Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Χημικών Μηχανικών

Διπλωματική Εργασία

Ανάπτυξη υπολογιστικών ροών για την εκπαίδευση μαθηματικών μοντέλων πρόβλεψης ιδιοτήτων υλικών με τεχνικές μηχανικής μάθησης

Ιάσων Σωτηρόπουλος

Επιβλέπων:
Καθ. Χαράλαμπος Σαρίμβεης
Καθηγητής ΕΜΠ

Συνεπιβλέπουσα:
Δρ. Μαριάννα Κοτζαμπασάκη
Ερευνήτρια, ΕΜΠ

Οκτώβριος, 2020

National Technical University of Athens
School of Chemical Engineering

Diploma Thesis

Development of computational pipelines for training mathematical models predicting properties of materials using machine learning techniques

Iason Sotiropoulos

Supervisor:
Prof. Haralambos Sarimveis
Professor, School of Chemical Engineering, NTUA

Co-supervisor:
Dr. Marianna Kotzabasaki
Researcher, School of Chemical Engineering, NTUA

October, 2020

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Χημικών Μηχανικών

Διπλωματική Εργασία

Ανάπτυξη υπολογιστικών ροών για την εκπαίδευση μαθηματικών μοντέλων πρόβλεψης ιδιοτήτων υλικών με τεχνικές μηχανικής μάθησης

Ιάσων Σωτηρόπουλος

Επιβλέπων:
Καθ. Χαράλαμπος Σαρίμβεης
Καθηγητής ΕΜΠ

Συνεπιβλέπουσα:
Δρ. Μαριάννα Κοτζαμπασάκη
Ερευνήτρια, ΕΜΠ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την ___η Οκτωβρίου 2020

Καθηγητής ΕΜΠ	Καθηγητής ΕΜΠ	Καθηγήτρια ΕΜΠ
Χαράλαμπος Σαρίμβεης	Χρήστος Κυρανούδης	Ευαγγελία Παυλάτου
……..	……..	……..

Περίληψη

Η παρούσα διπλωματική εργασία εστίασε στις επιστημονικές περιοχές της εξόρυξης δεδομένων (Data mining) και της μηχανικής μάθησης (machine learning), με στόχο την ανάπτυξη υπολογιστικών ροών που συνδέουν την τοξικότητα νανοϋλικών με τις φυσικοχημικές τους ιδιότητες. Συγκεκριμένα, στόχος της εργασίας αυτής ήταν η κατασκευή στατιστικών μοντέλων και μοντέλων μηχανικής μάθησης που περιγράφουν τη συσχέτιση της τοξικότητας με τις ιδιότητες των υλικών (*Quantitative structure–activity relationship, QSAR*). Έτσι, εφαρμόσαμε μια σειρά αλγορίθμων μηχανικής μάθησης σε δύο διαφορετικά σύνολα δεδομένων με στόχο την κατηγοριοποίηση άγνωστων δειγμάτων ως προς την τοξικότητά τους. Το πρώτο σετ δεδομένων κατασκευάσθηκε αντλώντας πληροφορίες από τη βιβλιογραφία για την τοξικότητα, τις φυσικοχημικές ιδιότητες και τις πειραματικές μετρήσεις 15 νανοσωλήνων άνθρακα πολλαπλών τοιχωμάτων (*Multi-Walled Carbon Nanotubes, MWCNTs*). Σε αυτό το σύνολο δεδομένων εξετάστηκαν αρκετά μοντέλα μηχανικής μάθησης, και συνδυασμοί αυτών με στόχο την εξαγωγή μιας συναινετικής πρόβλεψης (consensus prediction). Παράλληλα, αναπτύχθηκε μέθοδος ενισχυτικής μάθησης, που στηρίζεται στη μπεϋζιανή στατιστική, με στόχο τη βέλτιστη επιλογή των παραμέτρων βαθμονόμησης των μοντέλων. Το δεύτερο σετ δεδομένων απαρτίζεται από φυσικοχημικές ιδιότητες και δύο κλάσεις τοξικότητας για 16 υπέρ-παραμαγνητικά νανοσωματίδια οξειδίων σιδήρου (*Super-Paramagnetic Iron Oxide Nanoparticles, SPIONs*). Για την συλλογή αυτών των δεδομένων διεξήχθη μια ενδελεχής έρευνα στην διεθνή βιβλιογραφία, αντλώντας πληροφορίες από 12 διαφορετικές μελέτες. Στην περίπτωση αυτή, χρησιμοποιήθηκε εξειδικευμένο λογισμικό αλγορίθμου γενετικού προγραμματισμού, για την επιλογή του κατάλληλου προβλεπτικού μοντέλου για την τοξικότητα των οξειδίων σιδήρου. Και στα δύο διαφορετικά σύνολα δεδομένων, πριν τη μοντελοποίηση προηγήθηκε η προεπεξεργασία των δεδομένων και ο μετασχηματισμός τους, ώστε να είναι δυνατή η εισαγωγή τους στους αλγορίθμους μηχανικής μάθησης. Τελευταίο στάδιο στην ανάλυση των δεδομένων αποτέλεσε η μείωση του μεγέθους των μοντέλων και ο υπολογισμός του πεδίου εφαρμογής τους. Στη διπλωματική αυτή εργασία αναπτύχθηκαν υπολογιστικές ροές που υλοποιούν και αυτοματοποιούν την ανωτέρω διαδικασία. Τα αποτελέσματα αυτής της διπλωματικής εργασίας έδειξαν πως οι μέθοδοι μηχανικής μάθησης μπορούν να δώσουν απαντήσεις σε προβλήματα τοξικότητας νανοϋλικών, με την ανάπτυξη μοντέλων που καταλήγουν σε υψηλότερη ακρίβεια από αυτή των συμβατικών στατιστικών μεθόδων.

Λέξεις-κλειδιά

Μηχανική μάθηση, Εξόρυξη δεδομένων, νανοσωλήνες άνθρακα, οξείδια σιδήρου, Τυχαίο Δάσος, Μηχανές Διανυσμάτων Στήριξης, Λογιστική Παλινδρόμηση, Γραμμική Παλινδρόμηση, Αφελής Bayes, Παλινδρόμηση Μερικών Ελαχίστων Τετραγώνων, Ανάλυση Κύριων Συνιστωσών, Ιεραρχική Συσταδοποίηση

(η σελίδα αυτή αφέθηκε σκόπιμα λευκή)

Abstract

This diploma thesis focused on the scientific areas of Data Mining and Machine Learning, with aim the development of computational pipelines that connect the toxicity of nanomaterials with their physicochemical properties. The primary aim of this thesis was the construction of several statistical models that describe the connection of toxicity endpoint of materials to their properties (*Quantitative structure–activity relationship, QSAR*). In order to achieve this, we applied a group of machine learning methods, on two different datasets with respect to the toxicity endpoint. The first dataset was constructed by collecting from the literature toxicity, physicochemical and experimental data of 15 Multi-Walled Carbon Nanotubes (MWCNTs). This dataset contained data for in the literature, in similar studies. Several machine learning models, and combinations of these models were trained on the dataset in order to extract a consensus prediction. A reinforcement learning technique, based in Bayesian statistics, was developed in order to tune the hyperparameters of the models. The second dataset contained data of physicochemical properties and a two-class toxicity endpoint for 16 observations of Super-Paramagnetic Iron Oxide Nanoparticles (SPIONs). An extensive literature search took place and the data were manually extracted from 12 studies. In the case of SPIONs, α genetic programming algorithm was applied to find the appropriate predictive model for the toxicity of the nanoparticles. In both datasets, first step of the pipeline was preprocessing the data in order to fix problematic values and to transform the data to "machine" understandable terms. The last step of the data analysis workflow was the reduction of the models' dimensionality and the calculation of their domain of applicability. Computational pipelines were developed in order to automate the above analysis. The results of this thesis showed that the machine learning methods are capable of providing solutions to nanomaterials' toxicity problems, by building models of higher accuracy compared to conventional statistical methods.

Keywords

Machine Learning, Data Mining, MWCNTs, SPIONs, Random Forest, RF, Support Vectors Machines, SVM, Logistic Regression, LR, Linear Regression, OLS, Naïve Bayes, NB, Partial Least Squares, PLS, Principal Component Analysis, PCA, Hierarchical Clustering, HC

(this page was intentionally left blank)

Περιεχόμενα

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Πρόλογος και Ευχαριστίες

Η παρούσα διπλωματική εργασία εκπονήθηκε στην μονάδα Αυτόματης Ρύθμισης και Πληροφορικής της Σχολής Χημικών Μηχανικών ΕΜΠ. Η διπλωματική εργασία αποτελεί, χωρίς καμία αμφιβολία, μία πρώτη επαφή κάθε φοιτητή με την επιστημονική έρευνα. Ειδικότερα, όταν η ερευνητική ομάδα με την οποία θα συνεργαστεί ο φοιτητής έχει μεράκι και ενδιαφέρον για την επιστήμη, του δίνεται το ερέθισμα να ασχοληθεί εκτενέστερα. Καθώς η διεθνής κοινότητα στρέφεται ολοένα και περισσότερο προς την σύνδεση της «συμπεριφοράς» των υλικών με τις ιδιότητες τους, αντιλαμβάνεται τις δυνατότητες που μπορούν να της προσφέρουν οι τεχνικές μηχανικής μάθησης. Το εργαστήριο μας, έχοντας από καιρό εφαρμόσει αλγορίθμους μηχανικής μάθησης στον τομέα της βιοπληροφορικής και της νανοπληροφορικής, και έχοντας πρόσβαση στους μελλοντικούς χημικούς μηχανικούς, μπορεί να ακολουθήσει μία λαμπρή πορεία σε αυτόν τον τομέα. Ελπίζω η εργασία αυτή να αποτελέσει κίνητρο στους συμφοιτητές μου, ώστε να ασχοληθούν με αντίστοιχα θέματα στις δικές τους εργασίες.

Αρχικά, θα ήθελα να ευχαριστήσω τον καθηγητή Χαράλαμπο Σαρίμβεη, πρωτίστως για την αποδοχή του και την ευκαιρία που μου έδωσε να γίνω μέλος της ομάδας του, αλλά και διότι με την στάση του, τόσο ως ερευνητής όσο και ως καθηγητής αποτελεί παράδειγμα για εμάς, τους νέους Χημικούς Μηχανικούς. Θα ήθελα ακόμη να τον ευχαριστήσω για την υπομονή του και την άψογη συνεργασία μας από την πλευρά του, η οποία ευελπιστώ να συνεχιστεί στο μέλλον. Μαζί με αυτόν θα ήθελα να ευχαριστήσω τα μέλη της ομάδας του Παντελή Καρατζά και Περικλή Τσίρο για την καθοδήγηση, τον χρόνο και την όρεξη τους στα πρώτα βήματα μου. Ακόμα, θα ήθελα να ευχαριστήσω θερμά την Δρ. Μαριάννα Κοτζαμπασάκη για την ευγενική προσφορά της, την άψογη συνεργασία μας και για το ενδιαφέρον που έδειξε στην διπλωματική μου. Τέλος, ένα μεγάλο ευχαριστώ οφείλω στην οικογένεια μου, στους φίλους μου και στη σύντροφό μου, τους «αφανείς ήρωες» που με την αμέριστη κατανόηση τους στάθηκαν τα στηρίγματα μου αυτήν την περίοδο.

Κλείνοντας τον κύκλο των προπτυχιακών σπουδών μου κλείνει ένα μεγάλο κεφάλαιο στην έως τώρα πορεία μου. Ευελπιστώ, ωστόσο, λαμβάνοντας ελπίδα από τους ανθρώπους μου και έχοντας ως παράδειγμα τους ερευνητές με τους οποίους είχα την τιμή να συνεργαστώ, το τέλος αυτό να αποτελέσει την αρχή για κάτι μεγαλύτερο.

Εισαγωγή

Ένα βασικό χαρακτηριστικό που καθορίζει τις σύγχρονες επιστημονικές εξελίξεις είναι η αναζήτηση και αξιοποίηση των διαθέσιμων πληροφοριών και δεδομένων, ο όγκος των οποίων αναπτύσσεται ραγδαία. Με την εφαρμογή κατάλληλων αλγορίθμων και διαδικασιών, η επεξεργασία των πληροφοριών μπορεί να οδηγήσει σε εξαγωγή χρήσιμης γνώσης. Το διαδίκτυο παίζει σημαντικό ρόλο σε αυτήν την τεχνολογική εξέλιξη, καθώς τα δεδομένα μπορούν να είναι διαθέσιμα και να επικαιροποιούνται σε πραγματικό χρόνο. Είναι έντονη μάλιστα τα τελευταία χρόνια η τάση για ανοικτά δεδομένα και ανοικτά εργαλεία που μπορούν να αξιοποιηθούν από όλη την ερευνητική κοινότητα μέσω του διαδικτύου. Σε αυτό το πλαίσιο κινούνται και πολλοί σύγχρονοι κλάδοι πληροφορικής όπως η βιοπληροφορική, (bioinformatics), η χημειοπληροφορική (cheminformatics) και πιο πρόσφατα η νανοπληροφορική (nanoinformatics). Η υπολογιστική διαδικασία της αναγνώρισης προτύπων και μοτίβων σε σύνολα δεδομένων ονομάζεται εξόρυξη δεδομένων και αποτελεί έναν κόμβο συνάντησης πολλών επιστημονικών κλάδων, με βασικότερο τη μηχανική μάθηση (machine learning).

Η εργασία αυτή εντάσσεται στο πεδίο της νανοπληροφορικής. Εστίασε στην εξόρυξη δεδομένων για τα νανοϋλικά και την αξιοποίησή τους με σκοπό την ανάπτυξη μοντέλων που να είναι σε θέση να προβλέπουν ιδιότητες τοξικότητας των νανοϋλικών. Τα μοντέλα αυτά μπορούν να υποστηρίξουν ή ακόμη και να αντικαταστήσουν τα πειράματα, κυρίως εκείνα που διεξάγονται σε ζώα. Μπορούν επίσης να βοηθήσουν στην κατανόηση του μηχανισμού και των αιτιών που προκαλούν ένα βιολογικό φαινόμενο όπως αυτό της τοξικότητας. Τέλος, αποτελούν ιδιαίτερα χρήσιμα εργαλεία για τη βιομηχανία, καθώς δίνουν κατευθύνσεις στο σχεδιασμό νέων προϊόντων έτσι ώστε αυτά να είναι λειτουργικά αλλά και ασφαλή στη χρήση τους.

Συγκεκριμένα, συλλέχθηκαν δεδομένα τοξικότητας για δύο διαφορετικές κατηγορίες νανοϋλικών, νανοσωλήνες άνθρακα πολλαπλών τοιχωμάτων (*Multi-Walled Carbon Nanotubes, MWCNTs*) και υπερπαραμαγνητικά νανοσωματίδια οξειδίων σιδήρου (*Super-Paramagnetic Iron Oxide Nanoparticles, SPIONs*). Τα υλικά αυτά έχουν ήδη ευρεία χρήση σε ιατρικές εφαρμογές, η οποία φαίνεται να αυξάνεται, καθώς χάρη στη δομή τους και στις ιδιότητές τους, διαθέτουν χαρακτηριστικά σπάνια και ιδιαιτέρως χρήσιμα. Ειδικότερα, το έντονο ενδιαφέρον γύρω από τους MWCNTs έγκειται στο υψηλό μέτρο ελαστικότητας, στην ηλεκτρική και θερμική αγωγιμότητά τους, καθώς και στη χαμηλή διαλυτότητά τους, ενώ τα SPIONs χαρακτηρίζονται για τις καλές μαγνητικές τους ιδιότητες και για την χαμηλή τους τοξικότητα. Τα νανοϋλικά αυτά, ωστόσο, έχει παρατηρηθεί πως υπό ορισμένες προϋποθέσεις εμφανίζουν τοξική συμπεριφορά στον ανθρώπινο οργανισμό. Είναι φανερή επομένως, η χρησιμότητα ανάπτυξης μοντέλων που να συσχετίζουν την τοξικότητα των νανοϋλικών αυτών με τις φυσικοχημικές τους ιδιότητες (QSAR).

Τα δεδομένα που συλλέχθηκαν προέρχονται από *in vitro* και *in vivo* έρευνες που στόχευαν στη σύνδεση της τοξικότητας των νανοϋλικών με τα φυσικοχημικά χαρακτηριστικά τους. Η αποθήκευση τους έγινε σε μια δομή, χαρακτηριστική και συνήθη στους τομείς εξόρυξης και ανάλυσης δεδομένων δηλαδή σε πίνακες δεδομένων, στους οποίους κάθε γραμμή αναπαριστά ένα διαφορετικό δείγμα (στην περίπτωσή μας ένα MWCNT ή ένα SPION) και κάθε στήλη φιλοξενεί τιμές μίας χαρακτηριστικής μεταβλητής για τα δείγματα (φυσικοχημικές ιδιότητες, τιμές που προκύπτουν από προηγούμενα πειράματα κ.ο.κ.). Σε αυτό το σύνολο δεδομένων, προστίθεται μία στήλη, η στήλη του τελικού σημείου της ανάλυσης, μέσω της οποίας αντιστοιχίζεται κάθε δείγμα σε μία κλάση (π.χ. «τοξικό» ή «μη τοξικό»). Σε αντίθεση με άλλες επιστημονικές περιοχές όπου υπάρχει πληθώρα διαθέσιμων δεδομένων (big data), στην περιοχή της νανοπληροφορικής ο κανόνας είναι η μικρή διαθεσιμότητα δεδομένων, κάτι που λαμβάνεται υπόψη και στην επιλογή των αλγορίθμων μοντελοποίησής τους.

Ο βασικός στόχος αυτής της διπλωματικής εργασίας ήταν η ανάπτυξη ολοκληρωμένων υπολογιστικών ροών για την κατασκευή μοντέλων που να μπορούν με ικανοποιητική ακρίβεια να προβλέψουν την τοξικότητα αυτών των τύπων νανοϋλικών. Οι αλγόριθμοι που δοκιμάστηκαν καλύπτουν ένα ευρύ φάσμα κλασσικών μεθόδων μηχανικής μάθησης, αλλά και συνδυασμού αυτών και τη βελτιστοποίηση όλης της διαδικασίας, δηλαδή της κατάλληλης προεπεξεργασίας των δεδομένων, την επιλογή του κατάλληλου αλγόριθμου και της βέλτιστης επιλογής των παραμέτρων βαθμονόμησης. Για το σκοπό, αυτό προγραμματίστηκε υπολογιστική ροή που βασίζεται σε Μπεϋζιανή στατιστική αλλά και αλγόριθμοι αυτοματοποιημένης μηχανικής μάθησης (automated machine learning) που βασίζονται στο γενετικό προγραμματισμό. Λόγω του μικρού όγκου δεδομένων δεν χρειάστηκε κάποιο ισχυρό υπολογιστικό σύστημα. Αντιθέτως, όλες οι αναλύσεις διεξήχθησαν σε προσωπικό φορητό Η/Υ, με λειτουργικό σύστημα Windows 8.1, σε γλώσσα Python (version 3.6) και προγραμματιστικό περιβάλλον Jupyter notebook.

Τα μοντέλα που προέκυψαν χαρακτηρίστηκαν από πολύ καλή ακρίβεια και απόδοση στη διαδικασία αξιολόγησης και στις δύο μελέτες περίπτωσης. Τα αποτελέσματα συμβάλλουν στην ανάδειξη της επιστήμης των δεδομένων και της μηχανικής μάθησης ως σημαντικών εργαλείων στην περιοχή της νανοπληροφορικής για το σχεδιασμό και την ανάπτυξη νέων αποτελεσματικών προϊόντων νανοτεχνολογίας που ταυτόχρονα να είναι ασφαλή στη χρήση τους. Από τα αποτελέσματα αυτά προέκυψε μια δημοσίευση σε έγκριτο επιστημονικό περιοδικό:

Kotzabasaki, M. I.; Sotiropoulos, I.; Sarimveis, H. QSAR Modeling of the Toxicity Classification of Superparamagnetic Iron Oxide Nanoparticles (SPIONs) in Stem-Cell Monitoring Applications: An Integrated Study from Data Curation to Model Development. *RSC Adv* **2020**, *10* (9), 5385–5391. https://doi.org/10.1039/C9RA09475J.[1]

Κεφάλαιο 1

Ποσοτικές Σχέσεις Δομής-Δράσης (QSAR)

Οι Ποσοτικές Σχέσεις Δομής-Δράσης (Quantitative Structure Activity Relationships, QSAR) στοχεύουν στην ανάπτυξη εξισώσεων ή μοντέλων που προβλέπουν τη βιολογική δράση από τη δομή των νανοϋλικών, με την αξιοποίηση και τη μοντελοποίηση των διαθέσιμων δεδομένων. Στο κεφάλαιο αυτό θα γίνει μία αναφορά στην εξέλιξη των QSAR μοντέλων μέσα στην ιστορία, θα τοποθετηθούν οι βασικοί άξονες γύρω από τους οποίους αναπτύσσονται αξιόπιστα μοντέλα QSAR, και θα παρουσιαστεί μια εισαγωγή στον τρόπο με τον οποίο δομείται η ροή εργασίας για μια μελέτη QSAR. Οι μαθηματικοί αλγόριθμοι και οι τεχνικές μοντελοποίησης θα αναπτυχθούν εκτενέστερα στο Κεφάλαιο 2.

1.1 Ιστορική αναδρομή

Ο ισχυρισμός πως η βιολογική δράση συνδέεται με τη δομή των χημικών ενώσεων διατυπώθηκε για πρώτη φορά το 1868 από τους Brown και Fraser.[2] Μερικές δεκαετίες αργότερα, στα τέλη του 19ου αιώνα, οι Meyer και Overton κατάφεραν να διατυπώσουν για πρώτη φορά μια σχέση που εκφράζει μία δράση (αναισθητική δράση) με μία φυσικοχημική ιδιότητα (λιποφιλία).[3,4] Σημαντική συνεισφορά, στην ποσοτικοποίηση της επίδρασης των φυσικοχημικών ιδιοτήτων στη δράση, ήταν οι εξισώσεις του Hammett. Το 1937 ο Hammett εισήγαγε τις ηλεκτρονιακές σταθερές σ οι οποίες εκφράζουν την επίδραση των οργανικών μορίων στον ρυθμό μίας χημικής αντίδρασης.[5] Στη συνέχεια, μερικά χρόνια αργότερα, ο Taft, εισάγοντας την στερική σταθερά E_σ, απέρριψε την παραδοχή του Hammett για σταθερή επιρροή του υποκαταστάτη κι έδειξε πως η επίδραση των υποκαταστατών σε μια αντίδραση μπορεί να ποσοτικοποιηθεί.[6]

Η προσέγγιση των Hammett και Taft ενέπνευσε πολλούς επιστήμονες και ερευνητές, φτάνοντας στους Hansch και Fujita, οι οποίοι το 1964 ανέπτυξαν το πρώτο QSAR μοντέλο. Οι ανωτέρω ερευνητές μελέτησαν την βιολογική δράση ως συνάρτηση τριών φυσικοχημικών ιδιοτήτων, της λιποφιλίας, των ηλεκτρονιακών ιδιοτήτων και των στερικών ιδιοτήτων (εξίσωση 1.1).[7]

$$logBR = -a(logP)^2 + blogP + \rho\sigma + \delta E_\sigma + c \quad (1.1)$$

Στην εξίσωση 3.1 $logBR$ είναι γενική έκφραση του λογαρίθμου της βιολογικής απόκρισης, $logP$ είναι ο συντελεστής μερισμού που αποτελεί μέτρο της λιποφιλίας, σ η ηλεκτρονιακή σταθερά του Hammett, που εκφράζει τις ηλεκτρονιακές ιδιότητες των υποκαταστατών και E_σ η στερική σταθερά του Taft, που εκφράζει τις στερικές ιδιότητες των υποκαταστατών. Οι συντελεστές a, b, ρ, δ και ο σταθερός όρος c προκύπτουν από πολλαπλή γραμμική ανάλυση παλινδρόμησης.[7] Έκτοτε, τεράστια βήματα και εξελίξεις έχουν γίνει στην προσπάθεια δημιουργίας Ποσοτικών Σχέσεων Δομής-Δράσης.

1.2 Συστατικά μιας κλασικής QSAR μελέτης

Η εξαγωγή Ποσοτικών Σχέσεων Δομής-Δράσης είναι μια σύνθετη ανάλυση με ανάμειξη πολλών και διαφορετικών τομέων (εικόνα 1.1). Τα κλασσικά QSAR μοντέλα καταλήγουν σε γραμμικές εξισώσεις με γενικό τύπο την εξίσωση 1.2 μεταξύ μιας συγκεκριμένης βιολογικής απόκρισης και των δομικών χαρακτηριστικών μιας σειράς συγγενών (δομικά) μορίων.

$$Βιολογική\ Δράση = \alpha_0 + \alpha_1 P_1 + \alpha_2 P_2 + \cdots + \alpha_n P_n \quad (1.2)$$

$P_1 - P_n$: παράμετροι που αφορούν τη δομή
$\alpha_1 - \alpha_n$: συντελεστές που προκύπτουν κατά την διαδικασία εξαγωγής της έκφρασης.

Για μια αξιόπιστη μελέτη QSAR είναι απαραίτητο:[8,9]

(α) Να επιλεγεί μία σειρά δομικώς συγγενών ουσιών, οι οποίες να έχουν καλά καταμερισμένες τιμές βιολογικής δράσης και να επαναλαμβάνονται ως προς έναν υποδοχέα στόχο. Να διαχωριστούν οι ουσίες σε μια ομάδα εκμάθησης και μια ομάδα επαλήθευσης. Οι ομάδες είναι απαραίτητο να είναι αντιπροσωπευτικές και να περιέχουν ποικιλία χημικών δομών.

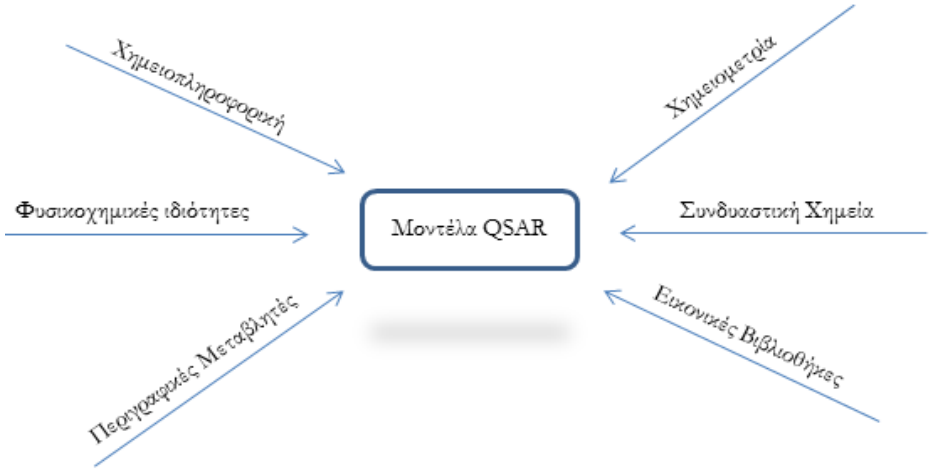
(β) Να προσδιοριστούν χαρακτηριστικά που προκύπτουν είτε από πειραματικές μελέτες ή από υπολογιστικές διαδικασίες για να περιγράψουν τη δομή των ουσιών.

(γ) Να επιλεγεί κατάλληλη στατιστική μέθοδος ή μέθοδος μηχανικής μάθησης για την εκπαίδευση του προβλεπτικού μοντέλου QSAR.

(δ) Να χρησιμοποιηθούν στατιστικές μέθοδοι αξιολόγησης για να εξασφαλιστεί η προβλεπτική ικανότητα και ακρίβεια του μοντέλου QSAR κυρίως στα δεδομένα επαλήθευσης.

(ε) Να καθοριστεί το πεδίο εφαρμογής του μοντέλου και να αξιολογηθεί ως προς την συνεισφορά του στην σύνθεση άλλων καινοτόμων αναλόγων.

(στ) Να επικαιροποιείται το μοντέλο τακτικά με την εισαγωγή νέων δεδομένων.



Εικόνα 1.1 **Ενδεικτικοί τομείς που συνδέονται με τα QSAR μοντέλα.**

1.2.1 Μεταβλητές απόκρισης και Περιγραφικές μεταβλητές

Τα βιολογικά δεδομένα που χρησιμοποιούνται για την εξαγωγή μοντέλων πρέπει να πληρούν τις παρακάτω προϋποθέσεις:[8]

(α) Οι ενώσεις είναι απαραίτητο να δρουν όλες στον ίδιο υποδοχέα και με τον ίδιο μηχανισμό δράσης.

(β) Τα αριθμητικά δεδομένα πρέπει να εκφράζουν την βιολογική δράση και να αντιστοιχούν σε μοριακές συγκεντρώσεις. Τα βιολογικά δεδομένα που δεν είναι εκφρασμένα αριθμητικά αλλά ομαδοποιούν τις ενώσεις ανάλογα με τη δράση τους, σε αδρανείς, δραστικές και πολύ δραστικές είναι δυνατόν να χρησιμοποιηθούν για την εξαγωγή μοντέλων ταξινόμησης.

(γ) Τα δεδομένα βιολογική δράσης πρέπει να έχουν καλή κατανομή και να καλύπτουν ένα μεγάλο εύρος τιμών.

(δ) Τα βιολογικά πειράματα είναι απαραίτητο να συνοδεύονται με πληροφορίες για την φερεγγυότητά τους και την δυνατότητα να επαναληφθούν. Τα στατιστικά στοιχεία του μοντέλου δεν θα πρέπει να είναι καλύτερα από τα στατιστικά στοιχεία των βιολογικών πειραμάτων.

Οι μεταβλητές απόκρισης αφορούν τη δράση που προσπαθεί να προσδιορίσει το QSAR μοντέλο. Σε ένα τέτοιο μοντέλο η μεταβλητή απόκρισης μπορεί να είναι είτε συνεχής (μοντέλα παλινδρόμησης-regression) είτε κατηγορική (μοντέλα ταξινόμησης-classification).[10] Σε μοντέλα παλινδρόμησης η μεταβλητή απόκρισης συνήθως χρησιμοποιείται σε λογαριθμική μορφή ώστε τα

βιολογικά μεγέθη να αυξάνουν αριθμητικά με την αύξηση της δράσης.[9] Αντίθετα, στα μοντέλα ταξινόμησης οι μεταβλητές απόκρισης μπορούν να πάρουν συγκεκριμένες τιμές, κάθε μία χαρακτηρίζει μία κατηγορία ή κλάση, επομένως ένας λογαριθμικός μετασχηματισμός των μεταβλητών αυτών δεν έχει κάποιο ιδιαίτερο νόημα.

Οι περιγραφικές μεταβλητές συνήθως αφορούν τη δομή του μορίου ή του υλικού, μπορούν ωστόσο να είναι και αποτελέσματα πειραμάτων που έχουν γίνει για το μόριο ή το υλικό αυτό. Η δομή ενός μορίου μπορεί να εκφραστεί με ένα πλήθος περιγραφικών μεταβλητών (descriptors), οι οποίες υπολογίζονται από κατάλληλα λογισμικά (όπως το Dragon). Για τον υπολογισμό των περιγραφικών μεταβλητών με τα διάφορα λογισμικά η δομή των μορίων εισάγεται στον υπολογιστή σχεδιαστικά ή με την γραφή SMILES (Simplified Molecular Input Line Entry Structure).[11] Σύμφωνα με τη γραφή αυτή το μόριο παριστάνεται σε μια γραμμή θεωρώντας όλα τα άτομα (πλην του υδρογόνου), τα οποία απεικονίζονται με τα σύμβολά τους στο περιοδικό σύστημα. Μέσω των SMILES είναι δυνατόν να υπολογισθούν περισσότερες από 1600 περιγραφικές μεταβλητές που προκύπτουν από διάφορες θεωρίες βασιζόμενες στη φυσικοχημεία, στην κβαντική χημεία, στα μαθηματικά, στην τοπολογία, στην χημειοπληροφορική, στην οργανική χημεία και σε άλλους επιστημονικούς κλάδους.[12,13] Εκτός, όμως, από αυτές τις μεταβλητές μπορούν να χρησιμοποιηθούν και πειραματικά αποτελέσματα από προηγούμενες έρευνες, όπως τοξικολογικές μελέτες, χορήγηση ενώσεων σε πειραματόζωα και κάθε άλλου τύπου μελέτη.

1.3 Επισκόπηση συχνών στατιστικών μεθόδων για τα QSAR μοντέλα

Οι στατιστικές μέθοδοι που χρησιμοποιούνται συνήθως σε μελέτες QSAR είναι η Πολλαπλή Γραμμική Ανάλυση Παλινδρόμησης (Multiple Linear Regression Analysis, MLRA) και η Πολυμεταβλητή Ανάλυση Δεδομένων (Multivariate Data Analysis, MVDA). Η MVDA διακρίνεται στην Ανάλυση Κυρίων Συνιστωσών (Principal Component Analysis, PCA) και στην Ανάλυση Μερικών Ελαχίστων Τετραγώνων (Partial Least Squares, PLS). Οι στατιστικές αυτές τεχνικές καταλήγουν σε γραμμικά μοντέλα. Εάν η γραμμικότητα δεν είναι όμως επιθυμητή τότε με μετασχηματισμό της μεταβλητής που περιγράφει τη βιολογική απόκριση, όπως αλλαγή σε λογαριθμική κλίμακα, είναι δυνατό να αποφευχθεί η γραμμικότητα. Ακόμα, υπάρχουν στατιστικές μέθοδοι που οδηγούν σε μη γραμμικά μοντέλα όπως τα Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks, ANN), οι Μηχανές Διανυσμάτων Στήριξης (Support Vector Machines SVM), τα Δένδρα Απόφασης (Decision Trees) και άλλες.

1.3.1 Πολλαπλή Γραμμική Ανάλυση Παλινδρόμησης (MLRA)

Η MLRA επιτυγχάνει τη συσχέτιση μία εξαρτημένης μεταβλητής y με ένα σύνολο ανεξάρτητων μεταβλητών x_i καταλήγοντας σε μία εξίσωση που περιγράφει το φαινόμενο και είναι στη μορφή της εξίσωσης 1.2. Τόσο για την μέθοδο αυτή, όσο και για κάθε άλλη QSAR μεθοδολογία, αρχικά σχηματίζεται ένας πίνακας δεδομένων, όπου οι στήλες αντιστοιχούν σε διαφορετικούς descriptors και οι σειρές σε διαφορετικές ενώσεις ή μόρια ή υλικά. Ανάμεσα στις στήλες υπάρχει και η εξαρτημένη μεταβλητή y. Ο υπολογισμός των συντελεστών α_i γίνεται με την μέθοδο των ελαχίστων τετραγώνων μεταξύ των υπολογιστικών και των αναμενόμενων τιμών. Πιο συγκεκριμένα, για κάθε αναμενόμενη τιμή της y μεταβλητής, υπολογίζεται μία υπολογιστική τιμή $y_{υπολ.}$, όπου:[14]

$$y_{υπολ.} = f(x_1, x_2, \cdots, x_n) = a_0 + a_1x_1 + a_2x_2 + \cdots + \alpha_nx_n \quad (1.3)$$

Η ακρίβεια της υπολογισμένης αυτής τιμής υπολογίζεται από την διαφορά r από την πραγματική τιμή y. Για τον υπολογισμό των α_i, η μέθοδος των ελαχίστων τετραγώνων ελαχιστοποιεί το άθροισμα S (εξίσωση 1.4) των τετραγώνων των διαφορών r, όπως φαίνεται στην εξίσωση 1.5.

$$S = \sum_i^n r_i^2 = \sum_i^n \left(y_i - y_{\upsilon\pi o\lambda.i}\right)^2 = \sum_i^n (y_i - a_0 - \alpha_i x_i)^2 \quad (1.4)$$

$$\nabla S = 0 \rightarrow \begin{cases} \dfrac{\partial S}{\partial a_0} = 0 \\ \dfrac{\partial S}{\partial a_1} = 0 \\ \vdots \\ \dfrac{\partial S}{\partial a_n} = 0 \end{cases} \rightarrow \begin{cases} 2\sum_i^n r_i \dfrac{\partial r_i}{\partial a_0} = 0 \\ 2\sum_i^n r_i \dfrac{\partial r_i}{\partial a_1} = 0 \\ \vdots \\ 2\sum_i^n r_i \dfrac{\partial r_i}{\partial a_n} = 0 \end{cases} \rightarrow \begin{cases} -2\sum_i^n r_i \dfrac{\partial f(x_1, x_2, \cdots, x_n)}{\partial a_0} = 0 \\ -2\sum_i^n r_i \dfrac{f(x_1, x_2, \cdots, x_n)}{\partial a_1} = 0 \\ \vdots \\ -2\sum_i^n r_i \dfrac{\partial f(x_1, x_2, \cdots, x_n)}{\partial a_n} = 0 \end{cases} \quad (1.5)$$

Το σύστημα που δημιουργείται στην εξίσωση 1.5 αποτελείται από $n+1$ αγνώστους και εξισώσεις.[15] Σύνηθες μέτρο για την αξιολόγηση της MLRA είναι ο συντελεστής συσχέτισης R ο οποίος πρέπει να τείνει στο $|1|$. Είθισται, να χρησιμοποιείται ακόμα πιο συχνά όμως το τετράγωνο του συντελεστή συσχέτισης R^2 (εξίσωση 1.6) το οποίο εκφράζει το ποσοστό των περιπτώσεων που ερμηνεύει το QSAR μοντέλο.

$$R^2 = 1 - \frac{\sum_{i=1}^m \left(y_i - y_{\upsilon\pi o\lambda.\iota}\right)^2}{\sum_{i=1}^m \left(y_i - y_{\mu\acute{\varepsilon}\sigma o}\right)^2} \quad (1.6)$$

$$R^2_{adj.} = 1 - \frac{m-1}{m-(n+1)}(1 - R^2) \quad (1.7)$$

Ο συντελεστής συσχέτισης, ωστόσο, είναι ιδιαίτερα ευαίσθητος στο πλήθος των περιγραφικών μεταβλητών, καθώς σύμφωνα με τις εξισώσεις 1.4 - 1.5 ο συντελεστής κάθε μεταβλητής που προστίθεται προσαρμόζεται με τέτοιο τρόπο ώστε να ελαχιστοποιούνται τα σφάλματα, άρα να εκφράζεται μεγαλύτερο ποσοστό περιπτώσεων κι επομένως να αυξάνεται το R^2. Αυτή η συνθήκη όμως εγκυμονεί ιδιαίτερα υψηλούς κινδύνους για το QSAR μοντέλο, καθώς συνεπάγεται πως οποιαδήποτε μεταβλητή κι αν προστεθεί στα δεδομένα, ακόμα και τυχαία, είναι αδύνατον να οδηγήσει σε χαμηλότερο R^2, γεγονός που δεν θα έπρεπε να ισχύει. Για τον λόγο αυτό, χρησιμοποιείται ο προσαρμοσμένος συντελεστής συσχέτισης (adjusted R^2) στην αξιολόγηση μιας MLRA. Ο τύπος του προσαρμοσμένου R^2 περιγράφεται στην εξίσωση 1.7, όπου m το πλήθος των περιπτώσεων και n το πλήθος των μεταβλητών (εξαρτημένων και ανεξάρτητης).[16] Το πλήθος των παραμέτρων που μπορεί να περιλαμβάνει ένα QSAR μοντέλο που έχει προέλθει από την MLRA εξαρτάται από τους βαθμούς ελευθερίας $m-n$, όπου m και n τα ίδια με αυτά της εξίσωσης 1.7. Ιδιαζόντως σημαντικό για ένα μοντέλο είναι ο αριθμός βαθμών ελευθερίας να είναι όσο το δυνατόν μεγαλύτερος. Η συνθήκη αυτή οδήγησε στην ανάγκη δημιουργίας του προσαρμοσμένου R^2 το οποίο «προσαρμόζει» το R^2 στους βαθμούς ελευθερίας του μοντέλου.[17]

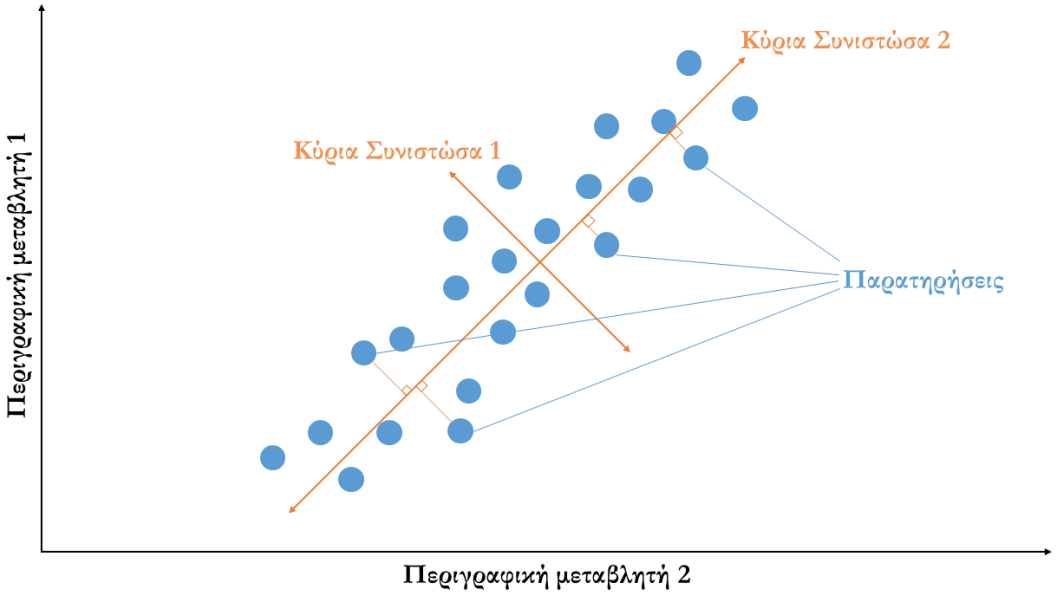
1.3.2 Πολυμεταβλητή ανάλυση δεδομένων (MVDA)

Η Πολυμεταβλητή Ανάλυση Δεδομένων (MVDA), όπως υποδηλώνει και το όνομα της, δίνει τη δυνατότητα διαχείρισης μεγάλου αριθμού σχετιζόμενων μεταξύ τους περιγραφικών μεταβλητών, ενώ επιτρέπει την ανάλυση, περισσότερων της μιας, μεταβλητών απόκρισης. Η μέθοδος αυτή στηρίζεται στην προβολή των σημείων από ένα χώρο πολλών διαστάσεων σε ένα χώρο με λιγότερες διαστάσεις. Η Πολυμεταβλητή Ανάλυση Δεδομένων περιλαμβάνει την Ανάλυση Κυρίων Συνιστωσών (Principal Component Analysis, PCA) και την Ανάλυση Ελαχίστων Τετραγώνων (Partial Least Squares,

PLS). [18]

Η Ανάλυση Κυρίων Συνιστωσών είναι μέθοδος ενός ενιαίου πίνακα περιγραφικών μεταβλητών. Αν θεωρήσουμε πως ο πίνακας αυτός διαθέτει n στήλες (ή μεταβλητές) τότε η PCA, όπως και οι περισσότερες μέθοδοι που πραγματοποιούν μετασχηματισμό των αρχικών δεδομένων, θεωρεί τις ενώσεις (ή μόρια ή υλικά) ως σημεία σε ένα χώρο n-διαστάσεων. Τα σημεία προβάλλονται σε ένα λιγότερων διαστάσεων χώρο, που οριοθετείται από τις Κύριες Συνιστώσες (Principal Components). Κάθε μία από τις αυτές αποτελεί έναν γραμμικό συνδυασμό των αρχικών μεταβλητών. Οι Κύριες Συνιστώσες έχουν το χαρακτηριστικό γνώρισμα να είναι κάθετες και ανεξάρτητες μεταξύ τους. Με τον μετασχηματισμό αυτό, δημιουργούνται νέοι άξονες, στους οποίους προβάλλονται τα σημεία. Έτσι έχουμε τη δημιουργία δυο νέων πινάκων, ο πίνακας T των συντεταγμένων των ενώσεων στις Κύριες Συνιστώσες και ο πίνακας P των φορτίων, δηλαδή των γωνιών που σχηματίζουν οι Κύριες Συνιστώσες ως προς τις αρχικές.[19]

Στην παρακάτω εικόνα (Εικόνα 1.2) περιγράφεται γραμμικά η αρχική απεικόνιση των σημείων σε ένα ορθογώνιο καρτεσιανό σύστημα συντεταγμένων και η προβολή των σημείων στις Κύριες Συνιστώσες όπως και η διαμόρφωση των φορτίων (διευθύνσεων) των Κυρίων Συνιστωσών ως προς τις αρχικές μεταβλητές.



