Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Χημικών Μηχανικών

Διπλωματική Εργασία

Ανάπτυξη μοντέλου μαθηματικού προγραμματισμού για την ομαδοποίηση υλικών και τον υπολογιστικό προσδιορισμό ανεπιθύμητων ιδιοτήτων

Νικολέττα Μαρία Κουτρούμπα

Επιβλέπων:

Καθηγητής ΕΜΠ Χαράλαμπος Σαρίμβεης

Αθήνα 2019

Διπλωματική εργασία
Ανάπτυξη μοντέλου μαθηματικού προγραμματισμού για την ομαδοποίηση υλικών και τον υπολογιστικό προσδιορισμό ανεπιθύμητων ιδιοτήτων
Σχολή Χημικών Μηχανικών, ΕΜΠ
Τομέας ΙΙ: Ανάλυσης, Σχεδιασμού και Ανάπτυξης Διεργασιών και Συστημάτων
Μονάδα Αυτόματης Ρύθμισης και Πληροφορικής
Επιβλέπων καθηγητής: Χαράλαμπος Σαρίμβεης

Νικολέττα Μαρία Κουτρούμπα

Στοιχεία επικοινωνίας:

Email: nikikoutroumpa@gmail.com

Περίληψη

Στα πλαίσια της παρούσας Διπλωματικής Εργασίας αναπτύχθηκε ένα μοντέλο μαθηματικού προγραμματισμού με στόχο την πρόβλεψη της τοξικότητας νανοσωματιδίων σύμφωνα με τη μεθοδολογία read-across. Η μεθοδολογία αυτή εντάσσεται στο πλαίσιο των μη πειραματικών τεχνικών με τις οποίες προβλέπεται η τοξικότητα και άλλες ανεπιθύμητες ιδιότητες νέων υλικών αποφεύγοντας τη χρήση πειραματόζωων. Η τεχνική βασίζεται στην εκτίμηση των ανεπιθύμητων ιδιοτήτων των άγνωστων υλικών χρησιμοποιώντας διαθέσιμα δεδομένα τοξικότητας από παρόμοια υλικά. Σύμφωνα με το διάγραμμα ροής εργασιών των μεθόδων read-across που προτάθηκε από τον Ευρωπαϊκό Οργανισμό Χημικών Προϊόντων (European Chemicals Agency, ECHA), ακολουθείται μία διαδικασία δοκιμής και σφάλματος των υποθέσεων ομαδοποίησης των υλικών, μέχρι να προσδιοριστεί εκείνη η υπόθεση που οδηγεί σε ακριβείς προβλέψεις. Ωστόσο, η διαδικασία αυτή είναι χρονοβόρα και δεν οδηγεί απαραίτητα σε βέλτιστα μοντέλα πρόβλεψης. Σκοπός της Διπλωματικής Εργασίας είναι η αυτοματοποίηση της διαδικασίας εύρεσης της υπόθεσης ομαδοποίησης. Το μοντέλο μαθηματικού προγραμματισμού που αναπτύχθηκε κατατάσσει τα νανοσωματίδια σε καθορισμένες περιοχές του πολυδιάστατου χώρου που ορίζεται από τις «μεταβλητές» (ιδιότητες) τους και προβλέπει την τοξικότητα τους μέσω ενός μοντέλου γραμμικής παλινδρόμησης μοναδικό για κάθε περιοχή. Με αυτό τον τρόπο είναι δυνατή η πρόβλεψη άγνωστων νανοσωματιδίων ανάλογα με την περιοχή που ανήκουν στον χώρο.

Τα δεδομένα που χρησιμοποιήθηκαν προέρχονται από τις δημοσιεύσεις των Gajewicz *et al*. (2015), Walkey *et al*. (2014), Xia *et al.* (2011) και Fourches *et al.* (2010) και αναφέρονται σε νανοσωματίδια μεταλλικών οξειδίων, νανοσωματίδια χρυσού, πολυεπίπεδους νανοσωλήνες άνθρακα και τροποποιημένα μεταλλικά νανοσωματίδια αντίστοιχα. Για τα τέσσερα σύνολα είναι γνωστές ορισμένες ιδιότητες και ένας δείκτης τοξικότητας.

Σε πρώτο βήμα, η ομαδοποίηση των νανοσωματιδίων σε περιοχές έγινε επιλέγοντας -μέσα από τη διαδικασία αριστοποίησης- μία ιδιότητα από το σύνολο των ανεξάρτητων μεταβλητών ως μεταβλητή διχοτόμησης («επίλυση σε μια διάσταση») και τα διαθέσιμα δείγματα χωρίστηκαν με βάση αυτή τη μεταβλητή σε δυο ή περισσότερες περιοχές. Η μεταβλητή που επιλέγεται κάθε φορά όπως και τα όρια διάσπασης συνιστούν την υπόθεση ομαδοποίησης. Και τα τέσσερα σύνολα δεδομένων στα οποία εφαρμόστηκε η μεθοδολογία οδήγησαν σε μοντέλα με αρκετά ακριβείς προβλέψεις.

Για τα σύνολα δεδομένων τα οποία διέθεταν διαφορετικά είδη ανεξάρτητων μεταβλητών, έγινε κατηγοριοποίηση αυτών ανάλογα με το είδος τους και οι περιοχές καθορίστηκαν από δυο μεταβλητές, μία για κάθε κατηγορία («επίλυση σε δύο διαστάσεις»). Το σύνολο των Gajewicz *et al*. (2015) που περιέχει κβαντομηχανικές και γεωμετρικές ιδιότητες και των Walkey *et al*. (2014) με φυσικοχημικές και βιολογικές ιδιότητες μελετήθηκαν με επίλυση σε δύο διαστάσεις και οδήγησαν σε μοντέλα με μεγαλύτερη ευαισθησία και ακρίβεια. Η αξιολόγηση των προβλέψεων έγινε χρήσει του δείκτη εξωτερικής ερμηνεύσιμης διακύμανσης Q^2_{test}, ο οποίος βελτιώθηκε και για τα δύο σύνολα δεδομένων με την επίλυση σε δύο διαστάσεις. Πιο συγκεκριμένα, για το πρώτο σύνολο η επίλυση σε μία διάσταση οδήγησε σε $Q^2_{test} = 0.65$ ενώ για διάσπαση σε δύο διαστάσεις προέκυψε $Q^2_{test} = 0.80$. Στο δεύτερο σύνολο αυξήθηκε η ακρίβεια των προβλέψεων από $Q^2_{test} = 0.86$ σε $Q^2_{test} = 0.93$.

Για την ανάπτυξη του μοντέλου βελτιστοποίησης και την ανάλυση των αποτελεσμάτων αναπτύχθηκε κώδικας σε γλώσσα MATLAB και χρησιμοποιήθηκε η εργαλειοθήκη YALMIP με την οποία συνδέθηκαν οι επιλύτες Mosek και Gurobi.

Από την παρούσα Εργασία προέκυψε η ανακοίνωση «*Read-across automated grouping and hazard endpoint predictions of nanoparticles based on mathematical optimization*» η οποία παρουσιάστηκε στα πλαίσια του επιστημονικού συνεδρίου 1st International Young Scientist Forum, το οποίο έλαβε χώρα στο Salzburg της Αυστρίας στις 9 και 10/09/2019, υπό την αιγίδα της Γερμανικής Εταιρείας Χημικών Συστημάτων, DECHEMA.

Λέξεις κλειδιά

Νανοπληροφορική, νανοσωματίδια, τοξικότητα, read-across, γραμμική παλινδρόμηση, μαθηματικός προγραμματισμός, αριστοποίηση

Abstract

Development of a Mathematical Programming Model for Material Grouping and Computational Estimation of their Adverse Effects

In this diploma thesis, a mathematical programming model is developed based on read-across methodology in order to predict toxicity related endpoints of nanoparticles. The read-across approach is an alternative, non-testing strategy that has been successfully used for the prediction of nanoparticles' toxicity. Its concept is based on the empirical knowledge that the estimation of the hazardous effects of untested chemicals can be achieved using the available data of similar chemicals. The European Chemicals Agency (ECHA) has presented a specific workflow for grouping and read-across methods that follows a trial-and-error process until the grouping hypothesis produce successful read-across predictions. However, it is time consuming and may not encounter the optimal read-across models. The main purpose of the present work is to automate the procedure of searching for the optimal grouping hypothesis. The developed mathematical programming model sorts the nanoparticles into regions and toxicities are predicted by a linear regression model that is unique to each region. Thus, non-tested nanoparticles' toxicity can be predicted pursuant to the region they belong.

Four datasets were considered for analysis, derived by Gajewicz *et al*. (2015), Walkey *et al*. (2014), Xia *et al.* (2011) and Fourches *et al.* (2010) which refer to metal oxide nanoparticles, gold nanoparticles, multiwalled carbon nanotubes and manufactured nanoparticles. These datasets also consist of several descriptors and a toxicity index.

Initially, the algorithm divides the domain into regions and groups the nanoparticles in these regions by selecting one feature of the available data that corresponds to the best model as the partition feature ("one-dimension problem"). The partition feature and the breakpoints resulting from the optimization problem form the optimal read-across grouping hypothesis. This methodology was applied in all four different datasets and produced accurate predictions.

Two of the datasets included different types of descriptors. For these, the descriptors were categorized into sets and the algorithm selected two partition features to define the regions; one of each descriptor set ("two-dimension problem"). The dataset by Gajewicz *et al*. (2015) included quantum-mechanical and image descriptors while the dataset by Walkey *et al*. (2014) included physicochemical and biological descriptors. The results of grouping the descriptors and solving the two-dimensional problem led to more accurate models. The reliability predictions of these models were validated using external explained variance Q^2_{test}, which was increased in comparison to the results from one-dimension problem. The external explained variance Q^2_{test} for the first dataset was increased from 0.65 to 0.80 while these values for the second dataset reached up to 0.93, whereas when solving the problem in one dimension, to 0.86.

The analysis code for the optimization problem was developed in MATLAB programming language. YALMIP toolbox and Mosek and Gurobi softwares were also used to solve the mathematical programming problem.

Results of this work were included in the publication entitled «*Read-across Automated Grouping and Hazard Endpoint Predictions of Nanoparticles based on Mathematical Optimization*» presented orally at the 1st International Young Scientist

Forum, that took place at Salzburg in Austria on 9th - 10th September 2019 under the auspices of the German Society for Chemical Apparatus, DECHEMA.

Key words

Nanoinformatics, nanoparticles, toxicity, read-across, linear regression, mathematical programming, optimization

Πίνακας περιεχομένων

Κατάλογος σχημάτων

Κατάλογος πινάκων

Κατάλογος διαγραμμάτων

Πρόλογος και ευχαριστίες

Η Διπλωματική Εργασία με τίτλο «*Ανάπτυξη μοντέλου μαθηματικού προγραμματισμού για την ομαδοποίηση υλικών και τον υπολογιστικό προσδιορισμό ανεπιθύμητων ιδιοτήτων*» εκπονήθηκε στη Μονάδα Αυτόματης Ρύθμισης και Πληροφορικής του Εθνικού Μετσόβιου Πολυτεχνείου, υπό την επίβλεψη του Καθηγητή ΕΜΠ Χαράλαμπου Σαρίμβει κατά το Ακαδημαϊκό Έτος 2018-19. Η εργασία αυτή σηματοδοτεί την ολοκλήρωση των σπουδών μου στη Σχολή Χημικών Μηχανικών και ως εκ τούτου θα ήθελα να ευχαριστήσω όσους συνέβαλαν στην επίτευξη αυτού του στόχου.

Θα ήθελα να ευχαριστήσω τον επιβλέποντά μου, καθηγητή Χ. Σαρίμβει για την ανάθεση ενός τόσο ενδιαφέροντος θέματος και την καθοδήγηση του καθ' όλη τη διάρκεια εκπόνησης αυτής της εργασίας. Ευχαριστώ θερμά την Δήμητρα Δανάη Βάρσου, υποψήφια διδάκτορα της Μονάδας Αυτόματης Ρύθμισης και Πληροφορικής, για την καθημερινή βοήθεια που μου πρόσφερε και την πολύτιμη υποστήριξή της καθ' όλη τη διάρκεια της διεξαγωγής και συγγραφής της διπλωματικής μου εργασίας.

Ακόμη, θα ήθελα να ευχαριστήσω τα μέλη της τριμελούς εξεταστικής επιτροπής για τον χρόνο που διέθεσαν στην ανάγνωση της διπλωματικής μου εργασίας και την τιμή που μου έκαναν να συμμετάσχουν στην αξιολόγησή της.

Τέλος, θα ήθελα να εκφράσω την ευγνωμοσύνη μου στην οικογένεια μου και τους φίλους μου για την υποστήριξή τους σε κάθε βήμα της πορείας μου.

Εισαγωγή

Η Νανοτεχνολογία αποτελεί έναν ραγδαία εξελισσόμενο τομέα στον οποίο πρωταγωνιστούν υλικά σε νανοκλίμακα. Τα υλικά με δομές που έχουν τουλάχιστον μία διάσταση μεταξύ 1-100 nm χαρακτηρίζονται ως νανοϋλικά και αξιοποιούνται σε πολλές εφαρμογές και επιστημονικά πεδία. Οι μοναδικές ιδιότητες των νανοσωματιδίων σχετίζονται άμεσα με το μέγεθος τους και τα διαφοροποιούν από τα αντίστοιχα υλικά σε μακροκλίμακα λόγω κβαντικών και επιφανειακών φαινομένων. Οι ξεχωριστές φυσικοχημικές και ηλεκτρικές ιδιότητες των νανοσωματιδίων μονοπωλούν το ενδιαφέρον σε μία πληθώρα εφαρμογών στην ιατρική, την ηλεκτρονική, τη βιοτεχνολογία κ.α. Για παράδειγμα, λόγω του μικρού τους μεγέθους, μπορούν να εισέλθουν στον ανθρώπινο οργανισμό, να διασχίσουν διάφορα βιολογικά εμπόδια και να αλληλεπιδράσουν με αυτά, δυνατότητες οι οποίες αξιοποιούνται στον τομέα της Ιατρικής για τη διάγνωση και θεραπεία ασθενειών.

Αν και έχει σημειωθεί σημαντική εξέλιξη στην νανοεπιστήμη και τις εφαρμογές της, η κατανόηση των κινδύνων που εγκυμονεί η χρήση νανοσωματιδίων για την ανθρώπινη υγεία είναι σε πρώιμο στάδιο. Όταν τα νανοσωματίδια εισέρχονται στον ανθρώπινο οργανισμό, αλληλεπιδρούν με το βιολογικό περιβάλλον και προσελκύουν βιομόρια τα οποία προσδένονται στην επιφάνεια τους, μεταβάλλοντας τόσο τα βιομόρια όσο και τις επιφανειακές τους ιδιότητες με άγνωστο πολλές φορές τρόπο. Πρόσφατες έρευνες σχετικά με τη νανοτοξικότητα (nanotocixity) αναφέρουν πιθανές δυσμενείς επιπτώσεις της έκθεσης των ζωντανών οργανισμών σε νανοσωματίδια, συμπεριλαμβανομένου του οξειδωτικού στρες, καταστροφής του DNA, καταστροφής της κυτταρικής μεμβράνης, ακόμα και λύσης του κυττάρου. Επομένως, είναι απαραίτητη η ανάπτυξη μεθόδων χαρακτηρισμού των νανοσωματιδίων και εκτίμησης των επιπτώσεων της έκθεσης σε αυτά τόσο στο περιβάλλον όσο και στην ανθρώπινη υγεία.

Η μελέτη των τοξικών επιδράσεων όλων των κατηγοριών των νανοσωματιδίων οδηγεί στην ανάγκη περισσότερων πειραμάτων και δοκιμών τα οποία αυξάνουν το κόστος σε χρήματα και χρόνο και απαιτούν τη χρήση πειραματόζωων, θέτοντας δεοντολογικά ζητήματα. Τα τελευταία χρόνια ενθαρρύνεται η ανάπτυξη εναλλακτικών, μη δοκιμαστικών μεθόδων για την εκτίμηση της τοξικότητας των νανοσωματιδίων που περιλαμβάνουν *in vitro* και *in silico* τεχνικές, οι οποίες θα μειώσουν την ανάγκη δοκιμών σε πειραματόζωα και θα δώσουν επιπλέον πληροφορίες για τους μηχανισμούς τοξικότητας. Σύμφωνα με τον κανονισμό (Regulation (EC) No 1907/2006) REACH (Registration, Evaluation, Authorization and Restriction of Chemicals) του Ευρωπαϊκού Οργανισμού Χημικών Προϊόντων (European Chemicals Agency, ECHA) και της Ευρωπαϊκής Ένωσης, οι εταιρείες που παράγουν ή χρησιμοποιούν χημικές ουσίες θα πρέπει να αξιολογούν τον κίνδυνο τοξικότητας, δημιουργώντας μία βιβλιοθήκη δεδομένων τοξικότητας η οποία θα είναι προσβάσιμη από οποιονδήποτε ενδιαφερόμενο, προτρέποντας τη χρήση εναλλακτικών μεθόδων για την αξιολόγηση της τοξικότητας και μειώνοντας τις δοκιμές χημικών προϊόντων σε πειραματόζωα στο ελάχιστο δυνατό.

Ο πρόσφατα εξελισσόμενος τομέας της Νανοπληροφορικής (Nanoinformatics) περιλαμβάνει νέες, υπολογιστικές τεχνικές που προσφέρουν αξιόπιστες προβλέψεις της συμπεριφοράς των νανοσωματιδίων. Μία πετυχημένη προσέγγιση είναι η χρήση των μοντέλων [Q]SARs (Qualitative/Quantitative Structure-Activity Relationships), που εφαρμόστηκαν επιτυχώς στον τομέα της Χημειοπληροφορικής. Ωστόσο, τα μοντέλα

αυτά παρουσιάζουν αδυναμίες. Απαιτούν μεγάλο σύνολο δεδομένων για να εκπαιδευτούν ενώ τα δεδομένα για νανοσωματίδια είναι περιορισμένα και επιπλέον, θεωρούν έναν κοινό μηχανισμό τοξικότητας, παραδοχή η οποία δεν είναι αποδεκτή καθώς οι μηχανισμοί τοξικότητας ανά είδος νανοσωματιδίου μπορεί να ποικίλουν. Μία άλλη προσέγγιση αφορά την πρόβλεψη της τοξικότητας συγκεκριμένων νανοσωματιδίων από παρόμοια νανοσωματίδια για τα οποία έχουν ήδη πραγματοποιηθεί πειραματικές μελέτες και που αναμένεται να έχουν παρόμοιες φυσικοχημικές ιδιότητες (μεθοδολογία read-across).

Στα πλαίσια της παρούσας Διπλωματικής Εργασίας αναπτύχθηκε ένα μοντέλο μαθηματικού προγραμματισμού μέσω του οποίου επιτυγχάνεται η ομαδοποίηση των νανοσωματιδίων σε περιοχές και η πρόβλεψη της τοξικότητάς τους ανά περιοχή, σύμφωνα με τη μεθοδολογία read-across. Σε κάθε περιοχή, με βάση ένα σύνολο εκπαίδευσης με γνωστές ιδιότητες και τοξικότητα, αναπτύσσεται ένα μοντέλο γραμμικής παλινδρόμησης το οποίο στη συνέχεια δύναται να προβλέψει την τοξικότητα άλλων νανοσωματιδίων, με άγνωστη τιμή τοξικότητας. Στη διαδικασία ομαδοποίησης, ανάλογα με τα διαθέσιμα δεδομένα, μπορούν να ληφθούν υπόψιν πολλαπλά κριτήρια χαρακτηρισμού των νανοσωματιδίων, αυξάνοντας κατ' επέκταση την ευαισθησία της μεθόδου.

Κεφάλαιο 1

Νανοϋλικά και Νανοπληροφορική

Λόγω των μοναδικών φυσικοχημικών ιδιοτήτων τους, τα νανοϋλικά αξιοποιούνται σε μία πληθώρα εφαρμογών, στην ιατρική, την ηλεκτρονική, σε δομικές κατασκευές κ.α. Ωστόσο, η αλληλεπίδραση των νανοσωματιδίων με κύτταρα ή βιομόρια όταν αυτά εισέλθουν σε βιολογικό περιβάλλον, καθώς και οι τοξικές επιδράσεις που παρατηρούνται, αποτελούν ανασταλτικό παράγοντα στη χρήση τους. Κρίνεται αναγκαία η μελέτη της τοξικότητας των νανοσωματιδίων καθώς και η ανάπτυξη εναλλακτικών μεθόδων εκτίμησης τοξικότητας που θα μειώσουν τις δοκιμές σε πειραματόζωα και θα δώσουν επιπλέον πληροφορίες για τους μηχανισμούς τοξικότητας των νανοσωματιδίων.

1.1 Νανοϋλικά και βιολογικό περιβάλλον

Υλικά με δομές που έχουν τουλάχιστον μία διάσταση στο εύρος 1-100 nm χαρακτηρίζονται ως νανοϋλικά. Λόγω των διαστάσεων αυτών μεταξύ των ατόμων τους, αποκτούν ιδιαίτερες οπτικές, ηλεκτρομαγνητικές, καταλυτικές και μηχανικές ιδιότητες. Τα νανοϋλικά συμπεριφέρονται διαφορετικά από τα ίδια υλικά σε μακροκλίμακα λόγω επιφανειακών και κβαντικών φαινομένων. Τα επιφανειακά φαινόμενα εκδηλώνονται λόγω του αυξημένου ποσοστού ατόμων στην επιφάνεια των νανοσωματιδίων σε σύγκριση με το εσωτερικό τους. Αυτά τα φαινόμενα περιλαμβάνουν αυξημένη χημική δραστικότητα και μείωση του σημείου τήξεως των νανοσωματιδίων σε σύγκριση με το αντίστοιχο υλικό σε μακροκλίμακα. Τα νανοσωματίδια έχουν μεγάλο λόγο επιφάνειας προς όγκο, ως εκ τούτου έχουν μεγάλη διαθέσιμη επιφάνεια για χημικές αντιδράσεις σε σύγκριση με μικροσωματίδια ή σωματίδια μεγαλύτερου μεγέθους. Λόγω των μοναδικών ιδιοτήτων τους, αξιοποιούνται σε μία πληθώρα εφαρμογών, στην ιατρική και τη βιοτεχνολογία ως φορείς φαρμάκων, σε κατασκευές ως δομικά υλικά, καθώς με τη σύνδεση τους με άλλα υλικά προσδίδουν νέες επιθυμητές ιδιότητες σε αυτά ή στην ηλεκτρονική για την παραγωγή νέου ηλεκτρονικού εξοπλισμού.[1,2]

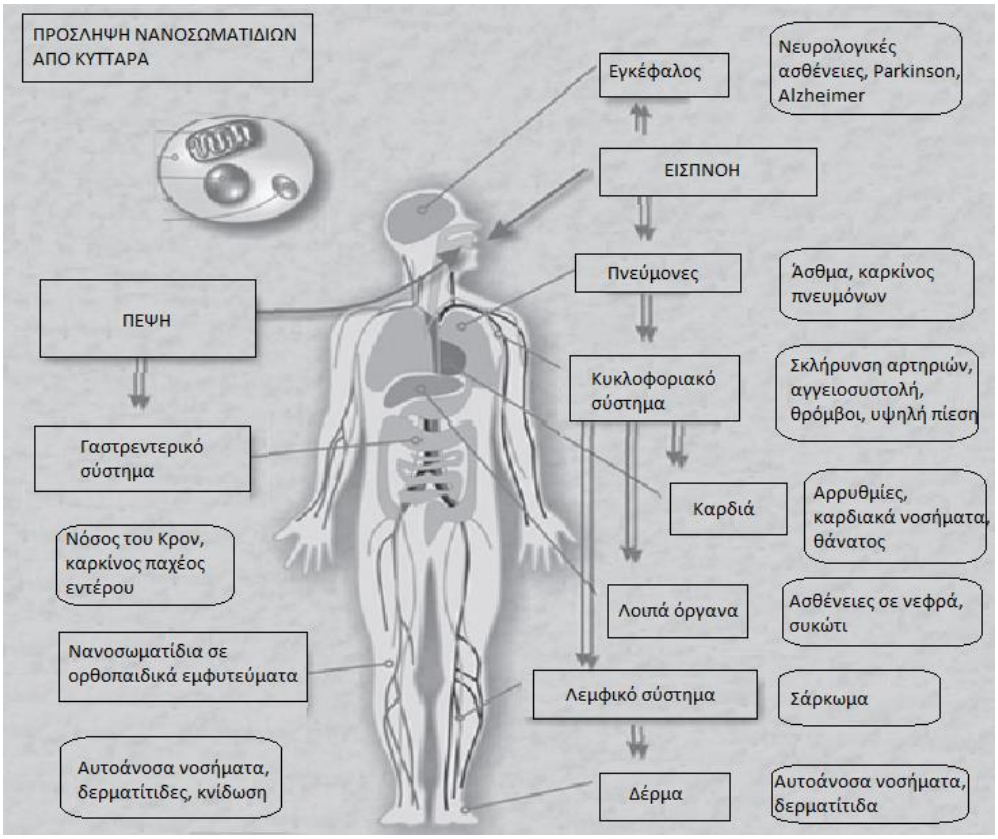
Τόσο τα φυσικά νανοσωματίδια, όπως τα σωματίδια που προέρχονται από ηφαιστειακές εκρήξεις, φυσική ή χημική διάβρωση πετρωμάτων ή ατελείς καύσεις, όσο και τα συνθετικά μπορεί να αποβούν τοξικά λόγω των αλληλεπιδράσεων τους με ενδοκυτταρικά οργανίδια, πρωτεΐνες ή και ολόκληρα γονίδια. Λόγω των διαστάσεων τους, τα νανοσωματίδια μπορούν να εισέλθουν στον ανθρώπινο οργανισμό με την εισπνοή, την κατάποση και την επαφή με το δέρμα. Η έκθεση των ανθρώπων σε αυτά μπορεί να έχει αρνητικές επιπτώσεις στην υγεία του οργανισμού (Σχήμα 1.1). Ο Πίνακας 1.1 συνοψίζει ορισμένες τοξικές δράσεις νανοσωματιδίων.[3,4]

Πίνακας 1.1: Νανοσωματίδια και τοξικότητα.[1]

Νανοσωματίδια	Εφαρμογές	Τοξικότητα
Οξειδίου του αργιλίου	Κυψελίδες καυσίμου Χρώματα Επικαλύψεις Βιοϋλικά	Διαταράσσουν τη βιωσιμότητα των κυττάρων Μεταβάλλουν τη λειτουργία των μιτοχονδρίων Προκαλούν οξειδωτικό στρες
Οξειδίου του χαλκού	Ημιαγωγοί Αντιμικροβιακά αντιδραστήρια	Διαταράσσουν την ακεραιότητα της κυτταρικής μεμβράνης Προκαλούν οξειδωτικό στρες
Αργύρου	Λόγω των αντιμικροβιακών ιδιοτήτων σε επικαλύψεις χειρουργικών εργαλείων και προσθετικά μέλη	Παραγωγή οξειδωτικών ριζών (ROS) Παραγωγή γαλακτικής αφυδρογονάσης (LDH)
Χρυσού	Φορείς φαρμάκων Αντιμετώπιση καρκίνου	Γενικά μη τοξικά, όμως η κυτταροτοξικότητα σχετίζεται με τη δόση που λαμβάνεται και το μέγεθος τους
Οξειδίου του σιδήρου	Φορείς φαρμάκων	Λύση κυττάρου Φλεγμονές Μεταβολή του συστήματος πήξης του αίματος
Πυριτίου	Φορείς φαρμάκων	Παραγωγή οξειδωτικών ριζών (ROS) Παραγωγή γαλακτικής αφυδρογονάσης (LDH) Παραγωγή μηλονικής διαλδεϋδης (MDA)

Οι τοξικολογικές επιπτώσεις των νανοϋλικών οφείλονται σε διακριτά φαινόμενα που συμβαίνουν κατά την είσοδο τους στον οργανισμό. Η παραγωγή δραστικών μορφών οξυγόνου (Reactive Oxygen Species, ROS), μέσα ή έξω από τα κύτταρα, είναι σημαντικός παράγοντας της τοξικολογικής συμπεριφοράς των νανοσωματιδίων. Τα νανοσωματίδια εισέρχονται στον οργανισμό και παράγουν δραστικά προϊόντα οξείδωσης. Η παραγωγή ελευθέρων ριζών, προκαλεί οξειδωτικό στρες, οδηγώντας σε βλάβες στο DNA ή τις πρωτεΐνες. Η διάλυση των σωματιδίων όταν εισέρχονται σε βιολογικό περιβάλλον οδηγεί στην απελευθέρωση τοξικών ιόντων που επηρεάζουν την λειτουργία των κυττάρων, προκαλώντας μηχανικές βλάβες σε κυτταρικά οργανίδια, στα λυσοσώματα, το ενδοπλασματικό δίκτυο ή τον πυρήνα. Μεγάλα σωματίδια μπορεί να προκαλέσουν μόνιμη βλάβη στην κυτταρική μεμβράνη ενώ μικρά σωματίδια μπορεί να διαπεράσουν την κυτταρική μεμβράνη και να βλάψουν το εσωτερικό των κυττάρων. Οι αλλαγές στην επιφανειακή ηλεκτρονιακή δομή των νανοσωματιδίων κατά την είσοδο τους στο βιολογικό περιβάλλον θα καθορίσει το επίπεδο αλληλεπίδρασης τους με αυτό. Ο σχηματισμός του πρωτεϊνικού στέμματος μπορεί να επιφέρει αρνητικές επιπτώσεις στις

πρωτεΐνες που συνδέονται στην επιφάνεια, με το ξεδίπλωμά τους, την απώλεια της δομής τους και τέλος την απώλεια της ενζυματικής τους δράσης.[2,5]



Σχήμα 1.1: Πρόσληψη νανοσωματιδίων και επιπτώσεις στον ανθρώπινο οργανισμό.[3]

Τα νανοσωματίδια υποβάλλονται σε τοξικολογικές μελέτες και μελετώνται ως προς την ανοσοτοξικότητα (immunotoxicity), κυτταροτοξικότητα (cytotoxicity) και γονιδιοτοξικότητα (genetoxicity) πριν από την χρήση τους στον ανθρώπινο οργανισμό. Οι τοξικές δράσεις των νανοσωματιδίων αξιολογούνται και συγκεντρώνονται δεδομένα τα οποία θα μπορούν να χρησιμοποιηθούν για την πρόβλεψη της τοξικότητας κάθε είδους νανοσωματιδίου. Ο σχεδιασμός, η προσαρμογή και η επικύρωση μοντέλων πρόβλεψης αποτελεί σκοπό των ερευνητών για την κάλυψη των κενών των γνώσεων που υπάρχουν σχετικά με την τοξικότητα των νανοσωματιδίων. Η κυτταρική συσχέτιση (cell association) αποτελεί έναν πρότυπο παράγοντα που χρησιμοποιείται για την πρόβλεψη της τοξικότητας, καθώς σύμφωνα με *in vivo* μελέτες σχετίζεται άμεσα με φλεγμονώδεις αποκρίσεις, βιοκατανομή και τοξικότητα.[1,6]

1.2 Μέτρηση ιδιοτήτων νανοσωματιδίων

Οι φυσικοχημικές ιδιότητες των νανοσωματιδίων καθώς και οι αλληλεπιδράσεις τους με βιομόρια όταν εισέρχονται σε βιολογικά συστήματα διερευνώνται με διάφορες μεθόδους ενόργανης ανάλυσης. Υπάρχουν αρκετές τεχνικές υπολογισμού των ιδιοτήτων των νανοσωματιδίων με τις πιο συνηθισμένες να αναφέρονται παρακάτω. Για το χαρακτηρισμό της δομής και την εύρεση του μέσου μεγέθους των νανοσωματιδίων χρησιμοποιείται η τεχνική της ηλεκτρονιακής μικροσκοπίας μετάδοσης (Transmission Electron Microscopy, TEM). Εναλλακτικές μέθοδοι είναι η ηλεκτρονική φασματοσκοπία απώλειας ενέργειας (Electron Energy Loss Spectrometry, EELS) ή η φασματοσκοπία

ενέργειας-διασποράς ακτινών Χ, (Energy-Dispersive-X-ray spectrometry, EDX). Με τη μέθοδο δυναμικής σκέδασης του φωτός (Dynamic Light Scattering, DLS) προσδιορίζεται η κινητικότητα των σωματιδίων και μετριέται η υδροδυναμική τους διάμετρος (Hydrodynamic Diameter, HD), που εκφράζει τη συμπεριφορά του νανοσωματιδίου μέσα σε υδατικό διάλυμα. Παρόμοιες μετρήσεις γίνονται και με την ανάλυση ανίχνευσης νανοσωματιδίων (Nanoparticle Tracking Analysis, NTA), η οποία δίνει καλύτερα αποτελέσματα από την DLS για πληθυσμούς σωματιδίων σε πολυ-διασκορπισμένα δείγματα καθώς μπορεί να γίνει διάκριση κάθε μεμονωμένου νανοσωματιδίου, σε αντίθεση με την DLS η οποία λαμβάνει μία μέτρηση για όλα τα σωματίδια ταυτόχρονα. Επιπλέον, χρησιμοποιούνται τεχνικές φασματοσκοπίας απορρόφησης (Absorbance Spectrophotometry, AS) για τη μέτρηση του δείκτη επιφανειακού συντονισμού πλασμονίων (Localized Surface Plasmon Resonance index, LSPRi), ο οποίος δίνει πληροφορίες για την αλληλεπίδραση των ηλεκτρονίων της επιφάνειας των νανοσωματιδίων όταν αυτά δέχονται ηλεκτρομαγνητική ακτινοβολία. Η μέτρηση του ζ-δυναμικού (Zeta Potential, ZP), αποτελεί ένδειξη του επιφανειακού φορτίου των νανοσωματιδίων όταν διασπείρονται σε πολικό μέσο και υπολογίζεται με ηλεκτροφόρηση σε πηκτή αγαρόζης (agarose gel electrophoresis), με τεχνικές σκέδασης φωτός (light scattering) ή με ρυθμιζόμενη ανίχνευση ανθεκτικών παλμών (Tunable Resistive Pulse Sensing, TRPS).

Όταν τα νανοσωματίδια εισέρχονται σε βιολογικό περιβάλλον παρουσιάζουν κολλοειδή συμπεριφορά και παρατηρούνται αλλαγές στην επιφάνεια τους. Τα συστατικά του βιολογικού μέσου τείνουν να προσροφηθούν στις επιφάνειες των νανοσωματιδίων, σχηματίζοντας ένα «πρωτεϊνικό στέμμα» (protein corona) όπως αναφέρεται στην Ενότητα 1.3. Η σύνθεση του πρωτεϊνικού στέμματος χαρακτηρίζεται ποιοτικά με τη μέθοδο ηλεκτροφόρησης σε πολυακριλαμίδιο (poly-acrylamide gel electrophoresis, PAGE) και ημι-ποσοτικά με συζευγμένη υγρή χρωματογραφία-φασματοσκοπία μάζας (LC-MS/MS). Άλλες τεχνικές αποτελούν η μειωμένη συνολική ανάκλαση-φασματοσκοπία υπέρυθρου μετασχηματισμού Fourier (Attenuated Total Reflectance Fourier Transform-Infrared, ATR-FT-IR), με την οποία μελετώνται οι αλληλεπιδράσεις νανοσωματιδίων με βιομόρια και ο κυκλικός διχρωισμός (Circular Dichroism, CD), όπου υπολογίζονται ποσοτικά οι πρωτεΐνες παρατηρώντας την απώλεια της δευτεροταγούς δομής τους όταν συνδέθηκαν στην επιφάνεια του νανοσωματιδίου.[6,7]

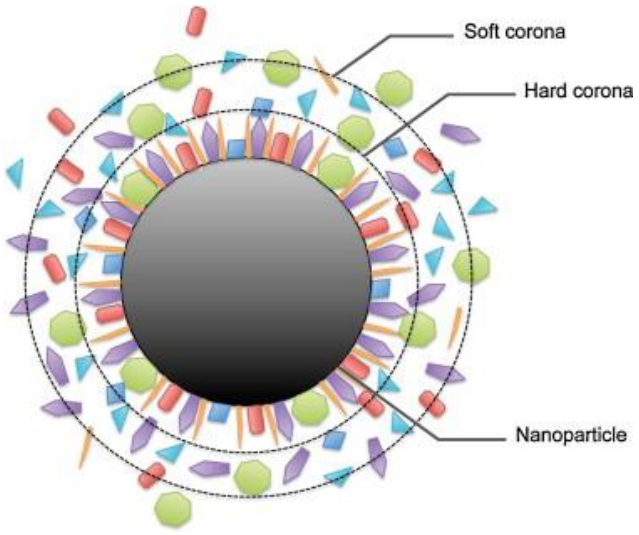
1.3 Πρωτεϊνικό στέμμα

Λόγω του μεγάλου λόγου επιφάνειας προς όγκο και της μεγάλης ελεύθερης ενέργειας στην επιφάνεια τους, όταν τα νανοσωματίδια εισέρχονται σε ένα βιολογικό περιβάλλον τείνουν να αλληλεπιδράσουν με βιομόρια που βρίσκονται σε αυτό, όπως πρωτεΐνες, νουκλεϊκά οξέα, λιπίδια, σάκχαρα ή μεταβολίτες. Μεγάλης σημασίας είναι η προσρόφηση ορισμένων πρωτεϊνών στην επιφάνεια ενός νανοσωματιδίου σχηματίζοντας μία δυναμική διεπιφάνεια γύρω από το σωματίδιο, που ονομάζεται «πρωτεϊνικό στέμμα». Η σύνδεση νανοσωματιδίου-πρωτεΐνης αποτελείται από τις εξής περιοχές: την επιφάνεια του σωματιδίου, τη διεπιφάνεια στερεού-υγρού και την περιοχή στέμματος-βιολογικού μέσου. Όταν τα νανοσωματίδια εισέρχονται σε υδατικά βιολογικά συστήματα, η επιφάνειά τους μεταβάλλεται λόγω της διαλυτοποίησης ή της προσρόφησης μικρών μορίων. Έτσι σχηματίζεται η διεπιφάνεια στερεού-υγρού που καθορίζει την συγγένεια

και την εκλεκτικότητα των βιομορίων που θα συνδεθούν στο νανοσωματίδιο και θα σχηματίσουν το πρωτεϊνικό στέμμα. Η σύνδεση των πρωτεϊνών οφείλεται σε διάφορους δεσμούς, όπως δυνάμεις Van der Waals, ηλεκτροστατικές δυνάμεις, υδρόφοβες αλληλεπιδράσεις, δεσμούς υδρογόνου ή δυνάμεις διαλυτοποίησης.[8,9]

Ο σχηματισμός του πρωτεϊνικού στέμματος προσδίδει μία νέα «βιολογική ταυτότητα» στο νανοσωματίδιο. Η αλληλεπίδραση νανοσωματιδίου-πρωτεΐνης ενδέχεται να μεταβάλλει το μέγεθος, το σχήμα, την επιφανειακή φόρτιση και την κατάσταση συσσωμάτωσης του νανοσωματιδίου. Τα νανοσωματίδια συνήθως αυξάνονται σε μέγεθος κατά 20-70 nm, υποδηλώνοντας ότι το στέμμα αποτελείται από πολλαπλές στρώσεις πρωτεϊνών. Αντίθετα, λιπιδικά νανοσωματίδια μειώνονται σε μέγεθος λόγω ωσμωτικών δυνάμεων που αναπτύσσονται καθώς η λιπιδική μεμβράνη είναι αδιαπέραστη σε πρωτεΐνες. Ο σχηματισμός πρωτεϊνικού στέμματος μπορεί να προκαλέσει συσσωμάτωση νανοσωματιδίων μέσω γεφυρών πρωτεΐνης ή να σταθεροποιήσει τα σωματίδια και να αποτρέψει την συσσωμάτωση λόγω των μεταβολών στο επιφανειακό φορτίο των σωματιδίων.[6,10]

Το πρωτεϊνικό στέμμα αποτελεί μια πολύπλοκη δομή πάχους περίπου 20-30 nm που αποτελείται από μία μαλακή και μία σκληρή στρώση, το «μαλακό» πρωτεϊνικό στέμμα (soft corona) και το «σκληρό» πρωτεϊνικό στέμμα (hard corona), όπως φαίνεται στο Σχήμα 1.2. Το «μαλακό» πρωτεϊνικό στέμμα αποτελείται από πρωτεΐνες με χαμηλής συγγένειας αλληλεπιδράσεις με την επιφάνεια του νανοσωματιδίου και χαρακτηρίζεται από συνεχή ανταλλαγή μακρομορίων μεταξύ περιβάλλοντος μέσου και επιφάνειας νανοσωματιδίου ενώ το «σκληρό» πρωτεϊνικό στέμμα χαρακτηρίζεται από αλληλεπιδράσεις υψηλής συγγένειας και μακρομόρια που συνδέονται στην επιφάνεια των νανοσωματιδίων με ισχυρούς δεσμούς, προκαλώντας μετουσίωση των πρωτεϊνών.[11,12]



Σχήμα 1.2: Σχηματισμός «μαλακού» και «σκληρού» πρωτεϊνικού στέμματος.[10]

Η σύνθεση του πρωτεϊνικού στέμματος ποικίλει ανάλογα με το υλικό, το σχήμα, το μέγεθος και την επιφανειακή φόρτιση του νανοσωματιδίου. Η προσρόφηση μίας πρωτεΐνης εξαρτάται επίσης από τη συγγένεια της προς την επιφάνεια του σωματιδίου και την ικανότητα της να συνδεθεί στην επιφάνειά του. Μόνο συγκεκριμένες ομάδες πρωτεϊνών με υψηλή συγγένεια ως προς την επιφάνεια του νανοσωματιδίου μπορούν να προσροφηθούν σε αυτή και να μείνουν συνδεδεμένες για μεγάλο χρονικό διάστημα. Για παράδειγμα, υδρόφοβα νανοσωματίδια (π.χ. νανοσωλήνες άνθρακα) προσελκύουν πρωτεΐνες με υδρόφοβα τμήματα, νανοσωματίδια με μεγάλο μέγεθος χαρακτηρίζονται

από αυξημένη προσρόφηση πρωτεϊνών λόγω της μειωμένης καμπυλότητας της επιφάνειας τους που επιτρέπει στις πρωτεΐνες να κινηθούν στην επιφάνεια και να αλληλεπιδράσουν με αυτή.[8,10–12]

Με την προσρόφηση πρωτεϊνών και τις αλλαγές σε ορισμένα χαρακτηριστικά των νανοσωματιδίων, ενδέχεται να μεταβληθεί η λειτουργικότητα ή η τοξικότητά τους. Είναι σημαντικό να αξιολογηθεί η ασφάλεια των νανοσωματιδίων έχοντας ως παράγοντα τον σχηματισμό πρωτεϊνικού στέμματος. Οι αλληλεπιδράσεις μεταξύ των συστατικών του βιολογικού μέσου και των νανοσωματιδίων μπορεί να οδηγήσουν στην εμφάνιση τοξικότητας. Η νέα «βιολογική ταυτότητα» του νανοσωματιδίου αποτελεί τη νέα του μορφή που επηρεάζεται από τα συστατικά του βιολογικού συστήματος με τα οποία συνδέεται. Η σύνδεση νανοσωματιδίου και πρωτεϊνών μπορεί να οδηγήσει σε αλλαγή της δομής των πρωτεϊνών με αποτέλεσμα το σχηματισμό πρωτεϊνών με λανθασμένες δευτεροταγείς και τριτοταγείς δομές, που είναι δυσλειτουργικές ή συχνά σχετίζονται με την εμφάνιση διαφόρων ασθενειών.

