ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΒΙΟΪΑΤΡΙΚΗΣ

Μελέτες ταξινόμησης φυσικών προϊόντων με εργαλεία Μηχανικής Μάθησης και Χημειοπληροφορικής

ΠΑΝΑΓΙΩΤΟΠΟΥΛΟΣ ΒΑΣΙΛΗΣ
Αριθμός Μητρώου: 16080

Επιβλέπων Καθηγητής
Διονύσιος Κάβουρας, Ομότιμος Καθηγητής

Αθήνα 22/7/2021

Η Τριμελής Εξεταστική Επιτροπή

Επιβλέπων		
Διονύσιος Κάβουρας	Παντελής Ασβεστάς	Παναγιώτης Ζουμπουλάκης
Ομότιμος Καθηγητής	Αναπληρωτής Καθηγητής	Αναπληρωτής Καθηγητής
[ΨΗΦΙΑΚΗ ΥΠΟΓΡΑΦΗ]	[ΨΗΦΙΑΚΗ ΥΠΟΓΡΑΦΗ]	[ΨΗΦΙΑΚΗ ΥΠΟΓΡΑΦΗ]
Dionysios Kavouras Digitally signed by Dionysios Kavouras Date: 2021.07.29 14:38:39 +03'00'	Digitally signed by Panteleimon Asvestas	Panagiotis Zoumpoulakis Digitally signed by Panagiotis Zoumpoulakis Date: 2021.07.29 14:49:09 +03'00'

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Ο κάτωθι υπογεγραμμένος **Παναγιωτόπουλος Βασίλειος** του **Αποστόλου**, με αριθμό μητρώου **16080** φοιτητής του Τμήματος **Μηχανικών Βιοϊατρικής** της Σχολής Μηχανικών του Πανεπιστημίου Δυτικής Αττικής, δηλώνω υπεύθυνα ότι:
«Είμαι συγγραφέας αυτής της διπλωματικής εργασίας και κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος.
Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου».

Ημερομηνία

22/7/2021

Ο Δηλών

ΠΕΡΙΛΗΨΗ

Σκοπός
Τα φυσικά προϊόντα είναι χημικές ενώσεις οι οποίες απαντώνται στην φύση. Τα τελευταία χρόνια η αξιοποίηση τους στον χώρο της φαρμακοβιομηχανίας έχει αυξηθεί ραγδαία. Επιπλέον, γίνονται ολοένα και περισσότερες προσπάθειες να συνδυαστούν αλγόριθμοι Μηχανικής Μάθησης και μέθοδοι Χημειοπληροφορικής, ώστε χρησιμοποιώντας είτε φυσικοχημικές, είτε βιολογικές ιδιότητες χημικών ενώσεων, να βοηθήσουν προς την κατεύθυνση καλύτερων προβλέψεων στην διαδικασία ανακάλυψης νέων φαρμάκων. Η παρούσα διπλωματική εργασία έχει ως στόχο να μελετήσει την ικανότητα αλγορίθμων ταξινόμησης Μηχανικής Μάθησης να διαχωρίσουν ενώσεις φυσικών προϊόντων σε δύο κατηγορίες. Για να επιτευχθεί αυτό αξιοποιούνται δεδομένα που προκύπτουν από μεθόδους Χημειοπληροφορικής.

Μεθοδολογία
Δύο φυσικά προϊόντα, η κουρκουμίνη και η ρεσβερατρόλη, επιλέχθηκαν ως ενώσεις αναφοράς. Υπολογίστηκαν τα μοριακά αποτυπώματά τους και μέσα από μία βάση δεδομένων φυσικών προϊόντων επιλέχθηκαν ενώσεις των οποίων η χημική δομή είναι παρόμοια με την χημική δομή των ενώσεων αναφοράς. Η αναζήτηση έγινε με βάση τον αλγόριθμο ομοιότητας Tanimoto. Από τις όμοιες ενώσεις που προέκυψαν δημιουργήθηκαν δύο σύνολα δεδομένων, ένα για την κουρκουμίνη και τις ενώσεις που ήταν δομικά παρόμοιες με αυτή και ένα για την ρεσβερατρόλη αντίστοιχα. Στην συνέχεια, για την κάθε ένωση υπολογίστηκαν 208 μοριακοί περιγραφείς. Μετά από κατάλληλη επεξεργασία τα τελικά σύνολα δεδομένων αποτελούνταν από 79 ενώσεις με 64 μοριακούς περιγραφείς για το σύνολο δεδομένων της κουρκουμίνης και 78 ενώσεις με 64 μοριακούς περιγραφείς για το σύνολο δεδομένων της ρεσβερατρόλης. Έπειτα, ακολούθησε η εκπαίδευση και η αξιολόγηση 10 αλγορίθμων ταξινόμησης Μηχανικής Μάθησης, η οποία χρίστηκες σε δύο περιπτώσεις. Στην πρώτη περίπτωση χρησιμοποιήθηκαν όλοι οι 64 μοριακοί περιγραφείς για την εκπαίδευση και την αξιολόγηση όλων των ταξινομητών και ακολούθως αξιοποιήθηκαν συνδυασμοί αυτών ανά 63, 62, 61 και 60 περιγραφέων για τον ταξινομητή που σημείωσε την καλύτερη επίδοση. Στην δεύτερη περίπτωση έγινε επιλογή 26 βέλτιστων μοριακών περιγραφέων μέσα από έναν αλγόριθμο μείωσης χαρακτηριστικών. Αντίστοιχα με την προηγούμενη περίπτωση χρησιμοποιήθηκαν όλοι οι βέλτιστοι περιγραφείς για την εκπαίδευση και αξιολόγηση όλων των ταξινομητών και στην συνέχεια αξιοποιηθήκαν όλοι οι πιθανοί συνδυασμοί ανά 25, 24, 23, 22, 21, 20 και ανά 19 μοριακοί περιγραφείς για τον καλύτερο ταξινομητή.

Αποτελέσματα
Τα αποτελέσματα από την αναζήτηση ομοιότητας Tanimoto στην βάση δεδομένων έδειξαν ότι για τιμές Tanimoto μεγαλύτερες ή ίσες από 0,75 υπάρχουν 79 ενώσεις με χημική δομή παρόμοια με την δομή της κουρκουμίνης και 79 ενώσεις με χημική δομή παρόμοια με την δομή της ρεσβερατρόλης. Την καλύτερη επίδοση σημείωσε ο ταξινομητής Random Forest, τόσο για τον συνδυασμό των 64 μοριακών περιγραφέων όσο και για τον συνδυασμό των 26 βέλτιστων μοριακών περιγραφέων, ενώ ο ταξινομητής MLP σημείωσε τις χειρότερες επιδόσεις και στις δύο περιπτώσεις.

Συμπεράσματα

Η αξιοποίηση όσο το δυνατόν περισσότερων μοριακών περιγραφέων παρέχει καλύτερα αποτελέσματα αξιοποιώντας περισσότερη χρήσιμη φυσικοχημική πληροφορία. Κρίνεται απαραίτητη η μείωση των χαρακτηριστικών, εν προκειμένω των μοριακών περιγραφέων και η επιλογή βέλτιστων μοριακών περιγραφέων, καθώς με αυτόν τον τρόπο οι ταξινομητές πετυχαίνουν υψηλότερα ποσοστά ακρίβειας και μειώνεται το ενδεχόμενο υπερπροσαρμογής. Χρειάζεται να γίνουν περισσότερες μελέτες με μεγαλύτερο αριθμό δεδομένων ώστε να εξαχθούν πιο αποτελεσματικά συμπεράσματα.

Λέξεις Κλειδιά: Χημειοπληροφορική, Μηχανική Μάθηση, φυσικά προϊόντα, κουρκουμίνη, ρεσβερατρόλη, ομοιότητα Tanimoto, μοριακοί περιγραφείς, μοριακά αποτυπώματα

Classification of natural products employing Machine Learning algorithms and Chemoinformatics methods

ABSTRACT

Purpose

Natural products are chemical compounds that can be found in nature. The utilization of natural products in pharmaceutical industry has risen over the past years. In addition, more efforts are being made to combine Machine Learning algorithms and Chemoinformatics methods in order to lean towards better predictions in the process of drug discovery, by using either physicochemical or biological properties of chemical compounds. This thesis aims to study the capability of Machine Learning classification algorithms to discriminate natural product compounds into two classes. To achieve this goal, data extracted with Chemoinformatics methods are being used.

Methods

Curcumin and resveratrol are two natural products, which are being used as reference compounds. Their molecular fingerprints are being calculated and then compounds with similar chemical structure are being selected through a natural product database. Tanimoto similarity algorithm is being used for similarity search inside the database. Two datasets are being created from the similar compounds, one dataset for curcumin and its similars and one dataset for resveratrol and its equivalents. Subsequently, 208 molecular descriptors are being calculated for every compound. After proper processing, final datasets contain 79 compounds with 64 descriptors for curcumin dataset and 78 compounds with 64 descriptors for resveratrol dataset. The process of training and evaluation of 10 Machine Learning classification algorithms is divided into two different cases. On the first case, 64 descriptors used for training and evaluation of all classifiers and then combinations per 63, 62, 61 and 60 descriptors used in accordance with the best classifier. On the second case, 26 best descriptors were selected throughout a feature elimination algorithm. As before, all 26 best descriptors used for training and evaluation of all classifiers and then combinations per 25, 24, 23, 22, 21, 20 and 19 descriptors used in accordance with the best classifier.

Results

The results from the Tanimoto similarity search inside the database showed that there are 79 structure similar compounds with curcumin and 79 structure similar compounds with resveratrol, for Tanimoto values greater or equal to 0,75. The classifier with the best accuracy results for the combination of 64 descriptors and the combination of 26 best descriptors was Random Forest classifier. On the other hand, MLP classifier had the worst accuracy results for both cases.

Conclusion

Employing as many as possible molecular descriptors provides sufficient results taking into account more useful physicochemical information. Feature reduction, in which case molecular descriptors and selection of best descriptors is crucial, because in this way classifiers perform better and provide higher accuracy results and also the chance of overfitting is being avoided. More studies with larger amount of data must be conducted in order to extract more sufficient results.

Keywords: Chemoinformatics, Machine Learning, natural products, curcumin, resveratrol, Tanimoto similarity, molecular descriptors, molecular fingerprints

Ευχαριστίες

Θα ήθελα να ευχαριστήσω αρχικά τον επιβλέποντα καθηγητή κ. Διονύση Κάβουρα για τις γνώσεις που μου μετέδωσε, αλλά και για τις ευκαιρίες που μου έδωσε κατά την διάρκεια των σπουδών μου. Επίσης, θα ήθελα να ευχαριστήσω θερμά τον κ. Μίνω Ματσούκα για τις συμβουλές και την καθοδήγηση που μου παρείχε κατά την συγγραφή της παρούσας εργασίας και για την εξαιρετική συνεργασία που είχαμε στα πλαίσια της πρακτικής μου άσκησης. Ακόμα, οφείλω να ευχαριστήσω τον συνάδελφο Σωτήρη Ουζούνη για την συνεισφορά και την βοήθεια που μου παρείχε όποτε την χρειάστηκα. Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου και τους φίλους μου για την στήριξη και την υπομονή που έδειξαν σε όλη την διάρκεια των σπουδών μου.

ΠΕΡΙΕΧΟΜΕΝΑ

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Κατάλογος συντομογραφιών	
ECFP	Extended Connectivity Fingerprint (Αποτύπωμα εκτεταμένης σύνδεσης)
FCFP	Functional Class Fingerprint (Αποτύπωμα λειτουργικής κλάσης)
SMILES	Simplified Molecular Input Line Entry System (Σύστημα απλοποιημένης μοριακής γραμμικής γραφής)
SMARTS	SMILES Arbitrary Target Specification
InChi	International Chemical Identifier (Διεθνές χημικό αναγνωριστικό)
IUPAC	International Union of Pure and Applied Chemistry (Διεθνής Ένωση Καθαρής και Εφαρμοσμένης Χημείας)
QSAR	Quantitative Structure-Activity Relationship (ποσοτική σχέση δομής-δραστικότητας)
QSPR	Quantitative Structure-Property Relationship (ποσοτική σχέση δομής-ιδιότητας)
KNN	k-Nearest Neighbors (ταξινομητής k-πλησιέστερων γειτόνων)
LogReg	Logistic Regression (ταξινομητής λογιστική παλινδρόμησης)
LDA	Linear Discriminant Analysis (ανάλυση γραμμικής διάκρισης)
ANOVA	Analysis of Variance (Ανάλυση διακύμανσης)
MDC	Minimum Distance Classifier (ταξινομητής ελάχιστης απόστασης)
SVM	Support Vector Machine (μηχανή διανυσματικής στήριξης)
RBF	Radial Basis Function (συνάρτηση ακτινικής βάσης)
CART	Classification And Regression Tree (δένδρο ταξινόμησης και παλινδρόμησης)
RF	Random Forest (ταξινομητής τυχαίου δάσους)
ANN	Artificial Neural Network (Τεχνητό Νευρωνικό Δίκτυο)
MLP	Multi Layer Perceptron (Perceptron Πολλαπλών Επιπέδων)
PCA	Principal Component Analysis (Ανάλυση Πρωτευόντων Συστατικών)
RFE	Recursive Feature Elimination (αλγόριθμος επαναλαμβανόμενου αποκλεισμού χαρακτηριστικών)
LOO	Leave One Out (μέθοδος παράλειψης ενός προτύπου)
ECV	External Cross Validation (μέθοδος εξωτερικής διασταυρωμένης επικύρωσης)

Εισαγωγή

Με την πάροδο των χρόνων η Τεχνολογία συνεχώς αναπτύσσεται και προσφέρει λύσεις σε ζητήματα τα οποία νωρίτερα δεν μπορούσαν να επιλυθούν. Τις τελευταίες δεκαετίες η Χημειοπληροφορική έχει συνεισφέρει σημαντικά στον κλάδο της Χημείας και της Ιατρικής, παρέχοντας δεδομένα και πληροφορίες, τα οποία απαιτούσαν μεγάλη υπολογιστική ισχύ, σε σύντομο χρονικό διάστημα. Μερικούς από τους τομείς στους οποίους εφαρμόζονται τα εργαλεία της Χημειοπληροφορικής είναι η εξαγωγή δεδομένων από χημικές δομές, η αναζήτηση χημικών ενώσεων σε βάσεις δεδομένων, καθώς και η ανακάλυψη νέων φαρμάκων. Σημαντικό ρόλο στην ανάπτυξη της Χημειοπληροφορικής στον φαρμακευτικό τομέα έχει παίξει και η Μηχανική Μάθηση. Με την αξιοποίηση των αλγορίθμων της Μηχανικής Μάθησης μπορούν να προβλεφθούν και να μοντελοποιηθούν διάφορες ιδιότητες των φαρμάκων, όπως είναι η τοξικότητα, η αλληλεπίδραση με άλλα φάρμακα και η καρκινογένεση. Η μοντελοποίηση αυτή γίνεται συνήθως με μοντέλα που εξετάζουν την ποσοτική σχέση δομής-δραστικότητας (QSAR) τα οποία βρίσκουν ευρεία εφαρμογή στην ανακάλυψη νέων φαρμάκων, καθώς μέσα από αυτά μπορεί και προβλέπεται η βιολογική συμπεριφορά των ενώσεων μετά από πιθανές χημικές τροποποιήσεις. Αρκετά από τα φάρμακα που κυκλοφορούν στην αγορά προέρχονται από χημικές ενώσεις που απαντώνται στην φύση, τα λεγόμενα φυσικά προϊόντα. Κάποιες μελέτες έχουν αξιοποιήσει τα μοντέλα QSAR σε συνδυασμό με αλγορίθμους Μηχανικής Μάθησης, ώστε να μπορέσουν να διακρίνουν ενώσεις φυσικών προϊόντων από συνθετικές ενώσεις. Σε αυτές τις περιπτώσεις αξιοποιείται δομική πληροφορία από την χημική ένωση μαζί με φυσικοχημική πληροφορία που προκύπτει από τον υπολογισμό ορισμένων μοριακών περιγραφέων. Έτσι αναπτύσσονται μέθοδοι οι οποίες μπορούν να διαχωρίσουν αποτελεσματικά τα φυσικά προϊόντα από τις συνθετικές ενώσεις, οι οποίες μπορούν να αξιοποιηθούν περαιτέρω για την ανακάλυψη νέων φαρμάκων.

Σκοπός της παρούσας διπλωματικής εργασίας είναι να μελετήσει διαφορετικούς αλγορίθμους ταξινόμησης Μηχανικής Μάθησης ως προς την ικανότητα τους να διαχωρίσουν ενώσεις φυσικών προϊόντων σε δύο κλάσεις, με βάση την πληροφορία που προκύπτει από την χημική τους δομή. Για να επιτευχθεί αυτό, αξιοποιούνται μέθοδοι Χημειοπληροφορικής για την εξαγωγή δεδομένων από μια βάση δεδομένων φυσικών προϊόντων και στην συνέχεια για την ανάκτηση της χρήσιμης πληροφορίας από κάθε χημική ένωση. Ως ενώσεις αναφοράς για τις δύο κλάσεις χρησιμοποιούνται η κουρκουμίνη και η ρεσβερατρόλη, δύο φυσικά προϊόντα με αρκετά διαφορετικές χημικές δομές αλλά και θεραπευτικές ιδιότητες. Ακολούθως, δημιουργούνται δύο σύνολα δεδομένων με ενώσεις που μοιάζουν δομικά στις ενώσεις αναφοράς και υπολογίζονται κάποια φυσικοχημικά χαρακτηριστικά για την κάθε ένωση. Στην συνέχεια χρησιμοποιούνται αλγόριθμοι Μηχανικής Μάθησης για να γίνει εκπαίδευση και αξιολόγηση του κάθε ταξινομητή που θα κληθεί να ταξινομήσει την κάθε ένωση στη σωστή κλάση.

Το κεφάλαιο 1 περιγράφει θεωρητικές έννοιες και ορισμούς για τις μεθόδους Χημειοπληροφορικής, Βάσεων Δεδομένων και Μηχανικής Μάθησης που αξιοποιήθηκαν στην εργασία, καθώς και κάποια θεωρητικά στοιχεία για τα φυσικά προϊόντα και πιο συγκεκριμένα για την κουρκουμίνη και την ρεσβερατρόλη. Ειδικότερα, αναφέρεται το πώς ορίζεται ο όρος φυσικό προϊόν και συνοπτικά ο ρόλος των φυσικών προϊόντων στην Ιατρική τα τελευταία χρόνια. Επίσης, παρουσιάζεται η δομή και οι θεραπευτικές ιδιότητες της κουρκουμίνης και τις ρεσβερατρόλης, οι οποίες αποτελούν τις ενώσεις αναφοράς. Στην συνέχεια περιγράφονται βασικές έννοιες της Χημειοπληροφορικής και της απεικόνισης μορίων, που αφορούν τους μοριακούς περιγραφείς, τα μοριακά αποτυπώματα, την αναζήτηση ομοιότητας και τα SMILES. Τέλος, γίνεται μια σύντομη αναφορά για την δομή και τον ρόλο των βάσεων δεδομένων στην Χημειοπληροφορική, αναφέρονται τα είδη Μηχανικής Μάθησης και περιγράφεται επιγραμματικά το θεωρητικό υπόβαθρο των αλγορίθμων ταξινόμησης Μηχανικής Μάθησης που χρησιμοποιήθηκαν στην εργασία. Επιπλέον αναφέρονται διάφορες μέθοδοι για την ελάττωση των χαρακτηριστικών καθώς και μέθοδοι αξιολόγησης ενός συστήματος ταξινόμησης Μηχανικής Μάθησης.

Το κεφάλαιο 2 πραγματεύεται την μεθοδολογία και την διαδικασία που ακολουθήθηκε για την ανάκτηση των δεδομένων και την λήψη των αποτελεσμάτων. Εκτενέστερα, περιγράφεται η διαδικασία με την οποία έγινε επιλογή των ενώσεων από την βάση δεδομένων και παρατίθεται η γλώσσα προγραμματισμού και τα υπολογιστικά παρκέτα που χρησιμοποιήθηκαν. Επιπλέον, αναλύεται η μεθοδολογία για τον υπολογισμό των μοριακών περιγραφέων και την δημιουργία των συνόλων δεδομένων. Τέλος, παρουσιάζεται ο τρόπος με τον οποίο χρησιμοποιήθηκαν τα σύνολα δεδομένων για να τροφοδοτήσουν τους αλγορίθμους Μηχανικής Μάθησης, ώστε να πραγματοποιηθεί η εκπαίδευση και η αξιολόγηση 10 ταξινομητών καθώς και οι παράμετροι που επιλέχθηκαν για τον κάθε ταξινομητή.