Εικόνα 1.2 **Ανάλυση Κύριων Συνιστωσών, PCA**

Στην PCA το πλήθος των Κύριων Συνιστωσών που θα χρησιμοποιηθούν το επιλέγει ο μελετητής, σύμφωνα με τις γνώσεις του για το αντικείμενο της μελέτης του αλλά και με το πόσο θέλει να ελαττωθεί η διαστατικότητα του αρχικού πίνακα δεδομένων του. Στις Κύριες Συνιστώσες παρουσιάζεται μία φθίνουσα συμπεριφορά όσον αφορά την πληροφορία. Πιο συγκεκριμένα, η πρώτη Κύρια Συνιστώσα περιέχει το μεγαλύτερο ποσοστό πληροφορίας, η δεύτερη Κύρια Συνιστώσα περιέχει το αμέσως μεγαλύτερο ποσοστό πληροφορίας, ενώ αντίστοιχα, η τελευταία Κύρια Συνιστώσα περιέχει την λιγότερη πληροφορία από όλες τις άλλες.[20] Ιδιαιτέρως σημαντικές είναι επομένως οι πληροφορίες που μας δίδουν οι απεικονίσεις που αφορούν τις πρώτες δύο συντεταγμένων και των φορτίων των δυο πρώτων Κυρίων Συνιστωσών, οι οποίες ερμηνεύουν το μεγαλύτερο μέρος της διακύμανσης των δεδομένων. Η απεικόνιση των συντεταγμένων των δυο πρώτων Κυρίων Συνιστωσών μας πληροφορεί για τη σχέση των γραμμών μεταξύ τους και εκφράζει μια πρώτη και επεξηγηματική εικόνα των δεδομένων.[21]

Η ανάλυση των Μερικών Ελαχίστων Τετραγώνων περιλαμβάνει δύο πίνακες, έναν πίνακα αποκρίσεων και έναν πίνακα περιγραφικών μεταβλητών. Η PLS είναι επέκταση της PCA και σε αντίθεση με αυτή, αποτελεί ένα προβλεπτικό μοντέλο, καθώς με εφαρμογή παλινδρόμησης καταλήγει σε προβλέψεις. Πριν από αυτό το στάδιο, πραγματοποιείται ανάλυση PCA τόσο στον πίνακα των

περιγραφικών μεταβλητών όσο και στον πίνακα των αποκρίσεων. Με τον τρόπο αυτόν προκύπτουν οι πίνακες των συντεταγμένων T και των φορτίων P που δημιουργούν τον πίνακα των descriptors και οι αντίστοιχοι πίνακες συντεταγμένων U και φορτίων C που δημιουργούν τον πίνακα των αποκρίσεων.[22]

Οι πίνακες T και U προσαρμόζονται με τέτοιο τρόπο, ώστε να υπάρχει μία μεγάλη εσωτερική σχέση μεταξύ τους. Αυτό επιτυγχάνεται αποδίδοντας ένα «βάρος» στις αρχικές μεταβλητές με το οποίο πολλαπλασιάζεται ο πίνακας των φορτίων P, δημιουργώντας έτσι έναν νέο πίνακα, τον πίνακα βαρών W. Αυτή είναι μία πολύ χρήσιμη ιδιότητα της PLS καθώς με χρήση αυτών των τιμών είναι δυνατός ο υπολογισμός της σπουδαιότητας των αρχικών περιγραφικών μεταβλητών. Ειδικότερα, η σπουδαιότητα υπολογίζεται μέσω του VIP (Variable Importance in Projection), όπως δείχνει η εξίσωση 1.8, και εκφράζει τη συμβολή της κάθε μεταβλητής στις Κύριες Συνιστώσες.[23]

$$VIP_j = \sqrt{\frac{J \sum_{f=1}^{F} w_{jf}^2 SSY_f}{SSY_{total} F}} \quad (1.8)$$

όπου VIP_j η VIP τιμή μίας αρχικής j μεταβλητής, J το πλήθος των αρχικών μεταβλητών, w_{fj} η τιμή του βάρους της j μεταβλητής για την f συνιστώσα, SSY_f το άθροισμα ων τετραγώνων της διακύμανσης της f συνιστώσας, SSY_{total} το άθροισμα ων τετραγώνων της διακύμανσης της εξαρτημένης μεταβλητής (ή μεταβλητής απόκρισης) και F το πλήθος των συνιστωσών. Μια μεταβλητή j είναι ιδιαίτερα σημαντική, εάν $VIP_j > 1$. Συνήθως το όριο που τίθεται για τη σπουδαιότητα της μεταβλητής είναι ~ 0.7-0.8.[24] Αυτή η ικανότητα της PLS την καθιστά μία ιδιαιτέρως σημαντική μέθοδο καθώς μπορεί να οδηγήσει σε πολύ σημαντικά συμπεράσματα που βοηθούν τον μελετητή να κατανοήσει καλύτερα τα δεδομένα του και αρκετά γρήγορα.

1.4 Περιορισμοί των QSAR μοντέλων

Τα QSAR μοντέλα, καθώς και όλες οι μέθοδοι ανάλυσης δεδομένων. βασίζονται στην υπόθεση της ομοιογένειας των δεδομένων (homogeneity) και της έλλειψης σημαντικών έκτροπων τιμών (outliers). Για τον λόγο αυτό, το σύστημα που μελετάται πρέπει να εμφανίζει παρόμοιες ιδιότητες και ο μηχανισμός, βάσει του οποίου ο πίνακας των περιγραφικών μεταβλητών επιδρά στη μεταβλητή απόκρισης πρέπει να είναι ο ίδιος. Κατά συνέπεια, δεν είναι αποδεκτό να εμφανίζονται σημαντικές έκτροπες τιμές ή μεγάλη ομαδοποίηση. Στην περίπτωση που υπάρχει μεγάλη ομαδοποίηση στα δεδομένα, η εξαγωγή ενός μόνο μοντέλου καθιστά δυσχερή την ανάλυση, για το λόγο ότι ένα τέτοιο μοντέλο περιγράφει μόνο τη συστηματική διακύμανση μεταξύ των ομάδων αλλά είναι αδύνατον να αποσαφηνίσει το τι συμβαίνει μέσα σε κάθε ομάδα. Επίσης, η σαφής εμφάνιση ομάδων στα δεδομένα αθετεί την υπόθεση της ομοιογένειας, διότι αν το σύνολο των δεδομένων ομαδοποιείται με υψηλό βαθμό διαχωρισμού μεταξύ των ομάδων, δεν παρουσιάζει πλέον ομοιογενή κατανομή. Σε αυτές τις περιπτώσεις, συνιστάται η επεξεργασία κάθε ομάδας ανεξάρτητα και η εξαγωγή μοντέλων QSAR για κάθε ομοιογενή ομάδα.[25]

Με την παρουσία έκτροπων τιμών, συνδέεται άμεσα το πεδίο εφαρμογής του μοντέλου. Σαν πεδίο εφαρμογής (Domain of Applicability, DOA) ενός QSAR μοντέλου μπορούμε να θεωρήσουμε το εύρος μέσα στο οποίο το μοντέλο μπορεί να ανεχθεί μία νέα παρατήρηση. Ένας από τους τρόπους για να προσδιοριστεί το εύρος ενός QSAR μοντέλου είναι με βάση τον συντελεστή μόχλευσης (Leverage) που εφαρμόζει μια ένωση στο μοντέλο. Οι τιμές του συντελεστή μόχλευσης είναι δυνατόν να υπολογιστούν για τις ενώσεις της ομάδας εκμάθησης όπως και για νέες ενώσεις. Στην πρώτη περίπτωση, είναι χρήσιμο μέγεθος προκειμένου να εντοπισθούν ενώσεις οι οποίες επηρεάζουν σε σημαντικό βαθμό τις παραμέτρους του μοντέλου και οδηγούν σε ασταθές μοντέλο (δηλαδή, αφαίρεσή τους συνεπάγεται κατάρρευση του μοντέλου) Στη δεύτερη περίπτωση, είναι χρήσιμες για να ελεγχθεί το πεδίο εφαρμογής του μοντέλου. Οι μόνες προβλέψεις που μπορούν να θεωρηθούν αξιόπιστες είναι αυτές που αναφέρονται σε ενώσεις που ανήκουν στο φυσικοχημικό χώρο της ομάδας εκμάθησης.[26]

Ένας ακόμα παράγοντας επίδρασης ενός QSAR μοντέλου είναι το πλήθος των μεταβλητών. Συγκεκριμένα, ορισμένοι descriptors έχουν μεγάλες διαφορές στην τάξη μεγέθους τους, γεγονός που μπορεί να επηρεάσει ένα QSAR μοντέλο, δίνοντας μεγαλύτερη στατιστική σημασία σε έναν descriptor που έχει υψηλότερες τιμές από τους άλλους. Κάτι τέτοιο όμως, μπορεί να μην είναι θεμιτό. Για τον λόγο αυτό, γίνεται κανονικοποίηση των δεδομένων, η οποία πραγματοποιείται κατά τρόπο ώστε να διασφαλισθεί πως όλες οι μεταβλητές έχουν την ίδια πιθανότητα να επηρεάσουν το μοντέλο. Οι συνηθέστερες μέθοδοι κανονικοποίησης είναι η «unit-variance scaling», κατά την οποία η διακύμανση κάθε μεταβλητής τίθεται ίση με 1 και η «min-max scaling» κατά την οποία από κάθε στήλη αφαιρείται η ελάχιστη τιμή και το αποτέλεσμα της αφαίρεσης διαιρείται με την διαφορά της μέγιστης μείον της ελάχιστης τιμής, καταλήγοντας έτσι σε στήλες που όλες οι τιμές είναι μεταξύ του 0 και του 1. Αυτός ο τρόπος προ-επεξεργασίας είναι επίσης χρήσιμος όταν οι μεταβλητές που χρησιμοποιούνται για την εξαγωγή του μοντέλου προέρχονται από διαφορετικές πηγές και εμφανίζουν σε μεγάλο βαθμό διαφορετικό αριθμητικό εύρος.[25]

Κεφάλαιο 2

Μηχανική Μάθηση

Είναι ευρέως αποδεκτό πως η σύγχρονη εποχή χαρακτηρίζεται από την πληθώρα πληροφοριών που είναι διαθέσιμες τόσο στο διαδίκτυο όσο και σε άλλα ηλεκτρονικά μέσα. Η συνθήκη αυτή έχει ως αποτέλεσμα διαθέσιμους όγκους δεδομένων, με πολλές και διαφορετικές παρατηρήσεις για ανάλυση κι επομένως πιο ικανοποιητικά αποτελέσματα. Ωστόσο, για να υλοποιηθούν οι QSAR μεθοδολογίες σε αυτά τα δείγματα είναι απαραίτητη η χρήση των Η/Υ, ώστε να διαχειριστεί ο ερευνητής ευκολότερα το πλήθος των παρατηρήσεων και να εκπαιδεύσει τα προβλεπτικά μοντέλα. Η διαδικασία αυτή ονομάζεται Μηχανική Μάθηση (machine learning) καθώς ο ερευνητής «μαθαίνει» στην «μηχανή» του να αναλύει τα μοτίβα που «κρύβονται» μέσα στα δεδομένα του. Το machine learning αποτελεί βασικό πεδίο της τεχνητής νοημοσύνης, και η κεντρική ιδέα είναι η εκμάθηση της μηχανής μέσα από παραδείγματα. Σκοπός του κεφαλαίου αυτού είναι η παράθεση και ανάλυση των διαδικασιών που πραγματοποιήθηκαν για την υλοποίηση και την ανάπτυξη σχετικά με την επίλυση των προβλημάτων που παρουσιάζονται στο Κεφάλαιο 3. Παρακάτω, γίνεται αναλυτική παρουσίαση, τόσο από την σκοπιά των μαθηματικών όσο και του προγραμματισμού, των μεθόδων που χρησιμοποιήθηκαν.

2.1 Ιστορική αναδρομή

Η μηχανική μάθηση αποτέλεσε κινητήρια δύναμη για την ανάπτυξη της επιστημονικής περιοχής Εξόρυξης Δεδομένων (Data Mining) και αναπόσπαστο τμήμα της. Ιδιαίτερα τα τελευταία χρόνια έχει παρατηρηθεί μία μεταστροφή σε αυτό τον κλάδο της πληροφορικής και των μαθηματικών από πληθώρα ερευνητών που ανήκουν σε διαφορετικά επιστημονικά πεδία. Παρά το γεγονός, όμως, πως τα τελευταία χρόνια υπάρχει αυτή η τάση, τα θεμέλια του τομέα αυτού μπήκαν τον 18ο αιώνα από τον Βρετανό μαθηματικό Thomas Bayes. Ο Bayes πίστεψε πως ήταν δυνατόν να προβλέψει ένα μελλοντικό συμβάν με βάση αντίστοιχα παραδείγματα που έχουν ήδη συμβεί. Στη συνέχεια, αρκετοί μαθηματικοί, συμπεριλαμβανομένου και του Laplace, συνέβαλλαν στην ολοκλήρωση της Μπεϋζιανής στατιστικής.[27]

Οι επόμενοι σταθμοί της ιστορίας του machine learning έρχονται αρκετά αργότερα. Στα τέλη του 19ου αιώνα και μισά του 20ου αιώνα εμφανίζονται για πρώτη φορά οι κυρίαρχες, μέχρι και σήμερα, μέθοδοι της μηχανικής μάθησης:

- Γραμμική Παλινδρόμηση (Linear Regression)[28]
- Νευρωνικά Δίκτυα (Neural Networks)[29]
- Δέντρα Απόφασης (Decision Trees)[30]
- Ταξινομητές «K» πλησιέστερων γειτόνων (K-Nearest Neighbors)[31]

Στη συνέχεια η μεγάλη αύξηση του όγκου των δεδομένων που ξεκίνησε το 1960, σε συνδυασμό με την ραγδαία εξέλιξη της τεχνολογίας οδήγησε στην σταδιακή εξέλιξη του τομέα της μηχανικής μάθησης με την δημιουργία τεχνικών και μεθόδων που αποτελούν ακόμη και σήμερα σε πολλά προβλήματα τις αποδοτικότερες προσεγγίσεις. Οι πιο αντιπροσωπευτικές από αυτές τις μεθόδους είναι οι Γενετικοί Αλγόριθμοι (Genetic Algorithms), οι Μηχανές Διανυσματικής Στήριξης (Support Vector Machines) και η Συσταδοποίηση (Clustering) που εμφανίστηκαν μέχρι το τέλος του 20ου αιώνα.

Αν και οι βασικοί πυλώνες της θεωρίας αυτής θεσπίστηκαν τον 18ο αιώνα, η διαρκής εξέλιξη της πληροφορικής και των μαθηματικών συνεπάγεται την εξέλιξη της θεωρίας. Η μηχανική μάθηση

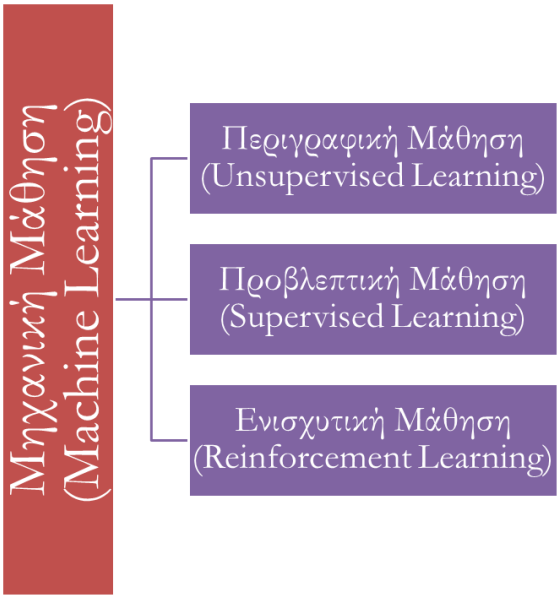
είναι ένα επιστημονικό πεδίο, με τεράστιο πεδίο βελτίωσης και εξέλιξης και για αυτόν τον λόγο ερευνητές από όλους τους επιστημονικούς κλάδους επιλέγουν να ασχοληθούν με αυτόν τον τομέα.

2.2 Κατηγοριοποίηση Αλγορίθμων Μηχανικής Μάθησης

Οι αλγόριθμοι μηχανικής μάθησης έχουν ως στόχο την προσαρμογή ενός (ή περισσοτέρων) μοντέλων στα υπάρχοντα δεδομένα, με τρόπο τέτοιο ώστε να αντιλαμβάνονται τα μοτίβα και οι συσχετίσεις που υπάρχουν σε αυτά, και να μπορούν να εξαχθούν χρήσιμα συμπεράσματα. Ανάλογα με το συμπέρασμα που μπορεί να προκύψει από την εφαρμογή ενός μοντέλου σε ένα σύνολο δεδομένων, τα μοντέλα, που χρησιμοποιούνται, διακρίνονται κυρίως σε δύο μεγάλες κατηγορίες (εικόνα2.1), στα μοντέλα περιγραφικής μάθησης (unsupervised learning) και τα μοντέλα προβλεπτικής μάθησης (supervised learning). Τα μοντέλα προβλεπτικής μάθησης ακολουθούν την γενική εξίσωση 2.1 και εκπαιδεύονται σε ορισμένες μεταβλητές εισόδου (X) και αποκρίσεις (y) με στόχο να «μάθουν» την συσχέτιση τους (f) και να προβλέψουν άγνωστες ή μελλοντικές τιμές σε μεταβλητές ενδιαφέροντος.

$$f(X) = y \quad (2.1)$$

Αντίθετα, τα μοντέλα περιγραφικής μάθησης είναι μία σύνθετη διαδικασία για τον προσδιορισμό κατανοητών σχέσεων ή προτύπων στα δεδομένα. Τέτοιου τύπου μοντέλα δρουν ως μέσο διερεύνησης των ιδιοτήτων των υπό εξέταση δεδομένων και δεν προβλέπουν νέες τιμές. Πέραν των προβλεπτικών και των περιγραφικών μοντέλων, ορισμένα μοντέλα ανήκουν στην τρίτη ευρεία κατηγορία, της ενισχυτικής μάθησης (reinforcement learning). Τέτοιου τύπου μοντέλα μαθαίνουν μία «στρατηγική ενεργειών» μέσα από άμεση αλληλεπίδραση με το περιβάλλον και χρησιμοποιούνται πολύ συχνά σε προβλήματα Σχεδιασμού ή βελτιστοποίησης.[32]

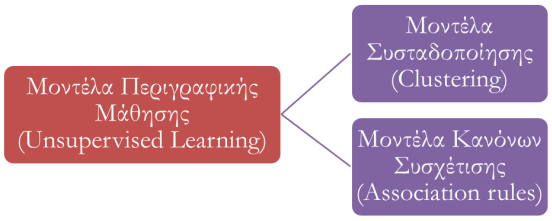


Εικόνα 2.1 **Βασική κατηγοριοποίηση των μεθόδων μηχανικής μάθησης**

2.2.1 Περιγραφική Μάθηση (Unsupervised learning)

Τα μοντέλα περιγραφικής μάθησης διακρίνονται σε μοντέλα συσταδοποίησης ή ομαδοποίησης (clustering) και σε μοντέλα κανόνων συσχέτισης (association rules), όπως φαίνεται και στην εικόνα 2.2. Τα πρώτα επιχειρούν να διαχωρίσουν τα δεδομένα σε άτυπες ομάδες, ενώ τα μοντέλα που

ανήκουν στην δεύτερη κατηγορία διερευνούν τα δεδομένα για την ανακάλυψη κανόνων συσχέτισης που τα χαρακτηρίζουν. Πιο συγκεκριμένα, συσταδοποίηση είναι η διαδικασία του καταμερισμού ενός ετερογενούς πληθυσμού σε ένα σύνολο περισσότερων ομογενών συστάδων (clusters). Η θεμελιώδης διαφορά που έχουν τέτοιου τύπου μοντέλα από τα μοντέλα ταξινόμησης (που ανήκουν στην προβλεπτική μάθηση) είναι πως οι συστάδες είναι άτυπες, δηλαδή δεν είναι προκαθορισμένες κατηγορίες. Αντίθετα, οι μέθοδοι ταξινόμησης εκπαιδεύονται σε ένα σύνολο δεδομένων όπου κάθε στοιχείο έχει ανατεθεί σε μία προκαθορισμένη κατηγορία κι έπειτα, για ένα τυχαίο δείγμα, το μοντέλο μπορεί να αποφανθεί σε ποια από τις κατηγορίες αυτές ανήκει.[33] Τα μοντέλα εξαγωγής κανόνων συσχέτισης παρουσιάζουν από πολλούς ερευνητές ένα μεγάλο ενδιαφέρον καθώς τους παρέχουν πληροφορίες που στοχεύουν στην κατανόηση των δεδομένων τους. Τα μοντέλα αυτά παρέχουν χρήσιμες και εύληπτες πληροφορίες σχετικά με τον τρόπο που αλληλεπιδρούν μεταξύ τους τα δεδομένα. Ο λόγος για τον οποίο η εξαγωγή κανόνων συσχέτισης είναι ευρέως διαδεδομένη είναι πως αυτές οι σχέσεις συνήθως δεν είναι εμφανείς κι επομένως οι ερευνητές εξοικονομούν πολύ χρόνο από την ανάλυση τους.[34]

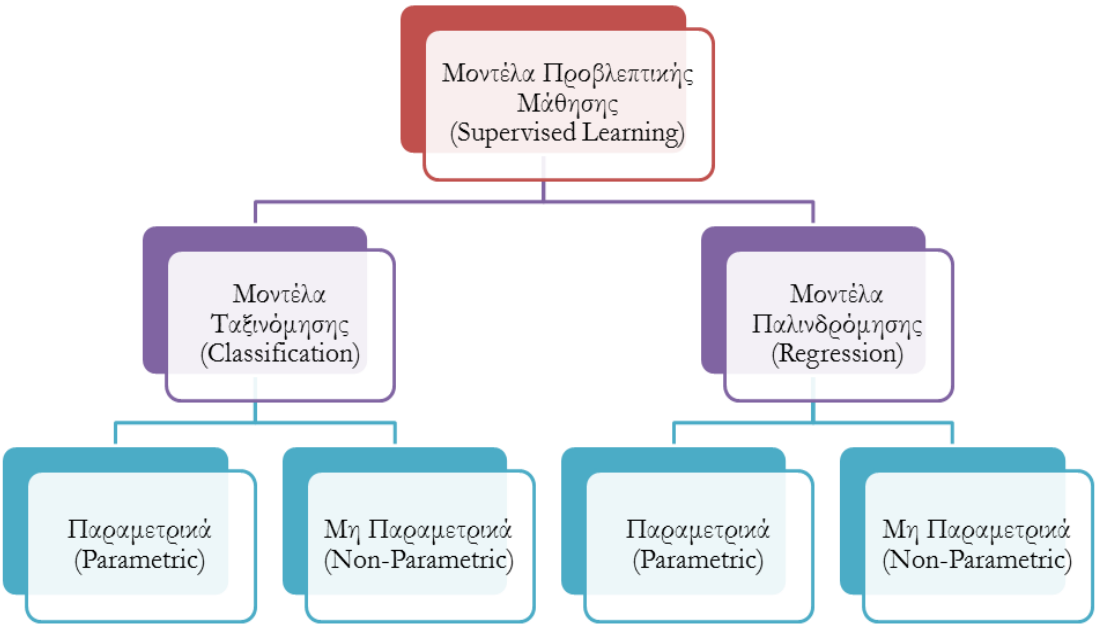


Εικόνα 2.2 Κατηγορίες μοντέλων περιγραφικές μάθησης

2.2.2 Προβλεπτική μάθηση (Supervised learning)

Σύμφωνα με τον Tom M. Mitchell θεωρείται πως η μηχανή «μαθαίνει» όταν από την εμπειρία E, σε σχέση με μια κατηγορία εργασιών T και μια απόδοση P, το σύστημα βελτιώνει την απόδοσή του.[35] Τα μοντέλα προβλεπτικής μάθησης, στα οποία ανήκουν και τα QSAR μοντέλα (Κεφάλαιο 1) διακρίνονται σε δύο κυρίως κατηγορίες ανάλογα το είδος της πρόβλεψης τους, στα μοντέλα ταξινόμησης (classification) και στα μοντέλα παλινδρόμησης (regression). Φυσικά τα μοντέλα και των δύο αυτών κατηγοριών ακολουθούν τον γενικό τύπο της εξίσωσης 2.1, ωστόσο η διαφορά τους είναι πως στην πρώτη κατηγορία η μεταβλητή y είναι μία διακριτή κατηγορική μεταβλητή ενώ στη δεύτερη μία συνεχής μεταβλητή. Για παράδειγμα, ένα μοντέλο το οποίο προβλέπει εάν μία ουσία είναι τοξική για τον άνθρωπο είναι μοντέλο κατηγοριοποίησης (τοξική / όχι τοξική), ενώ ένα μοντέλο το οποίο προβλέπει την συγκέντρωση πάνω από την οποία η ουσία αυτή είναι τοξική για τον άνθρωπο είναι ένα μοντέλο παλινδρόμησης.

Τα μοντέλα προβλεπτικής μάθησης, επίσης, κατηγοριοποιούνται σε παραμετρικά και μη παραμετρικά. Στα παραμετρικά μοντέλα προσδιορίζεται η συσχέτιση ανάμεσα στις μεταβλητές εισόδου X στην μεταβλητή εξόδου y με τη χρήση αλγεβρικών εξισώσεων. Στις εξισώσεις αυτές υπάρχουν παράμετροι, που είναι απροσδιόριστες, και η εκτίμηση τους γίνεται μέσω της εκπαίδευσης του μοντέλου πάνω στα δεδομένα που του δίνονται. Στην περίπτωση αυτή η δομή του μοντέλου θεωρείται εξ' αρχής καθορισμένη. Αντίθετα, τα μη παραμετρικά μοντέλα δεν περιέχουν παραμέτρους και δεν υπάρχει θεωρημένο μοντέλο αλλά καθοδηγείται από τα δεδομένα. Δηλαδή, δεν υπάρχουν εξισώσεις για το μοντέλο αλλά προσαρμόζεται το μοντέλο μέσω των δεδομένων. Χαρακτηριστικά παραδείγματα παραμετρικών μεθόδων είναι η Γραμμική και η Λογιστική Παλινδρόμηση ενώ παραδείγματα μεθόδων μη παραμετρικών αποτελούν τα νευρωνικά δίκτυα, τα δένδρα αποφάσεων και οι γενετικοί αλγόριθμοι. Παρακάτω παρατίθεται μία εικόνα με την κατηγοριοποίηση των μεθόδων της προβλεπτικής μάθησης (εικόνα 2.3).[36]



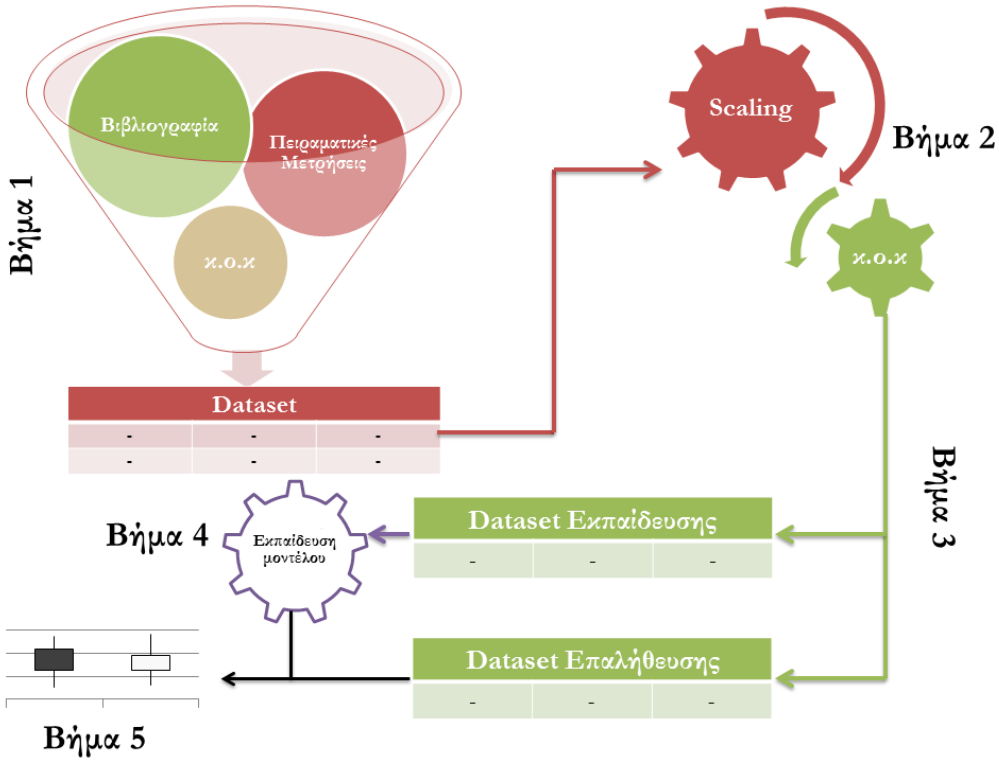
Εικόνα 2.3 **Κατηγορίες μοντέλων προβλεπτικής μάθησης**

2.2.3 Ενισχυτική μάθηση (Reinforcement learning)

Η ενισχυτική μάθηση αποτελείται από ένα σύνολο αλγορίθμων, η φιλοσοφία των οποίων στηρίζεται στον τρόπο που μαθαίνουν τα έμβια όντα. Ο στόχος της ενισχυτικής μάθησης είναι να βρει την βέλτιστη συσχέτιση μεταξύ καταστάσεων και ενεργειών, μαθαίνοντας μέσα από την εξερεύνηση στο χώρο των πιθανών ζευγαριών κατάστασης-ενέργειας. Μια κατάσταση θα πρέπει να είναι ιδιαιτέρως περιγραφική και να περιλαμβάνει όλες τις απαραίτητες πληροφορίες έτσι ώστε, κάθε κατάσταση ενός συστήματος να περιγράφεται με μοναδικό τρόπο κι επομένως η αντίστοιχη ενέργεια να είναι έκδηλη για το σύστημα. Ωστόσο, η προσθήκη παραμέτρων στις καταστάσεις μπορεί να οδηγήσει στη δημιουργία ενός αχανούς χώρου καταστάσεων που δεν είναι εύκολο να χειριστεί από το σύστημα. Από την άλλη πλευρά, η ανεπαρκής πληροφορία μπορεί να «θολώσει» την ικανότητα του συστήματος να μπορεί να διακρίνει τις διάφορες καταστάσεις, οδηγώντας το σε ανεπαρκή ικανότητα λήψης αποφάσεων. Ενέργειες που έχουν καλό αποτέλεσμα σε μια συγκεκριμένη κατάσταση «ανταμείβονται» από το σύστημα ενώ οι ενέργειες με κακό αποτέλεσμα «τιμωρούνται» από αυτό, μέσω του σήματος ανταμοιβής. Ο στόχος της ενισχυτικής μάθησης είναι να ανακαλύψει τα ζευγάρια καταστάσεων-ενεργειών που μεγιστοποιούν την ανταμοιβή του συστήματος. Οι μέθοδοι ενισχυτικής μάθησης χρησιμοποιούνται ευρέως σε βελτιστοποιήσεις αλγορίθμων ή σε ζητήματα Σχεδιασμού.[37]

2.3 Διαδικασία ορθής εκπαίδευσης μοντέλου μηχανικής μάθησης

Η συνήθης διαδικασία που ακολουθείται για την ορθή εκπαίδευση μοντέλων μηχανικής μάθησης φαίνεται στην εικόνα 2.4 και παρακάτω περιγράφονται αναλυτικά τα στάδια της.



Εικόνα 2.4 **Διάγραμμα συνήθους διαδικασίας για την ορθή εκπαίδευση μοντέλου μηχανικής μάθησης**

Βήμα 1) ***Επιλογή Δεδομένων και δημιουργία σετ δεδομένων***

Το πρόβλημα εκπαίδευσης ενός μοντέλου μηχανικής μάθησης μπορεί να αναχθεί σε ένα πρόβλημα προσέγγισης μίας συνάρτησης-στόχου, της συνάρτησης της εξίσωσης (2.1). Σε κάποια μοντέλα (π.χ. λογιστική παλινδρόμηση) η μορφή της συνάρτησης είναι γνωστή, ενώ σε ορισμένα άλλα δεν είναι. Σε κάθε περίπτωση το πεδίο ορισμού αυτής της συνάρτησης είναι ένα σύνολο οντοτήτων σε κάποια δομημένη αναπαράσταση, η οποία αποτελεί το χώρο των στοιχείων του προβλήματος. Η πλέον συνηθισμένη αναπαράσταση είναι αυτή που παρέχει το μοντέλο του διανυσματικού χώρου, το σετ δεδομένων.[38] Σύμφωνα με αυτό το μοντέλο, οι παρατηρήσεις του πληθυσμού αναπαρίστανται ως διανύσματα, των οποίων οι συνιστώσες αναπαριστούν τα χαρακτηριστικά (features) των παρατηρήσεων, που έχουν επιλεγεί για την συγκεκριμένη ανάλυση. Τα χαρακτηριστικά μπορούν να παίρνουν συμβολικές ή αριθμητικές τιμές, ενώ οι τιμές της συνάρτησης-στόχου μπορεί να είναι πρακτικά οτιδήποτε, αριθμητικές ή συμβολικές, διακριτές ή συνεχείς, βαθμωτές ή διανυσματικές, κ.ο.κ. Επομένως, η κατασκευή του σετ δεδομένων αποτελεί ένα ιδιαιτέρως σημαντικό κομμάτι για την δημιουργία ενός μοντέλου μηχανικής. Τα μοντέλα είτε παραμετρικά είτε μη παραμετρικά εκπαιδεύονται για να μαθαίνουν τα μοτίβα και τις ιδιαιτερότητες των δεδομένων, άρα τα στοιχεία ενός σετ δεδομένων και όλα τα δεδομένα του γενικότερα θα πρέπει να μην είναι αμφιβόλου ποιότητας.

Βήμα 2) ***Προεπεξεργασία δεδομένων (Pre-processing)***

Το βήμα αυτό αποτελεί το βραχύ στάδιο της διαδικασίας εκπαίδευσης μοντέλων μηχανικής μάθησης. Συχνά, τα δεδομένα που έρχονται για επεξεργασία έχουν προβληματικές τιμές στα πεδία τους. Τέτοιου τύπου προβληματικές τιμές μπορεί να είναι ελλιπή δεδομένα ή λεκτικά πεδία κ.ο.κ. Η διαδικασία της προεπεξεργασίας

δεν έχει σαφή κανόνες, αλλά επαφίεται στην διακριτική ευχέρεια του ερευνητή. Για παράδειγμα, για να μπορέσει ένα μοντέλο να εκπαιδευτεί θα πρέπει στο σετ δεδομένων να μην υπάρχουν κενές τιμές, ωστόσο ο ερευνητής επιλέγει αν θα κρατήσει τις παρατηρήσεις με τα ελλιπή δεδομένα ή αν τις κρατήσει, με ποια μέθοδος θα καλύψει τα κενά. Εκτός όμως από τις μηχανογραφικές επεμβάσεις που μπορεί να απαιτούνται να γίνουν στο σετ δεδομένων, στο βήμα αυτό συμπεριλαμβάνονται και οι διαδικασίες μετασχηματισμού των δεδομένων με τέτοιο τρόπο, ώστε να διευκολύνουν το μοντέλο να εκπαιδευτεί ορθά. Συνήθης τέτοιες περιπτώσεις είναι:

- Μείωση της διαστατικότητας του σετ δεδομένων,
- Μετατροπή λεκτικών πεδίων σε αριθμητικά με ανάθεση αριθμών στις λεκτικές τιμές (Encoding),
- Κανονικοποίηση των δεδομένων,
- Μετατροπή συνεχών αριθμητικών τιμών σε διακριτές.

Όπως φαίνεται κι από το Βήμα 1 οι ποιότητα των δεδομένων είναι μια ιδιαιτέρως σημαντική παράμετρος από την οποία εξαρτάται η τελική απόδοση του μοντέλου, επομένως το Βήμα αυτό απαιτεί μεγάλη προσοχή από τους ερευνητές καθώς είναι πιθανό με έναν λάθος χειρισμό να καταλήξουν σε παράλογα συμπεράσματα.

Βήμα 3) ***Διαχωρισμός του αρχικού Σετ δεδομένων σε εκπαίδευσης και επαλήθευσης***

Η διαδικασία αυτή, γνωστή ως Train-Test Split, γίνεται ώστε να «δημιουργηθούν» δεδομένα επαλήθευσης του μοντέλου μηχανικής μάθησης που θα εκπαιδευθεί. Πιο συγκεκριμένα, το αρχικό σετ δεδομένων διαχωρίζεται σε δύο μικρότερα, το σετ δεδομένων εκπαίδευσης (συνήθως κοντά στο 80% των παρατηρήσεων) και το σετ δεδομένων επαλήθευσης (συνήθως κοντά στο 20% των παρατηρήσεων) και το μοντέλο εκπαιδεύεται στο σετ δεδομένων εκπαίδευσης. Οι παρατηρήσεις που υπάρχουν στο δεύτερο σετ δεδομένων ήταν «κρυμμένες» από το εκπαιδευμένο μοντέλο, επομένως κάνοντας προβλέψεις σε αυτές τις παρατηρήσεις, και γνωρίζοντας την σωστή απάντηση, ο ερευνητής είναι σε θέση να εκτιμήσει την απόδοση του μοντέλου του. Υπάρχουν διάφοροι μέθοδοι για τον ανωτέρω διαχωρισμό όπως η μέθοδος των Kennard –Stone[39] που στοχεύει στην δημιουργία ενός σετ δεδομένων επαλήθευσης χωρίς σημαντικές έκτροπες τιμές, η μέθοδος τυχαίου train-test split, η μέθοδος του 80-10-10 που διαχωρίζει το αρχικό σετ δεδομένων σε τρία, το train (80% των παρατηρήσεων), το validation (10% των παρατηρήσεων) και το test σετ δεδομένων (10% των παρατηρήσεων), και άλλες. Φυσικά, πρέπει να σημειωθεί πως το βήμα αυτό δεν είναι απαραίτητο, εξαρτάται από τον ερευνητή αν θα επιλέξει να εκτελέσει αυτό το Βήμα.

Βήμα 4) ***Επιλογή και εκπαίδευση μοντέλου μηχανικής μάθησης***

Ανάλογα με το είδος της γνώσης που το μοντέλο της μηχανικής μάθησης θα αναζητήσει επιλέγεται και το κατάλληλο μοντέλο. Πρόκειται για ένα καθαρά υπολογιστικό στάδιο στο οποίο εκπαιδεύεται το μοντέλο και αναγνωρίζει τα μοτίβα και τις συσχετίσεις που υπάρχουν στο εκπαιδευτικό σετ δεδομένων.

Βήμα 5) ***Αξιολόγηση του εκπαιδευμένου μοντέλου***

Στο τελευταίο Βήμα, γίνεται η αξιολόγηση του εκπαιδευμένου μοντέλου μέσω ορισμένων μετρικών τιμών ή μέτρων (metrics). Ειδικότερα, οι μετρικές τιμές υπολογίζονται στο επαληθευτικό σετ για να βοηθήσουν τον ερευνητή να αντιληφθεί τον βαθμό που το μοντέλο έχει μάθει τα δεδομένα. Στην διαδικασία της αξιολόγησης του μοντέλου δεν χρησιμοποιούνται δεδομένα από το εκπαιδευτικό

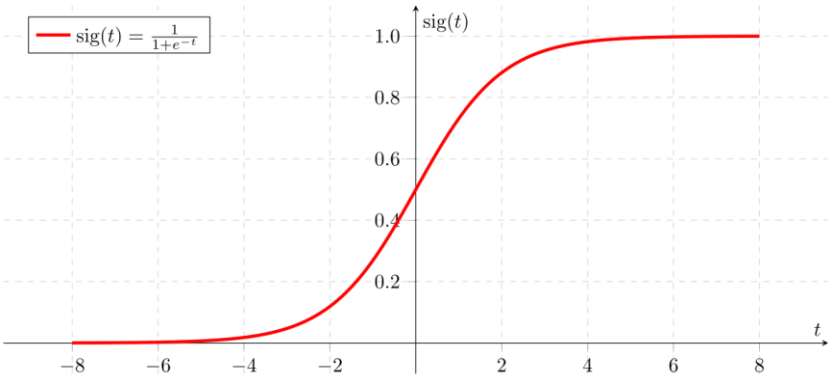
σετ καθώς τα αποτελέσματα θα είναι προσαρμοσμένα στα δεδομένα εκπαίδευσης του μοντέλου και σαν επακόλουθο η αξιολόγησή του δεν θα είναι αξιόπιστη και αντικειμενική. Οι μετρητικές τιμές χρησιμοποιούνται συχνά για να αναγνωρίσουν περιπτώσεις υπέρ-προσαρμογής (overfitting), όπου το μοντέλο έχει υπεραναλύσει τις συσχετίσεις και τα μοτίβα που επικρατούν στο εκπαιδευτικό σετ δεδομένων, χάνοντας έτσι την ουσία. Τέλος, υπάρχει και η περίπτωση το μοντέλο να εμφανίζει υπό-προσαρμογή (underfitting) στο εκπαιδευτικό σετ, γεγονός που θα σημαίνει πως το μοντέλο δεν έχει μάθει να αναγνωρίζει τον τρόπο που εξαρτάται η μεταβλητή απόκρισης από τις περιγραφικές μεταβλητές εισόδου. Το Βήμα 5 μπορεί να ανατροφοδοτήσει την διαδικασία εκπαίδευσης ώστε να εκπαιδευτεί ένα πιο στιβαρό και καλό μοντέλο.

2.4 Μοντέλα Προβλεπτικής Μάθησης (Supervised Learning)

Τα μοντέλα προβλεπτικής μάθησης, ή αλλιώς μάθησης με επίβλεψη (Supervised Learning), είναι μέθοδοι που αποσκοπούν στην πρόβλεψη ιδιοτήτων των στοιχείων. Οι ιδιότητες αυτές ονομάζονται τιμές-στόχος (target) ή τελικό σημείο (endpoint). Παρακάτω γίνεται μία σύντομη περιγραφή και μαθηματική θεμελίωση των μεθόδων που χρησιμοποιήθηκαν στα πλαίσια της παρούσης διπλωματικής εργασίας.

2.4.1 Λογιστική Παλινδρόμηση (Logistic Regression, LR)[40]

Η LR είναι ένα παραμετρικό μοντέλο που χρησιμοποιείται κυρίως σε προβλήματα δυαδικής ταξινόμησης (binary classification), δηλαδή σε προβλήματα που το τελικό σημείο είναι μία διακριτή μεταβλητή με δύο μόνο δυνατές τιμές ή κλάσεις, το «0» και το «1». Συνήθως στα προβλήματα ταξινόμησης κάθε κλάση αντιστοιχεί σε μία συγκεκριμένη συνθήκη, ενώ σε προβλήματα δυαδικής ταξινόμησης ακολουθείται η σύμβαση της θεωρίας των υπολογιστών όπου το «0» αντιστοιχεί στην τιμή «όχι» και το «1» στην τιμή «ναι».



Εικόνα 2.5 **Διάγραμμα σιγμοειδούς συνάρτησης απόφασης.**

Ειδικότερα, το μοντέλο αυτό ανήκει στο μικρό υποσύνολο μεθόδων όπου η συνάρτηση απόφασης είναι γνωστή. Η συνάρτηση απόφασης (decision function) είναι μία σιγμοειδής συνάρτηση, όπως φαίνεται και στο διάγραμμα της εικόνας 2.5, και ο τύπος της, για ένα δυσδιάστατο πρόβλημα, φαίνεται στην ακόλουθη εξίσωση:

$$p(y=1) = \frac{e^{a_0+a_1x}}{1+e^{a_0+a_1x}} = \frac{1}{1+e^{-a_0-a_1x}} \quad (2.2)$$

Η συνάρτηση αυτή υπολογίζει την πιθανότητα ενός στοιχείου με ιδιότητα x να ανήκει στην κλάση «1». Η συνάρτηση p είναι άνω και κάτω φραγμένη μεταξύ του 0 και του 1 και η τιμή της είναι η πιθανότητα μίας παρατήρησης να έχει τελικό σημείο το «ναι». Επομένως, αν για παράδειγμα ένα τέτοιο μοντέλο εξέταζε αν αύριο θα έχει καλό καιρό και έδινε τιμή $prob(x) = 0.85$, τότε η πιθανότητα αύριο να έχει καλό καιρό είναι 0.85 και η πιθανότητα να μην έχει είναι 0.25. Άρα είναι πιο πιθανό να έχει καλό καιρό κι επομένως το τελικό σημείο αυτής της παρατήρησης θα είναι το «1» ή καλύτερα, «Ναι, θα έχει καλό καιρό αύριο». Για να μπορέσει όμως αυτή η συνάρτηση να αποφανθεί για την κλάση στην οποία ανήκει η παρατήρηση θα πρέπει να υπολογισθούν οι a_i παράμετροι, και σε αυτό το σημείο εισχωρεί το κομμάτι της μηχανικής μάθησης.

Σε έναν πολυδιάστατο χώρο η συνάρτηση απόφασης δεν είναι τόσο απλή όσο η 2.2. Έστω ένα χώρος n διαστάσεων, τότε η πιθανότητα p ενός στοιχείου να ανήκει στην κλάση «1» δίνεται από την εξίσωση 2.3α:

$$p(y=1) = p = \frac{e^{a_0+a_1x_1+a_2x_2+\cdots+a_nx_n}}{1+e^{a_0+a_1x_1+a_2x_2+\cdots+a_nx_n}} \rightarrow \qquad (2.3\alpha)$$

$$(1+e^{a_0+a_1x_1+a_2x_2+\cdots+a_nx_n}) \cdot p = e^{a_0+a_1x_1+a_2x_2+\cdots+a_nx_n} \rightarrow \qquad (2.3\beta)$$

$$p = (1-p) \cdot e^{a_0+a_1x_1+a_2x_2+\cdots+a_nx_n} \rightarrow \qquad (2.3\gamma)$$

$$\frac{p}{1-p} = e^{a_0+a_1x_1+a_2x_2+\cdots+a_nx_n} \rightarrow \qquad (2.3\delta)$$

$$ln\left(\frac{p}{1-p}\right) = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n \qquad (2.4)$$

Όπως φαίνεται κι από την εξίσωση 2.3α ο υπολογισμός των a_i παραμέτρων είναι πιο πολύπλοκος κι απαιτείται η χρήση υπολογιστών ώστε να μπορέσει να γίνει. Η λογαριθμική μορφή της συνάρτησης απόφασης στην εξίσωση 2.4 δείχνει πως ο λογάριθμος του λόγου των συμπληρωματικών πιθανοτήτων (log-odds) είναι μία γραμμική συνάρτηση των ιδιοτήτων του στοιχείου. Ειδικότερα, φαίνεται πως κατά την εφαρμογή ενός μοντέλου LR, σχηματίζεται ένα μοντέλο γραμμικής παλινδρόμησης ως προς τα log-odds. Οι παράμετροι στο μοντέλο της λογιστικής παλινδρόμησης υπολογίζονται εφαρμόζοντας τον αλγόριθμο σύγκλισης με ελάττωση των βαθμίδων (Gradient Descent) σε μία συνάρτηση κόστους $J(a_i)$ που ορίζεται ως ο αρνητικός λογάριθμος της συνάρτησης εκτίμησης της μέγιστης πιθανότητας (Maximum Likelihood Estimator). Για τον υπολογισμό των παραμέτρων ελαχιστοποιείται η συνάρτηση κόστους. Η κατασκευή της συνάρτησης κόστους παρουσιάζεται στο Παράρτημα Α. Στην διαδικασία ελαχιστοποίησης προστίθεται ένας παράγοντας κανονικοποίησης (regularization) με στόχο να αποφευχθεί ο κίνδυνος της υπέρ-προσαρμογής (εξισώσεις 2.5α-2.5β). Από μαθηματικής άποψης ο όρος αυτός προστίθεται ώστε να εμποδίσει τις παραμέτρους a_i να προσαρμοστούν τόσο «καλά» στα δεδομένα που το τελικό μοντέλο να είναι υπέρ-προσαρμοσμένο. Υπάρχουν δύο παράγοντες κανονικοποίησης, η $L1$ (εξίσωση 2.5α) και η $L2$ (εξίσωση 2.5β), που χρησιμοποιούνται. Περισσότερη ανάλυση για την εκτίμηση των παραμέτρων ξεφεύγει από τους σκοπούς της παρούσης διπλωματικής εργασίας.