Αντίθετα, ο σχηματισμός πρωτεϊνικού στέμματος μπορεί να μετριάσει την κυτταροτοξικότητα των νανοσωματιδίων, καθώς κατά την αλληλεπίδραση νανοσωματιδίων με την κυτταρική μεμβράνη ενός κυττάρου, η πρωτεϊνική επικάλυψη μπορεί να αποτρέψει την καταστροφή του κυττάρου. Τέλος, οι πρωτεΐνες του στέμματος μπορεί να κωδικοποιούν πληροφορίες για την επαφή νανοσωματιδίου και κυττάρου αλλάζοντας έτσι την κυτταρική απάντηση από τα κύτταρα στόχους.[6,10]

1.4 Νανοπληροφορική

Οι μοναδικές ιδιότητες των νανοϋλικών έχουν αποδειχτεί χρήσιμες σε πολλές εφαρμογές. Ως βιοδείκτες (biomarkers) και παράγοντες αντίστροφης απεικόνισης (imaging contrast agents), τα νανοσωματίδια μπορούν να συνδεθούν με διαγνωστικούς παράγοντες που απευθύνονται σε συγκεκριμένους υποδοχείς, γεγονός που τους επιτρέπει την παρακολούθηση της κατανομής τους και τους δίνει τη δυνατότητα εφαρμογής ως βιοαισθητήρες. Η μεγάλη ελεύθερη επιφάνεια των νανοϋλικών, συνδεδεμένη με ειδικά σύμπλοκα (ligands), παρέχει υψηλή ευαισθησία για ανίχνευση μοριακών στόχων, όπως DNA, πρωτεϊνών, παθογόνων μικροοργανισμών, κυττάρων και ενζύμων. Ως συστήματα διανομής (delivery systems), τα νανοσωματίδια μπορούν να χρησιμοποιηθούν για την συστηματική απελευθέρωση φαρμάκων σε έναν συγκεκριμένο στόχο, αυξάνοντας την αποτελεσματικότητα του φαρμάκου και βελτιώνοντας την απόδοση της θεραπείας.[13]

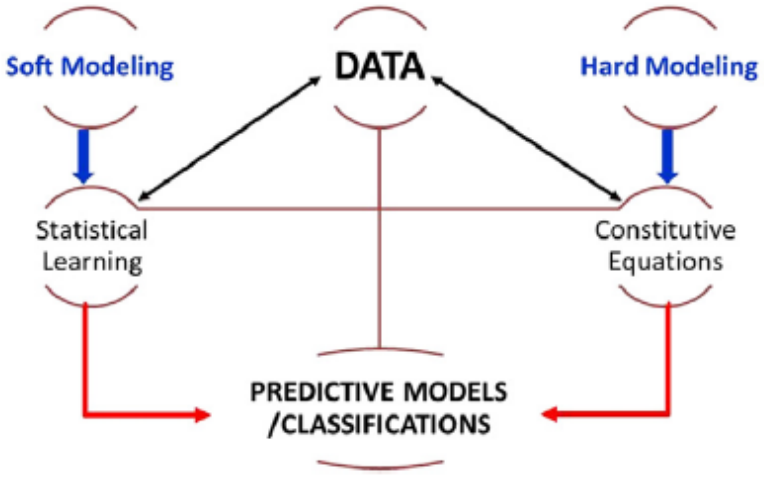
Οι πολύπλοκες και χρήσιμες εφαρμογές των νανοσωματιδίων οδήγησαν στην ανάγκη για επιπλέον κατανόηση των φυσικοχημικών ιδιοτήτων τους, των αλλαγών που υφίστανται αυτές οι ιδιότητες κατά τη εισαγωγή τους στο βιολογικό περιβάλλον καθώς και των επιπτώσεων στους οργανισμούς και στα φυσικά συστήματα (ατμοσφαιρικά, βιοχημικά), και κατ' επέκταση στην ανάγκη κατανόησης των γενικότερων κινδύνων των νανοϋλικών, σε όλη τη διάρκεια της ζωής τους, από την σύνθεση και την παραγωγή μέχρι την ανάκτηση και επαναχρησιμοποίησή τους.[5]

Ωστόσο, είναι απαραίτητη η ανάγκη μείωσης του χρόνου και του κόστους απόκτησης βασικών γνώσεων για τα νανοϋλικά, των αναλύσεων σε επιστημονικά πεδία και της κλινικής τους εφαρμογής. Η ανησυχία σχετικά με την τοξικότητα των νανοσωματιδίων ή νανοτοξικότητα, αποτελεί εμπόδιο στην εφαρμογή τους. Το πρόβλημα θα μπορούσε να αντιμετωπιστεί με την παροχή πληροφοριών μέσω ενός συστήματος ανταλλαγής και κατανομής διαθέσιμων τοξικών και φυσικοχημικών δεδομένων για τον αποτελεσματικό προσδιορισμό της σχέσης μεταξύ έκθεσης σε

νανοϋλικά και των παρενεργειών. Τα δεδομένα αυτά μπορούν να προσφέρουν χρήσιμες πληροφορίες για τα νανοσωματίδια μέσα από τον τομέα της Νανοπληροφορικής.[13]

Η Νανοπληροφορική είναι η επιστήμη και η πρακτική προσδιορισμού των πληροφοριών που σχετίζονται με την νανοτεχνολογία και ανάπτυξης και εφαρμογής αποτελεσματικών μηχανισμών συλλογής, επικύρωσης, αποθήκευσης, ανάλυσης και μοντελοποίησης αυτών των πληροφοριών. Είναι απαραίτητη για την ανάπτυξη και το χαρακτηρισμό των νανοϋλικών, για το σχεδιασμό και τη χρήση βελτιωμένων νανοδομών και νανοσυστημάτων και για την ανάπτυξη προηγμένων οργάνων και μεθόδων παραγωγής. Επιπλέον, η νανοπληροφορική προάγει την ανακάλυψη επιπλέον πληροφορίας για τα νανοϋλικά μέσω τεχνικών εξόρυξης δεδομένων και μηχανικής μάθησης.[14]

Η Νανοπληροφορική διαθέτει ένα ευρύ φάσμα εργαλείων που κυμαίνονται από συστηματικούς, συνδυαστικούς πειραματισμούς σε εξελιγμένα μοντέλα. Οι προσπάθειες μοντελοποίησης μπορούν να χωριστούν σε δύο κατηγορίες, όπως παρουσιάζονται στο Σχήμα 1.3: τη «βαριά μοντελοποίηση» (hard modeling) και την «ήπια μοντελοποίηση» (soft modeling). Η «βαριά μοντελοποίηση» περιλαμβάνει υπολογιστικές στρατηγικές διακριτοποίησης και παράλληλους αλγορίθμους. Μεταξύ αυτών των προσεγγίσεων είναι τα ατομικά μοντέλα, θερμοδυναμικά μοντέλα, προσομοίωση πεδίου φάσης και προσομοίωση πεπερασμένων στοιχείων σε επίπεδο μικροδομής. Η «ήπια μοντελοποίηση» σχετίζεται με προσεγγίσεις στατιστικής, ανεξάρτητες από το μοντέλο. Μεταξύ αυτών είναι η χρήση παλινδρόμησης, νευρωνικών δικτύων, γενετικών αλγορίθμων και αλγορίθμων ταξινόμησης. Η χρήση αυτών των τεχνικών αποσκοπεί στην ταξινόμηση της πληροφορίας και ανακάλυψη νέας, καθώς και στην ανάπτυξη αναλυτικών εργαλείων πρόβλεψης που βασίζονται σε μεθόδους στατιστικής μάθησης. Η «ήπια μοντελοποίηση» αποτελεί ένα ισχυρό μέσο για ανακάλυψη νέων συσχετίσεων και καθιστά εφικτό τον σχεδιασμό μοντέλων πρόβλεψης.[5]



Σχήμα 1.3: Μέθοδοι μοντελοποίησης στη Νανοπληροφορική.[5]

Η Νανοπληροφορική στοχεύει στην κάλυψη των κενών στα δεδομένα νανοτεχνολογίας ή στην εξόρυξη πληροφοριών από υπάρχοντα δεδομένα νανοτεχνολογίας. Η εξόρυξη πληροφοριών από τα σύνολα δεδομένων περιλαμβάνει τα ακόλουθα βήματα:[13]

i. Τη συλλογή δεδομένων
ii. Την ανάλυση δεδομένων και το σχολιασμό τους

iii. Την εξαγωγή πληροφοριών
iv. Την επικύρωση των δεδομένων

Μεγάλη δυσκολία παρατηρείται στη συλλογή δεδομένων, καθώς υπάρχουν μειωμένα πειραματικά αποτελέσματα και είναι δύσκολα προσβάσιμα. Επιπλέον, μερικά πειραματικά αποτελέσματα ενδεχομένως να περιέχουν αριθμητικά σφάλματα ή να μην έχουν περιγραφεί σωστά από άποψη πρωτοκόλλων, πειραματικών παραμέτρων και τεχνικών. Ως εκ τούτου, προκειμένου να εξαχθούν υψηλής ποιότητας πληροφορίες από τα σύνολα δεδομένων, είναι απαραίτητη η επικύρωσή τους και η διαγραφή μη έγκυρων ή περιττών δεδομένων.[13]

1.5 Μεθοδολογία read-across

Τα τελευταία χρόνια, ενθαρρύνεται η ανάπτυξη εναλλακτικών μεθόδων για την εκτίμηση της τοξικότητας των νανοσωματιδίων όπως οι μέθοδοι *in vitro* (σε ελεγχόμενες εργαστηριακές συνθήκες) και *in silico* (με τη χρήση υπολογιστή), οι οποίες θα μειώσουν την ανάγκη δοκιμών σε πειραματόζωα και θα δώσουν επιπλέον πληροφορίες για τους μηχανισμούς τοξικότητας. Όλες οι μέθοδοι έχουν κύριο σκοπό την αντιμετώπιση των προκλήσεων της χρήσης νανοσωματιδίων και την αξιολόγηση των κινδύνων με ακρίβεια και αποτελεσματικότητα.[15]

Στις πετυχημένες υπολογιστικές προσεγγίσεις συγκαταλέγονται και οι μέθοδοι της ποσοτικής/ποιοτικής σχέσης δομής-ιδιοτήτων ([Q]SARs), οι οποίες εφαρμόστηκαν τα προηγούμενα χρόνια επιτυχώς στο πεδίο της Χημειοπληροφορικής. Πρόκειται για μαθηματικές τεχνικές που συσχετίζουν τη δομή μίας ουσίας με την παρουσία ή την απουσία μίας ιδιότητας ή δραστηριότητας. Οι μέθοδοι [Q]SAR βασίζονται στην εξάρτηση μεταξύ της διακύμανσης των μοριακών δομών, που χαρακτηρίζονται από τους λεγόμενους «περιγραφείς» (descriptors) και της διακύμανσης της βιολογικής δραστηριότητας σε ένα σύνολο παρόμοιων χημικών ουσιών. Τα ειδικά διαμορφωμένα για τα νανοσωματίδια μοντέλα της μεθόδου [Q]SAR, χαρακτηρίζονται ως nano-[Q]SARs ή [Q]NARs (Quantitative Nanostructure-Activity Relationships). Ωστόσο, οι προσεγγίσεις [Q]NARs απαιτούν μεγάλα πειραματικά σύνολα δεδομένων. Το περιορισμένο μέγεθος των συνόλων δεδομένων για τα νανοσωματίδια αποτελεί ένα σημαντικό εμπόδιο στη χρήση τέτοιων τεχνικών, επομένως απαιτούνται εναλλακτικές τεχνικές που να επιτρέπουν την ανάπτυξη μοντέλων που να στηρίζονται σε περιορισμένα σύνολα δεδομένων, όπως οι μεθοδολογίες read-across.[15,16]

Οι μεθοδολογίες read-across αποτελούν εναλλακτικές μεθόδους συμπλήρωσης αγνώστων ιδιοτήτων των χημικών ουσιών και προέκυψαν πρωτίστως από τον κανονισμό REACH του Ευρωπαϊκού Οργανισμού Χημικών Προϊόντων (ECHA). Πιο συγκεκριμένα, πρόκειται για διαδικασίες εκτίμησης εξαρτημένων μεταβλητών, όπως η τοξικότητα, μίας άγνωστης χημικής ουσίας που καλείται «στόχος» (target chemical) με βάση τα αποτελέσματα που έχουν προκύψει από μία «πηγαία» ουσία η οποία είναι γνωστή, έχει ήδη μελετηθεί (source chemical) και είναι παρόμοια σε επίπεδο δομής και φυσικοχημικών ιδιοτήτων με την άγνωστη. Οι ομοιότητες μπορεί να βασίζονται σε κοινές λειτουργικές ομάδες, κοινά συστατικά ή πιθανότητα εμφάνισης κοινών πρόδρομων ουσιών. Επομένως, είναι δυνατή η χρήση της γνωστής ουσίας για την πρόβλεψη μίας ιδιότητας της άγνωστης ουσίας εάν έχουν παρόμοια χημική δομή.[17]

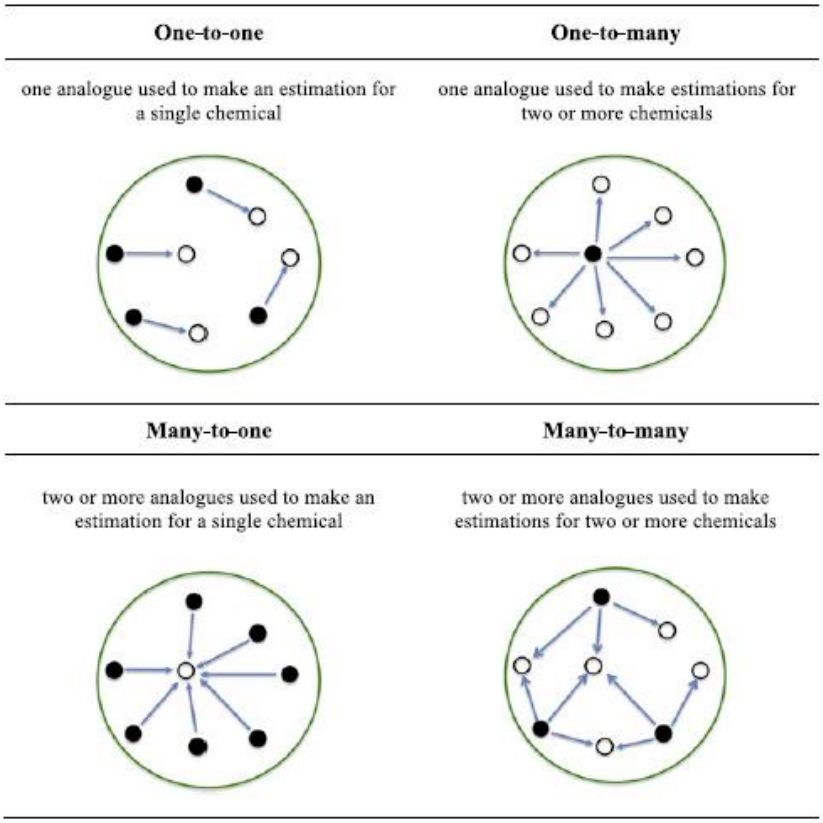
Υπάρχουν δύο προσεγγίσεις όσον αφορά το πλαίσιο εφαρμογής της μεθόδου read-across, η προσέγγιση κατηγοριών ή ομάδων (grouping approach) και η προσέγγιση ανάλογων (analogue approach). Οι ορισμοί των προσεγγίσεων παρουσιάζονται παρακάτω (Πίνακας 1.2).

Πίνακας 1.2: Ορισμοί προσέγγισης ανάλογων και προσέγγισης κατηγοριών.[15]

Προσέγγιση κατηγοριών	**Προσέγγιση ανάλογων**
Σύμφωνα με τον Οργανισμό Οικονομικής Συνεργασίας και Ανάπτυξης, ΟΟΣΑ (Organization for Economic Co-operation Development, OECD), ως χημική κατηγορία ορίζεται η «ομάδα χημικών ουσιών των οποίων οι φυσικοχημικές και τοξικολογικές ιδιότητες είναι πιθανό να είναι παρόμοιες ή να ακολουθούν τακτικά μοτίβα ως αποτέλεσμα των δομικών ομοιοτήτων τους». Ο όρος «προσέγγιση κατηγοριών» χρησιμοποιείται όταν γίνεται σύγκριση μεταξύ διάφορων ουσιών με δομικές ομοιότητες. Αυτές οι ουσίες ομαδοποιούνται βάσει των ομοιοτήτων και των διαφορών που υπάρχουν μεταξύ τους και η εκτίμηση των τοξικών ιδιοτήτων μιας ομάδας είναι η ίδια για όλα τις ουσίες που ανήκουν σε αυτή.	Σύμφωνα με τον OECD, ως ανάλογο ορίζεται η «χημική ουσία της οποίας οι φυσικοχημικές, περιβαλλοντικές ή τοξικολογικές ιδιότητες ενδέχεται να είναι παρόμοιες με τις ιδιότητες μίας άλλης ουσίας λόγω ομοιοτήτων σε δομικές και φυσικοχημικές ιδιότητες». Ο όρος «προσέγγιση ανάλογων» χρησιμοποιείται όταν η εκτίμηση της τοξικής συμπεριφοράς τους περιλαμβάνει ένα μικρό αριθμό χημικών ουσιών και δεν παρατηρούνται εμφανείς τάσεις ή τακτικά μοτίβα στις ιδιότητές τους.

Στην προσέγγιση ανάλογων, η πρόβλεψη περιορίζεται σε μία μικρή περιοχή του χώρου των δεδομένων και εφαρμόζεται τοπικά μια διαδικασία εκτίμησης των ιδιοτήτων. Ως αποτέλεσμα των ομοιοτήτων στη δομή, η γνωστή τοξικολογική ιδιότητα ενός «πηγαίου» νανοσωματιδίου μπορεί να χρησιμοποιηθεί για την πρόβλεψη της ίδιας ιδιότητας σε ένα νανοσωματίδιο «στόχο». Η απλούστερη περίπτωση της προσέγγισης ανάλογων αφορά τη σύγκριση ενός «πηγαίου» νανοσωματιδίου και ενός «στόχου». Εάν χρησιμοποιούνται περισσότερα από ένα «πηγαία» νανοσωματίδια ή «στόχοι», η αξιολόγηση της προσέγγισης πρέπει να επαναληφθεί για κάθε νανοσωματίδιο ξεχωριστά.[18]

Η προσέγγιση των ανάλογων μπορεί να εφαρμοστεί για δύο ή περισσότερες ουσίες με τέσσερις διαφορετικούς τρόπους όπως φαίνεται στο Σχήμα 1.4: Μία γνωστή ουσία να χρησιμοποιηθεί για την πρόβλεψη μίας άγνωστης (one-to-one), πολλές γνωστές ουσίες να χρησιμοποιηθούν για την πρόβλεψη μίας άγνωστης (many-to-one), μία γνωστή προς την πρόβλεψη πολλών αγνώστων (one-to-many) και τέλος πολλές γνωστές ουσίες για την πρόβλεψη πολλών αγνώστων (many-to-many).



Σχήμα 1.4: Τεχνικές προσέγγισης read-across.[19]

Η προσέγγιση κατηγοριών χρησιμοποιείται όταν γίνεται σύγκριση μεταξύ διαφόρων νανοσωματιδίων με δομικές ομοιότητες. Η ομαδοποίηση καθορίζεται από τις ομοιότητες και τις διαφορές σε δομικό επίπεδο και τα νανοσωματίδια που εντάσσονται σε μια ομάδα αντιμετωπίζονται ως ένα. Δηλαδή ως αποτέλεσμα των ομοιοτήτων, οι τοξικολογικές ιδιότητες των νανοσωματιδίων θα είναι όμοιες ή θα ακολουθούν ένα μοτίβο. Οι ομάδες των νανοσωματιδίων μπορούν να χωριστούν σε περαιτέρω υποομάδες βάσει των αλληλεξαρτήσεων των περιγραφέων και ο σχηματισμός αυτών των υποομάδων να οδηγήσει σε πιο ικανοποιητικές προβλέψεις.[18]

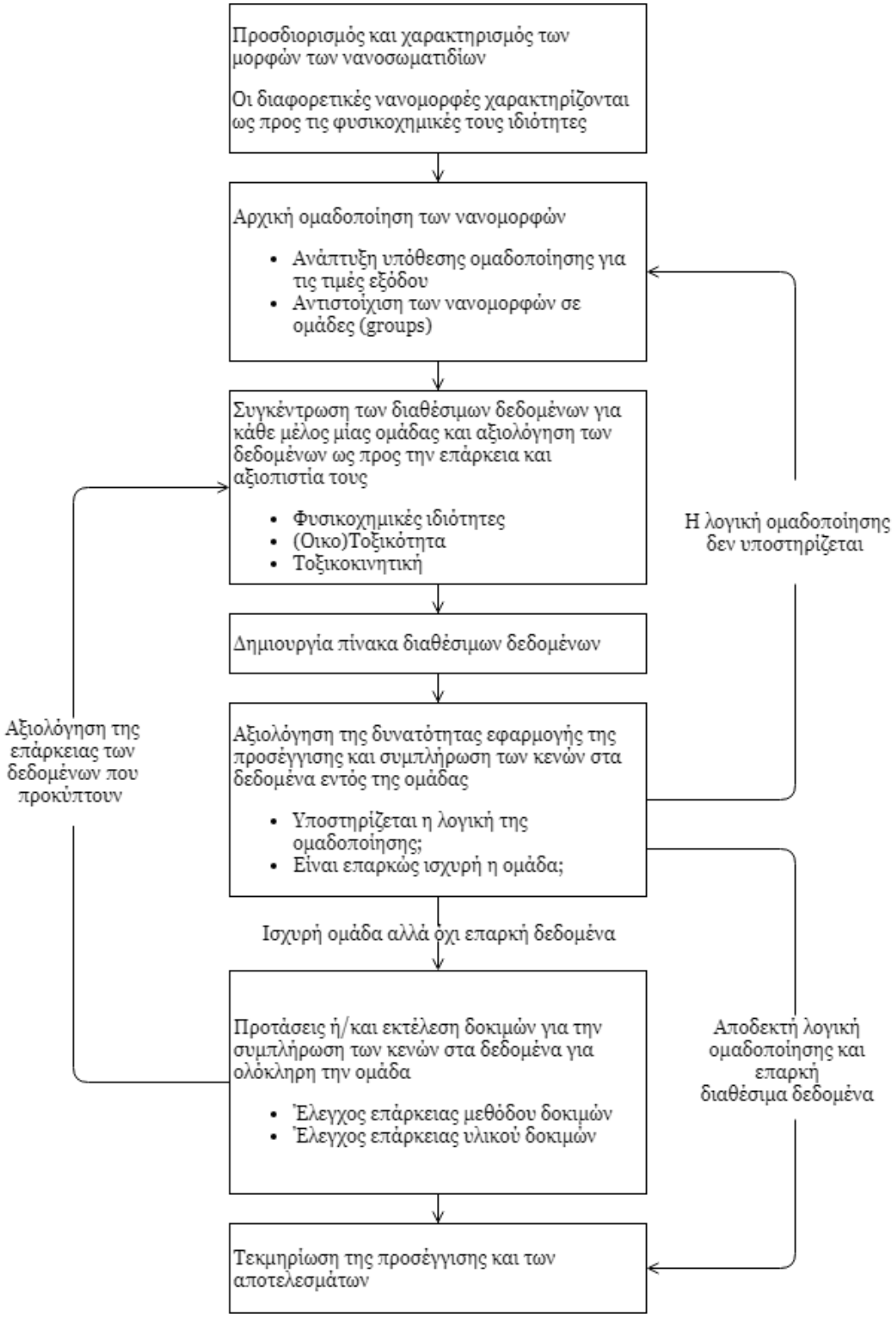
Διάφορες μελέτες έχουν διερευνήσει τη δυνατότητα ομαδοποίησης με άλλες τεχνικές που δε βασίζονται σε εμφανείς δομικές ομοιότητες. Με την ιεραρχική συσσώρευση (Hierarchical Clustering, HC) είναι δυνατός ο εντοπισμός πιθανών ομάδων ή ομάδων ανάλογων νανοσωματιδίων, με την εφαρμογή της ανάλυσης κύριων συνιστωσών (Principal Component Analysis, PCA) είναι δυνατός ο προσδιορισμός φυσικοχημικών ιδιοτήτων που διαφοροποιούν τα νανοσωματίδια καθώς και με την μέθοδο random forest, που αποτελεί μία επιτηρούμενη τεχνική για την εύρεση των πιο σχετικών ιδιοτήτων των νανοσωματιδίων.[16,18]

1.5.1 Ροή εργασιών για την ανάπτυξη μεθόδων read-across

Ο ECHA ανέπτυξε το 2017 το πλαίσιο αξιολόγησης της τεχνικής read-across (Read-Across Assessment Framework, RAAF) για την προσέγγιση των κατηγοριών με σκοπό την καθοδήγηση για την ανάλυση της αβεβαιότητας των προτεινόμενων μεθοδολογιών.

Συνοπτικά, η ροή εργασιών που έχει προταθεί παρουσιάζεται στο Σχήμα 1.5 περιέχει τα εξής βήματα:[20]

1. Τον χαρακτηρισμό των νανομορφών μέσω των ιδιοτήτων τους όπως τη σύνθεση, το μέγεθος, το σχήμα και την επιφανειακή χημεία τους.
2. Την ανάπτυξη υπόθεσης για ομαδοποίηση σύμφωνα με τις τιμές εξόδου (π.χ. δείκτης τοξικότητας), λαμβάνοντας υπόψη τις πληροφορίες από τις φυσικοχημικές ιδιότητες και τις γνωστές τιμές εξόδου, την παρατήρηση ομοιοτήτων μεταξύ των ανάλογων νανομορφών καθώς και την αντιστοίχιση τους σε ομάδες.
3. Τη συλλογή δεδομένων για κάθε νανοσωματίδιο που εντάσσεται σε μία ομάδα και την αξιολόγηση ως προς τη συνάφεια και την εφαρμογή τους για περαιτέρω ενίσχυση της υπόθεσης ομαδοποίησης. Τα δεδομένα μπορεί να περιλαμβάνουν φυσικοχημικές ιδιότητες, περιβαλλοντικές παραμέτρους και (οικο)τοξικολογικά αποτελέσματα.
4. Τη δημιουργία ενός πίνακα που να περιέχει όλα τα διαθέσιμα δεδομένα, τις φυσικοχημικές ιδιότητες και τα τοξικολογικά αποτελέσματα για κάθε μέλος μίας ομάδας.
5. Την συνολική αξιολόγηση των πληροφοριών που συγκεντρώθηκαν για κάθε ομάδα για να ελεγχθεί αν υποστηρίζεται η υπόθεση της ομαδοποίησης του βήματος 2, δηλαδή αν η ομάδα εμφανίζει επαρκείς ομοιότητες στις φυσικοχημικές ιδιότητες με βάση τα διαθέσιμα δεδομένα που συγκεντρώθηκαν καθώς και τον έλεγχο της ισχύος της ομάδας, δηλαδή την ύπαρξη επαρκών, σχετικών και αξιόπιστων πληροφοριών για κάθε μέλος της.
6. Τις προτάσεις για επιπλέον δοκιμές στην περίπτωση που η ομάδα δεν περιέχει επαρκείς πληροφορίες για κάθε μέλος της.
7. Την τελική τεκμηρίωση της ομαδοποίησης και την αξιολόγηση της μεθόδου.



Σχήμα 1.5: Ροή εργασιών μεθόδων read-across σύμφωνα με τις οδηγίες του ECHA.[20]

Κεφάλαιο 2

Μαθηματικός προγραμματισμός

Η Επιχειρησιακή Έρευνα (Operations Research) χαρακτηρίζεται ως η επιστημονική προσέγγιση της λήψης αποφάσεων διαχείρισης, εφαρμόζοντας μαθηματικές μεθόδους και τις δυνατότητες των σύγχρονων υπολογιστών. Ο μαθηματικός προγραμματισμός αποτελεί μία από τις σημαντικότερες μεθόδους λήψης αποφάσεων στον χώρο της Επιχειρησιακής Έρευνας. Πρόκειται για τη βέλτιστη κατανομή περιορισμένων πόρων b_i η οποία προκύπτει μέσα από μία σειρά m περιορισμών που επιβάλλονται από τη φύση του προβλήματος που μελετάται και από τη βελτιστοποίηση ενός στόχου-μίας αντικειμενικής συνάρτησης (objective function). Η αντικειμενική συνάρτηση εκφράζει τη σχέση ανάμεσα στις μεταβλητές x_j του προβλήματος (n μεταβλητές απόφασης) και το σκοπό του συστήματος (μεγιστοποίηση ή ελαχιστοποίηση). Στη συσχέτιση αυτή συμμετέχουν οι συντελεστές της αντικειμενικής συνάρτησης c_j που εκφράζουν τη μεταβολή της τιμής της αντικειμενικής συνάρτησης όταν η αντίστοιχη μεταβλητή απόφασης μεταβάλλεται κατά μία μονάδα. Οι περιορισμοί κωδικοποιούνται ως ανισότητες ή ισότητες που εκφράζουν την κατανομή των πόρων στους οποίους συμμετέχουν οι μεταβλητές απόφασης με κάποιο τεχνολογικό συντελεστή a_{ij} στις διάφορες δραστηριότητες.[21]

Όταν η μαθηματική αναπαράσταση του προβλήματος αποτελείται αποκλειστικά από γραμμικές συναρτήσεις ως προς τις άγνωστες μεταβλητές οι οποίες είναι συνεχείς, τότε πρόκειται για πρόβλημα Γραμμικού Προγραμματισμού (Linear Programming, LP). Στην περίπτωση γραμμικών συναρτήσεων και ακέραιων μεταβλητών, πρόκειται για Ακέραιο Γραμμικό Προγραμματισμό (Integer Linear Programming, ILP), ενώ εάν η μαθηματική αναπαράσταση του προβλήματος αποτελείται από μη γραμμικές συναρτήσεις, τότε είναι πρόβλημα Μη Γραμμικού Προγραμματισμού (Nonlinear Programming, NLP).[22]

Ο γραμμικός προγραμματισμός είναι ένα χρήσιμο βοήθημα για την αντιμετώπιση περίπλοκων προβλημάτων αποφάσεων. Οι αποφάσεις εξαρτώνται απόλυτα από την ακρίβεια της περιγραφής του προβλήματος και από την καταλληλότητα του μοντέλου. Οι προϋποθέσεις που πρέπει να ισχύουν για να διατυπωθεί ένα μοντέλο γραμμικού προγραμματισμού είναι η γραμμικότητα, δηλαδή όλες οι συναρτήσεις του προβλήματος (αντικειμενική συνάρτηση και περιορισμοί) να είναι γραμμικές ως προς τις άγνωστες μεταβλητές, η διαιρετότητα, δηλαδή όλες οι μεταβλητές να είναι συνεχείς και τέλος η βεβαιότητα, όπου όλες οι παράμετροι του προβλήματος είναι γνωστές με απόλυτη βεβαιότητα. Μία τυπική μορφή ενός μοντέλου γραμμικού προγραμματισμού περιλαμβάνει την αντικειμενική συνάρτηση η οποία επιθυμείται να ελαχιστοποιηθεί ή μεγιστοποιηθεί

$$\min \ or \max \ Z = \sum_{j=1}^{n} c_j x_j \qquad [2.1]$$

Και τους περιορισμούς:

$$\sum_{j=1}^{\mathrm{n}} a_{ij}x_j \leq b_i \qquad (i = 1,2,\dots,\mathrm{m}) \qquad [2.2]$$

$$x_j \geq 0 \qquad (j = 1,2,\dots,\mathrm{n}) \qquad [2.3]$$

x_j συνεχής μεταβλητή

Στην περίπτωση όπου οι μεταβλητές απόφασης παίρνουν ακέραιες τιμές, τότε η μορφή του προβλήματος ακέραιου γραμμικού προγραμματισμού είναι η ακόλουθη:

$$\min \ or \max \ Z = \sum_{j=1}^{\mathrm{n}} c_j x_j \qquad [2.4]$$

Και οι περιορισμοί:

$$\sum_{j=1}^{\mathrm{n}} a_{ij}x_j \leq b_i \qquad (i = 1,2,\dots,\mathrm{m}) \qquad [2.5]$$

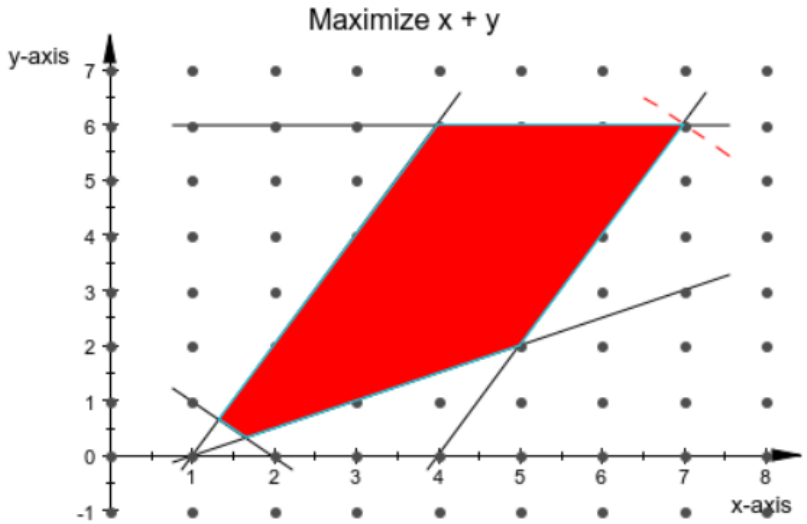
$$x_j \geq 0 \qquad (j = 1,2,\dots,\mathrm{n}) \qquad [2.6]$$

x_j ακέραια μεταβλητή

Όταν ορισμένες μεταβλητές απόφασης παίρνουν ακέραιες τιμές τότε πρόκειται για πρόβλημα μικτού ακέραιου γραμμικού προγραμματισμού (Mixed Integer Linear Programming, MILP).[22]

2.1 Επίλυση προβλημάτων γραμμικού προγραμματισμού

Στα προβλήματα γραμμικού προγραμματισμού, επιθυμείται η βελτιστοποίηση μίας γραμμικής συνάρτησης λαμβάνοντας υπόψιν μία σειρά γραμμικών περιορισμών. Τυπικά, ένα πρόβλημα γραμμικού προγραμματισμού αποτελεί ένα πρόβλημα ελαχιστοποίησης ή μεγιστοποίησης της γραμμικής συνάρτησης. Οι μεταβλητές απόφασης που ικανοποιούν όλους τους περιορισμούς αποτελούν εφικτές λύσεις του προβλήματος (feasible solution). Το σύνολο των εφικτών λύσεων ορίζει το πεδίο εφικτών λύσεων ή την επιτρεπτή περιοχή. Αν πρόκειται για πρόβλημα δύο διαστάσεων, μπορεί να γίνει αναπαράσταση των περιορισμών σε καρτεσιανό σύστημα. Το σύνολο των εφικτών λύσεων τότε ορίζει ένα κυρτό πολύγωνο που ονομάζεται σύνολο εφικτού πεδίου (feasible region). Η άριστη λύση του προβλήματος (optimal solution) είναι η λύση που αριστοποιεί την αντικειμενική συνάρτηση. Στον γραμμικό προγραμματισμό, η άριστη λύση είναι και λύση ακραίου σημείου, δηλαδή λύση που αντιστοιχεί σε μία γωνία του κυρτού πολυγώνου.[23]



Σχήμα 2.1: Αναπαράσταση συνόλου εφικτού πεδίου για πρόβλημα γραμμικού προγραμματισμού.[24]

2.1.1 Μέθοδος επίλυσης προβλήματος γραμμικού προγραμματισμού

Η πλέον γνωστή και περισσότερο χρησιμοποιημένη μέθοδος για την επίλυση ενός γενικού προβλήματος γραμμικού προγραμματισμού είναι η μέθοδος Simplex. Η μέθοδος αυτή αποτελεί μία αλγεβρική επαναληπτική διαδικασία κατά την οποία ο αλγόριθμος μεταβαίνει από τη μία εφικτή λύση στην επόμενη, βελτιώνοντας σε κάθε βήμα την τιμή της αντικειμενικής συνάρτησης. Η μέθοδος τερματίζει όταν δε μπορεί να βελτιωθεί επιπλέον η τιμή της αντικειμενικής συνάρτησης.

Δύο χαρακτηριστικά της μεθόδου Simplex την καθιστούν υπολογιστικό εργαλείο με ευρεία αποδοχή για επίλυση γραμμικών προβλημάτων. Πρώτον, πρόκειται για μία ισχυρή μέθοδο που λύνει κάθε γραμμικό πρόβλημα, ανιχνεύει περιττούς περιορισμούς και λύνει προβλήματα με μία ή περισσότερες άριστες λύσεις. Επιπλέον, η μέθοδος δίνει πληροφορίες όχι μόνο για τη βέλτιστη λύση, αλλά και τον τρόπο που επηρεάζεται η βέλτιστη λύση ως συνάρτηση των δεδομένων του προβλήματος. Οι πληροφορίες αυτές σχετίζονται με ένα γραμμικό πρόβλημα που καλείται δυϊκό πρόβλημα (dual problem), το οποίο επιλύεται αυτόματα μαζί με το δεδομένο πρόβλημα από τη μέθοδο Simplex.[22]

Εάν ένα πρόβλημα γραμμικού προγραμματισμού περιέχει περιορισμούς ισότητας ή μεγαλύτερου ίσου, τότε η αρχική βασική εφικτή λύση μπορεί να μην είναι εμφανής. Η μέθοδος του μεγάλου M (big M method) αποτελεί μία μέθοδο επίλυσης τέτοιων προβλημάτων στηριζόμενη στην μέθοδο Simplex κατά την οποία προστίθενται «τεχνητές» μεταβλητές (artificial variables) στο πρόβλημα. Η αντικειμενική συνάρτηση ωστόσο, πρέπει να τροποποιηθεί ώστε οι τεχνητές μεταβλητές να ισούνται με μηδέν κατά την ολοκλήρωση της μεθόδου. Ο όρος «big M» αναφέρεται σε μεγάλους θετικούς αριθμούς οι οποίοι σχετίζονται με τις τεχνητές μεταβλητές ώστε οι μεταβλητές αυτές να μην αποτελούν μέρος της εφικτής λύσης. Η τιμή του όρου M πρέπει να είναι μεγάλη ώστε να κυριαρχεί υπό των υπολοίπων μεταβλητών, αλλά όχι υπερβολικά υψηλή καθώς μπορεί να προκαλέσει σοβαρά υπολογιστικά σφάλματα.[25,26]

2.2 Προεπεξεργασία δεδομένων

Πριν γίνει η ανάλυση των δεδομένων οφείλεται να γίνει μία προεπεξεργασία σε αυτά ώστε τα παραγόμενα μοντέλα να είναι αποδοτικότερα. Μοντέλα που προκύπτουν από δεδομένα που έχουν επεξεργαστεί λάθος ή και καθόλου μπορεί να είναι προβληματικά ή μη αποδεκτά. Είναι πιθανό να υπάρχουν ελλιπή δεδομένα με άγνωστες-κενές τιμές σε κάποια δείγματα. Οι άγνωστες τιμές πρέπει να συμπληρώνονται από τιμές που προκύπτουν με διάφορες μεθόδους (όπως από τον μέσο όρο των αντιστοίχων μεταβλητών). Επιπλέον, τα δεδομένα πρέπει να καθαρίζονται από παραπλανητικά δείγματα ή ακραίες τιμές. Δείγματα με ακραίες τιμές μπορεί να προκαλέσουν εμπόδια στη δημιουργία αποδοτικών μοντέλων. Η περιοχή μέσα στην οποία εκπαιδεύεται το μοντέλο πρέπει να καθορίζεται και μόνο για τα δείγματα που ανήκουν σε αυτή την περιοχή, η οποία υπολογίζεται ως το πεδίο εφαρμογής του μοντέλου (Domain of Applicability), όπως αναφέρεται στην Ενότητα 2.5.4, το μοντέλο δίνει αξιόπιστα αποτελέσματα.

2.2.1 Κανονικοποίηση

Αρχικά, τα δείγματα περνούν από μία διαδικασία κανονικοποίησης (normalization). Ο σκοπός της κανονικοποίησης είναι μετατροπή των τιμών του συνόλου των δεδομένων σε μία κοινή κλίμακα, χωρίς να διαταράσσονται οι διαφορές στα εύρη τιμών. Με αυτόν τον τρόπο οι μεταβλητές συμβάλλουν με τον ίδιο βαθμό στη μοντελοποίηση. Υπάρχουν διάφορες μέθοδοι κανονικοποίησης, στην παρούσα Εργασία χρησιμοποιήθηκε η κανονικοποίηση μεγίστου-ελαχίστου, σύμφωνα με την Εξίσωση [2.7], ώστε όλες οι μεταβλητές να παίρνουν τιμές στο διάστημα [0,1] με μέγιστη τιμή μεταβλητής να είναι 1 και ελάχιστη 0.

$$X_{ij}^{n} = \frac{X_{ij} - X_{j,min}}{X_{j,max} - X_{j,min}} \qquad [2.7]$$

Όπου X_{ij} και X_{ij}^{n} αντιστοιχούν στη μη κανονικοποιημένη και τη κανονικοποιημένη τιμή της μεταβλητής j ($j=1,\ldots,$K) του δείγματος i ($i=1,\ldots,$N) αντίστοιχα, ενώ $X_{j,min}$ και $X_{j,max}$ στην ελάχιστη και μέγιστη τιμή της μεταβλητής j.[27]

2.3 Μοντέλα πρόβλεψης

Για την πρόβλεψη ενός μεγέθους μέσω ενός μοντέλου απαιτούνται δεδομένα της μορφής $(\boldsymbol{X}, y)$. Μία από τις μεταβλητές εκφράζεται ως συνάρτηση των υπολοίπων, επιτρέποντας την πρόβλεψη της τιμής της όταν είναι γνωστές οι υπόλοιπες. Η εξαρτημένη μεταβλητή (response variable), ο οποία υπολογίζεται μέσω μίας συνάρτησης, συμβολίζεται με y, ενώ οι γνωστές ιδιότητες (predictor variables) συμβολίζονται με $(x_1, x_2, \ldots, x_f)$. Το μοντέλο θα προβλέψει την τιμή της εξαρτημένης μεταβλητής μέσα από μία εξίσωση $\hat{y} = f(x_1, \ldots, x_f, \theta)$, όπου $\hat{y}$ αντιστοιχεί στην προβλεπόμενη τιμή του μοντέλου και θ αντιστοιχεί στις παραμέτρους του μοντέλου.[28]

2.4 Μοντέλα γραμμικής παλινδρόμησης

Η ανάλυση παλινδρόμησης (regression analysis) πρόκειται για μία στατιστική τεχνική που στοχεύει στην εύρεση της σχέσης μεταξύ των μεταβλητών $\boldsymbol{X}$ και y. Με την ανάλυση παλινδρόμησης προκύπτει μία εξίσωση ή ένα μοντέλο μεταξύ των μεγεθών $\boldsymbol{X}$ και y, το οποίο χρησιμοποιείται για την εκτίμηση της εξαρτημένης μεταβλητής y_i από τις ανεξάρτητες μεταβλητές x_i. Αν το μοντέλο που προκύπτει είναι τέτοιας μορφής ώστε η εξαρτημένη μεταβλητή y να είναι γραμμική συνάρτηση των παραμέτρων του μοντέλου, τότε ονομάζεται γραμμική παλινδρόμηση (linear regression).[29]

2.4.1 Μοντέλο απλής γραμμικής παλινδρόμησης

Στην περίπτωση του μοντέλου απλής γραμμικής παλινδρόμησης, η σχέση των μεταβλητών x και y είναι μία ευθεία γραμμή της μορφής

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad [2.8]$$

όπου β_0 είναι η αποτέμνουσα, β_1 η κλίση και ε ένα τυχαίο σφάλμα. Τα σφάλματα μπορεί να θεωρηθούν ότι έχουν μηδενική τιμή και άγνωστη διακύμανση σ^2. Επιπλέον, τα σφάλματα θεωρείται ότι δεν συσχετίζονται μεταξύ τους, δηλαδή η τιμή ενός σφάλματος δεν εξαρτάται από την τιμή ενός άλλου.[29]

Οι παράμετροι β_0 και β_1ονομάζονται και συντελεστές παλινδρόμησης. Η κλίση β_1εκφράζει την αλλαγή της μέσης τιμής της κατανομής y που προκύπτει από αλλαγή μίας μονάδας της μεταβλητής x. Αν στο σύνολο τιμών της μεταβλητής x περιλαμβάνεται η τιμή μηδέν, τότε για $x=0$ η αποτέμνουσα β_0 εκφράζει την μέση τιμή της κατανομής της εξαρτημένης μεταβλητής.

Οι παράμετροι β_0 και β_1 υπολογίζονται συνήθως με την μέθοδο των ελαχίστων τετραγώνων, ώστε το άθροισμα των τετραγώνων των διαφορών μεταξύ των παρατηρήσεων y_i και της ευθείας γραμμής να είναι ελάχιστο. Υποθέτοντας S ζεύγη σημείων $(y_1, x_1), \dots, (y_S, x_S)$, η γενική εξίσωση της γραμμικής παλινδρόμησης είναι της μορφής:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1,2, \dots, \mathrm{S} \qquad [2.9]$$

Η Εξίσωση [2.8] μπορεί να χαρακτηριστεί ως μοντέλο παλινδρόμησης για τον πληθυσμό ενώ η Εξίσωση [2.9] ως μοντέλο παλινδρόμησης για το δείγμα i. Το κριτήριο ελαχίστων τετραγώνων είναι:

$$\min E(\beta_0, \beta_1) = \sum_{i=1}^{\mathrm{S}} (y_i - \beta_0 - \beta_1 x_i)^2 \qquad [2.10]$$

Βρίσκοντας τις παραγώγους του E ως προς β_0 και ως προς β_1οι οποίες πρέπει να είναι ίσες με μηδέν, προκύπτουν δύο εξισώσεις με αγνώστους β_0 και β_1.

$$S\widehat{\beta_0} + \widehat{\beta_1}\sum_{i=1}^{S} x_i = \sum_{i=1}^{S} y_i \qquad [2.11]$$

$$\widehat{\beta_0}\sum_{i=1}^{S} x_i + \widehat{\beta_1}\sum_{i=1}^{S} {x_i}^2 = \sum_{i=1}^{S} y_i\, x_i \qquad [2.12]$$

Επιλύοντας το σύστημα των Εξισώσεων [2.11] και [2.12], βρίσκονται οι συντελεστές παλινδρόμησης ως εξής:

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x} \qquad [2.13]$$

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{S} y_i\, x_i - \frac{\left(\sum_{i=1}^{S} y_i\right)\left(\sum_{i=1}^{S} x_i\right)}{S}}{\sum_{i=1}^{S} {x_i}^2 - \frac{\left(\sum_{i=1}^{S} x_i\right)^2}{S}} \qquad [2.14]$$

Όπου $\bar{y} = \frac{1}{S}\sum_{i=1}^{S} y_i$ και $\bar{x} = \frac{1}{S}\sum_{i=1}^{S} x_i$

2.4.2 Μοντέλο πολλαπλής γραμμικής παλινδρόμησης

Στην περίπτωση όπου η μεταβλητή εξάρτησης y σχετίζεται με f γνωστές μεταβλητές, το μοντέλο έχει τη μορφή της Εξίσωσης [2.15] και καλείται μοντέλο πολλαπλής γραμμικής παλινδρόμησης (Multiple linear regression model) με *f* ανεξάρτητες μεταβλητές.

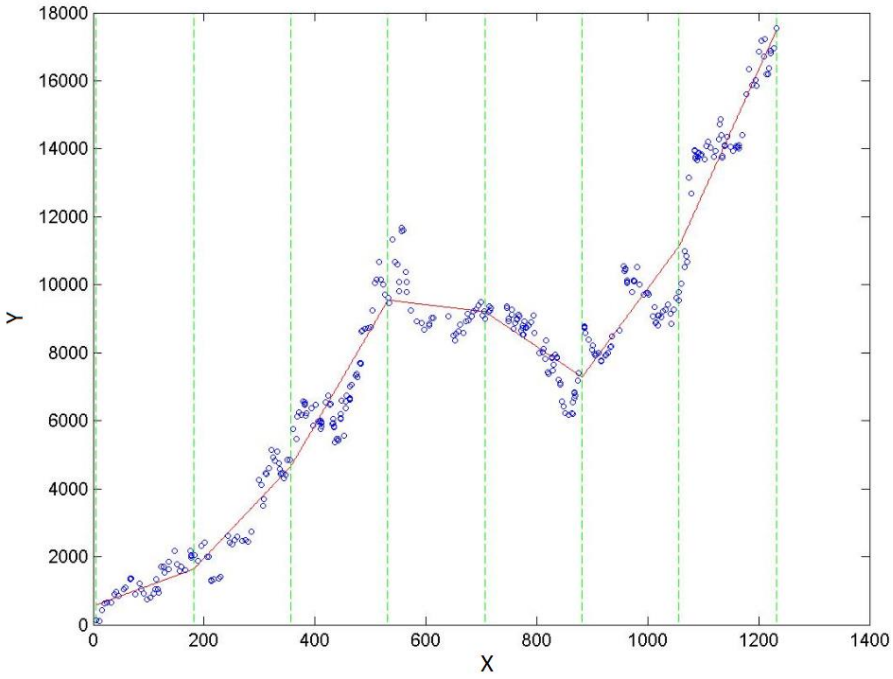
$$\hat{y} = \beta_0 + \sum_{i=1}^{f} \beta_i x_i + \varepsilon \qquad [2.15]$$

Αυτό το μοντέλο περιγράφει ένα υπερεπίπεδο (hyperplane) στο χώρο *f*-διαστάσεων των ανεξάρτητων μεταβλητών x_i. Η παράμετρος β_j, όπως και προηγουμένως, αντιπροσωπεύει την αλλαγή στην μέση τιμή της εξαρτημένης μεταβλητής y με αλλαγή μίας μονάδας της μεταβλητής x_j όταν όλες οι ανεξάρτητες μεταβλητές x_i $(i \neq j)$ παραμένουν σταθερές.[29]

2.4.3 Μοντέλο τμηματικής γραμμικής παλινδρόμησης

Επιπλέον γενίκευση του βασικού γραμμικού μοντέλου μπορεί να επιτευχθεί αν υποτεθεί ότι η εξαρτημένη μεταβλητή είναι «τοπικά» γραμμική στο πεδίο των ανεξάρτητων μεταβλητών x_i, με διαφορετική εξάρτηση σε κάθε περιοχή στον χώρο των x_i, δίνοντας ένα τμηματικά γραμμικό μοντέλο. Γεωμετρικά, η δομή του μοντέλου αποτελείται από ένα σύνολο διαφορετικών *f*-διαστάσεων υπερεπιπέδων, καθένα από τα οποία ανήκει σε

διαφορετική περιοχή των ανεξάρτητων μεταβλητών x_i. Οι παράμετροι της δομής του μοντέλου περιέχουν τις τοπικές παραμέτρους για κάθε υπερεπίπεδο, καθώς και τα όρια των υπερεπιπέδων. Για ανεξάρτητη μεταβλητή x μίας διάστασης και για r γραμμικά τμήματα, το μοντέλο αποτελείται από μία καμπύλη με r γραμμικά τμήματα.



Σχήμα 2.2: Τμηματική γραμμική παλινδρόμηση για μονοδιάστατη ανεξάρτητη μεταβλητή.[30]

Στο Σχήμα 2.2 παρατηρείται ότι η καμπύλη είναι μια συνεχής γραμμή που ενώνει τα τμήματα r. Η συνέχεια της γραμμής δεν είναι αναγκαία για τέτοιου είδους μοντέλα. Μπορούν να δημιουργηθούν μοντέλα με ασυνέχεια στα τμήματα, ωστόσο ορισμένες φορές αυτές οι ασυνέχειες μπορεί να δημιουργούν προβλήματα και να μην είναι επιθυμητά λόγω της απότομης αλλαγής της εξαρτημένης μεταβλητής y με απειροελάχιστη αλλαγή της ανεξάρτητης μεταβλητής x.

2.5 Αξιολόγηση μοντέλου

Η εξόρυξη δεδομένων (data mining) αποτελεί μια διαδικασία εξαγωγής χρήσιμων πληροφοριών από μία μεγάλη βάση δεδομένων. Υπάρχουν διάφορες τεχνικές εξόρυξης δεδομένων και συνήθως χρησιμοποιούνται παραπάνω από μία τεχνικές για τη μελέτη μίας βάσης δεδομένων. Ωστόσο, η εφαρμογή διαφορετικών μεθόδων σε μεγάλα σύνολα δεδομένων μπορεί να οδηγήσει σε υπολογιστικά προβλήματα. Μία λύση σε αυτό το πρόβλημα είναι η μείωση του μεγέθους των δεδομένων, επιλέγοντας ένα αντιπροσωπευτικό υποσύνολο αυτών. Δύο κύριες κατηγορίες μεθόδων βασίζονται σε τεχνικές συσταδοποίησης (cluster-base designs) και τεχνικές ομοιόμορφου σχεδιασμού (uniform designs). Οι τεχνικές συσταδοποίησης συσπειρώνουν το σύνολο δεδομένων και στη συνέχεια, σύμφωνα με την ομαδοποίηση τους, επιλέγονται τα αντιπροσωπευτικά σύνολα. Με τις τεχνικές ομοιότητας, τα αντιπροσωπευτικά σύνολα επιλέγονται ώστε να καλύπτουν ομοιόμορφα τον χώρο των δεδομένων.[31]

Η διάσπαση των δεδομένων (data splitting) σε υποσύνολα μπορεί να χρησιμοποιηθεί και ως τεχνική αξιολόγησης ενός μοντέλου, με τη διαίρεση των δεδομένων σε σύνολο εκπαίδευσης (training set) και σύνολο δοκιμών/επικύρωσης (test set). Το μοντέλο τροφοδοτείται με το σύνολο εκπαίδευσης και «χτίζεται» με έναν εκπαιδευτικό αλγόριθμο, ενώ στη συνέχεια η ισχύς της πρόβλεψης του μοντέλου αξιολογείται με βάση τη διαφορά προβλεπόμενης και πειραματικής τιμής των δεδομένων του συνόλου δοκιμών. Η μέθοδος αυτή ονομάζεται εξωτερική επικύρωση (external validation). Μόνο τα σωστά εκπαιδευμένα και επικυρωμένα μοντέλα είναι ικανά να παρέχουν αξιόπιστες προβλέψεις για άγνωστα δείγματα.[32]

2.5.1 Αλγόριθμος Kennard and Stone

Η πιο γνωστή τεχνική ομοιόμορφου σχεδιασμού στη χημειομετρία είναι η Kennard-Stone.[33] Επιλέγεται ένα υποσύνολο δειγμάτων *f*-διαστάσεων, κατανεμημένων ομοιόμορφα στο χώρο των δεδομένων. Όλα τα δείγματα *s* από το σύνολο δεδομένων X(*s,f*) εξετάζονται ως αντιπροσωπευτικά για το υποσύνολο. Το δείγμα με τιμή κοντινότερη στη μέση τιμή του συνόλου δεδομένων θεωρείται το πιο αντιπροσωπευτικό και επιλέγεται πρώτο στο υποσύνολο. Η επιλογή του επόμενου δείγματος γίνεται υπολογίζοντας την Ευκλείδεια απόσταση μεταξύ αυτού και του δείγματος που έχει επιλεγεί στο υποσύνολο, όπως ορίζεται παρακάτω, για δείγματα i και j:

$$d_{ij}^2 = \|x_i - x_j\| = \sum_k (x_{ik} - x_{jk})^2 \qquad [2.16]$$

Το κάθε δείγμα $s_1, s_2, \dots, s_k$ που επιλέγεται στο υποσύνολο είναι το πιο απομακρυσμένο από το προηγούμενο που ανήκει σε αυτό. Επομένως, το αποτέλεσμα της εφαρμογής του αλγορίθμου Kennard-Stone είναι η επιλογή του κεντρικού δείγματος και ορισμένων απομακρυσμένων από αυτό που να βρίσκονται στα όρια των δεδομένων και στη συνέχεια η προσθήκη υπολοίπων που βρίσκονται στον χώρο των δεδομένων.[31]

2.5.2 Έλεγχος αξιοπιστίας παραγόμενου μοντέλου

Μία συνηθισμένη συνθήκη για την αξιολόγηση ενός μοντέλου παλινδρόμησης είναι ο συντελεστής συσχέτισης r να είναι όσο το δυνατόν πιο κοντά στη μονάδα και το τυπικό σφάλμα εκτίμησης μικρό. Ωστόσο, αυτή η συνθήκη δείχνει την ικανότητα προσαρμογής του μοντέλου, δηλαδή πόσο ικανό είναι το μοντέλο να προβλέπει ικανοποιητικά την εξαρτημένη μεταβλητή για τα δεδομένα εκπαίδευσης και αποτελεί ανεπαρκή προϋπόθεση για την εγκυρότητα του μοντέλου. Οι παράμετροι αυτές δείχνουν την ποιότητα της προσαρμογής (quality of the fit) μεταξύ της πρόβλεψης και της πειραματικής τιμής, ενώ δε μπορούν να δώσουν στοιχεία για την ικανότητα του μοντέλου να παρέχει αξιόπιστες προβλέψεις για καινούρια δεδομένα.[34]

Ο συντελεστής συσχέτισης Pearson (Pearson correlation coefficient) μεταξύ δύο διανυσμάτων X και Y είναι ένας αδιάστατος αριθμός που υπολογίζεται σύμφωνα με την παρακάτω σχέση:[35]

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad [2.17]$$

Όπου x_i το στοιχείο i του διανύσματος x,
y_i το στοιχείο i του διανύσματος y,
$\bar{x}$ η μέση τιμή των στοιχείων του διανύσματος x,
$\bar{y}$ η μέση τιμή των στοιχείων του διανύσματος y.

