Πρόβλεψη Βιοδραστηριότητας Μικρών Χημικών Μορίων με Χρήση Μηχανικής Μάθησης και Στατιστικών Προσεγγίσεων

Εθνικό Μετσόβιο Πολυτεχνείο

Ορέστης Οικονόμου

Επιβλέπουσα : Καθηγήτρια Καρώνη Χρυσής

ΑΘΗΝΑ 2020

Πρόβλεψη Βιοδραστηριότητας Μικρών Χημικών Μορίων με Χρήση Μηχανικής Μάθησης και Στατιστικών Προσεγγίσεων

Εθνικό Μετσόβιο Πολυτεχνείο

Ορέστης Οικονόμου

Τριμελής Επιτροπή Εξέτασης

Καρώνη Χρυσηίς	Κουκουβίνος Χρήστος	Στεφανέας Πέτρος
(Καθηγήτρια Ε.Μ.Π)	(Καθηγητής Ε.Μ.Π)	(Επίκουρος Καθηγητής Ε.Μ.Π)
.....	.....	.....

ΑΘΗΝΑ 2020

Ευχαριστίες

Θα ήθελα να ευχαριστήσω την κυρία Καρώνη που σε όλη τη διάρκεια εκπόνησης της διπλωματικής μου η επικοινωνία μας ήταν ιδιαίτερα ανθρώπινη και άμεση και οι συμβουλές της ιδιαίτερα στοχευμένες και κατατοπιστικές.
Εννοείται πως σε αυτό το ταξίδι πάντα ήταν δίπλα μου οι γονείς μου στους οποίους είμαι ευγνώμων για την αμέριστη στήριξή τους. Ιωάννα και Ευάγγελε σας ευχαριστώ για όλα.
Τέλος σε όλους μου τους φίλους που με συνοδεύουν αυτά τα χρόνια των σπουδών αλλά και πριν από αυτά να ξέρετε ότι σας είμαι ευγνώμων, οι χαρούμενες στιγμές κάναν αυτό το ταξίδι κατά πολύ πιο ενδιαφέρον κι ευχάριστο.

Περίληψη

Αδιαμφισβήτητα, διανύουμε την εποχή της πληροφορίας του τεράστιου όγκου δεδομένων τα οποία σε συνδυασμό με την εξέλιξη του πεδίου της μηχανικής μάθησης, έχουν επηρεάσει πολλούς επιστημονικούς και βιομηχανικούς τομείς. Ένας τέτοιος τομέας, είναι και η χημειοπληροφορική που χρησιμοποιείται στην ανακάλυψη και στο σχεδιασμό νέων φαρμάκων. Η παρούσα μελέτη, είναι άρρηκτα συνδεδεμένη με τους προαναφερθέντες τομείς, αφού περιγράφει τη δημιουργία μοντέλων ταξινόμησης μέσω μεθόδων εποπτευόμενης μάθησης, χρησιμοποιώντας ως δεδομένα χημικά μόρια που έχει ελεγχθεί η δυνατότητά τους να αναστείλουν της πρόσδεσης ισταμίνης στον υποδοχέα H1. Συγκεκριμένα, για τη δημιουργία μοντέλων από τις αρχικές παρατηρήσεις δημιουργήθηκαν δύο σύνολα δεδομένων, ένα που περιγράφει τη βιοδραστηριότητα των χημικών μορίων με δύο κατηγορίες και ένα που την περιγράφει με τρεις. Έτσι, για την πρώτη περίπτωση χρησιμοποιήθηκαν οι μέθοδοι της πολλαπλής λογιστικής παλινδρόμησης, της λογιστικής παλινδρόμησης με μέθοδο συρρίκνωσης Lasso, της γραμμικής διακριτικής ανάλυσης και του δένδρου απόφασης. Για τη δεύτερη περίπτωση χρησιμοποιήθηκαν οι μέθοδοι της πολυωνυμικής λογιστικής παλινδρόμησης, των μηχανών διανυσμάτων υποστήριξης, του δένδρου απόφασης και του τυχαίου δάσους. Έπειτα, συγκρίθηκε η προβλεπτική ικανότητα των ταξινομητών, μέσω πληθώρας αριθμητικών μετρητών. Στην περίπτωση που η βιοδραστηριότητα των χημικών μορίων περιγράφεται από δύο κατηγορίες, επικρατέστερο ήταν το μοντέλο λογιστικής παλινδρόμησης με Lasso, με τις μεταβλητές που το περιγράφουν να αποτελούν τους κυριότερους φυσικοχημικούς δείκτες για την πρόβλεψη της ικανότητας ενός χημικού μορίου να αναστείλει την ισταμίνη. Ενώ, στην περίπτωση, που τα χημικά μόρια ταξινομούνται σε τρεις κατηγορίες, σύμφωνα με τη βιοδραστηριότητά τους, επικρατέστερο ήταν το μοντέλο που παρήχθη από τη μέθοδο τυχαίου δάσους. Τα μοντέλα αυτά, μελλοντικά, θα μπορούσαν, να χρησιμοποιηθούν για τον έλεγχο νέων χημικών μορίων, να αναστείλουν την πρόσδεση ισταμίνης στον υποδοχέα H1, οδηγώντας έτσι γρηγορότερα σε εργαστηριακή τους μελέτη με σκοπό την παραγωγή νέων αντιισταμινικών φαρμάκων.

Λέξεις Κλειδιά: μηχανική μάθηση, χημειοπληροφορική, σχεδιασμός φαρμάκων, εποπτευόμενη μάθηση, λογιστική παλινδρόμηση, τυχαία δάση, ισταμίνη, αντιισταμινικά, υποδοχέας H1

Abstract

Undoubtedly, this is the era of enormous data, which, combined with the evolution of machine learning, have influenced many scientific and industrial fields. One such area is computational chemistry, which is a major factor in discovery and design of new drugs. The present study is inextricably linked to the aforementioned areas, as it describes the creation of classification models through supervised learning methods, using as data chemical molecules that have been tested for their ability to inhibit histamine binding to the H1 receptor. Specifically, to create models from the initial observations, two sets of data were created, one describing the bioactivity of chemical molecules using two categories and one describing it with three. Thus, for the first case, the following methods were used: multiple logistic regression, logistic regression with Lasso shrinkage, linear discriminant analysis and decision trees. In the second case, the following methods were used: polynomial logistic regression, support vector machines, decision trees and random forests. The predictive ability of the classifiers was compared using various numerical metrics. In the case where the bioactivity of the chemical molecules was described by two categories, the Lasso logistic regression model was the most successful, with the variables describing it being the main physicochemical indicators for predicting the ability of a chemical molecule to inhibit histamine. On the other hand, in the case that the chemical molecules were classified into three categories according to their bioactivity, the model produced by the random forest method was the best. In the future, these models could be used for testing new chemical molecules to inhibit histamine binding to the H1 receptor, leading to a faster laboratory study to produce new antihistamines.

Keywords: machine learning, chemoinformatics, drug design, supervised learning, logistic regression, random forest, histamine, antihistamines, receptor H1

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

1.1 Μηχανική Μάθηση

Η τεχνητή νοημοσύνη ως ακαδημαϊκό πεδίο γεννήθηκε τη δεκαετία του 1950. Ως τεχνητή νοημοσύνη ορίστηκε η θεωρία και η ανάπτυξη υπολογιστικών συστημάτων που μπορούν να εκτελούν εργασίες για τις οποίες κατά κανόνα απαιτείται ανθρώπινη νοημοσύνη. Η τεχνητή νοημοσύνη ως επιστημονικός κλάδος θα μπορούσε να χωριστεί στα πεδία της Μηχανικής Μάθησης (Machine Learning), των Τεχνητών Νευρωνικών Δικτύων (Artificial Neural Networks), της Βαθιάς Μάθησης (Deep Learning), της Μηχανικής Όρασης (Computer Vision) και της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing). Η παρούσα μελέτη επικεντρώνεται στο πεδίο της μηχανικής μάθησης και σε διάφορες μεθόδους που αποτελούν μέρος της.

1.1.1 Τι είναι Μηχανική Μάθηση (Machine Learning)

Η Μηχανική Μάθηση, αποτελεί υποπεδίο της τεχνητής νοημοσύνης και σχετίζεται με την εξαγωγή γνώσης από δεδομένα. Σύμφωνα με τον Arthur Lee Samuel (1959) ορίστηκε ώς " *το πεδίο μελέτης που δίνει στους υπολογιστές την ικανότητα να μαθαίνουν, χωρίς να έχουν ρητά προγραμματιστεί* ". Είναι ένα πεδίο έρευνας που προκύπτει από την τομή των επιστημονικών πεδίων της στατιστικής, της τεχνητής νοημοσύνης και της επιστήμης των υπολογιστών. Αναπτύχθηκε από τη μελέτη της αναγνώρισης προτύπων και της υπολογιστικής θεωρίας στην τεχνητή νοημοσύνη.

Κύριος στόχος της μηχανικής μάθησης, είναι η μελέτη και η κατασκευή αλγορίθμων, που μπορούν να μάθουν από πληθώρα δεδομένων και να κάνουν προβλέψεις πάνω σε αυτά, χωρίς να ακολουθούν στατικά γραμμένες εντολές κώδικα. Καθίσταται λοιπόν σημαντικό να προσδιορισθεί τι είναι αλγόριθμος και τι το ξεχωριστό έχουν οι αλγόριθμοι μηχανικής μάθησης.

Οι αλγόριθμοι είναι μια ακολουθία πληθώρας οδηγιών, που ως στόχο έχουν την επίτευξη ενός συγκεκριμένου αποτελέσματος. Οι τεχνολογίες και ο προηγμένος ψηφιακός κόσμος του σήμερα, δομούνται κυρίως από αλγορίθμους που κατασκευάζονται από έναν φυσικό παράγοντα (προγραμμα-

τιστής) και δίνουν σε έναν υπολογιστή, ακριβείς οδηγίες, για να ακολουθήσει με στόχο την επίλυση ενός απλού ή μέχρι και ενός ιδιαίτερα σύνθετου προβλήματος. Όμως, η βασική διαδικασία της μηχανικής μάθησης είναι να δώσει δεδομένα κατάρτισης/εκπαίδευσης σε έναν αλγόριθμο μάθησης και αυτός με τη σειρά του να παράγει ένα νέο σύνολο κανόνων, βασισμένο σε συμπεράσματα από τα δεδομένα. Με αυτόν τον τρόπο, έχει καταφέρει να κάνει τους αλγορίθμους μηχανικής μάθησης να είναι αυτοί ουσιαστικά που δημιουργούν τους κανόνες και όχι κάποιος προγραμματιστής. Η δεδομενοκεντρική αυτή προσέγγιση δίνει τη δυνατότητα στους υπολογιστές να μπορούν να χρησιμοποιηθούν για νέες, πολύπλοκες εργασίες που δεν είναι εφικτό να προγραμματιστούν χειροκίνητα, καθιστώντας την κατασκευή νέων διαφορετικών μοντέλων μία σαφώς πιο εύκολη διαδικασία.

Η εισαγωγή νέων οδηγιών από τα δεδομένα, είναι η βασική δύναμη της μηχανικής μάθησης, ενώ υπογραμμίζει επίσης τον κρίσιμο ρόλο των δεδομένων, αφού όσο περισσότερα δεδομένα διατίθενται για την κατάρτιση του αλγορίθμου, τόσο περισσότερο αυτός μαθαίνει. Στην πραγματικότητα, πολλές πρόσφατες εξελίξεις στην τεχνητή νοημοσύνη δεν οφείλονται σε ριζικές καινοτομίες στους αλγορίθμους μάθησης, αλλά στην τεράστια ποσότητα δεδομένων που είναι προσβάσιμα στο διαδίκτυο.

1.1.2 Είδη Μηχανικής Μάθησης

Η μηχανική μάθηση χωρίζεται συνήθως σε δύο κύριες κατηγορίες. Η πρώτη κατηγορία είναι η **εποπτευόμενη ή προβλεπτική μάθηση (supervised learning)**[1], στόχος της οποίας είναι ένας αλγόριθμος να μάθει να απεικονίζει τα δεδομένα εισόδου (inputs) x στις εξόδους/προβλέψεις (outputs) y. Η κατηγορία αυτή από τη σκοπιά της στατιστικής μηχανικής μάθησης, περιλαμβάνει τη δημιουργία ενός στατιστικού μοντέλου για την πρόβλεψη ή την εκτίμηση μιας εξόδου, που βασίζεται σε μία ή περισσότερες εισόδους.

Ένα παράδειγμα προβλήματος εποπτευόμενης μηχανικής μάθησης θα μπορούσε να είναι ο προσδιορισμός ενός όγκου ως καλοήθη ή κακοήθη με βάση μία ιατρική εικόνα. Η είσοδος εδώ για τον αλγόριθμο μάθησης, είναι η δοσμένη εικόνα και η πρόβλεψη/αποτέλεσμα είναι το αν ο όγκος είναι καλοήθης ή κακοήθης.

Η δεύτερη κατηγορία μηχανικής μάθησης είναι η **μη εποπτευόμενη ή περιγραφική μάθηση (unsupervised learning)**[1]. Στα προβλήματα μη εποπτευόμενης μάθησης δίνονται μόνο είσοδοι δεδομένων (inputs), χωρίς καθορισμένες εξόδους εποπτείας. Μπορούμε όμως, να μάθουμε σχέσεις από αυτά τα δεδομένα, άλλωστε ο στόχος των συστημάτων επίλυσης προβλημάτων μη εποπτευόμενης μάθησης είναι να βρεθούν "ενδιαφέροντα μοτίβα" στα δεδομένα.

Ένα παράδειγμα προβλήματος μη εποπτευόμενης μηχανικής μάθησης θα μπορούσε να είναι η τμηματοποίηση πελατών σε ομάδες με παρόμοιες προτιμήσεις. Συγκεκριμένα, για ένα σύνολο δεδομένων με αρχεία πελατών, ίσως να ήταν επιθυμητός ο προσδιορισμός των πελατών που είναι όμοιοι σύμφωνα με τις αγοραστικές τους επιλογές. Φέρ'ειπείν για ένα διαδικτυακό κατάστημα αγορών, αυτές οι ομάδες μπορεί να είναι "γονείς", "παιδιά" ή "παίκτες ηλεκτρονικών παιχνιδιών".

[1]Οι κατηγορίες εποπτευόμενης και μη εποπτευόμενης μάθησης καθώς και των αλγορίθμων τους που χρησιμοποιήθηκαν στην παρούσα έρευνα παρουσιάζονται εκτενέστερα στο Κεφάλαιο 2.

Όμως, λόγω του ότι δεν είναι εκ των προτέρων γνωστές ποιες είναι ακριβώς αυτές οι ομάδες που θα προκύψουν από τα δεδομένα που επεξεργάστηκε ο εκάστοτε αλγόριθμος μη εποπτευόμενης μάθησης, ή ακόμα και ποιο είναι το πλήθος αυτών των ομάδων, δεν υπάρχουν προκαθορισμένες επιθυμητές προβλέψεις για τελικό αποτέλεσμα.

Είναι σημαντικό όμως να σημειωθεί ότι εκτός από τα δύο κύρια είδη μηχανικής μάθησης υπάρχει και ένα τρίτο είδος που έχει αναπτυχθεί ιδιαίτερα τα τελευταία χρόνια. Η κατηγορία αυτή είναι γνωστή ως **μάθηση ενίσχυσης (reinforcement learning)**[2].

Ένα σύστημα μάθησης ενίσχυσης εξετάζει μακροπρόθεσμες πολιτικές μάθησης, λόγω του ότι προσπαθεί να μάθει μέσα από την αλληλεπίδρασή του με το περιβάλλον. Ως πολιτική μάθησης ορίζεται μία απεικόνιση από την κατάσταση που βρίσκεται το σύστημά μας ως προς την επιθυμητή ενέργεια που πρέπει να επιτύχει να πραγματοποιήσει. Μάλιστα, τα συστήματα που ανήκουν στη μάθηση ενίσχυσης συνήθως δεν αντιλαμβάνονται τα οφέλη μίας ενέργειας που διετέλεσαν άμεσα, αλλά τα αποτελέσματα αυτής επιστρέφουν από το περιβάλλον ως αποτέλεσμα ανταμοιβή στο σύστημα, που πλέον έχει εκπαιδευτεί μετά από την πραγματοποίηση πολλών παρόμοιων ενεργειών. Είναι έτσι προφανές, ότι στόχος και βέλτιστη "πολιτική" ενός τέτοιου αλγοριθμικού συστήματος είναι η μεγιστοποίηση του αθροίσματος των εισπραχθέντων ανταμειβών του. Τυπικά, ένα πρόβλημα μάθησης ενίσχυσης θα μπορούσε να περιγραφεί, ως μία Μαρκοβιανή Διαδικασία Απόφασης με ένα σύνολο καταστάσεων S, ένα σύνολο δυνατών ενεργειών A, πιθανότητες μετάβασης $(p'|s,a)$ και μία συνάρτηση ανταμοιβής $r(s,a)$, αφού πολλοί αλγόριθμοι ενισχυτικής μάθησης χρησιμοποιούν τεχνικές δυναμικού προγραμματισμού.

Ένα παράδειγμα προβλήματος μάθησης ενίσχυσης θα αποτελούσε το ηλεκτρονικό παιχνίδι Super Mario, όπου το σύστημα ενίσχυσης που εκπαιδεύεται και αμείβεται είναι ο αλγόριθμος μάθησης και το περιβάλλον είναι το παιχνίδι (πιθανότατα ένα συγκεκριμένο επίπεδο του παιχνιδιού). Ο αλγόριθμός έχει μια σειρά ενεργειών. Συγκεκριμένα, αυτές θα είναι τα κουμπιά που πατάει κάποιος χρήστης για να παίξει το παιχνίδι, ενώ η ενημερωμένη κατάσταση θα είναι κάθε καρέ του παιχνιδιού καθώς ο χρόνος περνά, ενώ το σήμα ανταμοιβής που δίνεται στο σύστημα μάθησης θα είναι η αλλαγή στο σκορ. Εφόσον συνδεθούν όλα αυτά τα στοιχεία μαζί, θα έχει δημιουργηθεί ένα σύστημα μάθησης ενισχύσεων για να παιχτεί το παιχνίδι Super Mario.

1.1.3 Η Μηχανική Μάθηση και η Θεωρία Στατιστικής Μάθησης (Statistical Learning Theory)

Αρκετές φορές η μηχανική μάθηση έχει αντιμετωπιστεί ως μία εξιδανικευμένη μορφή του πεδίου της στατιστικής ενώ δεν είναι λίγες και οι περιπτώσεις που συγχέεται με τον όρο της θεωρίας στατιστικής μάθησης. Η σύγχυση αυτή είναι λογική, μιας και η μηχανική μάθηση και η στατιστική σαν επιστημονικά πεδία είναι άρρηκτα συνδεδεμένα. Όμως, είναι σημαντικό να διευκρινιστεί ότι η θεωρία της στατιστικής μάθησης είναι επί της ουσίας, το ευρύτερο πλαίσιο για τη μελέτη της έννοιας του συμπεράσματος τόσο στην εποπτευόμενη όσο και στη μη εποπτευόμενη μηχανική μάθηση.

[2]Υπογραμμίζεται ότι δε θα γίνει περαιτέρω ανάλυση της μάθησης ενίσχυσης, λόγω του ότι η παρούσα μελέτη δεν ασχολείται με αυτό το είδος μηχανικής μάθησης.

Η στατιστική συμπερασματολογία λοιπόν καλύπτει ολόκληρο το φάσμα της μηχανικής μάθησης, από την απόκτηση γνώσεων, την πρόβλεψη ή τη λήψη αποφάσεων μέχρι και την κατασκευή μοντέλων από ένα σύνολο δεδομένων με ή χωρίς προσδιοριστικές ετικέτες. Ακόμη, η διαδικασία που υλοποιείται στον τομέα της στατιστικής μάθησης, αναφέρεται σε ένα στατιστικό πλαίσιο, με κάθε υπόθεση να δηλώνεται μαθηματικά ως μηδενική ή εναλλακτική υπόθεση.

Πρακτικός στόχος της προσέγγισης της στατιστικής μάθησης είναι να καταστήσει την μηχανική μάθηση πιο ακριβή (αξιόπιστα αναπαραγωγίσιμη) και να δημιουργήσει νέους ή βελτιωμένους αλγορίθμους μοντελοποίησης. Αυτό επιτυγχάνεται κυρίως, με την παροχή ενός τυπικού, στατιστικού ορισμού των αφηρημένων εννοιών, όπως η εκμάθηση, η γενίκευση, η υπερφόρτωση και η απόδοση και στη συνέχεια με την εξέταση αυτών των υποθέσεων ανά μία παράμετρο κάθε φορά.

Η γενική προσέγγιση/διαδικασία μάθησης στη στατιστική θεωρία μάθησης είναι η ίδια με οποιοδήποτε άλλον επιστημονικό τομέα:

1. Παρατήρηση ενός φαινομένου
2. Κατασκευή ενός μοντέλου αυτού του φαινομένου
3. Πραγματοποίηση προβλέψεων χρησιμοποιώντας αυτό το μοντέλο

Όμως, στη στατιστική μηχανική μάθηση, η όλη διαδικασία πρέπει να αυτοματοποιηθεί για ένα πρόγραμμα ηλεκτρονικού υπολογιστή, ώστε να καταστεί δυνατό για το δεύτερο να μπορέσει να μάθει από αυτή.

Επομένως, κάθε βήμα της επιστημονικής μεθόδου θεωρείται ότι διέπεται από ένα πιθανοτικό μοντέλο του φαινομένου (ή από τη διαδικασία δημιουργίας δεδομένων). Πιο απλά, αυτό σημαίνει, ότι αν όλες οι παρελθούσες και οι μελλοντικές παρατηρήσεις ληφθούν τυχαία και ανεξάρτητα με συνεχή στατιστική δοκιμασία υποθέσεων, τότε οι πληροφορίες, που σχετίζονται με το υποβόσκον φαινόμενο (ή κατανομή πιθανοτήτων), μπορούν να συναχθούν αξιόπιστα. Αυτό είναι ιδιαίτερα σημαντικό, αφού επιτρέπει σε μία μηχανή να κατασκευάζει αλγορίθμους μάθησης, όπως οι k-πλησιέστεροι γείτονες με κατάλληλο k, οι οποίοι είναι συνεπείς (αναπαραγώγιμοι). Έτσι, καθίσταται δυνατή και η γενική ιδέα της βαθιάς μάθησης, καθώς όσο περισσότερα δεδομένα επεξεργάζονται, τόσο οι προβλέψεις του αλγορίθμου πλησιάζουν τις βέλτιστες λύσεις.

1.1.4 Η μηχανική μάθηση στο σήμερα

Λόγω της ραγδαίας αύξησης δεδομένων, το λεγόμενο "big data", η μηχανική μάθηση αποκτά ολοένα και περισσότερο έδαφος στο σήμερα. Συγκεκριμένα, οι αλγόριθμοι μάθησης έχουν τη δυνατότητα να μοντελοποιήσουν ακόμα πολυπλοκότερα και πιο ακριβή συστήματα. Αυτό είναι εύκολα αντιληπτό, αφού πλέον ο τομέας της μηχανικής μάθησης έχει οδηγήσει σε εφαρμογές που βελτιώνουν τόσο βιομηχανικές διαδικασίες όσο και την καθημερινή ζωή των ανθρώπων. Χαρακτηριστικά, στην παρούσα φάση, το επιστημονικό αυτό πεδίο έχει χρησιμοποιηθεί για παράδειγμα στα εξής:

1. Αναγνώριση εικόνας. Υπάρχουν πολλές καταστάσεις, όπου μπορεί να κατηγοριοποιηθεί ένα αντικείμενο ως ψηφιακή εικόνα και να αντιμετωπισθεί ως πρόβλημα κατηγοριοποίησης. Για παράδειγμα, στην περίπτωση ασπρόμαυρης εικόνας, η ένταση κάθε εικονοστοιχείου αντιμετωπίζεται ως μία από τις μετρήσεις, ενώ για έγχρωμες εικόνες, κάθε εικονοστοιχείο παρέχει 3 μετρήσεις εντάσεων σε τρία διαφορετικά χρώματα - κόκκινο, πράσινο και μπλε. Σε πιο ειδικές εφαρμογές, η αναγνώριση εικόνας μέσω αλγορίθμων μάθησης, μπορεί να οδηγήσει σε εφαρμογές για ανίχνευση προσώπου σε μια εικόνα, καθώς και για αναγνώριση χαρακτήρων ώστε, να διακρίνει ένα μηχάνημα χειρόγραφα καθώς και έντυπα γράμματα.

2. Αναγνώριση ομιλίας, η μετάφραση δηλαδή ομιλίας σε κείμενο. Εδώ, μια εφαρμογή λογισμικού μπορεί, να αναγνωρίσει τις λέξεις που ομιλούνται σε ένα κλιπ ή ένα αρχείο ήχου και στη συνέχεια να μετατρέψει τον ήχο σε αρχείο κειμένου. Η αναγνώριση ομιλίας χρησιμοποιείται σε εφαρμογές, όπως διεπαφή φωνητικού χρήστη, φωνητικές αναζητήσεις και πολλά άλλα (Alexa, Siri). Οι διεπαφές φωνητικού χρήστη περιλαμβάνουν τη φωνητική κλήση, τη δρομολόγηση κλήσεων και τον έλεγχο της συσκευής. Μπορεί, επίσης, να χρησιμοποιηθεί για μια απλή καταχώρηση δεδομένων και για προετοιμασία δομημένων εγγραφών.

3. Ιατρική διάγνωση. Η μηχανική μάθηση μπορεί να χρησιμοποιηθεί με τεχνικές και εργαλεία που δίνουν τη δυνατότητα, να βοηθήσουν στη διάγνωση ασθενειών. Χρησιμοποιείται για την ανάλυση των κλινικών παραμέτρων και του συνδυασμού τους για την πρόγνωση, την πρόβλεψη της εξέλιξης της νόσου, την εξαγωγή ιατρικής γνώσης, την έρευνα των αποτελεσμάτων, καθώς και για τον προγραμματισμό της θεραπείας και την παρακολούθηση των ασθενών.

4. Χημειοπληροφορική και ανακάλυψη φαρμάκων. Η μηχανική μάθηση με τη μέθοδό των Τεχνητών Νευρωνικών Δικτύων, οδήγησε στη βαθιά μάθηση (deep learning), η χρήση της οποίας εξέλιξε τον τομέα ανακάλυψης φαρμάκων μιας και κοπιώδεις διεργασίες όπως ο σχεδιασμός νέων χημικών ενώσεων ή η μελέτη της βιοδραστηριότητας πληθώρας χημικών μορίων καθίστανται πλέον πιο εφικτές και άμεσες. Μάλιστα, τεχνολογικοί γίγαντες όπως η IBM και η Google, νεοσύστατες εταιρείες βιοτεχνολογίας καθώς και ακαδημαϊκά κέντρα δεν παρέχουν μόνο υπολογιστικές υπηρεσίες που βασίζονται σε cloud, αλλά λειτουργούν επίσης στον φαρμακευτικό χώρο και στον χώρο υγειονομικής περίθαλψης με τους εταίρους της βιομηχανίας.

Η παρούσα μελέτη αξίζει να σημειωθεί, ότι ασχολείται με ένα παράδειγμα εφαρμογής στατιστικών μεθόδων και μηχανικής μάθησης στην χημειοπληροφορική που εστιάζει στη βιοδραστηριότητα χημικών μορίων με σκοπό τον εντοπισμό της κατάλληλης μεθόδου, που προβλέπει μελλοντικά χημικά μόρια, που θα μπορούσαν να χρησιμοποιηθούν για κάποια συγκεκριμένη λειτουργία που ήδη υλοποιούν άλλες ουσίες. Ωστόσο, εκτενής περιγραφή της περίπτωσης παρατίθεται στο Κεφάλαιο 4.

1.2 Χημειοπληροφορική και ανακάλυψη φαρμάκων

1.2.1 Από την παραδοσιακή διαδικασία ανακάλυψης φαρμάκων στη χημειοπληροφορική

Υπάρχουν επτά βήματα στη διαδικασία ανεύρεσης φαρμάκων. Πιο συγκεκριμένα, η επιλογή της προς μελέτη ασθένειας, η υπόθεση στόχου, η ταυτοποίηση των πιθανών ωφέλιμων χημικών ενώσεων (διαλογή), η βελτιστοποίηση των χημικών ενώσεων, η προ-κλινική δοκιμή, η κλινική δοκιμή και φαρμακογονιδιακή βελτιστοποίηση. Παραδοσιακά, τα βήματα αυτά εκτελούνταν διαδοχικά (Augen, 2002), ενώ εάν ένα από αυτά αργούσε, τότε θα επιβράδυνε ολόκληρη τη διαδικασία. Αυτά τα αργά βήματα αποτελούν σημεία συμφόρησης για τη συνολική διαδικασία. Τέτοια εμπόδια θα μπορούσαν να χαρακτηριστούν ο χρόνος και το κόστος δημιουργίας ή πειραματικής μελέτης νέων χημικών ενώσεων. Έτσι, οι φαρμακοβιομηχανίες στην προσπάθεια για μείωση του κόστους μελέτης νέων χημικών ενώσεων προσπάθησαν να ανακαλύψουν νέες τεχνολογίες για τη σύνθεση και τη μελέτη αυτών.

Λόγω της αυξημένης ζήτησης για νέες χημικές ενώσεις, ανακαλύφθηκαν υπολογιστικές χημικές τεχνολογίες για την παραγωγή αυτών σε μικρότερη χρονική περίοδο. Ειδικότερα, η συνδυαστική χημεία σαν κλάδος εστιάζει συστηματικά και επαναλαμβανόμενα στην παραγωγή μεγάλης ποικιλίας ενώσεων από σύνολα διαφορετικών τύπων αντιδραστηρίων που αποκαλούνται "δομικά στοιχεία". Από το 2000, πολλές στρατηγικές συνδυαστικής χημείας διαλυματο-φάσης και στερεο-φάσης ήταν καλά ανεπτυγμένες (Hall, Manku, and Wang, 2001), για παράδειγμα οι τεχνικές παράλληλης σύνθεσης χρησιμοποιούνται σήμερα σε όλες τις μεγάλες φαρμακευτικές εταιρείες. Αν και με την αύξηση των δυνατοτήτων παραγωγής και δοκιμής ενώσεων, υπήρχε η ελπίδα ότι η διαδικασία ανακάλυψης φαρμάκων θα μπορούσε να επιταχυνθεί δραματικά, δυστυχώς αυτό δεν έγινε.

Αναζητώντας τους λόγους των απογοητευτικών αποτελεσμάτων, η επιστημονική κοινότητα οδηγήθηκε στο συμπέρασμα ότι η αύξηση της χημικής ποικιλομορφίας των βιβλιοθηκών των ενώσεων θα ενίσχυε τη διαδικασία ανεύρεσης φαρμάκων. Έτσι, ήρθε η ανάπτυξη του τομέα της χημειοπληροφορικής προκειμένου να βελτιστοποιηθεί η χημική ποικιλομορφία των βιβλιοθηκών που περιέχουν τις διάφορες χημικές ενώσεις. Εκτενέστερη περιγραφή του κλάδου της χημειοπληροφορικής και της συμβολής του στο σχεδιασμό νέων φαρμάκων παρουσιάζεται παρακάτω.

1.2.2 Χημειοπληροφορική

Οι τεχνολογίες που πέτυχε να δημιουργήσει η συνδυαστική χημεία οδήγησαν στην παραγωγή εκατομμύρια ενώσεων, πολλές από τις οποίες όμως δεν ήταν ικανές για να είναι υποψήφιες για μελλοντικά φάρμακα. Προκειμένου, να αποφευχθεί η πληθώρα άσκοπων ανακαλύψεων πιστεύεται, ότι ήταν προτιμότερο να κατασκευαστούν χημικά διαφορετικές βιβλιοθήκες ενώσεων. Για να δημιουργηθεί όμως, μια βιβλιοθήκη ενώσεων με μεγάλη χημική ποικιλομορφία χρειάστηκε να δημιουργηθεί και να εφαρμοστεί μια ποικιλία τεχνολογιών δομικής επεξεργασίας για την ανάλυση ποικιλομορφίας. Αυτές οι υπολογιστικές προσεγγίσεις λοιπόν αποτέλεσαν τις συνιστώσες της χημειοπληροφορικής. Επομένως,

η χημειοπληροφορική συνδυάζει τους επιστημονικούς τομείς της χημείας, της επιστήμης των υπολογιστών, της στατιστικής και της επιστήμης της πληροφορίας, ενώ μπορεί να θεωρηθεί ως εκείνο το κομμάτι της υπολογιστικής χημείας, του οποίου τα μοντέλα δεν βασίζονται στην αναπαραγωγή της φυσικής και της χημείας με την οποία λειτουργεί ο πραγματικός κόσμος στη μοριακή κλίμακα.

Η χημειοπληροφορική, απλούστερα, έχει ως στόχο, να παράγει χρήσιμα μοντέλα που μπορούν να προβλέψουν τις χημικές και βιολογικές ιδιότητες των ενώσεων, δεδομένης της δισδιάστατης χημικής δομής ενός μορίου. Μάλιστα, σαν επιστημονικός κλάδος ξεκίνησε να δημιουργεί τοπικά μοντέλα, τυπικά για ποσοτικές σχέσεις δομής-δραστικότητας (QSAR) ή για ποσοτικές σχέσεις δομής-ιδιότητας (QSPR). Πιο συγκεκριμένα, ο σχεδιασμός για τα πρώτα μοντέλα βασίστηκε σε γραμμική και αργότερα σε πολλαπλή γραμική παλινδρόμηση. Τα μοντέλα αυτά κατασκευάστηκαν χρησιμοποιώντας πολύ λίγα χαρακτηριστικά και ήταν έγκυρα μόνο για μια μικρή σειρά από ενώσεις που έμοιαζαν πολύ δομικά.

Είναι ενδιαφέρον, ότι η αναγνώριση προτύπων με τη χρήση αλγορίθμων μηχανικής μάθησης έχει σχέση με τη χημεία, που ξεκινάει περισσότερο από τέσσερις δεκαετίες πριν, με μεθόδους όπως η γραμμική μηχανική μάθηση, η οποία εφαρμόζεται σε προβλήματα σαν αυτό της ερμηνείας των φασματοσκοπικών δεδομένων, όπως αναφέρθηκε σε μια πρώιμη ανασκόπηση από τον Kowalski (1974). Σε αντίθεση με το μικρό αριθμό τομέων εφαρμογής των πρώιμων μελετών για την QSAR, πρόσφατες έρευνες έχουν επικεντρωθεί σε παγκόσμια μοντέλα, τα οποία αποτελούν μοντέλα που εκπαιδεύονται και επομένως ισχύουν για ένα ευρύ φάσμα οργανικών ενώσεων ή φαρμάκων.

Όπως αναφέρθηκε και στο λόγο άνθησης της μηχανικής μάθησης, έτσι και στον τομέα της χημειοπληροφορικής, η αλματώδης αύξηση διαθέσιμων δεδομένων για μόρια που καλύπτουν ένα πολύ ευρύτερο χημικό χώρο, η δυνατότητα για χρήση μιας μεγάλης και ποικίλης επιλογής περιγραφικών στοιχείων και η ανάπτυξη περίπλοκων μη γραμμικών αλγορίθμων μηχανικής μάθησης, έχουν αυξήσει τη χρήση τέτοιων παγκόσμιων μοντέλων τα τελευταία χρόνια.

1.2.3 Σύγχρονοι σχεδιασμοί φαρμάκων και η συμβολή της μηχανικής μάθησης

Είναι γεγονός, ότι το εκτιμώμενο κόστος για την ανακάλυψη και ανάπτυξη φαρμάκων, ολοένα και αυξάνεται με την πάροδο των χρόνων. Λιγότερο από μία εικοσαετία πρίν, η ανάπτυξη ενός φαρμάκου χρειαζόταν κατά μέσο όρο 12 χρόνια, είχε κόστος κάτω από ένα δισεκατομμύριο δολάρια και οι μεγαλύτερες προκλήσεις ήταν οι αποτυχίες που οφείλονταν στην αναποτελεσματικότητα ή στην επαγόμενη από την τοξικότητα φθορά (Kola and Landis, 2004). Στη σημερινή εποχή τα προβλήματα και οι πιέσεις είναι πιθανό να διαφέρουν σε κάθε φαρμακευτική εταιρεία, ανάλογα με την αγορά που στοχεύουν και τους διαθέσιμους πόρους, όμως τα κύρια σημεία συμφόρησης της διαδικασίας δημιουργίας νέων φαρμάκων είναι κοινά (Wagner et al., 2018). Το μεγαλύτερο όμως, σε κόστος και πιο χρονοβόρο στάδιο που σχετίζεται με την ανάπτυξη φαρμάκων είναι η διεξαγωγή κλινικών δοκιμών. Έτσι, αναπτύχθηκαν νέες διαδικασίες ανακάλυψης φαρμάκων που συνοψίζονται παρακάτω.

Μία από τις πιο σύγχρονες διαδικασίες σχεδιασμού φαρμάκων σήμερα είναι η **de novo** σχεδίαση χημικών μορίων με χρήση βαθιάς μάθησης. Πιο συγκεκριμένα, ο de novo σχεδιασμός φαρμάκων

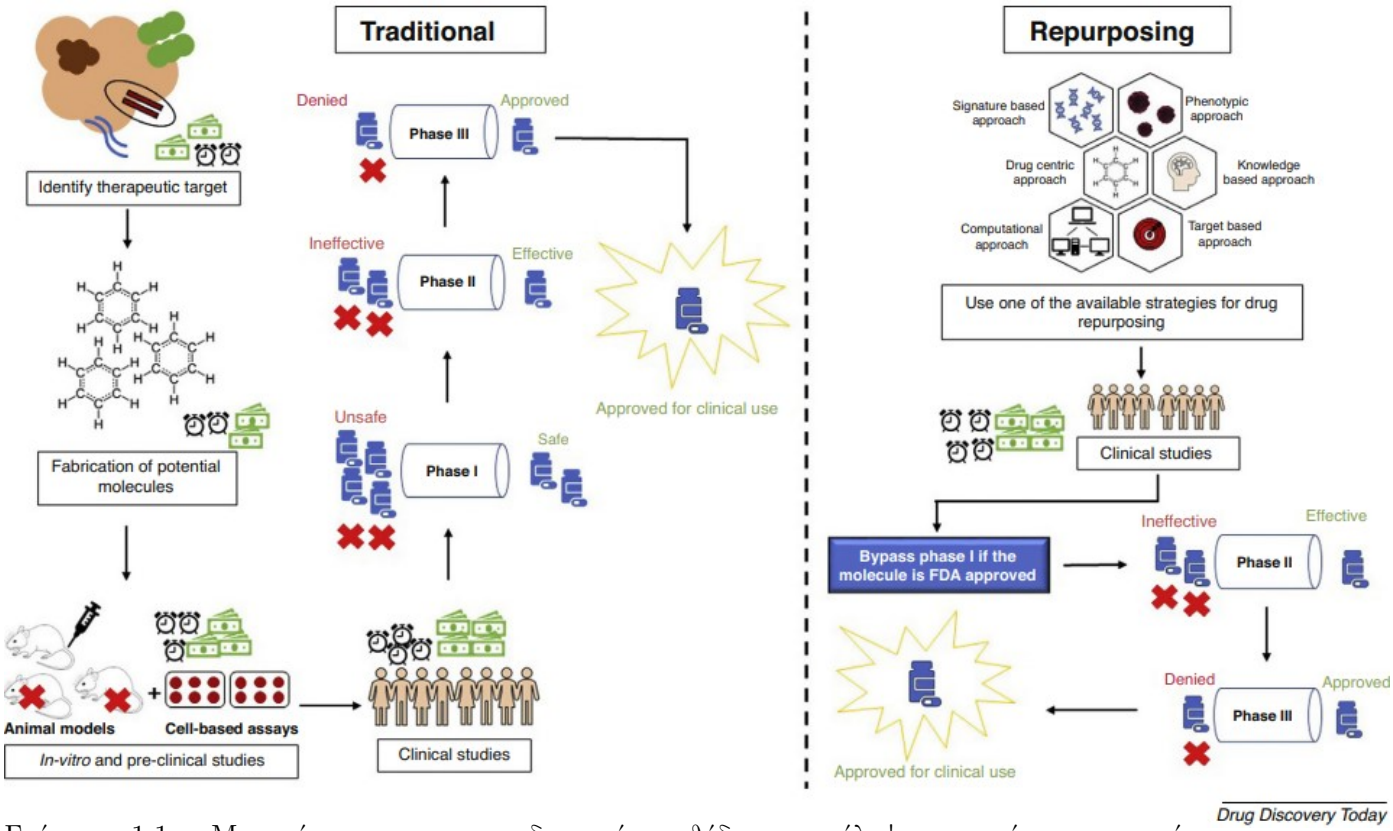
είναι μια επαναληπτική διαδικασία, στην οποία η τρισδιάστατη δομή ενός υποδοχέα (χημικές δομές, αποτελούμενες από πρωτεΐνες, που λαμβάνουν και μεταδίδουν σήματα που μπορούν να ενσωματωθούν σε βιολογικά συστήματα) χρησιμοποιείται για το σχεδιασμό νεότερων μορίων. Χαρακτηριστικά, περιλαμβάνει τον προσδιορισμό της δομής των συμπλεγμάτων του κύριου στόχου καταπολέμησης και το σχεδιασμό κύριων τροποποιήσεων χρησιμοποιώντας εργαλεία μοριακής μοντελοποίησης. Μπορεί επίσης, να χρησιμοποιηθεί για τον σχεδιασμό νέων κατηγοριών χημικών ενώσεων που παρουσιάζουν παρόμοιους υποκαταστάτες (ένα άτομο ή μια ομάδα ατόμων που λαμβάνει τη θέση ενός άλλου ατόμου ή ομάδας ή καταλαμβάνει μια συγκεκριμένη θέση σε ένα μόριο) στο στόχο καταπολέμησης χρησιμοποιώντας ένα πρότυπο ή μία "σκαλωσιά" (scaffold), το οποίο είναι χημικά διαφορετικό από παλιότερα ορισμένους οδηγούς.

Ο σχεδιασμός λοιπόν de novo θα μπορούσε να θεωρηθεί, ότι έχει τρία καθήκοντα: 1) γέννηση μορίων, 2) μοριακή βαθμολόγηση και 3) βελτιστοποίηση μορίων. Κάθε μία από αυτές τις πτυχές μπορεί, να πραγματοποιηθεί είτε ξεχωριστά είτε συλλογικά από ανθρώπους ή μηχανές. Με την εμφάνιση επομένως προηγμένων τεχνικών μηχανικής μάθησης έγινε εφικτή η αυτοματοποιημένη δημιουργία νέων χημικών οντοτήτων με τις επιθυμητές ιδιότητες (Schneider, 2018). Μάλιστα, σε πρωτοπόρες εφαρμογές, αλγόριθμοι μάθησης έχουν επιδείξει μοναδικές ικανότητες, όπως για παράδειγμα να συναρμολογήσουν μόρια από βασικά δομικά στοιχεία (άτομα, θραύσματα), ενώ ταυτόχρονα εξετάζουν παράλληλα πολλαπλές ιδιότητες και βιολογικές δραστηριότητες. Για παράδειγμα στην προσπάθειά τους για επίτευξη σχεδιασμού de novo των ενώσεων, οι Popova, Isayev και Tropsha ανέπτυξαν ένα τεχνητό νευρωνικό δίκτυο (artificial neural network) βαθιάς μάθησης ονομαζόμενο ReLeaSE (Popova, Isayev, and Tropsha, 2018) που στόχος του είναι να δημιουργήσει νέες δομές που έχουν φαρμακευτικές ιδιότητες. Χαρακτηριστικά, η μέθοδος αυτή βασίζεται στην ενσωμάτωση των προσεγγίσεων βαθιάς μάθησης και μάθησης ενίσχυσης, σε δύο νευρωνικά δίκτυα βαθιάς μάθησης, που εκπαιδεύτηκαν χωριστά. Το σημαντικό αυτής της μεθόδου είναι ότι όταν συνδυάζονται αυτά τα δύο τεχνητά νευρωνικά δίκτυα, μπορούν να χρησιμοποιηθούν, για να προτείνουν νέες χημικές ενώσεις, βασισμένα αποκλειστικά σε απλουστευμένες συμβολοσειρές εισόδου από μοριακά δεδομένα (Popova, Isayev, and Tropsha, 2018).

Ένας ακόμη τομέας που αναπτύσσεται όλο και περισσότερο σήμερα χάρη και στην εξέλιξη της μηχανικής μάθησης είναι η **επαναστόχευση φαρμάκων (drug repurposing)**. Η διαδικασία αυτή, λοιπόν, περιλαμβάνει χρήση φαρμάκων που έχουν εγκριθεί από ρυθμιστικούς οργανισμούς όπως ο Οργανισμός Τροφίμων και Φαρμάκων (FDA) των Η.Π.Α., ο Ευρωπαϊκός Οργανισμός Φαρμάκων (EMA) και η Ρυθμιστική Υπηρεσία Φαρμάκων και Προϊόντων Υγείας (MHRA) του Ηνωμένου Βασιλείου και ως στόχο έχει την παρατήρηση νέων ενδείξεων. Συγκεκριμένα, για να μπορέσει ένα νέο φάρμακο να εισέλθει στην αγορά, πρέπει να τηρήσει αυστηρά μία σειρά κανονισμών. Ο εντοπισμός ενός φαρμάκου και η περαιτέρω ανάπτυξη του απαιτεί σημαντική επένδυση, κυρίως ως αποτέλεσμα των ποικίλων φυσικοχημικών ιδιοτήτων των χημικών οντοτήτων και της πολυπλοκότητας της αύξησης της παραγωγής (Vaidya et al., 2019). Ο περιορισμός αυτός ωθεί όλο και περισσότερες φαρμακευτικές εταιρείες ή ακαδημαϊκά κέντρα να χρησιμοποιούν γρήγορα και αποτελεσματικά ήδη εγκεκριμένα φάρμακα για μια νέα ένδειξη (άλλου είδους ασθένεια από την προκαθορισμένη), για την οποία μέχρι πρότινος δεν υπήρχαν διαθέσιμες φαρμακευτικές αγωγές για την αντιμετώπισή της.

Ένα καλό ξεκίνημα για εφαρμογή αυτής της μεθόδου, λοιπόν, θα ήταν τα μόρια που σε ερευνητικό στάδιο απέτυχαν να δείξουν αποτελεσματικότητα για μία προκαθορισμένη ένδειξη, έτσι ώστε να αναγεννηθούν ως φάρμακα μέσω επαναστόχευσης. Μπορούν έτσι, τέτοια μόρια να σχεδιαστούν ξανά, για μια νέα ασθένεια αυτή τη φορά, καταλήγοντας να αποτελούν βιώσιμες θεραπείες, ιδιαίτερα χρήσιμες σε περιπτώσεις σπάνιων νόσων, οι οποίες παρουσιάζουν σημαντικές προκλήσεις στη διάγνωση, στη θεραπεία και την εύρεση των απαιτούμενων πόρων. Για παράδειγμα, ορισμένες αυτοάνοσες διαταραχές, βακτηριακές λοιμώξεις και οι σπάνιες μορφές καρκίνου δεν κληρονομούνται, καθιστώντας έτσι πιο δύσκολη τη θεραπεία επειδή είναι ιδιοπαθείς στη φύση τους (Lewis, Snyder, and Hyatt-Knorr, 2017). Η επαναστόχευση/αναπροσαρμογή όμως, ενός υπάρχοντος φαρμάκου, ούσα μια λιγότερο δαπανηρή και βραχύτερη προσέγγιση, μπορεί να φέρει πιο αποτελεσματικές θεραπείες σε ασθενείς με τα παραπάνω προβλήματα σε σύγκριση με τις δυσκίνητες παραδοσιακές διαδικασίες ανακάλυψης και ανάπτυξης. Επιπλέον, ο τρόπος αυτός ανακάλυψης και σχεδιασμού φαρμάκων βοηθάει να μειωθεί το κόστος εξόδων για την ανάπτυξη τους, μειώνοντας έτσι το κόστος για τους ασθενείς και τελικά και το πραγματικό συνολικό κόστος θεραπείας. Για να γίνουν αντιληπτά τα προαναφερθέντα, αρκεί να αναλογιστεί κανείς, ότι για ένα νέο ερευνητικό μόριο τα δεδομένα ασφάλειας και αποτελεσματικότητας του δεν είναι διαθέσιμα, όπως είναι για μόρια που συνιστούν ένα εγκεκριμένο φάρμακο. Έτσι υπάρχει η πιθανότητα μεγαλύτερης ζημίας για μία φαρμακευτική εταιρεία αφού κατά τη διάρκεια της διαδικασίας ανακάλυψης του φαρμάκου, μπορεί η φθορά να είναι μεγαλύτερη από ότι προβλεπόταν, οδηγώντας έτσι σε περισσότερες αποτυχίες, από ότι επιτυχίες εν δυνάμει φαρμάκων (Arrowsmith, 2011b) .

Εν αντιθέσει, όπως προαναφέρθηκε, με τα νέα ερευνητικά μόρια, όλα τα δεδομένα ασφάλειας, προκλινικής μελέτης και αποτελεσματικότητας είναι άμεσα διαθέσιμα για ένα επαναστοχευμένο μόριο, επιτρέποντας έτσι σε έναν ερευνητή να λάβει τεκμηριωμένη απόφαση σε κάθε στάδιο ανάπτυξης ενός φαρμάκου (Arrowsmith, 2011b). Η διαθεσιμότητα προηγούμενων γνώσεων σχετικά με την ασφάλεια, την αποτελεσματικότητα και την κατάλληλη διαδικασία χορήγησης μειώνουν σημαντικά το κόστος ανάπτυξης και τον χρόνο σχεδιασμού, με αποτέλεσμα να απαιτείται λιγότερη προσπάθεια για την επιτυχή διάθεση στην αγορά ενός επαναστοχευμένου φαρμάκου (Padhy and Gupta, 2011). Μάλιστα, η έλευση των νέων τεχνολογιών, όπως των εφαρμογών μηχανικής μάθησης, καθώς και άλλων υπολογιστικών εργαλείων, έκανε την ανακάλυψη φαρμάκων πιο προσιτή στην περίπτωση που κάποιος αρχίζει με ένα ήδη εγκεκριμένο φάρμακο (Cumming, 2016). Αυτή η νέα προσέγγιση της διαδικασίας σχεδιασμού είναι υπεύθυνη για το 30%, περίπου, όλων των νέων εγκεκριμένων φάρμακων από τον Οργανισμό Τροφίμων και Φαρμάκων (FDA) ("Drug repurposing: a promising tool to accelerate the drug discovery process" 2019). Σημαντική περιοχή ενδιαφέροντος για την επαναστόχευση ενός φαρμάκου θα μπορούσαν να είναι οι σπάνιες διαταραχές, καθώς αποτελούν ασθένειες που δεν έχουν ικανοποιητικά αντιμετωπιστεί λόγω της περιορισμένης διαθεσιμότητας των τυπικών θεραπειών, αλλά και εξαιτίας της επιδείνωσης των κλινικών αποτελεσμάτων (Gatta et al., 2011). Το Γράφημα 1.1 συνοψίζει βασικές διαφορές και οφέλη των προσεγγίσεων επαναστόχευσης φαρμάκων σε σύγκριση με τις παραδοσιακές προσεγγίσεις ανακάλυψης τους.



Γράφημα 1.1: Μια σύγκριση της παραδοσιακής μεθόδου ανακάλυψης φαρμάκων σε σχέση με την επαναστόχευση του φαρμάκου ("Drug repurposing: a promising tool to accelerate the drug discovery process" 2019)

Ένας γνωστός αλγόριθμος μηχανικής μάθησης για υπολογιστική επαναστόχευση φαρμάκων είναι η **"κανονικοποίηση βασικής γραμμής" (baseline regularization)** (Kuang et al., 2019). Η μέθοδος αυτή, στην ουσία αποτελεί μία κανονικοποιημένη μέθοδο ελαχίστων τετραγώνων με ποινή "σύντηξης lasso" (ελεγχόμενη από λ1) και lasso ποινή (ελεγχόμενη από λ2). Περισσότερες πληροφορίες[3] σχετικά με τη συγκεκριμένη μέθοδο δεν παρέχονται μιας και δεν αποτελούν αντικείμενο της παρούσας μελέτης. Ωστόσο, αξίζει να σημειωθεί ότι ο συγκεκριμένος αλγόριθμος χρησιμοποιήθηκε για την εύρεση φαρμάκων που με επαναστόχευση θα μπορούσαν να θέσουν υπό έλεγχο τα επίπεδα ζάχαρου, δείχνοντας μάλιστα ότι ουσίες που δεν είναι προκαθορισμένες για να ελέγχουν το ζάχαρο στο αίμα, όπως η βεραπαμίλη HCL, θα μπορούσαν να επαναστοχευτούν για αυτό το σκοπό.

1.3 Αντικείμενο Διπλωματικής

Όπως αναφέρθηκε και προηγουμένως παρατηρείται μια μεγάλη αύξηση των καταγραμμένων δεδομένων γύρω από χημικά μόρια και ενώσεις και ειδικά όσων αφορούν στην εξέλιξη του τομέα της ανακάλυψης νέων φαρμάκων. Στην παρούσα μελέτη μέσα από βάσεις δεδομένων πληθώρας χημικών μορίων, που μελετήθηκε η ικανότητά τους να αναστείλουν την πρόσδεση ισταμίνης με τον υποδοχέα H1, θα εξεταστούν ποια φυσικοχημικά χαρακτηριστικά είναι αυτά που αντιπροσωπεύουν χημικά μόρια που πραγματοποιούν την αναστολή αυτή.

[3]Το άρθρο των (Kuang et al. 2019) ωστόσο, παρέχει σημαντικό υλικό για περαιτέρω διερεύνηση

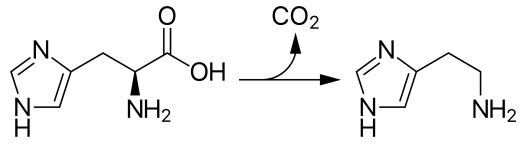
Για τη μελέτη αυτή κύριο σημείο αναφοράς, αποτελεί η βιοδραστηριότητα των διαφόρων χημικών μορίων, αφού θα αποτελέσει τη μεταβλητή απόκρισης για τα μοντέλα που θα δημιουργηθούν με χρήση αλγορίθμων μηχανικής μάθησης. Μάλιστα θα συγκριθούν τα αποτελέσματα των διαφορετικών αλγορίθμων μάθησης μέσω της ακρίβειας των παραγόμενων μοντέλων και άλλων περιγραφικών μετρητών και θα δοθεί μία εκτίμηση, ποιο μοντέλο είναι το καλύτερο. Το μοντέλο αυτό θα κρίνει ποια διαδικασία μηχανικής μάθησης οδηγεί στα πιο αξιόπιστα αποτελέσματα ταξινόμησης των χημικών μορίων, ενώ οι επεξηγηματικές μεταβλητές του τελικού αυτού μοντέλου (φυσικοχημικά χαρακτηριστικά χημικών μορίων) θα καθορίσουν ποιες είναι αυτές οι ιδιότητες που θα κρίνουν αν ένα νέο χημικό μόριο μπορεί να χρησιμοποιηθεί μελλοντικά για την καταπολέμηση της πρόσδεσης ισταμίνης στον υποδοχέα H1.

Το τελικό μοντέλο επί της ουσίας θα αποτελεί μία φόρμουλα εξέτασης χημικών μορίων και ενώσεων για το αν ενεργοποιούνται στην προσπάθειά τους για αναστολή πρόσδεσης ισταμίνης ή όχι. Εκτενέστερη παρουσίαση των αποτελεσμάτων της μελέτης και περιγραφή αυτών βρίσκεται στα Κεφάλαιο 4. Επιπλέον, είναι χρήσιμο να διευκρινιστεί, ότι με τη φράση "αναστολή της λειτουργίας του υποδοχέα ισταμίνης H1" εννοείται η διαδικασία που υλοποιούν τα γνωστά στο ευρύ κοινό αντιισταμινικά, καθιστώντας έτσι τα αποτελέσματα της παρούσας μελέτης ένα εργαλείο ανίχνευσης εν δυνάμει νέων αντιισταμινικών-H1 με σκοπό την μελλοντική ανακάλυψη νέων φαρμάκων. Αναλυτικότερη περιγραφή για την ισταμίνη και τα αντιισταμινικά γίνεται παρακάτω.