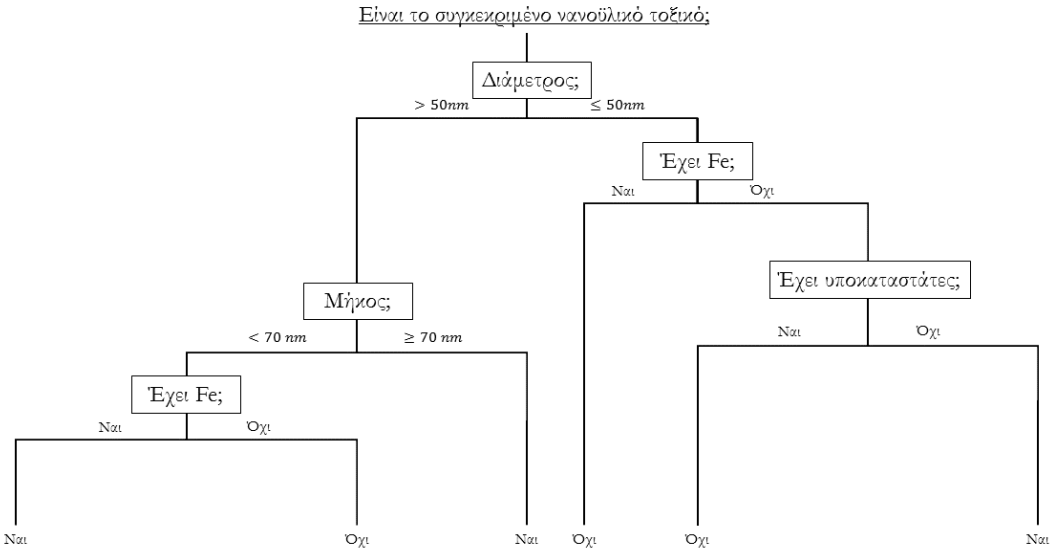
$$\begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} = arg\min_{a_1,\cdots,a_n} J(a_i) + L1 = arg\min_{a_1,\cdots,a_n} J(a_i) + \frac{1}{C}\sum_{i=1}^{n}|a_i| \quad (2.5a)$$

$$\begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} = arg\min_{a_1,\cdots,a_n} J(a_i) + L2 = arg\min_{a_1,\cdots,a_n} J(a_i) + \frac{1}{C}\sum_{i=1}^{n} a_i^2 \quad (2.5\beta)$$

Οι παράμετροι α_i στην πραγματικότητα αντιπροσωπεύουν την σημαντικότητα της κάθε x_i μεταβλητής αλλά και τον τρόπο με τον οποίο επιδρά στην λήψη της απόφασης. Μεταβλητές με παράμετρο να τείνει στο 0 είναι οι πιο ασήμαντες για το μοντέλο, καθώς ο όρος τους στον εκθέτη θα μηδενιστεί και η συνάρτηση της εξίσωσης 2.3α θα πάρει την ίδια απόφαση ανεξάρτητα από την τιμή της μεταβλητής. Επομένως η απόφαση του μοντέλου μένει ανεπηρέαστη από τις μεταβλητές με α_i «κοντά» στο 0, ενώ μεταβλητές των οποίων οι παράμετροι κατά απόλυτη τιμή είναι μακριά από το 0 επηρεάζουν σημαντικά την απόφαση του μοντέλου.

2.4.2 Τυχαίο Δάσος (Random Forest, RF)[40,41]

Το μοντέλου του τυχαίου δάσους, όπως υπονοεί και η ονομασία του, είναι μια ensemble μέθοδος που απαρτίζεται από αρκετά δέντρα απόφασης (Decision Trees). Τα δέντρα απόφασης θεωρούνται από τα πιο διαδεδομένα μη γραμμικά μοντέλα που έχουν εφαρμογή και σε προβλήματα ταξινόμησης και σε προβλήματα παλινδρόμησης. Ένα χαρακτηριστικό, και συνάμα απλό, παράδειγμα ταξινομητή δέντρου απόφασης φαίνεται στην εικόνα 2.6, όπου εξετάζεται ως προς την τοξικότητα του ένα νανοϋλικό. Στην εικόνα αυτή φαίνεται πως ρόλο για τον χαρακτηρισμό των νανοϋλικών του παραδείγματος ως τοξικά, παίζουν το μήκος, η διάμετρος, η ύπαρξη σιδήρου κι η ύπαρξη υποκατάστατών στην επιφάνεια του. Ένα δέντρο απόφασης διαβάζεται από πάνω προς τα κάτω (top-down) και σε κάθε κόμβο (node) ανοίγουν 2 κλαδιά ανάλογα με την απόφαση που παίρνει το μοντέλο στον κόμβο. Όπως φαίνεται και στην εικόνα 2.6, το κριτήριο σε κάθε κόμβο είναι

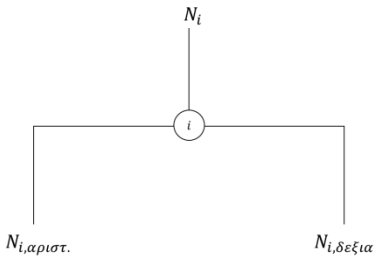


Εικόνα 2.6 Δομή ενός δέντρο απόφασης

τέτοιο ώστε οι απαντήσεις που μπορούν να δοθούν να είναι διακριτές και δύο στο πλήθος. Για αυτό τον λόγο δυαδικές μεταβλητές στο σετ δεδομένων εκπαίδευσης είναι ιδιαιτέρως σημαντικές για τα μοντέλα που ανήκουν στην οικογένεια των δέντρων. Ωστόσο, μεταβλητές που είναι συνεχείς, κατά την εκπαίδευση, διακριτοποιούνται από το μοντέλο. Για παράδειγμα, η διάμετρος και το μήκος είναι

συνεχείς μεταβλητές, ωστόσο το μοντέλο της εικόνας 2.6 εξετάζει μόνο αν η διάμετρος είναι μεγαλύτερη των 50nm ή μικρότερη, κι αν το μήκος είναι μεγαλύτερο ή μικρότερο των 70nm, καθώς κατά την εκπαίδευση του έμαθε πως αυτή η πληροφορία του χρειάζεται για να προβλέψει την τοξικότητα.

Ο τρόπος με τον οποίο ένα δέντρο απόφασης αποφασίζει να δημιουργήσει τα κλαδιά του είναι σύνθετος. Αν θεωρήσουμε ένα σετ δεδομένων εκπαίδευσης n διαστάσεων τότε κάθε κόμβος ενός δέντρου στοχεύει στην δημιουργία τμημάτων ή «κουτιών» στον n-διαστάσεων χώρο που περιλαμβάνουν όσο το δυνατό περισσότερα στοιχεία μιας κλάσης. Ειδικότερα, στόχος του μοντέλου είναι να δημιουργήσει J τμήματα που ελαχιστοποιούν ένα συγκεκριμένο κριτήριο. Η διαδικασία υποδιαίρεσης του αρχικού σετ συνεχίζεται μέχρις ότου κάθε κόμβος απαρτίζεται από στοιχεία μίας μόνο κλάσης. Οι κόμβοι αυτοί ονομάζονται τερματικοί κόμβοι (terminal nodes) ή φύλλα (leaves). Τα κριτήρια που χρησιμοποιούνται είναι διάφορα και ποικίλουν ανάλογα με το είδος του προβλήματος (ταξινόμησης ή παλινδρόμησης), με το είδος του σετ δεδομένων κ.ο.κ. Σε προβλήματα ταξινόμησης τα δύο πιο διαδεδομένα κριτήρια είναι το «Gini Impurity» και η «cross-entropy», ωστόσο στα πλαίσια της παρούσης διπλωματικής εργασίας θα ασχοληθούμε μόνο με το πρώτο.



Εικόνα 2.7 **Παράδειγμα κόμβου i ενός δέντρου απόφασης**

Αν θεωρούμε πως ο n-διαστάσεων χώρος αποτελείται από N στο πλήθος παρατηρήσεις και πως η μεταβλητή στόχος είναι διακριτή με δύο πιθανές κλάσεις τότε, κατά την εκπαίδευση ενός δέντρου, σε έναν κόμβο i «φτάνουν» N_i παρατηρήσεις, ορισμένες από τις οποίες ανήκουν στην κλάση C_1 κι οι υπόλοιπες στην κλάση C_2. Από τον κόμβο αυτό θα δημιουργηθούν δύο υποσύνολα του N_i, τα $N_{i,αριστ.}$ και $N_{i,δεξια}$, όπως φαίνεται και στην εικόνα 2.7. Η τιμή του Gini Impurity για τον κάθε νεοσύστατο κόμβο ορίζεται ως:

$$g(N_{i,αριστ.}) = \sum_{j=1}^{2} \hat{P}(C_j|N_{i,αριστ.})(1 - \hat{P}(C_j|N_{i,αριστ.})) \quad (2.6α)$$

$$g(N_{i,δεξια}) = \sum_{j=1}^{2} \hat{P}(C_j|N_{i,δεξια})(1 - \hat{P}(C_j|N_{i,δεξια})) \quad (2.6β)$$

Στις εξισώσεις 2.6α και 2.6β τα $\hat{P}(C_j|N_{i,αριστ.})$ και $\hat{P}(C_j|N_{i,δεξια})$ ορίζονται ως η συχνότητα της C_j κλάσης στα $N_{i,αριστ.}$ και $N_{i,δεξια}$ υποσύνολα αντίστοιχα και είναι:

$$\hat{P}(C_j|N_{i,k}) = \frac{N_{i,k} \cap C_j}{N_{i,k}} \quad (2.7)$$

Από τις εξισώσεις 2.6α και 2.6β φαίνεται πως στην πραγματικότητα η τιμή του Gini Impurity αντιπροσωπεύει την διακύμανση που υπάρχει στα $N_{i,k}$ υποσύνολα και η τιμή του αυξάνεται όταν παρατηρήσεις υπάρχουν παρατηρήσεις κι από τις δύο κλάσεις στο ίδιο υποσύνολο. Φυσικά, μεγιστοποιείται όταν οι δύο κλάσεις είναι εξίσου κατανεμημένες σε ένα υποσύνολο, ενώ παίρνει την ελάχιστη του τιμή όταν το υποσύνολο αποτελείται από παρατηρήσεις που ανήκουν όλες στην ίδια κλάση, ή καλύτερα, όταν ο κόμβος που καταλήγει είναι τερματικός.

Το κριτήριο Gini έχει το πλεονέκτημα πως επιτρέπει τον υπολογισμό της σημαντικότητας κάθε μεταβλητής του σετ εκπαίδευσης, δίνοντας έτσι μία πολύ καλή κατανόηση των συσχετισμών

που επικρατούν στο σετ δεδομένων. Ειδικότερα, ορίζεται η τιμή Gini Index (G_i) για τον κόμβο N_i, που συμβολίζει την σημαντικότητα του κόμβου στην λήψη της απόφασης:

$$G_i = \hat{P}(N_{i,αριστ.})g(N_{i,αριστ.}) + \hat{P}(N_{i,δεξια})g(N_{i,δεξια}) \quad (2.8)$$

όπου $\hat{P}(N_{i,αριστ.})$ και $\hat{P}(N_{i,δεξια})$ η αναλογία των $N_{i,αριστ.}$ και $N_{i,δεξια}$ στοιχείων στο N_i σύνολο αντίστοιχα, και είναι:

$$\hat{P}(N_{i,k}) = \frac{N_{i,k}}{N_i} \quad (2.9)$$

Η σημαντικότητα της κάθε μεταβλητής, ωστόσο, υπολογίζεται από την εξίσωση 2.10, σύμφωνα με την οποία η σημαντικότητα f_i είναι ίση με το πηλίκο του αθροίσματος των G_i τιμών των κόμβων που διαχωρίζονται με βάση την μεταβλητή, προς το άθροισμα όλων των G_i τιμών.

$$f_i = \frac{\sum_{k:κόμβοι\ που\ διαχωρίζονται\ με\ βάση\ την\ μεταβλητή\ i} G_{i_k}}{\sum_j^J G_{i_j}} \quad (2.10)$$

Συνήθως χρησιμοποιείται η μορφή των κανονικοποιημένων f_i, όπως αυτή ορίζεται από την εξίσωση 2.11:

$$normf_i = \frac{f_i}{\sum_k^n f_i} \quad (2.11)$$

Όπως ειπώθηκε και στην αρχή της παραγράφου ένα τυχαίο δάσος αποτελείται από ένα σύνολο δέντρων αποφάσεων. Πιο συγκεκριμένα, το τυχαίο δάσος εκπαιδεύει ένα σύνολο δέντρων αποφάσεων με την ιδιαιτερότητα πως κάθε δέντρο εκπαιδεύεται σε ένα τυχαίο υποσύνολο των αρχικών μεταβλητών. Επομένως, κάθε οι διαστάσεις στις οποίες εκπαιδεύεται κάθε δέντρο είναι μικρότερες από την αρχική διάσταση n του σετ δεδομένων. Αυτό έχει ως αποτέλεσμα να εκπαιδεύεται ένα σύνολο ταξινομητών χαμηλότερης απόδοσης, και όταν το τυχαίο δάσος εφαρμόζεται, κάθε ένα από τα δέντρα να κάνει μια πρόβλεψη για την κλάση του στοιχείου προς πρόβλεψη. Τελικά, η πρόβλεψη που δίνει το τυχαίο δάσος είναι η πλειοψηφική των δέντρων που το απαρτίζουν. Ο υπολογισμός της σημαντικότητας της κάθε μεταβλητής σε ένα τυχαίο δάσος γίνεται μέσω του πηλίκου του αθροίσματος των f_i κάθε δέντρου προς το σύνολο των δέντρων (T) (εξίσωση 2.12):

$$Random\ Forest\ Importance\ of\ i = \frac{\sum_{j\ \epsilon\ σε\ όλα\ τα\ δέντρα} normf_i}{T} \quad (2.12)$$

2.4.3 Αφελής Bayes (Naïve Bayes, NB)[41]

Η αφελής Bayes είναι μία μέθοδος ταξινόμησης που στηρίζεται, όπως είναι το αναμενόμενο, στο θεώρημα Bayes, με την «αφελή» θεώρηση πως όλες οι μεταβλητές του σετ δεδομένων είναι μεταξύ τους ανεξάρτητες. Θεωρώντας ένα διάνυσμα X με συνιστώσες τις τιμές των μεταβλητών μίας παρατήρησης και μία διακριτή μεταβλητή C_k, που συμβολίζει την κλάση της παρατήρησης, το θεώρημα του Bayes δηλώνει πως:

$$P(C_k|X) = \frac{P(X|C_k)P(C_k)}{P(X)} \quad (2.13)$$

όπου k ή κλάση, $P(C_k|X)$ η πιθανότητα η παρατήρηση με διάνυσμα μεταβλητών X να ανήκει στην κλάση k, $P(X|C_k)$ η πιθανότητα ένα στοιχείο που ανήκει στην κλάση k να έχει διάνυσμα

μεταβλητών X, $P(C_k)$ η πιθανότητα της κλάσης k και $P(X)$ η πιθανότητα ύπαρξης παρατήρησης με διάνυσμα μεταβλητών X. Με χρήση του κανόνα αλυσίδας η $P(X|C_k)$ μπορεί να γραφεί και ως:

$$P(X|C_k) = P(x_1|C_k, x_2, x_3, \dots, x_n)P(x_2|C_k, x_3, x_4, \dots, x_n) \cdots P(x_{n-1}|C_k, x_n)P(x_n|C_k) \quad (2.14)$$

Η παραδοχή της ανεξαρτησίας μεταξύ των μεταβλητών έχει ως αποτέλεσμα κάθε όρος της εξίσωσης 2.14 να μπορεί να γραφτεί με την μορφή της εξίσωσης 2.15α και τελικά η $P(X|C_k)$ δίνεται από την εξίσωση 2.15β:

$$P(x_i|C_k, x_{i+1}, x_{i+2}, \dots, x_n) = P(x_i|C_k) \quad (2.15a)$$

$$P(X|C_k) = \prod_{i=1}^{n} P(x_i|C_k) \quad (2.15\beta)$$

Επομένως, η εξίσωση 2.13 μπορεί να γραφεί με την μορφή της εξίσωσης 2.16α κι επειδή η $P(X)$ είναι σταθερή από τα δεδομένα εισόδου, η εξίσωση 2.16α μπορεί να εκφραστεί με την μορφή της εξίσωσης 2.16β:

$$P(C_k|X) = \frac{P(C_k)\prod_{i=1}^{n} P(x_i|C_k)}{P(X)} \quad (2.16\alpha)$$

$$P(C_k|X) \propto P(C_k)\prod_{i=1}^{n} P(x_i|C_k) \quad (2.16\beta)$$

Η εξίσωση 2.16β δείχνει πως η πιθανότητα ενός στοιχείου με δεδομένα X να ανήκει σε κάποια κλάση είναι ανάλογη της ποσότητας $P(C_k)\prod_{i=1}^{n} P(x_i|C_k)$ και η κλάση που θα επιλέξει το μοντέλο είναι η πιο πιθανή, δηλαδή η κλάση για την οποία η ποσότητα αυτή θα μεγιστοποιεί την πιθανότητα $P(C_k|X)$. Δεδομένου ότι τα x_i είναι γνωστά από τα δεδομένα εκπαίδευσης, η ανωτέρω πρόταση μπορεί να γραφεί με την μορφή εξίσωσης ως εξής (εξίσωση 2.17):

$$\hat{C} = \underset{C_k}{argmax}\, P(C_k)\prod_{i=1}^{n} P(x_i|C_k) \quad (2.17)$$

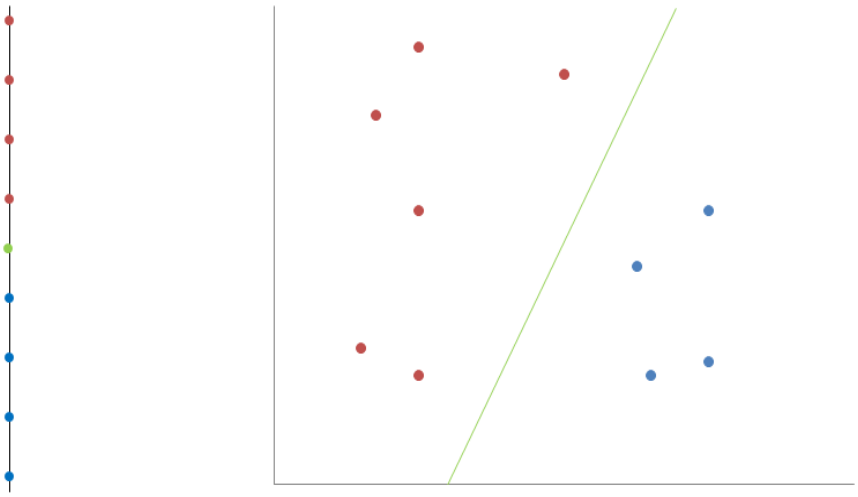
όπου $\hat{C}$ η κλάση που καταλήγει ο ταξινομητής. Από την εξίσωση 2.17, η $P(C_k)$ μπορεί να υπολογισθεί ως η σχετική συχνότητα της κλάσης C_k στα δεδομένα εκπαίδευσης του μοντέλου, και μένει να προσδιορισθεί η $P(x_i|C_k)$. Η αφελής Bayes αποτελεί στην πραγματικότητα ένα σύνολο μεθόδων (Gaussian Naive Bayes, Bernoulli Naive Bayes, Multinomial Naive Bayes κ.ο.κ.), κάθε μία από τις οποίες διακρίνεται για τον τρόπο που προσεγγίζει την πιθανότητα $P(x_i|C_k)$. Στην παρούσα διπλωματική εργασία χρησιμοποιήθηκε η Gaussian Naive Bayes, η οποία υποθέτει πως οι μεταβλητές x_i ακολουθούν κανονική κατανομή, με κέντρο το μ και διακύμανση σ και προσδιορίζει την συνάρτηση πυκνότητας πιθανότητας $P(x_i|C_k)$ σύμφωνα με την εξίσωση 2.18:

$$P(x_i|C_k) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2} \quad (2.18)$$

2.4.4 Μηχανές Διανυσμάτων Στήριξης(Support Vector Machines, SVM)[42]

Η μέθοδος των μηχανών διανυσμάτων στήριξης είναι μια γενικευμένη μορφή του απλού ταξινομητή μέγιστου περιθωρίου (maximal margin classifier) για μη γραμμικά όρια. Θεωρώντας έναν διανυσματικό χώρο n-διαστάσεων, όπου n είναι ο αριθμός των μεταβλητών του εκπαιδευτικού σετ, ο

ταξινομητής μέγιστου περιθωρίου επιτυγχάνει δυαδική ταξινόμηση δημιουργώντας μία «υπερ-επιφάνεια» (hyperplane) $n-1$ διαστάσεων στον χώρο και ανάλογα με την θέση των στοιχείων σε σχέση με την υπέρ-επιφάνεια αυτή. Για παράδειγμα, αν όλα τα στοιχεία του σετ δεδομένων



Εικόνα 2.8 Παραδείγματα τοποθέτησης των υπέρ-επιφανειών. Τα χρώματα των σημείων ξεχωρίζουν ανάλογα με τις κλάσεις, ενώ με πράσινο χρώμα συμβολίζεται η υπέρ-επιφάνεια. Αριστερή: Τα στοιχεία του σετ δεδομένων διαθέτουν μία μόνο μεταβλητή ($n=1$) άρα η υπέρ-επιφάνεια θα είναι ένα σημείο. Δεξιά: Τα στοιχεία του σετ δεδομένων διαθέτουν δύο μεταβλητές ($n=2$) άρα η υπέρ-επιφάνεια θα είναι μία ευθεία.

εκπαίδευσης είναι στοιχεία μιας γραμμής (1-διάσταση) τότε η υπέρ-επιφάνεια είναι ένα «σημείο-κατώφλι» (καμία διάσταση) της γραμμής αυτής, αν τα σημεία είναι απλωμένα σε έναν δυσδιάστατο χώρο τότε η υπέρ-επιφάνεια είναι μια γραμμή (1 διάσταση), ενώ αν τα σημεία απλώνονται σε έναν χώρο τριών διαστάσεων η υπέρ-επιφάνεια είναι μια επιφάνεια δύο διαστάσεων (εικόνα 2.8).

Για την μαθηματική ανάλυση της SVM, θα θεωρήσουμε πως το σετ εκπαίδευσης είναι ένας $m \times n$ πίνακας X που αποτελείται από m σε πλήθος στοιχεία στον n-διαστάσεων χώρο,

$$X_1 = \begin{pmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1n} \end{pmatrix}, \cdots, X_m = \begin{pmatrix} x_{m1} \\ x_{m2} \\ \vdots \\ x_{mn} \end{pmatrix} \quad (2.19)$$

κάθε ένα από τα οποία διαθέτει μια τιμή-στόχο $y_i \epsilon \{-1,1\}$, όπου η τιμή -1 αντιπροσωπεύει την μία κλάση και η τιμή 1 την άλλη. Θεωρούμε ακόμα πως υπάρχει τουλάχιστον μία υπέρ-επιφάνεια που μπορεί να διαχωρίσει τα στοιχεία της κλάσης -1 από τα στοιχεία της κλάσης 1. Η εξίσωση της υπέρ-επιφάνειας θα είναι της μορφής της εξίσωσης 2.20α και θα έχει τις ιδιότητες των εξισώσεων 2.20β για τα στοιχεία με $y_i = 1$ και 2.20γ για τα στοιχεία με $y_i = -1$:

$$f(X_i) = A \cdot X_i = \alpha_0 + a_1 x_{i1} + a_2 x_{i2} + \cdots + a_n x_{in} = 0 \quad (2.20a)$$

$$\alpha_0 + a_1 x_{i1} + a_2 x_{i2} + \cdots + a_n x_{in} > 0 \quad (2.20\beta)$$

$$\alpha_0 + a_1 x_{i1} + a_2 x_{i2} + \cdots + a_n x_{in} < 0 \quad (2.20\gamma)$$

Ισοδύναμα, η υπέρ-επιφάνεια έχει την ιδιότητα της εξίσωσης 2.21 για όλα τα $i = 1,2,\ldots,m$:

$$y_i(\alpha_o + a_1 x_{i1} + a_2 x_{i2} + \cdots + a_n x_{in}) > 0 \quad (2.21)$$

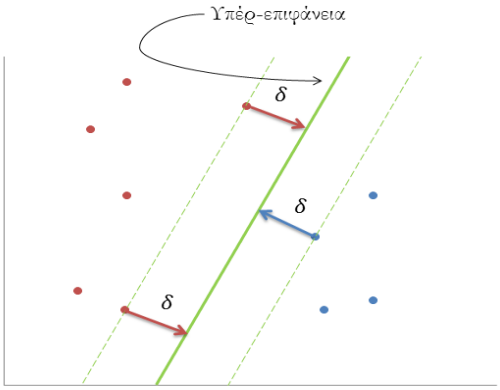
Η κλάση y^*, για ένα στοιχείο επαλήθευσης $X^* = (x_1^*, x_2^*, \dots, x_n^*)^T$, δίνεται με την εφαρμογή των εξισώσεων 2.20α και 2.21:

$$f(X^*) = a_0 + a_1 x_1^* + a_2 x_2^* + \cdots + a_n x_n^* \qquad (2.22\alpha)$$

$$y^* f(X^*) > 0 \rightarrow \begin{cases} y^* < 0 \rightarrow y^* = -1, & για\ f(X^*) < 0 \\ y^* > 0 \rightarrow y^* = 1, & για\ f(X^*) > 0 \end{cases} \qquad (2.22\beta)$$

Επομένως, η ταξινόμηση ενός στοιχείου εξαρτάται απόλυτα από την εξίσωση της υπέρ-επιφάνειας κι άρα, η εκπαίδευση του μοντέλου στην ουσία είναι η διαδικασία εύρεσης της εξίσωσης υπέρ-επιφάνειας. Αν υπάρχει, όμως, τουλάχιστον μια υπέρ-επιφάνεια που να διαχωρίζει τα δεδομένα με βάση την κλάση τους, τότε υπάρχουν άπειρες καθώς μία ελάχιστη μετακίνηση αυτής της επιφάνειας δημιουργεί μια διαφορετική. Για τον λόγο αυτό, κατά την εκπαίδευση του μοντέλου αναζητείται η βέλτιστη υπέρ-επιφάνεια, αυτή δηλαδή που θα οδηγήσει στο ακριβέστερο μοντέλο.

Η διαδικασία εύρεσης της βέλτιστης υπέρ-επιφάνειας είναι αρκετά σύνθετη και ξεφεύγει από τα πλαίσια της διπλωματικής αυτής εργασίας και για τον λόγο αυτό θα παρουσιαστούν συνοπτικά ορισμένα σημεία-κλειδιά της ανάλυσης που θα χρειαστούν παρακάτω. Πρώτο βήμα της διαδικασίας είναι ο υπολογισμός όλων των κάθετων αποστάσεων των στοιχείων από την υπέρ-επιφάνεια και στην συνέχεια επιλέγονται τα στοιχεία της κάθε κλάσης τα οποία έχουν την μικρότερη απόσταση δ, όπως φαίνεται και στην εικόνα 2.9. Η βέλτιστη υπέρ-επιφάνεια θεωρείται αυτή για την οποία οι αποστάσεις δ γίνονται μέγιστες. Στην διαδικασία αυτή προστίθεται άλλη μία συνθήκη που επιτρέπει στο μοντέλο να κάνει λάθη στην ταξινόμηση των δεδομένων του εκπαιδευτικού σετ. Αυτό συμβαίνει διότι συνήθως τα δεδομένα δεν είναι πλήρως διαχωρίσιμα από μία υπέρ-επιφάνεια και ώστε να αποφευχθεί ο κίνδυνος της υπέρ-προσαρμογής. Πιο συγκεκριμένα, οι έκτροπες τιμές που υπάρχουν στα δεδομένα εκπαίδευσης τείνουν να δημιουργήσουν μια υπέρ-επιφάνεια της οποίας οι δ αποστάσεις είναι ιδιαιτέρως μικρές με αποτέλεσμα το μοντέλο να μην έχει ευρεία εφαρμογή εκτός του συγκεκριμένου σετ δεδομένων. Τα παραπάνω βήματα συμπυκνώνονται στις εξισώσεις 2.23α – 2.23γ



Εικόνα 2.9 **Τρόπος επιλογής υπέρ-επιφάνειας**

$$A = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = \underset{\alpha_0, \alpha_1, \dots, \alpha_n}{argmax}\, \delta \qquad (2.23\alpha)$$

$$y_i(\alpha_o + a_1 x_{i1} + a_2 x_{i2} + \cdots + a_n x_{in}) > \delta(1 - \varepsilon_i) \qquad (2.23\beta)$$

$$\varepsilon_i \geq 0, \sum_{i=1}^{m} \varepsilon_i \leq C \qquad (2.23\gamma)$$

όπου C μια θετική ρυθμιστική παράμετρος και ε_i οι χαλαρές μεταβλητές (slack variables) που επιτρέπουν μεμονωμένες παρατηρήσεις να βρίσκονται στην λάθος πλευρά της υπέρ-επιφάνειας.

Η ρυθμιστική παράμετρος C ορίζεται ως το άθροισμα των χαλαρών μεταβλητών και λειτουργεί ως ρυθμιστής της ανταλλαγής σφάλματος-διακύμανσης (bias-variance trade off). Όταν η παράμετρος αυτή έχει χαμηλή τιμή τότε το μοντέλο θα προσαρμοστεί πολύ καλά στα δεδομένα,

επομένως θα έχει λίγα σφάλματα στο σετ εκπαίδευσης (χαμηλό bias) αλλά υψηλή διακύμανση (variance). Αντίθετα, όταν η παράμετρος C έχει υψηλή τιμή τότε επιτρέπονται στο μοντέλο περισσότερες «παραβάσεις» κι επομένως θα έχει περισσότερα σφάλματα αλλά χαμηλότερη διακύμανση.

Η βελτιστοποίηση των εξισώσεων 2.23α- 2.23γ έχει μια πολύ ενδιαφέρουσα ιδιότητα. Σύμφωνα με τις εξισώσεις αυτές, μόνο τα στοιχεία που είτε βρίσκονται μέσα στην περιοχή 2δ (εικόνα 2.9) είτε βρίσκονται στην λάθος πλευρά της υπέρ-επιφάνειας επηρεάζουν την εξίσωσης της κι επομένως τον ταξινομητή. Για τον λόγο αυτό, οι παρατηρήσεις αυτές ονομάζονται διανύσματα στήριξης (support vectors) καθώς επί της ουσίας «στηρίζουν» την υπέρ-επιφάνεια ενώ οι υπόλοιπες παρατηρήσεις δεν έχουν καμία επιρροή στην εξίσωση της. Το γεγονός πως τα διανύσματα στήριξης είναι λιγότερα από το σύνολο των παρατηρήσεων εκτός από υπολογιστική ευχέρεια καθιστά και το μοντέλο πιο αποτελεσματικό σε παρατηρήσεις που βρίσκονται μακριά από την υπέρ-επιφάνεια.

Η λύση των εξισώσεων 2.23α- 2.23γ οδηγεί στην εξίσωση 2.24α στην οποία υπάρχει ένας όρος με εσωτερικό γινόμενο μεταξύ του στοιχείου επαλήθευσης X^* και όλων των παρατηρήσεων επαλήθευσης. Δεδομένου όμως ότι μόνο τα διανύσματα στήριξης συμβάλλουν στην εκπαίδευση του μοντέλου, η εξίσωση 2.24α μπορεί να γραφεί στην μορφή της εξίσωσης 2.24β:

$$f(X^*) = a_0 + \sum_{i=1}^{m} a_i \langle x_i^*, x_i \rangle \quad (2.24a)$$

$$f(X^*) = a_0 + \sum_{i \,\epsilon\, S} a_i \langle x_i^*, x_i \rangle \quad (2.24\beta)$$

όπου S το σύνολο των διανυσμάτων στήριξης. Η παραπάνω ανάλυση αφορά το μοντέλο ενός απλού ταξινομητή μέγιστου περιθωρίου όπου υπάρχουν γραμμικά όρια για τον διαχωρισμό των παρατηρήσεων. Όπως αναφέρθηκε και στην αρχή αυτής της παραγράφου, η SVM αποτελεί μία γενίκευση αυτού του μοντέλου ώστε να μπορεί να εφαρμοστεί και σε προβλήματα όπου οι παρατηρήσεις δεν διαχωρίζονται με γραμμικό τρόπο. Αυτό επιτυγχάνεται με αντικατάσταση του όρου με το εσωτερικό γινόμενο $\langle x_i^*, x_i \rangle$ με μια γενικευμένη μορφή του εσωτερικού γινομένου (εξίσωση 2.25):

$$K(x_i^*, x_i) \quad (2.25)$$

όπου K μία συνάρτηση που ανήκει στην κατηγορία των kernel συναρτήσεων. Ορισμένες από τις πιο διαδεδομένες kernel συναρτήσεις που χρησιμοποιούνται στην SVM είναι:

- Γραμμική kernel: $K(x_i^*, x_i) = \sum_{j=1}^{m} x_{ij}^* x_{ij} \quad (2.26a)$
- Πολυώνυμικη kernel: $K(x_i^*, x_i) = \left(1 + \sum_{j=1}^{m} x_{ij}^* x_{ij}\right)^d \quad (2.26\beta)$
- Γκαουσιανή kernel (RBF): $K(x_i^*, x_i) = exp\left(-\gamma \sum_{j=1}^{m} \left(x_{ij}^* - x_{ij}\right)^2\right) \quad (2.26\gamma)$

2.5 Μοντέλα Περιγραφικής Μάθησης (Unsupervised Learning)

Τα μοντέλα περιγραφικής μάθησης, ή αλλιώς μάθησης χωρίς επίβλεψη (Unsupervised Learning), είναι μέθοδοι που αποσκοπούν στην καλύτερη κατανόηση των δεδομένων. Παρακάτω γίνεται μία υποτυπώδης μαθηματική θεμελίωση των δύο μεθόδων που χρησιμοποιήθηκαν στα πλαίσια της παρούσης διπλωματικής εργασίας.

2.5.1 Ανάλυση Κυρίων Συνιστωσών (Principal Component Analysis, PCA)[40]

Η PCA όπως προαναφέρθηκε και στο Κεφάλαιο 1 είναι μια μέθοδος που στόχος της είναι να μετασχηματίσει τα αρχικά δεδομένα. Η χρήση της είναι ευρεία καθώς ένας τέτοιος μετασχηματισμός μπορεί να καταλήξει σε ένα σετ δεδομένων λιγότερων διαστάσεων, δίνοντας έτσι την ευχέρεια στη μηχανή να κάνει πιο σύνθετους υπολογισμούς, αλλά και σε ένα σετ δεδομένων όπου οι μεταβλητές του είναι ανεξάρτητες μεταξύ τους. Η γενική ιδέα πίσω από την PCA είναι από ένα σετ δεδομένων με m, στο πλήθος, παρατηρήσεις και n στο πλήθος μεταβλητές να καταλήξει σε ένα νέο σετ $m \times k$ διαστάσεων, όπου k το πλήθος των νέων μεταβλητών, με $k \leq n$, κάθε μια από τις οποίες είναι ένας γραμμικός συνδυασμός των n μεταβλητών. Αυτό επιτυγχάνεται με την ακόλουθη διαδικασία.

Ας θεωρήσουμε και πάλι τον πίνακα δεδομένων X, όπως αυτός περιγράφεται από την εξίσωση 2.19. Πρώτο βήμα της διαδικασίας είναι να υπολογισθεί ο πίνακας συνδιακύμανσης C (covariance matrix) που περιγράφει τις συσχετίσεις μεταξύ των μεταβλητών και έχει την μορφή της εξίσωσης 2.27α ενώ ο υπολογισμός του γίνεται από τις εξισώσεις 2.27β και 2.27γ:

$$C = \begin{pmatrix} var(x_1) & cov(x_1, x_2) & \cdots & cov(x_1, x_n) \\ cov(x_2, x_1) & var(x_2) & \dots & cov(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ cov(x_n, x_1) & cov(x_2, x_n) & \dots & var(x_n) \end{pmatrix} \quad (2.27a)$$

$$var(x_i) = \frac{1}{m-1}\sum_{j=1}^{m}\left(x_{ij} - \overline{x_i}\right)^2 \quad (2.27\beta)$$

$$cov(x_i, x_i') = \frac{1}{m-1}\sum_{j=1}^{m}\left(x_{ij} - \overline{x_i}\right)\left(x_{ij}' - \overline{x_i'}\right) \quad (2.27\gamma)$$

όπου x_i μία μεταβλητή του πίνακα X, $var(x_i)$ η διακύμανση της μεταβλητής, $cov(x_i, x_i')$ η συνδιακύμανση δύο μεταβλητών του πίνακα X, x_{ij} η τιμή μιας μεταβλητής x_i για το στοιχείο j και $\overline{x_i}$ η μέση τιμή της μεταβλητής. Ο πίνακας C είναι ένας τετραγωνικός συμμετρικός πίνακας, τα στοιχεία της διαγωνίου του οποίου είναι οι διακυμάνσεις των μεταβλητών του X και περιγράφει τον τρόπο με τον οποίο κάθε μεταβλητή επηρεάζει τις υπόλοιπες. Η συμμετρία του C μπορεί εύκολα να γίνει αντιληπτή από την εξίσωση 2.27γ.

Επόμενο βήμα της διαδικασίας είναι ο υπολογισμός των ιδιοτιμών του C. Κάθε ιδιοτιμή λ_i αφορά την μία στήλη του πίνακα C και υπολογίζεται από τις εξισώσεις 2.28α-2.28β:

$$det(C - \lambda I) = 0 \rightarrow \quad (2.28a)$$

$$det\begin{pmatrix} var(x_1) - \lambda_1 & cov(x_1, x_2) & \cdots & cov(x_1, x_n) \\ cov(x_2, x_1) & var(x_2) - \lambda_2 & \dots & cov(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ cov(x_n, x_1) & cov(x_2, x_n) & \dots & var(x_n) - \lambda_n \end{pmatrix} = 0 \quad (2.28\beta)$$

όπου I ο μοναδιαίος πίνακας. Η εξίσωση 2.28β καταλήγει σε μία πολυωνυμική έκφραση με n ρίζες, που είναι και οι ιδιοτιμές του C. Ο μετασχηματισμός της PCA γίνεται με τον πολλαπλασιασμό του αρχικού πίνακα X με τον πίνακα των ιδιοδιανυσμάτων του C. Οι συνιστώσες των ιδιοδιανυσμάτων $e_i = (e_{i1}, e_{i2}, \cdots, e_{in})^T$ υπολογίζονται από τις εξισώσεις 2.29α-2.29γ:

$$C \cdot e_i = \lambda_i \cdot e_i \rightarrow \quad (2.29\alpha)$$

$$\begin{pmatrix} var(x_1) & \cdots & cov(x_1, x_n) \\ \vdots & \ddots & \vdots \\ cov(x_n, x_1) & \cdots & var(x_n) \end{pmatrix} \begin{pmatrix} e_{i1} \\ \vdots \\ e_{in} \end{pmatrix} = \lambda_i \begin{pmatrix} e_{i1} \\ \vdots \\ e_{in} \end{pmatrix} \rightarrow \quad (2.29\beta)$$

$$\begin{cases} e_{i1}var(x_1) + e_{i2}cov(x_1, x_2) + \cdots + cov(x_1, x_n) = \lambda_i e_{i1} \\ e_{i1}cov(x_2, x_1) + e_{i2}var(x_2) + \cdots + cov(x_2, x_n) = \lambda_i e_{i2} \\ \vdots \\ e_{i1}cov(x_n, x_1) + cov(x_2, x_n) + \cdots + var(x_n) = \lambda_i e_{in} \end{cases} \quad (2.29\gamma)$$

Το σύστημα της εξίσωσης 2.29γ δεν λύνεται με τις γνωστές διαδικασίες καθώς πάντα θα καταλήγει σε μία ταυτοτική σχέση. Για τον λόγο αυτό, υπολογίζονται από το σύστημα όλες οι συνιστώσες συναρτήσει μίας συγκεκριμένης, η οποία τίθεται ίση με την μονάδα καταλήγοντας έτσι σε μία λύση της μορφής της εξίσωσης 2.30α. Κατόπιν υπολογίζεται το μέτρο του διανύσματος e_iκαι το τελικό ιδιοδιάνυσμα e_i^{final} ορίζεται όπως στην εξίσωση 2.30β. Η διαδικασία αυτή (εξισώσεις 2.29α-2.30β) επαναλαμβάνεται για όλα τα λ_i.

$$e_i = \begin{pmatrix} \beta_1 e_{in} \\ \beta_2 e_{in} \\ \vdots \\ e_{in} \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ 1 \end{pmatrix} \quad (2.30\alpha)$$

$$e_i^{final} = \begin{pmatrix} \beta_1 / |e_i| \\ \beta_2 / |e_i| \\ \vdots \\ 1 / |e_i| \end{pmatrix} \quad (2.30\beta)$$

Η επιλογή μεταβλητών, στην περίπτωση που η PCA εφαρμόζεται για ελάττωση της διαστατικότητας, γίνεται με βάση τις ιδιοτιμές λ_i. Οι ιδιοτιμές αυτές δηλώνουν πόσο «πληροφορία» για την κατανομή των δεδομένων περιέχει κάθε ιδιοδιάνυσμα. Επομένως, η επιλογή των k τελικών μεταβλητών σημαίνει την επιλογή των k, στο πλήθος, ιδιοδιανυσμάτων που περιέχουν την περισσότερη πληροφορία. Τελικά η διαμόρφωση του μετασχηματισμένου σετ δεδομένων P, αν θεωρήσουμε πως S το σύνολο των δεικτών των ιδιοδιανυσμάτων που ο ερευνητής επιλέγει να κρατήσει, γίνεται σύμφωνα με τον μηχανισμό που περιγράφεται στις εξισώσεις 2.31α-2.31β:

$$W = (e_{S_1}, e_{S_2}, \cdots, e_{S_k}) = \begin{pmatrix} e_{S_1,1} & e_{S_2,1} & \cdots & e_{S_k,1} \\ e_{S_1,2} & e_{S_2,2} & \cdots & e_{S_k,2} \\ \vdots & \vdots & \ddots & \vdots \\ e_{S_1,n} & e_{S_2,n} & \cdots & e_{S_k,n} \end{pmatrix} \quad (2.31a)$$

$$P = W \cdot X \quad (2.31\beta)$$

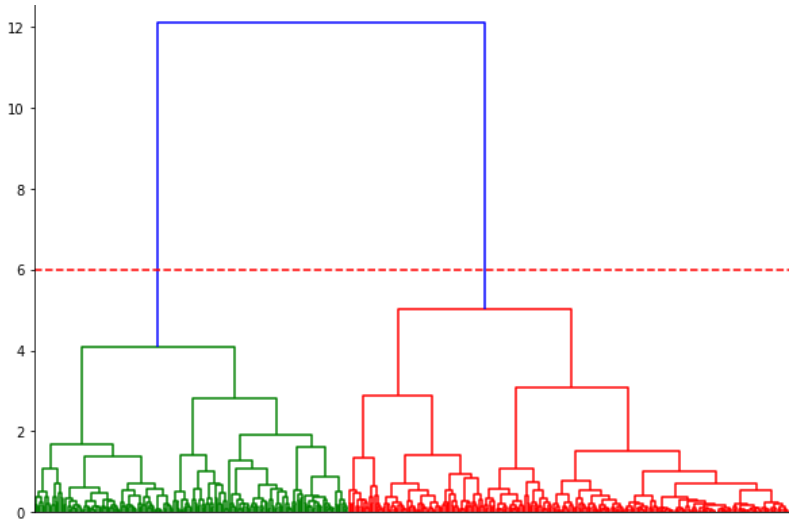
2.5.2 Ιεραρχική Συσταδοποίηση (Hierarchical Clustering, HC)[42]

Η ιεραρχική συσταδοποίηση, όπως γίνεται αμέσως αντιληπτό, ανήκει στην κατηγορία των μεθόδων συσταδοποίησης και χρησιμοποιείται για την ταξινόμηση αταξινόμητων στοιχείων (χωρίς τελικό σημείο). Στόχος της μεθόδου αυτής είναι η κατασκευή ενός «ανάποδου δέντρου», που ονομάζεται δενδρόγραμμα, σύμφωνα με το οποίο τα στοιχεία ταξινομούνται σε συστάδες. Υπάρχουν δύο προσεγγίσεις για την κατασκευή του δενδρογράμματος:

- «Από κάτω προς τα πάνω» (bottom-up ή agglomerative)
- «Από πάνω προς τα κάτω» (top-bottom ή divisive)

Η ανάλυση που θα ακολουθήσει αφορά την «από κάτω προς τα πάνω» προσέγγιση καθότι αυτή χρησιμοποιήθηκε στην παρούσα διπλωματική εργασία.

Στο δενδρόγραμμα που κατασκευάζεται, κάθε «φύλλο» (χαμηλότερο σημείο του δέντρου) θεωρείται ξεχωριστή συστάδα και αναπαριστά μια παρατήρηση του σετ δεδομένων. Όσο, όμως, κινούμαστε προς τα «πάνω» οι συστάδες αυτές συγχωνεύονται σε μεγαλύτερες συστάδες (κλαδιά), μέχρις ότου όλες οι παρατηρήσεις να ανήκουν σε μία. Οι συγχωνεύεις αυτές αντιστοιχούν σε παρατηρήσεις που είναι όμοιες μεταξύ τους. Όσο πιο «χαμηλά» στο δέντρο γίνει η συγχώνευση των συστάδων τόσο πιο όμοιες είναι αυτές μεταξύ τους. Επομένως, η ομοιότητα τους φαίνεται από το σημείο στο οποίο συγχωνεύθηκαν στο ίδιο κλαδί. Για να αναγνωριστούν τις διάφορες συστάδες που είναι χωρισμένα τα δεδομένα χαράσσεται μία οριζόντια γραμμή και οι διακριτές ομάδες κάτω από την γραμμή αυτή είναι οι συστάδες. Το ύψος στο οποίο χαράσσεται η γραμμή επαφίεται στη διακριτική ευχέρεια του ερευνητή, καθώς δεν υπάρχει κάποια μέθοδος που να προσδιορίζει την βέλτιστη γραμμή. Ένα χαρακτηριστικό παράδειγμα δενδρογράμματος παρουσιάζεται στην εικόνα 2.10. Η HC μέθοδος στηρίζεται σε αυτόν τον εξαιρετικά απλοϊκό αλγόριθμο, μένει όμως να διευκρινιστούν δύο σημαντικά θέματα, με ποιόν τρόπο εξετάζεται η ομοιότητα των παρατηρήσεων και πως γίνεται η σύνδεση μεταξύ δύο συστάδων.



