. Έτσι, αν στο αρχικό σύνολο δειγμάτων υπάρχουν α δείγματα κλάσης 1 και β δείγματα κλάσης 2 δημιουργούνται α και β συστάδες των κλάσεων 1 και 2, αντίστοιχα.

Σε κάθε συστάδα i, το κέντρο ορίζεται με βάση τις ιδιότητες του δείγματος στο οποίο βασίζεται. Έτσι, για το δείγμα i και τον σχηματισμό του C_i:

$$o_i = x_i \tag{4.1.1}$$

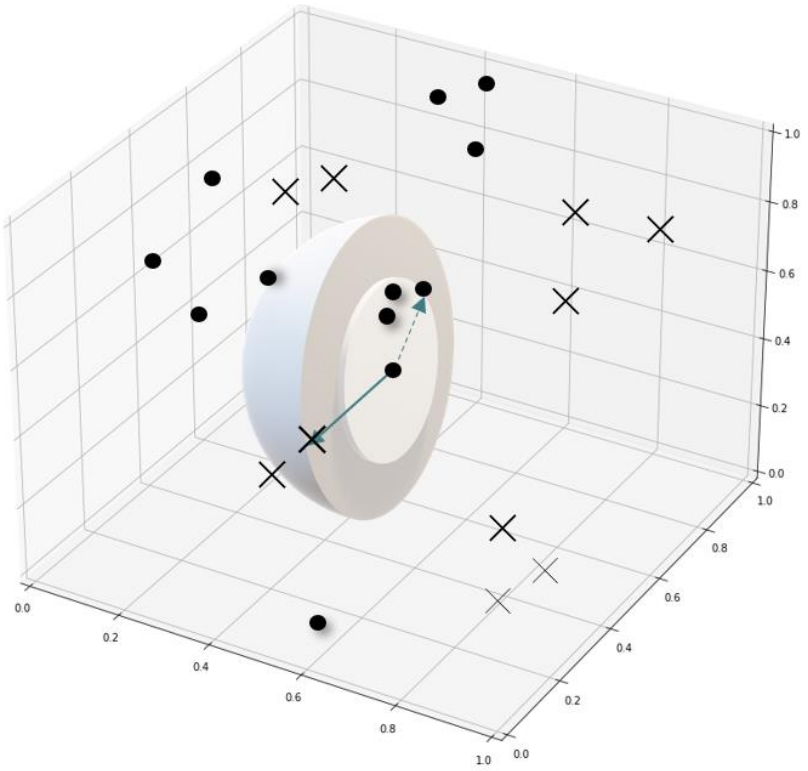
Λαμβάνοντας υπόψιν όλα τα δείγματα της αντίθετης κλάσης, υπολογίζεται η ακτίνα b_i ως η μικρότερη απόσταση μεταξύ του νέου o_i και των x των δειγμάτων της αντίθετης κλάσης. Δηλαδή:

$$b_i = \min\{d(o_i\ , x_j)\}\ \text{όπου}\ j \in S^{opposite} \tag{4.1.2}$$

Σημειώνεται πως για τον υπολογισμό κάθε απόστασης, χρησιμοποιείται η ευκλείδεια απόσταση μεταξύ των συντεταγμένων/ιδιοτήτων των δειγμάτων με βάση τη σχέση:

$$d(o_i, x_j) = \sqrt{\sum (o_{i,f} - x_{j,f})^2} \tag{4.1.3}$$

όπου f οι αριθμητικές τιμές των ιδιοτήτων των δειγμάτων.



Σχήμα 4.1.3: Απεικόνιση ενδεικτικού δειγματικού χώρου και του cluster C_s που προκύπτει από το δείγμα s στον τρισδιάστατο χώρο. Κάθε διάσταση αποτελεί και μια ιδιότητα των δειγμάτων με αποτέλεσμα δεδομένα με έως και τρείς ιδιότητες να μπορούν αποτυπωθούν γραφικά.

Στη συνέχεια, κάθε δείγμα j που ανήκει στην ίδια κλάση k με το υπό εξέταση δείγμα i ελέγχεται ως προς την απόσταση του από το o_i. Κάθε δείγμα j του οποίου η απόσταση από το o_i είναι μικρότερη από την ακτίνα που ορίζει το b_i, προστίθεται στο σύνολο M_1 της συστάδας C_i. Η λογική αυτή εξάρτηση, μπορεί να εκφραστεί:

$$M_{1,i} = \{j\ \in S\ \text{ώστε}\ d(o_i\ , x_j) < b_i\} \tag{4.1.4}$$

Αφού οριστεί το $M_{1,i}$, για κάθε συστάδα i, εντοπίζονται δυο περιπτώσεις. Αν σε αυτό δεν εμφανίζονται άλλα δείγματα πέραν του i τότε η συστάδα C_i κατονομάζεται singleton (μονάδα) και η ακτίνα r_i υπολογίζεται ως το ήμισυ του

b_i, χωρίς περαιτέρω αναζήτηση. Αντίθετα, αν στο $M_{1,i}$ εμφανίζονται περισσότερα του ενός δείγματα, το r_i υπολογίζεται ως η μέγιστη εκ των αποστάσεων του o_i από τα x_j των υπόλοιπων δειγμάτων του $M_{1,i}$.

Οι σχέσεις αυτές εκφράζονται μαθηματικά ως εξής:

$$r_i = \begin{cases} \dfrac{b_i}{2}, & M_{1,i} = \{i\} \\ \max\{d(o_i, x_j)\} \text{ όπου } j \in M_{1,i}, & \text{σε κάθε άλλη περίπτωση} \end{cases} \quad (4.1.5)$$

Εξ ορισμού, σε αυτό το στάδιο διαμόρφωσης των συστάδων, η ακτίνα εφαρμογής τους δεν ξεπερνά την ελάχιστη απόσταση από οποιοδήποτε δείγμα αντίθετης κλάσης. Έτσι, κάθε δείγμα που εντοπίζεται εντός της συστάδας C_i κλάσης k θα ανήκει, εξίσου, στην ίδια κλάση ενώ, ταυτόχρονα, δεν θα υπάρχουν δείγματα στο $M_{2,i}$.
Άρα η καθαρότητα του εκάστοτε cluster C_i και το $M_{2,i}$ ορίζονται:

$$h_i \leftarrow 1$$
$$M_{2,i} \leftarrow \emptyset$$

Έχοντας, πλέον, καθορίσει πλήρως τις ιδιότητες της συστάδας C_i για το δείγμα i και προσθέσει το index (αριθμό σειράς) του στο σύνολο P, η διαδικασία επαναλαμβάνεται έως ότου αρχικοποιηθούν όλες τις συστάδες, ένα για κάθε δείγμα εκάστοτε κλάσης.

Καθώς, μέχρι στιγμής, κάθε δείγμα αντιστοιχεί και σε μια συστάδα, τα σύνολα P και S ταυτίζονται, και για τις δύο κλάσεις.

Στάδιο 2ο: Αναδιαμόρφωση των συστάδων

Σε αυτό το στάδιο, έχοντας οριοθετηθεί τα C και οριστεί τα δείγματα που συμμετέχουν σε αυτά, οι συστάδες αναδιαμορφώνονται, επεκτείνοντας τα σύνορά τους και μετακινώντας το κέντρο τους, ώστε να συμπεριλάβουν όσο το δυνατόν περισσότερα δείγματα της ίδιας κλάσης k, χωρίς να «μολυνθούν» από δείγματα αντίθετης κλάσης (η καθαρότητα h_i παραμένει ίση με 1).

Στο στάδιο, αυτό καθοριστικό ρόλο έχει η αλληλεπίδραση του κέντρου o_i με τα συμμετέχοντα στο C_i δείγματα και τις ιδιότητές τους. Έτσι, δεν γίνεται καμία επεξεργασία στις συστάδες που περιλαμβάνουν μόνο ένα δείγμα, τα $singletons$ (με $|M_{1,i}| = 1$).

Για όλα τα υπόλοιπα C των συνόλων P των δύο κλάσεων με $|M_{1,i}| > 1$, προσεγγίζεται εκ νέου το κέντρο o_i, λαμβάνοντας υπόψιν όλα τα δείγματα που εντοπίζονται στο $M_{1,i}$ και υπολογίζοντας το κέντρο βάρους των ιδιοτήτων τους (βλέπε Σχήμα 4.1.4):

$$o_i' = \frac{\sum_{j \in M_{1,i}} x_j}{|M_{1,i}|} \quad (4.1.6)$$

Ομοίως με το προηγούμενο στάδιο, το νέο συντηρητικό σύνορο της συστάδας C_i, r_i' υπολογίζεται ως:

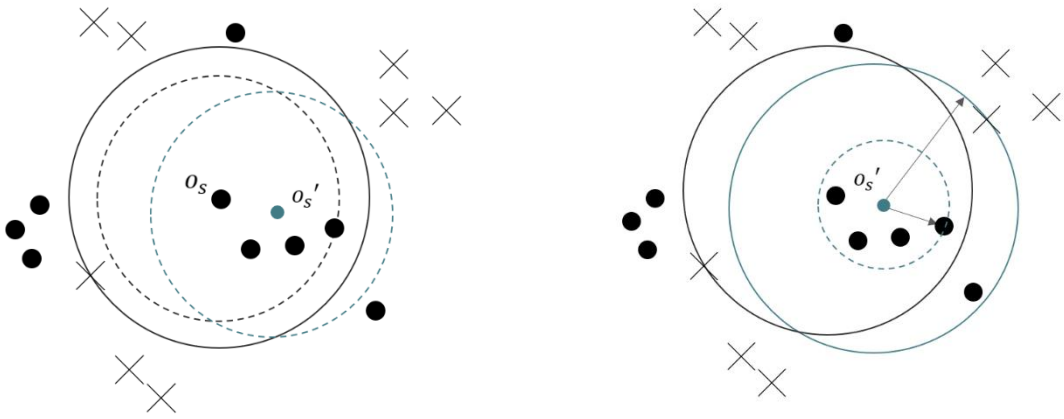
$$r_i' = \max\{d(o_i', x_j)\} \text{ όπου } j \in M_{1,i} \quad (4.1.7)$$

Καθώς πλέον τα όρια της συστάδας C_i έχουν διευρυνθεί, εξετάζεται η μόλυνση του από δείγματα αντίθετης κλάσης. Έτσι, αναζητώντας τα δείγματα αντίθετης κλάσης που απέχουν από το νέο κέντρο o_i' απόσταση μικρότερη του r_i', υπολογίζεται εκ νέου το $M_{2,i}$:

$$M_{2,i} = \{j \in S^{opposite} \text{ ώστε } d(o_i', x_j) < r_i\} \quad (4.1.8)$$

Χωρίς να ληφθεί υπόψιν το σύνολο των δειγμάτων που ανήκουν στο $M_{2,i}$, υπολογίζονται οι αποστάσεις του νέου κέντρου από τα x_j των j δειγμάτων αντίθετης κλάσης και το b_i' παίρνει την μικρότερη τιμή εξ αυτών. Αυτό εκφράζεται και ως:

$$b_i' = \min\{d(o_i', x_j)\} \text{ όπου } j \in S^{opposite} \backslash M_{2,i} \quad (4.1.9)$$



Σχήμα 4.1.4: Μετακίνηση του κέντρου o_s του cluster C_s λαμβάνοντας υπόψιν τα δείγματα που περιλαμβάνονται στο $M_{1,s}$ και τροποποίηση των r_s και b_s.

Σε περίπτωση που η λίστα αυτή είναι κενή, το b_i' παίρνει την τιμή: $b_i' = 1.5 \cdot r_i'$. Έχοντας επανυπολογίσει το διευρυμένο σύνορο καθενός C_i, ορίζεται εκ νέου και το $M_{1,i}$ ως:

$$M_{1,i} = \{j \in S \text{ ώστε } d(o_i', x_j) < b_i\} \quad (4.1.10)$$

Τέλος, για τον πλήρη καθορισμό της συστάδας C_i, υπολογίζεται και η καθαρότητά του ως το πηλίκο του αριθμού των δειγμάτων ίδια κλάσης, $|M_{1,i}|$, προς τον συνολικό αριθμό δειγμάτων και των δύο κλάσεων που ανήκουν σε αυτό, $|M_{1,i}| + |M_{2,i}|$.

$$h_i' = \frac{|M_{1,i}|}{|M_{1,i}| + |M_{2,i}|} \tag{6.1.11}$$

Στο σημείο αυτό, αναγνωρίζονται δυο περιπτώσεις. Αν η επέκταση των συνόρων δεν έχει οδηγήσει στην εισροή δειγμάτων αντίθετης κλάσης και η καθαρότητα παραμένει ίση με τη μονάδα, τότε η συστάδα διατηρεί την ομοιογένειά του και η αναδιαμόρφωση του ολοκληρώνεται επιτυχώς.
Αντίθετα, στην περίπτωση που η επέκταση οδηγεί σε ενσωμάτωση δειγμάτων αντίθετων δειγμάτων στη συστάδα, ακολουθείται εναλλακτική επαναληπτική διαδικασία μετακίνησης του νέου κέντρου προς την προηγούμενη θέση του (με μέγιστο αριθμό επαναλήψεων $n = 10$ ακόμα και αν δεν επιτευχθεί η ζητούμενη καθαρότητα), όπως αυτή περιγράφεται παρακάτω.

Το νέο κέντρο μετακινείται στη διεύθυνση $d := o_i - o_i{}'$ ή εναλλακτικά, $o_i' \leftarrow o_i' + \lambda d$. Για τον υπολογισμό του λ το οποίο κυμαίνεται σε τιμές μεταξύ 0 και 1, απαιτούνται δυο δείγματα που συμπεριλαμβάνονται στο C_i, το πλησιέστερο στο νέο κέντρο δείγμα αντίθετης κλάσης και το πιο απομακρυσμένο στο νέο κέντρο δείγμα ίδιας κλάσης. Μαθηματικά, αυτό μπορεί να γραφεί ως:

$$u = j \in M_{1,i}{}' \text{ ώστε } \{d(o_i{}', x_j) = max\} \tag{4.1.12}$$
$$w = j \in M_{2,i}{}' \text{ ώστε } \{d(o_i{}', x_j) = min\} \tag{4.1.13}$$

Κατά την εύρεση του w, ζητείται το διάνυσμα μεταξύ του υπολογιζόμενου σημείου και του κέντρου να βρίσκεται σε αμβλεία γωνία σε σχέση με το διάνυσμα d. Μετά από έλεγχο του $cos(\theta)$ μεταξύ των δύο διανυσμάτων, τα $j \in M_{2,i}{}'$ φιλτράρονται ώστε να διατηρούνται μόνο τα δείγματα που ικανοποιούν αυτό τον περιορισμό.

Αφού βρεθούν τα w και u, με βάση το σύνολο ιδιοτήτων των x_u και x_w, εντοπίζεται νέο σημείο m με $m \leftarrow o_i' + \lambda d$ το οποίο απέχει εξίσου από τα u και w. Μαθηματικά, σε μορφή γινομένου διανυσμάτων η απαίτηση αυτή εκφράζεται ως:

$$\left(m - \frac{x_u + x_w}{2}\right) \cdot (x_u + x_w) = 0$$

ή

$$\lambda = \frac{\left(\frac{x_u + x_w}{2} - o_i{}'\right) \cdot (x_u - x_w)}{d \cdot (x_u - x_w)} \tag{4.1.14}$$

Σε περίπτωση που μέσω των υπολογισμών προκύψουν τιμές ‘inf’ ή ‘NaN’, διατηρείται το προηγούμενο κέντρο. Με γνωστό το λ και προσθέτοντας ένα μικρό θετικό παράγοντα ε, υπολογίζεται το νέο κέντρο ως:

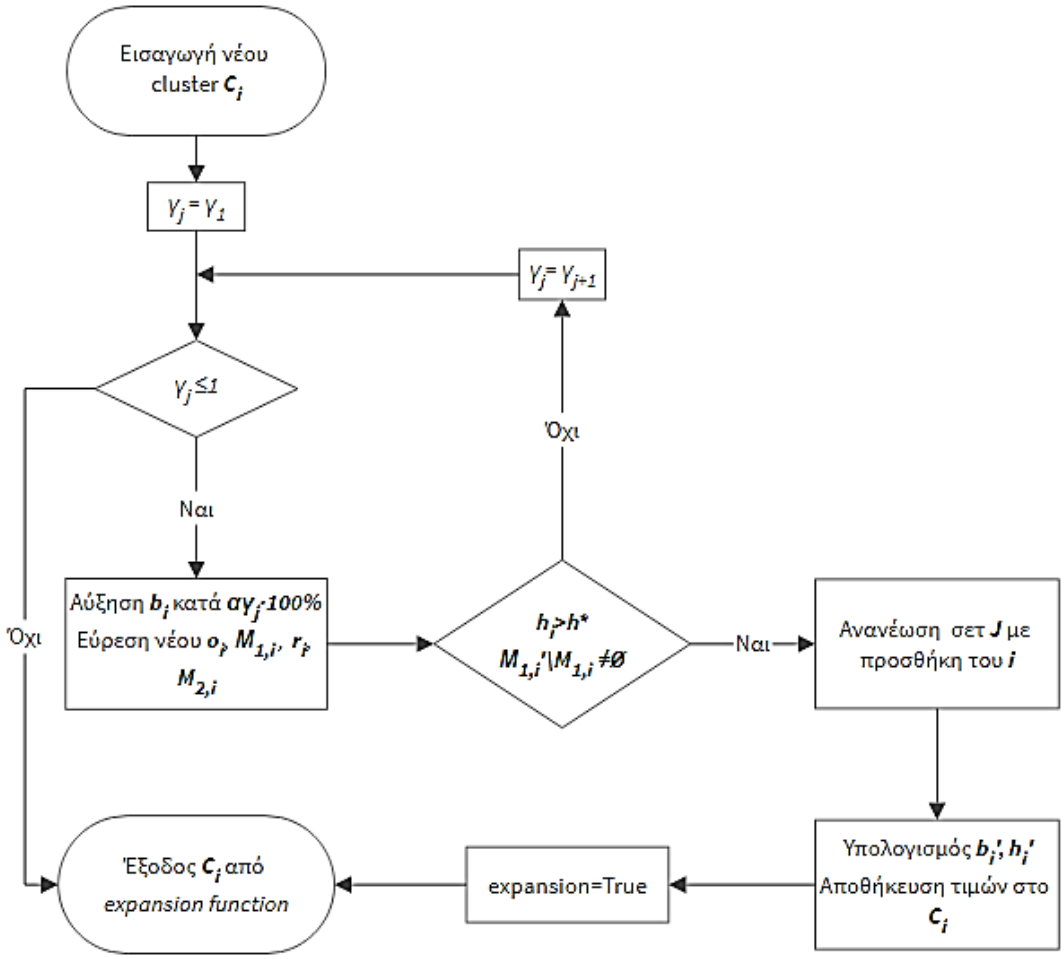
$$o_i' \leftarrow o_i' + (\lambda + \varepsilon) \cdot (o_i - o_i') \tag{4.1.15}$$

Γνωρίζοντας το νέο κέντρο, υπολογίζονται εκ νέου τα δύο σύνορα του κάθε cluster καθώς και τα συμμετέχοντα στα $M_{1,i}'$ και $M_{2,i}'$ δείγματα. Πραγματοποιείται έλεγχος της καθαρότητας και η διαδικασία προσέγγισης του νέου κέντρου επαναλαμβάνεται έως ότου $h_i \leftarrow 1$. Σε κάθε επανάληψη, το νέο κέντρο o_i' προσεγγίζει το προηγούμενο κέντρο o_i ώστε να αναιρεθεί η μόλυνση της συστάδας με δείγματα αντίθετης κλάσης, η οποία πραγματοποιήθηκε κατά την διεύρυνση του.

Αφού επιτευχθεί πλήρης καθαρότητα, ελέγχεται αν η αναδιαμόρφωση του κέντρου και των συνόρων της συστάδας C_i οδήγησε στην εισροή νέων δειγμάτων ίδιας κλάσης. Στην περίπτωση που $M_{1,i}'\backslash M_{1,i} \neq \emptyset$, υπολογίζεται εκ νέου το r_i' και για κάθε l στο $M_{1,i}'$ ανανεώνονται τα σύνολα J: $J_l = J_l \cup \{i\}$. Η διαδικασία ολοκληρώνεται με την ανανέωση των ιδιοτήτων της συστάδας με τα νέα o_i', b_i', r_i', h_i', $M_{1,i}'$, $M_{2,i}'$ και επαναλαμβάνεται για κάθε δείγμα και των δύο κλάσεων.

Στάδιο 3ο: Επέκταση των συστάδων

Στο στάδιο αυτό, γίνεται η προσπάθεια επέκτασης των ορίων των συστάδων θυσιάζοντας την απόλυτη καθαρότητα και επιτρέποντας της μόλυνση από

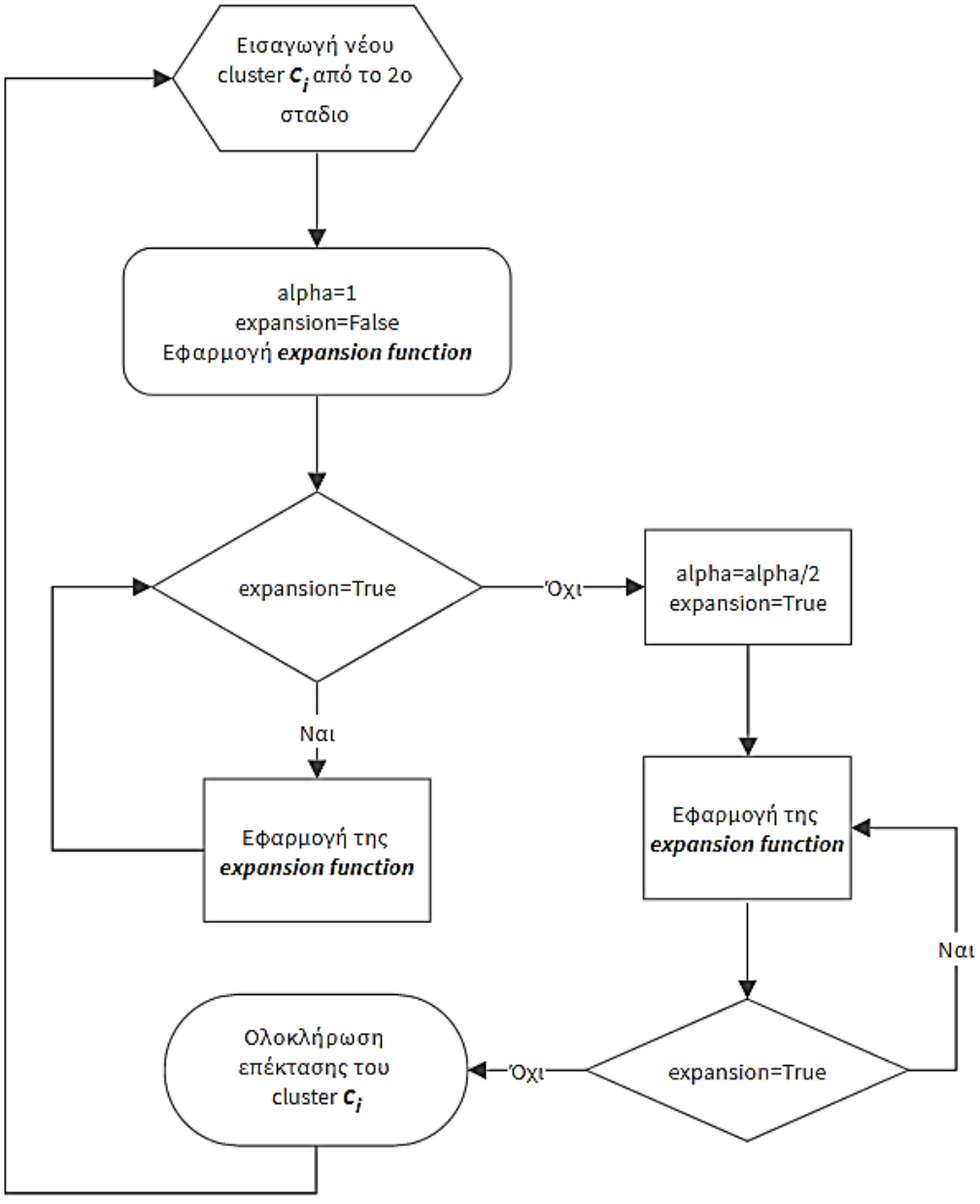


Σχήμα 4.1.5: Διαγραμματική ανάλυση της αλγοριθμικής λογικής του expansion function

δείγματα αντίθετης κλάσης με σκοπό την ενίσχυση των δυνατότητων γενίκευσης και της ευρωστίας της μεθόδου.

Σε αυτό το πλαίσιο, μέσω επαναληπτικής διαδικασίας επιχειρείται η μεγιστοποίηση του b_i εφόσον $h_i \geq h^*$ όπου h^* τιμή καθορισμένη από το χρήστη με βάση τις απαιτήσεις σε ομοιογένεια των συστάδων. Για την επέκταση κάθε C_i, χρησιμοποιείται η βοηθητική παράμετρος α η οποία αρχικοποιείται με την τιμή 1.

Το b_i αυξάνεται διαδοχικά κατά $a\gamma \cdot 100\%$ όπου $\gamma \in \{0.2, 0.4, 0.6, 0.8, 1\}$. Σε κάθε αύξηση, ανανεώνεται το σετ $M'_{1,i}$ ώστε να εντοπιστούν τα πιθανά νέα ενσωματωμένα δείγματα και με βάση αυτά, υπολογίζονται εκ νέου τα $o'_i, r'_i, M'_{2,i}$.



Σχήμα 4.1.6: Διαγραμματική απεικόνιση της αλγοριθμικής λογικής του σταδίου 3 της μεθοδολογίας

Αν η καθαρότητα παραμένει μεγαλύτερη του h^*, δηλαδή αν $\frac{|M_{1,i'}|}{|M_{1,i'}|+|M_{2,i'}|} \geq h^*$, και η επέκταση της συστάδας είναι επιτυχής, δηλαδή $M_{1,i}{}' \backslash M_{1,i} \neq \emptyset$, τότε για κάθε l στο $M_{1,i}{}'$:

$$J_l = J_l \cup \{i\}$$

ενώ ταυτόχρονα ανανεώνονται τα b_i' , h_i' και αποθηκεύονται οι νέες τιμές στις ιδιότητες του C_i. Τέλος, ορίζεται η λέξη-κλειδί expansion σε True, καθώς η επέκταση του cluster ήταν επιτυχής. Η αλγοριθμική διαδικασία που περιεγράφηκε στην παράγραφο ορίζεται ως expansion function (βλέπε Σχήμα 4.1.5).

Έτσι, για κάθε cluster C_i, αρχικά ορίζεται $alpha = 1$, $expansion = False$ και εφαρμόζεται η expansion function. Κατά την ολοκλήρωση της εφαρμογής της, ελέγχεται η κατάσταση της μεταβλητής expansion. Αν $expansion = True$, το cluster επεκτάθηκε επιτυχώς και εισέρχεται εκ νέου σε επαναληπτική διαδικασία εφαρμογής της expansion function έως ότου δεν μπορεί πλέον να επεκταθεί χωρίς να αλλοιωθεί σημαντικά η καθαρότητα του C_i. Αντίθετα, αν $expansion = False$, το cluster για $alpha = 1$ δεν μπορεί να επεκταθεί. Έτσι, ορίζεται $alpha = \frac{alpha}{2}$, $expansion = True$ και ακολουθείται εκ νέου η διαδικασία που ακολουθήθηκε παραπάνω με επανάληψη της expansion function έως ότου, πλέον, ολοκληρωθεί η διαδικασία και επιβεβαιωθεί η αδυναμία περαιτέρω επέκτασης, όπως φαίνεται στο Σχήμα 4.1.6.

Σημειώνεται πως, ο έλεγχος $expansion = False$ έπεται του $expansion = True$ και οι δύο έλεγχοι είναι ανεξάρτητοι μεταξύ τους. Έτσι, στην περίπτωση που το C_i εξέλθει από την πρώτη επανάληψη έχοντας επεκταθεί, μπορεί να εισέλθει ξανά στην δεύτερη επανάληψη, μειώνοντας το $alpha$ κατά το ήμισυ.

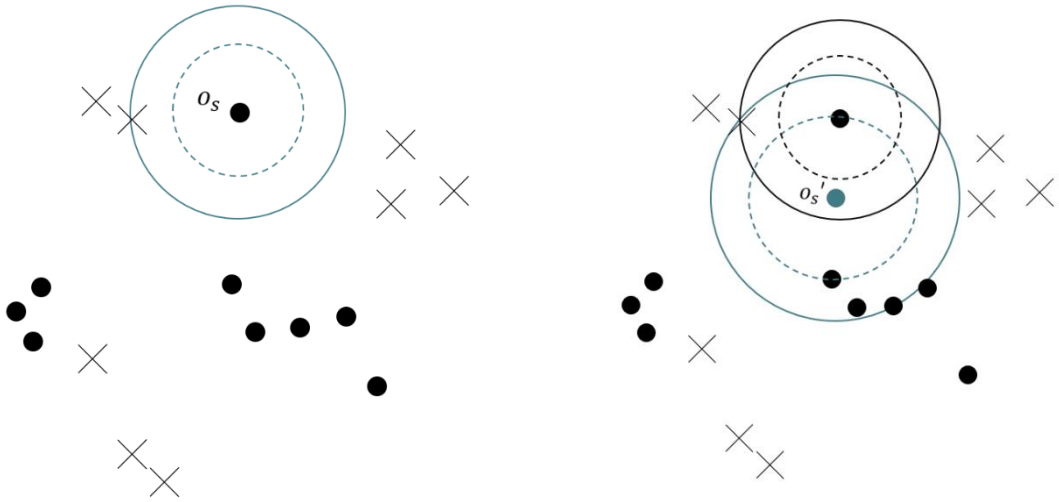
Η διαδικασία επαναλαμβάνεται για κάθε δείγμα και των δύο κλάσεων, ολοκληρώνοντας το τρίτο στάδιο επεξεργασίας των συστάδων.

Στάδιο 4ο: Διαχείριση των singleton συστάδων

Στο στάδιο αυτό, μελετώνται οι συστάδες τα οποία δεν κατάφεραν να επεκταθούν επαρκώς ($M_{1,i} = \{i\}$) και διατηρούν ένα μοναδικό δείγμα εντός των συνόρων τους, αυτό από το οποίο προήλθαν.

Ανάλογα τη διαμόρφωση των συστάδων πλησίον τους, εμφανίζονται δυο πιθανές περιπτώσεις. Στην περίπτωση που $|J_i| \geq 2$, το δείγμα καλύπτεται και από άλλο cluster πέραν του C_i. Τότε η παρουσία του C_i δεν είναι απαραίτητη καθώς αποτελεί υποσύνολο κάποιου άλλου C_j που περικλείει, μεταξύ άλλων, και το i. Έτσι, καθίσταται δυνατή η διαγραφή του ως cluster από το σύνολο P αλλά η διατήρηση του x_i ως δείγμα του C_j.

Ωστόσο, στην περίπτωση που $|J_i| = \{i\}$, εμφανίζονται δυο δυνατότητες.



Σχήμα 4.1.7: Μετακίνηση του κέντρου του singleton cluster στο μέσο της απόστασης μεταξύ του παλαιού κέντρου και του πλησιέστερου δείγματος ίδιας κλάσης με αναπροσαρμογή των b_s και r_s

<u>*1η Υποπερίπτωση:*</u> Υπάρχει κάποιο πλησίον δείγμα j ίδιας κλάσης με τέτοιο σύνολο ιδιοτήτων x_j ώστε αν το o_i μετακινηθεί στο μέσο της απόστασης $|x_i - x_j|$, τότε το αναθεωρημένο C_i, δεν είναι πλέον $singleton$ αλλά περιλαμβάνει και το δείγμα j, τουλάχιστον (βλέπε Σχήμα 4.1.7). Δηλαδή:

$$u = j \in S \text{ ώστε } \{d(x_i, x_j) = min\} \quad (4.1.16)$$

$$o_i' = \frac{x_u + x_i}{2} \quad (4.1.17)$$

Αφού υπολογιστούν εκ νέου τα $b_i', r_i', M_{1,i}', M_{2,i}'$, υπολογίζεται και ελέγχεται η καθαρότητα του διαμορφωμένου C_i. Αν αυτή παραμένει μεγαλύτερη του ορίου h^*, αποθηκεύονται οι νέες τιμές στο C_i και ανανεώνεται το σετ των J ώστε για κάθε l στο $M_{1,i}'$:

$$J_l = J_l \cup \{i\}$$

<u>*2η Υποπερίπτωση:*</u> Ακολουθώντας την ίδια διαδικασία με προηγουμένως, υπολογίζεται η καθαρότητα του διαμορφωμένου C_i, η οποία προκύπτει μικρότερη του κατώτατου επιθυμητού ορίου, οδηγώντας στο συμπέρασμα πως υπάρχουν περισσότερα δείγματα αντίθετης κλάσης κοντά στο i, παρά ίδιας.

Έτσι, θεωρώντας την ύπαρξή του στα όρια του στατιστικού λάθους, καθίσταται δόκιμη η διαγραφή του δείγματος i:

$$P = P \setminus \{i\}$$

Στάδιο 5ο: Επιλογή των τελικών συστάδων

Αφού έχει ολοκληρωθεί η διαμόρφωση των συστάδων, η επέκτασή τους ώστε να καλύπτουν όσο το δυνατόν περισσότερα δείγματα ίδιας κλάσης, και η

τροποποίηση ή διαγραφή αυτών που αδυνατούν να καλύψουν άλλα δείγματα, ακολουθεί διαδικασία γραμμικού προγραμματισμού με σκοπό την εύρεση του βέλτιστου συνδυασμού επιλεγμένων συστάδων ώστε να καλύπτεται όλος ο δειγματικός χώρος με τον ελάχιστο αριθμό συστάδων.

Για το σκοπό αυτό δημιουργείται βοηθητική παράμετρος w_i για κάθε C_i η οποία αντιπροσωπεύει τη 'σχετική βαρύτητα' (weight) του C_i, αξιολογώντας την ακτίνα του, τον αριθμό των δειγμάτων που περιλαμβάνει και την καθαρότητά του με βάση τη σχέση:

$$w_i = \frac{r_i}{\left|M_{1,i}\right| \cdot h_i} \tag{4.1.18}$$

Ομοίως δημιουργείται και πίνακας διαστάσεων (j, i) όπου j ο αριθμός των δειγμάτων σε κάθε κλάση. Τα στοιχεία του πίνακα αποκτούν την τιμή 1 αν το j καλύπτεται από το C_i, δηλαδή αν το j υπάρχει στο $M_{1,i}$ και την τιμή 0, αν όχι. Δηλαδή:

$$A_{ji} = \begin{cases} 1, \alpha\nu\ j \in\ M_{1,i} \\ 0, \sigma\varepsilon\ \kappa\acute{\alpha}\theta\varepsilon\ \acute{\alpha}\lambda\lambda\eta\ \pi\varepsilon\rho\acute{\imath}\pi\tau\omega\sigma\eta \end{cases} \tag{4.1.19}$$

Μέσω μεικτού ακέραιου γραμμικού προγραμματισμού (Mixed Integer Linear Programming MILP), ζητείται η ελαχιστοποίηση της σχέσης:

$$\sum w_i y_i$$

Το y_i αποτελεί τη δυαδική μεταβλητή επιλογής, η οποία θα πάρει την τιμή 1 αν επιλεχθεί το C_j και την τιμή 0, αν όχι.

$$y_j = \begin{cases} 1, \alpha\nu\ \varepsilon\pi\iota\lambda\varepsilon\chi\theta\varepsilon\acute{\imath}\ to\ cluster\ C_j \\ 0, \sigma\varepsilon\ \kappa\acute{\alpha}\theta\varepsilon\ \acute{\alpha}\lambda\lambda\eta\ \pi\varepsilon\rho\acute{\imath}\pi\tau\omega\sigma\eta \end{cases} \tag{4.1.20}$$

Βασικός γραμμικός περιορισμός για την επιλογή των συστάδων και τον καθορισμό των y_i είναι η σχέση:

$$\sum A_{ji} y_i \geq 1\ \forall\ j \in S \tag{4.1.21}$$

Βάσει αυτής, τα y οφείλουν να καθοριστούν με τέτοιο τρόπο ώστε, στο σύνολό τους, οι συστάδες που επιλέγονται να περιλαμβάνουν όλα τα δείγματα κάθε κλάσης τουλάχιστον μια φορά. Συμπληρωματικά, σχηματίζονται δυο ακόμα περιορισμοί για την εξασφάλιση της διατήρησης όλων των δειγμάτων κατά την διαγραφή των συστάδων. Για το σύνολο J κάθε δείγματος, θα πρέπει να επιλέγεται ένα τουλάχιστον εκ των συστάδων στα οποία περιέχεται δηλαδή:

$$\sum_{j\ \epsilon\ J_i} y_j \geq 1\ \forall\ i \in P \tag{6.1.22}$$

Ταυτόχρονα, εντοπίζονται τα δείγματα τα οποία εμφανίζονται μόνο σε μία και μοναδική συστάδα και αποθηκεύονται στο σύνολο $[ind]$. Για κάθε ένα δείγμα του $[ind]$, ζητείται η διατήρηση της συστάδας στο οποίο ανήκουν θέτοντας τον περιορισμό:

$$y_j = y_{single_j} \; \forall \; j \in [ind] \tag{4.1.23}$$

$$y_{single_j} \geq 1 \; \forall \; j \in S \tag{4.1.24}$$

Αντικειμενική συνάρτηση	$min \sum w_i y_i$
Περιορισμοί	$\sum A_{ji} y_i \geq 1 \; \forall \; j \in S$ $y_j = y\_single_j \; \forall \; j \in [ind]$ $y\_single_j \geq 1 \; \forall \; j \in S$ $\sum y_j \geq 1 \; \forall \; j \in J_i \, , i \in P$
Μεταβλητές	$y_i \, , \;\; i \in P \; και \; y_i \; \in \{0,1\}$ $y\_single_i \, , \;\; i \in P \; και \; y_i \; \in \{0,1\}$

Πίνακας 4.1.1: Συνοπτική παρουσίαση του προβλήματος μικτού ακέραιου γραμμικού προγραμματισμού για την εύρεση του βέλτιστου συνδυασμού cluster

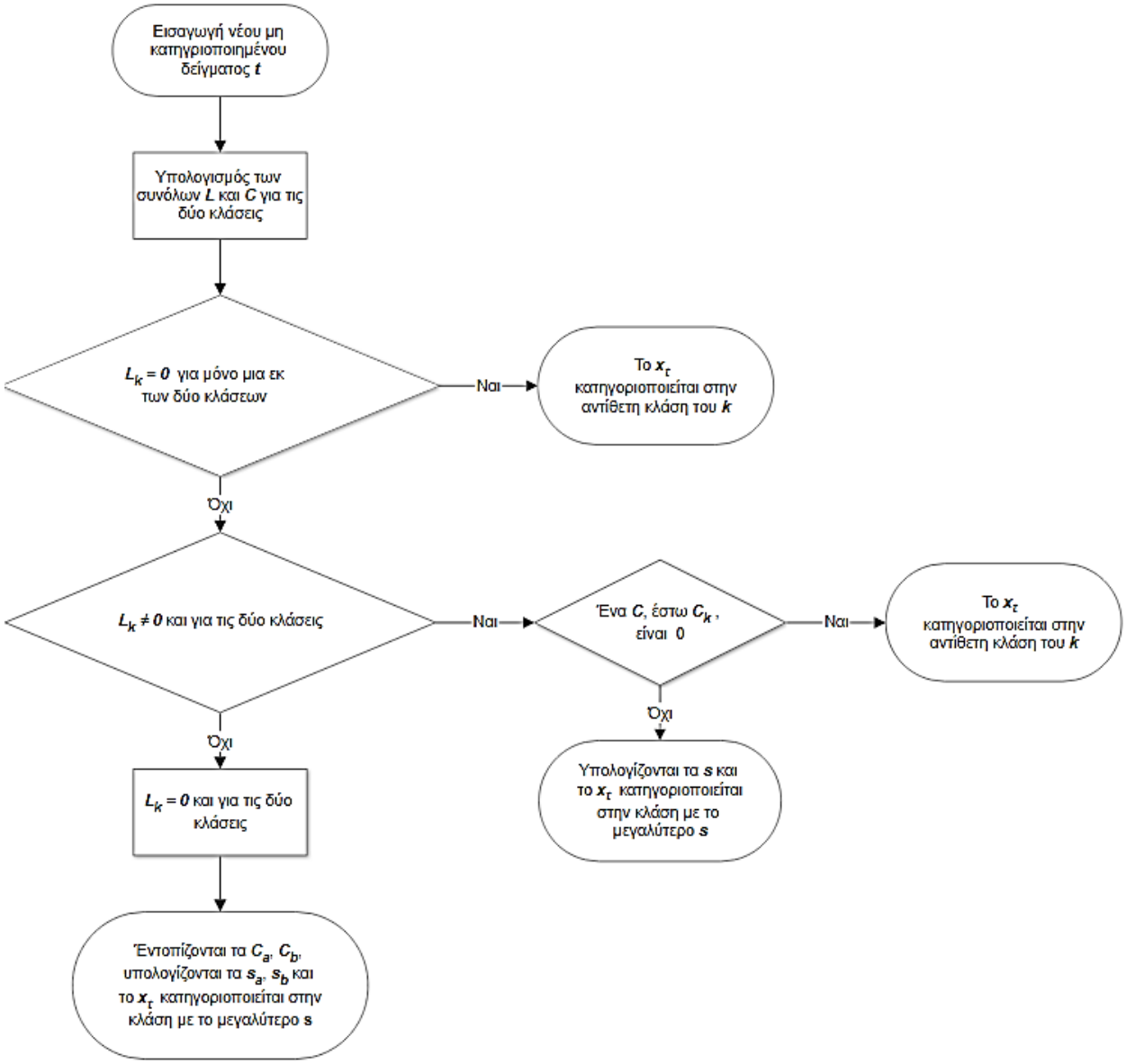
Ο αλγόριθμος ελαχιστοποίησης πραγματοποιείται δύο φορές, οδηγώντας σε δυο σετ λύσεων των y_i, βάσει των οποίων οι συστάδες φιλτράρονται και διαμορφώνεται η τελική επιλογή αυτών.