Το κεφάλαιο 3 παραθέτει τα αποτελέσματα που προέκυψαν από τις υπολογιστικές διαδικασίες που πραγματοποιήθηκαν. Πιο αναλυτικά, παρατίθενται κάποια διαγράμματα αναφορικά με την κατανομή των ενώσεων που προέκυψαν μετά από τον έλεγχο ομοιότητας τόσο στην βάση δεδομένων όσο και στις ενώσεις που επιλέχθηκαν για την δημιουργία των συνόλων δεδομένων. Επιπλέον, παρουσιάζονται συγκεντρωτικοί πίνακες με τα αποτελέσματα των επιδόσεων των ταξινομητών.

Το κεφάλαιο 4 σχολιάζει και αναλύει τα αποτελέσματα που προέκυψαν και προτείνει πιθανές μελλοντικές επεκτάσεις του θέματος. Συγκεκριμένα, σχολιάζονται τα αποτελέσματα που προέκυψαν από την αναζήτηση ομοιότητας στην βάση δεδομένων και εξηγείται ο λόγος της επιλογής συγκεκριμένων ενώσεων για την δημιουργία των συνόλων δεδομένων. Ακόμα, αναλύονται τα αποτελέσματα της ταξινόμησης των χημικών ενώσεων στις δύο κλάσεις και εξάγονται συμπεράσματα για την επίδοση των ταξινομητών. Εν κατακλείδι, προτείνονται ενδεχόμενες επεκτάσεις του θέματος για περαιτέρω γενίκευση της διαδικασίας.

Κεφάλαιο 1: Θεωρητικό υπόβαθρο

1.1 Φυσικά προϊόντα και Ιατρική

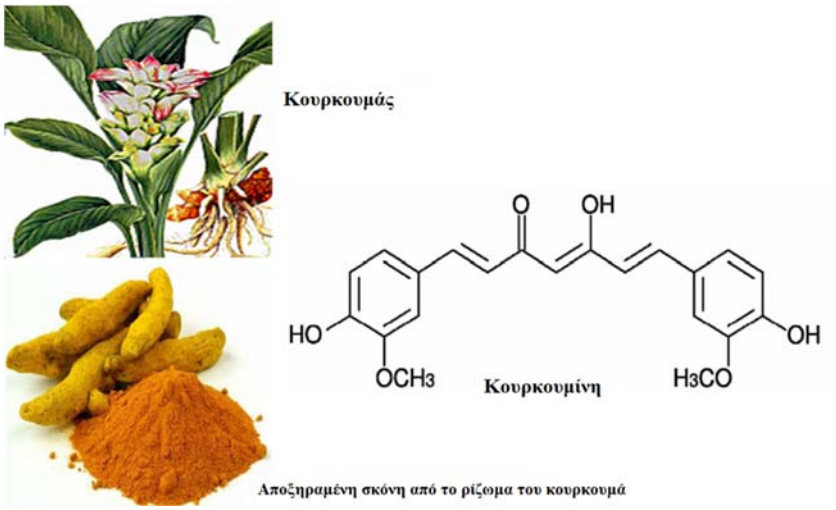
1.1.1 Φυσικά προϊόντα

Με τον όρο φυσικό προϊόν αναφερόμαστε σε μια χημική ένωση ή ένα σύνολο χημικών ουσιών το οποίο προέρχεται από ζωντανούς οργανισμούς (μικροοργανισμούς, φυτά, ζώα κλπ.) και απαντάται στην φύση. Μια ουσία που καλείται ως φυσικό προϊόν μπορεί να παρασκευαστεί και με χημική σύνθεση εργαστηριακά, αρκεί το τελικό αποτέλεσμα να είναι χημικά ισοδύναμο με την ένωση που προέρχεται από την φύση [1]. Από την αρχαιότητα αρκετοί πολιτισμοί βασίστηκαν στην φύση για τις βασικές τους ανάγκες, όπου μία εξ αυτών ήταν και η αντιμετώπιση διάφορων ασθενειών. Τα φυτά αποτέλεσαν την βάση για την δημιουργία θεραπευτικών τεχνικών που μπορούσαν να αντιμετωπίσουν ένα ευρύ φάσμα παθήσεων. Οι πρώτες αναφορές στην χρήση φυτικών εκχυλισμάτων ως θεραπευτικών παραγόντων για αντιμετώπιση λοιμώξεων, φλεγμονών και κρυολογημάτων ξεκινούν από τους αρχαίους πολιτισμούς της Μεσοποταμίας. Οι αρχαίοι πολιτισμοί της Αιγύπτου και της Κίνας έπαιξαν σημαντικό ρόλο στην καταγραφή και διατήρηση θεραπευτικών τεχνικών που στηρίζονται σε εκχυλίσματα φυτών. Ωστόσο, πιο σύγχρονοι πολιτισμοί όπως οι αρχαίοι Έλληνες και οι Ρωμαίοι συνεισέφεραν σημαντικά στην διάδοση της χρήσης φαρμάκων φυτικής προέλευσης στο αρχαίο δυτικό κόσμο. Έκτος όμως από τα φυσικά προϊόντα φυτικής προέλευσης, τα οποία έχουν σημαντικό ρόλο στην εξέλιξη των σημερινών φαρμάκων, οι θαλάσσιοι οργανισμοί και μικροοργανισμοί αποτελούν και αυτοί με την σειρά τους πηγές εύρεσης φυσικών προϊόντων. Δεδομένου ότι περισσότερο από το 70% της επιφάνειας του πλανήτη αποτελείται από ωκεανούς, τα θαλάσσια οικοσυστήματα αποτελούν και αυτά τεράστιους πόρους ανακάλυψης θεραπευτικών παραγόντων [2].

Από την ανακάλυψη της πενικιλίνης το 1928 και έπειτα, τα φυσικά προϊόντα έχουν παίξει καθοριστικό ρόλο στην σύγχρονη φαρμακοβιομηχανία, καθότι έχουν θέσει τα θεμέλια για την ανακάλυψη αρκετών αντιβιοτικών φαρμάκων και όχι μόνο [3][4]. Αρκετά από τα φάρμακα που χρησιμοποιούνται για την θεραπεία διαφόρων τύπων καρκίνου, καρδιαγγειακών παθήσεων, διαβήτη και άλλων ασθενειών, προέρχονται από φυσικά προϊόντα ή παράγωγά τους [5]. Ενδεικτικά αναφέρεται ότι το διάστημα από το 1981 έως το 2014 πάνω από το 50% των νέων φαρμάκων, αναπτύχθηκαν από φυσικά προϊόντα [6]. Για τον σκοπό της παρούσας εργασίας επιλέχθηκαν δύο φυσικά προϊόντα, η κουρκουμίνη και η ρεσβερατρόλη, τα οποία έχουν διαφορετική χημική δομή και διαφορετικές θεραπευτικές ιδιότητες.

1.1.2 Κουρκουμίνη

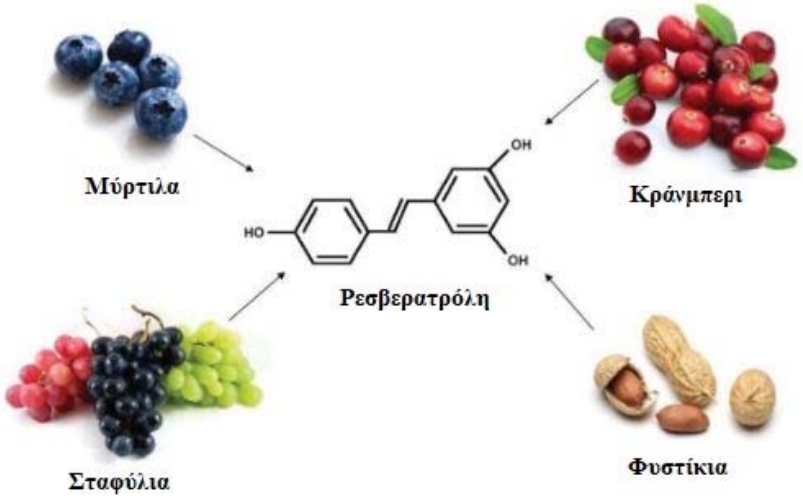
Η κουρκουμίνη προέρχεται από το ρίζωμα του κουρκουμά (*Curcuma longa*) και πρωτοανακαλύφθηκε το 1815 από δύο Γερμανούς επιστήμονες, τους Vogel και Pelletier. Ωστόσο οι πρώτες κλινικές μελέτες της κουρκουμίνης φαίνεται να καταγράφονται το 1937 [7]. Οι θεραπευτικές ιδιότητες του κουρκουμά ήταν γνωστές για χιλιάδες χρόνια, αλλά οι μηχανισμοί δράσης του και οι βιοδραστικές του ιδιότητες μελετήθηκαν τα τελευταία χρόνια. Ο κουρκουμάς χρησιμοποιούνταν σαν θεραπευτικό βότανο σε πολλές χώρες τις Ασίας εξαιτίας της αντιοξειδωτικής, της αντιφλεγμονώδους, της αντιμικροβιακής και της αντικαρκινικής δράσης του [8]. Τα τελευταία χρόνια η δυτική ιατρική έχει ασχοληθεί αρκετά με τις θεραπευτικές ιδιότητες που μπορεί να έχει η κουρκουμίνη στον ανθρώπινο οργανισμό και τα ευρήματα των μελετών δείχνουν ότι η δράση της είναι κυρίως αντιοξειδωτική και αντιφλεγμονώδης. Δεδομένου ότι οι φλεγμονές έχουν μείζονα ρόλο σε χρόνιες παθήσεις, όπως είναι η νόσος Alzheimer, η νόσος Parkinson ή η ρευματοειδής αρθρίτιδα, οι αντιφλεγμονώδεις παράγοντες της κουρκουμίνης δείχνουν να έχουν ανασταλτική δράση στην ανάπτυξη αυτών των ασθενειών [9]. Τα κύρια μειονεκτήματα της κουρκουμίνης είναι η χαμηλή βιοδιαθεσιμότητα και η χαμηλή διαλυτότητά της στο νερό. Μελέτες έχουν δείξει ότι παρόλο που η κουρκουμίνη έχει κάποιες αντικαρκινικές ιδιότητες, η χαμηλή της βιοδιαθεσιμότητα δεν την καθιστά ικανό μέσο αντιμετώπισης του καρκίνου [10]. Ωστόσο, κάποια από τα παράγωγα της κουρκουμίνης φαίνεται ξεπερνούν τους παραπάνω περιορισμούς και να έχουν αρκετές προοπτικές στην αποτελεσματική θεραπεία του καρκίνου του παχέως εντέρου, του μαστού και του προστάτη [11].



Εικόνα 1.1: *Το φυτό προέλευσης του κουρκουμά και η χημική δομή της κουρκουμίνης. (αναπροσαρμοσμένο από πηγή)* [12]

1.1.3 Ρεσβερατρόλη

Η ρεσβερατρόλη είναι ένα φυσικό προϊόν το οποίο εμπεριέχεται κυρίως στον φλοιό και στα κουκούτσια των σταφυλιών, στα όσπρια και στο φύλλωμα των πεύκων. Υπολογίζεται ότι αυτή η χημική ένωση βρίσκεται σε περισσότερα από 70 φυτικά προϊόντα, ενώ στην καθημερινότητα συναντάται σε ικανοποιητικές ποσότητες στους ξηρούς καρπούς, στο ρόδι, στα μούρα και στο κόκκινο κρασί. Η πρώτη αναφορά στην ρεσβερατρόλη έγινε το 1939 από τον Ιάπωνα ερευνητή M. Takaoka, ο οποίος κατάφερε να την απομονώσει από την ρίζα του φυτού λευκού Ελλέβορου (*Veratrum grandiflorum*) [13],[14]. Οι θεραπευτικές ιδιότητες της ρεσβερατρόλης σχετίζονται κυρίως με καρδιαγγειακές παθήσεις και με την πρόληψη του καρκίνου, ωστόσο φαίνεται να έχει επίσης αντιοξειδωτική, αντιφλεγμονώδη και αντιγηραντική δράση [15]. Η αντίληψη ότι η ρεσβερατρόλη έχει θετική επίδραση στο ανθρώπινο καρδιαγγειακό σύστημα προέκυψε από επιδημιολογικά δεδομένα καθώς και το αποκαλούμενο «Γαλλικό παράδοξο» (French paradox). Σύμφωνα με ευρήματα, ο γαλλικός πληθυσμός φαίνεται να παρουσιάζει χαμηλή πιθανότητα ανάπτυξης κάποιας καρδιαγγειακής νόσου, παρόλη την υψηλή πρόσληψη κορεσμένων λιπαρών. Το γεγονός αυτό αποδόθηκε στην σχετικά υψηλή κατανάλωση κρασιού [16]. Παρόλα αυτά έχουν γίνει αρκετές μελέτες που επιβεβαιώνουν ότι η ρεσβερατρόλη μπορεί να έχει προληπτικά αποτελέσματα στην ανάπτυξη παθήσεων του καρδιαγγειακού συστήματος [17],[18]. Κύρια μειονεκτήματα της ρεσβερατρόλης είναι η χαμηλή βιοδιαθεσιμότητα και η χαμηλή διαλυτότητά της στο νερό, γεγονός που μειώνει την απορρόφησή της από τον οργανισμό. Επίσης σε μεγάλες ποσότητες παρουσιάζει τοξικότητα, για αυτό και συστήνεται η λήψη πολλών μικρών δόσεων παρά μίας μεγάλης δόσης [19],[20]. Αρκετές μελέτες έχουν δείξει ότι η ρεσβερατρόλη έχει αντικαρκινική ιδιότητα σε αρκετούς τύπους καρκίνου, καθώς έχει αποδειχθεί ότι μπορεί να αναστείλει κάποια γεγονότα που συνδέονται με την ανάπτυξη όγκων. Ωστόσο, τα δεδομένα από κλινικές μελέτες είναι περιορισμένα και θα πρέπει να διεξαχθούν περαιτέρω κλινικές δοκιμές για να υπάρξει ξεκάθαρο συμπέρασμα [21], [22].



Εικόνα 1.2: *Χημική δομή της ρεσβερατρόλης και τροφές στις οποίες εμπεριέχεται. (αναπροσαρμοσμένο από πηγή)* [23]

1.2 Χημειοπληροφορική και βασικές έννοιες

1.2.1 Χημειοπληροφορική

Από τη αρχή της ιστορίας της, η Χημεία βασιζόταν σε δεδομένα που προέκυπταν από τις παρατηρήσεις πειραμάτων. Μετά από αρκετά χρόνια η θεωρητική Χημεία κατάφερε να φτάσει σε τέτοιο βαθμό που να μπορεί σε μερικές περιπτώσεις να εξάγει δεδομένα με υπολογιστικούς τρόπους. Ωστόσο, μια τέτοια διαδικασία απαιτούσε αρκετούς υπολογισμούς και ήταν αρκετά χρονοβόρα. Ακόμα, αρκετά χημικά φαινόμενα είναι αρκετά περίπλοκα για να μπορέσουν να ερμηνευτούν με παραδοσιακούς υπολογιστικούς τρόπους. Έτσι από την δεκαετία του 1960 άρχισαν να γίνονται προσπάθειες ώστε να χρησιμοποιηθεί η υπολογιστική ισχύς των ηλεκτρονικών υπολογιστών για την μοντελοποίηση και την αποσαφήνιση των χημικών φαινομένων [24]. Αρκετοί επιστήμονες άρχισαν να χρησιμοποιούν μεθόδους πληροφορικής για να επεξεργάζονται τον μεγάλο όγκο των δεδομένων που προέκυπταν, με αποτέλεσμα να δημιουργηθεί με την πάροδο του χρόνου το πεδίο της χημειοπληροφορικής. Ο όρος χημειοπληροφορική συγχωνεύει τρείς διαφορετικές επιστήμες, τη Χημεία, την Πληροφορική και τα Μαθηματικά, σε ένα κοινό επιστημονικό πεδίο. Ο επικρατέστερος ορισμός του όρου χημειοπληροφορική είναι ότι: *πρόκειται για το πεδίο το οποίο εφαρμόζει μεθόδους της Πληροφορικής για να επιλύσει χημικά προβλήματα* [25]. Τέτοια προβλήματα μπορεί να είναι η εξαγωγή δεδομένων από χημικές δομές, η αναζήτηση χημικών ενώσεων σε βάσεις δεδομένων, ακόμα και η ανακάλυψη νέων φαρμάκων [26].

1.2.2 Απεικόνιση μορίων

Στις μέρες μας, πολλές χημικές εφαρμογές παίρνουν ως δεδομένο ότι οι φυσικοχημικές και βιολογικές ιδιότητες μιας χημικής ουσίας απορρέουν από την χημική της δομή. Προκειμένου να γίνει καλύτερα αντιληπτή η έννοια της χημικής δομής με στόχο την εξαγωγή πληροφορίας από αυτή σε επόμενο στάδιο, δημιουργήθηκε μια διαδικασία απεικόνισης των μορίων. Ανάλογα με την χημική πληροφορία που θέλουμε να πάρουμε από μία ένωση, επιλέγουμε και την πολυπλοκότητα αυτής της απεικόνισης, η οποία μπορεί να έχει τις ακόλουθες μορφές [27]:

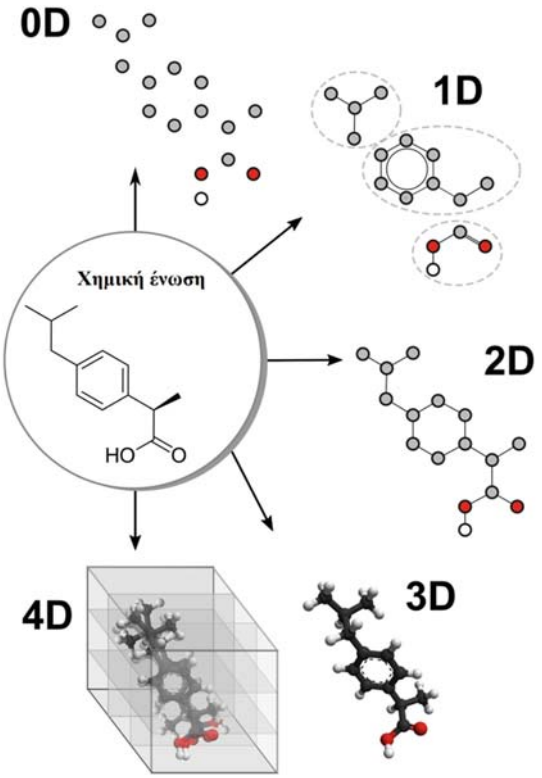
- **Μηδενικής διάστασης (0D):** Η απλούστερη μορφή μοριακής απεικόνισης είναι ο χημικός τύπος μιας ένωσης, όπου πρόκειται για μια λίστα με την αλληλουχία των διαφορετικών χημικών στοιχείων που αποτελούν την χημική ένωση. Για παράδειγμα, ο χημικός τύπος της ιβουπροφένης είναι $C_{13}H_{18}O_2$, ο οποίος μας δείχνει την παρουσία 13 ατόμων άνθρακα, 18 ατόμων υδρογόνου και 2 ατόμων οξυγόνου στο μόριο της ιβουπροφένης. Αυτός ο τρόπος απεικόνισης δεν παρέχει κάποια πληροφορία σχετικά με την μοριακή δομή ή την συνδεσιμότητα των ατόμων της ένωσης.

- **Μονοδιάστατη (1D):** Τα μόρια αναπαριστώνται ανά τμηματικές δομές (substructures) ενδιαφέροντος, όπως είναι τα μοριακά θραύσματα, οι χαρακτηριστικές ομάδες, ή δομές υποκατάστασης. Αυτή η απεικόνιση δεν προϋποθέτει την ακριβή γνώση της χημικής δομής.

- **Δισδιάστατη (2D):** Η συγκεκριμένη απεικόνιση λαμβάνει υπόψη την σύνδεση των ατόμων, δείχνει δηλαδή την παρουσία και τον τύπο των χημικών δεσμών μεταξύ των ατόμων της ένωσης. Συνήθως το μόριο παρουσιάζεται σαν ένα γράφημα του οποίου οι ακμές είναι οι δεσμοί και οι κορυφές είναι τα άτομα.