Εικόνα 2.10 **Παράδειγμα δενδρογράμματος. Στο παράδειγμα της εικόνας αυτής έχει χαραχθεί η οριζόντια γραμμή $y = 6$ και κάτω από αυτήν υπάρχουν δύο συστάδες. Ένα συμπέρασμα που μπορούμε ακόμα να αντλήσουμε από το παράδειγμα αυτό είναι πως οι συστάδες αυτές είναι αρκετά απομακρυσμένες επομένως είναι πολύ πιθανό να είναι γραμμικώς διαχωρίσιμες.**

Ο τρόπος με τον οποίον καταλήγει η μέθοδος στο συμπέρασμα πως δύο συστάδες θα πρέπει να συγχωνευθούν αποτελεί μια βασική παράμετρο για την μέθοδο αυτή και εξαρτάται από τον ερευνητή και τη φύση των δεδομένων που θα εφαρμοστεί η μέθοδος. Η ευκλείδεια απόσταση των παρατηρήσεων είναι ο πιο συχνός τρόπος έκφρασης ομοιότητας όταν εφαρμόζεται η HC καθώς

δίνεται η δυνατότητα έτσι να εντοπισθούν έκτροπες τιμές που θα μπορούσαν να επιφέρουν μεγάλα προβλήματα στην ανάλυση των δεδομένων. Ένας ακόμα συχνός τρόπος εξέτασης της ομοιότητας των συστάδων είναι η απόσταση βάσει συσχέτισης (correlation-based distance). Σύμφωνα με τον τρόπο αυτό, δύο παρατηρήσεις θεωρούνται όμοιες όταν υπάρχει συσχέτιση μεταξύ των μεταβλητών τους. Φυσικά, ο σκοπός για τον οποίον υλοποιείται η ιεραρχική συσταδοποίηση συνήθως καθιστά ξεκάθαρο τον τρόπο με τον οποίον θα πρέπει να χαρακτηρίζονται δύο συστάδες όμοιες.

Στην αρχή του αλγορίθμου, όλα τα στοιχεία αποτελούν μία συστάδα και τα στοιχεία που έχουν τις μικρότερες ευκλείδειες αποστάσεις μεταξύ τους συσταδοποιούνται. Μετά την πρώτη συγχώνευση όμως σχεδόν κάθε συστάδα περιέχει περισσότερες από μία παρατηρήσεις, επομένως θα πρέπει να υπάρξει ένας τρόπος όπου θα ελέγχεται το κριτήριο ομοιότητας για μία ομάδα περισσότερων παρατηρήσεων. Αυτή η επέκταση επιτυγχάνεται με την ανάπτυξη της ιδέας των συνδέσμων (linkages) μεταξύ ομάδων παρατηρήσεων. Οι τέσσερις πιο γνωστές μέθοδοι σύνδεσης είναι οι average, complete, single και centroid και παρουσιάζονται σύντομα στον πίνακα 1.

Σύνδεσμοι	Περιγραφή
Average	Υπολογισμός όλων των ομοιοτήτων κατά ζεύγη μεταξύ των παρατηρήσεων των συστάδων C_1 και C_2 και καταγραφή της μέσης τιμής των ομοιοτήτων.
Complete	Υπολογισμός όλων των ομοιοτήτων κατά ζεύγη μεταξύ των παρατηρήσεων των συστάδων C_1 και C_2 και καταγραφή της μέγιστης τιμής των ομοιοτήτων.
Single	Υπολογισμός όλων των ομοιοτήτων κατά ζεύγη μεταξύ των παρατηρήσεων των συστάδων C_1 και C_2 και καταγραφή της ελάχιστης τιμής των ομοιοτήτων.
Centroid	Υπολογισμός όλων της ομοιότητας μεταξύ των κέντρων των συστάδων C_1 και C_2. Η μέθοδος αυτή μπορεί να καταλήξει σε αθέμιτα αποτελέσματα.

Πίνακας 2.1 **Μέθοδοι συνδέσμων συστάδων με περισσότερες από μία παρατηρήσεις.**

2.6 Δείκτες αξιολόγησης (Metrics)[43]

Όπως προαναφέρθηκε και στο βήμα 5 της διαδικασίας της παραγράφου 2.3 ένα ιδιαζόντως σημαντικό βήμα για την εκπαίδευση ενός μοντέλου προβλεπτικής μάθησης είναι η αξιολόγηση του μέσω των μετρικών τιμών. Στην παράγραφο αυτή θα προσδιορισθούν ορισμένες δημοφιλής επαληθευτικές τιμές για προβλήματα ταξινόμησης, οι οποίες είναι κομμάτι και της υπολογιστικής διαδικασίας της παρούσας διπλωματικής εργασίας.

2.6.1 Confusion Matrix

Πρόκειται για έναν πίνακα που επιτρέπει την οπτικοποίηση της απόδοσης ενός μοντέλου. Για την ακρίβεια, όπως δηλώνει και το όνομα της συγκεκριμένης επαληθευτικής τιμής, μέσω αυτής φαίνεται αν το μοντέλο έχει «συγχύσει» τις κλάσεις και τις συσχετίσεις τους με τα δεδομένα. Ο πίνακας αυτός

		Πραγματικές Τιμές	
		Κλάση 0	**Κλάση 1**
Προβλέψεις Μοντέλου	**Κλάση 0**	True Negative (TN)	False Positive (FP)
	Κλάση 1	False Negative (FN)	True Positive (TP)

Εικόνα 2.11 **Η μορφή ενός Confusion πίνακα**

είναι της μορφής της εικόνας 2.11 και κάθε γραμμή αναπαριστά τις προβλέψεις του μοντέλου ενώ κάθε στήλη τις πραγματικές τιμές.

Όπως φαίνεται στην εικόνα 2.11 η κύρια διαγώνιος του πίνακα αποτελείται από τις σωστές τιμές που προέβλεψε το μοντέλο (TN και TP) ενώ η δευτερεύουσα διαγώνιος από τις λάθος (FN και FP). Πιο συγκεκριμένα, οι τιμές αυτές αποτελούν τις βασικές τιμές οι οποίες χαρακτηρίζουν την αποτελεσματικότητα των μοντέλων ταξινόμησης και ορίζονται ως:

- True Negative: Πλήθος στοιχείων που ανήκουν στην κλάση 0 και το μοντέλο προέβλεψε κλάση 0,
- False Positive: Πλήθος στοιχείων που ανήκουν στην κλάση 1 και το μοντέλο προέβλεψε κλάση 0,
- False Negative: Πλήθος στοιχείων που ανήκουν στην κλάση 0 και το μοντέλο προέβλεψε κλάση 1 και
- True Positive: Πλήθος στοιχείων που ανήκουν στην κλάση 1 και το μοντέλο προέβλεψε κλάση 1.

Οι τιμές αυτές χρησιμοποιούνται και στον υπολογισμό των ακόλουθων μετρικών τιμών.

2.6.2 Accuracy

Ο δείκτης αξιολόγησης accuracy αναφέρεται συνολικά στα στοιχεία που κατηγοριοποιήθηκαν σωστά στην κλάση τους. Η μετρική τιμή accuracy έχει μέγιστη τιμή το 1 και υπολογίζεται ως το πηλίκο του αθροίσματος των σωστών προβλέψεων προς τις συνολικές:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.32)$$

2.6.3 Precision

Η μετρική τιμή precision αναφέρεται στα στοιχεία που κατηγοριοποιήθηκαν σωστά στην κλάση ενδιαφέροντος (κλάση 1) ως προς το συνολικό αριθμό στιγμιότυπων που κατηγοριοποιήθηκαν σε αυτήν. Όσο η τιμή του πλησιάζει στο 1, τόσο μεγαλύτερη θεωρείται η ακρίβεια του μοντέλου για την κλάση 1:

$$Precision = \frac{TP}{TP + FP} \quad (2.33)$$

2.6.4 Recall

Η ανάκληση (recall) αναφέρεται στον αριθμό των στοιχείων που κατηγοριοποιήθηκαν σωστά στην κλάση 1 ως προς το συνολικό αριθμό στιγμιότυπων που θα έπρεπε να είχαν κατηγοριοποιηθεί σε αυτήν. Όσο η τιμή της ανάκλησης πλησιάζει στο 1, τόσο μεγαλύτερη είναι η ακρίβεια του μοντέλου:

$$Recall = \frac{TP}{TP + FN} \quad (2.34)$$

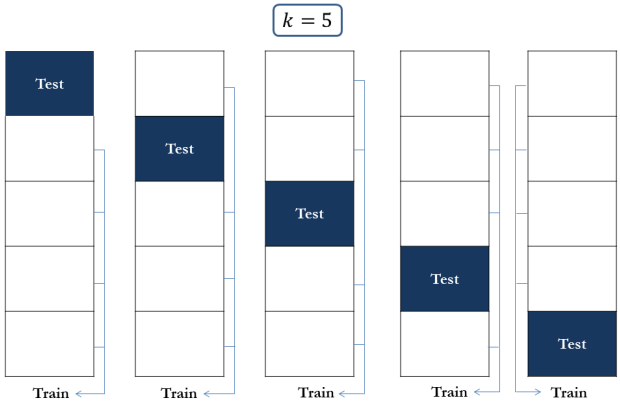
2.6.5 F-Score ή F-Measure

Ο αρμονικός μέσος (F-Score) υπολογίζεται βάσει των μετρικών τιμών precision και recall. Όσο πιο κοντά βρίσκονται οι δύο τιμές μεταξύ τους, τόσο πλησιάζει η τιμή του αρμονικού μέσου στο 1 και άρα τόσο πιο ακριβές το μοντέλο ταξινόμησης:

$$F - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (2.35)$$

2.6.6 Cross-Validation

Η μετρική τιμή cross-validation είναι παρόμοια με την accuracy, με την διαφορά πως το cross-validation δεν εφαρμόζεται στο επαληθευτικό σετ δεδομένων, αλλά σε ολόκληρο το σετ δεδομένων.



Εικόνα 2.12 **Οπτικοποίηση της μεθόδου που ακολουθεί το cross validation για την αποφυγή προβλημάτων υπέρ-προσαρμογής και υπό-προσαρμογής. Στο παράδειγμα αυτό, το σετ δεδομένων έχει διαχωριστεί σε 5 Folds.**

Πιο συγκεκριμένα, η εφαρμογή αυτής της μετρικής τιμής προϋποθέτει πως δεν έχει προηγηθεί διαχωρισμός των δεδομένων σε εκπαιδευτικά και επαληθευτικά. Χρησιμοποιείται για την αποφυγή περιπτώσεων υπέρ-προσαρμογής και υπό-προσαρμογής, και αυτό το επιτυγχάνει χρησιμοποιώντας κάθε παρατήρηση του σετ δεδομένων για την εκπαίδευση και την επαλήθευση του μοντέλου. Ειδικότερα, το σετ δεδομένων διαχωρίζεται σε k τμήματα (CVs ή Folds) και το μοντέλο εκπαιδεύεται k φορές, στα $k-1$ τμήματα και επαληθεύεται (accuracy) στο τμήμα που έμεινε εκτός κατά την εκπαίδευση (εικόνα 2.12). Με τον τρόπο αυτό, όλες οι παρατηρήσεις έχουν χρησιμοποιηθεί ως εκπαιδευτικές και ως επαληθευτικές. Τελικά, από την εφαρμογή αυτής της μετρικής τιμής προκύπτουν k accuracy τιμές και συνήθως, η μέση τιμή αυτών των παρουσιάζεται ως η «ακρίβεια του μοντέλου».

2.7 Μπεϋζιανή Βελτιστοποίηση (Bayesian Optimization)

Συνήθως τα μοντέλα μηχανικής μάθησης διαθέτουν παραμέτρους που ρυθμίζουν την αποδοτικότερη και γρηγορότερη εκπαίδευση τους. Στις μεθόδους μάθησης με επίβλεψη που παρουσιάστηκαν στην παράγραφο 2.4, για παράδειγμα, εμφανίζονται ορισμένες τέτοιες παράμετροι, όπως η παράμετρος C στην SVM, ο αριθμός των δένδρων στην RF, η παράμετρος C στην συνάρτηση κόστους της LR κ.ο.κ. Αυτές οι παράμετροι συχνά αποτελούν πρόβλημα κατά την εκπαίδευση για τον ερευνητή καθώς η τιμή που θα πρέπει να πάρουν δεν ορίζεται σαφώς, καταλήγοντας έτσι σε έναν τεράστιο παράγοντα αβεβαιότητας για την απόδοση του μοντέλου. Οι ερευνητές συχνά επιλέγουν τις

παραμέτρους αυτές τυχαία και μέχρι να καταλήξουν σε ένα ικανοποιητικό μοντέλο. Ένας πιο αποτελεσματικός τρόπος είναι να αυτοματοποιηθεί η διαδικασία επιλογής τιμών για τις παραμέτρους. Αυτό μπορεί να επιτευχθεί εάν η διαδικασία αυτή αντιμετωπιστεί σαν μια διαδικασία βελτιστοποίησης μιας άγνωστης συνάρτησης και χρησιμοποιηθούν οι αντίστοιχοι αλγόριθμοι. Ένας τέτοιος αλγόριθμος, ο οποίος χρησιμοποιήθηκε και στην υπολογιστική διαδικασία αυτής της διπλωματικής εργασίας, είναι η Μπεϋζιανή Βελτιστοποίηση (Bayesian Optimization).

Η μπεϋζιανή βελτιστοποίηση είναι μια ιδιαίτερα ισχυρή στρατηγική για την βελτιστοποίηση άγνωστων αντικειμενικών συναρτήσεων. Η στατιστική αυτή μέθοδος συχνά προτιμάται από τις υπόλοιπες, όπως τη σύγκλιση με ελάττωση της παραγώγου (gradient descent) που χρησιμοποιείται συχνά, και ιδιαίτερα όταν οι υπολογισμοί της αντικειμενικής συνάρτησης κοστίζουν αρκετά ή όταν δεν γνωρίζουμε τις παραγώγους της. Η τεχνική αυτή είναι από τις πιο αποτελεσματικές προσεγγίσεις όσον αφορά το πλήθος των φορών που απαιτείται εκτίμηση τιμής από την αντικειμενική συνάρτηση. Η δυνατότητα αυτή πηγάζει από το γεγονός πως η μπεϋζιανή βελτιστοποίηση χρησιμοποιεί την «πρότερη γνώση» (prior belief) για να ρυθμίσει την ανταλλαγή «εξερεύνησης-εκμετάλλευσης» (exploration and exploitation trade off). Το όνομα της τεχνικής αυτής πηγάζει από το θεώρημα Bayes (εξίσωση 2.13) που δηλώνει πως η ύστερη πιθανότητα (posterior probability) μίας υπόθεσης M, δεδομένης εμπειρίας E, είναι ανάλογη του γινόμενου της πιθανότητας της εμπειρίας E δεδομένης της υπόθεσης M επί την πρότερη πιθανότητα της υπόθεσης M (prior probability):

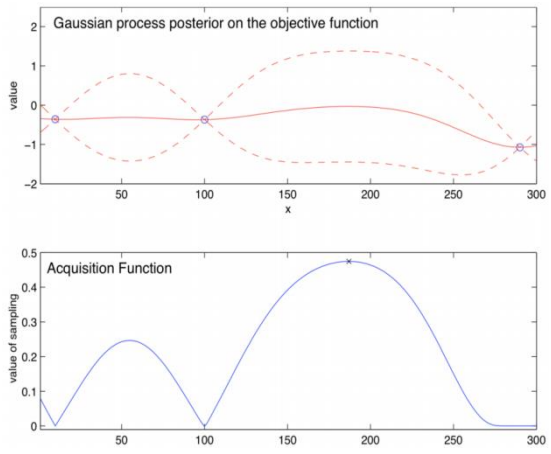
$$P(M|E) \propto P(E|M)P(M) \quad (2.36)$$

Η εξίσωση 2.36, είναι μία εναλλακτική μορφή της εξίσωσης 2.13 και παρά την απλότητα της περιέχει το κλειδί για την βελτιστοποίηση της άγνωστης αντικειμενικής συνάρτησης. Στην μπεϋζιανή βελτιστοποίηση ο όρος «πρότερη γνώση» αφορά την γνώση που έχει ή σταδιακά αποκτά ο αλγόριθμος για τον χώρο όλων των πιθανών αντικειμενικών συναρτήσεων.[44]

Αν ορίσουμε x_i ένα δείγμα και $f(x_i)$ μία παρατήρηση (τιμή της αντικειμενικής συνάρτησης για το x_i τότε όσο συγκεντρώνονται ζεύγη δειγμάτων-παρατηρήσεων $D = \{x_i, f(x_i)\}$ τόσο η πρότερη κατανομή συνδυάζεται με την συνάρτηση πιθανότητας $P(D|f)$. Στην ουσία της η $P(D|f)$ απαντά στην ερώτηση *«δεδομένων αυτών που γνωρίζουμε ήδη για την πρότερη κατανομή f, πόσο πιθανά είναι τα δεδομένα που έχουμε;»*. Επομένως, συνδυάζοντας όλα τα ανωτέρω με την εξίσωση 2.36 προκύπτει η έκφραση για την ύστερη κατανομή πιθανότητας για τις παρατηρήσεις $f(x_i)$ των δειγμάτων x_i (εξίσωση 2.37):

$$P(f|D) \propto P(D|f)P(f) \quad (2.37)$$

Αυτή η κατανομή, όπως και η μέση τιμή της $\mu(x_i)$, ή η προσέγγιση της f, και η διακύμανση της $\sigma(x_i)$, κατ' επέκταση και τα διαστήματα αξιοπιστίας (credible intervals) παρέχονται από μια Γκαουσιανή διαδικασία παλινδρόμησης (Gaussian Process Regression). Στην εικόνα 2.13[45]



Εικόνα 2.13 Παράδειγμα της μπεϋζιανής βελτιστοποίησης μία αντικειμενικής συνάρτησης μίας μεταβλητής. Πάνω: Παρατηρήσεις της αντικειμενικής συνάρτησης σε 3 σημεία, μία εκτίμηση της f με συνεχή κόκκινη γραμμή και τα διαστήματα αξιοπιστίας για την f μεταξύ των διακεκκομένων γραμμών. Κάτω: Η συνάρτηση απόκτησης για την f. Παρατηρείται πως για τα σημεία που έχει ήδη εξετάση, η τιμή της συνάρτησης απόκτησης είναι 0 ενώ για τα σημεία με μεγάλο διάστημα αξιοπιστίας και για τα σημεία που έχουν ύστερο μέσο υψηλο, η τιμή της συνάρτησης είναι υψηλή.

παρουσιάζεται η $\mu(x_i)$, ή η προσέγγιση της f, ως η συνεχής γραμμή ενώ τα διαστήματα αξιοπιστίας θεωρούνται τα διαστήματα μεταξύ των διακεκομμένων γραμμών.

Μετά τον υπολογισμό της κατανομής υπολογίζεται το επόμενο σημείο που θα γίνει η ανωτέρω διαδικασία. Το σημείο αυτό βρίσκεται μέσω της συνάρτησης απόκτησης «acquisition function» η οποία υπολογίζει το σημείο που είναι το πιο πιθανό να οδηγήσει στην μέγιστη τιμή της f. Αφού υπολογισθεί το σημείο αυτό, ενημερώνεται το σύνολο D και η διαδικασία επαναλαμβάνεται. Το τέλος της διαδικασίας είναι όταν γίνουν όλες οι επαναλήψεις που ορίσθηκαν στην αρχή.[45] Ο αλγόριθμος της διαδικασίας παρουσιάζεται στον πίνακα 2.2.

Αλγόριθμος για την Μπεϋζιανή Βελτιστοποίηση
Πρότερη Γκαουσιανή διαδικασία στην f χρησιμοποιώντας τα σημεία εξερεύνησης
Όρισμός N επαναλήψεων
Όσο $n \leq N$ κάνε (χρησιμοποιώντας τα σημεία εκμετάλλευσης):
Ενημέρωση της ύστερης κατανομής πιθανότητας της f χρησιμοποιώντας όλα τα δεδομένα
Επιλογή του x_n από την συνάρτηση απόκτησης με την τρέχουσα ύστερη κατανομή
Παρατήρηση της τιμής $y_n = f(x_n)$
Αύξηση του n
Τέλος Επανάληψης
Επιστροφή του μέγιστου f ή του $argmax_x f$

Πίνακας 2.2 Αλγόριθμος για την επιλογή του x που μεγιστοποιεί την f, μέσω της μπεϋζιανής βελτιστοποίησης

2.8 Tree-Based Pipeline Optimization Tool (TPOT)[46,47]

Το TPOT είναι ένα ανοικτό εργαλείο, που στηρίζεται στο γενετικό προγραμματισμό και ανήκει στην κατηγορία αλγορίθμων αυτόματης μηχανικής μάθησης (Automated machine learning, AutoML). Η AutoML είναι μια αυτοματοποιημένη διαδικασία αναζήτησης και επιλογής μοντέλων που περιγράφουν ένα σετ δεδομένων, ενώ παράλληλα ρυθμίζονται οι παράμετροι βαθμονόμησής τους. Το πακέτο TPOT αναζητά μέσα από όλα τα μοντέλα, που υπάρχουν διαθέσιμα από το πακέτο scikit-learn, εκείνο, ή το συνδυασμό εκείνων, που έχει την υψηλότερη απόδοση για τα εκπαιδευτικά δεδομένα.[48]

Ειδικότερα, όπως αναφέρθηκε, το TPOT στηρίζεται στο γενετικό προγραμματισμό, ο οποίος με την σειρά του στηρίζεται στη θεωρία της εξέλιξης των οργανισμών. Σύμφωνα με την εξελικτική θεωρία, οι οργανισμοί, από γενιά σε γενιά, αποκτούν χρήσιμα ή απορρίπτουν ανώφελα για την επιβίωση τους χαρακτηριστικά (μετάλλαξη). Φυσικά, υπάρχουν οργανισμοί οι οποίοι με το πέρας των αιώνων δεν κατάφεραν να επιβιώσουν και εξαφανίστηκαν (φυσική επιλογή). Με τον ίδιο τρόπο στο γενετικό προγραμματισμό ορίζονται οι γενιές (επαναλήψεις σε έναν βρόχο επανάληψης) κάθε μία από τις οποίες αποτελεί εξέλιξη της προηγούμενης. Έτσι, για g γενιές, κάθε γενιά g_i αποτελείται από έναν πληθυσμό x υπολογιστικών ροών (pipelines) και κάθε επόμενη γενιά g_{i+1} αποτελείται από τον ίδιο πληθυσμό x, με ορισμένες «μεταλλαγμένες» υπολογιστικές ροές της προηγούμενης γενιάς, ορισμένες ίδιες και καινούριες. Πιο συγκεκριμένα, το TPOT κατασκευάζει x υπολογιστικές ροές κάθε μία από τις οποίες διαθέτει έναν αριθμό βημάτων. Κάθε βήμα στις ροές αυτές θεωρείται ένας τελεστής (operator) ο οποίος μπορεί να μετασχηματίζει τα δεδομένα, να εφαρμόζει τεχνικές χωρίς επίβλεψη (unsupervised) ή τεχνικές με επίβλεψη (supervised). Η «μετάλλαξη» που μπορεί να συμβεί στις ροές αυτές, ορίζεται ως η αλλαγή στα βήματά τους.

Στόχος του TPOT είναι η εύρεση μιας υπολογιστικής ροής που έχει υψηλή ακρίβεια στα επαληθευτικά δεδομένα. Για να το επιτύχει αυτό, διαχωρίζει τα δεδομένα σε εκπαιδευτικά (75%) και επαληθευτικά (25%) και επιλέγει τυχαία την πρώτη γενιά των υπολογιστικών ροών. Με την

εφαρμογή των x στο πλήθος υπολογιστικών ροών στα εκπαιδευτικά και επαληθευτικά δεδομένα υπολογίζονται x, στο πλήθος, δείκτες αξιολόγησης «accuracy». Στη συνέχεια, το 10% των ροών με την υψηλότερη ακρίβεια υφίσταται «μεταλλάξεις» και συμπεριλαμβάνονται στη δεύτερη γενιά. Η ίδια διαδικασία επαναλαμβάνεται μέχρις ότου υπολογιστούν οι g γενιές, ή ξεπεραστεί ένα t χρονικό όριο.

Επομένως, φαίνεται πως για την εφαρμογή του, απαραίτητος είναι, ο προσδιορισμός των g γενιών, του x πληθυσμού και του t χρόνου εκτέλεσης. Τέλος, για την αποτελεσματικότερη διαδικασία, το TPOT έχει την επιλογή να χρησιμοποιηθεί η επαληθευτική τιμή διασταυρούμενης επαλήθευσης αντί της ακρίβειας στο 25% των παρατηρήσεων. Με τον τρόπο αυτό εξασφαλίζεται πως το μοντέλο που θα προκύψει δεν έχει υπέρ-προσαρμοστεί στα εκπαιδευτικά δεδομένα.

2.9 Εφαρμογή Jaqpot[49,50]

Πρόκειται για μία υπολογιστική πλατφόρμα που διαθέτει γραφικό περιβάλλον χρήστη (GUI) και διεπαφή προγραμματισμού εφαρμογών (API) για in silico μοντελοποιήσεις. Συγκεκριμένα, είναι μία online εφαρμογή ανεπτυγμένη σε Java, R και Python που επιτρέπει την ανάπτυξη, την χρήση και την διάθεση μοντέλων μηχανικής μάθησης. Το Jaqpot αναπτύχθηκε από το εργαστήριο Αυτόματης Ρύθμισης και Ελέγχου του ΕΜΠ στα πλαίσια του προγράμματος «FP7 OpenTox» το 2016 και έκτοτε λειτουργεί και εξελίσσεται, καταλήγοντας σε μία εφαρμογή που έχει πλήρη συμβατότητα με τις γλώσσες προγραμματισμού Python και R. Ειδικότερα, οι ενσωματωμένες Jaqpot βιβλιοθήκες στις γλώσσες Python και R, σε συνδυασμό με το πρωτόκολλο ανταλλαγής δεδομένων του Jaqpot (Jaqpot Protocol of Data Interchange, JPDI) επιτρέπει στους προγραμματιστές να χρησιμοποιήσουν την εφαρμογή μέσα από τον κώδικά τους, διαθέτοντας τα μοντέλα τους online. Τα μοντέλα αυτά μπορούν να χρησιμοποιηθούν είτε για επαλήθευση είτε για πρόβλεψη από τα μέλη της ομάδας που τα δημιούργησε, ανά πάσα στιγμή και από οπουδήποτε στον κόσμο. Άλλη μια ιδιαιτέρως σημαντική ιδιότητα της εφαρμογής είναι η δυνατότητα σύνδεσης πολλών ομάδων ή οργανισμών με τα μοντέλα. Με αυτόν τον τρόπο, διάφορες ομάδες διαθέτουν δικαιώματα στην χρήση του εκπαιδευμένου μοντέλου. Στην παρούσα εργασία, τα μοντέλα που εκπαιδεύτηκαν για την πρόβλεψη της τοξικότητας των διάφορων νανοϋλικών διατέθηκαν στο Jaqpot.

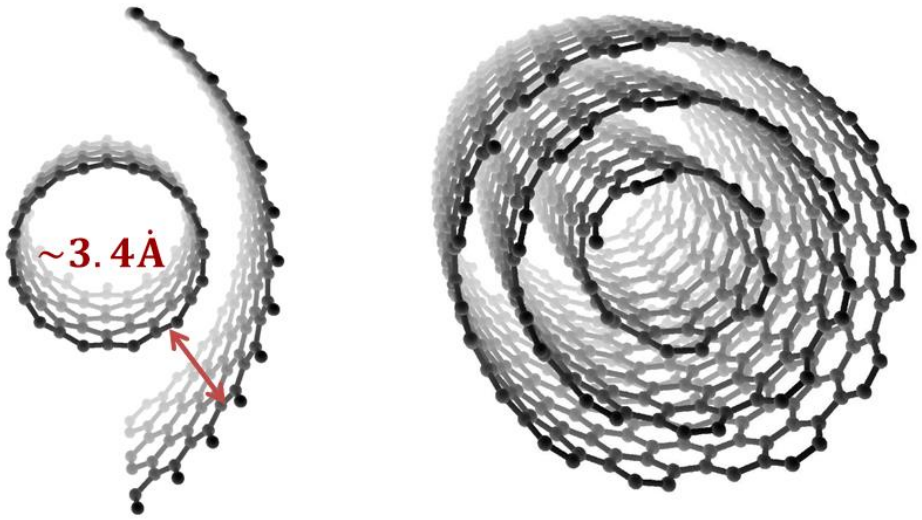
Κεφάλαιο 3

Υπό μελέτη συστήματα

Στην διπλωματική αυτή εργασία δημιουργήθηκαν και εκπαιδεύτηκαν μοντέλα που προβλέπουν την τοξικότητα νανοϋλικών, σύμφωνα με ορισμένες φυσικοχημικές ιδιότητές τους. Τα νανοϋλικά που εξετάσθηκαν ήταν νανοσωλήνες άνθρακα πολλαπλών τοιχωμάτων (Multi-Walled Carbon NanoTubes, MWCNTs) και υπέρ-παραμαγνητικά οξείδια σιδήρου (Super Paramagnetic Iron Oxides Nanoparticles, SPIONs). Τα νανοϋλικά αυτά έχουν ευρεία εφαρμογή στη βιομηχανία και παράλληλα πρόσφατες έρευνες έχουν καταλήξει στο συμπέρασμα πως υπό συνθήκες, τα υλικά αυτά είναι δυνατόν να είναι τοξικά για τους ανθρώπινους οργανισμούς. Στο κεφάλαιο αυτό θα παρουσιαστούν τα υλικά και θα γίνει μια εισαγωγή στην δομή τους και τις ιδιότητες τους. Επιπρόσθετα, θα παρουσιαστούν κάποιες από τις ευρείες εφαρμογές τους, ο κίνδυνος για τοξικότητα και η ανάγκη ύπαρξης μοντέλων πρόβλεψης της τοξικότητας τους.

3.1 Γενικά χαρακτηριστικά των MWCNTs

Οι νανοσωλήνες άνθρακα (Carbon Nanotubes, CNTs) αποτελούν τα πιο σημαντικά και πιο μελετημένα νανοϋλικά. Το έντονο ενδιαφέρον των επιστημόνων για τους νανοσωλήνες άνθρακα εκδηλώθηκε από την ανακάλυψή τους από τον Iijima το 1991, λόγω των ασυνήθιστων ιδιοτήτων τους (υψηλό μέτρο ελαστικότητας και ηλεκτρική/θερμική αγωγιμότητα). [51] Στη συνέχεια, οι Ebbesen και Ajayan από την εργαστηριακή ομάδα του Iijima απέδειξαν ότι τα CNTs μπορούν κάλλιστα να παραχθούν σε μεγάλες ποσότητες, μεταβάλλοντας τις συνθήκες εξάχνωσης του γραφίτη, δίνοντας έτσι την δυνατότητα στις βιομηχανίες να τους παράξουν μαζικά.[52] Οι νανοσωλήνες είναι αναδιπλωμένα φύλλα, παρόμοια με του γραφίτη, το κυλινδρικό τμήμα των οποίων αποτελείται αποκλειστικά από εξαγωνικούς δακτυλίους. Σε ορισμένες σπάνιες περιπτώσεις τα άκρα τους έχουν ημισφαιρικό σχήμα, όπως το μισό του μορίου του φουλερενίου. Συνήθως, όμως, έχουν επίπεδη μορφή, με ένα δακτύλιο από πεντάγωνα στο όριο μετάβασης προς το κυλινδρικό τμήμα του μορίου.



Εικόνα 3.1 **Δομή ενός MWCNT. Αριστερή: Η απόσταση μεταξύ των διαφορετικών τοιχωμάτων, ίση με 3.4 Å. Δεξιά: Ένας νανοσωλήνας άνθρακα τριών τοιχωμάτων.**

Υπάρχουν δύο ειδών CNTs: μονού τοιχώματος (Single-Walled Carbon Nanotubes, SWCNTs) και πολλαπλών τοιχωμάτων (Multi-Walled Carbon Nanotubes, MWCNTs). Τα SWCNTs έχουν ένα κυλινδρικό κέλυφος με πάχος όσο αυτό ενός μόνο ατόμου και μπορούν να θεωρηθούν ως η θεμελιώδης δομική μονάδα ενώ η διάμετρος τους παίρνει τιμές από 0.4 nm έως και 3.0 nm.[53] Οι MWCNTs, αποτελούνται από ένα πλήθος φύλλων γραφίτη, τα οποία αναδιπλώνονται σε ομοαξονικούς κυλίνδρους σχηματίζοντας ένα μεγάλο σωλήνα, με σταθερή απόσταση μεταξύ των τοιχωμάτων κάθε στρώματος, ίση με 3.4 Å.[51] Η δομή ενός νανοσωλήνα άνθρακα πολλαπλών τοιχωμάτων παρουσιάζεται στην εικόνα 3.1. Σε αυτήν την διπλωματική εργασία θα ασχοληθούμε με την δεύτερη κατηγορία CNTs, τους MWCNTs.

Η αύξηση του ερευνητικού και εμπορικού ενδιαφέροντος και η μαζική παραγωγή των MWCNTs, που ήδη έχει αρχίσει να συμβαίνει, έχει ως απόρροια την αλληλεπίδραση των MWCNTs με τον άνθρωπο και τους έμβιους οργανισμούς. Έτσι, η κατανόηση της τοξικότητας τους, του μηχανισμού με τον οποίον είναι τοξικά και των προϋποθέσεων για να είναι, αλλά και των επιπτώσεων στην υγεία του ανθρώπου, είναι ζωτικής σημασίας για την σύνθεση και την χρήση αυτών των υλικών.

3.2 Χρήσεις των CNTs

Οι CNTs εδώ και κάποια χρόνια βρίσκουν ευρεία χρήση σε βιολογικές και φαρμακευτικές εφαρμογές. Χαρακτηριστικό παράδειγμα εφαρμογής τους είναι η χρήση τους στον τομέα της βιολογίας ως αισθητήρες για την ανίχνευση του DNA και πρωτεϊνών, ως διαγνωστικές συσκευές για τη διάκριση των διαφορετικών πρωτεϊνών από δείγματα ορού, και ως φορείς πρωτεϊνών.[54] Η ισχυρή αντιμικροβιακή δράση που παρουσιάζουν κάποια είδη νανοσωλήνων άνθρακα αποτελεί κίνητρο για τη μελέτη των εφαρμογών τους. Η κεντρική ιδέα είναι να χρησιμοποιηθούν ως δραστικά δομικά στοιχεία για να σχηματίσουν νανοσύνθετα με διάφορες επιθυμητές μορφολογίες και λειτουργίες, οι οποίες μπορεί να είναι κατάλληλες για διάφορους ειδικούς σκοπούς. Στη συνέχεια αναλύονται κάποιες πιθανές χρήσεις τους.

3.2.1 Αντιρρυπαντικές επιφάνειες

Πολλές μελέτες έχουν επικεντρωθεί στην εφαρμογή των CNTs για την απολύμανση του νερού. Όταν οι CNTs εισάγονται σε ένα υδάτινο περιβάλλον, είναι δυνατό να αδρανοποιήσουν τα βακτήρια.[55] Ειδικότερα, επιφάνειες στις οποίες είχε εναποτεθεί στρώμα από νανοσωλήνες μονού τοιχώματος βρέθηκαν να είναι σε θέση να αναστείλουν το σχηματισμό βακτηριακού βιοφίλμ.[56] Τα βακτήρια και οι ιοί μπορούν επίσης να απομακρυνθούν αποτελεσματικά από τα φίλτρα μεμβράνης που περιέχουν MWCNTs.[57,58] Έτσι θα μπορούσαν να χρησιμοποιηθούν σαν φίλτρα με υψηλή απόδοση για βιομηχανική χρήση. Τέλος, η εισαγωγή MWCNTs σε μεμβράνες επηρεάζει το πορώδες και το μέσο μέγεθος των πόρων, ενώ μειώνεται η τραχύτητα της επιφάνειας. Έτσι, δύναται να χρησιμοποιηθούν σαν πρόσθετα σε ναυπηγικές επιστρώσεις αφού αυξάνουν την απόδοση και το χρόνο ζωής των αντιρρυπαντικών επιστρώσεων.[59]

3.2.2 Αντιμικροβιακή δράση

Έχει παρατηρηθεί πως οι νανοσωλήνες άνθρακα είναι πιθανό να παρουσιάζουν αντιμικροβιακή δράση, όπως φαίνεται και στην ανωτέρω παράγραφο (§ 3.2.1). Για τον λόγο αυτό δημιουργήθηκαν και μελετήθηκαν για πιθανές αντιμικροβιακές δράσεις, διάφορα σύνθετα CNT-πολυμερών. Οι μελέτες αυτές έδειξαν πως ορισμένα σύνθετα έχουν, πράγματι, αντιμικροβιακές ιδιότητες.

Χαρακτηριστικό παράδειγμα αποτελεί η κατασκευή μικροπορώδους φιλμ που αποτελείται από νανοσωλήνες άνθρακα και πολυβινυλοπυρρολιδόνη του ιωδίου (PVPI). Το ιώδιο, γνωστό για την αντισηπτική του δράση, ήταν διαθέσιμο στην επιφάνεια των CNTs οι οποίοι ήταν ενσωματωμένοι σε πολυμερές και αυτό το φιλμ βρήκε εφαρμογή ως αντισηπτικός επίδεσμος. Οι μελέτες έδειξαν πως παρουσίασε υψηλή αντιμικροβιακή δραστικότητα έναντι του βακτηρίου *E. coli*.[60] Ένα ακόμα παράδειγμα είναι αυτό του Nepal και των συνεργατών του, που κατασκεύασαν πολύ-λειτουργικό βιομιμητικό φιλμ αποτελούμενου από SWCNT, DNA και λυσοζύμη. Αυτή η συνθετική μεμβράνη, με υψηλό μέτρο ελαστικότητας κατά Young και ελεγχόμενη μορφολογία, έδειξε εξαιρετική μακροχρόνια αντιμικροβιακή δράση.[61]

3.2.3 Στοχευμένη αντιμικροβιακή δράση

Οι CNTs έχουν συχνά μελετηθεί ως εργαλείο για τη μεταφορά και την κυτταρική μετατόπιση θεραπευτικών μορίων.[62,63] Για παράδειγμα, σε μια *in vitro* μελέτη, τα φάρμακα που δεσμεύτηκαν στους νανοσωλήνες άνθρακα έδειξαν να μεταφέρονται πιο αποτελεσματικά εντός των κυττάρων σε σύγκριση με τα ελεύθερα φάρμακα και να έχουν ισχυρή αντιμυκητιακή δράση.[64,65] Οι πολλαπλές ομοιοπολικές τροποποιήσεις στο πλευρικό τοίχωμα ή στις άκρες των νανοσωλήνων άνθρακα τους επιτρέπει να μεταφέρουν πολλά μόρια ταυτόχρονα, στρατηγική η οποία παρέχει ένα θεμελιώδες πλεονέκτημα στη θεραπεία του καρκίνου.[66,67,68] Σε αντιμικροβιακά φάρμακα, είναι επιθυμητό οι CNTs να μπορούν να αναγνωρίσουν τα επιβλαβή μικρόβια και να μην παρουσιάζουν τοξικές επιδράσεις σε άλλα κύτταρα ή οργανισμούς, ωστόσο, λίγες μελέτες έχουν διερευνήσει τη στοχευμένη αντιμικροβιακή δράση μη τροποποιημένων ή επιφανειακά τροποποιημένων CNTs παρουσία μικροβιακών και άλλου είδους κυττάρων.

3.3 Χημική τροποποίηση των MWCNTs

Οι μεταβολές στη δομή των νανοσωλήνων άνθρακα πολλαπλών τοιχωμάτων έχει παρατηρηθεί πως επηρεάζουν τη δράση τους έναντι των μικροβίων. Τέτοιες μεταβολές μπορούν να επιτευχθούν με προσθήκη λειτουργικών ομάδων στην επιφάνεια των νανοσωλήνων με ομοιοπολικό δεσμό (ομοιοπολική τροποποίηση). Η χημική τροποποίηση των MWCNTs έχει, επομένως, ως αποτέλεσμα την επέκταση του εύρους των πιθανών εφαρμογών τους και για αυτό το λόγο παρουσιάζεται μεγάλο ενδιαφέρον από τους ερευνητές. Ιδιαίτερο ενδιαφέρον, σχετικά με την τοξικότητα, παρουσιάζει η ομοιοπολική τροποποίηση με προσθήκη λειτουργικών ομάδων στην επιφάνεια των νανοσωλήνων με ομοιοπολικό δεσμό. Η ανασκόπηση της βιβλιογραφίας υποδεικνύει ότι η τροποποίηση των MWCNTs περιπλέκει το μηχανισμό τοξικότητας σε σύγκριση με τα μη τροποποιημένα MWCNTs.[69]

Ο Zardini και οι συνεργάτες του παρατήρησαν πως σημαντική αντιμικροβιακή δράση έχει επιτευχθεί από τροποποιημένους MWCNTs με αργινίνη και λυσίνη. Αυτή η ενισχυμένη τοξικότητα αποδείχθηκε έναντι πολλών ειδών βακτηρίων, αρνητικών κατά Gram, συμπεριλαμβανομένων των *Salmonella typhimurium*, καθώς και το ανθεκτικό *Staphylococcus aureus*. Έχει προταθεί ότι οι θετικώς φορτισμένες λειτουργικές ομάδες αμινοξέων δύναται να προσροφηθούν στην αρνητικά φορτισμένη βακτηριακή μεμβράνη. Κατά αυτόν τον τρόπο, επέρχεται ξαφνική κυτταρική λύση. Παρέχουν δηλαδή μια οικονομικά βιώσιμη, λιγότερο τοξική αντιβακτηριακή εναλλακτική λύση, αντικαθιστώντας τα μεταλλικά νανοσωματίδια και τα SWCNTs.[70] Τα τροποποιημένα MWCNTs με -OH, -COOH και -NH2 σαν λειτουργικές ομάδες έχουν επίσης ερευνηθεί. Τα αποτελέσματα αυτής της έρευνας έδειξαν πως η τοξικότητα που παρουσιάζουν αυτά με την επιφανειακή προσθήκη αμινομάδων είναι μεγαλύτερη από αυτή των μη τροποποιημένων και αρκετά μικρότερη από αυτή που παρουσιάζουν οι CNTs με την επιφανειακή προσθήκη καρβοξυλομάδων.[70,71]

3.4 Τοξικότητα των MWCNTs

Στις προηγούμενες παραγράφους, είδαμε τη χημική δομή και τις ιδιότητες των MWCNTs, αλλά και τις εφαρμογές που έχουν βρεθεί μέχρι σήμερα. Παρά τις απαράμιλλες δυνατότητες που προσφέρουν οι MWCNTs, παρατηρήθηκε από τους ερευνητές πως οι νανοσωλήνες άνθρακα είναι ελάχιστα διαλυτοί σε πολλούς διαλύτες, δημιουργώντας κάποιους προβληματισμούς όσον αφορά την τοξικότητά τους. Ωστόσο, η χημική τροποποίησή τους μπορεί να τους καταστήσει υδατοδιαλυτούς και ικανούς να συνδεθούν με μια ποικιλία μορίων όπως πεπτίδια, πρωτεΐνες, νουκλεϊκά οξέα, και θεραπευτικούς παράγοντες.

Φυσικά, όπως προαναφέρθηκε, η αύξηση του εμπορικού ενδιαφέροντος και η μαζική παραγωγή των CNTs, έχει ως αποτέλεσμα την αυξημένη πιθανότητα για αλληλεπίδραση των CNTs με τον άνθρωπο. Έτσι, η κατανόηση των τοξικολογικών και περιβαλλοντικών επιπτώσεων των CNTs είναι ιδιάζουσας σημασίας για τις εφαρμογές τους. Πρώιμες μελέτες έχουν συνδέσει την τοξικότητα των CNTs με το μέγεθος και την επιφάνειά τους. Πιο συγκεκριμένα, μια από τις επικρατέστερες απόψεις που αφορούν την τοξικότητα των CNTs είναι πως καθώς το μέγεθος τους μειώνεται, αυξάνεται η ειδική επιφάνεια, οδηγώντας σε αυξημένη δυνατότητα για αλληλεπίδραση και πρόσληψη από τα ζωντανά κύτταρα. Αυτό το χαρακτηριστικό θα μπορούσε να οδηγήσει σε δυσμενείς βιολογικές επιδράσεις που διαφορετικά δεν θα ήταν δυνατόν με το ίδιο υλικό σε μία μεγαλύτερη μορφή.[72,73] Σε συμφωνία με αυτήν την άποψη, αρκετές μελέτες έχουν δείξει πως τα SWCNTs εμφανίζουν σημαντική κυτταροτοξικότητα σε κύτταρα ανθρώπων και ζώων, ενώ τα MWCNTs παρουσιάζουν πιο ήπια τοξικότητα.[73,74,75] Παρόλα αυτά, αυτό το αποτέλεσμα μπορεί να οφείλεται στη χρήση νανοσωλήνων διαφορετικής καθαρότητας και τροποποίησης ή στη χρήση διαφορετικών μέσων κυτταρικής καλλιέργειας, ίσως ακόμα και στη χρήση διαφορετικών κυτταρικών τύπων (π.χ. ευκαρυωτικά κύτταρα, βακτήρια, κτλ.).