4.2. Μεθοδολογίες Κατηγοριοποίησης νέων δειγμάτων

Αφού διαμορφωθούν οι συστάδες, αναπτύσσονται μέθοδοι - σύνολα κανόνων για την τοποθέτηση άγνωστων δειγμάτων σε αυτές. Η Μέθοδος Κατηγοριοποίησης 1 βασίζεται στη μεθοδολογία που παρουσιάζεται στη δημοσίευση των Ma *et al.* (2020) ενώ οι Μέθοδοι Κατηγοριοποίησης 2 και 3 αναπτύχθηκαν και δοκιμάστηκαν στην παρούσα εργασία.

4.2.1. Μέθοδος Κατηγοριοποίησης 1

Κάθε νέο δείγμα που τίθεται προς κατηγοριοποίηση σε κλάση, αξιολογείται βάσει συνόλου κριτηρίων και κανόνων διαμορφωμένους από τις επιλεγμένες συστάδες όπως αυτά παρουσιάζονται και στο Σχήμα 4.2.1.



Σχήμα 4.2.1: Διαγραμματική απεικόνιση της αλγοριθμικής λογικής της διαδικασίας κατηγοριοποίησης νέου δείγματος σε κλάση

Αν t είναι νέο δείγμα, του οποίου η κλάση είναι άγνωστη, με σύνολο ιδιοτήτων x_t, δύναται να σχηματιστούν δύο σύνολα L και C που αντιστοιχούν στις συστάδες οι οποίες καλύπτουν το x_t εντός των διευρυμένων (b_i) και των συντηρητικών (r_i) συνόρων τους, αντίστοιχα. Έτσι:

$$L = i \in P \text{ ώστε } \{d(o_i, x_t) \leq b_i\} \quad (4.2.1)$$

$$C = i \in L \text{ ώστε } \{d(o_i, x_t) \leq r_i\} \quad (4.2.2)$$

Καθώς οι συστάδες κάθε κλάσης εξετάζονται ξεχωριστά, δημιουργούνται δύο σύνολα L και C, ένα για κάθε κλάση. Βάσει των τεσσάρων διαμορφωμένων σετ cluster, διαμορφώνονται τρεις δυνατές εκδοχές.

1η Περίπτωση: Αν αυστηρά ένα εκ των δύο σετ L (έστω κλάσης k) είναι κενό ($L^k = \emptyset$), που υποδεικνύει πως το νέο μη κατηγοριοποιημένο δείγμα t βρίσκεται εκτός της επιρροής (των διευρυμένων συνόρων b_i) οποιουδήποτε C_i αυτής της κλάσης, τότε πραγματοποιείται η θεώρηση πως το δείγμα t ανήκει στην αντίθετη κλάση.

2η Περίπτωση: Αν κανένα εκ των δύο L δεν είναι κενό ($L^k \neq \emptyset \quad \forall\, k \in \{1,2\}$), υποδεικνύοντας πως το t καλύπτεται από διευρυμένες συστάδες και των δύο κλάσεων, εμφανίζονται δύο διαφορετικές υποπεριπτώσεις.

> *1η Υποπερίπτωση:* Αν αυστηρά ένα εκ των δύο C δεν είναι κενό (έστω της κλάσης $k, L^k \neq \emptyset$), τότε το t μπορεί να κατηγοριοποιηθεί ως δείγμα κλάσης k, αφού περιλαμβάνεται αποκλειστικά στα συντηρητικά σύνορα συστάδας τάξης k.
>
> *2η Υποπερίπτωση:* Σε οποιαδήποτε άλλη κατάσταση στην οποία η παραπάνω συνθήκη δεν ικανοποιείται, δηλαδή αν τα δύο σετ C είναι είτε αμφότερα κενά ή περιλαμβάνουν συστάδες, υπολογίζεται ένα νέο μέγεθος s.
>
> $$s = \sum_{i \in C} \{w_i(d(o_i, x_t) - r_i)\}^{-1} + \sum_{i \in L \setminus C} \{w_i(d(o_i, x_t) - b_i)\}^{-1} \quad (4.2.3)$$
>
> Το s, βασιζόμενο στη 'σχετική βαρύτητα' των συστάδων και τις ιδιότητές τους (o_i, b_i, r_i) αξιολογεί την σχέση (affinity) του νέου δείγματος με τα cluster των δύο κλάσεων. Προκύπτουν δύο τιμές s, μία για κάθε κλάση, οι οποίες συγκρίνονται μεταξύ τους και καθορίζουν την κλάση του t. Η κλάση για την οποία η τιμή s είναι μεγαλύτερη, θεωρείται πως αποτελεί και την κλάση του t.

3η Περίπτωση: Αν και τα δύο σετ L είναι κενά ($L^k = \emptyset \quad \forall\, k \in \{1,2\}$), γεγονός που υποδηλώνει πως το t δεν καλύπτεται ούτε από διευρυμένες συστάδες καμίας εκ των δύο κλάσεων, η διαδικασία κατηγοριοποίησης διαφέρει.

Υπολογίζοντας τις αποστάσεις του νέου δείγματος t από το συντηρητικό σύνορο εκάστοτε C_i και των δύο κλάσεων:

$$d(o_i, x_t) - r_i$$

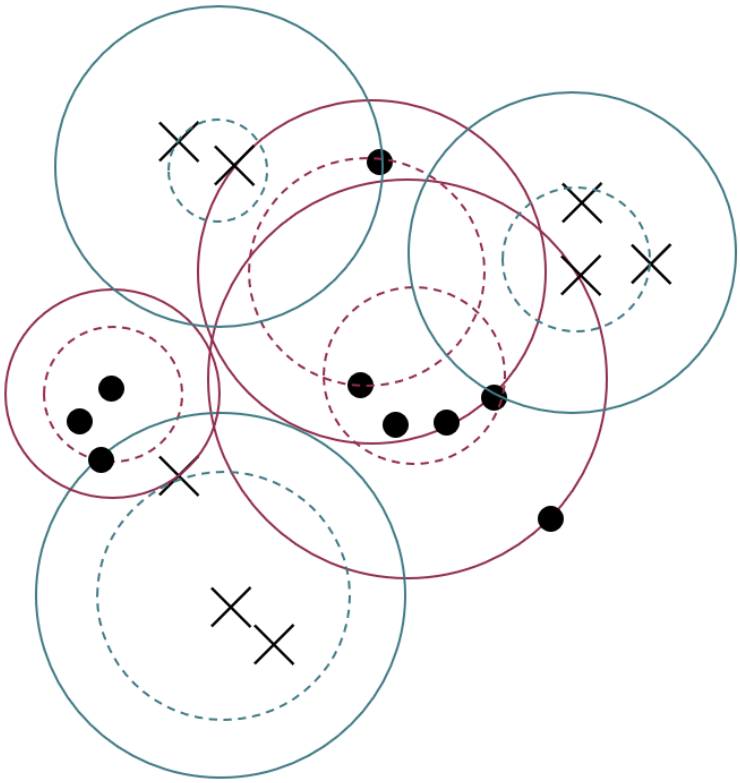
εντοπίζονται οι δύο μικρότερες τιμές εξ αυτών και οι σχετιζόμενες συστάδες, έστω C_a και C_b. Αν τα δύο αυτά cluster ανήκουν στην ίδια κλάση, έστω k, γίνεται η θεώρηση πως και το νέο δείγμα ανήκει στην k.

Αντίθετα, αν τα C_a και C_b ανήκουν σε διαφορετικές κλάσεις, υπολογίζονται δύο τιμές s με:

$$s_\alpha = \{w_\alpha(d(o_\alpha, x_t) - r_\alpha)\}^{-1}, \quad s_b = \{w_b(d(o_b, x_t) - r_b)\}^{-1}$$

Τα s_α, s_b τα οποία αξιολογούν την σχέση του t με τις δύο συστάδες, καθορίζουν την τάξη του νέου δείγματος καθώς το t αποκτά την κλάση του s με την μεγαλύτερη τιμή.

Έχοντας καλύψει όλες τις δυνατές περιπτώσεις για οποιοδήποτε δείγμα t, η διαδικασία σφαιρικής κατηγοριοποίησης νέων παρατηρήσεων έχει ολοκληρωθεί.



Σχήμα 4.2.2: Ενδεικτική απεικόνιση πιθανού συνδυασμού των τελικών cluster των δύο κλάσεων, μετά τη διαδικασία επιλογής

Στο πλαίσιο επέκτασης της μεθοδολογίας και πρότασης εναλλακτικών οδών κατηγοριοποίησης των νέων δειγμάτων, προτείνονται δυο νέοι αλγόριθμοι που αξιοποιούν τις σχετικές θέσεις των κέντρων των νέων συστάδων και τις αποστάσεις των άγνωστων δειγμάτων από αυτές. Λόγω της απλότητας της λογικής τους και της απουσίας σύνθετων υπολογισμών, το υπολογιστικό κόστος είναι μικρότερο συγκριτικά με τον αρχικό προτεινόμενο αλγόριθμο, ο οποίος εφεξής θα καλείται Method 1. Αντίστοιχα, οι δύο νέοι αλγόριθμοι θα καλούνται Method 2 και Method 3.

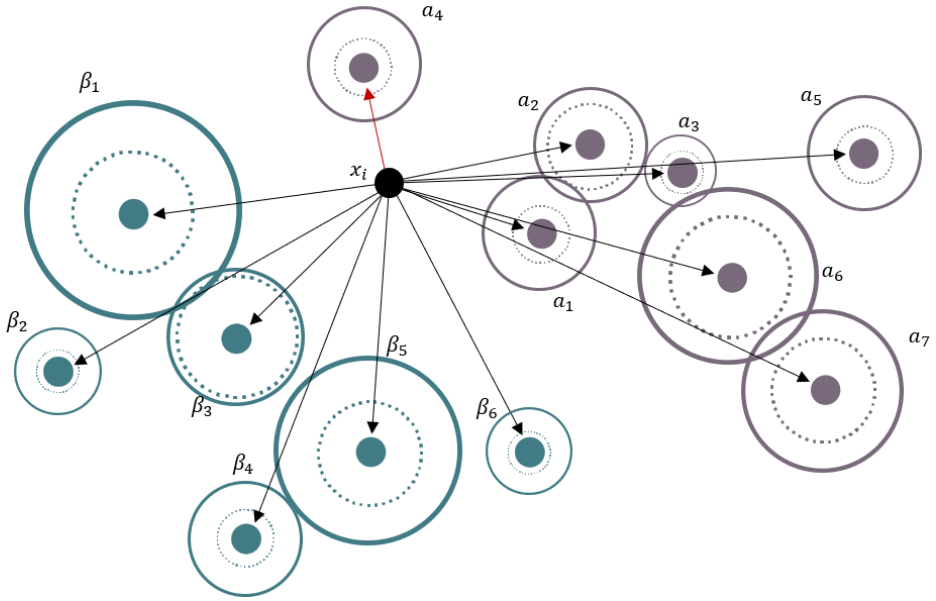
4.2.2. Μέθοδος Κατηγοριοποίησης 2

Έστω i το νέο δείγμα, άγνωστης κατηγορίας, και x_i το σύνολο των ιδιοτήτων-μεταβλητών του. Αρχικά, υπολογίζονται οι αποστάσεις του x_i από το κέντρο κάθε επιλεγμένη συστάδα της εκάστοτε κλάσης. Σχηματίζονται δυο σύνολα με τις υπολογισμένες αποστάσεις, U_1και U_2, ένα για κάθε κλάση. Αν c και d ο αριθμός των συστάδων των κλάσεων 1 και 2, αντίστοιχα, τότε:

$$U_1 = \{d(x_i, a_1), d(x_i, a_2), \dots, d(x_i, a_c)\} \quad (4.2.4)$$

$$U_2 = \{d(x_i, \beta_1), d(x_i, \beta_2), \dots, d(x_i, \beta_d)\} \quad (4.2.5)$$

Όπου $a_1, a_2, \dots, a_c$ τα κέντρα των σχηματισμένων συστάδων της κλάσης 1 και $\beta_1, \beta_2, \dots, \beta_d$ τα κέντρα των σχηματισμένων συστάδων της κλάσης 2.



Σχήμα 4.2.3: Απεικόνιση των cluster, του νέου δείγματος και των υπολογιζόμενων αποστάσεων μεταξύ αυτών στην περίπτωση της εναλλακτικής Μεθόδου 2. Οι δύο αποστάσεις που τίθενται προς σύγκριση είναι οι $d(x_i, a_4)$ και $d(x_i, \beta_3)$. Καθώς $d(x_i, a_4) < d(x_i, \beta_3)$ το δείγμα i κατηγοριοποιείται στην κλάση 1.

Από κάθε σύνολο επιλέγεται η μικρότερη τιμή, δηλαδή η συστάδα κάθε κλάσης, το οποίο απέχει λιγότερο από το νέο δείγμα. Πραγματοποιείται σύγκριση μεταξύ των δύο τελικών τιμών. Αν η τιμή που αφορά στη συστάδα της κλάσης 1 είναι μικρότερη το δείγμα κατηγοριοποιεί στην κλάση 1 και το αντίστροφο:

$$class_i = \begin{cases} 1, & \min(U_1) < \min(U_2) \\ 2, & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases} \quad (4.2.6)$$

4.2.3. Μέθοδος Κατηγοριοποίησης 3

Εναλλακτικά, υπολογίζονται τα αθροίσματα:

$$sum_1 = \frac{\sum_{k=1}^{n} \frac{1}{1 + d(x_i, a_k)}}{n} \tag{4.2.7}$$

$$sum_2 = \frac{\sum_{k=1}^{m} \frac{1}{1 + d(x_i, b_k)}}{m} \tag{4.2.8}$$

τα οποία εκτιμούν το αντίστροφο της σταθμισμένης απόστασης όλων των συστάδων της εκάστοτε κλάσης από το νέο δείγμα. Σε αυτή την περίπτωση, κριτήριο δεν αποτελεί η ελάχιστη απόσταση αλλά η μέγιστη δυνατή συγγένεια μεταξύ κάθε άγνωστου δείγματος και των συστάδων. Βάσει αυτού, αν η τιμή που αφορά στη συστάδα της κλάσης 1 είναι μεγαλύτερη, το δείγμα κατηγοριοποιείται στην κλάση 1 και το αντίστροφο:

$$class_i = \begin{cases} 1, & sum_1 > sum_2 \\ 2, & αλλιώς \end{cases} \tag{4.2.9}$$

Κεφάλαιο 5
Μελέτες περιπτώσεων

Για την μελέτη του μηχανισμού δράσης της προτεινόμενης μεθοδολογίας και της ικανότητάς της για πρόβλεψη ιδιοτήτων σε σύνολα δεδομένων (datasets), αξιοποιούνται οι δημοσιεύσεις των Forest *et al*. (2019), Liu *et al*. (2013), Kotzabasaki *et al*. (2020) και Papadiamantis *et al*. (2020). Στα datasets αυτά περιέχονται σύνολα δειγμάτων νανοσωματιδίων διαφόρων ιδιοτήτων, με δεδομένα τοξικότητας σε κύτταρα. Ο σκοπός της αναπτυσσόμενης μεθοδολογίας συνίσταται στην αξιοποίηση των ιδιοτήτων για την πρόβλεψη της τοξικότητας του κάθε δείγματος. Με βάση την τοξικότητά τους, τα δείγματα χωρίζονται σε δύο κλάσεις, 1 και 2, με τον συνήθη διαχωρισμό: 1 για τα μη τοξικά, 2 για τα τοξικά δείγματα. Σε μερικά σύνολα δεδομένων, η τοξικότητα μετράται (πειραματικά) αρχικά μέσω συνεχούς και όχι δυαδικής μεταβλητής. Έτσι, στις περιπτώσεις αυτές εφαρμόζεται φίλτρο με 'κατώφλι τιμών' (threshold) με βάση τις υποδείξεις της εκάστοτε δημοσίευσης, κατηγοριοποιώντας τα δείγματα στις δυο κλάσεις ανάλογα με την τιμή της μετρούμενης τοξικότητας. Αξίζει να σημειωθεί πως, αν και η αναπτυσσόμενη μεθοδολογία βασίζεται σε συνεχείς μεταβλητές εισόδου, η κανονικοποίηση των τιμών των ιδιοτήτων (στο πλαίσιο προεπεξεργασίας των dataset) επιτρέπει την παρουσία και αριθμημένων κατηγορικών μεταβλητών (της μορφής 0,1,2 κλπ.). Τα σύνολα των δεδομένων που χρησιμοποιήθηκαν παρουσιάζονται παρακάτω.

5.1. Σύνολο νανοσωματιδίων υδρ(οξειδείων) μετάλλων

Στη δημοσίευση '*Towards an alternative to nano-QSAR for nanoparticle toxicity ranking in case of small datasets*' [47], για τις ανάγκες μελέτης της μεθοδολογίας QSAR, συντίθενται 25 νανοσωματίδια από 6 οικογένειες οξειδίων: SiO_2, TiO_2, CeO_2, AlOOH, ZnO, $Ni(OH)_2$.

Οι ιδιότητες που μετρήθηκαν ή υπολογίστηκαν κατηγοριοποιούνται είτε ως μή διαστατικές, είτε ως διαστατικές. Στην πρώτη κατηγορία ανήκουν οι: $x(H_2O)$, n_oxy_M, r_cat *και* EN_M. Η ιδιότητα $x(H_2O)$ προκύπτει από τον τύπο $M(n^+)O_{\frac{n}{2}}.x(H_2O)$ και χρησιμοποιείται με σκοπό να υπάρξει διάκριση μεταξύ των οξειδίων και των υδροξειδίων. Οι n_oxy_M και r_cat εκφράζουν βαθμό οξείδωσης του μετάλλου και την ακτίνα του κατιόντος, αντίστοιχα, ενώ η EN_M την ηλεκτραρνητικότητα του Pauling.

Στη δεύτερη κατηγορία περιλαμβάνονται μετρούμενες ιδιότητες που εξαρτώνται άμεσα από το μέγεθος των σωματιδίων. Σε αυτή ανήκουν:

- Η μέγιστη και ελάχιστη ακτίνα του κάθε σωματίδιου ($dmax$ και $dmin$),
- ο συντελεστής σχήματος του σωματιδίου SF και ο διορθωμένος συντελεστής CSF εκφρασμένοι ως $SF = dmax/dmin$ και $CSF = log(d_{axis}/d_{perp})$ αντίστοιχα, όπου d_{axis} το μήκος του άξονα περιστροφής του νανοσωματιδίου και d_{perp} το μήκος του κάθετου σε αυτού άξονα,
- η διαλυτότητα των υλικών $s(\log Mtot)$ και το ηλεκτρικό δυναμικό $zeta$, μετρούμενα σε ουδέτερες συνθήκες pH,
- η ειδική επιφάνεια των σωματιδίων, (specific surface area, SSA) και
- ο βαθμός συσσωμάτωσης των σωματιδίων μετρούμενος σε κλίμακα 0 ως 2, ξεκινώντας από μηδενική συσσωμάτωση έως και πυκνά συσσωματώματα.

Συνολικά, το σετ δεδομένων περιέχει 12 φυσικοχημικές ιδιότητες για 25 διαφορετικά δείγματα, με σύσταση διαμορφωμένη έτσι, ώστε κάθε οικογένεια μετάλλων να περιλαμβάνει τουλάχιστον δύο διαφορετικά ως προς το μέγεθος ή το σχήμα σωματίδια, αποσκοπώντας στην ανεξαρτησία των χημικών και διαστατικών ιδιοτήτων.

Ως μεταβλητή στόχος, που εκφράζει την τοξικότητα των δειγμάτων, επιλέγεται και υπολογίζεται η απελευθέρωση της γαλακτικής αφυδρογονάσης (Lactate Dehydrogenase, LDH) σε κυτταρικό περιβάλλον χρησιμοποιώντας την πρότυπη μεθοδολογία αξιολόγησης CytoTox-96™ Homogeneous Membrane Integrity Assay. Στο πλαίσιο μέτρησης της τοξικότητας, πραγματοποιήθηκαν τρία ανεξάρτητα τετραπλά πειράματα. Από αυτά, προέκυψε μια μέση τιμή σήματος η οποία κανονικοποιήθηκε, βάσει των μετρήσεων με 'τυφλά' δείγματα (που δεν περιείχαν νανοσωματίδια). Για τις ανάγκες κατηγοριοποίησης των δειγμάτων ως τοξικά ή μη, εφαρμόζεται 'κατώφλι' στην τιμή της απελευθερούμενης LDH. Κάθε υπολογιζόμενη τιμή μικρότερη του 1.5 καθιστά το νανοσωματίδιο μη τοξικό, ενώ κάθε τιμή μεγαλύτερη του 1.5, τοξικό.

Στη συγκεκριμένη δημοσίευση, για τις ίδιες φυσικοχημικές ιδιότητες, υπολογίστηκαν δυο ακόμη μεταβλητές-στόχοι: η παραγωγή της κυτοσίνης του παράγοντα νέκρωσης όγκων α (TNF-a) και η παραγωγή δραστικών μορφών οξυγόνου (reactive oxygen species, ROS). Στο εξής, για λόγους συντομίας, το σύνολο θα καλείται 'MeHydOx'.

5.2. Σύνολο υπερ-παραμαγνητικών νανοσωματιδίων οξειδίου του σιδήρου

Ως δεύτερο σετ δεδομένων αξιοποιείται το σύνολο που έχει συσταθεί στο πλαίσιο της δημοσίευσης '*QSAR modeling of the toxicity classification of superparamagnetic*

iron oxide nanoparticles (SPIONs) in stem-cell monitoring applications: an integrated study from data curation to model development' των Kotzabasaki *et al* (2020) [37].

Αυτό αποτελείται από υπέρ-παραμαγνητικά νανοσωματίδια οξειδίου του σιδήρου (SPIONs) με πυρήνα μαγκεμίτη ($\gamma - Fe_2O_3$) ή μαγνητίτη (Fe_3O_4) και μέγεθος από 10 ως 100 nm. Η επιφάνειά τους είναι πιθανό να είναι επιστρωμένη με κάποιο οργανικό ή ανόργανο υλικό καθώς η ύπαρξη της επίστρωσης μειώνει την τοξικότητα των σωματιδίων, η οποία εξαρτάται κυρίως από το μέγεθος το φορτίο, τη δόση ή τις επιφανειακές ιδιότητες.

Για τη σύσταση του συνόλου δεδομένων που αξιοποιείται στη δημοσίευση μελετήθηκε πληθώρα βιβλιογραφικών αναφορών και μελετών και εντοπίστηκαν αυτές που περιείχαν μετρήσεις φυσικοχημικών ιδιοτήτων. Εξ αυτών, προέκυψαν δεδομένα για 15 νανοσωματίδια της κατηγορίας SPIONs. Οι ιδιότητες που εξετάζονται περιλαμβάνουν:

- τη φύση του πυρήνα (μαγνητίτης ή μαγκεμίτης), κωδικοποιημένη σε δυαδική μορφή, 0 και 1,
- το ηλεκτρικό δυναμικό *zeta*,
- το μέγεθος του νανοσωματιδίου σε nm,
- η δύναμη του μαγνητικού πεδίου που δημιουργείται, μετρούμενη σε T,
- η συγκέντρωση του σιδήρου σε κάθε κύτταρο σε pg, εκφράζοντας το βαθμό εισροής των SPIONs στα κύτταρα,
- το relaxivity, μετρούμενο σε $s^{-1}mM^{-1}$, που αντιστοιχεί στο βαθμό χαλάρωσης ως συνάρτηση της συγκέντρωσης.

Ως μεταβλητή στόχος, σχετιζόμενη με την τοξικότητα του εκάστοτε σωματιδίου, χρησιμοποιείται η βιωσιμότητα του κυττάρου (cell viability), δηλαδή η ποσοτικοποίηση του αριθμού των ζωντανών κυττάρων ελέγχου, μετά το πέρας των πειραμάτων. Για την εξυπηρέτηση του σκοπού της δημοσίευσης, κάθε δείγμα νανοσωματιδίου με βιωσιμότητα μεγαλύτερη του 75% θεωρείται μη τοξικό. Σε αντίθετη περίπτωση, θεωρείται τοξικό. Η κατηγοριοποίηση αυτή, κωδικοποιείται αριθμητικά αποδίδοντας της τιμή 1 και 2, αντίστοιχα. Στο εξής, το σύνολο θα καλείται 'SPIONs'.

5.3. Σύνολο μεταλλικών οξειδίων Α

Στη δημοσίευση '*Development of structure–activity relationship for metal oxide nanoparticles*' των Liu *et al*. (2013) [48], παρουσιάζεται μια μεθοδολογία τύπου nano-SAR πρόβλεψης τοξικότητας νανοσωματιδίων χρησιμοποιώντας ένα σετ δεδομένων ιδιοτήτων- τοξικότητας 23 νανοσωματιδίων μεταλλικών οξειδίων.

Στο αρχικό σετ εντοπίζονταν πληθώρα μετρήσεων τοξικολογικών αποκρίσεων τόσο σε επιθηλιακά ((BEAS-2B) όσο και σε μυελοειδή κύτταρα (RAW 264.7))

βάσει δοκιμών σε καθορισμένα πρότυπα. Οι μετρήσεις προέρχονται από μόνο- και πολυπαραμετρικές δοκιμές (MTS, ATP, LDH και HTS, αντίστοιχα). Συνδυάζοντας μετρήσεις 7 διαφορετικών δοκιμών για δυο κυτταρικές σειρές, μέσω στατιστικής ανάλυσης προκύπτει η μεταβλητή-στόχος που περιγράφει την τοξικότητα του εκάστοτε νανοσωματιδίου, η οποία και αξιοποιείται στη συγκεκριμένη δημοσίευση για την διαμόρφωση του μοντέλου υπολογισμού της πιθανότητας ένα νανοσωματίδιο να είναι τοξικό.

Μέσω ενός μοντέλου Support Vector Machine (SVM) και αξιοποιώντας το βέλτιστο συνδυασμό δύο μεταβλητών μεταξύ των διαθέσιμων, υπολογίστηκε η συνάρτηση διάκρισης (discriminant function) που συνδέει όλες τις μεταβλητές-περιγραφείς μαζί, $f(x)$:

$$f(x) = \sum_{\iota=1}^{6} \alpha_i e^{-2[(x_{i,1}-x_1)^2+(x_{i,2}-x_2)^2]} + b \qquad (5.3.1)$$

καθώς και η πιθανότητα τοξικότητας συναρτήσει της:

$$P(T|x) = \frac{1}{1+e^{-f(x)}} \qquad (5.3.2)$$

Με σκοπό την κατηγοριοποίηση της μεταβλητής στόχου σε δύο κλάσεις, 'Τοξικό' και 'Μη Τοξικό' δείγμα, χρησιμοποιείται το αριθμητικό 'κατώφλι' 0.5. Έτσι, κάθε δείγμα του οποίου η πιθανότητα να είναι τοξικό υπολογίζεται >0.5 από το μοντέλο, το δείγμα χαρακτηρίζεται τοξικό, και το αντίθετο.

Οι ιδιότητες που επιλέγονται σε κάθε μοντέλο και δοκιμή μπορεί να είναι φυσικοχημικές ή διαστατικές, βασιζόμενες στο είδος του νανοσωματίδιου. Μπορούν να διακριθούν στις εξής κατηγορίες:

- θεμελιώδεις ιδιότητες μεταλλικών οξειδίων. Σε αυτές εντάσσονται ο αριθμός των ατόμων μετάλλου και οξυγόνου στο μόριο, η ατομική μάζα του μετάλλου, το μοριακό βάρος του σωματιδίου καθώς και η ηλεκτραρνητικότητα του μετάλλου και του μεταλλικού οξειδίου
- ενεργειακές ιδιότητες των νανοσωματιδίων, οι οποίες περιλαμβάνουν το χημικό δυναμικό, τη σκληρότητα και το βαθμό ηλεκτροφιλίας τους
- το κύριο μέγεθος του μεταλλικού μορίου σε nm καθώς και 3 υδροδυναμικά μεγέθη μετρημένα σε διαφορετικά βιολογικά περιβάλλοντα: νερού, BEGM και DMEM ($d_{water}, d_{BEGM}, d_{DMEM}$)
- φυσικοχημικές ιδιότητες όπως: ενέργειες της ζώνης σθένους, της ζώνης αγωγιμότητας και η ενέργεια πλέγματος του οξειδίου ($E_V, E_C, \Delta H_{lat}$), ενέργεια ατομοποίησης και ιονισμού του μεταλλικού οξειδίου και του πρώτου μορίου του μετάλλου ($E_{Amz}, \Delta H_{IE}, \Delta H_{IE,1+}$), το ισοηλεκτρικό σημείο και το δυναμικό zeta μετρούμενο σε νερό με $pH = 7.4$ (IEP, ZP), η πρότυπη ενθαλπία σχηματισμού και εξάχνωσης ($\Delta H_{sub}, \Delta H_{sf}$) καθώς και ο ιονικός δείκτης του κατιόντος του μετάλλου (Z^2/r).

Στο εξής, το σύνολο θα καλείται ‘MeOx’.

5.4. Σύνολο μεταλλικών οξειδίων Β

Στη δημοσίευση ‘Predicting Cytotoxicity of Metal Oxide Nanoparticles Using Isalos Analytics Platform’ των Papadiamantis *et al*. [49], παρουσιάζεται ένα σύνολο δεδομένων νανοσωματιδίων μεταλλικών οξειδίων και χρησιμοποιείται για την ανάπτυξη ενός προβλεπτικού μοντέλου τοξικότητας τύπου QNAR. Είκοσι τέσσερα διαφορετικά μεταλλικά οξείδια ελέγχονται ως προς την τοξικότητα που προκαλούν σε δυο είδη κυττάρων, BEAS-2B και RAW 264.7.

Οι ανεξάρτητες μεταβλητές που χρησιμοποιούνται για την πρόβλεψη περιλαμβάνουν φυσικοχημικούς και δομικούς χαρακτηρισμούς όπως διάμετρος πυρήνα, ειδική επιφάνεια, z-δυναμικό, κρυσταλλική δομή ή υδροδυναμικό μέγεθος καθώς επίσης και ενέργειες σθένους ή σχηματισμού και ηλεκτραρνητικότητα. Οι παραπάνω παράμετροι, οι οποίες προέρχονται από σχετική δημοσίευση των Zhang *et al*. [50], εμπλουτίζονται περαιτέρω με ακόμα 62 υπολογιζόμενες ανεξάρτητες μεταβλητές βάσει αναπτυγμένης μεθοδολογίας υπολογισμού descriptors η οποία αξιοποιεί την κρυσταλλική δομή των υλικών. Οι τιμές που υπολογίστηκαν αφορούν την χημική σύνθεση, την ενέργεια δυναμικού, την τοπολογία, την ενέργεια πλέγματος και το μέγεθος των σωματιδίων. Συνολικά, το σύνολο δεδομένων που διαμορφώθηκε περιλάμβανε 77 descriptors. Εξ αυτών, οι 9 απομακρύνθηκαν λόγω εξαιρετικά χαμηλής διακύμανσης (low variance).

Μεταβλητή στόχος y για το σύνολο δεδομένων αποτελούν οι *in vitro* τοξικολογικές μετρήσεις τη απελευθέρωσης της γαλακτικής αφυδρογονάσης (Lactate Dehydrogenase, LDH) και της τριφωσφορικής αδενοσίνης (Adenosine Triphosphate, ATP), μια εκ των δύο σε κάθε καταγεγραμμένη δοκιμή. Ο τύπος μέτρησης που χρησιμοποιείται ως endpoint, LDH ή ATP, αξιοποιείται ως κατηγορική μεταβλητή-ιδιότητα καθώς εντοπίστηκε ισχυρή συσχέτιση μεταξύ αυτού και της υπολογιζόμενης τοξικότητας. Οι μετρήσεις των LDH ή ATP μεταφράζονται σε ποσοστό βιωσιμότητας (cell viability %). Για την κατηγοριοποίηση των δειγμάτων ως τοξικά ή μη τοξικά, χρησιμοποιείται το ‘κατώφλι’ (threshold) ύψους 50%: οποιαδήποτε τιμή κάτω από 50% χαρακτηρίζει το δείγμα ως τοξικό, ενώ τιμή άνω του 50% καθιστά το δείγμα μη τοξικό. Στο εξής, το σύνολο θα καλείται ‘Cytotox’.

Κεφάλαιο 6
Παρουσίαση αποτελεσμάτων

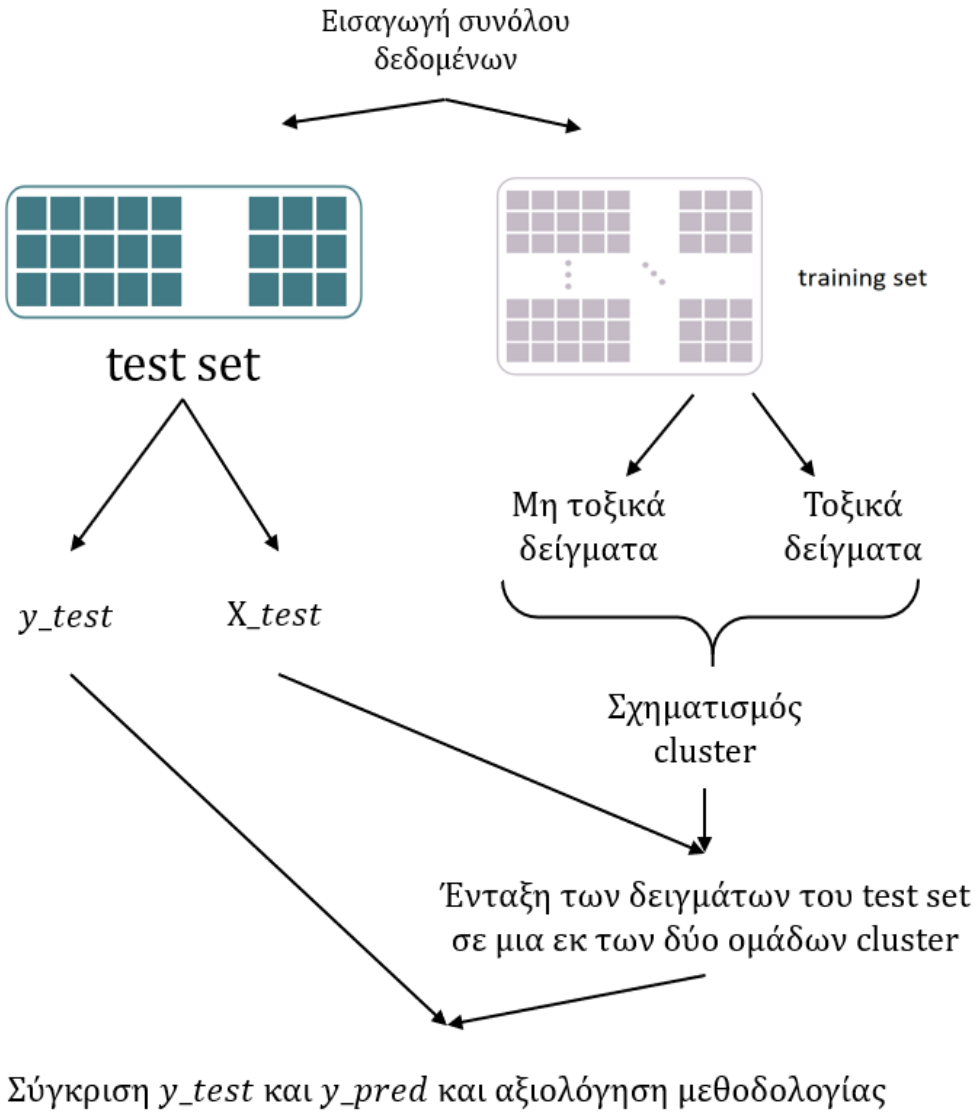
6.1. Πρόβλεψη τοξικότητας χρησιμοποιώντας όλες τις διαθέσιμες μεταβλητές

Η προτεινόμενη μεθοδολογία εφαρμόζεται σε κάθε σύνολο δεδομένων ξεχωριστά βάσει των βημάτων που περιεγράφηκαν και τις μεθόδους επικύρωσης που μπορούν να εφαρμοστούν. Σε αυτό το στάδιο αξιολόγησης της μεθόδου, τα σύνολα δεδομένων δεν υπόκεινται σε κάποια μείωση των διαστάσεών τους μέσω επιλογής μεταβλητών (Feature Selection) ή ανάλυση κυρίων συνιστωσών (Principal Component Analysis).

Η διαδικασία που ακολουθείται για κάθε σύνολο δεδομένων μπορεί να συνοψιστεί στα εξής βήματα και περιγράφεται γραφικά στο Σχήμα 6.1.1:

1. Εισαγωγή του νέου συνόλου υπό τη μορφή πίνακα. Οι γραμμές του πίνακα αντιστοιχούν στα διαφορετικά δείγματα του συνόλου και οι στήλες περιέχουν τις ανεξάρτητες μεταβλητές-ιδιότητες του κάθε δείγματος σε αυτό. Η πρώτη εκ των στηλών αντιστοιχεί στη μεταβλητή πρόβλεψης, την εξαρτημένη μεταβλητή-στόχο που εκφράζει την τοξικότητα του κάθε δείγματος νανοσωματιδίων, κωδικοποιημένη σε δυαδική μορφή, με τον συμβολισμό 1 για τα μη τοξικά δείγματα και 2 για τα τοξικά.
2. Το σύνολο των ιδιοτήτων κανονικοποιείται εντός των ορίων 0 και 1 και χωρίζεται σε δύο υποσύνολα, σύνολο εκπαίδευσης και ελέγχου με προκαθορισμένη μέθοδο (τυχαία ή Kennard Stone)
3. Το σύνολο εκπαίδευσης χωρίζεται βάσει της γνωστής μεταβλητής πρόβλεψης των δειγμάτων και δημιουργούνται δύο ομάδες δειγμάτων εκπαίδευσης, μια για τα τοξικά και μια για τα μη τοξικά δείγματα.
4. Βάσει των δύο σχηματιζόμενων ομάδων και μέσω των 5 βημάτων της προτεινόμενης μεθοδολογίας (§4.1), δημιουργούνται δύο ομάδες συστάδων που περιλαμβάνουν όλα τα δείγματα του συνόλου εκπαίδευσης και χαρακτηρίζονται από συντεταγμένες στο n-διάστατο χώρο, όπου n ο αριθμός των ανεξάρτητων μεταβλητών-ιδιοτήτων των δειγμάτων. Κάθε διάσταση αντιστοιχεί και σε μια κανονικοποιημένη ιδιότητα-μεταβλητή.
5. Από τις επιλεγμένες συστάδες στο τέλος του 5ου σταδίου της μεθοδολογίας, διαμορφώνονται τρεις διαφορετικές μέθοδοι κατηγοριοποίησης των δειγμάτων του συνόλου επικύρωσης (X_test), τα δείγματα του οποίου έχουν άγνωστο endpoint (y_test). Η κλάση του κάθε δείγματος ορίζεται βάσει των κριτηρίων μιας εκ των τριών μεθοδολογιών που περιεγράφηκαν στις § 4.2.1-4.2.3.

6. Οι προβλέψεις της τοξικότητας (y_pred) συγκρίνονται με την πραγματική κλάση (y_test) με σκοπό την αξιολόγηση της μεθόδου.



Σχήμα 6.1.1: Σχηματική απεικόνιση της διαδικασίας εξαγωγής αποτελεσμάτων χωρίς μείωση των διαστάσεων του συνόλου προς εξέταση

6.1.1. Σύνολο δεδομένων MeHydOx

Το σύνολο δεδομένων MeHydOx περιλαμβάνει 25 δείγματα με 12 μεταβλητές-ιδιότητες και ένα endpoint δυαδικής φύσης. Ο έλεγχος της προτεινόμενης μεθοδολογίας στο συγκεκριμένο σύνολο πραγματοποιείται μέσω εσωτερικής και εξωτερικής αξιολόγησης.

Εξωτερική αξιολόγηση

Το σύνολο δεδομένων χωρίζεται με χρήση της μεθόδου $Kennard\ Stone$ με διαφορετικούς λόγους διαμέρισης train:test set (train ratio) για λόγους ανάλυσης ευαισθησίας, που κυμαίνονται από 0.6 ως 0.8. Για κάθε χώρισμα και διαμόρφωση cluster, πραγματοποιούνται τρεις διαφορετικές προβλέψεις, βάσει των τριών εναλλακτικών μεθόδων κατηγοριοποίησης νέων δειγμάτων. Στον πίνακα σύγχυσης, ως $True$ θεωρούνται τα μη τοξικά δείγματα και ως $False$ τα τοξικά. Στους πίνακες 6.1.1-6.1.3 περιέχονται τα στατιστικά αποτελέσματα των δοκιμών για τις τρεις εναλλακτικές μεθόδους. Σημειώνεται πως σε κάθε σετ δοκιμών δεν αλλάζει κάτι πέρα από το τρόπο κατανομής των δειγμάτων στα δύο υποσύνολα εκπαίδευσης και επικύρωσης.