1.3.1 Ισταμίνη και Ανισταμινικά

Ισταμίνη

Η ισταμίνη είναι ένας μεσολαβητής πολλών φυσιολογικών και παθολογικών διεργασιών μέσα και έξω από το νευρικό σύστημα. Είναι ένα αγγελιοφόρος για το στομάχι, το δέρμα, το ανοσοποιητικό και το νευρικό σύστημα και ενεργεί τόσο ως ορμόνη όσο και ως νευροδιαβιβαστής, ανάλογα με τον ιστό που απελευθερώνεται. Ο οπίσθιος υποθάλαμος είναι η μόνη πηγή νευρώνων που περιέχουν ισταμίνη, οι οποίοι καλύπτουν ολόκληρο το κεντρικό νευρικό σύστημα και είναι ενεργοί αποκλειστικά κατά την αφύπνιση. Η ισταμίνη λοιπόν σχηματίζεται μέσα και απελευθερώνεται από τους νευρώνες αυτούς του κεντρικού νευρικού συστήματος, ενώ αποτελεί σημαντικό ρυθμιστή διαφόρων λειτουργιών του εγκεφάλου (Haas and Panula, 2003; Haas, Sergeeva, and Selbach, 2008). Μάλιστα, τρεις από τους τέσσερις γνωστούς μεταβοτροπικούς υποδοχείς ισταμίνης εκφράζονται ευρέως στον εγκέφαλο. Χαρακτηριστικά, οι H1 και H2 υποδοχείς είναι κυρίως διεγερτικοί ενώ ο H3 είναι ανασταλτικός αυτό- και ετερο-ϋποδοχέας. Αναφορικά, με τη σύνθεσή της η ισταμίνη προέρχεται από την αποκαρβοξυλίωση του αμινοξέος ιστιδίνη, διαδικασία που περιγράφεται στο Γράφημα 1.2 και αποτελεί μια αντίδραση που καταλύεται από το ένζυμο L-ιστιδίνη δεκαρβοξυλάση, ενώ είναι και μια υδρόφιλη αγγειοδραστική αμίνη.



Γράφημα 1.2: Μετατροπή ιστιδίνης σε ισταμίνη με αποκαρβοξυλίωση ιστιδίνης

Η ισταμίνη απελευθερώνεται επίσης, από τα μαστοκύτταρα και τα βασεόφιλα αντιγόνα, από ορισμένα πεπτίδια και μικρές βασικές ενώσεις. Συγκεκριμένα, τα μαστοκύτταρα είναι πολυάριθμα σε σημεία ενδεχόμενου τραυματισμού - της μύτης, του στόματος, των ποδιών, αλλά και των εσωτερικών επιφανειών του σώματος και των αιμοφόρων αγγείων. Σε αντίθεση με την ισταμίνη των μαστοκυττάρων, αυτή των μη ιστιοκυττάρων βρίσκεται σε αρκετούς ιστούς, συμπεριλαμβανομένου του εγκεφάλου. Ο σημαντικότερος παθοφυσιολογικός μηχανισμός απελευθέρωσης ισταμίνης από μαστοκύτταρα και βασεόφιλα είναι ανοσολογικού τύπου. Ωστόσο, υπάρχουν ορισμένες αμίνες και αλκαλοειδή, συμπεριλαμβανομένων φαρμάκων όπως η μορφίνη, που μπορούν να εκτοπίσουν την ισταμίνη σε κόκκους και να προκαλέσουν την απελευθέρωσή της, ενώ αντιβιοτικά όπως η πολυμυξίνη βρίσκονται επίσης, για να διεγείρουν την απελευθέρωση ισταμίνης.

Η ισταμίνη επιτελεί ιδιαίτερα σημαντικό ρόλο σε διάφορες λειτουργίες του σώματος. Ως φυσιολογικός μεσολαβητής, η ουσία αυτή είναι πιο γνωστή ως ενδογενές διεγερτικό της γαστρικής έκκρισης. Επιπλέον, συμμετέχει στη φλεγμονή και στη ρύθμιση της ανοσοαπόκρισης, ενώ καρδιακά αποθέματα ισταμίνης πιθανότατα δεν παίζουν φυσιολογικό ρόλο, αλλά μπορεί να έχουν παθολογική σημασία. Ακόμη, σε περιπτώση αναστολής του υποδοχέα ισταμίνης H2 με διάφορες ουσίες μπορεί να προκληθεί και πτώση της λίμπιντο. Όσον αφορά στον υποδοχέα H1, που θα αποτελέσει βασικό πυλώνα των δεδομένων της συγκεκριμένης πειραματικής μελέτης, αποτελεί μια κατηγορία υποδοχέων συζευγμένων με μια ενδοκυτταρική πρωτεΐνη G (Gq) που ενεργοποιεί τη φωσφολιπάση C και το "μονοπάτι" σηματοδότησης τριφωσφορικής ινοσιτόλης (IP3). Είναι χρήσιμο επίσης, να υπογραμμισθούν οι λειτουργίες του υποδοχέα αυτού, αφού αν προσδεθεί με μόρια ισταμίνης, τότε είναι υπεύθυνος για τις λειτουργίες που σχετίζονται με το κεντρικό νευρικό σύστημα, όπως ο κύκλος ύπνου-αφύπνισης ("προωθεί" την αφύπνιση), η θερμοκρασία του σώματος, η νοημοσύνη, η ενδοκρινική ομοιόσταση, ενώ ρυθμίζει και την όρεξη. Στην περίπτωση λειτουργιών πέραν του νευρικού κεντρικού συστήματος είναι υπεύθυνος για την πρόκληση βρογχοσυστολής, σύσπαση των βρογχικών λείων μυών, συστολές της ουροδόχου κύστης, αγγειοδιαστολή, αύξηση της υπεραλγισίας (σπλαχνική υπερευαισθησία), ενώ εμπλέκεται και στην ικανότητα αντίληψης της φαγούρας και της κνίδωσης (Thangam et al., 2018).

Αντιισταμινικά

Τα αντιισταμινικά χρησιμοποιούνται για την αναστολή της πρόσδεσης ισταμίνης με κάποιον υποδοχέα, διότι κάτι τέτοιο σε λανθασμένα επίπεδα θα μπορούσε να δημιουργήσει πληθώρα ανεπιθύμητων αποτελεσμάτων. Συγκεκριμένα, βοηθούν στην ανακούφιση ή την πρόληψη των συμπτωμάτων της αλλεργικής ρινίτιδας και άλλων τύπων αλλεργίας. Συνήθως, οι άνθρωποι παίρνουν τα αντιισταμινικά ως ένα φθηνό, γενικό φάρμακο, χωρίς την προϋπόθεση συνταγογράφησης, ικανό να προσφέρει ανακούφιση από τη ρινική συμφόρηση, το φτέρνισμα, τις κυψέλες που προκαλούνται από τη γύρη, τα ακάρεα σκόνης ή την αλλεργία στα ζώα με μικρές παρενέργειες. Τα αντιισταμινικά είναι συνήθως για βραχυχρόνια θεραπεία (Khelemsky, Gritsenko, and Maerz, 2017; Church and Church, 2013), ενώ οι χρόνιες αλλεργίες αυξάνουν τον κίνδυνο των προβλημάτων υγείας που ενδέχεται να μην μπορούν να αντιμετωπιστούν, συμπεριλαμβανομένου του άσθματος και της λοίμωξης του κατώτερου αναπνευστικού συστήματος.

Αν και η λέξη «αντιισταμινικά» έχει γίνει συνώνυμη με τα φάρμακα για τη θεραπεία αλλεργιών,

οι γιατροί και οι επιστήμονες χρησιμοποιούν τον όρο για να περιγράψουν μια κατηγορία φαρμάκων που αντιτίθεται στη δραστηριότητα πρόσδεσης ισταμίνης σε συγκεκριμένους υποδοχείς στο σώμα. Συνεπώς, χωρίζονται σε κατηγορίες σύμφωνα με τον υποδοχέα ισταμίνης, στον οποίον δρουν. Οι δύο μεγαλύτερες κατηγορίες αντιισταμινικών είναι τα H1-αντιισταμινικά και τα H2-αντιισταμινικά. Η παρούσα μελέτη ασχολείται με χημικά μόρια που θα μπορούσαν να δράσουν ως αντιισταμινικά-H1, δηλαδή να αναστείλουν την πρόσδεση ισταμίνης στον υποδοχέα H1. Τέτοια ουσίες χρησιμοποιούνται για τη θεραπεία αλλεργικών αντιδράσεων στη μύτη, ενώ μπορούν να χρησιμοποιηθούν και για την αντιμετώπιση της αϋπνίας ή του ιλίγγου που προκαλείται από προβλήματα στο εσωτερικό του αυτιού.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

Στο Κεφάλαιο αυτό, παρουσιάζονται οι αλγόριθμοι, οι τεχνικές και οι μεθοδολογίες που θα χρησιμοποιηθούν στην παρούσα διπλωματική, για την παραγωγή των διαφόρων μοντέλων. Η κατανόησή τους από τον αναγνώστη πριν από την παρουσίαση της ανάλυσης και σχεδίασης του συστήματος μάθησης είναι αναγκαία.

2.1 Διαδικασίες Μάθησης

Η μηχανική στατιστική μάθηση αναφέρεται σε ένα τεράστιο σύνολο εργαλείων για την κατανόηση των δεδομένων. Αυτά τα εργαλεία μπορούν να ταξινομηθούν όπως αναφέρθηκε στο Κεφάλαιο 1 ως εποπτευόμενα ή μη εποπτευόμενα (James et al., 2013). Παρακάτω, παρατίθεται μία σχηματική αναπαράσταση λειτουργίας των αλγορίθμων των δύο κατηγοριών ενώ αναλύονται και αλγόριθμοι κάθε κατηγορίας που χρησιμοποιήθηκαν στην παρούσα μελέτη.

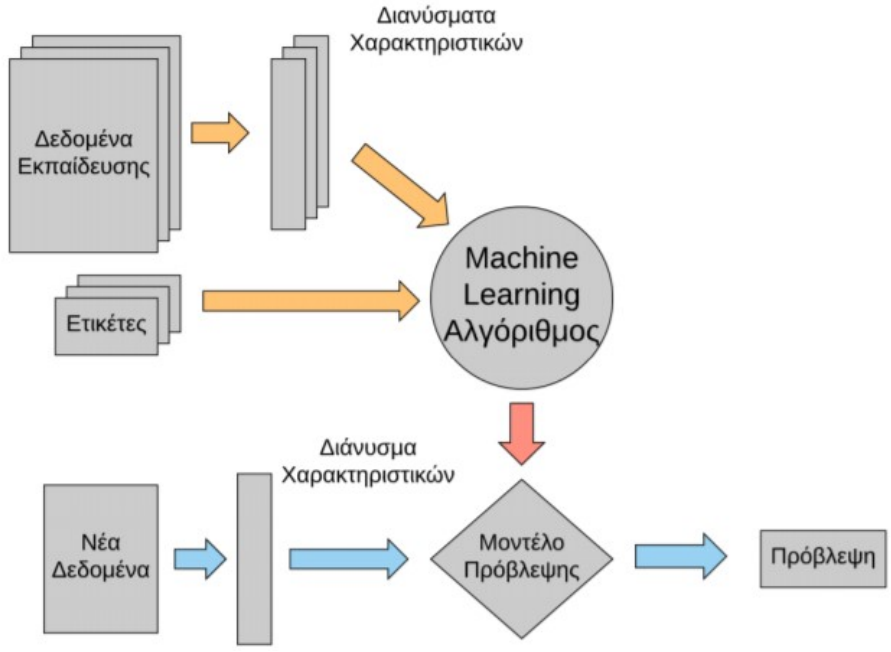
2.1.1 Εποπτευόμενη Μάθηση (Supervised Learning)

Όπως αναφέρθηκε και στην ενότητα (1.1.2) η εποπτευόμενη μάθηση είναι η διαδικασία μηχανικής μάθησης κατά την οποία γίνεται προσέγγιση μιας άγνωστης συνάρτησης, που απεικονίζει τα δεδομένα εισόδου (inputs) x στις εξόδους/προβλέψεις (outputs) y, δοσμένου ενός καθορισμένου συνόλου ζευγών εισόδου-εξόδου $D = \{x_i, y_i\}_{i=1}^{N}$. Ως D ονομάζεται το σύνολο δεδομένων εκπαίδευσης των αλγορίθμων μάθησης, ενώ N είναι το πλήθος των παραδειγμάτων που εκπαιδεύονται οι αλγόριθμοι.

Στην απλούστερη δυνατή εκδοχή του προβλήματος, κάθε δεδομένο εισόδου εκπαίδευσης x_i θα είναι ένα D-διαστατο διάνυσμα αριθμών, που μπορεί να αντιπροσωπεύει, για παράδειγμα, το βάρος και το ύψος ενός ατόμου. Αυτά τα στοιχεία ορίζονται ως **επεξηγηματικές μεταβλητές, χαρακτηριστικά, ιδιότητες ή συμμεταβλητές** και σε γενικότερο πλαίσιο ως x_i θα μπορούσε να οριστεί ένα πολύπλοκα δομημένο αντικείμενο, όπως μια εικόνα, μία πρόταση, ένα μήνυμα ηλεκτρονικού ταχυδρομείου, ένα μοριακό σχήμα, ένα γράφημα κλπ. Ομοίως, η μορφή της μεταβλητής y_i, που ορίζεται ως **μεταβλητή εξόδου ή απόκρισης** μπορεί κατά γενικό κανόνα να είναι οτιδήποτε. Επισημ-

μαίνεται βέβαια, ότι οι περισσότερες μέθοδοι μηχανικής μάθησης θεωρούν ότι το y_i είναι μια **κατηγορική ή ποιοτική** μεταβλητή προερχόμενη από κάποιο πεπερασμένο σύνολο, $y_i \in \{1, ..., M\}$ (για παράδειγμα δηλώνει το φύλλο αρσενικό ή θηλυκό) ή ότι είναι μια πραγματική συνεχής μεταβλητή (όπως για παράδειγμα επίπεδο εισοδήματος κάποιου). Συγκεκριμένα, όταν οι μεταβλητές y_i είναι κατηγορικές, το πρόβλημα είναι γνωστό ως **κατηγοριοποίηση (classification) ή αναγνώριση προτύπου (pattern recognition)**, ενώ όταν το y_i εκφράζεται με πραγματικές συνεχείς αριθμητικές τιμές, το πρόβλημα που θα κληθεί να επιλύσει ένας αλγόριθμος εποπτευόμενης μάθησης καλείται **παλινδρόμηση**. Στην ειδική περίπτωση που η μεταβλητή απόκρισης Y λαμβάνει τιμές από ένα σύνολο τιμών που εκφράζεται κάποια φυσική διάταξη (για παράδειγμα μικρό-μεσαίο-μεγάλο), ορίζουμε την παραλλαγή αυτή ως **τακτική παλινδρόμηση (ordinal regression)**.

Είναι σημαντικό να αναφερθεί, ότι οι αλγόριθμοι εποπτευόμενης μάθησης παράγουν μια συνάρτηση/στατιστικό μοντέλο αναλύοντας τα δεδομένα εκπαίδευσης, που μπορεί να χρησιμοποιηθεί και για την αντιστοίχηση νέων δειγμάτων. Ιδανικά, μετά την εκπαίδευσή τους θα μπορούσαν να προβλέψουν την κατηγορία δειγμάτων που δεν τους είχαν ξαναπαρουσιαστεί. Επομένως, αυτό προϋποθέτει ότι τέτοιοι αλγόριθμοι έχουν την ικανότητα να γενικεύουν από τα δεδομένα εκπαίδευσης σε δεδομένα που δεν είχαν αντιμετωπίσει ξανά. Στο Γράφημα 2.1 παρουσιάζεται μια σχηματική αναπαράσταση της διαδικασίας που υλοποιείται για επίλυση προβλημάτων με εποπτευόμενη μάθηση.

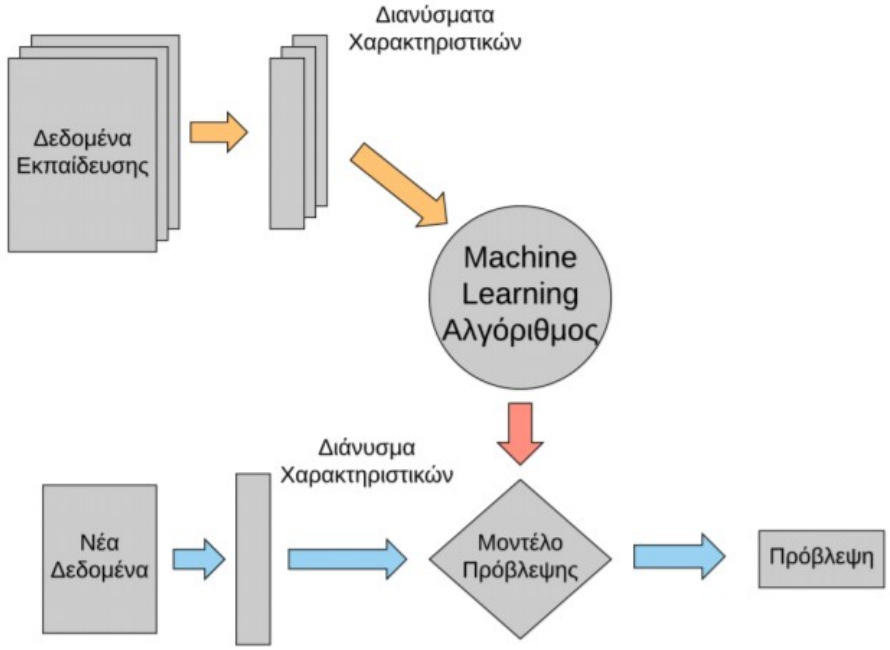


Γράφημα 2.1: Διαδικασία Εποπτευόμενης Μάθησης

2.1.2 Μη Εποπτευόμενη Μάθηση (Unsupervised Learning)

Στην ενότητα (1.1.2) έγινε αναφορά και στη μη εποπτευόμενη μάθηση, που αποτελεί τη δεύτερη μεγάλη κατηγορία προβλημάτων μηχανικής μάθησης. Το είδος αυτό περιγράφει την κάπως πιο δύσκολη κατάσταση στην οποία για κάθε παρατήρηση $i \in \{1, ..., N\}$ παρατηρούμε ένα διάνυσμα $D = \{x_i\}_{i=1}^{N}$ των μετρήσεων x_i αλλά χωρίς σχετική απόκριση y_i. Σε προβλήματα τέτοιου τύπου

ο αλγόριθμος κατά μία έννοια εργάζεται στα τυφλά, για αυτό και η διαδικασία που ακολουθείται αποκαλείται ορισμένες φορές και ανακάλυψη γνώσης. Επί της ουσίας, η μάθηση ονομάζεται μη εποπτευόμενη επειδή δεν υπάρχει μεταβλητή απόκρισης που μπορεί να επιβλέπει την εκάστοτε ανάλυση. Ειδικότερα, η μάθηση χωρίς επίβλεψη, αποτελείται από ένα σύνολο στατιστικών εργαλείων που προορίζονται για την περίπτωση που έχουμε μόνο ένα σύνολο p επεξηγηματικών μεταβλητών $X_1, X_2, ..., X_p$ για N παρατηρήσεις, ενώ λόγω έλλειψης συνδεδεμένης μεταβλητής απόκρισης Y στις επεξηγηματικές μεταβλητές στόχος δεν είναι η πρόβλεψη, αλλά να ανακαλυφθεί αν υπάρχουν υποομάδες μεταξύ των επεξηγηματικών μεταβλητών p ή μεταξύ των παρατηρήσεων i. Επίσης, γίνεται αντιληπτό ότι δεν υπάρχει προφανής μέτρηση σφάλματος που μπορεί να χρησιμοποιηθεί (σε αντίθεση με την εποπτευόμενη μάθηση, όπου μπορούμε να συγκρίνουμε την πρόβλεψή μας της μεταβλητής y για μία δοσμένη επεξηγηματική μεταβλητή x ως προς την παρατηρούμενη τιμή). Στο Γράφημα 2.2 παρουσιάζεται μια σχηματική αναπαράσταση της διαδικασίας που υλοποιείται για επίλυση προβλημάτων με μη εποπτευόμενη μάθηση.



Γράφημα 2.2: Διαδικασία Μη Εποπτευόμενης Μάθησης

2.2 Διαδικασίες Εποπτευόμενης Μάθησης

Στην παρούσα μελέτη χρησιμοποιήθηκαν συγκεκριμένες διαδικασίες εποπτευόμενης μάθησης που κυρίως σχετίζονται με προβλήματα κατηγοριοποίησης. Στα πλαίσια τέτοιων προβλημάτων θα έχουμε ένα σύνολο παρατηρήσεων εκπαίδευσης $(x_1, y_1), ..., (x_n, y_n)$ που μπορούμε να χρησιμοποιήσουμε για την κατασκευή έναν ταξινομητή, με τις μεταβλητές y να είναι ποιοτικού και όχι ποσοτικού χαρακτήρα. Θέλουμε μάλιστα, ο ταξινομητής μας να αποδίδει καλά όχι μόνο για τα δεδομένα στα οποία έχει εκπαιδευτεί αλλά και σε παρατηρήσεις δοκιμών που δεν χρησιμοποιήθηκαν για την εκπαίδευσή του.

2.2.1 Πολλαπλή Λογιστική Παλινδρόμηση (Multiple Logistic Regression)

Στην περίπτωση που η μεταβλητή απόκρισης Y είναι κατηγορική και για την πρόβλεψή της χρησιμοποιείται πλήθος επεξηγηματικών μεταβλητών ένας συνήθης τρόπος μοντελοποίησης ενός συστήματος που να τις περιγράφει είναι η χρήση Πολλαπλής Λογιστικής Παλινδρόμησης. Για ευκολία περιγραφής του μοντέλου θα θεωρηθεί ότι η Y εκφράζεται από τις τιμές 0/1 (βέβαια τα 0,1 είναι ενδεικτικές αναπαραστάσεις δύο κατηγοριών). Επομένως, η λογιστική παλινδρόμηση μοντελοποιεί την πιθανότητα η Y να ανήκει σε μία συγκεκριμένη κατηγορία. Για την μοντελοποίηση αυτή το λογιστικό μοντέλο χρησιμοποιεί την εξής λογιστική συνάρτηση:

$$p(Y=1) = \frac{e^X}{1+e^X} \text{ , } \mu\varepsilon\ X = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \tag{2.1}$$

Αναλυτικότερα, η μεταβλητή X εκπροσωπεί τη δράση μιας ομάδας ανεξάρτητων μεταβλητών ενώ η p(Y=1) προσδιορίζει την πιθανότητα ενός συγκεκριμένου αποτελέσματος λόγω της δράσης της ομάδας αυτής. Η μεταβλητή X (λογιστική) εκφράζει επίσης το μέτρο της ολικής συνεισφοράς όλων των συμμετεχουσών ανεξάρτητων μεταβλητών στο μοντέλο, όπου β_0 είναι το ύψος της κλίσης της γραμμής παλινδρόμησης και ισούται με την τιμή X όταν οι τιμές όλων των ανεξάρτητων μεταβλητών ισούνται με 0, ενώ β_i είναι οι συντελεστές παλινδρόμησης καθένας από τους οποίους εκφράζει το μέγεθος συνεισφοράς της αντίστοιχης μεταβλητής. Θετική τιμή του συντελεστή δηλώνει ότι η επεξηγηματική μεταβλητή αυξάνει την πιθανότητα της επιτυχημένης έκβασης (να συμβεί δηλαδή το γεγονός), αρνητική τιμή σημαίνει ότι η μεταβλητή μειώνει την πιθανότητα αυτής της έκβασης. Υψηλή τιμή του συντελεστή σημαίνει ότι η ανεξάρτητη μεταβλητή επηρεάζει πολύ ισχυρά την πιθανότητα να συμβεί το γεγονός ή μη, ενώ χαμηλή τιμή δηλώνει μικρή επίδραση της ανεξάρτητης μεταβλητής στην πιθανότητα εμφάνισης της ανάλογης έκβασης.

Μετά από μερικές παρεμβάσεις η (2.1), μετατρέπεται στην εξής σχέση:

$$\frac{p(Y=1)}{1-p(Y=1)} = e^X \text{ , } \mu\varepsilon\ X = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \tag{2.2}$$

Η ποσότητα $\frac{p(Y=1)}{1-p(Y=1)}$ ορίζεται ως λόγος συμπληρωματικών πιθανοτήτων (odds) και εκφράζει τις πιθανότητες που συγκλίνουν υπέρ της εμφάνισης ενός γεγονότος ή πρόθεσης του να συμβεί. Ειδικότερα, ο αριθμητής προσδιορίζει την πιθανότητα που έχει το προσδοκώμενο γεγονός να συμβεί και ο παρονομαστής την πιθανότητα να μη συμβεί, ενώ οι τιμές που μπορεί να δεχθεί η προαναφερθείσα ποσότητα είναι στο διάστημα $\{0, \infty\}$. Για παράδειγμα, θεωρώντας κατά μέσο όρο 1 στα 5 άτομα θα υπονοείται τιμή του λόγου ζεύγους ακέραιων τιμών (odds) της τάξης του 1/4, αφού το p(Y=1)= 1/5 = 0,2 υποδηλώνει λόγο συμπληρωματικών πιθανοτήτων (odds) $\frac{0.2}{1-0.2} = 1/4$.

Λογαριθμίζοντας τις δύο πλευρές λαμβάνεται η σχέση του λογαριθμοποιημένου λόγου συμπληρωματικών πιθανοτήτων (logit):

$$\log\left(\frac{p(Y=1)}{1-p(Y=1)}\right) = X \text{ , } \mu\varepsilon\ X = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \tag{2.3}$$

Η παραπάνω σχέση θα μπορούσε κάλλιστα να ενσωματωθεί σε ένα μοντέλο παλινδρόμησης.

Ακόμη, ο υπολογισμός των συντελεστών β_i του μοντέλου λογιστικής παλινδρόμησης γίνεται με τη βοήθεια της εκτίμησης της **Συνάρτησης Πιθανοφάνειας (Likelihood Estimate – LE)**, ως:

$$\ell(\beta_0, ..., \beta_N) = \prod_{i:y_i=1} p(y_i) \prod_{i':y_i'=0} (1 - p(y_i)) \tag{2.4}$$

Όπου οι εκτιμήσεις των $\hat{\beta_0}, ..., \hat{\beta_N}$ επιλέγονται να μεγιστοποιούν τη (2.4). Επίσης, στη λογιστική παλινδρόμηση για έλεγχο υποθέσεων έχουμε τον έλεγχο Wald για κάθε συντελεστή β_i που ισούται με $\frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$, όπου $SE(\hat{\beta}_i)$ είναι το σφάλμα της εκτίμησης του εκάστοτε συντελεστή.

2.2.2 Βελτιωτικές Μέθοδοι Λογιστικής Παλινδρόμησης

Βέβαια, πέρα από την κλασική λογιστική παλινδρόμηση υπάρχουν εναλλακτικές διαδικασίες προσαρμογής που μπορούν να αποδώσουν καλύτερη ακρίβεια προβλέψεων και αναπαράσταση μοντέλων παλινδρόμησης. Συγκεκριμένα, η ακρίβεια πρόβλεψης που θα αναλυθεί και στην ενότητα για τα μέτρα αξιολόγησης ανάλογα με τη σχέση του πλήθους παρατηρήσεων και επεξηγηματικών μεταβλητών μπορεί να επηρεαστεί αισθητά. Αυτό συμβαίνει, γιατί στην περίπτωση που οι παρατηρήσεις είναι υπερβολικά πολλές σε σχέση με το πλήθος επεξηγηματικών μεταβλητών τότε η διασπορά τείνει να είναι ιδιαίτερα μικρή ενώ στην αντίστροφη περίπτωση η διασπορά τείνει στο άπειρο. Έτσι, με χρήση μεθόδων που συρρικνώνουν τους συντελεστές του εκάστοτε μοντέλου, συχνά μειώνεται η διασπορά με κόστος μίας αμελητέας αύξησης της μεροληψίας. Αυτό μπορεί να οδηγήσει σε ουσιαστικές βελτιώσεις στην ακρίβεια με την οποία προβλέπεται η μεταβλητή απόκρισης μίας παρατήρησης που δεν έγινε χρήση της για την εκπαίδευση του μοντέλου.

Από την άλλη η αναπαράσταση των μοντέλων αρκετές φορές καθίσταται δυσκολότερη λόγω μεταβλητών που χρησιμοποιούνται σε ένα μοντέλο πολλαπλής παλινδρόμησης ενώ δεν συνδέονται ιδιαίτερα με την μεταβλητή απόκρισης. Η χρήση τέτοιων "ασήμαντων" μεταβλητών οδηγεί σε περιττή πολυπλοκότητα στο προκύπτον μοντέλο. Αφαιρώντας αυτές τις μεταβλητές -δηλαδή, θέτοντας τις αντίστοιχες εκτιμήσεις συντελεστών στο μηδέν- μπορεί να αποκτηθεί ένα μοντέλο που ερμηνεύεται πιο εύκολα. Παρακάτω, παρουσιάζονται μερικές προσεγγίσεις για αυτόματη επιλογή επεξηγηματικών μεταβλητών που σχετίζονται με τη μεταβλητή απόκρισης, καθώς και για συρρίκνωση των συντελεστών.

Μέθοδοι με βήματα

Ορισμένες μέθοδοι για την εύρεση υποομάδων προβλεπτικών όρων είναι η επιλογή του καλύτερου υποσυνόλου και οι μέθοδοι κατά βήματα. Η μέθοδος καλύτερου υποσυνόλου, που δεν αναλύεται στην παρούσα μελέτη, παρουσιάζει την εξής αδυναμία: Όσο μεγαλύτερος είναι ο χώρος αναζήτησης, τόσο μεγαλύτερη είναι η πιθανότητα εύρεσης μοντέλων που φαίνεται ότι περιγράφουν καλά τα δεδομένα

στα οποία εκπαιδεύτηκαν, παρόλο που μπορεί να μην έχουν καμία προβλεπτική ισχύ για μελλοντικά δεδομένα. Έτσι ένας τεράστιος χώρος αναζήτησης μπορεί να οδηγήσει σε υπερφόρτωση και μεγάλη διασπορά των εκτιμήσεων των συντελεστών. Για αυτούς τους λόγους, εφαρμόστηκαν στην έρευνά μας οι παρακάτω βηματικές μεθόδοι, οι οποίες διερευνούν ένα πολύ πιο περιορισμένο σύνολο μοντέλων, καταλήγοντας συχνά σε μοντέλα με λιγότερες επεξηγηματικές μεταβλητές.

- **Επιλογή με βήματα προς τα μπρός(Forward Stepwise Selection)** που περιγράφεται από τον παρακάτω αλγόριθμο.

1. Θεωρώ μοντέλο M_0 που υποδηλώνει το μηδενικό μοντέλο, το οποίο δεν περιέχει επεξηγηματικές μεταβλητές.
2. Για $k = 0, ..., p-1$:
   - Εξέταση όλων των μοντέλων $p-k$ που αυξάνουν τις επεξηγηματικές μεταβλητές στο M_k με μία επιπλέον επεξηγηματική μεταβλητή.
   - Επιλογή του καλύτερου μεταξύ αυτών των $p-k$ μοντέλων και ονομασία του ως M_{k+1}.
3. Επιλέγεται ένα μοναδικό καλύτερο μοντέλο από τα $M_0, ..., M_p$ χρησιμοποιώντας το δείκτη (AIC).

Επιλογή με βήματα προς τα πίσω(Backward Stepwise Selection)

1. Θεωρώ μοντέλο M_p που υποδηλώνει το μοντέλο με όλες τις p διαθέσιμες επεξηγηματικές μεταβλητές.
2. Για $k = p, ..., 1$:
   - Εξέταση όλων των μοντέλων k που περιέχουν όλες πλην μίας τις επεξηγηματικές μεταβλητές στο M_k, για συνολικά k-1 μεταβλητές.
   - Επιλογή του καλύτερου μεταξύ αυτών των k μοντέλων και ονομασία του ως M_{k-1}.
3. Επιλέγεται ένα μοναδικό καλύτερο μοντέλο από τα $M_0, ..., M_p$ χρησιμοποιώντας το δείκτη (AIC).

Και στις 2 μεθόδους, διαπερνόνται μόνο $1+\frac{p(p+1)}{2}$ μοντέλα, και έτσι μπορούν οι βηματικές μέθοδοι να εφαρμοστούν σε περιπτώσεις όπου το p είναι πολύ μεγάλο, για να εφαρμοστεί η καλύτερη επιλογή υποσυνόλου. Επίσης, και οι δύο μέθοδοι δεν εγγυώνται τη δημιουργία του καλύτερου δυνατού μοντέλου που περιέχει ένα υποσύνολο των προγνωστικών p.

Μέθοδος συρρίκνωσης Lasso

Γενικά οι μέθοδοι κατά βήματα που παρουσιάστηκαν πρωτύτερα βελτιώνουν σε ορισμένες περιπτώσεις κυρίως την ακρίβεια της πρόβλεψης, όπως όταν μόνο λίγες επεξηγηματικές μεταβλητές έχουν ισχυρή σχέση με το αποτέλεσμα, ενώ περιλαμβάνουν τη χρήση ελαχίστων τετραγώνων για την προσαρμογή

ενός γραμμικού μοντέλου που περιέχει ένα υποσύνολο των επεξηγηματικών μεταβλητών. Εναλλακτικά, ως λύση θα μπορούσε να προσαρμοστεί ένα μοντέλο που περιέχει όλους τους επεξηγηματικούς παράγοντες p χρησιμοποιώντας τεχνικές που περιορίζουν ή ρυθμίζουν τις εκτιμήσεις συντελεστών, ή ισοδύναμα, συρρικνώνουν τις εκτιμήσεις συντελεστών προς το μηδέν. Τέτοια μέθοδος είναι και η γνωστή μέθοδος **Lasso**, η οποία, βέβαια, εκτός από συρρίκνωση των συντελεστών πετυχαίνει και επιλογή των κατάλληλων επεξηγηματικών μεταβλητών για το τελικό μοντέλο (Tibshirani, 1996).

Ο εκτιμητής της λογιστικής Lasso παλινδρόμησης βασίζεται στην επιλογή μίας κατάλληλης παραμέτρου $\lambda \geq 0$, η οποία προσδιορίζεται ξεχωριστά. Οι εκτιμήσεις των συντελεστών του τελικού μοντέλου είναι οι τιμές που μεγιστοποιούν την ακόλουθη ελαφρώς διαφορετική από την (2.4) λογαριθμική συνάρτηση πιθανοφάνειας, όπου προστίθεται και η ποινή L1, καταλήγοντας στην εξής συνάρτηση μεγιστοποίησης με περιορισμούς (Hastie, Tibshirani, and Friedman, 2009):

$$\ell(\beta) = \sum_{i=1}^{n} [y_i x_i \beta - \log(1 + \exp x_i \beta))] - \lambda \sum_{j=1} p|\beta_j| \tag{2.5}$$

Ακόμη, η ποινή L1 που χρησιμοποιείται στη Lasso χρησιμοποιείται τόσο για την επιλογή των μεταβλητών όσο και για τη συρρίκνωση, καθώς έχει την ιδιότητα, όταν το λ είναι αρκετά μεγάλο, να αναγκάζει ορισμένες από τις εκτιμήσεις συντελεστών να είναι ακριβώς ίσες με το μηδέν (James et al., 2013). Η Lasso έχει το πλεονέκτημα, πως το τελικό μοντέλο μπορεί να περιλαμβάνει μόνο ένα υποσύνολο των προγνωστικών, το οποίο με τη σειρά του βελτιώνει την αναπαράσταση του μοντέλου.

2.2.3 Πολυωνυμική Λογιστική Παλινδρόμηση (Multinomial Logistic Regression)

Υπάρχει όμως η περίπτωση, να χρειαστεί να προβλεφθεί η τιμή μίας μεταβλητής απόκρισης, η οποία δε λαμβάνει την τιμή της από δύο πιθανά ενδεχόμενα αλλά υπάρχουν $k > 2$ το πλήθος διακριτές κατηγορίες, που θα μπορούσαν να την περιγράψουν.

Για την επίλυση ενός τέτοιου προβλήματος με χρήση λογιστικής παλινδρόμησης δεν επαρκούν μόνο όσα αναλύθηκαν παραπάνω και, έτσι γενικεύεται η απλή λογιστική παλινδρόμηση σε πολυωνυμική λογιστική παλινδρόμηση που εξηγείται παρακάτω. Συγκεκριμένα, στην πολυωνυμική λογιστική παλινδρόμηση η εκτίμηση των παραμέτρων γίνεται λαμβάνοντας μία από τις κατηγορίες σαν κατηγορία αναφοράς για να γίνουν ως προς αυτή συγκρίσεις. Μία παρατήρηση μπορεί να λάβει τιμή από $j \in \{1, ..., J\}$ διαθέσιμες κατηγορίες με τις εκ των προτέρων πιθανότητες κάθε κατηγορίας να ορίζονται ως,

$$\pi_j(x) = Pr\{Y = j|x\} \text{ , όπου x επεξηγηματικές μεταβλητές , } \sum_{j=1}^{J} \pi_{ij} \tag{2.6}$$

Ακόμη, η πιο συνήθης σχέση του λογαριθμοποιημένου λόγου συμπληρωματικών πιθανοτήτων (logit) στην πολυωνυμική λογιστική παλινδρόμηση ορίζεται ως,

$$log \frac{\pi_j(x)}{\pi_J(x)} = a_j + \beta'_j x \tag{2.7}$$

όπου $j = 1, ..., (J-1)$ και περιγράφει ταυτόχρονα τα αποτελέσματα επίδρασης των επεξηγηματικών μεταβλητών x στα J-1 logit μοντέλα. Ακόμη, οι παραπάνω λόγοι "ζευγαρώνουν" κάθε κατηγορία της μεταβλητής απόκρισης με την κατηγορία αναφοράς. Τέλος, η κατανομή πιθανότητας, για την περιγραφή των μεταβλητών απόκρισης των παρατηρήσεων ενός συνόλου δεδομένων δίνεται από την πολυωνυμική κατανομή,

$$\binom{Pr\{Y_{i1} = y_{i1}, ..., Y_{iJ} = y_{iJ}\} = n_i}{y_{i1}, ..., y_{iJ}\pi_{i1}^{y_i i1}...\pi_{iJ}^{y_i iJ}} \tag{2.8}$$

2.2.4 Γραμμική Διακριτική Ανάλυση (Linear Discriminant Analysis)

Εν αντιθέσει με τη λογιστική παλινδρόμηση, που μοντελοποιεί άμεσα μέσω της λογιστικής συνάρτησης την υπό συνθήκη κατανομή των μεταβλητών απόκρισης Y για δοσμένες επεξηγηματικές μεταβλητές X, υπάρχει μία πιο έμμεση επίτευξη του υπολογισμού των προς μοντελοποίηση πιθανοτήτων $Pr(Y = k|X = x)$, που θα μπορούσε να προτιμηθεί σε συγκεκριμένες περιπτώσεις. Η μέθοδος αυτή είναι η **γραμμική διακριτική ανάλυση (linear discriminant analysis)** και όπως θα αναλυθεί παρακάτω κάνει χρήση του θεωρήματος Bayes στην προσπάθειά της για υπολογισμό των $Pr(Y = k|X = x)$. Ακόμη, έχει αποδειχθεί ότι σαν μέθοδος δρα ιδιαίτερα καλά σε περιπτώσεις μεταβλητών απόκρισης με παραπάνω από δύο κατηγορίες, ενώ όταν οι κατηγορίες των μεταβλητών πληθυσμιακά είναι καλά χωρισμένες, δεν αντιμετωπίζει προβλήματα σταθερότητας στις εκτιμήσεις των παραμέτρων, όπως έχει παρατηρηθεί στην περίπτωση της λογιστικής παλινδρόμησης.

Θεώρημα Bayes και η χρήση του σε προβλήματα ταξινόμησης

Όπως, προαναφέρθηκε η μέθοδος LDA χρησιμοποιεί το θεώρημα Bayes, άρα η παρουσίασή του καθίσταται αναγκαία. Έστω λοιπόν, ότι γίνεται προσπάθεια ταξινόμησης μιας παρατήρησης σε μία από $K \geq 2$ πιθανές κατηγορίες, τότε για να μπορέσει να οριστεί το θεώρημα θεωρούνται αρχικά, τα εξής:

- Η εκ των προτέρων πιθανότητα (prior probability) π_k. Η ποσότητα αυτή, δηλώνει την πιθανότητα μία δοσμένη παρατήρηση να σχετίζεται με την k-στη κατηγορία της μεταβλητής απόκρισης Y.

- Η συνάρτηση πυκνότητας των επεξηγηματικών μεταβλητών X, για κάθε παρατήρηση προερχόμενη από την k-στη κατηγορία, $f_k(X) \equiv Pr(X = x|Y = k)$. Πρακτικά, μεγάλη τιμή της $f_k(x)$ προκύπτει στην περίπτωση, που υπάρχει υψηλή πιθανότητα για μία παρατήρηση στην k-στη κατηγορία να ισχύει ότι $X \approx x$. Το αντίστροφο ισχύει στην περίπτωση που η τιμή της $f_k(x)$ είναι μικρή.

Ορίζεται έτσι το θεώρημα Bayes ως:

$$p_k(X) = Pr(Y = k|X = x) = \frac{\pi_k f_k}{\sum_{l=1}^{K} \pi_l f_l(x)} \tag{2.9}$$

Άρα, αντί να υπολογίζεται άμεσα η τιμή του $p_k(X)$, όπως στη λογιστική παλινδρόμηση, η σχέση (2.9) δείχνει ότι αυτό μπορεί να γίνει απλά χρησιμοποιώντας τις εκτιμήσεις των $\pi_k, f_k(X)$. Συνήθως, η

εκτίμηση του π_k είναι εύκολη, αφού για ένα τυχαίο δείγμα μεταβλητών απόκρισης Y από το συνολικό τους πληθυσμό, υπολογίζεται απλά το κλάσμα των παρατηρήσεων εκπαίδευσης που ανήκουν στην τάξη k. Ωστόσο, η διαδικασία εκτίμησης της $f_k(X)$ τείνει να είναι πιο δύσκολη, εκτός και αν θεωρηθούν μερικές απλές μορφές για αυτές τις συναρτήσεις πυκνότητας. Επομένως, είναι χρήσιμο να βρεθεί τρόπος υπολογισμού της $f_k(X)$, ώστε να αναπτυχθεί μετά κατάλληλος ταξινομητής. Ιδανικά υπολογίζεται έτσι ώστε να προσεγγίζει τον ταξινομητή Bayes, ο οποίος έχει και το μικρότερο δείκτη σφάλματος από όλους τους ταξινομητές.

Γραμμική διακριτική ανάλυση για p>1

Η εκτίμηση λοιπόν του ταξινομητή LDA μετά τον ορισμό του θεωρήματος Bayes και της παρατήρησης για εύρεση κατάλληλης συνάρτησης πυκνότητας πιθανότητας συνοψίζεται παρακάτω. Αρχικά, θεωρείται ότι το πλήθος των επεξηγηματικών μεταβλητών $X = (X_1, ..., X_p)$ ακολουθεί πολυμεταβλητή κανονική κατανομή, με διακριτές μέσες τιμές ανά κατηγορία, αλλά κοινό πίνακα συνδιακύμανσης. Συμβολικά, όταν μία p-διαστατη τυχαία μεταβλητή X ακολουθεί πολυμεταβλητή Gaussian (κανονική) κατανομή, γράφεται $X \sim N(\mu_k, \mathbf{\Sigma})$, με $E(X) = \mu_k$ να αποτελεί τη μέση τιμή της k-στης κλάσης της X (διάνυσμα p συνιστωσών) και $Cov(X) = \mathbf{\Sigma}$ τον $p \times p$ πίνακα συνδιακύμανσης που είναι κοινός για όλες τις K κλάσεις. Η συνάρτηση πυκνότητας πιθανότητας της πολυμταβλητής Gaussian κατανομής ορίζεται ως,

$$f_k(X = x) = \frac{1}{\sqrt{2\pi}^p |\mathbf{\Sigma}|^{1/2}} \exp\big(-\frac{1}{2}(x - \mu_k)^T \mathbf{\Sigma}(x - \mu_k)\big), \tag{2.10}$$

Έπειτα, αντικαθιστώντας τη συνάρτηση πυκνότητας πιθανότητας (2.6) στη σχέση (2.5) και λογαριθμώντας τα δύο μέλη προκύπτει ότι ο ταξινομητής LDA λειτουργεί εναποθέτοντας την παρατήρηση $X = x$ μέσω του ταξινομητή Bayes στην κατηγορία εκείνη που η τιμή της διακριτικής συνάρτησης,

$$\delta_k(X = x) = x^T \mathbf{\Sigma}^{-1} \mu_k - \frac{1}{2} \mu_k^T \mathbf{\Sigma}^{-1} \mu_k + \log \pi_k \tag{2.11}$$

καθίσταται μεγαλύτερη. Επίσης, ο ταξινομητής LDA για να προσεγγίσει τον ταξινομητή Bayes, πρέπει να εκτιμηθούν και οι παράμετροι $\mu_i, i = 1, ..., k$,$\pi_i, i = 1, ..., k$ καθώς και ο πίνακας $\mathbf{\Sigma}$. Έτσι ορίζονται τα εξής:

$$\mu_k(X = x) = \frac{\sum_{i=1}^N I(y_i = k) x_i}{\sum_{i=1}^N I(y_i = k)} \tag{2.12}$$

$$\Sigma = \frac{\sum_{i=1}^N (x_i - \hat{\mu})(x_i - \hat{\mu})^T}{N}, \text{ με } \hat{\mu} = \frac{\sum_{i=1}^N x_i}{N} \tag{2.13}$$

$$\pi_k = \frac{\sum_{i=1}^N I(y_i = k)}{N} \tag{2.14}$$

Ακόμη, σημειώνεται ότι η συνάρτηση δ_k που αποτελεί τον κανόνα απόφασης για τον ταξινομητή LDA, είναι γραμμική για μία παρατήρηση x. Έτσι, αιτιολογείται η γραμμικότητα που υποστηρίζεται από το όνομα αυτού του ταξινομητή.

2.2.5 Δένδρα Απόφασης

Οι μέθοδοι που βασίζονται σε δέντρα περιλαμβάνουν τη στρωματοποίηση ή την κατάτμηση του χώρου των χαρακτηριστικών (επεξηγηματικών μεταβλητών) σε ένα πλήθος απλών περιοχών (σε ένα σύνολο

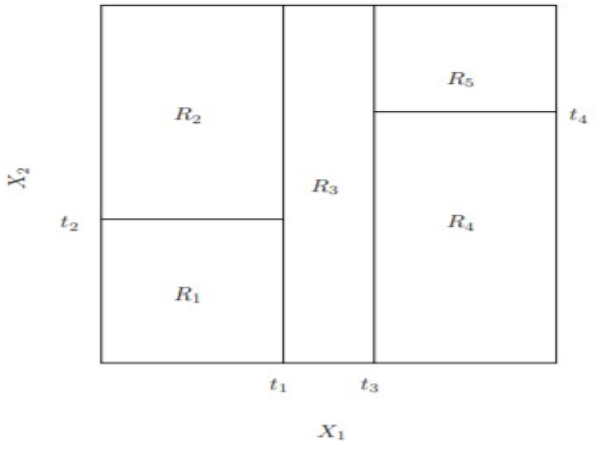
ορθογωνίων) χρησιμοποιώντας μία μεταβλητή κάθε φορά. Στη συνέχεια, κάθε τέτοια περιοχή αντιστοιχίζεται με μια τιμή για την προς κατηγοριοποίηση/ταξινόμηση μεταβλητή που βρίσκεται στο εσωτερικό του. Έτσι και τα δένδρα απόφασης, ακολουθώντας αυτή τη διαδικασία χρησιμοποιούνται ως μοντέλα πρόβλεψης. Ξεκινώντας από τις παρατηρήσεις για ένα αντικείμενο (οι οποίες αναπαρίστανται στα κλαδιά) μπορούν να εξάγουν συμπεράσματα για κάποια άλλα χαρακτηριστικά του αντικειμένου (τα οποία αναπαρίστανται στα φύλλα). Δημοφιλής μέθοδος για δενδρο-βασισμένη παλινδρόμηση και κατηγοριοποίηση (classification) αποτελεί η CART (Classification and Regression Trees) (Breiman et al., 1984), ενώ σημαντικός ανταγωνιστής της είναι η εξελιγμένη της μορφή η μέθοδος C4.5 (Quinlan, 2014). Η παρούσα μελέτη εστιάζεται κυρίως στη μέθοδο CART και ειδικότερα στην δημιουργία δένδρων κατηγοριοποίησης (classification trees).

Η διαδικασία λοιπόν δημιουργίας ενός δένδρου κατηγοριοποίησης περιγράφεται από τα εξής δύο βήματα:

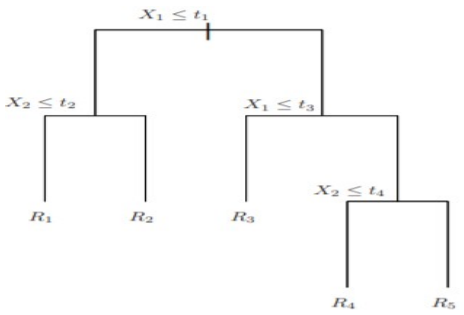
1. Διαχωρισμός του χώρου επεξηγηματικών μεταβλητών - αυτός ο χώρος είναι το σύνολο των πιθανών τιμών των $X_1, ..., X_p$ - σε J διακριτά και μη επικαλυπτόμμενα τμήματα $R_1, ..., R_J$.

2. Για κάθε παρατήρηση που εμπίπτει σε μία περιοχή R_j, προβλέπεται ότι η παρατήρηση αυτή ανήκει στην πιο συχνά εμφανιζόμενη κατηγορία των δεδομένων εκπαίδευσης στην περιοχή που αυτή ανήκει.

Για παράδειγμα, στο Γράφημα 2.3α παρατηρείται ότι από το βήμα 1 έχουν δημιουργηθεί 5 περιοχές. Αν η συχνότερα παρατηρούμενη τιμή απόκρισης των παρατηρήσεων εκπαίδευσης στην πρώτη περιοχή R_1 είναι 1, τότε για μια δεδομένη παρατήρηση X = x, αν $x \in R_1$ θα προβλεφθεί μια τιμή 1. Με αφορμή όμως το παράδειγμα αυτό αξίζει να υπογραμμιστεί ότι κατά την αναπαράσταση των αποτελεσμάτων σε ένα δέντρο ταξινόμησης, συχνά έχει αρκετά ενδιαφέρον όχι μόνο η πρόβλεψη της κλάσης που αντιστοιχεί σε μια συγκεκριμένη περιοχή τερματικού κόμβου, αλλά και το ποσοστό των κατηγοριών μεταξύ των παρατηρήσεων κατάρτισης που εμπίπτουν στην περιοχή αυτή.



(a) Διαμέριση δισδιάστατου χώρο χαρακτηριστικών με αναδρομικό δυαδικό διαχωρισμό



(b) Δέντρο που αντιστοιχεί στη διαμέριση του Γραφήματος 2.3α

Γράφημα 2.3: Αναπαράσταση δημιουργίας δένδρων

Βέβαια πρέπει να αναλυθεί και ο τρόπος με τον οποίο διαιρείται ο χώρος των χαρακτηριστικών στις περιοχές του βήματος 1. Συγκεκριμένα, αν και θεωρητικά, οι περιοχές θα μπορούσαν να έχουν οποιοδήποτε σχήμα επιλέγεται να αναπαρασταθούν ως ορθογώνια υψηλής διάστασης ή κουτιά, για απλότητα και ευκολία ερμηνείας του προκύπτοντος μοντέλου πρόβλεψης. Ο στόχος είναι να βρεθούν τα κουτιά $R_1, ..., R_J$ που ελαχιστοποιούν το ποσοστό σφάλματος ταξινόμησης (classification error rate). Το ποσοστό αυτό δεδομένου, ότι μια παρατήρηση σε μια δεδομένη περιοχή σκοπεύεται να αντιστοιχηθεί στην πιο συχνά εμφανιζόμενη τάξη παρατηρήσεων εκπαίδευσης αυτής της περιοχής, θα αποτελεί το λόγο των παρατηρήσεων εκπαίδευσης σε αυτήν την περιοχή, που δεν ανήκουν στην πιο συχνά παρατηρούμενη τάξη και θα ορίζεται ως

$$E = 1 - \max_k(\hat{p}_{mk}) \tag{2.15}$$

Στη σχέση (2.15) το $\hat{p}_{mk}$ αντιπροσωπεύει το ποσοστό των παρατηρήσεων "εκπαίδευσης" στη m-στη περιοχή από την k-στη κατηγορία.

Δυστυχώς, είναι υπολογιστικά ανέφικτο να ληφθούν υπόψη όλες οι πιθανές κατατμήσεις του χώρου χαρακτηριστικών σε J πλαίσια. Για αυτόν τον λόγο, εφαρμόζεται μια άπληστη προσέγγιση που είναι γνωστή ως αναδρομική δυαδική διάσπαση. Η άπληστη αυτή διαδικασία λειτουργεί επαγωγικά από πάνω προς τα κάτω (top-down approach), αφού αρχίζει στην κορυφή του δέντρου (στο σημείο εκείνο σημείο όπου όλες οι παρατηρήσεις ανήκουν σε μία μόνο περιοχή) και στη συνέχεια διαχωρίζει διαδοχικά το χώρο επεξηγηματικών μεταβλητών· κάθε διάσπαση υποδεικνύεται στο δέντρο μέσω δύο νέων κλάδων προς τα κάτω. Ακόμη, η αλγοριθμική αυτή προσέγγιση είναι άπληστη, επειδή σε κάθε βήμα της διαδικασίας δημιουργίας δένδρων, ο καλύτερος διαχωρισμός γίνεται σε αυτό το συγκεκριμένο βήμα, αντί να κοιτά ο αλγόριθμος μετέπειτα να επιλέξει μία διάσπαση που θα οδηγούσε σε ένα καλύτερο δέντρο σε κάποιο μελλοντικό βήμα.

Προκειμένου λοιπόν, να εκτελέσουμε αναδρομικό δυαδικό διαχωρισμό, επιλέγουμε πρώτα την μεταβλητή X_j και το σημείο "κοπής" (cutpoint) s έτσι ώστε να χωρίζει τον χώρο των προβλεπτικών μεταβλητών στις περιοχές $\{X|X_j < s\}$ και $\{X|X_j \geq s\}$, ενώ παράλληλα οδηγεί στη μέγιστη δυνατή μείωση του δείκτη σφάλματος ταξινόμησης (classification error rate).

Ωστόσο, πρέπει να τονιστεί, πως έχει αποδειχθεί ότι το σφάλμα ταξινόμησης δεν είναι επαρκώς ευαίσθητο για τη δημιουργία δένδρων (για την επιλογή μεταβλητής διαχωρισμού) και στην πράξη η ελαχιστοποίηση δύο άλλων μέτρων είναι προτιμότερη για το διαχωρισμό περιοχών. Αυτά τα μέτρα είναι ο **δείκτης Gini** και η **Διασταυρωμένη Εντροπία (cross entropy)** που συνοψίζονται παρακάτω.

Ο **δείκτης Gini** είναι ένας μετρητής συνολικής διασποράς σε όλες της K κλάσεις και ορίζεται ως:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}), \tag{2.16}$$

είναι λοιπόν εύκολα αντιληπτό ότι ο δείκτης Gini παίρνει μια μικρή τιμή αν όλα τα $\hat{p}_{mk}$ είναι κοντά στο μηδέν ή στο ένα. Για το λόγο αυτό ο δείκτης Gini αναφέρεται ως μετρητής της καθαρότητας (purity) ενός κόμβου, αφού μια μικρή τιμή του υποδεικνύει ότι ένας κόμβος περιέχει κυρίως παρατηρήσεις

από μια μόνη κλάση. Μια εναλλακτική λύση για τον δείκτη Gini όπως αναφέρθηκε αποτελεί η **διασταυρωμένη-εντροπία**, που δίνεται από

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log(\hat{p}_{mk}), \tag{2.17}$$

Επίσης, από τη στιγμή που ισχύει ότι $0 \leq \hat{p}_{mk} \leq 1$, προκύπτει ότι $0 \leq -\hat{p}_{mk} \log(\hat{p}_{mk})$. Οπότε εύκολα διαπιστώνει κάποιος ότι και η διασταυρωμένη εντροπία θα πάρει μια τιμή κοντά στο μηδέν, αν τα $\hat{p}_{mk}$ είναι όλα κοντά στο μηδέν ή κοντά στο ένα. Επομένως, όπως και ο δείκτης Gini, έτσι και η διασταυρωμένη εντροπία θα λαμβάνει μικρή τιμή εάν ο m-στος κόμβος είναι καθαρός (pure). Επομένως, αριθμητικά οι δύο μετρητές είναι παρόμοιοι, ενώ κατά τη δημιουργία ενός δένδρου ταξινόμησης, είτε ο δείκτης Gini είτε η διασταυρωμένη εντροπία (cross entropy) χρησιμοποιούνται για την αξιολόγηση της ποιότητας ενός συγκεκριμένου διαχωρισμού, καθώς αυτές οι δύο προσεγγίσεις είναι πιο ευαίσθητες στην καθαρότητα των κόμβων από ό,τι είναι ο δείκτης.