Παρά τη γενική συμφωνία από τους περισσότερους ερευνητές, σχετικά με την πιθανή τοξικότητα των MWCNTs, ο μηχανισμός τοξικότητας εξακολουθεί να μην είναι σαφής και εξακριβωμένος. Υπάρχουν αρκετοί μηχανισμοί που προτείνονται, ωστόσο, οι ακριβείς μηχανισμοί για την βακτηριακή απενεργοποίηση με παρουσία των MWCNTs δεν είναι ακόμη πλήρως κατανοητοί. Με δεδομένο το κόστος και την πολυπλοκότητα της εκτίμησης της τοξικότητας για κάθε πιθανό δείγμα MWCNT, γενικές συσχετίσεις μεταξύ της τοξικότητας και των φυσικοχημικών ιδιοτήτων των νανοσωλήνων θα είναι θεμελιώδης για τις εμπορικές εφαρμογές των νανοϋλικών. Στα πλαίσια αυτά εκπονήθηκε και η παρούσα διπλωματική εργασία, για να παρουσιάσει ένα μοντέλο QSAR που θα μπορεί να καταλήξει σε συμπέρασμα για την τοξικότητα των MWCNTs λαμβάνοντας ως περιγραφικές μεταβλητές τις φυσικοχημικές ιδιότητες των νανοσωλήνων.

3.5 Γενικά χαρακτηριστικά των SPIONs

Τα τελευταία χρόνια οι ερευνητές εστιάζουν το ενδιαφέρον τους στα νανοσωματίδια ή νανοκρυσταλλίτες οξειδίου του σιδήρου. Η βιοσυμβατότητα που παρουσιάζουν σε συνδυασμό με την υπερπαραμαγνητική τους φύση, λόγω του μεγέθους τους (<20 nm), είναι σημαντικές ιδιότητες που αξιοποιούνται τόσο σε βιοϊατρικές εφαρμογές όπως η μαγνητική στόχευση φαρμάκων (Magnetic Drug Targeting, MDT) και η θεραπεία του καρκίνου μέσω επαγόμενης μαγνητικής υπερθερμίας (Magnetic Hyperthermia Targeting, MHT), όσο και σε τεχνικές απεικόνισης όπως η απεικόνιση μαγνητικού συντονισμού (Magnetic Resonance Imaging, MRI).[76] Ο μαγνητίτης (Fe_3O_4) και ο μαγκεμίτης (γ-Fe_2O_3), που θα μελετηθούν και στην παρούσα διπλωματική εργασία, είναι τα κυριότερα οξείδια του σιδήρου, ενώ ο αιματίτης (α-Fe_2O_3) και ο βουστίτης (FeO) δεν μελετώνται κλινικά εξ αιτίας των ασθενών μαγνητικών ιδιοτήτων τους.

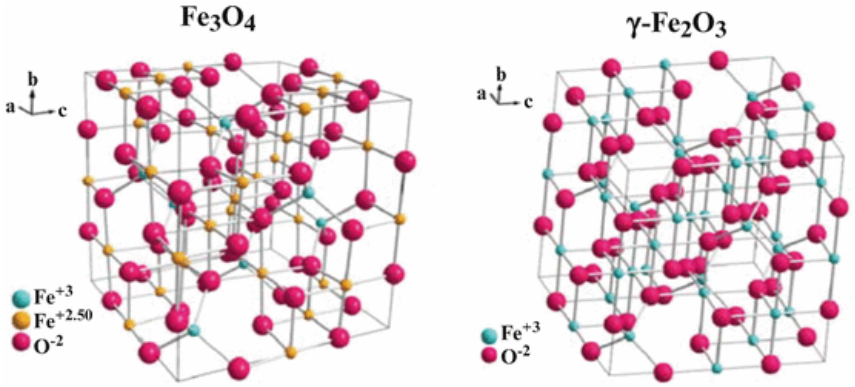
3.5.1 Μαγνητίτης (Magnetite)

Το πιο μαγνητικό ορυκτό που υπάρχει στη γη είναι ο μαγνητίτης (Fe_3O_4) που είναι ορυκτό του σιδήρου. Σε θερμοκρασία περιβάλλοντος είναι σιδηρομαγνητικό υλικό και έχει θερμοκρασία Curie (η θερμοκρασία αλλαγής από τη σιδηρομαγνητική συμπεριφορά στην παραμαγνητική) ίση με 805K. Ο άνθρωπος από την αρχαιότητα ανακάλυψε την έννοια του μαγνητισμού χάρη στην ιδιότητα που έχει ο μαγνητίτης να έλκει μικρά κομμάτια σιδήρου. Ο μαγνητίτης ανήκει σε μια αξιόλογη κατηγορία μεταλλικών οξειδίων, που ονομάζονται φερρίτες. Το σημαντικότερο δομικό χαρακτηριστικό του μαγνητίτη είναι ότι έχει ιόντα σιδήρου σε δισθενή και τρισθενή οξειδωτική βαθμίδα (εικόνα 3.2), και κατ' αυτόν τον τρόπο δύναται να λειτουργήσει και ως οξειδωτικό και ως αναγωγικό μέσο. Κρυσταλλώνεται σε ολοεδρία τριγωνικού σχήματος κατά το κυβικό σύστημα και έχει χρώμα σκούρο καφέ έως μαύρο και μεταλλική λάμψη. Απαντάται σε κοκκώδη και φλοιώδη συσσωματώματα με τη μορφή κόκκων με το όνομα «μαγνητίτης άμμος».[77]

Εικόνα 3.2 **Δομή Μαγνητίτη**

3.5.2 Μαγκεμίτης (Maghemite)

Ο μαγκεμίτης (γ-Fe_2O_3) έχει παρόμοια δομή με αυτή του μαγνητίτη, αλλά μόνο τα $^5/_6$ των συνολικών τετραεδρικών και οκταεδρικών θέσεων καταλαμβάνονται από άτομα σιδήρου και ανήκει στην οικογένεια των οξειδίων. Αποτελείται περίπου από $69.9\%wt.$ σίδηρο και $30.1\%wt.$ οξυγόνο.[77] Ο μηχανισμός οξείδωσης είναι αρκετά περίπλοκος. Η ομοιότητα του μαγνητίτη με τον μαγκεμίτη παρουσιάζεται στην εικόνα 3.3 και φαίνεται αν συλλογιστεί κανείς πως σε θερμοκρασία δωματίου, ακόμα, ο μαγνητίτης οξειδώνεται αργά σε μαγκεμίτη, οδηγώντας σε ασθενέστερο μαγνητισμό.[78] Το γεγονός ότι όλα τα κατιόντα σιδήρου του μαγκεμίτη βρίσκονται στην τρισθενή κατάσταση (Fe^{3+}) αποτελεί την βασική διαφορά του με τον μαγνητίτη.[77]



Εικόνα 3.3 **Ομοιότητα στην δομή του μαγνητίτη και του μαγκεμίτη**.

3.6 Παραμαγνητισμός και Υπερπαραμαγνητισμός

Ο μαγνητισμός είναι το φαινόμενο σύμφωνα με το οποίο, τόσο υλικά όσο και κινούμενα φορτισμένα σωματίδια ασκούν ελκτικές και απωθητικές δυνάμεις, καθώς και ροπές, σε άλλα υλικά ή φορτισμένα σωματίδια. Ένα άτομο χαρακτηρίζεται ως μαγνητικό, όταν έχει μη μηδενική μαγνητική ροπή. Κάποια γνωστά υλικά που έχουν εύκολα ανιχνεύσιμες μαγνητικές ιδιότητες (ονομάζονται μαγνήτες)

και είναι το νικέλιο, ο σίδηρος, το κοβάλτιο, το γαδολίνιο και τα κράματά τους. Όλα τα υλικά επηρεάζονται, σε μεγαλύτερο ή μικρότερο βαθμό, από την παρουσία ενός μαγνητικού πεδίου. Η ύπαρξη μαγνητικών φαινομένων αποδίδεται στην τροχιακή κίνηση και στο spin των ηλεκτρονίων.[79]

Οι μαγνητικές ροπές που σχετίζονται με τα ηλεκτρόνια και τα spin των παραμαγνητικών υλικών, μέσα σ' ένα άτομο δεν αλληλοαναιρούνται με την απουσία εξωτερικού μαγνητικού πεδίου. Κατά συνέπεια, κάθε άτομο έχει μια μικρή μαγνητική ροπή η διεύθυνση της οποίας όμως είναι τυχαία, γι' αυτό και η ολική μαγνητική ροπή μιας μεγάλης περιοχής δείγματος, όπως και η ολική μαγνήτιση, είναι μηδενικές απουσία εξωτερικού μαγνητικού πεδίου. Όταν εφαρμοστεί εξωτερικό πεδίο τότε τα μαγνητικά δίπολα προσανατολίζονται ελαφρώς κατά τη διεύθυνση του μαγνητικού πεδίου και δίνουν μία μη μηδενική μαγνήτιση παράλληλη προς αυτό. Τα παραμαγνητικά υλικά είναι αυτά που έλκονται ελαφρά από ένα μόνιμο μαγνήτη. Όπως ο διαμαγνητισμός, έτσι και ο παραμαγνητισμός είναι ένα αντιστρεπτό φαινόμενο. Στα παραμαγνητικά υλικά, οι μαγνητικές επιδεκτικότητες έχουν μικρές θετικές τιμές και οι σχετικές διαπερατότητες είναι λίγο μεγαλύτερες από τη μονάδα. Παράδειγμα παραμαγνητικών υλικών είναι το αλουμίνιο, το μαγγάνιο και ο λευκόχρυσος

Εξέλιξη του βασικού μαγνητισμού αποτελεί ο νανομαγνητισμός , με επίκεντρο τα σύγχρονα νανομαγνητικά υλικά και τα σχετικά με τις διαστάσεις τους φαινόμενα, όπως είναι ο υπερπαραμαγνητισμός, η σύζευξη ανταλλαγής και τα φαινόμενα υβριδισμού που οδηγούν σε εμφάνιση μαγνητικής ροπής σε μη μαγνητικά υλικά. Ο υπερπαραμαγνητισμός είναι μία μοναδική ιδιότητα που εμφανίζουν οι νανοδομές που είναι κατασκευασμένες από σιδηρομαγνητικό ή σιδηριμαγνητικό υλικό και βρίσκονται κάτω από ένα ορισμένο μέγεθος (10-20 nm). Ειδικά για την περίπτωση των οξειδίων του σιδήρου, όταν το μέγεθος αυτό παίρνει τιμές 20 nm εμφανίζουν σίδηρο- ή σιδηριμαγνητική συμπεριφορά. Σε ενδιάμεσες περιπτώσεις, η αύξηση του μεγέθους των νανοκρυσταλλίτων οδηγεί σε αύξηση των μαγνητικών ιδιοτήτων τους (αύξηση της μαγνήτισης κορεσμού).

3.7 Επιφανειακή τροποποίηση SPIONs

Η χημική τροποποίηση της επιφάνειας των μαγνητικών νανοσωματιδίων αποτελεί ένα σημαντικό στάδιο για την παρασκευή και τη χρήση τους, με στόχο να καταστούν όσο το δυνατόν λειτουργικότερα, στην εφαρμογή για την οποία προορίζονται. Η ενδεδειγμένη κατά περίπτωση, επιφανειακή τροποποίηση έχει ως αποτέλεσμα την ενίσχυση της κολλοειδούς τους σταθερότητας σε φυσιολογικές συνθήκες (pH, ιοντική ισχύς), την αύξηση της βιοσυμβατότητας των SPIONs, την ικανότητα εκλεκτικής προσκόλλησης σε συγκεκριμένες ουσίες-στόχους, τη δυνατότητα να αποφεύγουν βιολογικές διαδικασίες (φαγοκυττάρωση, οψωνινοποίηση) και να μεταφέρουν φαρμακευτικές ενώσεις. Ανάλογα με την επιθυμητή κάθε φορά εφαρμογή η παρουσία ενδεδειγμένων τερματικών χημικών ομάδων στην επιφάνεια των SPIONs, επιτρέπει την περαιτέρω ανάπτυξη και τροποποίηση των υβριδικών κολλοειδών. Ιδανικά ένα υβριδικό κολλοειδές για να δράσει με επιτυχία ως νανοφορέας, πρέπει η αρχιτεκτονική του να είναι τέτοια ώστε να προσδίδει στο σύστημα πολυλειτουργικότητα. Η επιφανειακή τροποποίηση είναι δυνατόν να πραγματοποιηθεί είτε κατά τη διάρκεια της σύνθεσης των νανοκρυσταλλιτών, με παρουσία κατάλληλων μονομερών ή πολυμερών, είτε μετά το τέλος της σύνθεσης. Η τροποποίηση της επιφάνειας των SPIONs μπορεί να γίνει άμεσα ή έμμεσα.

Η άμεση τροποποίηση περιλαμβάνει τη φυσική η χημική σύνδεση μορίων και πολυμερών στην επιφάνεια των μαγνητικών νανοσωματιδίων. Χαρακτηριστικά παραδείγματα πολυμερικών επιφανειακών τροποποιητών αποτελούν η πολυαιθυλενογλυκόλη (PEG), η δεξτράνη, η χιτοζάνη, τα αλγινικά και η πολυβινυλική αλκοόλη (PVA). Επίσης, μεταξύ άλλων, χρησιμοποιούνται τα πολυμερή: πολυακριλικό οξύ (PAA), πολυγαλακτικό οξύ (PLA), συμπολυμερές γαλακτικού-

γλυκολικού οξέος (PLGA) και πολυαιθυλενιμίνη (PEI).[80] Η PEG είναι το ευρύτερα χρησιμοποιούμενο πολυμερές, για το σκοπό αυτό, λόγω των αξιόλογων ιδιοτήτων της. Πρόκειται για ενα βιοσυμβατό και μη τοξικό πολυμερές, ενώ δεν επάγει ανοσοαποκρίσεις όσο άλλα πολυμερή. Λόγω του υδρόφιλου και μη ιοντικού χαρακτήρα της, προσδίδει μεγάλη κολλοειδή σταθερότητα, ενώ παρουσιάζει μικρότερη επιδεκτικότα για οψωνινοποίηση λόγω στερικής άπωσης με τα βιολογικά (μακρο)μόρια, με αποτέλεσμα την εξασφάλιση παρατεταμένου χρόνου κυκλοφορίας στο αίμα. Το γεγονός αυτό είναι ιδιαίτερα σημαντικό τόσο για τη παθητική αλλά και την ενεργητική στόχευση. Η τροποποίηση στην επιφάνεια των νανοσωματιδίων επιτυγχάνεται μέσω ομοιοπολικής ή ηλεκτροστατικής σύνδεσης. Σε άλλες περιπτώσεις, η πρόσδεση γίνεται μέσω χημικών ομάδων, οι οποίες αναπτύσσουν δεσμούς υδρογόνου με τις επιφάνειες των μαγνητικών νανοσωματιδίων.[81]

Η έμμεση τροποποίηση αφορά την ενκαψυλίωση των νανοσωματιδίων μέσα σε κατάλληλες διατάξεις όπως π.χ. τα λιποσώματα και τα μικκύλια. Ο εγκλωβισμός των μαγνητικών νανοσωματιδίων σε λιποσώματα, μικκύλια και τεχνητά καψίδια πολυηλεκτρολυτών αποτελεί μια ελπιδοφόρα προσέγγιση που μπορεί να λειτουργήσει ως εφαλτήριο για τη χρήση των SPIONs σε βιοϊατρικές εφαρμογές. Ένα παράδειγμα αυτής της προσέγγισης αποτελεί η μελέτη των J. Gao et al., οι οποίοι εγκλώβισαν επιτυχώς υδρόφοβα υπερπαραμαγνητικά νανοσωματίδια οξειδίου του σιδήρου (διαμέτρου 8nm) σε πολυμερικά μικκύλια (PLA-b-PEG), μαζί με το αντικαρκινικό φάρμακο Δοξορουβικίνη.[82]

3.8 Εφαρμογές και τοξικότητα SPIONs

Τα μαγνητικά νανοσωματίδια διαθέτουν, όπως έχει προαναφερθεί, ιδιότητες οι οποίες τα καθιστούν πολύτιμα εργαλεία στον τομέα της βιοϊατρικής. Αρχικά, έχουν ρυθμιζόμενο μέγεθος που κυμαίνεται από λίγα έως δεκάδες nm, δηλαδή μέγεθος συγκρινόμενο με εκείνο του κυττάρου, του ιού ή ενός γονιδίου. Η κύρια όμως ιδιότητα των μαγνητικών νανοκρυσταλλιτών, η οποία τους καθιστά πολύτιμο εργαλείο για χρήση στις βιοϊατρικές εφαρμογές, είναι ο μαγνητισμός. Χαρακτηριστικό παράδειγμα αποτελεί η συνδυαστική τεχνική διάγνωσης-θεραπείας (Theranostics) του καρκίνου με τη χρήση SPIONs. Η διάγνωση επιτυγχάνεται με τη τεχνική της μαγνητικής τομογραφίας πυρηνικού συντονισμού (Magnetic Resonance Imaging, MRI) και η θεραπεία μέσω επαγόμενης υπερθερμίας, ως αποτέλεσμα της έκθεσης σε εναλλασσόμενο μαγνητικό πεδίο (μαγνητική υπερθερμία). Μια ακόμα σπουδαία εφαρμογή αφορά τη δυνατότητα *in vivo* κατεύθυνσης των SPIONs στην επιθυμητή περιοχή-στόχο με τη χρήση εξωτερικού μαγνήτη (μαγνητική στόχευση), με σκοπό είτε την πρόκληση μαγνητικής υπερθερμίας είτε τη στοχευμένη μεταφορά φαρμακευτικών ουσιών στην πάσχουσα περιοχή. Τέλος, τα μαγνητικά νανοσωματίδια χρησιμοποιούνται και σε *in vitro* εφαρμογές όπως: ο μαγνητικός διαχωρισμός κυττάρων, η επισήμανση και απομόνωση DNA/RNA, η ανίχνευση και απομόνωση μικροοργανισμών.[83,84,85]

Όπως γίνεται άμεσα αντιληπτό, τα SPIONs έχουν αρχίσει και διαδραματίζουν ένα μεγάλο ρόλο στον τομέα της βιοϊατρικής. Η επαφή τους με τον ανθρώπινο οργανισμό είναι άμεση κι επομένως, αν και η γενική άποψη είναι πως τα υλικά αυτά, ως επί τω πλείστων, δεν είναι τοξικά ορισμένες μελέτες έχουν δείξει πως μπορούν να προκαλέσουν λύση των ανθρώπινων κυττάρων. Για τον λόγο αυτό μια εμπεριστατωμένη μέθοδος που να εξετάζει και να προβλέπει την τοξικότητα των νανοσωματιδίων αυτών αποτελεί ανάγκη για τους σκοπούς της σύγχρονης ιατρικής.

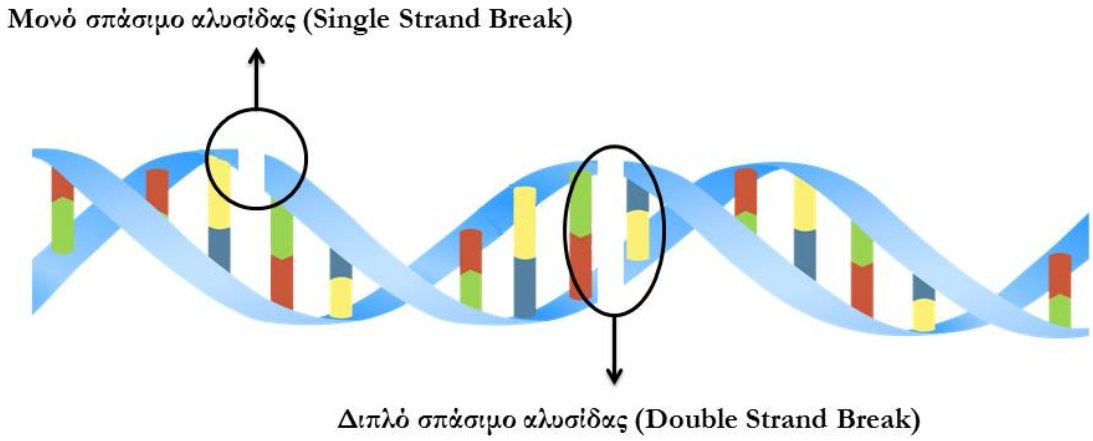
Κεφάλαιο 4

Μεθοδολογία και Αποτελέσματα

Όπως προκύπτει από το Κεφάλαιο 3, η ανάγκη μελέτης της τοξικότητας των MWCNTs και των SPIONs είναι μεγάλη, καθώς ο ανθρώπινος οργανισμός έρχεται σε συχνή επαφή με αυτά. Η ανάγκη αυτή μπορεί να ικανοποιηθεί με την ανάπτυξη ενός QSAR μοντέλου για το κάθε νανοϋλικό, χρησιμοποιώντας την θεωρία της μηχανικής μάθησης. Στο παρόν κεφάλαιο θα παρουσιαστεί αναλυτικά η μεθοδολογία που ακολουθήθηκε για την εκπαίδευση των QSAR μοντέλων που προβλέπουν την τοξικότητα των νανοϋλικών καθώς και τα αποτελέσματα των μοντέλων αυτών. Τα μοντέλα και οι αλγόριθμοι αυτής της διπλωματικής εργασίας αναπτύχθηκαν σε γλώσσα προγραμματισμού Python (version 3.6) και στο Jupyter notebook προγραμματιστικό περιβάλλον.[86,87]

4.1 Γενική μεθοδολογία για τα MWCNTs

Σύμφωνα με τις τελευταίες μελέτες, οι νανοσωλήνες άνθρακα παρουσιάζουν χαμηλή κυτταροτοξικότητα (cytotoxicity), ωστόσο, έχουν παρουσιαστεί αρκετές περιπτώσεις όπου έχει επηρεαστεί το γενετικό υλικό των οργανισμών. Για την ακρίβεια, αρκετές έρευνες έχουν δείξει πως η ύπαρξη MWCNTs μπορεί να οδηγήσει την DNA αλυσίδα να σπάσει (strand breaks), όπως φαίνεται στην εικόνα 4.1. Αυτή η τοξική συμπεριφορά ονομάζεται γονοτοξικότητα (genotoxicity).



Εικόνα 4.1 **Παραδείγματα σπασμένης αλυσίδας DNA**

Στην παρούσα εργασία αποφασίσαμε να εξετάσουμε την τοξικότητα των MWCNTs ως προς την γονοτοξικότητά τους. Τα δεδομένα που χρησιμοποιήθηκαν αντλήθηκαν από μελέτες που έχουν γίνει, τόσο *in vitro* όσο και *in vivo*, σε κύτταρα ποντικών. Κατόπιν, τα δεδομένα αποθηκεύτηκαν σε ένα σετ δεδομένων το οποίο χρησιμοποιήθηκε για την ανάλυση. Για την πιο αποτελεσματική ανάλυση των δεδομένων ως προς την γονοτοξικότητα των MWCNTs, θεωρήθηκαν και αναπτύχθηκαν δύο διαφορετικές μεθοδολογίες. Στην πρώτη μεθοδολογία εφαρμόστηκε στα δεδομένα η PCA με στόχο την ελάττωση των features του σετ, την αποφυγή θορύβου των πειραματικών μετρήσεων και τον εντοπισμό των συσχετίσεων που υπάρχουν, ενώ στη δεύτερη κάτι τέτοιο δεν συνέβη, με στόχο να υπολογισθούν οι πιο σημαντικές παράμετροι για την λήψη της απόφασης. Και στις δύο περιπτώσεις, εκπαιδεύθηκαν και βελτιστοποιήθηκαν μοντέλα προβλεπτικής μάθησης και στο τέλος της διαδικασίας επιλέχθηκε το μοντέλο που είχε τη βέλτιστη συμπεριφορά, ενώ υπολογίσθηκε και το πεδίο εφαρμογής του.

4.2 Κατασκευή του MWCNTs σετ δεδομένων

Τα δεδομένα που χρησιμοποιήθηκαν για την κατασκευή του σετ δεδομένων προήλθαν από μελέτες που συνδύαζαν την γονοτοξικότητα των νανοσωλήνων με διάφορες φυσικοχημικές τους ιδιότητες.[88,89,90,91] Το σετ δεδομένων που σχηματίστηκε (πίνακας 4.1) απαρτίζεται από 15 παρατηρήσεις (samples) και 50 μεταβλητές (features). Ιδιαίτερη έμφαση δόθηκε στην μελέτη της Jackson και των συνεργατών της και της Poulsen και των συνεργατών της, καθώς στις έρευνες αυτές παρουσιάστηκε ένα πλήθος φυσικοχημικών ιδιοτήτων για τους νανοσωλήνες που εξετάστηκαν.[89,90]

Σύμφωνα με προηγούμενες μελέτες, δέκα από τους δεκαπέντε νανοσωλήνες που επιλέχθηκαν έχουν κατηγοριοποιηθεί σε τρείς κατηγορίες (Group I - III) σύμφωνα με τις φυσικοχημικές τους ιδιότητες. Οι κατηγορίες αυτές είναι:

- Παχείς (thick), Group I
- Λεπτοί (thin), Group II
- Κοντοί (short), Group III

Κάθε μία από αυτές τις κατηγορίες περιλαμβάνει τρία δείγματα με διαφορετική επιφανειακή τροποποίηση, έναν καθαρό MWCNT, έναν MWCNT με λειτουργική ομάδα –OH και έναν με λειτουργική ομάδα –COOH. Στο Group III περιέχει έναν ακόμα νανοσωλήνα ο οποίος διαθέτει στην επιφάνεια του $–NH_2$ λειτουργική ομάδα. Τα πέντε δείγματα που απομένουν δεν είναι χημικώς τροποποιημένα και στην βιβλιογραφία ονομάζονται κανονικά υλικά (standard materials).[90]

Το τελικό σημείο (endpoint) της ανάλυσης θεωρήθηκε η γονοτοξικότητα και θεωρήθηκε πρόβλημα δυαδικής ταξινόμησης. Πιο συγκεκριμένα, τα μοντέλα που αναπτύχθηκαν αντιμετώπισαν την γονοτοξικότητα ως μια διακριτή μεταβλητή με δύο κλάσεις, την κλάση «0» για τους μη τοξικούς νανοσωλήνες και την κλάση «1» για τους τοξικούς. Τα δεδομένα της γονοτοξικότητας, που επιλέχθηκαν, μετρήθηκαν τόσο σε *in vivo* μελέτες, σύμφωνα με τις διασπάσεις της αλυσίδας DNA, όσο και σε *in vitro* μελέτες σύμφωνα με την γενετική μετάλλαξη κυττάρων.

Sample	Type	Length (nm)	Purity (%)	Zeta Potential	PdI	…	Genotoxicity
NRCWE- 040	PRISTINE	518.9	98.60	195.6	0.358	…	0
NRCWE- 041	OH	1005	99.20	218.7	0.501	…	0
NRCWE- 042	COOH	723.2	99.20	195.2	0.382	…	0
NRCWE- 043	PRISTINE	771.3	98.50	177.0	0.237	…	0
NRCWE- 044	OH	1330	98.60	202.2	0.236	…	1
NRCWE- 045	COOH	1553	96.30	192.4	0.227	…	1
NRCWE- 046	PRISTINE	717.2	98.70	182.9	0.434	…	0
NRCWE- 047	OH	532.5	98.70	200.1	0.402	…	0
NRCWE- 048	COOH	1604	98.80	182.2	0.397	…	0
NRCWE- 049	NH2	731.1	98.80	212.5	0.602	…	0
NM-400	PRISTINE	847	90.00	168.1	0.385	…	1
NM-401	PRISTINE	4048	99.19	923.8	0.323	…	0
NM-402	PRISTINE	1372	92.97	171.3	0.407	…	1
NM-403	PRISTINE	1373	90.00	200.2	0.407	…	1
NRCWE-006	PRISTINE	5700	99.00	633.3	0.256	…	1

Πίνακας 4.1 Δείγμα του σετ δεδομένων που κατασκευάστηκε. Κάθε γραμμή συμβολίζει τις συνιστώσες του διανύσματος που χαρακτηρίζει κάθε δείγμα. Τα υλικά μέσα στο κόκκινο πλαίσιο ανήκουν στο Group I, τα υλικά του μαύρου πλαισίου ανήκουν στο Group II και τα υλικά εντός του μωβ πλαισίου ανήκουν στο Group III, ενώ τα υπόλοιπα είναι τα κανονικά υλικά.

Όπως αναφέρθηκε και στο Κεφάλαιο 2, η αναπαράσταση των δεδομένων γίνεται με την θεώρηση διανυσμάτων για το κάθε στοιχείο (ή δείγμα). Οι συνιστώσες των διανυσμάτων είναι οι τιμές των στοιχείων για την κάθε ιδιότητα. Στο σετ δεδομένων τα διανύσματα τοποθετούνται στις οριζόντιες γραμμές του πίνακα ενώ οι στήλες συμβολίζουν τις ιδιότητες. Στον πίνακα παρουσιάζεται ένα δείγμα του σετ δεδομένων που κατασκευάστηκε για το πρόβλημα των MWCNTs, ενώ το ολοκληρωμένο σετ δεδομένων παρουσιάζεται στον πρώτο πίνακα του Παραρτήματος Β.

4.3 Προεπεξεργασία δεδομένων των MWCNTs

Η διαδικασία προεπεξεργασίας των δεδομένων αποτελεί ένα από τα πιο σημαντικά βήματα για την εκπαίδευση μοντέλων. Παρά την υψηλή ποιότητα των δεδομένων που συλλέχθηκαν, χρειάστηκαν ορισμένες παρεμβάσεις στο σετ δεδομένων των MWCNTs ώστε να μπορεί να εκπαιδευθεί ορθά ένα μοντέλο σε αυτά. Πιο συγκεκριμένα, χρειάστηκε να υπολογισθούν οι τιμές της μεταβλητής «Endotoxins (EU/mg)» για τα υλικά NRCWE-041 και NM-400, καθώς αυτές έλειπαν. Όπως φαίνεται και από τον πίνακα 4.1, το NRCWE-041 ανήκει στο Group I ενώ το NM-400 στα κανονικά νανοϋλικά. Επομένως, η κάλυψη του κενού για το NRCWE-041 έγινε με τον μέσο όρο των τιμών των ενδοτοξίνων για τα υλικά που ανήκουν στα Groups I-III, ενώ η τιμή του NM-400 υπολογίσθηκε ως ο μέσος όρος των τιμών των ενδοτοξίνων των κανονικών υλικών. Επιπρόσθετα, όπως φαίνεται κι από τον πίνακα Β1 στο Παράρτημα Β, οι νανοσωλήνες είχαν ορισμένες προσμίξεις στα εξής μέταλλα:

- Σίδηρος (Fe)
- Κοβάλτιο (Co)
- Νικέλιο (Ni)
- Μαγνήσιο (Mg)
- Μαγγάνιο (Mn)

Στα πλαίσια της ανάλυσης για την τοξικότητά τους, οι προσμίξεις αυτές συγκροτούν ένα ιδιαιτέρως ενδιαφέρον σημείο μελέτης, καθώς είναι πιθανό η συγκέντρωση αυτών να επιδρά στην γονοτοξικότητα. Ωστόσο, δεν υπήρχαν δεδομένα για όλους τους τύπους των προσμίξεων. Στα πλαίσια αυτά, δημιουργήθηκε μία νέα μεταβλητή, η «% Total Impurities», η οποία ορίστηκε ως το άθροισμα των ποσοστιαίων προσμίξεων μετάλλων στους MWCNTs, ενώ οι στήλες των επί μέρους προσμίξεων αφαιρέθηκαν από το σετ δεδομένων.

Τελευταίο βήμα αυτής της διαδικασίας «καθαρισμού» του σετ δεδομένων ήταν η αφαίρεση ορισμένων στηλών που θεωρήθηκαν περιττές για την πρόβλεψη της τελικής τιμής της ανάλυσης, δηλαδή της γονοτοξικότητας. Ειδικότερα, οι στήλες με δεδομένα του κατασκευαστή για το μήκος των νανοσωλήνων θεωρήθηκαν αμελητέας σημασίας καθώς χρησιμοποιήθηκε το μήκος που μετρήθηκε πειραματικά για τους νανοσωλήνες. Ακόμα, τα περιγραφικά δεδομένα για τα πειράματα που διεξήχθησαν για την μέτρηση της τοξικότητας, θεωρήθηκε πως δεν επηρεάζουν την πρόβλεψη της γονοτοξικότητας των MWCNTs, αφού για όλα τα νανοϋλικά (εκτός από ένα) είχαν την ίδια τιμή. Επομένως, η ελάχιστη διακύμανση αυτής της μεταβλητής συνεπάγεται την αδυναμία της να εντοπίσει μοτίβα στα δεδομένα. Για τον λόγο αυτό, αφαιρέθηκε και η στήλη με τα ποσοστά αναπαραγωγής των κυττάρων σε ελάχιστη δόση (σχεδόν $0\ \mu g/ml$) MWCNTs, αφού ήταν 100% για κάθε δείγμα νανοσωλήνα. Αντίστοιχα, αφαιρέθηκαν όλες οι στήλες με τις τυπικές αποκλίσεις των πειραματικών μετρήσεων καθώς οι μεταβλητές αυτές δεν έχουν κάποια συσχέτιση με την τοξικότητα των δειγμάτων κι επομένως δεν θα πρέπει να ληφθούν υπόψιν.

Έχοντας υλοποιήσει τις απαραίτητες παρεμβάσεις, η διαστατικότητα του σετ δεδομένων ελαττώθηκε ιδιαιτέρως, καθώς από 50 μεταβλητές καταλήξαμε σε 31. Οι μεταβλητές αυτές καθώς και η περιγραφή τους παρουσιάζονται στον πίνακα 4.2.

Μεταβλητές (Features)	Περιγραφή (Description)	Τύπος Μεταβλητής
Type	Τύπος λειτουργικής ομάδας	Διακριτή – Λεκτική
Length ave. (nm)	Μέσο μήκος , εκφρασμένο σε nm	Συνεχής
Diameter ave. (nm)	Μέση διάμετρος, εκφρασμένη σε nm	Συνεχής
Diameter max (nm)	Μέγιστη διάμετρος, εκφρασμένη σε nm	Συνεχής
Diameter min (nm)	Ελάχιστη διάμετρος, εκφρασμένη σε nm	Συνεχής
BET (m^2/g)	Εμβαδό επιφάνειας νανοσωλήνων, υπολογισμένο σύμφωνα με την BET μέθοδο.	Συνεχής
% Total Impurities	Ποσοστό προσμίξεων μετάλλων	Συνεχής
Purity (%)	Ποσοστό καθαρότητας MWCNTs	Συνεχής
Zave (batch)	Μέσο Zeta για batch διασπορά του μέσου	Συνεχής
PdI (batch)	Δείκτης πολυδιασποράς για batch διασπορά του μέσου	Συνεχής
Zave (12.5 ug/ml)	Μέσο Zeta για $12.5\,{}^{\mu g_{MWCNTs}}/_{ml}$	Συνεχής
PdI (12.5 ug/ml)	Δείκτης πολυδιασποράς για $12.5\,{}^{\mu g_{MWCNTs}}/_{ml}$	Συνεχής
Zave (200 ug/ml)	Μέσο Zeta για $200\,{}^{\mu g_{MWCNTs}}/_{ml}$	Συνεχής
PdI (200 ug/ml)	Δείκτης πολυδιασποράς για $200\,{}^{\mu g_{MWCNTs}}/_{ml}$	Συνεχής
ROS	Ικανότητα παραγωγής αντιδραστικών ειδών οξυγόνου	Συνεχής
Peak (ug/ml)	Συγκέντρωση MWCNTs ($\mu g/ml$) με μέγιστες OD τιμές	Συνεχής
0 (ug/ml)_via	Ποσοστό βιωσιμότητας σε συγκέντρωση κοντά στα $0\,{}^{\mu g_{MWCNTs}}/_{ml}$	Συνεχής
12.5 (ug/ml)_via	Ποσοστό βιωσιμότητας σε συγκέντρωση $12.5\,{}^{\mu g_{MWCNTs}}/_{ml}$	Συνεχής
25 (ug/ml)_via	Ποσοστό βιωσιμότητας σε συγκέντρωση $25\,{}^{\mu g_{MWCNTs}}/_{ml}$	Συνεχής
50 (ug/ml)_via	Ποσοστό βιωσιμότητας σε συγκέντρωση $50\,{}^{\mu g_{MWCNTs}}/_{ml}$	Συνεχής
100 (ug/ml)_via	Ποσοστό βιωσιμότητας σε συγκέντρωση $100\,{}^{\mu g_{MWCNTs}}/_{ml}$	Συνεχής
200 (ug/ml)_via	Ποσοστό βιωσιμότητας σε συγκέντρωση $200\,{}^{\mu g_{MWCNTs}}/_{ml}$	Συνεχής
12.5 (ug/ml)_prolif	Ποσοστό αναπαραγωγής σε συγκέντρωση $12.5\,{}^{\mu g_{MWCNTs}}/_{ml}$	Συνεχής
25 (ug/ml)_prolif	Ποσοστό αναπαραγωγής σε συγκέντρωση $25\,{}^{\mu g_{MWCNTs}}/_{ml}$	Συνεχής
50 (ug/ml)_prolif	Ποσοστό αναπαραγωγής σε συγκέντρωση $50\,{}^{\mu g_{MWCNTs}}/_{ml}$	Συνεχής
100 (ug/ml)_prolif	Ποσοστό αναπαραγωγής σε συγκέντρωση $100\,{}^{\mu g_{MWCNTs}}/_{ml}$	Συνεχής
200 (ug/ml)_prolif	Ποσοστό αναπαραγωγής σε συγκέντρωση $200\,{}^{\mu g_{MWCNTs}}/_{ml}$	Συνεχής
CEA: C.H.N.O (wt%)	Κατά βάρος συγκέντρωση των στοιχείων C, H, N και O	Συνεχής
OH mmol/g	Συγκέντρωση OH σε $mmol/g$	Συνεχής
COOH mmol/g	Συγκέντρωση COOH σε $mmol/g$	Συνεχής
Endotoxins (EU/mg)	Συγκέντρωση ενδοτοξίνων σε EU/mg	Συνεχής

Πίνακας 4.2 **Πίνακας μεταβλητών του σετ δεδομένων των MWCNTs**

Στην ανάλυση της προεπεξεργασίας στην παράγραφο 2.3 του κεφαλαίου 2 αναφέρθηκε πως για την εκπαίδευση μοντέλων απαιτείται όλες οι μεταβλητές να είναι αριθμητικές. Ωστόσο, όπως φαίνεται από τον πίνακα 4.2, η μεταβλητή Type είναι διακριτή και λεκτική κι επομένως οι τιμές της θα πρέπει να μετατραπούν σε αριθμητικές. Για τον σκοπό αυτό, εφαρμόστηκε η μέθοδος «One-Hot Encoding» σύμφωνα με την οποία η λεκτική αναπαράσταση των διακριτών τιμών μετατρέπεται σε αριθμητική, αποτρέποντας την πιθανότητα τυχαίας διάταξης. Ειδικότερα για την μέθοδο αυτή, για κάθε λεκτική τιμή που μπορεί να πάρει η μεταβλητή Type, δημιουργείται μία στήλη (ή μία νέα μεταβλητή) στο σετ δεδομένων που μπορεί να πάρει τιμή «0» ή «1». Κατόπιν, κάθε δείγμα έχει την τιμή «1» μόνο στη στήλη που συμβολίζει τη λειτουργική ομάδα που διαθέτει και την τιμή «0» στις υπόλοιπες (εικόνα 4.2). Η μέθοδος αυτή εξασφαλίζει πως δεν θα αντιμετωπισθούν με διαφορετικό στατιστικό βάρος οι διακριτές αυτές τιμές, ωστόσο αυξάνει την διαστατικότητα του σετ δεδομένων. Δεδομένου του μικρού όγκου δεδομένων που είχαμε στη διάθεσή μας και των τεσσάρων πιθανών τιμών της μεταβλητής Type (OH, COOH, NH2, PRISTINE), επιλέξαμε να εφαρμόσουμε αυτήν

Type
PRISTINE
OH
COOH
NH2
PRISTINE
COOH
OH
⋮

«One-Hot» Encode

Type PRISTINE	Type OH	Type COOH	Type NH2
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
1	0	0	0
0	0	1	0
0	1	0	0
⋮	⋮	⋮	⋮

Εικόνα 4.2 **Παράδειγμα διαδικασίας «One–Hot» Encoding**

την μέθοδο.

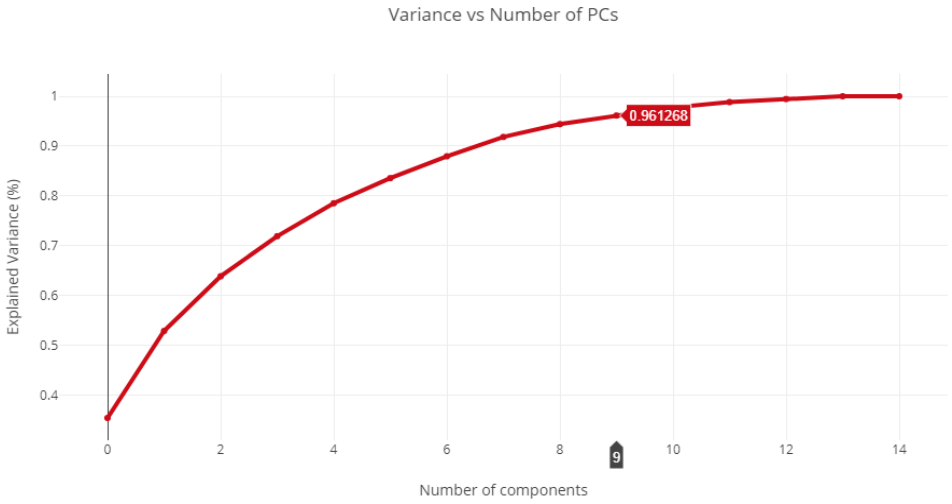
Παρατηρώντας τον πίνακα 4.1 φαίνεται πως μεταξύ των τιμών διαφορετικών μεταβλητών υπάρχει μία έντονη διαφορά στην τάξη μεγέθους. Για παράδειγμα, οι τιμές της στήλης του δείκτη της πολυδιασποράς δεν ξεπερνούν την μονάδα ενώ οι τιμές του μήκους μπορούν να ξεπεράσουν τα $5000\ nm$. Αυτή η συνθήκη μπορεί να έχει τραγικές επιπτώσεις στο μοντέλο που πρόκειται να κατασκευασθεί κι επομένως τα δεδομένα θα πρέπει να κανονικοποιηθούν. Για τον σκοπό αυτό επιλέχθηκε η μέθοδος «min-max scaling» σύμφωνα με την οποία κάθε τιμή x_i της στήλης x αντικαθίσταται από την τιμή z_i, σύμφωνα με την εξίσωση 4.1:

$$z_i = \frac{x_i - min(x)}{max(x) - min(x)} \quad (4.1)$$

Το τελικό σετ δεδομένων που προέκυψε από την ανωτέρω διαδικασία παρουσιάζεται στο Παράρτημα Β (πίνακας Β2). Όπως προαναφέρθηκε στην αρχή του κεφαλαίου (§ 4.1) για τους σκοπούς της ανάλυσης, για την περίπτωση των MWCNTs, αναπτύχθηκαν δύο μεθοδολογίες, μία όπου εφαρμόστηκε η PCA και μία όπου δεν εφαρμόστηκε. Στην δεύτερη περίπτωση, το κομμάτι της προεπεξεργασίας σταματάει σε αυτό το σημείο. Από εδώ και κάτω, μέχρι το τέλος αυτής της παραγράφου, τα υπόλοιπα βήματα προεπεξεργασίας αφορούν την περίπτωση όπου τα δεδομένα μετασχηματίστηκαν με εφαρμογή της PCA και τα μοντέλα αναπτύχθηκαν στο μετασχηματισμένο σετ δεδομένων. Για να είναι ξεκάθαρες οι μεθοδολογίες και τα βήματα που ακολουθήθηκαν στην κάθε περίπτωση, μέχρι το τέλος του κεφαλαίου αυτού θα αναφερόμαστε στο μετασχηματισμένο σετ

δεδομένων με τον όρο «σετ δεδομένων κύριων συνιστωσών» ενώ στο μη τροποποιημένο σετ δεδομένων θα αναφερόμαστε με τον όρο «κανονικοποιημένο σετ δεδομένων».

Ένας από τους κυριότερους λόγους που οδήγησαν στην εφαρμογή της PCA είναι το γεγονός πως στην συντριπτική πλειοψηφία τους, τα δεδομένα για τους νανοσωλήνες προέρχονται από πειραματικές μετρήσεις. Τέτοιου τύπου δεδομένα, ωστόσο, είναι δυνατό να έχουν μεγάλα σφάλματα, τα οποία σε μία υπολογιστική διαδικασία μεταφράζονται σε «θόρυβο». Η Ανάλυση Κύριων Συνιστωσών δίνει την δυνατότητα να μειωθεί ο θόρυβος (έως και να εξαλειφθεί) καθώς οι τελικές διαστάσεις του μετασχηματισμένου σετ δεδομένων επιλέγονται σύμφωνα με το ποσοστό της διακύμανσης ή της πληροφορίας που θέλουμε να κρατήσουμε. Λόγω του μικρού όγκου δεδομένων, η επιλογή του αριθμού των Κύριων Συνιστωσών ήταν τέτοια ώστε να μην χαθεί πληροφορία, αλλά μόνο ο πιθανός θόρυβος. Όπως φαίνεται και στο διάγραμμα της εικόνας 4.3, διατηρώντας εννέα Κύριες Συνιστώσες, διατηρείται περίπου το 96% της αρχικής πληροφορίας. Το ποσοστό αυτό είναι αρκετά υψηλό, εξασφαλίζοντας έτσι πως δεν χάνεται χρήσιμη πληροφορία από το σετ δεδομένων. Το σετ δεδομένων κυρίων συνιστωσών παρουσιάζεται στον πίνακα Β3 του Παραρτήματος Β.