- **Τρισδιάστατη (3D):** Το μόριο αναπαριστάται σαν ένα γεωμετρικό αντικείμενο στον χώρο, όπου εκτός από την φύση και την σύνδεση των ατόμων του, περιγράφεται και η χωρική τους διαμόρφωση. Πιο συγκεκριμένα, το μόριο προσδιορίζεται αναλόγως των τύπων των ατόμων και τις καρτεσιανές τους συντεταγμένες x-y-z.

- **Τετραδιάστατη (4D):** Επιπρόσθετα με την γεωμετρία του μορίου, μία "τέταρτη διάσταση" αποσκοπεί στο να προσδιορίσει ποσοτικά τις αλληλεπιδράσεις ανάμεσα σε ένα μόριο και μία ενεργό περιοχή ενός υποδοχέα.



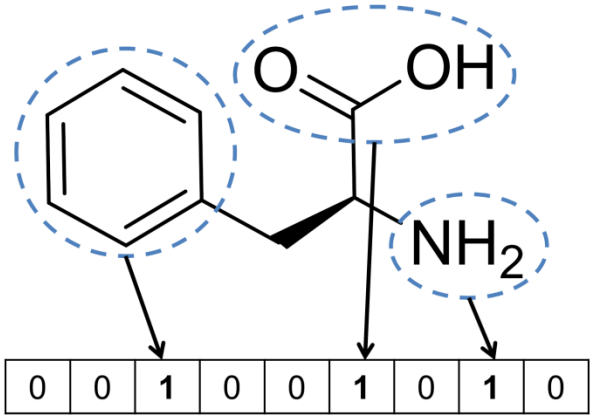
Εικόνα 1.3: *Οι διαφορετικές απεικονίσεις της ίδιας χημικής δομής (ιβουπροφένη). (αναπροσαρμοσμένο από πηγή)* [27]

1.2.3 Μοριακοί Περιγραφείς (Molecular Descriptors)

Σύμφωνα με τις παραπάνω απεικονίσεις μπορούμε να υπολογίσουμε διαφορετικούς μοριακούς περιγραφείς (Molecular Descriptors), ανάλογα με την πληροφορία που θέλουμε να εξάγουμε από το μόριο. Οι μοριακοί περιγραφείς είναι αριθμητικά χαρακτηριστικά τα οποία εξάγονται από την χημική δομή ενός μορίου και παρέχουν πληροφορίες για το συγκεκριμένο μόριο. Μπορούν να είναι μονοδιάστατοι (0D ή 1D), δισδιάστατοι (2D), τρισδιάστατοι (3D) ή τετραδιάστατοι (4D). Οι μονοδιάστατοι περιγραφείς αποτελούν μονόμετρα μεγέθη που παρέχουν μια συγκεντρωτική πληροφορία για το μόριο σύμφωνα με τον χημικό τύπο του. Τέτοιοι μοριακοί περιγραφείς είναι το μοριακό βάρος, ο αριθμός των ατόμων και ο αριθμός των δεσμών. Παρόλο που είναι εύκολο να υπολογιστούν, οι μονοδιάστατοι περιγραφείς αντιμετωπίζουν ορισμένα προβλήματα εκφυλισμού, κατά τα οποία διαφορετικά μόρια λαμβάνουν ίδιες τιμές για τον ίδιο περιγραφέα. Για τον λόγο αυτό, οι μονοδιάστατοι περιγραφείς συνήθως χρησιμοποιούνται σε συνδυασμό με περιγραφείς μεγαλύτερων διαστάσεων. Οι δισδιάστατοι μοριακοί περιγραφείς βασίζονται στην τοπολογία της δομής του μορίου, όπως είναι οι χαρακτηριστικές ομάδες ή τα μοριακά θραύσματα. Οι τρισδιάστατοι περιγραφείς εξάγουν πληροφορίες από την απεικόνιση των 3D συντεταγμένων του μορίου, ως εκ τούτου βασίζονται στην στη γεωμετρία του. Γνωστοί 3D περιγραφείς περιλαμβάνουν σταθερές ατόμων υποκατάστασης, περιγραφείς αυτοσυσχέτισης, περιγραφείς του λόγου επιφάνειας-όγκου και κβαντοχημικούς περιγραφείς. Τέλος, οι τετραδιάστατοι μοριακοί περιγραφείς αποτελούν μια επέκταση των τρισδιάστατων, όπου λαμβάνουν υπόψη πολλαπλές δομικές διαμορφώσεις ταυτόχρονα [26][28]. Ο υπολογισμός των μοριακών περιγραφέων προϋποθέτει την ύπαρξη μιας συγκεκριμένης κωδικοποίησης για το κάθε μόριο, προκειμένου ένα υπολογιστικό σύστημα να μπορεί να διακρίνει από ποια στοιχεία αποτελείται το εκάστοτε μόριο. Η κωδικοποίηση αυτή αντιστοιχεί στα μοριακά αποτυπώματα (molecular fingerprints), τα οποία θα περιγραφούν στο επόμενο κεφάλαιο.

1.2.4 Μοριακά Αποτυπώματα (Molecular Fingerprints)

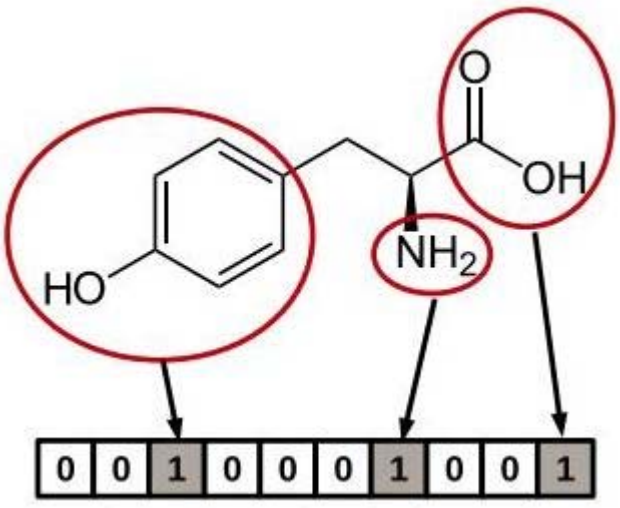
Όπως τα ανθρώπινα αποτυπώματα μπορούν να διαφοροποιήσουν έναν άνθρωπο από κάποιον άλλον, έτσι και τα μοριακά αποτυπώματα μιας χημικής δομής προσπαθούν να αναγνωρίσουν ένα μόριο σύμφωνα με κάποια ειδικά χαρακτηριστικά που διαθέτει. Οι χαρακτηριστικές αυτές ιδιότητες των μορίων μπορούν να περιγραφούν από την δομή ή αλλιώς τα δομικά κλειδιά, τα οποία υποδεικνύουν εάν μια συγκεκριμένη υποομάδα ή ένα μοριακό θραύσμα υπάρχει μέσα στο μόριο. Αυτές οι υποομάδες στην χημική δομή των μορίων μπορούν να κωδικοποιηθούν σε δυαδικά δομικά κλειδιά. Έτσι οι δομικές υποομάδες αναπαριστώνται ως αλληλουχίες ‘0’ και ‘1’ (bit strings), όπου το ‘0’ συμβολίζει την απουσία της συγκεκριμένης υποομάδας από την δομή του μορίου και αντίστοιχα το ‘1’ συμβολίζει την παρουσία της (Εικόνα 1.4). Αυτή η χαρακτηριστική αλληλουχία από ‘0’ και ‘1’ μιας χημικής δομής ονομάζεται μοριακό αποτύπωμα και τυπικά μπορεί να έχει μήκος από 150−2500 ψηφία (bits) [25].



Εικόνα 1.4: *Παράδειγμα δημιουργίας μοριακού αποτυπώματος. Τα ψηφία που έχουν την τιμή 1 δείχνουν ότι η συγκεκριμένη υποομάδα υπάρχει στην χημική ένωση.* [29]

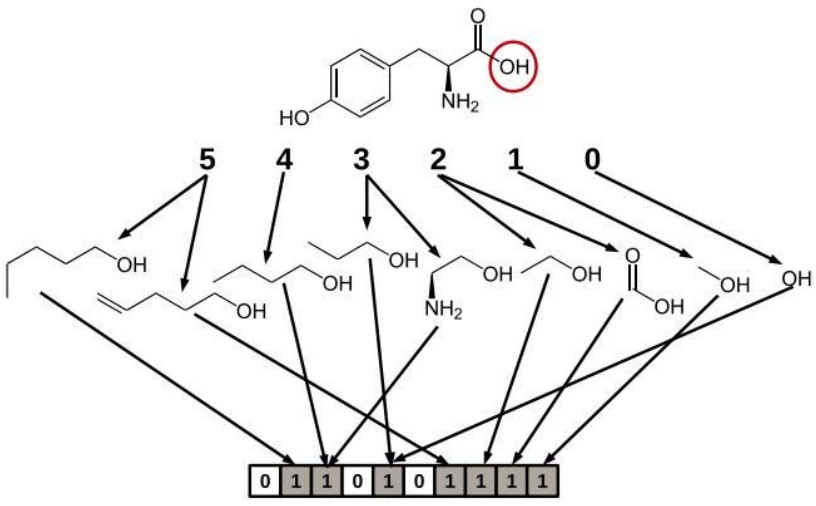
Υπάρχουν αρκετοί διαφορετικοί τύποι μοριακών αποτυπωμάτων οι οποίοι προκύπτουν ανάλογα με την μορφή της μοριακής απεικόνισης που θα εκφραστεί σε δυαδική αλληλουχία. Ο συχνότερα χρησιμοποιούμενος τύπος είναι αυτός που προκύπτει από την 2D μοριακή απεικόνιση και για το λόγο αυτό τα συγκεκριμένα αποτυπώματα ονομάζονται και 2D αποτυπώματα. Ωστόσο σε κάποιες περιπτώσεις υπάρχει και η δυνατότητα αποθήκευσης 3D πληροφορίας, με πιο σημαντικό παράδειγμα τα φαρμακοφόρα αποτυπώματα [30]. Τα 2D μοριακά αποτυπώματα μπορούν να χωριστούν σε τέσσερις μεγάλες κατηγορίες: τα αποτυπώματα που βασίζονται σε δομικά κλειδιά υποομάδων (substructure key-based), τα τοπολογικά αποτυπώματα (topological) ή αλλιώς τα αποτυπώματα που βασίζονται στην διαδρομή (path-based), τα κυκλικά αποτυπώματα (circular) και τέλος τα φαρμακοφόρα αποτυπώματα (pharmacophore) [31].

- Τα αποτυπώματα τα οποία είναι βασισμένα σε δομικά κλειδιά υποομάδων αφορούν δυαδικές συμβολοσειρές (bit strings), στις οποίες αντικατοπτρίζεται η παρουσία ορισμένων υποομάδων ή χαρακτηριστικών από μία δεδομένη λίστα δομικών κλειδιών στη χημική ένωση [30]. Μερικά παραδείγματα αυτών των αποτυπωμάτων είναι τα MACCS [32] αποτυπώματα, τα οποία έχουν δύο εκδοχές, μία με 960 δομικά κλειδιά και μία με 166 δομικά κλειδιά. Η δεύτερη εκδοχή είναι και η πιο συχνά χρησιμοποιούμενη, καθώς έχει μικρό μέγεθος (166 bits) και εμπεριέχει τα περισσότερα χημικά χαρακτηριστικά ενδιαφέροντος [30]. Τέλος υπάρχει το PubChem αποτύπωμα με 881 δομικά κλειδιά, το οποίο καλύπτει ένα ευρύ φάσμα υποομάδων και χαρακτηριστικών [33], και τα BCI αποτυπώματα με 1052 δομικά κλειδιά [34].



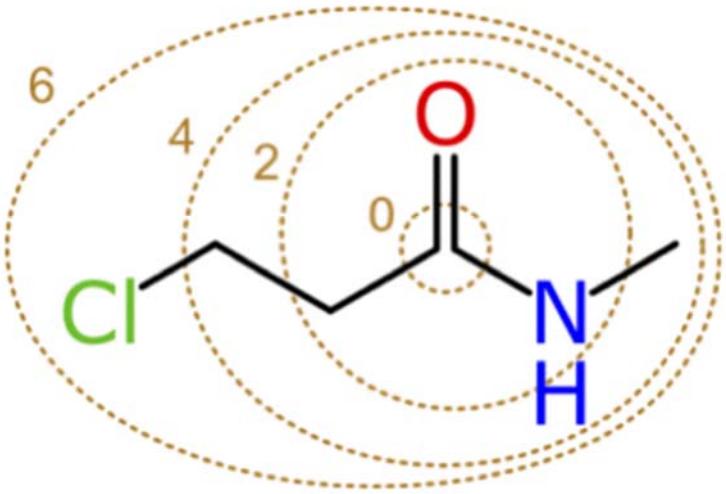
Εικόνα 1.5: *Παράδειγμα ενός υποθετικού αποτυπώματος 10-bit με βάση δομικά κλειδιά υποομάδων.* [30]

- Τα τοπολογικά αποτυπώματα αναλύουν όλες τις υποομάδες ενός μορίου ακολουθώντας συνήθως μια γραμμική διαδρομή μέχρι έναν συγκεκριμένο αριθμό χημικών δεσμών και την συνέχεια κωδικοποιούν τμηματικά καθεμία από αυτές τις διαδρομές για να δημιουργήσουν το αποτύπωμα [30]. Τα πιο διαδεδομένα αποτυπώματα σε αυτήν την κατηγορία είναι τα FP2 αποτυπώματα τα οποία έχουν μήκος 1024 bits [35], τα Daylight αποτυπώματα τα οποία αποτελούνται από 2048 bits και κωδικοποιούν όλες τις πιθανές διαδρομές συνδεσιμότητας στο μόριο [36] και τέλος τα αποτυπώματα ηλεκτροτοπολογικής κατάστασης, τα οποία βασίζονται στου δείκτες ηλεκτροτοπολογικής κατάστασης του μορίου [37].



Εικόνα 1.6: *Παράδειγμα ενός τοπολογικού αποτυπώματος βασισμένο σε γραμμικές διαδρομές.* [30]

- Τα κυκλικά αποτυπώματα είναι επίσης τοπολογικά αποτυπώματα τμηματοποίησης μόνο που αντί να αναλύουν διαδρομές στο μόριο, καταγράφουν το περιβάλλον γύρω από το κάθε άτομο του μορίου σε μια προκαθορισμένη ακτίνα [30]. Τα πιο χαρακτηριστικά κυκλικά αποτυπώματα είναι τα αποτυπώματα εκτεταμένης σύνδεσης (Extended Connectivity - ECFP) τα οποία αναπαριστούν κυκλικές περιοχές γύρω από τα άτομα και πρόκειται για αποτυπώματα με διάφορα μήκη. Συνήθως απαντώνται με την μορφή ECFP2 ή ECFP4, όπου το νούμερο στο τέλος υποδηλώνει την ακτίνα σε Angstrom (10^{-10} m) η οποία καθορίζει την κυκλική περιοχή γύρω από το κάθε άτομο. Ακόμα, μια παραλλαγή των ECFP αποτυπωμάτων είναι τα αποτυπώματα λειτουργικής κλάσης (Functional Class - FCFP), τα οποία αντί να καταγράφουν το περιβάλλον γύρω από ένα άτομο, όπως τα ECFP, καταγράφουν τον λειτουργικό ρόλο του ατόμου. Έτσι διαφορετικά άτομα με παρόμοιο λειτουργικό ρόλο δεν μπορούν να γίνουν διακριτά από αυτό το αποτύπωμα [30][38].



Εικόνα 1.7: *Παράδειγμα κυκλικού αποτυπώματος. Τα νούμερα καθορίζουν την ακτίνα σε Angstrom που ελέγχεται κάθε φορά.* [30]

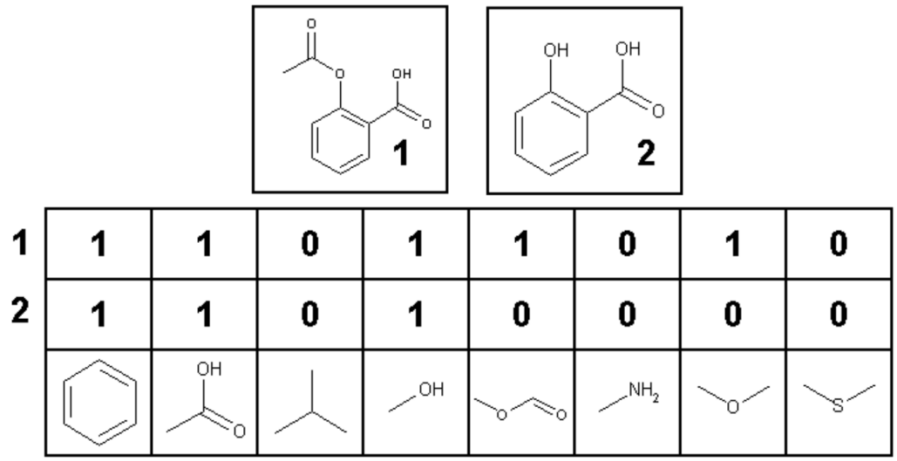
- Τα φαρμακοφόρα αποτυπώματα συνήθως εμπεριέχουν την πληροφορία από μία λίστα χαρακτηριστικών που παρουσιάζει το μόριο, κατά παρόμοιο τρόπο με τα αποτυπώματα που είναι βασισμένα σε δομικά κλειδιά υποομάδων. Η διαφορά είναι ότι λαμβάνεται υπόψη και η απόσταση που έχουν αυτά τα χαρακτηριστικά μεταξύ τους και έτσι με αυτόν τον τρόπο μεταφέρεται και 3D πληροφορία στο αποτύπωμα [39].

1.2.5 Μοριακή ομοιότητα και ομοιότητα Tanimoto

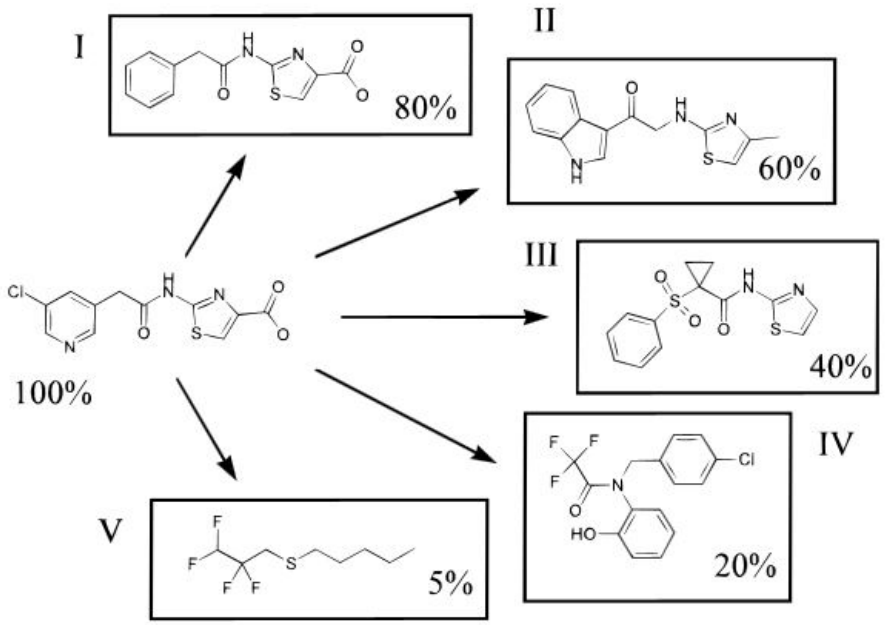
Η μοριακή ομοιότητα είναι ευρέως διαδεδομένη έννοια στην χημεία και παίζει καθοριστικό ρόλο στην φαρμακευτική χημεία [40]. Έχει ως στόχο την αναγνώριση και εύρεση ενώσεων που έχουν παρόμοιες χημικές δομές. Η βασική αρχή της μοριακής ομοιότητας είναι ότι χημικές ενώσεις με παρόμοια χημική δομή αναμένεται να παρουσιάζουν παρόμοιες φυσικές ιδιότητες ή παρόμοια βιολογική δράση [25][26]. Ωστόσο, ένα από τα μεγαλύτερα προβλήματα που προκύπτουν όταν πρόκειται να υπολογιστεί η ομοιότητα μεταξύ δύο μορίων είναι η πολυπλοκότητα που έχει αυτή η διαδικασία, η οποία εξαρτάται από την πολυπλοκότητα της μοριακή απεικόνισης που χρησιμοποιείται. Για να γίνει υπολογιστικά ευκολότερη η αναζήτηση ομοιότητας μεταξύ μορίων χρησιμοποιούνται συχνά τα μοριακά αποτυπώματα των ενώσεων και κυρίως τα 2D αποτυπώματα [30]. Το πιο διαδεδομένο μέγεθος που περιγράφει την ομοιότητα μεταξύ δύο μορίων είναι ο συντελεστής Tanimoto (Tc). Ο συντελεστής Tanimoto ποσοτικοποιεί την αλληλοεπικάλυψη των χαρακτηριστικών δύο μορίων, ως τον λόγο του πλήθους των κοινών χαρακτηριστικών προς το σύνολο των χαρακτηριστικών σε κάθε μοριακό αποτύπωμα και εκφράζεται με την ακόλουθη εξίσωση:

$$T_c = \frac{N_{ab}}{N_a + N_b - N_{ab}} \tag{1.1}$$

όπου το N_a είναι ο αριθμός των bit που είναι ‘1’ στο αποτύπωμα του ενός μορίου a, το N_b είναι ο αριθμός των bit που είναι ‘1’ στο αποτύπωμα του άλλου μορίου b και το N_{ab} είναι ο αριθμός των bit που είναι ‘1’ και στα δύο αποτυπώματα ταυτόχρονα. Η τιμή του συντελεστή Tanimoto κυμαίνεται από το 0 έως το 1 και μπορεί να ερμηνευτεί ως το ποσοστό των κοινών στοιχείων δύο ενώσεων [41][42]. Δύο ενώσεις θεωρούνται παρόμοιες όταν ο συντελεστής Tanimoto είναι μεγαλύτερος από 0.85 ή 85%, αλλά αυτό δεν συνεπάγεται ότι οι ενώσεις αυτές έχουν και παρόμοια βιολογική δράση [43].