2.2.6 Bootstrap Aggregation και Τυχαία Δάση (Random Forests)

Ένα Τυχαίο Δάσος (Breiman, 2001) είναι μια τεχνική ικανή να εκτελεί τόσο παλινδρόμηση όσο και ταξινόμηση/κατηγοριοποίηση με τη χρήση πολλαπλών δέντρων αποφάσεων και της τεχνικής που ονομάζεται Bootstrap Aggregation, κοινώς γνωστή ως bagging. Συνεπώς, πριν αναλυθεί η μέθοδος τυχαίων δασών είναι χρήσιμη η περιγραφή της μεθόδου bagging.

Bagging (Bootstrap aggregation)

Η τεχνική **Bagging (Bootstrap aggregation)** αποτελεί μία διαδικασία, που ως σκοπό έχει να βελτιώσει ένα σύνολο μοντέλων με μεγάλο σφάλμα διακύμανσης. Τα δέντρα απόφασης που παρουσιάστηκαν παραπάνω αποτελούν μια τέτοια περίπτωση, αφού το σφάλμα λόγω διακύμανσης αυξάνεται καθώς αυξάνεται το βάθος τους. Επομένως, ιδανικοί υποψήφιοι για τη μέθοδο bagging αποτελούν τα μεγάλα δέντρα χωρίς κλάδεμα. Η λογική της τεχνικής bootstrap aggregation είναι η εξής:

1. Κατασκευή B δειγμάτων με τη μέθοδο επαναδειγματοληψίας Bootstrap.
2. Κατασκευή B μοντέλων f_b ίδιου τύπου (μεγάλα δένδρα απόφασης κατά κύριο λόγο).
3. Ταξινόμηση των παρατηρήσεων στην πλειοψηφική κατηγορία των B δημιουργημένων μοντέλων, μέσω της διαδικασίας πλειοψηφικού ψηφίσματος :

$$\hat{f}_{bagging} = arg\max_{k} \sum_{i=1}^{B} I(\hat{f}_i = k) \tag{2.18}$$

Για τον υπολογισμό του σφάλματος του μοντέλου από τον ταξινομητή Bagging αποδεικνύεται, ότι καθένα από τα B μοντέλα χρησιμοποιούν περίπου τα 2/3 των παρατηρήσεων του αρχικού δείγματος

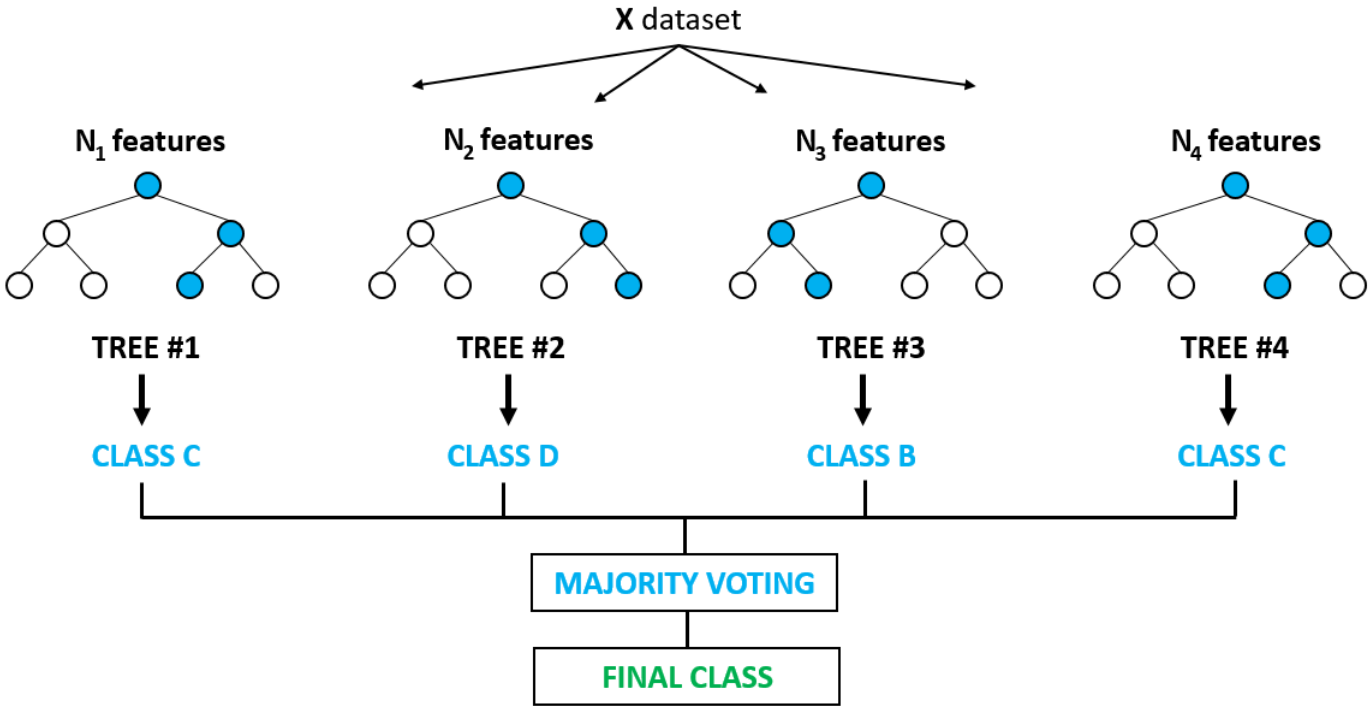
μιας και η μέθοδος Bootstrap κάνει δειγματοληψία με επανατοποθέτηση. Ακόμη, μία παρατήρηση που δεν χρησιμοποιήθηκε σε ένα από τα B μοντέλα καλείται out-of-bag παρατήρηση (OOB) για το εκάστοτε μοντέλο, ενώ υπάρχει ισχυρό ενδεχόμενο κάθε παρατήρηση να αποτελεί OOB παρατήρηση σε B/3 μοντέλα. Με τις προηγούμενες αναφορές εύκολα πλέον μπορεί να προβλεφθεί το σφάλμα λανθασμένης ταξινόμησης χρησιμοποιώντας ως πρόβλεψη για κάθε παρατήρηση την πλειοψηφία μόνο των μοντέλων στα οποία η παρατήρηση αυτή είναι OOB. Το πλεονέκτημα αυτής της διαδικασίας είναι ότι κοστίζει υπολογιστικά πολύ λιγότερο από διαδικασίες όπως η cross-validation που θα αναλυθεί μετέπειτα. Όμως, με τη διαδικασία αυτή δυστυχώς χάνονται δύο πολύ θετικά χαρακτηριστικά των δεντρικών μοντέλων. Πρώτον, χάνεται η διαισθητικότητα που χαρακτηρίζει τα δέντρα, μιας και πλέον έχουμε να εξετάσουμε μερικές εκατοντάδες δέντρα, δηλαδή δεν μπορούμε να αναπαραστήσουμε το μοντέλο μας σε μια μορφή κατανοητή από τον άνθρωπο. Δεύτερον, σε κάθε πρόβλεψη έχουμε πλέον υπολογιστικό κόστος, βέβαια σημειώνεται εδώ ότι το Bagging δεν αποτελεί μάθηση βασισμένη στη μνήμη.

Τυχαία Δάση (Random Forests)

Έχοντας ορίσει τη μέθοδο bagging είναι ευκολότερο πλέον να ορισθούν τα **τυχαία δάση**. Ένα τυχαίο δάσος (Breiman, 2001) αποτελεί και αυτό ένα συνδυαστικό ταξινομητή/κατηγοριοποιητή δημιουργημένο για να βελτιώσει ένα σύνολο μοντέλων με μεγάλο σφάλμα διακύμανσης. Ειδικότερα, τα τυχαία δάση χρησιμοποιούν δένδρα απόφασης ως δομικά στοιχεία για την κατασκευή ισχυρότερων μοντέλων πρόβλεψης και το θετικό τους είναι ότι μειώνουν τη συσχέτιση των δέντρων. Πιο συγκεκριμένα, όπως και στη μέθοδο bagging για την υλοποίηση ενός τυχαίου δάσους χτίζεται αρχικά ένα πλήθος δέντρων απόφασης για bootstraped δείγματα δεδομένων που έχουν δημιουργηθεί προηγουμένως. Όμως, πρέπει να σημειωθεί μια σημαντική διαφορά ανάμεσα στις δύο μεθόδους. Κατά την κατασκευή αυτών των δέντρων στα τυχαία δάση, κάθε φορά που πραγματοποιείται μία διάσπαση στο χώρο, ένα τυχαίο πλήθος m προβλεπτικών μεταβλητών επιλέγεται ως υποψήφιο να χωριστεί από το πλήρες σύνολο επεξηγηματικών μεταβλητών (predictors) p και η διαδικασία αφήνεται να επιλέξει μόνο μια από αυτές. Στη συνέχεια, ένα νέο δείγμα από m προβλεπτικές μεταβλητές λαμβάνεται σε κάθε νέα διάσπαση και συνήθως επιλέγεται $m \approx \sqrt{p}$, δηλαδή το πλήθος των predictors που λαμβάνονται υπόψη σε κάθε διαχωρισμό είναι περίπου ίσο με την τετραγωνική ρίζα του συνολικού αριθμού των predictors. Γενικά, η επιλογή μικρού m βοηθάει στις περιπτώσεις που το δείγμα μας έχει πολύ συσχετισμένες επεξηγηματικές μεταβλητές.

Στην οικοδόμηση, λοιπόν, ενός τυχαίου δάσους, σε κάθε διάσπαση στο δέντρο, ο αλγόριθμος δεν έχει την άδεια να εξετάσει την πλειοψηφία των διαθέσιμων predictors. Αυτό είναι μία έξυπνη τακτική του αλγορίθμου, αφού αν υποτεθεί ότι υπάρχει μία μεταβλητή με πολύ μεγάλη επεξηγηματική ισχύ στο σύνολο δεδομένων, μαζί με έναν αριθμό άλλων μέτρια ισχυρών predictors, τότε αποφεύγεται ο συνωστισμός υψηλά συσχετισμένων δέντρων που δε θα οδηγούσαν σε μείωση της διασποράς. Χαρακτηριστικά, τα τυχαία δάση επιβάλλοντας σε κάθε διαχωρισμό να λαμβάνει υπόψη μόνο ένα υποσύνολο των predictors καταφέρνουν κατά μέσο όρο, οι $\frac{(p-m)}{p}$ διαχωρισμοί να μη λαμβάνουν καν υπόψη τον ισχυρό predictor και έτσι οι υπόλοιποι πιο αδύναμοι predictors να έχουν περισσότερες

ευκαιρίες να ληφθούν κατά το διαχωρισμό οδηγώντας έτσι, σε δέντρα που έχουν πολύ διαφορετική δομή. Τέλος, το τελικό στάδιο της διαδικασίας δημιουργίας ενός τυχαίου δάσους, είναι ίδιο vm article με το στάδιο ολοκλήρωσης της τεχνικής bagging, δηλαδή εκτελείται ταξινόμηση σύμφωνα με την πλειοψηφική πρόβλεψη και ο υπολογισμός του σφάλματος λανθασμένης ταξινόμησης μέσω της διαδικασίας out-of-bag. Η διαδικασία των τυχαίων δασών συνοψίζεται στο Γράφημα 2.4. Σημαντική παρατήρηση για τις παραπάνω μεθόδους είναι ότι και στις δύο μεθόδους τα μεγάλα B σύνολα δεν υπερπροσαρμόζουν(overfit) τα δεδομένα.



Γράφημα 2.4: Δημιουργία Τυχαίου Δάσους

2.2.7 Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines)

Οι μηχανές διανυσμάτων υποστήριξης (SVM) (Cortes and Vapnik, 1995) είναι μια γενίκευση ενός απλού και διαισθητικού συστήματος ταξινομητή που ονομάζεται μέγιστος ταξινομητής περιθωρίου (maximal margin classifier). Αν και είναι κομψός και απλός, αποδεικνύεται ότι ο ταξινομητής αυτός δεν μπορεί να εφαρμοστεί στα περισσότερα σύνολα δεδομένων, αφού απαιτεί οι κλάσεις να μπορούν να διαχωριστούν από ένα γραμμικό όριο. Επίσης, είναι χρήσιμο για τη μετέπειτα περιγραφή της μεθόδου SVM να παρουσιαστεί ο ταξινομητής διανυσμάτων υποστήριξης (support vector classifier), που αποτελεί επέκταση του μέγιστου ταξινομητή περιθωρίου και μπορεί να εφαρμοστεί σε ένα ευρύτερο φάσμα περιπτώσεων. Μάλιστα, η μέθοδος SVM αποτελεί μια περαιτέρω επέκταση του ταξινομητή φορέα υποστήριξης, προκειμένου να μπορούν να αντιμετωπιστούν προβλήματα που είναι μη γραμμικά. Οι μηχανές διανυσμάτων υποστήριξης προορίζονται κατά κύριο λόγο για προβλήματα δυαδικής ταξινόμησης (υπάρχουν δύο τάξεις), όμως υπάρχουν επεκτάσεις της μεθόδου για την αντιμετώπιση περισσοτέρων από δύο κατηγορίες. Αναγκαίο είναι επίσης να υπογραμμισθεί ότι ο μέγιστος ταξινομητής περιθωρίου, ο ταξινομητής διανυσμάτων υποστήριξης και η μηχανή διανυσμάτων υποστήριξης, αν

και αναφέρονται συχνά ως "μηχανές διανυσμάτων υποστήριξης", υπόκεινται σε διάκριση, η οποία συνοψίζεται παρακάτω.

Μέγιστος Ταξινομητής Περιθωρίου (Maximal Margin Classifier)

Σε ένα χώρο p διαστάσεων, ένα υπερεπίπεδο είναι ένας επίπεδος συγγενής υπόχωρος διάστασης $p-1$. Για παράδειγμα, σε δύο διαστάσεις, ένα υπερεπίπεδο είναι ένας επίπεδος μονοδιάστατος υπόχωρος - με άλλα λόγια μια γραμμή. Ο μαθηματικός ορισμός ενός υπερεπιπέδου είναι αρκετά απλός. Σε p διαστάσεις, ένα υπερεπίπεδο ορίζεται από την εξίσωση

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p = 0 \tag{2.19}$$

Ειδικότερα, αν ένα σημείο $X = (X_1, X_2, ..., X_p)^T$ σε ένα p-διαστατο χώρο ικανοποιεί την (2.19), τότε το X πρόσκειται στο υπερεπίπεδο. Τώρα, ας υποθέσουμε ότι το X δεν ικανοποιεί την (2.19) αλλά ισχύει είτε

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p > 0 \tag{2.20}$$

είτε

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p < 0 \tag{2.21}$$

Τότε το X βρίσκεται είτε στη μία πλευρά του υπερεπιπέδου είτε στην άλλη του πλευρά. Άρα το υπερεπίπεδο θα μπορούσε να ορισθεί ως ο διαχωρισμός ενός χώρου p διαστάσεων σε δύο μισά και το πρόσημο της τιμής της $f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$ να καθορίζει την πλευρά που βρίσκεται το X ως προς το υπερεπίπεδο. Επομένως, ένα υπερεπίπεδο μπορεί να χρησιμοποιηθεί ως μοντέλο ταξινόμησης με μόνο κριτήριο το πρόσημο της f

Ακόμη, έχοντας ορίσει την έννοια του υπερεπιπέδου, αξίζει να σημειωθεί, πως μέσω ενός διαχωριστικού υπερεπιπέδου (separating hyperplane) θα μπορούσαν κάλλιστα να επιλυθούν και προβλήματα ταξινόμησης/κατηγοριοποίησης. Έστω λοιπόν ένας πίνακας δεδομένων X διαστάσεων $n \times p$ που περιέχει τις n παρατηρήσεις κατάρτισης (training observations) σε ένα p-διαστατο χώρο,

$$x_1 = \begin{pmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1p} \end{pmatrix}, ..., \; x_n = \begin{pmatrix} x_{n1} \\ x_{n2} \\ \vdots \\ x_{np} \end{pmatrix} \tag{2.22}$$

και ότι αυτές οι παρατηρήσεις εμπίπτουν σε δύο κατηγορίες - δηλαδή, $y_1, ..., y_n \in \{-1, 1\}$ όπου το -1 αντιπροσωπεύει μία κατηγορία και το 1 μία άλλη. Επίσης, υπάρχουν και παρατηρήσεις δοκιμής (test observations), που αναπαριστώνται ως p-διαστατα διανύσματα παρατηρούμενων επεξηγηματικών μεταβλητών της μορφής $x^* = ({x_1}^*...{x_p}^*)^T$. Στόχος, λοιπόν είναι η δημιουργία ενός μοντέλου ταξινόμησης βασισμένο στα δεδομένα εκπαίδευσης (training data) που θα κατηγοριοποιεί σωστά τις παρατηρήσεις δοκιμής (test observations) χρησιμοποιώντας τις τιμές των επεξηγηματικών μεταβλητών της. Αυτό μπορεί λοιπόν να επιτευχθεί με τη δημιουργία ενός υπερεπιπέδου που διαχωρίζει τις παρατηρήσεις εκπαίδευσης κατάλληλα με βάση τις ετικέτες κατηγορίας στην οποία ανήκουν. Το

υπερεπίπεδο αυτό ονομάζεται, όπως προειπώθηκε, ως διαχωριστικό (separating) και ανάλογα με τις τιμές των μεταβλητών απόκρισης $y_1, ..., y_n$ ορίζονται τα εξής :

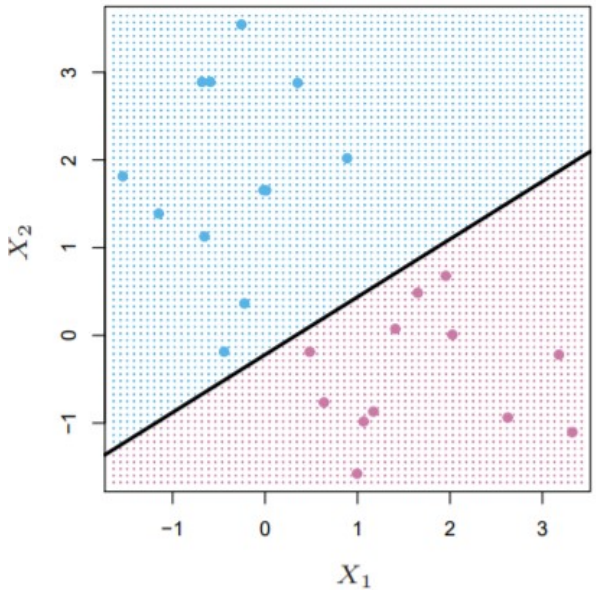
$$\beta_0 + \beta_1 X_i 1 + \beta_2 X_i 2 + ... + \beta_p X_i p > 0 \ , \ \alpha\nu \ y_i = 1 \tag{2.23}$$

είτε

$$\beta_0 + \beta_1 X_i 1 + \beta_2 X_i 2 + ... + \beta_p X_i p < 0 \ , \ \alpha\nu \ y_i = -1 \tag{2.24}$$

Ισοδύναμα, ένα διαχωριστικό υπερπίπεδο έχει την ιδιότητα

$$y_i(\beta_0 + \beta_1 X_i 1 + \beta_2 X_i 2 + ... + \beta_p X_i p) > 0 \ , \ \forall i = 1, ..., n \tag{2.25}$$

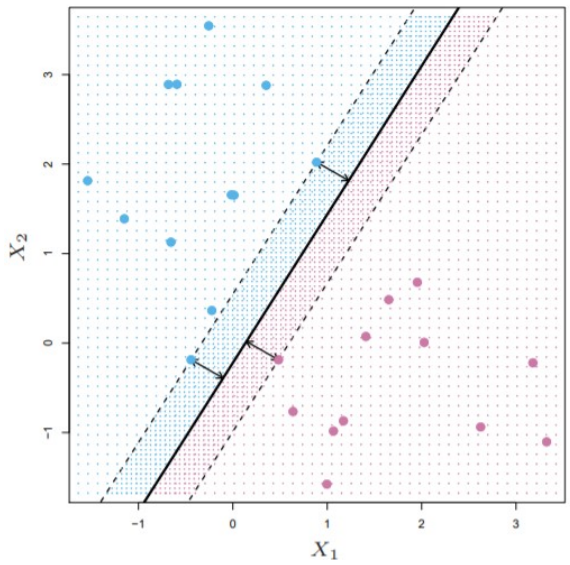


Γράφημα 2.5: Παράδειγμα Διαχωριστικού Υπερεπιπέδου με Πλέγματα Κατηγοριών

Στο Γράφημα 2.5 παρουσιάζεται ένα μοντέλο ταξινομητή (classifier) στην περίπτωση που υπάρχει ένα διαχωριστικό υπερεπίπεδο. Χαρακτηριστικά, ένα διαχωριστικό υπερεπίπεδο απεικονίζεται με μαύρο χρώμα, ενώ το μπλε και μωβ πλέγμα υποδηλώνουν τον κανόνα απόφασης (decision rule) που γίνεται από έναν ταξινομητή βασιζόμενο σε αυτό το διαχωριστικό υπερεπίπεδο. Συγκεκριμένα, μια παρατήρηση δοκιμής που βρίσκεται στο μπλε τμήμα του πλέγματος θα ανατεθεί στην μπλε κατηγορία, και μια παρατήρηση δοκιμής που πέφτει στο μωβ τμήμα του πλέγματος θα αποδοθεί στη μωβ τάξη. Ωστόσο, πρέπει να τονιστεί ότι ένας ταξινομητής, όπως αυτός του γραφήματος (2.5), που βασίζεται σε ένα διαχωριστικό υπερεπίπεδο οδηγεί σε γραμμικό φράγμα απόφαση (linear decision boundary).

Βέβαια, αν τα δεδομένα μας για ένα πρόβλημα διαχωρίζονται "τέλεια" με τη χρήση ενός υπερεπιπέδου, τότε θα υπάρχουν άπειρα τέτοια υπερεπίπεδα. Προκειμένου, λοιπόν, να κατασκευαστεί ένας ταξινομητής με βάση ένα υπερεπίπεδο διαχωρισμού, πρέπει να υπάρχει ένας λογικός τρόπος να αποφασιστεί ποιο από τα άπειρα υπερεπίπεδα θα χρησιμοποιηθεί. Ορίζοντας αρχικά, τη μικρότερη απόσταση του υπερεπιπέδου από τις παρατηρήσεις εκπαίδευσης ως περιθώριο M, η πιο διασθητική

επιλογή υπερεπιπέδου, είναι το διαχωριστικό εκείνο επίπεδο που απέχει περισσότερο από τις παρατηρήσεις εκπαίδευσης (training observations). Αυτό το υπερεπίπεδο, που απεικονίζεται στο Γράφημα 2.6 καλείται υπερεπίπεδο μέγιστου περιθωρίου (maximal margin hyperplane) και ανάλογα με την πλευρά του υπερεπιπέδου που εμπίπτει μία παρατήρηση δοκιμής μπορεί αυτή να κατηγοριοποιηθεί κατάλληλα. Η λειτουργία αυτή αποτελεί δυνατότητα του ταξινομητή μέγιστου περιθωρίου (maximal margin classifier). Επισημαίνεται όμως, ότι αν και ο ταξινομητής μέγιστου περιθωρίου συχνά κατηγοριοποιεί τις παρατηρήσεις δοκιμής επιτυχώς, μπορεί σε περιπτώσεις μεγάλου p (πλήθος επεξηγηματικών μεταβλητών) να οδηγήσει σε υπερφόρτωση/υπερπροσαρμογή (overfitting). Φυσικά για να πετύχουμε το μέγιστο περιθώριο θα πρέπει το υπερεπίπεδο να απέχει απόσταση M από τουλάχιστον δύο σημεία, ένα από κάθε κλάση. Τα σημεία που απέχουν απόσταση M από το υπερεπίπεδο καλούνται διανύσματα υποστήριξης (support vectors) μιας και είναι p-διαστατα διανύσματα και είναι τα μόνα που επηρεάζουν (υποστηρίζουν) το υπερεπίπεδο.



Γράφημα 2.6: Το υπερεπίπεδο μέγιστου περιθωρίου, σε ένα γραμμικώς διαχωρίσιμο δείγμα, δημιουργημένο από τον ταξινομητή μέγιστου περιθωρίου μέσω των τριών διανυσμάτων υποστήριξης

Η εύρεση ενός υπερεπιπέδου μέγιστου περιθωρίου, με βάση ένα σύνολο n παρατηρήσεων κατάρτισης $x_1, ..., x_n \in \mathbb{R}^p$ και σχετιζόμενες κατηγορικές $y_1, ..., y_n \in \{-1, 1\}$, αποτελεί τη λύση του εξής προβλήματος βελτιστοποίησης:

$$\underset{\beta_0,...,\beta_p}{\text{Μεγιστοποίηση}} \quad \text{περιθωρίου} \quad M \tag{2.26}$$

$$\text{Πρέπει να υπακούει στο άθροισμα} \sum_{j=1}^{p} {\beta_j}^2 = 1 \tag{2.27}$$

$$y_i(\beta_0 + \beta_1 X_i 1 + \beta_2 X_i 2 + ... + \beta_p X_i p) \geq M \; \forall \; i = 1, ..., n \tag{2.28}$$

Στο παραπάνω πρόβλημα βελτιστοποίησης, η σχέση (2.28) εξασφαλίζει ότι κάθε παρατήρηση θα είναι στη σωστή πλευρά του υπερεπιπέδου, δεδομένου ότι το M είναι θετικό. Ακόμη, η σχέση (2.27) ενισχύει την ιδιότητα της (2.28), αφού μπορεί να αποδειχθεί ότι με τον περιορισμό αυτό η κάθετη απόσταση από την i-στη παρατήρηση ως προς το υπερεπίπεδο δίνεται από την ποσότητα $y_i(\beta_0 + \beta_1 X_i 1 + \beta_2 X_i 2 + ... + \beta_p X_i p)$.

Ο μέγιστος ταξινομητής περιθωρίου αποτελεί αδιαμφισβήτητα έναν πολύ φυσικό τρόπο για να εκτελέσει κάποιος ταξινόμηση/κατηγοριοποίηση (classification), εάν υπάρχει ένα διαχωριστικό υπερεπίπεδο. Ωστόσο, σε πολλές περιπτώσεις υπερεπίπεδα τέτοιου είδους δεν υπάρχουν, άρα ούτε και ταξινομητής μέγιστου περιθωρίου. Σε αυτή την περίπτωση, λοιπόν το πρόβλημα βελτιστοποίησης (2.26) - (2.28) δεν έχει λύση με $M > 0$. Όμως, η ιδέα ενός διαχωριστικού υπερεπιπέδου μπορεί να επεκταθεί προκειμένου να αναπτυχθεί ένα υπερεπίπεδο που σχεδόν διαχωρίζει τις τάξεις της μεταβλητής απόκρισης, χρησιμοποιώντας το λεγόμενο μαλακό περιθώριο (soft margin). Η γενίκευση αυτή, του μέγιστου ταξινομητή περιθωρίου στην περίπτωση μη ύπαρξης διαχωριστικού επιπέδου είναι γνωστή ως ταξινομητής διανυσμάτων υποστήριξης (support vector classifier) και αναλύεται παρακάτω.

Ταξινομητής Διανυσμάτων Υποστήριξης (Support Vector Classifier)

Υπάρχουν λοιπόν, περιπτώσεις παρατηρήσεων που ταξινομούνται σε δύο κατηγορίες, αλλά δεν μπορούν να διαχωριστούν πλήρως από ένα υπερεπίπεδο. Μάλιστα, ακόμα κι αν υπάρχει ένα υπερεπίπεδο, είναι πιθανόν να υπάρχουν παρατηρήσεις για τις οποίες ένας ταξινομητής, βασισμένος στο υπερεπίπεδο αυτό να μην είναι ο κατάλληλος για την κατηγοριοποίηση των παρατηρήσεων. Συγκεκριμένα, ένας ταξινομητής βασισμένος σε ένα διαχωριστικό υπερεπίπεδο, αναγκαία θα κατηγοριοποιεί πλήρως όλες τις παρατηρήσεις που θα χρησιμοποιεί για την εκπαίδευση στατιστικών μοντέλων, γεγονός που τον καθιστά ευάλωτο σε μεμονωμένες παρατηρήσεις. Χαρακτηριστικά, εάν υπάρχει μία αποκλίνουσα παρατήρηση (outlier) στα δεδομένα, αυτή μπορεί να αλλάξει τον τρόπο καθορισμού των περιθωρίων που χρησιμοποιούνται για τον ταξινομητή διαχωριστικού υπερεπιπέδου, κάτι που πιθανώς να οδηγήσει σε εσφαλμένη ταξινόμηση παρατηρήσεων δοκιμής. Έτσι, είναι χρήσιμη η κατασκευή ενός ταξινομητή βασισμένου σε υπερεπίπεδο που δεν διαχωρίζει πλήρως τις δύο κατηγορίες των δεδομένων, αλλά ταξινομεί εσφαλμένα μερικές παρατηρήσεις εκπαίδευσης προκειμένου να γίνει καλύτερη δουλειά στην ταξινόμηση των υπόλοιπων παρατηρήσεων και να επιτευχθεί μεγαλύτερη ευρωστία (robustness) στις μεμονωμένες παρατηρήσεις. Ο **ταξινομητής διανυσμάτων υποστήριξης (support vector classifier)** λοιπόν, αποτελεί μία τέτοια περίπτωση, καθώς ταξινομεί μια παρατήρηση δοκιμής ανάλογα με την πλευρά του υπερεπιπέδου που βρίσκεται, ενώ επιτρέπεται σε κάποιες παρατηρήσεις εκπαίδευσης να είναι στην λανθασμένη πλευρά του περιθωρίου, ή ακόμα και στην εσφαλμένη πλευρά του υπερεπιπέδου. Ένα υπερεπίπεδο στο οποίο βασίζεται ένας τέτοιος ταξινομητής, αποτελεί λύση του παρακάτω προβλήματος βελτιστοποίησης:

$$\underset{\beta_0,\ldots,\beta_p,\varepsilon_1,\ldots,\varepsilon_n}{\text{Μεγιστοποίηση περιθωρίου}} \quad \text{M} \tag{2.29}$$

$$\text{Πρέπει να υπακούει στο άθροισμα } \sum_{j=1}^{p} {\beta_j}^2 = 1 \tag{2.30}$$

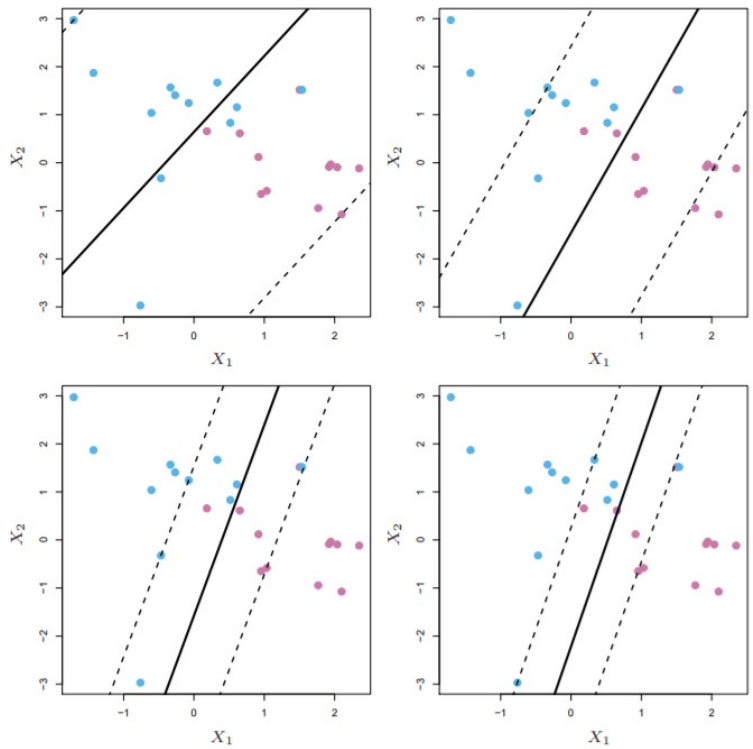
$$y_i(\beta_0 + \beta_1 X_i 1 + \beta_2 X_i 2 + \ldots + \beta_p X_i p) \geq M(1 - \varepsilon_i), \tag{2.31}$$

$$\varepsilon_i \geq 0,\ \sum_{i=1}^{n} \varepsilon_i \leq C, \tag{2.32}$$

με το C να αποτελεί μία παράμετρο συντονισμού, M το γνωστό πλάτος περιθωρίου, ενώ στη σχέση (2.31) οι $\varepsilon_1, \ldots, \varepsilon_n$ είναι μεταβλητές χαλάρωσης (slack variables) που επιτρέπουν σε μεμονωμένες παρατηρήσεις να είναι στην λανθασμένη πλευρά του περιθωρίου ή του υπερεπιπέδου.

Το πρόβλημα (2.29) - (2.32) φαίνεται περίπλοκο, αλλά μέσω μιας σειράς απλών παρατηρήσεων, που γίνονται παρακάτω, η κατανόησή του καθίσταται εύκολη. Αρχικά, η "χαλαρή" μεταβλητή ε_i περιγράφει που εντοπίζεται η i-στη παρατήρηση, σε σχέση με το υπερεπίπεδο και σε σχέση με το περιθώριο, δηλαδή για $\varepsilon_i = 0$ η παρατήρηση i βρίσκεται στη σωστή πλευρά του περιθωρίου. Αντίστοιχα, αν ισχύει $\varepsilon_i > 0$ τότε η i-στη παρατήρηση παραβιάζει το περιθώριο, δηλαδή βρίσκεται στη λάθος πλευρά του, ενώ για $\varepsilon_i > 1$ τότε βρίσκεται στη λάθος πλευρά του υπερεπιπέδου. Από την άλλη μεριά, η παράμετρος συντονισμού C φράσσει το άθροισμα των ε_i καθορίζοντας έτσι τον αριθμό και τη σοβαρότητα των παραβιάσεων στο περιθώριο (και στο υπερεπίπεδο) που είναι ανεκτές για το μοντέλο. Στη περίπτωση που ισχύει $C = 0$, τότε δεν μπορούν να υπάρξουν παραβιάσεις στο περιθώριο άρα το παραπάνω πρόβλημα συμπίπτει με το πρόβλημα βελτιστοποίησης του υπερεπιπέδου μέγιστου περιθωρίου.

Στην πράξη, το C αντιμετωπίζεται ως παράμετρος συντονισμού, που επιλέγεται γενικά μέσω διασταυρωμένης επικύρωσης (cross validation) και ελέγχει το αντιστάθμισμα μεροληψίας-διασποράς αυτής της τεχνικής στατιστικής μάθησης. Όταν για παράδειγμα το C είναι μικρό, αναζητούνται στενά περιθώρια που σπάνια παραβιάζονται, το οποίο ισοδυναμεί με έναν ταξινομητή πολύ κατάλληλο για τα δεδομένα, με πιθανώς χαμηλή μεροληψία αλλά μεγάλη διακύμανση. Από την άλλη, όταν το C είναι μεγαλύτερο, το περιθώριο είναι ευρύτερο και επιτρέπονται περισσότερες παραβιάσεις σε αυτό, έχοντας έτσι ευκολότερη προσαρμογή των δεδομένων με πιθανώς μικρότερη διασπορά και την απόκτηση ενός ταξινομητή που είναι δυνητικά πιο μεροληπτικός (biased).



Γράφημα 2.7: Ταξινομητές διανυσμάτων υποστήριξης για διαφορετικές τιμές της παραμέτρου συντονισμού C

Στην περιγραφή του παραπάνω προβλήματος βελτιστοποίησης πρέπει να προστεθεί και η ιδιότητα του: μία παρατήρηση που βρίσκεται αυστηρά στη σωστή πλευρά του περιθωρίου δεν καθίσταται δυνατή να επηρεάσει τον ταξινομητή διανυσμάτων υποστήριξης. Αντίθετα, τονίζεται, πως παρατηρήσεις που

βρίσκονται πάνω ακριβώς στο περιθώριο ή στη λάθος πλευρά του, σύμφωνα με την κλάση τους, επηρεάζουν τον ταξινομητή διανυσμάτων υποστήριξης και είναι γνωστές ως διανύσματα υποστήριξης (support vectors). Τέλος, το γεγονός ότι μόνο τα διανύσματα υποστήριξης επηρεάζουν τον ταξινομητή συνάδει με την προαναφερθείσα λειτουργία του C. Τα παραπάνω απεικονίζονται στο Γράφημα 2.7 όπου ένας ταξινομητής διανυσμάτων υποστήριξης προσαρμόστηκε χρησιμοποιώντας τέσσερις διαφορετικές τιμές της παραμέτρου C, με τη μεγαλύτερη τιμή της παραμέτρου συντονισμού να έχει χρησιμοποιηθεί στο πάνω αριστερό επιμέρους γράφημα ενώ οι χαμηλότερες τιμές χρησιμοποιήθηκαν στην κάτω δεξιά γραφική παράσταση. Όταν το C είναι μεγάλο, τότε υπάρχει μεγάλη ανοχή για τις παρατηρήσεις σε λάθος πλευρά του περιθωρίου, και έτσι το περιθώριο θα είναι μεγάλο. Πλέον, και γραφικά διαπιστώνεται πως, καθώς μειώνεται η τιμή C, η ανοχή για παρατηρήσεις που βρίσκονται σε λάθος πλευρά του περιθωρίου μειώνεται και το περιθώριο στενεύει.

Μηχανή Διανυσμάτων υποστήριξης (Support Vector Machine)

Στην περίπτωση, βέβαια που απαιτείται αντιμετώπιση προβλήματος μη γραμμικής ταξινόμησης όσα έχουν αναφερθεί παραπάνω δεν επαρκούν. Μόνη εξαίρεση αποτελεί η περίπτωση του ταξινομητή διανυσμάτων υποστήριξης, μεγαλώνοντας όμως τον χώρο επεξηγηματικών μεταβλητών με τη χρήση τετραγωνικών, κυβικών ή και μεγαλύτερης τάξης πολυωνυμικών συναρτήσεων των προβλεπτικών αυτών μεταβλητών. Αν και υπάρχουν πολλοί τρόποι για να διευρυνθεί ο χώρος των επεξηγηματικών μεταβλητών υπάρχει κίνδυνος δημιουργίας τεράστιου αριθμού μεταβλητών οδηγώντας έτσι σε μη διαχειρίσιμους υπολογισμούς. Για αυτό το λόγο, προτιμώνται οι **μηχανές διανυσμάτων υποστήριξης**, που αποτελούν επέκταση των ταξινομητών διανυσμάτων υποστήριξης, αφού διευρύνουν τον χώρο των προβλεπτικών μεταβλητών με τη μέθοδο των πυρήνων (kernels), έτσι ώστε οι υπολογισμοί που θα προκύψουν να είναι διαχειρίσιμοι.

Πιο συγκεκριμένα, αποδεικνύεται ότι:

- Εκφράζοντας το πρόβλημα του γραμμικού ταξινομητή διανυσμάτων υποστήριξης με τη βοήθεια του εσωτερικού γινομένου δύο παρατηρήσεων $x_i, x_{i'}$, δηλαδή $\langle x_i, x_{i'} \rangle = \sum_{j=1}^{p} x_{ij} x_{i'j}$ προκύπτει η εξής σχέση:

$$f(x) = \beta_o + \sum_{i=1}^{n} a_i \langle x, x_i \rangle \tag{2.33}$$

  όπου υπάρχουν n παράμετροι $\alpha_i, i = 1, ..., n$, ανά παρατήρηση εκπαίδευσης

- Ο υπολογισμός, των παραμέτρων $\alpha_1, ..., \alpha_n$ και β_0, χρειάζονται $\binom{n}{2}$ υπολογισμοί εσωτερικών γινομένων για κάθε ζεύγος παρατηρήσεων εκπαίδευσης.

Αντικαθιστώντας τώρα, το εσωτερικό γινόμενο της σχέσης (2.33) με τη γενίκευσή του $K(x_i, x_{i'})$ που αποτελεί τη συνάρτηση πυρήνα μπορούν, να προκύψουν πολλά ενδιαφέροντα αποτελέσματα. Μία συνάρτηση πυρήνα ποσοτικοποιεί την ομοιότητα δύο παρατηρήσεων ενώ ανάλογα με τη μορφή της τα αποτελέσματα ποικίλουν. Για παράδειγμα έχουμε τις εξής περιπτώσεις:

$$K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x_{i'j} \text{ Γραμμικός Πυρήνας} \tag{2.34}$$

$$K(x_i, x_{i'}) = (1 + \sum_{j=1}^{p} x_{ij} x_{i'j})^d \text{ Πολυωνυμικός Πυρήνας} \tag{2.35}$$

$$K(x_i, x_{i'}) = \exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2) \text{ Ακτινικός Πυρήνας} \tag{2.36}$$

Η σχέση (2.34) λοιπόν, με αντικατάστασή της στην (2.33) δίνει έναν ταξινομητή διανυσμάτων υποστήριξης, ενώ καλείται γραμμικός πυρήνας επειδή ο ταξινομητής αυτός είναι γραμμικός ως προς τις επεξηγηματικές μεταβλητές ενός προβλήματος. Ο γραμμικός πυρήνας ουσιαστικά ποσοτικοποιεί την ομοιότητα ενός ζεύγους παρατηρήσεων χρησιμοποιώντας (τυπική) συσχέτιση Pearson. Από την άλλη, αν επιλεγόταν η σχέση (2.35) να αντικαταστήσει το εσωτερικό γινόμενο στη (2.33), τότε θα είχε χρησιμοποιηθεί πολυωνυμικός πυρήνας βαθμού d, με το d θετικό. Η περίπτωση αυτή, ουσιαστικά ισοδυναμεί με το να προσαρμοστεί ένας ταξινομητής διανυσμάτων υποστήριξης σε ένα χώρο υψηλότερων διαστάσεων βασισμένο σε πολυώνυμα βαθμού d, αντί του αρχικού χώρου επεξηγηματικών μεταβλητών. Υπογραμμίζεται εδώ, ότι όταν ο ταξινομητής διανυσμάτων υποστήριξης συνδυάζεται με έναν μη γραμμικό πυρήνα όπως ο (2.35), ο προκύπτων ταξινομητής είναι γνωστός ως μηχανή διανυσμάτων υποστήριξης (SVM). Τέλος, γνωστή περίπτωση, ενός επίσης μη γραμμικού πυρήνα, αποτελεί ο ακτινικός που ορίζεται στη σχέση (2.36). Ο πυρήνας αυτός, για θετική σταθερά γ λειτουργεί με τον τρόπο που συνοψίζεται παρακάτω. Χαρακτηριστικά, αν μία δοσμένη παρατήρηση δοκιμής $x^* = (x_1{}^*...x_p{}^*)^T$ απέχει αρκετά από μία παρατήρηση εκπαίδευσης x_i, το οποίο υπολογίζεται από τη μεταξύ τους Ευκλείδεια απόσταση, τότε το άθροισμα $\sum_{j=1}^{p}(x_j{}^* - x_{ij})^2$ θα είναι μεγάλο και έτσι η τιμή της συνάρτησης ακτινικού πυρήνα θα είναι πολύ μικρή. Αυτό, σημαίνει ότι οι παρατηρήσεις εκπαίδευσης μακριά από την παρατήρηση δοκιμής X^* δε θα έχουν ιδιαίτερη σημασία για την προβλεπόμενη κλάση της X^*.

Τονίζεται, ότι η χρήση πυρήνων επιλέγεται έναντι της διεύρυνσης του χώρου των επεξηγηματικών μεταβλητών με συναρτήσεις των αρχικών αυτών χαρακτηριστικών, διότι τα πλεονεκτήματα είναι πολλά και κατά βάση υπολογιστικά. Συγκεκριμένα, πλέον απαιτείται μόνο ο υπολογισμός της συνάρτησης πυρήνα για τα $\binom{n}{2}$ διακριτά ζεύγη παρατηρήσεων εκπαίδευσης, χωρίς να χρειάζεται να εργαστεί κάποιος στο διευρυμένο χώρο επεξηγηματικών μεταβλητών.

Μηχανή Διανυσμάτων υποστήριξης (Support Vector Machine) για περισσότερες από δύο κατηγορικές κλάσεις

Τα παραπάνω περιορίζονται στην περίπτωση της δυαδικής ταξινόμησης, δηλαδή ταξινόμηση στην κατηγοριοποίηση δύο κλάσεων. Για να επεκταθούν τα SVM στην πιο γενική περίπτωση κάποιου αυθαίρετου αριθμού, $K > 2$, τάξεων μπορούν να ακολουθηθούν, δύο διαδικασίες. Η πρώτη διαδικασία γνωστή ως **ένας-εναντίον-ενός (one-versus-one)**, κατασκευάζει $\binom{K}{2}$ μηχανές διανυσμάτων υποστήριξης και κάθε μία από αυτές συγκρίνει ένα ζεύγος κλάσεων. Έπειτα, μια παρατήρηση δοκιμής ταξινομείται στην κατηγορία που πρέπει χρησιμοποιώντας κάθε έναν από τους $\binom{K}{2}$ ταξινομητές και μετρώντας τον αριθμό των φορών κατά τις οποίες η παρατήρηση της δοκιμής αντιστοιχεί σε κάθε μία από τις τάξεις K. Η τελική ταξινόμηση γίνεται με την ανάθεση στην παρατήρηση δοκιμής εκείνης της κλάσης, η οποία είχε ανατεθεί τις περισσότερες φορές στις $\binom{K}{2}$ κατηγοριοποιήσεις ζευγών.

Η δεύτερη διαδικασία, γνωστή ως **ένας-εναντίον-όλων (one-versus-all)**, για πλήθος K κατηγοριών, προσομοιώνει K SVMs, συγκρίνοντας κάθε φορά μία από τις K κλάσεις στις υπόλοιπες K-1. Θεωρώντας τώρα τις $\beta_{0k}, \beta_{1k}, ..., \beta_{pk}$ ως παραμέτρους που απορρέουν από την προσαρμογή μίας μηχανής διανυσμάτων στήριξης συγκρίνοντας την k-στη κλάση με τις υπόλοιπες και ως x^* μία παρατήρηση δοκιμής προκύπτει το επόμενο αποτέλεσμα. Η παρατήρηση δοκιμής, ανατίθεται στην κλάση εκείνη που η τιμή της σχέσης $\beta_{0k}, \beta_{1k}{x^*}_1, ..., \beta_{pk}{x^*}_p$ καθίσταται μέγιστη, καθώς αυτό ισοδυναμεί με υψηλό επίπεδο εμπιστοσύνης πως η παρατήρηση δοκιμής ανήκει στην τάξη k και όχι σε οποιαδήποτε άλλη τάξη

2.3 Αξιολόγηση Αλγορίθμων

Πέρα όμως από την εφαρμογή των αλγορίθμων που παρουσιάστηκαν παραπάνω, κρίθηκε αναγκαίο να αξιολογηθεί και η απόδοση μοντελοποίησης των δεδομένων που επεξεργάστηκαν. Για το λόγο αυτό χρησιμοποιήθηκε μία σειρά διαδικασιών και μετρητών, προκειμένου να ξεχωρίσουν τα πιο αποδοτικά μοντέλα.

2.3.1 Σύστημα αξιολόγησης

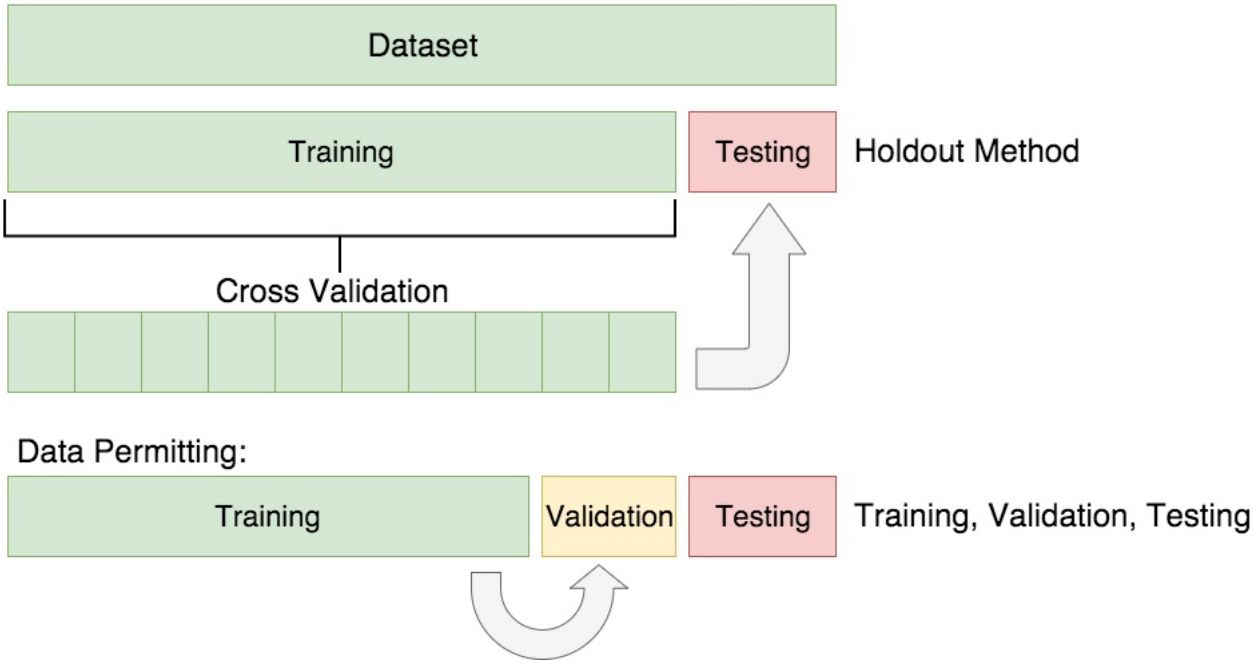
Στη διαδικασία δημιουργίας μοντέλων πρόβλεψης για ταξινόμηση η επικρατούσα πρακτική, που χρησιμοποιήθηκε και σε αυτή τη μελέτη, χωρίζει το δείγμα σε δυο υποδείγματα. Το πρώτο καλείται δείγμα εκπαίδευσης (train set) και στην παρούσα ανάλυση επιλέχθηκε να είναι το 70% του συνολικού δείγματος, που παρουσιαζόταν αρχικά σε κάθε αλγόριθμο. Το δεύτερο καλείται δείγμα ελέγχου (test set) και είναι το υπόλοιπο 30%, που παρέμεινε ως σύνολο εξέτασης του μοντέλου που είχε δημιουργηθεί μετά το πέρας της εκάστοτε διαδικασίας μάθησης.

Ο διαχωρισμός έγινε με απλή τυχαία δειγματοληψία χωρίς επανατοποθέτηση, ενώ το κάθε μοντέλο δημιουργήθηκε με τη χρήση μόνο του δείγματος εκπαίδευσης. Έτσι, είναι εφικτό να παρατηρηθεί πώς συμπεριφέρεται το μοντέλο τόσο στις παρατηρήσεις που το δημιούργησαν όσο και σε νέες παρατηρήσεις. Για την αξιολόγηση τώρα της συμπεριφοράς των παραγόμενων μοντέλων χρησιμοποιήθηκαν κάποιοι μετρητές που συνοψίζονται παρακάτω. Επίσης, επισημαίνεται ότι έχοντας γίνει απλός τυχαίος διαχωρισμός των δεδομένων, εξασφαλίστηκε για μεγάλο δείγμα ότι τα υποδείγματα που δημιουργήθηκαν προκύπτουν από την ίδια κατανομή.

Για επιπλέον αξιολόγηση, της προβλεπτικής ικανότητας, των μοντέλων που παρήχθησαν από τους διάφορους αλγορίθμους, δηλαδή της αποτελεσματικότητάς τους να κατηγοριοποιούν ορθώς παρατηρήσεις, (στις οποίες δεν έχει εκπαιδευτεί ένα μοντέλο) έγινε χρήση και της **τεχνικής διασταυρωμένης επικύρωσης k-τμημάτων (k-fold cross validation)**. Ειδικότερα, οι μέθοδοι διασταυρωμένης επικύρωσης χρησιμοποιούν μία διαδικασία κατά την οποία δημιουργούνται πάλι δύο υποδείγματα, αυτή τη φορά από το αρχικό σύνολο εκπαίδευσης που ορίστηκε και παραπάνω, όπου κάθε φορά το ένα χρησιμοποιείται για την εκπαίδευση του μοντέλου και το άλλο για την επικύρωσή του. Στην περίπτωση, τώρα, της διασταυρωμένης επικύρωσης k-τμημάτων το αρχικό train set

χωρίζεται σε k ομάδες ή τμήματα (folds) σχεδόν ίδιου μεγέθους. Το πρώτο τμήμα αντιμετωπίζεται ως δείγμα επικύρωσης και το μοντέλο προσαρμόζεται για τα υπόλοιπα k-1 τμήματα. Με τη δημιουργία του μοντέλου, το δείγμα επικύρωσης προσαρμόζεται στο παραγόμενο μοντέλο και υπολογίζεται ένας από τους μετρητές που αναλύονται παρακάτω (όπως το accuracy). Η διαδικασία, αυτή επαναλαμβάνεται k φορές και υπολογίζεται έτσι ο μέσος όρος της τιμής του επιλεγμένου μετρητή αξιολόγησης καθώς και το καλύτερο δυνατό μοντέλο. Στη συνέχεια, το δείγμα δοκιμής χρησιμοποιείται στο μοντέλο που έχει παραχθεί για να αξιολογηθεί εκ νέου ο μετρητής που χρησιμοποιήθηκε και για τον έλεγχο των εκάστοτε δειγμάτων επικύρωσης. Σημειώνεται, ότι η διαδικασία διασταυρωμένης επικύρωσης, θα μπορούσε να γίνει με τον ακριβώς ίδιο τρόπο, διαιρώντας το ολικό σύνολο δεδομένων δημιουργώντας k-1 δείγματα εκπαίδευσης και ένα δείγμα δοκιμής/επικύρωσης, γιατί στην προηγούμενη μορφή της μεθόδου επί της ουσίας δημιουργούμε ακόμα ένα δείγμα επικύρωσης. Αυτό γίνεται ώστε να αξιολογηθούν τα εκπαιδευμένα με cross validation μοντέλα για τις προβλέψεις τους σχετικά με τα δεδομένα του υποδείγματος δοκιμής, το οποίο περιέχει το 30% των συνολικών δεδομένων. Έτσι θα υπάρχει μεγαλύτερη συνάφεια και με τα υπόλοιπα αποτελέσματα.

Η παραπάνω διαδικασία περιγράφεται γραφικά στο Γράφημα 2.8.



Γράφημα 2.8: Περιγραφή μεθόδου διασταυρωμένης επικύρωσης k-τμημάτων

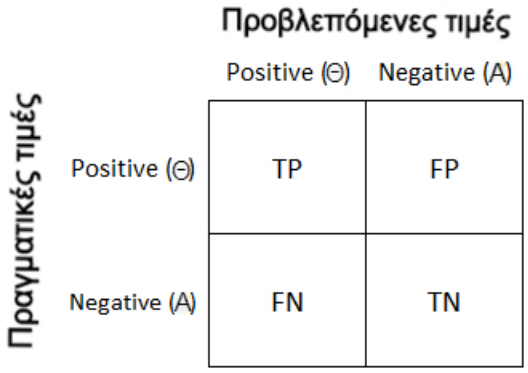
2.3.2 Παράμετροι αξιολόγησης

Δυαδική Ταξινόμηση

Εφ' όσον ορίστηκαν παραπάνω συστήματα αξιολόγησης, καθίσταται απαραίτητο να οριστούν και οι μετρητές εκείνοι που αξιοποιούνται από αυτά. Βασικός στόχος όπως θα παρουσιαστεί και στο Κεφάλαιο 4 είναι να υπάρξουν εκείνες οι κατάλληλες ποσότητες που θα ορίζουν πόσο καλά ένα μοντέλο κατηγοριοποιεί μία παρατήρηση που εξετάζει. Πριν οριστούν οι μετρητές για την αξιολόγηση των δυαδικών ταξινομήσεων είναι χρήσιμο να οριστούν οι παρακάτω μεταβλητές :

- **N**(Negative): Ο αριθμός των παρατηρήσεων που ανήκουν στην πρώτη κλάση, έστω Α.
- **P**(Positive): Ο αριθμός των παρατηρήσεων που ανήκουν στη δεύτερη δυνατή κλάση, έστω Θ.
- **TN**(True Negative): Ως αληθώς αρνητικές, ορίζονται οι παρατηρήσεις που ανήκουν στην κλάση Α και ορθώς ταξινομήθηκαν στην κλάση Α.
- **FN**(False Negative): Ως ψευδώς αρνητικές, ορίζονται οι παρατηρήσεις που ανήκουν στην κλάση Α και λανθασμένα ταξινομήθηκαν στην κλάση Θ.
- **TP**(True Positive): Ως αληθώς θετικές, ορίζονται οι παρατηρήσεις που ανήκουν στην κλάση Θ και ορθώς ταξινομήθηκαν στην κλάση Θ.
- **FP**(False Positive): Ως ψευδώς θετικές, ορίζονται οι παρατηρήσεις που ανήκουν στην κλάση Θ και λανθασμένα ταξινομήθηκαν στην κλάση Α.