Υψώνοντας στο τετράγωνο τον συντελεστή συσχέτισης Pearson, προκύπτει ο συντελεστής συσχέτισης R^2, που αποτελεί δείκτη αξιολόγησης του μοντέλου για το σύνολο εκπαίδευσης. Για τιμές του R^2 οι οποίες προσεγγίζουν την μονάδα, η πρόβλεψη θεωρείται πιο αξιόπιστη.

Η ικανότητα πρόβλεψης του μοντέλου παλινδρόμησης υπολογίζεται με τη χρήση ενός εξωτερικού συνόλου δεδομένων, του συνόλου δοκιμών, που δε χρησιμοποιήθηκε για την ανάπτυξη του μοντέλου. Χρησιμοποιώντας το μοντέλο που αναπτύχθηκε προκύπτουν οι τιμές πρόβλεψης για τα δείγματα του συνόλου δοκιμών, οι οποίες αξιολογούνται ως προς την αξιοπιστία τους μέσω ενός δείκτη εξωτερικής ερμηνεύσιμης διακύμανσης (external explained variance) Q^2_{test}. Τιμές του δείκτη Q^2_{test} οι οποίες προσεγγίζουν τη μονάδα δείχνουν ότι το μοντέλο είναι αξιόπιστο.

$$Q^2_{test} = 1 - \frac{\sum_{i=1}^{n_{test}}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n_{test}}(y_i - \bar{y})^2} \quad [2.18]$$

Όπου y_i η πειραματική τιμή της εξαρτημένης μεταβλητής του στοιχείου i,
$\hat{y_i}$ η προβλεπόμενη τιμή της εξαρτημένης μεταβλητής του στοιχείου i,
$\bar{y}$ η μέση τιμή της εξαρτημένης μεταβλητής για το σύνολο εκπαίδευσης

Σε αντίθεση με τον διασταυρούμενο συντελεστή συσχέτισης (cross-validated correlation coefficient, Q^2), στον συντελεστή εξωτερικής ερμηνεύσιμης διακύμανσης Q^2_{test}, το άθροισμα του τετραγώνου των διαφορών των προβλέψεων και των πειραματικών τιμών στον αριθμητή υπολογίζεται για τα δείγματα του συνόλου δοκιμών και το συνολικό άθροισμα αναφοράς των τετραγώνων στον παρονομαστή υπολογίζεται συγκρίνοντας την πειραματική τιμή απόκρισης για τα δείγματα του συνόλου δοκιμών με την μέση τιμή του συνόλου εκπαίδευσης.[34]

2.5.3 Έλεγχος τυχαίας επιλογής

Ο έλεγχος τυχαίας επιλογής (y-randomization ή y-scrambling) αποτελεί μια τεχνική για τον έλεγχο της ισχύος ενός μοντέλου. Σε αυτή τη διαδικασία, τα στοιχεία του πίνακα της εξαρτημένης μεταβλητής y που αντιστοιχούν σε κάθε δείγμα ανακατεύεται τυχαία και αναπτύσσεται ένα νέο μοντέλο χρησιμοποιώντας τις γνωστές μεταβλητές x. Η διαδικασία επαναλαμβάνεται αρκετές φορές. Αναμένεται ότι στα νέα μοντέλα οι συντελεστές συσχέτισης r και Q^2_{test} θα έχουν χαμηλές τιμές. Αν τα καινούρια μοντέλα που αναπτύσσονται έχουν χαμηλότερες τιμές συντελεστών συσχέτισης από το αρχικό μοντέλο που δημιουργήθηκε για τις πραγματικά κατανεμημένες τιμές εξαρτημένης μεταβλητής, το μοντέλο έχει αναπτυχθεί σωστά και μία καλή τιμή του συντελεστή

συσχέτισης δεν προκύπτει τυχαία. Αντίθετα, αν οι νέοι συντελεστές συσχέτισης έχουν υψηλές τιμές, τότε το μοντέλο δεν θεωρείται αξιόπιστο στις προβλέψεις του, τόσο λόγω των δεδομένων όσο και της μεθοδολογίας μοντελοποίησης.[34]

2.5.4 Πεδίο εφαρμογής μοντέλου

Ακόμη και αν το μοντέλο είναι πλήρως επικυρωμένο, οι προβλέψεις που παράγονται για ορισμένα δεδομένα εισόδου μπορεί να μην είναι αξιόπιστες. Για αυτό το λόγο, είναι σημαντικό να ορίζονται τα όρια του πεδίου εφαρμογής του μοντέλου. Μόνο οι προβλέψεις για τα δείγματα που ανήκουν στο πεδίο εφαρμογής μπορούν να θεωρηθούν αξιόπιστες. Το πεδίο εφαρμογής μπορεί να προσδιοριστεί υπολογίζοντας μετρήσεις ομοιότητας που βασίζονται στις Ευκλείδειες αποστάσεις μεταξύ των δεδομένων εκπαίδευσης και ελέγχου. Η απόσταση ενός δείγματος ελέγχου από το κοντινότερο γείτονα του που ανήκει στο σύνολο εκπαίδευσης συγκρίνεται με ένα προκαθορισμένο κατώφλι (APD) και η πρόβλεψη θεωρείται αξιόπιστη όταν η τιμή της απόστασης είναι μικρότερη από το κατώφλι. Ο υπολογισμός του προκαθορισμένου κατωφλιού APD γίνεται:

$$\mathrm{A}PD = \langle d \rangle + \mathrm{Z}\sigma \qquad [2.19]$$

Αρχικά υπολογίζονται οι Ευκλείδειες αποστάσεις μεταξύ όλων των δειγμάτων του συνόλου εκπαίδευσης. Στη συνέχεια, υπολογίζεται η νέα μέση τιμή $\langle d \rangle$ και η τυπική απόκλιση σ των αποστάσεων ενός νέου συνόλου δεδομένων που αποτελείται από τα δείγματα που έχουν αποστάσεις μικρότερες από τη μέση απόσταση του συνόλου εκπαίδευσης. Η σταθερά Z είναι μία εμπειρική τιμή αποκοπής (cutoff value), με συνήθη τιμή 0.5.[36]

Κεφάλαιο 3

Ανάπτυξη λογισμικού

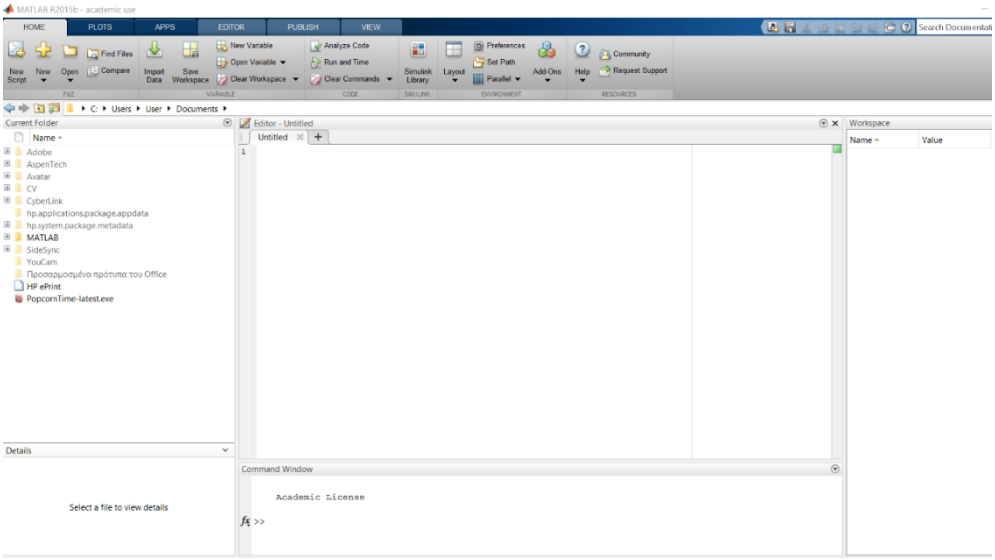
Για την ανάλυση και την επεξεργασία των δεδομένων στα πλαίσια της παρούσας Διπλωματικής Εργασίας, αναπτύχθηκε κώδικας σε γλώσσα MATLAB, έκδοση R2015b (επίσημη ιστοσελίδα: https://www.mathworks.com/products/matlab.html). Η MATLAB επιλέχθηκε επειδή είναι ένα ισχυρό υπολογιστικό πακέτο και δίνει τη δυνατότητα να ενταχθούν σε αυτή κατάλληλοι επιλύτες (solvers) για την επιτυχή «επίλυση» του προβλήματος μαθηματικού προγραμματισμού που αναπτύχθηκε στα πλαίσια της Εργασίας.

3.1 Λογισμικό MATLAB

Το MATLAB είναι ένα ισχυρό πακέτο λογισμικού κατάλληλο για τεχνικούς υπολογισμούς. Εκτός από τη γλώσσα προγραμματισμού, διαθέτει ένα φιλικό προς χρήστη περιβάλλον εργασίας για την ανάπτυξη του κώδικα καθώς και ένα ισχυρό σύστημα γραφικής απεικόνισης των αποτελεσμάτων. Διαθέτει ενσωματωμένες συναρτήσεις για την εκτέλεση πολλών λειτουργιών, ενώ υπάρχουν και εργαλειοθήκες (toolboxes) οι οποίες μπορούν να προστεθούν για να αυξηθούν αυτές οι συναρτήσεις. Επιπλέον, με ένα ειδικό υποπρόγραμμα, το SIMULINK, είναι δυνατή η μοντελοποίηση, ανάλυση και προσομοίωση δυναμικών συστημάτων. Κύριες χρήσεις του αφορούν την ανάλυση δεδομένων, οπτικοποίηση δεδομένων, μαθηματικούς υπολογισμούς, μοντελοποίηση, ανάπτυξη αλγορίθμων, ανάπτυξη εφαρμογών κ.α.[37,38]

Η λέξη MATLAB προέρχεται από τα αρχικά των λέξεων MATrix LABoratory, καθώς αρχικά φτιάχτηκε για να προσφέρει εύκολη χρήση στα λογισμικά LINPACK (Linear system package) και EISPACK (Eigen system package), τα οποία αποτελούν σύγχρονα λογισμικά για υπολογισμό πινάκων. Το MATLAB έχει πολλά πλεονεκτήματα σε σύγκριση με τις συμβατικές γλώσσες προγραμματισμού (C, Fortran) για την επίλυση τεχνικών προβλημάτων. Είναι ένα διαδραστικό σύστημα με βασικό στοιχείο δεδομένων έναν πίνακα που δεν απαιτεί δήλωση της διάστασής του εκ των προτέρων. Έτσι, είναι δυνατή η επίλυση μεγάλου αριθμού προβλημάτων, ειδικά αυτών που η διατύπωσή τους γίνεται σε διανυσματική μορφή σε σύντομο χρόνο.[37,38]

Το λογισμικό αυτό προσφέρει ένα ολοκληρωμένο Περιβάλλον Εργασίας (Desktop) στο οποίο περιλαμβάνονται ένα σύνολο εργαλείων σε γραφικό περιβάλλον για τη διαχείριση αρχείων, εντολών, μεταβλητών και εφαρμογών σχετικών με τη γλώσσα προγραμματισμού. Τα κυριότερα μέρη του περιβάλλοντος εργασίας είναι το παράθυρο εντολών (Command Window), ο χώρος εργασίας (Workspace) και το παράθυρο τρέχοντος φακέλου (Current Folder).[37]



Σχήμα 3.1: Αρχική μορφή περιβάλλοντος εργασίας MATLAB R2015b (desktop).[39]

3.2 Επιλύτες

Οι επιλύτες είναι κομμάτια μαθηματικού λογισμικού, συνήθως στη μορφή αυτόνομων προγραμμάτων υπολογιστή ή βιβλιοθηκών λογισμικού, που «λύνουν» ένα μαθηματικό πρόβλημα. Ένας επιλύτης παίρνει περιγραφές προβλημάτων σε κάποια γενική μορφή και υπολογίζει τη λύση τους. Για την επίλυση του προβλήματος βελτιστοποίησης που αναπτύχθηκε στην παρούσα Εργασία, χρειάστηκε να χρησιμοποιηθούν συγκεκριμένοι επιλύτες σε περιβάλλον MATLAB, οι οποίο παρουσιάζονται συνοπτικά στη συνέχεια.

3.2.1 YALMIP

Το YALMIP είναι μια διαθέσιμη εργαλειοθήκη η οποία αρχικά αναπτύχθηκε για επίλυση προβλημάτων εφαρμογών ημι-ορισμένου προγραμματισμού (semidefinite programming, SDP) σε συνδυασμό με εξωτερικά εργαλεία επίλυσης. Το YALMIP διευκολύνει την ανάπτυξη προβλημάτων βελτιστοποίησης και τον έλεγχο των SDP προβλημάτων. Η χρήση του αρχικά περιοριζόταν σε προβλήματα SDP και LMI (linear matrix inequalities) αλλά τα τελευταία χρόνια έχει επεκταθεί και σε άλλα προβλήματα. Περιέχει μια σειρά από εσωτερικούς επιλύτες, αλλά επίσης προσφέρει τη δυνατότητα να χρησιμοποιηθούν μέσω αυτού μια σειρά άλλων-εξωτερικών επιλυτών οι οποίοι είναι εμπορικά διαθέσιμοι ή παρέχονται δωρεάν μέσω ακαδημαϊκής άδειας. Έτσι είναι δυνατή η χρήση του για γραμμικό προγραμματισμό, τετραγωνικό προγραμματισμό (quadratic programming, QP), προγραμματισμό κώνου δεύτερης τάξης (second order cone programming, SOCP), MILP, πολυπαραμετρικό γραμμικό και τετραγωνικό προγραμματισμό κ.α.[40] Στη συγκεκριμένη εργασία χρησιμοποιήθηκε η έκδοση YALMIP R20190425.

3.2.2 MOSEK

Το MOSEK είναι ένα πακέτο λογισμικού για την επίλυση γραμμικών, μικτών ακέραιων γραμμικών, τετραγωνικών κωνικών και κυρτών τετραγωνικών προβλημάτων μαθηματικής βελτιστοποίησης. Είναι δυνατή η σύνδεση του με γλώσσες προγραμματισμού όπως η C, Java, .NET, Python, Matlab. Λόγω των σύγχρονων Interior-point μεθόδων για τη βελτιστοποίηση γραμμικών, τετραγωνικών και κωνικών προβλημάτων που χρησιμοποιεί, το MOSEK χρησιμοποιείται ευρέως στη βιομηχανία. Στη συγκεκριμένη εργασία χρησιμοποιήθηκε η έκδοση Mosek 9.0.[41]

3.2.3 GUROBI

Το GUROBI αποτελεί ένα πακέτο επίλυσης προβλημάτων βελτιστοποίησης LP, QP, περιορισμένου τετραγωνικού (QCP), MILP, μικτού ακέραιου τετραγωνικού (MIQP) και μικτού ακέραιου περιορισμένου τετραγωνικού προγραμματισμού (MIQCP). Υποστηρίζει μια μεγάλη ποικιλία γλωσσών προγραμματισμού όπως C++, C, Java, .NET, Python, Matlab, R. Στη συγκεκριμένη εργασία χρησιμοποιήθηκε η έκδοση Gurobi 811.[42]

Κεφάλαιο 4

Μεθοδολογία

Όπως αναφέρθηκε, τα τελευταία χρόνια γίνεται προσπάθεια ανάπτυξης εναλλακτικών μεθόδων αξιολόγησης των ανεπιθύμητων ιδιοτήτων των νανοσωματιδίων. Ο Ευρωπαϊκός Οργανισμός Χημικών Προϊόντων ανέπτυξε το πλαίσιο αξιολόγησης της τεχνικής read-across για την εκτίμηση της τοξικότητας των νανοσωματιδίων. Οι μεθοδολογίες read-across βασίζονται στην υπόθεση ότι παρόμοια υλικά μπορεί να παρουσιάζουν αντίστοιχες ιδιότητες, επομένως η εκτίμηση των ανεπιθύμητων ιδιοτήτων είναι εφικτή σε ένα σύνολο παρόμοιων νανοσωματιδίων. Μία απλοποιημένη μορφή της ροής εργασιών για την κατηγοριοποίηση των υλικών κατά την τεχνική read-across προτάθηκε από τους Lamon *et al*.[43] και περιλαμβάνει τα εξής βήματα: 1. Τον χαρακτηρισμό των νανομορφών, 2. Τη συλλογή δεδομένων, την αξιολόγηση τους και τη δημιουργία ενός πίνακα δεδομένων, 3. Την ανάπτυξη υπόθεσης ομαδοποίησης και 4. Την αξιολόγηση της υπόθεσης ομαδοποίησης και την συμπλήρωση κενών στα δεδομένα. Η ανάπτυξη υπόθεσης στις μεθοδολογίες read-across μπορεί να περιλαμβάνει την επιλογή σημαντικών μεταβλητών που είναι χρήσιμες για την πρόβλεψη της μεταβλητής απόκρισης και την επιλογή των «πηγαίων» νανοσωματιδίων που μπορούν να θεωρηθούν ικανά για την πρόβλεψη των νανοσωματιδίων «στόχων». Η διαδικασία αυτή περιλαμβάνει επαναλήψεις δοκιμής και σφάλματος μέχρι να βρεθεί η καλύτερη υπόθεση, συνεπώς είναι χρονοβόρα και ενδεχομένως να μην επιλέγει την καλύτερη υπόθεση ομαδοποίησης.[16,43]

Σκοπός της συγκεκριμένης εργασίας είναι η ανάπτυξη μίας μεθοδολογίας read-across η οποία αυτοματοποιεί τη διαδικασία επιλογής της βέλτιστης υπόθεσης ομαδοποίησης. Βασίζεται στην ανάπτυξη ενός προβλήματος μικτού ακέραιου γραμμικού προγραμματισμού που επιλέγει μία μεταβλητή την οποία θεωρεί μεταβλητή διχοτόμησης, κατά την οποία τα δεδομένα χωρίζονται σε περιοχές, δηλαδή ομαδοποιούνται σε κατηγορίες και στη συνέχεια γίνεται η πρόβλεψη των ανεπιθύμητων ιδιοτήτων σε κάθε κατηγορία νανοσωματιδίων.

4.1 Μαθηματικό Μοντέλο

Στην παρούσα Εργασία χρησιμοποιείται μία καινοτόμος μέθοδος τμηματικής γραμμικής παλινδρόμησης η οποία βασίζεται στην δημοσιευμένη εργασία των Cardoso-Silva *et al*. (2018)[44]. Κύρια ιδέα της μεθόδου είναι η επιλογή μίας μεταβλητής εισόδου από το σύνολο $\boldsymbol{X}$, η οποία χρησιμοποιείται για να χωρίσει το σύνολο των διαθέσιμων δειγμάτων (δηλαδή το πεδίο ορισμού) σε συνεχόμενες περιοχές (regions). Από κάθε περιοχή διέρχεται μία ξεχωριστή εξίσωση γραμμικής παλινδρόμησης. Η διαμέριση των δειγμάτων στις περιοχές και οι συντελεστές παλινδρόμησης για κάθε περιοχή υπολογίζονται ταυτόχρονα μέσα από μία διαδικασία που ελαχιστοποιεί μία αντικειμενική συνάρτηση η οποία εξαρτάται από έναν όρο σφαλμάτων και έναν όρο ομαλοποίησης και οδηγεί στη λύση με το μικρότερο σφάλμα πρόβλεψης.[44,45]

Ο αλγόριθμος βέλτιστης τμηματικής γραμμικής παλινδρόμησης *(Optimal Piecewise Linear Regression Algorithm, OPLRA)* που αποτελείται από μία γραμμική αντικειμενική συνάρτηση, συνεχείς και δυαδικές μεταβλητές και διάφορους γραμμικούς περιορισμούς λύνει προβλήματα Μικτού Ακέραιου Γραμμικού Προγραμματισμού χωρίζοντας το πεδίο ορισμού σε περιοχές στις οποίες η έξοδος υπολογίζεται από μια μοναδική γραμμική εξίσωση. Αρχικά, εφαρμόζεται γραμμική παλινδρόμηση για όλο το σύνολο δεδομένων ώστε να βρεθεί ένα αρχικό σφάλμα πρόβλεψης. Ο αλγόριθμος συνεχίζει επιλέγοντας σε κάθε επανάληψη μία μεταβλητή εισόδου του συνόλου δεδομένων ως μεταβλητή διχοτόμησης για δύο περιοχές (regions=2) και επιλέγεται η μεταβλητή διχοτόμησης η οποία δίνει το μικρότερο σφάλμα πρόβλεψης. Για τη μεταβλητή που επιλέχθηκε, ο αριθμός των περιοχών που χωρίζονται τα δείγματα αυξάνεται σε κάθε επανάληψη, μέχρι το σφάλμα πρόβλεψης μεταξύ των επαναλήψεων να γίνει μικρότερο από μία τιμή που ορίζεται εξωτερικά.

4.1.1 Επίλυση σε μία διάσταση

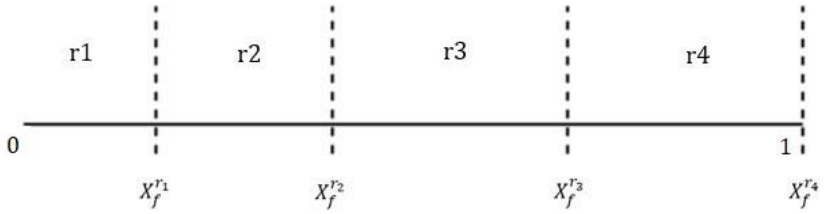
Οι δείκτες, παράμετροι και μεταβλητές που χρησιμοποιούνται στο μοντέλο που αναπτύχθηκε παρουσιάζονται παρακάτω:

Δείκτες	
s	*Δείγματα s=1,2,…,*S
f	*Μεταβλητές εισόδου/ιδιότητες f=1,2,…,*F
r	*Περιοχές διαμέρισης r=1,2,…,*R
f^*	*Μεταβλητή διχοτόμησης*
Παράμετροι	
A_{sf}	*Τιμή μεταβλητής f για το δείγμα s*
Y_s	*Τιμή εξόδου για το δείγμα s*
U, U'	*Αυθαίρετα μεγάλοι θετικοί αριθμοί*
λ	*Θετική παράμετρος για έλεγχο επίδρασης ομαλοποίησης*
β	*Παράμετρος απόκλισης σφαλμάτων μεταξύ επαναλήψεων με τιμή 0 έως 1 για αύξηση περιοχών r*
ε	*Θετικός αριθμός που εκφράζει την ελάχιστη διαφορά μεταξύ των σημείων καμπής*
Συνεχείς μεταβλητές	
X_{f*}^r	*Συντεταγμένες σημείου καμπής (break-point) μεταβλητής f^**
W_f^r	*Συντελεστές παλινδρόμησης για τη μεταβλητή f στην περιοχή r*
B^r	*Αποτέμνουσα στην περιοχή r*
$Pred_s^r$	*Προβλεπόμενη τιμή εξόδου για το δείγμα s στην περιοχή r*
E_s	*Σφάλμα τιμής πρόβλεψης και πειραματικής τιμής εξόδου για το δείγμα s*
E_s^r	*Σφάλμα τιμής πρόβλεψης και πειραματικής τιμής εξόδου για το δείγμα s στην περιοχή r*

Δυαδικές μεταβλητές

F_s^r	*1: αν το δείγμα s βρίσκεται στην περιοχή r*
	0: αν το δείγμα s δε βρίσκεται στην περιοχή r

Σε κάθε επανάληψη η μεταβλητή διχοτόμησης f^* και ο αριθμός των περιοχών r έχουν σταθερές τιμές. Τα σημεία διαμέρισης είναι συνεχή όπως φαίνεται από την Εξίσωση [4.1] και το Σχήμα 4.1 και απέχουν μεταξύ τους απόσταση μεγαλύτερη ή ίση με την ελάχιστη διαφορά ε.



Σχήμα 4.1: Διαμέριση του πεδίου ορισμού σε περιοχές.

$$X_{f*}^{r} \geq X_{f*}^{r-1} + \varepsilon \qquad \forall r \in \{2,3,\dots,\mathrm{R}\} \tag{4.1}$$

Η Εξίσωση [4.2] εξασφαλίζει ότι το πεδίο ορισμού θα χωριστεί σε περιοχές καθώς το πρώτο σημείο καμπής θα είναι μεγαλύτερο από την τιμή ε ώστε να τοποθετηθεί τουλάχιστον ένα δείγμα στη συγκεκριμένη περιοχή (το δείγμα με τιμή $A_{sf^*} = 0$), ενώ η Εξίσωση [4.3] εξασφαλίζει ότι το τελικό σημείο καμπής θα παίρνει την τιμή 1, συμπεριλαμβάνοντας και το δείγμα με τη μέγιστη τιμή 1, καθώς τα δεδομένα μας είναι κανονικοποιημένα σε εύρος τιμών [0,1].

$$X_{f*}^{r} \geq \varepsilon \qquad \gamma\iota\alpha\ r = 1 \tag{4.2}$$

$$X_{f*}^{r} = 1 \qquad \gamma\iota\alpha\ r = \mathrm{R} \tag{4.3}$$

Η Εξίσωση [4.4] δείχνει ότι ένα δείγμα μπορεί να ανήκει μόνο σε μία περιοχή. Κάθε δείγμα s θα βρίσκεται μόνο σε μία περιοχή r στην οποία θα ισχύει $F_s^r = 1$, ενώ σε όλες τις υπόλοιπες θα ισχύει $F_s^r = 0$. Επομένως το άθροισμα του F_s^r για κάθε δείγμα και σε όλες τις περιοχές διαμέρισης $r \in \{1,2,\dots,\mathrm{R}\}$ θα παίρνει την τιμή 1.

$$\sum_{r=1}^{\mathrm{R}} F_s^r = 1 \qquad \forall s \in \{1,2,\dots,\mathrm{S}\} \tag{4.4}$$

Οι Εξισώσεις [4.5] και [4.6] τοποθετούν τα δείγματα στις περιοχές ανάλογα με τα σημεία καμπής. Μεταξύ δύο διαδοχικών περιοχών $r-1$ και r, στην περίπτωση που η τιμή A_{sf^*} της μεταβλητής f^* του δείγματος s έχει τιμή μεγαλύτερη από την τιμή του σημείου καμπής της περιοχής διαμέρισης $r-1$ και μικρότερη του σημείου καμπής της περιοχής

r, τότε θα τοποθετήσει το δείγμα μεταξύ των δύο σημείων καμπής X_{f*}^{r-1} και X_{f*}^{r}. Σε αυτή την περιοχή θα ισχύει ότι $F_s^r = 1$ ενώ σε αντίθετη περίπτωση θα πάρει την τιμή 0.[i]

$$A_{sf^*} \geq X_{f*}^{r-1} + \varepsilon - U(1 - F_s^r) \qquad \forall s \in \{1,2,\dots,\mathrm{S}\}, r \in \{2,3,\dots,\mathrm{R}\} \qquad [4.5]$$

$$A_{sf^*} \leq X_{f*}^{r} - \varepsilon + U(1 - F_s^r) \qquad \forall s \in \{1,2,\dots,\mathrm{S}\}, r \in \{1,2,\dots,\mathrm{R}-1\} \qquad [4.6]$$

Η Εξίσωση [4.7] δίνει την προβλεπόμενη τιμή εξόδου, η οποία είναι της μορφής $y = \beta_1 x + \beta_0$, όπου W_f^r οι συντελεστές παλινδρόμησης για κάθε μεταβλητή f σε κάθε περιοχή r και B^r η αποτέμνουσα για κάθε περιοχή.

$$Pred_s^r = \sum_{f=1}^{\mathrm{F}} W_f^r A_{sf} + B^r \qquad \forall s \in \{1,2,\dots,\mathrm{S}\}, r \in \{1,2,\dots,\mathrm{R}\} \qquad [4.7]$$

Το σφάλμα πρόβλεψης δείγματος, E_s, ορίζεται ως μη αρνητικό (Εξίσωση [4.8]) και μεγαλύτερο ή ίσο από το αντίστοιχο σφάλμα του δείγματος σε μία συγκεκριμένη περιοχή, E_s^r (Εξίσωση [4.12]). Το σφάλμα πρόβλεψης δείγματος σε συγκεκριμένη περιοχή έχει ανώτερο όριο έναν μεγάλο θετικό αριθμό στην περίπτωση που το δείγμα ανήκει στην περιοχή ή την τιμή 0 εάν δεν ανήκει στην περιοχή (Εξίσωση [4.9]). Επιπλέον, υπολογίζεται στις Εξισώσεις [4.10] και [4.11] για κάθε δείγμα s σε κάθε περιοχή r, από τις πειραματικές τιμές Y_s, τις τιμές πρόβλεψης $Pred_s^r$ και την ύπαρξη ή μη του κάθε δείγματος στη συγκεκριμένη περιοχή.

Πιο συγκεκριμένα, στην περίπτωση που το δείγμα δεν ανήκει στην περιοχή r, δηλαδή για $F_s^r = 0$, τότε το απόλυτο σφάλμα E_s^r θα είναι μικρότερο από το 0, όπως θα προκύπτει από την Εξίσωση [4.9] και μεγαλύτερο από έναν μεγάλο αρνητικό αριθμό U' όπως προκύπτει από τις Εξισώσεις [4.10] και [4.11]. Ο συνδυασμός των Εξισώσεων [4.8] έως [4.12] θα «εξαναγκάσει» το σφάλμα πρόβλεψης του δείγματος να πάρει τιμή μεγαλύτερη ή ίση με 0. Εάν το δείγμα ανήκει στη συγκεκριμένη περιοχή, $F_s^r = 1$, τότε το απόλυτο σφάλμα E_s^r θα είναι μεγαλύτερο ή ίσο με τη διαφορά μεταξύ προβλεπόμενης και πραγματικής τιμής, ενώ οι Εξισώσεις [4.8] και [4.12] θα οδηγούν την τιμή του σφάλματος πρόβλεψης του συγκεκριμένου δείγματος να πάρει τιμή μεγαλύτερη ή ίση με την απόλυτη διαφορά μεταξύ προβλεπόμενης και πραγματικής τιμής. Λόγω της απαίτησης ελαχιστοποίησης της αντικειμενικής συνάρτησης, τα σφάλματα πρόβλεψης των δειγμάτων ανάλογα με την ύπαρξη ή μη του δείγματος στη συγκεκριμένη περιοχή, θα λαμβάνουν την τιμή της απόλυτης διαφοράς προβλεπόμενης και πραγματικής τιμής ή την τιμή 0 αντίστοιχα.

[i] Στο σημείο αυτό (και στους περιορισμούς που περιλαμβάνουν την παράμετρο U') εμφανίζεται μια άλλη χρήση του μετασχηματισμού bigM. Σε αυτές τις περιπτώσεις ο «αρκούντως μεγάλος» θετικός αριθμός, εξασφαλίζει την ισότητα των μεταβλητών που συμμετέχουν στους αντίστοιχους περιορισμούς (συνήθως εμφανίζονται σε ζεύγη) μόνο όταν μια συγκεκριμένη δυαδική μεταβλητή παίρνει μια τιμή, αλλά αφήνει τις μεταβλητές «ανοιχτές» εάν η δυαδική μεταβλητή παίρνει την αντίθετη τιμή.

$$E_s \geq 0 \qquad \forall s \in \{1,2,\dots,\mathrm{S}\} \qquad [4.8]$$

$$E_s^r \leq U' F_s^r \qquad \forall s \in \{1,2,\dots,\mathrm{S}\}, r \in \{1,2,\dots,\mathrm{R}\} \qquad [4.9]$$

$$E_s^r \geq Y_s - Pred_s^r - U'(1 - F_s^r) \qquad \forall s \in \{1,2,\dots,\mathrm{S}\}, r \in \{1,2,\dots,\mathrm{R}\} \qquad [4.10]$$

$$E_s^r \geq Pred_s^r - Y_s - U'(1 - F_s^r) \qquad \forall s \in \{1,2,\dots,\mathrm{S}\}, r \in \{1,2,\dots,\mathrm{R}\} \qquad [4.11]$$

$$E_s \geq E_s^r \qquad \forall s \in \{1,2,\dots,\mathrm{S}\}, r \in \{1,2,\dots,\mathrm{R}\} \qquad [4.12]$$

Για το μοντέλο μαθηματικού προγραμματισμού που αναπτύχθηκε, η αντικειμενική συνάρτηση που ελαχιστοποιείται περιλαμβάνει δύο όρους, το μέσο απόλυτο σφάλμα *MAE* και έναν όρο ομαλοποίησης *REG* (regularization) που υπολογίζεται ως το άθροισμα των απόλυτων τιμών των συντελεστών παλινδρόμησης και μειώνει τον κίνδυνο υπερπροσαρμογής του μοντέλου (overfitting). Οι τιμές *MAE* και *REG* υπολογίζονται από τις εξισώσεις [4.13] και [4.14]:

$$MAE = \frac{\sum_{s=1}^{\mathrm{S}} E_s}{|\mathrm{S}|} \qquad [4.13]$$

$$REG = \sum_{r=1}^{\mathrm{R}} \sum_{f=1}^{\mathrm{F}} W_f^{r+} \qquad [4.14]$$

Όπου οι απόλυτες τιμές των συντελεστών παλινδρόμησης W_f^r υπολογίζονται από τις εξισώσεις [4.15] και [4.16]:

$$W_f^{r+} \geq W_f^r \qquad \forall r \in \{1,2,\dots,\mathrm{R}\}, f \in \{1,2,\dots,\mathrm{F}\} \qquad [4.15]$$

$$W_f^{r+} \geq -W_f^r \qquad \forall r \in \{1,2,\dots,\mathrm{R}\}, f \in \{1,2,\dots,\mathrm{F}\} \qquad [4.16]$$

Οι παραπάνω εξισώσεις επιλύονται για ελαχιστοποίηση της τιμής της αντικειμενικής συνάρτησης [4.17].

$$z = MAE + \lambda\, REG \qquad [4.17]$$

όπου λ, μία παράμετρος που ορίζεται εξωτερικά από το χρήστη και ελέγχει την επίδραση της ομαλοποίησης.

Εκτός από την μείωση του κινδύνου υπερπροσαρμογής του μοντέλου, ο όρος της ομαλοποίησης έχει επίδραση και στην επιλογή των μεταβλητών. Με μηδενικό όρο ομαλοποίησης (λ=0), οι τιμές των συντελεστών παλινδρόμησης, που ορίζουν την τιμή REG (Εξίσωση [4.14]) μπορεί να πάρουν υψηλές τιμές, με αποτέλεσμα ακόμα και μικρές αποκλίσεις από δεδομένα εκπαίδευσης να οδηγούν σε μεγάλα σφάλματα πρόβλεψης. Το

μοντέλο που προκύπτει είναι υπερπροσαρμοσμένο στα δεδομένα εκπαίδευσης και δεν παρέχει ικανοποιητικές προβλέψεις για τα δεδομένα δοκιμών. Ωστόσο, όταν εφαρμόζεται ο όρος ομαλοποίησης, επιβάλλονται μικρότερες τιμές στους συντελεστές παλινδρόμησης με αποτέλεσμα να έχουν μικρότερη επίδραση στην ακρίβεια των προβλέψεων. Επιπλέον, με την ομαλοποίηση αναμένεται να μειωθούν οι μεταβλητές που επιλέγονται, υποδεικνύοντας τις σημαντικές για την πρόβλεψη μεταβλητές.

4.1.2 Επίλυση σε δύο διαστάσεις

Σε επόμενο βήμα, λόγω των πολλαπλών ιδιοτήτων των νανοσωματιδίων, γίνεται η θεώρηση ότι οι μεταβλητές εισόδου χωρίζονται σε δύο κατηγορίες, M και N. Επομένως, οι περιοχές διαμέρισης διασπώνται σε δύο διαστάσεις, όπου η μία διάσταση αφορά τις μεταβλητές F_m και η άλλη τις μεταβλητές F_n.

Οι δείκτες, παράμετροι και μεταβλητές που χρησιμοποιούνται στο μοντέλο των δύο διαστάσεων παρουσιάζονται παρακάτω:

Δείκτες

s	*Δείγματα s=1,2,…,*S
f	*Μεταβλητές εισόδου/ιδιότητες f=1,2,…,*F
f_m	*Μεταβλητές εισόδου/ιδιότητες* f_m*=1,2,…,*F_m
f_n	*Μεταβλητές εισόδου/ιδιότητες* f_n*=1,2,…,*F_n
r_m	*Περιοχές διαμέρισης* r_m*=1,2,…,*R_m
r_n	*Περιοχές διαμέρισης* r_n*=1,2,…,*R_n
f_m^*	*Μεταβλητή διχοτόμησης στη διάσταση m*
f_n^*	*Μεταβλητή διχοτόμησης στη διάσταση n*

Παράμετροι

A_{sf}	*Τιμή μεταβλητής f για το δείγμα s*
$A_{M_{sf_m}}$	*Τιμή μεταβλητής* f_m *για το δείγμα s*
$A_{N_{sf_n}}$	*Τιμή μεταβλητής* f_n *για το δείγμα s*
Y_s	*Τιμή εξόδου για το δείγμα s*
U, U'	*Αυθαίρετα μεγάλοι θετικοί αριθμοί*
λ	*Θετική παράμετρος για επίδραση ομαλοποίησης*
β	*Παράμετρος απόκλισης σφαλμάτων μεταξύ επαναλήψεων με τιμές 0 έως 1 για αύξηση των περιοχών r*
ε	*Θετικός αριθμός που εκφράζει την ελάχιστη διαφορά μεταξύ των σημείων καμπής*

Συνεχείς μεταβλητές

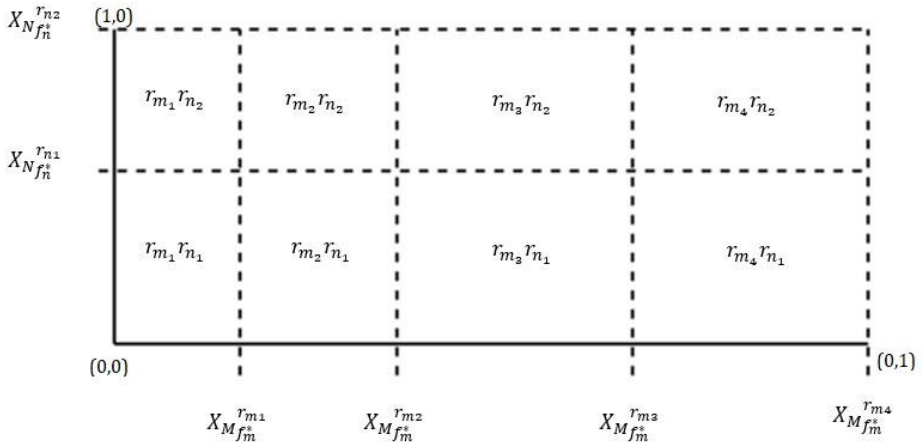
$X_{M_{f_m^*}}^{r_m}$	*Σημείο καμπής (break-point) μεταβλητής* f_m***
$X_{N_{f_n^*}}^{r_n}$	*Σημείο καμπής (break-point) μεταβλητής* f_n***
$W_f^{r_m r_n}$	*Συντελεστές παλινδρόμησης για τη μεταβλητή f στην περιοχή* $r_m r_n$
$B^{r_m r_n}$	*Αποτέμνουσα στην περιοχή* $r_m r_n$
$Pred_s^{r_m r_n}$	*Προβλεπόμενη τιμή εξόδου για το δείγμα s στην περιοχή* $r_m r_n$
E_s	*Σφάλμα τιμής πρόβλεψης και πειραματικής τιμής εξόδου για το δείγμα s*

$E_s^{r_m}$	*Σφάλμα τιμής πρόβλεψης και πειραματικής τιμής εξόδου για το δείγμα s στην περιοχή r της διάστασης m*
$E_s^{r_n}$	*Σφάλμα τιμής πρόβλεψης και πειραματικής τιμής εξόδου για το δείγμα s στην περιοχή r της διάστασης n*

Δυαδικές μεταβλητές

$F_s^{r_m}$	*1: αν το δείγμα s βρίσκεται στην περιοχή r της διάστασης m* *0: αν το δείγμα s δε βρίσκεται στην περιοχή r της διάστασης m*
$F_s^{r_n}$	*1: αν το δείγμα s βρίσκεται στην περιοχή r της διάστασης n* *0: αν το δείγμα s δε βρίσκεται στην περιοχή r της διάστασης n*

Σε κάθε επανάληψη οι μεταβλητές διχοτόμησης f_m^* και f_n^* και ο αριθμός των περιοχών σε κάθε διάσταση έχουν σταθερές τιμές. Τα σημεία διαμέρισης είναι συνεχή όπως φαίνεται στο Σχήμα 4.2 και τις Εξισώσεις [4.18] και [4.19] και απέχουν μεταξύ τους απόσταση μεγαλύτερη ή ίση με ε.



Σχήμα 4.2: Περιοχές διαμέρισης για δύο διαστάσεις m και n.

$$X_{M_{f_m^*}}^{r_m} \geq X_{M_{f_m^*}}^{r_m-1} + \varepsilon \qquad \forall r_m \in \{2,3,\dots,\mathrm{R}_m\} \tag{4.18}$$

$$X_{N_{f_n^*}}^{r_n} \geq X_{N_{f_n^*}}^{r_n-1} + \varepsilon \qquad \forall r_n \in \{2,3,\dots,\mathrm{R}_n\} \tag{4.19}$$

Οι Εξισώσεις [4.20] και [4.21] εξασφαλίζουν ότι το πεδίο ορισμού θα χωριστεί σε περιοχές καθώς το πρώτο σημείο καμπής σε κάθε διάσταση θα παίρνει τιμή μεγαλύτερη ή ίση με ε, επομένως θα τοποθετεί τουλάχιστον ένα δείγμα στην πρώτη περιοχή διαμέρισης. Οι Εξισώσεις [4.22] και [4.23] ορίζουν το τελικό σημείο καμπής ίσο με 1, ώστε να συμπεριληφθούν στη τελευταία περιοχή τα δείγματα με τις μέγιστες τιμές μεταβλητών, καθώς τα δεδομένα μας είναι κανονικοποιημένα στο εύρος [0,1].

$$X_{M_{f_m^*}}^{r_m} \geq \varepsilon \qquad \gamma\iota\alpha\ r_m = 1 \tag{4.20}$$

$$X_{N_{f_n^*}}^{r_n} \geq \varepsilon \qquad \gamma\iota\alpha\ r_n = 1 \tag{4.21}$$

$$X_{M_{f_m^*}}^{r_m} = 1 \qquad \gamma\iota\alpha\ r_m = \mathrm{R}_m \tag{4.22}$$

$$X_{N_{f_n^*}}^{r_n} = 1 \qquad \gamma\iota\alpha\ r_n = \mathrm{R}_n \tag{4.23}$$

Οι Εξισώσεις [4.24] και [4.25] δείχνουν ότι ένα δείγμα μπορεί να ανήκει μόνο σε μία περιοχή ανά διάσταση. Σε αναλογία με την εξίσωση [4.4], για κάθε δείγμα s που ανήκει στην περιοχή r_m ή r_nθα ισχύει $F_s^{r_m} = 1$ ή $F_s^{r_n} = 1$ αντίστοιχα ενώ στις υπόλοιπες περιοχές της κάθε διάστασης η μεταβλητή F_s^r θα παίρνει την τιμή 0. Έτσι, το άθροισμα των περιοχών σε κάθε διάσταση θα ισούται με 1.

$$\sum_{r_m}^{\mathrm{R}_m} F_s^{r_m} = 1 \qquad \forall s \in \{1,2,\dots,\mathrm{S}\} \tag{4.24}$$

$$\sum_{r_n}^{\mathrm{R}_n} F_s^{r_n} = 1 \qquad \forall s \in \{1,2,\dots,\mathrm{S}\} \tag{4.25}$$

Οι Εξισώσεις [4.26] και [4.27] τοποθετούν τα δείγματα στις περιοχές της διάστασης m, ανάλογα με τα σημεία καμπής, ενώ οι Εξισώσεις [4.28] και [4.29] τοποθετούν τα δείγματα στις περιοχές της διάστασης n, ανάλογα με τα σημεία καμπής. Σε αναλογία με τις Εξισώσεις [4.5] και [4.6], μεταξύ δύο διαδοχικών περιοχών διαμέρισης της κάθε διάστασης, τα δείγματα τοποθετούνται σε περιοχές ανάλογα με την γνωστή τιμή της κάθε μεταβλητής. Αν η τιμή αυτή είναι μικρότερη από το σημείο καμπής της περιοχής r και μεγαλύτερη του σημείο καμπής της περιοχής $r-1$ τότε το δείγμα θα ανήκει στην περιοχή r της αντίστοιχης διάστασης και θα ισχύει $F_s^{r_m} = 1$ ή $F_s^{r_n} = 1$. Εάν δεν ανήκει, τότε θα προκύπτει $F_s^{r_m} = 0$ ή $F_s^{r_n} = 0$.

$$A_{M_{sf_m^*}} \geq X_{M_{f_m^*}}^{r_m - 1} + \varepsilon - U\left(1 - F_s^{r_m}\right) \qquad \forall s \in \{1,2,\dots,\mathrm{S}\}, r_m \in \{2,3,\dots,\mathrm{R}_m\} \tag{4.26}$$

$$A_{M_{sf_m^*}} \leq X_{M_{f_m^*}}^{r_m} - \varepsilon + U\left(1 - F_s^{r_m}\right) \qquad \forall s \in \{1,2,\dots,\mathrm{S}\}, r_m \in \{1,2,\dots,\mathrm{R}_m - 1\} \tag{4.27}$$

$$A_{N_{sf_n^*}} \geq X_{N_{f_n^*}}^{r_n - 1} + \varepsilon - U\left(1 - F_s^{r_n}\right) \qquad \forall s \in \{1,2,\dots,\mathrm{S}\}, r_n \in \{2,3,\dots,\mathrm{R}_n\} \tag{4.28}$$

$$A_{N_{sf_n^*}} \leq X_{N_{f_n^*}}^{r_n} - \varepsilon + U\left(1 - F_s^{r_n}\right) \qquad \forall s \in \{1,2,\dots,\mathrm{S}\}, r_n \in \{1,2,\dots,\mathrm{R}_n - 1\} \tag{4.29}$$

Η εξίσωση [4.30] δίνει την προβλεπόμενη τιμή εξόδου, η οποία είναι της μορφής $y = \beta_1 x + \beta_0$, όπου $W_f^{r_m r_n}$ οι συντελεστές παλινδρόμησης και $B^{r_m r_n}$ η αποτέμνουσα για την περιοχή $r_m r_n$.

$$Pred_s^{r_m r_n} = \sum_{f=1}^{\mathrm{F}} W_f^{r_m r_n} A_{sf} + B^{r_m r_n} \qquad [4.30]$$
$$\forall s \in \{1,2,\dots,\mathrm{S}\}, r_m \in \{1,2,\dots,\mathrm{R}_m\}, r_n \in \{1,2,\dots,\mathrm{R}_n\}$$

Το σφάλμα πρόβλεψης δείγματος, E_s, ορίζεται ως μη αρνητικό (Εξίσωση [4.31]) και μεγαλύτερο ή ίσο από το αντίστοιχο σφάλμα του δείγματος σε μία συγκεκριμένη περιοχή της κάθε διάστασης, $E_s^{r_m}$ ή $E_s^{r_n}$. (Εξισώσεις [4.38], [4.39]). Τα σφάλματα πρόβλεψης δείγματος σε συγκεκριμένη περιοχή της κάθε διάστασης, έχουν ανώτερα όρια έναν μεγάλο θετικό αριθμό στην περίπτωση που το δείγμα ανήκει στην περιοχή της αντίστοιχης διάστασης ή την τιμή 0 εάν δεν ανήκει στην συγκεκριμένη περιοχή (Εξισώσεις [4.32], [4.33]). Επιπλέον, για τις περιοχές της διάστασης m το σφάλμα πρόβλεψης υπολογίζεται στις Εξισώσεις [4.34] και [4.35] για κάθε δείγμα s από τις πειραματικές τιμές Y_s, τις τιμές πρόβλεψης $Pred_s^{r_m r_n}$ και την ύπαρξη ή μη του κάθε δείγματος στη συγκεκριμένη περιοχή. Αντίστοιχα, υπολογίζεται το σφάλμα πρόβλεψης δείγματος σε συγκεκριμένη περιοχή της διάστασης n στις Εξισώσεις [4.36], [4.37].