Εικόνα 1.8: *Στην εικόνα παρουσιάζονται τα αποτυπώματα των ενώσεων 1 και 2, οπότε μπορεί να υπολογιστεί ο συντελεστής Tanimoto ως εξης:* N_a*=5,* N_b*=3 και* N_{ab}*=3, οπότε από την εξίσωση (1) προκύπτει ότι Tc=3/(5+3-3) = 0,6.* [44]



Εικόνα 1.9: *Μοριακή ομοιότητα πέντε ενώσεων (I-V) σε σχέση με μια συγκεκριμένη ένωση με βάση τον συντελεστή Tanimoto.* [45]

1.2.6 Κωδικοποίηση SMILES, SMARTS, InChi και InChi key

Εκτός από τα μοριακά αποτυπώματα που αναφέρθηκαν παραπάνω, υπάρχουν και κάποιοι πιο απλοί τρόποι αποτύπωσης της δομής ενός μορίου, με βασικότερο εξ αυτών την κωδικοποίηση SMILES (Simplified Molecular Input Line Entry System). Η κωδικοποίηση SMILES αφορά ένα σύστημα χημικής σημειογραφίας μικρού μεγέθους (σε σχέση με τα αποτυπώματα), το οποίο είναι εύκολο να διαβαστεί και να ερμηνευτεί. Επίσης περιγράφει με γραμμικό τρόπο την δομή ενός μορίου. Πρόκειται για μια συμβολοσειρά με αλφαριθμητικούς χαρακτήρες, η οποία περιγράφει την χημική δομή μιας ένωσης και διευκολύνει την αποτελεσματική αποθήκευση και γρήγορη επεξεργασία μεγάλου αριθμού ενώσεων. Για την κωδικοποίηση των μορίων σε SMILES χρησιμοποιούνται 5 βασικοί κανόνες [46][47].

1. Τα άτομα αναπαρίστανται από τα ατομικά τους γράμματα (πχ N για το άζωτο, O για το οξυγόνο). Τα ασθενή άτομα υδρογόνου δεν αναπαριστώνται αναλυτικά.

2. Τα γειτονικά άτομα τοποθετούνται το ένα δίπλα στο άλλο και οι χημικοί δεσμοί συμβολίζονται ως εξής: με '-' για τους απλούς δεσμούς, με '=' για τους διπλούς, με '#' για τους τριπλούς και με ':' για τους αρωματικούς. Οι απλοί και οι αρωματικοί δεσμοί συνήθως παραλείπονται.

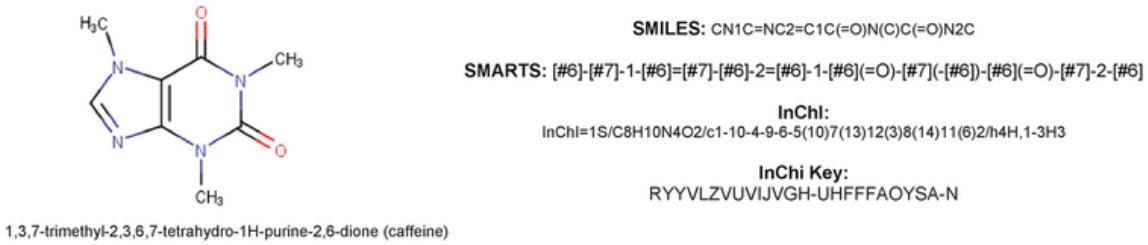
3. Ό,τι περικλείεται σε παρένθεση υποδηλώνει διακλάδωση στην χημική δομή του μορίου.

4. Για την γραμμική αναπαράσταση κυκλικών δομών διαχωρίζεται ένας δεσμός από κάθε δακτύλιο και τα 2 άτομα αυτού του δεσμού ακολουθούνται από το ίδιο αριθμητικό ψηφίο.

5. Τα άτομα στους αρωματικούς δακτυλίους αναπαριστώνται με πεζά γράμματα, γεγονός που προκαλεί προβλήματα μερικές φορές στην αντίληψη των αρωματικών δομών.

Μία παραλλαγή των SMILES είναι η κωδικοποίηση SMARTS (SMILES Arbitrary Target Specification). Πρόκειται για μια γλώσσα που αναπτύχθηκε για τον προσδιορισμό μοτίβων σε δομικές υποομάδες που χρησιμοποιούνται για να ταιριάζουν με μόρια και αντιδράσεις. Οι κανόνες που χρησιμοποιούνται για να επιτευχθεί ο προσδιορισμός αυτών των υποομάδων αποτελούν μια επέκταση των κανόνων που χρησιμοποιούνται για τα SMILES. Το πρόβλημα στην χρήση των SMILES εντοπίζεται στο ότι για την ίδια μοριακή απεικόνιση μπορούν να δημιουργηθούν διαφορετικά SMILES που να την περιγράφουν. Έτσι συχνά χρησιμοποιούνται κανονικοποιημένα SMILES προκειμένου να εξασφαλίσουν την μοναδικότητα ενός μορίου σε μία βάση δεδομένων. Όμως επειδή η κανονικοποίηση αυτή διαφέρει από λογισμικό σε λογισμικό δεν αποτρέπεται πλήρως η καταχώρηση διπλότυπων ενώσεων. Για τον λόγο αυτό προτείνεται η χρήση του InChi (International Chemical Identifier) και InChi key. Η κωδικοποίηση InChi δημιουργήθηκε υπό την αιγίδα της Διεθνής Ένωσης Καθαρής και Εφαρμοσμένης

Χημείας (IUPAC) και αποσκοπεί στην καθιέρωση ενός μοναδικού χαρακτηρισμού για κάθε ένωση, που θα επιτρέπει την καλύτερη οργάνωση των ενώσεων σε βάσεις δεδομένων. Η κωδικοποίηση αυτή επιλύει αρκετές από τις χημικές ασάφειες που τα SMILES αδυνατούν να αντιμετωπίσουν, όπως τα ταυτομερή και κάποια άλλα προβλήματα μοντέλων δραστικότητας. Ωστόσο στις περισσότερες περιπτώσεις τα InChi είναι δύσκολο να γίνουν απόλυτα κατανοητά από τον άνθρωπο. Από την άλλη μεριά ένα InChi key είναι μια συνεπτυγμένη ψηφιακή αναπαράσταση ενός InChi, η οποία διαθέτει ένα σταθερό προκαθορισμένο μήκος 27 χαρακτήρων. Τα InChi keys αναπτύχθηκαν ώστε να γίνονται πιο εύκολα οι διαδικτυακές αναζητήσεις για της χημικές δομές των ενώσεων [47].



Εικόνα 1.10: *Παράδειγμα των κωδικοποιήσεων SMILES, SMARTS, InChi και InChi key για την χημική δομή της καφεΐνης.* [47]

1.3 Βάσεις Δεδομένων και Μηχανική Μάθηση στην Χημειοπληροφορική

1.3.1 Βάσεις Δεδομένων

Ο όγκος της πληροφορίας που σχετίζεται με δεδομένα χημικών ενώσεων είναι τεράστιος και συνεχίζει να αυξάνεται με την πάροδο των χρόνων, γεγονός που καθιστά την επεξεργασία τους ακατόρθωτη με συμβατικές μεθόδους. Η αποθήκευση και η αναζήτηση δεδομένων αναφορικά με χημικές ενώσεις αποτέλεσε ένα από τα πρώτα, αν όχι το πρώτο, ζητήματα της χημειοπληροφορικής. Η διαχείριση όλης αυτής της πολυδιάστατης πληροφορίας σχετικά με τις χημικές ενώσεις μπορούσε να επιτευχθεί μόνο με ηλεκτρονικά μέσα και συγκεκριμένα τις βάσεις δεδομένων [25]. Μία βάση δεδομένων αποτελεί μια συλλογή από πληροφορίες ενός συγκεκριμένου θέματος (πχ. χημικές ενώσεις), οι οποίες συνήθως βρίσκονται σε πίνακες ή λίστες. Η ύπαρξη βάσεων δεδομένων βοηθάει στην καλύτερη οργάνωση της πληροφορίας, στην εύκολη πρόσβαση σε αυτήν και στην ανάκτηση της μέσα από απλές αναζητήσεις. Τα δεδομένα τοποθετούνται σε πίνακες οι οποίοι μπορούν να συνδέονται μεταξύ τους μέσα από κάποιο κοινό γνώρισμα, το οποίο και αναφέρεται ως πρωτεύον κλειδί. Μία βάση δεδομένων γενικά αποτελείται από πεδία (fields), εγγραφές (records), ερωτήματα (queries) και αναφορές (reports), τα οποία περιγράφονται επιγραμματικά παρακάτω [48]:

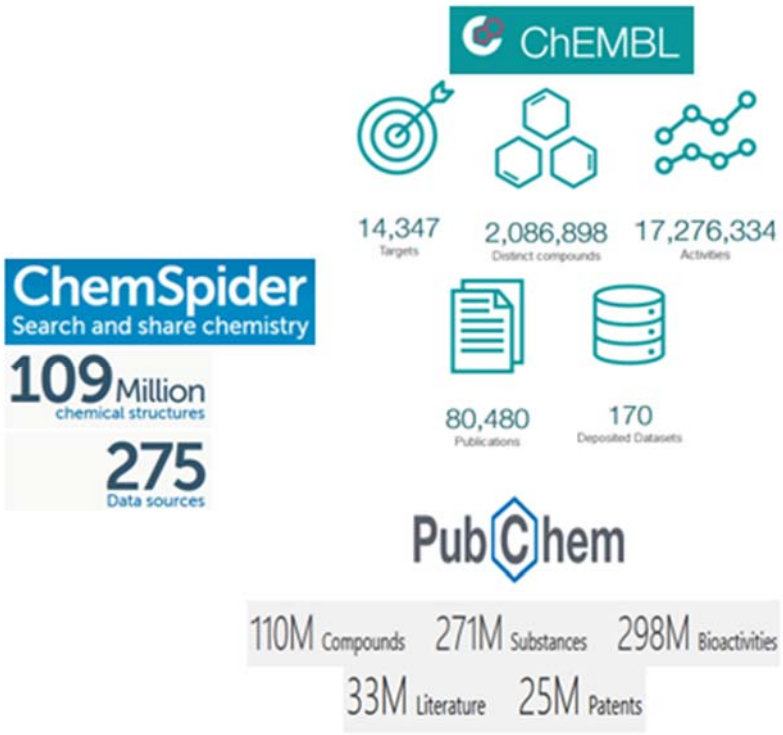
- **Πεδία:** Κατά την δημιουργία πινάκων σε μια βάση δεδομένων οι πληροφορίες τοποθετούνται κάτω από συγκεκριμένα πεδία, τα οποία θα πρέπει να έχουν μοναδικό όνομα για να είναι πιο εύκολη η ανάκτηση των εγγραφών που περιέχουν. Τέτοια πεδία, για παράδειγμα, μπορούν να είναι τα ονόματα των στηλών σε έναν πίνακα.

- **Εγγραφές:** Οι εγγραφές είναι συγκεκριμένα χαρακτηριστικά ενός αντικειμένου στην βάση δεδομένων. Ως αντικείμενο μπορεί να θεωρηθεί και ένας πίνακας, και οι γραμμές που θα έχει αυτός ο πίνακας αποτελούν και τις εγγραφές του.

- **Ερωτήματα:** Το ερώτημα (query) αφορά στην ουσία την αναζήτηση που κάνει ο χρήστης για να ανακτήσει την πληροφορία που επιθυμεί, η οποία είναι αποθηκευμένη στην βάση δεδομένων. Ένα παράδειγμα είναι όταν ένας χρήστης επιθυμεί να μάθει πληροφορίες για μια χημική ένωση, από μία σχετική βάση δεδομένων, χρησιμοποιώντας το όνομα της ένωσης για την αναζήτηση.

- **Αναφορές:** Οι αναφορές είναι τα αποτελέσματα που ανακτώνται μετά από ένα ερώτημα, μια αναζήτηση στην βάση δεδομένων. Οι αναφορές μπορούν να προσαρμοστούν ανάλογα με τις ανάγκες του κάθε χρήστη, έτσι ώστε η πληροφορία που θα λάβει να είναι περισσότερο χρήσιμη και λειτουργική.

Όπως προαναφέρθηκε, οι πίνακες και τα δεδομένα μέσα σε μία βάση μπορούν να συνδέονται μεταξύ τους και αυτό επιτυγχάνεται με την χρήση των πρωτευόντων και των ξένων κλειδιών. Ένα πρωτεύον κλειδί (primary key) είναι ένα χαρακτηριστικό, μία τιμή συνήθως, που χρησιμοποιείται για την εύρεση και αναγνώριση μιας συγκεκριμένης γραμμής σε έναν πίνακα. Κάθε γραμμή θα πρέπει να έχει μια ξεχωριστή μοναδική τιμή που θα λειτουργεί σαν πρωτεύον κλειδί για τον εντοπισμό της. Για να δημιουργηθεί μια σχέση μεταξύ δύο ή και περισσότερων πινάκων χρησιμοποιείται το ξένο κλειδί (foreign key). Το ξένο κλειδί είναι ένα πεδίο μέσα σε έναν πίνακα, το οποίο είναι πρωτεύον κλειδί σε έναν διαφορετικό πίνακα. Οι περισσότερες από τις σημερινές βάσεις δεδομένων αποτελούν σχεσιακές βάσεις δεδομένων, αποθηκεύουν δηλαδή τις πληροφορίες σε πίνακες που συνδέονται με κοινά χαρακτηριστικά (κλειδιά). Η αναζήτηση σε αυτές γίνεται χρησιμοποιώντας απλά το όνομα ενός πίνακα, το όνομα μιας ιδιότητας ή την τιμή από το πρωτεύον κλειδί ενός πίνακα [48]. Κάποιες από τις πιο διαδεδομένες βάσεις δεδομένων χημικών ενώσεων είναι η PubChem με 110 εκατομμύρια ενώσεις [49], η ChEMBL με 2.1 εκατομμύρια ενώσεις [50] και η ChemSpider με 106 εκατομμύρια χημικές ενώσεις [51].



Εικόνα 1.11: *Γνωστές βάσεις δεδομένων που χρησιμοποιούνται στην Χημειοπληροφορική και οι πληροφορίες που παρέχει η καθεμία.* [49][50][51]

1.3.2 Μηχανική Μάθηση (Machine Learning)

Η Μηχανική Μάθηση αποτελεί ένα επιστημονικό πεδίο το οποίο ασχολείται με τον τρόπο που ένας υπολογιστής μπορεί να μάθει και να εξελιχθεί μέσα από πειραματικά δεδομένα. Πρόκειται για μια συνένωση της επιστήμης των υπολογιστών και των μαθηματικών, καθώς με την χρήση στατιστικών μεθόδων διερευνάται η σχέση ανάμεσα στα δεδομένα που τροφοδοτούνται στο σύστημα. Στην συνέχεια μέσα από κατάλληλη επεξεργασία των δεδομένων και μέσα από αλγορίθμους κατασκευάζονται αντίστοιχα μοντέλα πρόβλεψης [52]. Κάθε σύνολο δεδομένων (dataset) περιέχει κάποια χαρακτηριστικά (features), τα οποία παρέχουν την πληροφορία. Η ποιότητα των χαρακτηριστικών που θα τροφοδοτήσουν το υπολογιστικό σύστημα παίζει καθοριστικό ρόλο για την ακρίβεια της πρόβλεψης. Διαφορετικοί συνδυασμοί χαρακτηριστικών παράγουν διαφορετικά αποτελέσματα ακρίβειας, οπότε η διαδικασία επαναλαμβάνεται αρκετές φορές μέχρι να βρεθεί το επιθυμητό αποτέλεσμα. Ένας αλγόριθμος Μηχανικής Μάθησης στην πράξη αποτελείται από 3 στάδια, την εκπαίδευση, την δοκιμή (τεστ) και την επαλήθευση. Το κάθε στάδιο είναι σημαντικό για την τελική ακρίβεια του συστήματος [53]. Οι αλγόριθμοι της Μηχανικής Μάθησης χωρίζονται σε 4 κύριες κατηγορίες: την επιβλεπόμενη μάθηση (Supervised learning), την μη επιβλεπόμενη μάθηση (Unsupervised learning), την

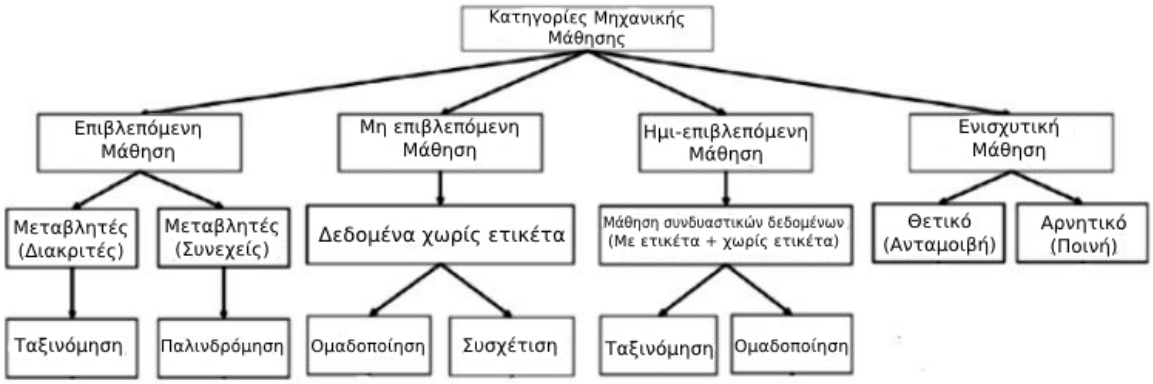
μάθηση με ημι-επίβλεψη (Semi-supervised learning) και την ενισχυτική μάθηση (Reinforcement learning) [54].

- **Επιβλεπόμενη μάθηση:** Στην επιβλεπόμενη μάθηση το σύστημα προσπαθεί να δημιουργήσει μια συνάρτηση η οποία θα μπορέσει να αντιστοιχίσει μία γνωστή είσοδο σε μία γνωστή έξοδο. Για να επιτευχθεί αυτό, παρέχεται στο σύστημα ένα γνωστό σύνολο δεδομένων εκπαίδευσης (training set), στο οποίο τα δεδομένα είναι χαρακτηρισμένα με ετικέτες (labels). Με την χρήση αλγορίθμων το σύστημα προσπαθεί να κατατάξει τα δεδομένα εισόδου κάθε φορά στην αντίστοιχη επιθυμητή έξοδο. Η επιβλεπόμενη μάθηση χρησιμοποιείται κυρίως για προβλήματα ταξινόμησης (classification) και παλινδρόμησης (regression) [54].