Πίνακας 6.1.1: Στατιστικά αποτελέσματα με εφαρμογή Kennard Stone για διαφορετικά training ratios στο σύνολο MeHydOx με τη μέθοδο 1

Train Ratio	Test samples	Accuracy	Sensitivity	Specificity	Error rate	MCC	Confusion Matrix
0.60	10	0.80	1.00	0.50	0.20	0.61	$\begin{bmatrix}6 & 2\\2 & 2\end{bmatrix}$
0.65	9	0.89	1.00	0.75	0.11	0.79	$\begin{bmatrix}5 & 0\\1 & 3\end{bmatrix}$
0.70	7	0.86	1.00	0.75	0.14	0.75	$\begin{bmatrix}3 & 0\\1 & 3\end{bmatrix}$
0.75	6	0.83	1.00	0.67	0.17	0.71	$\begin{bmatrix}3 & 0\\1 & 2\end{bmatrix}$
0.80	5	0.80	1.00	0.67	0.2	0.67	$\begin{bmatrix}2 & 0\\1 & 2\end{bmatrix}$

Η ακρίβεια των προβλέψεων κυμαίνεται σε τιμές [0.80, 0.89, 0.86, 0.83, 0.80] για λόγους διαμέρισης [0.60, 0.65, 0.70, 0.75, 0.80], αντίστοιχα. Η καλύτερη πρόβλεψη δίνεται για κλάσμα 0.65, δηλαδή όταν τα 2/3 του δειγματικού χώρου αξιοποιούνται για εκπαίδευση και το 1/3 για επικύρωση του μοντέλου. Ανεξάρτητα από το κλάσμα διαμέρισης, η ακρίβεια των προβλέψεων στην περίπτωση της Μεθόδου Κατηγοριοποίησης 1 είναι ικανοποιητική. Επιτυγχάνοντας την σωστή πρόβλεψη όλων των μη τοξικών δειγμάτων με απόλυτη ευαισθησία, η ειδικότητα κυμαίνεται σε χαμηλότερα επίπεδα, δυσχεραίνοντας την πρόβλεψη των τοξικών νανοσωματιδίων. Ο MCC, με τιμές [0.61, 0.79, 0.75, 0.71, 0.67], προσφέροντας μια συνολικότερη εικόνα για την προβλεπτική ικανότητα, υποδηλώνει πως το μοντέλο είναι αποδοτικό.

Πίνακας 6.1.2: Στατιστικά αποτελέσματα με εφαρμογή Kennard Stone για διαφορετικά training ratios στο σύνολο MeHydOx με τη μέθοδο 2

Train Ratio	Test samples	Accuracy	Sensitivity	Specificity	Error rate	MCC	Confusion Matrix
0.60	10	0.90	0.83	1.00	0.10	0.82	$\begin{bmatrix}5 & 1\\0 & 4\end{bmatrix}$

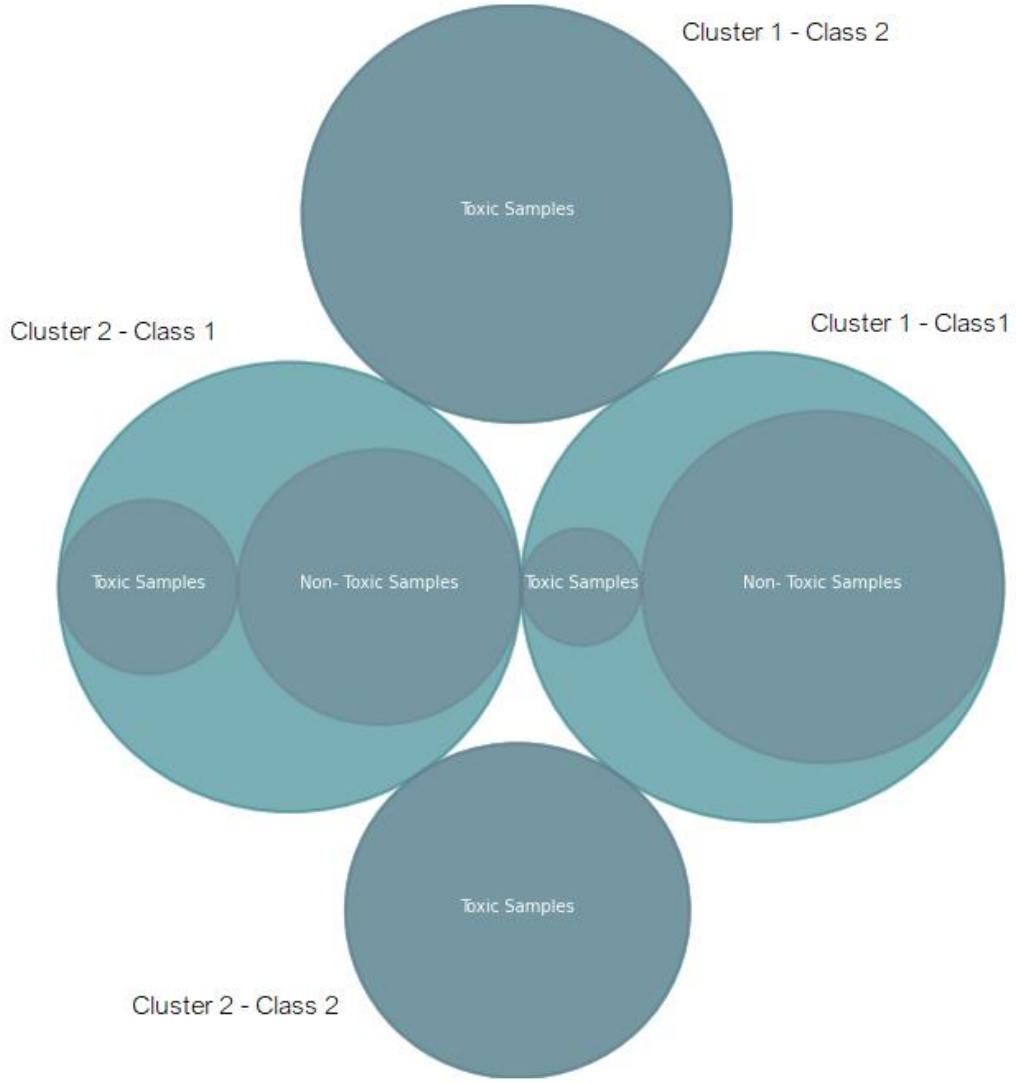
0.65	9	0.89	0.80	1.00	0.11	0.80	$\begin{bmatrix} 4 & 1 \\ 0 & 4 \end{bmatrix}$
0.70	7	0.86	0.67	1.00	0.14	0.73	$\begin{bmatrix} 2 & 1 \\ 0 & 4 \end{bmatrix}$
0.75	6	0.83	0.67	1.00	0.17	0.71	$\begin{bmatrix} 2 & 1 \\ 0 & 3 \end{bmatrix}$
0.80	5	0.80	0.5	1.00	0.20	0.61	$\begin{bmatrix} 1 & 1 \\ 0 & 3 \end{bmatrix}$

Στην περίπτωση της εναλλακτικής Μεθόδου Κατηγοριοποίησης 2, η ακρίβεια των προβλέψεων κυμαίνεται σε τιμές [0.90, 0.89, 0.86, 0.83, 0.80] για κλάσματα διαμέρισης [0.60, 0.65, 0.70, 0.75, 0.80], αντίστοιχα, με την καλύτερη πρόβλεψη να δίνεται για κλάσμα 0.6. Σε αντίθεση με την Μέθοδο 1, η ευαισθησία επιτυγχάνει χαμηλότερες τιμές και δεν χαρακτηρίζεται από πλήρη επιτυχία. Ωστόσο, η ειδικότητα αγγίζει την τιμή 1 σε κάθε διαφορετική δοκιμή, προβλέποντας τα μη τοξικά δείγματα με απόλυτη επιτυχία και καθιστώντας το μοντέλο ιδιαίτερα αποτελεσματικό. Αύξηση παρατηρείται και στις τιμές του MCC, υποδηλώνοντας καλύτερη προβλεπτική ικανότητα στην περίπτωση της Μεθόδου 2 με τιμές [0.82, 0.80, 0.73, 0.71, 0.61].

Πίνακας 6.1.3: Στατιστικά αποτελέσματα με εφαρμογή Kennard Stone για διαφορετικά training ratios στο σύνολο MeHydOx με τη μέθοδο 3

Train Ratio	Test samples	Accuracy	Sensitivity	Specificity	Error rate	MCC	Confusion Matrix
0.60	10	0.90	0.83	1.00	0.10	0.82	$\begin{bmatrix} 5 & 1 \\ 0 & 4 \end{bmatrix}$
0.65	9	0.89	0.80	1.00	0.11	0.80	$\begin{bmatrix} 4 & 1 \\ 0 & 4 \end{bmatrix}$
0.70	7	0.86	0.67	1.00	0.14	0.73	$\begin{bmatrix} 2 & 1 \\ 0 & 4 \end{bmatrix}$
0.75	6	0.83	0.67	1.00	0.17	0.71	$\begin{bmatrix} 2 & 1 \\ 0 & 3 \end{bmatrix}$
0.80	5	0.80	0.50	1.00	0.20	0.61	$\begin{bmatrix} 1 & 1 \\ 0 & 3 \end{bmatrix}$

Στην περίπτωση της εναλλακτικής Μεθόδου Κατηγοριοποίησης 3, η ακρίβεια των προβλέψεων κυμαίνεται σε τιμές [0.90, 0.89, 0.86, 0.83, 0.80] για κλάσματα διαμέρισης [0.60, 0.65, 0.70, 0.75, 0.80], αντίστοιχα, με την καλύτερη πρόβλεψη να δίνεται για κλάσμα 0.6. Οι τιμές ταυτίζονται πλήρως με τις τιμές της *Μεθόδου* 2, οδηγώντας στις ίδιες προβλέψεις και παρατηρήσεις.



Σχήμα 6.1.2: Απεικόνιση του δειγματικού χώρου σε σχέση με τα cluster των δυο κλάσεων στα οποία εντοπίζονται

Για την κατανόηση της λειτουργίας της προτεινόμενης μεθοδολογίας παρουσιάζεται ένα παράδειγμα. Υπενθυμίζεται ότι η δημιουργία των cluster είναι ανεξάρτητη από την κατάταξη των δειγμάτων του συνόλου επικύρωσης (3 εναλλακτικές μεθοδολογίες). Κατά την μοντελοποίηση με διαμέριση Kennard Stone, με λόγο 0.65, δημιουργούνται 4 cluster, 2 για κάθε κλάση (είδος τοξικότητας). Όπως παρουσιάζεται στην απεικόνιση της κατανομής του δειγματικού χώρου στα cluster (Σχήμα 6.1.2), τα cluster των τοξικών δειγμάτων περιλαμβάνουν αμιγώς τοξικά δείγματα. Αντίθετα, τα cluster των μη τοξικών δειγμάτων, καταλαμβάνουν περισσότερο χώρο και περιλαμβάνουν και τοξικά δείγματα, σε μικρή αναλογία.

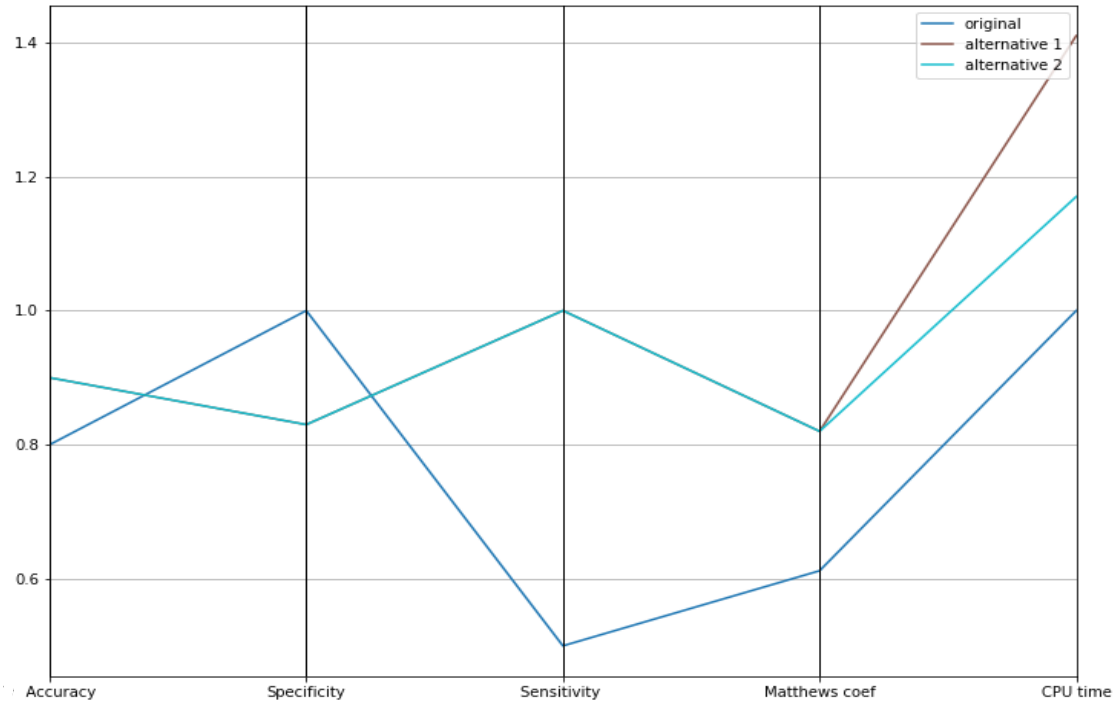
Διαφορές μεταξύ των μεθόδων παρατηρούνται και στον υπολογιστικό χρόνο (CPU Time) που απαιτεί κάθε πρόβλεψη. Η Μέθοδος 1, χαρακτηρίζεται από τους

μεγαλύτερους υπολογιστικούς χρόνους με μέση τιμή 1.31 δευτερόλεπτα. Ακολουθούν η Μέθοδος 2 και Μέθοδος 3 με 1.27 και 1.21 δευτερόλεπτα αντίστοιχα. Αν και οι Μέθοδοι 2 και 3 έχουν την ίδια αποτελεσματικότητα, η Μέθοδος 3 έχει μικρότερο υπολογιστικό κόστος.

Πίνακας 6.1.4: Υπολογιστικοί χρόνοι σε δευτερόλεπτα για την εφαρμογή του μοντέλου μέσω Kennard Stone στο σύνολο MeHydOx

Train Ratio	Μέθοδος 1	Μέθοδος 2	Μέθοδος 3
0.60	1.00	1.41	1.41
0.65	1.25	1.02	0.91
0.70	1.42	1.38	1.13
0.75	1.42	1.16	1.44
0.80	1.45	1.39	1.17

Η διαγραμματική σύγκριση των στατιστικών επιδόσεων των τριών εναλλακτικών μεθόδων κατηγοριοποίησης για κλάσμα διαμέρισης 0.6 παρουσιάζεται στο γράφημα 6.1.1.



Γράφημα 6.1.1: Σύγκριση στατιστικών επιδόσεων για τις τρεις εναλλακτικές μεθόδους με εφαρμογή στο σύνολο MeHydOx με Kennard Stone και training ratio=0.65

Εσωτερική αξιολόγηση

Για την εσωτερική αξιολόγηση, ακολουθώντας την διαδικασία K-Fold, το σύνολο χωρίζεται 3 φορές σε 3,4 και 5 υποσύνολα, αντίστοιχα. Σε κάθε διαφορετικό χώρισμα, ένα εκ των υποσυνόλων λειτουργεί ως σύνολο ελέγχου και τα υπόλοιπα

ως σύνολο εκπαίδευσης. Στους πίνακες 6.1.5-6.1.7, συνοψίζονται τα στατιστικά αποτελέσματα για τις τρεις δοκιμές διαμέρισης K-Fold για τις τρεις εναλλακτικές μεθοδολογίες κατηγοριοποίησης. Για την ορθή αξιολόγηση των τριών εναλλακτικών μεθοδολογιών και του μοντέλου, οι προβλέψεις κάθε μεθοδολογίας προκύπτουν από τα ίδια υποσύνολα εκπαίδευσης και επικύρωσης, άρα και τις ίδιες σχηματιζόμενες συστάδες.

Πίνακας 6.1.5: Στατιστικά αποτελέσματα με εφαρμογή 3-Fold στο σύνολο MeHydOx με τις 3 διαφορετικές Μεθόδους Κατηγοριοποίησης

Method	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.88	0.93	0.82	0.12	0.76	$\begin{bmatrix}13 & 1\\ 2 & 9\end{bmatrix}$
2	0.84	0.93	0.73	0.16	0.68	$\begin{bmatrix}13 & 1\\ 3 & 8\end{bmatrix}$
3	0.84	0.93	0.73	0.16	0.68	$\begin{bmatrix}13 & 1\\ 3 & 8\end{bmatrix}$

Στην περίπτωση της 3-Fold επικύρωσης, τα στατιστικά αποτελέσματα και για τις τρεις μεθοδολογίες είναι ικανοποιητικά, χωρίς μεγάλες αποκλίσεις μεταξύ των τιμών της ακρίβειας. Τις μικρότερες διακυμάνσεις σε συνδυασμό με την καλύτερη αθροιστική απόδοση επιδεικνύει η Μέθοδος 1, με ακρίβεια 0.88 και ειδικότητα 0.85. Από το σύνολο των τοξικών δειγμάτων νανοσωματιδίων, μόνο 2 δεν προβλέφθηκαν επιτυχώς στο σύνολο των τριών δοκιμών. Οι Μέθοδοι 2 και 3 , δεν προέβλεψαν την τοξικότητα 3 εκ του συνόλου των τοξικών δειγμάτων.

Ομοίως παρουσιάζονται τα αποτελέσματα για τις μεθόδους 4-Fold και 5-Fold.

Πίνακας 6.1.6: Στατιστικά αποτελέσματα με εφαρμογή 4-Fold στο σύνολο MeHydOx με τις 3 διαφορετικές Μεθόδους Κατηγοριοποίησης

Method	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.80	0.93	0.64	0.20	0.60	$\begin{bmatrix}13 & 1\\ 4 & 7\end{bmatrix}$
2	0.88	0.86	0.91	0.12	0.76	$\begin{bmatrix}12 & 2\\ 1 & 10\end{bmatrix}$
3	0.84	0.93	0.73	0.16	0.68	$\begin{bmatrix}13 & 1\\ 3 & 8\end{bmatrix}$

Στην περίπτωση της 4-Fold επικύρωσης, τα στατιστικά αποτελέσματα και για τις τρεις μεθοδολογίες είναι ικανοποιητικά, χωρίς μεγάλες αποκλίσεις μεταξύ των τιμών της ακρίβειας. Την καλύτερη αθροιστική απόδοση επιδεικνύει η Μέθοδος 2, με ακρίβεια 0.88. Από το σύνολο των τοξικών δειγμάτων νανοσωματιδίων, μόνο 1 δεν προβλέφθηκε επιτυχώς στο σύνολο των τριών δοκιμών. Οι Μέθοδοι 1 και 3 , δεν προέβλεψαν την τοξικότητα 4 και 3 αντίστοιχα εκ του συνόλου τω τοξικών δειγμάτων.

Πίνακας 6.1.7: Στατιστικά αποτελέσματα με εφαρμογή 5-Fold στο σύνολο MeHydOx με τις 3 διαφορετικές Μεθόδους Κατηγοριοποίησης

Method	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.80	0.93	0.64	0.20	0.60	$\begin{bmatrix} 13 & 1 \\ 4 & 7 \end{bmatrix}$
2	0.88	0.79	1.00	0.12	0.79	$\begin{bmatrix} 11 & 3 \\ 0 & 11 \end{bmatrix}$
3	0.88	0.93	0.82	0.12	0.76	$\begin{bmatrix} 13 & 1 \\ 2 & 9 \end{bmatrix}$

Στην περίπτωση της 5-Fold επικύρωσης, τα στατιστικά αποτελέσματα και για τις τρεις μεθοδολογίες είναι ικανοποιητικά. Την καλύτερη αθροιστική απόδοση επιδεικνύει η Μέθοδος 3, με ακρίβεια 0.88 και υψηλές τιμές ευαισθησίας και ειδικότητας. Από το σύνολο των τοξικών δειγμάτων νανοσωματιδίων, μόνο 2 δεν προβλέφθηκαν επιτυχώς στο σύνολο των τριών δοκιμών. Οι Μέθοδοι 1 και 2 , δεν προέβλεψαν την τοξικότητα 5 και 0 αντίστοιχα, εκ του συνόλου τω τοξικών δειγμάτων.

Ως προς το υπολογιστικό κόστος των τριών εναλλακτικών μεθόδων, τα αποτελέσματα συνοψίζονται στον πίνακα 6.1.8.

Πίνακας 6.1.8: Υπολογιστικοί χρόνοι σε δευτερόλεπτα για την εφαρμογή του μοντέλου μέσω K-Fold στο σύνολο MeHydOx

	Μέθοδος 1	**Μέθοδος 2**	**Μέθοδος 3**
3-Fold	1.28	3.09	3.17
4-Fold	1.35	1.79	2.50
5-Fold	1.33	1.66	3.24

Αν και υπάρχουν διακυμάνσεις ως προς την ταχύτερη μέθοδο στις τρεις δοκιμές, κατά μέσο όρο, η Μέθοδος 2 αποδεικνύεται γρηγορότερη.

Υπό το ίδιο σκεπτικό, βάσει της διαδικασίας Leave-One-Out, η μέθοδος εφαρμόζεται 25 φορές. Σε κάθε επανάληψη, ένα εκ των δειγμάτων δεν συμμετέχει στη διαμόρφωση του μοντέλου και χρησιμοποιείται για την τελική αξιολόγησή του. Τα αποτελέσματα για τις τρείς μεθοδολογίες συνοψίζονται στον πίνακα 6.1.9.

Πίνακας 6.1.9: Στατιστικά αποτελέσματα με εφαρμογή Leave-One-Out στο σύνολο MeHydOx για τις τρεις διαφορετικές μεθόδους

Μέθοδος	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.88	1.00	0.73	0.12	0.77	$\begin{bmatrix} 14 & 0 \\ 3 & 8 \end{bmatrix}$

2	0.88	0.86	0.91	0.12	0.76	$\begin{bmatrix} 12 & 2 \\ 1 & 10 \end{bmatrix}$
3	0.80	0.86	0.73	0.20	0.68	$\begin{bmatrix} 12 & 2 \\ 3 & 8 \end{bmatrix}$

Κατά τις δοκιμές Leave-One-Out, επιβεβαιώνεται η επιτυχία του μοντέλου με στατιστικά ακρίβειας [0.88, 0.88, 0.80] για τις μεθόδους 1, 2 και 3, αντίστοιχα. Μάλιστα, στην περίπτωση της Μεθόδου 2, μόνο ένα εκ των τοξικών δειγμάτων δεν εντοπίζεται.

Έλεγχος τυχαίας επιλογής

Ο έλεγχος y-scrambling πραγματοποιήθηκε με Kennard Stone και λόγο διαμέρισης 0.65, διαταράσσοντας τη σειρά των τιμών στη στήλη endpoint του training set, ώστε το μοντέλο να εκπαιδευτεί σε μη έγκυρα δεδομένα. Τα αποτελέσματα των 5 διαφορετικών τυχαίων δοκιμών που πραγματοποιήθηκαν, με αριθμό δειγμάτων εκπαίδευσης και ελέγχου 16 και 9 αντίστοιχα, συνοψίζονται στον πίνακα 6.1.10.

Πίνακας 6.1.10: Στατιστικά αποτελέσματα με εφαρμογή y-scrambling στο σύνολο MeHydOx

Accuracy	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
0.22	0.40	0.00	0.78	-0.67	$\begin{bmatrix} 2 & 3 \\ 4 & 0 \end{bmatrix}$
0.33	0.60	0.00	0.67	-0.48	$\begin{bmatrix} 3 & 2 \\ 4 & 0 \end{bmatrix}$
0.56	0.60	0.50	0.44	0.10	$\begin{bmatrix} 3 & 2 \\ 2 & 2 \end{bmatrix}$
0.22	0.00	0.50	0.78	-0.60	$\begin{bmatrix} 0 & 5 \\ 2 & 2 \end{bmatrix}$
0.22	0.40	0.00	0.78	-0.67	$\begin{bmatrix} 2 & 3 \\ 4 & 0 \end{bmatrix}$

Οι τιμές ακρίβειας [0.22, 0.33, 0.56, 0.22, 0.22] υποδηλώνουν ένα πλήρως αποτυχημένο μοντέλο το οποίο κάνει χειρότερες εκτιμήσεις συγκριτικά με ένα τυχαίο μοντέλο. Αυτό επιβεβαιώνει και την υπόθεση πως οι επιτυχημένες προβλέψεις οφείλονται στην σωστή δόμηση του μοντέλου και δεν είναι αποτέλεσμα τυχαιότητας.

Πεδίο εφαρμογής μοντέλου

Το αριθμητικό κατώφλι που αφορά στο πεδίου εφαρμογής του μοντέλου υπολογίζεται για τις 5 διαφορετικές περιπτώσεις χωρισμού του δειγματικού χώρου βάσει της Kennard Stone και κλάσματα διαμέρισης [0.60, 0.65, 0.70, 0.75, 0.80]. Οι τιμές συνοψίζονται στον πίνακα 6.1.11:

Πίνακας 6.1.11: Κατώφλι πεδίου εφαρμογής για το σύνολο MeHydOx

Train Ratio	***Κατώφλι***	***Δείγματα που ικανοποιούν το κατώφλι***
0.60	2.40	8/10
0.65	2.25	8/9
0.70	2.00	7/7
0.75	1.89	6/6
0.80	1.80	5/5

Για τα δείγματα με τιμή μόχλευσης μεγαλύτερη του ανώτατου ορίου που ορίζει το υπολογιζόμενο κατώφλι, οι προβλέψεις δεν κρίνονται αξιόπιστες.

Αποτίμηση της επίδοσης του μοντέλου

Η προβλεπτική ικανότητα του μοντέλου στο σύνολο δεδομένων MeHydOx είναι ικανοποιητική καθώς οι επιδόσεις σε εξωτερική και εσωτερική αξιολόγηση αλλά και ελέγχους στο πεδίο εφαρμογής και στην μέθοδο y-scrambling, είναι ενθαρρυντικές. Το μοντέλο που διαμορφώνεται συνδυάζει επαρκή ειδίκευση ώστε να εντοπίζει σε μεγάλο βαθμό τα τοξικά δείγματα αλλά και δυνατότητα γενίκευσης, καθιστώντας το ένα εύρωστο μοντέλο.

Την καλύτερη επίδοση μεταξύ όλων των ελέγχων φέρει η διαμέριση με Kennard Stone για training ratio 0.6 το οποίο χωρίζει το σύνολο δεδομένων σε 15 δείγματα εκπαίδευσης και 10 δείγματα ελέγχου. Εξ αυτών, όλα πλην ενός κατηγοριοποιούνται επιτυχώς με εφαρμογή των Μεθόδων Κατηγοριοποίησης 2 και 3, οι οποίες έχουν και την βέλτιστη επίδοση ως προς τον εντοπισμό των τοξικών δειγμάτων με ειδικότητα η οποία διατηρείται ίση με 1.

Οι Μέθοδοι 2 και 3 έχουν παρόμοιες αποδόσεις στους ελέγχους που πραγματοποιούνται, εμφανίζοντας διαφορά στους χρόνους που απαιτεί η εφαρμογή τους, με την Μέθοδο 2 να χαρακτηρίζεται από μικρότερους υπολογιστικούς χρόνους. Λαμβάνοντας υπόψιν και την επικύρωση Leave-One-Out, κατά την οποία η Μέθοδος 2 επιδεικνύει καλύτερη προβλεπτική ικανότητα, εντοπίζοντας 10 εκ των 11 τοξικών δειγμάτων και 12 εκ των 14 μη τοξικών, καταλληλότερη μέθοδος πρόβλεψης για το σύνολο δεδομένων MeHydOx χαρακτηρίζεται η Μέθοδος 2, ιδιαίτερα κατά την εφαρμογή εξωτερικής αξιολόγησης με Kennard Stone και training ratio 0.6.

6.1.2. Σύνολο δεδομένων SPIONs

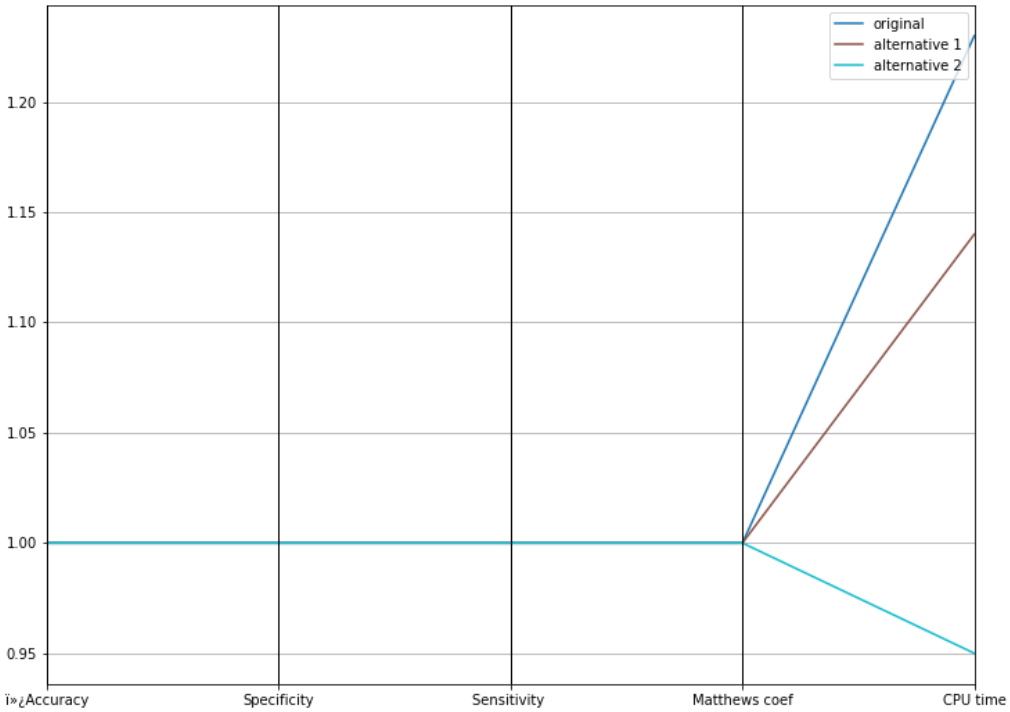
Το σύνολο δεδομένων των SPIONs περιλαμβάνει 15 δείγματα με 5 μεταβλητές-ιδιότητες και ένα endpoint δυαδικής φύσης. Ο έλεγχος στο συγκεκριμένο σύνολο πραγματοποιείται μέσω εσωτερικής και εξωτερικής αξιολόγησης.

Ακολουθείται η ίδια διαδικασία εξωτερικής αξιολόγησης με Kennard Stone για training ratio από 0.60 ως 0.80. Στον πίνακα 6.1.12 , περιέχονται τα στατιστικά αποτελέσματα των δοκιμών για τις τρεις εναλλακτικές μεθόδους.

Πίνακας 6.1.12: Στατιστικά αποτελέσματα με εφαρμογή Kennard Stone για διαφορετικά training ratios στο σύνολο SPIONs με τη μέθοδο 1

Train Ratio	Accuracy	Sensitivity	Specificity	Error rate	MCC
0.60,0.65,0.70,0.75,0.80	1.00	1.00	1.00	0.00	1.00

Για κάθε διαφορετικό κλάσμα διαμέρισης, οι προβλέψεις βάσει της Μεθόδου 1 είναι απόλυτα επιτυχείς με πλήρη ακρίβεια, ευαισθησία και ειδικότητα, χωρίς



Γράφημα 6.1.2: Σύγκριση στατιστικών επιδόσεων για τις τρεις εναλλακτικές μεθόδους με εφαρμογή στο σύνολο SPIONs με Kennard Stone και training ratio=0.65

εσφαλμένες προβλέψεις ενώ εντοπίζονται όλα τα τοξικά δείγματα νανοσωματιδίων. Οι προβλέψεις των μεθόδων 2 και 3 ταυτίζονται απόλυτα με αυτές της 1, δεν συμπεριλαμβάνονται και επιβεβαιώνουν την ισχυρή προβλεπτική ικανότητα του μοντέλου. Η διαγραμματική σύγκριση των στατιστικών επιδόσεων των τριών εναλλακτικών μεθόδων κατηγοριοποίησης για κλάσμα διαμέρισης 0.65 παρουσιάζεται στο γράφημα 6.1.2.

Διαφορές μεταξύ των μεθόδων παρατηρούνται στον υπολογιστικό χρόνο (CPU Time) που απαιτεί κάθε πρόβλεψη. Κατά μέσο όρο, γρηγορότερη εκ των τριών μεθόδων ορίζεται η Μέθοδος 3, ακολουθεί η 2 και η 1, με μικρή διαφορά.

Πίνακας 6.1.13: Υπολογιστικοί χρόνοι σε δευτερόλεπτα για την εφαρμογή του μοντέλου μέσω Kennard Stone στο σύνολο SPIONs

Train Ratio	**Μέθοδος 1**	**Μέθοδος 2**	**Μέθοδος 3**
0.60	1.23	1.14	0.84
0.65	0.91	0.77	0.83
0.70	0.70	0.89	0.88
0.75	0.84	0.75	0.77
0.80	0.81	0.95	0.75

Εσωτερική αξιολόγηση

Ακολουθώντας την ίδια διαδικασία K-Fold, το σύνολο χωρίζεται σε 3,4 και 5 υποσύνολα, αντίστοιχα. Σε κάθε διαφορετικό χώρισμα, ένα εκ των υποσυνόλων λειτουργεί ως σύνολο ελέγχου και τα υπόλοιπα ως σύνολο εκπαίδευσης. Στους πίνακες 6.1.14-6.1.16, συνοψίζονται τα στατιστικά αποτελέσματα για τις τρεις δοκιμές διαμέρισης K-Fold για τις τρεις εναλλακτικές μεθοδολογίες κατηγοριοποίησης.

Πίνακας 6.1.14: Στατιστικά αποτελέσματα με εφαρμογή 3-Fold στο σύνολο SPIONs για τις τρεις διαφορετικές μεθόδους

Μέθοδος	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.87	0.80	1.00	0.13	0.76	$\begin{bmatrix} 8 & 2 \\ 0 & 5 \end{bmatrix}$
2	0.93	0.90	1.00	0.07	0.87	$\begin{bmatrix} 9 & 1 \\ 0 & 5 \end{bmatrix}$
3	0.87	0.80	1.00	0.13	0.76	$\begin{bmatrix} 8 & 2 \\ 0 & 5 \end{bmatrix}$

Στην περίπτωση της $3-Fold$ επικύρωσης, τα στατιστικά αποτελέσματα και για τις τρεις μεθοδολογίες είναι ιδιαίτερα ικανοποιητικά, χωρίς μεγάλες αποκλίσεις μεταξύ των τιμών της ακρίβειας. Οι τρείς μέθοδοι κατηγοριοποίησης οδηγούν στις ίδιες προβλέψεις ως προς τα τοξικά νανοσωματίδια, καθώς από το σύνολο των τοξικών δειγμάτων νανοσωματιδίων, όλα προβλέφθηκαν επιτυχώς στο σύνολο των τριών δοκιμών. Ομοίως παρουσιάζονται τα αποτελέσματα για τις μεθόδους 4-Fold και 5-Fold.

Πίνακας 6.1.15: Στατιστικά αποτελέσματα με εφαρμογή 4-Fold στο σύνολο SPIONs για τις τρεις διαφορετικές μεθόδους

Μέθοδος	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.87	0.90	0.80	0.13	0.70	$\begin{bmatrix} 9 & 1 \\ 1 & 4 \end{bmatrix}$
2	0.80	0.90	0.60	0.20	0.53	$\begin{bmatrix} 9 & 1 \\ 2 & 3 \end{bmatrix}$

3	0.87	0.90	0.80	0.13	0.70	$\begin{bmatrix} 9 & 1 \\ 1 & 4 \end{bmatrix}$

Στην περίπτωση της 4-Fold επικύρωσης, τα στατιστικά αποτελέσματα και για τις τρεις μεθοδολογίες είναι ικανοποιητικά, αν και όχι εξίσου καλά με υψηλά με αυτά της 3-Fold επικύρωσης. Τις μικρότερες διακυμάνσεις σε συνδυασμό με την καλύτερη αθροιστική απόδοση επιδεικνύουν από κοινού οι Μέθοδοι 1 και 3, με ακρίβεια 0.87 και ειδικότητα 0.8. Από το σύνολο των τοξικών δειγμάτων νανοσωματιδίων, μόνο 1 δεν προβλέφθηκε επιτυχώς στο σύνολο των τριών δοκιμών. Η Μέθοδος 2 , δεν προέβλεψε την τοξικότητα 2 εκ του συνόλου τω τοξικών δειγμάτων.

Πίνακας 6.1.16: Στατιστικά αποτελέσματα με εφαρμογή 5-Fold στο σύνολο SPIONs για τις τρεις διαφορετικές μεθόδους

Μέθοδος	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.80	0.80	0.80	0.20	0.58	$\begin{bmatrix} 8 & 2 \\ 1 & 4 \end{bmatrix}$
2	0.87	0.90	0.80	0.20	0.70	$\begin{bmatrix} 9 & 1 \\ 1 & 4 \end{bmatrix}$
3	0.87	0.80	1.00	0.00	0.76	$\begin{bmatrix} 8 & 2 \\ 0 & 5 \end{bmatrix}$

Στην περίπτωση της 5-Fold επικύρωσης, τα στατιστικά είναι εξίσου ικανοποιητικά, με πολλές περιπτώσεις δοκιμών απόλυτα επιτυχείς και ειδικότητα ίση με τη μονάδα. Τις μικρότερες διακυμάνσεις σε συνδυασμό με την καλύτερη αθροιστική απόδοση επιδεικνύει η Μέθοδος 3, με ακρίβεια 0.87 και ειδικότητα 1.00. Από το σύνολο των τοξικών δειγμάτων νανοσωματιδίων, όλα προβλέφθηκαν επιτυχώς στο σύνολο των τριών δοκιμών. Οι Μέθοδοι 1 και 2 , δεν προέβλεψαν την τοξικότητα ενός δείγματος, εκ του συνόλου των τοξικών δειγμάτων.

Ως προς το υπολογιστικό κόστος των τριών εναλλακτικών μεθόδων, τα αποτελέσματα συνοψίζονται στον πίνακα 6.1.17. Μεγαλύτερο υπολογιστικό κόστος επιδεικνύει η Μέθοδος 1 ενώ οι Μέθοδοι 2 και 3 έχουν παρόμοιες αποδόσεις.

Πίνακας 6.1.17: Υπολογιστικοί χρόνοι σε δευτερόλεπτα για την εφαρμογή του μοντέλου μέσω K-Fold στο σύνολο SPIONs

	Μέθοδος 1	**Μέθοδος 2**	**Μέθοδος 3**
3-Fold	0.43	0.41	0.41
4-Fold	0.64	0.6	0.61
5-Fold	0.86	0.79	0.79

Υπό το ίδιο σκεπτικό, βάσει της διαδικασίας Leave-One-Out, η μέθοδος εφαρμόζεται 15 φορές. Τα αποτελέσματα για τις τρείς μεθοδολογίες συνοψίζονται στον πίνακα 6.1.18.

Πίνακας 6.1.18: Στατιστικά αποτελέσματα με εφαρμογή Leave-One-Out στο σύνολο SPIONs για τις τρεις διαφορετικές μεθόδους

Μέθοδος	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.87	0.9	0.8	0.13	0.7	$\begin{bmatrix} 9 & 1 \\ 1 & 4 \end{bmatrix}$
2	0.87	0.8	1	0.13	0.76	$\begin{bmatrix} 8 & 2 \\ 0 & 5 \end{bmatrix}$
3	0.87	0.8	1	0.13	0.76	$\begin{bmatrix} 8 & 2 \\ 0 & 5 \end{bmatrix}$

Κατά τις δοκιμές Leave-One-Out, επιβεβαιώνεται η επιτυχία του μοντέλου με στατιστικά ακρίβειας [0.87, 0.87, 0.87] για τις μεθόδους 1, 2 και 3, αντίστοιχα. Μάλιστα, στην περίπτωση των Μεθόδων 2 και 3, εντοπίζονται όλα τα τοξικά δείγματα. Η Μέθοδος 1 δεν εντοπίζει 1 τοξικό δείγμα. Σε κάθε περίπτωση η απόδοση του μοντέλου κρίνεται επιτυχής.

Έλεγχος τυχαίας επιλογής

Ο έλεγχος y-scrambling πραγματοποιήθηκε με Kennard Stone και κλάσμα διαμέρισης 0.65, επανατοποθετώντας τα endpoint του συνόλου εκπαίδευσης ώστε το μοντέλο να εκπαιδευτεί σε μη έγκυρα δεδομένα. Τα αποτελέσματα των 5 διαφορετικών τυχαίων δοκιμών που πραγματοποιήθηκαν, με αριθμό δειγμάτων training και test set 10 και 5 αντίστοιχα, συνοψίζονται στον πίνακα 6.1.19.