Στην περίπτωση που το δείγμα s ανήκει στην περιοχή $r_m r_n$ τότε $F_s^{r_m} = 1$ και $F_s^{r_n} = 1$. Οι Εξισώσεις [4.32], [4.33] περιορίζουν τα σφάλματα της συγκεκριμένης περιοχής μικρότερα από έναν μεγάλο θετικό αριθμό ενώ οι Εξισώσεις [4.34] έως [4.37] υπολογίζουν τα σφάλματα μεγαλύτερα ή ίσα την απόλυτη διαφορά μεταξύ πραγματικής και προβλεπόμενης τιμής. Από τις Εξισώσεις [4.31], [4.38], [4.39], το απόλυτο σφάλμα πρόβλεψης δείγματος προκύπτει μεγαλύτερο ή ίσο της διαφοράς μεταξύ πραγματικής και προβλεπόμενης τιμής. Στην περίπτωση που το δείγμα s δεν ανήκει στην περιοχή $r_m r_n$ τότε $F_s^{r_m} = 0$ και $F_s^{r_n} = 0$ τότε από τις Εξισώσεις [4.32] έως [4.37] τα σφάλματα πρόβλεψης στη συγκεκριμένη περιοχή παίρνουν τιμές μικρότερες ή ίσες του 0. Οι Εξισώσεις [4.31], [4.38], [4.39] εξασφαλίζουν ότι το σφάλμα πρόβλεψης του δείγματος, εφόσον δεν ανήκει στην περιοχή $r_m r_n$ θα πάρει τιμή μεγαλύτερη ή ίση από το 0. Λόγω της απαίτησης ελαχιστοποίησης της αντικειμενικής συνάρτησης, ανάλογα με την ύπαρξη ή μη του δείγματος στη συγκεκριμένη περιοχή, το σφάλμα πρόβλεψης θα «εξαναγκάζεται» να πάρει την τιμή μηδέν ή την απόλυτη διαφορά πραγματικής και προβλεπόμενης τιμής.

$$E_s \geq 0 \qquad \forall s \in \{1,2,\dots,\mathrm{S}\} \qquad [4.31]$$

$$E_s^{r_m} \leq U' F_s^{r_m} \qquad \forall s \in \{1,2,\dots,\mathrm{S}\}, r_m \in \{1,2,\dots,\mathrm{R}_m\} \qquad [4.32]$$

$$E_s^{r_n} \leq U' F_s^{r_n} \qquad \forall s \in \{1,2,\dots,\mathrm{S}\}, r_n \in \{1,2,\dots,\mathrm{R}_n\} \qquad [4.33]$$

$$E_s^{r_m} \geq Y_s - Pred_s^{r_m r_n} - U'\left(1 - F_s^{r_m}\right) - U'\left(1 - F_s^{r_n}\right)$$
$$\forall s \in \{1,2,\dots,\mathrm{S}\}, r_m \in \{1,2,\dots,\mathrm{R}_m\}, r_n \in \{1,2,\dots,\mathrm{R}_n\} \qquad [4.34]$$

$$E_s^{r_m} \geq Pred_s^{r_m r_n} - Y_s - U'(1 - F_s^{r_m}) - U'\left(1 - F_s^{r_n}\right)$$
$$\forall s \in \{1,2,\dots,\mathrm{S}\}, r_m \in \{1,2,\dots,\mathrm{R}_m\}, r_n \in \{1,2,\dots,\mathrm{R}_n\} \qquad [4.35]$$

$$E_s^{r_n} \geq Y_s - Pred_s^{r_m r_n} - U'\left(1 - F_s^{r_m}\right) - U'\left(1 - F_s^{r_n}\right)$$
$$\forall s \in \{1,2,\dots,\mathrm{S}\}, r_m \in \{1,2,\dots,\mathrm{R}_m\}, r_n \in \{1,2,\dots,\mathrm{R}_n\} \quad [4.36]$$

$$E_s^{r_n} \geq Pred_s^{r_m r_n} - Y_s - U'(1 - F_s^{r_m}) - U'\left(1 - F_s^{r_n}\right)$$
$$\forall s \in \{1,2,\dots,\mathrm{S}\}, r_m \in \{1,2,\dots,\mathrm{R}_m\}, r_n \in \{1,2,\dots,\mathrm{R}_n\} \quad [4.37]$$

$$E_s \geq E_s^{r_m} \qquad \forall s \in \{1,2,\dots,\mathrm{S}\}, r_m \in \{1,2,\dots,\mathrm{R}_m\} \quad [4.38]$$

$$E_s \geq E_s^{r_n} \qquad \forall s \in \{1,2,\dots,\mathrm{S}\}, r_n \in \{1,2,\dots,\mathrm{R}_n\} \quad [4.39]$$

Για το μοντέλο δύο διαστάσεων, η αντικειμενική συνάρτηση που ελαχιστοποιείται περιλαμβάνει τους δύο όρους που αναφέρθηκαν και στη μία διάσταση, το μέσο απόλυτο σφάλμα *MAE* και τον όρο ομαλοποίησης *REG*. Οι τιμές *MAE* και *REG* υπολογίζονται από τις εξισώσεις [4.13] και [4.40]:

$$MAE = \frac{\sum_{s=1}^{S} E_s}{|\mathrm{S}|} \quad [4.13]$$

$$REG = \sum_{r_m=1}^{\mathrm{R}_m} \sum_{r_n=1}^{\mathrm{R}_n} \sum_{f=1}^{\mathrm{F}} W_f^{r_m r_n +} \quad [4.40]$$

Όπου οι απόλυτες τιμές των συντελεστών παλινδρόμησης $W_f^{r_m r_n}$ υπολογίζονται από τις εξισώσεις [4.41] και [4.42]:

$$W_f^{r_m r_n +} \geq W_f^{r_m r_n} \qquad \forall r_m \in \{1,2,\dots,\mathrm{R}_m\}, r_n \in \{1,2,\dots,\mathrm{R}_n\}, f \in \{1,2,\dots,\mathrm{F}\} \quad [4.41]$$

$$W_f^{r_m r_n +} \geq -W_f^{r_m r_n} \qquad \forall r_m \in \{1,2,\dots,\mathrm{R}_m\}, r_n \in \{1,2,\dots,\mathrm{R}_n\}, f \in \{1,2,\dots,\mathrm{F}\} \quad [4.42]$$

Οι παραπάνω εξισώσεις επιλύονται για ελαχιστοποίηση της τιμής της αντικειμενικής συνάρτησης [4.17].

$$z = MAE + \lambda\, REG \quad [4.17]$$

4.2 Αξιολόγηση προτεινόμενης μεθοδολογίας

Η μέθοδος ανάπτυξης και αξιολόγησης του μοντέλου παρουσιάζεται στο Σχήμα 4.3. Τα δεδομένα υφίστανται μία διαδικασία προεπεξεργασίας που περιλαμβάνει την κανονικοποίηση τους σε εύρη τιμών [0,1]. Στη συνέχεια, χωρίζονται σε δύο υποσύνολα σύμφωνα με τον αλγόριθμο Kennard-Stone (Ενότητα 2.5.1). Το ένα υποσύνολο περιέχει το 75% των αρχικών δεδομένων και αποτελεί το σύνολο εκπαίδευσης, ενώ το υπόλοιπο

25% των δεδομένων αποτελεί το σύνολο δοκιμών. Το σύνολο εκπαίδευσης τροφοδοτεί το μοντέλο με δεδομένα που «χτίζουν» το μοντέλο παλινδρόμησης ενώ το σύνολο δοκιμών χρησιμοποιείται για την αξιολόγησή του.

Αρχικά, εφαρμόζεται η μέθοδος της απλής γραμμικής παλινδρόμησης για τα δεδομένα εκπαίδευσης και το μοντέλο που προκύπτει εφαρμόζεται στο σύνολο δοκιμών. Υπολογίζεται το σφάλμα πρόβλεψης για το σύνολο εκπαίδευσης και το σύνολο δοκιμών. Στη συνέχεια πραγματοποιείται επίλυση του αλγορίθμου βέλτιστης τμηματικής παλινδρόμησης (OPLRA) για το σύνολο εκπαίδευσης με δύο περιοχές διαμέρισης. Σε αυτό το βήμα βρίσκεται η μεταβλητή διχοτόμησης η οποία χωρίζει την περιοχή δεδομένων σε δυο τμήματα με το βέλτιστο τρόπο. Για κάθε μεταβλητή του συνόλου εκπαίδευσης οι παράμετροι του μοντέλου OPLRA με δύο περιοχές διαμέρισης εφαρμόζονται στο σύνολο δοκιμών και υπολογίζεται το σφάλμα πρόβλεψης. Η μεταβλητή από την οποία προκύπτει το μικρότερο σφάλμα πρόβλεψης, επιλέγεται ως βέλτιστη μεταβλητή διχοτόμησης.[ii]

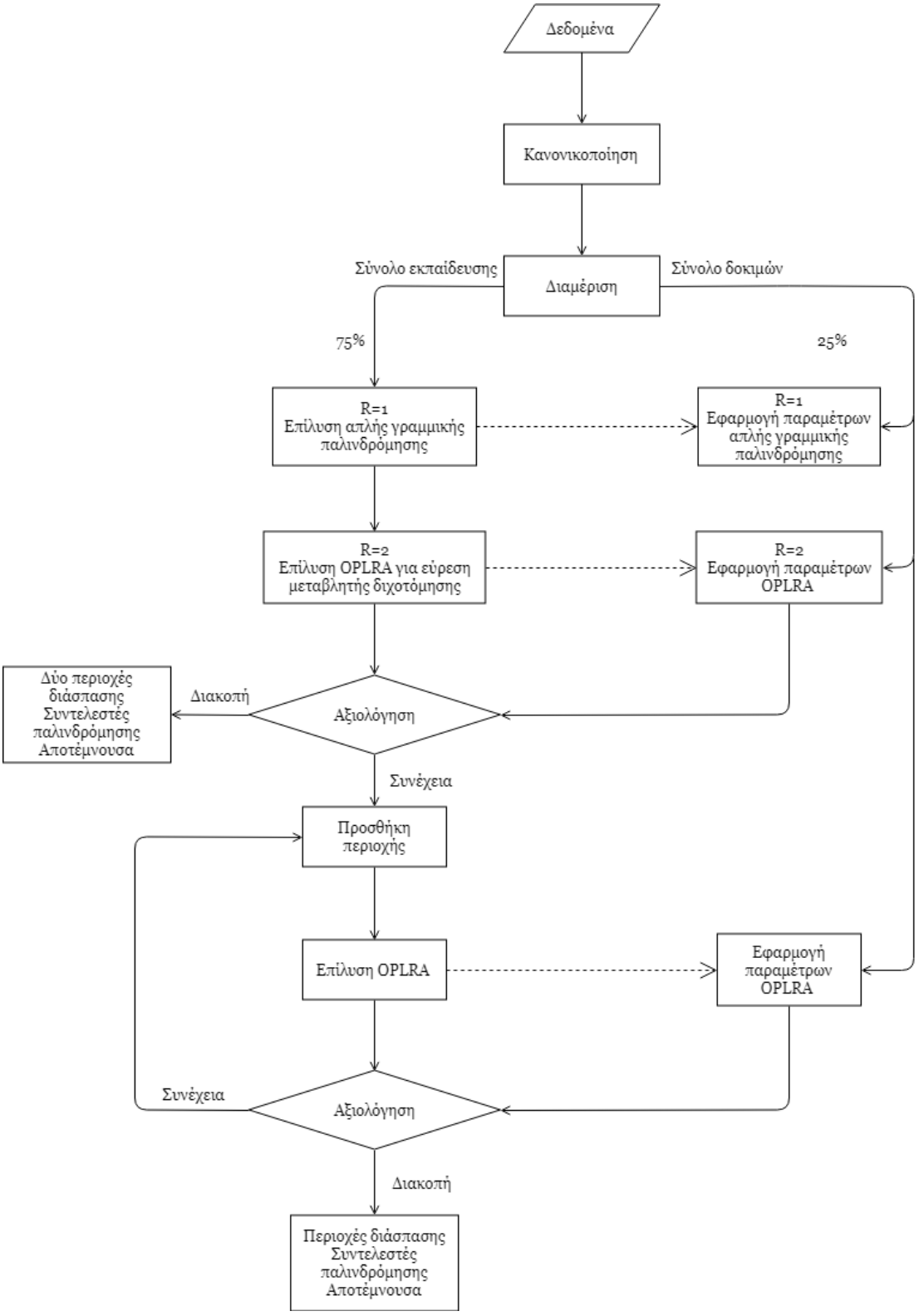
Στη συνέχεια, εάν η μεταβολή των σφαλμάτων για το σύνολο δοκιμών στην περίπτωση μίας και δύο περιοχών διαμέρισης αντίστοιχα, είναι μικρότερη από μία τιμή β που ορίζεται από το χρήστη, τότε προτίθεται άλλη μία περιοχή διαμέρισης.

$$ERROR\_test_R < (1-\beta) ERROR\_test_{R-1} \qquad [4.43]$$

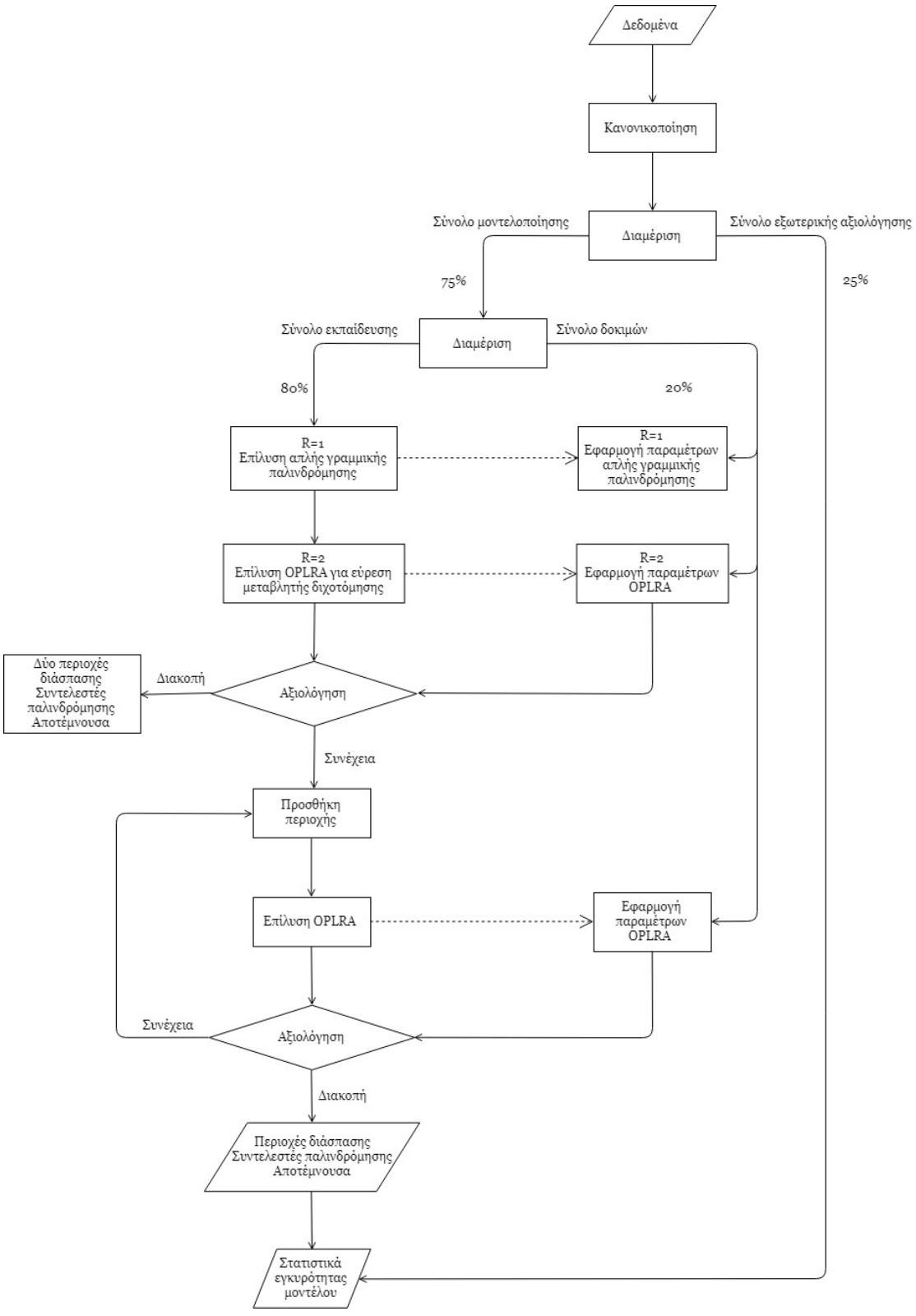
Ο αλγόριθμος βέλτιστης τμηματικής παλινδρόμησης επιλύεται με τα δεδομένα του συνόλου εκπαίδευσης για τρεις περιοχές διαμέρισης και στη συνέχεια εφαρμόζεται στο σύνολο δοκιμών και υπολογίζονται τα αντίστοιχα σφάλματα πρόβλεψης. Ελέγχεται ξανά η συνθήκη των σφαλμάτων για τα νέα σφάλματα πρόβλεψης τα οποία προκύπτουν από τα μοντέλα με δύο και τρεις περιοχές διαμέρισης αντίστοιχα. Εάν ικανοποιείται η συνθήκη, τότε προστίθεται και άλλη μία περιοχή διαμέρισης. Η διαδικασία προσθήκης περιοχών σταματάει όταν δεν ικανοποιείται η συνθήκη των σφαλμάτων και προκύπτει το τελικό μοντέλο τμηματικής παλινδρόμησης.

Για επιπλέον επικύρωση του μοντέλου, τα δεδομένα χωρίζονται δύο φορές, όπως φαίνεται στο Σχήμα 4.4. Στην πρώτη διαμέριση προκύπτει το σύνολο μοντελοποίησης και το σύνολο εξωτερικής αξιολόγησης σύμφωνα με τον αλγόριθμο Kennard-Stone με ποσοστά 75% και 25% των αρχικών δεδομένων. Στη δεύτερη διαμέριση προκύπτει το σύνολο εκπαίδευσης που περιέχει το 80% του συνόλου μοντελοποίησης και το σύνολο δοκιμών με το υπόλοιπο 20%. Το σύνολο εκπαίδευσης και το σύνολο δοκιμών λειτουργούν όπως στην προηγούμενη διαδικασία ανάπτυξης και αξιολόγησης του μοντέλου. Το σύνολο εξωτερικής αξιολόγησης δε συμμετέχει στη μοντελοποίηση και χρησιμοποιείται μόνο στο τελικό μοντέλο που προκύπτει. Αποτελεί «τυφλό» σύνολο δεδομένων και χρησιμοποιείται για επιπλέον επικύρωση του μοντέλου, αφού προσομοιάζονται οι ρεαλιστικές συνθήκες χρήσης του ανεπτυγμένου μοντέλου.

[ii] Για τις μεταβλητές τις οποίες το πρόβλημα βελτιστοποίησης δεν έχει εφικτή λύση (infeasible solution), ο επιλύτης μηδενίζει αυτόματα την αντικειμενική συνάρτηση. Επομένως, για την αποφυγή επιλογής μιας μεταβλητής διχοτόμησης που δεν δίνει εφικτή λύση, προστέθηκε στον κώδικα ένας περιορισμός ώστε η μεταβλητή που επιλέγεται να μη μηδενίζει την αντικειμενική συνάρτηση.



Σχήμα 4.3: Διάσπαση δεδομένων.



Σχήμα 4.4: Διάσπαση δεδομένων για εξωτερική επικύρωση.

Κεφάλαιο 5

Μελέτες περιπτώσεων

Στην παρούσα Διπλωματική Εργασία χρησιμοποιήθηκαν πειραματικά δεδομένα από τις δημοσιευμένες εργασίες των Gajewicz *et al.* (2015)[46], Walkey *et al.* (2014)[6], Xia *et al*. (2011)[47] και Fourches *et al.* (2010)[48]. Για λόγους συντομίας τα σύνολα από τις συγκεκριμένες δημοσιεύσεις θα ονομάζονται στο εξής «Μεταλλικά Οξείδια», «Νανοσωματίδια χρυσού», «Νανοσωλήνες άνθρακα» και «Επιφανειακά-τροποποιημένα νανοσωματίδια» αντίστοιχα. Όλα τα σύνολα δεδομένων περιέχουν νανοσωματίδια διαφόρων χαρακτηριστικών, τις μετρήσεις για ορισμένες ιδιότητες τους και μετρήσεις για ανεπιθύμητες ιδιότητες. Σκοπός της εργασίας είναι η ανάπτυξη ενός μοντέλου με τη χρήση των γνωστών ιδιοτήτων και της μεταβλητής απόκρισης για την πρόβλεψη της μεταβλητής απόκρισης σε νανοσωματίδια στα οποία δεν είναι γνωστή. Σε δύο σύνολα δεδομένων, στα «Μεταλλικά οξείδια» (Gajewicz *et al.*[46]) και τα «Νανοσωματίδια χρυσού» (Walkey *et al.*[6]) οι γνωστές ιδιότητες χωρίζονται σε δύο κατηγορίες, σε κβαντομηχανικές και περιγραφικές για το πρώτο σύνολο δεδομένων και σε φυσικοχημικές και βιολογικές μεταβλητές για το δεύτερο σύνολο δεδομένων. Σε αυτές τις περιπτώσεις, η επίλυση του αλγορίθμου πραγματοποιήθηκε τόσο σε μία διάσταση, θεωρώντας όλες τις γνωστές ιδιότητες ως μία ενιαία κατηγορία μεταβλητών, όσο και σε δύο διαστάσεις, όπου η κάθε διάσταση περιείχε τις δύο κατηγορίες ιδιοτήτων. Τα σύνολα δεδομένων που χρησιμοποιήθηκαν παρουσιάζονται παρακάτω.

5.1 Μεταλλικά οξείδια

Στη δημοσίευση των Gajewicz *et al*. (2015)[46] «*Towards understanding mechanisms governing cytotoxicity of metal oxides nanoparticles: Hints from nano-QSAR studies*» μελετάται η τοξικότητα 18 νανοσωματιδίων μεταλλικών οξειδίων στην κυτταρική σειρά ανθρώπινων κερατινοκυττάρων (Human Keratinocyte, HaCaT) κατά την έκθεση τους στο δέρμα. Με υπολογισμούς κβαντικής χημείας υπολογίστηκαν 18 κβαντομηχανικές ιδιότητες των νανοσωματιδίων και με τεχνική της ηλεκτρονιακής μικροσκοπίας μετάδοσης (TEM) προέκυψαν 11 γεωμετρικές ιδιότητες. Τα τοξικολογικά δεδομένα αφορούν την συγκέντρωση των νανοσωματιδίων που προκαλούν μείωση των κυττάρων σε ποσοστό 50% μετά από έκθεση αυτών σε νανοσωματίδια για 24 ώρες (LC_{50}). Οι τιμές του συγκεκριμένου δείκτη μετατράπηκαν σε αντίστροφο λογάριθμο $\log(LC_{50})^{-1}$.

5.2 Νανοσωματίδια χρυσού

Στη δημοσίευση των Walkey *et al*. (2014)[6] *«Protein Corona Fingerprinting Predicts the Cell Association of Gold and Silver Nanoparticles»* παρουσιάζονται τα αποτελέσματα που προέκυψαν κατά την ανάπτυξη ενός μοντέλου πρόβλεψης της κυτταρικής συσχέτισης

μέσω του χαρακτηρισμού του αποτυπώματος του πρωτεϊνικού στέμματος επικαλυμμένων νανοσωματιδίων χρυσού και αργύρου. Χρησιμοποιήθηκαν 105 νανοσωματίδια χρυσού διαμέτρων 15, 30 και 60 nm επικαλυμμένα με 67 οργανικούς επιφανειακούς προσδέτες (π.χ. μικρά μόρια, πολυμερή, λιπίδια ή πεπτίδια) τα οποία επιλέχθηκαν για να προσομοιάσουν την επιφανειακή χημεία διαφόρων νανοσωματιδίων. Οι επιφανειακοί προσδέτες χαρακτηρίστηκαν ως ουδέτεροι (neutral), ανιονικοί (anionic) ή κατιονικοί (cationic) ανάλογα με τη χημική τους δομή και το καθαρό φορτίο (net charge) σε φυσιολογικό pH. Από τα 105 νανοσωματίδια, αποκλείστηκαν τα 21 ουδέτερα νανοσωματίδια διότι δεν προσροφούν έντονα πρωτεΐνες στην επιφάνεια τους. Στη συνέχεια τα νανοσωματίδια μελετήθηκαν με συνδυασμούς τεχνικών ενόργανης ανάλυσης (Ενότητα 1.2) με σκοπό να μετρηθούν διάφοροι φυσικοχημικοί δείκτες. Συγκεκριμένα, μελετήθηκαν με την τεχνική της ηλεκτρονιακής μικροσκοπίας μετάδοσης, με τη μέθοδο δυναμικής σκέδασης του φωτός, με φασματοσκοπία απορρόφησης και με ηλεκτροφόρηση σε αγαρόζη.

Μετά τη σύνθεση και το χαρακτηρισμό τους, κάθε νανοσωματίδιο επωάστηκε σε μη αραιωμένο ανθρώπινο ορό αίματος στους 37°C για 1 ώρα και στη συνέχεια καθαρίστηκε με φυγοκέντρηση για απομάκρυνση των μη δεσμευμένων στο νανοσωματίδιο πρωτεϊνών. Ο ορός αίματος επιλέχθηκε για προσομοίωση του βιολογικού περιβάλλοντος που συναντά ένα νανοσωματίδιο μετά από ενδοφλέβια ένεση ή σε *in vitro* πειράματα κυτταρικής καλλιέργειας. Η σύνθεση του πρωτεϊνικού στέμματος χαρακτηρίστηκε ποιοτικά με ηλεκτροφόρηση γέλης πολυακριλαμιδίου και ημι-ποσοτικά με τη χρήση υγρής χρωματογραφίας-φασματομετρίας μάζας υψηλής ανάλυσης (LC-MS/MS). Από τις 785 πρωτεΐνες ορού που ταυτοποιήθηκαν, μόνο οι 129 θεωρήθηκαν αποδεκτές για σχετική ποσοτικοποίηση.

Μετά το χαρακτηρισμό του πρωτεϊνικού στέμματος, μελετήθηκε η συσχέτιση των νανοσωματιδίων με τα κύτταρα της σειράς A549 (καρκινικά επιθηλιακά κύτταρα ανθρώπινου πνεύμονα) σε καλλιέργεια μονοστιβάδας με την τεχνική τη φασματοσκοπίας ατομικής εκπομπής με πηγή επαγωγικά συζευγμένου πλάσματος (Inductively Coupled Plasma-Atomic Emission Spectroscopy ICP-AES). Η κυτταρική συσχέτιση y υπολογίστηκε χρησιμοποιώντας τον συντελεστή ψευδο-συμμετοχής όπως φαίνεται στη σχέση [5.1].

$$y = \frac{m_{cell}/m_{well}}{m_{cells}} \qquad [5.1]$$

Όπου m_{cell}, η περιεκτικότητα ατομικού χρυσού που συνδέεται με τα κύτταρα
m_{well}, η συνολική περιεκτικότητα ατομικού χρυσού
m_{cells}, η συνολική μάζα μαγνησίου ανά δείγμα.
Η κυτταρική συσχέτιση μετρήθηκε σε μονάδες mL/μg(Mg). Πριν τη μοντελοποίηση οι τιμές μετατράπηκαν σε λογαριθμική κλίμακα και αυτές οι τιμές χρησιμοποιήθηκαν και στη συγκεκριμένη Εργασία.

Το σύνολο των δεδομένων που χρησιμοποιήθηκε στην παρούσα Διπλωματική Εργασία περιέχει 84 νανοσωματίδια και δεδομένα για 40 φυσικοχημικούς δείκτες και 63 πρωτεΐνες. Από τις 129 πρωτεΐνες που ταυτοποιήθηκαν και υπολογίστηκαν ποσοτικά, οι Varsou *et al*. (2018)[49] επέλεξαν μόνο 63 πρωτεΐνες οι οποίες θεωρήθηκαν στατιστικά σημαντικές και ότι ανήκουν σε ενδιαφέροντα γονιδιακά σύνολα. Η επιλογή αυτή έγινε με ανάλυση εμπλουτισμού (Gene Set Enrichment Analysis, GSEA) και πιο συγκεκριμένα ανάλυση διακύμανσης συνόλου γονιδίων (Gene Set Variation Analysis, GSVA). Η ανάλυση εμπλουτισμού γονιδίων αποτελεί μία μέθοδο για τον εντοπισμό κατηγοριών γονιδίων ή πρωτεϊνών που υπερεκφράζονται ή υποεκφράζονται σε ένα μεγάλο σύνολο γονιδίων ή

πρωτεϊνών και μπορεί να έχουν σχέση με φαινοτύπους ασθενειών. Η ανάλυση διακύμανσης συνόλου γονιδίων αποτελεί μια μη επιτηρούμενη μέθοδο κατά την οποία υπολογίζεται η διακύμανση της δραστηριότητας βιολογικών μονοπατιών σε ένα δείγμα γονιδίων. [50,51]

5.3 Νανοσωλήνες άνθρακα

Στη δημοσίευση των Xia *et al.* (2011)[47] «*Mapping the surface adsorption forces of nanomaterials in biological systems*» παρουσιάζεται μία μέθοδος πρόβλεψης με πολλαπλή γραμμική παλινδρόμηση του δείκτη προσρόφησης βιολογικής επιφάνειας (Biological Surface Absorption Index , BSAI) σε νανοσωματίδια. Κάτω από ιδανικές βιολογικά συνθήκες, οι ιδιότητες προσρόφησης της επιφάνειας των νανοσωματιδίων μπορούν να μετρηθούν χρησιμοποιώντας ένα σύνολο ανιχνευτών με ποικίλες φυσικοχημικές ιδιότητες για την επιλογή νανο-περιγραφέων που αντιπροσωπεύουν τις μοριακές δυνάμεις αλληλεπίδρασης των υλικών με βιολογικά μόρια. Οι 5 περιγραφείς στη συγκεκριμένη δημοσίευση είναι η πολικότητα, η διασπορά London, τα μονήρη ηλεκτρόνια (lone-pair electrons), οι δότες και οι δέκτες δεσμού υδρογόνου. Ένα σύνολο 28 ενώσεων χρησιμοποιούνται ως ανιχνευτές σε πολυεπίπεδους νανοσωλήνες άνθρακα (Multiwalled Carbon Nanotubes, MWCNT) διαμέτρου 40 nm με επικαλύψεις παράγωγα υδροξυλίου. Η τιμή των συντελεστών προσρόφησης των ανιχνευτών στα νανοσωματίδια μετρήθηκε με μικροεκχύλιση στερεάς φάσης (Solid Phase Microextraction, SPME)-GC/MS και μετατράπηκε σε λογαριθμική κλίμακα.

5.4 Επιφανειακά-τροποποιημένα νανοσωματίδια

Στη δημοσίευση των Fourches *et al.* (2010)[48] «*Quantitative Nanostructrure-Activity Relationship (QNAR) Modeling*» παρουσιάζεται η μελέτη συνθετικών νανοσωματιδίων ως προς την πρόσληψη κυττάρων. Συγκεκριμένα, 109 νανοσωματίδια με τον ίδιο μεταλλικό πυρήνα και διαφορετικές επιφανειακές επικαλύψεις μελετήθηκαν ως προς την πρόσληψη μιας σειράς κυττάρων και συγκεκριμένα των ανθρώπινων καρκινικών κυττάρων του παγκρέατος (Pancreatic human cancer cells, PaCa2). Τα αρχικά δεδομένα περιείχαν τα 109 νανοσωματίδια με διαφορετικές οργανικές επικαλύψεις που κωδικοποιούνται σε μορφή SMILES. Έχει γίνει η υπόθεση ότι καθώς έχουν τον ίδιο μεταλλικό πυρήνα, οι διαφοροποιήσεις μεταξύ των νανοσωματιδίων οφείλονται μόνο στις επικαλύψεις τους. Μέσω του λογισμικού Mold2 υπολογίστηκαν 777 περιγραφείς των συγκεκριμένων επικαλύψεων. Το Mold2 πρόκειται για ένα ελεύθερο λογισμικό που χρησιμοποιείται για τον γρήγορο υπολογισμό περιγραφέων σύμφωνα με τη δισδιάστατη δομή των μορίων. Η τιμή της πρόσληψης των κυττάρων PaCa2 εκφράστηκε ως ο δεκαδικός λογάριθμος της συγκέντρωσης των νανοσωματιδίων προς τη συγκέντρωση των κυττάρων ($^{log_{10}[nanoparticles]}/_{cell\ pM}$). [48,52,53]

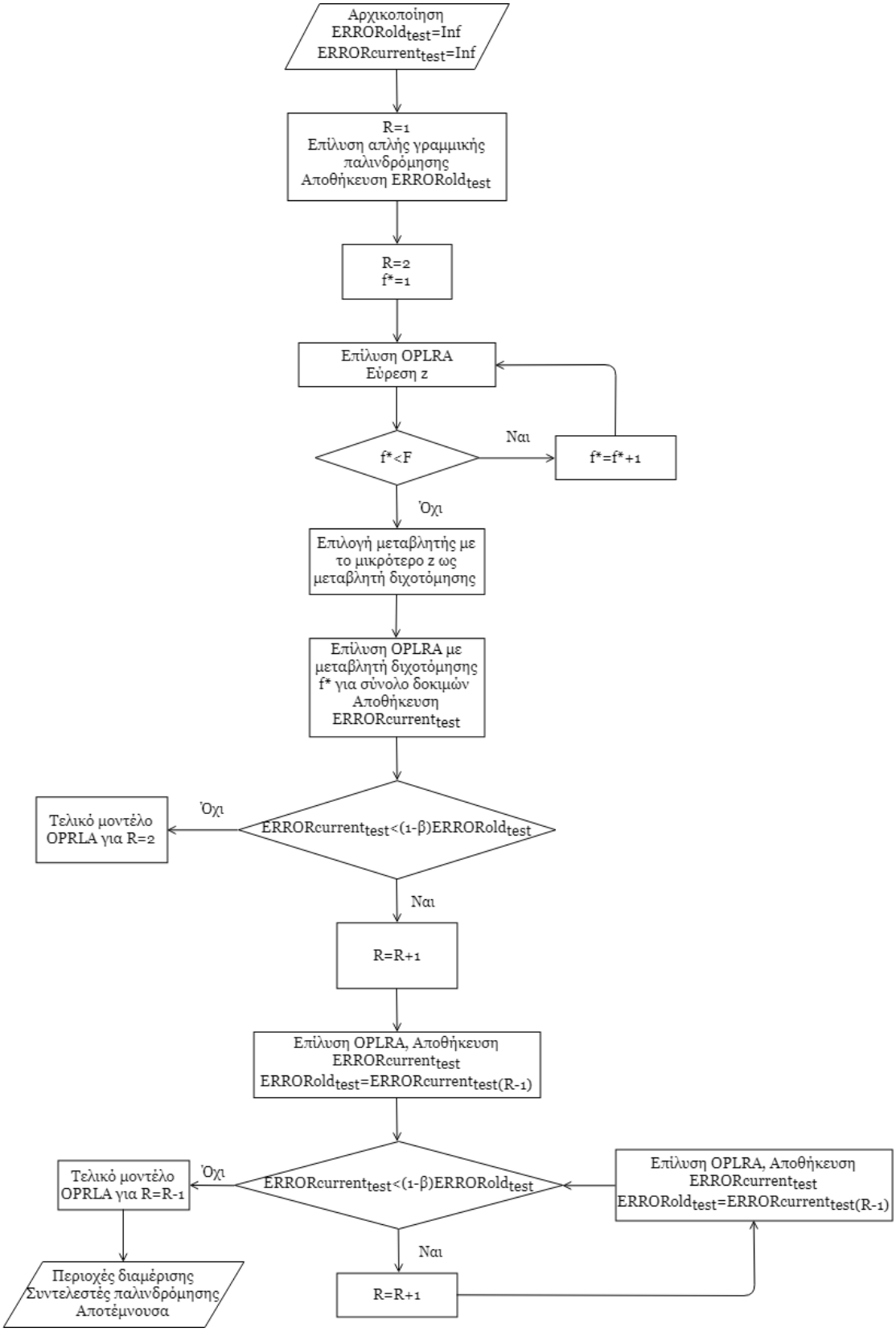
Κεφάλαιο 6

Αποτελέσματα

6.1 Αλγόριθμος επίλυσης σε μία διάσταση

Στο Σχήμα 6.1 συνοψίζονται τα βασικά βήματα που ακολουθούνται στην ανάλυση γραμμικής τμηματικής παλινδρόμησης όταν η διάσπαση του πεδίου ορισμού σε περιοχές γίνεται σε μία διάσταση, όπως παρουσιάζονται παρακάτω:

1. Αρχικοποιούνται οι τιμές των σφαλμάτων πρόβλεψης ως άπειρες.
2. Επιλύεται απλή γραμμική παλινδρόμηση ($R = 1$) και αποθηκεύονται οι τιμές των αντικειμενικών συναρτήσεων z τόσο για το σύνολο εκπαίδευσης όσο και κατά την εφαρμογή του μοντέλου στο σύνολο δοκιμών, $ERROROld$ και $ERROROld_{test}$ αντίστοιχα.
3. Γίνεται επίλυση του αλγορίθμου OPLRA στα δεδομένα εκπαίδευσης με δύο περιοχές διαμέρισης ($R = 2$) επιλέγοντας κάθε φορά μία μεταβλητή f από το σύνολο των μεταβλητών του συνόλου δεδομένων ως μεταβλητή διχοτόμησης.
4. Η μεταβλητή η οποία ελαχιστοποιεί την αντικειμενική συνάρτηση, επιλέγεται ως μεταβλητή διχοτόμησης f^*.
5. Στη συνέχεια εφαρμόζεται το μοντέλο OPLRA με δύο περιοχές διαμέρισης και μεταβλητή διχοτόμησης f^* στο σύνολο δοκιμών και αποθηκεύεται το σφάλμα πρόβλεψης $ERRORcurrent_{test} = z_{test}$.
6. Εξετάζεται η προσθήκη περιοχών διαμέρισης, $R = R + 1$, μέσω της συνθήκης $ERRORcurrent_{test} < (1 - \beta)ERROROld_{test}$ (όπου β, τιμή που ορίζεται από το χρήστη). Εάν δεν ισχύει η συνθήκη, τότε το τελικό μοντέλο θα αποτελείται από δύο περιοχές διαμέρισης.
7. Η αύξηση των περιοχών διαμέρισης πραγματοποιείται όταν η βελτίωση του σφάλματος μεταξύ των συνεχόμενων επαναλήψεων είναι μεγαλύτερη από την τιμή β. Στην περίπτωση αυτή, επιλύεται ο αλγόριθμος OPLRA για το σύνολο εκπαίδευσης με αριθμό περιοχών R και μεταβλητή διχοτόμησης f^*.
8. Εφαρμόζεται το μοντέλο που προκύπτει για συγκεκριμένο αριθμό περιοχών στο σύνολο δοκιμών και αποθηκεύονται τα σφάλματα $ERRORcurrent_{test} = z_{test}$ και $ERROROld_{test} = ERRORcurrent_{test\ (R-1)}$.
9. Εξετάζεται πάλι η συνθήκη προσθήκης περιοχών. Στην περίπτωση που ικανοποιείται, τότε προστίθεται και άλλη περιοχή, αλλιώς προκύπτει το τελικό μοντέλο με $R - 1$ περιοχές διαμέρισης.



Σχήμα 6.1: Αλγόριθμος OPLRA.

Στην περίπτωση που οι μεταβλητές εισόδου αποτελούν μία ενιαία κατηγορία, η διάσπαση του πεδίου ορισμού σε περιοχές διαμέρισης γίνεται σε μία διάσταση. Τα βήματα που ακολούθησαν παρουσιάζονται στο Σχήμα 6.1. Αρχικά, για την αξιολόγηση του μοντέλου επιλέχθηκε η διάσπαση των δεδομένων όπως παρουσιάζεται στο Σχήμα 4.3, όπου το σύνολο δεδομένων χωρίστηκε σε σύνολο εκπαίδευσης και σύνολο δοκιμών. Μελετήθηκε η επίδραση του όρου ομαλοποίησης και επιλύθηκε ο αλγόριθμος OPLRA για τιμές της παραμέτρου $\lambda = [0.000, 0.005, 0.010, 0.020]$ και με παράμετρο απόκλισης σφαλμάτων $\beta = 0.05$. Παρουσιάζεται η επίδραση της παραμέτρου ομαλοποίησης στη τιμή της αντικειμενικής συνάρτησης του συνόλου εκπαίδευσης και δοκιμών, z και z_{test}, ο συντελεστής συσχέτισης R^2 του μοντέλου καθώς και ο συντελεστής συσχέτισης R^2 σε κάθε περιοχή διαμέρισης, ο δείκτης εξωτερικής ερμηνεύσιμης διακύμανσης Q^2_{test}, οι περιοχές διαμέρισης, οι μεταβλητές που επιλέγονται, ο αριθμός των δειγμάτων του συνόλου εκπαίδευσης που απαρτίζουν την κάθε περιοχή διαμέρισης καθώς και ο υπολογιστικός χρόνος που απαιτείται για την επίλυση του προβλήματος.

6.1.1 Μεταλλικά οξείδια

Το σύνολο των δεδομένων διαθέτει 18 νανοσωματίδια μεταλλικών οξειδίων και 29 ιδιότητες. Η διαμέριση του συγκεκριμένου συνόλου έγινε σύμφωνα με την δημοσίευση των Gajewicz *et al.*[19,46] σε 10 δείγματα στο σύνολο εκπαίδευσης και 8 στο σύνολο δοκιμών ώστε να μπορεί να γίνει σύγκριση των αποτελεσμάτων με την αρχική δημοσίευση. Στον παρακάτω πίνακα (Πίνακας 6.1) παρουσιάζονται τα αποτελέσματα της επίδρασης του όρου ομαλοποίησης λ.

Πίνακας 6.1: Αποτελέσματα επίλυσης σε μία διάσταση για τα δεδομένα των Μεταλλικών οξειδίων.

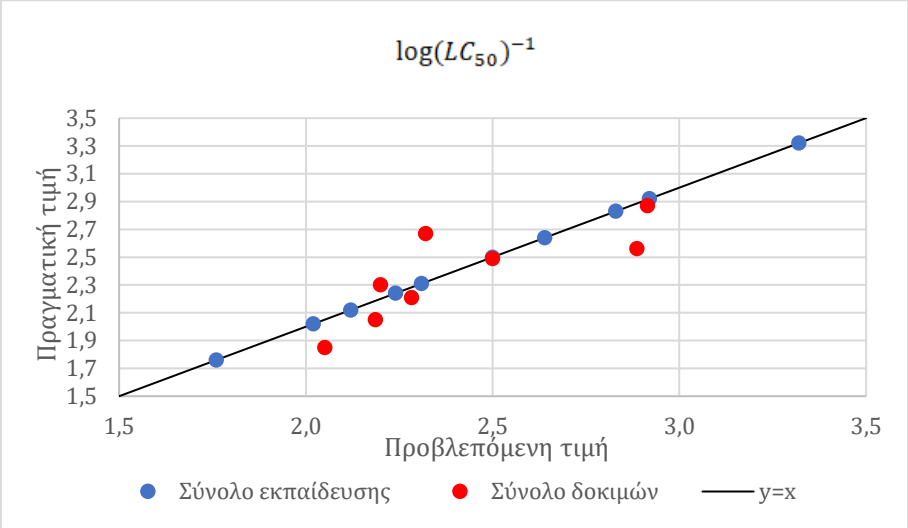
	Μεταλλικά οξείδια
z	
λ=0.000	$-2.1316 \cdot 10^{-15}$
λ=0.005	0.0074
λ=0.010	0.0149
λ=0.020	0.0298
z test	
λ=0.000	3.1576
λ=0.005	0.2315
λ=0.010	0.2389
λ=0.020	0.1849
MAE	
λ=0.000	$-2.1316 \cdot 10^{-15}$
λ=0.005	$-3.5527 \cdot 10^{-16}$
λ=0.010	$-3.5527 \cdot 10^{-16}$
λ=0.020	$-3.5527 \cdot 10^{-16}$
REG	
λ=0.000	$2.1242 \cdot 10^{2}$
λ=0.005	1.4899
λ=0.010	1.4899
λ=0.020	1.4899
R^2	
λ=0.000	1.0000
λ=0.005	1.0000

λ=0.010	1.0000
λ=0.020	1.0000
R^2 ανά περιοχή	
λ=0.000	Region 1: 1.0000 Region 2: -
λ=0.005	Region 1: - Region 2: 1.0000 Region 3: 1.0000
λ=0.010	Region 1: - Region 2: 1.0000 Region 3: 1.0000
λ=0.020	Region 1: - Region 2: 1.0000 Region 3: 1.0000
Q^2 test	
λ=0.000	−331.9893
λ=0.005	0.2338
λ=0.010	0.2338
λ=0.020	0.6452
Περιοχές	
λ=0.000	2
λ=0.005	3
λ=0.010	3
λ=0.020	3
Μεταβλητές	
λ=0.000	9
λ=0.005	7
λ=0.010	7
λ=0.020	7
Αριθμός δειγμάτων ανά περιοχή	
λ=0.000	Region 1: 9 Region 2: 1
λ=0.005	Region 1: 1 Region 2: 2 Region 3: 7
λ=0.010	Region 1: 1 Region 2: 2 Region 3: 7
λ=0.020	Region 1: 1 Region 2: 2 Region 3: 7
Υπολογιστικός χρόνος (min)	
λ=0.000	0.61
λ=0.005	0.68
λ=0.010	0.64
λ=0.020	0.72

Όταν όρος ομαλοποίησης είναι ίσος με μηδέν, η αντικειμενική συνάρτηση μηδενίζεται και ο συντελεστής συσχέτισης R^2 παίρνει την τιμή 1, ενώ οι αντίστοιχες τιμές για το σύνολο δοκιμών παίρνουν μη αποδεκτές τιμές, καθώς η αντικειμενική συνάρτηση αυξάνεται

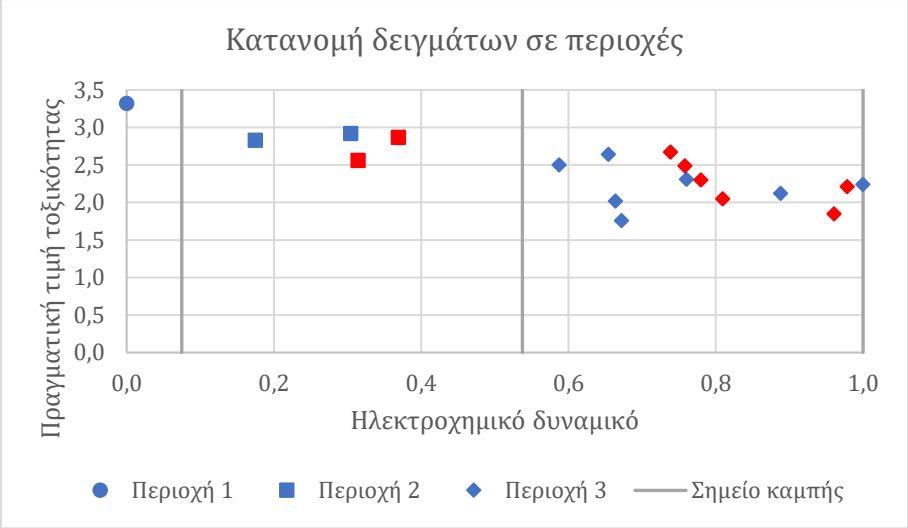
αρκετά ($z_{test} = 3.1576$) και ο δείκτης Q^2_{test} παίρνει αρνητική τιμή. Το σφάλμα πρόβλεψης MAE για τα δεδομένα εκπαίδευσης μηδενίζεται ενώ ο όρος ομαλοποίησης REG παίρνει πολύ μεγάλη τιμή. Επομένως, γίνεται υπερπροσαρμογή του μοντέλου στα δεδομένα εκπαίδευσης και οι προβλέψεις για τα δεδομένα δοκιμών δεν είναι αποδεκτές.