- **Μη επιβλεπόμενη μάθηση:** Στην μη επιβλεπόμενη μάθηση δεν υπάρχει η δυνατότητα δημιουργίας ενός συνόλου δεδομένων εκπαίδευσης, οπότε το σύστημα πρέπει να ομαδοποιήσει τα δεδομένα βάσει κάποιων κοινών στοιχείων. Οι αλγόριθμοι του συστήματος προσπαθούν να εντοπίσουν ομοιότητες και διαφορές στα δεδομένα με στόχο την δημιουργία ομάδων (clusters) ανάλογα με τα κοινά στοιχεία που θα αναγνωρίσουν. Επομένως το σύστημα παίρνει αποφάσεις χωρίς να έχει εκπαιδευτεί από κάποιο σύνολο δεδομένων. Οι αλγόριθμοι που χρησιμοποιούνται συνήθως στην μη επιβλεπόμενη μάθηση είναι κατά βάση αλγόριθμοι ομαδοποίησης (clustering algorithms) [53].

- **Ημι-επιβλεπόμενη μάθηση:** Η ημι-επιβλεπόμενη μάθηση μπορεί να θεωρηθεί ως ένας υβριδισμός μεταξύ της επιβλεπόμενης και της μη επιβλεπόμενης μάθησης, καθώς χρησιμοποιεί δεδομένα που έχουν χαρακτηριστεί με ετικέτες αλλά και δεδομένα που δεν έχουν χαρακτηριστεί με ετικέτες. Ο κύριος στόχος αυτής της μεθόδου είναι να παράξει όσο το δυνατόν καλύτερα αποτελέσματα πρόβλεψης, από την περίπτωση του να χρησιμοποιούνταν μόνο δεδομένα που είχαν χαρακτηριστεί με ετικέτες [54].

- **Ενισχυτική μάθηση:** Η ενισχυτική μάθηση χρησιμοποιεί μηχανισμούς ανάδρασης ώστε το σύστημα να αξιολογεί αυτόματα την επίδοσή του μέσα σε ένα συγκεκριμένο περιβάλλον, με απώτερο σκοπό να βελτιώσει την αποτελεσματικότητά του. Αυτός ο τρόπος μάθησης αξιοποιεί την μέθοδο της ανταμοιβής και ποινής για να διατηρήσει την γνώση που λαμβάνει το σύστημα από το περιβάλλον του, αυξάνοντας κάθε φορά την ανταμοιβή [54].



Εικόνα 1.12: *Διάφορες μέθοδοι Μηχανικής Μάθησης. (αναπροσαρμοσμένο από πηγή)* [54]

1.3.3 Αλγόριθμοι Ταξινόμησης επιβλεπόμενης Μάθησης και μοντέλα QSAR

Με την εξέλιξη της τεχνολογίας τα τελευταία χρόνια η Μηχανική Μάθηση έχει καθοριστικό ρόλο στον τομέα της Χημειοπληροφορικής και στην ανακάλυψη φαρμάκων. Μέσα από αλγορίθμους αναγνώρισης προτύπων διακρίνονται μαθηματικές σχέσεις ανάμεσα σε πειραματικά δεδομένα μορίων και με την βοήθεια των αλγορίθμων Μηχανικής Μάθησης, χρησιμοποιούνται για την πρόβλεψη φυσικοχημικών και βιολογικών ιδιοτήτων νέων ενώσεων. Σε σχέση με προηγούμενες μεθόδους, οι τεχνικές της Μηχανικής Μάθησης είναι περισσότερο αποτελεσματικές και προσαρμόζονται ευκολότερα σε μεγάλο όγκο δεδομένων και πληροφορίας. Ο κύριος στόχος των τεχνικών αυτών είναι η περαιτέρω κατανόηση και αξιοποίηση των σχέσεων μεταξύ της χημικής δομής και της βιολογικής δράσης διαφόρων ενώσεων. Τα μοντέλα που ερευνούν την ποσοτική σχέση δομής-δραστικότητας (QSAR), καθώς και τα μοντέλα που ερευνούν την ποσοτική σχέση δομής-ιδιότητας (QSPR) χρησιμοποιούν αλγορίθμους επιβλεπόμενης και μη επιβλεπόμενης Μηχανικής Μάθησης. Με την βοήθεια των μοντέλων QSAR μπορούν να αναπτυχθούν προγράμματα τεχνητής νοημοσύνης τα οποία να προβλέπουν *in silico* (μέσω υπολογιστή) με ακρίβεια το πώς κάποιες χημικές τροποποιήσεις επηρεάζουν την βιολογική συμπεριφορά των φαρμάκων. Μέσα από αυτές τις τεχνικές μπορούν να μοντελοποιηθούν αρκετές φυσικοχημικές ιδιότητες των φαρμάκων, όπως είναι η τοξικότητα, ο μεταβολισμός, η αλληλεπίδραση με άλλα φάρμακα και η καρκινογένεση. Τα πρώτα QSAR μοντέλα, όπως τα μοντέλα ανάλυσης από τους Hansch και Free-Wilson, χρησιμοποίησαν απλές τεχνικές παλινδρόμησης για να συσχετίσουν την δραστικότητα ενώσεων με διάφορα μοτίβα υποομάδων και κάποιες χημικές ιδιότητες, όπως είναι η διαλυτότητα, η υδροφοβικότητα και κάποιοι ηλεκτρονιακοί παράγοντες [26]. Ωστόσο, αρκετές μελέτες έχουν γίνει για την δημιουργία QSAR μοντέλων τα οποία να μπορούν να διαχωρίζουν αποτελεσματικά ενώσεις φυσικών προϊόντων από συνθετικές ενώσεις. Η μελέτη των Henkel et al. είναι ίσως από τις πρώτες μελέτες που εξέτασαν τις διαφορές των μοριακών ιδιοτήτων και των δομικών χαρακτηριστικών ανάμεσα σε φυσικά προϊόντα και συνθετικές ενώσεις [55]. Σε μια άλλη μελέτη των Stahura et al. εντοπίστηκαν κάποιοι

μοριακοί περιγραφείς οι οποίοι μπορούν να διαχωρίσουν ενώσεις φυσικών προϊόντων από συνθετικά μόρια, στηριζόμενοι στο μέγεθος της εντροπίας Shannon [56]. Στην μελέτη των Ertl et al. αναπτύχθηκε ένας δείκτης βαθμολόγησης για το κατά πόσο μια χημική ένωση μπορεί να κατηγοριοποιηθεί ως φυσικό προϊόν, αξιοποιώντας δομική πληροφορία και πληροφορία από μοριακούς περιγραφείς [57]. Στις παραπάνω μελέτες αξιοποιήθηκαν κατάλληλα αλγόριθμοι Μηχανικής Μάθησης. Στην παρούσα εργασία γίνεται μια μελέτη ως προς την αξιοποίηση αλγορίθμων ταξινόμησης επιβλεπόμενης μάθησης για των διαχωρισμό ενώσεων φυσικών προϊόντων σε δύο κατηγορίες.

Οι αλγόριθμοι ταξινόμησης της επιβλεπόμενης μάθησης, ή αλλιώς ταξινομητές, έχουν ως στόχο να διαχωρίσουν δύο ή περισσότερες κλάσεις σύμφωνα με μια εξίσωση διάκρισης. Υπάρχουν δύο κατηγορίες ταξινομητών, οι παραμετρικοί ταξινομητές και οι μη παραμετρικοί ταξινομητές. Οι παραμετρικοί ταξινομητές αξιοποιούν παραμέτρους από μια στατιστική κατανομή, την οποία θεωρείται ότι ακολουθούν και οι τιμές των χαρακτηριστικών των προτύπων. Από την άλλη οι μη παραμετρικοί ταξινομητές δεν αξιοποιούν παραμέτρους από κάποια στατιστική κατανομή [58]. Μερικοί από τους κυριότερους αλγορίθμους ταξινόμησης επιβλεπόμενης μάθησης περιγράφονται παρακάτω.

- **Naive Bayes:** Ο απλοϊκός Bayes ταξινομητής στηρίζεται στο θεώρημα του Bayes, το οποίο περιγράφει την πιθανότητα ενός γεγονότος βασισμένη στην προϋπάρχουσα γνώση για τις συνθήκες που σχετίζονται με αυτό το γεγονός. Ο ταξινομητής αυτός θεωρεί ότι κάθε χαρακτηριστικό που πρόκειται να ταξινομηθεί σε μία κλάση δεν σχετίζεται άμεσα με κανένα άλλο χαρακτηριστικό αυτής της κλάσης, παρόλο που τα χαρακτηριστικά που υπάρχουν σε αυτήν την κλάση θα μπορούσαν να έχουν κάποια αλληλεξάρτηση μεταξύ τους. Για κάθε νέα είσοδο σε αυτήν την μέθοδο, υπολογίζεται η πιθανοτική τιμή των κλάσεων σε σχέση με την συγκεκριμένη είσοδο και το χαρακτηριστικό θα ταξινομηθεί στην κλάση με την μεγαλύτερη πιθανολογική τιμή [53][59]. Η εξίσωση διάκρισης του Bayesian ταξινομητή, δηλαδή η εξίσωση που διαχωρίζει τις κλάσεις μεταξύ τους, περιγράφεται από την παρακάτω σχέση για την κλάση i:

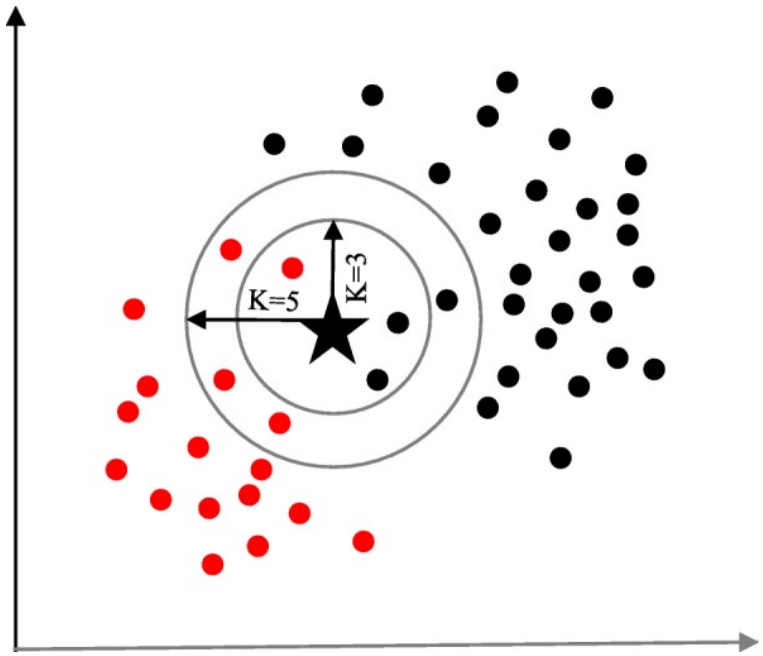
$$d_i = \ln(P_i) - \frac{1}{2}\ln|C_i| - \frac{1}{2}\left((\mathbf{x} - m_i)^T C_i^{-1}(\mathbf{x} - m_i)\right) \quad (1.2)$$

Όπου P_i είναι η πιθανότητα της κλάσης i, η οποία ορίζεται ως $P_i = \frac{N_i}{\sum_{i=1}^{c} N_i}$, στην οποία c είναι ο αριθμός των κλάσεων και N_i είναι ο αριθμός των δειγμάτων στην κλάση i. Το C_i είναι η μήτρα συνδιακύμανσης (covariance matrix) της κλάσης i και το $|C_i|$ είναι η ορίζουσα της μήτρας C_i. Η ορίζουσα υπολογίζεται ως εξής:

Αν $C_i = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ τότε $C_i = ad - bc$

Η ποσότητα $\left((\mathbf{x} - m_i)^T C_i^{-1}(\mathbf{x} - m_i)\right)$ είναι η mahalanobis απόσταση του x από το διάνυσμα των μέσων τιμών της κλάσης i. Το m_i είναι η μέση τιμή της κλάσης i και το x είναι το άγνωστο πρότυπο που πρόκειται να ταξινομηθεί [60].

- **K-Nearest Neighbors (KNN):** Ο ταξινομητής k-πλησιέστερων γειτόνων είναι ένας από τους πιο γνωστούς και απλούς αλγορίθμους ταξινόμησης. Προκειμένου να γίνει η ταξινόμηση ενός προτύπου σε μια κλάση, υπολογίζεται η Ευκλείδεια απόσταση ανάμεσα στο προς ταξινόμηση πρότυπο με το κάθε χαρακτηριστικό κάθε κλάσης. Το πρότυπο θα ταξινομηθεί στην κλάση στην οποία ανήκει το χαρακτηριστικό με την μικρότερη απόσταση από το πρότυπο. Το k δηλώνει τον αριθμό των πλησιέστερων γειτόνων που λαμβάνονται υπόψη για την ταξινόμηση. Η παραπάνω περίπτωση ισχύει για k=1, δηλαδή λαμβάνεται υπόψη μόνο ένας πλησιέστερος γείτονας ως προς το άγνωστο πρότυπο. Για τιμές του k μεγαλύτερες από 1, το πρότυπο ταξινομείται στην κλάση που ανήκουν τα περισσότερα χαρακτηριστικά που είναι πιο κοντά στο πρότυπο, ανάλογα με την τιμή του k. Ο ταξινομητής k-πλησιέστερων γειτόνων είναι ένας αλγόριθμος που καταναλώνει αρκετή υπολογιστική μνήμη και η ακρίβεια του εξαρτάται κάθε φορά από την τιμή που έχει το k. Δεν υπάρχει κάποιος κανόνας για το ποια θα πρέπει να είναι η τιμή του k, αλλά πολύ μικρές ή πολύ μεγάλες τιμές συχνά οδηγούν σε μη επιθυμητά αποτελέσματα [26][53][59].



Εικόνα 1.13: *Σχηματική αναπαράσταση του ταξινομητή k-πλησιέστερων γειτόνων. Όταν το k=3 το άγνωστο πρότυπο (αστέρι) ταξινομείται στην κλάση με τα μαύρα χαρακτηριστικά, ενώ όταν το k=5, το άγνωστο πρότυπο ταξινομείται στην κλάση με τα κόκκινα χαρακτηριστικά.* [59]

- **Logistic Regression:** Ο ταξινομητής λογιστικής παλινδρόμησης αφορά μια μέθοδο ταξινόμησης που έχει ως βάση την στατιστική και τις πιθανότητες. Χρησιμοποιεί την λογιστική ή αλλιώς σιγμοειδή συνάρτηση που φαίνεται στην σχέση (1.3) ,για να υπολογίσει τις πιθανότητες και να κατασκευάσει ένα μοντέλο ταξινόμησης, το οποίο ανταποκρίνεται αρκετά καλά σε σύνολα δεδομένων που διαχωρίζονται γραμμικά [54]. Με τον ταξινομητή λογιστικής παλινδρόμησης διερευνάται η πιθανότητα για το αν ένα νέο πρότυπο ανήκει σε κάποια συγκεκριμένη κλάση και εφόσον πρόκειται για πιθανότητες η έξοδος του υπολογισμού θα βρίσκεται μεταξύ του 0 και του 1. Οπότε θα πρέπει να υπάρχει ένα όριο, ένα κατώφλι, το οποίο θα διαχωρίζει τις κλάσεις μεταξύ τους. Για παράδειγμα, αν η τιμή της πιθανότητας για ένα πρότυπο εισόδου είναι μεγαλύτερη από 0,5 τότε αυτό θα κατηγοριοποιείται στην κλάση Α, διαφορετικά θα ανήκει στην κλάση Β [59]. Ο αλγόριθμος αυτός είναι μια μέθοδος η οποία χρειάζεται μια υπόθεση και μία συνάρτηση κόστους (cost function) και για την αποτελεσματική λειτουργία του κρίνεται σημαντική η βελτιστοποίηση της συνάρτησης κόστους [53]. Ακολούθως αναγράφεται η εξίσωση της σιγμοειδούς συνάρτησης (1.3) και η συνάρτηση κόστους (1.4).

$$d_j(x) = \frac{1}{\left(1+e^{-y}\right)} \tag{1.3}$$

Με $y = b + \sum_{i=1}^{N} x_i w_i$, όπου b είναι μία σταθερά, το x_i είναι το άγνωστο πρότυπο και w_i είναι βάρη που καθορίζονται από την μεγιστοποίηση της ακόλουθης συνάρτησης κόστους.

$$M(b,w) = \sum_{j=1}^{P} \left\{ L_j \log(s(y_j)) + (1 - L_j) \log(1 - s(y_j)) \right\} \quad (1.4)$$

Όπου L_j είναι ένδειξη της κλάσης του j προτύπου των P δεδομένων εκπαίδευσης του ταξινομητή [60].

- **Linear Discriminant Analysis (LDA):** Η ανάλυση γραμμικής διάκρισης πρόκειται για μία γραμμική μέθοδο ταξινόμησης η οποία αποτελεί μια γενίκευση της γραμμικής διάκρισης κατά Fisher, κατά την οποία το υπάρχων σύνολο δεδομένων προβάλλεται σε έναν χώρο λιγότερων διαστάσεων. Το μοντέλο ενός LDA ταξινομητή συνήθως προσαρμόζει τα δεδομένα κάθε κλάσης μέσα από ένα διάγραμμα Γκαουσιανής πυκνότητα υποθέτοντας ότι όλες οι κλάσεις έχουν την ίδια μήτρα συνδιακύμανσης (covariance matrix). Η ανάλυση γραμμικής διάκριση σχετίζεται με τις μεθόδους ANOVA (ανάλυση διακύμανσης) και ανάλυση παλινδρόμησης (regression analysis), οι οποίες προσπαθούν να εκφράσουν μια εξαρτημένη μεταβλητή ως έναν γραμμικό συνδυασμό άλλων χαρακτηριστικών ή μετρήσεων [54]. Η συνάρτηση διάκρισης του ταξινομητή LDA περιγράφεται από την σχέση (1.5).

$$d_i = x^T C^{-1} m_i - \frac{1}{2} m_i^T C^{-1} m_i + \ln\left(\frac{N_i}{N_{all}}\right) \quad (1.5)$$

Όπου m_i είναι το διάνυσμα της μέσης τιμής της κλάσης i, το x είναι το διάνυσμα του άγνωστου προτύπου, το N_i είναι το πλήθος των προτύπων της κλάσης i, το N_{all} είναι το πλήθος των προτύπων όλων των κλάσεων και C είναι η μήτρα συνδιακύμανσης, η οποία υπολογίζεται από την ακόλουθη σχέση:

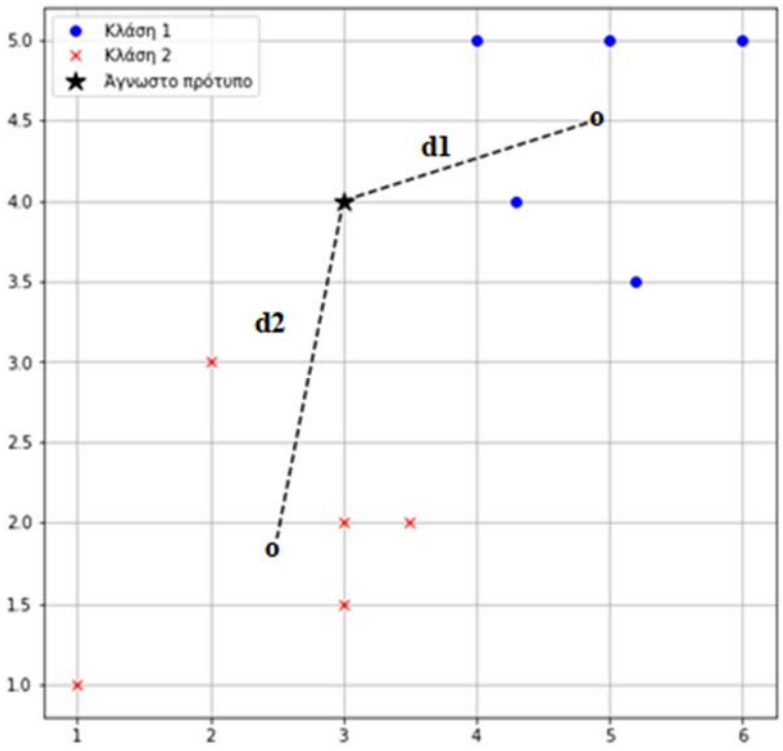
$$C_i = \left(\frac{1}{N_i} \sum \mathbf{z}\mathbf{z}^T \right) - m_i m_i^T \quad (1.6)$$

Όπου το z είναι το διάνυσμα του πρότυπου της κλάση. Το άγνωστο πρότυπο ταξινομείται τελικά στην κλάση με την μεγαλύτερη τιμή d_i [60].