Εικόνα 4.3 Ποσοστό της διακύμανσης που διατηρείται συναρτήσει των αριθμό των Κύριων Συνιστωσών κατά την εφαρμογή της PCA

4.4 Εκπαίδευση Προβλεπτικών Μοντέλων και Βελτιστοποίηση

Ένα κρίσιμο βήμα πριν την εκπαίδευση ενός μοντέλου για την πρόβλεψη της γονοτοξικότητας των MWCNTs είναι η «train-test split» διαδικασία, ώστε να διαχωριστεί το σετ δεδομένων στο εκπαιδευτικό σετ και στο επαληθευτικό σετ. Ο διαχωρισμός αυτός έγινε σύμφωνα με την μέθοδο των Kennard-Stone, καθώς η μέθοδος αυτή εξασφαλίζει πως δεν θα υπάρχουν έκτροπες τιμές στο επαληθευτικό σετ δεδομένων και με έναν τρόπο αυξάνει το πεδίο εφαρμογής του μοντέλου.[39] Αυτά τα δύο στοιχεία της μεθόδου ήταν ιδιαιτέρως σημαντικά τόσο για το σετ δεδομένων κύριων συνιστωσών των MWCNTs όσο και για το κανονικοποιημένο σετ δεδομένων, καθώς η έλλειψη περισσότερων δεδομένων καθιστά το μοντέλο που θα αναπτυχθεί ευαίσθητο σε έκτροπες τιμές. Η Kennard-Stone εφαρμόστηκε, αναζητώντας τους πέντε νανοσωλήνες που θα σχημάτιζαν το επαληθευτικό σετ δεδομένων κύριων συνιστωσών και το επαληθευτικό κανονικοποιημένο σετ δεδομένων αντίστοιχα. Αυτοί οι MWCNTs είναι οι:

- NRCWE – 040
- NRCWE – 045
- NRCWE – 048
- NM – 401
- NM – 402

Οι νανοσωλήνες αυτοί επιλέχθηκαν από την Kennard-Stone τόσο για το επαληθευτικό σετ δεδομένων κύριων συνιστωσών όσο και για το επαληθευτικό κανονικοποιημένο σετ δεδομένων. Το γεγονός αυτό είναι αναμενόμενο καθώς το σετ δεδομένων που κατασκευάσθηκε από την εφαρμογή της PCA διαθέτει την ίδια πληροφορία με το κανονικοποιημένο σετ δεδομένων χωρίς τον θόρυβο. Επομένως, είναι λογικό τα δείγματα που απαρτίζουν τα επαληθευτικά σετ δεδομένωνs να είναι κοινά. Τα δέκα υπόλοιπα στοιχεία συγκροτούν τα εκπαιδευτικά σετ δεδομένωνs.

Έχοντας διαχωρίσει τα σετ δεδομένων σε εκπαιδευτικά και επαληθευτικά, επόμενο βήμα της ανάλυσης είναι η εκπαίδευση των μοντέλων για τις δύο μεθοδολογίες (με PCA, χωρίς PCA). Τα μοντέλα που επιλέχθηκαν για να εκπαιδευθούν στα δεδομένα των MWCNTs παρουσιάζονται στον πίνακα 4.3. Σχεδόν όλα τα μοντέλα, όπως αναλύθηκε εκτενώς στο Κεφάλαιο 2, διαθέτουν ορισμένες παραμέτρους (υπέρ-παραμέτρους) οι οποίες στοχεύουν στην αποδοτικότερη εκπαίδευση των μοντέλων στα δεδομένα. Για τη ρύθμιση αυτών των υπέρ-παραμέτρων χρησιμοποιήθηκε η τεχνική της Μπεϋζιανής βελτιστοποίησης. Ειδικότερα, χρησιμοποιήθηκε το πακέτο Bayes_Opt της Python, όπου χρειάστηκε να τροποποιηθεί για τους σκοπούς της παρούσας διπλωματικής εργασίας. Το πακέτο Bayes_Opt είναι η υλοποίηση της μεθόδου που περιγράφηκε στην παράγραφο 2.7 και με την χρήση του είναι δυνατός ο υπολογισμός των ορισμάτων μίας άγνωστης συνάρτησης, για τα οποία η συνάρτηση αυτή μεγιστοποιείται. Στα πλαίσια αυτής της διπλωματικής εργασίας, χρειάστηκε να υλοποιηθεί μία επέκταση του πακέτου αυτού, ώστε να εφαρμόζεται στις μεθόδους που χρησιμοποιήθηκαν. Πιο συγκεκριμένα, κάθε μοντέλο θεωρήθηκε ως μία άγνωστη συνάρτηση πολλών μεταβλητών, η τιμή της οποίας αντιπροσωπεύει την ακρίβεια του μοντέλου (accuracy score), ενώ οι μεταβλητές, τις υπέρ-παραμέτρους του. Επομένως, με την επέκταση αυτή ήταν δυνατός ο προσδιορισμός των υπέρ-παραμέτρων έτσι ώστε η ακρίβεια του μοντέλου να γίνεται μέγιστη.

Μοντέλα που εκπαιδεύτηκαν στο κανονικοποιημένο σετ δεδομένων	**Μοντέλα που εκπαιδεύτηκαν στο σετ δεδομένων κύριων συνιστωσών**
SVM[37]	**SVM[37]**
RF[37,38]	**RF[37,38]**
LR[38]	**LR[38]**
Gaussian NB[38]	**Gaussian NB[38]**
PLS[20]	**-**

Πίνακας 4.3 Μοντέλα που εφαρμόστηκαν στα διαφορετικά σετ δεδομένωνs

Για την εφαρμογή της μπεϋζιανής βελτιστοποίησης απαραίτητος είναι ο ορισμός παραμέτρων που χαρακτηρίζουν την διαδικασία. Στην ανάλυση της μεθόδου, στο Κεφάλαιο 2, αναφέρθηκε πως κίνητρο για τη λειτουργία της μεθόδου και τον προσδιορισμό του επόμενου σημείου είναι η ρύθμιση της ανταλλαγής «εξερεύνησης-εκμετάλλευσης». Επομένως, κατά την εφαρμογή της θα πρέπει να ορισθούν οι παράμετροι που χαρακτηρίζουν αυτήν την ρύθμιση. Συγκεκριμένα ορίζονται το πλήθος των «σημείων εξερεύνησης» καθώς και το πλήθος των «σημείων

εκμετάλλευσης», δηλαδή το πόσες φορές επιτρέπεται στον αλγόριθμο να ερευνήσει μία «ανεξερεύνητη» περιοχή και το πόσες φορές να «επιμείνει» στην περιοχή πού φαίνεται να έχει υψηλές τιμές. Παράλληλα, ορισμένα από τα μοντέλα που τέθηκαν υπό βελτιστοποίηση διαθέτουν αρκετές υπέρ-παραμέτρους. Ο προσδιορισμός όλων αυτών, πέρα από το υψηλό κόστος των υπολογισμών, ενέχει τον κίνδυνο να υπέρ-προσαρμοστεί το μοντέλο στα εκπαιδευτικά δεδομένα. Για τον λόγο αυτό, δίνονται συγκεκριμένες υπέρ-παράμετροι για προσδιορισμό στον αλγόριθμο, αυτές που θεωρήθηκαν σημαντικές για την ορθή εκπαίδευση του μοντέλου.

4.4.1 Βελτιστοποίηση και εκπαίδευση στο σετ δεδομένων κύριων συνιστωσών

Όπως φαίνεται στον πίνακα 4.3, στο σετ δεδομένων κύριων συνιστωσών εφαρμόστηκε ένα λιγότερο μοντέλο. Ειδικότερα, δεν εκπαιδεύτηκε η PLS στο σετ δεδομένων αυτό καθώς όπως εξηγήθηκε στο Κεφάλαιο 1, η PLS είναι ένας συνδυασμός της PCA με την MLRA, επομένως ο μετασχηματισμός δεδομένων που έχουν ήδη μετασχηματιστεί, από την εφαρμογή της PCA, δεν είχε κάποιο νόημα και για τον λόγο αυτό και δεν εφαρμόστηκε.

Για την βελτιστοποίηση της SVM σημαντικές υπέρ-παράμετροι για την ορθή εκπαίδευση του μοντέλου θεωρήθηκαν οι:

- C
- Συνάρτηση kernel
- gamma

Οι δύο πρώτες υπέρ-παράμετροι αναφέρθηκαν και εξηγήθηκαν στο Κεφάλαιο 2 (§ 2.4.4) ωστόσο, η τελευταία εισάγεται τώρα για πρώτη φορά. Στην ουσία της η παράμετρος gamma διαδραματίζει ένα ρόλο αντίστοιχο με αυτόν της C, υπό το πρίσμα του ότι και οι δύο χρησιμοποιούνται για να ρυθμίσουν την ανταλλαγή σφάλματος-διακύμανσης. Η υπέρ-παράμετρος gamma χρησιμοποιείται όταν τα δεδομένα δεν είναι γραμμικώς διαχωρίσιμα (δηλαδή για συνάρτηση kernel διαφορετική της linear) και ορίζει πόσα δείγματα θα χρησιμοποιηθούν ως διανύσματα στήριξης. Επομένως, όσο μεγαλύτερη η τιμή της gamma, τόσο περισσότερα διανύσματα στήριξης άρα λιγότερα σφάλματα και μεγαλύτερη διακύμανση, ενώ όσο μικρότερη η τιμή της gamma, τόσο λιγότερα διανύσματα στήριξης, άρα περισσότερα σφάλματα και μικρότερη διακύμανση. Φυσικά, για τη γραμμική συνάρτηση kernel τα διανύσματα στήριξης είναι πολύ συγκεκριμένα, όπως αυτά περιγράφονται στο Κεφάλαιο 2.

Ο αλγόριθμος της μπεϋζιανής βελτιστοποίησης για την SVM εκτελέσθηκε με 2 σημεία εξερεύνησης (exploration) και 5 σημεία εκμετάλλευσης (exploitation) και οι επιτρεπτές τιμές των υπέρ-παραμέτρων που δόθηκαν στον αλγόριθμο για την SVM παρουσιάζονται στον πίνακα 4.4. Το εξαγόμενο από αυτήν την διαδικασία μοντέλο χρησιμοποιεί την linear kernel συνάρτηση, δείχνοντας πως τα δεδομένα του εκπαιδευτικού σετ δεδομένων κύριων συνιστωσών είναι γραμμικώς διαχωρίσιμα, και χρησιμοποιεί παράμετρο C ίση με 312.4 (πίνακας 4.4). Η gamma δεν χρειάζεται να προσδιορισθεί καθώς τα δεδομένα είναι γραμμικώς διαχωρίσιμα και για τον λόγο αυτόν θα τεθεί ίση με την προδοθείσα τιμή της (default).

Το μοντέλο RF είναι ένα αρκετά σύνθετο μοντέλο που διαθέτει ένα ευρύ πλήθος υπέρ-παραμέτρων. Για τους σκοπούς της βελτιστοποίησης του με την μπεϋζιανή τεχνική επιλέχθηκαν οι ακόλουθες για προσδιορισμό:

- n_estimators
- min_samples_split
- max_features

όπου η υπέρ-παράμετρος n_estimators είναι ένας θετικός ακέραιος αριθμός που αντιπροσωπεύει το πόσα δέντρα θα υπάρχουν στο δάσος, η min_samples_split είναι μία αριθμητική παράμετρος που αφορά τη διαδικασία εκπαίδευσης του μοντέλου και προσδιορίζει τον ελάχιστο αριθμό των δειγμάτων που απαιτούνται προκειμένου να δημιουργηθούν κλαδιά από έναν κόμβο και η max_features είναι επίσης μία αριθμητική παράμετρος που αφορά την εκπαίδευση του μοντέλου και ορίζει τον αριθμό των μεταβλητών που θα εκτιμήσει το κάθε δέντρο για να αποφασίσει τον βέλτιστο διαχωρισμό σε κλαδιά. Για την καλύτερη κατανόηση των δύο τελευταίων υπέρ-παραμέτρων, όπως περιγράφεται αναλυτικά στο Κεφάλαιο 2, σε κάθε κόμβο φτάνουν N_i παρατηρήσεις και κατά την εκπαίδευση υπολογίζεται η τιμή του Gini Impurity για ορισμένες μεταβλητές και τις παρατηρήσεις αυτές. Από τον κόμβο αυτόν θα προκύψουν δύο νέοι κόμβοι με N_{i1} παρατηρήσεις ο ένας και N_{i2} παρατηρήσεις ο άλλος. Επομένως, η min_samples_split ορίζει τον ελάχιστο αριθμό των παρατηρήσεων για τους νέους κόμβους και η max_features ορίζει τον μέγιστο αριθμό μεταβλητών για τις οποίες θα υπολογισθεί η τιμή του Gini Impurity.

Η μπεϋζιανή βελτιστοποίηση για την RF εκτελέσθηκε με 3 σημεία εξερεύνησης και 15 σημεία εκμετάλλευσης και οι επιτρεπτές τιμές των υπέρ-παραμέτρων που δόθηκαν στον αλγόριθμο για την RF παρουσιάζονται στον πίνακα 4.4. Το εξαγόμενο από αυτήν τη διαδικασία μοντέλο αποτελείται από 164 δέντρα απόφασης καθένα από τα οποία είχε στους κόμβους του τουλάχιστον το 26% των δειγμάτων και έλαβε υπ' όψιν του το πολύ το 79% των μεταβλητών, πριν την δημιουργία νέων κόμβων (πίνακας 4.4).

Η τελευταία μέθοδος που βελτιστοποιήθηκε και εφαρμόστηκε στο εκπαιδευτικό σετ δεδομένων κύριων συνιστωσών ήταν η LR. Για τη λογιστική παλινδρόμηση σημαντικές υπέρ-παράμετροι που τέθηκαν υπό βελτιστοποίηση θεωρήθηκαν οι παράγοντες κανονικοποίησης της συνάρτησης κόστους. Συγκεκριμένα, οι παράμετροι αυτοί ήταν ο παράγοντας C και η επιλογή της κατάλληλης νόρμας (L1 ή L2). Για την εκτέλεση της μπεϋζιανής μεθοδολογίας χρησιμοποιήθηκαν μόνο 2 σημεία εξερεύνησης και 3 σημεία εκμετάλλευσης και το τελικό εξαγόμενο μοντέλο χρησιμοποίησε για την κανονικοποίηση της συνάρτησης κόστους την L1 νόρμα και την παράμετρο C ίση με 100.1 (πίνακας 4.4).

Στο σετ δεδομένων κύριων συνιστωσών εφαρμόστηκε και το μοντέλο της Naive Bayes το οποίο όμως, όπως εύκολα προκύπτει κι από το Κεφάλαιο 2 δεν διαθέτει υπέρ-παραμέτρους. Τα δεδομένα του εκπαιδευτικού σετ δεδομένων ήταν συνεχείς αριθμητικές τιμές και ακολουθούσαν κανονική κατανομή. Για τον λόγο αυτό χρησιμοποιήθηκε η Γκαουσιανή κατανομή για τον προσδιορισμό της πιθανότητας $P(x_i|C_k)$.

Μοντέλα	Υπέρ-Παράμετροι	Επιτρεπτές Τιμές	Τελικές Τιμές
SVM	C	$[200-500]$	312.36
	gamma	$[1-2]$	–
	kernel	$\{linear, poly, sigmoid, rbf\}$	$linear$
RF	n_estimators	$[10-1000]$	164
	min_samples_split	$[0.1-0.9]$	0.2248
	max_features	$[0.5-0.9]$	0.7395
LR	C	$[10-200]$	81.16
	penalty	$\{L1, L2\}$	$L1$

Πίνακας 4.4 Επιτρεπτές τιμές των υπέρ-παραμέτρων που προσδιορίστηκαν μέσω της μπεϋζιανής βελτιστοποίησης και οι τελικές τιμές τους για το σετ δεδομένων κύριων συνιστωσών. Η παράμετρος gamma δεν έχει τιμή καθώς η συνάρτηση kernel είναι η γραμμική συνάρτηση.

4.4.2 Βελτιστοποίηση και εκπαίδευση στο κανονικοποιημένο σετ δεδομένων

Στο κανονικοποιημένο εκπαιδευτικό σετ δεδομένων εφαρμόστηκαν και βελτιστοποιήθηκαν οι ίδιες μέθοδοι με την περίπτωση του σετ δεδομένων κύριων συνιστωσών και μία επιπλέον, η PLS. Οι υπέρ-παράμετροι που επιλέχθηκαν ως οι σημαντικές για την εκπαίδευση των μοντέλων στα δεδομένα είναι οι ίδιες που επιλέχθηκαν στην προηγούμενη μεθοδολογία. Φυσικά τα διαστήματα επιτρεπτών τιμών για αυτές καθώς και το πλήθος των σημείων εξερεύνησης και εκμετάλλευσης είναι διαφορετικά.

Η διαδικασία βελτιστοποίησης της SVM έγινε με 5 σημεία εκμετάλλευσης και 3 σημεία εξερεύνησης. Το μοντέλο που προέκυψε από αυτήν τη διαδικασία χρησιμοποιεί την linear kernel συνάρτηση, και η τιμή του C είναι 249.8 και εδώ και πάλι ο προσδιορισμός του gamma δεν έχει κάποιο ιδιαίτερο νόημα. Η RF, όπως και στην περίπτωση του σετ δεδομένων κύριων συνιστωσών, χρειάστηκε περισσότερα σημεία εκμετάλλευσης (10) και εξερεύνησης (5) για να οδηγηθεί στο βέλτιστο μοντέλο για τις επιτρεπτές τιμές που δόθηκαν στον αλγόριθμο της μπεϋζιανής βελτιστοποίησης. Το δάσος στο οποίο κατέληξε ο αλγόριθμος απαρτίζεται από 185 δέντρα και για την εκπαίδευση του καθενός οι τιμές των min_samples_split και max_features ήταν 0.4803 και 0.3996 αντίστοιχα. Αντίστοιχα, ο αλγόριθμος για την βελτιστοποίηση της LR επαναλήφθηκε και για τους δύο διαφορετικούς τύπους κανονικοποίησης 8 φορές (5 σημεία εκμετάλλευσης και 3 σημεία εξερεύνησης) και η ιδανικότερη νόρμα φάνηκε να είναι η L1 με παράμετρο C ίση με 43.7.

Όπως έχει αναφερθεί αρκετές φορές μέχρι τώρα, ένα βασικό στοιχείο της PLS είναι πως εφαρμόζει ορθογωνικό μετασχηματισμό (PCA) στα δεδομένα προτού εφαρμοστεί η γραμμική παλινδρόμηση. Για τη βελτιστοποίηση και την εκπαίδευση της PLS επιλέχθηκε μόνο μία υπέρ-παράμετρος υψηλής σημασίας για την ταξινόμηση των νανοσωλήνων σε τοξικούς και μη τοξικούς, η n_components. Η παράμετρος αυτή δέχεται μία θετική ακέραια τιμή που αντιπροσωπεύει τον αριθμό των τελικών συνιστωσών. Όπως και στην PCA, έτσι και στην PLS επιλέγεται ο αριθμός των συνιστωσών που θα κρατήσει το μοντέλο. Φυσικά ο αριθμός αυτός δεν είναι δυνατό να υπερβαίνει το πλήθος μεταβλητών που διαθέτει το σετ δεδομένων. Ο αλγόριθμος της μπεϋζιανής βελτιστοποίησης εκτελέσθηκε με 3 σημεία εκμετάλλευσης και 2 σημεία εξερεύνησης και κατέληξε στο συμπέρασμα πως ο ελάχιστος αριθμός συνιστωσών είναι 4.

Στον ακόλουθο πίνακα (πίνακας 4.5) παρουσιάζονται τα αποτελέσματα της βελτιστοποίησης των υπέρ-παραμέτρων των μοντέλων που εκπαιδεύθηκαν στο κανονικοποιημένο εκπαιδευτικό σετ δεδομένων.

Μοντέλα	Υπέρ-Παράμετροι	Επιτρεπτές Τιμές	Τελικές Τιμές
SVM	C	$[200 - 1000]$	499.63
	gamma	$[0.1 - 1]$	–
	kernel	$\{linear, poly, sigmoid, rbf\}$	$linear$
RF	n_estimators	$[10 - 250]$	185
	min_samples_split	$[0.1 - 0.5]$	0.4803
	max_features	$[0.1 - 0.9]$	0.3996
LR	C	$[10 - 100]$	43.71
	penalty	$\{L1, L2\}$	$L1$
PLS	n_components	$[1 - 34]$	4

Πίνακας 4.5 Επιτρεπτές τιμές των υπέρ-παραμέτρων που προσδιορίστηκαν μέσω της μπεϋζιανής βελτιστοποίησης και οι τελικές τιμές τους για το κανονικοποιημένο σετ δεδομένων. Η παράμετρος gamma, και πάλι δεν έχει τιμή καθώς η συνάρτηση kernel είναι η γραμμική συνάρτηση.

4.5 Αποτελέσματα βελτιστοποίησης για τους MWCNTs

Έχοντας βελτιστοποιήσει και εκπαιδεύσει τα μοντέλα στα δύο σετ δεδομένωνs, θα πρέπει να αξιολογήσουμε τα αποτελέσματα. Όπως περιγράφεται αναλυτικά στο Κεφάλαιο 2, η αξιολόγηση των εκπαιδευμένων μοντέλων γίνονται με την εφαρμογή των μετρικών τιμών. Οι μετρικές τιμές που χρησιμοποιήθηκαν σε αυτό το στάδιο της ανάλυσης των MWCNTs είναι οι:

- Accuracy score
- Precision
- Recall
- F1-Score
- Confusion matrix

Στον πίνακα 4.6 παρουσιάζονται τα αποτελέσματα των μοντέλων που εφαρμόστηκαν στο κανονικοποιημένο σετ δεδομένων. Είναι φανερό πως όλα τα μοντέλα, πλην της Gaussian NB, απέδωσαν με ικανοποιητική ακρίβεια όλες τις συσχετίσεις που υπάρχουν στα δεδομένα μας. Οι μετρικές τιμές της PLS συγκεκριμένα, δείχνουν πως η εκπαίδευση της είναι όσο καλή θα μπορούσε να είναι, καθώς όχι μόνο κατόρθωσε να αφομοιώσει τα μοτίβα του εκπαιδευτικού κανονικοποιημένου σετ δεδομένων, αλλά χρησιμοποιώντας αυτή τη γνώση της μπόρεσε να προβλέψει στα άγνωστα δείγματα με 100% ακρίβεια. Ωστόσο, κάτι τέτοιο δεν συνέβη με τις υπόλοιπες. Οι SVM, RF και LR αν και έχουν ικανοποιητική ακρίβεια στα test δεδομένα, φαίνεται να ταξινόμησαν λάθος ένα μη τοξικό δείγμα. Το δείγμα αυτό ήταν κοινό και στις τρεις (το NM-401) γεγονός που σημαίνει πως το δείγμα αυτό αποτελείται από έκτροπες τιμές και παρά την εφαρμογή της Kennard-Stone πέρασε στις επαληθευτικές παρατηρήσεις. Η PLS λόγω του μετασχηματισμού που υλοποιεί, κατόρθωσε να ομαλοποιήσει τις τιμές του NM-401 με αποτέλεσμα να μην το ταξινομήσει λάθος. Πιο συγκεκριμένα, το συμπέρασμα που εξάγεται από αυτήν την έκβαση είναι πως οι πειραματικές μετρήσεις του NM-401 περιείχαν υψηλά ποσοστά θορύβου, καθιστώντας το «outlier», και κατά τη διαδικασία του μετασχηματισμού της PLS ο θόρυβος αυτός συρρικνώθηκε, ή εξαλείφθηκε, δίνοντας μία ακόμα σωστή πρόβλεψη στον αλγόριθμο. Επομένως, και τα τέσσερα μοντέλα θεωρείται πως περιγράφουν ορθά τις συσχετίσεις του τελικού σημείου με τις μεταβλητές.

Σε αντίθεση με τα ανωτέρω τέσσερα μοντέλα, η Gaussian NB δεν φαίνεται να ανταποκρίνεται στο στόχο του προβλήματος. Η ακρίβεια του μοντέλου στα test δεδομένα ήταν 40%, χειρότερη δηλαδή από την τυχαία ταξινόμηση των παρατηρήσεων. Όπως εξηγήθηκε στο Κεφάλαιο 2, το μοντέλο αυτό παίρνει ως παραδοχή πως δεν υπάρχουν συσχετίσεις μεταξύ των δεδομένων, επομένως η κακή του απόδοση μας οδηγεί στο συμπέρασμα πως η υπόθεση αυτή δεν ισχύει. Δεδομένου ότι η εφαρμογή της PCA μετασχηματίζει με τέτοιον τρόπο το σετ δεδομένων, ώστε να συρρικνώνεται ο θόρυβος και κάθε μία από τις κύριες συνιστώσες να είναι γραμμικώς ανεξάρτητη από την άλλη, περιμένουμε η Gaussian NB και οι SVM, RF, LR να αποδίδουν καλύτερα στο σετ δεδομένων κύριων συνιστωσών.

	SVM		**RF**		**LR**		**Gaussian NB**		**PLS**	
Accuracy Score	80%		80%		80%		40%		100%	
Precision	0.66		0.66		0.66		0.33		1.0	
Recall	1.0		1.0		1.0		0.50		1.0	
F1-Score	0.48		0.48		0.48		0.40		1.0	
Confusion Matrix	2	1	2	1	2	1	1	2	3	0
	0	2	0	2	0	2	1	1	0	2

Πίνακας 4.6 Πίνακας με τις μετρικές τιμές των μοντέλων που βελτιστοποιήθηκαν και εκπαιδεύτηκαν στο κανονικοποιημένο σετ δεδομένων.

Πράγματι, όπως φαίνεται από τον πίνακα 4.7, τα τέσσερα αυτά μοντέλα είχαν σημαντικά υψηλότερη ακρίβεια στα επαληθευτικά δεδομένα του μετασχηματισμένου σετ δεδομένων (+20% αύξηση). Η αύξηση αυτή οφείλεται στο NM-401 που ταξινόμησαν σωστά αυτήν τη φορά τα μοντέλα. Φυσικά, το accuracy score της Gaussian NB παραμένει χαμηλό, επομένως φαίνεται αυτή η μέθοδος να αδυνατεί γενικότερα να εκπαιδευθεί στα δεδομένα των νανοσωλήνων. Παρά το γεγονός αυτό όμως, τα υπόλοιπα τρία μοντέλα (SVM, RF και LR) που εφαρμόστηκαν στο σετ δεδομένων κύριων συνιστωσών των MWCNTs, κατάφεραν να εντοπίσουν τα μοτίβα πίσω από τις παρατηρήσεις με 100% ακρίβεια. Προκύπτει επομένως, πως η εφαρμογή της PCA, πράγματι οδήγησε στην αφαίρεση θορύβου και δεν επηρέασε τη χρήσιμη πληροφορία των δεδομένων των MWCNTs.

	SVM		**RF**		**LR**		**Gaussian NB**	
Accuracy Score	100%		100%		100%		60%	
Precision	1.0		1.0		1.0		0.66	
Recall	1.0		1.0		1.0		0.66	
F1-Score	1.0		1.0		1.0		0.66	
Confusion Matrix	3	0	3	0	3	0	2	1
	0	2	0	2	0	2	1	1

Πίνακας 4.7 Πίνακας με τις μετρικές τιμές των μοντέλων που βελτιστοποιήθηκαν και εκπαιδεύτηκαν στο σετ δεδομένων κύριων συνιστωσών.

Άλλο ένα ενδιαφέρον αποτέλεσμα από την ανωτέρω διαδικασία είναι το γεγονός πως η PLS κρατώντας 4 κύριες συνιστώσες κατάφερε να εκπαιδευτεί στα δεδομένα των MWCNTs. Με άλλα λόγια, με 4 κύριες συνιστώσες τα δεδομένα των MWCNTs γίνονται γραμμικώς διαχωρίσιμα. Από αυτό το γεγονός αυτό βγαίνει το συμπέρασμα πως οι τέσσερις κύριες συνιστώσες διαθέτουν όλη την πληροφορία που είναι απαραίτητη για την ταξινόμηση των νανοσωλήνων. Ωστόσο, οι διαθέσιμη πληροφορία για τους νανοσωλήνες είναι εκ των πραγμάτων περιορισμένη καθώς τα διαθέσιμα δεδομένα είναι λίγα, επομένως οι 4 κύριες συνιστώσες διαθέτουν περίπου όση πληροφορία περιείχε το αρχικό σετ δεδομένων. Το αποτέλεσμα αυτό σε συνδυασμό με το διάγραμμα που παρουσιάζεται στην εικόνα 4.3 οδηγούν στο συμπέρασμα πως περίπου το 72% της διακύμανσης του αρχικού σετ δεδομένων περιέχει όλη την πληροφορία που απαιτείται για την ταξινόμηση των νανοσωλήνων άνθρακα σε τοξικούς και μη τοξικούς, ή καλύτερα, περίπου το 28% της πληροφορίας των δεδομένων για τους MWCNTs αποτελεί θόρυβο.

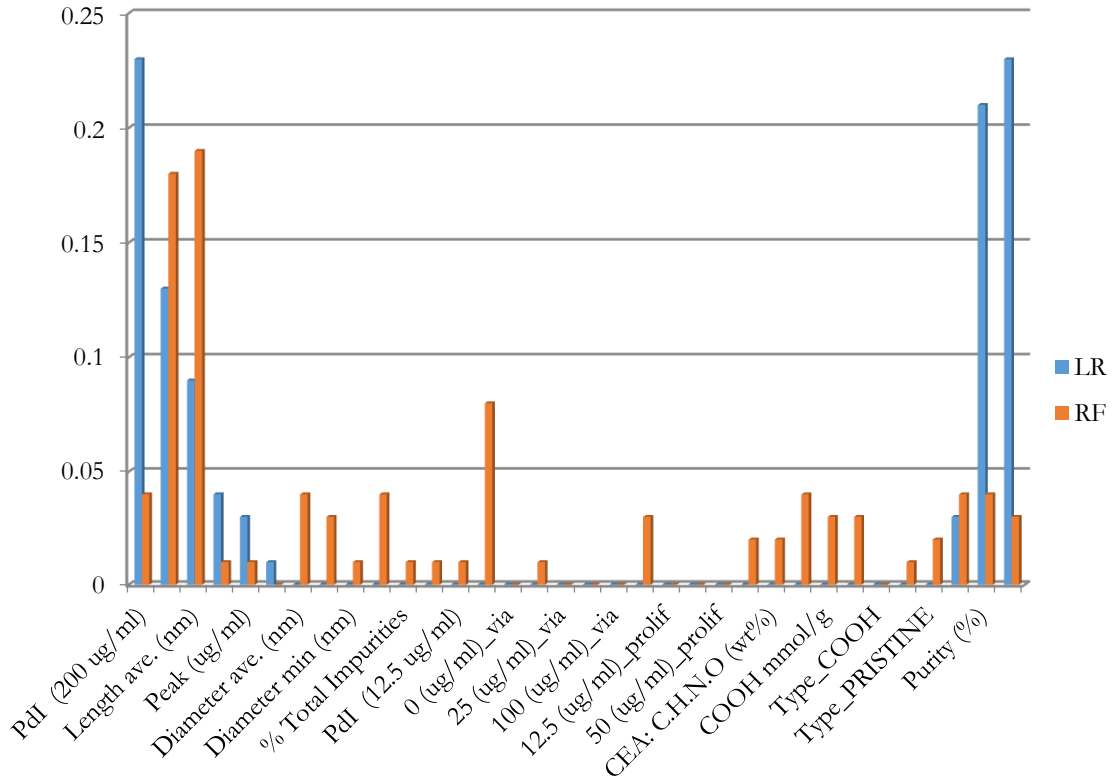
Τα αποτελέσματα αυτά είναι ενθαρρυντικά καθώς φαίνεται πως παρά τους περιορισμούς που θέτουν τα πειραματικά δεδομένα, είναι δυνατή η πρόβλεψη της γονοτοξικότητας των νανοσωλήνων με υψηλή ακρίβεια. Παρόλα αυτά, αν και τα αποτελέσματα δείχνουν να έχει εντοπιστεί η συσχέτιση του τελικού σημείου της ανάλυσης με τις μεταβλητές, τα μοντέλα αυτά είναι ιδιαιτέρως δύσχρηστα. Ο λόγος για τον οποίο συμβαίνει αυτό είναι πως το εκπαιδευτικό σετ δεδομένων και στις δύο μεθοδολογίες αποτελείται από έναν μεγάλο αριθμό στηλών. Επομένως, για να γίνει η πρόβλεψη της γονοτοξικότητας ενός τυχαίου νανοσωλήνα απαιτούνται 31 τιμές εισόδου γεγονός που δυσχεραίνει τη χρήση αυτών των μοντέλων. Μια ακόμη δυσκολία στη χρήση αυτών των μοντέλων αποτελεί το γεγονός πως ορισμένες από τις μεταβλητές εισόδου προέρχονται από ένα σύνολο πειραματικών μετρήσεων. Για παράδειγμα η στήλη «12.5 (ug/ml)_via» αντιπροσωπεύει το ποσοστό βιωσιμότητας συγκεκριμένου τύπου κυττάρων που εκτέθηκαν για συγκεκριμένο χρόνο σε συγκέντρωση $12.5\ {}^{\mu g_{MWCNT}}/_{ml}$ και σε συγκεκριμένες συνθήκες. Άρα για να προβλέψει κανείς εάν ένας νανοσωλήνας άνθρακα πολλαπλών τοιχωμάτων είναι τοξικός θα πρέπει να επαναλάβει τα πειράματα στις ίδιες συνθήκες και να υπολογίσει το ποσοστό βιωσιμότητας για τον συγκεκριμένο νανοσωλήνα. Γίνεται αμέσως αντιληπτή, λοιπόν, η ανάγκη ύπαρξης ενός πιο εύχρηστου προβλεπτικού μοντέλου.

4.6 Ελάττωση Διαστατικότητας Μοντέλου (RFE) και Βελτιστοποίηση

Η διαδικασία ελάττωσης των διαστάσεων ενός μοντέλου αποτελεί ένα ιδιαιτέρως σημαντικό κεφάλαιο της προβλεπτικής μάθησης και της QSAR μοντελοποίησης. Υπάρχουν αρκετές μέθοδοι σύμφωνα με τις οποίες καθίσταται δυνατή αυτή η διαδικασία. Στο παρών πρόβλημα εφαρμόστηκε η μέθοδος της Επαναλαμβανόμενης Κατάργησης Μεταβλητών (Recursive Feature Elimination, RFE).

Για την υλοποίηση της RFE απαραίτητος ήταν ο υπολογισμός της σπουδαιότητας κάθε μεταβλητής στο μοντέλο εφαρμογής της RFE. Η διαδικασία αυτή δεν αφορά τα μοντέλα που εκπαιδεύτηκαν στο σετ δεδομένων κύριων συνιστωσών, καθώς ο υπολογισμός της σπουδαιότητας κάθε κύριας συνιστώσας δεν έχει νόημα, δεδομένου πως πριν από κάθε πρόβλεψη για έναν τυχαίο νανοσωλήνα προηγείται ο μετασχηματισμός των χαρακτηριστικών του στις κύριες συνιστώσες και για αυτήν την διαδικασία απαιτούνται και πάλι όλες οι μεταβλητές. Έτσι η διαδικασία αυτή περιορίζεται στα μοντέλα που εκπαιδεύτηκαν στο κανονικοποιημένο σετ δεδομένων. Ακόμα, από αυτά τα πέντε μοντέλα, ο υπολογισμός της σπουδαιότητας είναι δυνατός μόνο στα LR, RF και PLS. Ωστόσο, η διαδικασία RFE εφαρμόστηκε στα δύο πρώτα μόνο λόγω της δυσκολίας της PLS.

Η διαδικασία με την οποία υπολογίζεται η σπουδαιότητα κάθε μεταβλητής περιγράφεται αναλυτικά στο Κεφάλαιο 2. Τα αποτελέσματα αυτού του υπολογισμού παρουσιάζονται στο παρακάτω διάγραμμα (εικόνα 4.4).



Εικόνα 4.4 **Διάγραμμα σπουδαιότητας μεταβλητών για τις RF και LR**

Όπως φαίνεται στο παραπάνω διάγραμμα οι περισσότερες μεταβλητές δεν επηρεάζουν καθόλου (ή επηρεάζουν ελάχιστα) τη λήψη της απόφασης των δύο αυτών μοντέλων. Επομένως, τα μοντέλα θα κατέληγαν στα ίδια συμπεράσματα και χωρίς να δέχονται αυτές τις μεταβλητές εισόδου. Αυτή είναι η κεντρική ιδέα της RFE. Σύμφωνα με τη μέθοδο αυτή, αφαιρείται η λιγότερο σημαντική μεταβλητή από το σετ δεδομένων και το μοντέλο βελτιστοποιείται και εκπαιδεύεται από την αρχή. Η διαδικασία αυτή είναι μία επαναληπτική διαδικασία η οποία σταματάει όταν το μοντέλο φτάσει

στις επιθυμητές διαστάσεις. Στο προκείμενο πρόβλημα, επιλέχθηκε ως κριτήριο τερματισμού της επαναληπτικής αυτής διαδικασίας η μείωση της ακρίβειας του μοντέλου. Η διαδικασία αυτή επαναλήφθηκε πολλές φορές και κατέληξε σε τέσσερις σημαντικές για την LR μεταβλητές και τέσσερις για την RF. Οι μεταβλητές που χαρακτηρίστηκαν ως οι πιο σημαντικές από την LR είναι οι:

- Purity (%)
- PdI (batch)
- Zave (12.5 ug/ml)
- Type_OH

ενώ οι πιο σημαντικές μεταβλητές για το μοντέλο της RF είναι οι:

- Length ave. (nm)
- Purity (%)
- Zave (12.5 ug/ml)
- Diameter ave. (nm)

Οι υπέρ-παράμετροι των τελικών μοντέλων που προέκυψαν από αυτήν τη διαδικασία παρουσιάζονται στον πίνακα 4.8.

Μοντέλα	Υπέρ-Παράμετροι	Τελικές Τιμές
RF	n_estimators	197
	min_samples_split	0.001
	max_features	0.500
LR	C	1374.54
	penalty	L1

Πίνακας 4.8 **Τελικές τιμές υπέρ-παραμέτρων των μοντέλων RF και LR μετά την ελάττωση της διαστατικότητας τους.**

4.7 Αποτελέσματα RFE και Επιλογή Μοντέλου

Η διαδικασία ελάττωσης της διαστατικότητας των μοντέλων οδήγησε στην εκπαίδευση δύο μοντέλων με τέσσερις μόνο μεταβλητές εισόδου. Όπως φαίνεται και στον πίνακα 4.9 και τα δύο μοντέλα έχουν 100% ακρίβεια στα test δεδομένα, γεγονός που σημαίνει πως και τα δύο μοντέλα παρά τις χαμηλές απαιτήσεις τους σε μεταβλητές εισόδου έχουν εκπαιδευτεί ορθά και αναγνωρίζουν τα μοτίβα που χαρακτηρίζουν έναν νανοσωλήνα τοξικό ή μη τοξικό. Ένα ενδιαφέρον αποτέλεσμα

	RF		LR	
Accuracy Score	100%		100%	
Precision	1.0		1.0	
Recall	1.0		1.0	
F1-Score	1.0		1.0	
Confusion Matrix	4	0	4	0
	0	1	0	1

Πίνακας 4.9 **Πίνακας με τις μετρικές τιμές των μοντέλων RF και LR μετά την RFE.**

της διαδικασίας αυτής είναι πως στο επαληθευτικό σετ δεδομένων για την μέθοδο της LR συμπεριλήφθηκε και το NM-401 και όπως φαίνεται στον πίνακα 4.9, το μοντέλο κατάφερε να το ταξινομήσει σωστά στην κλάση του. Επομένως, αν και μειώθηκαν οι μεταβλητές εισόδου του μοντέλου φαίνεται πως μεγάλωσε το πεδίο εφαρμογής του.

Οι σημαντικές, για την πρόβλεψη της τοξικότητας, μεταβλητές που προέκυψαν από την διαδικασία ήταν αναμενόμενες, γεγονός που ενισχύει την δύναμη του μοντέλου. Πιο συγκεκριμένα, όπως αναφέρεται στην παράγραφο 3.4, διάφορες έρευνες έχουν δείξει πως η τοξικότητα των νανοσωλήνων άνθρακα είναι πολύ πιθανό να συνδέεται με το μέγεθος τους (μήκος και διάμετρος) και με την επιφάνειά τους (επιφανειακή τροποποίηση και καθαρότητα). Η παρούσα ανάλυση έρχεται να επιβεβαιώσει τις έρευνες αυτές, καθώς στην προηγούμενη παράγραφο φαίνεται πως όλες αυτές οι φυσικοχημικές ιδιότητες αποτελούν σημαντικές παραμέτρους στην πρόβλεψη της τοξικότητας, βέβαια, δεν είναι όλες σημαντικές και για τα δύο μοντέλα. Ένα ιδιαιτέρως ενδιαφέρων αποτέλεσμα, είναι πως δεν φαίνεται να συσχετίζονται όλες οι λειτουργικές ομάδες των νανοσωλήνων με την πιθανή τους τοξικότητα, αλλά μόνο η –OH λειτουργική ομάδα.

Φυσικά τα δύο αυτά μοντέλα είναι παρόμοια ως προς τις απαιτήσεις τους σε δεδομένα εισόδου, επομένως η επιλογή του βέλτιστου μοντέλου θα γίνει με διαφορετικό τρόπο. Το QSAR μοντέλο που περιγράφει την τοξικότητα των MWCNTs εκτός από ακριβές και χαμηλών απαιτήσεων σε μεταβλητές, θα πρέπει ναι είναι και εύρωστο (robust). Για αυτήν την ανάλυση εφαρμόστηκε η μέθοδος «cross-validation» ώστε να επιβεβαιώσουμε την αποτελεσματικότητα των μοντέλων. Για τη μέθοδο αυτή, χρησιμοποιήθηκαν 4 «folds» και τα αποτελέσματα έδειξαν πως το μοντέλο της λογιστικής παλινδρόμησης τείνει να είναι πιο ανθεκτικό στις μεταβολές των δεδομένων από το μοντέλο του τυχαίου δάσους. Ειδικότερα, η μέθοδος «cross-validation» για το LR μοντέλο κατέληξε στην τιμή 0.81, δείχνοντας πως το μοντέλο είναι «robust», ενώ για το RF μοντέλο κατέληξε στην τιμή 0.62, δείχνοντας πως το μοντέλο αυτό έχει υπέρ-προσαρμοστεί στα εκπαιδευτικά δεδομένα καθώς όταν άλλαζαν, κατά την εφαρμογή της μεθόδου «cross-validation», το μοντέλο δεν μπορούσε να εκπαιδευτεί ορθά. Παράλληλα, υπολογίσθηκαν οι πιθανότητες με τις οποίες τα δύο αυτά μοντέλα ταξινόμησαν κάθε επαληθευτικό δείγμα. Οι πιθανότητες αυτές παρουσιάζονται στον πίνακα που ακολουθεί (πίνακας 4.10):

RFC			LR		
MWCNTs	**Πιθανότητα για μη-τοξικότητα**	**Πιθανότητα για τοξικότητα**	**MWCNTs**	**Πιθανότητα για μη-τοξικότητα**	**Πιθανότητα για τοξικότητα**
NRCWE-040	0.78	0.22	**NRCWE-040**	1.00	0.00
NRCWE-041	0.68	0.32	**NM-401**	0.76	0.24
NRCWE-048	0.50	0.50	**NM-402**	0.01	0.99
NRCWE-045	0.20	0.80	**NRCWE-047**	1.00	0.00
NRCWE-042	0.87	0.13	**NRCWE-042**	1.00	0.00

Πίνακας 4.10 Πιθανότητες ταξινόμησης των δειγμάτων του επαληθευτικού σετ δεδομένων για τα RF και LR μοντέλα μετά την εφαρμογή της RFE.

Όπως φαίνεται και στον πίνακα 4.10, το μοντέλο της RF αν και ταξινόμησε σωστά όλα τα επαληθευτικά δείγματα, η πιθανότητα σύμφωνα με την οποία τα ταξινόμησε είναι αρκετά χαμηλή, ενώ το μοντέλο της LR κατέληξε σε ιδιαιτέρως υψηλές πιθανότητες. Ο πίνακας 4.10, σε συνδυασμό με τα αποτελέσματα από την εφαρμογή της «cross-validation» οδηγούν στο συμπέρασμα πως το μοντέλο της λογιστικής παλινδρόμησης είναι αυτό που εκπαιδεύτηκε καλύτερα στα δεδομένα κι επομένως το βέλτιστο μοντέλο για το πρόβλημα αυτό.

Τελευταίο βήμα της διαδικασίας αυτής αποτελεί η εκπαίδευση του μοντέλου αυτού στο σύνολο των διαθέσιμων δεδομένων και ο υπολογισμός του πεδίου εφαρμογής. Η εκπαίδευση του

μοντέλου το σύνολο των δεδομένων πραγματοποιήθηκε με σκοπό την εκμετάλλευση όλης της διαθέσιμης πληροφορίας και κατέληξε στην συνάρτηση απόφασης που φαίνεται στην εξίσωση 4.3. Όπως έχει προαναφερθεί, το πεδίο εφαρμογής αποτελεί ένα από τα βασικότερα στοιχεία ενός QSAR μοντέλου. Ο υπολογισμός του εύρους του μοντέλου έγινε με βάση την τεχνική Leverage. Σύμφωνα με αυτήν την τεχνική υπολογίζεται μία τιμή-κατώφλι (εξίσωση 4.2α), $threshold$, και κατόπιν υπολογίζεται μία τιμή h_i για κάθε δείγμα για το οποίο πρόκειται να γίνει πρόβλεψη για την τοξικότητα του (εξίσωση 4.2β). Εάν η h_i είναι μεγαλύτερη του $threshold$ τότε, το δείγμα θεωρείται εκτός του πεδίου εφαρμογής του μοντέλου κι επομένως η πρόβλεψη δεν θεωρείται αξιόπιστη.

$$threshold = 3 \cdot \frac{(πλήθος\ μεταβλητών)}{(πλήθος\ δεδομένων\ εκπαίδευσης)} \quad (4.2α)$$

$$h_i = x_i^T (X^T X)^{-1} x_i \quad (4.2β)$$

όπου x_i το διάνυσμα του δείγματος για το οποίο πρόκειται να γίνει η πρόβλεψη και X ο πίνακας του εκπαιδευτικού σετ δεδομένων (πίνακας Β4 Παράρτημα Β). Η τιμή-κατώφλι για το τελικό μοντέλο προκύπτει ίση με 1.2, και επομένως για κάθε δείγμα με $h_i > 1.2$, η πρόβλεψη θεωρείται αναξιόπιστη.

Επομένως, το τελικό μοντέλο που προέκυψε από την όλη διαδικασία είναι ένα Logistic Regression μοντέλο, που χρησιμοποιεί ως παράγοντα κανονικοποίησης τον L1, με παράμετρο C ίση με 1374.54, το οποίο δέχεται σαν μεταβλητές εισόδου την καθαρότητα των νανοσωλήνων, τον δείκτη πολυδιασποράς, την μέση τιμή του Zeta Potential και την ύπαρξη –OH λειτουργικής ομάδας. Η συνάρτηση που υπολογίζει την πιθανότητα κάποιος νανοσωλήνας να είναι τοξικός παρουσιάζεται στην εξίσωση 4.3.