Παρατηρείται ότι για $\lambda = 0.02$ προκύπτει η καλύτερη πρόβλεψη με $z_{test} = 0.1849$ και $Q^2_{test} = 0.6452$. Ο αλγόριθμος επιλύθηκε για τιμές $\beta = [0.001, 0.025, 0.05, 0.1, 0.2]$ και δεν παρατηρήθηκαν βελτιώσεις στα αποτελέσματα. Στο Διάγραμμα 6.1 παρουσιάζεται η συσχέτιση των πραγματικών και των προβλεπόμενων τιμών της μεταβλητής απόκρισης, δηλαδή του αντίστροφου λογαρίθμου της συγκέντρωσης νανοσωματιδίων που προκαλούν μείωση των κυττάρων σε ποσοστό 50% ($\log(LC_{50})^{-1}$) στο σύνολο εκπαίδευσης και στο σύνολο δοκιμών.



Διάγραμμα 6.1: Πραγματικές και προβλεπόμενες τιμές ($\log(LC_{50})^{-1}$) των μεταλλικών οξειδίων με επίλυση σε μία διάσταση για τα σύνολα εκπαίδευσης και δοκιμών.

Τα 10 δείγματα του συνόλου εκπαίδευσης κατατάσσονται σε 3 περιοχές διαμέρισης όπως παρουσιάζεται στο Διάγραμμα 6.2 ενώ ως μεταβλητή διχοτόμησης επιλέγεται το ηλεκτροχημικό δυναμικό.



Διάγραμμα 6.2: Κατανομή δειγμάτων εκπαίδευσης και δοκιμών ανά περιοχή διαμέρισης και σημεία καμπής για τα δεδομένα των Μεταλλικών οξειδίων. Με μπλε χρώμα απεικονίζονται τα δείγματα εκπαίδευσης και με κόκκινο τα δείγματα δοκιμών.

Η τιμή εξόδου για το δείγμα της περιοχής 1 είναι ίση με την αποτέμνουσα στη συγκεκριμένη περιοχή (Εξίσωση [6.1]), καθώς δεν επιλέγεται καμία μεταβλητή για την πρόβλεψη. Για τα δείγματα της περιοχής 2 η τιμή εξόδου υπολογίζεται μέσω της εξίσωσης γραμμικής παλινδρόμησης [6.2] και τέλος για τα δείγματα της 3ης περιοχής μέσω της εξίσωσης [6.3].

$$\log(LC_{50})^{-1} = 3.3200 \quad [6.1]$$

$$\log(LC_{50})^{-1} = -0.0956Core + 2.9256 \quad [6.2]$$

$$\begin{aligned}\log(LC_{50})^{-1} = {} & 0.7225\Delta H_f^c - 0.3302TE - 0.0440S \\ & + 0.0062Volume.mass.diameter \\ & + 0.2370Volume.surface.diameter + 0.0545PorosityY \\ & + 1.8289\end{aligned} \quad [6.3]$$

Πεδίο εφαρμογής μοντέλου

Στη συνέχεια, ορίστηκε το πεδίο εφαρμογής του μοντέλου και υπολογίστηκε η τιμή του προκαθορισμένου κατωφλιού APD ίση με 1.5372. Από τα 8 δείγματα του συνόλου δοκιμών, μόνο ένα έχει απόσταση από τον κοντινότερο γείτονα του συνόλου εκπαίδευσης μεγαλύτερη από την τιμή του κατωφλιού APD, επομένως η πρόβλεψη για το συγκεκριμένο δείγμα δεν θεωρείται αξιόπιστη.

Έλεγχος τυχαίας επιλογής

Για επιπλέον έλεγχο της ισχύος του μοντέλου, πραγματοποιήθηκε έλεγχος τυχαίας επιλογής. Οι τιμές της εξαρτημένης μεταβλητής y ανακατεύτηκαν τυχαία και ο αλγόριθμος χρησιμοποίησε τις ανακατεμένες τιμές εξόδου για την εκπαίδευση του

μοντέλου. Τα αποτελέσματα που προέκυψαν από τον συγκεκριμένο έλεγχο παρουσιάζονται παρακάτω.

Πίνακας 6.2: Αποτελέσματα ελέγχου τυχαίας επιλογής με επίλυση σε μία διάσταση για τα δεδομένα των Μεταλλικών οξειδίων.

	Μεταλλικά οξείδια
z	0.0766
z test	0.4375
R^2	1.0000
R^2 ανά περιοχή	Region 1: 1.0000 Region 2: 1.0000
Q^2 test	-0.8571
Περιοχές	2
Μεταβλητές	8
Αριθμός δειγμάτων ανά περιοχή	Region 1: 5 Region 2: 5
Υπολογιστικός χρόνος (min)	0.53

Παρατηρείται ότι η αντικειμενική συνάρτηση για το σύνολο δοκιμών ($z_{test} = 0.4375$) παίρνει αρκετά μεγαλύτερη τιμή από την αντίστοιχη του συνόλου εκπαίδευσης ($z = 0.0766$) και ο δείκτης Q^2_{test} παίρνει μη αποδεκτή τιμή ($Q^2_{test} = -0.8571$). Τα αποτελέσματα του ελέγχου τυχαίας επιλογής επιβεβαιώνουν ότι το μοντέλο με τις τυχαία κατανεμημένες τιμές εξόδου δεν μπορεί να δώσει αξιόπιστες προβλέψεις για άγνωστα δεδομένα. Συνεπώς η πιθανότητα τυχαίας συσχέτισης δεδομένων-εξόδου στο αρχικό μοντέλο, έχει ελαχιστοποιηθεί.

Λόγω των περιορισμένων δειγμάτων, δεν πραγματοποιήθηκε εξωτερική επικύρωση στο συγκεκριμένο σύνολο δεδομένων καθώς αποτελείται μόνο από 18 δείγματα.

6.1.2 Νανοσωματίδια χρυσού

Το σύνολο δεδομένων αποτελείται από 84 νανοσωματίδια χρυσού-δείγματα και 103 μεταβλητές και χωρίστηκε μέσω της μεθόδου Kennard and Stone σε σύνολο εκπαίδευσης με 63 δείγματα και σύνολο δοκιμών με 21 δείγματα. Στον παρακάτω πίνακα (Πίνακας 6.3) παρουσιάζονται τα αποτελέσματα που προκύπτουν από την μελέτη επίδρασης της παραμέτρου ομαλοποίησης λ.[iii]

[iii] Λόγω των αυστηρών περιορισμών προσδιορισμού των σφαλμάτων, για $\lambda = 0$ γίνεται υπερπροσαρμογή των μοντέλων στα δεδομένα εκπαίδευσης και το σφάλμα E_s μηδενίζεται για κάθε πιθανή μεταβλητή διχοτόμησης. Επομένως λόγω του περιορισμού της τιμής της αντικειμενική συνάρτησης να είναι διάφορη του μηδενός, δεν επιλέγεται καμία μεταβλητή διχοτόμησης. Για την επίλυση του προβλήματος για $\lambda = 0$ για αυτά τα δεδομένα, αφαιρέθηκε αυτός ο περιορισμός και επιλέχτηκε από τον αλγόριθμο η $1^{η}$ μεταβλητή η οποία μηδένισε το σφάλμα πρόβλεψης.

Πίνακας 6.3: Αποτελέσματα επίλυσης σε μία διάσταση για τα δεδομένα των Νανοσωματιδίων χρυσού.

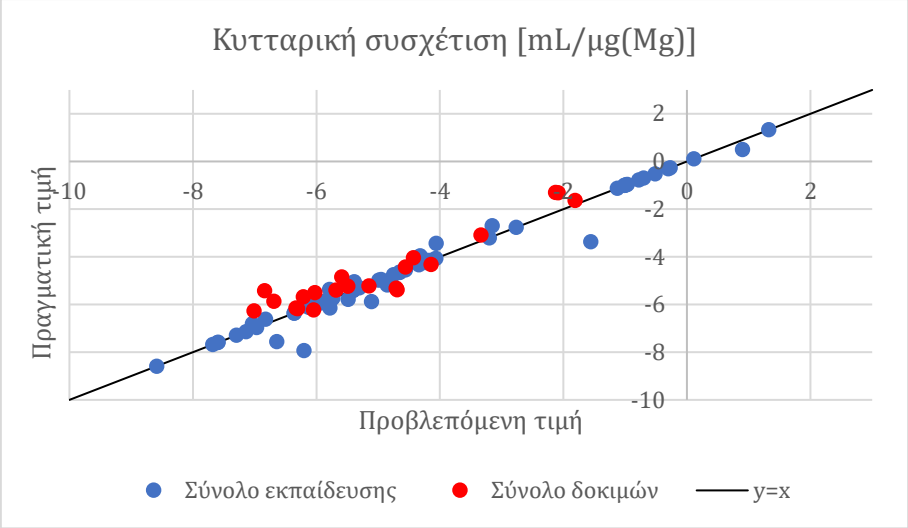
	Νανοσωματίδια χρυσού
z	
λ=0.000	0.0000
λ=0.005	0.3219
λ=0.010	0.5531
λ=0.020	0.8166
z test	
λ=0.000	218.6452
λ=0.005	0.9347
λ=0.010	0.8763
λ=0.020	0.7329
MAE	
λ=0.000	0.0000
λ=0.005	0.0345
λ=0.010	0.1503
λ=0.020	0.5232
REG	
λ=0.000	$3.3389 \cdot 10^4$
λ=0.005	57.4697
λ=0.010	40.2859
λ=0.020	14.6711
R^2	
λ=0.000	1.0000
λ=0.005	0.9959
λ=0.010	0.9755
λ=0.020	0.8178
R^2 ανά περιοχή	
λ=0.000	Region 1: 1.0000 Region 2: 1.0000
λ=0.005	Region 1: 0.9947 Region 2: 1.0000
λ=0.010	Region 1: 0.9687 Region 2: 0.9991
λ=0.020	Region 1: 0.9186 Region 2: 0.5352
Q^2 test	
λ=0.000	$-2.6881 \cdot 10^4$
λ=0.005	0.7117
λ=0.010	0.8622
λ=0.020	0.8457
Περιοχές	
λ=0.000	2
λ=0.005	2
λ=0.010	2
λ=0.020	2
Μεταβλητές	
λ=0.000	67
λ=0.005	54

λ=0.010	40
λ=0.020	22
Αριθμός δειγμάτων ανά περιοχή	
λ=0.000	Region 1: 23 Region 2: 40
λ=0.005	Region 1: 53 Region 2: 10
λ=0.010	Region 1: 53 Region 2: 10
λ=0.020	Region 1: 53 Region 2: 10
Υπολογιστικός χρόνος (min)	
λ=0.000	4.84
λ=0.005	5.09
λ=0.010	4.99
λ=0.020	5.51

Στην περίπτωση που ο όρος ομαλοποίησης είναι ίσος με μηδέν ($\lambda = 0.00$), επιλέγονται 67 από τις 102 μεταβλητές ως απαραίτητες για την πρόβλεψη και όπως αναμένεται γίνεται υπερπροσαρμογή του μοντέλου. Αυτό παρατηρείται καθώς η αντικειμενική συνάρτηση για το σύνολο εκπαίδευσης μηδενίζεται και ο συντελεστής συσχέτισης R^2 παίρνει την τιμή 1, ενώ για το σύνολο δοκιμών η αντικειμενική συνάρτηση z_{test} αυξάνεται αρκετά ($z_{test} = 218.6452$) και ο δείκτης Q^2_{test} παίρνει τιμή εκτός των επιθυμητών ορίων ($Q^2_{test} = -2.6881 \cdot 10^4$). Το σφάλμα πρόβλεψης MAE για τα δεδομένα εκπαίδευσης μηδενίζεται ενώ ο όρος ομαλοποίησης παίρνει πολύ μεγάλη τιμή. Επομένως, το μοντέλο που προκύπτει είναι ικανό να δώσει ικανοποιητικές προβλέψεις μόνο για το σύνολο εκπαίδευσης, αλλά αδυνατεί να προβλέψει δεδομένα δοκιμών, δηλαδή να εφαρμοστεί σε πραγματικά προβλήματα.

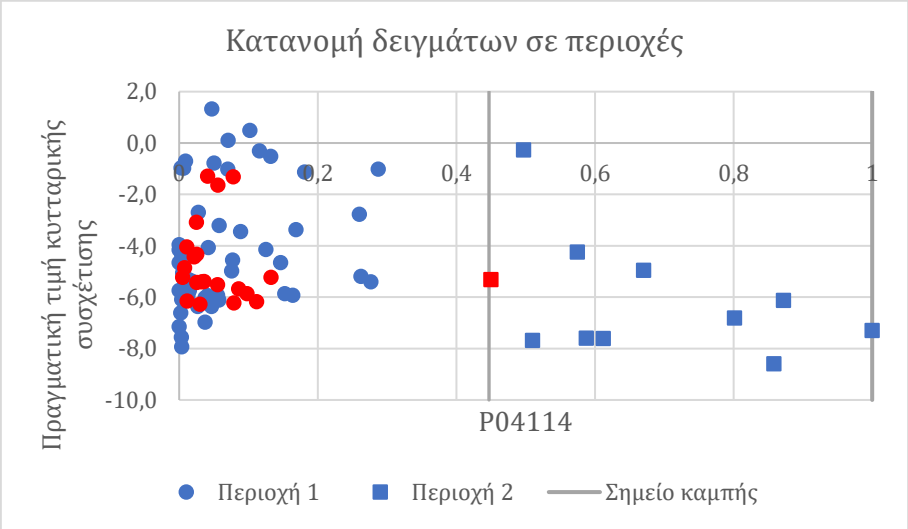
Από τις επαναλήψεις της επίλυσης του αλγορίθμου για διαφορετικές τιμές λ, καλύτερα αποτελέσματα προκύπτουν για $\lambda = 0.010$. Η αντικειμενική συνάρτηση για το σύνολο εκπαίδευσης παίρνει την τιμή $z = 0.5531$ ενώ για το σύνολο δοκιμών $z_{test} = 0.8763$. Ο συντελεστής συσχέτισης R^2 τόσο του μοντέλου όσο και ανά περιοχή διαμέρισης, $R^2_{regions}$ παίρνει τιμές που προσεγγίζουν τη μονάδα και ο δείκτης Q^2_{test} παίρνει την υψηλότερη τιμή από όλες τις επαναλήψεις ($Q^2_{test} = 0.8622$). Η παράμετρος του χρόνου παραμένει σχεδόν σταθερή για όλες τις επαναλήψεις. Στη συνέχεια, πραγματοποιήθηκε ανάλυση ευαισθησίας της μεθόδου στην παράμετρο β, για σταθερή τιμή παραμέτρου ομαλοποίησης $\lambda = 0.010$, όπως επιλέχθηκε από την προηγούμενη ανάλυση. Ο αλγόριθμος επιλύθηκε για τιμές $\beta = [0.001, 0.025, 0.05, 0.1, 0.2]$ και οι τιμές τόσο των σφαλμάτων πρόβλεψης, όσο και των δεικτών αξιολόγησης του μοντέλου δεν παρουσίασαν σημαντικές μεταβολές.

Στο Διάγραμμα 6.3 παρουσιάζεται η συσχέτιση πραγματικών και προβλεπόμενων τιμών της κυτταρικής συσχέτισης για το σύνολο εκπαίδευσης και το σύνολο δοκιμών. Στο σύνολο εκπαίδευσης, η πρόβλεψη γίνεται με μεγάλη ακρίβεια ($R^2 = 0.9755$) και η σχέση προβλεπόμενης και πραγματικής τιμής προσεγγίζει την ευθεία $y = x$. Αρκετά ικανοποιητική είναι και η πρόβλεψη για το σύνολο δοκιμών.



Διάγραμμα 6.3: Πραγματικές και προβλεπόμενες τιμές κυτταρικής συσχέτισης των νανοσωματιδίων χρυσού με επίλυση σε μία διάσταση για τα σύνολα εκπαίδευσης και δοκιμών.

Η μεταβλητή που επιλέγεται ως μεταβλητή διχοτόμησης για τη διαμέριση του πεδίου ορισμού σε περιοχές είναι η P04114. Πρόκειται για την Απολιποπρωτεΐνη Β-100 της οποίας οι βιολογικές δράσεις αφορούν την μεταφορά λιπιδίων, την πήξη και τις κυτταρικές αλληλεπιδράσεις[6]. Τα 63 δείγματα που ανήκουν στο σύνολο εκπαίδευσης χωρίζονται στις δύο περιοχές διαμέρισης σε 53 και 10 αντίστοιχα. Στο Διάγραμμα 6.4 παρουσιάζεται το σημείο καμπής X_{f*}^{r} της μεταβλητής διχοτόμησης, οι περιοχές διαμέρισης και η κατανομή των δειγμάτων σε αυτές.



Διάγραμμα 6.4: Κατανομή δειγμάτων εκπαίδευσης και δοκιμών ανά περιοχή διαμέρισης και σημεία καμπής για τα δεδομένα των Νανοσωματιδίων χρυσού. Με μπλε χρώμα απεικονίζονται τα δείγματα εκπαίδευσης και με κόκκινο τα δείγματα δοκιμών.

Από τα 63 δείγματα του συνόλου εκπαίδευσης, τα 53 που ανήκουν στην περιοχή διαμέρισης 1 δίνουν την εξίσωση γραμμικής παλινδρόμησης [6.4]. Τα υπόλοιπα 10 δείγματα, ανήκουν στην περιοχή διαμέρισης 2 με εξίσωση γραμμικής παλινδρόμησης [6.5]. Επιλέγονται μόνο 40 από τις 102 μεταβλητές ως απαραίτητες για την πρόβλεψη (33 μεταβλητές στην πρώτη περιοχή και 8 στη δεύτερη, καθώς μία μεταβλητή επιλέγεται

και στις δύο περιοχές), μειώνοντας την επίδραση των συντελεστών παλινδρόμησης στην πρόβλεψη.

Από τα 21 δείγματα του συνόλου δοκιμών, τα 20 ανήκουν στην περιοχή διαμέρισης 1 και η πρόβλεψη τους γίνεται με την επιλογή των 33 μεταβλητών που έχουν επιλεγεί ως σημαντικές στη συγκεκριμένη περιοχή διαμέρισης και με συντελεστές παλινδρόμησης και αποτέμνουσα όπως παρουσιάζονται στην εξίσωση γραμμικής παλινδρόμησης [6.4]. Τέλος, το ένα δείγμα που ανήκει στην δεύτερη περιοχή διαμέρισης προβλέπεται ικανοποιητικά από τις 8 μεταβλητές που παρουσιάζονται στην Εξίσωση [6.5] και έχουν επιλεγεί ως σημαντικές στην συγκεκριμένη περιοχή.

$$\begin{aligned} net.c = {} & 0.09 lspri.serum + 1.12 lspri.rel.cl + 0.50 zav.serum \\ & + 2.77 int.serum + 0.45 hdrel.serum - 0.64 vol.ch \\ & + 0.41 pdi.rel + 0.92 zp.serum - 2.30 zp.rel \\ & + 2.28 zp.synth.sign - 0.90 AS.total - 0.61 P01024 \\ & + 1.45 P02649 - 0.19 P04196 - 0.31 P05154 + 1.49 P19823 \\ & - 0.36 P49908 - 0.42 P68871 + 0.95 Q43866 + 0.63 P03951 \\ & + 2.41 P02654 - 0.19 P01011 + 0.07 P18428 + 0.51 P00736 \\ & + 0.10 P00742 + 0.25 P03950 - 0.39 P00450 + 0.63 P08567 \\ & + 0.24 P01019 + 0.57 P02671 + 0.23 P00451 - 0.03 P23528 \\ & + 0.47 Q99467 - 6.35 \end{aligned} \quad [6.4]$$

$$\begin{aligned} net.c = {} & -2.60 pdi.serum - 0.82 nt.rel + 0.05 zp.synth.mag + 2.34 P01009 \\ & + 0.93 P02749 - 3.19 P02655 + 4.72 P27169 + 0.75 P01019 \\ & - 5.33 \end{aligned} \quad [6.5]$$

Εξωτερική Επικύρωση

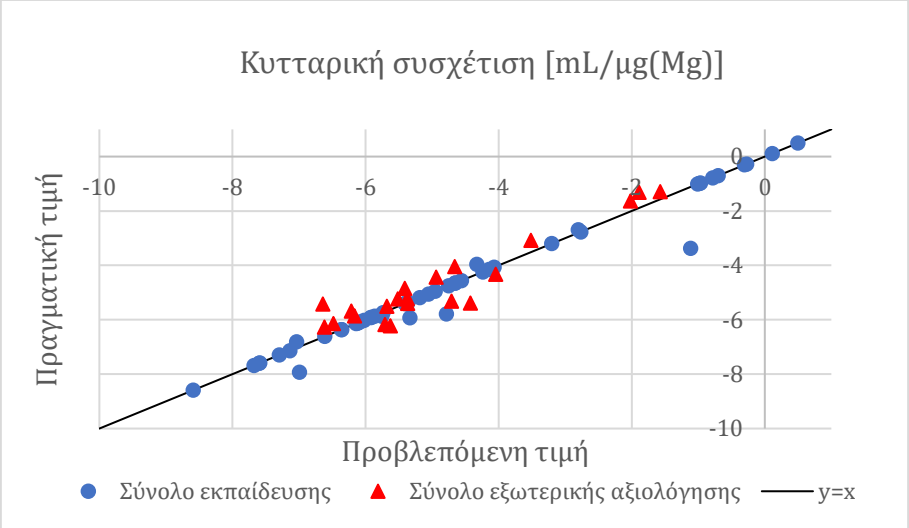
Στη συνέχεια, για επιπλέον επικύρωση του μοντέλου, τα δεδομένα χωρίστηκαν δύο φορές σε σύνολα, όπως παρουσιάζεται στο Σχήμα 4.4. Στην πρώτη διαμέριση προέκυψε το σύνολο μοντελοποίησης και το σύνολο εξωτερικής αξιολόγησης με ποσοστά 75% και 25% των αρχικών δεδομένων που περιείχαν 63 και 21 δείγματα αντίστοιχα. Στη δεύτερη διαμέριση προέκυψε το σύνολο εκπαίδευσης που περιείχε το 80% του συνόλου μοντελοποίησης (50 δείγματα) και το σύνολο δοκιμών με το υπόλοιπο 20% (13 δείγματα). Ο αλγόριθμος OPLRA επιλύθηκε για $\lambda = 0.01$ και $\beta = 0.05$ και τα αποτελέσματα που προέκυψαν παρουσιάζονται παρακάτω (Πίνακας 6.4).

Πίνακας 6.4: Αποτελέσματα εξωτερικής επικύρωσης με επίλυση σε μία διάσταση για τα δεδομένα των Νανοσωματιδίων χρυσού.

	Νανοσωματίδια χρυσού
z	0.4929
z test	1.1081
z external test	0.8398
R^2	0.9741
R^2 ανά περιοχή	Region 1: 0.9642 Region 2: 0.9991

Q^2 test	0.9061
Q^2 external test	0.8830
Περιοχές	2
Μεταβλητές	40
Αριθμός δειγμάτων ανά περιοχή	Region 1: 40 Region 2: 10
Αριθμός άγνωστων δειγμάτων ανά περιοχή	Region 1: 20 Region 2: 1
Υπολογιστικός χρόνος (min)	3.27

Παρατηρείται ότι η τιμή της αντικειμενικής συνάρτησης του συνόλου εξωτερικής αξιολόγησης παίρνει σχετικά χαμηλή τιμή και ο δείκτης Q^2_{ext} λαμβάνει τιμή αρκετά ικανοποιητική ($Q^2_{ext} = 0.8830$), που επαληθεύει ότι το μοντέλο δίνει αξιόπιστες προβλέψεις για σύνολα δεδομένων που δεν συμμετέχουν στην εκπαίδευσή του. Στο Διάγραμμα 6.5 οι προβλεπόμενες τιμές κυτταρικής συσχέτισης σε σχέση με τις πραγματικές για το σύνολο εκπαίδευσης και για το σύνολο εξωτερικής αξιολόγησης προσεγγίζουν μία ευθεία της μορφής $y = x$, επαληθεύοντας ότι οι τιμές της τοξικότητας προβλέπονται αρκετά ικανοποιητικά από το μοντέλο.



Διάγραμμα 6.5: Πραγματικές και προβλεπόμενες τιμές κυτταρικής συσχέτισης των νανοσωματιδίων χρυσού με επίλυση σε μία διάσταση για τα σύνολα εκπαίδευσης και εξωτερικής αξιολόγησης.

Πεδίο εφαρμογής μοντέλου

Στη συνέχεια, ορίστηκε το πεδίο εφαρμογής του μοντέλου και υπολογίστηκε η τιμή του προκαθορισμένου κατωφλιού APD ίση με 2.8277. Όλα τα δείγματα του συνόλου δοκιμών που χρησιμοποιήθηκαν για την επικύρωση του μοντέλου, έχουν απόσταση από τον κοντινότερο γείτονα του συνόλου εκπαίδευσης μικρότερη από την τιμή του κατωφλιού APD. Επομένως οι προβλέψεις για τα δείγματα ελέγχου μπορούν να θεωρηθούν αξιόπιστες.

Έλεγχος τυχαίας επιλογής

Πραγματοποιήθηκε έλεγχος τυχαίας επιλογής και τα αποτελέσματα που προέκυψαν από τον συγκεκριμένο έλεγχο παρουσιάζονται στον πίνακα που ακολουθεί (Πίνακας 6.5).

Πίνακας 6.5: Αποτελέσματα ελέγχου τυχαίας επιλογής με επίλυση σε μία διάσταση για τα δεδομένα των Νανοσωματιδίων χρυσού.

	Νανοσωματίδια χρυσού
z	1.1059
z test	2.1784
R^2	0.6753
R^2 ανά περιοχή	Region 1: 0.6650 Region 2: 0.6910
Q^2 test	-0.5281
Περιοχές	2
Μεταβλητές	38
Αριθμός δειγμάτων ανά περιοχή	Region 1: 41 Region 2: 22
Υπολογιστικός χρόνος (min)	4.45

Παρατηρείται ότι η αντικειμενική συνάρτηση για τα δείγματα ελέγχου παίρνει διπλάσια τιμή από την αντίστοιχη τιμή των δειγμάτων εκπαίδευσης και ο δείκτης Q^2_{test} παίρνει αρνητική, μη επιθυμητή τιμή. Το παραγόμενο μοντέλο από τυχαία κατανεμημένα δεδομένα εξόδου δεν μπορεί να δώσει αξιόπιστες προβλέψεις για άγνωστα δεδομένα.

6.1.3 Νανοσωλήνες άνθρακα

Το σύνολο δεδομένων διαθέτει 28 ενώσεις και 5 ιδιότητες και χωρίστηκε με τη μέθοδο Kennard and Stone σε σύνολο εκπαίδευσης με 21 δείγματα και σύνολο δοκιμών με 7 δείγματα. Ο Πίνακας 6.6 συνοψίζει τα αποτελέσματα της επίδρασης της παραμέτρου ομαλοποίησης λ.

Πίνακας 6.6: Αποτελέσματα επίλυσης σε μία διάσταση για τα δεδομένα των Νανοσωλήνων άνθρακα.

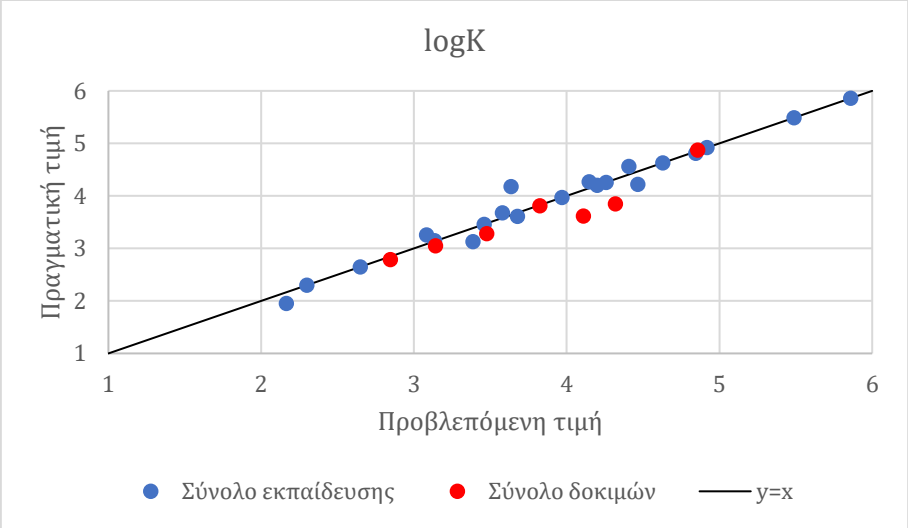
	Νανοσωλήνες άνθρακα
z	
λ=0.000	0.0917
λ=0.005	0.1614
λ=0.010	0.2091
λ=0.020	0.2811
z test	
λ=0.000	3.9399
λ=0.005	0.2597

λ=0.010	0.2935
λ=0.020	0.3630
MAE	
λ=0.000	0.0917
λ=0.005	0.0921
λ=0.010	0.1348
λ=0.020	0.1394
REG	
λ=0.000	269.4668
λ=0.005	13.8727
λ=0.010	7.4268
λ=0.020	7.0867
R^2	
λ=0.000	0.9663
λ=0.005	0.9718
λ=0.010	0.9570
λ=0.020	0.9578
R^2 ανά περιοχή	
λ=0.000	Region 1: 0.9614 Region 2: 1.0000
λ=0.005	Region 1: 0.9552 Region 2: 1.0000
λ=0.010	Region 1: 0,9568 Region 2: -
λ=0.020	Region 1: 0,9577 Region 2: -
Q^2 test	
λ=0.000	$-1.0578 \cdot 10^2$
λ=0.005	0.8539
λ=0.010	0.8572
λ=0.020	0.8615
Περιοχές	
λ=0.000	2
λ=0.005	2
λ=0.010	2
λ=0.020	2
Μεταβλητές	
λ=0.000	5
λ=0.005	5
λ=0.010	5
λ=0.020	4
Αριθμός δειγμάτων ανά περιοχή	
λ=0.000	Region 1: 16 Region 2: 5
λ=0.005	Region 1: 17 Region 2: 4
λ=0.010	Region 1: 20 Region 2: 1
λ=0.020	Region 1: 20 Region 2: 1

	Υπολογιστικός χρόνος (min)
λ=0.000	0.18
λ=0.005	0.14
λ=0.010	0.15
λ=0.020	0.15

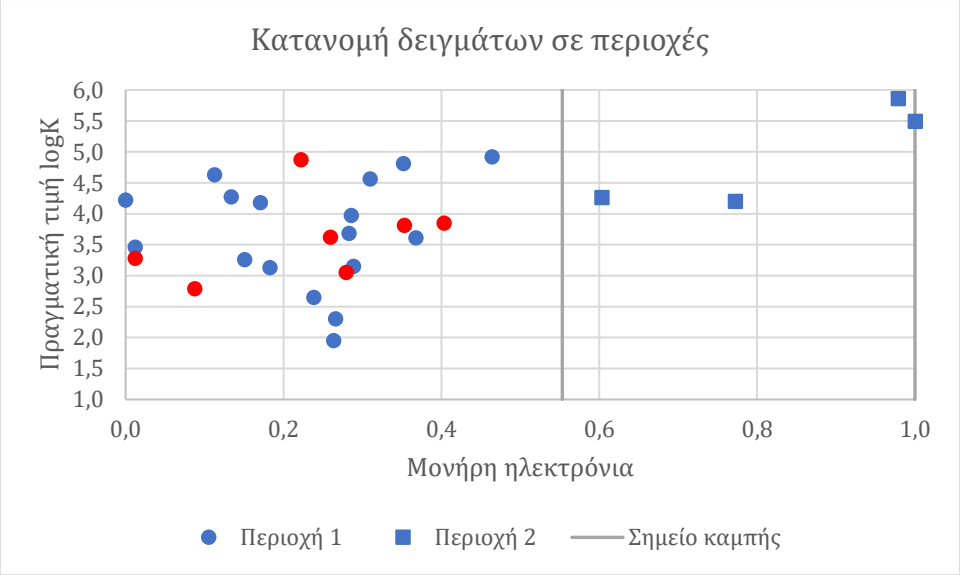
Όταν όρος ομαλοποίησης είναι ίσος με μηδέν, το σφάλμα πρόβλεψης MAE για το σύνολο εκπαίδευσης παίρνει την τιμή $MAE = 0.0917$ που ισούται με την τιμή της αντικειμενικής συνάρτησης z και ο συντελεστής συσχέτισης R^2 παίρνει την τιμή 0.9663. Για το σύνολο δοκιμών προκύπτει $z_{test} = 3.9399$ και ο δείκτης Q^2_{test} παίρνει μη αποδεκτή τιμή ($Q^2_{test} = -1.0578 \cdot 10^2$). Επομένως, το μοντέλο δεν μπορεί να δώσει σωστές προβλέψεις σε άγνωστα δεδομένα.

Παρατηρείται ότι για $\lambda = 0.005$ προκύπτουν ικανοποιητικά αποτελέσματα για το σύνολο δοκιμών με $z_{test} = 0.2597$ και $Q^2_{test} = 0.8539$. Επιλέγονται και οι 5 μεταβλητές ως απαραίτητες για την πρόβλεψη του δείκτη προσρόφησης της βιολογικής επιφάνειας (5 στην πρώτη περιοχή διαμέρισης και 3 στην δεύτερη) και 2 περιοχές διαμέρισης. Στο Διάγραμμα 6.6 παρουσιάζονται οι τιμές του λογαρίθμου του συντελεστή προσρόφησης που προκύπτουν από το μοντέλο και οι πραγματικές τιμές για το σύνολο εκπαίδευσης και το σύνολο δοκιμών.



Διάγραμμα 6.6: Πραγματικές και προβλεπόμενες τιμές του logK των νανοσωλήνων άνθρακα, με επίλυση σε μία διάσταση για τα σύνολα εκπαίδευσης και δοκιμών.

Η μεταβλητή που επιλέγεται ως μεταβλητή διχοτόμησης για τη διαμέριση του πεδίου ορισμού σε περιοχές είναι τα μονήρη ηλεκτρόνια. Πρόκειται για μία από τις 5 μεταβλητές οι οποίες εκφράζουν τις δυνάμεις προσρόφησης της επιφάνειας των νανοσωματιδίων όταν βρίσκονται σε βιολογικό περιβάλλον. Τα δείγματα εκπαίδευσης χωρίζονται στις περιοχές σε 17 και 4 αντίστοιχα. Στο Διάγραμμα 6.7 παρουσιάζεται το σημείο καμπής X^r_{f*} της μεταβλητής διχοτόμησης, οι περιοχές διαμέρισης και η κατανομή των δειγμάτων σε αυτές.



Διάγραμμα 6.7: Κατανομή δειγμάτων εκπαίδευσης και δοκιμών ανά περιοχή διαμέρισης και σημεία καμπής για τα δεδομένα των νανοσωλήνων άνθρακα. Με μπλε χρώμα απεικονίζονται τα δείγματα εκπαίδευσης και με κόκκινο τα δείγματα δοκιμών.

Από τα 21 δείγματα του συνόλου εκπαίδευσης, τα 17 δείγματα της περιοχής 1 προβλέπονται μέσω της εξίσωσης γραμμικής παλινδρόμησης [6.6], ενώ τα υπόλοιπα 4 δείγματα μέσω της εξίσωσης γραμμικής παλινδρόμησης [6.7]. Τα 7 δείγματα του συνόλου δοκιμών ανήκουν στην 1η περιοχή διαμέρισης και η πρόβλεψή τους γίνεται από την εξίσωση [6.6].

$$logK = 0.46R + 1.64\pi - 0.004\alpha - 2.51\beta + 3.27V + 2.17 \quad [6.6]$$

$$logK = 3.13R + 1.30\pi + 1.53V + 0.45 \quad [6.7]$$

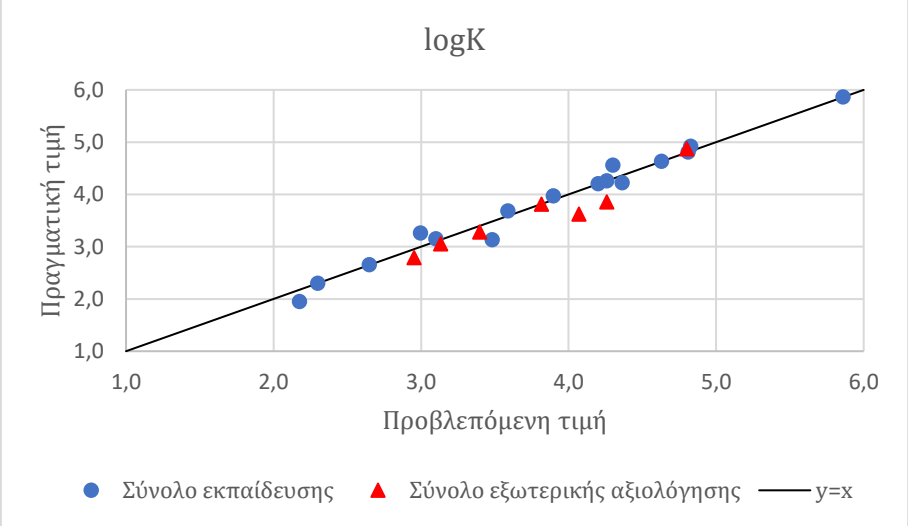
Εξωτερική επικύρωση

Το σύνολο δεδομένων χωρίστηκε αρχικά σε σύνολο μοντελοποίησης που περιείχε το 75% (21 ενώσεις) και σε σύνολο εξωτερικής αξιολόγησης με το 25% (7 ενώσεις). Το σύνολο μοντελοποίησης χωρίστηκε κατά το 75% σε σύνολο εκπαίδευσης με 16 ενώσεις και σε 5 ενώσεις που αποτέλεσαν το σύνολο δοκιμών. Για $\lambda = 0.01$ και $\beta = 0.05$ προκύπτει $Q^2_{ext} = 0.8664$. Επομένως, ο δείκτης απορρόφησης της βιολογικής επιφάνειας για άγνωστες ενώσεις προβλέπεται ικανοποιητικά. Στο Διάγραμμα 6.8 παρουσιάζονται οι γνωστές και οι προβλεπόμενες τιμές του δείκτη BSAI (logK) για τις ενώσεις του συνόλου εξωτερικής αξιολόγησης.

Πίνακας 6.7: Αποτελέσματα εξωτερικής επικύρωσης με επίλυση σε μία διάσταση για τα δεδομένα των Νανοσωλήνων άνθρακα.

	Νανοσωλήνες άνθρακα
z	0.1721
z test	0.3797
z external test	0.2607
R^2	0.9788

R^2 ανά περιοχή	Region 1: 0.9786 Region 2: -
Q^2 test	0.7845
Q^2 external test	0.8664
Περιοχές	2
Μεταβλητές	5
Αριθμός δειγμάτων ανά περιοχή	Region 1: 15 Region 2: 1
Αριθμός άγνωστων δειγμάτων ανά περιοχή	Region 1: 6 Region 2: 1
Υπολογιστικός χρόνος (min)	0.14



Διάγραμμα 6.8: Πραγματικές και προβλεπόμενες τιμές του logK των νανοσωλήνων άνθρακα, με επίλυση σε μία διάσταση για τα σύνολα εκπαίδευσης και εξωτερικής αξιολόγησης.

Πεδίο εφαρμογής μοντέλου

Στη συνέχεια, ορίστηκε το πεδίο εφαρμογής του μοντέλου και υπολογίστηκε η τιμή του προκαθορισμένου κατωφλιού APD ίση με 0.6637. Όλα τα δείγματα που χρησιμοποιήθηκαν για την επικύρωση του μοντέλου, έχουν απόσταση από τον κοντινότερο γείτονα του συνόλου εκπαίδευσης μικρότερη από την τιμή του κατωφλιού APD και οι προβλέψεις για αυτά θεωρούνται αξιόπιστες.

Έλεγχος τυχαίας επιλογής

Τα αποτελέσματα που προέκυψαν από τον έλεγχο τυχαίας επιλογής παρουσιάζονται στον παρακάτω πίνακα (Πίνακας 6.8).

Πίνακας 6.8: Αποτελέσματα ελέγχου τυχαίας επιλογής με επίλυση σε μία διάσταση για τα δεδομένα των Νανοσωλήνων άνθρακα.

	Νανοσωλήνες άνθρακα
z	0.4013
z test	0.9350
R^2	0.5876
R^2 ανά περιοχή	Region 1: 0.4631 Region 2: 0.8582
Q^2 test	-1.7815
Περιοχές	2
Μεταβλητές	5
Αριθμός δειγμάτων ανά περιοχή	Region 1: 15 Region 2: 6
Υπολογιστικός χρόνος (min)	0.15

Παρατηρείται ότι ο δείκτης Q^2_{test} παίρνει αρνητική τιμή επομένως επιβεβαιώνεται ότι το μοντέλο που προκύπτει από τυχαία κατανεμημένα δεδομένα εξόδου δε μπορεί να δώσει αξιόπιστες προβλέψεις για άγνωστα δεδομένα.

6.1.4 Επιφανειακά-τροποποιημένα νανοσωματίδια

Το σύνολο δεδομένων αποτελείται από 109 νανοσωματίδια με διαφορετικές οργανικές επικαλύψεις και 777 περιγραφείς των συγκεκριμένων νανοσωματιδίων. Λόγω όμοιων τιμών των μεταβλητών, μετά την κανονικοποίηση των δεδομένων σε εύρος τιμών [0,1] (Ενότητα 2.2.1) προκύπτουν 548 μεταβλητές. Το σετ δεδομένων χωρίζεται με τη μέθοδο Kennard and Stone σε σύνολο εκπαίδευσης με 82 δείγματα νανοσωματιδίων και σε σύνολο δοκιμών με 27 δείγματα. Τα αποτελέσματα της επίδρασης του όρου ομαλοποίησης παρουσιάζονται παρακάτω (Πίνακας 6.9).

Πίνακας 6.9: Αποτελέσματα επίλυσης σε μία διάσταση για τα δεδομένα των Επιφανειακά-τροποποιημένων νανοσωματιδίων.

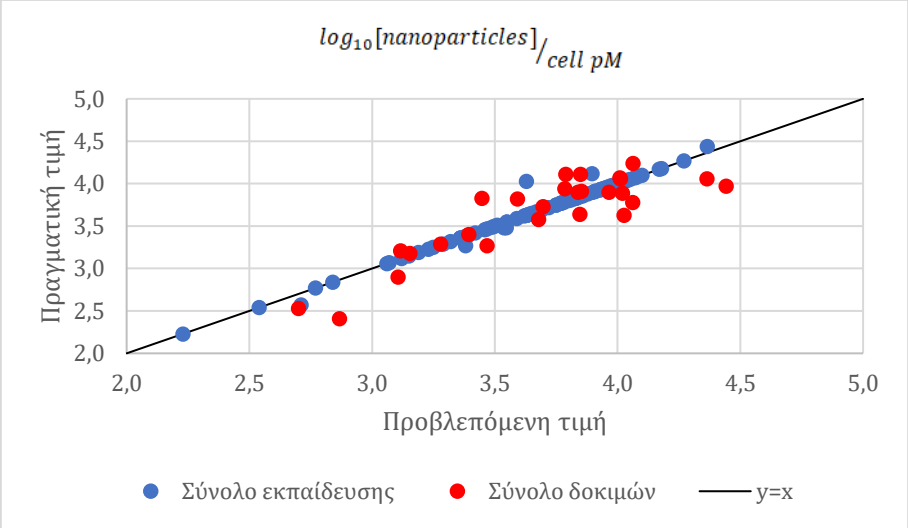
	Επιφανειακά-τροποποιημένα νανοσωματίδια
z	
λ=0.000	$-1.1066 \cdot 10^{-14}$
λ=0.005	0.0886
λ=0.010	0.1405
λ=0.020	0.1870
z test	
λ=0.000	7.3191
λ=0.005	0.2571
λ=0.010	0.3018
λ=0.020	0.2995

MAE	
λ=0.000	$-1.1066 \cdot 10^{-14}$
λ=0.005	0.0131
λ=0.010	0.0695
λ=0.020	0.1271
REG	
λ=0.000	$2.1609 \cdot 10^{4}$
λ=0.005	15.1026
λ=0.010	7.1013
λ=0.020	2.9925
R^2	
λ=0.000	1.0000
λ=0.005	0.9814
λ=0.010	0.8525
λ=0.020	0.6693
R^2 ανά περιοχή	
λ=0.000	Region 1: 1.0000 Region 2: 1.0000
λ=0.005	Region 1:0.9966 Region 2: 0.8383
λ=0.010	Region 1: 0.7620 Region 2: 0.6897
λ=0.020	Region 1: 0.5384 Region 2: 0.1711
Q^2 test	
λ=0.000	$-1.5177 \cdot 10^{3}$
λ=0.005	0.7645
λ=0.010	0.6461
λ=0.020	0.5595
Περιοχές	
λ=0.000	2
λ=0.005	2
λ=0.010	2
λ=0.020	2
Μεταβλητές	
λ=0.000	110
λ=0.005	69
λ=0.010	53
λ=0.020	27
Αριθμός δειγμάτων ανά περιοχή	
λ=0.000	Region 1: 76 Region 2: 6
λ=0.005	Region 1: 47 Region 2: 35
λ=0.010	Region 1: 61 Region 2: 21
λ=0.020	Region 1: 61 Region 2: 21
Υπολογιστικός χρόνος (min)	
λ=0.000	77.03

λ=0.005	79.31
λ=0.010	73.80
λ=0.020	74.45

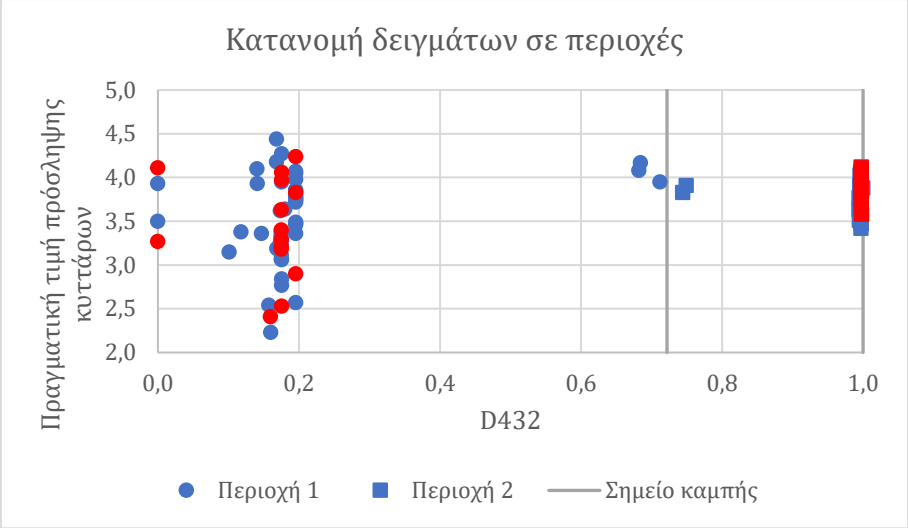
Όταν ο όρος της ομαλοποίησης απουσιάζει, για $\lambda = 0$, το σφάλμα πρόβλεψης MAE μηδενίζεται ενώ ο όρος ομαλοποίησης REG παίρνει την τιμή $REG = 2.1609 \cdot 10^4$. Η αντικειμενική συνάρτηση για το σύνολο εκπαίδευσης μηδενίζεται ενώ για το σύνολο δοκιμών παίρνει αρκετά υψηλή τιμή $z_{test} = 7.3191$. Ο δείκτης Q^2_{test} παίρνει μεγάλη αρνητική τιμή ($Q^2_{test} = -1.5177 \cdot 10^3$) που επιβεβαιώνει ότι το μοντέλο αδυνατεί να προβλέψει ικανοποιητικά άγνωστα δεδομένα.

Παρατηρείται ότι για $\lambda = 0.005$ γίνεται η καλύτερη πρόβλεψη για το σύνολο δοκιμών ($Q^2_{test} = 0.7645$). Επιλέγονται 69 από τις 548 μεταβλητές (42 στην πρώτη περιοχή διαμέρισης και 31 στη δεύτερη, 4 μεταβλητές επιλέχτηκαν και στις 2 περιοχές) ως απαραίτητες για την πρόβλεψη της πρόσληψης των κυττάρων PaCa2. Η πρόβλεψη της τιμής πρόσληψης των κυττάρων PaCa2 για τα δείγματα του συνόλου εκπαίδευσης ($R^2 = 0.9814$) και του συνόλου δοκιμών ($Q^2_{test} = 0.7645$) παρουσιάζεται στο Διάγραμμα 6.9.