- **Minimum Distance Classifier (MDC):** Ο ταξινομητής ελάχιστης απόστασης είναι ένας από τους πιο απλούς ταξινομητές που χρησιμοποιείται σε ένα ευρύ φάσμα εφαρμογών. Για κάθε κλάση υπολογίζεται το μέσο σημείο της στον χώρο σύμφωνα με τις θέσεις των χαρακτηριστικών που την απαρτίζουν. Στην συνέχεια υπολογίζεται η Ευκλείδεια απόσταση του προς ταξινόμηση προτύπου από το κέντρο της κάθε κλάσης και τελικά το πρότυπο ταξινομείται στην κλάση με την μικρότερη απόσταση [61][62]. Έστω ότι έχουμε ένα άγνωστο πρότυπο $X(x_1,x_2)$, τότε σε αυτόν τον ταξινομητή το πρότυπο ταξινομείται στην κλάση για την οποία η συνάρτηση διάκρισης της σχέσης (1.7) λαμβάνει την μεγαλύτερη τιμή.

$$g_i(X) = (\mu_1^i x_1 + \mu_2^i x_2 - \tfrac{1}{2}\left(\left[\mu_1^i\right]^2 + \left[\mu_2^i\right]^2\right) \quad \text{για } i=1,2,3,\ldots,N \quad (1.7)$$

Όπου το x είναι το άγνωστο πρότυπο και το μ^i η μέση τιμή της κλάσης i [60].



Εικόνα 1.14: *Σχηματική αναπαράσταση του ταξινομητή ελάχιστης απόστασης. Το άγνωστο πρότυπο ταξινομείται στην κλάση 1(μπλε) καθώς το άγνωστο πρότυπο βρίσκεται πιο κοντά στο κέντρο της κλάσης αυτής.*

- **Support Vector Machine (SVM):** Ο ταξινομητής που βασίζεται σε διανυσματικές μηχανές στήριξης μπορεί να χρησιμοποιηθεί σε γραμμικά διαχωρίσιμα και σε μη γραμμικά διαχωρίσιμα δεδομένα. Σε αυτόν τον αλγόριθμο τα δεδομένα παρουσιάζονται ως σημεία σε έναν n-διάστατο χώρο, όπου n ο αριθμός των χαρακτηριστικών, με την τιμή του κάθε χαρακτηριστικού να αντιστοιχεί στις συντεταγμένες του. Στην συνέχεια ο αλγόριθμος καλείται να εντοπίσει εκείνο το υπερεπίπεδο το οποίο να διαχωρίζει τις κατηγορίες αφήνοντας το μεγαλύτερο δυνατό περιθώριο (margin) μεταξύ των σημείων που βρίσκονται κοντά στα όρια της κάθε κλάσης. Δηλαδή το υπερεπίπεδο που θα είναι τελικά αυτό που θα κάνει τον διαχωρισμό θα πρέπει να έχει την μεγαλύτερη δυνατή κάθετη απόσταση από δυο υποθετικά υπερεπίδα, τα οποία θα είναι παράλληλα μεταξύ τους και το καθένα θα εφάπτεται στο πιο ακραίο σημείο στα όρια της κάθε κλάσης. Ο SVM ταξινομητής χρησιμοποιείται και σε μη γραμμικά διαχωρίσιμα δεδομένα με την χρήση κάποιων μαθηματικών συναρτήσεων που αναφέρονται ως συναρτήσεις πυρήνα (kernel functions). Μερικές από τις πιο διαδεδομένες συναρτήσεις πυρήνα είναι η γραμμική (linear), η πολυωνυμική (polynomial), η συνάρτηση ακτινικής βάσης (Radial Basis Function - RBF) και η σιγμοειδής (sigmoid) [53][54][59]. Η συνάρτηση διάκρισης του ταξινομητή περιγράφεται από την ακόλουθη σχέση:

$$d(\mathbf{x}) = \sum_{i=1}^{N} y_i a_i \, k(\mathbf{x}, \mathbf{x}_i) + b \tag{1.8}$$

Όπου x είναι το άγνωστο πρότυπο, x_i είναι το πρότυπο της κλάσης i, το N είναι το πλήθος των προτύπων της κλάσης i, το y_i είναι 1 ή -1 ανάλογα με την κλάση 1 ή 2 που ανήκει, a_i και b είναι συντελεστές και k είναι η κατάλληλη συνάρτηση πυρήνα. Μερικές από τις πιο γνωστές συναρτήσεις πυρήνα περιγράφονται παρακάτω.

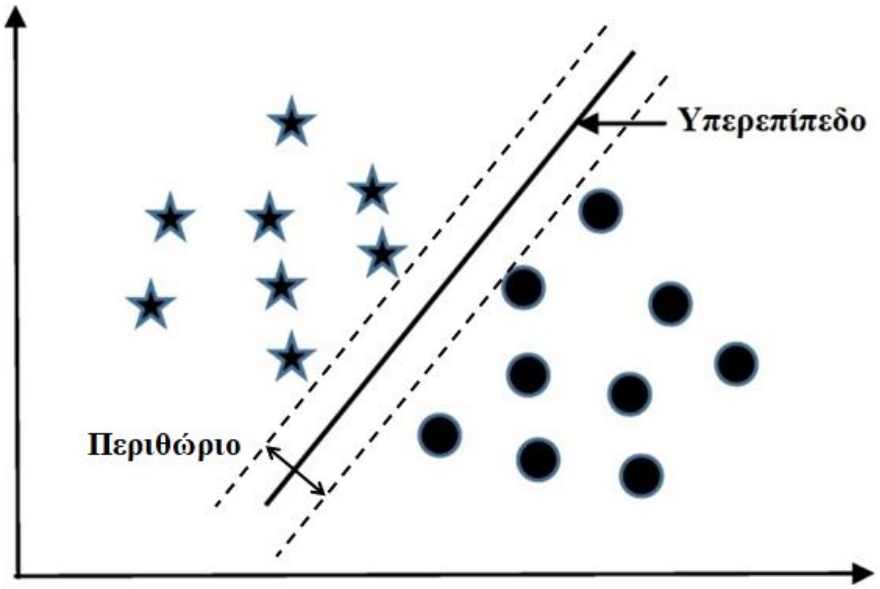
Γραμμική (linear): $k(x, y) = x^T y$

Πολυωνυμική (Polynomial): $k(x, y) = (x^T y + b)^n$

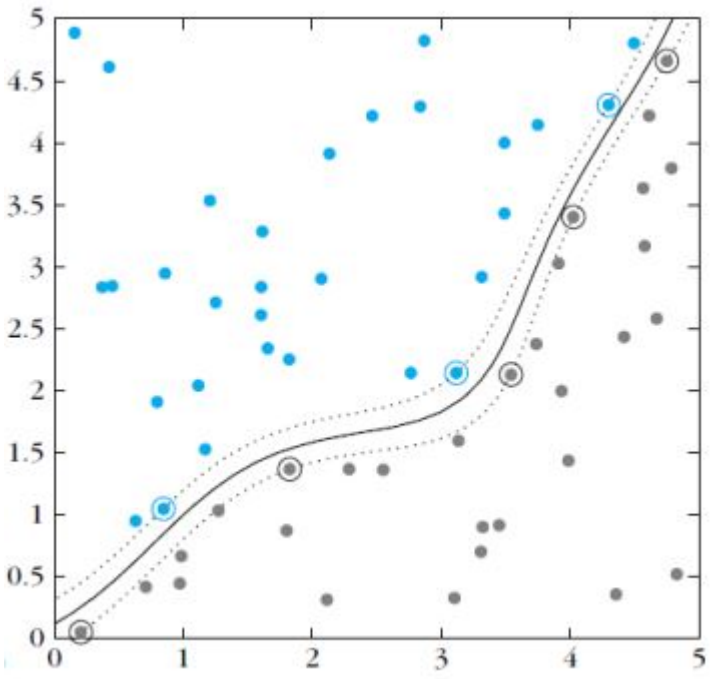
Ακτινικής Βάσης (RBF): $k(x, y) = \exp\left(-\frac{||x-y||^2}{2\sigma^2}\right)$

Σιγμοειδής (Sigmoid): $k(x, y) = \tanh\left(p(x^T y) + b\right)$

Όπου n,b,σ,b παράμετροι. Το άγνωστο πρότυπο ταξινομείται στην κλάση 1 αν η τιμή της συνάρτησης διάκρισης είναι μεγαλύτερη του μηδενός και στην κλάση 2 αν η τιμή της συνάρτησης είναι μικρότερη του μηδενός [58][60].

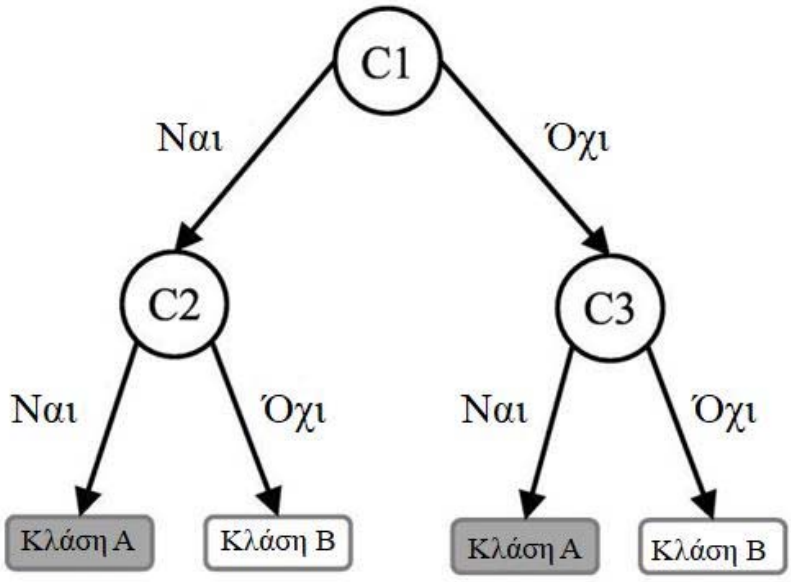


Εικόνα 1.15: *Σχηματική αναπαράσταση του ταξινομητή γραμμικού SVM, όπου διακρίνεται εκείνο το υπερεπίπεδο που αφήνει το μεγαλύτερο περιθώριο ανάμεσα στις 2 κλάσεις. (αναπροσαρμοσμένο από πηγή)* [59]



Εικόνα 1.16: *Σχηματική αναπαράσταση ενός μη γραμμικού SVM.* [58]

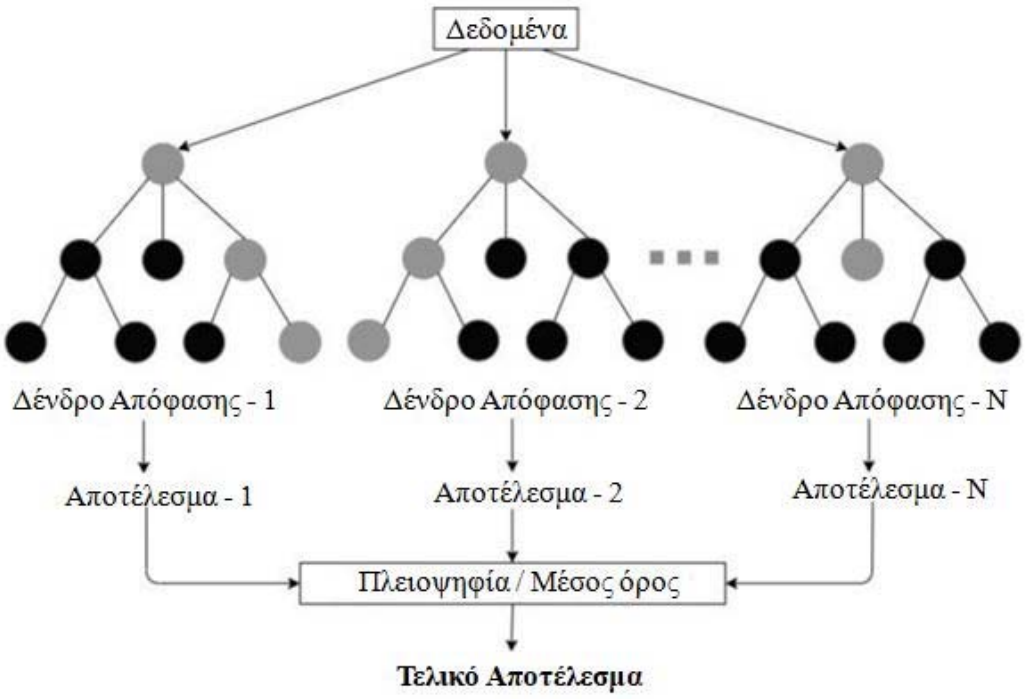
- **Decision Trees:** Οι ταξινομητές που στηρίζονται στους αλγορίθμους δένδρων αποφάσεων λαμβάνουν υπόψη τους τις αποφάσεις που παίρνουν, πιθανές επιπτώσεις, καθώς και το πιθανό αποτέλεσμα και μπορούν να παρασταθούν γραφικά με την μορφή ενός δένδρου. Ένα δένδρο απόφασης αποτελείται από κόμβους (nodes), διακλαδώσεις (branches) και το τελικό αποτέλεσμα (leaf node). Στους κόμβους λαμβάνονται αποφάσεις με την λογική ναι/όχι και κάθε διακλάδωση αποτελεί και ένα αποτέλεσμα. Ένα δένδρο απόφασης μπορεί να αποτελείται από αρκετά επίπεδα, όμως ο πρώτος κόμβος χαρακτηρίζεται ως ρίζα (root) του δένδρου. Οι εσωτερικοί κόμβοι αντιπροσωπεύουν ερωτήσεις για τις μεταβλητές εισόδου που ανάλογα με την απάντηση, ο αλγόριθμος προσανατολίζεται προς το κατάλληλο τελικό αποτέλεσμα το οποίο αποτελεί και το αποτέλεσμα της ταξινόμησης. Ένας από τους αλγορίθμους ταξινόμησης των δένδρων αποφάσεων είναι ο CART (Classification And Regression Tree). Ο CART αποτελεί ένα μοντέλο πρόβλεψης του οποίου το αποτέλεσμα στηρίζεται στις υπάρχουσες τιμές του ήδη κατασκευασμένου δένδρου. Ένα CART μοντέλο παρουσιάζεται ως ένα δυαδικό δένδρο του οποίου κάθε ρίζα αναπαριστά μία μόνο είσοδο και ένα συγκεκριμένο σημείο της μεταβλητής εισόδου. Τα τελικά στάδια (leaf nodes) περιέχουν ένα αποτέλεσμα το οποίο χρησιμοποιείται για την πρόβλεψη [53][59].



Εικόνα 1.17: *Σχηματική αναπαράσταση ενός δένδρου απόφασης. Οι κόμβοι αναπαριστώνται ως κύκλοι και τα αποτελέσματα ως ορθογώνια (αναπροσαρμοσμένο από πηγή)* [59]

- **Random Forest (RF):** Ένα τυχαίο δάσος είναι ένας αλγόριθμος ταξινόμησης ο οποίος αποτελείται από πολλά δένδρα απόφασης, όπως αντίστοιχα και ένα κανονικό δάσος αποτελείται από πληθώρα δένδρων. Το κάθε δένδρο απόφασης εκπαιδεύεται χρησιμοποιώντας διαφορετικά τυχαία τμήματα (subsets) από το σύνολο των δεδομένων εκπαίδευσης. Προκειμένου να ταξινομηθεί ένα άγνωστο πρότυπο θα πρέπει η είσοδος να φτάσει στο τελικό

στάδιο κάθε δένδρου. Κάθε ένα από τα δένδρα λαμβάνει υπόψη του διαφορετικό μέρος της αρχικής εισόδου και με βάση αυτό παρέχει το αποτέλεσμα. Η τελική ταξινόμηση γίνεται σύμφωνα με την πλειοψηφία ή τον μέσο όρο των αποτελεσμάτων που έχουν παράξει τα δένδρα απόφασης. Με αυτόν τον τρόπο μειώνεται το πρόβλημα της υπερπροσαρμογής των δεδομένων και αυξάνεται η ακρίβεια και ο έλεγχος της πρόβλεψης. Ως εκ τούτου, ένα μοντέλο τυχαίου δάσους με πολλά δένδρα απόφασης είναι πιο ακριβές από ένα μοντέλο που χρησιμοποιεί μόνο ένα δένδρο απόφασης [54][59].

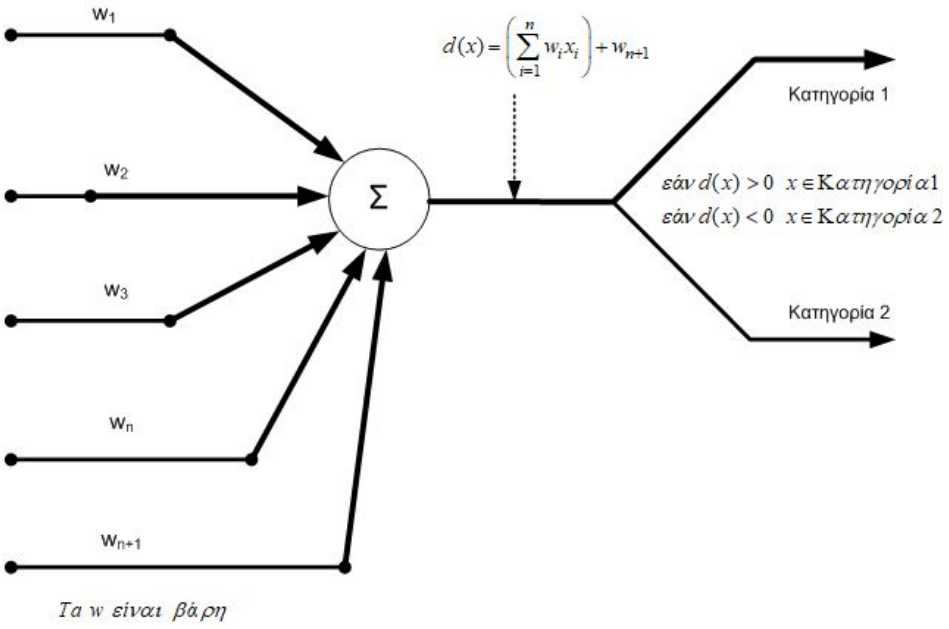


Εικόνα 1.18: *Σχηματική αναπαράσταση του αλγορίθμου τυχαίου δάσους αποτελούμενο από πολλαπλά δένδρα απόφασης. (αναπροσαρμοσμένο από πηγή)* [54]

- **Artificial Neural Network (ANN):** Τα τεχνητά νευρωνικά δίκτυα είναι μια κατηγορία αλγορίθμων Μηχανικής Μάθησης εμπνευσμένη από την λειτουργία των νευρώνων του ανθρώπινου εγκεφάλου. Σε ένα τεχνητό νευρωνικό δίκτυο κάθε νευρώνας δέχεται μια πληθώρα από σήματα εισόδου, όπως γίνεται και με τους δενδρίτες στους νευρώνες του ανθρώπου, και στην συνέχεια εκτελείται ένα άθροισμα αυτών των σημάτων υπό την επήρεια κάποιων συντελεστών, που αναφέρονται ως βάρη. Έτσι, μέσα από μία συνάρτηση ενεργοποίησης δημιουργείται ένα ερέθισμα το οποίο, αν είναι μεγαλύτερο από μια καθορισμένη τιμή κατωφλίου, θα περάσει στην έξοδο του νευρώνα που θα είναι και το τελικό αποτέλεσμα. Ο απλούστερος τεχνητός νευρώνας ονομάζεται perceptron και ο ταξινομητής που προκύπτει από αυτόν, καλείται ως ταξινομητής Perceptron [26][59][63]. Η συνάρτηση διάκρισης του ταξινομητή Perceptron περιγράφεται ακολούθως:

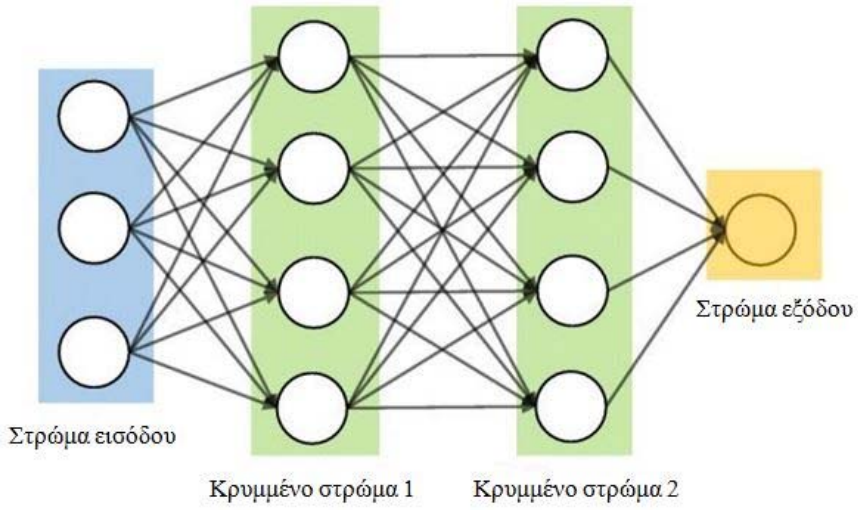
$$d(x)=\left(\sum_{i=1}^{n} w_i x_i\right)+w_{n+1} \quad \text{, όπου } w_{n+1}=1 \tag{1.9}$$

Το άγνωστο πρότυπο ταξινομείται σε μια από 2 κατηγορίες ανάλογα από το αν η τιμή που θα λάβει η συνάρτηση είναι πιο κοντά στο 1 ή στο -1 [60].