Πίνακας 6.1.19: Στατιστικά αποτελέσματα με εφαρμογή y-scrambling στο σύνολο SPIONs για την Μέθοδο 1

Accuracy	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
0.40	1.00	0.00	0.60	0.00	$\begin{bmatrix} 2 & 0 \\ 3 & 0 \end{bmatrix}$
0.40	1.00	0.00	0.60	0.00	$\begin{bmatrix} 2 & 0 \\ 3 & 0 \end{bmatrix}$
0.40	1.00	0.00	0.60	0.00	$\begin{bmatrix} 2 & 0 \\ 3 & 0 \end{bmatrix}$
0.20	0.50	0.00	0.80	-0.61	$\begin{bmatrix} 1 & 1 \\ 3 & 0 \end{bmatrix}$
0.40	1.00	0.00	0.60	0.00	$\begin{bmatrix} 2 & 0 \\ 3 & 0 \end{bmatrix}$

Οι τιμές ακρίβειας [0.4, 0.4, 0.4, 0.2, 0.4] υποδηλώνουν ένα πλήρως αποτυχημένο μοντέλο το οποίο κάνει χειρότερες εκτιμήσεις συγκριτικά με ένα τυχαίο μοντέλο. Αυτό επιβεβαιώνει και την υπόθεση πως οι επιτυχημένες προβλέψεις οφείλονται στην σωστή δόμηση του μοντέλου και δεν είναι αποτέλεσμα τυχαιότητας.

Πεδίο εφαρμογής μοντέλου

Το αριθμητικό κατώφλι που αφορά στο πεδίου εφαρμογής του μοντέλου υπολογίζεται για τις 5 διαφορετικές περιπτώσεις χωρισμού του δειγματικού χώρου βάσει τις Kennard Stone και κλάσματα διαμέρισης [0.6, 0.65, 0.7, 0.75, 0.8]. Οι τιμές συνοψίζονται στον πίνακα 6.1.20.

Πίνακας 6.1.20: Κατώφλι πεδίου εφαρμογής για το σύνολο SPIONs

Train Ratio	***Κατώφλι***	***Δείγματα που ικανοποιούν το κατώφλι***
0.60	1.67	6/6
0.65	1.5	5/5
0.70	1.5	5/5
0.75	1.36	4/4
0.80	1.25	3/3

Ο υπολογισμός της τιμής μόχλευσης κάθε δείγματος και για κάθε δοκιμή δεν αναδεικνύει κάποιο δείγμα με τιμή μόχλευσης μεγαλύτερη του ανώτατου ορίου που ορίζει το υπολογιζόμενο κατώφλι. Έτσι, το σύνολο των προβλέψεων θεωρείται αξιόπιστο.

Αποτίμηση της επίδοσης του μοντέλου

Η προβλεπτική ικανότητα του μοντέλου στο σύνολο δεδομένων SPIONs κρίνεται επιτυχημένη καθώς οι επιδόσεις σε εξωτερική και εσωτερική αξιολόγηση αλλά και ελέγχους στο πεδίο εφαρμογής και στην μέθοδο y-scrambling, είναι ιδιαίτερα υψηλές. Το μοντέλο εντοπίζει σχεδόν σε απόλυτο βαθμό τα τοξικά δείγματα σε κάθε έλεγχο στον οποίο υπόκειται, καθιστώντας το ένα εύρωστο μοντέλο.

Η εφαρμογή Kennard Stone κατηγοριοποιεί ορθά το σύνολο το δειγμάτων ανεξαρτήτων training ratio και Μεθόδου Κατηγοριοποίησης. Έτσι, κριτήριο για τον εντοπισμό της βέλτιστης Μεθόδου θα αποτελέσει η εσωτερική αξιολόγηση στην οποία η Μέθοδος 1 φαίνεται να έχει την καλύτερη απόδοση, βάσει και των τριών ελέγχων 3-Fold, 4-Fold, 5-Fold, καθώς στο σύνολο των δοκιμών, μόνο σε δύο περιπτώσεις δεν εντοπίζονται όλα τα τοξικά δείγματα. Ανασταλτικά για την εφαρμογή της Μεθόδου 1, λειτουργεί ο σχετικά μεγαλύτερος υπολογιστικός χρόνος που απαιτεί συγκριτικά με τις άλλες δύο μεθοδολογίες.

Λαμβάνοντας υπόψιν τη συνολική επίδοση σε όλους τους ελέγχους, καταλληλότερη μέθοδος πρόβλεψης για το σύνολο δεδομένων SPIONs χαρακτηρίζεται η Μέθοδος 1, ιδιαίτερα κατά την εφαρμογή εσωτερικής αξιολόγησης με επικύρωση 4-Fold και 5-Fold.

6.1.3. Σύνολο δεδομένων MeOx

Το σύνολο δεδομένων των $MeOx$ περιλαμβάνει 23 δείγματα με 24 μεταβλητές-ιδιότητες και ένα endpoint δυαδικής φύσης. Ο έλεγχος στο συγκεκριμένο σύνολο πραγματοποιείται μέσω εσωτερικής και εξωτερικής αξιολόγησης.

Εξωτερική αξιολόγηση

Στους πίνακες 6.1.21-6.1.23, περιέχονται τα στατιστικά αποτελέσματα των δοκιμών $Kennard\ Stone$ με διαφορετικά κλάσματα διαμέρισης που κυμαίνονται από 0.60 ως 0.80 για τις τρεις εναλλακτικές μεθόδους.

Πίνακας 6.1.21: Στατιστικά αποτελέσματα με εφαρμογή Kennard Stone για διαφορετικά training ratios στο σύνολο MeOx με τη μέθοδο 1

Train Ratio	Test samples	Accuracy	Sensitivity	Specificity	Error rate	MCC	Confusion Matrix
0.60	9	0.67	0.33	0.83	0.33	0.19	$\begin{bmatrix}1 & 2\\1 & 5\end{bmatrix}$
0.65	8	0.75	0.50	0.83	0.25	0.33	$\begin{bmatrix}1 & 1\\1 & 5\end{bmatrix}$
0.70	7	0.71	0.50	0.80	0.29	0.30	$\begin{bmatrix}1 & 1\\1 & 4\end{bmatrix}$
0.75	6	0.83	0.50	1.00	0.17	0.63	$\begin{bmatrix}1 & 1\\0 & 4\end{bmatrix}$
0.80	5	0.80	0.50	1.00	0.20	0.61	$\begin{bmatrix}1 & 1\\0 & 3\end{bmatrix}$

Η ακρίβεια των προβλέψεων κυμαίνεται σε τιμές $[0.67, 0.75, 0.71, 0.83, 0.80]$ για κλάσματα διαμέρισης $[0.60, 0.65, 0.70, 0.75, 0.80]$, αντίστοιχα. Η καλύτερη πρόβλεψη δίνεται για κλάσμα 0.75. Ανεξάρτητα από το κλάσμα διαμέρισης, η ακρίβεια των προβλέψεων στην περίπτωση της Μεθόδου Κατηγοριοποίησης 1 είναι ικανοποιητική. Επιτυγχάνοντας την σωστή πρόβλεψη όλων των μη τοξικών δειγμάτων με απόλυτη ειδικότητα σε δύο εκ των πέντε δοκιμών, η ευαισθησία κυμαίνεται σε χαμηλότερα επίπεδα, δυσχεραίνοντας την πρόβλεψη των μη τοξικών νανοσωματιδίων. Ο MCC, με τιμές $[0.19, 0.33, 0.30, 0.63, 0.61]$, προσφέροντας μια συνολικότερη εικόνα για την προβλεπτική ικανότητα, υποδηλώνει πως το μοντέλο είναι αποδοτικό αλλά με αρκετές αδυναμίες.

Πίνακας 6.1.22: Στατιστικά αποτελέσματα με εφαρμογή Kennard Stone για διαφορετικά training ratios στο σύνολο MeOx με τη μέθοδο 2

Train Ratio	Test samples	Accuracy	Sensitivity	Specificity	Error rate	MCC	Confusion Matrix
0.60	9	0.78	1.00	0.67	0.22	0.63	$\begin{bmatrix}3 & 0\\2 & 4\end{bmatrix}$
0.65	8	0.88	1.00	0.83	0.12	0.75	$\begin{bmatrix}2 & 0\\1 & 5\end{bmatrix}$
0.70	7	0.86	1.00	0.80	0.14	0.73	$\begin{bmatrix}2 & 0\\1 & 4\end{bmatrix}$

0.75	6	0.83	1.00	0.75	0.17	0.71	$\begin{bmatrix} 2 & 0 \\ 1 & 3 \end{bmatrix}$
0.80	5	0.80	1.00	0.67	0.20	0.67	$\begin{bmatrix} 2 & 0 \\ 1 & 2 \end{bmatrix}$

Στην περίπτωση της εναλλακτικής Μεθόδου Κατηγοριοποίησης 2, η ακρίβεια των προβλέψεων κυμαίνεται σε τιμές [0.78, 0.88, 0.86, 0.83, 0.80] για κλάσματα διαμέρισης [0.60, 0.65, 0.70, 0.75, 0.80], αντίστοιχα, με την καλύτερη πρόβλεψη να δίνεται για κλάσμα 0.65. Σε αντίθεση με την Μέθοδο 1, τα μη τοξικά δείγματα προβλέπονται σωστά στο σύνολό τους και το μοντέλο χαρακτηρίζεται από απόλυτη ευαισθησία. Στις 5 δοκιμές, η ειδικότητα λαμβάνει συγκριτικά μεγαλύτερες τιμές ενώ η αύξηση στις τιμές του MCC, [0.63, 0.75, 0.73, 0.71, 0.67] υποδηλώνει καλύτερη προβλεπτική ικανότητα σε σχέση με την Μέθοδο Κατηγοριοποίησης 1.

Πίνακας 6.1.23: Στατιστικά αποτελέσματα με εφαρμογή Kennard Stone για διαφορετικά training ratios στο σύνολο MeOx με τη μέθοδο 3

Train Ratio	Test samples	Accuracy	Sensitivity	Specificity	Error rate	MCC	Confusion Matrix
0.60	9	0.89	1.00	0.83	0.11	0.79	$\begin{bmatrix} 3 & 0 \\ 1 & 5 \end{bmatrix}$
0.65	8	0.75	1.00	0.67	0.25	0.58	$\begin{bmatrix} 2 & 0 \\ 2 & 4 \end{bmatrix}$
0.70	7	0.86	1.00	0.80	0.14	0.73	$\begin{bmatrix} 2 & 0 \\ 1 & 4 \end{bmatrix}$
0.75	6	0.83	1.00	0.75	0.17	0.71	$\begin{bmatrix} 2 & 0 \\ 1 & 3 \end{bmatrix}$
0.80	5	0.80	1.00	0.67	0.20	0.67	$\begin{bmatrix} 2 & 0 \\ 1 & 2 \end{bmatrix}$

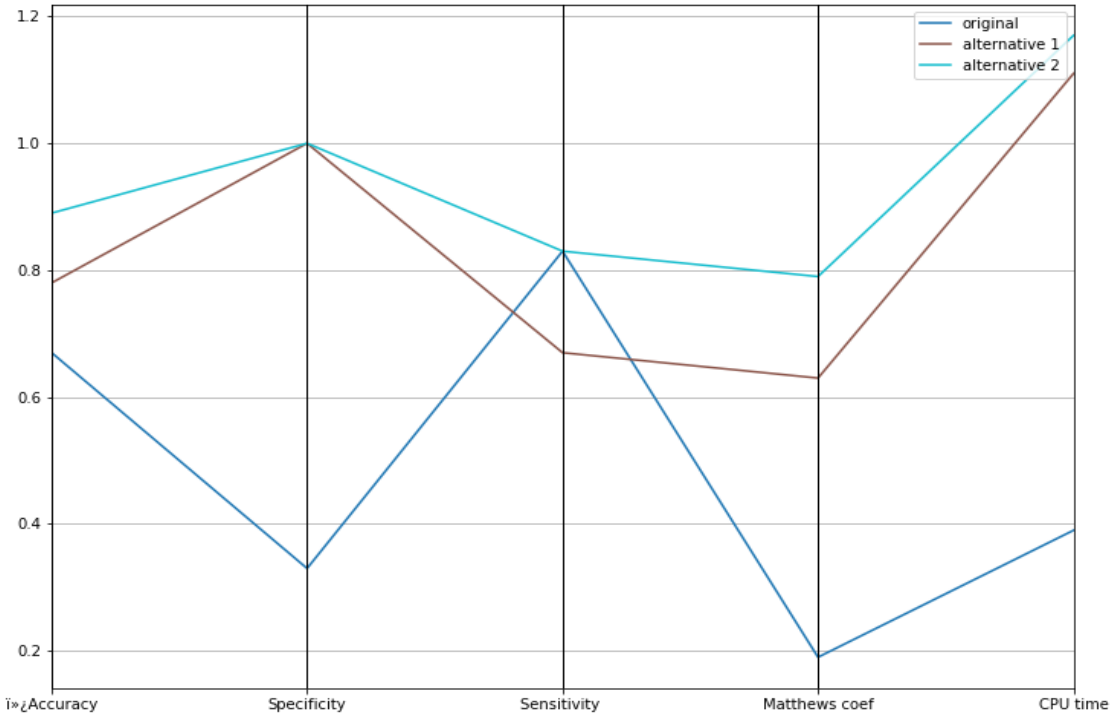
Στην περίπτωση της εναλλακτικής Μεθόδου Κατηγοριοποίησης 3, η ακρίβεια των προβλέψεων κυμαίνεται σε τιμές [0.89, 0.75, 0.86, 0.83, 0.80] για κλάσματα διαμέρισης [0.60, 0.65, 0.70, 0.75, 0.80], αντίστοιχα, με την καλύτερη πρόβλεψη να δίνεται για κλάσμα 0.60. Οι τιμές των στατιστικών μεγεθών είναι αντίστοιχες με αυτές της Μεθόδου 2, με μικρές διαφοροποιήσεις και συγκριτικά καλύτερες από αυτές της Μεθόδου 1, κυρίως στον εντοπισμό των τοξικών δειγμάτων.

Διαφορές μεταξύ των μεθόδων παρατηρούνται και στον υπολογιστικό χρόνο (CPU Time) που απαιτεί κάθε πρόβλεψη. Η Μέθοδος 1, χαρακτηρίζεται από τους μικρότερους υπολογιστικούς χρόνους με μέση τιμή 0.61 δευτερόλεπτα. Ακολουθούν η Μέθοδος 2 και Μέθοδος 3 με 1.20 και 1.28 δευτερόλεπτα αντίστοιχα. Αν και οι Μέθοδοι 2 και 3 έχουν παρόμοια προβλεπτική ικανότητα στο συγκεκριμένο σύνολο, η Μέθοδος 3 έχει μεγαλύτερο υπολογιστικό κόστος.

Πίνακας 6.1.24: Υπολογιστικοί χρόνοι σε δευτερόλεπτα για την εφαρμογή του μοντέλου με Kennard Stone στο σύνολο MeOx

Train Ratio	Μέθοδος 1	Μέθοδος 2	Μέθοδος 3
0.60	0.39	1.11	1.48
0.65	0.41	1.06	1.08
0.70	0.63	1.25	1.27
0.75	0.77	1.31	1.19
0.80	0.83	1.28	1.36

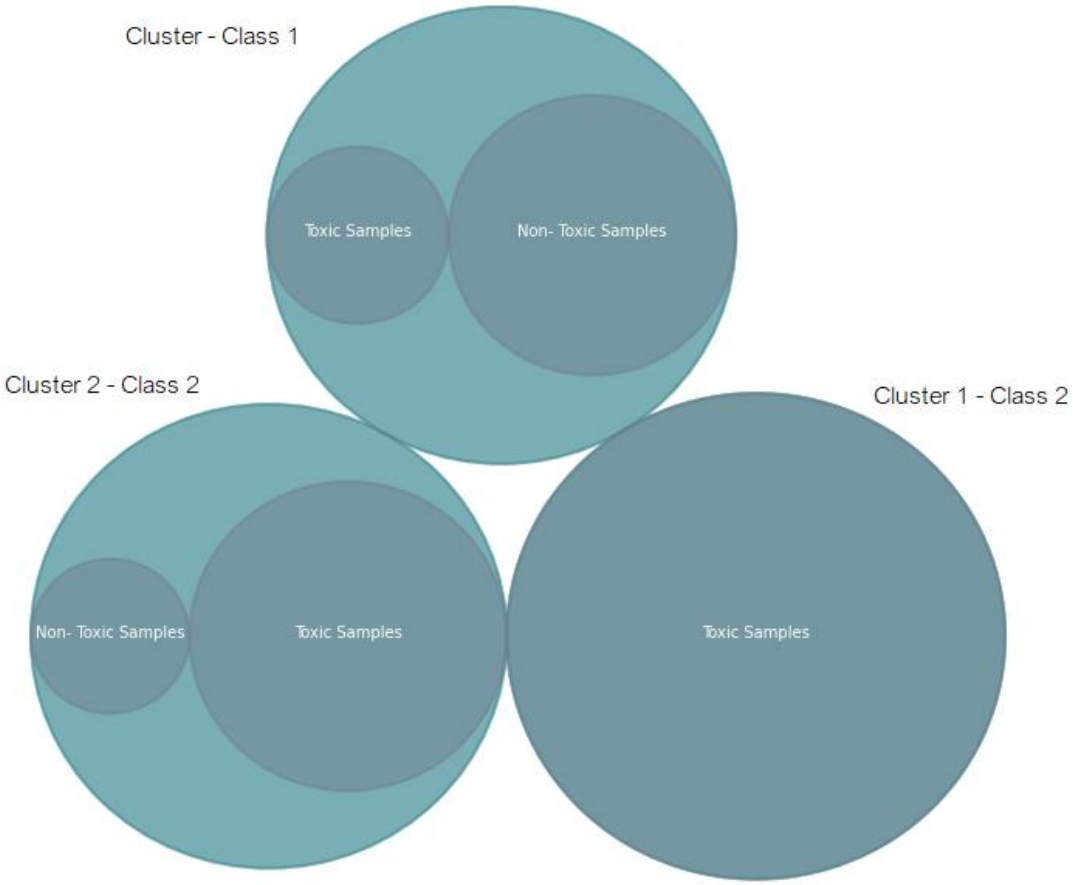
Η διαγραμματική σύγκριση των στατιστικών επιδόσεων των τριών εναλλακτικών μεθόδων κατηγοριοποίησης για κλάσμα διαμέρισης 0.65 παρουσιάζεται στο γράφημα 6.1.3. Συνολικά, και παρά το μεγαλύτερο υπολογιστικό κόστος, την καλύτερη πρόβλεψη λαμβάνοντας υπόψιν όλα τα στατιστικά μεγέθη των δειγμάτων νανοσωματιδίων πραγματοποιεί η Μέθοδος 3.



Γράφημα 6.1.3: Σύγκριση στατιστικών επιδόσεων για τις τρεις εναλλακτικές μεθόδους με εφαρμογή στο σύνολο MeOx με Kennard Stone και training ratio=0.65

Προς κατανόηση των αποτελεσμάτων παρουσιάζεται το ακόλουθο παράδειγμα. Κατά την μοντελοποίηση με Kennard Stone, για λόγο διαμέρισης 0.65, δημιουργούνται 3 συστάδες, 1 για τα μη τοξικά δείγματα και 2 για τα τοξικά. Όπως παρουσιάζεται στην απεικόνιση της κατανομής του δειγματικού χώρου στις συστάδες (Σχήμα 6.1.2), δύο εκ των συστάδων περιέχουν και δείγματα

αντίθετης κλάσης σε μικρότερη αναλογία και μόλις μια είναι αμιγώς συστάδα τοξικών δειγμάτων.



Σχήμα 6.1.2: Απεικόνιση του δειγματικού χώρου του συνόλου MeOx σε σχέση με τα cluster των δυο κλάσεων στα οποία εντοπίζονται

Εσωτερική αξιολόγηση

Ακολουθώντας την διαδικασία K-Fold, στους πίνακες 6.1.25-6.1.27, συνοψίζονται τα στατιστικά αποτελέσματα για τις τρεις δοκιμές διαμέρισης K-Fold για τις τρεις εναλλακτικές μεθοδολογίες κατηγοριοποίησης.

Πίνακας 6.1.25: Στατιστικά αποτελέσματα με εφαρμογή 3-Fold στο σύνολο MeOx για τις τρεις διαφορετικές μεθόδους

Μέθοδος	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.70	0.29	0.88	0.30	0.20	$\begin{bmatrix} 2 & 5 \\ 2 & 14 \end{bmatrix}$
2	0.91	1.00	0.88	0.09	0.82	$\begin{bmatrix} 7 & 0 \\ 2 & 14 \end{bmatrix}$
3	0.87	0.86	0.88	0.13	0.71	$\begin{bmatrix} 6 & 1 \\ 2 & 14 \end{bmatrix}$

Στην περίπτωση της 3-Fold επικύρωσης, τα στατιστικά αποτελέσματα για τις Μεθόδους 2 και 3 είναι ικανοποιητικά, χωρίς μεγάλες αποκλίσεις μεταξύ των τιμών της ακρίβειας. Η Μέθοδος 1 δεν αποδίδει το ίδιο αποτελεσματικά με ακρίβεια 0.7 και αδυναμία εντοπισμού των μη τοξικών δειγμάτων. Τις μικρότερες διακυμάνσεις σε συνδυασμό με την καλύτερη αθροιστική απόδοση επιδεικνύει η Μέθοδος 2, με ακρίβεια 0.91. Με τις Μεθόδους 2 και 3, από το σύνολο των τοξικών δειγμάτων νανοσωματιδίων, μόνο 2 δεν προβλέφθηκαν επιτυχώς στο σύνολο των τριών δοκιμών.

Ομοίως παρουσιάζονται τα αποτελέσματα για τις μεθόδους 4-Fold και 5-Fold.

Πίνακας 6.1.26: Στατιστικά αποτελέσματα με εφαρμογή 4-Fold στο σύνολο MeOx για τις τρεις διαφορετικές μεθόδους

Μέθοδος	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.78	0.29	1.00	0.22	0.47	$\begin{bmatrix} 2 & 5 \\ 0 & 16 \end{bmatrix}$
2	0.74	1.00	0.63	0.26	0.58	$\begin{bmatrix} 7 & 0 \\ 6 & 10 \end{bmatrix}$
3	0.69	1.00	0.56	0.31	0.53	$\begin{bmatrix} 7 & 0 \\ 7 & 9 \end{bmatrix}$

Στην περίπτωση της 4-Fold επικύρωσης, τα στατιστικά αποτελέσματα και για τις τρεις μεθοδολογίες είναι λιγότερο ικανοποιητικά συγκριτικά με την 3-Fold επικύρωση. Τις μικρότερες διακυμάνσεις σε συνδυασμό με την καλύτερη αθροιστική απόδοση επιδεικνύει η Μέθοδος 2, με ακρίβεια 0.74 αν και δεν προβλέπει επιτυχώς 6 εκ των τοξικών δειγμάτων νανοσωματιδίων. Οι Μέθοδοι 2 και 3 , έχουν παρόμοια επίδοση ενώ η Μέθοδος 1 επιδεικνύει πλήρη αδυναμία στη πρόβλεψη των μη τοξικών δειγμάτων παρουσιάζοντας πολύ μικρή ευαισθησία.

Πίνακας 6.1.27: Στατιστικά αποτελέσματα με εφαρμογή 5-Fold στο σύνολο MeOx για τις τρεις διαφορετικές μεθόδους

Μέθοδος	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.78	0.29	1.00	0.22	0.47	$\begin{bmatrix} 2 & 5 \\ 0 & 16 \end{bmatrix}$
2	0.69	0.86	0.63	0.31	0.44	$\begin{bmatrix} 6 & 1 \\ 6 & 10 \end{bmatrix}$
3	0.74	0.86	0.69	0.26	0.50	$\begin{bmatrix} 6 & 1 \\ 5 & 11 \end{bmatrix}$

Στην περίπτωση της 5-Fold επικύρωσης, τα στατιστικά αποτελέσματα και για τις τρεις μεθοδολογίες δεν είναι ικανοποιητικά, με πολλές περιπτώσεις δοκιμών απόλυτα αποτυχημένες και μη ικανές να προβλέψουν ορθά τα μη τοξικά και τοξικά δείγματα. Τις μικρότερες διακυμάνσεις σε συνδυασμό με την καλύτερη αθροιστική απόδοση επιδεικνύει η Μέθοδος 3, με ακρίβεια 0.74. Από το σύνολο των τοξικών δειγμάτων νανοσωματιδίων, 5 δεν προβλέφθηκαν επιτυχώς στο

σύνολο των τριών δοκιμών. Αντίθετα, η Μέθοδος 1, προέβλεψε απόλυτα επιτυχημένα τα μη τοξικά δείγματα αλλά επέδειξε απόλυτη αποτυχία στον εντοπισμό των μη τοξικών δειγμάτων με ποσοστό επιτυχίας 2 σε σύνολο 7 δειγμάτων.

Ως προς το υπολογιστικό κόστος των τριών εναλλακτικών μεθόδων, τα αποτελέσματα συνοψίζονται στον πίνακα 6.1.28.

Πίνακας 6.1.28: Υπολογιστικοί χρόνοι σε δευτερόλεπτα για την εφαρμογή του μοντέλου μέσω K-Fold στο σύνολο MeOx

	Μέθοδος 1	**Μέθοδος 2**	**Μέθοδος 3**
3-Fold	1.47	1.80	1.57
4-Fold	2.71	2.78	2.93
5-Fold	3.57	3.38	3.98

Αν και υπάρχουν διακυμάνσεις ως προς την ταχύτερη μέθοδο στις τρεις δοκιμές, κατά μέσο όρο, η Μέθοδος 1 αποδεικνύεται γρηγορότερη. Υπό το ίδιο σκεπτικό, βάσει της διαδικασίας Leave-One-Out, η μέθοδος εφαρμόζεται 23 φορές. Τα αποτελέσματα για τις τρείς μεθοδολογίες συνοψίζονται στον πίνακα 6.1.29.

Πίνακας 6.1.29: Στατιστικά αποτελέσματα με εφαρμογή Leave-One-Out στο σύνολο MeOx για τις τρεις διαφορετικές μεθόδους

Μέθοδος	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.70	0.29	0.88	0.30	0.20	$\begin{bmatrix} 2 & 5 \\ 2 & 14 \end{bmatrix}$
2	0.61	1.00	0.44	0.39	0.44	$\begin{bmatrix} 7 & 0 \\ 9 & 7 \end{bmatrix}$
3	0.61	1.00	0.44	0.39	0.44	$\begin{bmatrix} 7 & 0 \\ 9 & 7 \end{bmatrix}$

Κατά τις δοκιμές Leave-One-Out το μοντέλο εμφανίζει σημαντικές ελλείψεις στην ορθή πρόβλεψη των τοξικών και μη τοξικών δειγμάτων. Οι Μέθοδοι 2 και 3 παρουσιάζουν παρόμοια συμπεριφορά με μετρούμενη ακρίβεια και ειδικότητα 0.61 και 0.44, αντίστοιχα. Αντίθετα, η Μέθοδος 1 καταγράφει ειδικότητα 0.88 αλλά πολύ χαμηλότερη ευαισθησία στο 0.29.

Έλεγχος τυχαίας επιλογής

Ο έλεγχος y-scrambling πραγματοποιήθηκε με Kennard Stone και κλάσμα διαμέρισης 0.65, επανατοποθετώντας τη μεταβλητή πρόβλεψης του συνόλου ελέγχου ώστε το μοντέλο να εκπαιδευτεί σε μη έγκυρα δεδομένα. Τα αποτελέσματα των 5 διαφορετικών τυχαίων δοκιμών που πραγματοποιήθηκαν, με αριθμό δειγμάτων εκπαίδευσης και ελέγχου 15 και 8 αντίστοιχα, συνοψίζονται στον πίνακα 6.1.30.

Πίνακας 6.1.30: Στατιστικά αποτελέσματα με εφαρμογή y-scrambling στο σύνολο MeOx

Accuracy	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
0.38	0.00	0.50	0.62	-0.45	$\begin{bmatrix} 0 & 2 \\ 3 & 3 \end{bmatrix}$
0.75	0.00	1.00	0.25	0.00	$\begin{bmatrix} 0 & 2 \\ 0 & 6 \end{bmatrix}$
0.38	0.00	0.50	0.62	-0.45	$\begin{bmatrix} 0 & 2 \\ 3 & 3 \end{bmatrix}$
0.62	0.00	0.83	0.38	-0.22	$\begin{bmatrix} 0 & 2 \\ 1 & 5 \end{bmatrix}$
0.50	0.00	0.66	0.50	-0.33	$\begin{bmatrix} 0 & 2 \\ 2 & 4 \end{bmatrix}$

Οι τιμές ακρίβειας $[0.38, 0.75, 0.38, 0.62, 0.50]$ υποδηλώνουν ένα αποτυχημένο μοντέλο το οποίο κάνει χειρότερες εκτιμήσεις συγκριτικά με ένα τυχαίο μοντέλο. Αυτό επιβεβαιώνει και την υπόθεση πως οι επιτυχημένες προβλέψεις οφείλονται στην σωστή δόμηση του μοντέλου και δεν είναι αποτέλεσμα τυχαιότητας.

Πεδίο εφαρμογής μοντέλου

Το αριθμητικό κατώφλι που αφορά στο πεδίου εφαρμογής του μοντέλου σε *training set* καθορισμένο με Kennard Stone για train_ratio 0.70 υπολογίζεται 3.19.

Ο υπολογισμός της τιμής μόχλευσης καθενός εκ των 7 δειγμάτων δεν αναδεικνύει κάποιο δείγμα με τιμή μόχλευσης μεγαλύτερη του ανώτατου ορίου που ορίζει το υπολογιζόμενο κατώφλι. Έτσι, το σύνολο των προβλέψεων θεωρείται αξιόπιστο.[3]

Πίνακας 7.1.1: Κατώφλι πεδίου εφαρμογής για το σύνολο MeOx

Train Ratio	***Κατώφλι***	***Δείγματα που ικανοποιούν το κατώφλι***
0.60	5.14	9/9
0.65	4.80	8/8
0.70	4.50	7/7
0.75	4.24	6/6
0.80	4.00	5/5

Αποτίμηση της επίδοσης του μοντέλου

Η προβλεπτική ικανότητα του μοντέλου στο σύνολο δεδομένων MeOx είναι ικανοποιητική καθώς οι επιδόσεις σε εξωτερική και εσωτερική αξιολόγηση αλλά

[3] Στον υπολογισμό του κατωφλιού του πεδίου εφαρμογής και των h_i δεν έχουν ληφθεί υπόψιν τα descriptors που προκύπτουν μέσω υπολογισμών ή άλλων μετρούμενων φυσικοχημικών ιδιοτήτων.

και ελέγχους στο πεδίο εφαρμογής και στην μέθοδο y-scrambling, είναι σχετικά υψηλές. Το μοντέλο που διαμορφώνεται συνδυάζει επαρκή ειδίκευση ώστε να εντοπίζει σε μεγάλο βαθμό τα τοξικά δείγματα.

Την καλύτερη επίδοση μεταξύ όλων των ελέγχων φέρει η διαμέριση με Kennard Stone για κλάσμα διαμέρισης 0.60 το οποίο χωρίζει το σύνολο δεδομένων σε 14 δείγματα εκπαίδευσης και 9 δείγματα ελέγχου. Εξ αυτών, όλα πλην ενός κατηγοριοποιούνται επιτυχώς με εφαρμογή της Μεθόδου Κατηγοριοποίησης 3, η οποία έχει συνολικά και την βέλτιστη επίδοση ως προς τον εντοπισμό των τοξικών δειγμάτων με ειδικότητα η οποία διατηρείται κοντά στο 1.

Αν και κατά την εφαρμογή της εσωτερικής αξιολόγησης, η απόδοση του μοντέλου, ανεξαρτήτως της Μεθόδου Κατηγοριοποίησης που χρησιμοποιείται, δεν είναι ιδιαίτερα ενθαρρυντική αλλά χαρακτηρίζεται από έντονες διακυμάνσεις και μεγάλο εύρος ακρίβειας και ειδικότητας, καταλληλότερη μέθοδος πρόβλεψης για το σύνολο δεδομένων MeOx χαρακτηρίζεται η Μέθοδος 3, ιδιαίτερα κατά την εφαρμογή εξωτερικής αξιολόγησης με Kennard Stone και training ratio 0.60.

6.1.4. Σύνολο δεδομένων Cytotox

Το σύνολο Cytotox αποτελείται από 494 δείγματα και 65 descriptors. Λόγω του μεγέθους καθίσταται δύσκολη η διαχείριση του με επαναλαμβανόμενη εσωτερική αξιολόγηση και επιλέγεται η εκτενέστερη εξωτερική αξιολόγηση.

Οι συντελεστές της δημοσίευσης από την οποία προέρχεται το σύνολο δεδομένων το χωρίζουν σε σύνολο εκπαίδευσης και ελέγχου με τυχαίο τρόπο, διατηρώντας παρόμοιες αναλογίες τοξικών και μη τοξικών δειγμάτων στα δύο υποσύνολα. Αυτή η διαμέριση οδηγεί σε ένα σύνολο εκπαίδευσης που αποτελείται από 37 τοξικά και 308 μη τοξικά δείγματα καθώς και ένα σύνολο επικύρωσης το οποίο αποτελείται από 14 τοξικά και 135 μη τοξικά δείγματα. Βάσει αυτής της διαμέρισης (για λόγους σύγκρισης) και ακολουθώντας ακριβώς την ίδια μεθοδολογία προκύπτουν τα αποτελέσματα για τις τρεις μεθόδους στον πίνακα 6.1.32.

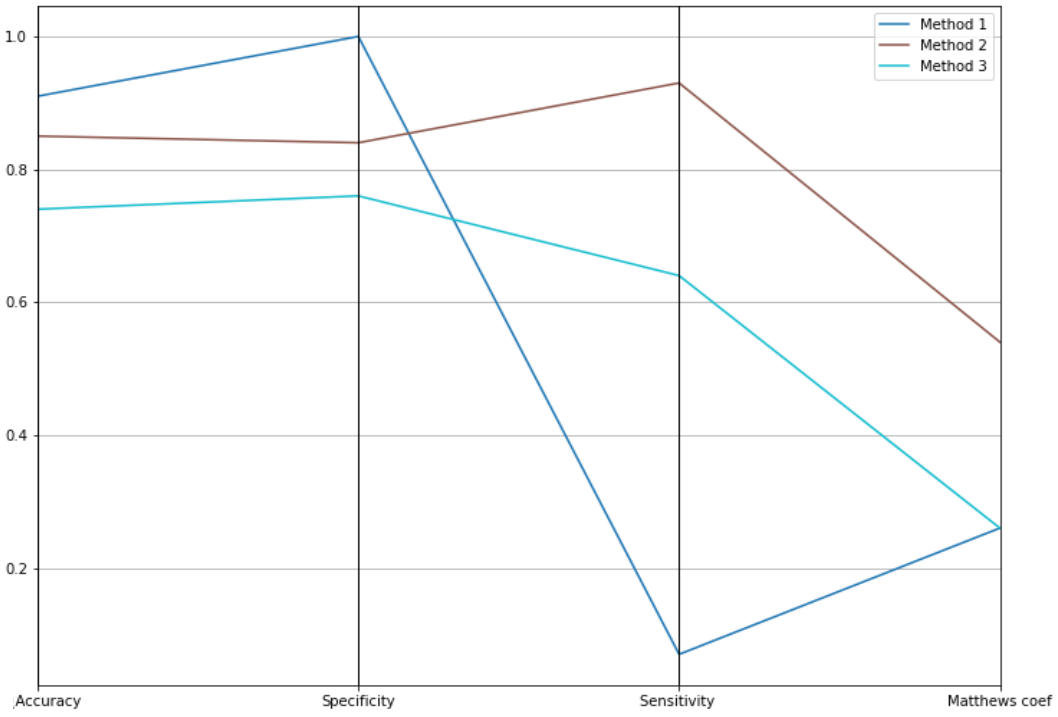
Πίνακας 6.1.32: Στατιστικά αποτελέσματα με εφαρμογή της προτεινόμενης διαμέρισης στο σύνολο Cytotox για τις Μεθόδους 1, 2 και 3

Μέθοδος	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.91	1.00	0.07	0.09	0.26	$\begin{bmatrix} 135 & 0 \\ 13 & 1 \end{bmatrix}$
2	0.85	0.84	0.93	0.15	0.54	$\begin{bmatrix} 114 & 21 \\ 1 & 13 \end{bmatrix}$
3	0.74	0.76	0.64	0.26	0.26	$\begin{bmatrix} 102 & 33 \\ 5 & 9 \end{bmatrix}$

Εφαρμόζοντας την προτεινόμενη διαμέριση, τα αποτελέσματα είναι ανάμεικτα. Η Μέθοδος 1 αποτυγχάνει πλήρως στον εντοπισμό των τοξικών δειγμάτων ενώ η

Μέθοδος 3 εμφανίζει αποδεκτά αποτελέσματα εντοπίζοντας 9 εκ των 14 δειγμάτων. Αντίθετα, η Μέθοδος 2 έχει την καλύτερη επίδοση εντοπίζοντας 13 εκ των 14 τοξικών και 114 εκ των 135 μη τοξικών δειγμάτων.

Συνολικά για τις τρείς μεθοδολογίες , οι ακρίβειες που καταγράφονται είναι [0.91,0.85,0.74], αντίστοιχα ενώ οι συντελεστές συσχέτισης Matthews είναι [0.26,0.54,0.24], επιβεβαιώνοντας την επιτυχία του μοντέλου με εφαρμογή της Μεθόδου Κατηγοριοποίησης 2.



Γράφημα 7.1.4: Σύγκριση στατιστικών επιδόσεων για τις τρεις εναλλακτικές μεθόδους με εφαρμογή στο σύνολο Cytotox βάσει της προτεινόμενης διαμέρισης

Ακολούθησε και η εξωτερική αξιολόγηση του μοντέλου στο συγκεκριμένο σύνολο δεδομένων βάσει της διαμέρισης *Kennard Stone*, τα αποτελέσματα της οποία παρουσιάζονται στους πίνακες 6.1.33-6.1.36.

Πίνακας 6.1.33: Στατιστικά αποτελέσματα με εφαρμογή Kennard Stone για training ratio 0.6 στο σύνολο Cytotox

Μέθοδος	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.55	0.56	0.20	0.45	-0.10	$\begin{bmatrix} 106 & 82 \\ 8 & 2 \end{bmatrix}$
2	0.95	0.97	0.60	0.05	0.55	$\begin{bmatrix} 183 & 5 \\ 4 & 6 \end{bmatrix}$
3	0.98	0.99	0.90	0.02	0.85	$\begin{bmatrix} 186 & 2 \\ 1 & 9 \end{bmatrix}$

Πίνακας 6.1.34: Στατιστικά αποτελέσματα με εφαρμογή Kennard Stone για training ratio 0.65 στο σύνολο Cytotox

Μέθοδος	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.95	1.00	0.00	0.05	0.00	$\begin{bmatrix} 164 & 0 \\ 9 & 0 \end{bmatrix}$
2	0.94	0.95	0.89	0.06	0.62	$\begin{bmatrix} 155 & 9 \\ 1 & 8 \end{bmatrix}$
3	0.84	0.85	0.67	0.16	0.31	$\begin{bmatrix} 140 & 24 \\ 3 & 6 \end{bmatrix}$

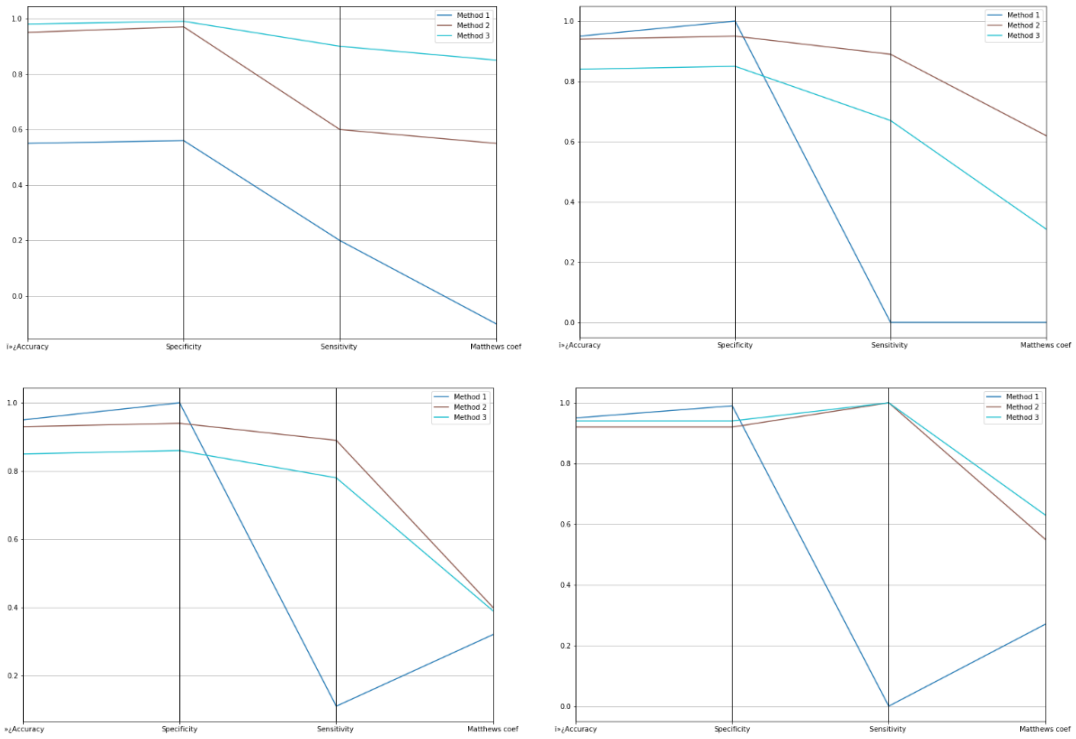
Πίνακας 6.1.35: Στατιστικά αποτελέσματα με εφαρμογή Kennard Stone για training ratio 0.7 στο σύνολο Cytotox

Μέθοδος	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.95	1.00	0.11	0.05	0.32	$\begin{bmatrix} 139 & 0 \\ 8 & 1 \end{bmatrix}$
2	0.93	0.94	0.89	0.07	0.40	$\begin{bmatrix} 130 & 9 \\ 1 & 8 \end{bmatrix}$
3	0.85	0.86	0.78	0.15	0.39	$\begin{bmatrix} 119 & 20 \\ 2 & 7 \end{bmatrix}$

Πίνακας 6.1.36: Στατιστικά αποτελέσματα με εφαρμογή Kennard Stone για training ratio 0.75 στο σύνολο Cytotox

Μέθοδος	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.95	0.99	0.00	0.05	0.27	$\begin{bmatrix} 118 & 1 \\ 5 & 1 \end{bmatrix}$
2	0.92	0.92	1.00	0.08	0.55	$\begin{bmatrix} 109 & 10 \\ 0 & 5 \end{bmatrix}$
3	0.94	0.94	1.00	0.06	0.63	$\begin{bmatrix} 112 & 7 \\ 0 & 5 \end{bmatrix}$

Η εφαρμογή της Kennard Stone για 4 διαφορετικά training ratios ακολουθεί παρόμοια συμπεριφορά με αυτή της προτεινόμενης διαμέρισης. Αν και η Μέθοδος 1 αποδεικνύεται πλήρως αποτυχημένη χωρίς καμία ικανότητα εντοπισμού των τοξικών δειγμάτων, οι Μέθοδοι 2 και 3 επιδεικνύουν καλύτερη συμπεριφορά. Μάλιστα, η Μέθοδος 2 έχει ακρίβεια και ειδικότητα για τα 4 διαφορετικά κλάσματα διαμέρισης [0.95, 0.94, 0.93, 0.92] και [0.60, 0.89, 0.89, 1.00] αντίστοιχα. Η Μέθοδος 3 έχει ικανοποιητικά αποτελέσματα με ακρίβεια και ειδικότητα για τα 4 διαφορετικά κλάσματα διαμέρισης [0.95, 0.84, 0.85, 0.94] και [0.90, 0.67, 0.78, 1.00] αντίστοιχα.