Διάγραμμα 6.9: Πραγματικές και προβλεπόμενες τιμές πρόσληψης κυττάρων των επιφανειακά-τροποποιημένων νανοσωματιδίων με επίλυση σε μία διάσταση για τα σύνολα εκπαίδευσης και δοκιμών.

Στο Διάγραμμα 6.10 τα δείγματα του συνόλου εκπαίδευσης χωρίζονται σε δύο περιοχές διαμέρισης όπου 47 ανήκουν στην πρώτη περιοχή και 35 στη δεύτερη περιοχή διαμέρισης. Η μεταβλητή που επιλέγεται για τη διαμέριση των περιοχών είναι η 432η μεταβλητή (Topological structure autocorrelation length-2 weighted by atomic Sanderson Electronegativities), η οποία υπολογίζεται με τη χρήση του λογισμικού Mold2 από τις αρχικές μορφές SMILES που δόθηκαν τα δεδομένα.[54]



Διάγραμμα 6.10: Κατανομή δειγμάτων εκπαίδευσης και δοκιμών ανά περιοχή διαμέρισης και σημεία καμπής για τα δεδομένα των Επιφανειακά-τροποποιημένων νανοσωματιδίων. Με μπλε χρώμα απεικονίζονται τα δείγματα εκπαίδευσης και με κόκκινο τα δείγματα δοκιμών.

Εξωτερική επικύρωση

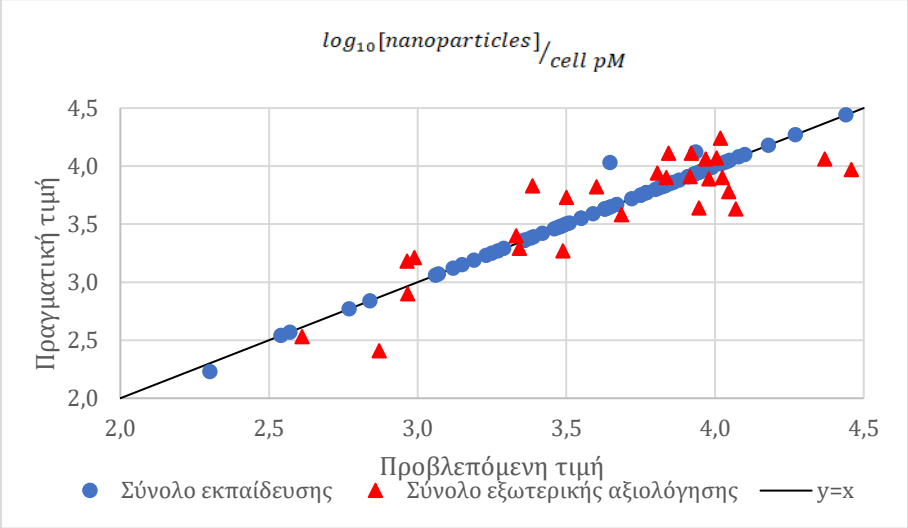
Το σύνολο δεδομένων χωρίστηκε αρχικά σε σύνολο μοντελοποίησης που περιείχε το 75% με 82 δείγματα και σε σύνολο εξωτερικής αξιολόγησης με 27 δείγματα. Το σύνολο μοντελοποίησης χωρίστηκε κατά το 90% σε σύνολο εκπαίδευσης με 74 δείγματα και σε 8 δείγματα που αποτέλεσαν το σύνολο δοκιμών. Για $\lambda = 0.005$ και $\beta = 0.05$ προκύπτουν τα αποτελέσματα (Πίνακας 6.10).

Πίνακας 6.10: Αποτελέσματα εξωτερικής επικύρωσης με επίλυση σε μία διάσταση για τα δεδομένα των Επιφανειακά-τροποποιημένων νανοσωματιδίων.

	Επιφανειακά-τροποποιημένα νανοσωματίδια
z	0.0861
z test	0.1962
z external test	0.2790
R^2	0.9857
R^2 ανά περιοχή	Region 1: 0.9996 Region 2: 0.8598
Q^2 test	0.7754
Q^2 external test	0.7317
Περιοχές	2
Μεταβλητές	68
Αριθμός δειγμάτων ανά περιοχή	Region 1: 42 Region 2: 32
Αριθμός άγνωστων δειγμάτων ανά περιοχή	Region 1: 14 Region 2: 13

Υπολογιστικός χρόνος (min)	77.68

Η πρόβλεψη της τιμής πρόσληψης κυττάρων για τα δείγματα του συνόλου εκπαίδευσης και του συνόλου εξωτερικής αξιολόγησης παρουσιάζεται στο Διάγραμμα 6.11.



Διάγραμμα 6.11: Πραγματικές και προβλεπόμενες τιμές πρόσληψης κυττάρων των επιφανειακά-τροποποιημένων νανοσωματιδίων με επίλυση σε μία διάσταση για τα σύνολα εκπαίδευσης και εξωτερικής αξιολόγησης.

Πεδίο εφαρμογής μοντέλου

Στη συνέχεια, ορίστηκε το πεδίο εφαρμογής του μοντέλου και υπολογίστηκε η τιμή του προκαθορισμένου κατωφλιού APD ίση με 5.2729. Όλα τα δείγματα που χρησιμοποιήθηκαν για την επικύρωση του μοντέλου δίνουν αξιόπιστες προβλέψεις.

Έλεγχος τυχαίας επιλογής

Τα αποτελέσματα από τον έλεγχο τυχαίας επιλογής παρουσιάζονται στον πίνακα που ακολουθεί (Πίνακας 6.11). Η αντικειμενική συνάρτηση για τα δείγματα ελέγχου παίρνει πολλαπλάσια τιμή από του συνόλου εκπαίδευσης και ο δείκτης Q^2_{test} παίρνει αρνητική τιμή. Τα αποτελέσματα του ελέγχου τυχαίας επιλογής επιβεβαιώνουν ότι το μοντέλο που προκύπτει από τυχαία κατανεμημένα δεδομένα εξόδου δε μπορεί να δώσει αξιόπιστες προβλέψεις για άγνωστα δεδομένα.

Πίνακας 6.11: Αποτελέσματα ελέγχου τυχαίας επιλογής με επίλυση σε μία διάσταση για τα δεδομένα των επιφανειακά-τροποποιημένων νανοσωματιδίων.

	Επιφανειακά-τροποποιημένα νανοσωματίδια
z	0.1165
z test	0.4924
R^2	0.9692
R^2 ανά περιοχή	Region 1: 0.8981 Region 2: 0.9978

Q^2 test	-0.2365
Περιοχές	2
Μεταβλητές	66
Αριθμός δειγμάτων ανά περιοχή	Region 1: 41 Region 2: 41
Υπολογιστικός χρόνος (min)	74.70

6.2 Αλγόριθμος επίλυσης σε δύο διαστάσεις

Όταν γίνεται η θεώρηση ότι οι μεταβλητές εισόδου χωρίζονται σε δύο κατηγορίες, M και N, η διάσπαση του πεδίου ορισμού σε περιοχές γίνεται σε δύο διαστάσεις, όπου η μία διάσταση αφορά τις μεταβλητές F_m και η άλλη τις μεταβλητές F_n. Το σύνολο εκπαίδευσης καθώς και το σύνολο δοκιμών χωρίζονται στα αντίστοιχα δύο σύνολα που αφορούν τις δύο κατηγορίες μεταβλητών. Στην περίπτωση αυτή, τα βήματα που ακολουθούνται παρουσιάζουν ορισμένες διαφορές από τα βήματα της Ενότητας 4.1.1. Από τα σύνολα δεδομένων που μελετήθηκαν στη συγκεκριμένη Εργασία, τα σύνολα των Μεταλλικών οξειδίων[46] και Νανοσωματιδίων χρυσού[6] διαθέτουν δύο κατηγορίες μεταβλητών, επομένως η μελέτη στις δύο διαστάσεις γίνεται μόνο στα συγκεκριμένα σύνολα. Στο πρώτο σύνολο, οι μεταβλητές χωρίζονται σε κβαντομηχανικές και γεωμετρικές/περιγραφικές ενώ στο δεύτερο σε φυσικοχημικές και βιολογικές. Επιπλέον, από την ανάλυση της επίλυσης σε μία διάσταση, παρατηρήθηκε ότι είναι απαραίτητος ο όρος της ομαλοποίησης. Στην περίπτωση που δεν υπάρχει (για $\lambda = 0$), το μοντέλο που προκύπτει εξαρτάται έντονα από τις μεταβλητές του συνόλου εκπαίδευσης και μικρές αποκλίσεις από αυτό οδηγούν σε μεγάλα σφάλματα για το σύνολο δοκιμών. Το μοντέλο αυτό είναι υπερπροσαρμοσμένο στα δεδομένα εκπαίδευσης και αδυνατεί να προβλέψει ικανοποιητικά άγνωστα δείγματα. Επομένως, για την επίλυση σε δύο διαστάσεις θεωρείται ότι η αντικειμενική συνάρτηση διαθέτει και τον όρο ομαλοποίησης. Τα βήματα που ακολουθούνται για την επίλυση στις δύο διαστάσεις αναλύονται παρακάτω.

1. Αρχικοποιούνται οι τιμές των σφαλμάτων πρόβλεψης ως άπειρες.
2. Επιλύεται απλή γραμμική παλινδρόμηση ($R = 1$) για το σύνολο εκπαίδευσης και αποθηκεύονται οι τιμές των σφαλμάτων πρόβλεψης τόσο για το σύνολο εκπαίδευσης όσο και κατά την εφαρμογή του μοντέλου απλής γραμμικής παλινδρόμησης στο σύνολο δοκιμών, $ERRORold$ και $ERROROld_{test}$ αντίστοιχα.
3. Όταν οι μεταβλητές διχοτόμησης επιλέγονται ταυτόχρονα, γίνεται επίλυση του αλγορίθμου OPLRA ταυτόχρονα στα δύο σύνολα εκπαίδευσης και με δύο περιοχές διαμέρισης ($R = 2$) σε κάθε διάσταση. Για κάθε μεταβλητή διχοτόμησης f_m από το σύνολο των μεταβλητών F_m αξιολογείται κάθε μεταβλητή διχοτόμησης f_n από το σύνολο των μεταβλητών F_n.
4. Ο συνδυασμός μεταβλητών διχοτόμησης f_m και f_n ο οποίος ελαχιστοποιεί την τιμή της αντικειμενικής συνάρτησης z, επιλέγεται ως το ζεύγος μεταβλητών διχοτόμησης f_m^* και f_n^*.
5. Στη συνέχεια εφαρμόζεται το μοντέλο OPLRA με δύο περιοχές διαμέρισης σε κάθε διάσταση και μεταβλητές διχοτόμησης f_m^* και f_n^* που προσδιορίστηκαν από το βήμα

4 στο σύνολο δοκιμών και αποθηκεύεται το σφάλμα πρόβλεψης $ERRORcurrent_{test} = z_{test}$.

6. Εξετάζεται η συνθήκη $ERRORcurrent_{test} < (1-\beta)ERRORold_{test}$ και στην περίπτωση που δεν ικανοποιείται προκύπτει το τελικό μοντέλο με δύο περιοχές διαμέρισης σε κάθε διάσταση, δηλαδή 4 περιοχές συνολικά.
7. Αν η συνθήκη του βήματος 6 ικανοποιείται, τότε εξετάζονται τρεις περιπτώσεις προσθήκης περιοχών διαμέρισης: αύξηση κατά μία περιοχή στη διάσταση m, αύξηση κατά μία περιοχή στη διάσταση n και αύξηση κατά μία περιοχή στη διάσταση m και μία στη διάσταση n ταυτόχρονα. Από τις διαφορετικές περιπτώσεις επιλέγεται εκείνη που δίνει το μικρότερο σφάλμα πρόβλεψης.
8. Εφαρμόζεται το μοντέλο που προκύπτει για συγκεκριμένο αριθμό περιοχών στο σύνολο δοκιμών και αποθηκεύονται τα σφάλματα $ERRORcurrent_{test} = z_{test}$ και $ERRORold_{test} = ERRORcurrent_{test\ (R-1)}$.
9. Εξετάζεται πάλι η συνθήκη προσθήκης περιοχών. Αν βελτιώνεται το σφάλμα, τότε η λύση αποθηκεύεται και εξετάζονται ξανά οι τρεις περιπτώσεις προσθήκης περιοχών και επιλέγεται η καλύτερη περίπτωση. Σε αντίθετη περίπτωση η λύση που προέκυψε προηγουμένως θεωρείται η τελική-βέλτιστη λύση.

Στην περίπτωση που οι μεταβλητές διχοτόμησης επιλέγονται διαδοχικά ή ανεξάρτητα, τα βήματα 3 και 4 του αλγορίθμου επίλυσης τροποποιούνται όπως παρουσιάζονται στη συνέχεια:

Διαδοχική επιλογή μεταβλητών διχοτόμησης

3. Γίνεται επίλυση του αλγορίθμου OPLRA στα δεδομένα εκπαίδευσης και με δύο περιοχές διαμέρισης ($R = 2$) επιλέγοντας κάθε φορά μία μεταβλητή f_m από το σύνολο των μεταβλητών F_m του συνόλου δεδομένων ως μεταβλητή διχοτόμησης στη διάσταση m. Η μεταβλητή f_m η οποία ελαχιστοποιεί την αντικειμενική συνάρτηση z επιλέγεται ως μεταβλητή διχοτόμησης f_m^*.
4. Με βάση τη μεταβλητή f_m^* αξιολογείται κάθε μεταβλητή διχοτόμησης f_n από το σύνολο των μεταβλητών F_n ως μεταβλητή διχοτόμησης στη διάσταση n. Ο συνδυασμός μεταβλητών διχοτόμησης f_m^* και f_n ο οποίος ελαχιστοποιεί την αντικειμενική συνάρτηση z, επιλέγεται ως το ζεύγος μεταβλητών διχοτόμησης f_m^* και f_n^*.

Ανεξάρτητη επιλογή μεταβλητών διχοτόμησης

3. Γίνεται επίλυση του αλγορίθμου OPLRA στα δεδομένα εκπαίδευσης και με δύο περιοχές διαμέρισης ($R = 2$) επιλέγοντας κάθε φορά μία μεταβλητή f_m από το σύνολο των μεταβλητών F_m του συνόλου δεδομένων ως μεταβλητή διχοτόμησης στη διάσταση m. Η διαδικασία επαναλαμβάνεται για τις μεταβλητές F_n στη διάσταση n.
4. Η μεταβλητή f_m για την οποία προκύπτει η μικρότερη τιμή της αντικειμενικής συνάρτησης z για τη διάσταση m, επιλέγεται ως μεταβλητή διχοτόμησης f_m^* ενώ η μεταβλητή f_n για την οποία προκύπτει η μικρότερη τιμή της αντικειμενικής συνάρτησης z για τη διάσταση n, επιλέγεται ως μεταβλητή διχοτόμησης f_n^*.

6.2.1 Μεταλλικά οξείδια

Το σύνολο των δεδομένων διαθέτει 18 νανοσωματίδια μεταλλικών οξειδίων και 29 μεταβλητές, οι οποίες χωρίζονται σε 18 κβαντομηχανικές ιδιότητες και 11 γεωμετρικές. Η διαμέριση του συγκεκριμένου συνόλου έγινε και στην περίπτωση των δύο διαστάσεων σύμφωνα με την δημοσίευση των Gajewicz *et al.*[46] σε 10 δείγματα στο σύνολο εκπαίδευσης και 8 στο σύνολο δοκιμών ώστε να μπορεί να γίνει σύγκριση των αποτελεσμάτων με την αρχική δημοσίευση.

6.2.1.1 Επίλυση σε δύο διαστάσεις: Ταυτόχρονη επιλογή μεταβλητών διχοτόμησης

Για τιμή της παραμέτρου ομαλοποίησης $\lambda = 0.02$ και παράμετρο απόκλισης σφαλμάτων $\beta = 0.05$ προκύπτουν τα αποτελέσματα που παρουσιάζονται παρακάτω.

Πίνακας 6.12: Αποτελέσματα ταυτόχρονης επιλογής μεταβλητών διχοτόμησης για τα δεδομένα των Μεταλλικών οξειδίων.

	Μεταλλικά οξείδια	
z	0.0274	
z test	0.4122	
R^2	1.0000	
R^2 ανά περιοχή	1.0000	-
	1.0000	-
Q^2 test	-1.6872	
Περιοχές στη διάσταση *m*	2	
Περιοχές στη διάσταση *n*	2	
Κβαντομηχανικές ιδιότητες	3	
Γεωμετρικές ιδιότητες	3	
Αριθμός δειγμάτων ανά περιοχή	6	1
	2	1
Υπολογιστικός χρόνος (min)	6.89	

Για το συγκεκριμένο σύνολο δεδομένων, η ταυτόχρονη επιλογή μεταβλητών διχοτόμησης στις δύο διαστάσεις δεν οδηγεί σε μοντέλο που μπορεί να δώσει ικανοποιητικές προβλέψεις για το σύνολο δοκιμών. Το σφάλμα πρόβλεψης για τα δεδομένα του συνόλου δοκιμών, ($z_{test} = 0.4122$) είναι πολύ μεγαλύτερο από του συνόλου εκπαίδευσης, ($z = 0.0274$) και ο δείκτης Q^2_{test} παίρνει αρνητική τιμή.

6.2.1.2 Επίλυση σε δύο διαστάσεις: Διαδοχική επιλογή μεταβλητών διχοτόμησης

Για τιμή της παραμέτρου ομαλοποίησης $\lambda = 0.02$ και παράμετρο απόκλισης σφαλμάτων $\beta = 0.05$ προκύπτουν τα αποτελέσματα του ακόλουθου πίνακα (Πίνακας 6.13). Το μοντέλο που προκύπτει δε μπορεί να δώσει ικανοποιητικές προβλέψεις για τα δείγματα ελέγχου. Επιπλέον παρατηρείται ότι η επίλυση με διαδοχική μεταβλητών διχοτόμησης καταλήγει στην ίδια βέλτιστη λύση με εκείνη που προκύπτει κατά την ταυτόχρονη επιλογή μεταβλητών.

Πίνακας 6.13: Αποτελέσματα διαδοχικής επιλογής μεταβλητών διχοτόμησης για τα δεδομένα των Μεταλλικών οξειδίων.

	Μεταλλικά οξείδια	
z	0.0274	
z test	0.4122	
R^2	1.0000	
R^2 ανά περιοχή	1.0000	-
	1.0000	-
Q^2 test	-1.6872	
Περιοχές στη διάσταση *m*	2	
Περιοχές στη διάσταση *n*	2	
Κβαντομηχανικές ιδιότητες	3	
Γεωμετρικές ιδιότητες	3	
Αριθμός δειγμάτων ανά περιοχή	6	1
	2	1
Υπολογιστικός χρόνος (min)	0.61	

6.2.1.3 Επίλυση σε δύο διαστάσεις: Ανεξάρτητη επιλογή μεταβλητών διχοτόμησης

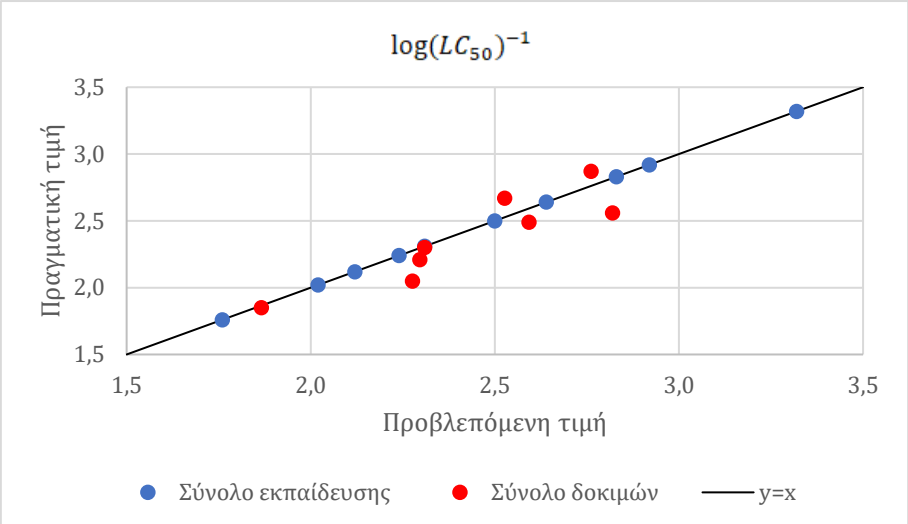
Για τιμή της παραμέτρου ομαλοποίησης $\lambda = 0.02$ και παράμετρο απόκλισης σφαλμάτων $\beta = 0.05$ προκύπτουν τα αποτελέσματα που παρουσιάζονται παρακάτω.

Πίνακας 6.14: Αποτελέσματα ανεξάρτητης επιλογής μεταβλητών διχοτόμησης για τα δεδομένα των Μεταλλικών οξειδίων.

	Μεταλλικά οξείδια
z	0.0528
z test	0.1720
R^2	1.0000

R^2 ανά περιοχή	-	1.0000
	-	-
Q^2 test	0.8030	
Περιοχές στη διάσταση *m*	2	
Περιοχές στη διάσταση *n*	2	
Κβαντομηχανικές ιδιότητες	4	
Γεωμετρικές ιδιότητες	3	
Αριθμός δειγμάτων ανά περιοχή	1	8
	0	1
Αριθμός άγνωστων δειγμάτων ανά περιοχή	1	7
	0	0
Υπολογιστικός χρόνος (min)	0.38	

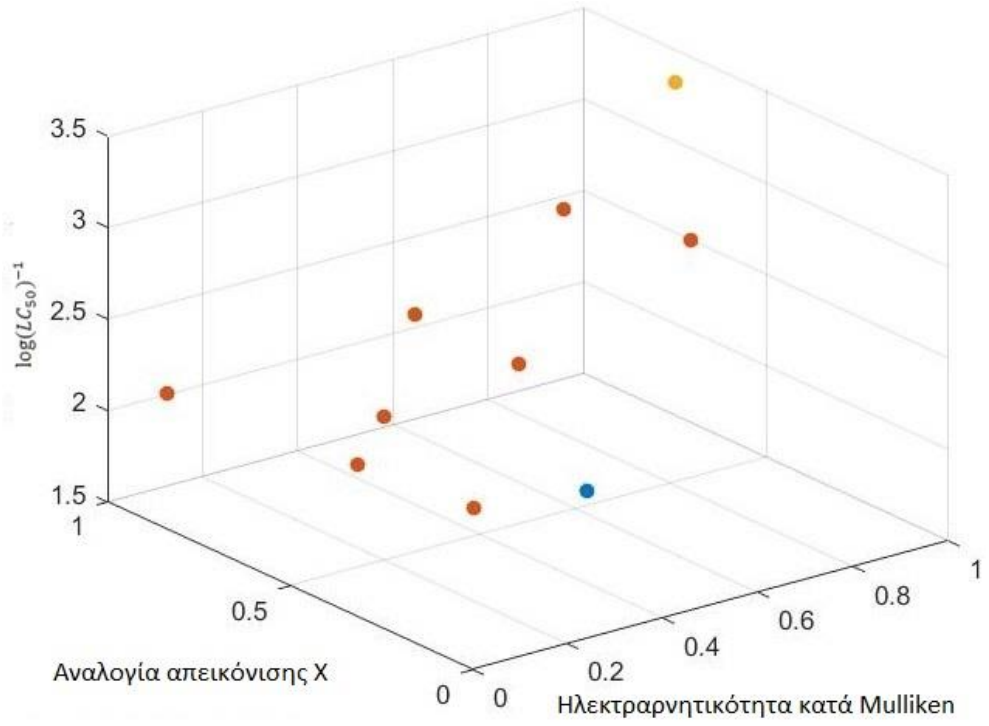
Στην περίπτωση της ανεξάρτητης επιλογής μεταβλητών διχοτόμησης, η αντικειμενική συνάρτηση προκύπτει ίση με $z = 0.0528$ και για το σύνολο δοκιμών $z_{test} = 0.1720$. Ο δείκτης Q^2_{test} παίρνει την τιμή 0.8030, επομένως προβλέπονται αρκετά καλά τα δείγματα του συνόλου δοκιμών. Επιλέγονται 4 κβαντομηχανικές και 3 γεωμετρικές μεταβλητές για την πρόβλεψη της τοξικότητας. Το Διάγραμμα 6.12 δείχνει τη συσχέτιση προβλεπόμενης και πραγματικής τιμής τοξικότητας για το σύνολο εκπαίδευσης και το σύνολο δοκιμών. Όπως φαίνεται και από το διάγραμμα, η πρόβλεψη για τα δείγματα δοκιμών γίνεται ικανοποιητικά καθώς οι προβλεπόμενες και οι πραγματικές τιμές προσεγγίζουν την ευθεία $y = x$.



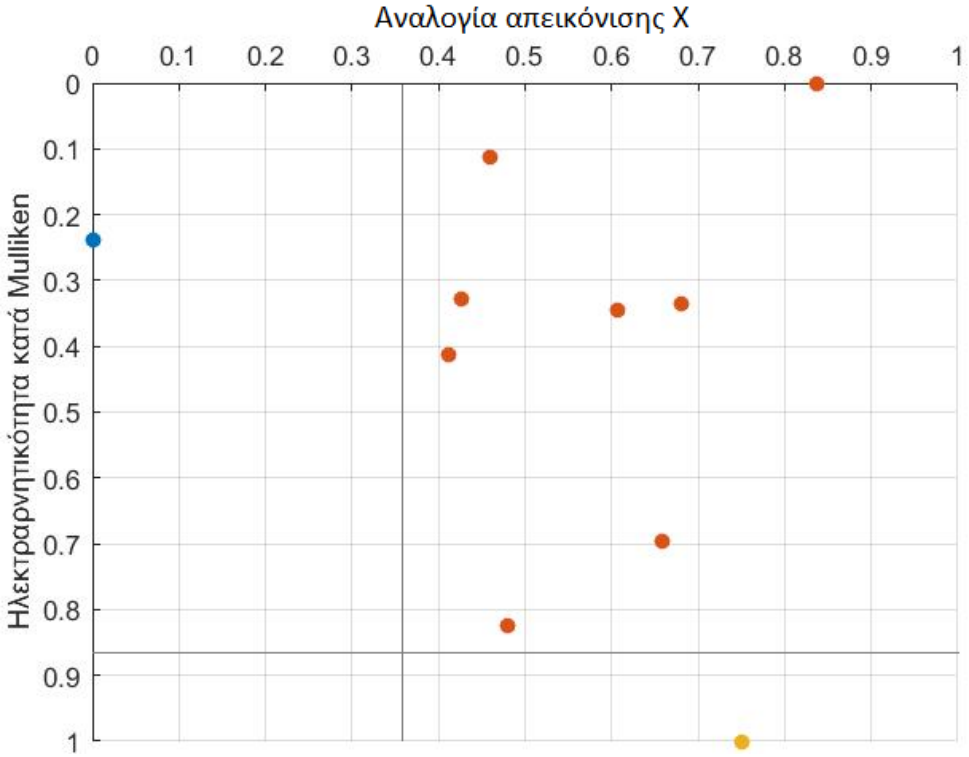
Διάγραμμα 6.12: Πραγματικές και προβλεπόμενες τιμές $\log(LC_{50})^{-1}$ των μεταλλικών οξειδίων με ανεξάρτητη επιλογή μεταβλητών διχοτόμησης για τα σύνολα εκπαίδευσης και δοκιμών.

Για την διάσπαση σε περιοχές, επιλέγεται από τις κβαντομηχανικές μεταβλητές η ηλεκτραρνητικότητα κατά Mulliken και από τις περιγραφικές η αναλογία απεικόνισης X (aspect ratio X). Τα δείγματα του συνόλου εκπαίδευσης τοποθετούνται στο χώρο όπως φαίνονται στο Διάγραμμα 6.13. Στο Διάγραμμα 6.14 και το Διάγραμμα 6.15

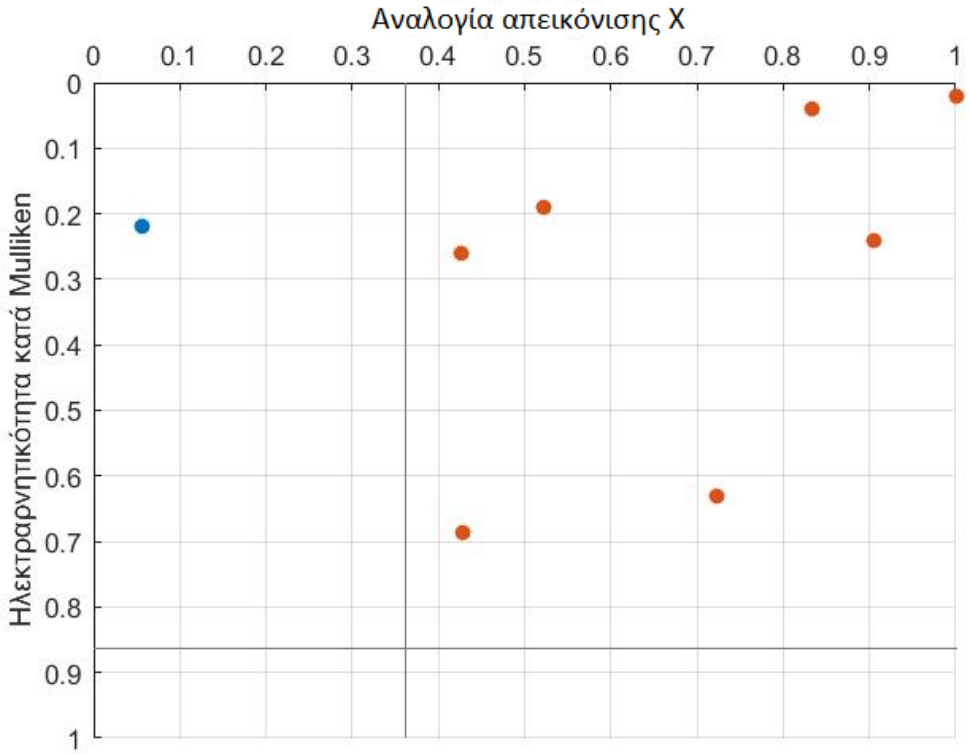
παρουσιάζεται η κατανομή των δειγμάτων εκπαίδευσης και δοκιμών σε περιοχές και τα σημεία καμπής των περιοχών διαμέρισης.



Διάγραμμα 6.13: Κατανομή των δειγμάτων εκπαίδευσης στο χώρο για τα δεδομένα των Μεταλλικών οξειδίων και με ανεξάρτητη επιλογή μεταβλητών διχοτόμησης.



Διάγραμμα 6.14: Κατανομή δειγμάτων εκπαίδευσης και σημεία καμπής των περιοχών διαμέρισης για τα δεδομένα των Μεταλλικών οξειδίων και με ανεξάρτητη επιλογή μεταβλητών διχοτόμησης.



Διάγραμμα 6.15: Κατανομή δειγμάτων δοκιμών και σημεία καμπής των περιοχών διαμέρισης για τα δεδομένα των Μεταλλικών οξειδίων και με ανεξάρτητη επιλογή μεταβλητών διχοτόμησης.

Για το σύνολο εκπαίδευσης, το ένα δείγμα που ανήκει στην περιοχή $(r_m = 1, r_n = 1)$ και το δείγμα που ανήκει στην περιοχή $(r_m = 2, r_n = 2)$ έχουν τιμή εξόδου που ισούται με το σταθερό όρο $B^{r_m r_n}$, καθώς δεν επιλέγεται καμία μεταβλητή για την πρόβλεψη (Εξισώσεις

[6.8] και [6.9] αντίστοιχα). Τα 8 δείγματα της περιοχή ($r_m = 1, r_n = 2$) υπολογίζονται μέσω της εξίσωσης γραμμικής παλινδρόμησης [6.10] για την οποία απαιτούνται 7 μεταβλητές, 4 κβαντομηχανικές και 3 περιγραφικές.

$$\log(LC_{50})^{-1} = 2.31 \quad [6.8]$$

$$\log(LC_{50})^{-1} = 3.32 \quad [6.9]$$

$$\begin{aligned}\log(LC_{50})^{-1} = {} & 0.96\Delta H_f^c - 0.59LUMO + 0.16S - 0.65Shift \\ & - 0.10volume.mass.diameter \\ & - 0.03volume.surface.diameter - 0.16PorosityY + 2.56\end{aligned} \quad [6.10]$$

Έλεγχος τυχαίας επιλογής

Πραγματοποιήθηκε έλεγχος τυχαίας επιλογής και τα αποτελέσματα του μοντέλου που προέκυψε με ανακατεμένες τις τιμές της τυχαίας μεταβλητής παρουσιάζονται στον ακόλουθο πίνακα (Πίνακας 6.15). Από τα αποτελέσματα που προκύπτουν, επαληθεύεται ότι με τις τυχαία ανακατεμένες τιμές εξόδου το μοντέλο που προκύπτει δεν μπορεί να προβλέψει άγνωστα δεδομένα, επομένως δεν υπάρχει πιθανότητα τυχαίας συσχέτισης των δεδομένων στο αρχικό μοντέλο.

Πίνακας 6.15: Αποτελέσματα ελέγχου τυχαίας επιλογής με ανεξάρτητη επιλογή μεταβλητών διχοτόμησης για τα δεδομένα των Μεταλλικών οξειδίων.

	Μεταλλικά οξείδια	
z	0.0437	
z test	0.3656	
R^2	1.0000	
R^2 ανά περιοχή	- 1.0000	- 1.0000
Q^2 test	-0.7022	
Περιοχές στη διάσταση *m*	2	
Περιοχές στη διάσταση *n*	2	
Κβαντομηχανικές ιδιότητες	2	
Γεωμετρικές ιδιότητες	4	
Αριθμός δειγμάτων ανά περιοχή	1 4	1 4
Υπολογιστικός χρόνος (min)	0.65	

6.2.2 Νανοσωματίδια χρυσού

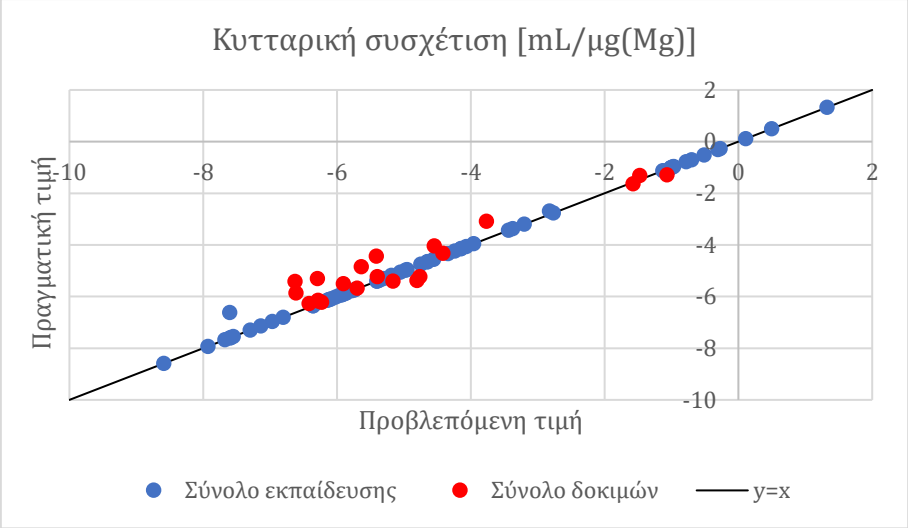
6.2.2.1 Επίλυση σε δύο διαστάσεις: Ταυτόχρονη επιλογή μεταβλητών διχοτόμησης

Για τιμή της παραμέτρου ομαλοποίησης $\lambda = 0.005$ και παράμετρο απόκλισης σφαλμάτων $\beta = 0.05$ προκύπτουν τα αποτελέσματα που παρουσιάζονται παρακάτω.

Πίνακας 6.16: Αποτελέσματα ταυτόχρονης επιλογής μεταβλητών διχοτόμησης για τα δεδομένα των νανοσωματιδίων χρυσού.

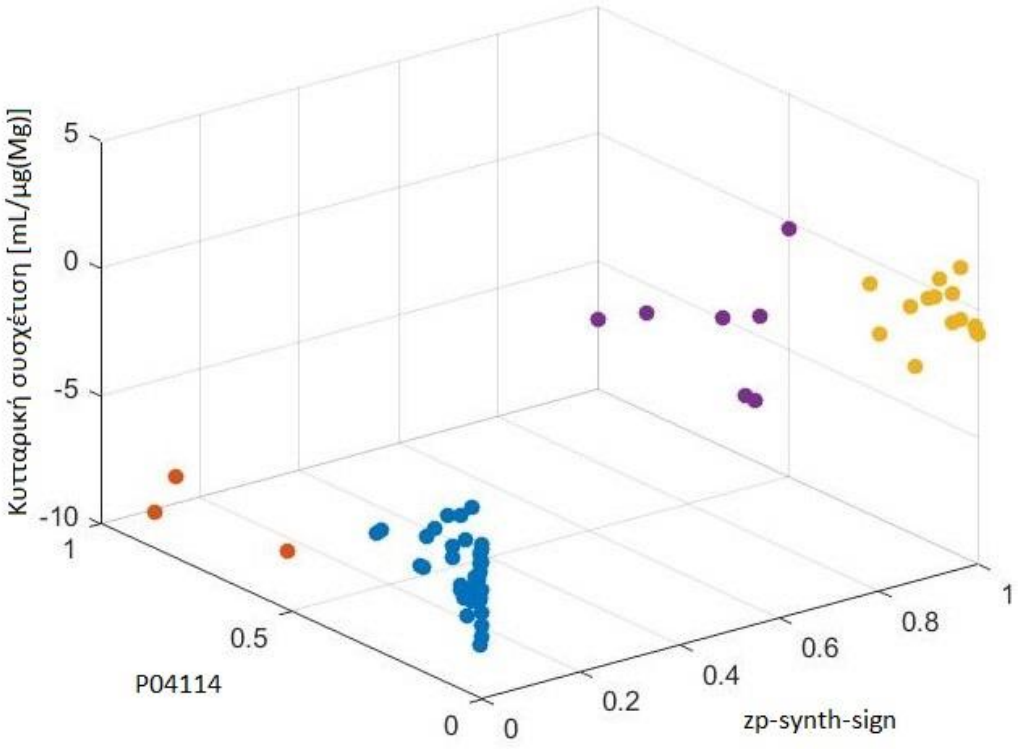
	Νανοσωματίδια χρυσού	
z	0.2746	
z test	0.6709	
R^2	0.9974	
R^2 ανά περιοχή	0.9819	1.0000
	1.0000	1.0000
Q^2 test	0.8777	
Περιοχές στη διάσταση *m*	2	
Περιοχές στη διάσταση *n*	2	
Φυσικοχημικές Μεταβλητές	13	
Βιολογικές Μεταβλητές	34	
Αριθμός δειγμάτων ανά περιοχή	39	3
	14	7
Υπολογιστικός χρόνος (min)	450.37	

Η πρόβλεψη της κυτταρικής συσχέτισης για τα δείγματα δοκιμών γίνεται ικανοποιητικά, όπως δείχνει και ο δείκτης Q^2_{test} που παίρνει την τιμή 0.8777. Η συσχέτιση προβλεπόμενης και πραγματικής τιμής για το σύνολο εκπαίδευσης και το σύνολο δοκιμών παρουσιάζεται στο Διάγραμμα 6.16.

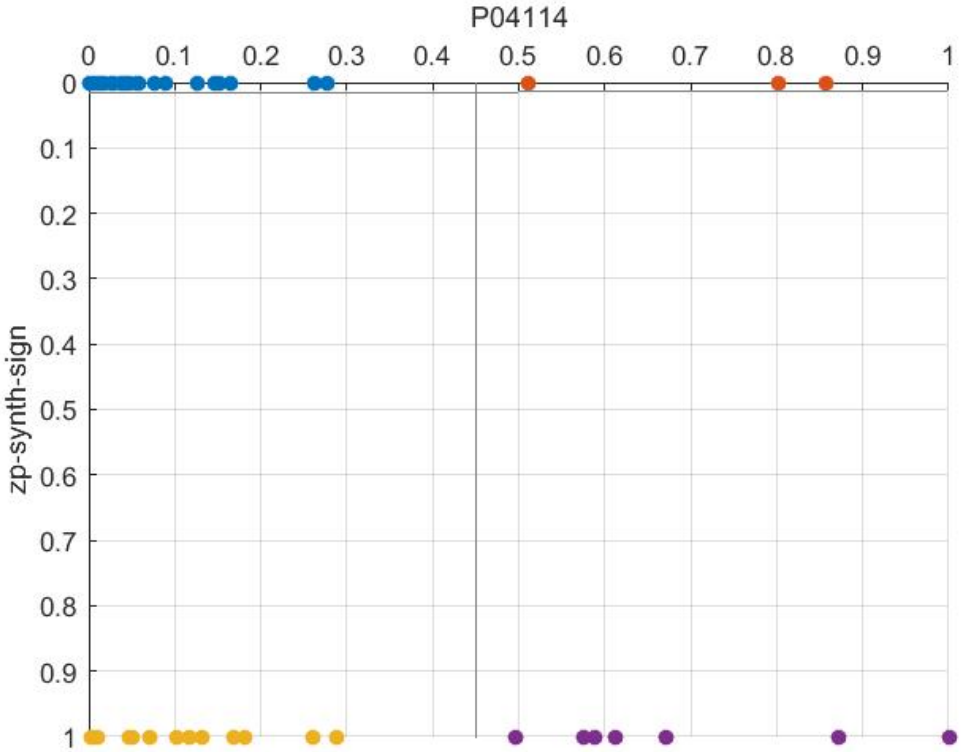


Διάγραμμα 6.16: Πραγματικές και προβλεπόμενες τιμές κυτταρικής συσχέτισης των νανοσωματιδίων χρυσού με επίλυση ταυτόχρονης επιλογής μεταβλητών διχοτόμησης για τα σύνολα εκπαίδευσης και δοκιμών.

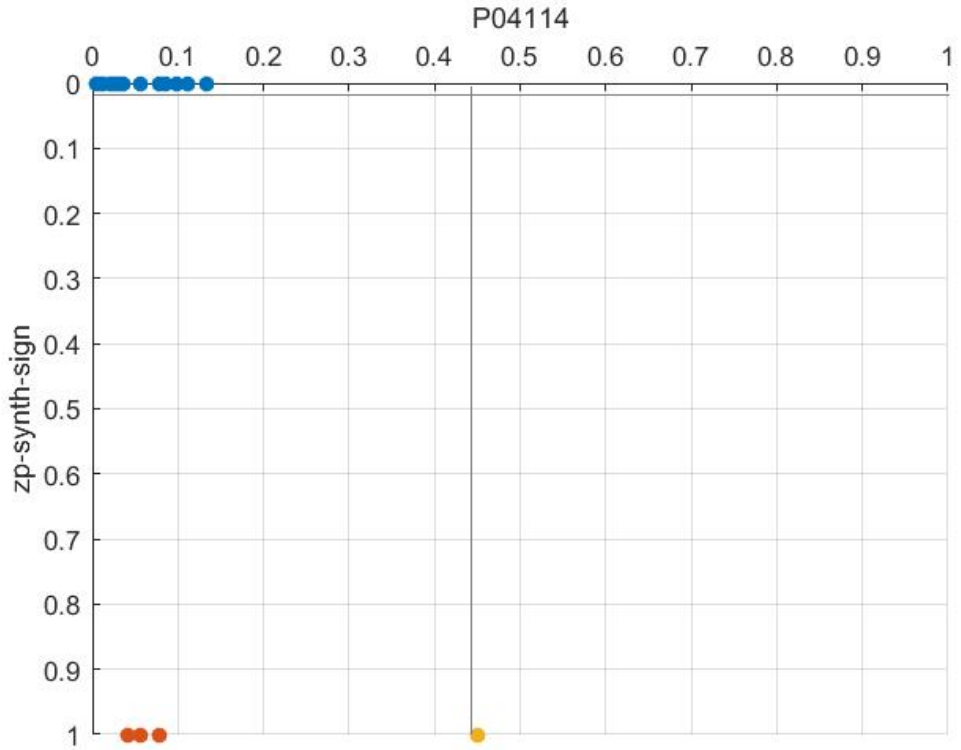
Τα δείγματα εκπαίδευσης κατανέμονται στο χώρο όπως φαίνεται στο Διάγραμμα 6.17. Οι μεταβλητές που επιλέγονται για τη διαμέριση των δειγμάτων σε περιοχές είναι η ένδειξη του ζ-δυναμικού σύνθεσης ($zp - synth - sign$) και η Απολιποπρωτεΐνη B-100 ($P04114$) και τα σημεία καμπής των περιοχών παρουσιάζονται στο Διάγραμμα 6.18 και το Διάγραμμα 6.19.



Διάγραμμα 6.17: Κατανομή των δειγμάτων εκπαίδευσης στο χώρο για τα δεδομένα των νανοσωματιδίων χρυσού και με ταυτόχρονη επιλογή μεταβλητών διχοτόμησης.



Διάγραμμα 6.18: Κατανομή των δειγμάτων εκπαίδευσης και σημεία καμπής των περιοχών διαμέρισης για τα δεδομένα των νανοσωματιδίων χρυσού και με ταυτόχρονη επιλογή μεταβλητών διχοτόμησης.



Διάγραμμα 6.19: Κατανομή των δειγμάτων δοκιμών και σημεία καμπής των περιοχών διαμέρισης για τα δεδομένα των νανοσωματιδίων χρυσού και με ταυτόχρονη επιλογή μεταβλητών διχοτόμησης.

Τα 39 δείγματα εκπαίδευσης που κατατάσσονται στην περιοχή $(r_m = 1, r_n = 1)$ προβλέπονται από τη γραμμική εξίσωση [6.11] η οποία επιλέγει 36 μεταβλητές, τα 3 δείγματα της περιοχής $(r_m = 1, r_n = 2)$ υπολογίζονται από την εξίσωση [6.12], τα 14 δείγματα της περιοχής $(r_m = 2, r_n = 1)$ από την [6.13] και τέλος τα 7 δείγματα της περιοχής $(r_m = 2, r_n = 2)$ από την [6.14].

$$\begin{aligned} net.c = {} & 0.23class + 1.66lspri.serum + 0.20zav.synth + 0.81zav.serum \\ & + 0.24num.synth + 3.14int.serum + 1.54pdi.rel \\ & + 0.55zp.serum - 0.03zp.synth.mag - 0.37AS.total \\ & - 0.65P0C0L4 + 1.10P02649 + 0.03Q1462 - 0.09P02743 \\ & + 0.97P10720 - 0.10P05546 - 0.41P49908 \\ & + 0.66Q03591 + 0.40Q43866 + 1.13P02749 \\ & + 2.98P02654 - 1.44P03952 - 0.87P01011 \\ & + 0.28P18428 + 0.25P00736 + 0.65P00748 \\ & - 0.26P02774 + 1.34P00751 + 0.15P03950 \\ & - 1.01P02790 + 0.70P18065 - 0.60P00450 \\ & + 0.45P08567 + 0.54P08709 + 0.57P00451 \\ & - 0.36P23528 - 8.64 \end{aligned} \quad [6.11]$$

$$net.c = 1.53zp.synth.mag + 1.31AS.total - 8.78 \quad [6.12]$$

$$\begin{aligned} net.c = {} & 0.57lspri.serum + 1.51vol.synth + 0.01zav.rel + 0.44vol.rel \\ & + 0.77P01009 + 0.83P00738 + 0.14P01011 \\ & - 1.93P18428 - 1.04P02655 + 0.66P08567 \\ & + 1.45P01019 + 0.44P02671 + 0.21Q99467 - 2.07 \end{aligned} \quad [6.13]$$

$$\begin{aligned} net.c = {} & 0.43P01009 + 0.16P68871 + 1.23P02749 - 0.12P02774 \\ & + 1.88P27169 + 7.95P01019 - 8.00 \end{aligned} \quad [6.14]$$

Εξωτερική επικύρωση

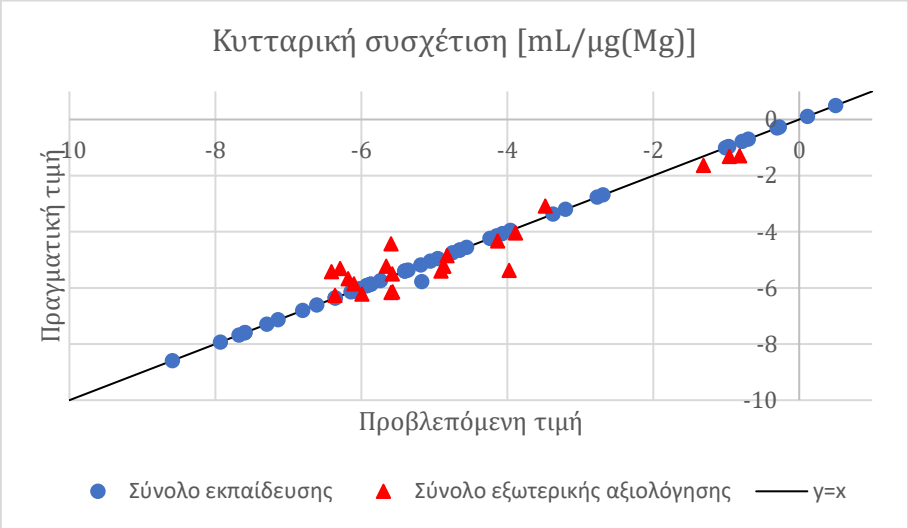
Στη συνέχεια, το αρχικό σύνολο δεδομένων χωρίζεται σε δύο φάσεις με τη μέθοδο Kennard and Stone, σε σύνολο μοντελοποίησης και σύνολο εξωτερικής αξιολόγησης με 63 και 21 δείγματα. Το πρώτο χωρίζεται επιπλέον σε σύνολο εκπαίδευσης με 50 δείγματα και σύνολο δοκιμών με 13 δείγματα. Για παράμετρο ομαλοποίησης $\lambda = 0.01$ προκύπτουν τα ακόλουθα αποτελέσματα.