Τα παραπάνω συνοψίζονται σε ένα πίνακα σύγχυσης (confusion matrix) που περιγράφεται στο Γράφημα 2.9.



Γράφημα 2.9: Πίνακας σύγχυσης (confusion matrix)

Ορίζονται τώρα οι εξής μετρητές:

- **Accuracy**: Η ακρίβεια, αποτελεί το ποσοστό σωστών προβλέψεων και ορίζεται ως, $\frac{TP+TN}{P+N}$.
- **Error Rate**: Ο δείκτης σφάλματος, αποτελεί το ποσοστό λανθασμένων προβλέψεων και ορίζεται ως, $1 - Accuracy$.
- **Precision**: Η ακρίβεια προσέγγισης, αποτελεί το ποσοστό παρατηρήσεων που ταξινομήθηκαν ως κλάσης Θ και είναι πραγματικά αυτής της κλάσης και ορίζεται ως, $\frac{TP}{TP+FP}$.
- **Sensitivity ή Recall**: Ως ευαισθησία ή ανάκληση, ορίζεται η ποσότητα $\frac{TP}{TP+FN} = \frac{TP}{P}$ και αποτελεί το ποσοστό πραγματικών παρατηρήσεων κλάσης Θ που ταξινομήθηκαν σωστά στην κλάση αυτή.

- **Specificity**: Ως ιδιαιτερότητα, ορίζεται η ποσότητα $\frac{TN}{TN+FP} = \frac{TN}{N}$ και αποτελεί το ποσοστό πραγματικών παρατηρήσεων κλάσης A που ταξινομήθηκαν σωστά στην κλάση αυτή.

- **False Negative Rate (FNR)**: Αποτελεί το ποσοστό παρατηρήσεων που ανήκουν κανονικά στην κλάση Θ και δεν κατάφεραν να εντοπιστούν από τον αλγόριθμο. Ορίζεται ως, $\frac{FN}{FN+TP} = \frac{FN}{P} = 1 - Recall$.

- **False Positive Rate (FPR)**: Αποτελεί το ποσοστό παρατηρήσεων που ανήκουν κανονικά στην κλάση A και και επισημάνθηκαν ως κλάσης Θ. Ορίζεται ως, $\frac{FP}{FP+TN} = \frac{FP}{N}$.

- **F1Score**: Αποτελεί τον αρμονικό μέσο της ακρίβειας και της ανάκλησης και ορίζεται ως $2\frac{Precission \cdot Recall}{Precission+Recall}$. Επομένως, αυτός ο μετρητής λαμβάνει υπόψη τόσο τα ψευδώς θετικά όσο και τα ψευδώς αρνητικά. Διαισθητικά δεν είναι τόσο εύκολη η κατανόηση της ακρίβειας, αλλά το F1 είναι συνήθως πιο χρήσιμο από την ακρίβεια, ειδικά στην περίπτωση μιας ανομοιογενούς κατανομής κατάταξης.

- **Kappa**: Ορίζεται ως $kappa = \frac{observedAccuracy - randomAccuracy}{1 - randomAccuracy}$ και αποτελεί έναν τρόπο μέτρησης της αξιοπιστίας της ακρίβειας ενός μοντέλου, αφού συγκρίνει την ακρίβεια του με αυτή ενός τυχαίου συστήματος.
  Ως randomAccuracy ορίζεται η ποσότητα $\frac{(TN+FP)*(TN+FN)+(TP+FN)*(TP+FP)}{Total\_Observations^2}$.

Ταξινόμηση πολλαπλών κατηγοριών

Όμοιοι μετρητές αξιολόγησης με αυτούς της δυαδικής ταξινόμησης ισχύουν και στην ταξινόμηση πολλαπλών κατηγοριών με τη διαφορά ότι εδώ αναφέρονται σε κάθε διαφορετική κατηγορία ξεχωριστά. Εξαιρείται ο μετρητής της ακρίβειας, η οποία δείχνει την επίδοση του μοντέλου ανεξάρτητα από την κατηγορία.

Στην παρούσα μελέτη έγινε κατά κύριο λόγο χρήση του μετρητή ακριβείας. Βέβαια, το πρόβλημα σε αυτή την πρακτική είναι ότι, στην περίπτωση που οι κλάσεις δεν είναι ισορροπημένες ως προς τον αριθμό των παρατηρήσεων, όπως παρατηρήθηκε και στο σύνολο δεδομένων που επεξεργάστηκε η παρούσα μελέτη, η ακρίβεια μπορεί να γίνει τελείως παραπλανητική. Για αυτό το λόγο έπρεπε να εξεταστούν όλα τα παραπάνω. Δεν αρκεί δηλαδή η υπεροχή ενός μοντέλου στα κριτήρια που ορίστηκαν έναντι κάποιου άλλου με χειρότερες επιδόσεις.

Κεφάλαιο 3

Αρχιτεκτονική Συστήματος

Στα πλαίσια αυτής της διπλωματικής, όπως έχει προαναφερθεί εκπαιδεύονται και αξιολογούνται μοντέλα, μέσω διάφορων στατιστικών και αλγοριθμικών μεθόδων μηχανικής μάθησης, τα οποία χρησιμοποιούνται για την πρόβλεψη της ικανότητας χημικών μορίων να αναστείλουν την πρόσδεση ισταμίνης στον υποδοχέα ισταμίνης H1, βάσει των διαφόρων φυσικοχημικών δεικτών που τα χαρακτηρίζουν. Το Γράφημα 3.1 παρουσιάζει την αρχιτεκτονική του συστήματος για την επίτευξη των προαναφερθέντων.

Γράφημα 3.1: Διάγραμμα Ροής του Συστήματος της Πειραματικής Μελέτης

3.1 Συλλογή Δεδομένων

Τα δεδομένα προέρχονται από την βάση δεδομένων ChEMBL(Gaulton et al., 2016). Η ChEMBL αποτελεί μια βάση δεδομένων βιοδραστικών μικρών μορίων (παρόμοιων με φάρμακα), και περιέχει τις δισδιάστατες δομές τους, υπολογισμένες ιδιότητες (π.χ. logP, μοριακό βάρος, παράμετροι Lipinski κ.λπ.) και αφηρημένες βιοδραστικές ιδιότητες (π.χ., σταθερές δέσμευσης, φαρμακολογία). Τα δεδομένα συλλέγονται από πληθώρα εργαστηρίων και βιβλιογραφικών αναφορών, ενώ η οργάνωσή τους πραγματοποιείται σύμφωνα με τη βασική επιστημονική βιβλιογραφία και καλύπτουν ένα σημαντικό μέρος του τομέα Σχέσης Δομής-Δραστηριότητας (SAR) (structure-activity relationship) και της ανακάλυψης σύγχρονων φαρμάκων. Ακολουθώντας τις τρέχουσες εξελίξεις, στο σύστημα της ChEMBL πραγματοποιείται ενσωμάτωση πρόσθετων δεδομένων σχετικά με την κλινική πρόοδο των διαφόρων ενώσεων.

Βέβαια η βάση δεδομένων της CheMBL περιέχει πολλές επιμέρους βάσεις δεδομένων ανάλογα με τους στόχους για τους οποίους χρησιμοποιούνται τα διάφορα χημικά μόρια. Η παρούσα μελέτη επεξεργάζεται χημικά μόρια που προέρχονται από τη βάση δεδομένων ChEMBL231, η οποία περιέχει πολλαπλές μετρήσεις από διαφορετικά εργαστήρια για μεγάλο πλήθος από χημικά μόρια, που έχουν ως στόχο την ουσία ισταμίνη και συγκεκριμένα προσπαθούν να αναστείλουν την πρόσδεσή της στον υποδοχέα H1. Για κάθε τέτοιο χημικό μόριο έχει μετρηθεί μεγάλο πλήθος παραμέτρων που αποτελούν φυσικοχημικούς δείκτες και παρουσιάζονται παρακάτω.

Σημειώνεται, ότι στη βάση δεδομένων υπάρχει για κάθε μόριο και μία σειρά από αποτυπώματα (fingerprints), τα οποία όντας δυαδικές αναπαραστάσεις της δομής του κάθε μορίου, δεν συγκαταλέγονται στους προαναφερθέντες φυσικοχημικούς δείκτες γι' αυτό και δεν αξιοποιήθηκαν στην παρούσα μελέτη. Όμως, είναι χρήσιμο να αναφερθεί πως αρκετά σύγχρονα εργαλεία μηχανικής μάθησης χρησιμοποιούνται για την ανακάλυψη μοντέλων και μοτίβων με στόχο την επαναστόχευση φαρμάκων ή και την από το μηδέν ανακάλυψη νέων φαρμάκων.

Εξάγοντας λοιπόν τα στοιχεία που περιέχονται στη βάση δεδομένων ChEMBL231 σε ένα αρχείο μορφής πίνακα, είναι δυνατή η αξιοποίηση του για να εποπτευθούν οι πληροφορίες των ποικίλων χημικών μορίων. Χαρακτηριστικά, με τη βοήθεια του στατιστικού πακέτου R και εκτελώντας την εντολή **read_excel()**, όπως παρουσιάζεται στο Πλαίσιο Α.1 (βλ. Παράρτημα Α), εισάγονται τα δεδομένα που θα χρησιμοποιήθούν στο περιβάλλον της R για επεξεργασία. Επίσης, λόγω της αναφοράς στο Πλαίσιο Α.1 επισημαίνεται, ότι όλα τα πλαίσια που περιέχουν τους κώδικες στους οποιούς αναφέρονται οι ενότητες του Κεφαλαίου 3 βρίσκονται στο Παράρτημα Α.

Έπειτα, με την εντολή **str()**, δίδεται μία αναλυτικότερη εικόνα των δεδομένων, η οποία παρουσιάζει ενδιαφέροντα αποτελέσματα που συνοψίζονται στο Πλαίσιο Α.2. Συγκεκριμένα, το αρχικό αρχείο έχει 16385 παρατηρήσεις χημικών μορίων και το καθένα περιγράφεται από 41 μεταβλητές. Οι μεταβλητές αυτές αποτελούν φυσικοχημικούς δείκτες που αναλύονται στη συνέχεια. Βέβαια, σε κάθε μόριο στη βάση δεδομένων συμπεριλαμβάνεται, όπως φαίνεται, τόσο το εργαστηριακό όνομά του όσο και η "χημική" του ονομασία. Τα διάφορα στοιχεία που απεικονίζονται στο Πλαίσιο Α.2 αναλύονται στους επόμενους δύο πίνακες.

Φυσικοχημικοί Δείκτες	**Εξήγηση**
SlogP	Το partition coefficient ανάμεσα σε octanol και water είναι το (logP) . Το SlogP υπολογίζει το logP συγκεντρώνοντας τη συνεισφορά των προσβάσιμων σε διαλύτη επιφανειακών περιοχών (SASA) και του διορθωτικού συντελεστή (correction factor)
SMR	Η μοριακή διαθλαστικότητα (συμπεριλαμβανομένων των σιωπηρών υδρογόνων)
TPSA	Ο συνολικός αριθμός των πολικών ατόμων επιφάνειας σε ένα μόριο. Αποτελεί πολύ καλό περιγραφικό δείκτη για prediction της παθητικής μοριακής μεταφοράς μέσω της μεμβράνης και παρέχει καλή συσχέτιση (correlation) με πειραματικά δεδομένα μεταφοράς
AMW	Μέσο μοριακό βάρος
ExactMW	Ακριβές μοριακό βάρος
LabuteASA	Η προσπελάσιμη επιφάνεια ύδατος υπολογίζεται με ακτίνα 1,4 A για το μόριο του νερού. Μια πολυεδρική αναπαράσταση χρησιμοποιείται για κάθε άτομο κατά τον υπολογισμό της επιφάνειας.
NumRotatableBonds/ NumHBD/ NumHBA/ NumAmideBonds/ NumLipinskiH-BA/NumLipinskiHBD	Οι δεσμοί υδρογόνου (hydrogen bonds) εμφανίζονται όταν ένα άτομο "δότη" δωρίζει το ομοιοπολικά συνδεδεμένο άτομο υδρογόνου, του σε ένα ηλεκτροαρνητικό "δέκτη" άτομο. Ένας αμιδικός δεσμός (amide bond) σχηματίζεται όταν, η καρβοξυλομάδα ενός αμινοξέος, συνδέεται με την αμινομάδα άλλου, για να σχηματίσει ένα πεπτίδιο HBD = Hydrogen bond donor (είναι ο αριθμός των ατόμων υδρογόνου που συνδέονται με άτομα - αζώτου και οξυγόνου) HBA = Hydrogen bond acceptor - Δεσμός Αποδοχής Υδρογόνου
NumHetero/ NumHeavy/ Num - Atoms	Αριθμός Heteroatoms/ Heavy ατόμων/ ατόμων (γενικά) ενός molecule.
Num/ NumAromat-ic/ NumSaturated/ NumAliphatic - Rings	Αριθμός συνολικών δαχτυλιδιών/ αρωματικών/ κορεσμένων/ αλιφατικών δακτυλίων ενός μορίου.
NumAromatic/ NumSaturated/ NumAliphatic - Heterocycles	Αριθμός αρωματικών/ κορεσμένων/ αλιφατικών Heterocycle. Heterocycle : Μία ετεροκυκλική ένωση ή δομή δακτυλίου, είναι μία κυκλική ένωση, η οποία έχει άτομα τουλάχιστον δύο διαφορετικών στοιχείων, ως μέλη του δακτυλίου (των δακτυλίων).
NumAromatic/ NumSaturated/ NumAliphatic - Carbocycles	Αριθμός αρωματικών/ κορεσμένων/ αλιφατικών κύκλων άνθρακα ενός μορίου.
Kier-Hall Connectivity and Kappa Shape Indices	Για ένα βαρύ άτομο i έστω $v_i = \frac{(p_i - h_i)}{(Z_i - p_i - 1)}$ όπου p_i είναι ο αριθμός των ηλεκτρονίων Z_i και p_i valence (χημική χωρητικότητα-σθένος) του ατόμου i. Οι δείκτες συνδεσιμότητας Kier και Hall Chi υπολογίζονται από τον βαθμό βαρέων ατόμων d_i (αριθμός βαρέων γειτόνων) και v_i. Οι δείκτες μοριακού σχήματος Kier και Hall Kappa (Hall and Kier, 1991) συγκρίνουν τη μοριακή γραφική παράσταση με ελάχιστες και μέγιστες μοριακές διαβαθμίσεις και προορίζονται να συλλάβουν διαφορετικές πλευρές μοριακού σχήματος.
Chi0v/ Chi1v/ Chi2v/ Chi3v/ Chi4v	Δείκτες συνδεσιμότητας ατομικού σθένους (σειρά 0). Υπολογίζονται ως το άθροισμα του λόγου $\frac{1}{\sqrt{v_i}}$ σε όλα τα βαρέα άτομα i με $v_i > 0$. Δείκτες συνδεσιμότητας ατομικού σθένους (σειρά 1). Υπολογίζονται, ως το άθροισμα $\frac{1}{\sqrt{v_i v_j}}$ πάνω από όλους τους δεσμούς μεταξύ βαρέων ατόμων i και j όπου $i < j$. Περιγράφηκαν επίσημα στη βιβλιογραφία, από(Hall and Kier, 1991; Kier and Hall, 1977)
kappa1/ kappa2/ kappa3	First kappa shape index: $\frac{(n-1)^2}{m^2}$ Second kappa shape index: $\frac{(n-1)^2}{m^2}$ Third kappa shape index: $\frac{(n-1)(n-3)^2}{p_3^2}$για μονά n, and $\frac{(n-3)(n-2)^2}{p_3^2}$ για ζυγά n. Όπου, n ο αριθμός των ατόμων στο γράφημα καταστολής του υδρογόνου το m είναι ο αριθμός των δεσμών στο γράφημα που καταστέλλεται με υδρογόνο και το a είναι το άθροισμα του $(r_i/r_c - 1)$ όπου r είναι η ομοιοπολική ακτίνα του ατόμου i , και r_c είναι η ομοιοπολική ακτίνα ενός ατόμου άνθρακα.

Μεταβλητές	**Εξήγηση**
chemblID	Το όνομα κάθε χημικού μορίου στη βάση δεδομένων ChEMBL.
value	Αποτελεί την τιμή IC_{50} του κάθε μορίου και δηλώνει την μέση μέγιστη ανασταλτική συγκέντρωση ενός χημικού μορίου που χρειάζεται για την αναστολή της πρόσδεσης ισταμίνης στον υποδοχέα Ισταμίνης H1. Η αριθμητική τιμή αυτή, λοιπόν, δηλώνει τη μέση μέγιστη συγκέντρωση ενός χημικού μορίου σε έναν υποδοχέα που καθίσταται ικανό ή όχι να δράσει ανασταλτικά για τη δράση της ισταμίνης ή όχι.
smiles	Το απλοποιημένο σύστημα εισαγωγής γραμμών μοριακής εισόδου (SMILES), είναι μια προδιαγραφή με τη μορφή γραμμής συμβόλων, για την περιγραφή της δομής των διάφορων χημικών ειδών, χρησιμοποιώντας συμβολοσειρές ASCII. Οι συμβολοσειρές SMILES μπορούν να αξιοποιηθούν από τους περισσότερους σχεδιαστές/εκδότες μορίων, για μετατροπή σε δισδιάστατα σχέδια ή τρισδιάστατα μοντέλα μορίων.

3.2 Προεπεξεργασία Δεδομένων

Προτού ξεκινήσει η αριθμητική και περιγραφική ανάλυση των δεδομένων, καθώς και η μοντελοποίηση τους, απαιτείται να πραγματοποιηθεί μία σειρά ενεργειών, ιδιαίτερα σημαντική στον τομέα εξόρυξης δεδομένων και συστημάτων μηχανικής μάθησης. Οι ενέργειες αυτές περιγράφονται συνολικά ως διαδικασίες προεπεξεργασίας δεδομένων. Συγκεκριμένα, περιλαμβάνουν διαδικασίες για την αξιολόγηση της ποιότητας δεδομένων (έλλειψη πληροφοριών, διπλοεγγραφές πληροφοριών), τη συγκεντρωτική παρουσίαση χαρακτηριστικών, τη δειγματοληψία στοιχείων από το σύνολο δεδομένων, για την εκπαίδευση μοντέλων, καθώς και την κωδικοποίηση χαρακτηριστικών.

Πρώτα απ' όλα, καθίσταται αναγκαίο να ελεγχθεί η ύπαρξη ελλιπών τιμών στο σύνολο δεδομένων. Κάτι τέτοιο, είναι πολύ συνηθισμένο και μπορεί να έχει συμβεί κατά τη συλλογή δεδομένων ή ίσως οφείλεται σε κάποιον κανόνα επικύρωσης δεδομένων. Ανεξάρτητα, όμως, από την αιτία που το προκάλεσε, πρέπει να ληφθούν υπόψη οι τιμές που λείπουν.

Με τη χρήση κατάλληλων εντολών, όπως η **is.na()** για τη διαπίστωση ύπαρξης κενών τιμών, παρατηρείται στο Πλαίσιο Α.3, ότι, στο προς επεξεργασία σύνολο δεδομένων, δεν υπάρχουν κενές τιμές, οπότε είναι δυνατή η περαιτέρω προετοιμασία του συνόλου δεδομένων. Όμως, όπως αναφέρθηκε και κατά την περιγραφή της συλλογής των δεδομένων για κάποια μόρια υπάρχουν πολλαπλές τιμές λόγω του ότι υπάρχει μέτρηση από διαφορετικά εργαστήρια. Για αυτό το λόγο, στα μόρια με πολλαπλές τιμές υπολογίστηκε ο μέσος όρος των μετρήσεων των στοιχείων τους.

Επίσης, η αριθμητική τιμή value αποτελώντας την τιμή IC_{50} εκφράζει τη βιοδραστηριότητα ενός μορίου. Πρακτικά, αυτό σημαίνει, ότι θα μπορούσε να μεταφραστεί σε μία κατηγορική μεταβλητή, που θα εξηγεί αν κάποιο μόριο, κατά την προσπάθεια αναστολής της πρόσδεσης ισταμίνης, είναι ενεργό, ανενεργό ή και ακόμα πολύ ενεργό. Η μετατροπή αυτή γίνεται εφικτή αν χρησιμοποιηθούν τα κατάλληλα όρια τιμών στην τιμή value. Βέβαια, πληθώρα μελετών δε χρησιμοποιεί καθολικά ένα μόνο κατώφλι, αλλά υπάρχουν δύο συχνές περιπτώσεις. Στη μία περίπτωση ορίζεται ως κατώφλι η τιμή των 10μmolar για το διαχωρισμό των χημικών μορίων σε ενεργά και ανενεργά (Drakakis, Wafford, et al., 2017), ενώ στην άλλη περίπτωση, προστίθεται μία ακόμα κατηγορία που τα χημικά

μόρια με τιμή IC_{50} κάτω από 1μmolar (Zheng et al., 2008) ορίζονται ως πολύ ενεργά.

Στα πλαίσια της παρούσας μελέτης, δημιουργήθηκαν μοντέλα κατηγοριοποίησης και για τις δύο παραπάνω περιπτώσεις. Αρχικά, δημιουργήθηκε η δίτιμη κατηγορική μεταβλητή (activity) στο σύνολο δεδομένων, που θα αποτελέσει μία εν δυνάμει μεταβλητή απόκρισης και χαρακτηρίζει active ή inactive ένα χημικό μόριο. Έπειτα, εισήχθη στο σύνολο δεδομένων και μία τρίτιμη κατηγορική μεταβλητή (activity_multi) (inactive, moderately active, highly active). Τα παραπάνω έγιναν με τη βοήθεια της βιβλιοθήκης **dplyr** (Wickham et al., 2019) του στατιστικού πακέτου R και παρουσιάζονται στο Πλαίσιο Α.4.

Έτσι, μετά από τη συγχώνευση των παρατηρήσεων, καθώς και τη δημιουργία δύο νέων μεταβλητών, όπως διαφαίνεται από το Πλαίσιο Α.4, το σύνολο δεδομένων αποτελείται πλέον από 44 μεταβλητές και 1237 παρατηρήσεις. Ωστόσο, αφού σκοπεύεται να δημιουργηθούν μοντέλα ταξινόμησης, η τιμή value καθίσταται περιττή, αφού ο ρόλος της μεταβλητής απόκρισης για τους ταξινομητές θα συμπληρωθεί από τις μεταβλητές activity και activity_multi. Συνεπώς, αφαιρέθηκαν από το χώρο χαρακτηριστικών η μεταβλητή value και οι μεταβλητές χαρακτήρων chemblId και smiles που δεν περιέχουν καμία πληροφορία για την πρόβλεψη της βιοδραστηριότητας τον μορίων. Οι παρεμβάσεις αυτές οδήγησαν στη μείωση των μεταβλητών σε 41, ενώ η όλη διαδικασία υλοποιείται ιδιαίτερα εύκολα και περιγράφεται στο Πλαίσιο Α.5.

Πρίν αξιοποιηθεί από τους αλγορίθμους μάθησης για την εκπαίδευση μοντέλων το ανανεωμένο σύνολο δεδομένων θα πραγματοποιηθεί μια περαιτέρω επεξεργασία του, καθώς και περιγραφική του ανάλυση. Βέβαια, όπως αναφέρθηκε και στην γραφική αναπαράσταση της αρχιτεκτονικής του συστήματος, για την δημιουργία ταξινομητών και την πραγματοποίηση προβλέψεων, χρειάζεται να δημιουργηθεί ένα υποδείγμα δεδομένων εκπαίδευσης για τη μοντελοποίηση των παρατηρήσεων, καθώς και ένα υποδείγμα δοκιμής, για την πραγματοποίηση προβλέψεων και αξιολόγησης των μοντέλων.

Η διάσπαση του συνόλου δεδομένων σε δύο επιμέρους σύνολα, κρίνεται αναγκαία, πριν επεξεργαστούν τα χαρακτηριστικά των δεδομένων που θα αποτελέσουν επεξηγηματικές μεταβλητές, έτσι ώστε να αποφευχθεί η πιθανότητα μεροληψίας στο δείγμα δοκιμής. Κάτι τέτοιο θα μπορούσε να οδηγήσει σε παραπλανητικά αποτελέσματα κατά την αξιολόγηση των μοντέλων. Τα δείγματα, λοιπόν, που δημιουργήθηκαν με τυχαία διαμέριση (random splitting) των παρατηρήσεων, είναι ένα υποδείγμα εκπαίδευσης, που περιέχει το 70% των συνολικών παρατηρήσεων και ένα υποδείγμα δοκιμής, με το 30% των παρατηρήσεων του συνόλου δεδομένων.

Σημειώνεται, πως λόγω της εκπαίδευσης μοντέλων, που θα προβλέπουν μία δίτιμη κατηγορική μεταβλητή απόκρισης, αλλά και μία τρίτιμη κατηγορική μεταβλητή, η παραπάνω διαδικασία πραγματοποιήθηκε δύο φορές. Για την υλοποίηση των όσων αναφέρθηκαν, έγινε χρήση του πακέτου **dataPreparation** (Toulemonde, 2020) και η κωδικοποίηση με βάση αυτό παρουσιάζεται στο Πλαίσιο Α.6.

3.3 Ανάλυση Δεδομένων

Παρακάτω γίνεται μία προσπάθεια να κατανοηθούν τα δεδομένα για να αξιοποιηθούν όσο το δυνατόν πιο αποτελεσματικά.

3.3.1 Φιλτράρισμα Περιττών Μεταβλητών

Το πακέτο **dataPreparation** προσφέρει μία σειρά εντολών που βοηθούν στον εντοπισμό "άχρηστων" για τη μοντελοποίηση μεταβλητών. Χαρακτηριστικά στο Πλαίσιο Α.7 ελέγχθηκε η περίπτωση ύπαρξης σταθερών μεταβλητών, διπλοεγγραφών καθώς και ισοδυναμιών 1-1, για τα διάφορα χαραλτηριστικά, με αρνητικά αποτελέσματα και για του τρεις ελέγχους.

3.3.2 Περιγραφικά Στατιστικά Μεγέθη

Έχοντας προετοιμάσει τα υποδείγματα εκπαίδευσης και δοκιμής καταλλήλως, η ομαδοποίηση των επεξηγηματικών μεταβλητών και απόκρισης κρίνεται ωφέλιμη, για την καλύτερη αριθμητική και γραφική περιγραφή των δεδομένων. Στο Πλαίσιο Α.8 πραγματοποιείται η δημιουργία ενιαίων υποδειγμάτων εκπαίδευσης και για τις δύο περιπτώσεις μεταβλητών απόκρισης, δηλαδή ένα $train\_binary$ για τη δίτιμη περιγραφή της βιοδραστηριότητας των μορίων και ένα $train\_triple$ για την τρίτιμη περιγραφή.

Με τα σύνολα εκπαίδευσης ενιαία και επεξεργασμένα, στο Πλαίσιο Α.9 παρουσιάζεται μία πρώτη προσπάθεια εκτίμησης των μεταβλητών, που περιγράφουν τα χημικά μόρια, για την περίπτωση που τα τελευταία κατηγοριοποιούνται σε δύο κατηγορίες, ανάλογα με το αν η μέση μέγιστη ποσότητα για επίτευξη της ανασταλτικής τους δράσης είναι περισσότερη ή λιγότερη των 10μmolar. Παρατηρείται, πως όλες οι μεταβλητές είναι αριθμητικές, πλην μίας και συγκεκριμένα της μεταβλητής activity, που περιγράφει την βιοδραστηριότητα των χημικών μορίων. Η μεταβλητή αυτή αποτελεί την μεταβλητή απόκρισης για τους διάφορους ταξινομητές, που θα μοντελοποιηθούν, ενώ συμπεραίνεται, ότι όλες οι επεξηγηματικές μεταβλητές των μοντέλων θα είναι αριθμητικού τύπου.

Έπειτα, η ίδια αριθμητική περιγραφή των μεταβλητών, που περιγράφουν τις διάφορες παρατηρήσεις μορίων, πραγματοποιείται στο Πλαίσιο Α.10, αυτή τη φορά για χημικά μόρια, που θα μοντελοποιηθούν για πρόβλεψη τριών διαφορετικών κατηγοριών βιοδραστηριότητας. Στο $train\_triple$, λοιπόν, η μεταβλητή απόκρισης είναι η $activity\_multi$ και οι υπόλοιπες μεταβλητές είναι επεξηγηματικές. Ιδιαίτερη αίσθηση δημιουργεί το γεγονός, ότι στα υποδείγματα $train\_binary$ και $train\_triple$ οι επεξηγηματικές μεταβλητές, που περιγράφουν το ίδιο μέγεθος, εμφανίζουν ίδιες τιμές στους αριθμητικούς περιγραφικούς δείκτες, όπως η μέγιστες και ελάχιστες τιμές.

Τα Πλαίσια Α.9 και Α.10 αναφορικά με τις κατηγορικές μεταβλητές δίνουν μόνο πληροφορίες πλήθους, για το πως κατανέμονται οι διάφορες παρατηρήσεις στις δυνατές κατηγορίες που εκφράζονται στο κάθε υποδείγμα. Έτσι, για καλύτερη κατανόηση των δεδομένων στα Πλαίσια Α.11 και Α.12 παρουσιάζεται η προσπάθεια ποσοτικοποίησης της κατανομής των παρατηρήσεων στις κατηγορίες που εκφράζουν τη βιοδραστηριότητα των χημικών μορίων.

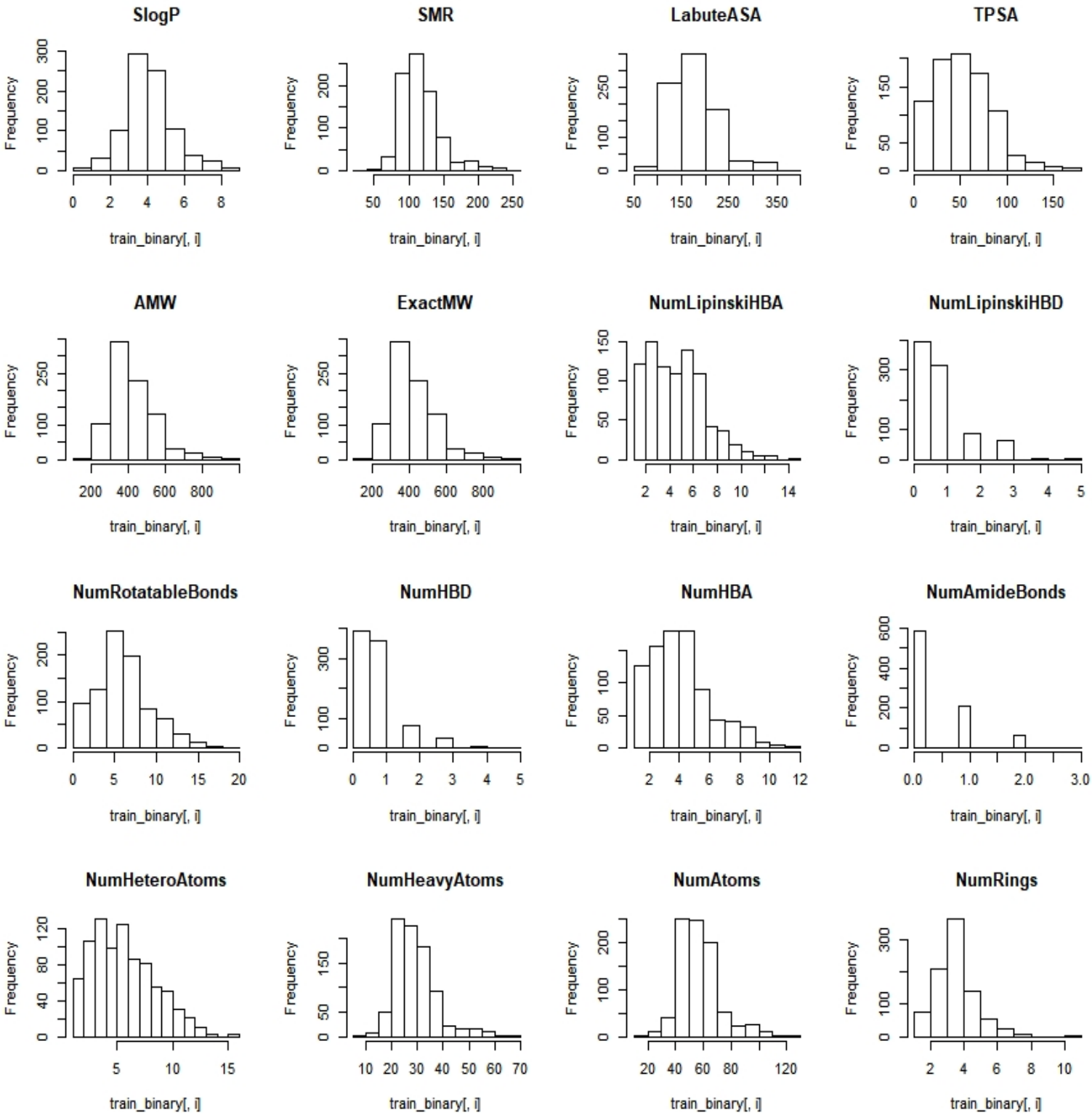
Επίσης, στο Πλαίσιο Α.11, φαίνεται ότι υπάρχει ανισορροπία στις τιμές της μεταβλητής activity. Συγκεκριμένα, υπάρχει πράγματι διαχωρισμός από 90.75% έως 9.25% για τις active - inactive τιμές της μεταβλητής activity που δεν είναι ισορροπημένη. Μάλιστα η ανισορροπία της μεταβλητής είναι αρκετά μεγάλη και θα πρέπει να μείνει ανοιχτό το ενδεχόμενο για επανεξισορρόπησή της, πριν αποφασιστεί ποιο είναι το ιδανικό καταληκτικό μοντέλο.

Για να διορθωθεί η ανισορροπία του πλήθους των κλάσεων των δεδομένων δημιουργείται ένα επιπλέον σύνολο δεδομένων για μελλοντική μελέτη χρησιμοποιώντας την τεχνική Up-Sampling/Over-Sampling (Chawla, 2009), όπου οι γραμμές από την κλάση των μειονοτήτων (δηλαδή την κλάση inactive), επανειλημμένα υποβάλλονται σε δειγματοληψία μέχρι να φτάσουν στο ίδιο μέγεθος με την κλάση πλειοψηφίας (δηλαδή την κλάση active). Αυτό σημαίνει ότι, κατά τη δημιουργία του συνόλου δεδομένων train_binary, οι γραμμές παρατηρήσεων με την inactive κλάση θα συλλέγονται περισσότερες φορές κατά τη διάρκεια της τυχαίας δειγματοληψίας. Ακόμη, όπως παρουσιάζεται και στο Πλαίσιο Α.13 για την εφαρμογή της τεχνικής UpSampling θα χρησιμοποιηθεί το πακέτο caret (Kuhn et al., 2019) της R. Ωστόσο, αν και στο Πλαίσιο Α.12 είναι εμφανής η υπεροχή την κατηγορίας highly_active έναντι των υπόλοιπων δύο στο υποδείγμα train _triple, προς το παρόν δεν καθίσταται το ίδιο αναγκαία η χρήση της μεθόδου upsampling για δημιουργία ενός νέου πιο ισορροπημένου μοντέλου.

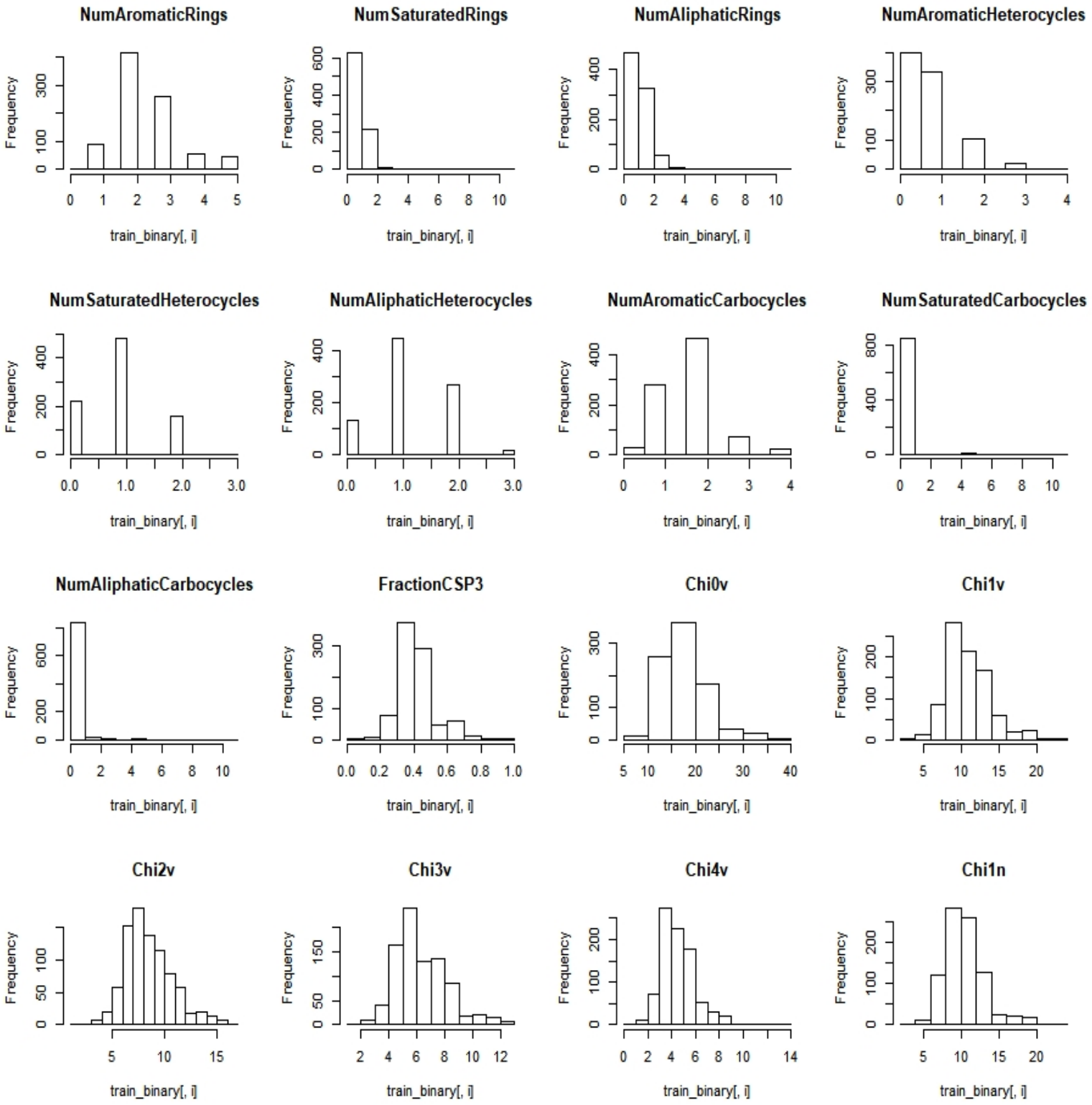
3.3.3 Απεικόνιση Δεδομένων

Στην προσπάθεια γραφικής αναπαράστασης των δεδομένων εκπαίδευσης δημιουργούνται αρχικά ιστογράμματα και θηκοδιαγράμματα για τις εν δυνάμει αριθμητικές επεξηγηματικές μεταβλητές των συνόλων δεδομένων. Συγκεκριμένα, στο Πλαίσιο Α.14 περιγράφεται η υλοποίηση των ιστογραμμάτων που απεικονίζονται στα Γραφήματα 3.2, 3.3, 3.4, ενώ στο Πλαίσιο Α.15 περιγράφεται η υλοποίηση των θηκοδιαγραμμάτων που απεικονίζονται στα Γραφήματα 3.5, 3.6, 3.7, 3.8.

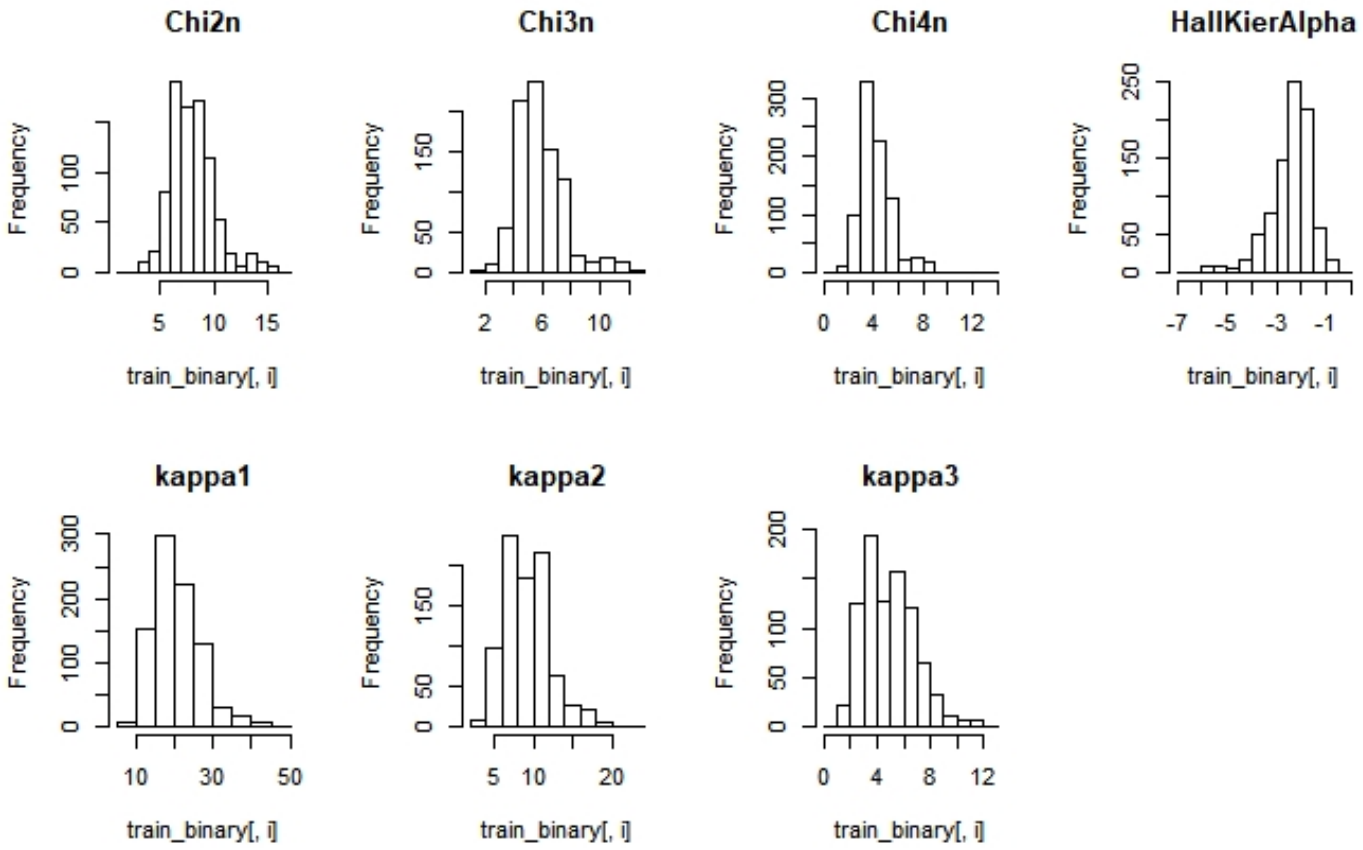
Ειδικότερα, στα Γραφήματα 3.2, 3.3, 3.4 παρατηρείται σε αρκετές μεταβλητές η εμφάνιση τιμών που εκπροσωπούν πολύ μικρό πλήθος παρατηρήσεων ενώ η κλίμακα που εκφράζει κάποιες μεταβλητές μπορεί να διαφέρει και μερικές εκατοντάδες. Για παράδειγμα, η μεταβλητή SlogP εκπροσωπείται κατά βάση από τιμές στο φάσμα 2 με 4 ενώ η μεταβλητή TSPA, κυρίως από τιμές στο διάστημα από 50-60. Βέβαια, αυτό οφείλεται στο γεγονός, ότι κάθε μεταβλητή περιγράφει μία διαφορετική φυσικοχημική ιδιότητα του εκάστοτε χημικού μορίου, καθιστώντας έτσι της μεταβλητές να μην μπορούν να εξισωθούν ποσοτικά πάντα μεταξύ τους. Από την άλλη, μεταβλητές που ανήκουν στην ίδια ομάδα φυσικοχημικών δεικτών, για περιγραφή χημικών ενώσεων όπως οι δείκτες kappa1,kappa2,kappa3, φαίνεται να παρουσιάζεται μεγαλύτερη συνέπεια στην κλίμακα που τους περιγράφει. Ακόμη, στα Γραφήματα 3.5, 3.6, 3.7, 3.8 επιβεβαιώνεται, πως αρκετές επεξηγηματικές μεταβλητές εκφράζονται από διαφορετική κλίμακα, εκτός από αυτές που ανήκουν στο ίδιου γκρουπ φυσικοχημικών δεικτών. Βέβαια, τώρα διαπιστώνεται πολύ πιο εύκολα, πως όλες οι επεξηγηματικές μεταβλητές, παρουσιάζουν έκτροπες τιμές συγκριτικά με τη μέση τιμή τους για το πλήθος των παρατηρήσεων. Κάτι τέτοιο θέλει προσοχή κατά τη μοντελοποίηση, για αποφυγή πιθανού επηρεασμού των αποτελεσμάτων.



Γράφημα 3.2: Ιστογράμματα των εν δυνάμει επεξηγηματικών μεταβλητών

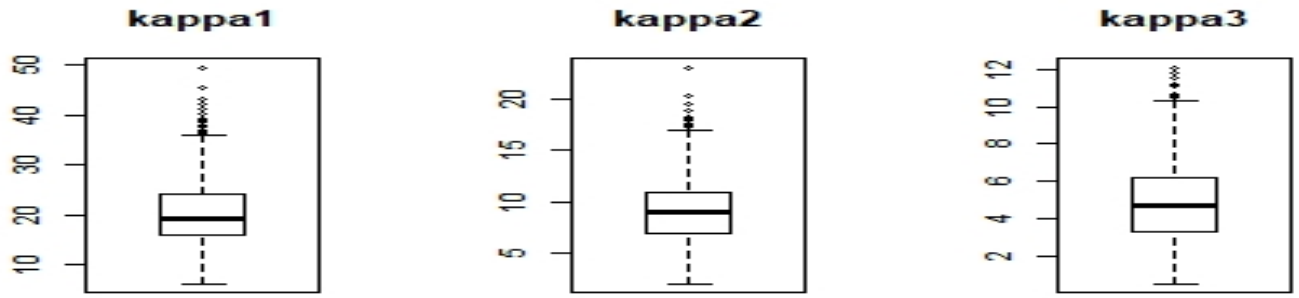


Γράφημα 3.3: Ιστογράμματα των εν δυνάμει επεξηγηματικών μεταβλητών

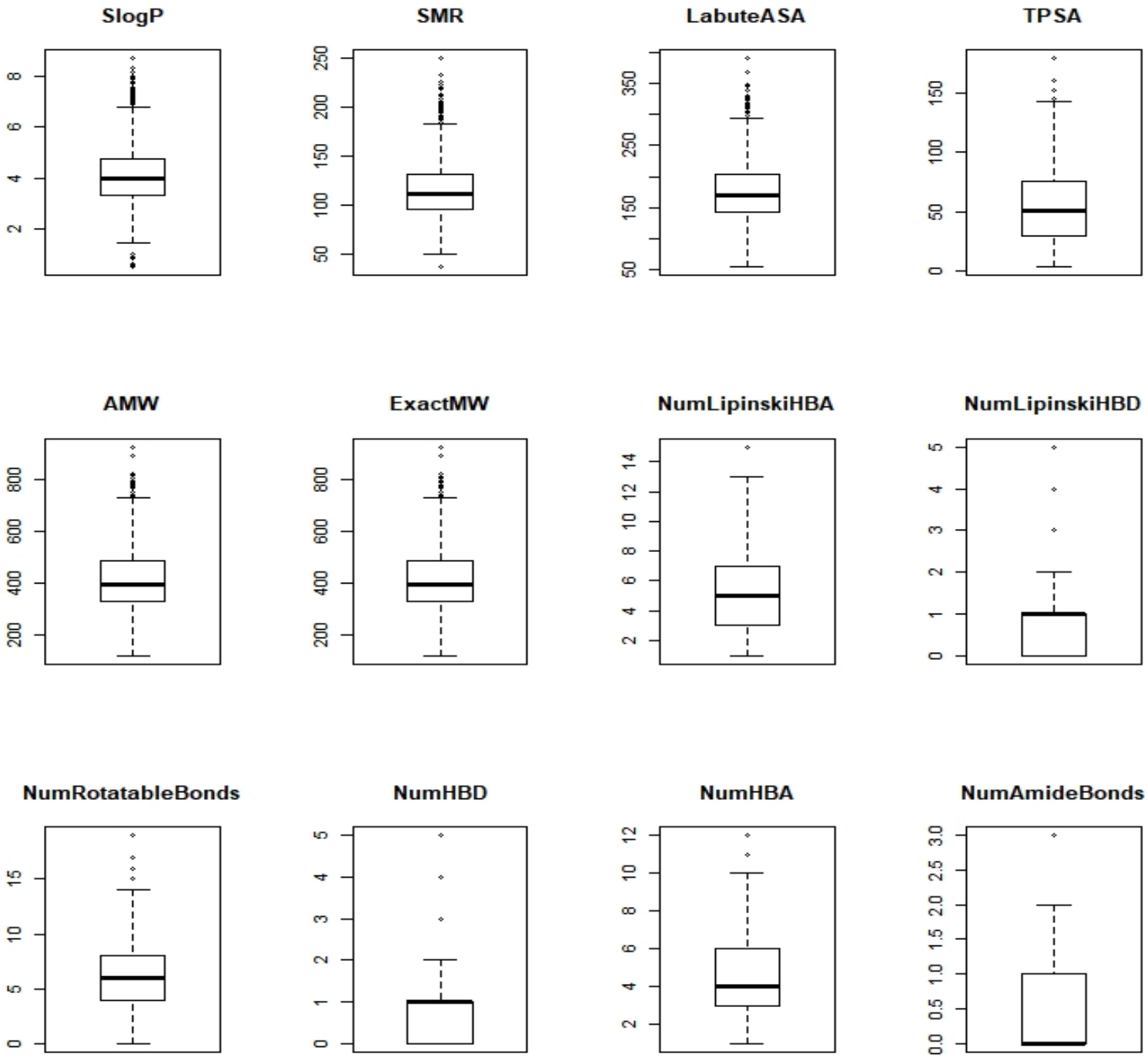


Γράφημα 3.4: Ιστογράμματα των εν δυνάμει επεξηγηματικών μεταβλητών

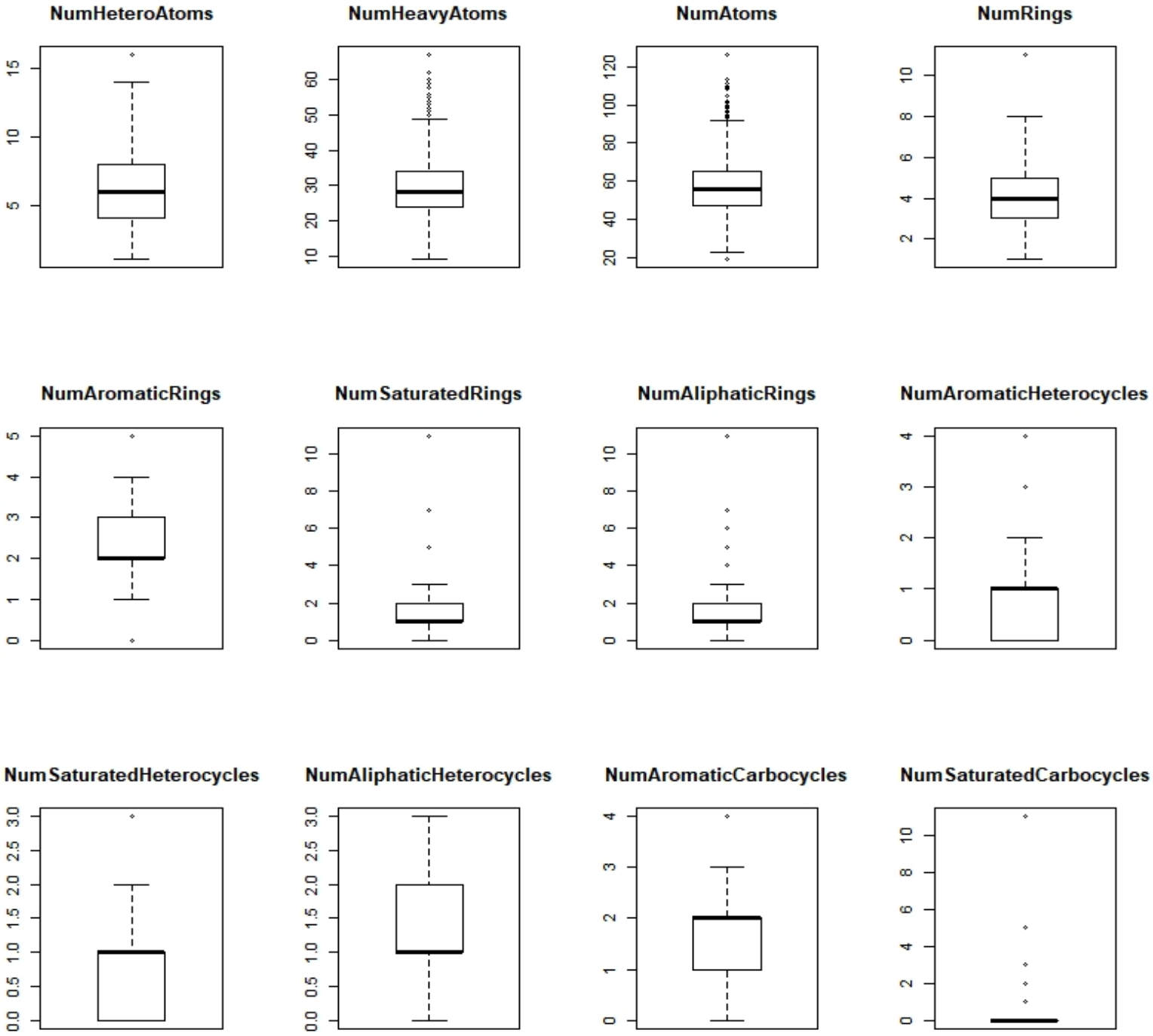
Επίσης, τα ιστογράμματα δημιουργούν την αίσθηση, ότι οι συναρτήσεις πυκνότητας των μεταβλητών δε θα είναι απαραίτητα κανονικής κατανομής, αλλά bimodal ή και multimodal, γεγονός που σχετίζεται και με την κατηγοριοποίηση των μεταβλητών σε 2 ή 3 κατηγορίες. Περισσότερες, όμως, πληροφορίες αναφορικά με τις συναρτήσεις πυκνότητας δίδονται στη συνέχεια που μοντελοποιούνται γραφικά. Προς το παρόν, συνεχίζεται η περιγραφή των επεξηγηματικών μεταβλητών μέσω της χρήσης θηκοδιαγραμμάτων.