Γράφημα 7.1.5: Σύγκριση στατιστικών επιδόσεων για τις τρεις εναλλακτικές μεθόδους με εφαρμογή στο σύνολο Cytotox με Kennard Stone και training ratio=0.60 (πάνω αριστερά), 0.65 (πάνω δεξιά), 0.70 (κάτω αριστερά) και 0.75 (κάτω δεξιά)

Έλεγχος τυχαίας επιλογής

Ο έλεγχος y-scrambling πραγματοποιήθηκε με δοκιμές στον προτεινόμενο τυχαίο τρόπο διαμέρισης, επανατοποθετώντας τη μεταβλητή πρόβλεψης του συνόλου ελέγχου ώστε το μοντέλο να εκπαιδευτεί σε μη έγκυρα δεδομένα. Τα αποτελέσματα των 5 διαφορετικών τυχαίων δοκιμών που πραγματοποιήθηκαν, με αριθμό δειγμάτων εκπαίδευσης και ελέγχου 345 και 149 αντίστοιχα, συνοψίζονται στον πίνακα 6.1.37.

Πίνακας 6.1.37: Στατιστικά αποτελέσματα με εφαρμογή y-scrambling στο σύνολο Cytotox για την Μέθοδο 2

Accuracy	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
0.20	0.16	0.57	0.80	-0.20	$\begin{bmatrix} 22 & 113 \\ 6 & 8 \end{bmatrix}$
0.49	0.47	0.70	0.51	0.11	$\begin{bmatrix} 64 & 71 \\ 4 & 10 \end{bmatrix}$
0.35	0.30	0.78	0.65	0.06	$\begin{bmatrix} 41 & 94 \\ 3 & 11 \end{bmatrix}$
0.44	0.45	0.36	0.56	-0.11	$\begin{bmatrix} 61 & 74 \\ 5 & 9 \end{bmatrix}$

0.17	0.11	0.79	0.87	-0.09	$\begin{bmatrix} 15 & 120 \\ 3 & 11 \end{bmatrix}$

Οι τιμές ακρίβειας $[0.20, 0.49, 0.35, 0.44, 0.17]$ υποδηλώνουν ένα αποτυχημένο μοντέλο το οποίο κάνει χειρότερες εκτιμήσεις συγκριτικά με ένα τυχαίο μοντέλο. Αυτό επιβεβαιώνει και την υπόθεση πως οι επιτυχημένες προβλέψεις οφείλονται στην σωστή δόμηση του μοντέλου και δεν είναι αποτέλεσμα τυχαιότητας.

Πεδίο εφαρμογής μοντέλου

Το αριθμητικό κατώφλι που αφορά στο πεδίου εφαρμογής του μοντέλου στο προτεινόμενο σύνολο εκαπίδευσης υπολογίζεται 0.57.

Ο υπολογισμός της τιμής μόχλευσης καθενός εκ των 149 δειγμάτων δεν αναδεικνύει κάποιο δείγμα με τιμή μόχλευσης μεγαλύτερη του ανώτατου ορίου που ορίζει το υπολογιζόμενο κατώφλι. Έτσι, το σύνολο των προβλέψεων θεωρείται αξιόπιστο.

Αποτίμηση της επίδοσης του μοντέλου

Η προβλεπτική ικανότητα του μοντέλου στο σύνολο δεδομένων Cytotox είναι ικανοποιητική καθώς οι επιδόσεις σε εξωτερική αξιολόγηση αλλά και ελέγχους στο πεδίο εφαρμογής και στην μέθοδο y-scrambling, είναι σχετικά υψηλές. Το μοντέλο που διαμορφώνεται εντοπίζει σε επαρκή βαθμό τα τοξικά δείγματα.

Την καλύτερη επίδοση μεταξύ όλων των ελέγχων φέρει η διαμέριση με Kennard Stone για κλάσμα διαμέρισης 0.75. Εξ των τοξικών δειγμάτων του συνόλου εκπαίδευσης όλα κατηγοριοποιούνται επιτυχώς με εφαρμογή της Μεθόδου Κατηγοριοποίησης 2 , η οποία έχει συνολικά και την βέλτιστη επίδοση ως προς τον εντοπισμό των τοξικών δειγμάτων με ειδικότητα η οποία διατηρείται κοντά στο 1, σε όλες τις δοκιμές.

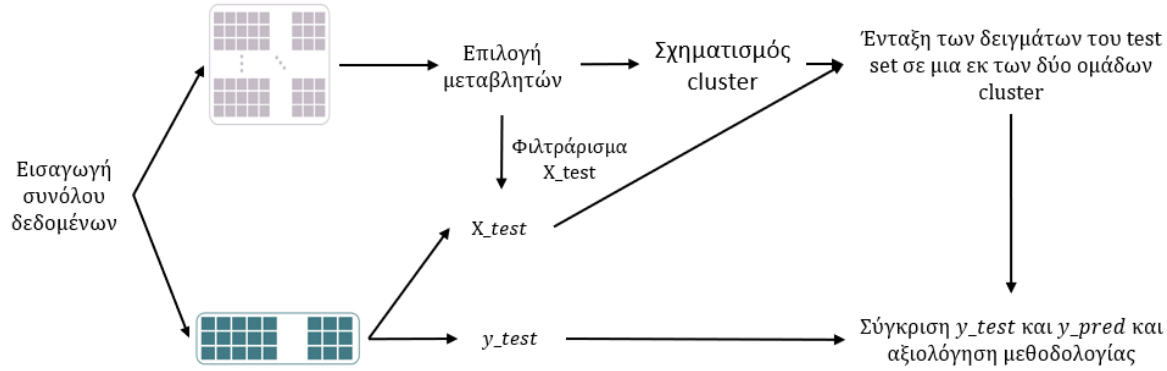
Λαμβάνοντας υπόψιν τη συνολική απόδοση στις δοκιμές και κυρίως την ικανότητα εντοπισμού των τοξικών δειγμάτων, καταλληλότερη μέθοδος πρόβλεψης για το σύνολο δεδομένων Cytotox χαρακτηρίζεται η Μέθοδος 2, ιδιαίτερα κατά την εφαρμογή εξωτερικής αξιολόγησης με Kennard Stone και training ratio 0.75.

6.2. Πρόβλεψη τοξικότητας εφαρμόζοντας επιλογή μεταβλητών

Με σκοπό την αύξηση της ευρωστίας του μοντέλου και την καλύτερη πρόβλεψη με μικρότερο χρονικό και υπολογιστικό κόστος, πραγματοποιούνται δοκιμές και

στην περίπτωση μείωσης των μεταβλητών-ιδιοτήτων που αξιοποιούνται από το μοντέλο. Οι μεταβλητές με την καλύτερη συσχέτιση με τη μεταβλητή πρόβλεψης διατηρούνται ενώ οι υπόλοιπες δεν λαμβάνονται υπόψιν στον σχηματισμό των cluster και την πρόβλεψη της κλάσης των δειγμάτων του συνόλου ελέγχου.

Η επιλογή των καλύτερα σχετιζόμενων μεταβλητών πραγματοποιείται μέσω των μεθόδων υπολογισμού συντελεστών συσχέτισης (correlation coefficient) chi2 και f_classif οι οποίες υπάρχουν ως εργαλεία της βιβλιοθήκης sklearn της Python και βασίζονται στις τιμές των δεικτών x-square και ANOVA αντίστοιχα. Η επιλογή των βέλτιστων μεταβλητών πραγματοποιείται βάσει των συσχετίσεων μεταξύ μεταβλητών και μεταβλητή πρόβλεψης στο σύνολο εκπαίδευσης, χωρίς να λαμβάνεται υπόψιν το σύνολο ελέγχου και να επηρεάσει την επιλογή. Στη συνέχεια, το σύνολο ελέγχου φιλτράρεται βάσει των ήδη επιλεγμένων ιδιοτήτων.



Σχήμα.6.2.1 : Σχηματική απεικόνιση της διαδικασίας εξαγωγής αποτελεσμάτων με επιλογή μεταβλητών του συνόλου προς εξέταση

Εξαιτίας του μεγάλου αριθμού των μεταβλητών και δειγμάτων, η διαδικασία επιλογής μεταβλητών εφαρμόζεται μόνο στο σύνολο Cytotox. Οι δοκιμές που πραγματοποιήθηκαν στα υπόλοιπα σύνολα δεδομένων όταν απομακρύνονταν κάποιες εκ των μεταβλητών, δεν απέδωσαν αξιόλογα αποτελέσματα και ως εκ τούτου δεν συμπεριλήφθηκαν.

6.2.1. Σύνολο δεδομένων Cytotox

Λόγω του μεγάλου αριθμού μεταβλητών και δειγμάτων για το σύνολο Cytotox ενδείκνυται η εφαρμογή μεθόδου επιλογής μεταβλητών πριν την μοντελοποίηση με την προτεινόμενη μεθοδολογία. Η εκπαίδευση του μοντέλου επιλέγεται να γίνει βάσει του προτεινόμενου συνόλου εκπαίδευσης που αποτελείται από 345 δείγματα, 37 εκ των οποίων είναι τοξικά. Ως προς τις μεθόδους επιλογής μεταβλητών, εξετάζονται διαφορετικοί αριθμοί επιλεγμένων μεταβλητών καθώς και διαφορετικά κριτήρια υπολογισμού της συσχέτισης μεταξύ των μεταβλητών και της μεταβλητής πρόβλεψης, όπως το ο έλεγχος chi-squared της βιβλιοθήκης

sklearn. Παράλληλα, δοκιμάζεται και η εφαρμογή του μοντέλου χωρίς τις υπολογισμένες παραμέτρους, οι οποίες δεν αντιστοιχούν σε κάποια φυσικοχημική ιδιότητα των νανοσωματιδίων (no computed variables). Τέλος, πραγματοποιείται δοκιμή που περιλαμβάνει μόνο τις προτεινόμενες από τους συγγραφείς μεταβλητές, οι οποίες είχαν τη βέλτιστη συμπεριφορά στο μοντέλο της δημοσίευσης των Papadiamantis *et al.* (2020) (selected variables). Τα στατιστικά αποτελέσματα των δοκιμών συνοψίζονται στον πίνακα 6.2.1.

Πίνακας 6.2.1: Στατιστικά αποτελέσματα με την προτεινόμενη διαμέριση και επιλογή μεταβλητών στο σύνολο Cytotox

Μέθοδος επιλογής μεταβλητών	***Μέθοδος Κατηγ.***	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Conf. Matrix***
chi squared n=6	1	0.88	0.95	0.21	0.12	0.19	$\begin{bmatrix}128 & 7\\ 11 & 3\end{bmatrix}$
	2	0.79	0.77	0.93	0.21	0.45	$\begin{bmatrix}104 & 31\\ 1 & 13\end{bmatrix}$
	3	0.85	0.89	0.50	0.15	0.32	$\begin{bmatrix}120 & 15\\ 7 & 7\end{bmatrix}$
chi squared n=12	1	0.91	1.00	0.00	0.09	0.00	$\begin{bmatrix}135 & 0\\ 14 & 0\end{bmatrix}$
	2	0.89	0.95	0.36	0.11	0.33	$\begin{bmatrix}128 & 7\\ 9 & 5\end{bmatrix}$
	3	0.81	0.82	0.71	0.19	0.37	$\begin{bmatrix}111 & 24\\ 4 & 10\end{bmatrix}$
no computed variables	1	0.88	0.95	0.21	0.12	0.19	$\begin{bmatrix}128 & 7\\ 11 & 3\end{bmatrix}$
	2	0.79	0.77	0.93	0.21	0.45	$\begin{bmatrix}104 & 31\\ 1 & 13\end{bmatrix}$
	3	0.85	0.89	0.5	0.15	0.32	$\begin{bmatrix}120 & 15\\ 7 & 7\end{bmatrix}$
selected variables	1	0.91	1.00	0.00	0.09	0.00	$\begin{bmatrix}135 & 0\\ 14 & 0\end{bmatrix}$
	2	0.69	0.67	0.93	0.31	0.36	$\begin{bmatrix}90 & 45\\ 1 & 13\end{bmatrix}$
	3	0.82	0.81	0.93	0.18	0.42	$\begin{bmatrix}109 & 26\\ 1 & 13\end{bmatrix}$

Μεταξύ των διαφορετικών μεθόδων κατηγοριοποίησης, η Μέθοδος 2 επιδεικνύει την καλύτερη προβλεπτική ικανότητα, εντοπίζοντας τα περισσότερα τοξικά δείγματα στις περισσότερες δοκιμές επιτυγχάνοντας ακρίβεια και ειδικότητα [0.79, 0.89, 0.79, 0.69] και [0.93, 0.36, 1.00, 0.93, 0.93], αντίστοιχα. Η Μέθοδος 1 επιδεικνύει την χειρότερη συμπεριφορά, αποτυγχάνοντας πλήρως στον εντοπισμό των τοξικών δειγμάτων με ειδικότητα [0.21, 0.00, 0.21, 0.00] ενώ η Μέθοδος 3 παρουσιάζει καλύτερες προβλέψεις με ειδικότητα [0.50, 0.71, 0.50, 0.93].

Τα καλύτερα αποτελέσματα αναφορικά με το κριτήριο επιλογής μεταβλητών επιδεικνύουν από κοινού οι δοκιμές στις οποίες συμμετέχουν μόνο οι φυσικοχημικές μεταβλητές καθώς και οι δοκιμές chi-squared με n=6, κατά τις

οποίες εντοπίζεται σχεδόν το σύνολο των τοξικών δειγμάτων του συνόλου ελέγχου με ακρίβεια, ευαισθησία και ειδικότητα 0.79, 0.77 και 0.93, αντίστοιχα. Τα αποτελέσματα είναι συγκρίσιμα με αυτά των δοκιμών που δεν περιλαμβάνουν επιλογή μεταβλητών, χωρίς κάποια αισθητή βελτίωση. Ως προς τον υπολογιστικό χρόνο που απαιτεί η μοντελοποίηση, οι δοκιμές που περιλαμβάνουν επιλογή μεταβλητών επιτυγχάνουν μικρότερους χρόνους, μειώνοντάς τους στο 2/3 του μεγέθους περίπου, όπως φαίνεται και στον πίνακα 6.2.2.

Πίνακας 6.2.2: Σύγκριση υπολογιστικών χρόνων σε δευτερόλεπτα με και χωρίς επιλογή μεταβλητών στο σύνολο Cytotox

Μέθοδος	***CPU time χωρίς επιλογή***	***CPU time chi squared n=6***	***CPU time chi squared n=12***	***CPU time no computed variables***	***CPU time selected variables***
1	365	263	290	263	193
2	364	262	288	261	191
3	367	262	288	260	193

Δοκιμές 10-Fold επικύρωσης

Για την περαιτέρω ανάλυση της βέλτιστης διαδικασίας μείωσης των μεγάλων διαστάσεων του συνόλου, δοκιμάζεται η 10-Fold εσωτερική αξιολόγηση. Διαιρώντας το σύνολο σε 10 υποσύνολα, πραγματοποιούνται 10 διαδοχικές μοντελοποιήσεις και προβλέψεις. Σε κάθε μία, ένα εκ των υποσυνόλων λειτουργεί ως σύνολο ελέγχου. Δοκιμάζονται οι διαδικασίες επιλογής μεταβλητών: **chi squared** με n=6 και **chi squared** με n=12. Μετά από τις 10 δοκιμές, διαμορφώνονται οι συνολικοί πίνακες των στατιστικών μεγεθών για τις τρεις μεθόδους κατηγοριοποίησης.

Πίνακας 6.2.3: Στατιστικά αποτελέσματα με εφαρμογή επιλογής μεταβλητών με δύο τρόπους στο σύνολο εκπαίδευσης του συνόλου Cytotox

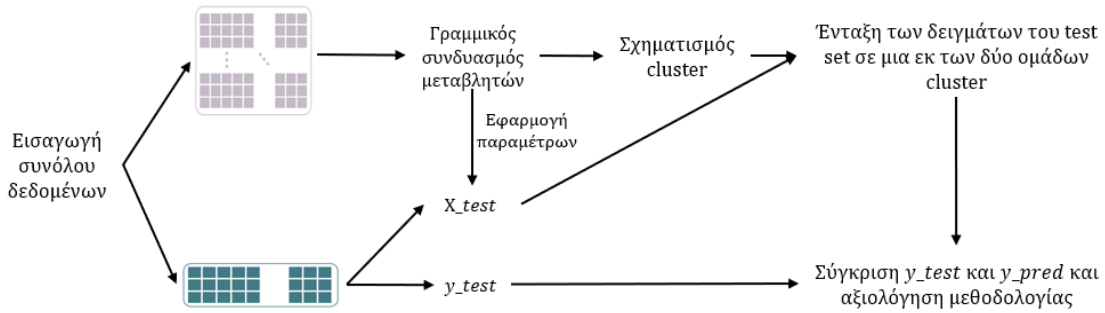
Μέθοδος επιλογής μεταβλητών	***Μέθοδος Κατηγ.***	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Conf. Matrix***
chi squared n=8	1	0.92	1.00	0.22	0.08	0.44	$\begin{bmatrix} 308 & 0 \\ 29 & 8 \end{bmatrix}$
	2	0.88	0.90	0.70	0.12	0.50	$\begin{bmatrix} 277 & 31 \\ 11 & 26 \end{bmatrix}$
	3	0.74	0.75	0.65	0.26	0.27	$\begin{bmatrix} 230 & 78 \\ 13 & 24 \end{bmatrix}$
chi squared n=12	1	0.92	1.00	0.16	0.09	0.38	$\begin{bmatrix} 308 & 0 \\ 31 & 6 \end{bmatrix}$
	2	0.89	0.91	0.73	0.11	0.55	$\begin{bmatrix} 281 & 27 \\ 10 & 27 \end{bmatrix}$
	3	0.68	0.68	0.65	0.33	0.21	$\begin{bmatrix} 209 & 99 \\ 13 & 24 \end{bmatrix}$

Η σύγκριση θα πραγματοποιηθεί στη συνέχεια ακολουθώντας αντίστοιχη διαδικασία με εφαρμογή ανάλυσης κυρίων συνιστωσών στο σύνολο.

6.3. Πρόβλεψη τοξικότητας εφαρμόζοντας ανάλυση κύριων συνιστωσών

Εναλλακτικά της επιλογής μεταβλητών δοκιμάζεται η ανάλυση κυρίων συνιστωσών PCA, βάσει της οποίας οι n ανεξάρτητες μεταβλητές-ιδιότητες συνδυάζονται γραμμικά και συνοψίζονται σε m<n με σκοπό την μείωση των διαστάσεων, τη συμπύκνωση της πληροφορίας και της εξοικονόμηση υπολογιστικού κόστους.

Ομοίως με τη διαδικασία επιλογής μεταβλητών, το σύνολο δεδομένων χωρίζεται σε σύνολο εκπαίδευσης και ελέγχου. Η ανάλυση κύριων συνιστωσών PCA εφαρμόζεται στο σύνολο εκπαίδευσης και το τροποποιεί κατάλληλα πριν την διαμόρφωση των συστάδων. Οι συντελεστές του γραμμικού συνδυασμού αποθηκεύονται και, στη συνέχεια, εφαρμόζονται στο σύνολο επικύρωσης, ώστε να υπάρχει ίδια προεπεξεργασία στα δύο υποσύνολα.

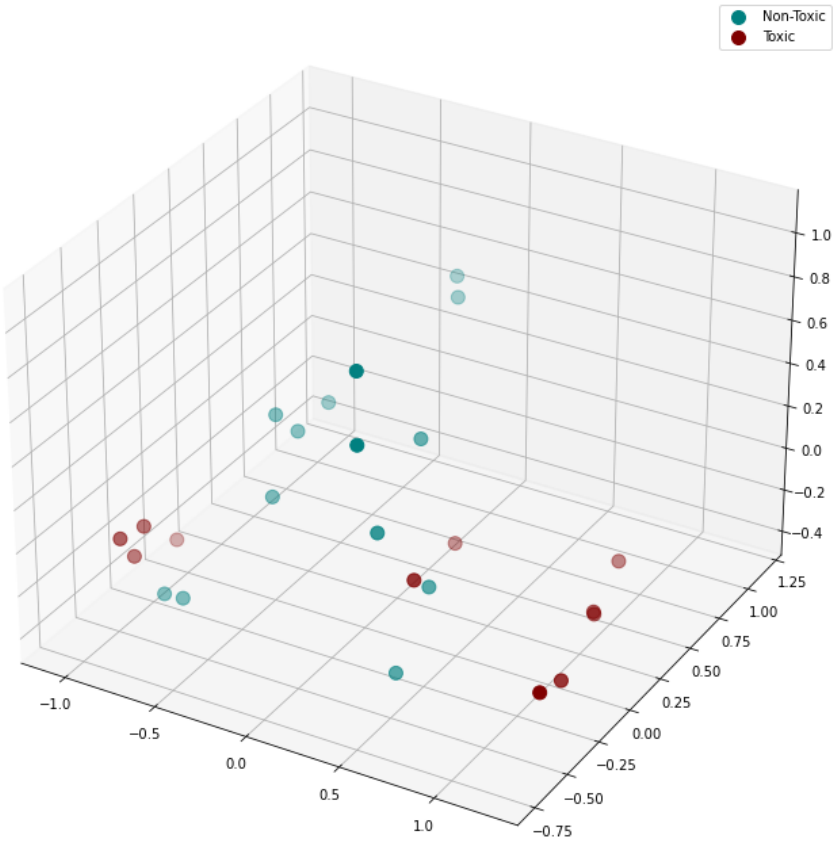


Σχήμα 6.3.1: Σχηματική απεικόνιση της διαδικασίας εξαγωγής αποτελεσμάτων με εφαρμογή της ανάλυσης κύριων συνιστωσών PCA στο σύνολο προς εξέταση

Η επικύρωση των συνόλων δεδομένων πραγματοποιείται με τυχαία διαμέριση και όχι με χρήση της Kennard Stone. Καθώς η Kennard Stone θα προηγούνταν της αναδιαμόρφωσης των εξαρτημένων μεταβλητών, η διαμέριση που θα είχε τελεστεί δεν θα αντιστοιχούσε στις νέες τιμές των μεταβλητών και πιθανώς να μην αποτελούσε αντιπροσωπευτικό δείγμα του νέου δειγματικού χώρου, το οποίο και χρησιμοποιείται στο μοντέλο. Απουσία άλλου κριτηρίου ή μεθόδου χωρισμού, η διαμέριση πραγματοποιείται τυχαία.

6.3.1. Σύνολο δεδομένων MeHydOx

Το σύνολο δεδομένων MeHydOx μετά από επεξεργασία PCA και μείωση των διαστάσεων των μεταβλητών του σε 3, διαμορφώνει ένα νέο δειγματικό χώρο (βλέπε Σχήμα 6.3.2).



Σχήμα 6.3.2: Δειγματικός χώρος του συνόλου MeHydOx με εφαρμογή PCA (n=3) στις μεταβλητές

Εξωτερική αξιολόγηση

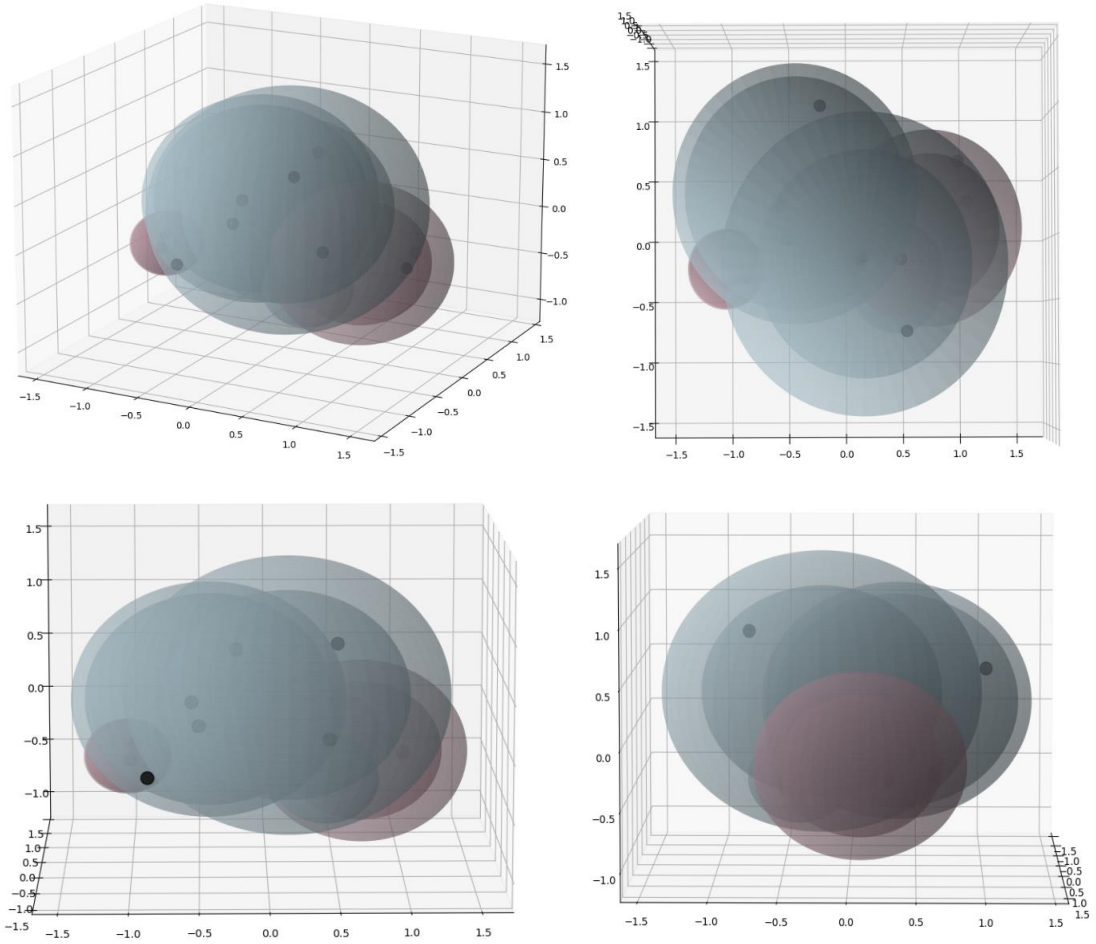
Το σύνολο δεδομένων χωρίζεται με χρήση της τυχαίας μεθόδου train_test_split της Python με κλάσμα διαμέρισης 0.7. Πραγματοποιούνται τρεις διαφορετικές προβλέψεις με την ίδια διαμέριση, βάσει των τριών εναλλακτικών μεθόδων κατηγοριοποίησης νέων δειγμάτων και παρουσιάζονται στον πίνακα 6.3.1.

Πίνακας 6.3.1: Στατιστικά αποτελέσματα με εφαρμογή τυχαίας διαμέρισης με κλάσμα 0.7 στο σύνολο MeHydOx με τις τρεις διαφορετικές μεθόδους

Μέθοδος	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.57	0.67	0.5	0.40	0.07	$\begin{bmatrix} 2 & 1 \\ 2 & 2 \end{bmatrix}$
2	1.00	1.00	1.00	0.00	1.00	$\begin{bmatrix} 3 & 0 \\ 0 & 4 \end{bmatrix}$

3	0.71	0.67	0.75	0.20	0.49	$\begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}$

Η ακρίβεια των προβλέψεων κυμαίνεται σε τιμές $[0.57, 1.00, 0.71]$ για τις τρεις μεθόδους αντίστοιχα. Η καλύτερη πρόβλεψη δίνεται με τη Μέθοδο 2, επιτυγχάνοντας την σωστή πρόβλεψη όλων των μη τοξικών δειγμάτων με απόλυτη ευαισθησία. Με τη Μέθοδο 1 το μοντέλο δεν αποδίδει καθόλου, με την ειδικότητα να κυμαίνεται σε χαμηλά επίπεδα, δυσχεραίνοντας την πρόβλεψη των τοξικών νανοσωματιδίων. Συγκριτικά με τις ίδιες δοκιμές χωρίς PCA, η Μέθοδος 1 έχει μικρότερη ακρίβεια, η Μέθοδος 2 μεγαλύτερη και η Μέθοδος 3 του ίδιου ύψους.



Γράφημα 6.3.1: Σχηματική αναπαράσταση του δειγματικού χώρου με τα σχηματιζόμενα cluster για το σύνολο MeHydOx με εφαρμογή της PCA (n=3)

Με τις ανεξάρτητες μεταβλητές συνεπτυγμένες σε τρεις νέες και για κλάσμα διαμέρισης 0.70, η εφαρμογή του μοντέλου δημιουργεί 4 συστάδες στον δειγματικό χώρο, δύο για κάθε κλάση, όπως παρουσιάζεται στο γράφημα 6.3.1.

Ο υπολογιστικός χρόνος στην περίπτωση εφαρμογής της ανάλυσης κύριων συνιστωσών, μειώνεται σημαντικά συγκριτικά με τους ίδιους υπολογισμούς όταν

συμμετέχει όλο το σύνολο δεδομένων. Η αριθμητική σύγκριση παρουσιάζεται στον πίνακα 6.3.2. Παρατηρείται πως με την εφαρμογή των ίδιων μεθόδων, ο χρόνος μειώνεται στο 1/3 περίπου.

Πίνακας 6.3.2: Σύγκριση υπολογιστικών χρόνων σε δευτερόλεπτα με τυχαία διαμέριση στο σύνολο MeHydOx, με και χωρίς εφαρμογή PCA

PCA	**Μέθοδος 1**	**Μέθοδος 2**	**Μέθοδος 3**
NAI	0.36	0.37	0.36
OXI	1.25	1.02	0.91

Εσωτερική αξιολόγηση

Ακολουθώντας την διαδικασία K-Fold, στους πίνακες 6.3.3-6.3.5, συνοψίζονται τα στατιστικά αποτελέσματα για τις τρεις δοκιμές διαμέρισης K-Fold για τις τρεις εναλλακτικές μεθοδολογίες κατηγοριοποίησης.

Πίνακας 6.3.3: Στατιστικά αποτελέσματα με εφαρμογή 3-Fold στο σύνολο MeHydOx με PCA για τις τρεις διαφορετικές μεθόδους

Μέθοδος	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.84	0.86	0.82	0.16	0.68	$\begin{bmatrix}12 & 2\\ 2 & 9\end{bmatrix}$
2	0.84	0.79	0.91	0.16	0.69	$\begin{bmatrix}11 & 3\\ 1 & 10\end{bmatrix}$
3	0.92	0.93	0.91	0.08	0.84	$\begin{bmatrix}13 & 1\\ 1 & 10\end{bmatrix}$

Στην περίπτωση της 3-Fold επικύρωσης, τα στατιστικά αποτελέσματα για τις Μεθόδους 1, 2 και 3 είναι ικανοποιητικά, χωρίς μεγάλες αποκλίσεις μεταξύ των τιμών της ακρίβειας. Τις μικρότερες διακυμάνσεις σε συνδυασμό με την καλύτερη αθροιστική απόδοση επιδεικνύει η Μέθοδος 3, ενώ από το σύνολο των τοξικών δειγμάτων νανοσωματιδίων, μόνο 1 δεν προβλέφθηκαν επιτυχώς στο σύνολο των τριών δοκιμών. Συγκριτικά με τις ίδιες δοκιμές που δεν περιλαμβάνουν ανάλυση PCA, το μοντέλο αποδίδει καλύτερα καθώς πετυχαίνει την ορθή πρόβλεψη περισσότερων τοξικών νανοσωματιδίων.

Πίνακας 6.3.4: Στατιστικά αποτελέσματα με εφαρμογή 4-Fold στο σύνολο MeHydOx με PCA για τις τρεις διαφορετικές μεθόδους

Μέθοδος	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.72	0.86	0.55	0.28	0.43	$\begin{bmatrix}12 & 2\\ 5 & 6\end{bmatrix}$

2	0.92	1.00	0.82	0.08	0.85	$\begin{bmatrix} 14 & 0 \\ 2 & 9 \end{bmatrix}$
3	0.88	1.00	0.73	0.12	0.77	$\begin{bmatrix} 14 & 0 \\ 3 & 8 \end{bmatrix}$

Στην περίπτωση της 4-Fold επικύρωσης, τα στατιστικά αποτελέσματα και για τις τρεις μεθοδολογίες είναι λιγότερο ικανοποιητικά συγκριτικά με την 3-Fold επικύρωση, κυρίως για τη Μέθοδο 1. Τις μικρότερες διακυμάνσεις σε συνδυασμό με την καλύτερη αθροιστική απόδοση επιδεικνύει η Μέθοδος 2, με ακρίβεια 0.92 αν και δεν προβλέπει επιτυχώς 2 εκ των τοξικών δειγμάτων νανοσωματιδίων. Συγκριτικά με τις ίδιες δοκιμές που δεν περιλαμβάνουν ανάλυση PCA, το μοντέλο αποδίδει καλύτερα με τη χρήση των Μεθόδων 2 και 3 καθώς πετυχαίνει την ορθή πρόβλεψη περισσότερων τοξικών νανοσωματιδίων, αλλά έχει παρόμοια επίδοση κατά τη εφαρμογή της Μεθόδου 1.

Πίνακας 6.3.5: Στατιστικά αποτελέσματα με εφαρμογή 5-Fold στο σύνολο MeHydOx με PCA για τις τρεις διαφορετικές μεθόδους

Μέθοδος	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.80	0.86	0.73	0.20	0.59	$\begin{bmatrix} 12 & 2 \\ 3 & 8 \end{bmatrix}$
2	0.88	0.93	0.73	0.12	0.68	$\begin{bmatrix} 13 & 1 \\ 3 & 8 \end{bmatrix}$
3	0.72	0.86	0.55	0.28	0.43	$\begin{bmatrix} 12 & 2 \\ 5 & 6 \end{bmatrix}$

Στην περίπτωση της 5-Fold επικύρωσης, τα στατιστικά αποτελέσματα και για τις τρεις μεθοδολογίες δεν είναι εξίσου ικανοποιητικά, με σχετικά χαμηλές τιμές ειδικότητας. Τις μικρότερες διακυμάνσεις σε συνδυασμό με την καλύτερη αθροιστική απόδοση επιδεικνύει η Μέθοδος 2, με ακρίβεια 0.88. Από το σύνολο των τοξικών δειγμάτων νανοσωματιδίων, 3 δεν προβλέφθηκαν επιτυχώς στο σύνολο των τριών δοκιμών. Όπως και στην περίπτωση της 4-Fold επικύρωσης, συγκριτικά με τις ίδιες δοκιμές που δεν περιλαμβάνουν ανάλυση PCA, το μοντέλο αποδίδει καλύτερα με τη χρήση των Μεθόδων 2 και 3 καθώς πετυχαίνει την ορθή πρόβλεψη περισσότερων τοξικών νανοσωματιδίων.

Ως προς το υπολογιστικό κόστος των τριών εναλλακτικών μεθόδων, τα αποτελέσματα συνοψίζονται στον πίνακα 6.3.6.

Πίνακας 6.3.6: Υπολογιστικοί χρόνοι σε δευτερόλεπτα για την εφαρμογή του μοντέλου μέσω K-Fold και PCA στο σύνολο MeHydOx

	Μέθοδος 1	**Μέθοδος 2**	**Μέθοδος 3**
3-Fold	0.89	0.87	0.87
4-Fold	1.29	1.28	1.26
5-Fold	1.85	1.82	1.82

Αν και υπάρχουν διακυμάνσεις ως προς την ταχύτερη μέθοδο στις τρεις δοκιμές, κατά μέσο όρο, η Μέθοδος 3 αποδεικνύεται γρηγορότερη. Σε σχέση με τις αντίστοιχες τιμές όταν δεν εφαρμοστεί PCA, όπως παρουσιάζεται στον πίνακα 7.1.8, οι υπολογισμοί και οι προβλέψεις είναι σημαντικά ταχύτεροι καθώς μειώνεται το υπολογιστικό κόστος CPU στο 1/3 της αρχικής τιμής, χωρίς να θυσιάζεται σημαντικά η ακρίβεια του προβλεπτικού μοντέλου.

Τα αποτελέσματα Leave-One-Out για τις τρείς μεθοδολογίες συνοψίζονται στον πίνακα 6.3.7.

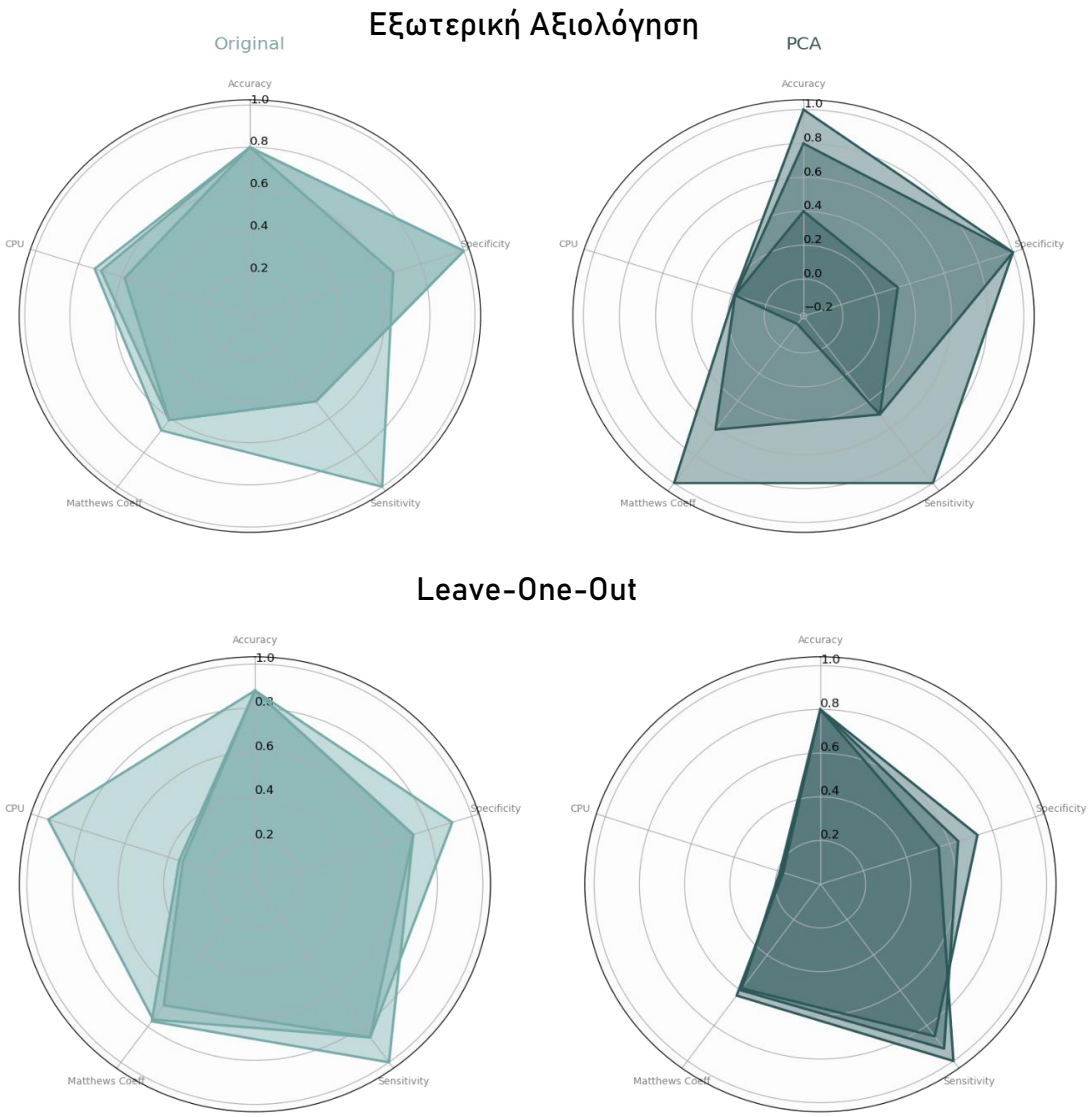
Πίνακας 6.3.7: Στατιστικά αποτελέσματα με εφαρμογή Leave-One-Out και PCA στο σύνολο MeHydOx για τις τρεις διαφορετικές μεθόδους

Μέθοδος	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.80	1.00	0.55	0.20	0.63	$\begin{bmatrix} 14 & 0 \\ 5 & 6 \end{bmatrix}$
2	0.80	0.86	0.73	0.20	0.59	$\begin{bmatrix} 12 & 2 \\ 3 & 8 \end{bmatrix}$
3	0.80	0.93	0.64	0.20	0.60	$\begin{bmatrix} 13 & 1 \\ 4 & 7 \end{bmatrix}$

Κατά τις δοκιμές Leave-One-Out, επιβεβαιώνεται η επιτυχία του μοντέλου με στατιστικά ακρίβειας $[0.8, 0.8, 0.8]$ για τις μεθόδους 1, 2 και 3, αντίστοιχα. Ωστόσο, , συγκριτικά με τις ίδιες δοκιμές που δεν περιλαμβάνουν ανάλυση PCA, η ειδικότητα επιτυγχάνει χαμηλότερες τιμές καθώς πολλά τοξικά δείγματα δεν εντοπίζονται από τη μέθοδο κατηγοριοποίησης.