Πίνακας 6.17: Αποτελέσματα εξωτερικής επικύρωσης με ταυτόχρονη επιλογή μεταβλητών διχοτόμησης για τα δεδομένα των νανοσωματιδίων χρυσού.

	Νανοσωματίδια χρυσού
z	0.4253
z test	1.1065

z external test	0.7970	
R^2	0.9925	
R^2 ανά περιοχή	0.9639	1.0000
	0.9785	1.0000
Q^2 test	0.8802	
Q^2 external test	0.8731	
Περιοχές στη διάσταση *m*	2	
Περιοχές στη διάσταση *n*	2	
Φυσικοχημικές Μεταβλητές	10	
Βιολογικές Μεταβλητές	25	
Αριθμός δειγμάτων ανά περιοχή	30	3
	10	7
Αριθμός άγνωστων δειγμάτων ανά περιοχή	17	0
	3	1
Υπολογιστικός χρόνος (min)	433.84	

Τόσο για τα δείγματα δοκιμής όσο και της εξωτερικής αξιολόγησης οι προβλέψεις γίνονται ικανοποιητικά όπως δείχνουν οι δείκτες Q^2_{test} και Q^2_{ext} οι οποίοι παίρνουν τιμές μεγαλύτερες από 0.85. Το ίδιο παρατηρείται και στο Διάγραμμα 6.20 που δείχνει τις προβλεπόμενες και τις πραγματικές τιμές κυτταρικής συσχέτισης του συνόλου εξωτερικής αξιολόγησης.



Διάγραμμα 6.20: Πραγματικές και προβλεπόμενες τιμές κυτταρικής συσχέτισης των νανοσωματιδίων χρυσού με επίλυση ταυτόχρονης επιλογής μεταβλητών διχοτόμησης για τα σύνολα εκπαίδευσης και εξωτερικής αξιολόγησης.

Πεδίο εφαρμογής μοντέλου

Ορίστηκε το πεδίο εφαρμογής του μοντέλου και υπολογίστηκε η τιμή του κατωφλιού APD ίση με 2.8977. Όλα τα δείγματα του συνόλου δοκιμών, έχουν απόσταση από τον κοντινότερο γείτονα του συνόλου εκπαίδευσης μικρότερη από την τιμή του κατωφλιού APD. Επομένως οι προβλέψεις για τα δείγματα ελέγχου μπορούν να θεωρηθούν αξιόπιστες. Το ίδιο ισχύει και στις περιπτώσεις διαδοχικής και ανεξάρτητης επιλογής μεταβλητών καθώς γίνεται η ίδια διαμέριση των δειγμάτων.

Έλεγχος τυχαίας επιλογής

Για έλεγχο του παραγόμενου μοντέλου, πραγματοποιήθηκε έλεγχος τυχαίας επιλογής. Προκύπτει ο δείκτης Q^2_{test} για το σύνολο δοκιμών ίσος με $Q^2_{test} = -2.2024$ Από τα αποτελέσματα που προέκυψαν (Πίνακας 6.18) συμπεραίνεται ότι δεν υπάρχει πιθανότητα τυχαίας συσχέτισης των δεδομένων και των τιμών εξόδου στο μοντέλο.

Πίνακας 6.18: Αποτελέσματα ελέγχου τυχαίας επιλογής με ταυτόχρονη επιλογή μεταβλητών διχοτόμησης για τα δεδομένα των νανοσωματιδίων χρυσού.

	Νανοσωματίδια χρυσού	
z	0.6939	
z test	3.1258	
R^2	0.9995	
R^2 ανά περιοχή	0.9994	1.0000
	1.0000	-
Q^2 test	-2.2024	
Περιοχές στη διάσταση *m*	2	
Περιοχές στη διάσταση *n*	2	
Φυσικοχημικές Μεταβλητές	13	
Βιολογικές Μεταβλητές	37	
Αριθμός δειγμάτων ανά περιοχή	45	5
	12	1
Υπολογιστικός χρόνος (min)	450.37	

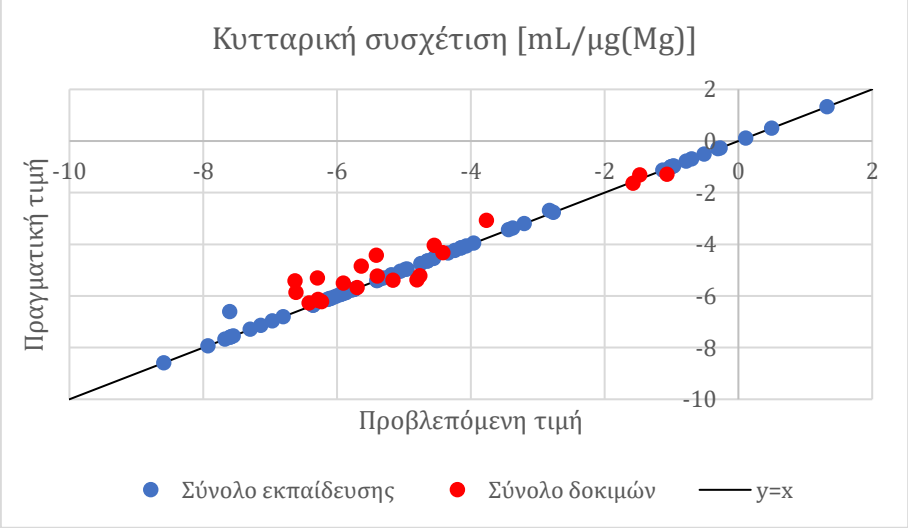
6.2.2.2 Επίλυση σε δύο διαστάσεις: Διαδοχική επιλογή μεταβλητών διχοτόμησης

Για τιμή της παραμέτρου ομαλοποίησης $\lambda = 0.005$ και παράμετρο απόκλισης σφαλμάτων $\beta = 0.05$ προκύπτουν τα αποτελέσματα που παρουσιάζονται παρακάτω.

Πίνακας 6.19: Αποτελέσματα διαδοχικής επιλογής μεταβλητών διχοτόμησης για τα δεδομένα των νανοσωματιδίων χρυσού.

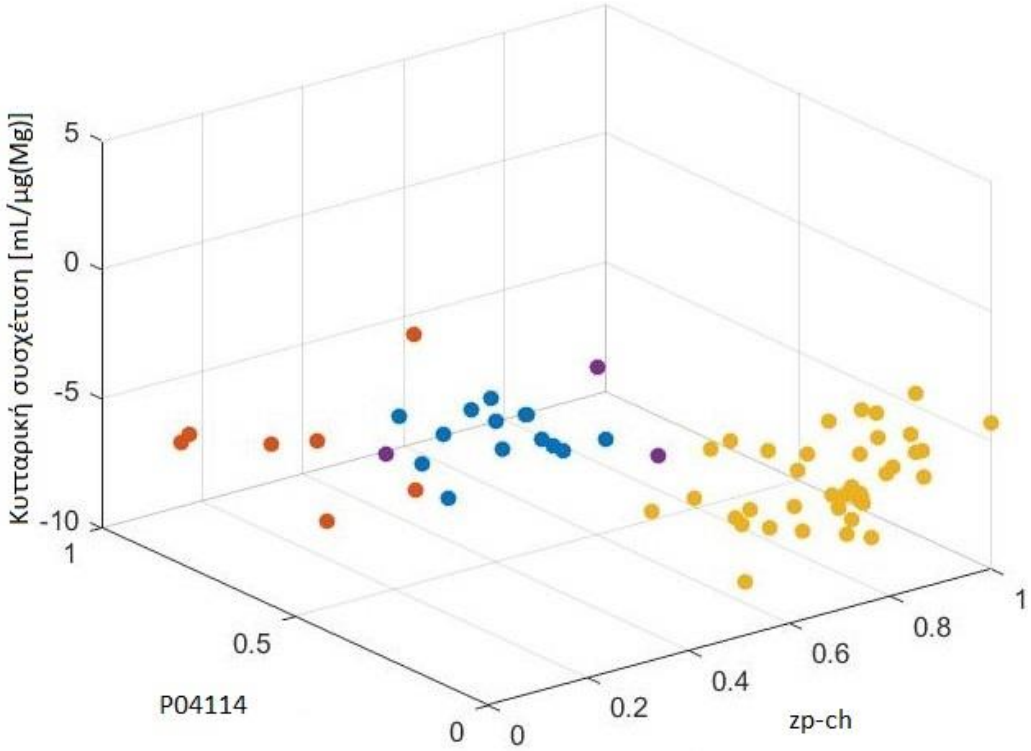
	Νανοσωματίδια χρυσού	
z	0.2746	
z test	0.6709	
R^2	0.9974	
R^2 ανά περιοχή	1.0000	1.0000
	0.9819	1.0000
Q^2 test	0.8777	
Περιοχές στη διάσταση *m*	2	
Περιοχές στη διάσταση *n*	2	
Φυσικοχημικές Μεταβλητές	13	
Βιολογικές Μεταβλητές	34	
Αριθμός δειγμάτων ανά περιοχή	14	7
	39	3
Υπολογιστικός χρόνος (min)	10.62	

Προκύπτει τιμή αντικειμενικής συνάρτησης για το σύνολο εκπαίδευσης ίση με $z = 0.2746$ ενώ για το σύνολο δοκιμών $z_{test} = 0.6709$. Ο δείκτης Q^2_{test} προκύπτει ίσος με 0.8777, επομένως, το μοντέλο δίνει ικανοποιητικές προβλέψεις για το σύνολο δοκιμών. Επιλέγονται 13 από τις 40 φυσικοχημικές μεταβλητές και 34 από τις 63 βιολογικές μεταβλητές για την πρόβλεψη της τοξικότητας. Στο Διάγραμμα 6.21 παρουσιάζεται η σχέση μεταξύ πραγματικών και προβλεπόμενων τιμών για το σύνολο εκπαίδευσης και για το σύνολο δοκιμών.

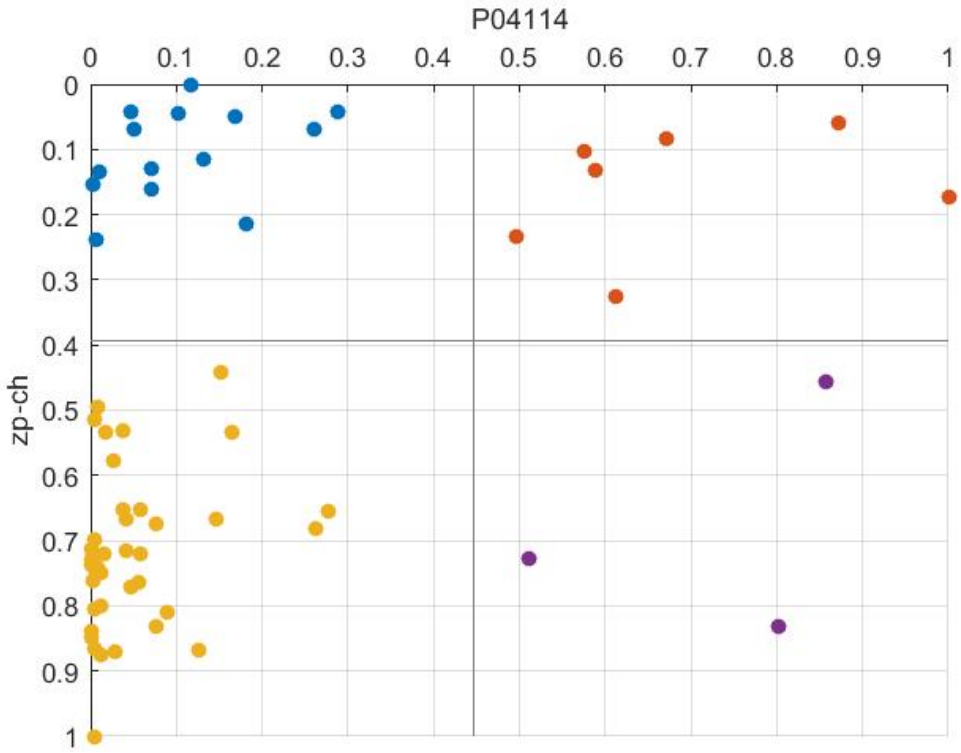


Διάγραμμα 6.21: Πραγματικές και προβλεπόμενες τιμές κυτταρικής συσχέτισης των νανοσωματιδίων χρυσού με επίλυση διαδοχικής επιλογής μεταβλητών διχοτόμησης για τα σύνολα εκπαίδευσης και δοκιμών.

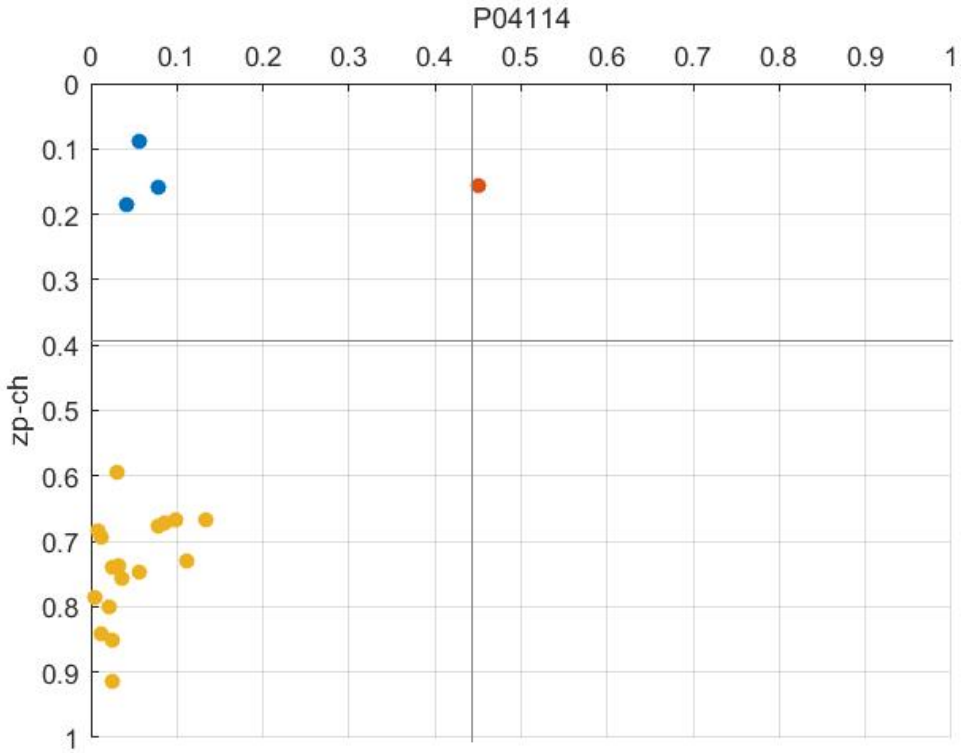
Τα δείγματα του συνόλου εκπαίδευσης τοποθετούνται στο χώρο όπως φαίνεται στο Διάγραμμα 6.22. Στο Διάγραμμα 6.23 και το Διάγραμμα 6.24 παρουσιάζονται τα δείγματα εκπαίδευσης και δοκιμής και τα σημεία καμπής των περιοχών διαμέρισης. Ως μεταβλητή διχοτόμησης από τις φυσικοχημικές μεταβλητές επιλέχτηκε το ζ δυναμικό ορού-σύνθεσης ($zp - ch$) ενώ ως μεταβλητή διχοτόμησης των βιολογικών μεταβλητών η Απολιποπρωτεΐνη B-100 ($P04114$).



Διάγραμμα 6.22: Κατανομή των δειγμάτων εκπαίδευσης στο χώρο για τα δεδομένα των νανοσωματιδίων χρυσού και με διαδοχική επιλογή μεταβλητών διχοτόμησης.



Διάγραμμα 6.23: Κατανομή των δειγμάτων εκπαίδευσης και σημεία καμπής των περιοχών διαμέρισης για τα δεδομένα των νανοσωματιδίων χρυσού και με διαδοχική επιλογή μεταβλητών διχοτόμησης.



Διάγραμμα 6.24: Κατανομή των δειγμάτων δοκιμών και σημεία καμπής των περιοχών διαμέρισης για τα δεδομένα των νανοσωματιδίων χρυσού και με διαδοχική επιλογή μεταβλητών διχοτόμησης.

Τα 14 δείγματα εκπαίδευσης που κατατάσσονται στην περιοχή $(r_m = 1, r_n = 1)$ προβλέπονται από τη γραμμική εξίσωση [6.15] η οποία επιλέγει 13 μεταβλητές, τα 7 δείγματα της περιοχής $(r_m = 1, r_n = 2)$ υπολογίζονται από την εξίσωση [6.16], τα 39 δείγματα της περιοχής $(r_m = 2, r_n = 1)$ από την [6.17] και τέλος τα 3 δείγματα της περιοχής $(r_m = 2, r_n = 2)$ από την [6.18].

$$\begin{aligned} net.c = {} & 0.57 lspri.serum + 1.51 vol.synth + 0.01 zav.rel + 0.44 vol.rel \\ & + 0.77 P01009 + 0.83 P00738 + 0.14 P01011 \\ & - 1.93 P18428 - 1.04 P02655 + 0.66 P08567 \\ & + 1.45 P01019 + 0.44 P02671 + 0.21 Q99467 - 2.07 \end{aligned} \quad [6.15]$$

$$\begin{aligned} net.c = {} & 0.43 P01009 + 0.16 P68871 + 1.23 P02749 - 0.12 P02774 \\ & + 1.88 P27169 + 7.95 P01019 - 8.00 \end{aligned} \quad [6.16]$$

$$\begin{aligned} net.c = {} & 0.23 class + 1.66 lspr.serum + 0.20 zav.synth + 0.81 zav.serum \\ & + 0.24 num.synth + 3.14 int.serum + 1.54 pdi.rel \\ & + 0.55 zp.serum - 0.03 zp.synth.mag - 0.37 AS.total \\ & - 0.65 P0C0L4 + 1.10 P02649 + 0.03 Q1462 - 0.09 P02743 \\ & + 0.97 P10720 - 0.10 P05546 - 0.41 P49908 \\ & + 0.66 Q03591 + 0.40 Q43866 + 1.13 P02749 \\ & + 2.98 P02654 - 1.44 P03952 - 0.87 P01011 \\ & + 0.28 P18428 + 0.25 P00736 + 0.65 P00748 \\ & - 0.26 P02774 + 1.34 P00751 + 0.15 P03950 \\ & - 1.01 P02790 + 0.70 P18065 - 0.60 P00450 \\ & + 0.45 P08567 + 0.54 P08709 + 0.57 P00451 \\ & - 0.36 P23528 - 8.09 \end{aligned} \quad [6.17]$$

$$net.c = 1.53 zp.synth.mag + 1.31 AS.total - 8.78 \quad [6.18]$$

Εξωτερική επικύρωση

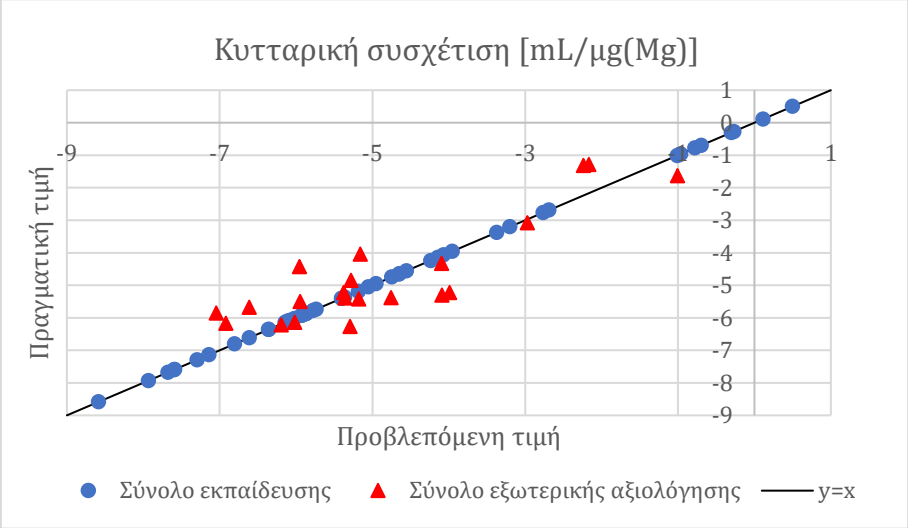
Στη συνέχεια, το αρχικό σύνολο δεδομένων χωρίζεται σε δύο φάσεις με τη μέθοδο Kennard and Stone, σε σύνολο μοντελοποίησης και σύνολο εξωτερικής αξιολόγησης με 63 και 21 δείγματα. Το πρώτο χωρίζεται επιπλέον σε σύνολο εκπαίδευσης με 50 δείγματα και σύνολο δοκιμών με 13 δείγματα. Για παράμετρο ομαλοποίησης $\lambda = 0.005$ προκύπτουν τα αποτελέσματα του ακόλουθου πίνακα (Πίνακας 6.20).

Πίνακας 6.20: Αποτελέσματα εξωτερικής επικύρωσης με διαδοχική επιλογή μεταβλητών διχοτόμησης για τα δεδομένα των νανοσωματιδίων χρυσού.

	Νανοσωματίδια χρυσού
z	0.2804
z test	1.2702
z external test	0.9370

R^2	1.0000	
R^2 ανά περιοχή	1.0000	1.0000
	1.0000	1.0000
Q^2 test	0.7977	
Q^2 external test	0.7345	
Περιοχές στη διάσταση *m*	2	
Περιοχές στη διάσταση *n*	2	
Φυσικοχημικές Μεταβλητές	14	
Βιολογικές Μεταβλητές	29	
Αριθμός δειγμάτων ανά περιοχή	4	3
	30	13
Αριθμός άγνωστων δειγμάτων ανά περιοχή	0	1
	9	11
Υπολογιστικός χρόνος (min)	9.96	

Στο Διάγραμμα 6.25 παρουσιάζεται η σχέση μεταξύ πραγματικών και προβλεπόμενων τιμών για το σύνολο εκπαίδευσης και για το σύνολο εξωτερικής αξιολόγησης.



Διάγραμμα 6.25: Πραγματικές και προβλεπόμενες τιμές κυτταρικής συσχέτισης των νανοσωματιδίων χρυσού με επίλυση διαδοχικής επιλογής μεταβλητών διχοτόμησης για τα σύνολα εκπαίδευσης και εξωτερικής αξιολόγησης.

Έλεγχος τυχαίας επιλογής

Για έλεγχο του παραγόμενου μοντέλου, πραγματοποιήθηκε έλεγχος τυχαίας επιλογής. Προκύπτει ο δείκτης Q^2_{test} για το σύνολο δοκιμών ίσος με $Q^2_{test} = -0.3391$ Από τα αποτελέσματα που προέκυψαν (Πίνακας 6.21) συμπεραίνεται ότι δεν υπάρχει πιθανότητα τυχαίας συσχέτισης των δεδομένων και των τιμών εξόδου στο μοντέλο.

Πίνακας 6.21: Αποτελέσματα ελέγχου τυχαίας επιλογής με διαδοχική επιλογή μεταβλητών διχοτόμησης για τα δεδομένα των νανοσωματιδίων χρυσού.

	Νανοσωματίδια χρυσού	
z	0.6715	
z test	2.0365	
R^2	0.9989	
R^2 ανά περιοχή	0.9983	1.0000
	1.0000	1.0000
Q^2 test	-0.3391	
Περιοχές στη διάσταση *m*	2	
Περιοχές στη διάσταση *n*	2	
Φυσικοχημικές Μεταβλητές	13	
Βιολογικές Μεταβλητές	37	
Αριθμός δειγμάτων ανά περιοχή	43	2
	15	3
Υπολογιστικός χρόνος (min)	10.74	

6.2.2.3 Επίλυση σε δύο διαστάσεις: Ανεξάρτητη επιλογή μεταβλητών διχοτόμησης

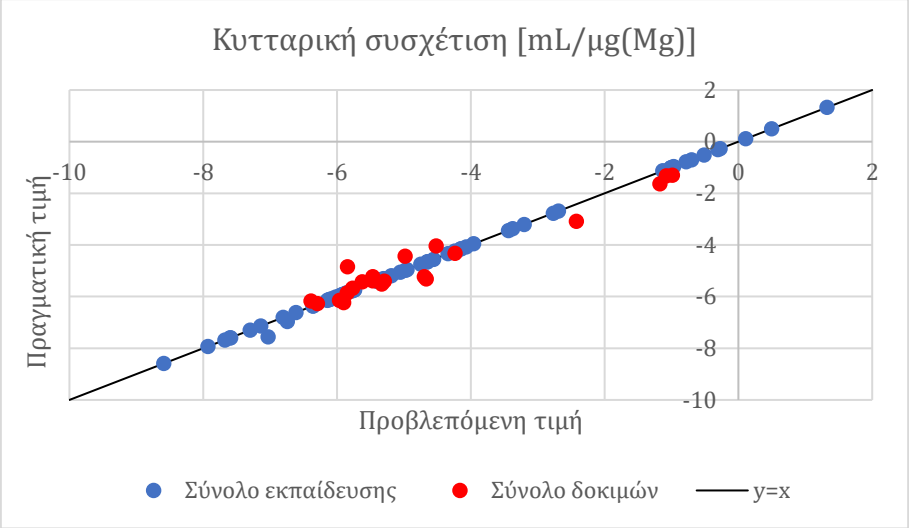
Για τιμή της παραμέτρου ομαλοποίησης $\lambda = 0.005$ και παράμετρο απόκλισης σφαλμάτων $\beta = 0.05$ προκύπτουν τα αποτελέσματα που παρουσιάζονται παρακάτω.

Πίνακας 6.22: Αποτελέσματα ανεξάρτητης επιλογής μεταβλητών διχοτόμησης για τα δεδομένα των νανοσωματιδίων χρυσού.

	Νανοσωματίδια χρυσού	
z	0.3112	
z test	0.6138	
R^2	0.9992	
R^2 ανά περιοχή	1.0000	1.0000
	0.9942	1.0000
Q^2 test	0.9339	
Περιοχές στη διάσταση *m*	2	
Περιοχές στη διάσταση *n*	2	
Φυσικοχημικές Μεταβλητές	16	
Βιολογικές Μεταβλητές	31	

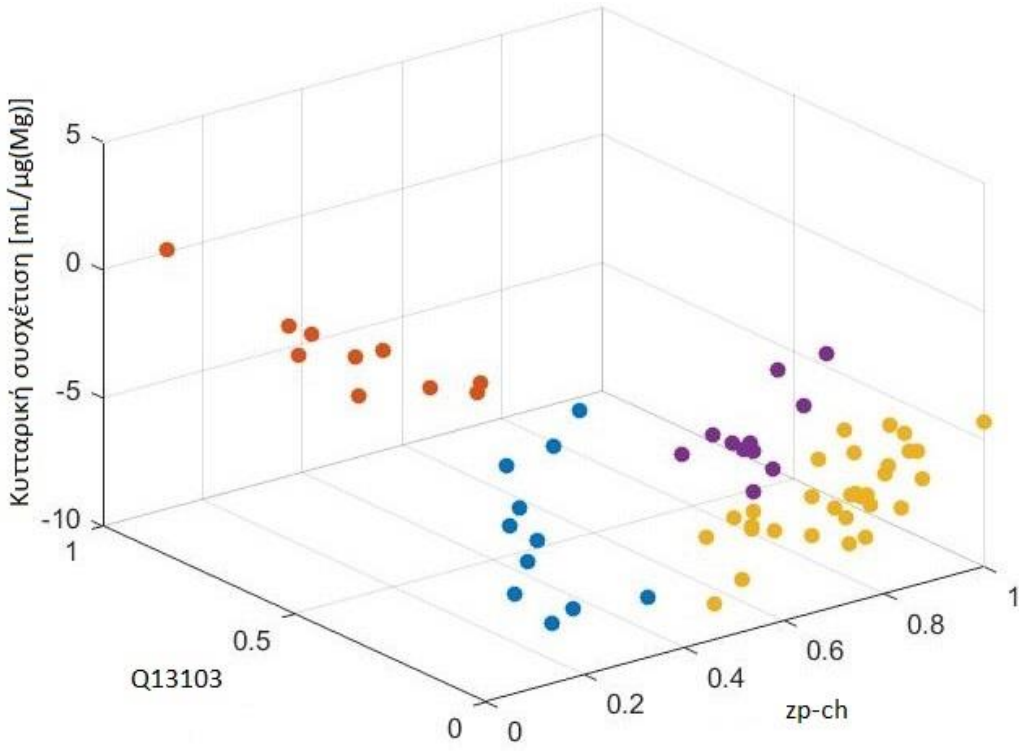
Αριθμός δειγμάτων ανά περιοχή	11 31	10 11	
Υπολογιστικός χρόνος (min)	5.55		

Κατά την ανεξάρτητη επιλογή μεταβλητών διχοτόμησης, η πρόβλεψη για το σύνολο δοκιμών γίνεται με μεγάλη ακρίβεια, όπως επιβεβαιώνεται και από τον δείκτη Q^2_{test} ο οποίος παίρνει τιμή που προσεγγίζει τη μονάδα ($Q^2_{test} = 0.9339$). Στο Διάγραμμα 6.26 παρουσιάζεται η σχέση μεταξύ πραγματικών και προβλεπόμενων τιμών κυτταρικής συσχέτισης για το σύνολο εκπαίδευσης και για το σύνολο δοκιμών.

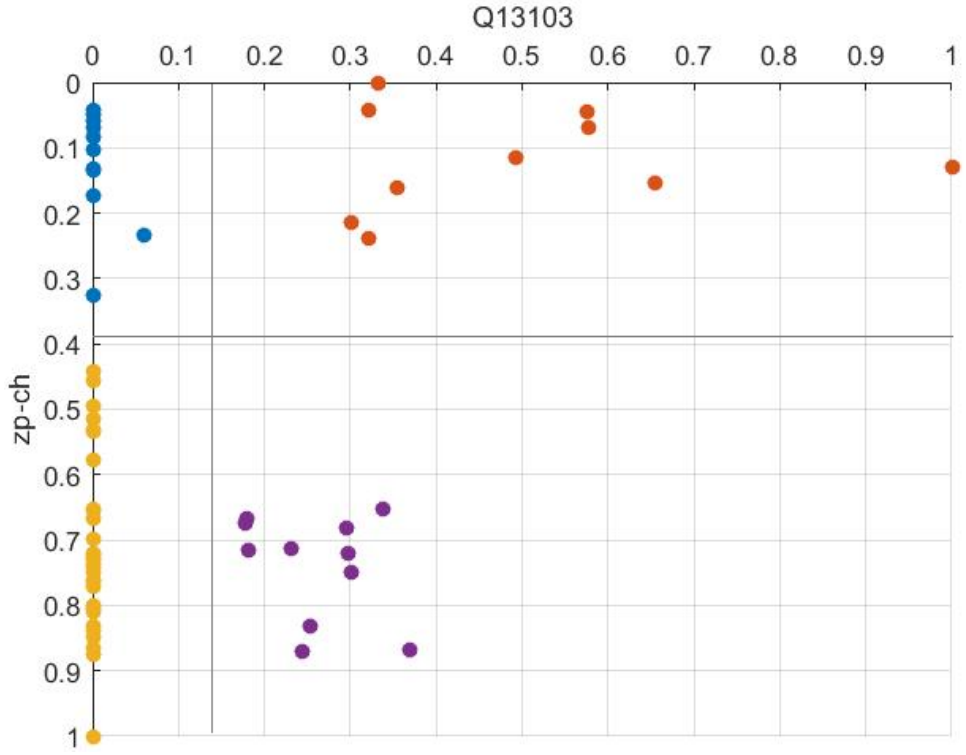


Διάγραμμα 6.26: Πραγματικές και προβλεπόμενες τιμές κυτταρικής συσχέτισης των νανοσωματιδίων χρυσού με επίλυση ανεξάρτητης επιλογής μεταβλητών διχοτόμησης για τα σύνολα εκπαίδευσης και δοκιμών.

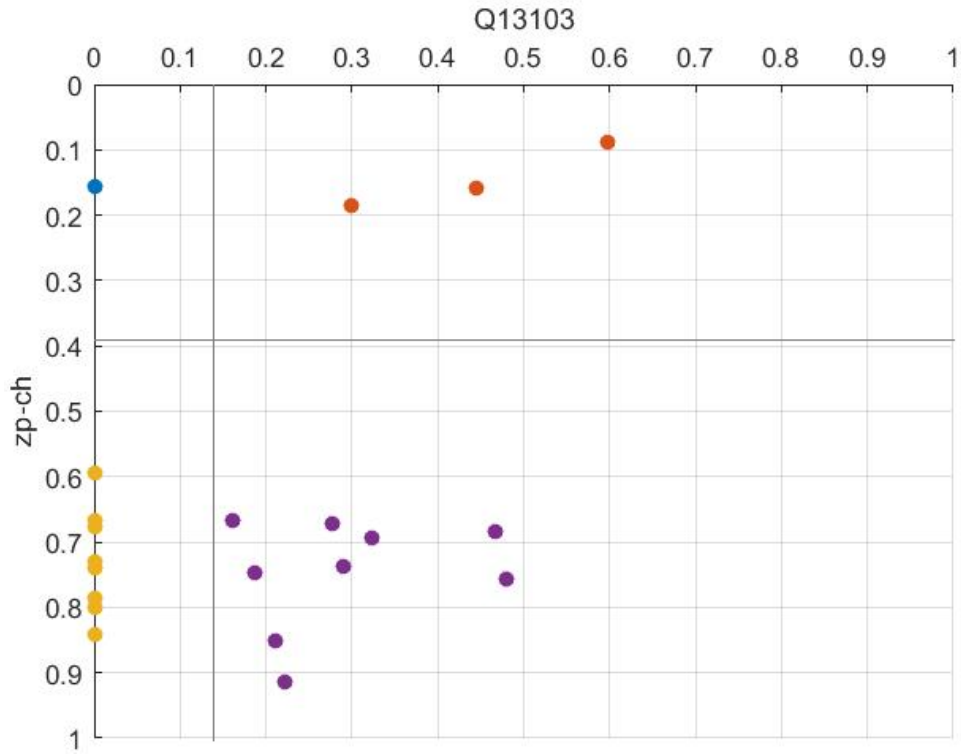
Επιλέγονται ως σημαντικές για την πρόβλεψη 16 φυσικοχημικές και 31 βιολογικές μεταβλητές. Τα δείγματα χωρίζονται σε 4 περιοχές, δύο σε κάθε διάσταση στις οποίες βρίσκονται 11, 10, 31 και 11 δείγματα αντίστοιχα. Τα δείγματα του συνόλου εκπαίδευσης κατανέμονται στο χώρο όπως φαίνεται στο Διάγραμμα 6.27. Στο Διάγραμμα 6.28 και το Διάγραμμα 6.29 παρουσιάζονται τα δείγματα εκπαίδευσης και δοκιμής στο χώρο και τα σημεία καμπής των περιοχών διαμέρισης. Ως μεταβλητή διχοτόμησης από τις φυσικοχημικές μεταβλητές επιλέχτηκε το ζ δυναμικό ορού-σύνθεσης ($zp - ch$) ενώ ως μεταβλητή διχοτόμησης των βιολογικών μεταβλητών η εκκρινόμενη φωσφοπρωτεΐνη 24 ($Q13103$).



Διάγραμμα 6.27: Κατανομή των δειγμάτων εκπαίδευσης στο χώρο για τα δεδομένα των νανοσωματιδίων χρυσού και με ανεξάρτητη επιλογή μεταβλητών διχοτόμησης.



Διάγραμμα 6.28: Κατανομή των δειγμάτων εκπαίδευσης και σημεία καμπής των περιοχών διαμέρισης για τα δεδομένα των νανοσωματιδίων χρυσού και με ανεξάρτητη επιλογή μεταβλητών διχοτόμησης.



Διάγραμμα 6.29: Κατανομή των δειγμάτων δοκιμών και σημεία καμπής των περιοχών διαμέρισης για τα δεδομένα των νανοσωματιδίων χρυσού και με ανεξάρτητη επιλογή μεταβλητών διχοτόμησης.

Τα 11 δείγματα εκπαίδευσης που κατατάσσονται στην περιοχή $(r_m = 1, r_n = 1)$ προβλέπονται από τη γραμμική εξίσωση [6.19] η οποία επιλέγει 10 μεταβλητές, τα 10 δείγματα της περιοχής $(r_m = 1, r_n = 2)$ υπολογίζονται από την εξίσωση [6.20], τα 31 δείγματα της περιοχής $(r_m = 2, r_n = 1)$ από την [6.21] και τέλος τα 11 δείγματα της περιοχής $(r_m = 2, r_n = 2)$ από την [6.22].

$$\begin{aligned} net.c = {} & -0.89 int.rel - 2.03 zp.serum + 0.27 P10909 + 2.47 P01009 \\ & - 4.05 P04114 + 1.64 P68871 - 1.83 P01011 \\ & + 1.86 P27169 + 1.28 P00450 + 5.84 Q99467 - 3.62 \end{aligned} \quad [6.19]$$

$$\begin{aligned} net.c = {} & 0.47 lspri.serum + 2.29 vol.synth + 0.34 vol.serum \\ & + 0.40 zav.rel + 0.20 P01023 - 0.07 O43866 + 0.44 P01011 \\ & + 0.59 P01019 + 0.20 Q99467 - 1.70 \end{aligned} \quad [6.20]$$

$$net.c = -0.79class + 0.57zav.synth + 2.22zav.serum + 0.12vol.synth + 0.76int.serum + 1.03pdi.rel - 2.21num.rel + 0.12zp.synth.mag + 0.35bca.density + 0.26AS.total + 0.86P02649 - 2.88P00739 + -0.16P12259 + 0.60Q03591 - 0.18O43866 + 1.48P02749 + 0.28P03951 + 3.00P02654 - 1.14P03952 + 1.27P00736 - 1.23P02774 + 1.18P03950 - 0.06P02790 - 0.33P02788 - 0.27P00450 + 0.16P00451 + 0.23P14618 - 0.57P23528 - 6.74 \quad [6.21]$$

$$net.c = 0.88lspri.serum + 1.64num.serum - 2.70AS.total + 0.87P10720 - 0.73Q03591 - 0.94P03952 - 0.25P00751 + 0.33P09871 + 0.62P20851 + 0.49P14618 - 4.53 \quad [6.22]$$

Εξωτερική επικύρωση

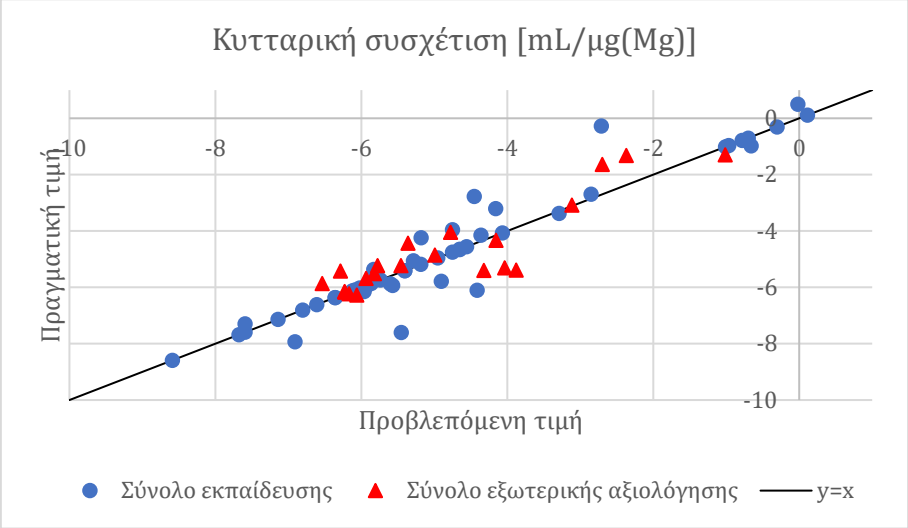
Στη συνέχεια, το σύνολο δεδομένων χωρίζεται σε σύνολο μοντελοποίησης και σύνολο εξωτερικής αξιολόγησης με 63 και 21 δείγματα. Το σύνολο μοντελοποίησης χωρίζεται επιπλέον σε σύνολο εκπαίδευσης με ποσοστό 80% (50 δείγματα) και σε σύνολο δοκιμών με ποσοστό 20% (13 δείγματα) . Για παράμετρο ομαλοποίησης $\lambda = 0.02$ προκύπτουν τα αποτελέσματα που παρουσιάζονται παρακάτω (Πίνακας 6.23).

Πίνακας 6.23: Αποτελέσματα εξωτερικής επικύρωσης με ανεξάρτητη επιλογή μεταβλητών διχοτόμησης για τα δεδομένα των νανοσωματιδίων χρυσού.

	Νανοσωματίδια χρυσού	
z	0.7286	
z test	1.1648	
z external test	0.9654	
R^2	0.9228	
R^2 ανά περιοχή	0.8458	0.8719
	0.4546	-
Q^2 test	0.8941	
Q^2 external test	0.7887	
Περιοχές στη διάσταση *m*	2	
Περιοχές στη διάσταση *n*	2	
Φυσικοχημικές Μεταβλητές	7	
Βιολογικές Μεταβλητές	17	
Αριθμός δειγμάτων ανά περιοχή	41	5
	3	1

Αριθμός άγνωστων δειγμάτων ανά περιοχή	20 0	1 0
Υπολογιστικός χρόνος (min)	3.87	

Στην περίπτωση που το μοντέλο εφαρμόζεται σε εντελώς άγνωστα δεδομένα, όπως αυτά του συνόλου εξωτερικής αξιολόγησης, η πρόβλεψη γίνεται αρκετά ικανοποιητικά όπως επαληθεύεται και από τον δείκτη Q^2_{ext} που παίρνει τιμή 0.7887. Η τιμή της αντικειμενικής συνάρτησης του συνόλου εξωτερικής αξιολόγησης δεν έχει μεγάλη διαφορά από την αντίστοιχη του συνόλου εκπαίδευσης και τέλος επιλέγονται 7 φυσικοχημικές και 17 βιολογικές μεταβλητές. Στο Διάγραμμα 6.30 παρουσιάζεται η σχέση μεταξύ πραγματικών και προβλεπόμενων τιμών κυτταρικής συσχέτισης για το σύνολο εκπαίδευσης και για το σύνολο εξωτερικής αξιολόγησης.



Διάγραμμα 6.30: Πραγματικές και προβλεπόμενες τιμές κυτταρικής συσχέτισης των νανοσωματιδίων χρυσού με επίλυση ανεξάρτητης επιλογής μεταβλητών διχοτόμησης για τα σύνολα εκπαίδευσης και εξωτερικής αξιολόγησης.

Έλεγχος τυχαίας επιλογής

Πραγματοποιήθηκε έλεγχος τυχαίας επιλογής και με την πρόβλεψη στα άγνωστα δεδομένα προέκυψε $z_{test} = 2.1060$, αρκετά μεγαλύτερη τιμή από την αντίστοιχη για το σύνολο εκπαίδευσης ($z = 0.6500$). Επιπλέον, ο δείκτης Q^2_{test} πήρε μη επιθυμητή τιμή, επομένως επαληθεύεται ότι με τις τυχαία ανακατεμένες τιμές εξόδου το μοντέλο που προκύπτει δεν μπορεί να προβλέψει άγνωστα δεδομένα.

Πίνακας 6.24: Αποτελέσματα ελέγχου τυχαίας επιλογής με ανεξάρτητη επιλογή μεταβλητών διχοτόμησης για τα δεδομένα των νανοσωματιδίων χρυσού.

	Νανοσωματίδια χρυσού
z	0.6500
z test	2.1060
R^2	0.9947

R^2 ανά περιοχή	0.9954	1.0000
	0.9913	1.0000
Q^2 test	-0.4931	
Περιοχές στη διάσταση *m*	2	
Περιοχές στη διάσταση *n*	2	
Φυσικοχημικές Μεταβλητές	11	
Βιολογικές Μεταβλητές	34	
Αριθμός δειγμάτων ανά περιοχή	37	8
	16	2
Υπολογιστικός χρόνος (min)	3.54	

Κεφάλαιο 7

Συμπεράσματα και προτάσεις για μελλοντική έρευνα

Τα νανοϋλικά χαρακτηρίζονται από αξιοσημείωτες και μοναδικές ιδιότητες που τα διαφοροποιούν από τα υλικά σε άλλες κλίμακες. Οι ιδιότητες αυτές έχουν αποδειχτεί χρήσιμες σε πολλές εφαρμογές και τα τελευταία χρόνια μονοπωλούν το ενδιαφέρον στην επιστήμη των υλικών. Έρευνες όμως σχετικά με την νανοτοξικότητα αναφέρουν ορισμένες δυσμενείς επιπτώσεις από τη χρήση των νανοϋλικών σε ζωντανούς οργανισμούς και στο περιβάλλον. Η συμβατική μελέτη και εκτίμηση των επιπτώσεων αυτών, απαιτεί τη διεξαγωγή δοκιμών και πειραμάτων σε ζώα. Ο εξελισσόμενος τομέας της Νανοπληροφορικής στοχεύει στην ανάπτυξη εναλλακτικών, υπολογιστικών μεθόδων και εργαλείων για την ανάλυση δεδομένων νανοτεχνολογίας, την πρόβλεψη ανεπιθύμητων ιδιοτήτων και την αξιολόγηση της επικινδυνότητας της χρήση νανοϋλικών.

Στην παρούσα Διπλωματική Εργασία αναπτύσσεται ένα μοντέλο μικτού ακέραιου γραμμικού προγραμματισμού για την πρόβλεψη δεικτών που σχετίζονται με την τοξικότητα των νανοσωματιδίων. Η ανάπτυξη των μοντέλων στηρίζεται στα βήματα της μεθοδολογίας read-across και στοχεύει στην αυτοματοποίηση της διαδικασίας ομαδοποίησης των νανοσωματιδίων και επιλογής των κατάλληλων μεταβλητών εισόδου.

Η διαδικασία που ακολουθείται περιλαμβάνει την ομαδοποίηση των νανοσωματιδίων σε περιοχές που ορίζονται από την βέλτιστα επιλεγμένη μεταβλητή διχοτόμησης και την πρόβλεψη της τοξικότητας από γραμμικές συναρτήσεις που διαμορφώνονται σε κάθε περιοχή. Έτσι επιλέγεται η «βέλτιστη» υπόθεση ομαδοποίησης η οποία προβλέπει με ακρίβεια τον δείκτη της τοξικότητας εντός της ομάδας νανοσωματιδίων που έχει δημιουργηθεί. Σε περίπτωση που είναι διαθέσιμα δεδομένα για δυο ή περισσότερες κατηγορίες ιδιοτήτων των νανοσωματιδίων, είναι δυνατόν να βρεθούν δυο ή περισσότερες ιδιότητες (μια από κάθε κατηγορία) που να διαχωρίζουν το πεδίο ορισμού (διαθέσιμα δείγματα) σε περιοχές και στις περιοχές αυτές κατά αντιστοιχία να εφαρμόζονται μοντέλα γραμμικής παλινδρόμησης.

Η επικύρωση του μοντέλου έγινε χρησιμοποιώντας τη μέθοδο εξωτερικής επικύρωσης. Υπολογίστηκε το πεδίο εφαρμογής του μοντέλου για να βρεθούν τα δείγματα για τα οποία οι προβλέψεις μπορούν να θεωρηθούν αξιόπιστες. Τα μοντέλα που αναπτύχθηκαν εξετάστηκαν ως προς την ικανότητα πρόβλεψής τους για δεδομένα που δεν συμμετείχαν στην διαδικασία εκπαίδευσης. Επιπλέον, πραγματοποιήθηκε ανάλυση ευαισθησίας ως προς την παράμετρο ομαλοποίησης λ, με βάση την οποία αποφεύχθηκε η υπερπροσαρμογή του μοντέλου στα δεδομένα εκπαίδευσης. Τέλος, πραγματοποιήθηκε έλεγχος τυχαίας επιλογής (y-randomization), για να αποκλειστεί το ενδεχόμενο τυχαίας συσχέτισης των δεδομένων εισόδου και της μεταβλητής απόκρισης. Για το μοντέλο μαθηματικού προγραμματισμού αναπτύχθηκε κώδικας στη γλώσσα προγραμματισμού MATLAB (όπως παρουσιάζεται στο Παράρτημα).