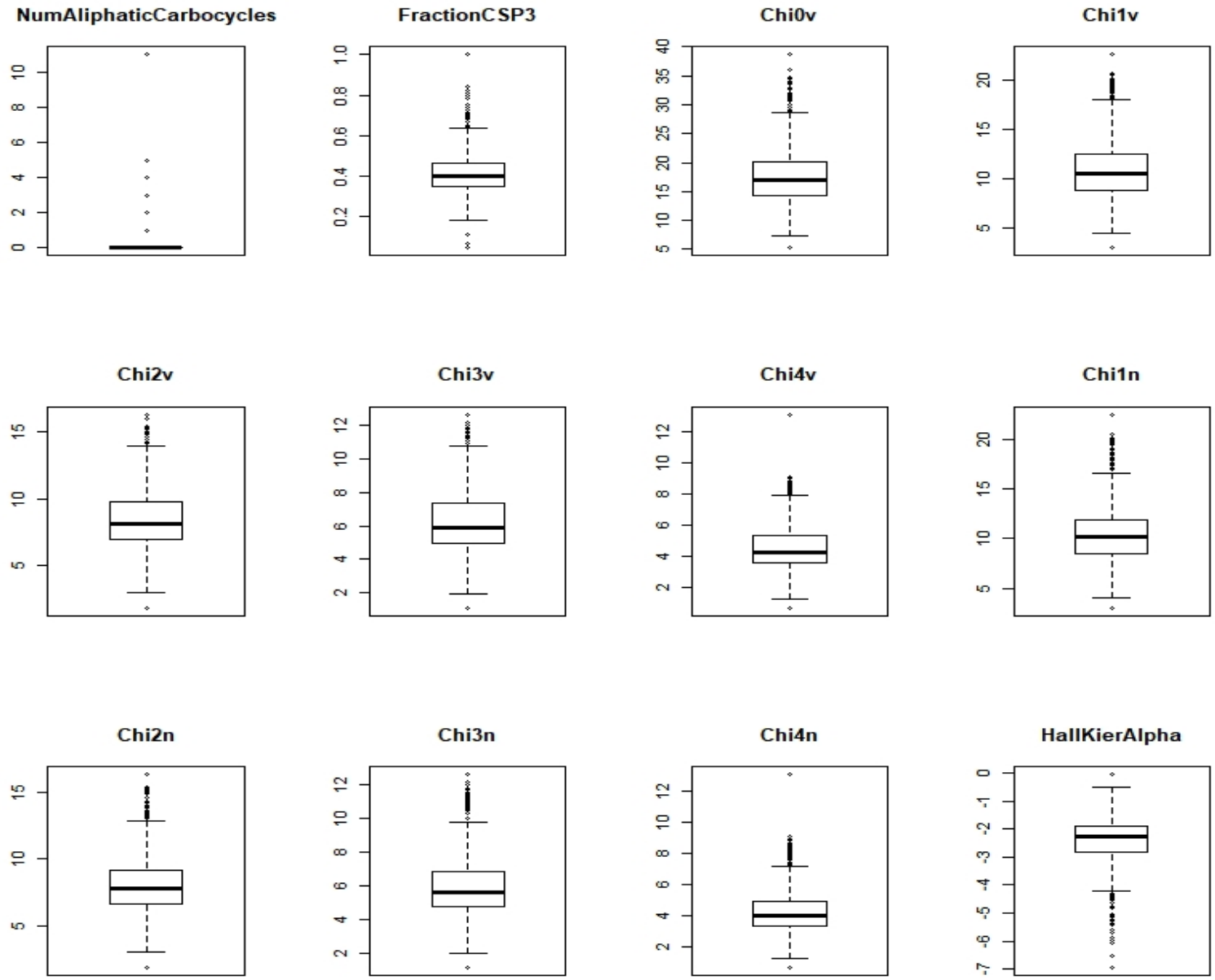
Γράφημα 3.5: Θηκοδιαγράμματα των εν δυνάμει επεξηγηματικών μεταβλητών



Γράφημα 3.6: Θηκοδιαγράμματα των εν δυνάμει επεξηγηματικών μεταβλητών



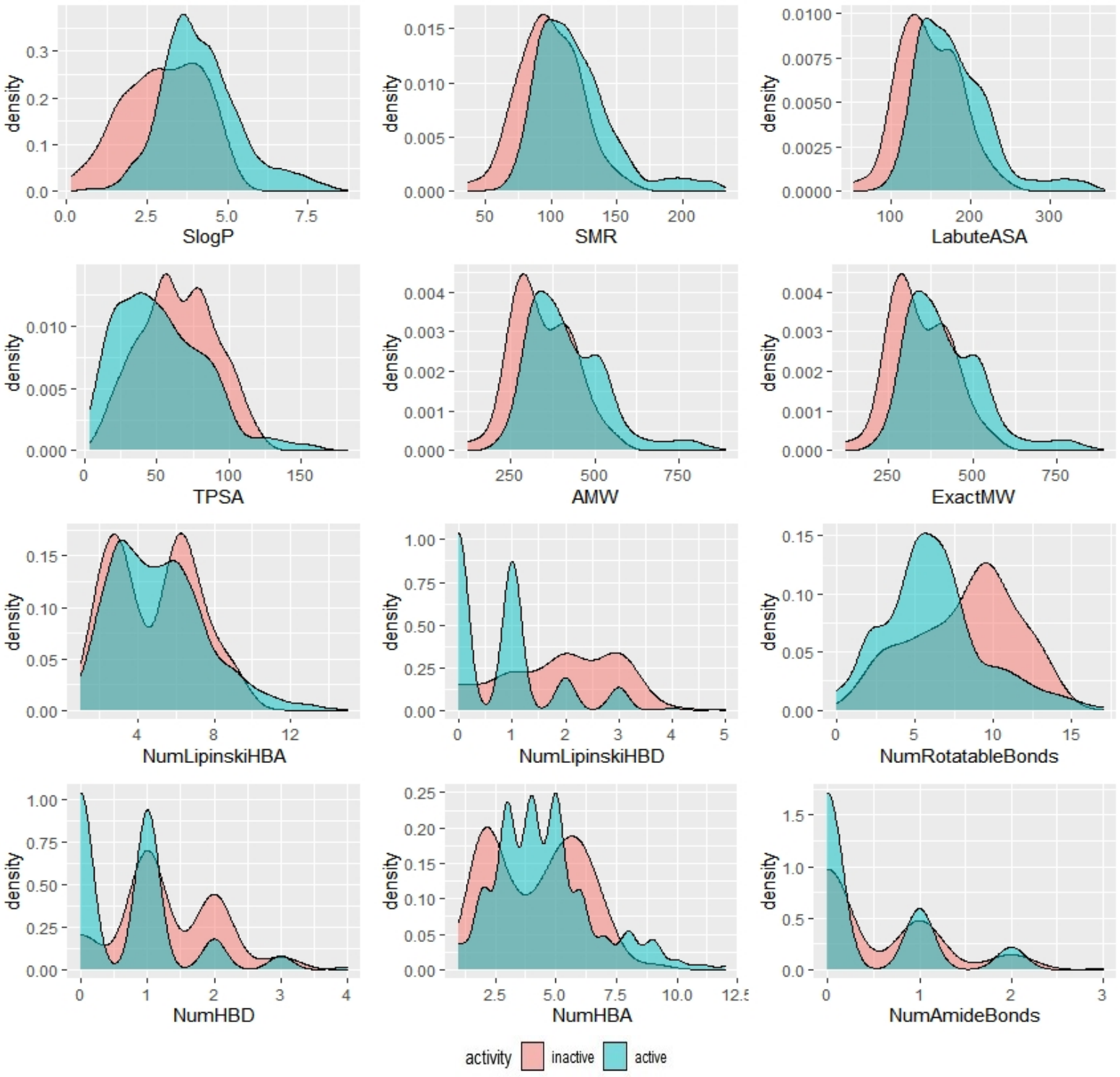
Γράφημα 3.7: Θηκοδιαγράμματα των εν δυνάμει επεξηγηματικών μεταβλητών



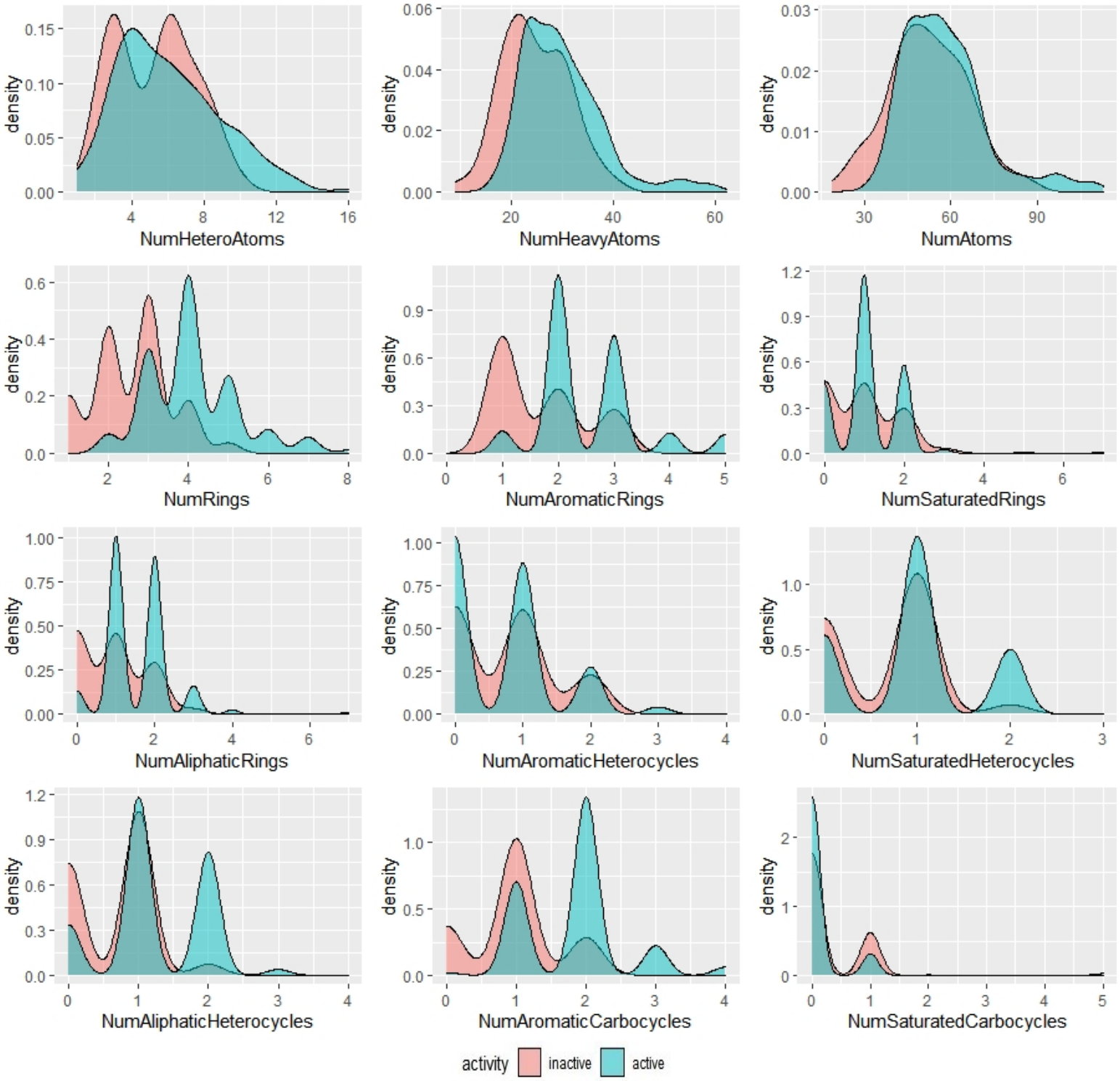
Γράφημα 3.8: Θηκοδιαγράμματα των εν δυνάμει επεξηγηματικών μεταβλητών

Όπως αναφέρθηκε και στην παρουσίαση των ιστογραμμάτων των διαφόρων μεταβλητών, παρατηρείται ότι η κατανομή, που πιθανόν τις περιγράφει, δεν είναι η κανονική, αλλά κατά πάσα πιθανότητα κάποια multimodal. Για αυτό το λόγο, κρίνεται αναγκαία η δημιουργία των γραφικών παραστάσεων για τη συνάρτηση πυκνότητας, που περιγράφει κάθε μεταβλητή.

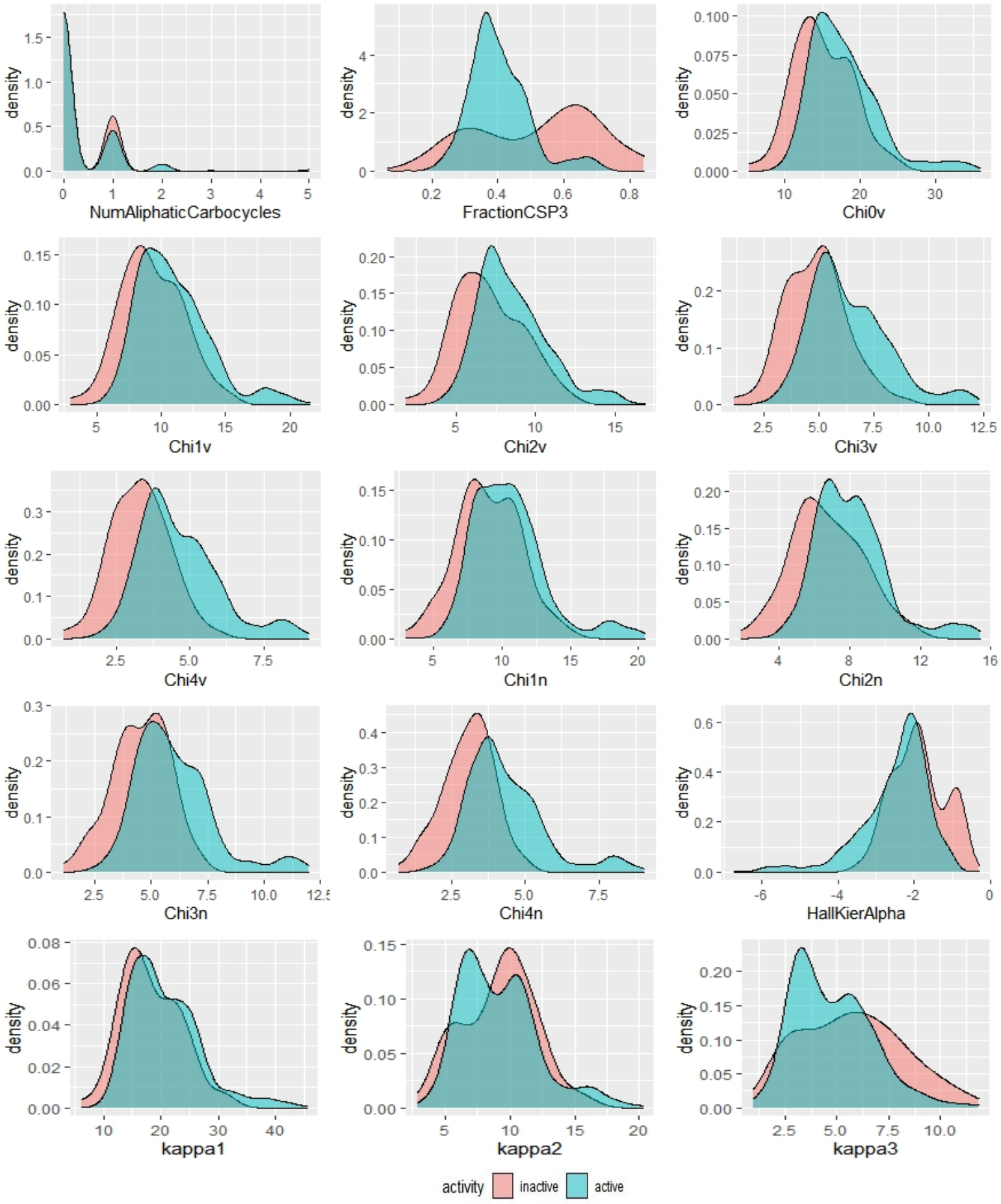
Καθώς στα δύο υποδείγματα εκπαίδευσης οι παρατηρήσεις ανήκουν σε μία κατηγορία, από 2 ή 3 δυνατές, κρίνεται χρήσιμη η γραφική αναπαράστασή της κατανομής πυκνότητας κάθε μεταβλητής με επικαλυπτόμμενες καμπύλες, ανάλογα με την κατηγορία στην οποία εμπίπτει η κάθε παρατήρηση. Για τη δημιουργία των γραφικών παραστάσεων, χρησιμοποιήθηκε η εντολή **ggpairs()** του πακέτου **GGally**(Schloerke et al., 2018), όπως εξηγείται στους πίνακες (A.16)-(A.17) και έτσι επιτεύχθηκε το αποτέλεσμα των Γραφημάτων 3.9 - 3.14.



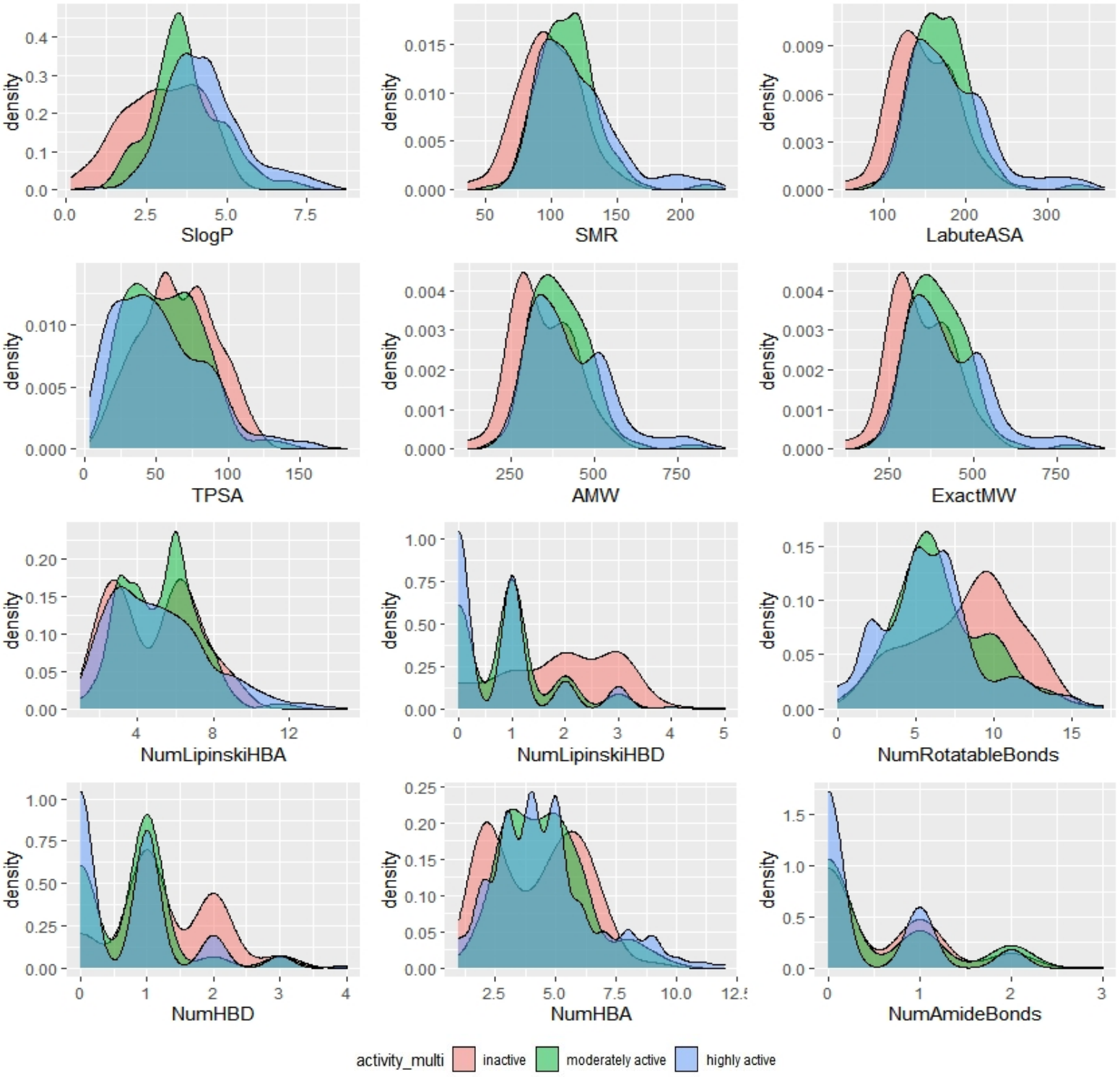
Γράφημα 3.9: Γραφήματα κατανομής πυκνότητας για τη δίτιμη μεταβλητή απόκρισης activity



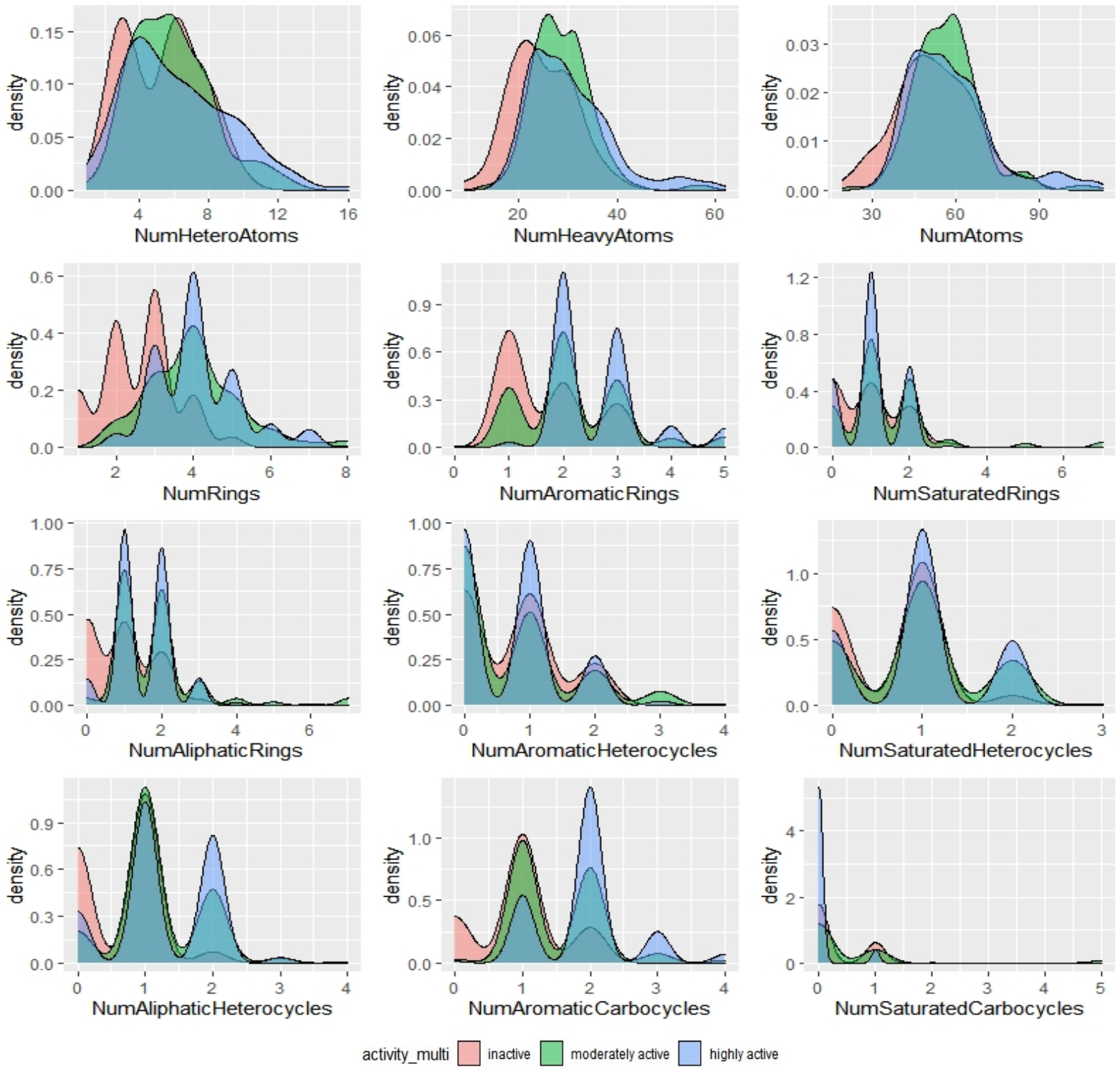
Γράφημα 3.10: Γραφήματα κατανομής πυκνότητας για τη δίτιμη μεταβλητή απόκρισης activity



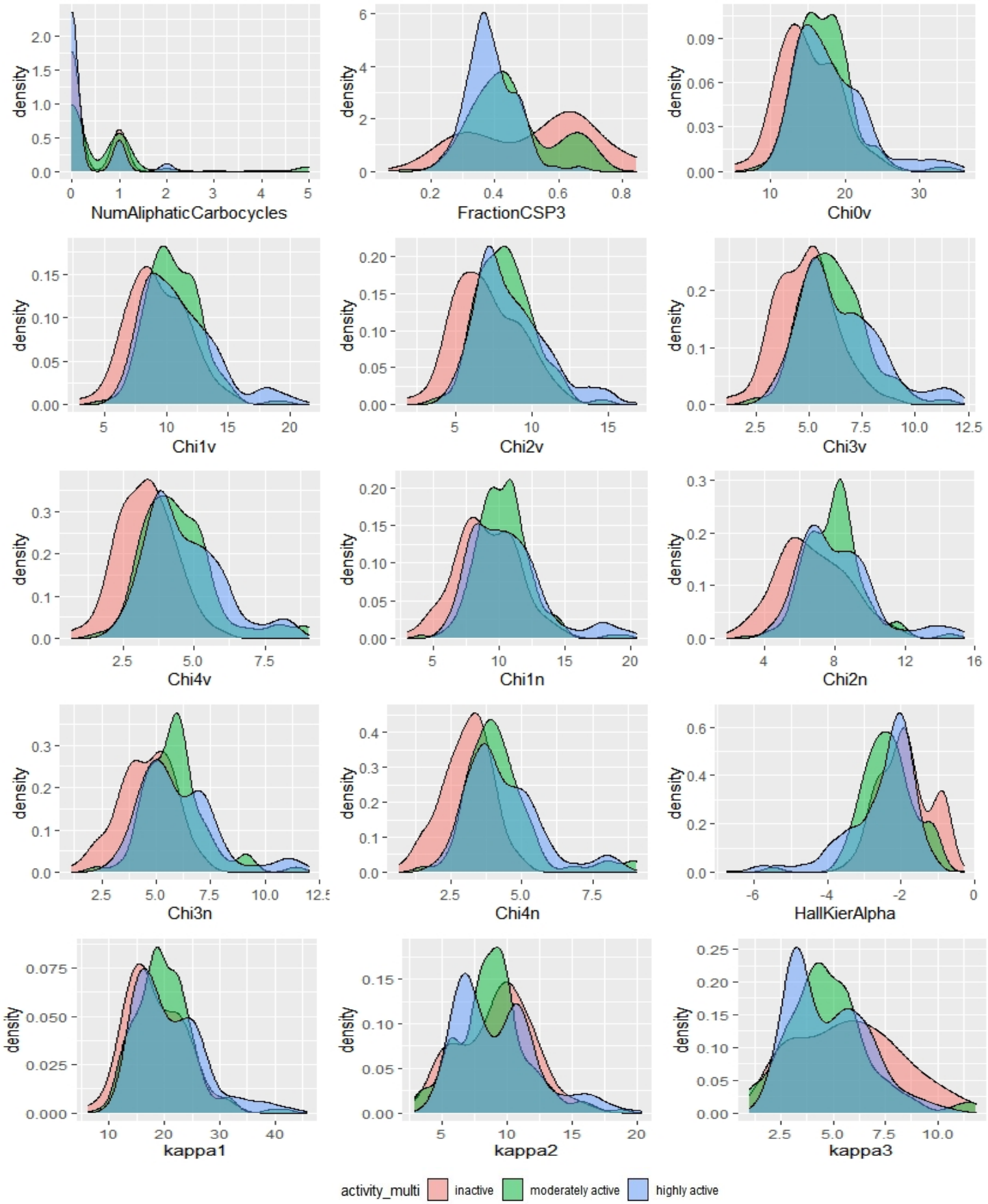
Γράφημα 3.11: Γραφήματα κατανομής πυκνότητας για τη δίτιμη μεταβλητή απόκρισης activity



Γράφημα 3.12: Γραφήματα κατανομής πυκνότητας για την τρίτιμη μεταβλητή απόκρισης activity_multi



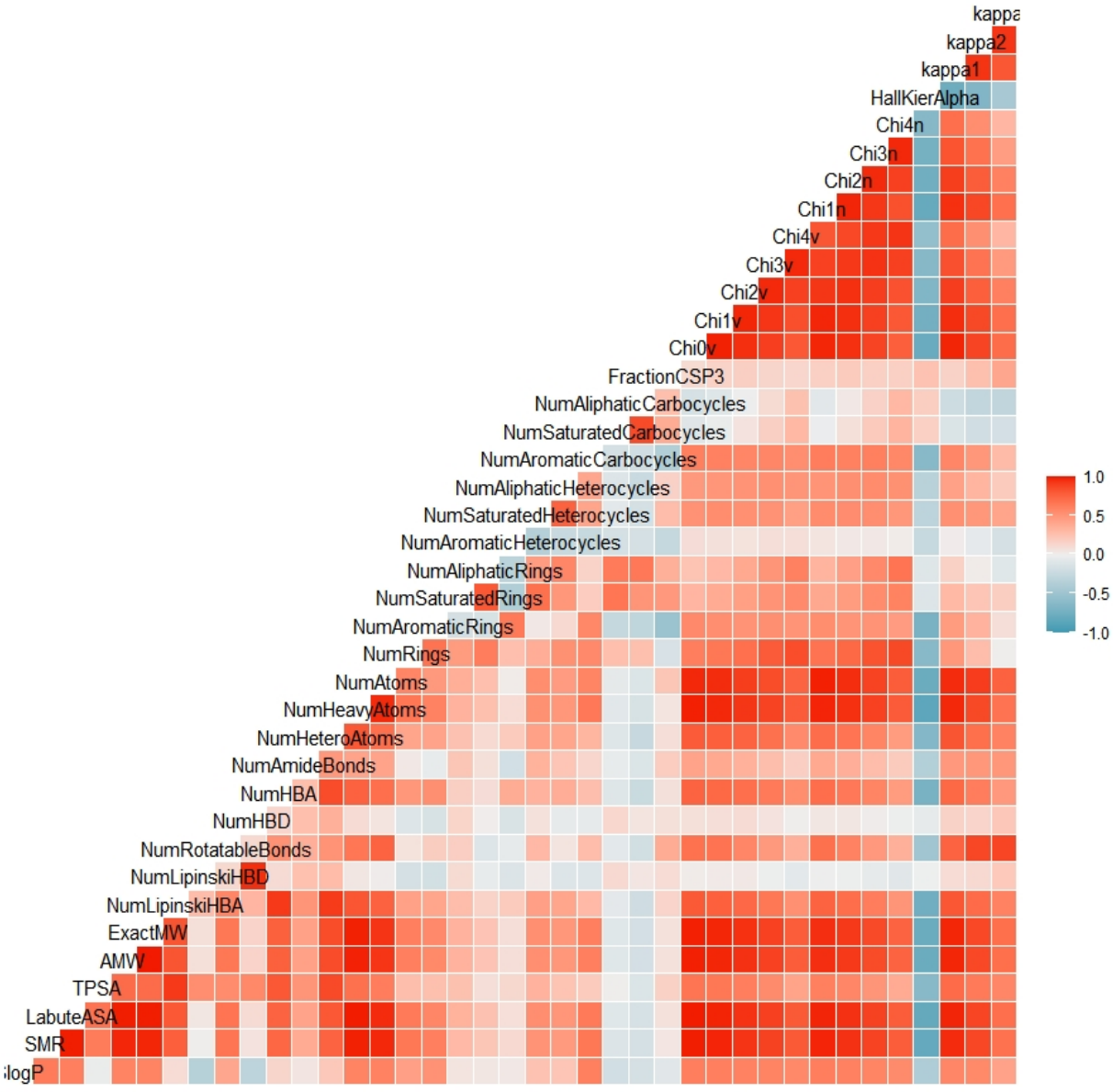
Γράφημα 3.13: Γραφήματα κατανομής πυκνότητας για την τρίτιμη μεταβλητή απόκρισης activity_multi



Γράφημα 3.14: Γραφήματα κατανομής πυκνότητας για την τρίτιμη μεταβλητή απόκρισης activity_multi

Στα Γραφήματα 3.9 - 3.14 επιβεβαιώνεται, πως οι συναρτήσεις πυκνότητας των επεξηγηματικών μεταβλητών, δεν ακολουθούν κανονικές κατανομές και για τα δύο υποδείγματα εκπαίδευσης. Μάλιστα, παρουσιάζονται πολλές κορυφές λόγω των τοπικών μεγίστων που εμφανίζουν οι μεταβλητές για κάθε κατηγορία που εμπίπτουν τα διάφορα χημικά μόρια. Συνεπώς, τα περισσότερα από ένα peaks(κορυφές) δηλώνουν πως οι κατανομές αυτές είναι όντως multimodal.

Επίσης, σημαντικό είναι να υπολογιστεί η συσχέτιση μεταξύ κάθε ζεύγους αριθμητικών μεταβλητών. Αυτό εύκολα υλοποιείται μέσω της εντολής **ggcorr()** του πακέτου **GGally** με τρόπο που περιγράφεται στο Πλαίσιο Α.18. Η ggcor() επιτρέπει την απεικόνιση της συσχέτισης κάθε ζεύγους μεταβλητών ως χρωματισμένο τετράγωνο ανάλογα με το βαθμό συσχέτισής τους. Αυτό, γίνεται πιο κατανοητό στο Γράφημα 3.15.



Γράφημα 3.15: Διάγραμμα συσχέτισης των χαρακτηριστικών των δεδομένων εκπαίδευσης

3.4 Εκπαίδευση Υποδειγμάτων και Δημιουργία Μοντέλων

Έχοντας πλέον παρουσιάσει και ελέγξει τα δεδομένα που θα χρησιμοποιηθούν για να εκπαιδευτούν οι αλγόριθμοι θα ξεκινήσει η παραγωγή των ταξινομητών που περιγράφηκαν στο Κεφάλαιο 2.

3.4.1 Εκπαίδευση μοντέλων για υποδείγμα με δίτιμη κατηγορική μεταβλητή απόκρισης

Η δημιουργία μοντέλων ξεκινά από μεθόδους δημιουργίας ταξινομητών για πρόβλεψη της δίτιμης κατηγοριοποίησης της βιοδραστηριότητας των μοντέλων με υποδείγμα εκπαίδευσης το *train_binary* και μεταβλητή απόκρισης την κατηγορική μεταβλητή activity.

3.4.1.1 Εκπαίδευση μοντέλου με απλή λογιστική παλινδρόμηση

Λόγω ύπαρξης μεταβλητής απόκρισης, που περιγράφεται από δύο κατηγορίες, η πιο συνήθης μέθοδος πρόβλεψης παρατηρήσεων, είναι αυτή της λογιστικής παλινδρόμησης και συγκεκριμένα της πολλαπλής λογιστικής παλινδρόμησης, καθώς τα δεδομένα που θα διαχειριστούν οι αλγόριθμοι περιγράφονται από μεγάλο πλήθος επεξηγηματικών μεταβλητών.

Πιο συγκεκριμένα, στην R με χρήση της εντολής **glm()** στο Πλαίσιο Α.19 γίνεται η μοντελοποίηση των παρατηρήσεων με τη βοήθεια λογιστικής παλινδρόμησης. Σημειώνεται, πως το όρισμα link="logit", εκφράζει πως η συνάρτηση σύνδεσης που θα χρησιμοποιηθεί από το μοντέλο είναι η συνάρτηση logit (σχέση 2.3). Συγκεκριμένα, μια συνάρτηση ζεύξης είναι απλά μια συνάρτηση του μέσου της μεταβλητής απόκρισης activity που χρησιμοποιείται ως απόκριση στο μοντέλο αντί για την ίδια τη μεταβλητή activity. Η συνάρτηση σύνδεσης logit όπως αυτή ορίστηκε στο Κεφάλαιο 2 χρησιμοποιείται για να μοντελοποιήσει το λόγο συμπληρωματικών πιθανοτήτων odds (σχέση 2.2). Ο σκοπός του συνδέσμου logit είναι να ληφθεί ένας γραμμικός συνδυασμός των επεξηγηματικών μεταβλητών και να μετατρέψει αυτές τις τιμές στην κλίμακα μιας πιθανότητας, δηλαδή $(0, 1)$.

Στο Πλαίσιο Α.19 απεικονίζεται το τελικό μοντέλο λογιστικής παλινδρόμησης με όλες τις επεξηγηματικές μεταβλητές του συνόλου δεδομένων train_binary να έχουν χρησιμοποιηθεί και να έχουν υπολογιστεί οι συντελεστές τους β_i, μέσω της μεγιστοποίησης της συνάρτησης πιθανοφάνειας (σχέση 2.4). Βέβαια, στις μεταβλητές NumAliphaticRings, NumAromaticCarbocycles, NumSaturatedCarbocycles, NumAliphaticCarbocycles, παρατηρείται η ένδειξη NA για τους συντελεστές που τις περιγράφουν. Η ένδειξη αυτή του κενού, προκλήθηκε πιθανώς λόγω της αυξημένης συσχέτισης που παρουσιάστηκε ανάμεσα σε αυτές τις μεταβλητές, όπως αυτή περιγράφεται στο Γράφημα 3.15.

Ακόμη, το Πλαίσιο Α.20, που εξετάζει μέσω της εντολής **vif()**, το συντελεστή διόγκωσης της διασποράς των επεξηγηματικών μεταβλητών του λογιστικού μοντέλου, παρουσιάζει κάποια ενδιαφέροντα αποτελέσματα. Συγκεκριμένα, ο συντελεστής (VIF) αποτελεί ένα δείκτη που μετρά πόσο αυξάνεται η διασπορά ενός εκτιμώμενου συντελεστή παλινδρόμησης, λόγω της πολυσυγραμμικότητας. Μάλιστα, η τιμή ενός παράγοντα VIF, μεταξύ 5 και 10, δείχνει υψηλή συσχέτιση, που μπορεί να είναι

προβληματική, ενώ αν υπερβεί την τιμή 10, μπορεί να υποτεθεί ότι οι συντελεστές παλινδρόμησης δεν υπολογίζονται επαρκώς λόγω της πολυσυγραμμικότητας.

Επομένως, γίνεται εύκολα αντιληπτό, ότι οι επεξηγηματικές μεταβλητές του μοντέλου εκφράζονται από υψηλές τιμές πολυσυγραμμικότητας, με αποτέλεσμα το τελικό μοντέλο λογιστικής παλινδρόμησης να μην είναι ισχυρός υποψήφιος για το καλύτερο μοντέλο περιγραφής των δεδομένων, αν και αυτό θα κριθεί στο Κεφάλαιο 4 μέσα από μία σειρά μετρήσεων για την αξιολόγησή του.

3.4.2 Εκπαίδευση μοντέλου λογιστικής παλινδρόμησης με τη μέθοδο βημάτων (Stepwise)

Ένας τρόπος να αξιοποιηθεί κάπως καλύτερα η μέθοδος της λογιστικής παλινδρόμησης, είναι μέσα από μία βελτιωτική μέθοδο με βήματα, όπως παρουσιάστηκε στην υποενότητα (2.2.2). Χαρακτηριστικά, μέσω της εντολής **stepAIC()** του πακέτου **MASS** (Venables and Ripley, 2013), εκτελείται λογιστική παλινδρόμηση με βήματα, αξιοποιώντας το παραπάνω λογιστικό μοντέλο. Σημειώνεται, ότι θα χρησιμοποιηθεί μία υβριδική μέθοδος με βήματα γι' αυτό και δόθηκε στην εντολή το όρισμα στην κατεύθυνση "both". Χαρακτηριστικά, η υβριδική αυτή μέθοδος κάνει χρήση της μεθόδου προς τα μπρος και της μεθόδου προς τα πίσω, όπως αυτές περιγράφηκαν στο Κεφάλαιο 2 , δίνοντας εν τέλει το λογιστικό μοντέλο που περιέχει εκείνες τις επεξηγηματικές μεταβλητές που ελαχιστοποιούν όσο αυτό είναι εφικτό τον δείκτη AIC.

Το εκπαιδευμένο μοντέλο με βήματα, όπως αυτό υπολογίστηκε από την R, παρουσιάζεται στο Πλαίσιο Α.21. Παρατηρείται, πως οι επεξηγηματικές μεταβλητές του καινούριου μοντέλου είναι εμφανώς μειωμένες στο πλήθος, αφού πλέον είναι 18 και όχι 39, ενώ ο δείκτης AIC, όπως ήταν αναμενόμενο, έχει πέσει από 243.56 σε 217.85 μονάδες. Γίνεται λόγος, εκ πρώτης όψεως, για ένα πιο διαχειρίσιμο μοντέλο, που υπάρχει πιθανότητα να έχει καλύτερη προβλεπτική ικανότητα από το απλό λογιστικό μοντέλο.

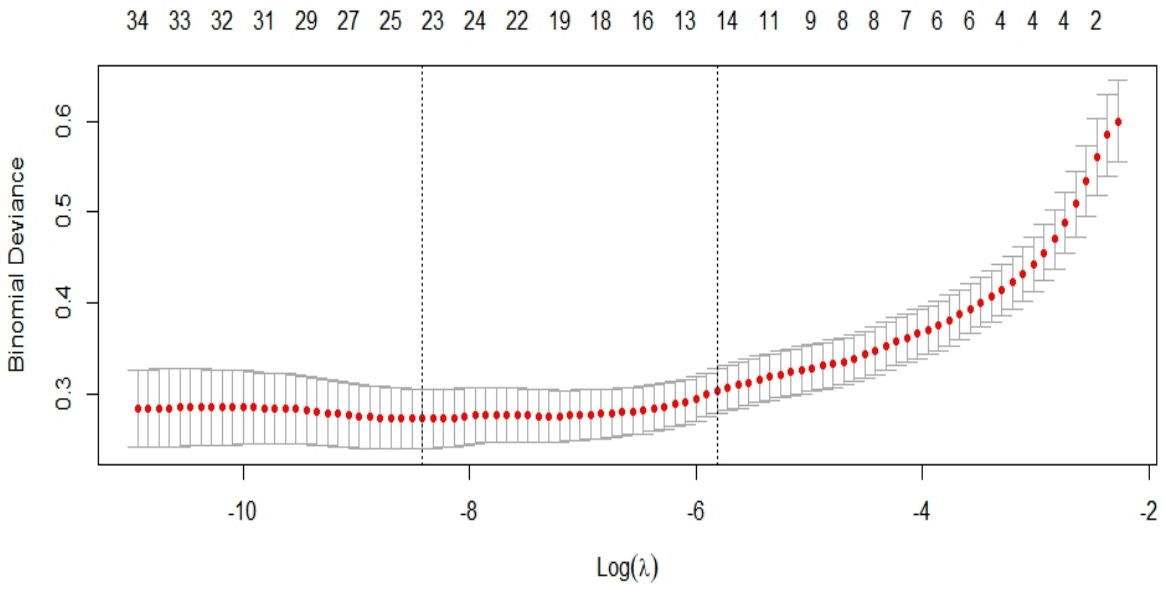
Στο Πλαίσιο Α.22 πραγματοποιήθηκε, και για το μοντέλο της μεθόδου stepwise, έλεγχος του παράγοντα vif των επεξηγηματικών μεταβλητών, με πιο ενθαρρυντικά αποτελέσματα αυτή τη φορά. Στην πλειοψηφία τους όμως, οι μεταβλητές φαίνεται να εκφράζονται από υψηλή πολυσυγραμμικότητα, γεγονός που δεν εξασφαλίζει την εγκυρότητα του μοντέλου.

Βέβαια, προτού υπολογιστούν μοντέλα με άλλες μεθόδους, πραγματοποιείται ένας πρώτος έλεγχος ανάλυσης διασποράς με τη μέθοδο **anova()**. Συγκεκριμένα, στο Πλαίσιο Α.23 μετά από τον έλεγχο anova, με χρήση του στατιστικού ελέγχου Chisq για σύγκριση των δύο παραπάνω μοντέλων, προκύπτει πως, αφού η τιμή του ελέγχου είναι $p = 0.9598 > 0.05$ τα δύο μοντέλα δε διαφέρουν ιδιαίτερα, οπότε οι επεξηγηματικές μεταβλητές που αφαιρέθηκαν από το λογιστικό μοντέλο κατά τη χρήση της μεθόδου stepwise σε ένα βαθμό ίσως είναι περιττές. Κοινώς, το μοντέλο με τη μέθοδο stepwise θα μπορούσε να προτιμηθεί έναντι του απλού λογιστικού μοντέλου.

3.4.3 Εκπαίδευση μοντέλου λογιστικής παλινδρόμησης με τη μέθοδο ποινής Lasso

Έχοντας, ήδη κάνει χρήση μίας βελτιωτικής μεθόδου για τη χρήση λογιστικής παλινδρόμησης, με μη ικανοποιητικά αποτελέσματα, αφού οι μεταβλητές του μοντέλου εμφάνισαν υψηλή πολυσυγγραμικότητα, κρίνεται χρήσιμη η εφαρμογή μίας άλλης μεθόδου συρρίκνωσης των συντελεστών του μοντέλου, αυτή της μεθόδου Lasso. Η μέθοδος Lasso, όπως ορίστηκε στην ενότητα (2.2.2) είναι μία μέθοδος ποινής που μπορεί να μειώσει στο μηδέν του συντελεστές της λογιστικής παλινδρόμησης.

Εξ' ορισμού η μέθοδος Lasso, όπως φαίνεται και στη σχέση (2.5) χρησιμοποιεί παραμέτρους λ για τη συρρίκνωση και την εύρεση των καλύτερων συντελεστών. Για αυτό το λόγο και για την εύρεση των καλύτερων δυνατών παραμέτρων λ, χρησιμοποιήθηκε η μέθοδος διασταυρωμένης επικύρωσης (cross validation) για τη δημιουργία πολλών μοντέλων Lasso. Έτσι, ήταν δυνατός ο υπολογισμός του λ εκείνου, για το οποίο επιτυγχάνεται η ελαχιστοποίηση του μέσου τετραγωνικού σφάλματος. Η επίτευξη αυτής της διαδικασίας έγινε με χρήση της εντολής **cv.glmnet()** του πακέτου **glmnet** (Simon et al., 2011b) και περιγράφεται στο Πλαίσιο Α.24.



Γράφημα 3.16: Cross-validation σφάλμα σύμφωνα με το log(λ)

Έχοντας υπολογίσει τους δείκτες λ, για τη μέθοδο Lasso, στο Γράφημα 3.16 απεικονίζεται η λογαριθμική τους τιμή συναρτήσει της μέσης διωνυμικής τυπικής απόκλισης (μείον το διπλάσιο της λογαριθμικής πιθανοφάνειας ως προς τα δεδομένα που έχουν παραλειφθεί από τη διασταυρωμένη επικύρωση (Simon et al., 2011b)). Η τιμή lambda.min δηλώνει την τιμή της παραμέτρου που ελαχιστοποιεί την παραπάνω τιμή και επιτυγχάνονται έτσι τα καλύτερα δυνατά αποτελέσματα για τους συντελεστές. Η lambda.1se από την άλλη, είναι η τιμή της παραμέτρου λ, που απέχει ένα τετραγωνικό σφάλμα, από την τιμή λ που επιτυγχάνει ελαχιστοποίηση.

Υπολογίζοντας τώρα, τα μοντέλα λογιστικής παλινδρόμησης με τη μέθοδο Lasso, αντικαθιστώντας τις δύο προαναφερθείσες τιμές λ στη σχέση (2.5) προκύπτουν οι τιμές των συντελεστών που παρουσιάζονται στα πλαίσια (A.26) και (A.27).

Συγκεκριμένα, για λ = 0.00022 οι επεξηγηματικές μεταβλητές του μοντέλου μειώνονται σε 25, αφού οι συντελεστές 14 εξ' αυτών έχουν συρρικνωθεί στο μηδέν. Ενώ, για λ=0.0029 συρρικνώνονται ακόμα περισσότεροι συντελεστές. (Βέβαια το κόστος αυτού του φαινομένου στην προβλεπτική ικανότητα του μοντέλου θα αναλυθεί κατά την αξιολόγηση των μοντέλων στο Κεφάλαιο 4).

3.4.4 Εκπαίδευση μοντέλου με γραμμική διακριτική ανάλυση (LDA)

Αφήνοντας το λογιστικό μοντέλο, μία μέθοδος που χρησιμοποιείται επίσης για τη δημιουργία ταξινομητών, είναι η γραμμική διακριτική ανάλυση. Συγκεκριμένα, με χρήση της εντολής **lda()** υπολογίζεται εύκολα το μοντέλο γραμμικής διακριτικής ανάλυσης και οι συντελεστές του παρουσιάζονται στο Πλαίσιο A.28. Παρατηρείται, όμως και ένα μήνυμα για "προσοχή" της εγκυρότητας των αποτελεσμάτων, λόγω ύπαρξης συγγραμμικών μεταβλητών. Όπως επισημάνθηκε και στη δημιουργία του λογιστικού μοντέλου αυτές οι μεταβλητές θα είναι οι NumAliphaticRings, NumAromaticCarbocycles, NumSaturatedCarbocycles, NumAliphaticCarbocycles. Αφαιρώντας, τις μεταβλητές αυτές υπολογίστηκε ξανά στο Πλαίσιο A.29 το μοντέλο γραμμικής διακριτικής ανάλυσης, χωρίς προειδοποιητικά μηνύματα αυτή τη φορά, αλλά, τελικά, με ίδιες τιμές στους συντελεστές των μεταβλητών που χρησιμοποιήθηκαν στο νέο μοντέλο. Είναι εμφανές, λοιπόν, ότι η αφαίρεση των μεταβλητών αυτών δεν έπαιξε κάποιο ρόλο στο μοντέλο. Ακόμη, τα μεγέθη group means δηλώνουν τη μέση τιμή κάθε μεταβλητής σε κάθε κατηγορία χημικού μορίου.

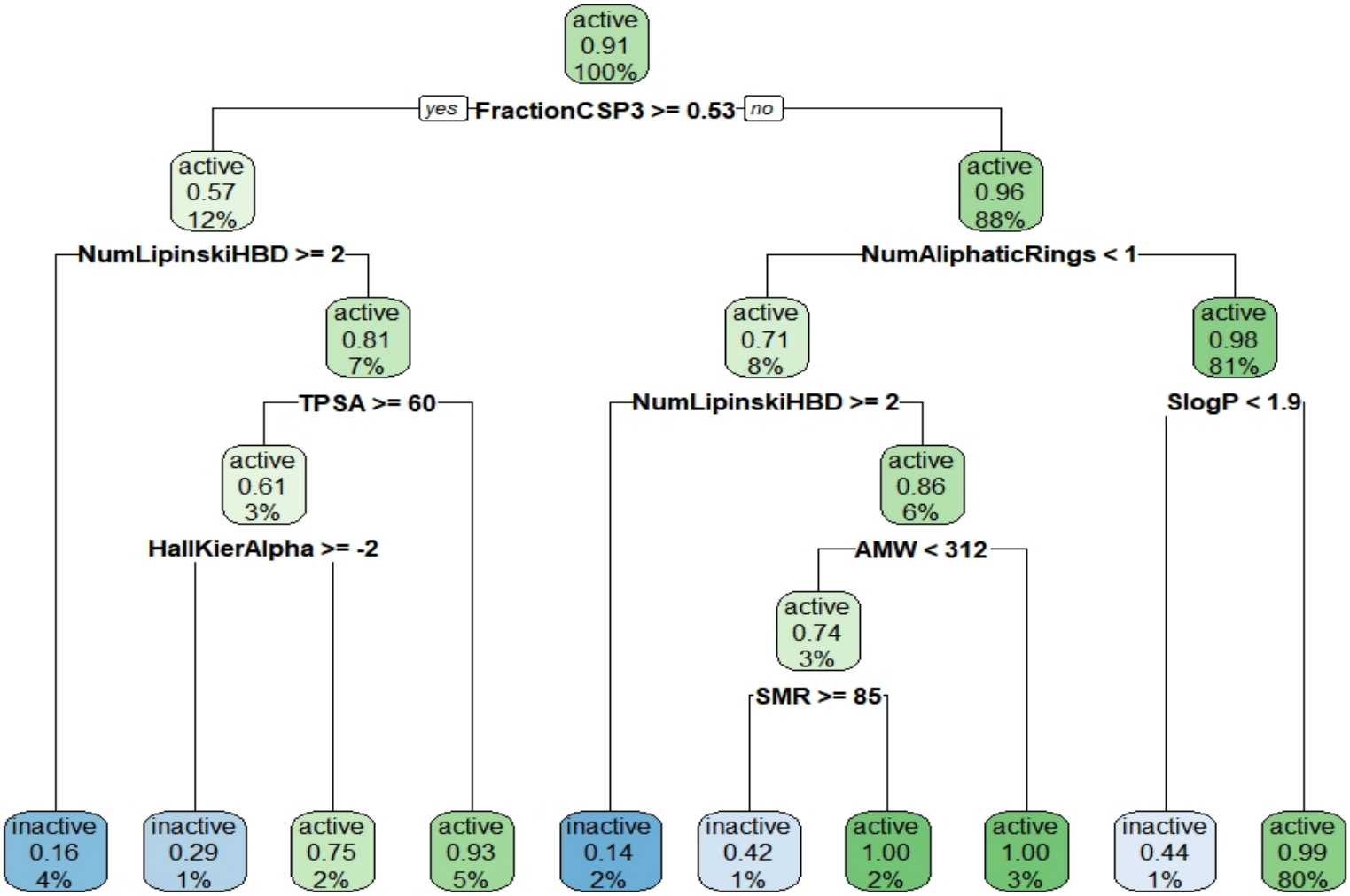
3.4.5 Εκπαίδευση μοντέλου με δένδρο απόφασης CART (decision tree)

Παιρνώντας σε χρήση γραφικών μεθόδων μοντελοποίησης των δεδομένων, δημιουργείται ένα δένδρο απόφασης CART, με βάση τη δίτιμη κατηγορική μεταβλητή activity. Όπως, φαίνεται και στο Πλαίσιο A.30 με χρήση της μεθόδου **rpart** του πακέτου **rpart** (Therneau and Atkinson, 2019) δημιουργείται ένα δένδρο απόφασης το οποίο απεικονίζεται στην τελική του μορφή στο Γράφημα 3.17 μέσω της εντολής **rpart.plot** του προαναφερθέντος πακέτου.

Το δένδρο απόφασης, αποτελείται από 8 επεξηγηματικές μεταβλητές, που ανάλογα με την τιμή που λαμβάνει κάθε χημικό μόριο σε αυτές, καθορίζεται το μονοπάτι εκείνο στο δένδρο που θα το οδηγήσει να προβλεφθεί ως ενεργό ή ανενεργό. Μάλιστα, τα φύλλα του δένδρου είναι 10, γεγονός που δηλώνει, πως για κάθε χημικό μόριο, μπορεί να προσδιορισθεί η βιοδραστηριότητά από 10 διαφορετικά μονοπάτια. Πρακτικά, αυτό σημαίνει πως, μόνο μία μεταβλητή δεν αρκεί για να καθορίσει ένα μοντέλο ως ενεργό ή ανενεργό, αλλά πρέπει να πληρείται μία σειρά από προδιαγραφές για τα φυσικοχημικά χαρακτηριστικά του, FractionCSP3, NumLipinskiHBD, NumAliphaticRings, TPSA, HallKierAlpha,

SlogP, AMW και SMR. Σημειώνεται, πως η μεταβλητή FractionCSP3 είναι αυτή που επιτυγχάνει τον καλύτερο κερματισμό του αρχικού συνόλου δεδομένων γι' αυτό και βρίσκεται στη ρίζα του δέντρου.

Προφανώς, είναι άμεσα αντιληπτό, πως το μοντέλο αυτό είναι πιο εύχρηστο, για να προβλεφθεί η βιοδραστηριότητα ενός χημικού μορίου, αλλά δεν είναι σίγουρο ότι εξασφαλίζεται και η ακρίβειά του. Κάτι τέτοιο θα εξετασθεί και θα αναλυθεί στο Κεφάλαιο 4.