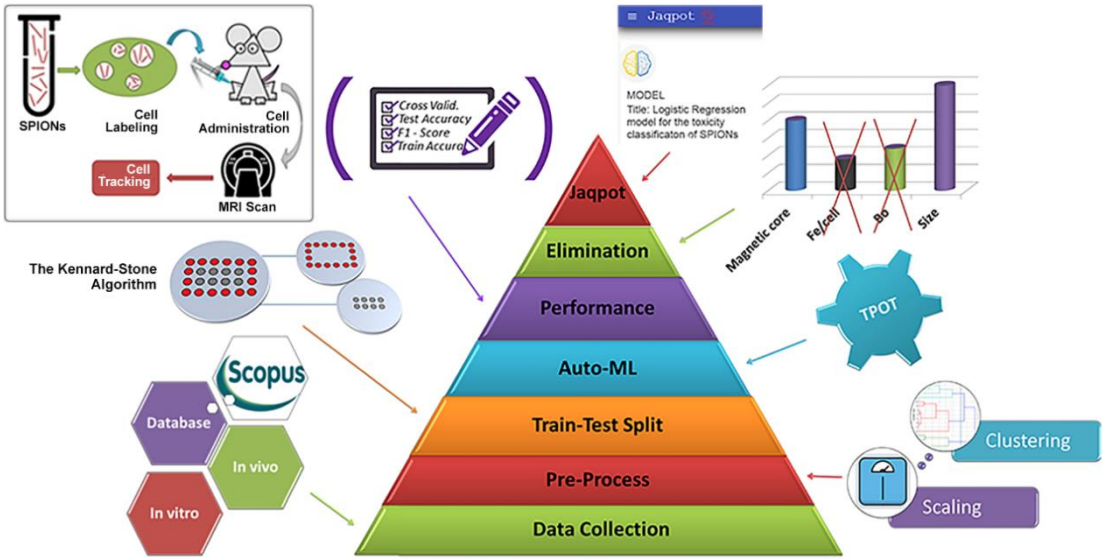
$$p(y=1) = \frac{e^{-30.73 \cdot scaled(Purity) - 26.39 \cdot scaled(PdI) + 15.85 \cdot scaled(Zeta) + 7.36 \cdot (-OH\ type) + 22.37}}{1 + e^{-30.73 \cdot scaled(Purity) - 26.39 \cdot scaled(PdI) + 15.85 \cdot scaled(Zeta) + 7.36 \cdot (-OH\ type) + 22.37}} \quad (4.3)$$

4.8 Γενική μεθοδολογία για τα SPIONs[1]

Τις τελευταίες δεκαετίες τα νάνο-βιοϋλικά έχουν βρει ευρεία εφαρμογή στον τομέα των αναγεννητικών φαρμάκων και στη θεραπεία διαφόρων ασθενειών. Ορισμένα από τα βασικά χαρακτηριστικά που πρέπει να διαθέτουν τα νανοϋλικά αυτά είναι να είναι βιοδιασπώμενα, χημικά σταθερά σε φυσιολογικές συνθήκες και μη τοξικά. Από τα νάνο-βιοϋλικά οι ερευνητές φαίνεται να έχουν δείξει ιδιαίτερο ενδιαφέρον στα SPIONs, όπως προκύπτει και από το Κεφάλαιο 3, κυρίως λόγω της χαμηλής τοξικότητάς τους. Μεγάλο ενδιαφέρον παρουσιάζει η χρήση τους σε εφαρμογές παρακολούθησης θεραπειών βασισμένων σε βλαστοκύτταρα. Για να χρησιμοποιηθούν τα SPIONs σε τέτοιες απεικονιστικές τεχνικές (MRI) πρέπει να παρουσιάζουν παραμαγνητικές ιδιότητες. Στην πραγματικότητα, τα SPIONs ενσωματώνονται στα βλαστικά κύτταρα, μεταφέροντάς τους τις μαγνητικές τους ιδιότητες. Τα σημασμένα με SPIONs βλαστικά κύτταρα στη συνέχεια εμφυτεύονται στον ανθρώπινο οργανισμό και απεικονίζονται με την MRI. Με αυτόν τον τρόπο, μπορεί να ελεγχθεί απολύτως το όργανο που έχουν μεταναστεύσει τα βλαστοκύτταρα στον οργανισμό καθώς και ο αριθμός τους σε πραγματικό χρόνο. Αν και η πληθώρα των βιβλιογραφικών πηγών χαρακτηρίζει τα SPIONs ως μη τοξικά, έχει παρατηρηθεί τοξικότητα σε ορισμένα εμπορικά SPIONs, όπως στο Ferridex και Revosit.[92] Ωστόσο, παρά το πλήθος των ερευνών που έχουν γίνει για την τοξικότητα των SPIONs, καμία μελέτη δεν βρέθηκε για την τοξικότητά τους ως σκιαγραφικά MRI στην παρακολούθηση των βλαστοκυττάρων. Για τον λόγο αυτό, στην παρούσα διπλωματική εργασία έγινε

μία προσπάθεια εκπαίδευσης ενός μοντέλου που θα προβλέπει την τοξικότητα των SPIONs σε τέτοιες εφαρμογές.

Η διαδικασία για την παραγωγή του QSAR μοντέλου για την περίπτωση των SPIONs είναι όμοια με αυτήν των MWCNTs. Η τοξικότητα των SPIONs εξετάστηκε ως προς τα ποσοστά βιωσιμότητας των βλαστικών κυττάρων τόσο σε *in vitro* όσο και σε *in vivo* μελέτες. Τα δεδομένα που χρησιμοποιήθηκαν αντλήθηκαν από ένα πλήθος ερευνών που συνέδεαν την τοξικότητα των οξειδίων με τις φυσικοχημικές τους ιδιότητες. Μετά την απαραίτητη προεπεξεργασία των δεδομένων χρησιμοποιήθηκε ένας γενετικός αλγόριθμος (Tree-Based Pipeline Optimization Tool, TPOT) με στόχο την εύρεση ενός QSAR μοντέλου που να περιγράφει τις συσχετίσεις της τοξικότητας και των ιδιοτήτων των οξειδίων.[93] Τελικό στάδιο της διαδικασίας αυτής ήταν η ελάττωση των διαστάσεων του μοντέλου και η διάθεσή του στο διαδίκτυο. Τα βήματα της διαδικασίας που ακολουθήθηκε για την εκπαίδευση ενός ικανοποιητικού μοντέλου παρουσιάζονται στην εικόνα 4.5.



Εικόνα 4.5 **Σχηματική αναπαράσταση του «wokflow» για την περίπτωση των SPIONs**

4.9 Κατασκευή του SPIONs σετ δεδομένων

Τα δεδομένα που χρησιμοποιήθηκαν σε αυτήν την μελέτη συλλέχθηκαν από μια εκτενή βιβλιογραφική έρευνα σε *in vitro* και *in vivo* μελέτες τοξικότητας των SPIONs, που χρησιμοποιούνται στην αναγεννησιακή ιατρική ως σκιαγραφικά σε MRI εξετάσεις για να ελέγξουν-κατευθύνουν θεραπείες με βλαστοκύτταρα. Από την διαδικτυακή βάση δεδομένων Scopus συγκεντρώθηκαν 71 μελέτες σχετικές με την τοξικότητα των οξειδίων αυτών.[94] Ωστόσο, λόγω έλλειψης πολλών φυσικοχημικών ιδιοτήτων το σετ δεδομένων κατασκευάστηκε με δεδομένα από 12 μελέτες.[95,96,97,98,99,100,101,102,103,104,105] Το σετ δεδομένων των οξειδίων αποτελείται από 16 δείγματα 6 ιδιότητες και μία μεταβλητή απόκρισης.

Η μεταβλητή απόκρισης επιλέχθηκε να είναι η βιωσιμότητα των βλαστικών κυττάρων. Η μεταβλητή αυτή ήταν και πάλι μία διακριτή μεταβλητή με 2 κλάσεις, την κλάση «0» που αντιστοιχεί στα μη τοξικά SPIONs και την κλάση «1» που αντιστοιχεί στα τοξικά SPIONs. Ειδικότερα, τα οξείδια που οδήγησαν σε ποσοστό βιωσιμότητας των βλαστικών κυττάρων χαμηλότερο από 75% θεωρήθηκαν τοξικά ενώ τα οξείδια που οδήγησαν σε ποσοστό βιωσιμότητας υψηλότερο από 75% θεωρήθηκαν μη τοξικά. Για το νανοϋλικό με αύξοντα αριθμό στο σετ δεδομένων ίσο με 13, τα

δεδομένα για την βιωσιμότητα δεν ήταν καλώς ορισμένα και για τον λόγο αυτό, το νανοϋλικό θεωρήθηκε μη τοξικό. Στον πίνακα 4.11 παρουσιάζεται το αρχικό σετ δεδομένων των SPIONs.

4.10 Προεπεξεργασία Δεδομένων των SPIONs

Όπως φαίνεται και στον πίνακα 4.11 τα δεδομένα που συλλέχθηκαν χρειάζεται να επεξεργασθούν προτού εφαρμοσθεί ο γενετικός αλγόριθμος. Αρχικά θα πρέπει να τροποποιηθούν οι λεκτικές-διακριτές μεταβλητές, δηλαδή οι «Magnetic core» και «Surface». Όσον αφορά την μεταβλητή «Magnetic core» φαίνεται πως τα δείγματα του σετ δεδομένων διαθέτουν ως πυρήνα είτε μαγκεμίτη είτε μαγνητίτη, ενώ τρίτη τιμή δεν υπάρχει. Επομένως στην περίπτωση αυτή η χρήση της «One-Hot» τεχνικής δεν είναι η αποδοτικότερη, καθώς θα δημιουργηθούν 2 στήλες με την ίδια ακριβώς πληροφορία. Αυτό το συμβάν θα είχε ως αποτέλεσμα μεγάλη συσχέτιση μεταξύ τους και κάτι τέτοιο θα είχε αντίθετα αποτελέσματα στο μοντέλο που θα εκπαιδευτεί (εικόνα 4.6).

Magnetic core
Magnetite
Maghemite
Magnetite
Magnetite
Maghemite
Magnetite
Maghemite

«One-Hot» Encode

Magnetic core_Magnetite	Megnetic core_Maghmite
1	0
0	1
1	0
1	0
0	1
1	0
0	1

Εικόνα 4.6 Αποτέλεσμα του «One-Hot Encoding» για την στήλη Magnetic core. Όπως φαίνεται οι δύο στήλες που προκύπτουν είναι συμπληρωματικές, υπό το πρίσμα πως διαθέτουν την ίδια πληροφορία. Για να έχει νόημα αυτή η μέθοδος απαιτούνται περισσότερες κατηγορίες στην αρχική στήλη, ειδάλλως, η μέθοδος «Label Encode» καταλήγει σε μία μόνο στήλη με την ίδια πληροφορία.

Επομένως, ο μετασχηματισμός της στήλης «Magnetic core» έγινε με την τεχνική «Label Encoding», σύμφωνα με την οποία μία αριθμητική τιμή αποδίδεται σε κάθε κατηγορία. Στο σετ δεδομένων των SPIONs αποδόθηκε η τιμή «1» για τον μαγκεμίτη και η τιμή «0» για τον μαγνητίτη. Παρατηρώντας τη δεύτερη στήλη με τα λεκτικά πεδία, για τη μεταβλητή «Surface», γίνεται αμέσως εμφανές το γεγονός πως σχεδόν κάθε παρατήρηση έχει διαφορετική τιμή για τη μεταβλητή αυτή, για την ακρίβεια 13 από τις 16 παρατηρήσεις έχουν διαφορετική τιμή. Αυτό σημαίνει πως σχεδόν κάθε ένα από τα οξείδια του παρόντος σετ δεδομένων έχει διαφορετική χημική τροποποίηση στην επιφάνειά του και επομένως δεν είναι δυνατό να προκύψει κάποιο μοτίβο ή κάποια συσχέτιση από αυτήν την πληροφορία. Άρα, η μεταβλητή αυτή δεν έχει κάποιο στατιστικό βάρος, παρά μόνο θόρυβο, και για τον λόγο αυτό αφαιρέθηκε από το σετ δεδομένων.

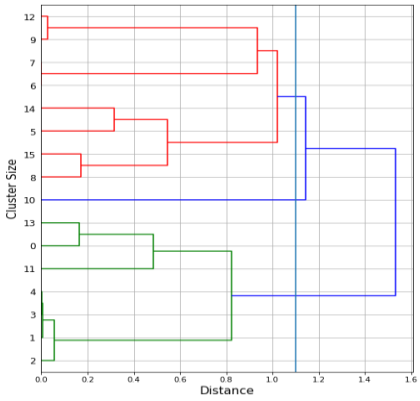
Επιπρόσθετα, στον πίνακα 4.11 φαίνεται πως υπάρχουν ορισμένα κενά στην στήλη «Relaxivity», τα οποία πρέπει να καλυφθούν. Για την κάλυψη αυτών των τιμών ακολουθήθηκε μία παρόμοια διαδικασία με αυτήν της περίπτωσης των MWCNTs. Συγκεκριμένα, οι τιμές αυτές συμπληρώθηκαν με τον μέσο όρο των τιμών των «γειτονικών» τους παρατηρήσεων στον φυσικοχημικό χώρο. Για τον ορισμό των «γειτονικών» παρατηρήσεων πραγματοποιήθηκε η μέθοδος της ιεραρχικής συσταδοποίησης. Δεδομένου πως κατά την εφαρμογή της, η μέθοδος αυτή υπολογίζει τις ευκλείδειες αποστάσεις των σημείων μεταξύ τους, πριν την υλοποίηση της τα δεδομένα κανονικοποιήθηκαν με τον ίδιο τρόπο όπως στην περίπτωση των MWCNTs (min-max scaling). Για την ακρίβεια, υλοποιήθηκε η «από κάτω προς τα πάνω» μέθοδος και τα αποτελέσματα της παρουσιάζονται στο δενδρόγραμμα της εικόνας 4.7

Για την εφαρμογή της ιεραρχικής συσταδοποίησης χρησιμοποιήθηκαν όλες οι μεταβλητές εκτός από την «Reaxivity», επομένως μόνο οι:

- Magnetic core
- Size
- B_0
- Fe/cell

A/A	Name	Magnetic core	Surface	Size (nm)	Relaxivity	B_0 (T)	Fe/cell	Cell Viablility
0	Fe2O3-PLL	Maghemite	PLL	15	-	1.5	64.51	**0**
1	Uncoated	Maghemite	Uncoated	10	549	4.7	29.3	**1**
2	D-mannose-coated Fe2O3	Maghemite	D-mannose	2	509	4.7	21.1	**1**
3	Fe2O3-PLL	Maghemite	PLL	10	492	4.7	24.5	**1**
4	PDMAAm-coated-Fe2O3	Maghemite	PDMMAm	10	89	4.7	23.2	**1**
5	N-dodecyl-PEI2k/SPIO	Magnetite	N-dodecyl-grafted PEI 2K	54.7	345	3	7.1	**0**
6	Iron oxide-loaded cationic nanovesicle	Magnetite	PEI-SA	150	343.1	1.5	50.02	**0**
7	Iron oxide-loaded cationic nanovesicle	Magnetite	PEG-PGA	150	343.1	1.5	50.02	**0**
8	CMCS-SPIONs	Magnetite	(Carboxymethyl)chitosan	55.4	160.5	1.5	26.7	**0**
9	ED-Pullulan coating SPIO	Magnetite	Ethylenediamine Pullulan	94	-	7	65	**0**
10	IONP-6PEG-HA	Magnetite	Amine-functionalizd six-armed PEG covalently linked to hyaluronic acid	75	454.5	3	1459	**0**
11	PDMAAm-coated-Fe2O3NPs	Maghemite	PDMMAm	77.8	27.26	0.5	36.9	**0**
12	Citrate SPION	Magnetite	Citrate	90.13	-	7	69.6	**0**
13	D-mannose coated SPIONs	Maghemite	D-mannose	6	140.4	0.5	51.7	**1**
14	SPIO@SiO2-NH2	Magnetite	SiO2-NH2	8.5	43.5	3	68.7	**0**
15	TAT-CLIO	Magnetite	Tat peptide functionalized cross-linking dextran	65.2	73.4	0.47	2.15	**0**

Όπως φαίνεται στο δενδρόγραμμα της εικόνας 4.7, σχηματίζονται 3 συστάδες, εκ των οποίον η μία αποτελείται από ένα μοναδικό στοιχείο. Τα δείγματα με αύξοντα αριθμό 9 και 12 βρίσκονται στην ίδια συστάδα και είναι εξαιρετικά κοντά, αφού ενώνονται πολύ νωρίς, γεγονός που φαίνεται κι από τις τιμές των μεταβλητών τους. Για τον λόγο αυτό τοποθετήθηκε η ίδια τιμή στην στήλη «Relaxivity» και για τα δύο αυτά δείγματα.



Εικόνα 4.7 Δενδρόγραμμα που προκύπτει από την Ιεραρχική Συσταδοποίηση των δεδομένων των SPIONs

Στην ίδια συστάδα με τα δείγματα 9 και 12 βρίσκονται και τα 5, 6, 7, 8, 14 και 15 ενώ το δείγμα 0 βρίσκεται στην ίδια συστάδα με τα 1, 2, 3, 4, 11 και 13. Επομένως οι τιμές τους θα είναι:

$$Rel(9) = \frac{Rel(5) + Rel(6) + Rel(7) + Rel(8) + Rel(14) + Rel(15)}{6} = 0.3658 \quad (4.4\alpha)$$

$$Rel(12) = Rel(9) = 0.3658 \quad (4.4\beta)$$

$$Rel(0) = \frac{Rel(1) + Rel(2) + Rel(3) + Rel(4) + Rel(11) + Rel(13)}{6} = 0.5249 \quad (4.4\gamma)$$

όπου $Rel(i)$ η κανονικοποιημένη τιμή της στήλης «Relaxivity» για το δείγμα με αύξοντα αριθμό i. Από την διαδικασία αυτή προέκυψε το κανονικοποιημένο σετ δεδομένων για τα SPIONs ο οποίο παρουσιάζεται στον πίνακα 4.12.

A/A	Magnetic core	Size (nm)	Relaxivity	B0 (T)	Fe/cell	Cell Viablility
0	1.0	0.088	0.525	0.158	0.043	0
1	1.0	0.054	1.000	0.648	0.019	1
2	1.0	0.000	0.923	0.648	0.013	1
3	1.0	0.054	0.891	0.648	0.015	1
4	1.0	0.054	0.118	0.648	0.014	1
5	0.0	0.356	0.609	0.387	0.003	0
6	0.0	1.000	0.605	0.158	0.033	0
7	0.0	1.000	0.605	0.158	0.033	0
8	0.0	0.361	0.255	0.158	0.017	0
9	0.0	0.622	0.366	1.000	0.043	0
10	0.0	0.493	0.819	0.387	1.000	0
11	1.0	0.512	0.000	0.005	0.024	0
12	0.0	0.595	0.366	1.000	0.046	0
13	1.0	0.027	0.217	0.005	0.034	1
14	0.0	0.044	0.031	0.387	0.046	0
15	0.0	0.427	0.088	0.000	0.000	0

Πίνακας 4.12 Κανονικοποιημένο σετ δεδομένων των δεδομένων των SPIONs

4.11 Εφαρμογή Γενετικού Αλγόριθμου TPOT και Εκπαίδευση Μοντέλου

Όπως και στην ανάλυση των MWCNTs, έτσι και εδώ, πριν την αναζήτηση και εφαρμογή του προβλεπτικού μοντέλου απαραίτητος είναι ο διαχωρισμός του σετ δεδομένων σε εκπαιδευτικό και σε επαληθευτικό. Για τον σκοπό αυτό χρησιμοποιήθηκε και πάλι η Kennard-Stone μεθοδολογία για να αποκλειστούν οι έκτροπες τιμές από το επαληθευτικό σετ δεδομένων. Η μέθοδος αυτή εκτελέστηκε αναζητώντας 5 δείγματα που θα συνθέσουν το επαληθευτικό σετ δεδομένων ενώ τα άλλα 11 αποτελούν το εκπαιδευτικό. Τα επαληθευτικά δείγματα που προέκυψαν είναι αυτά που έχουν αύξοντες αριθμούς 2, 3, 7, 8 και 12 στον Πίνακα 4.12. Για την εύρεση ενός μοντέλου που να περιγράφει ικανοποιητικά τις συσχετίσεις των μεταβλητών και του τελικού σημείου του σετ δεδομένων των οξειδίων χρησιμοποιήθηκε ένας από τους πιο διαδεδομένους γενετικούς αλγορίθμους, το TPOT.[46,47]

Το πακέτο TPOT αναζητά μέσα από όλα τα μοντέλα, που υπάρχουν διαθέσιμα, εκείνο, ή το συνδυασμό εκείνων, που έχει την υψηλότερη «cross-validation» τιμή στα εκπαιδευτικά δεδομένα. Για την εφαρμογή του απαραίτητος είναι ο προσδιορισμός του χρόνου εκτέλεσης και ο αριθμός των «folds» που θα χρησιμοποιηθεί για τη διασταυρούμενη επαλήθευση. Στο σετ δεδομένων των SPIONs η αναζήτηση του μοντέλου με την χρήση του TPOT διήρκησε 10 λεπτά και η επαλήθευση των μοντέλων έγινε χρησιμοποιώντας 3 «folds». Το μοντέλο που προέκυψε από αυτήν την διαδικασία ήταν ένα LR μοντέλο με L2 penalty και C ίση με 20.

4.12 Αποτελέσματα του εξαγόμενου LR Μοντέλου

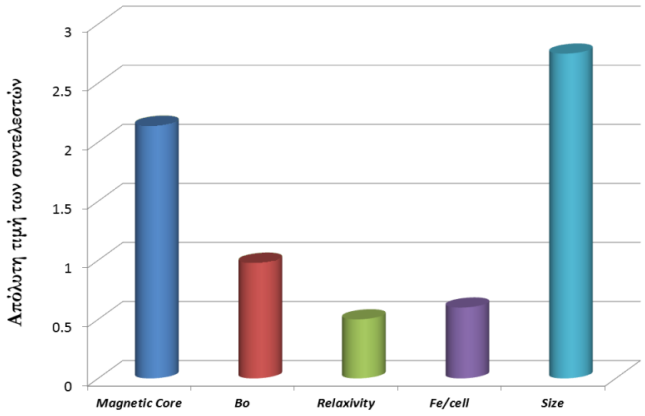
Το κριτήριο αξιολόγησης των μοντέλων που εξέτασε το TPOT, όπως προαναφέρθηκε στην προηγούμενη παράγραφο ήταν μια «3-fold cross validation». Η μετρική αυτή τιμή για το μοντέλο που προέκυψε από την διαδικασία ήταν 0.83 με τυπική απόκλιση ίση με 0.17. Η υψηλή αυτή τιμή δείχνει πως το μοντέλο αυτό δεν έχει υπέρ-προσαρμοστεί στα εκπαιδευτικά δεδομένα καθώς δεν επηρεάζεται με την αλλαγή τους. Οι υπόλοιπες μετρικές τιμές που χρησιμοποιήθηκαν για την αξιολόγηση του LR μοντέλου είναι οι ίδιες με αυτές που χρησιμοποιήθηκαν για την αξιολόγηση των μοντέλων των MWCNTs, και τα αποτελέσματα παρουσιάζονται στον πίνακα 4.13.

	Μετρικές τιμές στο Εκπαιδευτικό Σετ δεδομένων		**Μετρικές τιμές στο Επαληθευτικό Σετ δεδομένων**	
Accuracy Score	91%		100%	
Precision	0.75		1.0	
Recall	1.0		1.0	
F1-Score	0.86		1.0	
Confusion Matrix	7	1	3	0
	0	3	0	2

Πίνακας 4.13 Μετρικές τιμές του LR μοντέλου στο εκπαιδευτικό και στο επαληθευτικό σετ δεδομένων

Όπως φαίνεται το μοντέλο κατάφερε να αναγνωρίσει τα μοτίβα που υπάρχουν στο σετ δεδομένων. Συγκεκριμένα, φαίνεται πως οι προβλέψεις του μοντέλου ήταν αλάνθαστες στο επαληθευτικό σετ δεδομένων ενώ ταξινομήθηκε λάθος ένα μόνο δείγμα του εκπαιδευτικού σετ δεδομένων. Η λάθος αυτή ταξινόμηση δεν αποτελεί πρόβλημα καθώς όπως εξηγήθηκε στο Κεφάλαιο 2, με τον τρόπο αυτό εξασφαλίζεται μεγαλύτερη αποτελεσματικότητα του μοντέλου.

Όπως και στην προηγούμενη ανάλυση, έτσι και στο παρών μοντέλο υπολογίστηκαν οι τιμές της σπουδαιότητας κάθε μεταβλητής, με στόχο τον εντοπισμό των μεταβλητών που επηρεάζουν την πρόβλεψη του μοντέλου. Οι τιμές αυτές προκύπτουν από το συντελεστή της κάθε μεταβλητής στον εκθέτη της συνάρτησης απόφασης και παρουσιάζονται στο διάγραμμα της εικόνας 4.8. Φαίνεται, πως οι δύο πιο σημαντικές μεταβλητές είναι οι «Magnetic core» και «Size» και μάλιστα η διαφορά τους με τις υπόλοιπες είναι ιδιαιτέρως σημαντική. Για τον λόγο αυτό, κατασκευάστηκε και εκπαιδεύτηκε ένα νέο μοντέλο με μόνες μεταβλητές εισόδου αυτές τις δύο.



Εικόνα 4.8 **Διάγραμμα με τις τιμές της σπουδαιότητας των μεταβλητών του σετ δεδομένων των SPIONs**

4.13 Ελάττωση Διαστατικότητας και Αποτελέσματα

Τα αποτελέσματα της ως τώρα ανάλυσης δείχνουν πως το μαγνητικό πεδίο, οι χρόνοι χαλάρωσης και η προσροφημένη ποσότητα σιδήρου ανά κύτταρο δεν έχουν ενεργό ρόλο στην ταξινόμηση ενός οξειδίου ως προς την τοξικότητά του. Στα πλαίσια αυτά, οι πληροφορίες των μεταβλητών αυτών αφαιρέθηκαν από το σετ δεδομένων και το TPOT έτρεξε ξανά, αναζητώντας ένα νέο μοντέλο που να περιγράφει την συσχέτιση της τοξικότητας με τον τύπο του μαγνητικού πυρήνα και το μέγεθος των SPIONs. Το μοντέλο που προέκυψε από αυτήν τη διαδικασία ήταν παρόμοιο με το προηγούμενο μοντέλο, γεγονός που είναι λογικό αν σκεφτεί κανείς πως οι μεταβλητές που αφαιρέθηκαν δεν έπαιζαν σημαντικό ρόλο στην λήψη της απόφασης. Πιο συγκεκριμένα, το νέο μοντέλο είναι μία Λογιστική Παλινδρόμηση με L1 penalty και C ίσο με 40. Τα αποτελέσματα των μετρικών τιμών για το μοντέλο αυτό παρουσιάζονται στον πίνακα 4.14 και δείχνουν πως το μοντέλο όχι μόνο διατήρησε την απόδοση του στα επαληθευτικά δεδομένα αλλά κατάφερε και ταξινόμησε σωστά ακόμα και το δείγμα που στο προηγούμενο σετ δεδομένων απέτυχε να ταξινομήσει. Αυτό το αποτέλεσμα σημαίνει πως οι στήλες που αφαιρέθηκαν τελικά δεν περνούσαν απαρατήρητες από το μοντέλο αλλά προσέθεταν θόρυβο, με αποτέλεσμα να επηρεάζεται αρνητικά η προβλεπτική του ικανότητα.

	Μετρικές τιμές στο Εκπαιδευτικό Σετ δεδομένων		**Μετρικές τιμές στο Επαληθευτικό Σετ δεδομένων**	
Accuracy Score	100%		100%	
Precision	1.0		1.0	
Recall	1.0		1.0	
F1-Score	1.0		1.0	
Confusion Matrix	8	0	3	0
	0	3	0	2

Πίνακας 4.14 **Μετρικές τιμές του LR μοντέλου μετά την ελάττωση της διαστατικότητας του**

Στην εξίσωση 4.5 παρουσιάζεται η συνάρτηση πιθανότητας για το τελικό μοντέλο LR. Η απλοϊκή της μορφή γίνεται ακόμα απλούστερη αν σκεφτεί κανείς πως η μία από τις δύο μεταβλητές της (Magnetic core) είναι διακριτή με δύο μόνο δυνατές τιμές.

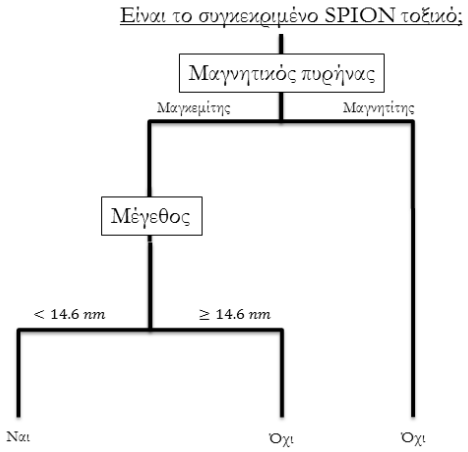
$$p(y=1) = \frac{e^{4.18 \times Magnetic\ core - 29.69 \times Scaled(size) - 1.63}}{1 + e^{4.18 \times Magnetic\ core - 29.69 \times Scaled(size) - 1.63}} \quad (4.5)$$

Η συνάρτηση αυτή δίνει πολύ καλά αποτελέσματα για τα δεδομένα που σετ δεδομένων των SPIONs, όπως φαίνεται και στον πίνακα 4.15.

Α/Α	Νανοϋλικό	Εκπαιδευτικό (Τ) Επαληθευτικό (V)	Πραγματική κλάση	Κλάση Πρόβλεψης	Πιθανότητα για κλάση «0»	Πιθανότητα για κλάση «1»
0	Fe2O3-PLL	T	0	0	0.52	0.48
1	Uncoated γ-Fe_2O_3	T	1	1	0.28	0.72
2	D-mannose-coated-γ-Fe2O3	T	1	1	0.07	0.93
3	Fe2O3-PLL	V	1	1	0.28	0.72
4	PDMAAm-coated-γ-Fe2O3-PLL	V	1	1	0.28	0.72
5	N-dodecyl-PEI2k/SPIO	T	0	0	1.00	0.00
6	iron oxide-loaded cationic nanovesicle	T	0	0	1.00	0.00
7	iron oxide-loaded cationic nanovesicle	V	0	0	1.00	0.00
8	CMCS-SPIONs	V	0	0	1.00	0.00
9	ED-Pullulan coating SPIO	T	0	0	1.00	0.00
10	IONP-6PEG-HA	T	0	0	1.00	0.00
11	PDMAAm-coated-γ-Fe2O3-PLL	T	0	0	1.00	0.00
12	Citrate SPION	V	0	0	1.00	0.00
13	D-mannose-coated SPIONs	T	1	1	0.15	0.85
14	SPIO@SiO2-NH2	T	0	0	0.95	0.05
15	TAT-CLIO	T	0	0	1.00	0.00

Πίνακας 4.15 **Προβλέψεις σε όλα τα δεδομένα του σετ δεδομένων και οι αντίστοιχες πιθανότητες για την κάθε κλάση**

Από τη συνάρτηση αυτή προκύπτει ένας απλός και ιδιαιτέρως εύχρηστος κανόνας, ο οποίος παρουσιάζεται με την μορφή δέντρου απόφασης στην εικόνα 4.9, όσον αφορά την τοξικότητα των SPIONs που χρησιμοποιούνται για την απεικόνιση βλαστικών κυττάρων. *Τα SPIONs με πυρήνα μαγνητίτη είναι μη τοξικά, ενώ τα SPIONs με πυρήνα μαγκεμίτη είναι τοξικά αν το μήκος τους είναι μικρότερο από 14.6 nm*. Είναι φανερό πως το αποτέλεσμα αυτό «ρίχνει φως» στη μελέτη της τοξικότητας των συγκεκριμένων SPIONs. Ο τρόπος με τον οποίο καταλήξαμε σε αυτό το συμπέρασμα παρουσιάζεται στο Παράρτημα Γ.



Εικόνα 4.9 Παρουσίαση του αποτελέσματος της LR για τα SPIONs σε δέντρο απόφασης.

Ο παραπάνω κανόνας είναι ένα ιδιαιτέρως σημαντικό αποτέλεσμα καθώς μας βοηθάει να κατανοήσουμε καλύτερα τον μηχανισμό με τον οποίο τα SPIONs είναι τοξικά. Ωστόσο θα πρέπει να εφαρμόζεται με προσοχή καθώς η εφαρμογή του σε ένα SPION εκτός του πεδίου εφαρμογής του μοντέλου μπορεί να έχει σημαντικές επιπτώσεις. Επομένως, τελικό στάδιο της ανάλυσης αυτής και πάλι ήταν ο υπολογισμός του πεδίου εφαρμογής του μοντέλου. Αυτό επετεύχθη και πάλι με την μέθοδο Leverage, και εδώ η τιμή του $threshold$ ήταν 0.55. *Άρα ο παραπάνω κανόνας είναι αξιόπιστος για τα SPIONs με τιμή Leverage μικρότερη από την τιμή 0.55*. Το τελικό μοντέλο «ανέβηκε» στην πλατφόρμα του Jaqpot και είναι διαθέσιμο για χρήση στους ερευνητές (https://app.jaqpot.org/model/DcWnWFp9GESI16R4o2av).

Κεφάλαιο 5

Συμπεράσματα και Προτάσεις για Μελλοντική Έρευνα

5.1 Συμπεράσματα

Στη διπλωματική αυτή εργασία εξετάσαμε δύο προβλήματα τοξικότητας νανοϋλικών. Στο πρώτο πρόβλημα, εξετάστηκε η τοξικότητά των MWCNTs ως προς τις διασπάσεις της αλυσίδας DNA (γονοτοξικότητά) και στο δεύτερο πρόβλημα εξετάστηκε η τοξικότητά των SPIONs ως προς τη βιωσιμότητα των βλαστικών κυττάρων στα οποία εμφυτεύθηκαν. Τόσο στο πρώτο πρόβλημα όσο και στο δεύτερο, λόγω του μικρού όγκου δεδομένων, επιλέχθηκε η ταξινόμηση των νανοϋλικών σε δύο μόνο κλάσεις, τα «τοξικά» και τα «μη τοξικά». Για την εύρεση του βέλτιστου αλγορίθμου εξετάστηκαν εναλλακτικές μεθοδολογίες. Στο πρόβλημα των MWCNTs φαίνεται πως τα δεδομένα ακολουθούν μία εκθετική κατανομή και για τον λόγο αυτό η Gaussian NB που υποθέτει Γκαουσιανή κατανομή δεν είχε καλά αποτελέσματα, σε αντίθεση με την LR. Στο πρόβλημα των SPIONs παρατηρείται επίσης εκθετική κατανομή, αφού φαίνεται η LR να δίνει και πάλι περισσότερο ακριβή και εύρωστα αποτελέσματα.

Το σετ δεδομένων των MWCNTs αρχικά αποτελείτο από 31 μεταβλητές, ένα πλήθος δυσανάλογο του αριθμού δειγμάτων που είχαμε. Για τη μοντελοποίησή των δεδομένων αυτών ακολουθήθηκαν δύο διαφορετικές υπολογιστικές ροές. Στην πρώτη υπολογιστική διαδικασία, εφαρμόστηκε μετασχηματισμός των δεδομένων σύμφωνα με την τεχνική PCA και το μοντέλο που προέκυψε παρουσίασε ακρίβεια 100% στα δεδομένα επαλήθευσης. Στη δεύτερη μεθοδολογία χρησιμοποιήθηκε αρχικά το σύνολο των κανονικοποιημένων δεδομένων και σε αυτά εφαρμόστηκε μια σειρά από αλγόριθμους μηχανικής μάθησης που βελτιστοποιήθηκαν, πλην της μεθόδου PLS με την μπεϋζιανή τεχνική. Τα μοντέλα αυτά δεν παρουσίασαν ακρίβεια 100%, αφού δεν κατάφεραν να ταξινομήσουν σωστά ένα συγκεκριμένο δείγμα. Στην συνέχεια εφαρμόστηκε η μεθοδολογία RFE για την απομάκρυνση των στατιστικά μη σημαντικών μεταβλητών. Στο μειωμένο σετ δεδομένων επιλέχθηκαν μόνο τέσσερις μεταβλητές και σε αυτό εκπαιδεύτηκε μοντέλο με τη μέθοδο LR που περιέγραψε με πολύ καλή ακρίβεια (100%) την τοξικότητά τους. Οι τέσσερις μεταβλητές που επέλεξε η μέθοδος RFE ήταν αναμενόμενες σύμφωνα με τη βιβλιογραφία, γεγονός που ενισχύει την αξιοπιστία του μοντέλου. Συγκεκριμένα, από αυτήν την υπολογιστική ροή αποδείχθηκε πως η τοξικότητα των MWCNTs συνδέεται με την επιφάνειά τους (ύπαρξη υδροξυλίου ως επιφανειακή λειτουργική ομάδα και καθαρότητα), ενώ σημαντικές μεταβλητές ήταν ακόμα το μέγεθος Z – average και ο δείκτης πολυδιασποράς. Το συγκεκριμένο μοντέλο αν και παρουσιάζει την ίδια ακρίβεια με το μοντέλο PCA έχει το σημαντικό πλεονέκτημα ότι απαιτεί πολύ λιγότερους δείκτες για να υπολογίσει την πρόβλεψη.

Για το σετ δεδομένων των SPIONs ακολουθήθηκε παρόμοια διαδικασία με αυτή των MWCNTs. Η διαφορά σε αυτήν την περίπτωση ήταν στον τρόπο με τον οποίο επιλέχθηκε ο κατάλληλος αλγόριθμος μηχανικής μάθησης. Το αρχικό σετ δεδομένων αποτελείτο από λιγότερες μεταβλητές, επομένως δεν υπήρχε εξαρχής το πρόβλημα των δυσανάλογα πολλών μεταβλητών. Για την επιλογή του καλύτερου μοντέλου εφαρμόστηκε ο αλγόριθμος γενετικού προγραμματισμού TPOT, που επέλεξε την LR ως την πιο κατάλληλη μέθοδο, και αξιολογήθηκε με *100% ακρίβεια* στα δεδομένα επαλήθευσης και *0.92 τιμή διασταυρούμενης επαλήθευσης* στα δεδομένα εκπαίδευσης. Αν και αυτό το μοντέλο ήταν ιδιαίτερα ικανοποιητικό, παρατηρώντας τις τιμές της σημαντικότητας των μεταβλητών φάνηκε να υπάρχει περιθώριο βελτίωσης. Πράγματι, το τελικό μοντέλο εκπαιδεύτηκε στις δύο πιο στατιστικά σημαντικές μεταβλητές, τον τύπο του μαγνητικού πυρήνα και το μέγεθος των οξειδίων, δίνοντας *100% ακρίβεια* στο σύνολο των δεδομένων και πολύ υψηλές πιθανότητες βεβαιότητας στις αποφάσεις του. Από το αποτέλεσμα αυτό, προκύπτει το συμπέρασμα πως μόνο με τον τύπο του πυρήνα και με το μέγεθος ενός SPION, είναι δυνατή η πρόβλεψη της τοξικότητάς του και μάλιστα με σημαντική στατιστική πιθανότητα.

Αναλύοντας τον απλό τύπο της συνάρτησης ταξινόμησης των SPIONs καταλήξαμε σε έναν ιδιαιτέρως απλό και εύχρηστο κανόνα, όσον αφορά την τοξικότητα των SPIONs. *Σύμφωνα με τον κανόνα αυτό, τα SPIONs με πυρήνα μαγνητίτη δεν είναι τοξικά, ενώ τα SPIONs με πυρήνα μαγκεμίτη είναι τοξικά στην περίπτωση όπου το μέγεθος τους είναι μικρότερο των 14.6 nm. Από αυτόν τον κανόνα προκύπτει το συμπέρασμα, πως όσο μικρότερα είναι τα SPIONs τόσο και πιο τοξικά είναι.* Η συμπεριφορά τους αυτή μπορεί να οφείλεται στο γεγονός πως όσο πιο μικρά είναι τόσο πιο εύκολα μπορούν να εισέλθουν στα οργανίδια του κυττάρου, προκαλώντας την καταστροφή τους.

5.2 Προτάσεις για Μελλοντική Έρευνα

Σκοπός αυτής της εργασίας ήταν η ανάπτυξη μοντέλων QSAR για τη συσχέτιση της τοξικότητας δύο τύπων νανοϋλικών, που χρησιμοποιούνται σε βιοϊατρικές εφαρμογές, με τις φυσικοχημικές τους ιδιότητες και τελικά την πρόβλεψη της τοξικότητας σε άγνωστα δείγματα. Αν και τα αποτελέσματα και συμπεράσματα που προέκυψαν από αυτήν την έρευνα έδειξαν πως τα μοντέλα που αναπτύχθηκαν έδωσαν πολύ καλή ακρίβεια στις προβλέψεις τους, αναδείχθηκε και σε αυτή τη μελέτη η περιορισμένη διαθεσιμότητα δεδομένων στο χώρο της νανοπληροφορικής. Μια πλουσιότερη συλλογή δεδομένων θα μπορέσει να βελτιώσει όλα τα στάδια εκπαίδευσης των μοντέλων μηχανικής μάθησης, να ενισχύσει την αξιοπιστία τους και να διευρύνει το πεδίο εφαρμογής τους. Σε αυτήν την κατεύθυνση μπορούν να συμβάλουν οι ανοιχτές διαδικτυακές βάσεις δεδομένων που ήδη αρχίζουν να αναπτύσσονται, στις οποίες συγκεντρώνονται, ενοποιούνται και εναρμονίζονται τα αποτελέσματα που παράγονται από τις πειραματικές ερευνητικές ομάδες με την παράλληλη ανάπτυξη σχετικών οντολογιών. Ακόμη, τα δεδομένα εκπαίδευσης των μοντέλων θα μπορούσαν να πολλαπλασιαστούν με τεχνικές «resampling», οι οποίες κατασκευάζουν εικονικά δεδομένα, που επιπλέον μπορούν να εξασφαλίσουν καλύτερες κατανομές. Τέλος προτείνεται η συνεργασία με τις πειραματικές μονάδες, καθώς τα μοντέλα QSAR μπορούν να προσφέρουν πολύτιμες κατευθύνσεις για το σχεδιασμό και τη διεξαγωγή περιορισμένων αλλά στοχευμένων πειραμάτων, που να είναι όμως σε θέση να προσφέρουν πλούσια σε χρήσιμες πληροφορίες δεδομένα.

Παράρτημα

Παράρτημα Α' Μελέτη της συνάρτησης κόστους της LR

Παράρτημα Β' Πίνακες δεδομένων για το MWCNTs σετ δεδομένων

Παράρτημα Γ' Μελέτη της συνάρτησης ταξινόμησης της LR για τα SPIONs

Παράρτημα Α'

Μελέτη της συνάρτησης κόστους της Logistic Regression

Για την κατασκευή της συνάρτησης κόστους θα θεωρήσουμε πρώτα την συνάρτηση σύμφωνα με την οποία η LR υπολογίζει την πιθανότητα ενός στοιχείου να ανήκει στην κλάση «1» για την γενική περίπτωση ενός συνόλου δεδομένων $m \times n$ διαστάσεων:

$$p(y_i = 1) = \frac{e^{a_0 + a_1 x_{i,1} + a_2 x_{i,2} + \cdots + a_{i,n-1} x_{i,n-1} + a_n x_{i,n}}}{1 + e^{a_0 + a_1 x_{i,1} + a_2 x_{i,2} + \cdots + a_{i,n-1} x_{i,n-1} + a_n x_{i,n}}} \rightarrow \quad A1$$

$$p(y_i = 1) = \frac{1}{1 + e^{-a_0 - a_1 x_{i,1} - a_2 x_{i,2} - \cdots - a_{n-1} x_{i,n-1} - a_n x_{i,n}}} \quad A2$$

όπου x_i οι τιμές των μεταβλητών της παρατήρησης i. Την εξίσωση Α2 θα την εκφράσουμε με στην μορφή της Α3:

$$p(y_i = 1) = \varphi(z_i) \quad A3$$

όπου $\varphi(t) = \frac{1}{1+e^{-t}}$ και $z_i = a_0 + \sum_{i=1}^{n} a_i x_i$.