7.1 Ανάλυση αποτελεσμάτων και συμπεράσματα

Η μεθοδολογία OPLRA εφαρμόστηκε και αξιολογήθηκε σε τέσσερα σύνολα δεδομένων από τις δημοσιευμένες εργασίες «*Towards understanding mechanisms governing cytotoxicity of metal oxides nanoparticles: Hints from nano-QSAR studies*»[46], *«Protein Corona Fingerprinting Predicts the Cell Association of Gold and Silver Nanoparticles»*[6], «*Mapping the surface adsorption forces of nanomaterials in biological systems*»[47] και «*Quantitative Nanostructrure-Activity Relationship (QNAR) Modeling*»[48]. Σε όλες τις περιπτώσεις, ο αλγόριθμος αξιολογήθηκε σε δεδομένα που δεν συμπεριλήφθηκαν στη διαδικασία εκπαίδευσης. Σε όλα τα προβλήματα τα αποτελέσματα αξιολόγησης ήταν συγκρίσιμα με τα καλύτερα αποτελέσματα της βιβλιογραφίας. Στο πιο σύνθετο πρόβλημα των νανοσωματιδίων χρυσού, ο δείκτης Q^2_{test} που προέκυψε από την προτεινόμενη μεθοδολογία έφτασε την τιμή 0.93 που είναι η υψηλότερη τιμή του δείκτη που έχει αναφερθεί στη βιβλιογραφία για το συγκεκριμένο πρόβλημα. Εκτός των αποτελεσμάτων που καταδεικνύουν την προβλεπτική ικανότητα των μοντέλων που αναπτύσσονται με την προτεινόμενη μέθοδο, θα πρέπει να τονιστεί ότι η μέθοδος είναι αυτοματοποιημένη και δεν απαιτεί την χρονοβόρα μέθοδο δοκιμής και σφάλματος που απαιτείται από άλλες μεθόδους για τη διαμέριση του χώρου των μεταβλητών εισόδου και την επιλογή των μεταβλητών.

Για τα σύνολα δεδομένων που διέθεταν διαφορετικά είδη μεταβλητών, έγινε επίλυση του αλγορίθμου OPLRA σε δύο διαστάσεις, μία για κάθε κατηγορία μεταβλητών. Σε αυτήν την περίπτωση, ο αλγόριθμος αναπτύχθηκε με τρεις διαφορετικούς τρόπους, με επιλογή των μεταβλητών διχοτόμησης της κάθε διάστασης ταυτόχρονα, με διαδοχική επιλογή των μεταβλητών και με ανεξάρτητη επιλογή μεταβλητών διχοτόμησης για κάθε διάσταση. Τα αποτελέσματα έδειξαν την υπεροχή της ανεξάρτητης επιλογής. Για το σύνολο δεδομένων «Μεταλλικά οξείδια» η ταυτόχρονη και η διαδοχική επιλογή μεταβλητών δεν οδήγησε σε μοντέλο ικανό να προβλέψει δείγματα δοκιμών. Ωστόσο, με την ανεξάρτητη επιλογή μεταβλητών διχοτόμησης προέκυψε μοντέλο που προέβλεψε τα δείγματα ελέγχου αρκετά ικανοποιητικά, φτάνοντας τον δείκτη εξωτερικής ερμηνεύσιμης διακύμανσης Q^2_{test} ίσο με 0.80 για τα δείγματα του συνόλου δοκιμών. Για το σύνολο «Νανοσωματίδια χρυσού» και οι τρεις εναλλακτικές επίλυσης έδωσαν μοντέλα με υψηλή ικανότητα πρόβλεψης του δείκτη τοξικότητας. Η ταυτόχρονη και η διαδοχική επίλυση οδήγησαν σε Q^2_{test} της τάξης του 0.88, ενώ με την ανεξάρτητη επίλυση που ο δείκτης Q^2_{test} έφτασε την τιμή 0.93 και την επιλογή ίδιου αριθμού μεταβλητών.

Σε όλα τα μοντέλα που αναπτύχθηκαν πραγματοποιήθηκε έλεγχος τυχαίας επιλογής για έλεγχο της ισχύος τους (y-randomization). Οι συντελεστές συσχέτισης των μοντέλων με τυχαία κατανεμημένες τιμές εξαρτημένης μεταβλητής ήταν χαμηλοί, επιβεβαιώνοντας ότι το μοντέλο που δημιουργήθηκε με τις πραγματικές τιμές μεταβλητών αναπτύχθηκε σωστά και δεν προέκυψε τυχαία η ακρίβεια των προβλέψεων. Όλα τα σύνολα δεδομένων και όλες οι μεθοδολογίες επίλυσης, τόσο σε μία διάσταση όσο και σε δύο διαστάσεις, ανταπεξήλθαν στον έλεγχο τυχαίας επιλογής οδηγώντας στο συμπέρασμα ότι τόσο τα δεδομένα όσο και η μεθοδολογία μοντελοποίησης μπορούν να θεωρηθούν αξιόπιστα.

7.2 Προτάσεις για μελλοντική έρευνα

Η μέθοδος που παρουσιάστηκε στην παρούσα διπλωματική εργασία καταλήγει στην ανάπτυξη γραμμικών συναρτήσεων που προβλέπουν το δείκτη τοξικότητας των νανοϋλικών σε κάθε περιοχή διαμέρισης. Προτείνεται η επέκταση της μεθοδολογίας με

τη δυνατότητα ανάπτυξης μη γραμμικών σχέσεων σε κάθε περιοχή. Αυτό μπορεί να συμβεί είτε με εφαρμογή μη γραμμικών μετασχηματισμών των μεταβλητών εισόδου είτε με την εφαρμογή μη γραμμικών αλγόριθμων μηχανικής μάθησης, όπως τα νευρωνικά δίκτυα. Σε αυτήν βέβαια την περίπτωση θα διαμορφωθεί πρόβλημα μη γραμμικού και ακέραιου προγραμματισμού (Mixed Integer Nonlinear Programming) που πιθανόν να μη μπορεί να επιλυθεί αποτελεσματικά από συμβατικούς αλγόριθμους και να απαιτηθεί η ανάπτυξη στοχαστικού αλγόριθμου, όπως για παράδειγμα ενός γενετικού αλγόριθμου. Τέλος, είναι σημαντική η εφαρμογή της μεθόδου και σε άλλα σύνολα δεδομένων, ώστε να διευρυνθεί το εύρος προβλημάτων και να αξιολογηθεί περαιτέρω η αποτελεσματικότητα, η αξιοπιστία και η αποδοχή του από την επιστημονική κοινότητα. Έτσι, η εφαρμογή του αλγορίθμου θα μπορέσει να συμβάλλει στην προσπάθεια για περιορισμό των πειραμάτων σε ζώα και το σχεδιασμό ασφαλών νανοδομών πριν φθάσουν στο στάδιο της παραγωγής τους (safety-by-design).

Παράρτημα

Κώδικας για επίλυση σε μία διάσταση

```
%Import files
data_file = 'dataset.csv';
data=readtable(data_file,'Delimiter',';','ReadRowNames',true);
data=data(:,:);
totalsamples=size(data,1);
data_sc=scaling(data);
data_sc(:,1) = table2array(data(:,1));

%external validation
samples=round(0.75*totalsamples);
test_samples=totalsamples-samples;
[training_samples,Test_samples]=kenstone(data_sc(:,2:end),samples);
training_data=data_sc(training_samples,:);
test_data=data_sc(Test_samples,:);
trainingsampleNames=sampleNames(training_samples,1);
exp_test=test_data(:,1);
Asf_test=test_data(:,2:end);
Asf=training_data(:,2:end);
exp=training_data(:,1);
features=size(Asf,2);

regions=1;
%parameters
U=10;
U2=abs(sum(exp));
l=0.005;
beta=0.05;
epsilon=0.01;
Z=0.5;

%simple linear regression for R=1 training set
simplelinear=fitlm(Asf,exp);
initialpredictions=predict(simplelinear,Asf);
MAE_in=sum(abs(initialpredictions-exp))/samples;
W=simplelinear.Coefficients.Estimate(2:end);
REG_in=sum(abs(W));
z_in=MAE_in+l*REG_in;
ERRORcurrent=z_in;
ERRORold=Inf;
ERRORtmp=Inf;
%simple linear regression for R=1 test set
initialpredictions_test=predict(simplelinear,Asf_test);
MAE_in_test=sum(abs(initialpredictions_test-exp_test))/test_samples;
z_in_test=MAE_in_test+l*REG_in;
ERRORcurrent_test=z_in_test;
ERRORold_test=Inf;
ERRORtmp_test=Inf;
regions=regions+1;
fbest=0;

%R21 training set
final1=[exp,initialpredictions];
R1=corrcoef(final1(:,2),final1(:,1));
R21=R1(1,2)^2;
%R21 test set
final1_test=[exp_test,initialpredictions_test];
R21_test=1-sum((final1_test(:,1)-
final1_test(:,2)).^2)/sum((final1_test(:,1)-mean(exp)).^2);

reliability=domain(Asf,Asf_test,Z);
```

```
for p_f=1:features
      p_f
   Frs=binvar(regions,samples);
   Br=sdpvar(regions,1);
   Pred=sdpvar(regions,samples);
   Xrf=sdpvar(regions,features);
   Es=sdpvar(samples,1);
   Ers=sdpvar(regions,samples);
   Wrf=sdpvar(regions,features);
   Wrfpos=sdpvar(regions,features);
   MAE=sdpvar(1);
   REG=sdpvar(1);

   constraints=[];
   constraints=[constraints,MAE==sum(Es)/samples];

   for re=1:regions
       for fe=1:features
           constraints=[constraints,Wrfpos(re,fe)>=Wrf(re,fe)];
           constraints=[constraints,Wrfpos(re,fe)>=-Wrf(re,fe)];
       end
   end

   constraints=[constraints, REG==sum(sum(Wrfpos,2))];

   for re=2:regions
       constraints=[constraints,Xrf(re,p_f)>=Xrf(re-1,p_f)+epsilon];
   end

   constraints=[constraints,Xrf(regions,p_f)==1];
   constraints=[constraints,Xrf(1,p_f)>=epsilon];

   for i=1:samples
       constraints=[constraints,sum(Frs(:,i))==1];
   end

   for re=2:regions
       for i=1:samples
        constraints=[constraints,Xrf(re-1,p_f)+epsilon-U*(1-   Frs(re,i))<=
         Asf(i,p_f)];
       end
   end

for re=1:regions-1
       for i=1:samples
        constraints=[constraints,Asf(i,p_f)<=Xrf(re,p_f)-epsilon+U*(1-
         Frs(re,i))];
       end
   end

   for re=1:regions
      for i=1:samples
           constraints=[constraints,Pred(re,i)==sum(Asf(i,:).*Wrf(re,:))+Br(re,
           1)]
      end
   end

   for re=1:regions
       for i=1:samples
         constraints=[constraints,Es(i,1)>=0];
         constraints=[constraints,Ers(re,i)<=U2*(Frs(re,i))];
         constraints=[constraints,Ers(re,i)>=exp(i,1)-Pred(re,i)-U2*(1-
          Frs(re,i))];
         constraints=[constraints,Ers(re,i)>=Pred(re,i)-exp(i,1)-U2*(1-
          Frs(re,i))];
         constraints=[constraints,Es(i,1)>=Ers(re,i)];
       end
```

```
   end

   z=MAE+l*REG;
   options=sdpsettings('solver','mosek');
   readAcross=optimize(constraints,z,options);

   if value(z)<ERRORtmp && value(z)~=0
       ERRORtmp=value(z);
       zopt=value(z);
       bestMAE=value(MAE);
       bestREG=value(REG);
       bestEs=value(Es);
       bestErs=value(Ers);
       fbest=p_f;
       bestFrs=value(Frs);
       bestPred=value(Pred);
       bestXrf=value(Xrf);
       bestWrf=value(Wrf);
       bestBr=value(Br);
   end
end

finalregions=regions;

z_test=OPLRA_1D_test(fbest,features,test_samples,exp_test,finalregions,
Asf_test,l,bestXrf,bestWrf,bestBr);

ERRORold=ERRORcurrent;
ERRORcurrent=ERRORtmp;
ERRORold_test=ERRORcurrent_test;
ERRORcurrent_test=z_test;

while ERRORcurrent_test<(1-beta)*ERRORold_test && ERRORcurrent_test~=0
       clear p_f, Frs, Xrf, Br, Pred, Wrf, Wrfpos, Es, Ers, MAE, REG

       regions=regions+1

   Frs=binvar(regions,samples);
   Br=sdpvar(regions,1);
   Pred=sdpvar(regions,samples);
   Xrf=sdpvar(regions,features);
   Es=sdpvar(samples,1);
   Ers=sdpvar(regions,samples);
   Wrf=sdpvar(regions,features);
   Wrfpos=sdpvar(regions,features);
   MAE=sdpvar(1);
   REG=sdpvar(1);

   constraints=[];
   constraints=[constraints,MAE==sum(Es)/samples];
   for re=1:regions
       for fe=1:features
           constraints=[constraints,Wrfpos(re,fe)>=Wrf(re,fe)];
           constraints=[constraints,Wrfpos(re,fe)>=-Wrf(re,fe)];
       end
   end

   constraints=[constraints, REG==sum(sum(Wrfpos,2))];

   for re=2:regions
       constraints=[constraints,Xrf(re,fbest)>=Xrf(re-1,fbest)+epsilon];
   end

   constraints=[constraints,Xrf(regions,fbest)==1];
   constraints=[constraints,Xrf(1,fbest)>=epsilon];

   for i=1:samples
       constraints=[constraints,sum(Frs(:,i))==1];
```

```
    end

    for re=2:regions
        for i=1:samples
       constraints=[constraints,Xrf(re-1,fbest)+epsilon-U*(1-Frs(re,i))<=
        Asf(i,fbest)];
        end
    end

    for re=1:regions-1
        for i=1:samples
       constraints=[constraints,Asf(i,fbest)<=Xrf(re,fbest)-epsilon+U*(1-
        Frs(re,i))];
        end
    end

    for re=1:regions
       for i=1:samples
            constraints=[constraints,Pred(re,i)==sum(Asf(i,:).*Wrf(re,:))+Br(re,1
            )]
       end
    end

    for re=1:regions
        for i=1:samples
       constraints=[constraints,Es(i,1)>=0];
       constraints=[constraints,Ers(re,i)<=U2*(Frs(re,i))];
       constraints=[constraints,Ers(re,i)>=exp(i,1)-Pred(re,i)-U2*(1-
        Frs(re,i))];
       constraints=[constraints,Ers(re,i)>=Pred(re,i)-exp(i,1)-U2*(1-
        Frs(re,i))];
       constraints=[constraints,Es(i,1)>=Ers(re,i)];
        end
    end

    z=MAE+l*REG;
    options=sdpsettings('solver','mosek');
    readAcross=optimize(constraints,z,options);


    z_test=OPLRA_1D_test(fbest,features,test_samples,exp_test,regions,Asf_test
    ,l,value(Xrf),value(Wrf),value(Br));

    ERRORold=ERRORcurrent;
    ERRORcurrent=value(z);
    ERRORold_test=ERRORcurrent_test;
    ERRORcurrent_test=z_test;


    if ERRORcurrent_test<(1-beta)*ERRORold_test && ERRORcurrent_test~=0
        bestFrs=value(Frs);
        bestPred=value(Pred);
        bestXrf=value(Xrf);
        bestWrf=value(Wrf);
        bestBr=value(Br);
        zopt=value(z);
        bestMAE=value(MAE);
        bestREG=value(REG);
        bestEs=value(Es);
        bestErs=value(Ers);
        finalregions=regions;
    end
end

Ps=value(bestFrs.*bestPred);
Ps=Ps';
Ps1=sum(Ps,2);
```

```
%R2 training
final=[exp,Ps1];
R=corrcoef(final(:,2),final(:,1));
R2=R(1,2)^2;

for re=1:finalregions
    if sum(bestFrs(re,:))<=1
        R2_re(re)=NaN;
    else
    clear final2, joint, exp_reduced
    exp_reduced=exp(:,1).*(bestFrs(re,:)');
    joint=[exp_reduced,Ps(:,re)];
    rowfinal2=1;
    for i=1:samples
    if bestFrs(re,i)==1
       final2(rowfinal2,:)=joint(i,:);
       rowfinal2=rowfinal2+1;
    end
    end
    R_re=corrcoef(final2(:,2),final2(:,1));
    R2_re(re)=R_re(1,2)^2;
    end
end

Frs_test=zeros(finalregions,test_samples);

for i=1:test_samples
    for re=2:finalregions-1
        if Asf_test(i,fbest)>=bestXrf(re-1,fbest) &&
            Asf_test(i,fbest)<=bestXrf(re,fbest)
            Frs_test(re,i)=1;
        end
    end
end

for i=1:test_samples
    re=1;
    if bestXrf(re,fbest )>Asf_test(i,fbest)
        Frs_test(re,i)=1;
    end
end

for i=1:test_samples
    re=finalregions;
    if bestXrf(re-1,fbest )<Asf_test(i,fbest) &&
    Asf_test(i,fbest)<=bestXrf(re,fbest)
        Frs_test(re,i)=1;
    end
end

for re=1:finalregions
   for i=1:test_samples
        Pred_test(re,i)=sum(Asf_test(i,:).*bestWrf(re,:))+bestBr(re,1);
   end
end

Ps_test=value(Frs_test.*Pred_test);
Ps_test=Ps_test';
Ps1_test=sum(Ps_test,2);

for i=1:test_samples
  E(i,1)=abs(exp_test(i,1)-Ps1_test(i,1));
end

MAE_test=sum(E)/test_samples;

for re=1:finalregions
    for fe=1:features
```

```
        Wpos_t(re,fe)=abs(bestWrf(re,fe));
    end
end

REG_test=sum(sum(Wpos_t,2));
z_test=MAE_test+l*REG_test;

%R2 test
final_test=[exp_test,Ps1_test];
R2_test=1-sum((final_test(:,1)-final_test(:,2)).^2)/sum((final_test(:,1)-
mean(exp)).^2);
```

OPLRA_1D_test

```
function
z_test=OPLRA_1D_test(fbest,features,test_samples,exp_test,finalregions,Asf_t
est,l,bestXrf,bestWrf,bestBr)

Frs_test=zeros(finalregions,test_samples);

for i=1:test_samples
    for re=2:finalregions-1
        if Asf_test(i,fbest)>=bestXrf(re-1,fbest) &&
Asf_test(i,fbest)<=bestXrf(re,fbest)
            Frs_test(re,i)=1;
        end
    end
end

for i=1:test_samples
    re=1 ;
    if bestXrf(re,fbest )>Asf_test(i,fbest)
        Frs_test(re,i)=1;
    end
end

for i=1:test_samples
    re=finalregions;
    if bestXrf(re-1,fbest )<Asf_test(i,fbest) &&
Asf_test(i,fbest)<=bestXrf(re,fbest)
        Frs_test(re,i)=1;
    end
end


for re=1:finalregions
   for i=1:test_samples
        Pred_test(re,i)=sum(Asf_test(i,:).*bestWrf(re,:))+bestBr(re,1);
   end
end

Ps_test=value(Frs_test.*Pred_test);
Ps_test=Ps_test';
Ps1_test=sum(Ps_test,2);

for i=1:test_samples
  E(i,1)=abs(exp_test(i,1)-Ps1_test(i,1));
end

MAE_test=sum(E)/test_samples;

for re=1:finalregions
    for fe=1:features
        Wpos_t(re,fe)=abs(bestWrf(re,fe));
    end
end
```

```
REG_test=sum(sum(Wpos_t,2));
z_test=MAE_test+l*REG_test;
```

Domain of Applicability

```
function [reliability]=domain(Asf,Asf_test,Z)

test_samples=size(Asf_test,1);
training_samples=size(Asf,1);
reliability=zeros(test_samples,1);

for i=1:training_samples
    for j=1:training_samples
        D(i,j)=norm(Asf(j,:)-Asf(i,:),2);
    end
end

M=sum(sum(D,1),2)/(size(D,1)*size(D,2));

A=zeros(training_samples,training_samples);
N1=0;
for i=1:training_samples
    for j=1:training_samples
        if D(i,j)<=M
            A(i,j)=D(i,j);
        else
            A(i,j)=NaN;
        end
        if isnan(A(i,j))==0
            N1=N1+1;
        end
    end
end

A=A(:)
A=A(any(isnan(A),2)==0,:)

d=sum(sum(A,1,'omitnan'),2,'omitnan')/N1;

s=std2(A)

APD=d+Z*s

for i=1:test_samples
    for j=1:training_samples
        dist(i,j)=norm(Asf(j,:)-Asf_test(i,:),2);
    end
end

neighbors=zeros(test_samples,1);
for i=1:test_samples
    neighbors(i,1)=min(dist(i,:));
end

neighbors(:,1)

for i=1:test_samples
    if APD>neighbors(i,1)
        reliability(i,1)=1;
    else
        reliability(i,1)=0;
    end
end
```

Κώδικας για επίλυση σε δύο διαστάσεις

```
%Import files
data_file = 'dataset.csv';
data=readtable(data_file,'Delimiter',';','ReadRowNames',true);
data=data(:,:);
totalsamples=size(data,1);
data_sc=scaling(data);
data_sc(:,1) = table2array(data(:,1));

%external validation
samples=round(0.75*totalsamples);
test_samples=totalsamples-samples;
[training_samples,Test_samples]=kenstone(data_sc(:,2:end),samples);
training_data=datasample(data_sc,samples,'Replace',false);
training_data=data_sc(training_samples,:);
test_data=data_sc(Test_samples,:);
exp_test=test_data(:,1);
Asf1_test=test_data(:,2:19);
Asf2_test=test_data(:,20:end);
Asf_test=test_data(:,2:end);
Asf1=training_data(:,2:19);
Asf2=training_data(:,20:end);
Asf=training_data(:,2:end);
exp=training_data(:,1);
features1=size(Asf1,2);
features2=size(Asf2,2);
features=size(Asf,2);

regions1=2;
regions2=2;
U=10;
U2=abs(sum(exp));
l=0.02;
beta=0.05;
epsilon=0.05;
Z=0.5;

%simple linear regression for R=1 training
simplelinear=fitlm(Asf,exp);
initialpredictions=predict(simplelinear,Asf);
MAE_in=sum(abs(initialpredictions-exp))/samples;
W=simplelinear.Coefficients.Estimate(2:end);
REG_in=sum(abs(W));
z_in=MAE_in+l*REG_in;
ERRORcurrent=z_in;
ERRORold=Inf;
ERRORtmp=Inf;
%simple linear regression for R=1 test
initialpredictions_test=predict(simplelinear,Asf_test);
MAE_in_test=sum(abs(initialpredictions_test-exp_test))/test_samples;
z_in_test=MAE_in_test+l*REG_in_test;
ERRORcurrent_test=z_in_test;
ERRORold_test=Inf;
ERRORtmp_test=Inf;

fbest1=0;
fbest2=0;

%R2 training
final1=[exp,initialpredictions];
R1=corrcoef(final1(:,2),final1(:,1));
R21=R1(1,2)^2;
%R21 test
final1_test=[exp_test,initialpredictions_test];
R21_test=1-sum((final1_test(:,1)-
final1_test(:,2)).^2)/sum((final1_test(:,1)-mean(exp)).^2);
```

```
reliability=domain(Asf,Asf_test,Z);

for p_f1=1:features1
    for p_f2=1:features2
        p_f1
        p_f2
     Fr1s=binvar(regions1,samples);
     Fr2s=binvar(regions2,samples);
     Br1r2=sdpvar(regions1,regions2,'full');
     Pred=sdpvar(regions1,regions2,samples,'full');
     Xrf1=sdpvar(regions1,features1);
     Xrf2=sdpvar(regions2,features2);
     Ers1=sdpvar(regions1,samples);
     Ers2=sdpvar(regions2,samples);
     Es=sdpvar(samples,1);
     Wrf=sdpvar(regions1,regions2,features,'full');
     Wrfpos=sdpvar(regions1,regions2,features,'full');
     MAE=sdpvar(1);
     REG=sdpvar(1);

     constraints=[];
     constraints=[constraints,MAE==sum(Es)/samples];

     for re1=1:regions1
         for re2=1:regions2
             for fe=1:features
     constraints=[constraints,Wrfpos(re1,re2,fe)>=Wrf(re1,re2,fe)];
     constraints=[constraints,Wrfpos(re1,re2,fe)>=-Wrf(re1,re2,fe)];
             end
         end
     end

     constraints=[constraints, REG==sum(sum(sum(Wrfpos)))];

     for re=2:regions1
     constraints=[constraints,Xrf1(re,p_f1)>=Xrf1(re-1,p_f1)+epsilon];
     end

     constraints=[constraints,Xrf1(regions1,p_f1)==1];
     constraints=[constraints,Xrf1(1,p_f1)>=epsilon];

     for re=2:regions2
     constraints=[constraints,Xrf2(re,p_f2)>=Xrf2(re-1,p_f2)+epsilon];
     end

     constraints=[constraints,Xrf2(regions2,p_f2)==1];
     constraints=[constraints,Xrf2(1,p_f2)>=epsilon];

     for i=1:samples
         constraints=[constraints,sum(Fr1s(:,i))==1];
     end

     for i=1:samples
         constraints=[constraints,sum(Fr2s(:,i))==1];
     end

     for re=2:regions1
         for i=1:samples
    constraints=[constraints,Xrf1(re-1,p_f1)+epsilon-U*(1-Fr1s(re,i))<=
     Asf1(i,p_f1)];
         end
     end

     for re=1:regions1-1
         for i=1:samples
```

```
constraints=[constraints,Asf1(i,p_f1)<=Xrf1(re,p_f1)-epsilon+U*(1-
 Fr1s(re,i))];
     end
 end

 for re=2:regions2
     for i=1:samples
constraints=[constraints,Xrf2(re-1,p_f2)+epsilon-U*(1-Fr2s(re,i))<=
 Asf2(i,p_f2)];
     end
 end

 for re=1:regions2-1
     for i=1:samples
constraints=[constraints,Asf2(i,p_f2)<=Xrf2(re,p_f2)-epsilon+U*(1-
 Fr2s(re,i))];
     end
 end


 for re1=1:regions1
     for re2=1:regions2
            for i=1:samples
             constraints=[constraints,Pred(re1,re2,i)==sum(Asf(i,:).*Wrf(re
             1,re2,:))+Br1r2(re1,re2)];
         end
     end
 end

 for re1=1:regions1
    for re2=1:regions2
         for i=1:samples
constraints=[constraints,Es(i,1)>=0];
constraints=[constraints,Ers1(re1,i)<=U2*(Fr1s(re1,i))];
constraints=[constraints,Ers2(re2,i)<=U2*(Fr2s(re2,i))];
constraints=[constraints,Ers1(re1,i)>=exp(i,1)-Pred(re1,re2,i)-U2*(1-
 Fr1s(re1,i))-U2*(1-Fr2s(re2,i))];
constraints=[constraints,Ers1(re1,i)>=Pred(re1,re2,i)-exp(i,1)-U2*(1-
 Fr1s(re1,i))-U2*(1-Fr2s(re2,i))];
constraints=[constraints,Ers2(re2,i)>=exp(i,1)-Pred(re1,re2,i)-U2*(1-
 Fr1s(re1,i))-U2*(1-Fr2s(re2,i))];
constraints=[constraints,Ers2(re2,i)>=Pred(re1,re2,i)-exp(i,1)-U2*(1-
 Fr1s(re1,i))-U2*(1-Fr2s(re2,i))];
constraints=[constraints,Es(i,1)>=Ers1(re1,i)];
constraints=[constraints,Es(i,1)>=Ers2(re2,i)];
         end
     end
 end

 z=MAE+l*REG;
 options=sdpsettings('solver','mosek');
 readAcross=optimize(constraints,z,options);


 if value(z)<ERRORtmp && value(z)~=0
     ERRORtmp=value(z);
     zopt=value(z);
     bestMAE=value(MAE);
     bestREG=value(REG);
     bestEs=value(Es);
     fbest1=p_f1;
     fbest2=p_f2;
     bestFr1s=value(Fr1s);
     bestFr2s=value(Fr2s);
     bestPred=value(Pred);
     bestXrf1=value(Xrf1);
     bestXrf2=value(Xrf2);
     bestWrf=value(Wrf);
```

```
            bestBr1r2=value(Br1r2);
        end
    end
    end

    finalregions1=regions1;
    finalregions2=regions2;

    [z_test,
    E_test]=OPLRA_2D_test_final(fbest1,fbest2,test_samples,exp_test,finalregions
    1,finalregions2,Asf_test,Asf1_test,Asf2_test,l,bestXrf1,bestXrf2,bestWrf,bes
    tBr1r2);

    ERRORold=ERRORcurrent;
    ERRORcurrent=ERRORtmp;

    ERRORold_test=ERRORcurrent_test;
    ERRORcurrent_test=z_test;


    bestF=zeros(regions1,regions2,samples);

    for i=1:samples
        for re1=1:regions1
            for re2=1:regions2
                if bestFr1s(re1,i)>=1 && bestFr2s(re2,i)>=1
                    bestF(re1,re2,i)=1;
                end
            end
        end
    end

    while ERRORcurrent_test<(1-beta)*ERRORold_test && ERRORcurrent_test~=0

    clear MAE, REG, Es, Ers1, Ers2, Pred, Wrf, Wrfpos, Fr1s, Fr2s, Xrf1, Xrf2

    r_a1=regions1+1;
    r_a2=regions2;
    [z_a,MAE_a,REG_a,Es_a,Fr1s_a,Fr2s_a,Br1r2_a,Wrf_a,Pred_a,Xrf1_a,Xrf2_a]=OPLRA_2D
    _final(fbest1,fbest2,exp,r_a1,r_a2,Asf,Asf1,Asf2,epsilon,U,U2,l)

    r_b1=regions1;
    r_b2=regions2+1;
    [z_b,MAE_b,REG_b,Es_b,Fr1s_b,Fr2s_b,Br1r2_b,Wrf_b,Pred_b,Xrf1_b,Xrf2_b]=OPLRA_2D
    _final(fbest1,fbest2,exp,r_b1,r_b2,Asf,Asf1,Asf2,epsilon,U,U2,l)

    r_c1=regions1+1;
    r_c2=regions2+1;
    [z_c,MAE_c,REG_c,Es_c,Fr1s_c,Fr2s_c,Br1r2_c,Wrf_c,Pred_c,Xrf1_c,Xrf2_c]=OPLRA_2D
    _final(fbest1,fbest2,exp,r_c1,r_c2,Asf,Asf1,Asf2,epsilon,U,U2,l)

    minz=Inf;
    if z_a<minz && z_a~=0
        minz=z_a;
        regions1=r_a1;
        regions2=r_a2;
        z=z_a;
        MAE=MAE_a;
        REG=REG_a;
        Es=Es_a;
        Fr1s=Fr1s_a;
        Fr2s=Fr2s_a;
        Pred=Pred_a;
        Xrf1=Xrf1_a;
        Xrf2=Xrf2_a;
        Wrf=Wrf_a;
        Br1r2=Br1r2_a;
    end
```

```
if z_b<minz && z_b~=0
    minz=z_b;
    regions1=r_b1;
    regions2=r_b2;
    z=z_b;
    MAE=MAE_b;
    REG=REG_b;
    Es=Es_b;
    Fr1s=Fr1s_b;
    Fr2s=Fr2s_b;
    Pred=Pred_b;
    Xrf1=Xrf1_b;
    Xrf2=Xrf2_b;
    Wrf=Wrf_b;
    Br1r2=Br1r2_b;
end

if z_c<minz && z_c~=0
    minz=z_c;
    regions1=r_c1;
    regions2=r_c2;
    z=z_c;
    MAE=MAE_c;
    REG=REG_c;
    Es=Es_c;
    Fr1s=Fr1s_c;
    Fr2s=Fr2s_c;
    Pred=Pred_c;
    Xrf1=Xrf1_c;
    Xrf2=Xrf2_c;
    Wrf=Wrf_c;
    Br1r2=Br1r2_c;
end

ERRORold=ERRORcurrent;
ERRORcurrent=minz;

F=zeros(regions1,regions2,samples);

for i=1:samples
    for re1=1:regions1
        for re2=1:regions2
            if Fr1s(re1,i)>=1 && Fr2s(re2,i)>=1
                F(re1,re2,i)=1;
            end
        end
    end
end

[z_test,
E_test]=OPLRA_2D_test_final(fbest1,fbest2,test_samples,exp_test,regions1,reg
ions2,Asf_test,Asf1_test,Asf2_test,l,Xrf1,Xrf2,Wrf,Br1r2);

ERRORold_test=ERRORcurrent_test;
ERRORcurrent_test=z_test;

if ERRORcurrent_test<(1-beta)*ERRORold_test
    clear bestF, zopt
    bestF=F;
    bestPred=Pred;
    zopt=z;
    bestMAE=MAE;
    bestREG=REG;
    bestEs=Es;
    bestXrf1=Xrf1;
    bestXrf2=Xrf2;
    bestWrf=Wrf;
```

```
    bestBr1r2=Br1r2;
    finalregions1=regions1;
    finalregions2=regions2;
end

end

Ps=value(bestF.*bestPred);
A=sum(sum(Ps,1));
Ps1(:,1)=A(1,1,:);

%R2 training
final=[exp,Ps1];
R=corrcoef(final(:,2),final(:,1));
R2=R(1,2)^2;

samplestoregions=zeros(samples,1);
for re1=1:finalregions1
    for re2=1:finalregions2
    if sum(bestF(re1,re2,:),3)<=1
        R2_re(re1,re2)=NaN;
    else
    clear final2, joint, exp_reduced
    samplestoregions(:,1)=bestF(re1,re2,:);
    exp_reduced=exp(:,1).*samplestoregions(:,1);
    Ps2(:,1)=Ps(re1,re2,:);
    joint=[exp_reduced,Ps2(:,1)];
    rowfinal2=1;
    for i=1:samples
    if bestF(re1,re2,i)==1
       final2(rowfinal2,:)=joint(i,:);
       rowfinal2=rowfinal2+1;
    end
    end
    R_re=corrcoef(final2(:,2),final2(:,1));
    R2_re(re1,re2)=R_re(1,2)^2;
    end
    end
end

F_test=zeros(finalregions1,finalregions2,test_samples);
Fr1s_test=zeros(finalregions1,test_samples);
Fr2s_test=zeros(finalregions2,test_samples);

for i=1:test_samples
    for re1=2:finalregions1-1
        if Asf1_test(i,fbest1)>=bestXrf1(re1-1,fbest1) &&
          Asf1_test(i,fbest1)<=bestXrf1(re1,fbest1)
            Fr1s_test(re1,i)=1;
        end
    end
end

for i=1:test_samples
    for re2=2:finalregions2-1
        if Asf2_test(i,fbest2)>=bestXrf2(re2-1,fbest2) &&
          Asf2_test(i,fbest2)<=bestXrf2(re2,fbest2)
            Fr2s_test(re2,i)=1;
        end
    end
end

for i=1:test_samples
    re=1;
    if bestXrf1(re,fbest1)>Asf1_test(i,fbest1)
        Fr1s_test(re,i)=1;
    end
end
```

```
for i=1:test_samples
    re=1;
    if bestXrf2(re,fbest2)>Asf2_test(i,fbest2)
        Fr2s_test(re,i)=1;
    end
end

for i=1:test_samples
    re=finalregions1;
    if bestXrf1(re-1,fbest1)<Asf1_test(i,fbest1) &&
    Asf1_test(i,fbest1)<=bestXrf1(re,fbest1)
        Fr1s_test(re,i)=1;
    end
end

for i=1:test_samples
    re=finalregions2;
    if bestXrf2(re-1,fbest2 )<Asf2_test(i,fbest2) &&
    Asf2_test(i,fbest2)<=bestXrf2(re,fbest2)
        Fr2s_test(re,i)=1;
    end
end

for re1=1:finalregions1
    for re2=1:finalregions2
      bestW(1,:)=bestWrf(re1,re2,:);
     for i=1:test_samples
Pred_test(re1,re2,i)=sum(Asf_test(i,:).*bestW(1,:))+bestBr1r2(re1,re2);
     end
    end
end

for i=1:test_samples
    for re1=1:finalregions1
        for re2=1:finalregions2
          if Fr1s_test(re1,i)>=1 && Fr2s_test(re2,i)>=1
               F_test(re1,re2,i)=1;
          end
       end
    end
end

Ps_test=value(F_test.*Pred_test);
B=sum(sum(Ps_test,1));
Ps1_test(:,1)=B(1,1,:);

for i=1:test_samples
  E(i,1)=abs(exp_test(i,1)-Ps1_test(i,1));
end

MAE_test=sum(E)/test_samples;

for re1=1:finalregions1
    for re2=1:finalregions2
        for fe=1:features
            Wpos_t(re1,re2,fe)=abs(bestWrf(re1,re2,fe));
        end
    end
end

REG_test=sum(sum(sum(Wpos_t)));
z_test=MAE_test+l*REG_test;

%R2 test
final_test=[exp_test,Ps1_test];
R2_test=1-sum((final_test(:,1)-final_test(:,2)).^2)/sum((final_test(:,1)-
mean(exp)).^2);
```

OPLRA_2D_test_final

```
function
[z_test,E_test]=OPLRA_2D_test_final(fbest1,fbest2,test_samples,exp_test,
finalregions1,finalregions2,Asf_test,Asf1_test,Asf2_test,l,bestXrf1,best
Xrf2,bestWrf,bestBr1r2)

features=size(Asf_test,2);

F_test=zeros(finalregions1,finalregions2,test_samples);

for i=1:test_samples
    for re1=2:finalregions1-1
        for re2=2:finalregions2-1
         if Asf1_test(i,fbest1)>=bestXrf1(re1-1,fbest1) &&
Asf1_test(i,fbest1)<=bestXrf1(re1,fbest1) &&
Asf2_test(i,fbest2)>=bestXrf2(re2-1,fbest2) &&
Asf2_test(i,fbest2)<=bestXrf2(re2,fbest2)
            F_test(re1,re2,i)=1;
        end
    end
    end
end

for i=1:test_samples
    re1=1;
    re2=1;
    if bestXrf1(re1,fbest1 )>Asf1_test(i,fbest1) &&
bestXrf2(re2,fbest2 )>Asf2_test(i,fbest2)
        F_test(re1,re2,i)=1;
    end
end

for i=1:test_samples
    re1=finalregions1;
    re2=finalregions2;
    if bestXrf1(re1-1,fbest1 )<Asf1_test(i,fbest1) &&
Asf1_test(i,fbest1)<=bestXrf1(re1,fbest1) && bestXrf2(re2-1,fbest2 )<Asf2_test(i,fbest2)
&& Asf2_test(i,fbest2)<=bestXrf2(re2,fbest2)
        F_test(re1,re2,i)=1;
    end
end

for re1=1:finalregions1
    for re2=1:finalregions2
        bestW(1,:)=bestWrf(re1,re2,:);
     for i=1:test_samples
Pred_test(re1,re2,i)=sum(Asf_test(i,:).*bestW(1,:))+bestBr1r2(re1,re2);
     end
    end
end

Ps_test=value(F_test.*Pred_test);
B=sum(sum(Ps_test,1));
Ps1_test(:,1)=B(1,1,:);

for i=1:test_samples
  E_test(i,1)=abs(exp_test(i,1)-Ps1_test(i,1));
end

MAE_test=sum(E_test)/test_samples;

for re1=1:finalregions1
    for re2=1:finalregions2
        for fe=1:features
            Wpos_t(re1,re2,fe)=abs(bestWrf(re1,re2,fe));
        end
```

```
    end
end

REG_test=sum(sum(sum(Wpos_t)));
z_test=MAE_test+l*REG_test;
```

Βιβλιογραφία

[1] Bahadar H., Maqbool F., Niaz K. & Abdollahi M. (2016) "Toxicity of Nanoparticles and an Overview of Current Experimental Models", *Iranian Biomedical Journal* **20**, 1–11

[2] Khan I., Saeed K. & Khan I. (2017)"Nanoparticles: Properties, applications and toxicities", *Arabian Journal of Chemistry*

[3] Vanner R. & Buzea C. (2012) "Toxicity of nanoparticles", *Woodhead Publishing Limited* 427-475

[4] Makino H. (2018) "Environmental and Safety Issues With Nanoparticles", *Nanoparticle Technology Handbook* 365–395

[5] Rajan K. (2018) "Chapter 6-Nanoinformatics: Data-Driven Materials Design for Health and Environmental Needs", *Nanotechnology Environmental Health and Safety* 119–150

[6] Walkey C. D. *et al.* (2014) "Protein corona fingerprinting predicts the cellular interaction of gold and silver nanoparticles", *ACS Nano* **8**, 2439–2455

[7] Gunsolus I. L. & Haynes C. L. (2016) "Analytical Aspects of Nanotoxicology", *Analytical Chemistry* **88**, 451–479

[8] Saptarshi S.R., Duschl A. & Lopata A.L. (2013) "Interaction of nanoparticles with proteins: Relation to bio-reactivity of the nanoparticle", *Journal of Nanobiotechnology* **11**, 1–12

[9] Xia X. R., Monteiro-Riviere N. A. & Riviere J. E. (2010) "An index for characterization of nanomaterials in biological systems", *Nature Nanotechnology* **5**, 671–675

[10] Wolfram J., Yang Y., Shen J., Moten A. & Chen C. (2014) "The nano-plasma interface: Implications of the protein corona", *Colloids Surfaces B Biointerfaces*, **124**, 17-24

[11] Nierenberg D., Khaled A. R. & Flores O. (2018) "Formation of a protein corona influences the biological identity of nanomaterials", *Reports of Practical Oncology and Radiotherapy* **23**, 300–308

[12] Zanganeh S., Ho J. Q., Aieneravaie M. & Erfanzadeh M. (2018) "Protein Corona: The Challenge at the Nanobiointerfaces", *Iron Oxide Nanoparticles for Biomedical Applications* 91–104

[13] De La Iglesia D., Cachau R. E., García-Remesal M. & Maojo V., (2013) "Nanoinformatics knowledge infrastructures: Bringing efficient information management to nanomedical research" *Computational Science and Discovery* **6**, 014011

[14] National Nanomanufacturing Network. Nanoinformatics. Διαθέσιμο: http://nanoinformatics.org/nanoinformatics/index.php/Main_Page. (Πρόσβαση: 10 Σεπτεμβρίου 2019)

[15] Gajewicz A., Jagiello K., Cronin M.T.D., Leszczynski J. & Puzyn T. (2017) "Addressing a bottle neck for regulation of nanomaterials: quantitative read-across (Nano-QRA) algorithm for cases when only limited data is available" *Environmental Science Nano* **4**, 346–358

[16] Varsou D., Afantitis A. & Melagraki G. (2019) "Read-across predictions of nanoparticle hazard endpoints: a mathematical optimization approach" *Nanoscale Advances,* **1**, 3485-3498

[17] White A. & Cronin M.T.D. (2015) "A strategy for structuring and reporting a read-across prediction of toxicity" *Regulatory Toxicology and Pharmacology* **72**, 586–601

[18] ECHA "Read-Across Assessment Framework (RAAF) - Human health effects & Environmental fate and effects" (2015)

[19] Gajewicz A., Cronin M.T.D., Rasulev B., Leszczynski J. & Puzyn T. (2015) "Novel approach for efficient predictions properties of large pool of nanomaterials based on limited set of species: Nano-read-across", *Nanotechnology* **26**, 15701

[20] European Chemicals Agency (2017) "Guidance on information requirements and chemical safety assessment: Appendix R.6-1 for nanomaterials applicable to the Guidance on QSARs and Grouping of Chemicals" *Version 1.0* 1–29

[21] Πραστάκος Γ.(2006) "Διοικητική Επιστήμη, Λήψη Επιχειρησιακών αποφάσεων στην κοινωνία της πληροφορίας", *Εκδόσεις Σταμούλης*

[22] Bradley S. P., Hax A. C., Magnanti T. L. (1977) "Applied Mathematical Programming", *Addison-Wesley*

[23] Cormen T. H., Leiserson C. E., Rivest R. L. & Stein C. (2002) "Introduction to Algorithms", *The MIT Press*

[24] MathWorks, "Plot the feasible region of a linear program" Διαθέσιμο: https://www.mathworks.com/help/symbolic/mupad_ref/linopt-plot_data.html (Πρόσβαση: 10 Σεπτεμβρίου 2019)

[25] Griva I., Nash S. G. & Sofer A. (2011) "Linear and Nonlinear Optimization. Linear and Nonlinear Optimization", *Society for Industrial and Applied Mathematics*

[26] Businessmanagementcourses.org. "The Big M Method". Διαθέσιμο: http://businessmanagementcourses.org/Lesson09TheBigMMethod.pdf (Πρόσβαση: 10 Σεπτεμβρίου 2019)

[27] Golbraikh A., Shen M., Xiao Z., Xiao Y. & Lee K. (2003) "Rational selection of training and test sets for the development of validated QSAR models" *Journal of Computer-Aided Molecular Design*, **17**, 241–253

[28] Hand D., Mannila H., Smyth P., (2001) "Principles of Data Mining", *The MIT Press*

[29] Montgomery D. C., Peck E. A., Vining G. C. (2001) "Introduction to linear regression analysis", *John Wiley & Sons*

[30] Piecewise linear curve fitting. *MathWorks Blogs* Διαθέσιμο: https://blogs.mathworks.com/videos/2012/03/02/piecewise-linear-curve-fitting/ (Πρόσβαση: 10 Σεπτεμβρίου 2019)

[31] Daszykowski M., Walczak B. & Massart D. L. (2002) "Representative subset selection". *Analytica Chimica Acta* **468**, 91–103

[32] Puzyn T., Mostrag-Szlichtyng A., Gajewicz A., Skrzyński M. & Worth A. P. "Investigating the influence of data splitting on the predictive ability of QSAR/QSPR models" *Journal of Structural Chemistry* **22**, 795–804

[33] Stone L A, Kennard R. W. (1969) "Computer Aided Design of Experiments"

Technometrics **11**, 137–148

[34] OECD Environment Health and Safety Publications Series on Testing and Assessment No . 69 GUIDANCE DOCUMENT ON THE VALIDATION OF (QUANTITATIVE) STRUCTURE-ACTIVITY RELATIONSHIP [(Q)SAR] MODELS Environment Directorate. (2007)

[35] Rodgers J. L. & Nicewander W. A. (2008) "Thirteen Ways to Look at the Correlation Coefficient", *The American Statistician* **42**, 59–66

[36] Melagraki G. & Afantitis A. (2013) "Enalos KNIME nodes: Exploring corrosion inhibition of steel in acidic medium", *Chemometrics and Intelligent Laboratory Systems* **123**, 9–14

[37] Χ.Ν.Στεφανάκος (2011) "Προγραμματίζοντας σε Matlab", *Εκδόσεις Συμμετρία*

[38] Houcque D. "Introduction to Matlab for engineering students", *Northwester University* (2005). Διαθέσιμο: https://www.mccormick.northwestern.edu/documents/students/undergraduate/introduction-to-matlab.pdf (Πρόσβαση: 10 Σεπτεμβρίου 2019)

[39] Desktop basics. *MATLAB & Simulink, MathWorks* Διαθέσιμο: https://ch.mathworks.com/help/matlab/learn_matlab/desktop.html. (Πρόσβαση: 10 Σεπτεμβρίου 2019)

[40] Löfberg J. (2004) "YALMIP: A toolbox for modeling and optimization in MATLAB" in *2004 IEEE International Conference on Robotics and Automation* 284–289

[41] Discover Mosek. Διαθέσιμο: https://www.mosek.com/discover/. (Πρόσβαση: 10 Σεπτεμβρίου 2019)

[42] Gurobi Optimizer. Διαθέσιμο: http://www.gurobi.com/products/gurobi-optimizer. (Πρόσβαση: 10 Σεπτεμβρίου 2019)

[43] Lamon L., Asturiol D., Richarz A., Joossens E., Graepel R., Aschberger K., Worth A. (2018) "Grouping of nanomaterials to read-across hazard endpoints: From data collection to assessment of the grouping hypothesis by application of chemoinformatic techniques", *Particle and Fibre Toxicology* **15**, 1–17

[44] Cardoso-Silva J., Papadatos G., Papageorgiou L. G. & Tsoka S. (2018) "Optimal Piecewise Linear Regression Algorithm for QSAR Modelling", *Molecular Informatics* **38**, 1–14

[45] Yang L., Liu S., Tsoka S. & Papageorgiou L. G. (2016) "Mathematical Programming for Piecewise Linear Regression Analysis", *Expert Systems with Applications* **44**, 156–146

[46] Gajewicz A. *et al.* (2015) "Towards understanding mechanisms governing cytotoxicity of metal oxides nanoparticles: Hints from nano-QSAR studies" *Nanotoxicology* **9**, 313–325

[47] Xia X. *et al.* (2011) "Mapping the Surface Adsorption Forces of Nanomaterials in Biological Systems", *ACS Nano* **5**, 1–6

[48] Fourches D., Pu D., Tassa C., Weissleder R., Shaw S.Y., Mumper R.J., & Trospha A. (2010) "Quantitative Nanostructure-Activity Relationship (QNAR) Modeling", *ACS Nano* **5**, 1–7

[49] Varsou D. D. *et al.* (2018) "ToxFlow: A Web-Based Application for Read-Across Toxicity Prediction Using Omics and Physicochemical Data", *Journal of Chemical*

Information and Mondeling **58**, 543–549

[50] Subramanian A. *et al.* (2005) "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles", *Proceedings of the National Academy of Science U. S. A.* **102**, 15545–50

[51] Hänzelmann S., Castelo R. & Guinney J. (2013), "GSVA: gene set variation analysis for microarray and RNA-seq data", *BMC Bioinformatics* **14**, 7

[52] Mold2. *U.S. Food & Drug Administration* Ιστοσελίδα: https://www.fda.gov/science-research/bioinformatics-tools/mold2, (Πρόσβαση: 10 Σεπτεμβρίου 2019)

[53] Weissleder R., Kelly K., Sun E. Y., Shtatland T. & Josephson L. (2005) "Cell-specific targeting of nanoparticles by multivalent attachment of small molecules", *Nature Biotechnology* **23**, 1418–1423

[54] Dehmer M. & Varmuza K., (2012) "Statistical Modelling of Molecular Descriptors in QSAR/QSPR", *Wiley-Blackwell*