Γράφημα 3.17: Διάγραμμα συσχέτισης των χαρακτηριστικών των δεδομένων εκπαίδευσης

3.4.6 Εκπαίδευση μοντέλων για εξισορροπημένο υποδείγμα δίτιμης κατηγορικής μεταβλητής απόκρισης, με μέθοδο Upsampling,

Όπως αναφέρθηκε και κατά την επεξεργασία των δεδομένων, το σύνολο δεδομένων εκπαίδευσης train_binary παρουσιάζει μεγάλη ανομοιογένεια. Γι' αυτόν το λόγο, υπάρχει περίπτωση, κατά την πρόβλεψη παρατηρήσεων, μέσω των μοντέλων που εκπαιδεύτηκαν παραπάνω, τα αποτελέσματα να μη χαρακτηρίζονται από αμεροληψία. Κρίνεται, έτσι χρήσιμο, να εκπαιδευτούν τα παραπάνω μοντέλα στο ισορροπημένο σύνολο δεδομένο up_train, που υπολογίστηκε στο Πλαίσιο Α.13.

Οι διαδικασίες εκπαίδευσης των μοντέλων παρουσιάζονται στο Πλαίσιο A.31, δίχως λεπτομέρειες για τα παραγόμενα μοντέλα. Ο σκοπός του να παρουσιαστεί μόνον η εκπαίδευση των ταξινομητών, σε ένα πιο ισορροπημένο υποδείγμα, είναι πως περισσότερη σημασία έχει η ουσιαστικότερη αξιολόγηση των παραπάνω ταξινομητών, λόγω της κακής αναλογίας των δεδομένων (διαδικασία που πραγματοποιείται στο Κεφάλαιο 4) και όχι η παρουσίαση εκ νέου του τρόπου δημιουργίας των εκπαιδευμένων ταξινομητών, με τις ίδιες μεθόδους.

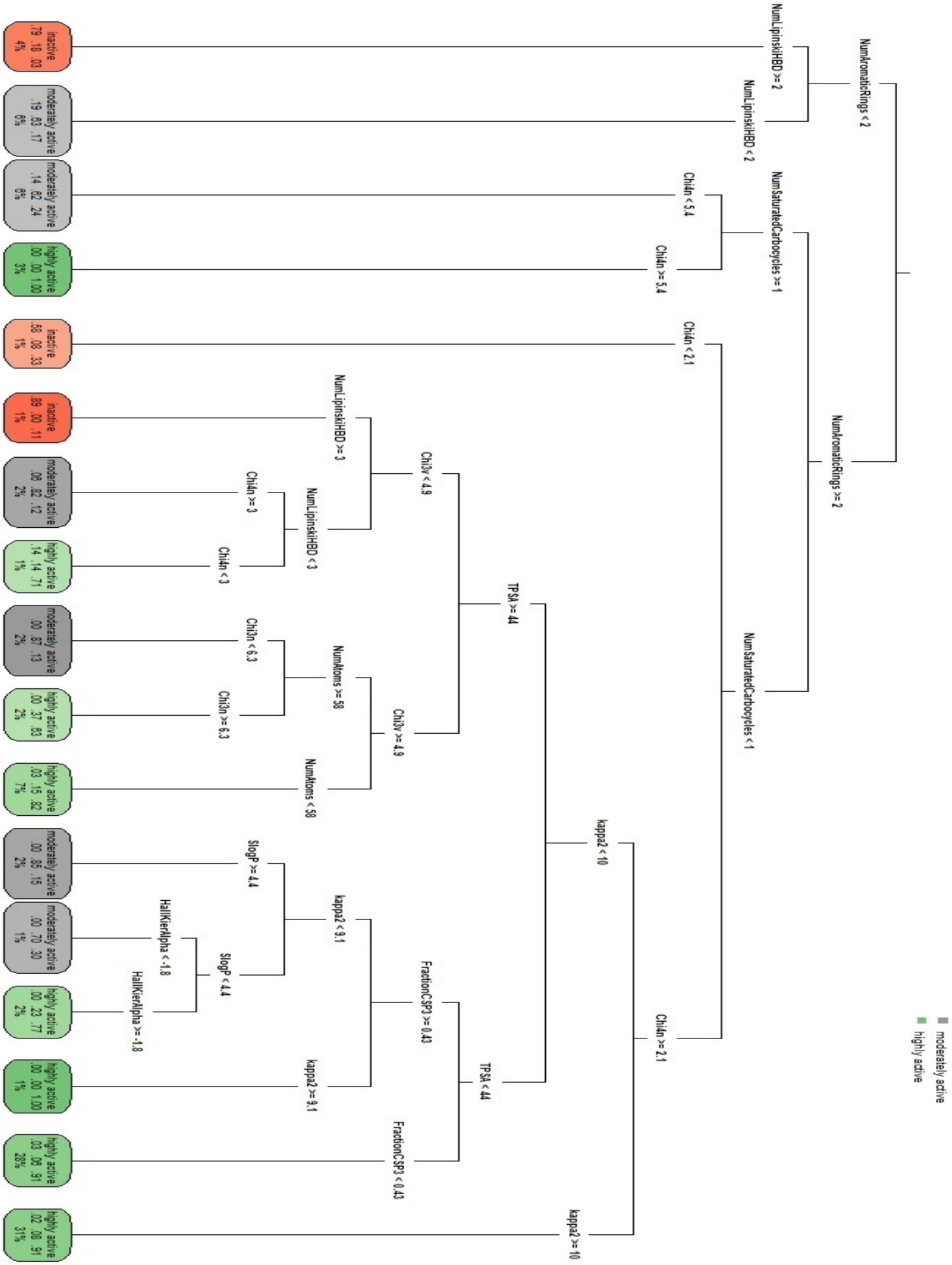
3.4.7 Εκπαίδευση μοντέλων για υποδείγμα με τρίτιμη κατηγορική μεταβλητή απόκρισης

Έχοντας μοντελοποιήσει, παραπάνω, τα δεδομένα των χημικών μορίων για την περίπτωση που διαχωρίζονται ανάλογα με την τιμή της μέσης μέγιστης ανασταλτικής τους ιδιότητας σε δύο κατηγορίες, παρακάτω, θα γίνει προσπάθεια μοντελοποίησης των δεδομένων, στην περίπτωση που η βιοδραστηριότητα των χημικών οντοτήτων περιγράφεται από τρεις διαφορετικές κατηγορίες .

3.4.7.1 Εκπαίδευση μοντέλου με δένδρο απόφασης CART (decision tree)

Δημιουργώντας ένα δένδρο απόφασης, όπως και στην ενότητα 3.5.5, με τη βοήθεια της εντολής rpart() αυτή τη φορά με υποδείγμα εκπαίδευσης το train_triple, προκύπτει το αποτέλεσμα του πλαισίου (A.32).

Το συγκεκριμένο δένδρο δείχνει την πολυπλοκότητα, που απέκτησε η δομή, του συγκριτικά με το δένδρο απόφασης με δίτιμη μεταβλητή απόκρισης. Αυτό συμβαίνει γιατί, ο αλγόριθμος που περιγράφηκε στην ενότητα (2.2.5) κλήθηκε να διαχωρίζει τα δεδομένα βάσει 3 κατηγοριών αυτή τη φορά, άρα για την επιλογή του κατάλληλου cutpoint υπήρχαν παραπάνω τιμές για την επικρατέστερη κατηγορία σε ένα υποσύνολο. Έτσι, το συνολικό δένδρο για την καλύτερη πρόβλεψη της κατηγορίας των χημικών μορίων, που χαρακτηρίζονται από τις κατηγορίες, inactive, moderately active και highly active, περιγράφεται από περισσότερες μεταβλητές, 16 το πλήθος ενώ τα φύλλα του δένδρου είναι 17. Άρα, υπάρχουν 17 διαφορετικές σειρές προδιαγραφών, για να προβλεφθεί η κατηγορία μίας χημικής ένωσης ή ενός χημικού μορίου.

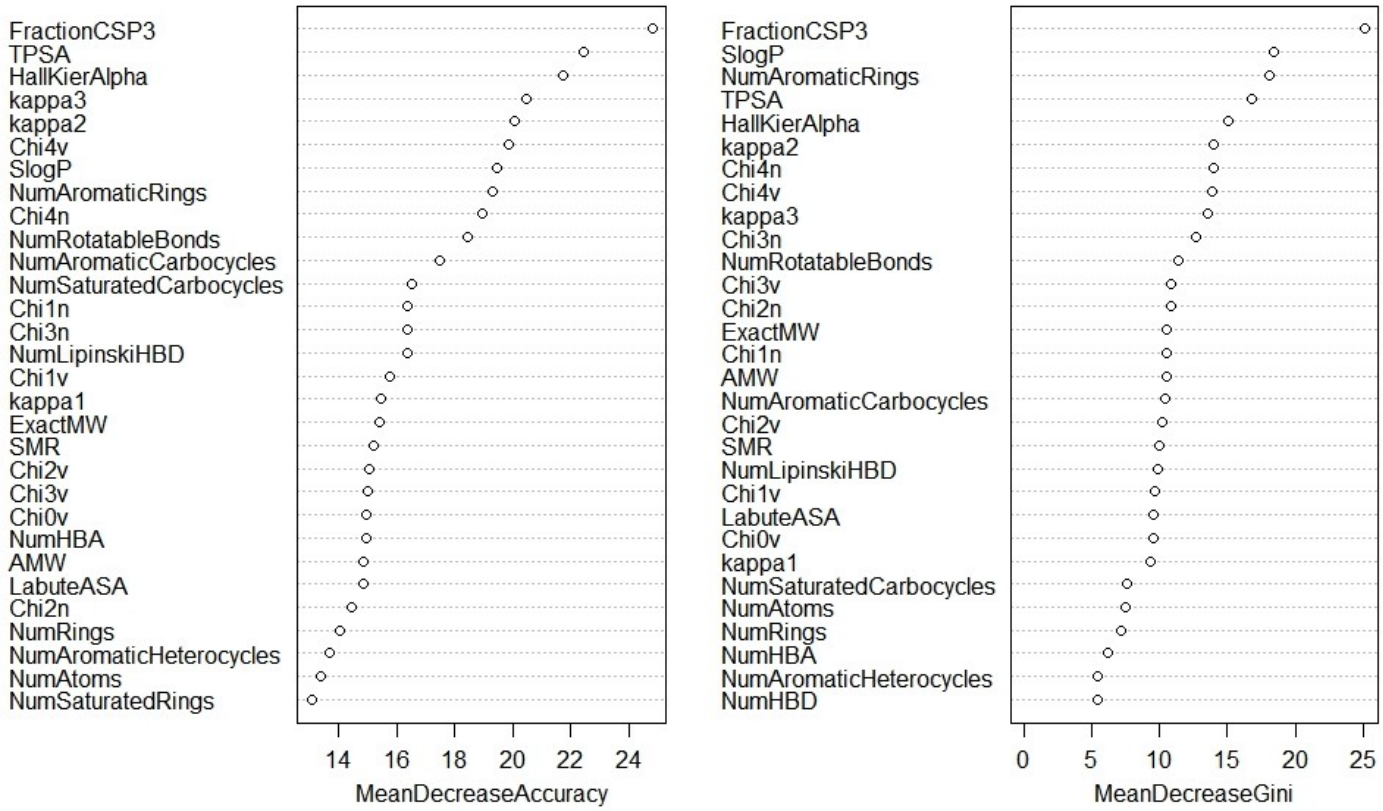


Γράφημα 3.18: Διάγραμμα τυχαίου δάσους για το υποδείγμα train_triple

3.4.7.2 Εκπαίδευση μοντέλου με τυχαίο δάσος (random forest)

Έχοντας αναλύσει τη δημιουργία μοντέλου από ένα τυχαίο δάσος, είναι εύκολο τώρα να δημιουργηθεί ένα μοντέλο, από την ανάπτυξη πολλών τέτοιων δένδρων, με τη γνωστή αλγοριθμική διαδικασία του τυχαίου δάσους (ενότητα 2.2.6). Συγκεκριμένα, στο Πλαίσιο Α.33, δημιουργείται ένα τυχαίο δάσος, μέσω της εντολής **randomForest()**του πακέτου **randomForest**(Liaw, Wiener, et al., 2002).

Παρατηρείται πως, κατά την ανάπτυξη του τυχαίου δάσους, δημιουργήθηκαν 500 δένδρα, ενώ ο αριθμός των μεταβλητών σε κάθε διαχωρισμό των δεδομένων ήταν 6. Ακόμη, ο υπολογισμός του σφάλματος λανθασμένης ταξινόμησης μέσω της διαδικασίας out-of-bag, έδωσε τιμή σφάλματος της τάξης του 16.88%. Επίσης, στο Γράφημα 3.19 μέσω της εντολής **varImpPlot()** του πακέτου **randomForest** παρέχεται μία αναπαράσταση της σημαντικότητας της κάθε επεξηγηματικής μεταβλητής του μοντέλου που έχει παραχθεί από το τυχαίο δάσος. Ειδικότερα, όσο ψηλότερα είναι στη λίστα του γραφήματος μία μεταβλητή τόσο πιο σημαντική είναι.



Γράφημα 3.19: Διαγράμματα σημαντικότητας μεταβλητών μοντέλου τυχαίου δάσους για το υποδείγμα train_triple

3.4.7.3 Εκπαίδευση μοντέλου με πολυωνυμική λογιστική παλινδρόμηση (multinomial logistic regression)

Επιστρέφοντας, σε μεθόδους εκπαίδευσης των μοντέλων που απέχουν από τη δημιουργία δέντρων, χρησιμοποιείται για ακόμα μία φορά λογιστική παλινδρόμηση. Βέβαια, λόγω της τρίτιμης κατηγορικής μεταβλητής απόκρισης για τα δεδομένα του υποδείγματος train_triple, η απλή λογιστική παλινδρόμηση δεν επαρκεί για τη μοντελοποίηση των δεδομένων. Έτσι, επιστρατεύεται η μέθοδος της πολυωνυμικής λογιστικής παλινδρόμησης, όπως αυτή παρουσιάστηκε στην ενότητα (2.2.3).

Η δημιουργία του μοντέλου, επιτεύχθηκε μέσω της R με την εντολή **multinom** του πακέτου **nnet**(Venables and Ripley, 2013), ενώ ως κατηγορία αναφοράς για τη δημιουργία των λόγων συμπληρωματικών πιθανοτήτων logit (σχέση 2.7), χρησιμοποιήθηκε η κατηγορία highly active.

3.4.8 Εκπαίδευση μοντέλου με μηχανές διανυσμάτων υποστήριξης (SVM)

Η τελευταία μέθοδος που θα χρησιμοποιηθεί για τη δημιουργία ταξινομητή των παρατηρήσεων είναι αυτή των μηχανών διανυσμάτων υποστήριξης που μοντελοποιούν με μεγάλη ευχέρεια δεδομένα, των οποίων η μεταβλητή απόκρισής τους μπορεί να περιγραφεί από παραπάνω από δύο ενδεχόμενες καταστάσεις.

Στο Πλαίσιο Α.35, γίνεται χρήση της εντολής **svm()** του πακέτου **e1071** (Meyer et al., 2019), για τη δημιουργία διανυσμάτων υποστήριξης με σκοπό την εύρεση του βέλτιστου υπερεπιπέδου διαχωρισμού των παρατηρήσεων. Τελικά, υπολογίστηκαν 433 διανύσματα υποστήριξης για την εύρεση του υπερεπιπέδου, με χρήση ακτινικού πυρήνα (σχέση 2.36), για την ποσοτικοποίηση της ομοιότητας ανάμεσα σε δύο παρατηρήσεις.

Για την καλύτερη ανάπτυξη του μοντέλου με χρήση μηχανών διανυσμάτων υποστήριξης, αφού στη σχέση (2.33) χρησιμοποιήθηκε ως πυρήνας ο ακτινικός, θα ήταν εφικτό μέσω της R να υπολογιστεί η κατάλληλη παράμετρος γ του πυρήνα. Ακόμα, είναι δυνατόν να υπολογιστεί και η κατάλληλη παράμετρος συντονισμού C της σχέσης (2.32) για την εύρεση του κατάλληλου ταξινομητή διανυσμάτων υποστήριξης. Τα προαναφερθέντα γίνονται άμεσα, μέσω της εντολής **tune()** του πακέτου e1071, όπως αναλύεται και στο Πλαίσιο Α.36. Έτσι, αυτή τη φορά, με συντονισμένες κατάλληλα τις παραμέτρους, τα διανύσματα υποστήριξης που δημιουργήθηκαν για την εύρεση του κατάλληλου υπερεπιπέδου διαχωρισμού των παρατηρήσεων είναι 681.

Κεφάλαιο 4

Αξιολόγηση Απόδοσης

Στο Κεφάλαιο 3, αναλύθηκε η πλειοψηφία των βημάτων της αρχιτεκτονικής του συστήματος, για την πρόβλεψη της βιοδραστηριότητας χημικών μορίων, με σκοπό την αναστολή της πρόσδεσης της ισταμίνης στον υποδοχέα H1. Μάλιστα, πραγματοποιήθηκαν όλες οι διαδικασίες μέχρι και την εκπαίδευση των μοντέλων που χρησιμοποιούνται για την ταξινόμηση των χημικών ουσιών. Έτσι, σε αυτό το σημείο είναι αναγκαίο να αξιολογηθούν τα μοντέλα αυτά, ως προς την ικανότητά τους να ταξινομούν τα χημικά μόρια στη σωστή κατηγορία που εκφράζει τη βιοδραστηριότητά τους και να επιλεχθεί το μοντέλο εκείνο που πραγματοποιεί πιο έγκυρες προβλέψεις.

4.1 Αξιολόγηση απόδοσης μοντέλων για τα υποδείγματα εκπαίδευσης

Το πιο σημαντικό για την αξιολόγηση ενός μοντέλου, είναι η ικανότητά του να προβλέπει ορθά νέες παρατηρήσεις, τις οποίες δεν χρησιμοποίησε για να εκπαιδευτεί σε αυτές. Όμως, σαν προκαταρκτικό βήμα, θα ήταν ιδανικό να αναλυθούν οι αποδόσεις των μοντέλων για την πρόβλεψη παρατηρήσεων τις οποίες χρησιμοποίησαν οι αλγόριθμοι για να δημιουργήσουν τους ταξινομητές. Κάτι, τέτοιο -όπως αναλύεται και στα Πλαίσια B.1 - B.12 (βλ. Παράρτημα B)- επιτυγχάνεται με χρήση της εντολής **predict()** του πακέτου **caret** (Kuhn et al., 2019).

Αρχικά, από το υποδείγμα εκπαίδευσης train_binary δημιουργήθηκαν 6 διαφορετικοί ταξινομητές και τα μέτρα αξιολόγησης της προβλεπτικής τους ικανότητας, ως προς τις παρατηρήσεις αυτού του υποδείγματος, παρουσιάζονται στον Πίνακα 4.1. Συγκεκριμένα, την μεγαλύτερη ακρίβεια σωστής πρόβλεψης των παρατηρήσεων τη σημείωσε ο ταξινομητής που δημιουργήθηκε με λογιστική παλινδρόμηση και ποινή Lasso για τον υπολογισμό των συντελεστών του μοντέλου. Διευκρινίζεται, ότι η τιμή λ της μεθόδου Lasso στο συγκεκριμένο μοντέλο είναι αυτή που ελαχιστοποιεί το δείκτη σφάλματος πρόβλεψης της διασταυρωμένης επικύρωσης. Χαρακτηριστικά, επιτεύχθηκε ακρίβεια 95.95%. Βέβαια, πολύ καλές τιμές ακρίβειας (95.84%) καθώς και του συντελεστή kappa (0.7464) σημείωσε και η μέθοδος του δένδρου απόφασης, ενώ υψηλότερο ποσοστό ακριβείας σημείωσε και στην πρόβλεψη των ενεργών χημικών μορίων έναντι του λογιστικού μοντέλου Lasso. Η διαφορά, της ακρίβειας

στην πρόβλεψη των ενεργών χημικών ουσιών, άγγιξε, μάλιστα, το 33%.

Εν συνεχεία, στον Πίνακα 4.2 παρουσιάζονται τα μέτρα αξιολόγησης της ικανότητας σωστών προβλέψεων των παρατηρήσεων του υποδείγματος train_triple, από τους 5 ταξινομητές, που εκπαιδεύτηκαν στα δεδομένα του ίδιου υποδείγματος.

Συγκεκριμένα, τη μεγαλύτερη ακρίβεια σωστής πρόβλεψης των παρατηρήσεων τη σημείωσε ο ταξινομητής που δημιουργήθηκε με τυχαίο δάσος (99.19%), προβλέποντας δηλαδή ορθά την κατηγορία σχεδόν όλων των παρατηρήσεων. Ακόμη, σημείωσε και το μεγαλύτερο δείκτη kappa, καθιστώντας έτσι το μοντέλο τυχαίου δάσους τον πιο ισχυρό υποψήφιο για πρόβλεψη μορίων, που μπορούν να κατηγοριοποιηθούν με τρεις διαφορετικούς τρόπους. Επίσης, σύμφωνα με τους υπολογισμούς των δεικτών sensitivity και specificity, ο ταξινομητής τυχαίου δάσους υπολόγισε ορθά τα μόρια στις κατηγορίες που όντως ανήκουν με ποσοστά επιτυχίας που ξεπερνούν το 90%. Εν αντιθέσει, η πολυωνυμική λογιστική παλινδρόμηση φαίνεται πως οδηγεί στα πιο παραπλανητικά αποτελέσματα, καθώς μόνο το 42% των χημικών μορίων, που κατηγοριοποίησε ως moderately active, άνηκαν όντως σε αυτή την κατηγορία.

Μεγέθη Αξιολόγησης	**Μέθοδοι**					
	Λογιστική	**Stepwise**	**Lasso(1se) Λογιστική**	**Lasso(min) Λογιστική**	**LDA**	**Δένδρο Απόφασης**
Accuracy	0.9179	0.9364	0.9538	0.9595	0.9445	0.9584
Sensitivity	0.915	1	0.9924	0.64935	0.9708	0.9759
Specificity	0.9481	0.2857	0.5584	0.98985	0.6753	0.7792
Kappa	0.6304	0.4216	0.6589	0.7193	0.6538	0.7464

Πίνακας 4.1: Αξιολόγηση πρόβλεψης παρατηρήσεων υποδείγματος train_binary

Μεγέθη Αξιολόγησης	**Κατηγορίες Χημικού Μορίου**	**Μέθοδοι**				
		Δένδρο Απόφασης	**Τυχαίο Δάσος**	**Πολυωνυμική Λογιστική**	**SVM**	**SVM (tuned)**
Accuracy		0.8486	0.9919	0.8335	0.9665	0.985
Kappa		0.6424	0.9819	0.5838	0.924	0.9664
Sensitivity	**inactive**	0.58442	0.97403	0.74026	0.90909	0.94805
	moderately active	0.6337	0.9826	0.42442	0.9186	0.9767
	highly active	0.9416	0.9968	0.9594	0.987	0.9919
Specificity	**inactive**	0.98223	0.99492	0.9835	0.99619	0.99619
	moderately active	0.9307	0.9957	0.95527	0.9827	0.9885
	highly active	0.7229	1	0.5984	0.9438	0.992

Πίνακας 4.2: Αξιολόγηση πρόβλεψης παρατηρήσεων υποδείγματος train_triple

4.2 Αξιολόγηση απόδοσης μοντέλων για τα υποδείγματα δοκιμής

Έχοντας μία πρώτη εικόνα για την αποδοτικότητα των μοντέλων ως προς την πρόβλεψη των παρατηρήσεων στις οποίες εκπαιδεύτηκαν, κρίνεται σημαντικό και αναγκαίο να αξιολογηθεί η ικανότητα των μοντέλων να προβλέψουν ορθά την κατηγορία παρατηρήσεων για τις οποίες δεν είχαν καμία πληροφορία κατά τη διάρκεια ανάπτυξής τους. Κάτι, τέτοιο όπως αναλύεται και στα Πλαίσια B.13 - B.24 (βλ. Παράρτημα B) είναι ιδιαίτερα εύκολο με χρήση της εντολής predict.

Αρχικά, αξιολογήθηκε η ικανότητα των 6 διαφορετικών ταξινομητών -που δημιουργήθηκαν από το υποδείγμα εκπαίδευσης train_binary- να προβλέψουν παρατηρήσεις του υποδείγματος δοκιμής test_binary. Τα αποτελέσματα των μετρητών αξιολόγησης αυτών των μοντέλων παρουσιάζονται στον Πίνακα 4.3. Συγκεκριμένα, μια ενδιαφέρουσα εξέλιξη δείχνει ότι τη μεγαλύτερη ακρίβεια σωστής πρόβλεψης των παρατηρήσεων τη σημείωσε ξανά ο ταξινομητής που δημιουργήθηκε με λογιστική παλινδρόμηση και ποινή Lasso, υπολογισμένη με παράμετρο λ την lambda.min, για την εύρεση των συντελεστών του μοντέλου. Χαρακτηριστικά, επιτεύχθηκε ακρίβεια 93.28%. Βέβαια, αυτή τη φορά οι πολύ καλές τιμές ακρίβειας (92.47%) καθώς και sensitivity (0.96) και specificity (0.62) που σημείωσε η μέθοδος του δένδρου απόφασης, πρέπει να ληφθούν πιο σοβαρά υπόψη, αφού το μοντέλο CART υπολογίζει ορθά ως inactive μεγαλύτερο ποσοστό χημικών μορίων έναντι της μεθόδου με ποινή Lasso. Συγκεκριμένα, το μοντέλο της λογιστικής παλινδρόμησης με μέθοδο συρρίκνωσης Lasso σημείωσε, παρά την υψηλή συνολική του ακρίβεια, πολύ χαμηλά ποσοστά πρόβλεψης ανενεργών χημικών μορίων, αφού το 53% των μορίων που είναι κανονικά ανενεργά μόρια υπολογίστηκαν από τη μέθοδο ως ενεργά. Η τελευταία παρατήρηση προκύπτει από τη γραμμή specificity για τη στήλη της μεθόδου Lasso, όπου σημειώνεται specificity, δηλαδή ορθή αναγνώρισή των ανενεργών μορίων, της τάξης του 43%.

Εν συνεχεία, στον Πίνακα 4.2, παρουσιάζονται τα μέτρα αξιολόγησης της ικανότητας των 5 ταξινομητών, που μοντελοποιήθηκαν με χρήση του υποδείγματος train_triple, να προβλέπουν σωστά τις παρατηρήσεις του υποδείγματος δοκιμής test_triple.

Για ακόμη μία φορά, τη μεγαλύτερη ακρίβεια σωστής πρόβλεψης των παρατηρήσεων με τρίτιμη μεταβλητή απόκρισης, τη σημείωσε ο ταξινομητής που δημιουργήθηκε με τυχαίο δάσος (82.53%). Ακόμη, σημείωσε και το μεγαλύτερο δείκτη kappa, καθιστώντας έτσι το μοντέλο τυχαίου δάσους τον πιο ισχυρό υποψήφιο για πρόβλεψη μορίων που μπορούν να κατηγοριοποιηθούν με τρεις διαφορετικούς τρόπους. Επίσης, σύμφωνα με τους υπολογισμούς των δεικτών sensitivity και specificity, ο ταξινομητής τυχαίου δάσους υπολόγισε ορθά τα μόρια στις κατηγορίες που όντως ανήκουν με ποσοστά που ξεπερνούν αυτά των υπόλοιπων μεθόδων για όλες τις πιθανές κατηγορίες. Το δυσάρεστο, όμως, είναι πως όλα τα μοντέλα εμφάνισαν πολύ χαμηλά ποσοστά στην ακρίβειά τους, να υπολογίζουν ως ανενεργά, τα χημικά μόρια που όντως είναι. Αυτό φαίνεται εύκολα από την παρατήρηση του μετρητή sensitivity για την κατηγορία inactive. Πρακτικά, υπάρχει κίνδυνος να προσδιοριστούν χημικά μόρια ως εν δυνάμει περιπτώσεις αντιισταμινικών ουσιών, ενώ στην πραγματικότητα δεν έχουν τη δυνατότητα να αναστείλλουν την πρόσδεση ισταμίνης στον υποδοχέα H1.

Μεγέθη Αξιολόγησης	**Μέθοδοι**					
	Λογιστική	**Stepwise**	**Lasso(1se) Λογιστική**	**Lasso(min) Λογιστική**	**LDA**	**Δένδρο Απόφασης**
Accuracy	0.8844	0.8817	0.9274	0.9328	0.9059	0.9247
Sensitivity	0.8964	0.8964	0.9852	0.9793	0.9467	0.9556
Specificity	0.7647	0.7353	0.3529	0.4706	0.5	0.6176
Kappa	0.4872	0.4701	0.4362	0.5265	0.4409	0.5585

Πίνακας 4.3: Αξιολόγηση πρόβλεψης παρατηρήσεων υποδείγματος test_binary

Μεγέθη Αξιολόγησης	**Κατηγορίες Χημικού Μορίου**	**Μέθοδοι**				
		Δένδρο Απόφασης	**Τυχαίο Δάσος**	**Πολυωνυμική Λογιστική**	**SVM**	**SVM (tuned)**
Accuracy		0.75	0.8253	0.7608	0.8145	0.7823
Kappa		0.435	0.5752	0.4182	0.5709	0.4405
Sensitivity	**inactive**	0.47059	0.5	0.47059	0.5	0.35294
	moderately active	0.46753	0.5325	0.3247	0.6104	0.41558
	highly active	0.8697	0.954	0.9272	0.9157	0.9464
Specificity	**inactive**	0.97929	0.98817	0.97337	0.99112	0.99408
	moderately active	0.85085	0.939	0.9017	0.9017	0.93559
	highly active	0.6216	0.6126	0.5405	0.6667	0.4595

Πίνακας 4.4: Αξιολόγηση πρόβλεψης παρατηρήσεων υποδείγματος test_triple

4.2.1 Αξιολόγηση προβλέψεων με ταξινομητές μοντελοποιημένους για εξισορροπημένο υποδείγμα εκπαίδευσης

Όπως, αναφέρθηκε και στο Κεφάλαιο 3, δημιουργήθηκαν μοντέλα εκπαιδευμένα σε εξισορροπημένο υποδείγμα δεδομένων, για να μην υπάρχει μεροληψία έναντι της κατηγορίας inactive, λόγω αραιής εμφάνισης ανενεργών μορίων στις παρατηρήσεις. Έτσι, στο σημείο αυτό θα αξιολογηθούν, για την πρόβλεψη των παρατηρήσεων του υποδείγματος δοκιμής, οι 6 ταξινομητές που μοντελοποίησαν το ισορροπημένο σύνολο δεδομένων.

Μεγέθη Αξιολόγησης	**Μέθοδοι**					
	Λογιστική	**Stepwise**	**Lasso(1se) Λογιστική**	**Lasso(min) Λογιστική**	**LDA**	**Δένδρο Απόφασης**
Accuracy	0.8172	0.8038	0.8629	0.8952	0.8522	0.8871
Sensitivity	0.8166	0.8047	0.8728	0.9083	0.858	0.926
Specificity	0.8235	0.7941	0.7647	0.7647	0.7941	0.5
Kappa	0.3677	0.3364	0.4358	0.516	0.4228	0.3853

Πίνακας 4.5: Αξιολόγηση εκπαιδευμένων μοντέλων σε εξισορροπημένο υποδείγμα για τη δίτιμη μεταβλητή απόκρισης activity

Στον Πίνακα 4.5, τη μεγαλύτερη ακρίβεια σωστής πρόβλεψης των παρατηρήσεων τη σημείωσε ο ταξινομητής που δημιουργήθηκε με λογιστική παλινδρόμηση και ποινή Lasso και παράμετρο λ= lambda.min. Για πολλοστή φορά, σημείωσε ακρίβεια περίπου της τάξης του 90%. Σημαντικό, επίσης, είναι το γεγονός, ότι για ακόμα μία φορά, το μοντέλο που παράγει το δένδρο απόφασης έχει τα δεύτερα καλύτερα αποτελέσματα ακριβείας -μετά την λογιστική παλινδρόμηση με μέθοδο συρρίκνωσης Lasso- καθιστώντας το επίσης έναν καλό υποψήφιο για τελικό μοντέλο περιγραφής χημικών μορίων με δύο κατηγορίες περιγραφής της βιοδραστηριότητάς τους. Επίσης, είναι εμφανές, πως ακόμα και με καλύτερη αναλογία ανάμεσα στους κατηγορικούς πληθυσμούς των χημικών μορίων, η λογιστική παλινδρόμηση με βήματα παρουσιάζει τους μικρότερους δείκτες στα μέτρα αξιολόγησης των αποτελεσμάτων της. Αυτό πρακτικά σημαίνει πως η μείωση του αριθμού των επεξηγηματικών μεταβλητών του μοντέλου, παρά τη δημιουργία ενός ιδιαίτερα εύχρηστου μοντέλου, οδηγεί σε πιο παραπλανητικά αποτελέσματα.

4.2.2 Αξιολόγηση προβλέψεων με ταξινομητές μοντελοποιημένους με διασταυρωμένη επικύρωση k-τμημάτων

Έχοντας, αξιολογήσει τα διάφορα μοντέλα για τους δύο διαφορετικούς τύπους δεδομένων, δηλαδή για τις παρατηρήσεις που περιγράφονται από δύο κατηγορίες και από αυτές που περιγράφονται από τρεις, θεωρητικά είναι δυνατόν να προσδιορισθούν τα τελικά μοντέλα. Ωστόσο, για ισχυροποίηση των παραπάνω αποτελεσμάτων στο Πλαίσιο Β.32 (βλ. Παράρτημα Β) εκπαιδεύτηκαν εκ νέου μοντέλα με τις προαναφερθείσες μεθόδους δημιουργίας ταξινομητών, με χρήση της διασταυρωμένη επικύρωσης με k-τμήματα.

Ειδικότερα, τα υποδείγματα εκπαίδευσης χωρίστηκαν σε 10 τμήματα για την εκπαίδευσή τους ενώ, στα Πλαίσια Β.33 - Β.40 (βλ. Παράρτημα Β) υπολογίστηκαν για κάθε μοντέλο τα μέτρα αξιολόγησης για τις προβλέψεις των παρατηρήσεων του υποδείγματος δοκιμής που αντιστοιχεί σε κάθε περίπτωση.

Έτσι, στον Πίνακα 4.6, παρουσιάζονται τα αποτελέσματα για τα μοντέλα που προβλέπουν χημικά μόρια, τα οποία μπορούν να αντιστοιχηθούν σε δύο κατηγορίες βιοδραστηριότητας. Τα αποτελέσματα ήταν πιο ενθαρρυντικά για όλα τα μοντέλα, με εκείνο που δημιουργήθηκε μέσω λογιστικής παλινδρόμησης με Lasso να παρουσιάζει τη μεγαλύτερη ακρίβεια στις συνολικές του προβλέψεις. Έκπληξη, αποτέλεσε το μοντέλο πολλαπλής λογιστικής παλινδρόμησης, που επέδειξε καλύτερα αποτελέσματα σε όλους τους μετρητές του, συγκριτικά με τις προηγούμενες φορές, χωρίς βέβαια αυτό να το καθιστά καλύτερο από το μοντέλο με Lasso. Σημαντικό πρόβλημα όμως, αποτελεί η έλλειψη αρκετών ανενεργών χημικών μορίων μιας και τα επίπεδα του μετρητή specificity για όλα τα μοντέλα είναι ιδιαίτερα χαμηλά, γεγονός που μαρτυρά πως αρκετά ανενεργά χημικά μόρια προβλέφθηκαν λανθασμένα ως ενεργά.

Από την άλλη, στον Πίνακα 4.7, παρουσιάζονται τα αποτελέσματα για τα μοντέλα που προβλέπουν χημικά μόρια, τα οποία μπορούν να αντιστοιχηθούν σε τρεις κατηγορίες βιοδραστηριότητας. Για ακόμα μία φορά, η υπεροχή του τυχαίου δάσους είναι εμφανής καθιστώντας το τον ισχυρότερο υποψήφιο για πρόβλεψη χημικών μορίων, των οποίων η βιοδραστηριότητα περιγράφεται από τις κατη-

γορίες inactive, moderately active, highly active. Σημειώνεται ακόμη, πως το δένδρο απόφασης δεν είναι ιδιαίτερα καλή επιλογή για πρόβλεψη μοντέλων που η μέση μέγιστη ανασταλτική συγκέντρωση στον υποδοχέα ισταμίνη H1 είναι (1-10μm), καθώς υπολογίζουν ορθά ως moderately_active μόνο το 15.54% των χημικών μορίων.

Μεγέθη Αξιολόγησης	**Μέθοδοι**			
	Λογιστική	**Lasso Λογιστική**	**Δένδρο Απόφασης**	**LDA**
Accuracy	0.9247	0.9301	0.922	0.9059
Sensitivity	0.9704	0.9852	0.9882	0.9467
Specificity	0.4706	0.3824	0.2647	0.5
Kappa	0.4932	0.4662	0.3501	0.4409

Πίνακας 4.6: Αξιολόγηση προβλεπτικής ικανότητας binary μοντέλων εκπαιδευμένων με k-fold cross validation

Μεγέθη Αξιολόγησης	**Κατηγορίες Χημικού Μορίου**	**Μέθοδοι**			
		Δένδρο Απόφασης	**Τυχαίο Δάσος**	**SVM**	**Πολυωνυμική Λογιστική**
Accuracy		0.7366	0.8226	0.8011	0.7661
Kappa		0.2699	0.5586	0.4639	0.4284
Sensitivity	**inactive**	0.32353	0.5	0.5	0.5
	moderately active	0.15584	0.48052	0.28571	0.2987
	highly active	0.9617	0.9655	0.9923	0.9387
Specificity	**inactive**	0.98225	0.98817	0.99112	0.97041
	moderately active	0.95254	0.94915	0.97288	0.90847
	highly active	0.2973	0.5766	0.4324	0.5495

Πίνακας 4.7: Αξιολόγηση προβλεπτικής ικανότητας triple μοντέλων εκπαιδευμένων με k-fold cross validation

Κεφάλαιο 5

Συζήτηση

5.1 Συμπεράσματα

Αυτή η μελέτη επικεντρώθηκε στη σύγκριση μοντέλων που παρήχθησαν μέσω μεθόδων μηχανικής μάθησης για τον προσδιορισμό της βιοδραστηριότητας χημικών μορίων, σύμφωνα με φυσικοχημικούς δείκτες που τα περιγράφουν. Απώτερος σκοπός είναι ο προσδιορισμός του καλύτερου μοντέλου και κατ' επέκταση η εύρεση των φυσικοχημικών δεικτών (επεξηγηματικές μεταβλητές του καλύτερο μοντέλο), που προσδιορίζουν την ικανότητα ενός χημικού μορίου να αναστείλει την πρόσδεση ισταμίνης στον υποδοχέα H1. Βέβαια, όπως επισημάνθηκε και στο Κεφάλαιο 3 στα πλαίσια του προσδιορισμού της βιοδραστηριότητας των χημικών μορίων δεν υπάρχει καθολικό όριο της ποσότητας IC_{50} για το διαχωρισμό των χημικών ουσιών σε κατηγορίες σύμφωνα με τη βιοδραστηριότητά τους. Έτσι, με τα ίδια δεδομένα δημιουργήθηκαν και συγκρίθηκαν μοντέλα μηχανικής μάθησης δύο περιπτώσεων. Στην πρώτη περίπτωση, η μεταβλητή απόκρισης, δηλαδή η βιοδραστηριότητα ενός χημικού μορίου, περιγραφόταν από τις εξής δύο κατηγορίες: (α) ικανό και (β) καθόλου ικανό, να αναστείλει την πρόσδεση ισταμίνης στον υποδοχέα H1. Ενώ στη δεύτερη περίπτωση τα μοντέλα κατά τη φάση δημιουργίας τους λάμβαναν υπόψη για τα χημικά μόρια μία επιπλέον κατηγορία που τα κατατάσσει ως μερικώς ικανά να αναστείλουν την ισταμίνη.

Τα ευρήματα των συγκρίσεων λοιπόν για την πρώτη περίπτωση -που τα μοντέλα κατηγοριοποιούνται σε δύο ομάδες- παρουσιάζονται στον Πίνακα 4.3. Συγκεκριμένα, ως ιδανικότερη μέθοδος για την ταξινόμηση παρατηρήσεων που δεν είχε χρησιμοποιήσει το μοντέλο κατά τη φάση δημιουργίας του, αποδείχτηκε η μέθοδος λογιστικής παλινδρόμησης με μέθοδο συρρίκνωσης Lasso. Όμως, όπως επισημάνθηκε και στο Κεφάλαιο 3, τα μοντέλα αυτά, κατά τη φάση δημιουργίας τους, χρησιμοποίησαν παρατηρήσεις το πλήθος, των οποίων δεν ήταν ισορροπημένο στις δύο κατηγορίες της μεταβλητής απόκρισης. Γι' αυτό το λόγο, οι μέθοδοι που δημιούργησαν τα μοντέλα του Πίνακα 4.3 χρησιμοποιήθηκαν εκ νέου για την παραγωγή και τη σύγκριση μοντέλων, που έχουν εκπαιδευτεί αυτή τη φορά σε εξισορροπημένο πλήθος δεδομένων. Σημειώνεται, πως η εξισορρόπηση του πλήθους των παρατηρήσεων πραγματοποιήθηκε μέσω της μεθόδου upsampling. Τα τελικά αποτελέσματα και σε αυτή την περίπτωση, όπως παρουσιάζονται και στον Πίνακα 4.5 υπέδειξαν για ακόμα μία φορά

ως ιδανικότερο μοντέλο περιγραφής των χημικών μορίων εκείνο που είχε δημιουργηθεί μέσω της μεθόδου λογιστικής παλινδρόμησης με μέθοδο συρρίκνωσης Lasso. Επομένως, σύμφωνα με το μοντέλο της λογιστικής παλινδρόμησης με Lasso οι φυσικοχημικοί δείκτες, που αποτελούν σημαντικούς παράγοντες για την περιγραφή της ικανότητας ενός χημικού μορίου να αναστείλει την πρόσδεση ισταμίνης στον υποδοχέα H1 είναι οι εξής: SlogP, SMR, TPSA, NumLipinskiHBA, NumLipinskiHBD, NumRotatableBonds, NumHBD, NumHBA, NumAmideBonds, NumHeteroAtoms, NumAtoms, NumSaturatedRings, NumAliphaticRings, NumAromaticHeterocycles, NumSaturatedHeterocycles, NumAromaticCarbocycles, NumAliphaticCarbocycles, FractionCSP3, Chi2v, Chi3v, Chi4v, Chi1n, Chi3n, HallKierAlpha, kappa2.

Από την άλλη, στα πλαίσια της εύρεσης του καταλληλότερου μοντέλου ταξινόμησης των χημικών παρατηρήσεων σε τρεις διαφορετικές κατηγορίες σύμφωνα με τη βιοδραστηριότητά τους, πραγματοποιήθηκαν συγκρίσεις των οποίων τα ευρήματα παρουσιάζονται στον Πίνακα 4.4. Συγκεκριμένα, ως ιδανικότερη μέθοδος για την ταξινόμηση παρατηρήσεων που δεν είχε χρησιμοποιήσει το μοντέλο κατά τη φάση δημιουργίας του, αποδείχτηκε η μέθοδος τυχαίου δάσους. Άρα, σύμφωνα με το μοντέλο τυχαίου δάσους, το οποίο δομήθηκε από τη δημιουργία 500 δένδρων απόφασης, οι φυσικοχημικοί δείκτες που αποτελούν βασικότερους παράγοντες για την περιγραφή της ικανότητας ενός χημικού μορίου να αναστείλει τη πρόσδεση ισταμίνης στον υποδοχέα H1 είναι οι ακόλουθοι: FractionCSP3, TPSA, HallKierAlpha, Kappa3, kappa2, Chi4v, SlogP, NumAromaticRings, Chi4n, NumRotatableBonds.

Ακόμη, για επαλήθευση των παραπάνω χρησιμοποιήθηκε στην Ενότητα 4.2.2 η μέθοδος διασταυρωμένης επικύρωσης k-τμημάτων για την παραγωγή ακόμα πιο αξιόπιστων μοντέλων, τα οποία συγκρίθηκαν μεταξύ τους μέσω των ευρημάτων των Πινάκων 4.6 και 4.7. Στους δύο αυτούς πίνακες που παρουσιάζονται οι μετρητές αξιολόγησης των διαφόρων μοντέλων ταξινόμησης για τους δύο διαφορετικούς τύπους δεδομένων οι επικρατέστερες σε κάθε κατηγορία ήταν για ακόμα μία φορά η λογιστική παλινδρόμηση με Lasso και το μοντέλο του τυχαίου δάσους. Συνεπώς, σε περίπτωση που υπάρχει χημικό μόριο, για το οποίο έχουν μετρηθεί οι φυσικοχημικοί δείκτες του Κεφαλαίου 3, ο προσδιορισμός της δυνατότητάς του να ενεργοποιηθεί ή όχι κατά τη την προσπάθειά του να αναστείλει την πρόσδεση ισταμίνης στον υποδοχέα H1, καθίσταται ακριβέστερος με χρήση του μοντέλου ταξινόμησης, που δομήθηκε με λογιστική παλινδρόμηση με μέθοδο συρρίκνωσης Lasso. Από την άλλη στην περίπτωση που είναι επιθυμητή η κατάταξή του χημικού μορίου σε μία πιο λεπτομερή κλίμακα τριών επιπέδων -ανάλογα με την ικανότητά του να αναστέλλει σίγουρα, σχεδόν σίγουρα ή καθόλου την πρόσδεση της ισταμίνης- τότε το καταλληλότερο μοντέλο, που μπορεί να χρησιμοποιηθεί είναι αυτό, που δημιουργήθηκε από τη σύνθεση ενός τυχαίου δάσους.

5.2 Παρατηρήσεις και Μελλοντικές επεκτάσεις

Αν και τα παραπάνω, που αναφέρθηκαν επιβεβαιώνουν τη σίγουρη ύπαρξη δύο μοντέλων με αρκετά καλή προβλεπτική ικανότητα πρέπει να λαμβάνεται υπόψιν η πιθανότητα σφάλματος των ταξινομητών λόγω μικρής εμφάνισης παρατηρήσεων στο σύνολο των δεδομένων με ποσότητα IC_{50} μεγαλύτερη

των 10μM. Επισημαίνεται, ότι η παρούσα μελέτη χρησιμοποίησε το προαναφερθέν κατώφλι για το διαχωρισμό των ανενεργών από τα ενεργά ή σχεδόν ενεργά χημικά μόρια στην προσπάθεια τους να αναστείλουν την έκκριση ισταμίνης. Σύμφωνα με τη διεθνή βιβλιογραφία, ενώ αποτελεί ένα πολύ καλό όριο επιθυμητής ποσότητας IC_{50} για τον ανθρώπινο οργανισμό, υπάρχουν και οι περιπτώσεις χημικών ουσιών, όπως η ολοπαταδίνη που έχει $IC_{50} = 559\mu M$ και αποτελεί γνωστό ανταγωνιστή της ισταμίνης που προσδένεται στον υποδοχέα H1, αλλά και σταθεροποιητή των μαστοκυττάρων. Πρακτικά, στην περίπτωση που η ολοπαταδίνη βρισκόταν στο δείγμα που χρησιμοποιήθηκε για τη δημιουργία των μοντέλων θα είχε ταξινομηθεί λανθασμένα ως ανενεργή ουσία επηρεάζοντας έτσι το τελικό μοντέλο.

Όμως, ο ικανοποιητικός αριθμός μορίων, που είχε ελεγχθεί σε εργαστηριακή φάση η δυνατότητά τους να αναστείλουν την ισταμίνη στον υποδοχέα H1, καθώς και πως για αρκετά από τα μόρια της έρευνας είχαν μετρηθεί οι φυσικοχημικοί του δείκτες, από περισσότερα από ένα εργαστήρια, οδηγούν στα αρκετά αξιόπιστα μοντέλα της παρούσας μελέτης. Μοντέλα, που μελλοντικά θα μπορούσαν να βοηθήσουν να εξεταστεί η ικανότητα νέων χημικών μορίων ή ήδη υπαρχόντων που δεν έχουν χρησιμοποιηθεί στα πλαίσια ανταγωνισμού της ισταμίνης, να καταφέρουν να αναστείλουν την πρόσδεσή της στον υποδοχέα H1. Αυτό, είναι ιδιαίτερα βοηθητικό, αφού επιτρέπει μία πιο γρήγορη διαλογή χημικών μορίων ή ενώσεων που θα μπορούσε να ελεγχθεί εργαστηριακά η ικανότητά τους να ανταγωνίζονται την ισταμίνη στην προσπάθειά της να "εισχωρήσει" στον υποδοχέα H1. Έτσι, είναι πιθανό να δημιουργηθεί πιο γρήγορα ένα νέο αντιισταμινικό που θα μπορούσε να βοηθήσει σε αλλεργικές παθήσεις, μιας και όπως υπογραμμίστηκε στο Κεφάλαιο 1 είναι τα συχνότερα αποτελέσματα της πρόσδεσης ισταμίνης συγκεκριμένα στον υποδοχέα H1.