Εικόνα 1.19: *Βασικό στοιχείο του ταξινομητή Perceptron.* [60]

Εκτός από τον Perceptron υπάρχει και ο Perceptron πολλαπλών επιπέδων (Multi Layer Perceptron – MLP), ο οποίος αποτελείται από ένα δίκτυο νευρώνων. Ένα δίκτυο νευρώνων αποτελείται από ένα στρώμα εισόδου (input layer), ένα ή περισσότερα κρυμμένα στρώματα (hidden layers) τα οποία αποτελούνται από κόμβους (nodes) και το στρώμα εξόδου (output layer). Το στρώμα εισόδου λαμβάνει το σήμα εισόδου, στην συνέχεια το σήμα αποστέλλεται σε κάθε νευρώνα του επόμενου επιπέδου πολλαπλασιασμένο με ένα κατάλληλο βάρος. Η διαδικασία συνεχίζεται για όλα τα εσωτερικά στρώματα του δικτύου μέχρι να καταλήξει στο στρώμα εξόδου, από το οποίο θα προκύψει το αποτέλεσμα της ταξινόμησης [63][64].



Εικόνα 1.20: *Σχηματική αναπαράσταση ενός τεχνητού νευρωνικού δικτύου (ANN) με 2 κρυμμένα στρώματα (hidden layers). (αναπροσαρμοσμένο από πηγή)* [59]

1.3.4 Κανονικοποίηση δεδομένων και μείωση χαρακτηριστικών

Η κανονικοποίηση των δεδομένων αποτελεί βασικό και αναπόσπαστο στάδιο της διαδικασίας σχεδιασμού ενός συστήματος Μηχανικής Μάθησης. Οι αριθμητικές τιμές των διαφόρων χαρακτηριστικών που αποτελούν τα σύνολα δεδομένων (datasets) μπορεί να έχουν τιμές που να είναι αρκετά διαφορετικές μεταξύ τους. Για παράδειγμα, ένα χαρακτηριστικό όπως είναι το μοριακό βάρος μιας χημικής ένωσης μπορεί να έχει τιμές από μηδέν μέχρι και μερικές χιλιάδες, ανάλογα με τον αριθμό των ατόμων που υπάρχουν στην ένωση. Από την άλλη, ένα άλλο χαρακτηριστικό όπως είναι ο δεκαδικός λογάριθμος του συντελεστή λιποφιλικότητας (logP) μπορεί να λάβει θετικές αλλά και αρνητικές τιμές. Είναι προφανές λοιπόν ότι τα χαρακτηριστικά που χρησιμοποιούνται σε ένα σύστημα ενδέχεται να έχουν αρκετά διαφορετικά εύρη τιμών, τα οποία μπορούν να επηρεάσουν αρνητικά το σύστημα. Για τον λόγο αυτό, οι τιμές των χαρακτηριστικών πρέπει να κανονικοποιηθούν, ώστε να υπάρχει ομοιογένεια και επηρεάζουν ισότιμα την λήψη αποφάσεων του συστήματος. Ένας συχνά χρησιμοποιούμενος τρόπος κανονικοποίησης περιγράφεται από την ακόλουθη σχέση:

$$\overline{f_i} = \frac{(f_i - \mu)}{\sigma} \tag{1.10}$$

Όπου $\overline{f_i}$ είναι το κανονικοποιημένο χαρακτηριστικό, μ είναι η μέση τιμή και σ είναι η τυπική απόκλιση του χαρακτηριστικού f_i. Για τον υπολογισμό της τυπικής απόκλισης και της μέσης τιμής λαμβάνονται υπόψη όλες οι τιμές του συγκεκριμένου χαρακτηριστικού σε όλες τις κλάσεις [60].

Σε περιπτώσεις που το πλήθος των χαρακτηριστικών είναι αρκετά μεγάλο εφαρμόζονται αλγόριθμοι ελάττωσης χαρακτηριστικών ώστε να μειωθεί το υπολογιστικό κόστος, να αποφευχθεί το φαινόμενο της υπερπροσαρμογής (overfitting) και να απορριφθούν περιττά χαρακτηριστικά, τα οποία δεν προσφέρουν κάποια χρήσιμη πληροφορία για το σύστημα. Η μείωση των χαρακτηριστικών είναι μια διαδικασία κατά την οποία επιλέγεται ένα υποσύνολο μοναδικών χαρακτηριστικών από τα υπάρχοντα χαρακτηριστικά, το οποίο χρησιμοποιείται τελικά για τον σχεδιασμό του συστήματος. Η διαδικασία αυτή μειώνει την πολυπλοκότητα του συστήματος και συνεπώς τον χρόνο των υπολογισμών, απορρίπτοντας χαρακτηριστικά τα οποία θεωρείται ότι δεν παρέχουν κάποια χρήσιμη πληροφορία [54]. Τα χαρακτηριστικά τα οποία θα χρησιμοποιηθούν τελικά στο σύστημα είναι επιθυμητό να έχουν όσο το δυνατό μικρότερη συσχέτιση μεταξύ τους, διότι χαρακτηριστικά με υψηλή συσχέτιση δεν συμβάλουν στην διαχωρισιμότητα μεταξύ των κλάσεων. Οι αλγόριθμοι που χρησιμοποιούνται για την μείωση των χαρακτηριστικών ιεραρχούν τα χαρακτηριστικά με βάση κάποια κριτήρια και στην συνέχεια επιλέγουν τα χαρακτηριστικά με την καλύτερη βαθμολογία. Η ιεράρχηση αυτή μπορεί να πραγματοποιηθεί με βάση την συσχέτιση των χαρακτηριστικών, αξιοποιώντας κάποιο κατώφλι διακύμανσης (variance threshold) ή κάποιον συντελεστή συσχέτισης, όπως τον συντελεστή συσχέτισης Pearson. Ακόμα, τα χαρακτηριστικά μπορούν να ιεραρχηθούν με βάση τη στατιστική διαφορά που παρουσιάζουν μεταξύ τους, αξιοποιώντας στατιστικές δοκιμασίες όπως το t-test, το Wilcoxon test, το ANOVA test και το chi-squared test. Επιπλέον, μπορεί να γίνει συνδυασμός των παραπάνω μεθόδων και τα χαρακτηριστικά να ιεραρχηθούν με βάση τον συνδυασμό συσχέτισης και στατιστικής διαφοράς. Επιπροσθέτων μπορούν να αξιοποιηθούν μαθηματικοί μετασχηματισμοί, όπως είναι η μέθοδος PCA, καθώς και μέθοδοι συνολικότητας (wrapper methods), όπως είναι ο αλγόριθμός επαναλαμβανόμενου αποκλεισμού χαρακτηριστικών (Recursive Feature Elimination - RFE) σε συνδυασμό με μεθόδους στατιστικής εξάρτησης (regression), όπως είναι ο Random Forest Regressor, η Logistic Regression και ο Decision Trees Regressor [54][58][60]. Παρακάτω περιγράφονται μερικές από τις προηγούμενες διαδικασίες.

- **Κατώφλι Διακύμανσης (Variance Threshold):** Η μέθοδος αυτή απορρίπτει τα χαρακτηριστικά που παρουσιάζουν χαμηλή διακύμανση, για παράδειγμα τα χαρακτηριστικά που έχουν διακύμανση μικρότερη από το κατώφλι. Επιπλέον, απορρίπτει εξ αρχής τα χαρακτηριστικά με μηδενική διακύμανση, δηλαδή τα χαρακτηριστικά εκείνα που έχουν την ίδια τιμή σε όλα τα πρότυπα [54].

- **Συσχέτιση Pearson (Pearson correlation):** Αυτή η μέθοδος εξετάζει την συσχέτιση που έχουν τα χαρακτηριστικά μεταξύ τους ανά δύο, μέσα από την εξίσωση του συντελεστή Pearson (σχέση 1.11). Οι τιμές του συντελεστή Pearson βρίσκονται στο διάστημα [-1,1], με το -1 να συμβολίζει την απόλυτη αρνητική συσχέτιση, το +1 δείχνει την απόλυτη θετική συσχέτιση και το 0

συμβολίζει ότι τα δύο χαρακτηριστικά δεν έχουν γραμμική συσχέτιση. Για δύο τυχαία χαρακτηριστικά X και Y η τιμή του συντελεστή Pearson περιγράφεται από την ακόλουθη σχέση:

$$r(X,Y) = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \quad (1.11)$$

Όπου X_i και Y_i η τιμή του χαρακτηριστικού X και Y αντίστοιχα στο κάθε πρότυπο i, το n είναι το πλήθος των προτύπων και $\bar{X}, \bar{Y}$ είναι η μέση τιμή του χαρακτηριστικού X και Y αντίστοιχα [54].

- **ANOVA:** Η ανάλυση διακύμανσης (Analysis of Variance) είναι μια μέθοδος στατιστικής, η οποία εξετάζει αν υπάρχουν στατιστικά σημαντικές διαφορές ανάμεσα στις μέσες τιμές δύο ή περισσότερων ομάδων δεδομένων. Για να ελέγξει την στατιστική σχέση των παραπάνω μέσων τιμών, η μέθοδος ANOVA αξιοποιεί το στατιστικό κριτήριο F, το οποίο ορίζεται ως ο λόγος της μέσης διακύμανσης των διασπορών μεταξύ των ομάδων προς την μέση διακύμανση των διασπορών ανάμεσα σε κάθε ομάδα. Η τιμή του αποτελέσματος του κριτηρίου F μπορεί να χρησιμοποιηθεί για την μείωση των χαρακτηριστικών, καθώς τα χαρακτηριστικά που δεν παρουσιάζουν στατιστικά σημαντική διαφορά μπορούν να παραλειφθούν. Στην περίπτωση που η μέθοδος ANOVA χρησιμοποιηθεί μόνο για δύο ομάδες δειγμάτων, τότε είναι ισοδύναμη με το στατιστικό t-test ανάμεσα σε δύο ομάδες δεδομένων [54][60].

- **Recursive Feature Elimination (RFE):** Ο αλγόριθμος επαναλαμβανόμενου αποκλεισμού χαρακτηριστικών (RFE) συνήθως χρησιμοποιείται σε συνδυασμό με κάποιο μοντέλο στατιστικής εξάρτησης όπως είναι η λογιστική παλινδρόμηση (Logistic Regression) ή η παλινδρόμηση δένδρων απόφασης (Decision Trees Regressor). Η μέθοδος αυτή εφαρμόζει το εκάστοτε στατιστικό μοντέλο και κάθε φορά απορρίπτει το πιο αδύναμο χαρακτηριστικό μέχρι να φτάσει στο επιθυμητό πλήθος χαρακτηριστικών που έχει οριστεί. Τα χαρακτηριστικά βαθμολογούνται από συντελεστές ή δείκτες σημαντικότητας του κάθε μοντέλου. Ο αλγόριθμος RFE αποσκοπεί στο να εξαλείψει τυχόν συσχετίσεις χαρακτηριστικών και συγγραμικότητα στο μοντέλο που χρησιμοποιείται, αφαιρώντας επαναλαμβανόμενα έναν μικρό αριθμό χαρακτηριστικών σε κάθε επανάληψη [54].

- **Principal Component Analysis (PCA):** Η ανάλυση πρωτευόντων συστατικών (PCA) είναι μια μαθηματική τεχνική, η οποία μετασχηματίζει μια ομάδα δεδομένων με συσχετισμένες μεταβλητές σε μια νέα ομάδα δεδομένων με νέες μεταβλητές, οι οποίες καλούνται πρωτεύοντα συστατικά (principal

components). Τα νέα δεδομένα αποκτούν μέγιστη διακύμανση και οι νέες μεταβλητές είναι ασυσχέτιστες. Με την εφαρμογή της μεθόδου PCA επιτυγχάνεται μεγιστοποίηση της διακύμανσης, καθώς το εύρος των τιμών των δεδομένων μειώνεται στους υπόλοιπους άξονες και αυξάνεται σε έναν από τους νέους άξονες. Επιπλέον, δεν υπάρχει συσχέτιση των νέων μεταβλητών μεταξύ τους, οπότε είναι πιθανό να είναι και ανεξάρτητες. Ακόμα, μετά την εφαρμογή της μεθόδου PCA η κατανόηση και η οπτικοποίηση των δεδομένων γίνεται ευκολότερα, καθώς τα νέα δεδομένα αναπαριστώνται σε λιγότερες διαστάσεις από ότι τα αρχικά, γεγονός που καθιστά ευκολότερη και την επεξεργασία τους [54][58].

1.3.5 Αξιολόγηση Ταξινομητών

Η αξιολόγηση ενός ταξινομητή είναι ένα σημαντικό στάδιο της σχεδίασης ενός συστήματος ταξινόμησης. Η έννοια της αξιολόγησης ενός ταξινομητή στην ουσία αναφέρεται στην εύρεση του ποσοστού επιτυχίας του ταξινομητή. Υπάρχουν δύο τρόποι αξιολόγησης ενός συστήματος ταξινόμησης, η εσωτερική αξιολόγηση και η εξωτερική αξιολόγηση. Στην εσωτερική αξιολόγηση γίνεται σχεδιασμός του συστήματος με δεδομένα εκπαίδευσης (training data) και στην συνέχεια το σύστημα αξιολογείται με ένα υποσύνολο των δεδομένων αυτών (validation set), ώστε να βρεθούν οι κατάλληλοι παράμετροι. Αφότου έχει σχεδιαστεί το σύστημα με τις βέλτιστες παραμέτρους, στην συνέχεια αξιολογείται με τα δεδομένα ελέγχου (test data), τα οποία δεν έχουν χρησιμοποιηθεί στο στάδιο προηγούμενο στάδιο σχεδίασης. Η αξιολόγηση του συστήματος με άγνωστα για αυτό δεδομένα ελέγχου καλείται εξωτερική αξιολόγηση [58]. Για την εκτίμηση της επίδοσης ενός συστήματος ταξινόμησης χρησιμοποιείται συχνά ο πίνακας αληθείας (confusion matrix ή truth table, contingency table).

Πίνακας 1.1: *Πίνακας αληθείας για 2 κλάσεις K_1 και K_2.* [60]

	Πίνακας Αληθείας		
Πειραματική κατηγοριοποίηση	Κατηγοριοποίηση με ΗΥ		
		K_1	**K_2**
	K_1	**n_{11}**	**n_{12}**
	K_2	**n_{21}**	**n_{22}**

Στον Πίνακα 1.1 φαίνεται η δομή του πίνακα αληθείας για δύο κλάσεις K_1 και K_2. Με n_{11} και n_{22} συμβολίζεται ο αριθμός των σωστά ταξινομημένων προτύπων και με n_{12} και n_{21} συμβολίζεται ο αριθμός των λανθασμένα ταξινομημένων προτύπων. Από τον πίνακα αληθείας μπορούν να υπολογιστούν τα παρακάτω μεγέθη [60][64]:

- Ανάκληση 1ης κλάσης = $\frac{n_{11}}{n_{11}+n_{12}} \times 100$
- Ανάκληση 2ης κλάσης = $\frac{n_{21}}{n_{21}+n_{22}} \times 100$
- Ακρίβεια 1ης κλάσης = $\frac{n_{11}}{n_{11}+n_{21}} \times 100$
- Ακρίβεια 2ης κλάσης = $\frac{n_{22}}{n_{12}+n_{22}} \times 100$
- Συνολική ακρίβεια = $\frac{n_{11}+n_{22}}{n_{11}+n_{12}+n_{21}+n_{22}} \times 100$

Για την αξιολόγηση του σχεδιασμού ενός συστήματος ταξινόμησης υπάρχουν διάφορες μέθοδοι. Μερικές από αυτές τις μεθόδους περιγράφονται ακολούθως.

- **Resubstitution Method:** Η μέθοδος επαναχρησιμοποίησης (resubstitution method) χρησιμοποιεί το ίδιο σύνολο δεδομένων για την εκπαίδευση και την αξιολόγηση του ταξινομητή. Η συγκεκριμένη μέθοδος παρουσιάζει αρκετά υψηλό ποσοστό επιτυχίας, το οποίο όμως δεν είναι απόλυτα αξιόπιστο για την εκτίμηση τις επίδοσης του ταξινομητή σε άγνωστα δεδομένα. Το πλεονέκτημα αυτής της μεθόδου είναι η ταχύτητα, αφού ο ταξινομητής εκπαιδεύεται μόνο μία φορά για να ταξινομήσει ακολούθως όλα τα πρότυπα [58][65].

- **Holdout Method:** Με την μέθοδο διαμέρισης του συνόλου δεδομένων (Holdout method) το σύνολο δεδομένων χωρίζεται σε δύο υποσύνολα (π.χ. 70% - 30%), από τα οποία το ένα χρησιμοποιείται για την εκπαίδευση του ταξινομητή (training set) και το άλλο χρησιμοποιείται για την αξιολόγησή του (test set). Σημαντικά μειονεκτήματα αυτής της μεθόδου είναι ότι μειώνει το μέγεθος του συνόλου εκπαίδευσης και του συνόλου αξιολόγησης, καθώς και ότι θα πρέπει να αποφασιστεί πόσα από τα συνολικά δεδομένα θα αξιοποιηθούν για ο σύνολο εκπαίδευσης και πόσα για το σύνολο αξιολόγησης. Στην περίπτωση που υπάρχουν αρκετά διαθέσιμα δεδομένα, τότε ο διαχωρισμός σε δεδομένα εκπαίδευσης και αξιολόγησης γίνεται μία φορά. Σε διαφορετική περίπτωση η διαδικασία επαναλαμβάνεται αρκετές φορές με τυχαίο διαχωρισμό των δεδομένων σε σύνολα εκπαίδευσης και αξιολόγησης και η συνολική επίδοση του ταξινομητή εκτιμάται από την μέση τιμή και την τυπική απόκλιση των επιδόσεων [58][60][65].

- **Leave-one-out Method (LOO):** Με την μέθοδο παράλειψης ενός προτύπου (Leave-one-out method) χρησιμοποιούνται όλα τα δεδομένα για την εκπαίδευση του ταξινομητή εκτός από ένα πρότυπο, το οποίο στην συνέχεια ταξινομείται. Η συγκεκριμένη διαδικασία επαναλαμβάνεται τόσες φορές όσο είναι το πλήθος των προτύπων, χρησιμοποιώντας κάθε φορά διαφορετικό πρότυπο για την αξιολόγηση. Η επίδοση του ταξινομητή υπολογίζεται από τον συνολικό αριθμό των προτύπων που έχουν ταξινομηθεί σωστά. Ένα σημαντικό χαρακτηριστικό αυτής της μεθόδου είναι ότι το πρότυπο που καλείται κάθε φορά να ταξινομηθεί θεωρείται άγνωστο για το σύστημα, καθώς δεν συμμετέχει στην εκπαίδευσή του, έτσι επιτυγχάνεται ανεξαρτησία μεταξύ των συνόλων εκπαίδευσης και δοκιμής. Το κύριο μειονέκτημα είναι το υπολογιστικό κόστος, καθώς το σύστημα εκπαιδεύεται τόσες φορές όσα είναι και τα πρότυπα. Οπότε αυξάνεται η πολυπλοκότητα, με αποτέλεσμα η συνολική διαδικασία να είναι αρκετά αργή [58][60][65].