Στο γράφημα 6.3.2, καθίσταται εμφανής η μείωση του υπολογιστικού χρόνου με την εφαρμογή της PCA. Στην περίπτωση της εξωτερικής επικύρωσης, με την εφαρμογή PCA, βελτιώνονται σημαντικά και οι στατιστικές επιδόσεις πρόβλεψης, κυρίως κατά την εφαρμογή της Μεθόδου 2 η οποία αγγίζει τη μονάδα σε ακρίβεια, ειδικότητα και ευαισθησία. Αντίθετα, στην περίπτωση της εσωτερικής επικύρωσης Leave-One-Out, η μείωση του υπολογιστικού χρόνου δεν επιφέρει και την βελτίωση των στατιστικών επιδόσεων του μοντέλου και παρατηρείται μείωση και για τις τρεις μεθόδους στην ακρίβεια και την ειδικότητα.

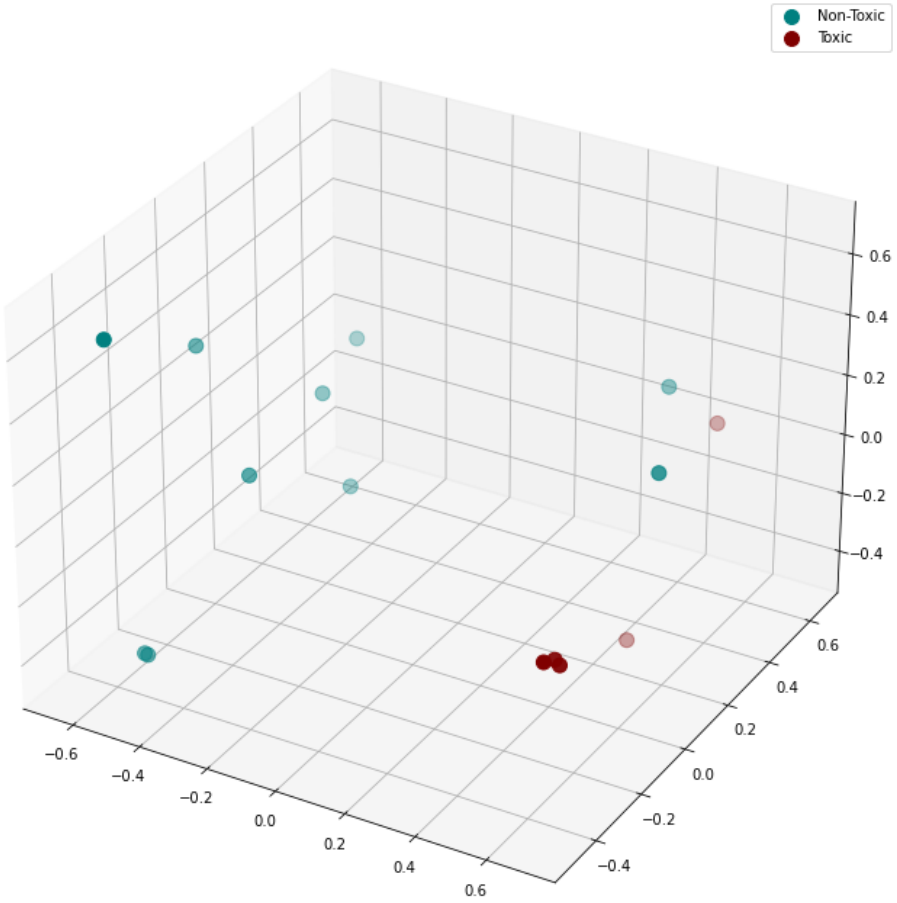
Και στην περίπτωση εφαρμογής PCA η εξωτερική αξιολόγηση του μοντέλου με Kennard Stone επιδεικνύει πολύ καλά αποτελέσματα, με ταυτόχρονη μείωση του υπολογιστικού χρόνου, για αυτό και κρίνεται ως η βέλτιστη επιλογή για επικύρωση του μοντέλου.



Γράφημα 6.3.2: Συγκριτικό διάγραμμα των στατιστικών για το μοντέλο MeHydOx στην περίπτωση εφαρμογής και μη PCA με εξωτερική αξιολόγηση και Leave-One-Out

6.3.2. Σύνολο δεδομένων SPIONs

Το σύνολο δεδομένων SPIONs μετά από επεξεργασία PCA και μείωση των διαστάσεων των μεταβλητών του σε 3, διαμορφώνει ένα νέο δειγματικό χώρο.



Σχήμα 6.3.3: Δειγματικός χώρος του συνόλου SPIONs με εφαρμογή PCA (n=3) στις μεταβλητές

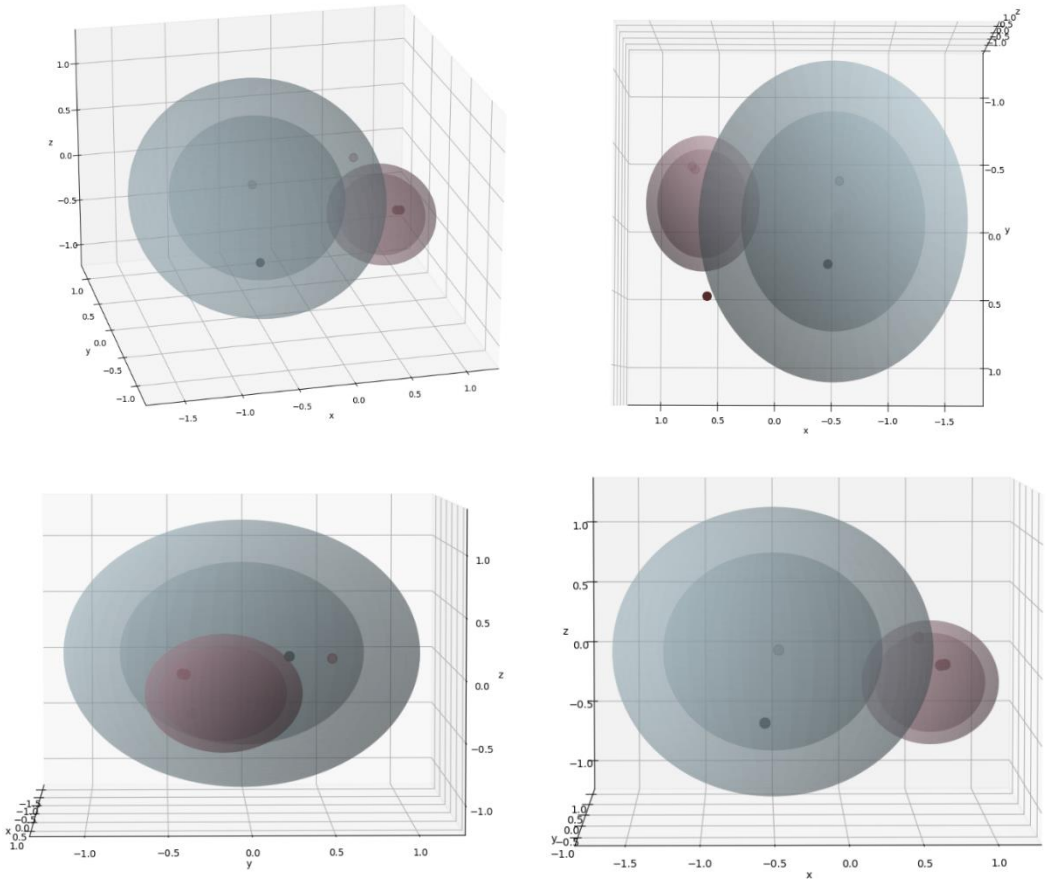
Εξωτερική αξιολόγηση

Το σύνολο δεδομένων χωρίζεται με χρήση της μεθόδου train_test_split της Python με κλάσμα διαμέρισης 0.70. Πραγματοποιούνται τρεις διαφορετικές προβλέψεις με την ίδια τυχαία διαμέριση, βάσει των τριών εναλλακτικών μεθόδων κατηγοριοποίησης νέων δειγμάτων και παρουσιάζονται στον πίνακα 6.3.8.

Πίνακας 6.3.8: Στατιστικά αποτελέσματα με εφαρμογή τυχαίας διαμέρισης με κλάσμα 0.7 στο σύνολο SPIONs με τις τρεις διαφορετικές μεθόδους

Μέθοδος	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.80	1.00	0.67	0.20	0.67	$\begin{bmatrix} 2 & 0 \\ 1 & 2 \end{bmatrix}$
2	0.80	1.00	0.67	0.20	0.67	$\begin{bmatrix} 2 & 0 \\ 1 & 2 \end{bmatrix}$
3	0.80	1.00	0.67	0.20	0.67	$\begin{bmatrix} 2 & 0 \\ 1 & 2 \end{bmatrix}$

Η ακρίβεια των προβλέψεων κυμαίνεται σε τιμές [0.80, 0.80 , 0.80] για τις τρεις μεθόδους αντίστοιχα. Και οι τρεις μέθοδοι χαρακτηρίζονται από τα ίδια ποσοστά επιτυχίας, με μόνο μια εσφαλμένη πρόβλεψη τοξικού δείγματος νανοσωματιδίου. Στην περίπτωση μη εφαρμογής της PCA, με Kennard Stone 0.70, το αντίστοιχο ποσοστό επιτυχίας ήταν 100%, υποδηλώνοντας μια σχετική μείωση στην επίδοση του μοντέλου.



Σχήμα 6.3.6: Σχηματική αναπαράσταση του δειγματικού χώρου με τα σχηματιζόμενα cluster για το σύνολο SPIONs με εφαρμογή της PCA (n=3)

Με τις ανεξάρτητες μεταβλητές συνεπτυγμένες σε τρεις νέες και για κλάσμα διαμέρισης 0.70, η εφαρμογή του μοντέλου δημιουργεί 2 cluster στον δειγματικό χώρο, ένα για κάθε κλάση, όπως παρουσιάζεται στο σχήμα 6.3.6.

Ο υπολογιστικός χρόνος στην περίπτωση εφαρμογής της ανάλυσης κύριων συνιστωσών, μειώνεται σημαντικά συγκριτικά με τους ίδιους υπολογισμούς όταν συμμετέχει όλο το σύνολο δεδομένων. Η αριθμητική σύγκριση παρουσιάζεται στον πίνακα. Παρατηρείται πως με την εφαρμογή των ίδιων μεθόδων, ο χρόνος μειώνεται στο 1/4 περίπου.

Πίνακας 6.3.9: Σύγκριση υπολογιστικών χρόνων με τυχαία διαμέριση στο σύνολο SPIONs, με και χωρίς εφαρμογή PCA

PCA	Μέθοδος 1	Μέθοδος 2	Μέθοδος 3
NAI	0.18	0.17	0.17
OXI	0.70	0.89	0.88

Εσωτερική αξιολόγηση

Ακολουθώντας την διαδικασία K-Fold, στους πίνακες 6.3.10-6.3.12, συνοψίζονται τα στατιστικά αποτελέσματα για τις τρεις δοκιμές διαμέρισης K-Fold για τις τρεις εναλλακτικές μεθοδολογίες κατηγοριοποίησης.

Πίνακας 6.3.10: Στατιστικά αποτελέσματα με εφαρμογή 3-Fold στο σύνολο SPIONs με PCA για τις τρεις διαφορετικές μεθόδους

Μέθοδος	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.87	0.80	1.00	0.13	0.76	$\begin{bmatrix} 8 & 2 \\ 0 & 5 \end{bmatrix}$
2	0.93	0.90	1.00	0.07	0.87	$\begin{bmatrix} 9 & 1 \\ 0 & 5 \end{bmatrix}$
3	0.87	0.80	1.00	0.07	0.76	$\begin{bmatrix} 8 & 2 \\ 0 & 5 \end{bmatrix}$

Στην περίπτωση της 3-Fold επικύρωσης, τα στατιστικά αποτελέσματα για τις Μεθόδους 1, 2 και 3 είναι ικανοποιητικά, χωρίς μεγάλες αποκλίσεις μεταξύ των τιμών της ακρίβειας. Τις μικρότερες διακυμάνσεις σε συνδυασμό με την καλύτερη αθροιστική απόδοση επιδεικνύει η Μέθοδος 2, ενώ από το σύνολο των τοξικών δειγμάτων νανοσωματιδίων, όλα προβλέφθηκαν επιτυχώς στο σύνολο των τριών δοκιμών. Συγκριτικά με τις ίδιες δοκιμές που δεν περιλαμβάνουν ανάλυση PCA, το μοντέλο αποδίδει εξίσου καλά. Ακολουθούν παρόμοιοι πίνακες για τις δοκιμές 4-Fold και 5-Fold.

Πίνακας 6.3.11: Στατιστικά αποτελέσματα με εφαρμογή 4-Fold στο σύνολο SPIONs με PCA για τις τρεις διαφορετικές μεθόδους

Μέθοδος	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.87	0.90	0.80	0.13	0.70	$\begin{bmatrix} 9 & 1 \\ 1 & 4 \end{bmatrix}$
2	0.80	0.90	0.60	0.20	0.53	$\begin{bmatrix} 9 & 1 \\ 2 & 3 \end{bmatrix}$
3	0.87	0.90	0.80	0.13	0.70	$\begin{bmatrix} 9 & 1 \\ 1 & 4 \end{bmatrix}$

Στην περίπτωση της 4-Fold επικύρωσης, τα στατιστικά αποτελέσματα και για τις τρεις μεθοδολογίες είναι λιγότερο ικανοποιητικά συγκριτικά με την 3-Fold επικύρωση, καθώς και σε σύγκριση με τις ίδιες δοκιμές που δεν περιλαμβάνουν PCA. Τις μικρότερες διακυμάνσεις σε συνδυασμό με την καλύτερη αθροιστική απόδοση επιδεικνύουν εξίσου οι Μέθοδοι 1 και 3, με ακρίβεια 0.87 αν και δεν προβλέπουν επιτυχώς 1 εκ των τοξικών δειγμάτων νανοσωματιδίων. Στις αντίστοιχες δοκιμές χωρίς PCA οι ίδιες μέθοδοι προέβλεπαν σωστά όλα τα τοξικά δείγματα.

Πίνακας: 6.3.12: Στατιστικά αποτελέσματα με εφαρμογή 5-Fold στο σύνολο SPIONs με PCA για τις τρεις διαφορετικές μεθόδους

Μέθοδος	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.80	0.80	0.80	0.80	0.58	$\begin{bmatrix} 8 & 2 \\ 1 & 4 \end{bmatrix}$
2	0.87	0.90	0.80	0.13	0.70	$\begin{bmatrix} 9 & 1 \\ 1 & 4 \end{bmatrix}$
3	0.87	0.80	1.00	0.13	0.76	$\begin{bmatrix} 8 & 2 \\ 0 & 5 \end{bmatrix}$

Στην περίπτωση της 5-Fold επικύρωσης, παρατηρείται παρόμοιο μοτίβο συμπεριφορών και οι τιμές κυμαίνονται στα ίδια επίπεδα. Τις μικρότερες διακυμάνσεις σε συνδυασμό με την καλύτερη αθροιστική απόδοση επιδεικνύει η Μέθοδος 3, με ακρίβεια 0.87. Από το σύνολο των τοξικών δειγμάτων νανοσωματιδίων στο σύνολο των τριών δοκιμών, όλα προβλέφθηκαν επιτυχώς. Συγκριτικά με τις ίδιες δοκιμές που δεν περιλαμβάνουν ανάλυση PCA, το μοντέλο αποδίδει εξίσου με τη χρήση των Μεθόδων 2 και 3 και καλύτερα με τη χρήση της Μεθόδου 1 καθώς πετυχαίνει την ορθή πρόβλεψη περισσότερων τοξικών νανοσωματιδίων.

Ως προς το υπολογιστικό κόστος των τριών εναλλακτικών μεθόδων, τα αποτελέσματα συνοψίζονται στον πίνακα 6.3.13. Αν και υπάρχουν διακυμάνσεις ως προς την ταχύτερη μέθοδο στις τρεις δοκιμές, κατά μέσο όρο, η Μέθοδος 2 αποδεικνύεται γρηγορότερη. Σε σχέση με τις αντίστοιχες τιμές όταν δεν εφαρμοστεί PCA, και ενάντια στις έως τώρα παρατηρήσεις, οι υπολογισμοί καθυστερούν σημαντικά, ιδιαίτερα στην επικύρωση 5-Fold.

Πίνακας 6.3.13: Υπολογιστικοί χρόνοι σε δευτερόλεπτα για την εφαρμογή του μοντέλου μέσω K-Fold και PCA στο σύνολο SPIONs

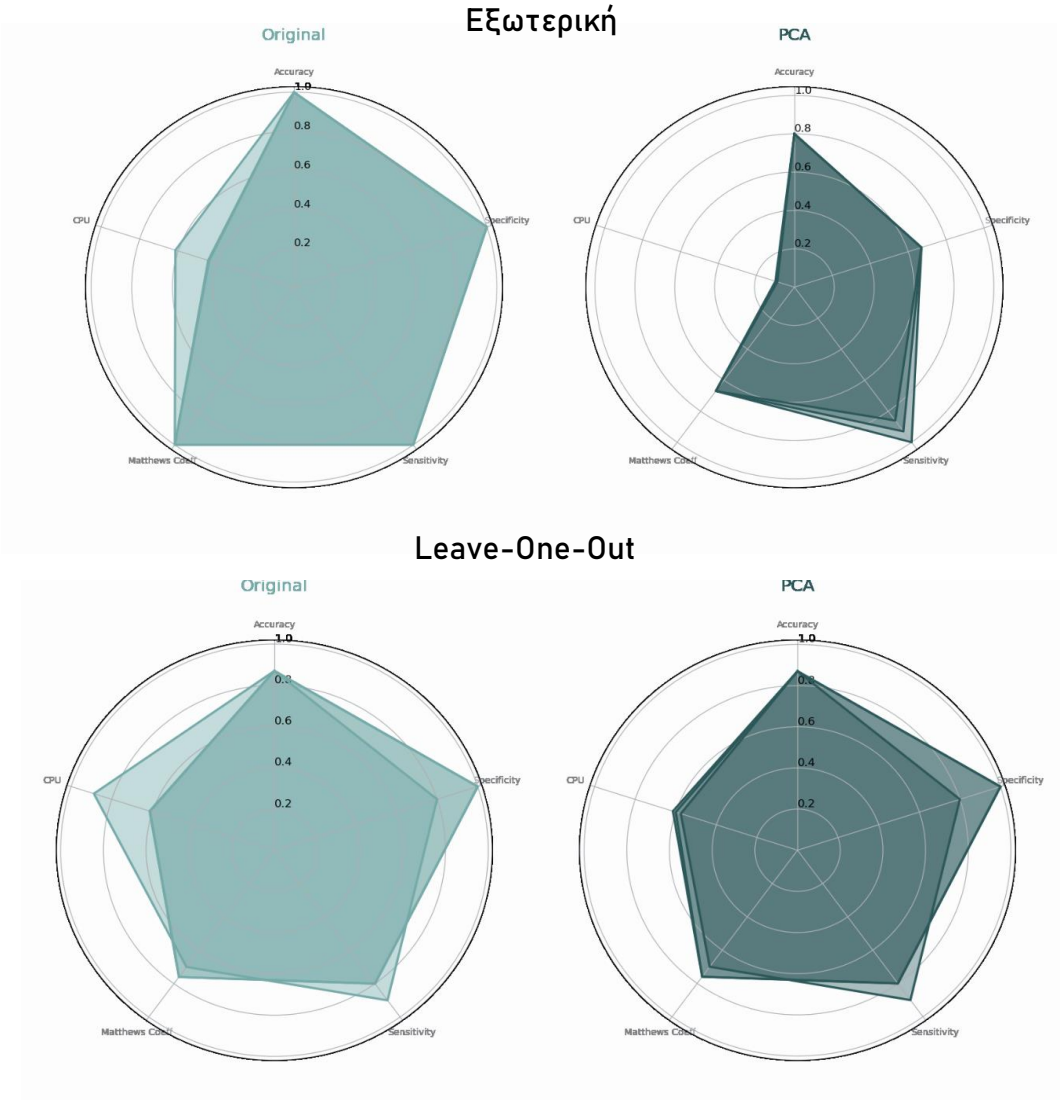
	Μέθοδος 1	**Μέθοδος 2**	**Μέθοδος 3**
3-Fold	0.55	0.52	0.725
4-Fold	0.92	0.91	1.00
5-Fold	1.25	1.23	1.29

Τα αποτελέσματα Leave-One-Out για τις τρείς μεθοδολογίες συνοψίζονται στον πίνακα 6.3.14.

Πίνακας 6.3.14: Στατιστικά αποτελέσματα με εφαρμογή Leave-One-Out και PCA στο σύνολο SPIONs για τις τρεις διαφορετικές μεθόδους

Μέθοδος	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.87	0.80	1.00	0.13	0.76	$\begin{bmatrix} 8 & 2 \\ 0 & 5 \end{bmatrix}$
2	0.87	0.90	0.80	0.14	0.70	$\begin{bmatrix} 9 & 1 \\ 1 & 4 \end{bmatrix}$
3	0.87	0.80	1.00	0.13	0.76	$\begin{bmatrix} 8 & 2 \\ 0 & 5 \end{bmatrix}$

Κατά τις δοκιμές Leave-One-Out, επιβεβαιώνεται η επιτυχία του μοντέλου με στατιστικά ακρίβειας [0.87, 0.89, 0.87] για τις μεθόδους 1, 2 και 3, αντίστοιχα.



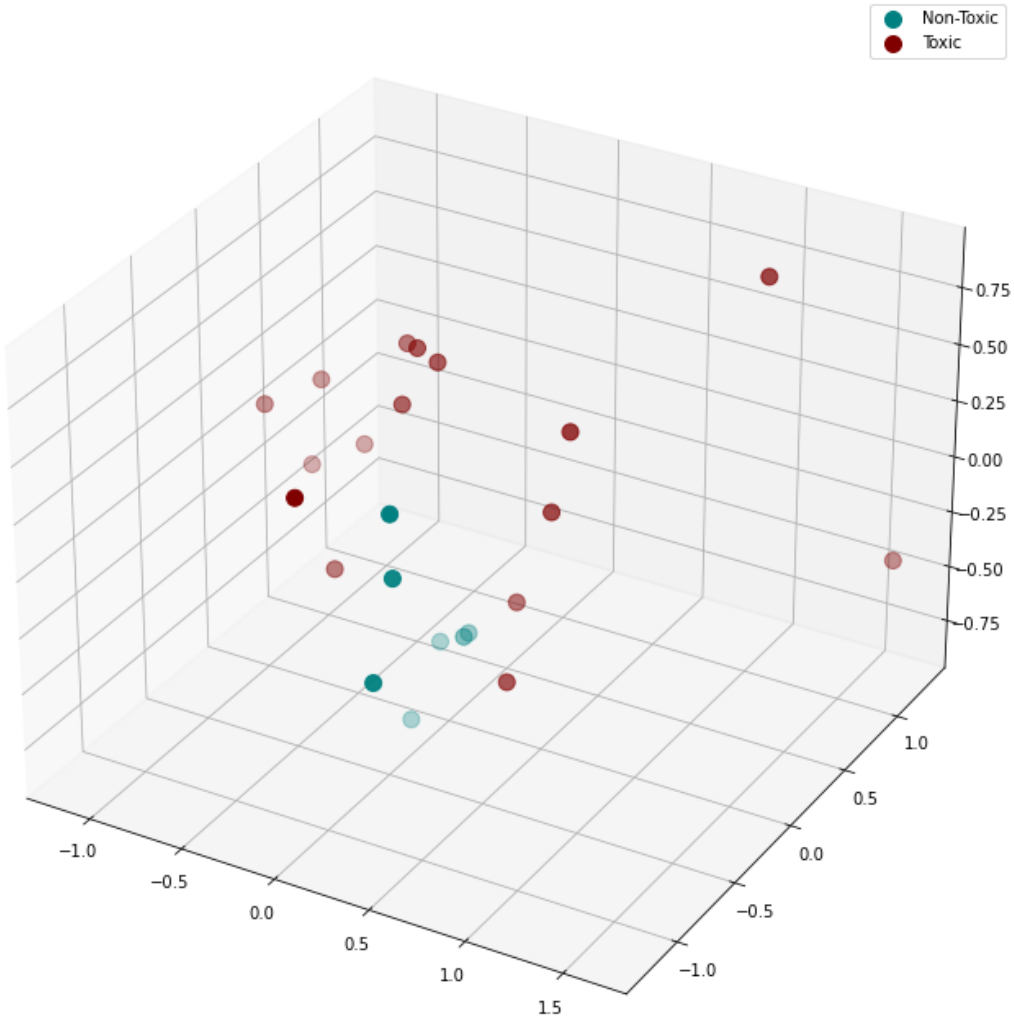
Γράφημα 6.3.3: Συγκριτικό διάγραμμα των στατιστικών για το μοντέλο SPIONs στην περίπτωση εφαρμογής και μη PCA με εξωτερική αξιολόγηση και Leave-One-Out

Συγκριτικά με τις ίδιες δοκιμές που δεν περιλαμβάνουν ανάλυση PCA, η ειδικότητα επιτυγχάνει παρόμοιες τιμές και μόνο 1 τοξικό δείγμα δεν εντοπίζεται από τη Μέθοδο 2.

Όπως φαίνεται και στο γράφημα 6.3.3, η μείωση του υπολογιστικού χρόνου με την εφαρμογή PCA είναι κοινή και στις δύο διαδικασίες επικύρωσης. Ωστόσο, σε αντίθεση με την περίπτωση του συνόλου MeHydOx, η απόδοση της εξωτερικής αξιολόγησης μειώνεται, τόσο ως προς την ακρίβεια, όσο και ως προς την ειδικότητα, ενώ οι στατιστικές επιδόσεις κατά την εφαρμογή Leave-One-Out παραμένουν ίδιες, παρά τη μείωση του υπολογιστικού χρόνου. Έτσι, για το σύνολο δεδομένων SPIONs, δεν ενδείκνυται η εφαρμογή PCA καθώς εντοπίζεται μείωση στις τιμές βασικών δεικτών αξιολόγησης του μοντέλου και ο αριθμός των descriptors είναι ήδη επαρκώς μικρός.

6.3.3. Σύνολο δεδομένων MeOx

Το σύνολο δεδομένων SPIONs μετά από επεξεργασία PCA και μείωση των διαστάσεων των μεταβλητών του σε 3, διαμορφώνει ένα νέο δειγματικό χώρο.



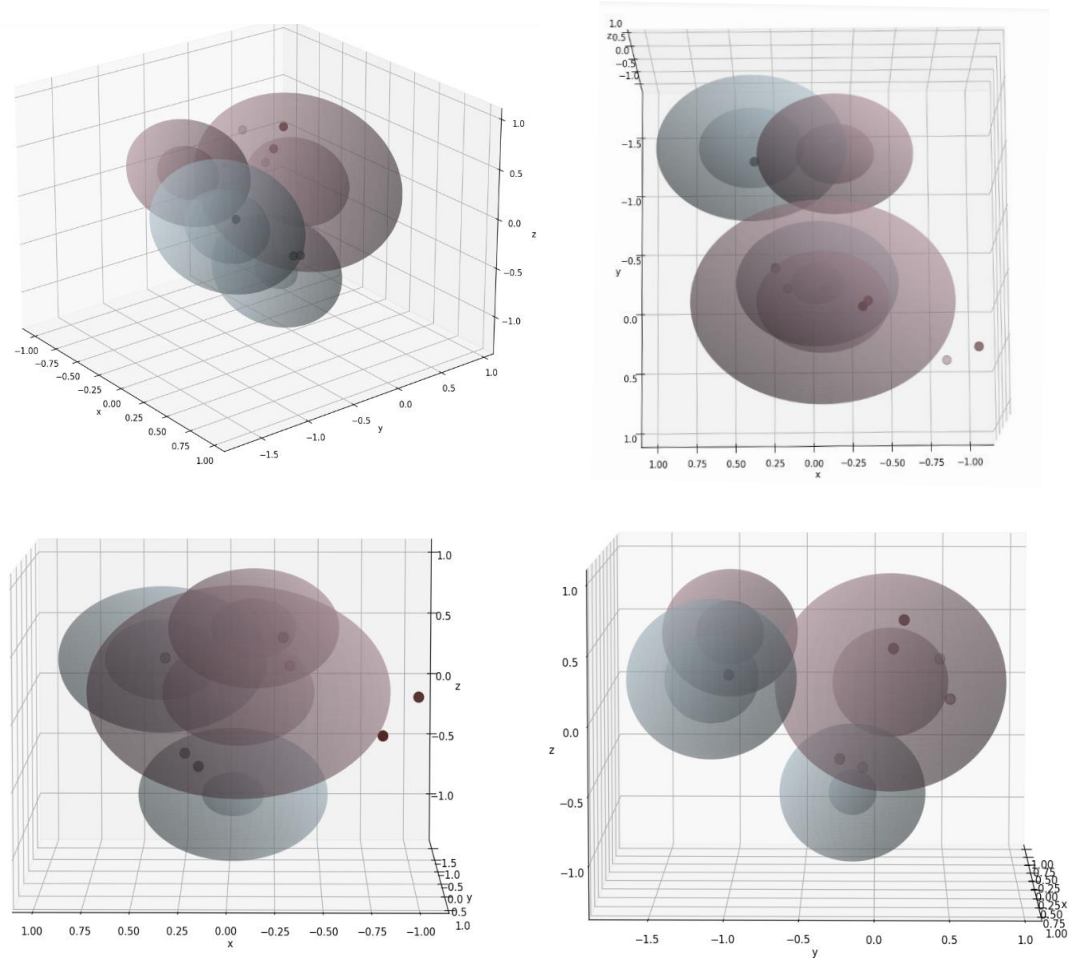
Σχήμα 6.3.7: Δειγματικός χώρος του συνόλου MeOx με εφαρμογή PCA (n=3) στις μεταβλητές

Το σύνολο δεδομένων χωρίζεται με χρήση της μεθόδου train_test_split της Python με κλάσμα διαμέρισης 0.70. Πραγματοποιούνται τρεις διαφορετικές προβλέψεις με την ίδια τυχαία διαμέριση, βάσει των τριών εναλλακτικών μεθόδων κατηγοριοποίησης νέων δειγμάτων και παρουσιάζονται στον πίνακα 6.3.15.

Πίνακας 6.3.15: Στατιστικά αποτελέσματα με εφαρμογή τυχαίας διαμέρισης με κλάσμα 0.7 στο σύνολο MeOx με τις τρεις διαφορετικές μεθόδους

Μέθοδος	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.80	0.00	1.00	0.20	0.00	$\begin{bmatrix} 0 & 1 \\ 0 & 4 \end{bmatrix}$
2	1.00	1.00	1.00	0.00	1.00	$\begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}$
3	1.00	1.00	1.00	0.00	1.00	$\begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}$

Η ακρίβεια των προβλέψεων κυμαίνεται σε τιμές [0.80, 1.00 , 1.00] για τις τρεις μεθόδους αντίστοιχα. Και οι τρεις μέθοδοι χαρακτηρίζονται από παρόμοια ποσοστά επιτυχίας, χωρίς καμία εσφαλμένη πρόβλεψη τοξικού δείγματος



Σχήμα 6.3.8: Σχηματική αναπαράσταση του δειγματικού χώρου με τα σχηματιζόμενα cluster για το σύνολο MeOx με εφαρμογή της PCA (n=3)

νανοσωματιδίου, ομοίως με την περίπτωση μη εφαρμογής της PCA, με Kennard Stone 0.70. Με τις ανεξάρτητες μεταβλητές συνεπτυγμένες σε τρεις νέες και για κλάσμα διαμέρισης 0.70, η εφαρμογή του μοντέλου δημιουργεί 4 cluster στον δειγματικό χώρο, 2 για κάθε κλάση, όπως παρουσιάζεται στο σχήμα 6.3.8.

Ο υπολογιστικός χρόνος στην περίπτωση εφαρμογής της ανάλυσης κύριων συνιστωσών, μειώνεται σημαντικά συγκριτικά με τους ίδιους υπολογισμούς όταν συμμετέχει όλο το σύνολο δεδομένων. Η αριθμητική σύγκριση παρουσιάζεται στον πίνακα. Παρατηρείται πως με την εφαρμογή των ίδιων μεθόδων, ο χρόνος μειώνεται στο 1/3 περίπου.

Πίνακας 6.3.16: Σύγκριση υπολογιστικών χρόνων με τυχαία διαμέριση στο σύνολο MeOx , με και χωρίς εφαρμογή PCA

PCA	**Μέθοδος 1**	**Μέθοδος 2**	**Μέθοδος 3**
NAI	0.38	0.36	0.39
OXI	0.83	1.28	1.36

Εσωτερική αξιολόγηση

Ακολουθώντας την διαδικασία K-Fold, στους πίνακες 6.3.17-6.3.19, συνοψίζονται τα στατιστικά αποτελέσματα για τις τρεις δοκιμές διαμέρισης K-Fold για τις τρεις εναλλακτικές μεθοδολογίες κατηγοριοποίησης.

Πίνακας 6.3.17: Στατιστικά αποτελέσματα με εφαρμογή 3-Fold στο σύνολο MeOx με PCA για τις τρεις διαφορετικές μεθόδους

Μέθοδος	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.70	0.29	0.88	0.30	0.20	$\begin{bmatrix} 2 & 5 \\ 2 & 14 \end{bmatrix}$
2	0.87	0.86	0.88	0.13	0.71	$\begin{bmatrix} 6 & 1 \\ 2 & 14 \end{bmatrix}$
3	0.78	1.00	0.67	0.22	0.82	$\begin{bmatrix} 7 & 0 \\ 5 & 11 \end{bmatrix}$

Στην περίπτωση της 3-Fold επικύρωσης, τα στατιστικά αποτελέσματα για τις Μεθόδους 1, 2 και 3 δεν είναι ιδιαίτερα ικανοποιητικά, και εντοπίζονται αποκλίσεις μεταξύ των τιμών της ακρίβειας, κυρίως κατά την εφαρμογή της Μεθόδου 1. Τις μικρότερες διακυμάνσεις σε συνδυασμό με την καλύτερη αθροιστική απόδοση επιδεικνύει η Μέθοδος 2, ενώ από το σύνολο των τοξικών δειγμάτων νανοσωματιδίων, 2 δεν προβλέφθηκαν επιτυχώς στο σύνολο των τριών δοκιμών. Συγκριτικά με τις ίδιες δοκιμές που δεν περιλαμβάνουν ανάλυση PCA, το μοντέλο αποδίδει χειρότερα καθώς πετυχαίνει την ορθή πρόβλεψη λιγότερων τοξικών νανοσωματιδίων.

Πίνακας 6.3.18: Στατιστικά αποτελέσματα με εφαρμογή 4-Fold στο σύνολο MeOx με PCA για τις τρεις διαφορετικές μεθόδους

Μέθοδος	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.74	0.29	0.94	0.26	0.30	$\begin{bmatrix} 2 & 5 \\ 1 & 15 \end{bmatrix}$
2	0.70	0.71	0.69	0.26	0.42	$\begin{bmatrix} 5 & 2 \\ 5 & 11 \end{bmatrix}$
3	0.83	0.71	0.88	0.17	0.59	$\begin{bmatrix} 5 & 2 \\ 2 & 14 \end{bmatrix}$

Στην περίπτωση της 4-Fold επικύρωσης, τα στατιστικά αποτελέσματα και για τις τρεις μεθοδολογίες είναι λιγότερο ικανοποιητικά συγκριτικά με την 3-Fold επικύρωση. Συγκριτικά με τις ίδιες δοκιμές που δεν περιλαμβάνουν ανάλυση PCA, το μοντέλο αποδίδει εξίσου κατά την εφαρμογή των Μεθόδων 2 και 3 αλλά χειρότερα κατά την εφαρμογή της Μεθόδου 1, καθώς πετυχαίνει την ορθή πρόβλεψη λιγότερων τοξικών νανοσωματιδίων.

Πίνακας 6.3.19: Στατιστικά αποτελέσματα με εφαρμογή 5-Fold στο σύνολο MeOx με PCA για τις τρεις διαφορετικές μεθόδους

Μέθοδος	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.70	0.14	0.94	0.30	0.13	$\begin{bmatrix} 1 & 6 \\ 1 & 15 \end{bmatrix}$
2	0.78	1.00	0.69	0.22	0.63	$\begin{bmatrix} 7 & 0 \\ 5 & 11 \end{bmatrix}$
3	0.74	0.14	0.69	0.26	0.50	$\begin{bmatrix} 6 & 1 \\ 5 & 11 \end{bmatrix}$

Στην περίπτωση της 5-Fold επικύρωσης, τα στατιστικά αποτελέσματα και για τις τρεις μεθοδολογίες δεν είναι εξίσου ικανοποιητικά, με σχετικά χαμηλές τιμές ειδικότητας. Τις μικρότερες διακυμάνσεις σε συνδυασμό με την καλύτερη αθροιστική απόδοση επιδεικνύει η Μέθοδος 2, με ακρίβεια 0.78. Από το σύνολο των τοξικών δειγμάτων νανοσωματιδίων, 5 δεν προβλέφθηκαν επιτυχώς στο σύνολο των τριών δοκιμών. Όπως και στην περίπτωση της 4-Fold επικύρωσης, συγκριτικά με τις ίδιες δοκιμές που δεν περιλαμβάνουν ανάλυση PCA, το μοντέλο αποδίδει εξίσου κατά την εφαρμογή των Μεθόδων 1,2 και 3 .

Ως προς το υπολογιστικό κόστος των τριών εναλλακτικών μεθόδων, τα αποτελέσματα συνοψίζονται στον πίνακα 6.3.20.

Πίνακας 6.3.20: Υπολογιστικοί χρόνοι σε δευτερόλεπτα για την εφαρμογή του μοντέλου μέσω K-Fold και PCA στο σύνολο MeOx

	Μέθοδος 1	**Μέθοδος 2**	**Μέθοδος 3**
3-Fold	1.18	1.14	1.08

4-Fold	1.36	1.30	1.29
5-Fold	2.22	2.49	2.36

Αν και υπάρχουν διακυμάνσεις ως προς την ταχύτερη μέθοδο στις τρεις δοκιμές, κατά μέσο όρο, η Μέθοδος 3 αποδεικνύεται γρηγορότερη. Σε σχέση με τις αντίστοιχες τιμές όταν δεν εφαρμοστεί PCA, οι υπολογισμοί και οι προβλέψεις επιταχύνονται σημαντικά μειώνοντας κατά το ήμισυ τον υπολογιστικό χρόνο CPU.