Αν η πιθανότητα ενός στοιχείου να ανήκει στην κλάση «1» είναι $\varphi(z_i)$ τότε η πιθανότητα του ίδιου στοιχείου να ανήκει στην κλάση «0» θα είναι $1 - \varphi(z_i)$. Η συνάρτηση εκτίμησης της μέγιστης πιθανότητας (Maximum Likelihood Estimator, MLE) ορίζεται ως το γινόμενο των πιθανοτήτων των στοιχείων ενός συνόλου δεδομένων να ανήκουν στην πραγματική τους κλάση. Για παράδειγμα, σε ένα σετ δεδομένων που αποτελείται από 5 παρατηρήσεις, 3 από τις οποίες ανήκουν στην κλάση «1» και 2 στην κλάση «0», όπως αυτό του ακόλουθου πίνακα, η συνάρτηση MLE θα είναι:

Α/Α	Μεταβλητές	Πραγματική κλάση
1	…	1
2	…	1
3	…	0
4	…	1
5	…	0

$$MLE = p(y_1 = 1) \cdot p(y_2 = 1) \cdot p(y_3 = 0) \cdot p(y_4 = 1) \cdot p(y_5 = 0) \quad A4$$

Γνωρίζοντας τις τιμές των μεταβλητών για κάθε παρατήρηση, από τις εξισώσεις Α3 και Α4, είναι φανερό πως οι ανεξάρτητες μεταβλητές της MLE είναι οι συντελεστές α_i των μεταβλητών. Για τον λόγο αυτό την συνάρτηση MLE θα την συμβολίζουμε $L(a)$, όπου α θεωρείται ένα διάνυσμα με τις συνιστώσες a_i. Για να κατασκευάσουμε έναν καθολικό τύπο που να εκφράζει την $L(a)$ θα πρέπει να λάβουμε υπ' όψιν μας πως όταν η πραγματική κλάση μίας παρατήρησης είναι «1» τότε η πιθανότητα θα είναι $\varphi(z_i)$, ενώ όταν η πραγματική κλάση μίας παρατήρησης είναι «0» τότε η πιθανότητα θα είναι $1 - \varphi(z_i)$.

$$L(a) = \prod_{i=1}^{m} [\varphi(z_i)^{y_i}(1-\varphi(z_i))^{1-y_i}] \quad A5$$

Η συνάρτηση κόστους $J(a)$ υπολογίζεται ως ο αρνητικός νεπέριος λογάριθμος της $L(a)$ και είναι:

$$J(a) = -ln\big(L(a)\big) \rightarrow \quad A6$$

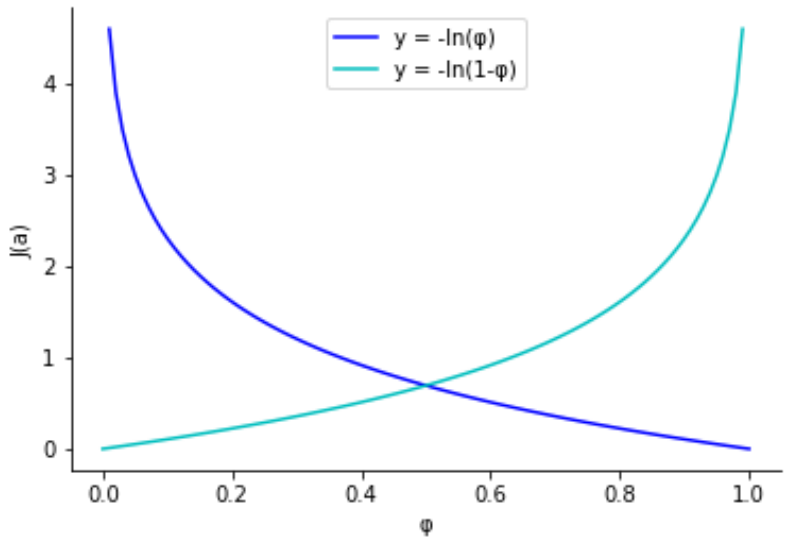
$$J(a) = -ln\left(\prod_{i=1}^{m} [\varphi(z_i)^{y_i}(1-\varphi(z_i))^{1-y_i}]\right) \rightarrow \quad A7$$

$$J(a) = -\sum_{i=1}^{m} ln\left(\varphi(z_i)^{y_i}\big(1-\varphi(z_i)\big)^{1-y_i}\right) \rightarrow \quad A8$$

$$J(a) = \sum_{i=1}^{m} \big(-y_i \ln \varphi(z_i) - (1-y_i)\ln\big(1-\varphi(z_i)\big)\big) \quad A9$$

Η συνάρτηση κόστους τελικά δίνεται από την εξίσωση Α9. Για την σχηματική της απεικόνιση θα θεωρήσουμε πως έχουμε ένα στοιχείο, δηλαδή $m = 1$. Τότε η συνάρτηση δίνεται από την εξίσωση 10 και στο σχήμα που ακολουθεί απεικονίζεται γραφικά.

$$J(a) = -y \cdot ln(\varphi) - (1-y) \cdot ln(1-\varphi) = \begin{cases} -ln(\varphi) & ,\alpha\nu\ y = 1 \\ -ln(1-\varphi) & ,\alpha\nu\ y = 0 \end{cases} \quad A10$$



Παράρτημα Β'

Πίνακες δεδομένων για το MWCNTs σετ δεδομένων

Στο παράρτημα αυτό παρουσιάζονται οι σχετικού με την ανάλυση των MWCNTs πίνακες. Συγκεκριμένα, οι πίνακες που υπάρχουν σε αυτό το παράρτημα είναι:

B1: Ολοκληρωμένο σετ δεδομένων για τους MWCNTs, πριν οποιαδήποτε προεπεξεργασία.

B2: Κανονικοποιημένο σετ δεδομένων για τους MWCNTs

B3: Σετ δεδομένων κύριων συνιστωσών για τους MWCNTs

Πίνακας B1

A/A	Code	Type	Length ave.	Diameter ave.	Diameter max	Diameter min	BET	%Fe	%Co	%Ni
0	**NRCWE - 040**	PRISTI-NE	518.9	22.1	29.9	14.3	150	0.2	0.001	518.9
1	**NRCWE - 041**	OH	1005	26.9	37	16.8	152	0.13	0.001	1005
2	**NRCWE - 042**	COOH	723.2	30.2	44.4	16	141	0.08	0	723.2
3	**NRCWE - 043**	PRISTI-NE	771.3	55.6	73.7	37.5	82	0.008	0.001	771.3
4	**NRCWE - 044**	OH	1330	32.7	46.3	19.1	74	0.004	0.002	1330
5	**NRCWE - 045**	COOH	1553	30.2	45.8	14.6	119	1.17	0.25	1553
6	**NRCWE - 046**	PRISTI-NE	717.2	29.1	45.2	13	223	0.008	0.25	717.2
7	**NRCWE - 047**	OH	532.5	22.6	32.7	12.5	216	0.007	0.25	532.5
8	**NRCWE - 048**	COOH	1604	17.9	35.8	0	185	0.007	0.24	1604
9	**NRCWE - 049**	NH2	731.1	14.9	20.5	9.3	199	0.004	0.25	731.1
10	**NM-400**	PRISTI-NE	847	11	14	8	254	0.2607	0.1063	847
11	**NM-401**	PRISTI-NE	4048	67	91	43	18	0.05	-	4048
12	**NM-402**	PRISTI-NE	1372	11	14	8	226	1.31	-	1372
13	**NM-403**	PRISTI-NE	1373	13	16	10	227	2.31	-	1373
14	**NRCWE -006**	PRISTI-NE	5700	65	90	40	26	0.680	-	5700
15	**NRCWE - 040**	PRISTI-NE	518.9	22.1	29.9	14.3	150	0.2	0.001	518.9

(συνεχίζεται)

A/A	%Mg	%Mn	Exposure	Day	Deposited Dose(μg)	Purity (%)	Manuf. Length ave.	Manuf. Length max.	Manuf. Length min.	Zave batch
0	0.01	0.002	Intratracheal Instillation	1, 28, 92	6, 18, 54	98.60	30.0	50	10	195.6
1	0.02	0.001	Intratracheal Instillation	1, 28, 92	6, 18, 54	99.20	30.0	50	10	218.7
2	0.03	0.001	Intratracheal Instillation	1, 28, 92	6, 18, 54	99.20	30.0	50	10	195.2
3	0.01	-	Intratracheal Instillation	1, 28, 92	6, 18, 54	98.50	15.0	20	10	177.0
4	0.02	-	Intratracheal Instillation	1, 28, 92	6, 18, 54	98.60	15.0	20	10	202.2
5	0.02	0.002	Intratracheal Instillation	1, 28, 92	6, 18, 54	96.30	15.0	20	10	192.4
6	0.22	0.3	Intratracheal Instillation	1, 28, 92	6, 18, 54	98.70	6.5	12	1	182.9
7	0.22	0.3	Intratracheal Instillation	1, 28, 92	6, 18, 54	98.70	6.5	12	1	200.1
8	0.19	0.28	Intratracheal Instillation	1, 28, 92	6, 18, 54	98.80	6.5	12	1	182.2
9	0.19	0.29	Intratracheal Instillation	1, 28, 92	6, 18, 54	98.80	6.5	12	1	212.5
10	-	-	Intratracheal Instillation	92	54	90.00	1.5	-	-	168.1
11	0.015	-	Intratracheal Instillation	92	54	99.19	10	15	5	923.8
12	0.001	0.001	Intratracheal Instillation	1, 28, 92	6, 18, 54	92.97	5.5	10	1	171.3
13	1.001	1.001	Intratracheal Instillation	1, 28, 93	6, 18, 55	90.00	5.05	10	0.1	200.2
14	-	-	-	-	-	99.00	10.00	19.00	1.00	633.3

A/A	σ	PdI batch	Zave (12,5)	σ	PdI (12,5)	σ	Zave (200)	σ	PdI (200)	σ
0	4.7	0.358	0.018	159	3	0.398	0.018	262	0.408	0.014
1	5.6	0.501	0.012	236	8	0.484	0.039	235	0.496	0.025
2	3.2	0.382	0.020	155	1	0.314	0.020	190	0.425	0.018
3	1.4	0.237	0.008	145	2	0.254	0.009	154	0.306	0.020
4	1.7	0.236	0.011	437	13	0.478	0.065	605	0.580	0.033
5	2.0	0.227	0.008	336	28	0.477	0.100	295	0.391	0.026
6	5.8	0.434	0.045	598	28	0.504	0.083	1159	0.497	0.052
7	4.6	0.402	0.014	244	6	0.446	0.044	307	0.484	0.030
8	3.3	0.397	0.018	152	3	0.465	0.024	128	0.372	0.011
9	4.3	0.602	0.028	197	6	0.867	0.052	217	0.639	0.095
10	1.6	0.385	0.015	274	11	0.582	0.125	270	0.592	0.119
11	33.9	0.323	0.049	570	29	0.463	0.035	652	0.502	0.058
12	3.1	0.407	0.022	456	6	0.482	0.040	213	0.465	0.016
13	3.9	0.407	0.010	587	18	0.514	0.017	291	0.506	0.021
14	10.8	0.256	0.022	832	24	0.459	0.042	2102	0.448	0.090

(συνεχίζεται)

A/A	ROS	Peak	0 µg/ml viab.	12.5 µg/ml viab.	25 µg/ml viab.	50 µg/ml viab.	100 µg/ml viab.	200 µg/ml viab.	0 µg/ml prolif.	12.5 µg/ml prolif.
0	1.7	12.50	94	95	96	97	97	99	100	95
1	3.0	12.50	93	94	96	96	96	98	100	79
2	0.8	12.50	94	95	96	97	98	99	100	89
3	2.3	12.50	96	96	98	98	99	99	100	105
4	1.7	12.50	96	96	97	98	98	99	100	99
5	1.5	12.50	94	97	96	98	98	98	100	95
6	6.6	6.25	96	98	98	98	99	99	100	93
7	2.1	6.25	97	99	99	99	99	99	100	99
8	2.1	6.25	97	97	98	98	99	100	100	95
9	3.4	12.50	96	97	97	99	99	98	100	95
10	5.1	5.60	97	97	98	99	99	98	100	105
11	0.0	11.25	96	96	94	94	95	95	100	77
12	7.7	12.50	97	97	98	98	99	96	100	113
13	1.9	12.50	96	97	98	98	99	98	100	83
14	0.3	11.25	95	94	93	93	91	94	100	75

A/A	25 µg/ml prolif.	50 µg/ml prolif.	100 µg/ml prolif.	200 µg/ml prolif.	CEA: C,H,O,N (wt%)	OH mmol/g	COOH mmol/g	OH mmol/m2	COOH mmol/m2	Endotoxins (EU/mg)
0	85	60	67	62	96.00	0.35	0.18	0.0023	0.0012	0.18
1	62	50	54	59	97.00	1.69	0.84	0.0110	0.0056	–
2	96	72	79	56	96.00	4.09	2.04	0.0290	0.0069	0.26
3	115	114	111	101	96.00	0.18	0.09	0.0022	0.0011	0.25
4	96	95	91	77	97.00	0.23	0.11	0.0030	0.0015	0.27
5	100	104	79	29	93.00	0.63	0.31	0.0053	0.0011	0.34
6	88	112	97	10	96.00	0.63	0.32	0.0028	0.0014	0.19
7	105	90	79	19	97.00	0.26	0.13	0.0012	0.0006	0.01
8	97	84	81	42	96.00	0.58	0.29	0.0031	0.0012	0.03
9	80	83	79	20	97.00	0.33	0.16	0.0016	0.0008	0.05
10	105	121	102	115	88.00	0.79	0.40	0.0032	0.0016	ND
11	64	52	50	47	98.00	0.03	0.02	0.0017	0.0011	0.42
12	100	99	104	47	92.00	0.28	0.14	0.0012	0.0006	0.01
13	94	77	78	14	97.00	0.19	0.09	0.0014	0.0007	0.01
14	73	55	55	55	98.00	0.08	0.04	0.0031	0.0015	0.51

A/A	Geno-toxicity
0	0
1	0
2	0
3	0
4	1
5	1
6	0
7	0
8	0
9	0
10	1
11	0
12	1
13	1
14	1

Πίνακας Β2

A/A	Length ave.	Diameter ave.	Diameter max	Diameter min	BET	% Total Impurities	Purity (%)	Zave batch	PdI batch	Zave (12.5)
0	0.0000	0.1982	0.2065	0.3326	0.559	0.1349	0.9348	0.0364	0.3493	0.0204
1	0.0938	0.2839	0.2987	0.3907	0.5678	0.0757	1.0000	0.0670	0.7307	0.1325
2	0.0394	0.3429	0.3948	0.3721	0.5212	0.0498	1.0000	0.0359	0.4133	0.0146
3	0.0487	0.7964	0.7753	0.8721	0.2712	0.2199	0.9239	0.0118	0.0267	0.0000
4	0.1565	0.3875	0.4195	0.4442	0.2373	0.1907	0.9348	0.0451	0.0240	0.4250
5	0.1996	0.3429	0.4130	0.3395	0.4280	0.5177	0.6848	0.0322	0.0000	0.2780
6	0.0383	0.3232	0.4052	0.3023	0.8686	0.1368	0.9456	0.0196	0.5520	0.6594
7	0.0026	0.2071	0.2429	0.2907	0.8390	0.1364	0.9456	0.0423	0.4667	0.1441
8	0.2094	0.1232	0.2831	0.0000	0.7076	0.1250	0.9565	0.0187	0.4533	0.0102
9	0.0410	0.0696	0.0844	0.2163	0.7669	0.1282	0.9565	0.0587	1.0000	0.0757
10	0.0633	0.0000	0.0000	0.1861	1.0000	0.0577	0.0000	0.0000	0.4213	0.1878
11	0.6811	1.0000	1.0000	1.0000	0.0000	0.0000	0.9989	1.0000	0.2560	0.6187
12	0.1647	0.0000	0.0000	0.1861	0.8814	0.2378	0.3228	0.0042	0.4800	0.4527
13	0.1649	0.0357	0.0260	0.2326	0.8856	1.0000	0.0000	0.0425	0.4800	0.6434
14	1.0000	0.9643	0.9870	0.9302	0.0339	0.1172	0.9783	0.6156	0.0773	1.0000

A/A	PdI (12.5)	Zave (200)	PdI (200)	ROS	Peak	0 μg/ml viab.	12.5 μg/ml viab.	25 μg/ml viab.	50 μg/ml viab.	100 μg/ml viab.
0	0.2349	0.0679	0.3063	0.2208	1.0000	0.2500	0.2000	0.5000	0.6667	0.7500
1	0.3752	0.0542	0.5706	0.3896	1.0000	0.0000	0.0000	0.5000	0.5000	0.6250
2	0.0979	0.0314	0.3574	0.1039	1.0000	0.2500	0.2000	0.5000	0.6667	0.8750
3	0.0000	0.0132	0.0000	0.2987	1.0000	0.7500	0.4000	0.8333	0.8333	1.0000
4	0.3654	0.2416	0.8228	0.2208	1.0000	0.7500	0.4000	0.6667	0.8333	0.8750
5	0.3638	0.0846	0.2553	0.1948	1.0000	0.2500	0.6000	0.5000	0.8333	0.8750
6	0.4078	0.5223	0.5736	0.8571	0.0942	0.7500	0.8000	0.8333	0.8333	1.0000
7	0.3132	0.0907	0.5345	0.2727	0.0942	1.0000	1.0000	1.0000	1.0000	1.0000
8	0.3442	0.0000	0.1982	0.2727	0.0942	1.0000	0.6000	0.8333	0.8333	1.0000
9	1.0000	0.0451	1.0000	0.4416	1.0000	0.7500	0.6000	0.6667	1.0000	1.0000
10	0.5351	0.0719	0.8589	0.6623	0.0000	1.0000	0.6000	0.8333	1.0000	1.0000
11	0.3409	0.2654	0.5886	0.0000	0.8188	0.7500	0.4000	0.1667	0.1667	0.5000
12	0.3719	0.043	0.4775	1.0000	1.0000	1.0000	0.6000	0.8333	0.8333	1.0000
13	0.4241	0.0826	0.6006	0.2467	1.0000	0.7500	0.6000	0.8333	0.8333	1.0000
14	0.3344	1.0000	0.4264	0.0390	0.8188	0.5000	0.0000	0.0000	0.0000	0.0000

A/A	200 μg/ml viab	12.5 μg/ml prolif.	25 μg/ml prolif.	50 μg/ml prolif.	100 μg/ml prolif.	200 μg/ml prolif.	CEA: C,H,O, N (wt%)	OH mmol/g	COOH mmol/g	Endo-toxins EU/mg
0	0.8333	0.5263	0.4340	0.1408	0.2787	0.4952	0.8000	0.0788	0.0792	0.3400
1	0.6667	0.1053	0.0000	0.0000	0.0656	0.4667	0.9000	0.4089	0.4059	0.3400
2	0.8333	0.3684	0.6415	0.3099	0.4754	0.4381	0.8000	1.0000	1.0000	0.5000
3	0.8333	0.7895	1.0000	0.9014	1.0000	0.8667	0.8000	0.0369	0.0346	0.4800
4	0.8333	0.6316	0.6415	0.6338	0.6721	0.6381	0.9000	0.0493	0.0445	0.5200
5	0.6667	0.5263	0.7170	0.7606	0.4754	0.1809	0.5000	0.1478	0.1436	0.6600
6	0.8333	0.4737	0.4906	0.8732	0.7705	0.0000	0.8000	0.1478	0.1485	0.3600
7	0.8333	0.6316	0.8113	0.563	0.4754	0.0857	0.9000	0.0566	0.0545	0.0000
8	1.0000	0.5263	0.6604	0.4789	0.5082	0.3048	0.8000	0.1355	0.1337	0.0400
9	0.6667	0.5263	0.3396	0.4648	0.4754	0.0952	0.9000	0.0739	0.0693	0.0800
10	0.6667	0.7895	0.8113	1.0000	0.8525	1.0000	0.0000	0.1872	0.1881	0.4600
11	0.1667	0.0526	0.0377	0.0282	0.0000	0.3524	1.0000	0.0000	0.0000	0.8200
12	0.3333	1.0000	0.7170	0.6901	0.8852	0.3524	0.4000	0.0616	0.0594	0.0000
13	0.6667	0.2105	0.6038	0.3803	0.4590	0.0381	0.9000	0.0394	0.0346	0.0000
14	0.0000	0.0000	0.2075	0.0704	0.0820	0.4286	1.0000	0.0123	0.0099	1.0000

(συνεχίζεται)

A/A	Type COOH	Type NH2	Type OH	Type PRISTINE	Geno-toxicity
0	0.0000	0.0000	0.0000	1.0000	0
1	0.0000	0.0000	1.0000	0.0000	0
2	1.0000	0.0000	0.0000	0.0000	0
3	0.0000	0.0000	0.0000	1.0000	0
4	0.0000	0.0000	1.0000	0.0000	1
5	1.0000	0.0000	0.0000	0.0000	1
6	0.0000	0.0000	0.0000	1.0000	0
7	0.0000	0.0000	1.0000	0.0000	0
8	1.0000	0.0000	0.0000	0.0000	0
9	0.0000	1.0000	0.0000	0.0000	0
10	0.0000	0.0000	0.0000	1.0000	1
11	0.0000	0.0000	0.0000	1.0000	0
12	0.0000	0.0000	0.0000	1.0000	1
13	0.0000	0.0000	0.0000	1.0000	1
14	0.0000	0.0000	0.0000	1.0000	1

Πίνακας Β3

Α/Α	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9	Genotoxicity
0	0.2187	0.0456	0.0425	-0.4852	-0.4652	0.1204	0.1000	0.5607	-0.2241	0
1	0.5249	1.1784	-0.6954	-0.1097	-0.6085	0.5401	-0.3300	0.1612	0.0525	0
2	0.1061	1.3023	0.7826	-0.6204	0.1976	0.3805	0.0527	0.0528	0.2026	0
3	0.0572	-0.5460	1.2166	0.3645	-0.7096	-0.3147	0.4530	0.3280	-0.0164	0
4	0.0367	0.3926	0.0256	0.7897	-0.6642	-0.2099	-0.2244	-0.5362	0.1277	1
5	-0.0509	0.5487	0.7410	-0.3366	0.3305	-0.5234	-0.2004	-0.5444	0.1450	1
6	-0.5025	-0.7042	-0.2013	0.3150	0.5228	0.2102	0.0785	0.4440	0.7402	0
7	-0.9193	0.3723	-0.3978	1.0885	0.1443	-0.1407	-0.3306	0.2361	-0.1325	0
8	-0.7851	0.6034	0.3461	0.2191	0.9078	-0.0598	0.0154	0.2250	-0.4000	0
9	-0.6650	0.4813	-1.0894	-0.2062	0.0315	-0.2615	0.9981	-0.3187	0.0155	0
10	-1.2646	-1.0787	0.2122	-0.0467	0.0366	0.9070	-0.0900	-0.4971	-0.3159	1
11	2.1077	-0.4449	-0.2175	0.1864	0.2671	-0.0380	0.2368	0.0530	-0.4026	0
12	-0.9519	-0.9321	-0.0955	-0.4278	-0.2666	-0.0227	0.0228	-0.1122	0.1385	1
13	-0.4856	-0.6157	-0.5491	-0.7773	-0.0428	-0.6935	-0.6374	0.1789	-0.1151	1
14	2.5738	-0.6031	-0.1206	0.0466	0.3185	0.1060	-0.1443	-0.2312	0.1846	1

Παράρτημα Γ'

Μελέτη της συνάρτησης ταξινόμησης της LR για τα SPIONs

Η συνάρτηση κατανομής πιθανότητας της LR ($prob$), για την κλάση 1, έχει την εξής μορφή:

$$p(y=1) = prob(x_1, x_2) = \frac{e^{ax_1+bx_2+c_0}}{1+\ e^{ax_1+bx_2+c_0}}$$

Όπου: x_1: $Magnetic\ core$, x_2: $Size$ και $a = 4.17639361938979$, $b = -29.962451148168253$, $c_0 = -1.626392858524227$, ενώ οι τιμές που μπορούν να πάρουν τα x_1 και x_2 είναι scaled μεταξύ του 0 και 1.

Ισχύει πως $x_1 = \begin{cases} 0, & \text{άν μαγνητίτης } Magnetic\ core \\ 1, & \text{άν μαγκεμίτης } Magnetic\ core \end{cases}$, επομένως έχουμε για την $prob$:

$$prob(x_1, x_2) = \begin{cases} prob(0, x_2) \\ prob(1, x_2) \end{cases} = \begin{cases} \dfrac{e^{bx_2+c_0}}{1+\ e^{bx_2+c_0}} \\ \dfrac{e^{bx_2+c_0+a}}{1+\ e^{bx_2+c_0+a}} \end{cases} = \begin{cases} \dfrac{e^{bx_2+c_0}}{1+\ e^{bx_2+c_0}} \\ \dfrac{e^{bx_2+c_0'}}{1+\ e^{bx_2+c_0'}} \end{cases}$$

Παρατηρούμε πως οι δύο κλάδοι της συνάρτησης καταλήγουν σε μια αντίστοιχη μορφή, με διαφορετική μόνο τη σταθερά c_0 στο εκθετικό. Επομένως, θέτουμε και μελετάμε μια παραμετρική συνάρτηση με ανεξάρτητη μεταβλητή την x, παράμετρο $c \in \mathbb{R}$ και τη μορφή:

$$f(x) = \frac{e^{bx+c}}{1+\ e^{bx+c}}\ , \text{όπου } x \in [0,1]$$

Η f είναι συνεχής και παραγωγίσιμη στο διάστημα $[0,1]$ άρα έχουμε:

$$f'(x) = \frac{be^{bx+c}\left(1+\ e^{bx+c}\right) - be^{bx+c}e^{bx+c}}{(1+\ e^{bx+c})^2} = \frac{be^{bx+c}}{(1+\ e^{bx+c})^2} < 0, \ \ \alpha\varphi\text{ού } b < 0$$

Άρα η f είναι γνησίως φθίνουσα συνάρτηση κι έτσι για το σύνολο τιμών της θα ισχύει:

$$D_f = f([0,1]) = [f(1), f(0)]$$

Έτσι έχουμε:

<u>Για $c = c_0$:</u>

$$D_f = [f(1), f(0)] = [1.9 \times 10^{-14}, 0.164]$$

Για $c = c_0'$:

$$D_f = [f(1), f(0)] = [1.2 \times 10^{-12}, 0.928$$

Άρα το σύνολο τιμών της $prob(x_1, x_2)$ είναι:

$$D_{prob} = [1.9 \times 10^{-14}, 0.928]$$

Άρα παρατηρούμε πως η $prob(x_1, x_2)$ παρουσιάζει ολικό μέγιστο το $prob(1,0) = 0.928$ και πως εάν $x_1 = 0$ τότε $prob(0, x_2) < 0.5,\ \forall\ x_2 \in [0,1]$.

Ακόμα για να χαρακτηριστεί κάποια παρατήρηση πως ανήκει στην κλάση 1 πρέπει

$$prob(x_1, x_2) > 0.5 \rightarrow prob(1, x_2) > 0.5 \rightarrow \frac{e^{bx_2 + c_0'}}{1 + e^{bx_2 + c_0'}} > 0.5 \rightarrow e^{bx_2 + c_0'} > \frac{1}{2} + \frac{1}{2} e^{bx_2 + c_0'} \rightarrow$$

$$\rightarrow e^{bx_2 + c_0'} > 1 \rightarrow bx_2 + c_0' > 0 \rightarrow (b < 0) \rightarrow x_2 < -\frac{c_0'}{b}$$

Άρα, για να χαρακτηρισθεί μία παρατήρηση στην κλάση 1 πρέπει $x_1 = 1$ καθώς και $x_2 < x_0 = 0.085106547133126$

ή καλύτερα, θα πρέπει να έχει *πυρήνα μαγκεμίτη και μέγεθος < 14.6nm.*

Βιβλιογραφία

1. Kotzabasaki MI, Sotiropoulos I, Sarimveis H. QSAR modeling of the toxicity classification of superparamagnetic iron oxide nanoparticles (SPIONs) in stem-cell monitoring applications: an integrated study from data curation to model development. *RSC Adv*. 2020;10(9):5385-5391. doi:10.1039/C9RA09475J

2. Brown AC, Fraser TR. V.—On the Connection between Chemical Constitution and Physiological Action. Part. I.—On the Physiological Action of the Salts of the Ammonium Bases, derived from Strychnia, Brucia, Thebaia, Codeia, Morphia, and Nicotia. *Trans R Soc Edinburgh*. Published online 1868. doi:10.1017/S0080456800028155

3. Meyer H. Zur Theorie der Alkoholnarkose. *Arch für Exp Pathol und Pharmakologie*. Published online 1901. doi:10.1007/bf01978064

4. Ernst Overton. Studien uber die Narcose, Zugleich ein Beitrag zur allgemeinen Pharmakologie. *Science (80-)*. Published online 1901. doi:10.1126/science.14.361.845

5. Hammett LP. The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. *J Am Chem Soc*. Published online 1937. doi:10.1021/ja01280a022

6. Taft RW. Polar and Steric Substituent Constants for Aliphatic and o-Benzoate Groups from Rates of Esterification and Hydrolysis of Esters. *J Am Chem Soc*. Published online 1952. doi:10.1021/ja01132a049

7. Hansch C, Fujita T. ϱ-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J Am Chem Soc*. Published online 1964. doi:10.1021/ja01062a035

8. Cherkasov A, Muratov EN, Fourches D, et al. QSAR modeling: Where have you been? Where are you going to? *J Med Chem*. Published online 2014. doi:10.1021/jm4004285

9. Tropsha A. Best practices for QSAR model development, validation, and exploitation. *Mol Inform*. Published online 2010. doi:10.1002/minf.201000061

10. ORCHESTRA. *Theory, Guidance and Applications on QSAR and REACH*. 1st ed. (Benfenati E, ed.).; 2012. http://home.deib.polimi.it/gini/papers/orchestra.pdf

11. Weininger D. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *J Chem Inf Comput Sci*. Published online 1988. doi:10.1021/ci00057a005

12. Mauri A, Consonni V, Pavan M, Todeschini R. DRAGON software: An easy approach to molecular descriptor calculations. *Match*. Published online 2006.

13. Todeschini R, Consonni V. *Handbook of Molecular Descriptors*.; 2000. doi:10.1002/9783527613106

14. Alexopoulos EC. Introduction to multivariate regression analysis. *Hippokratia*. Published online 2010. doi:10.2307/2287933

15. Hartley R. *Multivariable Calculus with Vectors*. Prentice Hall; 1998.

16. Miles J. R Squared, Adjusted R Squared. In: *Wiley StatsRef: Statistics Reference Online*. ; 2014. doi:10.1002/9781118445112.stat06627

17. Frost J. Multiple regression analysis: Use adjusted R-squared and predicted R-squared to

include the correct number of variables. *Minitab Blog*. Published online 2013.

18. Eriksson L, Johansson E, Kettaneh-Wold N, Trygg C, Wikström C, Wold S. *Multi- and Megavariate Data Analysis Part I. Basic Principles and Applications*.; 2006.

19. Pearson K. LIII. On lines and planes of closest fit to systems of points in space . *London, Edinburgh, Dublin Philos Mag J Sci*. Published online 1901. doi:10.1080/14786440109462720

20. Abdi H, Williams LJ. Principal component analysis. *Wiley Interdiscip Rev Comput Stat*. Published online 2010. doi:10.1002/wics.101

21. Smith LI. A tutorial on Principal Components Analysis Introduction. *Statistics (Ber)*. Published online 2002.

22. Wold S, Sjöström M, Eriksson L. PLS-regression: A basic tool of chemometrics. In: *Chemometrics and Intelligent Laboratory Systems*. ; 2001. doi:10.1016/S0169-7439(01)00155-1

23. Farrés M, Platikanov S, Tsakovski S, Tauler R. Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation. *J Chemom*. Published online 2015. doi:10.1002/cem.2736

24. Matteo C, Francesca G, Cassotti M, Grisoni F. Variable selection methods: an introduction. *Mol Descriptors*. Published online 2012.

25. Tropsha A, Gramatica P, Gombar VK. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. In: *QSAR and Combinatorial Science*. ; 2003. doi:10.1002/qsar.200390007

26. Sushko I, Novotarskyi S, Körner R, et al. Applicability domains for classification problems: Benchmarking of distance to models for ames mutagenicity set. *J Chem Inf Model*. Published online 2010. doi:10.1021/ci100253r

27. A History of Bayes' Theorem. Published 2011. https://www.lesswrong.com/posts/RTt59BtFLqQbsSiqd/a-history-of-bayes-theorem

28. Galton F. Regression Towards Mediocrity in Hereditary Stature. *J Anthropol Inst Gt Britain Irel*. Published online 1886. doi:10.2307/2841583

29. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys*. Published online 1943. doi:10.1007/BF02478259

30. Belson WA. Matching and Prediction on the Principle of Biological Classification. *Appl Stat*. Published online 1959. doi:10.2307/2985543

31. Silverman BW, Jones MC. E. Fix and J.L. Hodges (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951). *Int Stat Rev / Rev Int Stat*. Published online 1989. doi:10.2307/1403796

32. Kubat M. *An Introduction to Machine Learning*.; 2017. doi:10.1007/978-3-319-63913-0

33. Everitt B. *Cluster Analysis 5th Edition*.; 2011. doi:10.4135/9781412983648

34. Liu B, Hsu W, Ma Y, Ma B. Integrating Classification and Association Rule Mining. *Knowl Discov Data Min*. Published online 1998. doi:10.1.1.48.8380

35. Mitchell TM. The Discipline of Machine Learning. *Mach Learn*. Published online 2006. doi:10.1080/026404199365326

36. Bishop C. *Pattern Recognition and Machine Learning*. (Jordan M, Kleinberg J, Scholkopf B, eds.).; 2006.

37. Russell S, Norvig P. *Artificial Intelligence A Modern Approach Third Edition*.; 2010. doi:10.1017/S0269888900007724

38. Salton G, McGill MJ. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc.; 1986.

39. Kennard RW, Stone LA. Computer Aided Design of Experiments. *Technometrics*. Published online 1969. doi:10.1080/00401706.1969.10490666

40. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning*. (Casella G, Fienberg S, Olkin I, eds.). Springer Publishing Company, Incorporated; 2014. doi:10.1007/978-1-4614-7138-7

41. Mitchell TM. *Machine Learning*. McGraw-Hill Science/Engineering/Math; 1997.

42. Hastie T, Tibshirani R, Friedman J. *Elements of Statistical Learning 2nd Ed*. Springer Publishing Company, Incorporated; 2009. doi:10.1007/978-0-387-84858-7

43. Learning M, Zheng A. *Evaluating Machine Learning Models\nA Beginner's Guide to Key Concepts and Pitfalls*.; 2015.

44. Snoek J, Larochelle H, Adams RP. Practical Bayesian optimization of machine learning algorithms. In: *Advances in Neural Information Processing Systems*. ; 2012.

45. Frazier PI. Bayesian Optimization. In: *Recent Advances in Optimization and Modeling of Contemporary Problems*. ; 2018. doi:10.1287/educ.2018.0188

46. Olson RS, Bartley N, Urbanowicz RJ, Moore JH. Evaluation of a tree-based pipeline optimization tool for automating data science. In: *GECCO 2016 - Proceedings of the 2016 Genetic and Evolutionary Computation Conference*. ; 2016. doi:10.1145/2908812.2908918

47. Olson RS, Moore JH. TPOT: A Tree-Based Pipeline Optimization Tool for Automating Machine Learning. In: ; 2019. doi:10.1007/978-3-030-05318-5_8

48. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res*. Published online 2011.

49. Chomenidis C, Drakakis G, Tsiliki G, et al. Jaqpot Quattro: A Novel Computational Web Platform for Modeling and Analysis in Nanoinformatics. *J Chem Inf Model*. Published online 2017. doi:10.1021/acs.jcim.7b00223

50. Jaqpot: https://app.jaqpot.org/

51. Iijima S. Helical microtubules of graphitic carbon. *Nature*. Published online 1991. doi:10.1038/354056a0

52. Ebbesen TW, Ajayan PM. Large-scale synthesis of carbon nanotubes. *Nature*. Published online 1992. doi:10.1038/358220a0

53. Popov VN. Carbon nanotubes: Properties and application. *Mater Sci Eng R Reports*. Published online 2004. doi:10.1016/j.mser.2003.10.001

54. Bianco A, Kostarelos K, Partidos CD, Prato M. Biomedical applications of functionalised carbon nanotubes. *Chem Commun*. Published online 2005. doi:10.1039/b410943k

55. Kang S, Mauter MS, Elimelech M. Physicochemical determinants of multiwalled carbon nanotube bacterial cytotoxicity. *Environ Sci Technol*. Published online 2008. doi:10.1021/es8010173

56. Rodrigues DF, Elimelech M. Toxic effects of single-walled carbon nanotubes in the development of E. Coli biofilm. *Environ Sci Technol*. Published online 2010. doi:10.1021/es1005785

57. Brady-Estévez AS, Schnoor MH, Vecitis CD, Saleh NB, Elimelech M. Multiwalled carbon nanotube filter: Improving viral removal at low pressure. *Langmuir*. Published online 2010. doi:10.1021/la102783v

58. Brady-Estévez AS, Schnoor MH, Kang S, Elimelech M. SWNT-MWNT hybrid filter attains high viral removal and bacterial inactivation. *Langmuir*. Published online 2010. doi:10.1021/la103776y

59. Vatanpour V, Madaeni SS, Moradian R, Zinadini S, Astinchap B. Fabrication and characterization of novel antifouling nanofiltration membrane prepared from oxidized multiwalled carbon nanotube/polyethersulfone nanocomposite. *J Memb Sci*. Published online 2011. doi:10.1016/j.memsci.2011.03.055

60. Simmons TJ, Lee SH, Park TJ, Hashim DP, Ajayan PM, Linhardt RJ. Antiseptic single wall carbon nanotube bandages. *Carbon N Y*. Published online 2009. doi:10.1016/j.carbon.2009.02.005

61. Nepal D, Balasubramanian S, Simonian AL, Davis VA. Strong antimicrobial coatings: Single-walled carbon nanotubes armored with biopolymers. *Nano Lett*. Published online 2008. doi:10.1021/nl080522t

62. Bianco A, Kostarelos K, Prato M. Applications of carbon nanotubes in drug delivery. *Curr Opin Chem Biol*. Published online 2005. doi:10.1016/j.cbpa.2005.10.005

63. Kostarelos K, Bianco A, Prato M. Promises, facts and challenges for carbon nanotubes in imaging and therapeutics. *Nat Nanotechnol*. Published online 2009. doi:10.1038/nnano.2009.241

64. Wu W, Wieckowski S, Pastorin G, et al. Targeted delivery of amphotericin B to cells by using functionalized carbon nanotubes. *Angew Chemie - Int Ed*. Published online 2005. doi:10.1002/anie.200501613

65. Pastorin G, Wu W, Wieckowski S, et al. Double functionalisation of carbon nanotubes for multimodal drug delivery. *Chem Commun*. Published online 2006. doi:10.1039/b516309a

66. Liu Z, Cai W, He L, et al. In vivo biodistribution and highly efficient tumour targeting of carbon nanotubes in mice. *Nat Nanotechnol*. Published online 2007. doi:10.1038/nnano.2006.170

67. McDevitt MR, Chattopadhyay D, Kappel BJ, et al. Tumor targeting with antibody-functionalized, radiolabeled carbon nanotubes. *J Nucl Med*. Published online 2007. doi:10.2967/jnumed.106.039131

68. Bhirde AA, Patel V, Gavard J, et al. Targeted killing of cancer cells in vivo and in vitro with EGF-directed carbon nanotube-based drug delivery. *ACS Nano*. Published online 2009. doi:10.1021/nn800551s

69. Kang S, Herzberg M, Rodrigues DF, Elimelech M. Antibacterial effects of carbon nanotubes: Size does matter! *Langmuir*. Published online 2008. doi:10.1021/la800951v

70. Zardini HZ, Amiri A, Shanbedi M, Maghrebi M, Baniadam M. Enhanced antibacterial activity of amino acids-functionalized multi walled carbon nanotubes by a simple method. *Colloids Surfaces B Biointerfaces*. Published online 2012. doi:10.1016/j.colsurfb.2011.11.045

71. Zhu Y, Ran T, Li Y, Guo J, Li W. Dependence of the cytotoxicity of multi-walled carbon nanotubes on the culture medium. *Nanotechnology*. Published online 2006. doi:10.1088/0957-4484/17/18/024

72. Nel A, Xia T, Mädler L, Li N. Toxic potential of materials at the nanolevel. *Science (80-)*.

Published online 2006. doi:10.1126/science.1114397

73. Jia G, Wang H, Yan L, et al. Cytotoxicity of carbon nanomaterials: Single-wall nanotube, multi-wall nanotube, and fullerene. *Environ Sci Technol*. Published online 2005. doi:10.1021/es048729l

74. Shvedova A, Castranova V, Kisin E, et al. Exposure to carbon nanotube material: Assessment of nanotube cytotoxicity using human keratinocyte cells. *J Toxicol Environ Heal - Part A*. Published online 2003. doi:10.1080/713853956

75. Chen X, Tam UC, Czlapinski JL, et al. Interfacing carbon nanotubes with living cells. *J Am Chem Soc*. Published online 2006. doi:10.1021/ja060276s

76. Bietenbeck M, Florian A, Faber C, Sechtem U, Yilmaz A. Remote magnetic targeting of iron oxide nanoparticles for cardiovascular diagnosis and therapeutic drug delivery: Where are we now? *Int J Nanomedicine*. Published online 2016. doi:10.2147/IJN.S110542

77. Yamazaki M, Ito T. Deformation and Instability in Membrane Structure of Phospholipid Vesicles Caused by Osmophobic Association: Mechanical Stress Model for the Mechanism of Poly(ethylene glycol)-Induced Membrane Fusion. *Biochemistry*. Published online 1990. doi:10.1021/bi00457a029

78. Van Driessche AE., KellermeierLiane M, Benning LG, Gebauer D. *New Perspectives on Mineral Nucleation and Growth*. Springer Publishing Company, Incorporated; 2017. doi:10.1007/978-3-319-45669-0

79. Lowry MB, Duchemin AM, Robinson JM, Anderson CL. Functional separation of pseudopod extension and particle internalization during Fcγ receptor-mediated phagocytosis. *J Exp Med*. Published online 1998. doi:10.1084/jem.187.2.161

80. Mahmoudi M, Sant S, Wang B, Laurent S, Sen T. Superparamagnetic iron oxide nanoparticles (SPIONs): Development, surface modification and applications in chemotherapy. *Adv Drug Deliv Rev*. Published online 2011. doi:10.1016/j.addr.2010.05.006

81. Shkilnyy A, Munnier E, Hervé K, et al. Synthesis and evaluation of novel biocompatible super-paramagnetic iron oxide nanoparticles as magnetic anticancer drug carrier and fluorescence active label. *J Phys Chem C*. Published online 2010. doi:10.1021/jp9112188

82. Nasongkla N, Bey E, Ren J, et al. Multifunctional polymeric micelles as cancer-targeted, MRI-ultrasensitive drug delivery systems. *Nano Lett*. Published online 2006. doi:10.1021/nl061412u

83. Neuberger T, Schöpf B, Hofmann H, Hofmann M, Von Rechenberg B. Superparamagnetic nanoparticles for biomedical applications: Possibilities and limitations of a new drug delivery system. In: *Journal of Magnetism and Magnetic Materials*. ; 2005. doi:10.1016/j.jmmm.2005.01.064

84. Xu H, Aguilar ZP, Yang L, et al. Antibody conjugated magnetic iron oxide nanoparticles for cancer cell separation in fresh whole blood. *Biomaterials*. Published online 2011. doi:10.1016/j.biomaterials.2011.08.076

85. Jeníkovâ G, Pazlarovâ J, Demnerovâ K. Detection of Salmonella in food samples by the combination of immunomagnetic separation and PCR assay. *Int Microbiol*. Published online 2000. doi:10.2436/im.v3i4.9284

86. Python. https://www.python.org/

87. Jupyter. https://www.jupyter.org/

88. Kato T, Totsuka Y, Ishino K, et al. Genotoxicity of multi-walled carbon nanotubes in both in vitro and in vivo assay systems. *Nanotoxicology*. Published online 2013.

doi:10.3109/17435390.2012.674571

89. Jackson P, Kling K, Jensen KA, et al. Characterization of genotoxic response to 15 multiwalled carbon nanotubes with variable physicochemical properties including surface functionalizations in the FE1-Muta(TM) mouse lung epithelial cell line. *Environ Mol Mutagen*. Published online 2015. doi:10.1002/em.21922

90. Poulsen SS, Jackson P, Kling K, et al. Multi-walled carbon nanotube physicochemical properties predict pulmonary inflammation and genotoxicity. *Nanotoxicology*. Published online 2016. doi:10.1080/17435390.2016.1202351

91. Aschberger K, Asturiol D, Lamon L, Richarz A, Gerloff K, Worth A. Grouping of multi-walled carbon nanotubes to read-across genotoxicity: A case study to evaluate the applicability of regulatory guidance. *Comput Toxicol*. Published online 2019. doi:10.1016/j.comtox.2018.10.001

92. Li L, Jiang W, Luo K, et al. Superparamagnetic iron oxide nanoparticles as MRI contrast agents for non-invasive stem cell labeling and tracking. *Theranostics*. Published online 2013. doi:10.7150/thno.5366

93. Olson RS, Urbanowicz RJ, Andrews PC, Lavender NA, Kidd LC, Moore JH. Automating biomedical data science through tree-based pipeline optimization. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. ; 2016. doi:10.1007/978-3-319-31204-0_9

94. Scopus Database. https://www.scopus.com/

95. Ju S, Teng G, Zhang Y, Ma M, Chen F, Ni Y. In vitro labeling and MRI of mesenchymal stem cells from human umbilical cord blood. *Magn Reson Imaging*. Published online 2006. doi:10.1016/j.mri.2005.12.017

96. Babič M, Horák D, Trchová M, et al. Poly(L-lysine)-modified iron oxide nanoparticles for stem cell labeling. *Bioconjug Chem*. Published online 2008. doi:10.1021/bc700410z

97. Horák D, Babič M, Jendelová P, et al. Effect of different magnetic nanoparticle coatings on the efficiency of stem cell labeling. *J Magn Magn Mater*. Published online 2009. doi:10.1016/j.jmmm.2009.02.082

98. Liu G, Wang Z, Lu J, et al. Low molecular weight alkyl-polycation wrapped magnetite nanoparticle clusters as MRI probes for stem cell labeling and in vivo imaging. *Biomaterials*. Published online 2011. doi:10.1016/j.biomaterials.2010.08.099

99. Guo RM, Cao N, Zhang F, et al. Controllable labelling of stem cells with a novel superparamagnetic iron oxide-loaded cationic nanovesicle for MR imaging. *Eur Radiol*. Published online 2012. doi:10.1007/s00330-012-2509-z

100. Huang C, Neoh KG, Kang ET, Shuter B. Surface modified superparamagnetic iron oxide nanoparticles (SPIONs) for high efficiency folate-receptor targeting with low uptake by macrophages. *J Mater Chem*. Published online 2011. doi:10.1039/c1jm11270h

101. Jo J ichiro, Aoki I, Tabata Y. Design of iron oxide nanoparticles with different sizes and surface charges for simple and efficient labeling of mesenchymal stem cells. *J Control Release*. Published online 2010. doi:10.1016/j.jconrel.2009.11.014

102. Chung HJ, Lee H, Bae KH, et al. Facile synthetic route for surface-functionalized magnetic nanoparticles: Cell labeling and magnetic resonance imaging studies. *ACS Nano*. Published online 2011. doi:10.1021/nn201198f

103. Horák D, Babič M, Jendelová P, et al. D-mannose-modified iron oxide nanoparticles for stem cell labeling. *Bioconjug Chem*. Published online 2007. doi:10.1021/bc060186c

104. Andreas K, Georgieva R, Ladwig M, et al. Highly efficient magnetic stem cell labeling with citrate-coated superparamagnetic iron oxide nanoparticles for MRI tracking. *Biomaterials*. Published online 2012. doi:10.1016/j.biomaterials.2012.02.064

105. Song M, Woo KM, Kim Y, Lim D, Song IC, Yoon BW. Labeling efficacy of superparamagnetic iron oxide nanoparticles to human neural stem cells: Comparison of ferumoxides, monocrystalline iron oxide, cross-linked iron oxide (CLIO)-NH2 and tat-CLIO. *Korean J Radiol*. Published online 2007. doi:10.3348/kjr.2007.8.5.365