Παραρτήματα

Παράρτημα Α

Κώδικες Κεφαλαίου 3

```
#Load Data
library("readxl")
physchem <- read_excel("C:/Users/orestis/Downloads/Orestis-HRH1-PhysChem.xlsx")
```

Πλαίσιο Α.1: Φόρτωση Δεδομένων

```
str(physchem)
Class 'data.frame': 16385 obs. of 42 variables:
$ chemblId                : chr "CHEMBL263881" "CHEMBL199824" "CHEMBL137781" ...
$ smiles                  : chr "CCN(CC)C(=O)[C@H]1CN(C)[C@@H]2Cc3c[nH]c4cccc(C2=C1)c34" ...
$ value                   : num 1540 150 500 160 19 ...
$ SlogP                   : num 2.91 1.71 3.05 2.84 2.84 ...
$ SMR                     : num 97.8 122.1 100.3 101.6 101.6 ...
$ LabuteASA               : num 143 181 148 148 148 ...
$ TPSA                    : num 39.3 72.6 39.3 51.4 51.4 ...
$ AMW                     : num 323 423 335 338 338 ...
$ ExactMW                 : num 323 422 335 338 338 ...
$ NumLipinskiHBA          : num 4 8 4 5 5 5 5 4 2 2 ...
$ NumLipinskiHBD          : num 1 0 1 2 2 2 2 1 1 0 ...
$ NumRotatableBonds       : num 3 7 1 3 3 3 3 1 3 5 ...
$ NumHBD                  : num 1 0 1 2 2 2 2 1 1 0 ...
$ NumHBA                  : num 2 8 2 2 2 2 2 2 3 2 ...
$ NumAmideBonds           : num 1 0 1 2 2 2 2 1 0 0 ...
$ NumHeteroAtoms          : num 4 8 4 5 5 5 5 4 4 3 ...
$ NumHeavyAtoms           : num 24 31 25 25 25 25 25 25 22 23 ...
$ NumAtoms                : num 49 60 50 51 51 51 51 50 45 46 ...
$ NumRings                : num 4 4 5 4 4 4 4 5 3 3 ...
$ NumAromaticRings        : num 2 3 2 2 2 2 2 2 2 2 ...
$ NumSaturatedRings       : num 0 1 1 0 0 0 0 1 1 0 ...
$ NumAliphaticRings       : num 2 1 3 2 2 2 2 3 1 1 ...
$ NumAromaticHeterocycles : num 1 1 1 1 1 1 1 1 0 1 ...
$ NumSaturatedHeterocycles: num 0 1 1 0 0 0 0 1 1 0 ...
$ NumAliphaticHeterocycles: num 1 1 2 1 1 1 1 2 1 0 ...
$ NumAromaticCarbocycles  : num 1 2 1 1 1 1 1 1 2 1 ...
$ NumSaturatedCarbocycles : num 0 0 0 0 0 0 0 0 0 0 ...
$ NumAliphaticCarbocycles : num 1 0 1 1 1 1 1 1 0 1 ...
$ FractionCSP3            : num 0.45 0.435 0.476 0.45 0.45 ...
$ Chi0v                   : num 14.7 18.2 15.1 15.2 15.2 ...
$ Chi1v                   : num 8.94 10.74 9.38 9.19 9.19 ...
$ Chi2v                   : num 6.94 7.95 8.11 7.05 7.05 ...
$ Chi3v                   : num 5.66 6.02 6.47 5.63 5.63 ...
$ Chi4v                   : num 4.38 4.12 5.25 4.3 4.3 ...
$ Chi1n                   : num 8.94 10.74 9.38 9.19 9.19 ...
$ Chi2n                   : num 6.94 7.95 8.11 7.05 7.05 ...
$ Chi3n                   : num 5.66 6.02 6.47 5.63 5.63 ...
$ Chi4n                   : num 4.38 4.12 5.25 4.3 4.3 ...
$ HallKierAlpha           : num -2.07 -3.13 -2.07 -2.27 -2.27 -2.27 -2.27 ...
$ kappa1                  : num 15.5 21.1 15.2 16.2 16.2 ...
$ kappa2                  : num 5.78 9.34 5.21 6.23 6.23 ...
$ kappa3                  : num 2.2 4.64 2.07 2.63 2.63 ...
```

Πλαίσιο Α.2: Αρχικό Σύνολο Δεδομένων

```
if(sum(is.na(physchem))==0){print (" No missing values")}
[1] "No missing values"
```

Πλαίσιο Α.3: Έλεγχος Ελλιπών Παρατηρήσεων

```
library("dplyr")
physchem <-physchem %>%
  group_by(chemblId) %>%
  summarize_all( funs(mean))
physchem$activity <- ifelse(physchem$value > 10000, "inactive","active")
physchem <- mutate(d,activity_multi = case_when( value > 10000 ~ "inactive",  value <= 1000 ~ "
    highly active",  value <= 10000 & value >  1000 ~ "moderately active"))
physchem$activity_multi<-factor(physchem$activity_multi,levels=c("inactive","active"))
physchem$activity_multi<-factor(physchem$activity_multi,levels=c("inactive","moderately active","
    highly active"))
str(physchem)
Class and 'data.frame': 1237 obs. of  44 variables:
 $ chemblId                : chr  "CHEMBL1000" "CHEMBL100454"  ...
 $ smiles                  : chr  "CCN(CC)C(=O)[C@H]1CN(C)[C@@H]2Cc3c[nH]c4cccc(C2=C1)c34" ...
 $ value                   : num  19.6 6 11000 9100 501.2 ...
 $ SlogP                   : num  3.15 2.62 3.68 3.1 3.22 ...
 $ SMR                     : num  106.2 95.2 124.2 115.2 101.4 ...
 $ LabuteASA               : num  165 139 191 178 154 ...
 $ TPSA                    : num  53 51.1 84.4 75.7 57.8 ...
 $ AMW                     : num  389 328 451 429 357 ...
 $ ExactMW                 : num  388 328 450 428 357 ...
 $ NumLipinskiHBA          : num  5 5 7 6 6 7 6 6 9 4 ...
 $ NumLipinskiHBD          : num  1 2 1 1 1 1 1 1 3 1 ...
 $ NumRotatableBonds       : num  8 1 9 10 7 10 8 10 10 6 ...
 $ NumHBD                  : num  1 2 1 1 1 1 1 1 3 1 ...
 $ NumHBA                  : num  4 6 7 5 4 6 5 5 7 4 ...
 $ NumAmideBonds           : num  0 0 1 0 1 0 1 0 2 0 ...
 $ NumHeteroAtoms          : num  6 6 8 7 6 8 6 8 9 4 ...
 $ NumHeavyAtoms           : num  27 23 32 30 26 32 27 31 32 25 ...
 $ NumAtoms                : num  52 43 58 58 53 59 52 58 67 49 ...
 $ NumRings                : num  3 4 4 4 3 4 3 4 2 3 ...
 $ NumAromaticRings        : num  2 2 3 2 2 2 2 2 1 2 ...
 $ NumSaturatedRings       : num  1 1 1 2 1 2 1 2 1 1 ...
 $ NumAliphaticRings       : num  1 2 1 2 1 2 1 2 1 1 ...
 $ NumAromaticHeterocycles : num  0 1 2 0 1 0 1 0 0 0 ...
 $ NumSaturatedHeterocycles: num  1 1 1 1 1 1 1 1 1 1 ...
 $ NumAliphaticHeterocycles: num  1 2 1 1 1 1 1 1 1 1 ...
 $ NumAromaticCarbocycles  : num  2 1 1 2 1 2 1 2 1 2 ...
 $ NumSaturatedCarbocycles : num  0 0 0 1 0 1 0 1 0 0 ...
 $ NumAliphaticCarbocycles : num  0 0 0 1 0 1 0 1 0 0 ...
 $ FractionCSP3            : num  0.381 0.353 0.333 0.435 0.45 ...
 $ Chi0v                   : num  16 13.9 18.8 17.8 15.3 ...
 $ Chi1v                   : num  9.57 8.68 11.59 12.06 9.26 ...
 $ Chi2v                   : num  7.13 7.12 8.68 9.8 6.35 ...
 $ Chi3v                   : num  5.1 5.4 6.29 7.16 4.57 ...
 $ Chi4v                   : num  3.62 3.94 4.47 5.08 3.09 ...
 $ Chi1n                   : num  9.2 7.87 10.71 10.58 9.26 ...
 $ Chi2n                   : num  6.69 5.98 7.82 8.09 6.35 ...
 $ Chi3n                   : num  4.85 4.36 5.4 5.62 4.57 ...
 $ Chi4n                   : num  3.49 3.01 3.68 3.79 3.09 ...
 $ HallKierAlpha           : num  -1.92 -1.89 -3.1 -2.22 -2.47 -2.73 ...
 $ kappa1                  : num  19.8 14.7 22.1 21 18.3 ...
 $ kappa2                  : num  9.85 5.96 10.48 9.29 9.29 ...
 $ kappa3                  : num  5.78 2.73 5.77 5.47 5.02 ...
 $ activity                : Factor w/ 2 levels "inactive","active": 1 2 ...
 $ activity_multi          : Factor w/ 3 levels "inactive","moderately active",..: 3 3 ...
```

Πλαίσιο Α.4: Τελικό Σύνολο Δεδομένων

```
physchem[,c(1:3)]<- NULL
```

Πλαίσιο Α.5: Διαγραφή περιττών για την επεξεργασία μεταβλητών

```
# Random sample indexes with 70-30 split
train_index <- sample(1:nrow(physchem), 0.7 * nrow(physchem))
test_index <- setdiff(1:nrow(physchem), train_index)
# Build X_train, y_train, X_test, y_test for Binary Response
X_train_binary <- physchem[train_index, -c(40:41)]
y_train_binary <- physchem[train_index, "activity"]
X_test_binary <- physchem[test_index, -(40:41)]
y_test_binary <- physchem[test_index, "activity"]
# Build X_train, y_train, X_test, y_test for Triple Response
X_train_triple <- physchem[train_index, -c(40:41)]
y_train_triple <- physchem[train_index, "activity_multi"]
X_test_triple <- physchem[test_index, -(40:41)]
y_test_triple <- physchem[test_index, "activity_multi"]
```

Πλαίσιο Α.6: Διαχωρισμός Του Συνόλου Δεδομένων σε training και test sets με split 70-30

```
 constant_cols <- whichAreConstant(physchem)
[1] "whichAreConstant: it took me 0.29s to identify 0 constant column(s)"
 double_cols <- whichAreInDouble(physchem)
[1] "whichAreInDouble: it took me 0.06s to identify 0 column(s) to drop."
 bijections_cols <- whichAreBijection(physchem)
[1] "whichAreBijection: it took me 14.61s to identify 0 column(s) to drop."
```

Πλαίσιο Α.7: Φιλτράρισμα "Αχρηστών" Μεταβλητών

```
train_binary <- X_train_binary
train_binary$activity <- y_train_binary
train_triple <- X_train_triple
train_triple$activity_multi <- y_train_triple
```

Πλαίσιο Α.8: Προετοιμασία συνόλων για αριθμητική και γραφική απεικόνιση

```
summary(train_binary)
```

```
     SlogP             SMR            LabuteASA          TPSA             AMW
 Min.   :0.4816   Min.   : 37.00   Min.   : 54.72   Min.   :  3.24   Min.   :122.2
 1st Qu.:3.3277   1st Qu.: 96.07   1st Qu.:142.87   1st Qu.: 29.02   1st Qu.:327.4
 Median :3.9765   Median :112.03   Median :169.75   Median : 51.02   Median :392.5
 Mean   :4.1127   Mean   :116.98   Mean   :177.53   Mean   : 53.63   Mean   :415.0
 3rd Qu.:4.7380   3rd Qu.:131.38   3rd Qu.:204.43   3rd Qu.: 75.71   3rd Qu.:488.0
 Max.   :8.7153   Max.   :249.45   Max.   :389.20   Max.   :179.19   Max.   :923.0

    ExactMW      NumLipinskiHBA   NumLipinskiHBD   NumRotatableBonds   NumHBD
 Min.   :122.1   Min.   : 1.000   Min.   :0.0000   Min.   : 0.000   Min.   :0.0000
 1st Qu.:327.2   1st Qu.: 3.000   1st Qu.:0.0000   1st Qu.: 4.000   1st Qu.:0.0000
 Median :392.2   Median : 5.000   Median :1.0000   Median : 6.000   Median :1.0000
 Mean   :414.5   Mean   : 5.059   Mean   :0.8208   Mean   : 6.465   Mean   :0.7249
 3rd Qu.:487.2   3rd Qu.: 7.000   3rd Qu.:1.0000   3rd Qu.: 8.000   3rd Qu.:1.0000
 Max.   :922.4   Max.   :15.000   Max.   :5.0000   Max.   :19.000   Max.   :5.0000

     NumHBA        NumAmideBonds    NumHeteroAtoms   NumHeavyAtoms    NumAtoms
 Min.   : 1.000   Min.   :0.0000   Min.   : 1.000   Min.   : 9.00   Min.   : 19.00
 1st Qu.: 3.000   1st Qu.:0.0000   1st Qu.: 4.000   1st Qu.:24.00   1st Qu.: 47.00
 Median : 4.000   Median :0.0000   Median : 6.000   Median :28.00   Median : 56.00
 Mean   : 4.547   Mean   :0.4023   Mean   : 6.053   Mean   :29.62   Mean   : 57.63
 3rd Qu.: 6.000   3rd Qu.:1.0000   3rd Qu.: 8.000   3rd Qu.:34.00   3rd Qu.: 65.00
 Max.   :12.000   Max.   :3.0000   Max.   :16.000   Max.   :67.00   Max.   :126.00

    NumRings       NumAromaticRings   NumSaturatedRings   NumAliphaticRings
 Min.   : 1.000   Min.   :0.000      Min.   : 0.000      Min.   : 0.000
 1st Qu.: 3.000   1st Qu.:2.000      1st Qu.: 1.000      1st Qu.: 1.000
 Median : 4.000   Median :2.000      Median : 1.000      Median : 1.000
 Mean   : 3.976   Mean   :2.475      Mean   : 1.098      Mean   : 1.501
 3rd Qu.: 5.000   3rd Qu.:3.000      3rd Qu.: 2.000      3rd Qu.: 2.000
 Max.   :11.000   Max.   :5.000      Max.   :11.000      Max.   :11.000

 NumAromaticHeterocycles  NumSaturatedHeterocycles  NumAliphaticHeterocycles
 Min.   :0.0000           Min.   :0.0000           Min.   :0.000
 1st Qu.:0.0000           1st Qu.:0.0000           1st Qu.:1.000
 Median :1.0000           Median :1.0000           Median :1.000
 Mean   :0.7225           Mean   :0.9341           Mean   :1.191
 3rd Qu.:1.0000           3rd Qu.:1.0000           3rd Qu.:2.000
 Max.   :4.0000           Max.   :3.0000           Max.   :3.000

 NumAromaticCarbocycles  NumSaturatedCarbocycles  NumAliphaticCarbocycles  FractionCSP3
 Min.   :0.000           Min.   : 0.0000          Min.   : 0.0000          Min.   :0.04545
 1st Qu.:1.000           1st Qu.: 0.0000          1st Qu.: 0.0000          1st Qu.:0.35000
 Median :2.000           Median : 0.0000          Median : 0.0000          Median :0.40000
 Mean   :1.753           Mean   : 0.1642          Mean   : 0.3098          Mean   :0.41737
 3rd Qu.:2.000           3rd Qu.: 0.0000          3rd Qu.: 0.0000          3rd Qu.:0.46667
 Max.   :4.000           Max.   :11.0000          Max.   :11.0000          Max.   :1.00000

     Chi0v            Chi1v            Chi2v            Chi3v            Chi4v
 Min.   : 5.248   Min.   : 3.032   Min.   : 1.860   Min.   : 1.127   Min.   : 0.6939
 1st Qu.:14.349   1st Qu.: 8.765   1st Qu.: 6.929   1st Qu.: 5.001   1st Qu.: 3.5783
 Median :16.923   Median :10.467   Median : 8.109   Median : 5.893   Median : 4.2687
 Mean   :17.717   Mean   :10.858   Mean   : 8.474   Mean   : 6.275   Mean   : 4.5236
 3rd Qu.:20.101   3rd Qu.:12.440   3rd Qu.: 9.779   3rd Qu.: 7.365   3rd Qu.: 5.3125
 Max.   :38.665   Max.   :22.514   Max.   :16.304   Max.   :12.650   Max.   :13.0642

     Chi1n            Chi2n            Chi3n            Chi4n          HallKierAlpha
 Min.   : 3.032   Min.   : 1.860   Min.   : 1.127   Min.   : 0.6939   Min.   :-6.940
 1st Qu.: 8.495   1st Qu.: 6.567   1st Qu.: 4.733   1st Qu.: 3.3658   1st Qu.:-2.810
 Median :10.161   Median : 7.805   Median : 5.607   Median : 3.9767   Median :-2.240
 Mean   :10.466   Mean   : 8.019   Mean   : 5.882   Mean   : 4.2364   Mean   :-2.431
 3rd Qu.:11.852   3rd Qu.: 9.105   3rd Qu.: 6.838   3rd Qu.: 4.9141   3rd Qu.:-1.870
 Max.   :22.514   Max.   :16.304   Max.   :12.650   Max.   :13.0642   Max.   :-0.040

     kappa1           kappa2           kappa3          activity
 Min.   : 6.233   Min.   : 2.136   Min.   : 0.5136   inactive: 80
 1st Qu.:16.091   1st Qu.: 6.927   1st Qu.: 3.3017   active  :785
 Median :19.279   Median : 9.004   Median : 4.6816
 Mean   :20.579   Mean   : 9.236   Mean   : 4.9065
 3rd Qu.:24.145   3rd Qu.:10.920   3rd Qu.: 6.1597
 Max.   :49.493   Max.   :22.989   Max.   :12.1005
```

Πλαίσιο Α.9: Υπολογισμός Αριθμητικών Περιγραφικών Δεικτών Των Μεταβλητών για υποδείγμα εκπαίδευσης δίτιμης μεταβλητής απόκρισης activity

```
summary(train_triple)
     SlogP             SMR            LabuteASA          TPSA             AMW
 Min.   :0.4816   Min.   : 37.00   Min.   : 54.72   Min.   :  3.24   Min.   :122.2
 1st Qu.:3.3277   1st Qu.: 96.07   1st Qu.:142.87   1st Qu.: 29.02   1st Qu.:327.4
 Median :3.9765   Median :112.03   Median :169.75   Median : 51.02   Median :392.5
 Mean   :4.1127   Mean   :116.98   Mean   :177.53   Mean   : 53.63   Mean   :415.0
 3rd Qu.:4.7380   3rd Qu.:131.38   3rd Qu.:204.43   3rd Qu.: 75.71   3rd Qu.:488.0
 Max.   :8.7153   Max.   :249.45   Max.   :389.20   Max.   :179.19   Max.   :923.0

    ExactMW        NumLipinskiHBA   NumLipinskiHBD   NumRotatableBonds    NumHBD
 Min.   :122.1   Min.   : 1.000   Min.   :0.0000   Min.   : 0.000   Min.   :0.0000
 1st Qu.:327.2   1st Qu.: 3.000   1st Qu.:0.0000   1st Qu.: 4.000   1st Qu.:0.0000
 Median :392.2   Median : 5.000   Median :1.0000   Median : 6.000   Median :1.0000
 Mean   :414.5   Mean   : 5.059   Mean   :0.8208   Mean   : 6.465   Mean   :0.7249
 3rd Qu.:487.2   3rd Qu.: 7.000   3rd Qu.:1.0000   3rd Qu.: 8.000   3rd Qu.:1.0000
 Max.   :922.4   Max.   :15.000   Max.   :5.0000   Max.   :19.000   Max.   :5.0000

     NumHBA         NumAmideBonds    NumHeteroAtoms   NumHeavyAtoms    NumAtoms
 Min.   : 1.000   Min.   :0.0000   Min.   : 1.000   Min.   : 9.00   Min.   : 19.00
 1st Qu.: 3.000   1st Qu.:0.0000   1st Qu.: 4.000   1st Qu.:24.00   1st Qu.: 47.00
 Median : 4.000   Median :0.0000   Median : 6.000   Median :28.00   Median : 56.00
 Mean   : 4.547   Mean   :0.4023   Mean   : 6.053   Mean   :29.62   Mean   : 57.63
 3rd Qu.: 6.000   3rd Qu.:1.0000   3rd Qu.: 8.000   3rd Qu.:34.00   3rd Qu.: 65.00
 Max.   :12.000   Max.   :3.0000   Max.   :16.000   Max.   :67.00   Max.   :126.00

    NumRings       NumAromaticRings   NumSaturatedRings   NumAliphaticRings
 Min.   : 1.000   Min.   :0.000      Min.   : 0.000      Min.   : 0.000
 1st Qu.: 3.000   1st Qu.:2.000      1st Qu.: 1.000      1st Qu.: 1.000
 Median : 4.000   Median :2.000      Median : 1.000      Median : 1.000
 Mean   : 3.976   Mean   :2.475      Mean   : 1.098      Mean   : 1.501
 3rd Qu.: 5.000   3rd Qu.:3.000      3rd Qu.: 2.000      3rd Qu.: 2.000
 Max.   :11.000   Max.   :5.000      Max.   :11.000      Max.   :11.000

 NumAromaticHeterocycles NumSaturatedHeterocycles NumAliphaticHeterocycles
 Min.   :0.0000          Min.   :0.0000           Min.   :0.000
 1st Qu.:0.0000          1st Qu.:0.0000           1st Qu.:1.000
 Median :1.0000          Median :1.0000           Median :1.000
 Mean   :0.7225          Mean   :0.9341           Mean   :1.191
 3rd Qu.:1.0000          3rd Qu.:1.0000           3rd Qu.:2.000
 Max.   :4.0000          Max.   :3.0000           Max.   :3.000

 NumAromaticCarbocycles NumSaturatedCarbocycles NumAliphaticCarbocycles FractionCSP3
 Min.   :0.000          Min.   : 0.0000         Min.   : 0.0000         Min.   :0.04545
 1st Qu.:1.000          1st Qu.: 0.0000         1st Qu.: 0.0000         1st Qu.:0.35000
 Median :2.000          Median : 0.0000         Median : 0.0000         Median :0.40000
 Mean   :1.753          Mean   : 0.1642         Mean   : 0.3098         Mean   :0.41737
 3rd Qu.:2.000          3rd Qu.: 0.0000         3rd Qu.: 0.0000         3rd Qu.:0.46667
 Max.   :4.000          Max.   :11.0000         Max.   :11.0000         Max.   :1.00000

     Chi0v            Chi1v            Chi2v            Chi3v            Chi4v
 Min.   : 5.248   Min.   : 3.032   Min.   : 1.860   Min.   : 1.127   Min.   : 0.6939
 1st Qu.:14.349   1st Qu.: 8.765   1st Qu.: 6.929   1st Qu.: 5.001   1st Qu.: 3.5783
 Median :16.923   Median :10.467   Median : 8.109   Median : 5.893   Median : 4.2687
 Mean   :17.717   Mean   :10.858   Mean   : 8.474   Mean   : 6.275   Mean   : 4.5236
 3rd Qu.:20.101   3rd Qu.:12.440   3rd Qu.: 9.779   3rd Qu.: 7.365   3rd Qu.: 5.3125
 Max.   :38.665   Max.   :22.514   Max.   :16.304   Max.   :12.650   Max.   :13.0642

     Chi1n            Chi2n            Chi3n            Chi4n          HallKierAlpha
 Min.   : 3.032   Min.   : 1.860   Min.   : 1.127   Min.   : 0.6939   Min.   :-6.940
 1st Qu.: 8.495   1st Qu.: 6.567   1st Qu.: 4.733   1st Qu.: 3.3658   1st Qu.:-2.810
 Median :10.161   Median : 7.805   Median : 5.607   Median : 3.9767   Median :-2.240
 Mean   :10.466   Mean   : 8.019   Mean   : 5.882   Mean   : 4.2364   Mean   :-2.431
 3rd Qu.:11.852   3rd Qu.: 9.105   3rd Qu.: 6.838   3rd Qu.: 4.9141   3rd Qu.:-1.870
 Max.   :22.514   Max.   :16.304   Max.   :12.650   Max.   :13.0642   Max.   :-0.040

     kappa1           kappa2           kappa3          activity_multi
 Min.   : 6.233   Min.   : 2.136   Min.   : 0.5136   inactive        : 80
 1st Qu.:16.091   1st Qu.: 6.927   1st Qu.: 3.3017   moderately active:176
 Median :19.279   Median : 9.004   Median : 4.6816   highly active   :609
 Mean   :20.579   Mean   : 9.236   Mean   : 4.9065
 3rd Qu.:24.145   3rd Qu.:10.920   3rd Qu.: 6.1597
 Max.   :49.493   Max.   :22.989   Max.   :12.1005
```

Πλαίσιο Α.10: Υπολογισμός Αριθμητικών Περιγραφικών Δεικτών Των Μεταβλητών για υποδείγμα εκπαίδευσης τρίτιμης μεταβλητής απόκρισης activity_multi

```
# class distribution
cbind(freq=table(train_binary$activity), percentage=prop.table(table(train_binary$
    activity))*100)
          freq percentage
inactive    80   9.248555
active     785  90.751445
```

Πλαίσιο Α.11: Ποσοστά εμφάνισης του κάθε επιπέδου της μεταβλητής activity

```
#class distribution
cbind(freq=table(train_triple$activity_multi), percentage=prop.table(table(train_
    triple$activity_multi))*100)
                   freq percentage
inactive             77   8.901734
moderately active   172  19.884393
highly active       616  71.213873
```

Πλαίσιο Α.12: Ποσοστά εμφάνισης του κάθε επιπέδου της μεταβλητής activity_multi

```
# Up Sample.
library(caret)
'%ni%' <- Negate('%in%')  # define 'not in' func
options(scipen=999)  # prevents printing scientific notations.
set.seed(100)
up_train <- upSample(x = train_binary[, colnames(train_binary) %ni% "activity"], y
    = train_binary$activity)
cbind(freq=table(up_train$activity), percentage=prop.table(table(up_train$activity
    ))*100)
          freq percentage
inactive   785         50
active     785         50
```

Πλαίσιο Α.13: Εξισορρόπηση με μέθοδο Up-Sampling

```
# histogram for each attribute
par(mfrow=c(3,3))
for(i in 1:39) {  hist(train_binary[,i], main=names(train_binary)[i])}
```

Πλαίσιο Α.14: Υπολογισμός Ιστογραμμάτων

```
# boxplots for each attribute
par(mfrow=c(3,3))
for(i in 1:39) {  boxplot(train_binary[,i], main=names(train_binary)[i])}
```

Πλαίσιο Α.15: Υπολογισμός Θηκοδιαγραμμάτων

```
require(GGally)
p <- train_binary %>% ggpairs(.,columns = 1:39, mapping = ggplot2::aes(color=
    activity),
                              lower = list(continuous = wrap("smooth", alpha = 0.3,
                                  size=0.1),
                              discrete = "blank", combo="blank"),
                              diag = list(discrete="barDiag", continuous = wrap("
                                  densityDiag", alpha=0.5 )),
                              upper = list(combo = wrap("box_no_facet", alpha=0.5),
                                  continuous = wrap("cor", size=4, alignPercent=0.8))
                                  ) +  theme(panel.grid.major = element_blank())     #
                                   remove gridlines
plots <- apply(expand.grid(i=1:39, j=1:p$ncol), 1, function(ij) getPlot(p, i=ij
    [1], j=ij[2]))
ggmatrix(plots, byrow=FALSE, nrow = 39, ncol=p$ncol, xAxisLabels = p$xAxisLabels
    [1:p$ncol], yAxisLabels = p$yAxisLabels,
         title="Density Functions, Corellation,Scatter")
```

Πλαίσιο Α.16: Γραφήματα κατανομής πυκνότητας για τη δίτιμη μεταβλητή απόκρισης activity

```
p <- train_triple %>% ggpairs(.,columns = 1:39, title = "A Nice Plot Using ggpairs
    ",
                mapping = ggplot2::aes(color=activity_multi), lower = list(
                    continuous = wrap("smooth", alpha = 0.3, size=0.1),
                                discrete = "blank", combo="blank"),
                    diag = list(discrete="barDiag",
                                continuous = wrap("densityDiag", alpha=0.5 )),
                    upper = list(combo = wrap("box_no_facet", alpha=0.5),
                                continuous = wrap("cor", size=4, alignPercent
                                    =0.8))) +
  theme(panel.grid.major = element_blank())     # remove gridlines
plots <- apply(expand.grid(i=1:39, j=1:p$ncol), 1, function(ij) getPlot(p, i=ij
    [1], j=ij[2]))
ggmatrix(plots, byrow=FALSE, nrow = 39, ncol=p$ncol, xAxisLabels = p$xAxisLabels
    [1:p$ncol],  yAxisLabels = p$yAxisLabels,
         title="Density Functions, Corellation,Scatter")
```

Πλαίσιο Α.17: Γραφήματα κατανομής πυκνότητας για την τρίτιμη μεταβλητή απόκρισης activity_multi

```
ggcorr(train_binary, method = c("everything", "pearson"))
```

Πλαίσιο Α.18: Υπολογισμός συσχέτισης μεταβλητών

```
logistic = glm(activity~., data=train_binary,family=binomial(link='logit'))
summary(logistic)
Call:
glm(formula = activity ~ ., family = binomial(link = "logit"),
    data = train_binary)
 Deviance Residuals:
     Min        1Q    Median        3Q       Max
-3.12384   0.00000   0.00264   0.10658   1.91926
Coefficients: (4 not defined because of singularities)
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              2.002e-01  7.509e+00   0.027   0.9787
SlogP                    2.328e+00  9.289e-01   2.507   0.0122 *
SMR                      7.037e-01  3.934e-01   1.789   0.0736 .
LabuteASA               -4.783e-01  5.348e-01  -0.894   0.3711
TPSA                    -4.194e-01  2.295e-01  -1.827   0.0677 .
AMW                     -7.197e+00  7.739e+00  -0.930   0.3524
ExactMW                  7.199e+00  7.797e+00   0.923   0.3559
NumLipinskiHBA           9.650e+00  4.963e+00   1.944   0.0519 .
NumLipinskiHBD           2.692e+00  3.044e+00   0.884   0.3764
NumRotatableBonds       -2.660e-01  3.475e-01  -0.765   0.4440
NumHBD                  -3.582e-01  1.334e+00  -0.268   0.7884
NumHBA                  -9.440e-01  8.037e-01  -1.175   0.2401
NumAmideBonds           -1.243e+00  9.978e-01  -1.245   0.2130
NumHeteroAtoms          -3.105e+00  2.695e+00  -1.152   0.2491
NumHeavyAtoms            3.389e+00  4.565e+00   0.742   0.4579
NumAtoms                -3.225e+00  1.864e+00  -1.730   0.0836 .
NumRings                 1.675e+01  2.048e+03   0.008   0.9935
NumAromaticRings        -1.821e+01  2.048e+03  -0.009   0.9929
NumSaturatedRings       -1.829e+01  2.048e+03  -0.009   0.9929
NumAliphaticRings               NA         NA      NA       NA
NumAromaticHeterocycles -2.193e+00  9.227e-01  -2.377   0.0175 *
NumSaturatedHeterocycles -1.380e+00 2.438e+03  -0.001   0.9995
NumAliphaticHeterocycles -4.439e-01 2.438e+03   0.000   0.9999
NumAromaticCarbocycles          NA         NA      NA       NA
NumSaturatedCarbocycles         NA         NA      NA       NA
NumAliphaticCarbocycles         NA         NA      NA       NA
FractionCSP3             9.041e+00  1.044e+01   0.866   0.3867
Chi0v                   -5.526e-01  3.601e+00  -0.153   0.8781
Chi1v                    2.908e+00  4.028e+00   0.722   0.4703
Chi2v                   -7.743e-03  4.840e+00  -0.002   0.9987
Chi3v                   -8.214e+00  5.113e+00  -1.607   0.1081
Chi4v                    4.664e+00  3.837e+00   1.216   0.2242
Chi1n                   -1.817e+00  6.754e+00  -0.269   0.7879
Chi2n                    3.540e+00  4.872e+00   0.727   0.4675
Chi3n                    1.144e+01  5.845e+00   1.956   0.0504 .
Chi4n                   -5.731e+00  4.561e+00  -1.257   0.2089
HallKierAlpha            8.174e+00  6.097e+00   1.341   0.1800
kappa1                   2.410e+00  3.404e+00   0.708   0.4789
kappa2                   2.282e+00  1.571e+00   1.453   0.1462
kappa3                  -1.053e+00  1.024e+00  -1.028   0.3041
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 519.45  on 864  degrees of freedom
Residual deviance: 171.56  on 829  degrees of freedom
AIC: 243.56
Number of Fisher Scoring iterations: 20
```

Πλαίσιο Α.19: Υπολογισμός μοντέλου με λογιστική παλινδρόμηση

```
vif(glm(activity~.-NumAromaticCarbocycles-NumSaturatedCarbocycles-NumAliphaticCarbocycles
    -NumAliphaticRings, data=train_binary,family=binomial(link='logit')))

SlogP                       SMR                LabuteASA          TPSA               AMW
2.356208e+01                1.717902e+03       8.556583e+03       1.067129e+03       1.084865e+07

ExactMW                     NumLipinskiHBA     NumLipinskiHBD     NumRotatableBonds
1.099272e+07                2.954049e+03       2.387565e+02       3.100048e+01

NumHBD                      NumHBA             NumAmideBonds      NumHeteroAtoms
2.903822e+01                4.188631e+01       1.282838e+01       8.632060e+02

NumHeavyAtoms               NumAtoms           NumRings           NumAromaticRings
1.826871e+04                1.608892e+04       6.605327e+07       6.055003e+07

NumSaturatedRings           NumAromaticHeterocycles  NumSaturatedHeterocycles
7.031548e+07                1.311087e+01             4.552427e+07

NumAliphaticHeterocycles    FractionCSP3       Chi0v              Chi1v
4.552424e+07                6.852768e+01       4.396714e+03       2.107796e+03

Chi2v                       Chi3v              Chi4v              Chi1n
2.269129e+03                1.057227e+03       3.127972e+02       5.640650e+03

Chi2n                       Chi3n              Chi4n
2.077931e+03                1.055572e+03       3.010109e+02

HallKierAlpha               kappa1                   kappa2             kappa3
3.094256e+02                8.028400e+03             5.111822e+02       1.609724e+02
```

Πλαίσιο Α.20: Έλεγχος πολυσυγγραμικότητας επεξηγηματικών μεταβλητών λογιστικού μοντέλου

```
require(MASS)
stepwise <- stepAIC(logistic,direction = "both")
summary(stepwise)
Call:
glm(formula = activity ~ SlogP + SMR + TPSA + AMW + ExactMW +
    NumLipinskiHBA + NumHBA + NumAmideBonds + NumAtoms + NumRings +
    NumAromaticRings + NumSaturatedRings + NumAromaticHeterocycles +
    Chi3v + Chi4v + Chi3n + Chi4n + HallKierAlpha, family = binomial(link = "logit"),
    data = train_binary)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3877   0.0000   0.0027   0.1045   2.0635
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -1.3149     1.8515  -0.710 0.477575
SlogP                        1.7319     0.6065   2.856 0.004296 **
SMR                          0.8513     0.2067   4.118 3.82e-05 ***
TPSA                        -0.1692     0.0285  -5.935 2.93e-09 ***
AMW                        -10.8994     3.5886  -3.037 0.002387 **
ExactMW                     10.9496     3.6025   3.039 0.002370 **
NumLipinskiHBA               5.4363     0.9903   5.490 4.03e-08 ***
NumHBA                      -1.6318     0.5094  -3.203 0.001358 **
NumAmideBonds               -1.3463     0.7132  -1.888 0.059061 .
NumAtoms                    -1.9337     0.3762  -5.140 2.75e-07 ***
NumRings                    15.5780  1157.2996   0.013 0.989260
NumAromaticRings           -17.3627  1157.2996  -0.015 0.988030
NumSaturatedRings          -18.4437  1157.2994  -0.016 0.987285
NumAromaticHeterocycles     -2.6065     0.7062  -3.691 0.000224 ***
Chi3v                       -9.4373     2.9416  -3.208 0.001336 **
Chi4v                        5.1981     2.7951   1.860 0.062928 .
Chi3n                       12.2240     3.5336   3.459 0.000541 ***
Chi4n                       -4.6297     3.0476  -1.519 0.128732
HallKierAlpha                8.3550     2.4175   3.456 0.000548 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 519.45  on 864  degrees of freedom
Residual deviance: 179.85  on 846  degrees of freedom
AIC: 217.85
Number of Fisher Scoring iterations: 20
```

Πλαίσιο Α.21: Υπολογισμός μοντέλου λογιστικής παλινδρόμησης με χρήση μεθόδου με βήματα

```
vif(stepwise)
SlogP                      SMR                      TPSA                     AMW
1.006373e+0                5.127395e+02             1.800932e+01             2.528209e+06

ExactMW                    NumLipinskiHBA           NumHBA                   NumAmideBonds
2.542654e+06               1.251759e+02             1.839151e+01             6.410233e+00

NumAtoms                   NumRings                 NumAromaticRings         NumSaturatedRings
6.860387e+02               2.353799e+07             2.029730e+07             2.537343e+07

NumAromaticHeterocycles    Chi3v                    Chi4v                    Chi3n
 7.975694e+00              3.963151e+02             1.941468e+02             4.264067e+02

 Chi4n                     HallKierAlpha
 1.615640e+02              5.154160e+01
```

Πλαίσιο Α.22: Έλεγχος πολυσυγγραμικότητας επεξηγηματικών μεταβλητών λογιστικού μοντέλου με βήματα

```
anova(logistic,stepwise,test="Chisq")
Analysis of Deviance Table
Model 1: activity ~ SlogP + SMR + LabuteASA + TPSA + AMW + ExactMW +
    NumLipinskiHBA +
    NumLipinskiHBD + NumRotatableBonds + NumHBD + NumHBA + NumAmideBonds +
    NumHeteroAtoms + NumHeavyAtoms + NumAtoms + NumRings + NumAromaticRings +
    NumSaturatedRings + NumAliphaticRings + NumAromaticHeterocycles +
    NumSaturatedHeterocycles + NumAliphaticHeterocycles + NumAromaticCarbocycles +
    NumSaturatedCarbocycles + NumAliphaticCarbocycles + FractionCSP3 +
    Chi0v + Chi1v + Chi2v + Chi3v + Chi4v + Chi1n + Chi2n + Chi3n +
    Chi4n + HallKierAlpha + kappa1 + kappa2 + kappa3
Model 2: activity ~ SlogP + SMR + TPSA + AMW + ExactMW + NumLipinskiHBA +
    NumHBA + NumAmideBonds + NumAtoms + NumRings + NumAromaticRings +
    NumSaturatedRings + NumAromaticHeterocycles + Chi3v + Chi4v +
    Chi3n + Chi4n + HallKierAlpha
  Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
1       829     171.56
2       846     179.85 -17  -8.2949   0.9598
```

Πλαίσιο Α.23: Σύγκριση μοντέλων λογιστικής παλινδρόμησης και κατά βήματα

```
require(glmnet)
bioact<-as.factor(y_train_binary)
xlasso<-model.matrix(bioact~.,X_train_binary)[,-1]
set.seed(123)
bioact<-as.factor(bioact)
cv.lasso<-cv.glmnet(xlasso,bioact,alpha=1,family="binomial",type="class")
plot(cv.lasso)
```

Πλαίσιο Α.24: Lasso παλινδρόμηση με cross-validation

```
cv.lasso$lambda.min
[1] 0.0002205826

cv.lasso$lambda.1se
[1] 0.002984588
```

Πλαίσιο Α.25: Κυριότεροι δείκτες λ

```
  coef(cv.lasso,cv.lasso$lambda.min)
40 x 1 sparse Matrix of class "dgCMatrix"
                                      1
(Intercept)                -2.8369905677
SlogP                       0.2987887021
SMR                         0.1545846927
LabuteASA                   .
TPSA                       -0.1251983386
AMW                         .
ExactMW                     .
NumLipinskiHBA              1.0442555629
NumLipinskiHBD             -1.1111343585
NumRotatableBonds          -0.3342704701
NumHBD                      1.3719071694
NumHBA                     -0.0899897865
NumAmideBonds              -0.4148804433
NumHeteroAtoms              0.7412730890
NumHeavyAtoms               .
NumAtoms                   -0.4256625578
NumRings                    .
NumAromaticRings            .
NumSaturatedRings          -3.3525241828
NumAliphaticRings           2.9252465200
NumAromaticHeterocycles    -1.0771456297
NumSaturatedHeterocycles   -0.7308830769
NumAliphaticHeterocycles    .
NumAromaticCarbocycles      0.5612878420
NumSaturatedCarbocycles     .
NumAliphaticCarbocycles     0.2282036595
FractionCSP3                2.2415559264
Chi0v                       .
Chi1v                       .
Chi2v                      -0.2412545634
Chi3v                      -1.1030935663
Chi4v                       0.0004068327
Chi1n                      -0.0001213647
Chi2n                       .
Chi3n                       2.8685786111
Chi4n                       .
HallKierAlpha               0.2264977808
kappa1                      .
kappa2                      0.5602836431
kappa3                      .
```

Πλαίσιο Α.26: Συντελεστές μοντέλου για λ=lambda.min

```
coef(cv.lasso,cv.lasso$lambda.1se)
40 x 1 sparse Matrix of class "dgCMatrix"
                                      1
(Intercept)                -0.70184478
SlogP                       0.27334031
SMR                         .
LabuteASA                   .
TPSA                       -0.02395650
AMW                         .
ExactMW                     .
NumLipinskiHBA              .
NumLipinskiHBD             -1.32847987
NumRotatableBonds          -0.28899365
NumHBD                      0.99309145
NumHBA                      .
NumAmideBonds               .
NumHeteroAtoms              0.34864421
NumHeavyAtoms               .
NumAtoms                    .
NumRings                    .
NumAromaticRings            .
NumSaturatedRings          -0.64227609
NumAliphaticRings           0.75509477
NumAromaticHeterocycles     .
NumSaturatedHeterocycles    .
NumAliphaticHeterocycles    .
NumAromaticCarbocycles      0.98539194
NumSaturatedCarbocycles     .
NumAliphaticCarbocycles     .
FractionCSP3               -0.94359997
Chi0v                       .
Chi1v                       .
Chi2v                      -0.06268227
Chi3v                       .
Chi4v                       .
Chi1n                       .
Chi2n                       .
Chi3n                       .
Chi4n                       0.79792649
HallKierAlpha               .
kappa1                      .
kappa2                      0.04746081
kappa3                      .
```

Πλαίσιο Α.27: Συντελεστές μοντέλου για λ=lambda.1se

```
lda.model = lda (activity~., data=train_binary)
Warning message:
In lda.default(x, grouping, ...) : variables are collinear
Call:
Prior probabilities of groups:
  inactive     active
0.08901734 0.91098266
Group means:
             SlogP       SMR LabuteASA      TPSA      AMW ExactMW NumLipinskiHBA    NumLipinskiHBD
inactive  2.996580  98.72613  148.6115  64.82364 344.9320 344.6350       4.909091         1.8571429
active    4.182342 118.44977  179.7495  53.09797 419.9152 419.3998       5.081218         0.7690355

          NumRotatableBonds    NumHBD    NumHBA NumAmideBonds NumHeteroAtoms NumHeavyAtoms NumAtoms
inactive           8.246753 1.272727  4.233766      0.4805195       5.272727       24.70130 52.20779
active             6.189086 0.713198  4.545685      0.4175127       6.093909       30.00254 57.91117


                      NumRings                NumAromaticRings                NumSaturatedRings
inactive              2.584416                        1.675325                        0.9090909
active                4.105330                        2.538071                        1.1230964

                NumAliphaticRings          NumAromaticHeterocycles          NumSaturatedHeterocycles
inactive                0 .9090909                        0.7272727                        0.6493506
active                   1.5672589                        0.7068528                        0.9619289

          NumAliphaticHeterocycles          NumAromaticCarbocycles          NumSaturatedCarbocycles
inactive                 0.6493506                        0.9480519                        0.2597403
active                   1.2398477                        1.8312183                        0.1611675


          NumAliphaticCarbocycles FractionCSP3      Chi0v     Chi1v    Chi2v    Chi3v    Chi4v
inactive               0.2597403    0.5168118   15.05821  9.290437 7.048281 4.841572 3.300979
active                 0.3274112    0.4035995   17.90119 10.952550 8.563481 6.386772 4.629051

                Chi1n    Chi2n    Chi3n    Chi4n HallKierAlpha   kappa1   kappa2   kappa3
inactive     9.014534 6.712272 4.538265 3.058235    -1.890130 18.30013 9.206698 5.750724
active      10.559536 8.113417 5.999068 4.345282    -2.497265 20.71508 9.189050 4.776957

Coefficients of linear discriminants:

                                LD1
SlogP                  0.170212049
SMR                    0.049327673
LabuteASA             -0.054904488
TPSA                  -0.074251502
AMW                    0.768718523
ExactMW               -0.794592940
NumLipinskiHBA         1.315894148
NumLipinskiHBD        -1.970664773
NumRotatableBonds     -0.130023439
NumHBD                 2.318214428
NumHBA                 0.058335425
NumAmideBonds          0.290014477
NumHeteroAtoms        -0.280848196
NumHeavyAtoms          0.714007584
NumAtoms              -0.394811496
NumRings              -0.236117969
NumAromaticRings      -0.222261589
NumSaturatedRings     -0.163060598
NumAliphaticRings     -0.156276782
NumAromaticHeterocycles   -0.294607195
NumSaturatedHeterocycles  -0.183383866
NumAliphaticHeterocycles  -0.330066500
NumAromaticCarbocycles    -0.006198007
NumSaturatedCarbocycles   -0.122217650
NumAliphaticCarbocycles    0.082150927
FractionCSP3          -3.955397452
Chi0v                 -0.528180104
Chi1v                  2.653815553
Chi2v                 -1.889911815
Chi3v                 -1.207127662
Chi4v                  1.238474843
Chi1n                 -1.556364531
Chi2n                  2.195253733
Chi3n                  1.678096790
Chi4n                 -1.034302917
HallKierAlpha          1.471002089
kappa1                 0.486978032
kappa2                 0.475124609
kappa3                -0.491247124
```

Πλαίσιο Α.28: Υπολογισμός μοντέλου γραμμικής διακριτικής ανάλυσης

```
lda.model = lda (activity~.-NumAromaticCarbocycles-NumSaturatedCarbocycles-
    NumAliphaticCarbocycles-NumAliphaticRings, data=train_binary)
Call:
lda(activity ~ . - NumAromaticCarbocycles - NumSaturatedCarbocycles - NumAliphaticCarbocycles -
    NumAliphaticRings, data = train_binary)

Prior probabilities of groups:
  inactive     active
0.08901734 0.91098266

Group means:
             SlogP       SMR LabuteASA      TPSA      AMW ExactMW NumLipinskiHBA NumLipinskiHBD
inactive 2.996580  98.72613  148.6115 64.82364 344.9320 344.6350       4.909091       1.8571429
active   4.182342 118.44977  179.7495 53.09797 419.9152 419.3998       5.081218       0.7690355


          NumRotatableBonds      NumHBD     NumHBA NumAmideBonds NumHeteroAtoms NumHeavyAtoms
inactive          8.246753    1.272727   4.233766     0.4805195       5.272727       24.70130
active            6.189086    0.713198   4.545685     0.4175127       6.093909       30.00254


                  NumAtoms                NumRings          NumAromaticRings
inactive          52.20779                2.584416                  1.675325
active            57.91117                4.105330                  2.538071

             NumSaturatedRings NumAromaticHeterocycles NumSaturatedHeterocycles
inactive             0.9090909               0.7272727                0.6493506
active               1.1230964               0.7068528                0.9619289


        NumAliphaticHeterocycles FractionCSP3     Chi0v     Chi1v     Chi2v     Chi3v     Chi4v
inactive                0.6493506    0.5168118  15.05821  9.290437  7.048281  4.841572  3.300979
active                  1.2398477    0.4035995  17.90119 10.952550  8.563481  6.386772  4.629051

                Chi1n    Chi2n    Chi3n    Chi4n HallKierAlpha    kappa1    kappa2    kappa3
inactive     9.014534 6.712272 4.538265 3.058235    -1.890130  18.30013  9.206698  5.750724
active      10.559536 8.113417 5.999068 4.345282    -2.497265  20.71508  9.189050  4.776957

Coefficients of linear discriminants:

                                LD1
SlogP                  0.170212049
SMR                    0.049327673
LabuteASA             -0.054904488
TPSA                  -0.074251502
AMW                    0.768718523
ExactMW               -0.794592940
NumLipinskiHBA         1.315894148
NumLipinskiHBD        -1.970664773
NumRotatableBonds     -0.130023439
NumHBD                 2.318214428
NumHBA                 0.058335425
NumAmideBonds          0.290014477
NumHeteroAtoms        -0.280848196
NumHeavyAtoms          0.714007584
NumAtoms              -0.394811496
NumRings              -0.236117969
NumAromaticRings      -0.222261589
NumSaturatedRings     -0.163060598
NumAliphaticRings     -0.156276782
NumAromaticHeterocycles   -0.294607195
NumSaturatedHeterocycles  -0.183383866
NumAliphaticHeterocycles  -0.330066500
NumAromaticCarbocycles    -0.006198007
NumSaturatedCarbocycles   -0.122217650
NumAliphaticCarbocycles    0.082150927
FractionCSP3          -3.955397452
Chi0v                 -0.528180104
Chi1v                  2.653815553
Chi2v                 -1.889911815
Chi3v                 -1.207127662
Chi4v                  1.238474843
Chi1n                 -1.556364531
Chi2n                  2.195253733
Chi3n                  1.678096790
Chi4n                 -1.034302917
HallKierAlpha          1.471002089
kappa1                 0.486978032
kappa2                 0.475124609
kappa3                -0.491247124
```

Πλαίσιο Α.29: Υπολογισμός μοντέλου γραμμικής διακριτικής ανάλυσης με αφαίρεση συγγραμικών μεταβλητών

```
require(rpart)
require(rpart.plot)
decision_tree <- rpart(activity~., data = train_binary, method = 'class')
par(xpd = NA) # Avoid clipping the text in some device
rpart.plot(decision_tree)
```

Πλαίσιο Α.30: Υπολογισμός δένδρου απόφασης CART με χρήση του συνόλου train_binary

```
#Logistic
logistic_up <- glm(activity~., data=up_train,family=binomial(link='logit'),maxit
    =100)

#Stepwise
spewise_up <-stepAIC(logistic_up,direction="both")

#Lasso
bioact_up<-as.factor(up_train$activity)
x_up<-up_train[,-40]
xlasso_up<-model.matrix(bioact~.,X_train_binary)[,-1]
set.seed(123)
cv.lasso<-cv.glmnet(xlasso_up,bioact_up,alpha=1,family="binomial")

#LDA
lda.model_up = lda (activity~., data=up_train)

#Decision Tree
decision_tree <- rpart(activity~., data = up_trian, method = 'class')
```

Πλαίσιο Α.31: Υπολογισμός μοντέλων με χρήση του εξισορροπημένου συνόλου up_train

```
d3_tree<-rpart(activity_multi~.,data=train_triple,method="class")
rpart.plot(d3_tree)
d3_tree
n= 865
node), split, n, loss, yval, (yprob)
      * denotes terminal node

  1) root 865 249 highly active (0.08901734 0.19884393 0.71213873)
    2) NumAromaticRings< 1.5 90  50 inactive (0.44444444 0.44444444 0.11111111)
      4) NumLipinskiHBD>=1.5 38   8 inactive (0.78947368 0.18421053 0.02631579) *
      5) NumLipinskiHBD< 1.5 52  19 moderately active (0.19230769 0.63461538 0.17307692) *
    3) NumAromaticRings>=1.5 775 169 highly active (0.04774194 0.17032258 0.78193548)
      6) NumSaturatedCarbocycles>=0.5 80  38 highly active (0.08750000 0.38750000 0.52500000)
       12) Chi4n< 5.36035 50  19 moderately active (0.14000000 0.62000000 0.24000000) *
       13) Chi4n>=5.36035 30   0 highly active (0.00000000 0.00000000 1.00000000) *
      7) NumSaturatedCarbocycles< 0.5 695 131 highly active (0.04316547 0.14532374 0.81151079)
       14) Chi4n< 2.130959 12   5 inactive (0.58333333 0.08333333 0.33333333) *
       15) Chi4n>=2.130959 683 123 highly active (0.03367496 0.14641288 0.81991215)
         30) kappa2< 10.10466 418  98 highly active (0.04545455 0.18899522 0.76555024)
           60) TPSA>=43.78 129  56 highly active (0.09302326 0.34108527 0.56589147)
            120) Chi3v< 4.906799 33  18 moderately active (0.30303030 0.45454545 0.24242424)
              240) NumLipinskiHBD>=2.5 9   1 inactive (0.88888889 0.00000000 0.11111111) *
              241) NumLipinskiHBD< 2.5 24   9 moderately active (0.08333333 0.62500000
                  0.29166667)
                482) Chi4n>=2.96217 17   3 moderately active (0.05882353 0.82352941 0.11764706) *
                483) Chi4n< 2.96217 7   2 highly active (0.14285714 0.14285714 0.71428571) *
            121) Chi3v>=4.906799 96  31 highly active (0.02083333 0.30208333 0.67708333)
              242) NumAtoms>=57.5 34  14 moderately active (0.00000000 0.58823529 0.41176471)
                484) Chi3n< 6.34314 15   2 moderately active (0.00000000 0.86666667 0.13333333) *
                485) Chi3n>=6.34314 19   7 highly active (0.00000000 0.36842105 0.63157895) *
              243) NumAtoms< 57.5 62  11 highly active (0.03225806 0.14516129 0.82258065) *
           61) TPSA< 43.78 289  42 highly active (0.02422145 0.12110727 0.85467128)
            122) FractionCSP3>=0.431677 45  21 highly active (0.00000000 0.46666667 0.53333333)
              244) kappa2< 9.070506 36  15 moderately active (0.00000000 0.58333333 0.41666667)
                488) SlogP>=4.4007 13   2 moderately active (0.00000000 0.84615385 0.15384615) *
                489) SlogP< 4.4007 23  10 highly active (0.00000000 0.43478261 0.56521739)
                  978) HallKierAlpha< -1.85 10   3 moderately active (0.00000000 0.70000000
                      0.30000000) *
                  979) HallKierAlpha>=-1.85 13   3 highly active (0.00000000 0.23076923
                      0.76923077) *
              245) kappa2>=9.070506 9   0 highly active (0.00000000 0.00000000 1.00000000) *
            123) FractionCSP3< 0.431677 244  21 highly active (0.02868852 0.05737705 0.91393443)
                *
         31) kappa2>=10.10466 265  25 highly active (0.01509434 0.07924528 0.90566038) *
```

Πλαίσιο Α.32: Εκπαίδευση μοντέλου με τη μέθοδο υπολογισμού δένδρου απόφασης CART με χρήση του συνόλου train_triple

```
rf<-randomForest(activity_multi~.,data=train_triple,importance=TRUE)
rf
Call:
 randomForest(formula = activity_multi ~ ., data = train_triple)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 6

        OOB estimate of  error rate: 16.88%
Confusion matrix:
                 inactive moderately active highly active
inactive               48                15             14
moderately active      14                79             79
highly active           3                21            592
                 class.error
inactive          0.37662338
moderately active 0.54069767
highly active     0.03896104

varImpPlot(rf)
```

Πλαίσιο Α.33: Εκπαίδευση μοντέλου με τη μέθοδο τυχαίου δάσους

```
require(nnet)
require(reshape2)
train_triple$activity_multi1 <- relevel(train_triple$activity_multi, ref = "highly active")
multi_logistic<- multinom(activity_multi1~.-activity_multi, data = train_triple)
# weights:  123 (80 variable)
initial  value 950.299630
iter  10 value 554.188904
iter  20 value 496.303995
iter  30 value 399.639775
iter  40 value 372.089425
iter  50 value 360.574596
iter  60 value 358.404741
iter  70 value 357.886709
iter  80 value 357.748164
iter  90 value 357.714053
iter 100 value 357.699521
final  value 357.699521
stopped after 100 iterations

summary(multi_logistic)
Call:
multinom(formula = activity_multi1 ~ . - activity_multi, data = train_triple)

Coefficients:
                 (Intercept)      SlogP        SMR LabuteASA     TPSA       AMW
inactive           -5.492651 -2.970741 -0.6488007 0.6206521 0.5968279  9.029041
moderately active  -8.901406 -0.240873  0.2163152 0.3024376 0.2263367 -2.394340

                   ExactMW    NumLipinskiHBA  NumLipinskiHBD NumRotatableBonds NumHBD
inactive         -9.071319  -13.671101       -3.7368857       0.6468759       0.1396776
moderately active  2.376938  -3.992617        -0.8258357       0.6414679      -0.7219588

                   NumHBA     NumAmideBonds NumHeteroAtoms NumHeavyAtoms  NumAtoms
inactive           0.7620404  1.66808793       4.996635      -4.717855     4.2548112
moderately active -0.6942074 -0.04064859       2.077616      -2.851027     0.7943569

                   NumRings  NumAromaticRings NumSaturatedRings  NumAliphaticRings
inactive           0.418830  2.9545146        6.4938893          -2.535685
moderately active  1.739783  0.3673978        0.7700125           1.372386
                 NumAromaticHeterocycles NumSaturatedHeterocycles NumAliphaticHeterocycles
inactive                       2.7652686                3.9120756              -1.196325
moderately active              0.4897074               -0.2379883               1.291652

                 NumAromaticCarbocycles NumSaturatedCarbocycles NumAliphaticCarbocycles
inactive                      0.1892461               2.581814              -1.33936002
moderately active            -0.1223096               1.008001               0.08073356

                 FractionCSP3     Chi0v     Chi1v     Chi2v     Chi3v     Chi4v
inactive           -3.959211 1.2835553 -1.26455 -2.726090  8.661689 -4.568732
moderately active   7.699909 0.6652304  3.00189 -4.093887 -1.625229  1.043647

                     Chi1n      Chi2n       Chi3n     Chi4n   HallKierAlpha
inactive         -0.7137892 -3.707732 -12.2889451  4.449838      -11.118795
moderately active -1.6752491 -0.501430   0.2172374 -2.044548       -2.715455

                   kappa1      kappa2   kappa3
inactive         -2.75348490 -4.509812 2.171307
moderately active  0.09981641 -3.983131 1.919381
Residual Deviance: 715.399
AIC: 859.399
```

Πλαίσιο Α.34: Εκπαίδευση μοντέλου με τη μέθοδο πολυωνυμικής λογιστικής παλινδρόμησης

```
require(e1071)
svm1 <- svm(activity_multi~., data=train_triple,method="C-classification",
            kernal="radial",gamma=0.1,cost=10,scale=TRUE)
print(svm1)
Call:
svm(formula = activity_multi ~ ., data = train_triple, method = "C-classification"
    ,
    kernal = "radial", gamma = 0.1, cost = 10, scale = TRUE)
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  10
Number of Support Vectors:  433
```

Πλαίσιο Α.35: Υπολογισμός μοντέλου με χρήση μηχανών διανυσμάτων υποστήριξης

```

svm_tune <- tune(svm, train.x=x, train.y=y,  kernel="radial",
                 ranges=list(cost=10^(-1:2), gamma=c(.5,1,2)))
print(svm_tune)
Parameter tuning of 'svm':
sampling method: 10-fold cross validation
best parameters:
 cost gamma
   10   0.5
best performance: 0.2214381
svmTune <- svm(activity_multi~., data=train_triple,method="C-classification",
               kernal="radial",gamma=0.5, cost=10,scale=TRUE)
print(svmTune)

Call:
svm(formula = activity_multi ~ ., data = train_triple, method = "C-classification"
    , kernal = "radial", gamma = 0.5, cost = 10, scale = TRUE)

Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  10

Number of Support Vectors:  681
```

Πλαίσιο Α.36: Συντονισμός παραμέτρων διανυσμάτων υποστήριξης και εκ νέου υπολογισμός μοντέλου με χρήση μηχανών διανυσμάτων υποστήριξης

Παράρτημα Β

Κώδικες Κεφαλαίου 4

```
probsss=predict(logistic,type="response")
preddd=ifelse(probsss<0.90,"inactive","active")
train_binary$activity<-ifelse(train_binary$activity==0,"inactive","active")
> confusionMatrix(as.factor(preddd), as.factor(train_binary$activity))
Confusion Matrix and Statistics

          Reference
Prediction active inactive
  active      721        4
  inactive     67       73
               Accuracy : 0.9179
                 95% CI : (0.8976, 0.9353)
    No Information Rate : 0.911
    P-Value [Acc > NIR] : 0.2587
                  Kappa : 0.6304
 Mcnemar's Test P-Value : 1.866e-13
            Sensitivity : 0.9150
            Specificity : 0.9481
         Pos Pred Value : 0.9945
         Neg Pred Value : 0.5214
             Prevalence : 0.9110
         Detection Rate : 0.8335
   Detection Prevalence : 0.8382
      Balanced Accuracy : 0.9315
       'Positive' Class : active
```

Πλαίσιο Β.1: Υπολογισμός μέτρων αξιολόγησης του μοντέλου λογιστικής παλινδρόμησης για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης train_binary

```
probsss=predict(stepwise,type="response")
preddd=ifelse(probsss<0.90,"inactive","active")
> preddd=relevel(preddd,"inactive")
> train_binary$activity=relevel(train_binary$activity,"active")
> train_binary$activity=relevel(train_binary$activity,"inactive")
> confusionMatrix(as.factor(preddd),as.factor(act),positive = "active")
Confusion Matrix and Statistics

          Reference
Prediction inactive active
  inactive       22      0
  active         55    788

               Accuracy : 0.9364
                 95% CI : (0.918, 0.9517)
    No Information Rate : 0.911
    P-Value [Acc > NIR] : 0.003761
                  Kappa : 0.4216

 Mcnemar's Test P-Value : 3.305e-13

            Sensitivity : 1.0000
            Specificity : 0.2857
         Pos Pred Value : 0.9348
         Neg Pred Value : 1.0000
             Prevalence : 0.9110
         Detection Rate : 0.9110
   Detection Prevalence : 0.9746
      Balanced Accuracy : 0.6429
       'Positive' Class : active
```

Πλαίσιο Β.2: Υπολογισμός μέτρων αξιολόγησης του μοντέλου λογιστικής παλινδρόμησης με βήματα για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης train_binary

```
predicted_lda <-(predict(lda.model ,type="class")$class)
> caret::confusionMatrix(predicted_lda,as.factor(train_binary$activity),positive="
    active")
Confusion Matrix and Statistics

          Reference
Prediction inactive active
  inactive       52     23
  active         25    765

               Accuracy : 0.9445
                 95% CI : (0.9271, 0.9588)
    No Information Rate : 0.911
    P-Value [Acc > NIR] : 0.0001515
                  Kappa : 0.6538
 Mcnemar's Test P-Value : 0.8852339
            Sensitivity : 0.9708
            Specificity : 0.6753
         Pos Pred Value : 0.9684
         Neg Pred Value : 0.6933
             Prevalence : 0.9110
         Detection Rate : 0.8844
   Detection Prevalence : 0.9133
      Balanced Accuracy : 0.8231
       'Positive' Class : active
```

Πλαίσιο Β.3: Υπολογισμός μέτρων αξιολόγησης του μοντέλου γραμμικής διακριτικής ανάλυσης για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης train_binary

```
 predicted_lda1 <-(predict(lda.model1 ,type="class")$class)
> caret::confusionMatrix(predicted_lda1,as.factor(train_binary$activity),positive=
    "active")
Confusion Matrix and Statistics

          Reference
Prediction inactive active
  inactive       52     23
  active         25    765

               Accuracy : 0.9445
                 95% CI : (0.9271, 0.9588)
    No Information Rate : 0.911
    P-Value [Acc > NIR] : 0.0001515
                  Kappa : 0.6538
 Mcnemar's Test P-Value : 0.8852339
            Sensitivity : 0.9708
            Specificity : 0.6753
         Pos Pred Value : 0.9684
         Neg Pred Value : 0.6933
             Prevalence : 0.9110
         Detection Rate : 0.8844
   Detection Prevalence : 0.9133
      Balanced Accuracy : 0.8231
       'Positive' Class : active
```