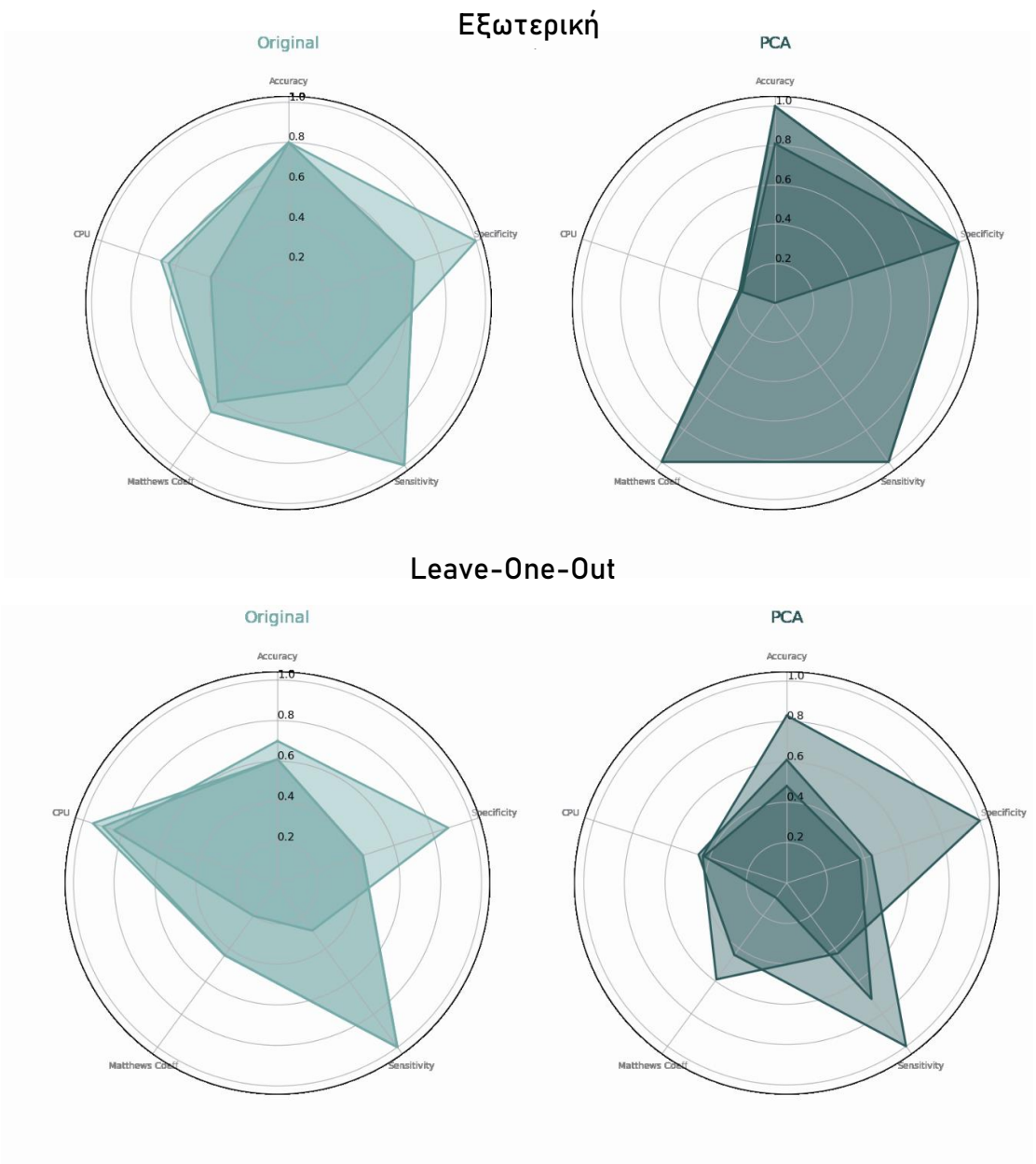
Τα αποτελέσματα Leave-One-Out για τις τρείς μεθοδολογίες συνοψίζονται στον πίνακα 6.3.21.

Πίνακας 6.3.21: Στατιστικά αποτελέσματα με εφαρμογή Leave-One-Out και PCA στο σύνολο MeOx για τις τρεις διαφορετικές μεθόδους

Μέθοδος	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Confusion Matrix***
1	0.83	0.43	1.00	0.17	0.59	$\begin{bmatrix} 3 & 4 \\ 0 & 16 \end{bmatrix}$
2	0.61	1.00	0.44	0.39	0.44	$\begin{bmatrix} 7 & 0 \\ 9 & 7 \end{bmatrix}$
3	0.48	0.71	0.38	0.52	0.09	$\begin{bmatrix} 5 & 2 \\ 10 & 6 \end{bmatrix}$

Κατά τις δοκιμές Leave-One-Out, δοκιμάζεται η επιτυχία του μοντέλου με στατιστικά ακρίβειας [0.83, 0.61, 0.48] για τις μεθόδους 1, 2 και 3, αντίστοιχα. Αν και η Μέθοδοι 2 και 3 αποτυγχάνουν στην ορθή πρόβλεψη, η Μέθοδος 1 αποδεικνύεται ιδιαίτερα επιτυχής καθώς εντοπίζει όλα τα τοξικά δείγματα. Ωστόσο, , συγκριτικά με τις ίδιες δοκιμές που δεν περιλαμβάνουν ανάλυση PCA, η Μέθοδος 1 έχει καλύτερη απόδοση, η Μέθοδος 2 την ίδια, ενώ η Μέθοδος 3 έχει χειρότερη προβλεπτική ικανότητα.

Όπως φαίνεται και στο γράφημα 6.3.4, ο υπολογιστικός χρόνος μειώνεται σημαντικά και στις δύο μεθόδους επικύρωσης. Αν και οι στατιστικές επιδόσεις στο σύνολό τους κατά μέσο όρο δεν βελτιώνονται με ή χωρίς PCA, η Μέθοδος 2 επιδεικνύει πολύ μεγάλη βελτίωση στην περίπτωση της PCA με ακρίβεια και ειδικότητα να αγγίζουν τη μονάδα.



Γράφημα 6.3.4: Συγκριτικό διάγραμμα των στατιστικών για το μοντέλο MeOx στην περίπτωση εφαρμογής και μη PCA με εξωτερική αξιολόγηση και Leave-One-Out

6.3.4. Σύνολο δεδομένων Cytotox

Μελετώντας την συμπεριφορά του συνόλου Cytotox με εφαρμογή PCA , επιλέγεται η εκπαίδευση του μοντέλου να γίνει βάσει του προτεινόμενου training set που αποτελείται από 345 δείγματα, 37 εκ των οποίων είναι τοξικά. Ως προς την PCA, εξετάζονται διαφορετικοί αριθμοί των συμπτυγμένων μεταβλητών οι οποίοι κυμαίνονται μεταξύ n= 4 και n=20 (εκ των 65 αρχικών). Τα στατιστικά αποτελέσματα των δοκιμών συνοψίζονται στον πίνακα 6.3.22.

Πίνακας 6.3.22: Στατιστικά αποτελέσματα με την προτεινόμενη διαμέριση και εφαρμογή PCA στο σύνολο Cytotox

	Μέθοδος	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Conf. Matrix***
n=4	1	0.93	0.96	0.64	0.07	0.61	$\begin{bmatrix}130 & 5\\ 5 & 9\end{bmatrix}$
	2	0.70	0.70	0.71	0.30	0.25	$\begin{bmatrix}94 & 41\\ 4 & 10\end{bmatrix}$
	3	0.70	0.72	0.50	0.30	0.14	$\begin{bmatrix}97 & 38\\ 7 & 7\end{bmatrix}$
n=6	1	0.91	0.99	0.07	0.09	0.16	$\begin{bmatrix}134 & 1\\ 13 & 1\end{bmatrix}$
	2	0.86	0.87	0.71	0.14	0.45	$\begin{bmatrix}118 & 17\\ 4 & 10\end{bmatrix}$
	3	0.57	0.54	0.86	0.43	0.23	$\begin{bmatrix}73 & 62\\ 2 & 12\end{bmatrix}$
n=8	1	0.91	1.00	0.07	0.09	0.26	$\begin{bmatrix}135 & 0\\ 13 & 1\end{bmatrix}$
	2	0.94	0.93	1.00	0.06	0.75	$\begin{bmatrix}126 & 9\\ 0 & 14\end{bmatrix}$
	3	0.85	0.86	0.71	0.15	0.42	$\begin{bmatrix}116 & 19\\ 4 & 10\end{bmatrix}$
n=10	1	0.89	0.97	0.07	0.11	0.07	$\begin{bmatrix}131 & 4\\ 13 & 1\end{bmatrix}$
	2	0.81	0.80	0.93	0.19	0.48	$\begin{bmatrix}108 & 27\\ 1 & 13\end{bmatrix}$
	3	0.83	0.84	0.71	0.17	0.40	$\begin{bmatrix}114 & 21\\ 4 & 10\end{bmatrix}$
n=15	1	0.91	0.99	0.07	0.09	0.16	$\begin{bmatrix}134 & 1\\ 13 & 1\end{bmatrix}$
	2	0.84	0.83	0.93	0.16	0.52	$\begin{bmatrix}112 & 23\\ 1 & 13\end{bmatrix}$
	3	0.79	0.80	0.71	0.21	0.35	$\begin{bmatrix}108 & 27\\ 4 & 10\end{bmatrix}$
n=20	1	0.92	1.00	0.14	0.08	0.36	$\begin{bmatrix}135 & 0\\ 12 & 2\end{bmatrix}$
	2	0.83	0.82	0.93	0.17	0.51	$\begin{bmatrix}111 & 24\\ 1 & 13\end{bmatrix}$
	3	0.81	0.81	0.71	0.19	0.36	$\begin{bmatrix}110 & 25\\ 4 & 10\end{bmatrix}$

Μεταξύ των διαφορετικών μεθόδων κατηγοριοποίησης, η Μέθοδος 2 επιδεικνύει την καλύτερη προβλεπτική ικανότητα, εντοπίζοντας τα περισσότερα τοξικά δείγματα στις περισσότερες δοκιμές επιτυγχάνοντας ακρίβεια και ειδικότητα [0.7, 0.86, 0.94, 0.81, 0.84, 0.83] και [0.71, 0.71, 1.00, 0.93, 0.93, 0.93], αντίστοιχα, για n=4, 6, 8, 10, 15 και 20. Η Μέθοδος 1 επιδεικνύει την χειρότερη συμπεριφορά, αποτυγχάνοντας πλήρως στον εντοπισμό των τοξικών δειγμάτων με ειδικότητα [0.64, 0.07, 0.07, 0.07, 0.07, 0.14] ενώ η Μέθοδος 2 παρουσιάζει καλύτερες προβλέψεις με ειδικότητα [0.50, 0.86, 0.71, 0.71, 0.71, 0.71].

Τα καλύτερα αποτελέσματα αναφορικά με τον αριθμό των μεταβλητών επιδεικνύουν οι δοκιμές για n=8 η οποία εντοπίζει το σύνολο των τοξικών δειγμάτων του συνόλου ελέγχου με ακρίβεια, ευαισθησία και ειδικότητα 0.94,

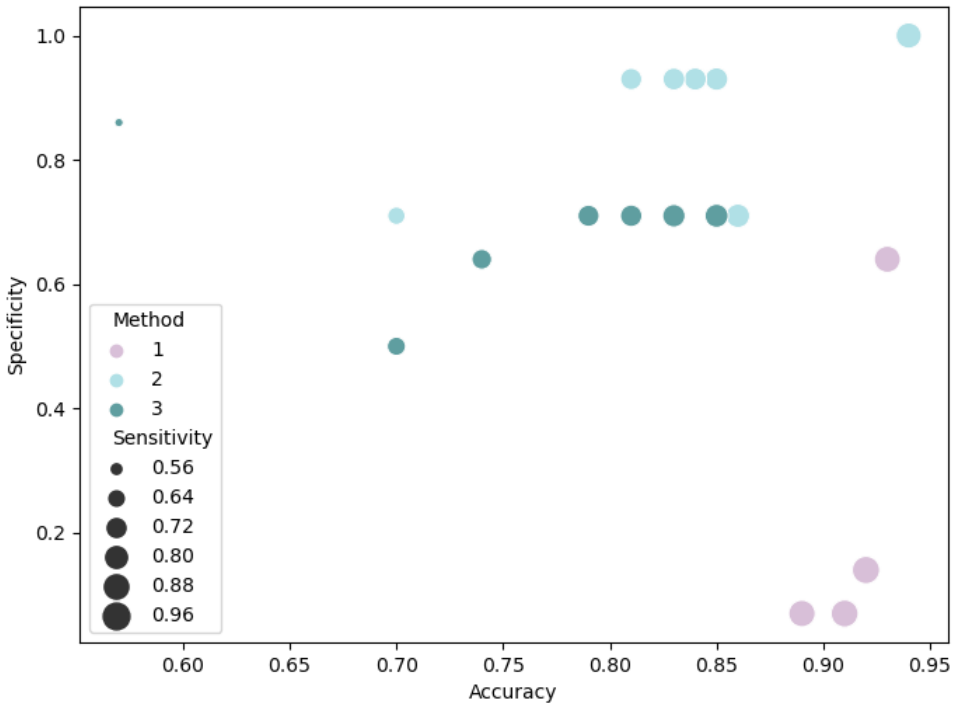
0.93 και 1, αντίστοιχα. Συγκριτικά με τις δοκιμές που δεν περιλαμβάνουν PCA, για την ίδια διαμέριση, τα στατιστικά μεγέθη βελτιώνονται για τις Μεθόδους 2 και 3 και κατηγοριοποιούνται σωστά περισσότερα δείγματα. Η συμπεριφορά αυτή παρατηρείται για n=8,10,15 και 20 ενώ για n=4,6 τα αποτελέσματα είναι παρόμοια με αυτά των δοκιμών που δεν περιλαμβάνουν ανάλυση κύριων συνιστωσών.

Ως προς τον υπολογιστικό χρόνο που απαιτεί η μοντελοποίηση, οι δοκιμές που περιλαμβάνουν εφαρμογή PCA έχουν καλύτερη απόδοση επιτυγχάνοντας μικρότερους χρόνους, μειώνοντάς τους στο 1/3 του μεγέθους περίπου, όπως φαίνεται και στον πίνακα 7.3.23. Αυξάνοντας τον αριθμό των μεταβλητών, αυξάνεται και ο υπολογιστικός χρόνος και των τριών μεθόδων εξίσου.

Πίνακας 6.3.23: Σύγκριση υπολογιστικών χρόνων σε δευτερόλεπτα με και χωρίς εφαρμογή PCA στο σύνολο Cytotox

Μέθοδος	***CPU time χωρίς PCA***	***CPU time PCA n=4***	***CPU time PCA n=6***	***CPU time PCA n=8***	***CPU time PCA n=10***
1	365	139	135	143	159
2	364	138	136	143	156
3	367	138	136	142	157

Για το σύνολο δεδομένων Cytotox, την καλύτερη επίδοση επιδεικνύει η Μέθοδος Κατηγοριοποίησης 2, με την υψηλότερη ακρίβεια και ειδικότητα. Η συμπεριφορά



Γράφημα 6.3.5: Συγκριτικό διάγραμμα απόδοσης (ακρίβεια, ευαισθησία και ειδικότητα) των τριών μεθοδολογιών κατηγοριοποίησης για το σύνολο Cytotox, με και χωρίς PCA

αυτή υποδεικνύεται και στο σχετικό γράφημα 6.3.5 , όπου καθίσταται εμφανής η πλήρης αποτυχία της Μεθόδου Κατηγοριοποίησης 1 με ή χωρίς εφαρμογή PCA.

Δοκιμές 10-Fold επικύρωσης

Σε συνέχειας της διαδικασίας που ακολουθήθηκε για την μελέτη της επιλογής μεταβλητών, δοκιμάζεται η $10 - Fold$ εσωτερική αξιολόγηση. Διαιρώντας το training set σε 10 υποσύνολα, πραγματοποιούνται 10 διαδοχικές μοντελοποιήσεις και προβλέψεις. Σε κάθε μία, ένα εκ των υποσυνόλων λειτουργεί ως σύνολο ελέγχου. Δοκιμάζονται οι διαδικασίες επιλογής μεταβλητών: PCA με n=8 και PCA με n=4. Μετά από τις 10 δοκιμές, διαμορφώνεται ο συνολικός πίνακας των στατιστικών μεγεθών για τις τρεις μεθόδους κατηγοριοποίησης.

Πίνακας 6.3.24: Στατιστικά αποτελέσματα με εφαρμογή ανάλυσης κύριων συνιστωσών με δύο τρόπους στο σύνολο εκπαίδευσης του συνόλου Cytotox

Μέθοδος επιλογής μεταβλητών	***Μέθοδος Κατηγ.***	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Conf. Matrix***
PCA n=8	1	0.89	1.00	0.00	0.11	0.00	$\begin{bmatrix} 308 & 0 \\ 37 & 0 \end{bmatrix}$
	2	0.92	0.95	0.68	0.07	0.62	$\begin{bmatrix} 294 & 14 \\ 12 & 25 \end{bmatrix}$
	3	0.84	0.89	0.46	0.15	0.30	$\begin{bmatrix} 274 & 14 \\ 20 & 17 \end{bmatrix}$
PCA n=4	0.90	0.97	0.32	0.12	0.38	0.90	$\begin{bmatrix} 299 & 9 \\ 25 & 12 \end{bmatrix}$
	0.78	0.79	0.70	0.21	0.34	0.78	$\begin{bmatrix} 242 & 66 \\ 11 & 26 \end{bmatrix}$
	0.79	0.83	0.38	0.25	0.17	0.79	$\begin{bmatrix} 257 & 51 \\ 23 & 14 \end{bmatrix}$

Μελετώντας τα αποτελέσματα συνδυαστικά με αυτά του πίνακα 6.2.3 ο οποίος παρουσιάζει αποτελέσματα 10-Fold επικύρωσης με μεθόδους επιλογής μεταβλητών, την καλύτερη επίδοση συνολικά με ιδιαίτερη βάση στην ειδικότητα, φαίνεται να επιδεικνύουν οι δοκιμές με τη PCA με n=8. Όπως και σε κάθε δοκιμή του συνόλου, η Μέθοδος 1 έχει ιδιαίτερα απογοητευτικά αποτελέσματα αλλά οι Μέθοδοι 2 και 3 , δίνουν επαρκώς καλά αποτελέσματα και ως βέλτιστη επιλογή μεταξύ των δυνατών τρόπων μείωσης των διαστάσεων του συνόλου, δοκιμάζεται η εφαρμογή τους και στο εξωτερικό σύνολο ελέγχου. Η επίδοση του μοντέλου σε αυτό, για τις τρεις μεθοδολογίες παρουσιάζεται στον πίνακα 6.3.25.

Πίνακας 6.3.25: Στατιστικά αποτελέσματα με εφαρμογή της βέλτιστης μεθοδολογίας μείωσης διαστάσεων (PCA με n=8) στο test set του συνόλου Cytotox

Μέθοδος Κατηγ.	***Accuracy***	***Sensitivity***	***Specificity***	***Error rate***	***MCC***	***Conf. Matrix***
1	0.91	1.00	0.07	0.09	0.26	$\begin{bmatrix} 135 & 0 \\ 13 & 1 \end{bmatrix}$
2	0.94	0.93	1.00	0.06	0.75	$\begin{bmatrix} 126 & 9 \\ 0 & 14 \end{bmatrix}$
3	0.85	0.86	0.71	0.15	0.42	$\begin{bmatrix} 116 & 19 \\ 4 & 10 \end{bmatrix}$

Η Μέθοδος Κατηγοριοποίησης 2 στέφεται με απόλυτη επιτυχία εντοπίζονται με απόλυτη ειδικότητα όλα τα τοξικά δείγματα, διατηρώντας υψηλά την ευαισθησία. Η εφαρμογή του μοντέλου στο σύνολο ελέγχου με PCA για n=8, είναι επιτυχής και συστήνεται ως η βέλτιστη μέθοδος μείωσης των διαστάσεων του συνόλου με σκοπό να καταστεί το μοντέλο ταχύτερο και πιο εύρωστο.

6.4. Συνολική αποτίμηση του μοντέλου και των μεθοδολογιών προεπεξεργασίας

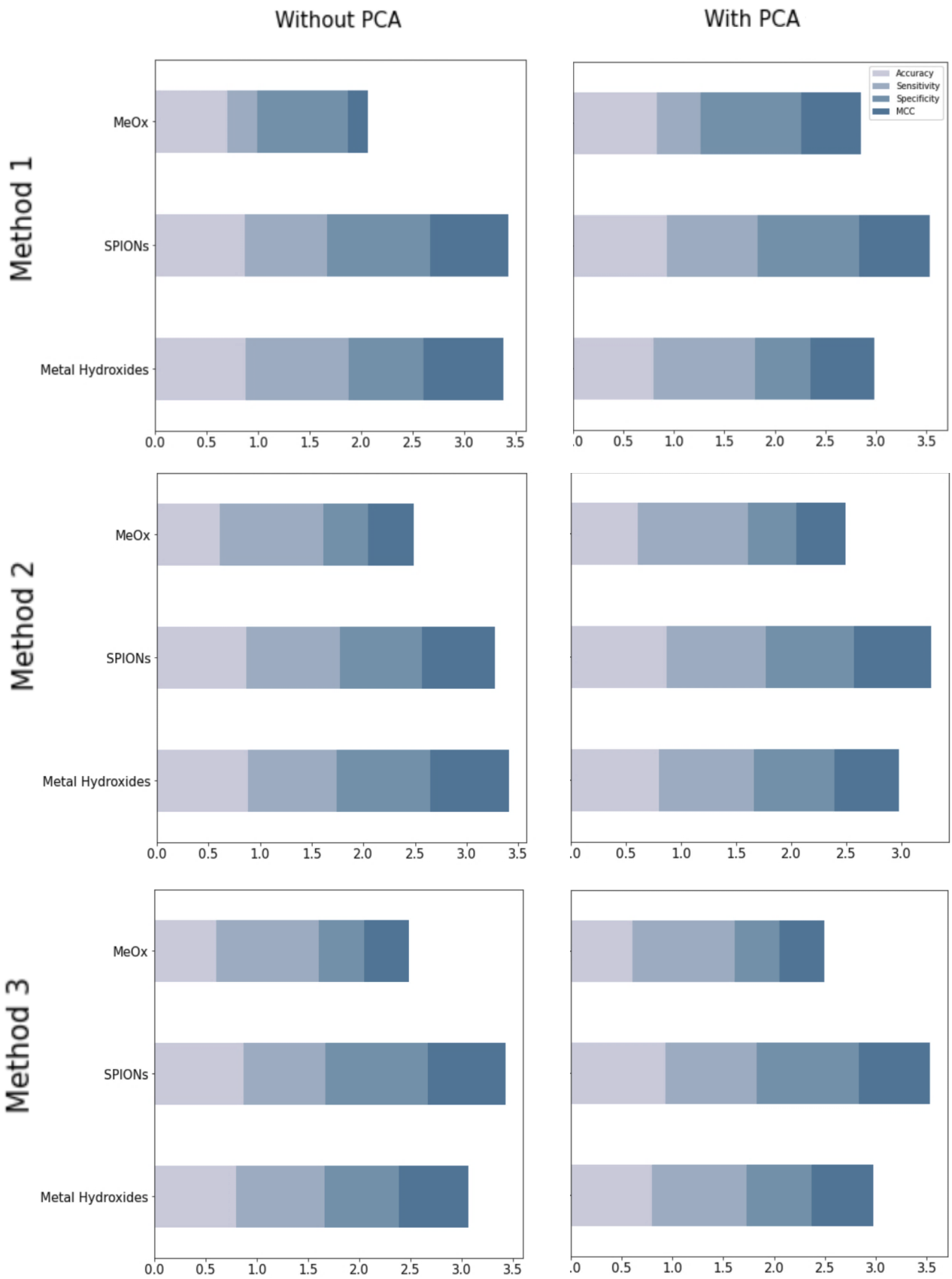
6.4.1. Σύγκριση μεταξύ των Μεθόδων Κατηγοριοποίησης και των συνόλων δεδομένων

Για την συγκριτική απεικόνιση της ανταπόκρισης του μοντέλου στα διαφορετικά σύνολα δεδομένων, με και χωρίς εφαρμογή PCA, συνίσταται ο πίνακας 6.4.1.

Πίνακας 6.4.1: Συγκριτικός πίνακας με τις τιμές του Matthews Correlation Coefficient που προκύπτει για δοκιμές εξωτερικής αξιολόγησης σε όλα τα σύνολα δεδομένων [4]

	Χωρίς PCA			**Με PCA**		
	1	2	3	1	2	3
MeHydOx	0.71	0.61	0.61	-0.16	1.00	0.61
SPIONs	1.00	1.00	1.00	1.00	1.00	1.00
MeOx	0.61	0.67	0.67	0.00	1.00	1.00
Cytotox	0.24	0.54	0.26	0.26	0.75	0.42

[4] Για τα σύνολα MeHydOx, SPIONs και MeOx, η διαμέριση έχει πραγματοποιηθεί με Kennard Stone για training ratio 0.8 ενώ για το σύνολο Cytotox βάσει της προτεινόμενης τυχαιοποιημένης διαμέρισης. Οι δοκιμές PCA για τα σύνολα MeHydOx, SPIONs και MeOx πραγματοποιήθηκαν με n=3 ενώ για το σύνολο Cytotox με n=8.



Γράφημα 6.4.1: Διαγραμματική ποιοτική σύγκριση της επίδοσης των τριών Μεθόδων του μοντέλου σε Leave-One-Out επικύρωση με και χωρίς PCA για όλα τα σύνολα δεδομένων

Καθίσταται εμφανές πως η καλύτερη απόδοση του μοντέλου αντιστοιχεί στο σύνολο SPIONs, για το οποίο κατηγοριοποιεί ορθά το σύνολο των δειγμάτων με κάθε μέθοδο με και χωρίς PCA. Ικανοποιητικά αποτελέσματα εντοπίζονται και στην περίπτωση του συνόλου MeHydOx, αν και δεν υπάρχουν σημαντικές διαφορές μεταξύ εφαρμογής και μη της PCA, σε αντίθεση με το σύνολο MeOx, του οποίου οι προβλέψεις βελτιώνονται σημαντικά με εφαρμογή PCA. Οι τιμές MCC για το σύνολο Cytotox, διατηρούνται συνολικά σε χαμηλότερα επίπεδα αλλά η βελτίωση με εφαρμογή PCA είναι καθολική για όλες τις Μεθόδους Κατηγοριοποίησης. Συγκρίνοντας τις μεθόδους, αθροιστικά για όλα τα σύνολα δεδομένων, η Μέθοδος Κατηγοριοποίησης 1 εμφανίζει τη χαμηλότερη προβλεπτική ικανότητα, η οποία μάλιστα χειροτερεύει με εφαρμογή PCA. Αντίθετα, οι Μέθοδοι 2 και 3, βελτιώνουν τις προβλέψεις που πραγματοποιούν όταν εφαρμόζεται μείωση μεταβλητών.

Μεταξύ των δύο, ξεχωρίζει η Μέθοδος 2 η οποία επιτυγχάνει να εντοπίσει όλα τα τοξικά δείγματα και στο σύνολο Cytotox, ικανότητα στην οποία αποτυγχάνουν σε μεγάλο βαθμό οι υπόλοιπες προβλέψεις που πραγματοποιούν όταν εφαρμόζεται μείωση μεταβλητών.

Με σκοπό την περαιτέρω σύγκριση της απόκρισης του μοντέλου, εξετάζεται σε η απόδοση του σε εσωτερική αξιολόγηση και συγκεκριμένα σε επικύρωση $Leave - One - Out$. Βάσει των γραφημάτων 6.4.1 που διαμορφώνονται, εντοπίζονται κάποιες σημαντικές παρατηρήσεις.

Το σύνολο MeOx χαρακτηρίζεται από την χαμηλότερη απόδοση στο σύνολο των δοκιμών, αν και υπάρχει βελτίωση στις στατιστικές επιδόσεις με εφαρμογή PCA ενώ οι Μέθοδοι 2 και 3 έχουν παρόμοια συμπεριφορά στα σύνολα και δεν παρατηρούνται αισθητές βελτιώσεις με μείωση των μεταβλητών. Λαμβάνοντας υπόψιν όλες τις τιμές, για το σύνολο MeOx, την καλύτερη απόδοση εμφανίζει η Μέθοδος 1, για το σύνολο SPIONs η 3 και για το σύνολο MeHydOx η 2.

Μέσω των συνεχών δοκιμών επικύρωσης αναδεικνύονται και καταγράφονται οι βέλτιστες Μέθοδοι Κατηγοριοποίησης και μείωσης των διαστάσεων για κάθε σύνολο δεδομένων, με σημαντικότερα κριτήρια τους στατιστικούς δείκτες απόδοσης, όπως αυτοί παρουσιάζονται στον πίνακα 6.4.2.

Από την παραπάνω ανάλυση είναι εμφανές πως η εφαρμογή PCA στα 2 εκ των 4 συνόλων δεδομένων συμβάλλει, όχι μόνο στην μείωση του υπολογιστικού χρόνου, αλλά και στην αύξηση της επίδοσης του μοντέλου. Η μείωση της επίδοσης του μοντέλου κατά την εφαρμογή του με PCA στο σύνολο SPIONs πιθανώς να οφείλεται στον ήδη μικρό αριθμό ιδιοτήτων που χαρακτηρίζει το σύνολο. Αντίθετα, σε σύνολο με υψηλό αριθμό ιδιοτήτων όπως το Cytotox ή το MeOx, η μεγάλη μείωση του μέσω PCA αυξάνει την αποτελεσματικότητα και ευρωστία του μοντέλου.

Πίνακας 6.4.2: Συγκριτικός πίνακας με τις τιμές του Matthews Correlation Coefficient που προκύπτει για το βέλτιστο συνδυασμό Μεθόδου Κατηγοριοποίησης και μείωσης διαστάσεων

	Βέλτιστη Μέθοδος	Μέθοδος Επικύρωσης	Μέθοδος Μείωσης Διαστάσεων	MCC
MeHydOx	2	Εξωτερική επικύρωση	Δεν εφαρμόζεται	1.00
SPIONs	1-2-3[5]	Εξωτερική επικύρωση	Δεν εφαρμόζεται	1.00
MeOx	2-3[6]	Εξωτερική επικύρωση	PCA (n=3)	1.00
Cytotox	2	Εξωτερική επικύρωση	PCA (n=8)	0.75

Παράλληλα, αξίζει να σημειωθεί πως η Μέθοδος Κατηγοριοποίησης 2 αποδεικνύεται η πλέον αποδοτικότερη για την ορθή πρόβλεψη της τοξικότητας με εφαρμογή του προτεινόμενου μοντέλου καθώς επιφέρει την καλύτερη απόδοση κατά την εφαρμογή της σε όλα τα σύνολα, συχνά εξίσου καλή με κάποια εκ των άλλων 2 μεθόδων.

Κατά την εφαρμογή των βέλτιστων μεθόδων σχηματίζονται περισσότερα του ενός cluster για κάθε κλάση. Ο αριθμός τους παρουσιάζεται στον πίνακα 6.4.3.

Πίνακας 6.4.3: Αριθμός των σχηματιζόμενων cluster σε κάθε κλάση τοξικότητας για τα σύνολα δεδομένων MeHydOx, SPIONs και MeOx βάσει της βέλτιστης τεχνικής μείωσης των διαστάσεων

	Αριθμός cluster μη τοξικών δειγμάτων	Αριθμός cluster τοξικών δειγμάτων
MeHydOx	2	2
SPIONs	1	1
MeOx	2	1

Φυσικά, αύξηση του αριθμού των δειγμάτων οδηγεί σε μεγαλύτερο αριθμό συστάδων καθώς οι συσχετίσεις και οι σχετικές αποστάσεις των δειγμάτων στον πολυδιάστατο χώρο περιπλέκονται. Όπως ήταν αναμενόμενο, το σύνολο SPIONs σχηματίζει εμφανώς λιγότερες συστάδες, ένα για κάθε κλάση, συγκριτικά με το σύνολο MeHydOx που σχηματίζει δύο cluster για κάθε κλάση.

[5] Όλες οι Μεθοδολογίες Κατηγοριοποίησης οδηγούν σε ίδιες προβλέψεις, απόλυτα επιτυχημένες ως προς την ορθότητά τους. Εμφανίζονται μικρές διαφορές ως προς τον υπολογιστικό χρόνο, ωστόσο, λόγω της μεταβλητότητάς τους, δεν μπορεί να επιλεχθεί μια εξ αυτών ως η βέλτιστη λόγω ταχύτητας.

[6] Ομοίως με το προηγούμενο σύνολο, οι Μεθοδολογίες Κατηγοριοποίησης 2 και 3 οδηγούν σε ίδιες προβλέψεις, απόλυτα επιτυχημένες ως προς την ορθότητά τους. Λόγω της μεταβλητότητας των υπολογιστικών χρόνων, δεν μπορεί να επιλεχθεί μια εξ αυτών ως η βέλτιστη.

6.4.2. Σύγκριση επιδόσεων με βιβλιογραφικές δοκιμές μοντέλων

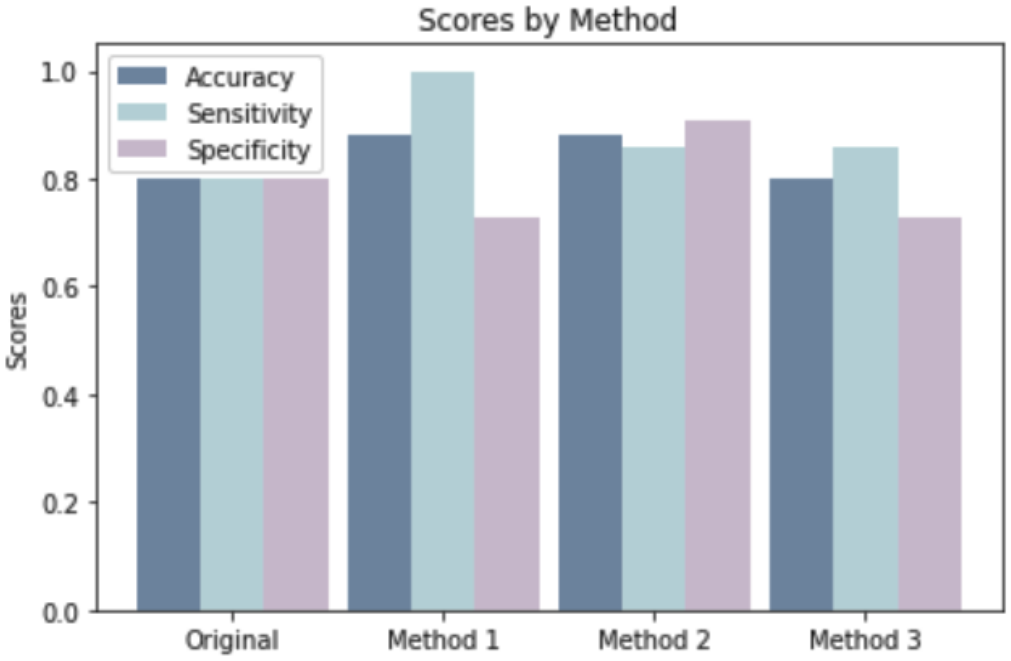
Στη συνέχεια παρουσιάζονται τα αποτελέσματα της μοντελοποίησης της προτεινόμενης μεθοδολογίας σε σύγκριση με άλλες μεθοδολογίες που χρησιμοποιούν αυτά τα σύνολα δεδομένων. Σημειώνεται ότι για το σύνολο Cytotox, στη δημοσίευση από όπου αντλήθηκαν τα δεδομένα παρουσιάζει ένα μοντέλο παλινδρόμησης και όχι κατηγοριοποίησης, οπότε δε δύναται να πραγματοποιηθεί σύγκριση με την παρούσα μεθοδολογία.

MeHydOx

Το προβλεπτικό μοντέλο που περιγράφεται στη βιβλιογραφία δεν εφαρμόζει κάποια QSAR μεθοδολογία λόγω του μικρού αριθμού δειγμάτων. Αντιθέτως, επιλέγει την ανάλυση μέσω Μερικής Παλινδρόμησης Ελαχίστων Τετραγώνων και εντοπίζει τις ιδιότητες με την ισχυρότερη επίδραση στο endpoint με χρήση Δέντρου Αποφάσεων. Πραγματοποιώντας δοκιμές εσωτερικής αξιολόγησης Leave-One-Out για το σύνολο των 25 δειγμάτων επιτυγχάνεται ακρίβεια, ευαισθησία και ειδικότητα ίσες με 0.8 ενώ μόνο 2 εκ των τοξικών δειγμάτων δεν κατηγοριοποιείται ορθά [47].

Οι αντίστοιχες δοκιμές του προτεινόμενου μοντέλου καταλήγουν σε παρόμοια στατιστικά μεγέθη. Οι Μέθοδοι 1,2 και 3 επιτυγχάνουν ακρίβεια, ευαισθησία και ειδικότητα [0.88, 1, 0.73], [0.88, 0.86, 0.91], [0.8, 0.86, 0.73], αντίστοιχα. Αν και οι Μέθοδοι 1 και 3 επιδεικνύουν παρόμοια αποτελέσματα με αυτά της βιβλιογραφίας, η Μέθοδος 2 έχει ακόμα καλύτερη επίδοση εντοπίζοντας όλα πλην

Γράφημα 6.4.2: Σύγκριση ακρίβειας, ευαισθησίας και ειδικότητας μεταξύ των μεθόδων της βιβλιογραφίας και των τριών προτεινόμενων μέσω δοκιμών Leave-One-Out



ενός τα τοξικά δείγματα με ειδικότητα 0.91 και ακρίβεια 0.88. Τα στοιχεία αυτά επιβεβαιώνουν την επιτυχία του μοντέλου στην ορθή πρόβλεψη των τοξικών δειγμάτων του συνόλου MeHydOx, η οποία δύναται να ξεπεράσει την επίδοση της βιβλιογραφίας.

SPIONs

Αναφορικά με το σύνολο SPIONs και παρά το μικρό μέγεθός του, το προβλεπτικό μοντέλο που περιγράφεται στη βιβλιογραφία επιλέγει την εφαρμογή μοντέλου QSAR. Μάλιστα, επικεντρώνεται σε μοντέλα Αυτοματοποιημένης Μηχανικής Μάθησης και συγκεκριμένα σε μεθόδους Λογιστικής Παλινδρόμησης. Στις βιβλιογραφικές δοκιμές, το σύνολο χωρίζεται με τη μέθοδο της Kennard Stone σε 10 δείγματα εκπαίδευσης και 5 δείγματα επικύρωσης. Στο σύνολο επικύρωσης, επιτυγχάνεται ακρίβεια, ευαισθησία και ειδικότητα ίση με τη μονάδα [37]. Η ίδια επιτυχία πρόβλεψης προκύπτει και με εφαρμογή οποιασδήποτε εκ των τριών Μεθόδων Κατηγοριοποίησης του προτεινόμενου μοντέλου της παρούσας εργασίας.

Η υψηλή απόδοση του, η οποία ταυτίζεται με αυτή της βιβλιογραφίας, επιβεβαιώνει την επιτυχία του μοντέλου στο σύνολο SPIONs.

MeOx

Το προβλεπτικό μοντέλο που περιγράφεται στη βιβλιογραφία, μετά από ποικίλες δοκιμές nano-SAR μεθοδολογιών, επιλέγεται να βασίζεται σε τεχνική Μηχανής Διανυσματικής Υποστήριξης (Support Vector Machine). Τα διαθέσιμα στοιχεία για την απόδοση του αναφορικά με το σύνολο MeOx περιλαμβάνουν μόνο την ακρίβεια. Συγκεκριμένα, η διαίρεση του μοντέλου σε σύνολο εκπαίδευσης και επικύρωσης πραγματοποιείται μέσω της bootstrap μεθόδου 0.632 χωρίς να προσδιορίζεται ο ακριβής αριθμός των δειγμάτων. Ανάλογα το μοντέλο που χρησιμοποιείται η ακρίβεια που επιτυγχάνεται κυμαίνεται σε τιμές μεταξύ 0.85 και 0.94 [48].

Συγκρίνοντας τις τιμές αυτές με το προτεινόμενο μοντέλο, επιλέγεται ως συγκρίσιμο μέθοδος η εξωτερική αξιολόγηση με διαμέριση μέσω Kennard Stone και κλάσμα διαμέρισης 0.7. Σε αυτές τις δοκιμές οι Μέθοδοι Κατηγοριοποίησης επιτυγχάνουν ακρίβεια 0.71, 0.86, 0.86. Αν και η Μέθοδος 1 αποδίδει εμφανώς χειρότερα συγκριτικά με το βιβλιογραφικό μοντέλο, οι Μέθοδοι 2 και 3 βρίσκονται εντός των ορίων των επιδόσεων του, γεγονός που καθιστά το μοντέλο μια αποτελεσματική εναλλακτική μέθοδος πρόβλεψης τοξικότητας και για το σύνολο MeOx.

Κεφάλαιο 7

Συμπεράσματα και προτάσεις για περαιτέρω έρευνα

Η εξέλιξη της τεχνολογίας και της μηχανικής, συνδυαστικά με την ανάγκη εύρεσης νέων υλικών με συγκεκριμένες ιδιότητες, έχει συμβάλει καθοριστικά στην ραγδαία ανάπτυξη των νανουλικών. Αυτά, όχι μόνο εμφανίζουν ξεχωριστές ιδιότητες στη νάνο-κλίμακα, αλλά επιδέχονται και δομικές τροποποιήσεις ώστε να αποκτήσουν επιθυμητά χαρακτηριστικά. Ωστόσο, παρά τη μεγάλη ζήτηση σε πληθώρα εφαρμογών, σημαντικοί προβληματισμοί έχουν προκύψει αναφορικά με τις αρνητικές επιπτώσεις τις οποίες μπορεί να έχουν τόσο στον ανθρώπινο οργανισμό, όσο και σε οποιοδήποτε βιολογικό σύστημα. Υπό αυτό το πρίσμα, καθίσταται αναγκαία η εις βάθος μελέτη της συμπεριφοράς των νανουλικών σε τέτοια συστήματα, μέσω πειραμάτων.

Στην προσπάθεια σταδιακής ελάττωσης της χρήσης πειραματόζωων, αλλά και για την εξοικονόμηση χρόνου και πόρων, τέτοιες δοκιμές μπορούν να διεξαχθούν με τη βοήθεια της Νανοπληροφορικής, η οποία προσφέρει υπολογιστικές μεθόδους και αλγοριθμικά εργαλεία για την ανάλυση και μετέπειτα πρόβλεψη της συμπεριφοράς των νανοδομών ως προς την τοξικότητά τους. Ταυτόχρονα η ανάπτυξη του πεδίου της υπολογιστικής μελέτης της τοξικότητας των νανοϋλικών, επιτρέπει την ανάπτυξη εκ των προτέρων ασφαλών νανοδομών. Συγκεκριμένα, καθίστανται δυνατοί ο σχεδιασμός νανοϋλικών «στο χαρτί» και η εικονική μελέτη των ανεπιθύμητων επιπτώσεών τους με βάση τις ιδιότητες που προκύπτουν από τη δομή τους (virtual screening). Στη συνέχεια, είναι εφικτό να πραγματοποιηθούν διορθωτικές μετατροπές στη δομή τους -και πάλι εικονικά- έως ότου επιτευχθεί ο συνδυασμός επιθυμητών ιδιοτήτων και ασφαλών δομών. Συνεπώς στην παραγωγική διαδικασία θα προωθούνται εξ' αρχής ασφαλείς νανοδομές και η ανάγκη για την πραγματοποίηση πολλών πειραμάτων αξιολόγησης κινδύνων θα μειωθεί (safety-by-design).

Στην παρούσα Διπλωματική Εργασία αναπτύχθηκε ένα μοντέλο συσταδοποίησης με σκοπό την πρόβλεψη του τοξικού χαρακτήρα νανοσωματιδίων. Η μεθοδολογία που προτείνεται βασίζεται στην τεχνική read-across, επιδιώκοντας την αυτοματοποίηση της διαδικασίας ομαδοποίησης των νανοσωματιδίων και τον χαρακτηρισμό άγνωστων νανοσωματιδίων βάσει αυτής. Το μοντέλο ακολουθεί μια σειρά πέντε διαδοχικών βημάτων. Τα τέσσερα πρώτα βήματα τοποθετούν τα νανοσωματίδια στον πολυδιάστατο χώρο με άξονα τις ιδιότητές τους οι οποίες λειτουργούν ως συντεταγμένες. Στη συνέχεια, δημιουργούν κατηγοριοποιημένες σφαίρες γύρω από κάθε γνωστό σωματίδιο, χαρακτηριζόμενες από δύο σφαιρικά σύνορα, ένα συντηρητικό και ένα ελεύθερο. Μέσω βελτιστοποίησης με γραμμικό προγραμματισμό, στο 5^{ο} και τελευταίο βήμα, επιλέγεται ο βέλτιστος συνδυασμός

εξ αυτών, ώστε να καλύπτεται όλος ο δειγματικός χώρος. Βάσει των διαμορφωμένων συστάδων και των σχετικών αποστάσεων από αυτές, άγνωστα νανοσωματίδια κατηγοριοποιούνται σε μια εκ των δύο ομάδων, τοξικών νανοσωματιδίων ή μη. Η κατηγοριοποίηση πραγματοποιήθηκε βάσει τριών λογικών μεθόδων, βασιζόμενες σε συγκριτικές αποστάσεις συστάδων και δειγμάτων. Οι δύο εξ αυτών, οι Μέθοδοι Κατηγοριοποίησης 2 και 3, προτείνονται για πρώτη φορά στην παρούσα Εργασία.

Στο πλαίσιο αξιολόγησης του μοντέλου, διενεργήθηκαν έλεγχοι εσωτερικής και εξωτερικής επικύρωσης. Σε κάθε έλεγχο, το σύνολο δεδομένων χωρίστηκε σε σύνολο εκπαίδευσης και ελέγχου, με το δεύτερο να μην συμμετέχει στη διαδικασία εκπαίδευσης του μοντέλου. Για την εκτίμηση της απόδοσης επιστρατεύτηκαν στατιστικά εργαλεία όπως η ακρίβεια, η ευαισθησία, η ειδικότητα και ο πίνακας σύγχυσης των προβλέψεων, τα οποία αποτέλεσαν βασικό κριτήριο αξιολόγησης στα δεδομένα του συνόλου ελέγχου. Όλοι οι έλεγχοι συνοδεύτηκαν από τον υπολογισμό του πεδίου εφαρμογής, ώστε να καθοριστεί ο δειγματικός χώρος μέσα στον οποίο παράγονται αξιόπιστες προβλέψεις. Τέλος πραγματοποιήθηκε έλεγχος τυχαίας επιλογής (y-scrambling), ο οποίος επικυρώνει τη μέθοδο, αποκλείοντας την πιθανότητα η καλή επίδοση να είναι απόρροια τυχαιότητας.

7.1. Ανάλυση αποτελεσμάτων και συμπεράσματα

Η προτεινόμενη μεθοδολογία εφαρμόστηκε σε τέσσερα σύνολα δεδομένων όπως αυτά περιλαμβάνονται στις δημοσιεύσεις *'Towards an alternative to nano-QSAR for nanoparticle toxicity ranking in case of small datasets'* των Forest *et al.* (2019), *'QSAR modeling of the toxicity classification of superparamagnetic iron oxide nanoparticles (SPIONs) in stem-cell monitoring applications: an integrated study from data curation to model development'* των Kotzabasaki *et al.* (2020), '*Development of structure–activity relationship for metal oxide nanoparticles*' των Liu *et al.* (2013) *και 'Predicting Cytotoxicity of Metal Oxide Nanoparticles Using Isalos Analytics Platform'* των Papadiamantis *et al*. (2020).

Η αξιολόγηση της προβλεπτικής ικανότητας της μεθοδολογίας σε κάθε σύνολο δεδομένων πραγματοποιήθηκε μέσω διαχωρισμού του σε σύνολο εκπαίδευσης και ελέγχου. Η επικύρωση πραγματοποιήθηκε στο δεύτερο σύνολο, το οποίο δεν συμμετείχε σε κανένα στάδιο της εκπαίδευσης.

Σύνολο δεδομένων	**Κωδική ονομασία**	**Αριθμός δειγμάτων**	**Αριθμός ιδιοτήτων**
'Towards an alternative to nano-QSAR for nanoparticle toxicity ranking in case of small datasets' των Forest et al (2019)	*MeHydOx*	*25*	*12*

'QSAR modeling of the toxicity classification of superparamagnetic iron oxide nanoparticles (SPIONs) in stem-cell monitoring applications: an integrated study from data curation to model development' των Kotzabasaki et al (2020)	*SPIONs*	*15*	*6*
'Development of structure–activity relationship for metal oxide nanoparticles' των Liu et al (2013)	*MeOx*	*23*	*24*
'Predicting Cytotoxicity of Metal Oxide Nanoparticles Using Isalos Analytics Platform' των Papadiamantis et al (2020)	*Cytotox*	*494*	*65*

Τα σύνολα εκπαίδευσης των τεσσάρων συνόλων δεδομένων δημιούργησαν συστάδες, στις οποίες στη συνέχεια κατηγοριοποιήθηκαν τα αντίστοιχα σύνολα ελέγχου, και εκτιμήθηκε η ακρίβεια των προβλέψεων, στην περίπτωση εσωτερικής και εξωτερικής αξιολόγησης. Στις περισσότερες δοκιμές, τα στατιστικά μεγέθη της ακρίβειας, της ειδικότητας και της ευαισθησίας διατηρούνται υψηλότερα του 0.8 ενώ σε όλα τα σύνολα τα οποία εξετάστηκαν, τα αποτελέσματα που προέκυψαν ήταν παρόμοιας τάξης με αυτά της βιβλιογραφίας. Μάλιστα, στην περίπτωση του συνόλου των Forest *et al.* (2019), τα αποτελέσματα βελτιώνονται εμφανώς συγκριτικά με αυτά της δημοσίευσης [47] , καθώς, ιδιαίτερα με τη Μέθοδο Κατηγοριοποίησης 2, επιτυγχάνεται ακρίβεια, ειδικότητα και ευαισθησία ίση με 0.88, 0.86 και 0.91 αντίστοιχα.

Τα τέσσερα σύνολα δεδομένων μοντελοποιήθηκαν ως έχουν χρησιμοποιώντας όλες τις διαθέσιμες ιδιότητές τους αλλά και με εφαρμογή της μεθόδου Ανάλυσης Κύριων Συνιστωσών (PCA) η οποία επιτρέπει τη μείωση του αριθμού ανεξάρτητων μεταβλητών με σύμπτυξή τους. Οι παράμετροι που υπολογίστηκαν για τη μετατροπή του συνόλου εκπαίδευσης σε κάθε δοκιμή χρησιμοποιήθηκαν για τη μετατροπή του συνόλου ελέγχου, εξασφαλίζοντας μηδενική συμμετοχή του συνόλου ελέγχου στην εκπαίδευση του μοντέλου. Η σύγκριση μεταξύ των δύο τεχνικών προεπεξεργασίας υποδεικνύει βελτίωση του χρήσει PCA, κυρίως αναφορικά με τα σύνολα MeHydOx, Cytotox και MeOx, δύο εκ των οποίων επιτυγχάνουν ακρίβεια ίση με 1. Το σύνολο SPIONs επέφερε συγκρίσιμα αποτελέσματα και με τις δύο τεχνικές οι οποίες κρίθηκαν απόλυτα επιτυχημένες με ακρίβεια ίση με 1. Το σύνολο δεδομένων SPIONs λόγω του επαρκούς μεγέθους του, δύναται να δεχθεί και προεπεξεργασία επιλογής μεταβλητών (feature selection). Τα αποτελέσματα που προκύπτουν είναι συγκρίσιμα με αυτά των δύο άλλων τεχνικών με τιμές ακρίβειας που κυμαίνονται σε τιμές 0.7 ως 0.9.

Με σκοπό να επιβεβαιωθεί η ευρωστία των μοντέλων και να αποκλειστεί το ενδεχόμενο η επιτυχία τους να οφείλεται σε τυχαίους συσχετισμούς μεταξύ ανεξάρτητων μεταβλητών και της μεταβλητής στόχου, πραγματοποιείται έλεγχος τυχαίας επιλογής ο οποίος επιφέρει τα επιθυμητά αποτελέσματα. Οι αποτυχημένες προβλέψεις υπερισχύουν των επιτυχημένων, καθιστώντας το μοντέλο ορθά δομημένο. Η αξιοπιστία των προβλέψεων επιβεβαιώνεται και μέσω

υπολογισμού του πεδίου εφαρμογής των μοντέλων. Ως επί το πλείστον, τα δείγματα του συνόλου ελέγχου βρίσκονται εντός του πεδίου εφαρμογής, επικυρώνοντας την εγκυρότητά των προβλέψεων.

Τέλος η μεθοδολογία που αναπτύχθηκε, διατίθεται ελεύθερα υπό την μορφή συνάρτησης και κλάσης της γλώσσας προγραμματισμού Python προς χρήση από όλη την επιστημονική κοινότητα. Ο κώδικας, συνοδευόμενος από οδηγίες χρήσης. είναι προσβάσιμος από το αποθετήριο GitHub. Οφείλεται να τονιστεί ότι παρόλο που η μεθοδολογία εφαρμόστηκε σε δεδομένα νανοτοξικότητας, αυτό δεν αποκλείει την εφαρμογή του σε οποιοδήποτε τομέα της επιστήμης δεδομένων και της μηχανικής μάθησης, όπου χρειάζεται κάποια εφαρμογή ενός μοντέλου κατηγοριοποίησης.

7.2. Προτάσεις για μελλοντική έρευνα

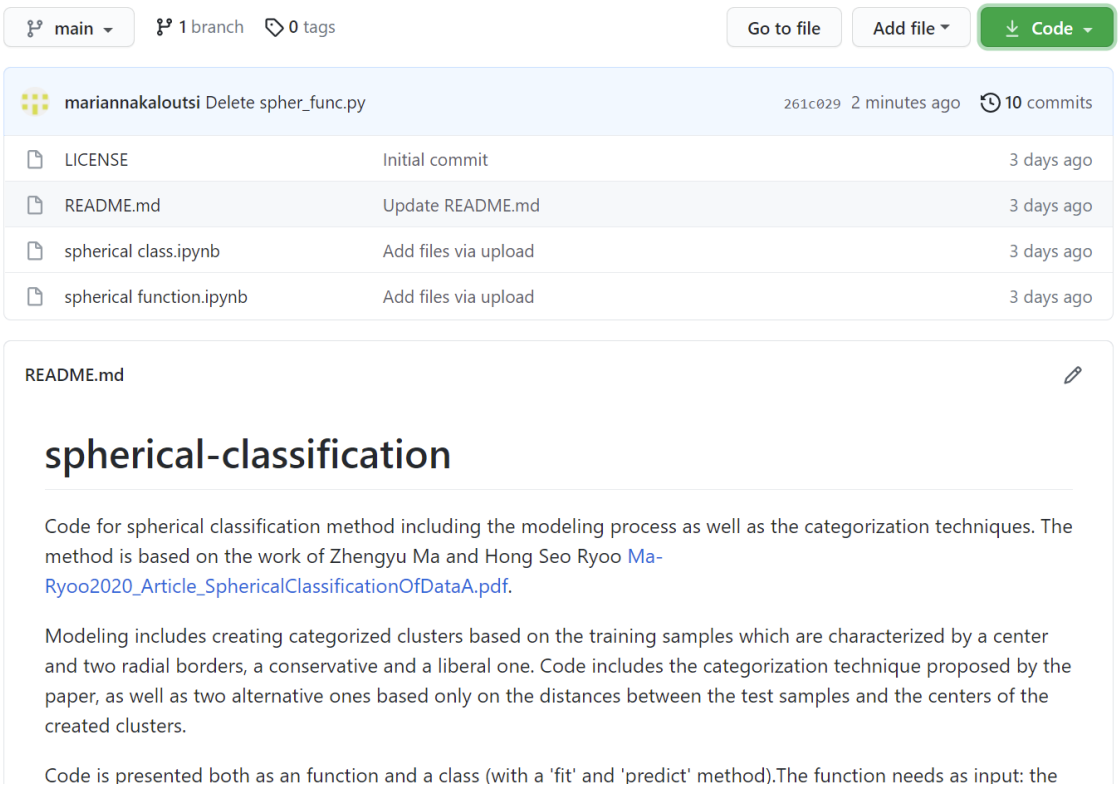
Η προτεινόμενη μεθοδολογία πρόβλεψης τοξικότητας που παρουσιάστηκε σε αυτή τη Διπλωματική Εργασία αποτελείται από δύο σκέλη, την συσταδοποίηση γνωστών δειγμάτων και την κατηγοριοποίηση άγνωστων δειγμάτων, και τα δύο εκ των οποίων επιδέχονται βελτιώσεις και περαιτέρω δοκιμές. Ως προς το σκέλος της συσταδοποίησης, προτείνεται η έρευνα να επικεντρωθεί σε τρόπους επαναδιαμόρφωσης της αρχιτεκτονικής των βημάτων με σκοπό να μειωθεί το υπολογιστικό κόστος. Αντίστοιχα, ως προς το σκέλος της κατηγοριοποίησης, ενθαρρύνεται η αναζήτηση εναλλακτικών πολύ-παραμετρικών μεθόδων κατηγοριοποίησης, οι οποίες θα λαμβάνουν υπόψιν τους περισσότερα χαρακτηριστικά των μη κατηγοριοποιημένων δειγμάτων και όχι μόνο συγκριτικές αποστάσεις. Μια τέτοια πολύπλοκη μέθοδος θα μπορούσε να αναπτυχθεί μέσω ενός προβλήματος βελτιστοποίησης με ακέραιο γραμμικό προγραμματισμό.

Με σκοπό τη μείωση του υπολογιστικού κόστους και αύξησης της ευρωστίας, συστήνεται και η ενσωμάτωση στο μοντέλου μιας επαναληπτικής μεθόδου μείωσης των ανεξάρτητων μεταβλητών-ιδιοτήτων, στα πλαίσια βελτιστοποίησης του. Φυσικά, η περαιτέρω δοκιμή του προτεινόμενου μοντέλου σε άλλα σύνολα δεδομένων κρίνεται καθοριστικής σημασίας ώστε να αξιολογηθεί βαθύτερα η απόδοσή του και να σκιαγραφηθεί η συμπεριφορά του. Απώτερο σκοπό αποτελεί η αποδοχή του από την επιστημονική κοινότητα και η εφαρμογή του σε πραγματικά προβλήματα, ως μια ανταγωνιστική εναλλακτική των *in vivo* πειραμάτων.

Κεφάλαιο 8

Διάθεση του κώδικα στην επιστημονική κοινότητα

Η προτεινόμενη μέθοδος, κωδικοποιημένη στην γλώσσα προγραμματισμού Python, διατίθεται στην επιστημονική κοινότητα μέσω του αποθετηρίου GitHub μέσω του συνδέσμου [mariannakaloutsi/spherical-classification: Code for a new spherical classification method based on the creation of clusters (github.com)] προς χρήση, επεξεργασία, μετατροπή και βελτίωση (βλέπε Εικόνα 8.1.1). Η πρόσβαση είναι ελεύθερη και χωρίς περιορισμούς για κάθε ενδιαφερόμενο εγγεγραμμένο ή μη χρήστη.



Εικόνα 8.1.1: Περιβάλλον GitHub στο οποίο φιλοξενείται ο κώδικας

Ο κώδικας έχει τροποποιηθεί κατάλληλα ώστε να λαμβάνει δυο πιθανές μορφές, συνάρτησης (Python Function) και κλάσης (Python Class). Και στις δύο περιπτώσεις τα απαιτούμενα δεδομένα που πρέπει να δώσει ο χρήστης για την εφαρμογή τους παρουσιάζονται στον πίνακα 8.1.1.

Πίνακας 8.1.1: Απαιτούμενα δεδομένα προς ορισμό από τον χρήστη για εφαρμογή της μεθοδολογίας και πρόβλεψη των δειγμάτων ελέγχου

variable	details	Input form	default
dataset	datafile με path όπως "C:/Users/marianna/spherical/dataset/ dataset.csv"	File path	
endpoint_col	Ο αριθμός της στήλης του dataset που περιλαμβάνει το endpoint	number	
feat_sel	Διαδικασία επιλογής μεταβλητών με δυνατότητα επιλογής μεταξύ True or False	True or False	*False*
feat_num	Αριθμός των βέλτιστων μεταβλητών μόνο αν feat_sel=True	number	*len(dataset.columns)/3*
feat_func	Μέθοδος επιλογής μεταβλητών μεταξύ των chi2, f_classif,mutual_info_classif	name	*chi2*
solver_name	Επιλύτης που χρησιμοποιείται για το πρόβλημα γραμμικού προγραμματισμού κατά την επιλογή των cluster[7]	name (string)	*'cbc'*
t_ratio	Αναλογία training και test set κατά την διαμόρφωση του μοντέλου	number	*0.7*
pred	Μέθοδος κατηγοριοποίησης των νέων δειγμάτων μεταξύ των τριών διαθέσιμων, original [1], alternative_1 [2] και alternative_2 [3]	list	*[1,2,3]*
print_report	Εμφάνιση των στατιστικών αποτελεσμάτων στο test set	True or False	*True*

Για την ορθή εισαγωγή του συνόλου δεδομένων, το αρχείο πρέπει να είναι σε μορφή CSV κατάλληλα επεξεργασμένο σε πίνακα, χωρίς να έχει απομακρυνθεί η στήλη με τα ονόματα των δειγμάτων. Η μορφή του εισαγόμενου συνόλου είναι προκαθορισμένη. Ο πίνακας πρέπει να απαρτίζεται από στήλες που περιέχουν τις

[7] Πληροφορίες για πιθανούς επιλύτες μπορούν να αντληθούν από τον ακόλουθο σύνδεσμο: https://docs.python-mip.com/en/latest/intro.html

	A	B	C	D	E	F	G
1	Material ID	Toxicity	Magnetic core	Overall size (nm)	Relaxivity (s-1/mmol Fe)(r2)	B0 (T)	Fe/cell (pg)
2	0	1	1	15	301.11	1.5	64.51
3	1	2	1	10	549	4.7	29.3
4	2	2	1	2	509	4.7	21.1
5	3	2	1	10	492	4.7	24.5
6	4	2	1	10	89	4.7	23.2
7	5	1	0	54.7	345	3	7.1
8	6	1	0	150	343.1	1.5	50.02
9	8	1	0	55.4	160.5	1.5	26.7
10	9	1	0	94	193.1	7	65
11	10	1	0	75	454.5	3	1459
12	11	1	1	77.8	27.26	0.5	36.9
13	12	1	0	90.13	193.1	7	69.6
14	13	2	1	6	140.4	0.5	51.7
15	14	1	0	8.5	43.5	3	68.7
16	15	1	0	65.2	73.4	0.47	2.15

Εικόνα 8.1.2: Ενδεικτική μορφή του αρχείου .csv του συνόλου SPIONs

ανεξάρτητες μεταβλητές και τη μεταβλητή πρόβλεψης, καθώς και γραμμές, μια για κάθε δείγμα (βλέπε Εικόνα 8.1.2). Ο πίνακας δεν πρέπει να περιλαμβάνει κενά κελιά, χαρακτήρες, infs ή NaNs. Η στήλη της μεταβλητής πρόβλεψης πρέπει να περιέχει μόνο αριθμούς: 1 για τα μη τοξικά δείγματα και 2 για τα τοξικά.

```
data_file = #"C:/Users... dataset file directory
y_predicted=spherical(data_file,
                      endpoint_col=1,
                      t_ratio=0.75,
                      feat_sel=False,
                      feat_num=4,
                      feat_func=chi2,
                      print_report=False,
                      solver_name='grb',
                      pred=[1,2,3])
```

```
train_file = #"C:/Users... training dataset file directory
test_file = #"C:/Users... test dataset file directory
model=spher_model(train_file,
                  test_file,
                  1,
                  True,
                  4,
                  chi2,
                  'grb',
                  [1,2,3])
clusters=model.fit()
predictions=model.predict(clusters)
```

Εικόνα 9.1.3: Απόσπασμα κώδικα για την εφαρμογή της μεθόδου ως function (πάνω) και ως class (κάτω)

Για την εφαρμογή της μεθοδολογίας, ο χρήστης πρέπει να μεταφέρει τον κώδικα του αρχείου spherical function.ipynb που βρίσκεται στο αποθετήριο του GitHub στην πλατφόρμα εκτέλεσης κώδικα Python που χρησιμοποιεί, τροποποιώντας κατάλληλα το input του function, βάσει των απαιτήσεών του, όπως φαίνεται στην

Εικόνα 8.1.3. Ως έξοδος, λαμβάνονται 1, 2 ή 3 λίστες, ανάλογα τις επιθυμητές προβλέψεις του χρήστη (ανά μέθοδο τελικής κατηγοριοποίησης), κάθε μια εκ των οποίων περιλαμβάνει την προβλεπόμενη κλάση καθενός εκ των δειγμάτων του συνόλου επικύρωσης σε δυαδική μορφή (τιμή 1 για τα μη τοξικά δείγματα και 2 για τα τοξικά).

Αν έναντι της εφαρμογής function, ο χρήστης επιθυμεί τη δημιουργία κλάσης, αυτό καθίσταται δυνατό μέσω του κώδικα του αρχείου spherical class.ipynb. Σε αυτή την περίπτωση, τα σύνολα εκπαίδευσης και ελέγχου εισάγονται ξεχωριστά, ακολουθώντας την τυποποίηση της μορφής όπως αυτή περιεγράφηκε παραπάνω. Ωστόσο, από το σύνολο επικύρωσης, θα πρέπει να απουσιάζει η στήλη που περιλαμβάνει τη μεταβλητή πρόβλεψης. Από τα απαιτούμενα δεδομένα απουσιάζει το train_ratio καθώς τα δύο σύνολα είναι ήδη καθορισμένα.

Μετά την εισαγωγή των δεδομένων και τη διαμόρφωση του μοντέλου, με την ιδιότητα .fit(), δημιουργούνται τα clusters βάσει των δειγμάτων του συνόλου εκπαίδευσης ενώ με την ιδιότητα .predict(clusters), πραγματοποιούνται οι προβλέψεις των τριών μεθοδολογιών για το σύνολο επικύρωσης. Η χρήση της κλάσης ενδείκνυται τόσο για την ανάπτυξη ενός μοντέλου και την αξιολόγησή του μέσω του συνόλου επικύρωσης, όσο και για την χρήση του μοντέλου σε άγνωστα δείγματα.

Παράρτημα

Ανάπτυξη λογισμικού

Γλώσσα προγραμματισμού Python

Η μεθοδολογία που αναπτύχθηκε κωδικοποιήθηκε στην γλώσσα προγραμματισμού Python (έκδοση 3.8.5, [38]). Το συγκεκριμένο εργαλείο επιλέχθηκε λόγω της ευκολίας εκμάθησης και αναγνωσιμότητάς του. Όντας μια αντικειμενοστραφής γλώσσα προγραμματισμού, η Python, έχει μεγάλη ευελιξία και εύρος δυνατοτήτων ως προς την επίλυση σύνθετων αλγοριθμικών προβλημάτων. Ταυτόχρονα, ωστόσο, είναι γλώσσα ανοικτού κώδικα, εύκολα προσβάσιμη από το χρήστη και χωρίς μεγάλες απαιτήσεις σε υπολογιστική ισχύ του συστήματος στο οποίο αναπτύσσεται.

Η ικανότητά της να συνδυάζεται με άλλα στοιχεία λογισμικού και να ενσωματώνει διαφορετικά υπολογιστικά εργαλεία στον κώδικα συμβάλλει στην ευρεία χρήση της στο πεδίο της Επιστήμης των Δεδομένων (Data Science) και της Μηχανικής Μάθησης (Machine Learning). Συγκεκριμένα, υπάρχει διαθέσιμη πληθώρα βιβλιοθηκών που προσφέρουν υπολογιστικές λύσεις στη διαχείριση και επεξεργασία των δεδομένων, μερικές από τις οποίες αξιοποιούνται στο πλαίσιο ανάπτυξης της συγκεκριμένης μεθοδολογίας και περιγράφονται συνοπτικά παρακάτω.

Για την ανάπτυξη του κώδικα χρησιμοποιήθηκε το ελεύθερο λογισμικό Anaconda (έκδοση 4.9.2, [39]),το οποίο λειτουργεί ως διανομή της Python, προσφέροντας πλήρη πρόσβαση στα εργαλεία και τις βιβλιοθήκες της. Η επεξεργασία του πραγματοποιήθηκε στο διαδικτυακό υπολογιστικό περιβάλλον Jupyter Notebooks, το οποίο υποστηρίζεται μέσω του Anaconda. Τα Jupyter Notebooks (έκδοση 6.1.7, [40]) επιλέχθηκαν ως εργαλείο λόγω του φιλικού γραφικού περιβάλλοντος (interface) και της δυνατότητάς τους να διαχειρίζονται και να εκτελούν κομμάτια του κώδικα ξεχωριστά.

Βιβλιοθήκες

NumPy

Βασικό εργαλείο για την διαχείριση συνόλων δεδομένων στην Python αποτελεί η βιβλιοθήκη NumPy (έκδοση 1.19.2, [41]). Προσφέροντας υποστήριξη για

διαχείριση μεγάλων, πολυδιάστατων πινάκων και διανυσμάτων, διευκολύνει τις μαθηματικές πράξεις μεταξύ αυτών και επιλύει γρήγορα προβλήματα στο πεδίο του Data Science. Στην περίπτωση γραφικής απεικόνισης των δεδομένων προσφέρεται συμπληρωματικά και η βιβλιοθήκη Matplotlib (έκδοση 3.3.2, [42]), η οποία προσεγγίζει τα εργαλεία απεικόνισης του MATLAB ως προς τις δυνατότητές και τις λειτουργίες.

Για την ενσωμάτωση της βιβλιοθήκης στον κώδικα, απαιτείται να πραγματοποιηθεί εισαγωγή της μέσω του περιβάλλοντος των Jupyter Notebooks (βλέπε Απόσπασμα κώδικα 1):

```
import numpy as np
```

Απόσπασμα κώδικα 1: Εισαγωγή βιβλιοθήκης NumPy

Pandas

Παρόμοιο σκοπό επιτελεί και η χρήση της βιβλιοθήκης Pandas (έκδοση 1.1.3, [43]), η οποία προσφέρει ακόμα περισσότερες λειτουργίες στην ανάλυση των δεδομένων και την φιλική προς τον χρήστη παρουσίασή τους. Στο πλαίσιο του Data Analysis, αξιοποιούνται τα dataframes, δομές σε μορφή αριθμημένων πινάκων που υποστηρίζουν εισαγωγή δεδομένων από πληθώρα τύπων αρχείου (όπως CSV, XLS, JSON).

Και σε αυτή την περίπτωση, για την ενσωμάτωση της βιβλιοθήκης στον κώδικα, απαιτείται η εισαγωγή της μέσω του περιβάλλοντος των Jupyter Notebooks (βλέπε Απόσπασμα κώδικα 2):

```
import pandas as pd
```

Απόσπασμα κώδικα 2: Εισαγωγή βιβλιοθήκης Pandas

scikit-learn

Για την εις βάθος διείσδυση σε προβλήματα μηχανικής μάθησης, η Python προσφέρει τη βιβλιοθήκη ανοικτού κώδικα scikit-learn (ή sklearn, έκδοση 0.23.2, [44]).

Αυτή περιλαμβάνει πρόσβασης σε εργαλεία κατηγοριοποίησης, παλινδρόμησης και συσταδοποίησης όπως k-means clustering ή random forests, όντας σχεδιασμένη να λειτουργεί παράλληλα με συμπληρωματικές βιβλιοθήκες ανάλυσης δεδομένων της Python όπως οι προαναφερθείσες Numpy και Pandas.

Στο πλαίσιο ανάπτυξης της μεθοδολογίας, αξιοποιούνται ιδιαιτέρως τα εργαλεία feature_selection, model_selection και metrics.

Ομοίως με τις υπόλοιπες βιβλιοθήκες απαιτείται η εισαγωγή της μέσω των Jupyter Notebooks.

MIP

Η εφαρμογή της προτεινόμενης μεθοδολογίας απαιτεί την επίλυση προβλημάτων μικτού ακέραιου γραμμικού προγραμματισμού με χρήση δυαδικών μεταβλητών.

Βασικό εργαλείο στην μοντελοποίηση τέτοιων προβλημάτων αποτελεί το πακέτο MIP (Mixed Integer Programming , έκδοση 1.13.0, [45]) το οποίο διευκολύνει την αριθμητική επίλυση με γρήγορη πρόσβαση σε ποικιλία συμβατών επιλυτών (solvers) όπως ο Gurobi (έκδοση 9.0.3 , [46]) και άμεση παραμετροποίησή τους.

Βιβλιογραφία

[1] H. Bahadar, F. Maqbool, K. Niaz, and M. Abdollahi, "Toxicity of nanoparticles and an overview of current experimental models," *Iran. Biomed. J.*, vol. 20, no. 1, pp. 1–11, 2016, doi: 10.7508/ibj.2016.01.001.

[2] C. Buzea, I. I. Pacheco, and K. Robbie, "Nanomaterials and nanoparticles: Sources and toxicity," *Biointerphases*, vol. 2, no. 4, pp. MR17–MR71, 2007, doi: 10.1116/1.2815690.

[3] T. A. Saleh, "Nanomaterials: Classification, properties, and environmental toxicities," *Environ. Technol. Innov.*, vol. 20, p. 101067, 2020, doi: 10.1016/j.eti.2020.101067.

[4] I. Khan, K. Saeed, and I. Khan, "Nanoparticles: Properties, applications and toxicities," *Arab. J. Chem.*, vol. 12, no. 7, pp. 908–931, 2019, doi: 10.1016/j.arabjc.2017.05.011.

[5] D. Chimene, D. L. Alge, and A. K. Gaharwar, "Two-Dimensional Nanomaterials for Biomedical Applications: Emerging Trends and Future Prospects," *Adv. Mater.*, vol. 27, no. 45, pp. 7261–7284, 2015, doi: 10.1002/adma.201502422.

[6] F. Barbero *et al.*, "Formation of the Protein Corona: The Interface between Nanoparticles and the Immune System," *Semin. Immunol.*, vol. 34, no. July, pp. 52–60, 2017, doi: 10.1016/j.smim.2017.10.001.

[7] B. Kharazian, N. L. Hadipour, and M. R. Ejtehadi, "Understanding the nanoparticle-protein corona complexes using computational and experimental methods," *Int. J. Biochem. Cell Biol.*, vol. 75, pp. 162–174, 2016, doi: 10.1016/j.biocel.2016.02.008.

[8] Y. W. Huang, M. Cambre, and H. J. Lee, "The Toxicity of Nanoparticles Depends on Multiple Molecular and Physicochemical Mechanisms," *Int. J. Mol. Sci.*, vol. 18, no. 12, 2017, doi: 10.3390/ijms18122702.

[9] A. B. Raies and V. B. Bajic, "In silico toxicology: computational methods for the prediction of chemical toxicity," *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, vol. 6, no. April, pp. 147–172, 2016, doi: 10.1002/wcms.1240.

[10] S. Panneerselvam and S. Choi, "Nanoinformatics: Emerging databases and available tools," *Int. J. Mol. Sci.*, vol. 15, no. 5, pp. 7158–7182, 2014, doi: 10.3390/ijms15057158.

[11] B. Saini and S. Srivastava, "Nanotoxicity prediction using computational modelling - Review and future directions," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 348, no. 1, pp. 0–9, 2018, doi: 10.1088/1757-899X/348/1/012005.

[12] S. E. Escher *et al.*, "Towards grouping concepts based on new approach methodologies in chemical hazard assessment: the read-across approach of the EU-ToxRisk project," *Arch. Toxicol.*, vol. 93, no. 12, pp. 3643–3667, 2019, doi: 10.1007/s00204-019-02591-7.

[13] A. Gajewicz, K. Jagiello, M. T. D. Cronin, J. Leszczynski, and T. Puzyn, "Addressing a bottle neck for regulation of nanomaterials: quantitative read-across (Nano-QRA) algorithm for cases when only limited data is available," *Environ. Sci. Nano*, vol. 4, no. 2, pp. 346–358, 2017, doi: 10.1039/c6en00399k.

[14] OECD, "GUIDANCE ON GROUPING OF CHEMICALS," *Organ. Econ. Co-operation Dev.*, vol. 33, no. November 2007, pp. 1–16, 2009, [Online]. Available: http://www.oecd.org/officialdocuments/displaydocumentpdf?cote=env/jm/mono(2010)46&doclanguage=en.

[15] T. W. Schultz *et al.*, "A strategy for structuring and reporting a read-across prediction of toxicity," *Regul. Toxicol. Pharmacol.*, vol. 72, no. 3, pp. 586–601, 2015, doi: 10.1016/j.yrtph.2015.05.016.

[16] N. M. Nawi, W. H. Atomi, and M. Z. Rehman, "The Effect of Data Pre-processing on Optimized Training of Artificial Neural Networks," *Procedia Technol.*, vol. 11, no. Iceei, pp. 32–39, 2013, doi: 10.1016/j.protcy.2013.12.159.

[17] K. Jajuga and M. Walesiak, "Standardisation of Data Set under Different Measurement Scales," no. January, pp. 105–112, 2000, doi: 10.1007/978-3-642-57280-7_11.

[18] S. B. Kotsiantis and D. Kanellopoulos, "Data preprocessing for supervised leaning," *Int. J. …*, vol. 1, no. 2, pp. 1–7, 2006, doi: 10.1080/02331931003692557.

[19] C. K. Yoo and M. Shahlaei, "The applications of PCA in QSAR studies: A case study on CCR5 antagonists," *Chem. Biol. Drug Des.*, vol. 91, no. 1, pp. 137–152, 2018, doi: 10.1111/cbdd.13064.

[20] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," *Int. Jt. Conf. Artif. Intell.*, no. June, 1995.

[21] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tool and Techniques*, vol. 53, no. 9. 1981.

[22] C. Patino and J. Ferreira, "Research methods knowledge base: Qualitative validity," *J Bras Pneumol.*, vol. 44, no. 3, p. 183, 2018.

[23] "Cross-validation: evaluating estimator performance." https://scikit-learn.org/stable/modules/cross_validation.html (accessed Mar. 18, 2021).

[24] M. Daszykowski, B. Walczak, and D. L. Massart, "Representative subset selection," *Anal. Chim. Acta*, vol. 468, no. 1, pp. 91–103, 2002, doi: 10.1016/S0003-2670(02)00651-7.

[25] E. W. Steyerberg, S. E. Bleeker, H. A. Moll, D. E. Grobbee, and K. G. M. Moons, "Internal and external validation of predictive models: A simulation study of bias and precision in small samples," *J. Clin. Epidemiol.*, vol. 56, no. 5, pp. 441–447, 2003, doi: 10.1016/S0895-4356(03)00047-7.

[26] A. Saptoro, M. O. Tadé, and H. Vuthaluru, "A modified Kennard-Stone

algorithm for optimal division of data for developing artificial neural network models," *Chem. Prod. Process Model.*, vol. 7, no. 1, 2012, doi: 10.1515/1934-2659.1645.

[27] R. K. H. Galvão, M. C. U. Araujo, G. E. José, M. J. C. Pontes, E. C. Silva, and T. C. B. Saldanha, "A method for calibration and validation subset partitioning," *Talanta*, vol. 67, no. 4, pp. 736–740, 2005, doi: 10.1016/j.talanta.2005.03.025.

[28] "sklearn.metrics.confusion_matrix." scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html (accessed Mar. 18, 2021).

[29] H. M and S. M.N, "A Review on Evaluation Metrics for Data Classification Evaluations," *Int. J. Data Min. Knowl. Manag. Process*, vol. 5, no. 2, pp. 01–11, 2015, doi: 10.5121/ijdkp.2015.5201.

[30] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *BBA - Protein Struct.*, vol. 405, no. 2, pp. 442–451, 1975, doi: 10.1016/0005-2795(75)90109-9.

[31] "sklearn.metrics.matthews_corrcoef." https://scikit-learn.org/stable/modules/generated/sklearn.metrics.matthews_corrcoef.html (accessed Mar. 18, 2021).

[32] R. Kiralj and M. M. C. Ferreira, "Basic validation procedures for regression models in QSAR and QSPR studies: Theory and application," *J. Braz. Chem. Soc.*, vol. 20, no. 4, pp. 770–787, 2009, doi: 10.1590/S0103-50532009000400021.

[33] C. Rucker, G. Rucker, and M. Meringer, "Y-Randomization and its Variants in QSPR/QSAR," *J. Chem. Inf. Model.*, vol. 47, no. 6, pp. 1–42, 2007, doi: 10.1021/ci700157b.

[34] Y. Xu and R. Goodacre, "On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning," *J. Anal. Test.*, vol. 2, no. 3, pp. 249–262, 2018, doi: 10.1007/s41664-018-0068-2.

[35] J. A. Stendal, M. Bambach, M. Eisentraut, I. Sizova, and S. Weiß, "Applying machine learning to the phenomenological flow stress modeling of TNM-B1," *Metals (Basel).*, vol. 9, no. 2, pp. 1–18, 2019, doi: 10.3390/met9020220.

[36] K. Roy, S. Kar, and P. Ambure, "On a simple approach for determining applicability domain of QSAR models," *Chemom. Intell. Lab. Syst.*, vol. 145, pp. 22–29, 2015, doi: 10.1016/j.chemolab.2015.04.013.

[37] M. I. Kotzabasaki, I. Sotiropoulos, and H. Sarimveis, "QSAR modeling of the toxicity classification of superparamagnetic iron oxide nanoparticles (SPIONs) in stem-cell monitoring applications: An integrated study from data curation to model development," *RSC Adv.*, vol. 10, no. 9, pp. 5385–5391, 2020, doi: 10.1039/c9ra09475j.

[38] J. S. Schwarz, C. Chapman, E. M. Feit, J. S. Schwarz, C. Chapman, and E. McDonnell Feit, "Welcome to Python," *Python for Marketing Research and Analytics*, 2020. https://www.python.org/ (accessed Mar. 18, 2021).

[39] "Anaconda - The World's Most Popular Data Science Platform," 2013. www.anaconda.com (accessed Mar. 18, 2021).

[40] T. Kluyver *et al.*, "Project Jupyter | Home," *Jupyter Notebooks -- a publishing format for reproducible computational workflows*, 2016. https://jupyter.org/ (accessed Mar. 18, 2021).

[41] "NumPy - The fundamental package for scientific computing with Python." www.numpy.org (accessed Mar. 18, 2021).

[42] John D. Hunter, "Matplotlib: Python plotting," *Matplotlib*, 2003. www.matplotlib.org (accessed Mar. 18, 2021).

[43] W. (AQR) McKinney, "Pandas : a Python Data Analysis Library," *New York*, 2009. www.pandas.pydata.org (accessed Mar. 18, 2021).

[44] F. Pedregosa, G. Varoquaux, L. Buitinck, G. Louppe, O. Grisel, and A. Mueller, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, 2015. https://scikit-learn.org/stable/ (accessed Mar. 18, 2021).

[45] "Python MIP (Mixed-Integer Linear Programming) Tools - PyPI." https://www.python-mip.com/ (accessed Mar. 18, 2021).

[46] "Gurobi - The fastest solver - Gurobi." www.gurobi.com (accessed Mar. 18, 2021).

[47] V. Forest *et al.*, "Towards an alternative to nano-QSAR for nanoparticle toxicity ranking in case of small datasets," *J. Nanoparticle Res.*, vol. 21, no. 5, 2019, doi: 10.1007/s11051-019-4541-2.

[48] R. Liu *et al.*, "Development of structure-activity relationship for metal oxide nanoparticles," *Nanoscale*, vol. 5, no. 12, pp. 5644–5653, 2013, doi: 10.1039/c3nr01533e.

[49] A. G. Papadiamantis *et al.*, "Predicting cytotoxicity of metal oxide nanoparticles using isalos analytics platform," *Nanomaterials*, vol. 10, no. 10, pp. 1–19, 2020, doi: 10.3390/nano10102017.

[50] A. M. Marlena S. Fejzoa Frederic Paik Schoenbergb, Kimber MacGibbonc, Patrick Mullind, Roberto Romeroe, f, and Khalil Tabsha aUniversity, "Use of Metal Oxide Nanoparticle Band Gap to Develop a Predictive Paradigm for Oxidative Stress and Acute Pulmonary Inflammation," *ACS Nano*, vol. 23, no. 1, pp. 1–7, 2008, doi: 10.1021/nn3010087.Use.

[51] Z. Ma and H. S. Ryoo, "Spherical Classification of Data, a New Rule-Based Learning Method," *J. Classif.*, 2020, doi: 10.1007/s00357-019-09355-z.