Πλαίσιο Β.4: Υπολογισμός μέτρων αξιολόγησης του μοντέλου γραμμικής διακριτικής ανάλυσης χωρίς συγγραμικές μεταβλητές για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης train_binary

```
 predlaso1<-as.factor(predict(cv.lasso, newx = xlasso, s =c(cv.lasso$lambda.1se),
    type="class"))
> predlaso1<-relevel(predlaso1,"inactive")
> cm<-confusionMatrix(predlaso1,bioact,positive="active")
> cm
Confusion Matrix and Statistics

          Reference
Prediction inactive active
  inactive       43      6
  active         34    782

               Accuracy : 0.9538
                 95% CI : (0.9376, 0.9668)
    No Information Rate : 0.911
    P-Value [Acc > NIR] : 1.080e-06
                  Kappa : 0.6589
 Mcnemar's Test P-Value : 1.963e-05
            Sensitivity : 0.9924
            Specificity : 0.5584
         Pos Pred Value : 0.9583
         Neg Pred Value : 0.8776
             Prevalence : 0.9110
         Detection Rate : 0.9040
   Detection Prevalence : 0.9434
      Balanced Accuracy : 0.7754
       'Positive' Class : active
```

Πλαίσιο Β.5: Υπολογισμός μέτρων αξιολόγησης του μοντέλου λογιστικής παλινδρόμησης με μέθοδο Lasso και λ=lamda.1se για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης train_binary

```
 predlasom<-as.factor(predict(cv.lasso, newx = xlasso, s =c(cv.lasso$lambda.min),
    type="class"))
> predlasom<-relevel(predlasom,"inactive")
> cmm<-confusionMatrix(predlasom,bioact)
> cmm
Confusion Matrix and Statistics

          Reference
Prediction inactive active
  inactive       50      8
  active         27    780

               Accuracy : 0.9595
                 95% CI : (0.9442, 0.9717)
    No Information Rate : 0.911
    P-Value [Acc > NIR] : 2.196e-08
                  Kappa : 0.7193
 Mcnemar's Test P-Value : 0.002346
            Sensitivity : 0.64935
            Specificity : 0.98985
         Pos Pred Value : 0.86207
         Neg Pred Value : 0.96654
             Prevalence : 0.08902
         Detection Rate : 0.05780
   Detection Prevalence : 0.06705
      Balanced Accuracy : 0.81960
       'Positive' Class : inactive
```

Πλαίσιο Β.6: Υπολογισμός μέτρων αξιολόγησης του μοντέλου λογιστικής παλινδρόμησης με μέθοδο Lasso και λ= lambda.min για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης train_binary

```
 predicted_d2_tree <-predict(decision_tree ,train_binary,na.action=na.pass,type="
    class")
> caret::confusionMatrix(predicted_d2_tree,as.factor(train_binary$activity),
    positive="active")
Confusion Matrix and Statistics

          Reference
Prediction inactive active
  inactive       60     19
  active         17    769

               Accuracy : 0.9584
                 95% CI : (0.9428, 0.9707)
    No Information Rate : 0.911
    P-Value [Acc > NIR] : 5.055e-08
                  Kappa : 0.7464
 Mcnemar's Test P-Value : 0.8676
            Sensitivity : 0.9759
            Specificity : 0.7792
         Pos Pred Value : 0.9784
         Neg Pred Value : 0.7595
             Prevalence : 0.9110
         Detection Rate : 0.8890
   Detection Prevalence : 0.9087
      Balanced Accuracy : 0.8776
       'Positive' Class : active
```

Πλαίσιο Β.7: Υπολογισμός μέτρων αξιολόγησης του μοντέλου με δένδρο απόφασης για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης train_binary

```
predicted_d3_tree <-predict(d3_tree,train_triple,na.action=na.pass,type="class")
caret::confusionMatrix(predicted_d3_tree,as.factor(train_triple$activity_multi))
Confusion Matrix and Statistics
                    Reference
Prediction           inactive moderately active highly active
  inactive                 45                 8             6
  moderately active        18               109            30
  highly active            14                55           580

Overall Statistics
               Accuracy : 0.8486
                 95% CI : (0.8229, 0.8718)
    No Information Rate : 0.7121
    P-Value [Acc > NIR] : < 2.2e-16
                  Kappa : 0.6424
 Mcnemar's Test P-Value : 0.002409
Statistics by Class:

                     Class: inactive Class: moderately active Class: highly active
Sensitivity                  0.58442                   0.6337                0.9416
Specificity                  0.98223                   0.9307                0.7229
Pos Pred Value               0.76271                   0.6943                0.8937
Neg Pred Value               0.96030                   0.9110                0.8333
Prevalence                   0.08902                   0.1988                0.7121
Detection Rate               0.05202                   0.1260                0.6705
Detection Prevalence         0.06821                   0.1815                0.7503
Balanced Accuracy            0.78332                   0.7822                0.8322
```

Πλαίσιο Β.8: Υπολογισμός μέτρων αξιολόγησης του μοντέλου με δένδρο απόφασης για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης train_triple

```
predicted_randomf <-predict(rf,train_triple,na.action=na.pass,type="class")
caret::confusionMatrix(as.factor(predicted_randomf),as.factor(train_triple$
    activity_multi))
Confusion Matrix and Statistics
                    Reference
Prediction           inactive moderately active highly active
  inactive                 75                 3             1
  moderately active         2               169             1
  highly active             0                 0           614

Overall Statistics
               Accuracy : 0.9919
                 95% CI : (0.9834, 0.9967)
    No Information Rate : 0.7121
    P-Value [Acc > NIR] : <2e-16
                  Kappa : 0.9819
 Mcnemar's Test P-Value : 0.5319

Statistics by Class:
                     Class: inactive Class: moderately active Class: highly active
Sensitivity                  0.97403                   0.9826                0.9968
Specificity                  0.99492                   0.9957                1.0000
Pos Pred Value               0.94937                   0.9826                1.0000
Neg Pred Value               0.99746                   0.9957                0.9920
Prevalence                   0.08902                   0.1988                0.7121
Detection Rate               0.08671                   0.1954                0.7098
Detection Prevalence         0.09133                   0.1988                0.7098
Balanced Accuracy            0.98447                   0.9891                0.9984
```

Πλαίσιο Β.9: Υπολογισμός μέτρων αξιολόγησης του μοντέλου με τυχαίο δάσος για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης train_triple

```
predicted_svm1 <-predict(svm1,train_triple,na.action=na.pass,type="class")
caret::confusionMatrix(predicted_svm1,as.factor(train_triple$activity_multi))
Confusion Matrix and Statistics
                    Reference
Prediction           inactive moderately active highly active
  inactive                 70                 3             0
  moderately active         4               158             8
  highly active             3                11           608

Overall Statistics
               Accuracy : 0.9665
                 95% CI : (0.9522, 0.9774)
    No Information Rate : 0.7121
    P-Value [Acc > NIR] : <2e-16
                  Kappa : 0.924
 Mcnemar's Test P-Value : 0.306

Statistics by Class:
                     Class: inactive Class: moderately active Class: highly active
Sensitivity                  0.90909                   0.9186                0.9870
Specificity                  0.99619                   0.9827                0.9438
Pos Pred Value               0.95890                   0.9294                0.9775
Neg Pred Value               0.99116                   0.9799                0.9671
Prevalence                   0.08902                   0.1988                0.7121
Detection Rate               0.08092                   0.1827                0.7029
Detection Prevalence         0.08439                   0.1965                0.7191
Balanced Accuracy            0.95264                   0.9506                0.9654
```

Πλαίσιο Β.10: Υπολογισμός μέτρων αξιολόγησης του μοντέλου με μηχανές διανυσμάτων υποστήριξης για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης train_triple

```
predicted_svmTune <-predict(svmTune,train_triple,na.action=na.pass,type="class")
caret::confusionMatrix(predicted_svmTune,as.factor(train_triple$activity_multi))
Confusion Matrix and Statistics
                    Reference
Prediction           inactive moderately active highly active
  inactive                 73                 2             1
  moderately active         4               168             4
  highly active             0                 2           611

Overall Statistics
               Accuracy : 0.985
                 95% CI : (0.9744, 0.992)
    No Information Rate : 0.7121
    P-Value [Acc > NIR] : <2e-16
                  Kappa : 0.9664
 Mcnemar's Test P-Value : 0.5062

Statistics by Class:
                     Class: inactive Class: moderately active Class: highly active
Sensitivity                  0.94805                   0.9767                0.9919
Specificity                  0.99619                   0.9885                0.9920
Pos Pred Value               0.96053                   0.9545                0.9967
Neg Pred Value               0.99493                   0.9942                0.9802
Prevalence                   0.08902                   0.1988                0.7121
Detection Rate               0.08439                   0.1942                0.7064
Detection Prevalence         0.08786                   0.2035                0.7087
Balanced Accuracy            0.97212                   0.9826                0.9919
```

Πλαίσιο Β.11: Υπολογισμός μέτρων αξιολόγησης του μοντέλου με συντονισμένες μηχανές διανυσμάτων υποστήριξης για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης train_triple

```
predicted_multilogistic <-predict(multi_logistic,train_triple,na.action=na.pass,
    type="class")
caret::confusionMatrix(predicted_multilogistic,as.factor(train_triple$activity_
    multi))
Confusion Matrix and Statistics
                    Reference
Prediction           inactive moderately active highly active
  inactive                 57                 9             4
  moderately active        10                73            21
  highly active            10                90           591

Overall Statistics
                Accuracy : 0.8335
                  95% CI : (0.807, 0.8578)
    No Information Rate : 0.7121
    P-Value [Acc > NIR] : < 2.2e-16
                   Kappa : 0.5838
 Mcnemar's Test P-Value : 7.188e-10

Statistics by Class:
                     Class: inactive Class: moderately active Class: highly active
Sensitivity                  0.74026                  0.42442               0.9594
Specificity                  0.98350                  0.95527               0.5984
Pos Pred Value               0.81429                  0.70192               0.8553
Neg Pred Value               0.97484                  0.86991               0.8563
Prevalence                   0.08902                  0.19884               0.7121
Detection Rate               0.06590                  0.08439               0.6832
Detection Prevalence         0.08092                  0.12023               0.7988
Balanced Accuracy            0.86188                  0.68984               0.7789
```

Πλαίσιο B.12: Υπολογισμός μέτρων αξιολόγησης του μοντέλου με πολυωνυμική λογιστική παλινδρόμηση για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης train_triple

```
probsss=predict(logistic,newdata = test_binary ,type="response")
preddd=as.factor(ifelse(probsss<0.90,"inactive","active"))
preddd=relevel(preddd,"inactive")
confusionMatrix(as.factor(preddd), as.factor(test_binary$activity),positive="
    active")
Confusion Matrix and Statistics
          Reference
Prediction inactive active
  inactive       26     35
  active          8    303

               Accuracy : 0.8844
                 95% CI : (0.8475, 0.9151)
    No Information Rate : 0.9086
    P-Value [Acc > NIR] : 0.9524
                  Kappa : 0.4872
 Mcnemar's Test P-Value : 7.341e-05
            Sensitivity : 0.8964
            Specificity : 0.7647
         Pos Pred Value : 0.9743
         Neg Pred Value : 0.4262
             Prevalence : 0.9086
         Detection Rate : 0.8145
   Detection Prevalence : 0.8360
      Balanced Accuracy : 0.8306
       'Positive' Class : active
```

Πλαίσιο B.13: Υπολογισμός μέτρων αξιολόγησης του μοντέλου λογιστικής παλινδρόμησης για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης test_binary

```
probsss=predict(stepwise,newdata = test_binary ,type="response")
preddd=as.factor(ifelse(probsss<0.90,"inactive","active"))
preddd=relevel(preddd,"inactive")
confusionMatrix(as.factor(preddd), as.factor(test_binary$activity),positive="
    active")
Confusion Matrix and Statistics
          Reference
Prediction inactive active
  inactive       25     35
  active          9    303

               Accuracy : 0.8817
                 95% CI : (0.8445, 0.9127)
    No Information Rate : 0.9086
    P-Value [Acc > NIR] : 0.966735
                  Kappa : 0.4701
 Mcnemar's Test P-Value : 0.000164
            Sensitivity : 0.8964
            Specificity : 0.7353
         Pos Pred Value : 0.9712
         Neg Pred Value : 0.4167
             Prevalence : 0.9086
         Detection Rate : 0.8145
   Detection Prevalence : 0.8387
      Balanced Accuracy : 0.8159
       'Positive' Class : active
```

Πλαίσιο Β.14: Υπολογισμός μέτρων αξιολόγησης του μοντέλου λογιστικής παλινδρόμησης με βήματα για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης test_binary

```
predicted_lda <-(predict(lda.model, newdata = test_binary ,type="class")$class)
caret::confusionMatrix(predicted_lda,as.factor(test_binary$activity),positive="
    active")
Confusion Matrix and Statistics
          Reference
Prediction inactive active
  inactive       17     18
  active         17    320

               Accuracy : 0.9059
                 95% CI : (0.8716, 0.9336)
    No Information Rate : 0.9086
    P-Value [Acc > NIR] : 0.615
                  Kappa : 0.4409
 Mcnemar's Test P-Value : 1.000
            Sensitivity : 0.9467
            Specificity : 0.5000
         Pos Pred Value : 0.9496
         Neg Pred Value : 0.4857
             Prevalence : 0.9086
         Detection Rate : 0.8602
   Detection Prevalence : 0.9059
      Balanced Accuracy : 0.7234
       'Positive' Class : active
```

Πλαίσιο Β.15: Υπολογισμός μέτρων αξιολόγησης του μοντέλου γραμμικής διακριτικής ανάλυσης για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης test_binary

```
predicted_lda1 <-(predict(lda.model1, newdata = test_binary ,type="class")$class)
caret::confusionMatrix(predicted_lda1,as.factor(test_binary$activity),positive="
    active")
Confusion Matrix and Statistics

          Reference
Prediction inactive active
  inactive       17     18
  active         17    320

               Accuracy : 0.9059
                 95% CI : (0.8716, 0.9336)
    No Information Rate : 0.9086
    P-Value [Acc > NIR] : 0.615
                  Kappa : 0.4409
 Mcnemar's Test P-Value : 1.000
            Sensitivity : 0.9467
            Specificity : 0.5000
         Pos Pred Value : 0.9496
         Neg Pred Value : 0.4857
             Prevalence : 0.9086
         Detection Rate : 0.8602
   Detection Prevalence : 0.9059
      Balanced Accuracy : 0.7234
       'Positive' Class : active
```

Πλαίσιο Β.16: Υπολογισμός μέτρων αξιολόγησης του μοντέλου γραμμικής διακριτικής ανάλυσης χωρίς συγγραμικές μεταβλητές για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης test_binary

```
bioact<-as.factor(y_test_binary)
xlassotest<-model.matrix(bioact~.,X_test_binary)[,-1]
predlaso1<-as.factor(predict(cv.lasso, newx = xlassotest, s =c(cv.lasso$lambda.1se
    ),type="class"))
predlaso1<-relevel(predlaso1,"inactive")
cm<-confusionMatrix(predlaso1,bioact,positive="active")
cm
Confusion Matrix and Statistics
          Reference
Prediction inactive active
  inactive       12      5
  active         22    333

               Accuracy : 0.9274
                 95% CI : (0.8962, 0.9516)
    No Information Rate : 0.9086
    P-Value [Acc > NIR] : 0.118922
                  Kappa : 0.4362
 Mcnemar's Test P-Value : 0.002076
            Sensitivity : 0.9852
            Specificity : 0.3529
         Pos Pred Value : 0.9380
         Neg Pred Value : 0.7059
             Prevalence : 0.9086
         Detection Rate : 0.8952
   Detection Prevalence : 0.9543
      Balanced Accuracy : 0.6691
       'Positive' Class : active
```

Πλαίσιο Β.17: Υπολογισμός μέτρων αξιολόγησης του μοντέλου λογιστικής παλινδρόμησης με Lasso και λ=lambda.1se για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης test_binary

```
predlasom<-as.factor(predict(cv.lasso, newx = xlassotest, s =c(cv.lasso$lambda.min
    ),type="class"))
predlasom<-relevel(predlasom,"inactive")
cmm<-confusionMatrix(predlasom,bioact,positive = "active")
cmm
Confusion Matrix and Statistics
          Reference
Prediction inactive active
  inactive       16      7
  active         18    331

               Accuracy : 0.9328
                 95% CI : (0.9024, 0.956)
    No Information Rate : 0.9086
    P-Value [Acc > NIR] : 0.05847
                  Kappa : 0.5265
 Mcnemar's Test P-Value : 0.04550
            Sensitivity : 0.9793
            Specificity : 0.4706
         Pos Pred Value : 0.9484
         Neg Pred Value : 0.6957
             Prevalence : 0.9086
         Detection Rate : 0.8898
   Detection Prevalence : 0.9382
      Balanced Accuracy : 0.7249
       'Positive' Class : active
```

Πλαίσιο Β.18: Υπολογισμός μέτρων αξιολόγησης του μοντέλου λογιστικής παλινδρόμησης με Lasso και λ=lambda.min για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης test_binary

```
predicted_d2_tree <-predict(decision_tree ,test_binary,na.action=na.pass,type="
    class")
caret::confusionMatrix(predicted_d2_tree,as.factor(test_binary$activity),positive=
    "active")
Confusion Matrix and Statistics
          Reference
Prediction inactive active
  inactive       21     15
  active         13    323

               Accuracy : 0.9247
                 95% CI : (0.8931, 0.9494)
    No Information Rate : 0.9086
    P-Value [Acc > NIR] : 0.1611
                  Kappa : 0.5585
 Mcnemar's Test P-Value : 0.8501
            Sensitivity : 0.9556
            Specificity : 0.6176
         Pos Pred Value : 0.9613
         Neg Pred Value : 0.5833
             Prevalence : 0.9086
         Detection Rate : 0.8683
   Detection Prevalence : 0.9032
      Balanced Accuracy : 0.7866
       'Positive' Class : active
```

Πλαίσιο Β.19: Υπολογισμός μέτρων αξιολόγησης του μοντέλου με δένδρο απόφασης για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης test_binary

```
predicted_d3_tree predict(d3_tree,test_triple,na.action=na.pass,type="class")
caret::confusionMatrix(predicted_d3_tree,as.factor(test_triple$activity_multi))
Confusion Matrix and Statistics
                    Reference
Prediction           inactive moderately active highly active
  inactive                 16                 5             2
  moderately active        12                36            32
  highly active             6                36           227

Overall Statistics
               Accuracy : 0.75
                 95% CI : (0.7028, 0.7932)
    No Information Rate : 0.7016
    P-Value [Acc > NIR] : 0.02233
                  Kappa : 0.435
 Mcnemar's Test P-Value : 0.16338

Statistics by Class:
                     Class: inactive Class: moderately active Class: highly active
Sensitivity                  0.47059                  0.46753                0.8697
Specificity                  0.97929                  0.85085                0.6216
Pos Pred Value               0.69565                  0.45000                0.8439
Neg Pred Value               0.94842                  0.85959                0.6699
Prevalence                   0.09140                  0.20699                0.7016
Detection Rate               0.04301                  0.09677                0.6102
Detection Prevalence         0.06183                  0.21505                0.7231
Balanced Accuracy            0.72494                  0.65919                0.7457
```

Πλαίσιο Β.20: Υπολογισμός μέτρων αξιολόγησης του μοντέλου με δένδρο απόφασης για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης test_triple

```
rf<-randomForest(activity_multi~.,data=train_triple)
predicted_randomf <-predict(rf,test_triple,na.action=na.pass,type="class")
caret::confusionMatrix(as.factor(predicted_randomf),as.factor(test_triple$activity
    _multi))
Confusion Matrix and Statistics
                    Reference
Prediction           inactive moderately active highly active
  inactive                 17                 3             1
  moderately active         7                41            11
  highly active            10                33           249

Overall Statistics
               Accuracy : 0.8253
                 95% CI : (0.7828, 0.8625)
    No Information Rate : 0.7016
    P-Value [Acc > NIR] : 3.09e-08
                  Kappa : 0.5752
 Mcnemar's Test P-Value : 0.0001727
Statistics by Class:
                     Class: inactive Class: moderately active Class: highly active
Sensitivity                  0.50000                   0.5325                0.9540
Specificity                  0.98817                   0.9390                0.6126
Pos Pred Value               0.80952                   0.6949                0.8527
Neg Pred Value               0.95157                   0.8850                0.8500
Prevalence                   0.09140                   0.2070                0.7016
Detection Rate               0.04570                   0.1102                0.6694
Detection Prevalence         0.05645                   0.1586                0.7849
Balanced Accuracy            0.74408                   0.7357                0.7833
```

Πλαίσιο Β.21: Υπολογισμός μέτρων αξιολόγησης του μοντέλου με τυχαίο δάσος για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης test_triple

```
predicted_svm1 <-predict(svm1,test_triple,na.action=na.pass,type="class")
caret::confusionMatrix(predicted_svm1,as.factor(test_triple$activity_multi))
Confusion Matrix and Statistics
                    Reference
Prediction           inactive moderately active highly active
  inactive                 17                 2             1
  moderately active         8                47            21
  highly active             9                28           239

Overall Statistics
               Accuracy : 0.8145
                 95% CI : (0.7712, 0.8527)
    No Information Rate : 0.7016
    P-Value [Acc > NIR] : 4.563e-07
                  Kappa : 0.5709
 Mcnemar's Test P-Value : 0.01173

Statistics by Class:
                     Class: inactive Class: moderately active Class: highly active
Sensitivity                  0.50000                   0.6104                0.9157
Specificity                  0.99112                   0.9017                0.6667
Pos Pred Value               0.85000                   0.6184                0.8659
Neg Pred Value               0.95170                   0.8986                0.7708
Prevalence                   0.09140                   0.2070                0.7016
Detection Rate               0.04570                   0.1263                0.6425
Detection Prevalence         0.05376                   0.2043                0.7419
Balanced Accuracy            0.74556                   0.7560                0.7912
Warning message:
In confusionMatrix.default(predicted_svm1, as.factor(test_triple$activity_multi))
    :
  Levels are not in the same order for reference and data. Refactoring data to
     match.
```

Πλαίσιο Β.22: Υπολογισμός μέτρων αξιολόγησης του μοντέλου με μηχανές διανυσμάτων υποστήριξης για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης test_triple

```
predicted_svmTune <-predict(svmTune,test_triple,na.action=na.pass,type="class")
caret::confusionMatrix(predicted_svmTune,as.factor(test_triple$activity_multi))
Confusion Matrix and Statistics
                    Reference
Prediction           inactive moderately active highly active
  inactive                 12                 2             0
  moderately active         5                32            14
  highly active            17                43           247

Overall Statistics
               Accuracy : 0.7823
                 95% CI : (0.7368, 0.8232)
    No Information Rate : 0.7016
    P-Value [Acc > NIR] : 0.0002961
                  Kappa : 0.4405
 Mcnemar's Test P-Value : 0.0000003159
Statistics by Class:

Statistics by Class:
                     Class: inactive Class: moderately active Class: highly active
Sensitivity                  0.35294                   0.41558               0.9464
Specificity                  0.99408                   0.93559               0.4595
Pos Pred Value               0.85714                   0.62745               0.8046
Neg Pred Value               0.93855                   0.85981               0.7846
Prevalence                   0.09140                   0.20699               0.7016
Detection Rate               0.03226                   0.08602               0.6640
Detection Prevalence         0.03763                   0.13710               0.8253
Balanced Accuracy            0.67351                   0.67559               0.7029
```

Πλαίσιο Β.23: Υπολογισμός μέτρων αξιολόγησης του μοντέλου με συντονισμένες μηχανές διανυσμάτων υποστήριξης για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης test_triple

```
predicted_multilogistic predict(multi_logistic,test_triple,na.action=na.pass,type=
    "class")
caret::confusionMatrix(predicted_multilogistic,as.factor(test_triple$activity_
    multi))
Confusion Matrix and Statistics
                    Reference
Prediction           inactive moderately active highly active
  inactive                 16                 8             1
  moderately active        11                25            18
  highly active             7                44           242

Overall Statistics
                Accuracy : 0.7608
                  95% CI : (0.7141, 0.8032)
    No Information Rate : 0.7016
    P-Value [Acc > NIR] : 0.006587
                   Kappa : 0.4182
 Mcnemar's Test P-Value : 0.001202

Statistics by Class:
                     Class: inactive Class: moderately active Class: highly active
Sensitivity                  0.47059                  0.3247                0.9272
Specificity                  0.97337                  0.9017                0.5405
Pos Pred Value               0.64000                  0.4630                0.8259
Neg Pred Value               0.94813                  0.8365                0.7595
Prevalence                   0.09140                  0.2070                0.7016
Detection Rate               0.04301                  0.0672                0.6505
Detection Prevalence         0.06720                  0.1452                0.7876
Balanced Accuracy            0.72198                  0.6132                0.7339
```

Πλαίσιο Β.24: Υπολογισμός μέτρων αξιολόγησης του μοντέλου με πολυωνυμική λογιστική παλινδρόμηση για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης test_triple

```
probsss=predict(logistic_up,newdata = test_binary ,type="response")
preddd=as.factor(ifelse(probsss<0.90,"inactive","active"))
preddd=relevel(preddd,"inactive")
confusionMatrix(as.factor(preddd), as.factor(test_binary$activity),positive="
    active")
Confusion Matrix and Statistics
          Reference
Prediction inactive active
  inactive       28     62
  active          6    276

               Accuracy : 0.8172
                 95% CI : (0.7741, 0.8552)
    No Information Rate : 0.9086
    P-Value [Acc > NIR] : 1
                  Kappa : 0.3677
 Mcnemar's Test P-Value : 0.00000000002563
            Sensitivity : 0.8166
            Specificity : 0.8235
         Pos Pred Value : 0.9787
         Neg Pred Value : 0.3111
             Prevalence : 0.9086
         Detection Rate : 0.7419
   Detection Prevalence : 0.7581
      Balanced Accuracy : 0.8200
       'Positive' Class : active
```

Πλαίσιο Β.25: Υπολογισμός μέτρων αξιολόγησης του εξισορροπημένου μοντέλου με λογιστική παλινδρόμηση για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης test_binary

```
probsss=predict(stepwise_up,newdata = test_binary ,type="response")
preddd=as.factor(ifelse(probsss<0.90,"inactive","active"))
preddd=relevel(preddd,"inactive")
confusionMatrix(as.factor(preddd), as.factor(test_binary$activity),positive="
    active")
Confusion Matrix and Statistics
          Reference
Prediction inactive active
  inactive       27     66
  active          7    272

               Accuracy : 0.8038
                 95% CI : (0.7597, 0.8429)
    No Information Rate : 0.9086
    P-Value [Acc > NIR] : 1
                  Kappa : 0.3364
 Mcnemar's Test P-Value : 0.00000000001134
            Sensitivity : 0.8047
            Specificity : 0.7941
         Pos Pred Value : 0.9749
         Neg Pred Value : 0.2903
             Prevalence : 0.9086
         Detection Rate : 0.7312
   Detection Prevalence : 0.7500
      Balanced Accuracy : 0.7994
       'Positive' Class : active
```

Πλαίσιο Β.26: Υπολογισμός μέτρων αξιολόγησης του εξισορροπημένου μοντέλου με λογιστική παλινδρόμηση κατα βήματα για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης test_binary

```
 predicted_lda_up <-(predict(lda.model_up, newdata = test_binary ,type="class")$
    class)
> caret::confusionMatrix(predicted_lda_up,as.factor(test_binary$activity),positive
   ="active")
Confusion Matrix and Statistics
          Reference
Prediction inactive active
  inactive       27     48
  active          7    290

               Accuracy : 0.8522
                 95% CI : (0.8119, 0.8866)
    No Information Rate : 0.9086
    P-Value [Acc > NIR] : 0.9998
                  Kappa : 0.4228
 Mcnemar's Test P-Value : 0.00000006906
            Sensitivity : 0.8580
            Specificity : 0.7941
         Pos Pred Value : 0.9764
         Neg Pred Value : 0.3600
             Prevalence : 0.9086
         Detection Rate : 0.7796
   Detection Prevalence : 0.7984
      Balanced Accuracy : 0.8261
       'Positive' Class : active
```

Πλαίσιο Β.27: Υπολογισμός μέτρων αξιολόγησης του εξισορροπημένου μοντέλου με γραμμική διακριτική ανάλυση για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης test_binary

```
predicted_lda1 <-(predict(lda.model1_up, newdata = test_binary ,type="class")$
    class)
caret::confusionMatrix(predicted_lda1,as.factor(test_binary$activity),positive="
    active")
Confusion Matrix and Statistics
          Reference
Prediction inactive active
  inactive       27     48
  active          7    290

               Accuracy : 0.8522
                 95% CI : (0.8119, 0.8866)
    No Information Rate : 0.9086
    P-Value [Acc > NIR] : 0.9998
                  Kappa : 0.4228
 Mcnemar's Test P-Value : 0.00000006906
            Sensitivity : 0.8580
            Specificity : 0.7941
         Pos Pred Value : 0.9764
         Neg Pred Value : 0.3600
             Prevalence : 0.9086
         Detection Rate : 0.7796
   Detection Prevalence : 0.7984
      Balanced Accuracy : 0.8261
       'Positive' Class : active
```

Πλαίσιο Β.28: Υπολογισμός μέτρων αξιολόγησης του εξισορροπημένου μοντέλου με γραμμική διακριτική ανάλυση χωρίς συγγραμικές μεταβλητές για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης test_binary

```
bioact<-as.factor(y_test_binary)
xlassotest<-model.matrix(bioact~.,X_test_binary)[,-1]
predlaso1<-as.factor(predict(cv.lasso, newx = xlassotest, s =c(cv.lasso$lambda.1se
    ),type="class"))
predlaso1<-relevel(predlaso1,"inactive")
cm<-confusionMatrix(predlaso1,bioact,positive="active")
cm
Confusion Matrix and Statistics
          Reference
Prediction inactive active
  inactive       26     43
  active          8    295

               Accuracy : 0.8629
                 95% CI : (0.8237, 0.8962)
    No Information Rate : 0.9086
    P-Value [Acc > NIR] : 0.9985
                  Kappa : 0.4358
 Mcnemar's Test P-Value : 0.000001927
            Sensitivity : 0.8728
            Specificity : 0.7647
         Pos Pred Value : 0.9736
         Neg Pred Value : 0.3768
             Prevalence : 0.9086
         Detection Rate : 0.7930
   Detection Prevalence : 0.8145
      Balanced Accuracy : 0.8187
       'Positive' Class : active
```

Πλαίσιο Β.29: Υπολογισμός μέτρων αξιολόγησης του εξισορροπημένου μοντέλου με λογιστική παλινδρόμηση με Lasso και λ=lambda.1se για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης test_binary

```
predlasom<-as.factor(predict(cv.lasso, newx = xlassotest, s =c(cv.lasso$lambda.min
    ),type="class"))
predlasom<-relevel(predlasom,"inactive")
cmm<-confusionMatrix(predlasom,bioact,positive = "active")
cmm
Confusion Matrix and Statistics
          Reference
Prediction inactive active
  inactive       26     31
  active          8    307

               Accuracy : 0.8952
                 95% CI : (0.8595, 0.9244)
    No Information Rate : 0.9086
    P-Value [Acc > NIR] : 0.839133
                  Kappa : 0.516
 Mcnemar's Test P-Value : 0.000427
            Sensitivity : 0.9083
            Specificity : 0.7647
         Pos Pred Value : 0.9746
         Neg Pred Value : 0.4561
             Prevalence : 0.9086
         Detection Rate : 0.8253
   Detection Prevalence : 0.8468
      Balanced Accuracy : 0.8365
       'Positive' Class : active
```

Πλαίσιο Β.30: Υπολογισμός μέτρων αξιολόγησης του εξισορροπημένου μοντέλου με λογιστική παλινδρόμηση με Lasso και λ=lambda.min για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης test_binary

```
decision_tree_up <- rpart(activity~., data = up_train, method = 'class')
predicted_d2_tree <-predict(decision_tree_up ,test_binary,na.action=na.pass,type="
    class")
caret::confusionMatrix(predicted_d2_tree,as.factor(test_binary$activity),positive=
    "active")
Confusion Matrix and Statistics
          Reference
Prediction inactive active
  inactive       17     25
  active         17    313

               Accuracy : 0.8871
                 95% CI : (0.8505, 0.9174)
    No Information Rate : 0.9086
    P-Value [Acc > NIR] : 0.9333
                  Kappa : 0.3853
 Mcnemar's Test P-Value : 0.2801
            Sensitivity : 0.9260
            Specificity : 0.5000
         Pos Pred Value : 0.9485
         Neg Pred Value : 0.4048
             Prevalence : 0.9086
         Detection Rate : 0.8414
   Detection Prevalence : 0.8871
      Balanced Accuracy : 0.7130
       'Positive' Class : active
```

Πλαίσιο Β.31: Υπολογισμός μέτρων αξιολόγησης του εξισορροπημένου μοντέλου με δένδρο απόφασης για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης test_binary

```
#Run algorithms using 10-fold cross validaiton
#Prepare the test harness for evaluating algorithms
require(mlbench)
trainControl <- trainControl(method="cv",number=10)
metric<-"Accuracy"


#GLM
set.seed(7)
fit.glm<-train(activity~.,data=train_binary,method="glm",metric=metric,trControl=
    trainControl)


#GLMNET
set.seed(7)
fit.glmnet<-train(activity~.,data=train_binary,method="glmnet",metric=metric,
    trControl=trainControl)

#LDA
set.seed(7)
fit.lda<-train(activity~.,data=train_binary,method="lda",metric=metric,trControl=
    trainControl)

#CART2
set.seed(7)
fit.d2.tree<-train(activity~.,data=train_binary,method="rpart",metric=metric,
    trControl=trainControl)


#CART3
set.seed(7)
fit.d3.tree<-train(activity_multi~.,data=train_triple,method="rpart",metric=metric
    ,trControl=trainControl)

#SVM
set.seed(7)
fit.svm<-train(activity_multi~.,data=train_triple,method="svmRadial",metric=metric
    ,trControl=trainControl)

#RandomForest
set.seed(7)
fit.rf<-train(activity_multi~.,data=train_triple,method="rf",metric=metric,
    trControl=trainControl)


#Multinomial
set.seed(7)
fit.multinom<-train(activity_multi~.,data=train_triple,method="multinom",metric=
    metric,trControl=trainControl)
```

Πλαίσιο B.32: Υπολογισμός μοντέλων με χρήση διασταυρωμένης επικύρωσης k-τμημάτων

```
pred<-predict(fit.glm,test_binary)
prediction from a rank-deficient fit may be misleading
confusionMatrix(pred,test_binary$activity,positive = "active")
Confusion Matrix and Statistics
          Reference
Prediction inactive active
  inactive       16     10
  active         18    328

               Accuracy : 0.9247
                 95% CI : (0.8931, 0.9494)
    No Information Rate : 0.9086
    P-Value [Acc > NIR] : 0.1611
                  Kappa : 0.4932
 Mcnemar's Test P-Value : 0.1859
            Sensitivity : 0.9704
            Specificity : 0.4706
         Pos Pred Value : 0.9480
         Neg Pred Value : 0.6154
             Prevalence : 0.9086
         Detection Rate : 0.8817
   Detection Prevalence : 0.9301
      Balanced Accuracy : 0.7205
       'Positive' Class : active
```

Πλαίσιο Β.33: Υπολογισμός μέτρων αξιολόγησης του εκπαιδευμένου με διασταυρωμένη επικύρωση μοντέλου λογιστικής παλινδρόμησης για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης test_binary

```
pred<-predict(fit.glmnet,test_binary)
confusionMatrix(pred,test_binary$activity,positive = "active")
Confusion Matrix and Statistics

          Reference
Prediction inactive active
  inactive       13      5
  active         21    333

               Accuracy : 0.9301
                 95% CI : (0.8993, 0.9538)
    No Information Rate : 0.9086
    P-Value [Acc > NIR] : 0.084879
                  Kappa : 0.4662
 Mcnemar's Test P-Value : 0.003264
            Sensitivity : 0.9852
            Specificity : 0.3824
         Pos Pred Value : 0.9407
         Neg Pred Value : 0.7222
             Prevalence : 0.9086
         Detection Rate : 0.8952
   Detection Prevalence : 0.9516
      Balanced Accuracy : 0.6838
       'Positive' Class : active
```

Πλαίσιο Β.34: Υπολογισμός μέτρων αξιολόγησης του εκπαιδευμένου με διασταυρωμένη επικύρωση μοντέλου λογιστικής παλινδρόμησης με Lasso και λ=lambda.min για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης test_binary

```
pred<-predict(fit.lda,test_binary)
confusionMatrix(pred,test_binary$activity,positive = "active")
Confusion Matrix and Statistics
          Reference
Prediction inactive active
  inactive       17     18
  active         17    320

               Accuracy : 0.9059
                 95% CI : (0.8716, 0.9336)
    No Information Rate : 0.9086
    P-Value [Acc > NIR] : 0.615
                  Kappa : 0.4409
 Mcnemar's Test P-Value : 1.000
            Sensitivity : 0.9467
            Specificity : 0.5000
         Pos Pred Value : 0.9496
         Neg Pred Value : 0.4857
             Prevalence : 0.9086
         Detection Rate : 0.8602
   Detection Prevalence : 0.9059
      Balanced Accuracy : 0.7234
       'Positive' Class : active
```

Πλαίσιο Β.35: Υπολογισμός μέτρων αξιολόγησης του εκπαιδευμένου με διασταυρωμένη επικύρωση μοντέλου γραμικής διακριτικής ανάλυσης για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης test_binary

```
pred<-predict(fit.d2.tree,test_binary)
confusionMatrix(pred,test_binary$activity,positive = "active")
Confusion Matrix and Statistics
          Reference
Prediction inactive active
  inactive        9      4
  active         25    334

               Accuracy : 0.922
                 95% CI : (0.89, 0.9472)
    No Information Rate : 0.9086
    P-Value [Acc > NIR] : 0.2114622
                  Kappa : 0.3501
 Mcnemar's Test P-Value : 0.0002041
            Sensitivity : 0.9882
            Specificity : 0.2647
         Pos Pred Value : 0.9304
         Neg Pred Value : 0.6923
             Prevalence : 0.9086
         Detection Rate : 0.8978
   Detection Prevalence : 0.9651
      Balanced Accuracy : 0.6264
       'Positive' Class : active
```

Πλαίσιο Β.36: Υπολογισμός μέτρων αξιολόγησης του εκπαιδευμένου με διασταυρωμένη επικύρωση μοντέλου δένδρου απόφασης για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης test_binary

```
pred<-predict(fit.d3.tree,test_triple)
confusionMatrix(pred,test_triple$activity_multi,positive = "active")
Confusion Matrix and Statistics
                    Reference
Prediction           inactive moderately active highly active
  inactive                 11                 4             2
  moderately active         6                12             8
  highly active            17                61           251

Overall Statistics
               Accuracy : 0.7366
                 95% CI : (0.6887, 0.7806)
    No Information Rate : 0.7016
    P-Value [Acc > NIR] : 0.07711
                  Kappa : 0.2699
 Mcnemar's Test P-Value : 0.00000000001877

Statistics by Class:
                     Class: inactive Class: moderately active Class: highly active
Sensitivity                  0.32353                  0.15584               0.9617
Specificity                  0.98225                  0.95254               0.2973
Pos Pred Value               0.64706                  0.46154               0.7629
Neg Pred Value               0.93521                  0.81214               0.7674
Prevalence                   0.09140                  0.20699               0.7016
Detection Rate               0.02957                  0.03226               0.6747
Detection Prevalence         0.04570                  0.06989               0.8844
Balanced Accuracy            0.65289                  0.55419               0.6295
```

Πλαίσιο Β.37: Υπολογισμός μέτρων αξιολόγησης του εκπαιδευμένου με διασταυρωμένη επικύρωση μοντέλου δένδρου απόφασης για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης test_triple

```
pred<-predict(fit.svm,test_triple)
confusionMatrix(pred,test_triple$activity_multi,positive = "active")
Confusion Matrix and Statistics

                    Reference
Prediction           inactive moderately active highly active
  inactive                 17                 2             1
  moderately active         7                22             1
  highly active            10                53           259

Overall Statistics
               Accuracy : 0.8011
                 95% CI : (0.7568, 0.8404)
    No Information Rate : 0.7016
    P-Value [Acc > NIR] : 0.0000089808579073
                  Kappa : 0.4639
 Mcnemar's Test P-Value : 0.0000000000005287

Statistics by Class:
                     Class: inactive Class: moderately active Class: highly active
Sensitivity                  0.50000                  0.28571               0.9923
Specificity                  0.99112                  0.97288               0.4324
Pos Pred Value               0.85000                  0.73333               0.8043
Neg Pred Value               0.95170                  0.83918               0.9600
Prevalence                   0.09140                  0.20699               0.7016
Detection Rate               0.04570                  0.05914               0.6962
Detection Prevalence         0.05376                  0.08065               0.8656
Balanced Accuracy            0.74556                  0.62930               0.7124
```

Πλαίσιο Β.38: Υπολογισμός μέτρων αξιολόγησης του εκπαιδευμένου με διασταυρωμένη επικύρωση μοντέλου μηχανών διανυσμάτων υποστήριξης για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης test_triple

```
pred<-predict(fit.rf,test_triple)
confusionMatrix(pred,test_triple$activity_multi,positive = "active")
Confusion Matrix and Statistics
                    Reference
Prediction           inactive moderately active highly active
  inactive                 17                 3             1
  moderately active         7                37             8
  highly active            10                37           252

Overall Statistics
                Accuracy : 0.8226
                  95% CI : (0.7799, 0.86)
    No Information Rate : 0.7016
    P-Value [Acc > NIR] : 0.0000000622
                   Kappa : 0.5586
 Mcnemar's Test P-Value : 0.0000042960

Statistics by Class:
                     Class: inactive Class: moderately active Class: highly active
Sensitivity                  0.50000                  0.48052                0.9655
Specificity                  0.98817                  0.94915                0.5766
Pos Pred Value               0.80952                  0.71154                0.8428
Neg Pred Value               0.95157                  0.87500                0.8767
Prevalence                   0.09140                  0.20699                0.7016
Detection Rate               0.04570                  0.09946                0.6774
Detection Prevalence         0.05645                  0.13978                0.8038
Balanced Accuracy            0.74408                  0.71484                0.7710
```

Πλαίσιο Β.39: Υπολογισμός μέτρων αξιολόγησης του εκπαιδευμένου με διασταυρωμένη επικύρωση μοντέλου τυχαίου δάσους για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης test_triple

```
pred<-predict(fit.multinom,test_triple)
confusionMatrix(pred,test_triple$activity_multi,positive = "active")
Confusion Matrix and Statistics
                    Reference
Prediction           inactive moderately active highly active
  inactive                 17                 9             1
  moderately active        12                23            15
  highly active             5                45           245

Overall Statistics
                Accuracy : 0.7661
                  95% CI : (0.7197, 0.8082)
    No Information Rate : 0.7016
    P-Value [Acc > NIR] : 0.0033010
                   Kappa : 0.4284
 Mcnemar's Test P-Value : 0.0004204

Statistics by Class:
                     Class: inactive Class: moderately active Class: highly active
Sensitivity                  0.50000                  0.29870                0.9387
Specificity                  0.97041                  0.90847                0.5495
Pos Pred Value               0.62963                  0.46000                0.8305
Neg Pred Value               0.95072                  0.83230                0.7922
Prevalence                   0.09140                  0.20699                0.7016
Detection Rate               0.04570                  0.06183                0.6586
Detection Prevalence         0.07258                  0.13441                0.7930
Balanced Accuracy            0.73521                  0.60359                0.7441
```

Πλαίσιο Β.40: Υπολογισμός μέτρων αξιολόγησης του εκπαιδευμένου με διασταυρωμένη επικύρωση μοντέλου πολυωνυμικής λογιστικής παλινδρόμησης για πρόβλεψη παρατηρήσεων του υποδείγματος εκπαίδευσης test_triple

Λίστα Γραφημάτων

Λίστα Πινάκων

Λίστα Πλαισίων

Βιβλιογραφία

Arrowsmith, J. (2011a). "Phase II failures: 2008–2010". *Nature Reviews Drug Discovery 2011* 10 : 328–329.

— (2011b). "Phase III and submission failures: 2007–2010". *Nature Reviews Drug Discovery* 10 : 87.

Augen, J. (2002). "The evolving role of information technology in the drug discovery process". *Drug Discovery Today* 7 : 315–323.

Breiman, L. (2001). "Random forests". *Machine Learning* 45 : 5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. Monterey, CA: Wadsworth and Brooks.

Chawla, N. V. (2009). "Data mining for imbalanced datasets: An overview". In: *Data Mining and Knowledge Discovery Handbook*. Ed. by R. L. Maimon O. Boston, MA: Springer US.

Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., and Blaschke, T. (2018). "The rise of deep learning in drug discovery". *Drug Discovery Today* 23 : 1241–1250.

Church, M. K. and Church, D. S. (2013). "Pharmacology of antihistamines". *Indian Journal of Dermatology* 58 : 219.

Cortes, C. and Vapnik, V. (1995). "Support-vector networks". *Machine Learning* 20 : 273–297.

Cumming, S. (2016). *Drug Repositioning Quickly Meeting Demand for Novel Therapeutics at Lower Cost*. URL: `https://www.bccresearch.com/pressroom/phm/drug-repositioning-quickly-meeting-demand-for-novel-therapeutics-at-lower-cost`.

Diseases | Genetic and Rare Diseases Information Center (GARD) – an NCATS Program (n.d.). URL: `https://rarediseases.info.nih.gov/diseases`.

Drakakis, G., Koutsoukas, A., Brewerton, S. C., Evans, D. D., and Bender, A. (2013). "Using machine learning techniques for rationalising phenotypic readouts from a rat sleeping model". *Journal of Cheminformatics* 5 : 1–1.

Drakakis, G., Wafford, K. A., Brewerton, S. C., Bodkin, M. J., Evans, D. A., and Bender, A. (2017). "Polypharmacological in silico bioactivity profiling and experimental validation uncovers sedative-hypnotic effects of approved and experimental drugs in rat". *ACS Chemical Biology* 12 : 1593–1602.

"Drug repurposing: a promising tool to accelerate the drug discovery process" (2019). *Drug Discovery Today* 24 : 2076–2085.

Ekins, S., Puhl, A. C., Zorn, K. M., Lane, T. R., Russo, D. P., Klein, J. J., Hickey, A. J., and Clark, A. M. (2019). "Exploiting machine learning for end-to-end drug discovery and development". *Nature Materials* 18 : 435.

Ferreira, L. and Andricopulo, A. (2019). "From chemoinformatics to deep learning: An open road to drug discovery". *Future Medicinal Chemistry* 11 : 371–374.

Gatta, G., Van Der Zwan, J. M., Casali, P. G., Siesling, S., Dei Tos, A. P., Kunkler, I., Otter, R., Licitra, L., Mallone, S., Tavilla, A., et al. (2011). "Rare cancers are not so rare: The rare cancer burden in Europe". *European Journal of Cancer* 47 : 2493–2511.

Gaulton, A., Hersey, A., -l Nowotka, M., Patrícia Bento, A., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrí an-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Paula MagariñosMagari, M., Overington, J. P., Papadatos, G., Smit, I., and Leach, A. R. (2016). "The ChEMBL database in 2017". *Nucleic Acids Research* 45 : 945–954. URL: `https://www.ebi.ac.uk/chembl`.

Haas, H. and Panula, P. (2003). "The role of histamine and the tuberomamillary nucleus in the nervous system". *Nature Reviews Neuroscience* 4 : 121–130.

Haas, H., Sergeeva, O., and Selbach, O. (Aug. 2008). "Histamine in the nervous system". *Physiological Reviews* 88 : 1183–241.

Hall, D. G., Manku, S., and Wang, F. (2001). "Solution-and solid-phase strategies for the design, synthesis, and screening of libraries based on natural product templates: a comprehensive survey". *Journal of Combinatorial Chemistry* 3 : 125–150.

Hall, L. H. and Kier, L. B. (1991). "The molecular connectivity chi indexes and kappa shape indexes in structure-property modeling". *Reviews in Computational Chemistry* : 367–422.

Harrell Jr, F. E. (2015). *Regression Modeling Strategies: with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Heidelberg: Springer.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer Science & Business Media.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Vol. 103. Springer Texts in Statistics. New York, NY: Springer New York. URL: http://link.springer.com/10.1007/978-1-4614-7138-7.

Khelemsky, Y., Gritsenko, K., and Maerz, D. (2017). "Antihistamines". In: *Pain Medicine: An Essential Review*. Ed. by R. Yong, M. Nguyen, E. Nelson, and R. Urman. Springer International Publishing.

Kier, L. B. and Hall, L. H. (1977). "The nature of structure-activity relationships and their relation to molecular connectivity". *European Journal of Medicinal Chemistry* 12 : 307–312.

Kola, I. and Landis, J. (2004). "Can the pharmaceutical industry reduce attrition rates?" *Nature Reviews Drug Discovery* 3 : 711–716. URL: http://www.nature.com/articles/nrd1470.

Koutsoukas, A., Simms, B., Kirchmair, J., Bond, P. J., Whitmore, A. V., Zimmer, S., Young, M. P., Jenkins, J. L., Glick, M., Glen, R. C., et al. (2011). "From in silico target prediction to multi-target drug design: current databases, methods and applications". *Journal of Proteomics* 74.12 : 2554–2574.

Kowalski, B. R. (1974). "Pattern recognition in chemical research". In: *Computers in Chemical and Biochemical Research*. Ed. by C. L. W. Charles E. Klopfenstein. Elsevier.

Kuang, Z., Bao, Y., Thomson, J., Caldwell, M., Peissig, P., Stewart, R., Willett, R., and Page, D. (2019). "A machine-learning-based drug repurposing approach using baseline regularization". In: *Computational Methods for Drug Repurposing*. Ed. by V. Quentin. New York: Humana Press.

Kuhn, M., Wing, Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., R Core Team, the, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., and Hunt., T. (2019). *caret: Classification and regression training*. R package version 6.0-84. URL: https://CRAN.R-project.org/package=caret.

Kulkarni, N. S., Guererro, Y., Gupta, N., Muth, A., and Gupta, V. (2019). "Exploring potential of quantum dots as dual modality for cancer therapy and diagnosis". *Journal of Drug Delivery Science and Technology* 49 : 352–364.

Kulkarni, N. S., Parvathaneni, V., Shukla, S. K., Barasa, L., Perron, J. C., Yoganathan, S., Muth, A., and Gupta, V. (2019). "Tyrosine kinase inhibitor conjugated quantum dots for non-small cell lung cancer (NSCLC) treatment". *European Journal of Pharmaceutical Sciences* 133 : 145–159.

Leurs, R., Hough, L. B., Blandina, P., and Haas, H. L. (2012). "Histamine". In: *Basic Neurochemistry*. Elsevier.

Lewis, J., Snyder, M., and Hyatt-Knorr, H. (2017). "Marking 15 years of the Genetic and Rare Diseases Information Center". *Translational Science of Rare Diseases* 2 : 77–88.

Liaw, A., Wiener, M., et al. (2002). "Classification and regression by randomForest". *R News* 2 : 18–22.

Lo, Y.-C., Rensi, S. E., Torng, W., and Altman, R. B. (2018). "Machine learning in chemoinformatics and drug discovery". *Drug Discovery Today* 23 : 1538–1546.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2019). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.7-2. URL: `https://CRAN.R-project.org/package=e1071`.

Murphy, K. P. (2012). *Machine learning : a probabilistic perspective*. Cambridge, Mass. [u.a.]: MIT Press.

Padhy, B. and Gupta, Y. (2011). "Drug repositioning: Re-investigating existing drugs for new therapeutic indications". *Journal of Postgraduate Medicine* 57 : 153.

Paolini, G. V., Lyons, R. A., and Laflin, P. (2010). "How desirable are your IC50s? A way to enhance screening-based decision making". *Journal of Biomolecular Screening* 15 : 1183–1193.

Pollard, C. E., Skinner, M., Lazic, S. E., Prior, H. M., Conlon, K. M., Valentin, J.-P., and Dota, C. (2017). "An analysis of the relationship between preclinical and clinical QT interval-related data". *Toxicological Sciences* 159 : 94–101.

Popova, M., Isayev, O., and Tropsha, A. (2018). "Deep reinforcement learning for de novo drug design". *Science Advances* 4 : eaap7885.

Quinlan, J. R. (2014). *C4. 5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers.

Ramasubramanian, K. and Singh, A. (2016). *Machine Learning Using R*. Berkeley, CA: Apress Media LLC.

Rognan, D. (2010). "Structure-based approaches to target fishing and ligand profiling". *Molecular Informatics* 29.3 : 176–187.

Schloerke, B., Crowley, J., Cook, D., Briatte, F., Marbach, M., Thoen, E., Elberg, A., and Larmarange, J. (2018). *GGally: Extension to 'ggplot2'*. R package version 1.4.0. URL: `https://CRAN.R-project.org/package=GGally`.

Schneider, G. (2018). "Generative models for artificially-intelligent molecular design". *Molecular Informatics* 37 : 1880131.

Schneider, G. and Clark, D. E. (2019). *Automated de novo drug design: are we nearly there yet?* URL: http://doi.wiley.com/10.1002/anie.201814681.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011a). "Regularization paths for cox's proportional hazards model via coordinate descent". *Journal of Statistical Software* 45 : 945–954.

— (2011b). "Regularization paths for generalized linear models via coordinate descent". *Journal of Statistical Software* 45 : D945–D954. URL: http://www.jstatsoft.org/v33/i01/.

Tabarean, I. (2013). "Functional pharmacology of H1 histamine receptors expressed in mouse preoptic/anterior hypothalamic neurons". *British Journal of Pharmacology* 170 : 415–425.

Thangam, E. B., Jemima, E. A., Singh, H., Baig, M. S., Khan, M., Mathias, C. B., Church, M. K., and Saluja, R. (2018). "The role of histamine and histamine receptors in mast cell-mediated allergy and inflammation: The hunt for new therapeutic targets". *Frontiers in Immunology* 9 : 1873.

Therneau, T. and Atkinson, B. (2019). *rpart: Recursive partitioning and regression trees*. R package version 4.1-15. URL: https://CRAN.R-project.org/package=rpart.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso". *Journal of the Royal Statistical Society: Series B (Methodological)* 58 : 267–288.

Toulemonde, E.-L. (2020). *dataPreparation: Automated data preparation*. R package version 0.4.3. URL: https://CRAN.R-project.org/package=dataPreparation.

Vaidya, B., Parvathaneni, V., Kulkarni, N. S., Shukla, S. K., Damon, J. K., Sarode, A., Kanabar, D., Garcia, J. V., Mitragotri, S., Muth, A., et al. (2019). "Cyclodextrin modified erlotinib loaded PLGA nanoparticles for improved therapeutic efficacy against non-small cell lung cancer". *International Journal of Biological Macromolecules* 122 : 338–347.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., et al. (2019). "Applications of machine learning in drug discovery and development". *Nature Reviews Drug Discovery* 18.6 : 463–477.

Venables, W. N. and Ripley, B. D. (2013). *Modern Applied Statistics with S-PLUS*. 4th ed. New York: Springer. URL: http://www.stats.ox.ac.uk/pub/MASS4.

Wagner, J. A., Dahlem, A. M., Hudson, L. D., Terry, S. F., Altman, R. B., Gilliland, C. T., DeFeo, C., and Austin, C. P. (2018). "Application of a dynamic map for learning, communicating, navigating, and improving therapeutic development". *Clinical and Translational Science* 11 : 166–174.

Wickham, H., François, R., Henry, L., and Müller, K. (2019). *dplyr: A grammar of data manipulation*. R package version 0.8.3. URL: https://CRAN.R-project.org/package=dplyr.

Xu, J. and Hagler, A. (2002). "Chemoinformatics and drug discovery". *Molecules* 7.8 : 566–600.

Zheng, W., Johnson, S. R., Baskin, I., Bajorath, J., Horvath, D., Laggner, C., Langer, T., Schneider, G., Filimonov, D., Poroikov, V., Tetko, I., Van De Waterbeemd, H., Oprea, T., Radchenko, E., Palyulin, V., Zefirov, N., Peltason, L., Wolber, G., Schuster, D., Kirchmair, J., Proschak, E., and Tanrikulu, Y. (2008). *Chemoinformatics Approaches to Virtual Screening*. Ed. by A. T. Alexandre Varnek. Cambridge, UK: Royal Society of Chemistry.