- **k-fold method:** Η μέθοδος k-Ομάδων (k-fold method) είναι μια πιο γενική περίπτωση της μεθόδου παράλειψης ενός προτύπου (leave-one-out). Εδώ τα δεδομένα χωρίζονται σε k ομάδες ίδιου μεγέθους, όπου τα δεδομένων των k-1 ομάδων χρησιμοποιούνται για την εκπαίδευση του ταξινομητή (training set) και τα δεδομένα από την ομάδα που απομένει χρησιμοποιούνται για την αξιολόγηση (test set). Η διαδικασία επαναλαμβάνεται διαδοχικά k φορές για όλες τις ομάδες και η τελική επίδοση του ταξινομητή εκτιμάται από την σωστή ταξινόμηση για όλα τα δεδομένα [58][60].

- **External-Cross-Validation method (ECV):** Κατά την μέθοδο εξωτερικής διασταυρωμένης επικύρωσης (ECV) τα δεδομένα χωρίζονται με τυχαίο τρόπο σε 3 υποσύνολα (π.χ. 50% - 30% - 20%) και το σύστημα σχεδιάζεται 2 φορές. Η μια καλείται εσωτερική σχεδίαση και η άλλη καλείται εξωτερική σχεδίαση. Κατά την εσωτερική σχεδίαση, ο ταξινομητής σχεδιάζεται με το υποσύνολο που αποτελείται από το 50% των συνολικών δεδομένων και στην συνέχεια αξιολογείται από το υποσύνολο που αποτελείται από το 30% των δεδομένων. Η διαδικασία αυτή επαναλαμβάνεται για διάφορους συνδυασμούς χαρακτηριστικών μέχρι να εντοπιστεί ο συνδυασμός χαρακτηριστικών που θα σημειώσει την καλύτερη ακρίβεια ταξινόμησης. Το σύστημα που είχε την καλύτερη επίδοση στην εσωτερική σχεδίαση χρησιμοποιείται κατά την εξωτερική σχεδίαση ώστε να ταξινομήσει το υπόλοιπο 20% των δεδομένων που απομένει. Η συνολική διαδικασία, δηλαδή η εσωτερική και η εξωτερική σχεδίαση, επαναλαμβάνεται τουλάχιστον 10 φορές και η τιμή της μέσης ακρίβειας ταξινόμησης του τρίτου υποσυνόλου δεδομένων, εν προκειμένω του 20% των δεδομένων, αποτελεί μια ένδειξη της ακρίβειας του συστήματος σε άγνωστα δεδομένα. Η συγκεκριμένη μέθοδος αξιολόγησης είναι αρκετά αντικειμενική, γεγονός που την καθιστά ως την πιο προτιμητέα σε σχέση με άλλες μεθόδους. Βέβαια, η αξιοποίηση της μεθόδου εξωτερικής διασταυρωμένη επικύρωσης (ECV) προϋποθέτει την διαθεσιμότητα αρκετά μεγάλου αριθμού δεδομένων, ώστε να χωριστούν στα αντίστοιχα υποσύνολα. Η τελική ακρίβεια αυτής της μεθόδου είναι συνήθως 5% - 10% μικρότερη από την ακρίβεια άλλων μεθόδων, αλλά θεωρείται πιο αξιόπιστη [60].

- **Bootstrap method:** Η μέθοδος Bootstrap είναι παρόμοια με την μέθοδο Holdout, αλλά διαφέρει από αυτή ως προς τον τρόπο με τον οποίο επιλέγονται τα δεδομένα για το σύνολο εκπαίδευσης. Αρχικά, σε αυτήν την μέθοδο τα συνολικά δεδομένα χωρίζονται με τυχαίο τρόπο σε δύο υποσύνολα, ένα για εκπαίδευση (training set) και ένα για αξιολόγηση (test set), για παράδειγμα τα 2/3 των δεδομένων χρησιμοποιούνται για εκπαίδευση και το 1/3 για αξιολόγηση. Η τυχαία δειγματοληψία του συνόλου εκπαίδευσης σε αυτήν την μέθοδο γίνεται με επανατοποθέτηση (re-substitution), δηλαδή τα πρότυπα που επιλέγονται για το σύνολο εκπαίδευσης μπορούν να επιλεχθούν περισσότερες από μια φορές. Εάν ο αριθμός των δεδομένων είναι μικρός, τότε η διαδικασία επαναλαμβάνεται για συγκεκριμένο αριθμό επαναλήψεων, ο οποίος αναφέρεται ως εποχή (epoch). Η συνολική ακρίβεια του συστήματος προκύπτει από την μέση τιμή και την τυπική απόκλιση όλων των επιδόσεων. Η μέθοδος αυτή προτιμάται όταν τα συνολικά διαθέσιμα δεδομένα είναι λίγα [60].

Κεφάλαιο 2: Περιγραφή Διαδικασίας - Μεθοδολογία

2.1 Εισαγωγή

Υπό το πρίσμα όσων αναφέρθηκαν στα προηγούμενα κεφάλαια, για την συγκεκριμένη εργασία επιλέχθηκαν 2 ομάδες δεδομένων από χημικές ενώσεις φυσικών προϊόντων. Η μια ομάδα αφορά την κουρκουμίνη και ενώσεις που έχουν παρόμοια χημική δομή με αυτήν και η άλλη ομάδα αφορά την ρεσβερατρόλη και ενώσεις με παρόμοια χημική δομή. Η κουρκουμίνη και η ρεσβερατρόλη έχουν αρκετά διαφορετικές χημικές δομές, το οποίο ήταν και ζητούμενο για την επιλογή τους. Οι ενώσεις που επιλέχθηκαν για την δημιουργία των 2 ομάδων δεδομένων προέκυψαν με βάση τον έλεγχο της ομοιότητας Tanimoto, που πραγματοποιήθηκε μέσα από μία βάση δεδομένων περίπου 2,5 εκατομμυρίων ενώσεων. Στην συνέχεια υπολογίστηκαν 1D και 2D μοριακοί περιγραφείς για καθεμία από τις ενώσεις που αποτελούσαν τις 2 ομάδες δεδομένων. Ο στόχος της παρούσας εργασίας ήταν να γίνει μια αξιολόγηση κάποιων τεχνικών ταξινόμησης Μηχανικής Μάθησης σε ότι αφορά τον διαχωρισμό χημικών ενώσεων σε 2 κατηγορίες με βάση τους μοριακούς περιγραφείς. Η μια κατηγορία αφορά ενώσεις που ταιριάζουν περισσότερο στο χημικό προφίλ της κουρκουμίνης και η άλλη κατηγορία αφορά ενώσεις που ταιριάζουν στο χημικό προφίλ της ρεσβερατρόλης.

2.2 Ανάκτηση Δεδομένων

Οι πληροφορίες, όπως τα SMILES και τα InChi keys, για τις ενώσεις που χρησιμοποιήθηκαν, ανακτήθηκαν από την βάση δεδομένων Cnatural [66]. Η Cnatural είναι μια υπό κατασκευή πλατφόρμα που περιλαμβάνει μια βάση δεδομένων φυσικών προϊόντων, τα οποία μπορούν να χρησιμοποιηθούν ως συστατικά σε συμπληρώματα διατροφής. Επίσης, σαν πλατφόρμα, θα παρέχει την δυνατότητα άντλησης δεδομένων για περαιτέρω επεξεργασία ή σύνδεση με εξωτερικά λογισμικά. Περιέχει περίπου 2,5 εκατομμύρια χημικές ενώσεις, εκ των οποίων περίπου οι 300 χιλιάδες είναι φυσικά προϊόντα που έχουν συλλεχθεί από άλλες βάσεις δεδομένων όπως η ChEMBL, η ChEBI, η NPASS, η CTD και η Zinc. Επιπλέον, παρέχει πληροφορίες για την φυτική προέλευση της κάθε ένωσης, βιβλιογραφία για τις in vitro και in vivo δοκιμές που έχουν πραγματοποιηθεί, καθώς και συσχετίσεις ενώσεων με ασθένειες που έχουν καταγραφεί. Τέλος, η Cnatural προσφέρει πληροφορίες για την 2D και 3D απεικόνιση των μορίων αλλά και για κάποιους μοριακούς περιγραφείς για την κάθε ένωση, όπως είναι η λιποφιλικότητα, το μοριακό βάρος, η διαλυτότητα κ.α.

2.3 Υπολογισμός μοριακών αποτυπωμάτων και έλεγχος ομοιότητας

Για την εργασία αυτή όλες οι ενέργειες ανάκτησης δεδομένων και οι υπολογισμοί έγιναν σε περιβάλλον Python 3 σε απομακρυσμένο διακομιστή με επεξεργαστή 64 πυρήνων, 132 GB μνήμη RAM και 2 κάρτες γραφικών. Αρχικά μέσω της βιβλιοθήκης Psycopg2 [67] ανακτήθηκαν τα SMILES και τα InChi keys για την κουρκουμίνη και την ρεσβερατρόλη. Στην συνέχεια ανακτήθηκαν τα SMILES και τα InChi keys όλων των ενώσεων της Cnatural με τον ίδιο τρόπο. Η βιβλιοθήκη Psycopg2 προσαρμόζει εντολές της γλώσσας PostgreSQL σε περιβάλλον Python, για τον λόγο αυτό χρησιμοποιήθηκε καταλλήλως για την σύνδεση στην βάση Cnatural, την αναζήτηση μέσα σε αυτήν και την ανάκτηση των επιθυμητών δεδομένων. Έπειτα χρησιμοποιήθηκε το πακέτο RDKit για τον υπολογισμό των μοριακών αποτυπωμάτων, τόσο για την κουρκουμίνη και την ρεσβερατρόλη, όσο και για όλες τις υπόλοιπες ενώσεις. Το RDKit είναι ένα εργαλείο χημειοπληροφορικής ανοιχτού κώδικα το οποίο χρησιμοποιείται σε πολλές εφαρμογές της χημειοπληροφορικής, από την 2D και 3D απεικόνιση και επεξεργασία χημικών ενώσεων, μέχρι και τον υπολογισμό μοριακών περιγραφέων για χρήση σε μεθόδους Μηχανικής Μάθησης [68]. Ειδικότερα χρησιμοποιήθηκε η συνάρτηση Chem.MolFromSmiles για την μετατροπή των SMILES σε μορφή κατάλληλη για να διαβαστεί από το πρόγραμμα και στην συνέχεια αξιοποιήθηκε η συνάρτηση Chem.RDKFingerprint για την δημιουργία των μοριακών αποτυπωμάτων των ενώσεων. Από τις αρκετές επιλογές μοριακών αποτυπωμάτων του παρέχει το RDKit επιλέχθηκαν τα αποτυπώματα του ίδιου του πακέτου, τα οποία είναι βασισμένα στα αποτυπώματα Daylight. Πρόκειται δηλαδή για τοπολογικά αποτυπώματα 2048 bit.

Εφόσον είχαν υπολογιστεί τα μοριακά αποτυπώματα για όλες τις ενώσεις, με την χρήση της συνάρτησης DataStructs.FingerprintSimilarity έγινε ο έλεγχος της ομοιότητας Tanimoto, αρχικά της κουρκουμίνης με όλες τις ενώσεις της Cnatural και έπειτα της ρεσβερατρόλης με όλες τις ενώσεις. Για την δημιουργία των 2 συνόλων δεδομένων (datasets) επιλέχθηκαν οι ενώσεις όπου είχαν ομοιότητα Tanimoto μεγαλύτερη ή ίση με 0,75 ($Tc \geq 0,75$). Οπότε, προέκυψαν 79 ενώσεις με χημική δομή παρόμοια με την κουρκουμίνη και 79 ενώσεις με χημική δομή παρόμοια με την ρεσβερατρόλη για $Tc \geq 0,75$.

2.4 Υπολογισμός μοριακών περιγραφέων και δημιουργία datasets

Για τις συνολικά 158 ενώσεις υπολογίστηκαν 1D και 2D μοριακοί περιγραφείς από την βιβλιοθήκη του RDKit. Πιο συγκεκριμένα με την συνάρτηση ML.Descriptors.MoleculeDescriptors.CalcDescriptors υπολογίστηκαν 208 μοριακοί περιγραφείς για κάθε ένωση και αποθηκεύτηκαν σε 2 αρχεία Excel, στα οποία οι στήλες αποτελούσαν τους μοριακούς περιγραφείς και οι γραμμές τις χημικές ενώσεις. Στην συγκεκριμένη εργασία οι μοριακοί περιγραφείς αποτελούν τα χαρακτηριστικά (features) των 2 κλάσεων που χρησιμοποιήθηκαν αργότερα στο σημείο της Μηχανικής Μάθησης. Από τους 208 περιγραφείς αφαιρέθηκαν από τα σύνολα

δεδομένων οι περιγραφείς οι οποίοι είχαν μηδενικές ή κενές τιμές. Ακόμα, αφαιρέθηκε μια ένωση από το σύνολο δεδομένων των ενώσεων που η χημική τους δομή είναι παρόμοια με την δομής της ρεσβερατρόλης, διότι υπήρχαν αρκετές μηδενικές τιμές στους μοριακούς περιγραφείς αυτής της ένωσης. Γεγονός που υποδείκνυε ότι υπήρχε κάποιο πρόβλημα με τον υπολογισμό τους. Έτσι, τα 2 σύνολα δεδομένων που προέκυψαν περιείχαν 64 μοριακούς περιγραφείς και 79 ενώσεις, για το σύνολο δεδομένων της κουρκουμίνης και 64 μοριακούς περιγραφείς και 78 ενώσεις για το σύνολο δεδομένων της ρεσβερατρόλης. Ο Πίνακας 2.1 περιέχει τους μοριακούς περιγραφείς που συμπεριλήφθηκαν στα τελικά σύνολα δεδομένων. Οι συγκεκριμένοι μοριακοί περιγραφείς παρέχουν πληροφορίες σχετικά με τις φυσικοχημικές ιδιότητες των ενώσεων και τον τρόπο με τον οποίο συνδέονται και αλληλεπιδρούν τα άτομα της κάθε ένωσης [69].

Πίνακας 2.1: *Οι 64 μοριακοί περιγραφείς που συμπεριλήφθηκαν στα σύνολα δεδομένων.*

Μοριακοί Περιγραφείς (Molecular Descriptors)			
MaxEStateIndex	BCUT2D_MWHI	Chi2n	SMR_VSA10
MinEStateIndex	BCUT2D_MWLOW	Chi2v	SMR_VSA7
MaxAbsEStateIndex	BCUT2D_CHGHI	Chi3n	SMR_VSA9
MinAbsEStateIndex	BCUT2D_CHGLO	Chi3v	SlogP_VSA11
Qed	BCUT2D_LOGPHI	Chi4n	SlogP_VSA2
MolWt	BCUT2D_LOGPLOW	Chi4v	SlogP_VSA5
HeavyAtomMolWt	BCUT2D_MRHI	HallKierAlpha	SlogP_VSA6
ExactMolWt	BCUT2D_MRLOW	Ipc	SlogP_VSA8
NumValenceElectrons	BalabanJ	Kappa1	TPSA
MaxPartialCharge	BertzCT	Kappa2	VSA_EState4
MinPartialCharge	Chi0	Kappa3	VSA_EState5
MaxAbsPartialCharge	Chi0n	LabuteASA	VSA_EState6
MinAbsPartialCharge	Chi0v	PEOE_VSA1	VSA_EState7
FpDensityMorgan1	Chi1	PEOE_VSA6	HeavyAtomCount
FpDensityMorgan2	Chi1n	PEOE_VSA7	MolLogP
FpDensityMorgan3	Chi1v	SMR_VSA1	MolMR

2.5 Μηχανική Μάθηση

Οι αλγόριθμοι που χρησιμοποιήθηκαν για το μέρος της Μηχανικής Μάθησης ανακτήθηκαν από την βιβλιοθήκη sklearn της γλώσσας προγραμματισμού Python [70]. Αρχικά ανακτήθηκαν από το πρόγραμμα τα δεδομένα από τα 2 αρχεία Excel και χωρίστηκαν σε 2 κλάσεις. Η κλάση Α περιείχε τις τιμές δεδομένων από τις ενώσεις που η χημική τους δομή έμοιαζε περισσότερο με την δομή της κουρκουμίνης και η κλάση Β περιείχε τις τιμές δεδομένων από τις ενώσεις που έμοιαζαν περισσότερο με την ρεσβερατρόλη. Στην συνέχεια χρησιμοποιήθηκε η συνάρτηση

preprocessing.normalize για να κανονικοποιηθούν τα δεδομένα και να λάβουν τιμές που βρίσκονται στο διάστημα [-1,1]. Με την κανονικοποίηση των δεδομένων επιτυγχάνεται ότι τα δεδομένα θα προσαρμοστούν κατάλληλα και θα λάβουν τιμές σε κοινό διάστημα. Σε διαφορετική περίπτωση, ενδέχεται τα δεδομένα να είχαν τιμές σε διαφορετικά διαστήματα, το οποίο πιθανώς να οδηγούσε σε παραπλανητικά αποτελέσματα κατά την ταξινόμηση [65].

Ο στόχος της εργασίας ήταν να γίνει η ταξινόμηση λαμβάνοντας υπόψη όσο το δυνατόν περισσότερα χαρακτηριστικά. Για το λόγο αυτό η διαδικασία της ταξινόμησης πραγματοποιήθηκε για 2 περιπτώσεις. Στην πρώτη περίπτωση χρησιμοποιήθηκαν όλα τα χαρακτηριστικά που φαίνονται στον Πίνακα 2.1 και στην συνέχεια χρησιμοποιήθηκαν συνδυασμοί χαρακτηριστικών μειώνοντας το πλήθος των χαρακτηριστικών κάθε φορά κατά ένα. Δηλαδή αρχικά αξιοποιήθηκε ο συνδυασμός 64 χαρακτηριστικών, έπειτα όλοι οι πιθανοί συνδυασμοί ανά 63 χαρακτηριστικά, κατόπιν ανά 62, ανά 61 και ανά 60. Οι παραπάνω συνδυασμοί χρησιμοποιήθηκαν για τον ταξινομητή που σημείωσε την καλύτερη επίδοση στον συνδυασμό των 64 μοριακών περιγραφέων. Στην δεύτερη περίπτωση έγινε επιλογή 26 βέλτιστων χαρακτηριστικών μέσω της μεθόδου Recursive Feature Elimination (RFE) χρησιμοποιώντας την συνάρτηση feature_selection.RFE από την βιβλιοθήκη sklearn και το μοντέλο ExtraTreesClassifier για 20 εκτιμητές (estimators). Σε αυτήν την περίπτωση αξιοποιήθηκε για την ταξινόμηση ο συνδυασμός των 26 βέλτιστων χαρακτηριστικών σε όλους τους ταξινομητές και έπειτα όλοι οι πιθανοί συνδυασμοί ανά 25, ανά 24, ανά 23, ανά 22, ανά 21, ανά 20 και ανά 19 χαρακτηριστικά για τον ταξινομητή με την καλύτερη επίδοση. Στον Πίνακα 2.2 αναφέρονται τα ονόματα των 26 βέλτιστων χαρακτηριστικών.

Πίνακας 2.2: *Οι 26 βέλτιστοι μοριακοί περιγραφείς που προέκυψαν μετά από την ελάττωση χαρακτηριστικών.*

Μοριακοί Περιγραφείς (Molecular Descriptors)		
MinEStateIndex	BCUT2D_MWLOW	Kappa3
Qed	BCUT2D_LOGPHI	PEOE_VSA1
NumValenceElectrons	BCUT2D_MRHI	SMR_VSA7
MinPartialCharge	BCUT2D_MRLOW	SMR_VSA9
MaxAbsPartialCharge	BalabanJ	SlogP_VSA11
MinAbsPartialCharge	Chi2v	SlogP_VSA5
FpDensityMorgan2	Chi3v	SlogP_VSA8
FpDensityMorgan3	Ipc	TPSA
BCUT2D_MWHI	Kappa2	

Για την ταξινόμηση χρησιμοποιήθηκαν 10 αλγόριθμοι από την βιβλιοθήκη sklearn. Η αξιολόγηση της ακρίβειας του συστήματος ταξινόμησης έγινε με την μέθοδο Bootstrap, η οποία εσωτερικά χρησιμοποίησε την μέθοδο αξιολόγησης k-fold. Κατά την αξιολόγηση k-fold τα δεδομένα χωρίζονται σε k ίσα μέρη, από τα οποία τα k-1 χρησιμοποιούνται για να σχεδιαστεί το σύστημα ταξινόμησης (training set) και το ένα

χρησιμοποιείται για την αξιολόγηση (test set). Η διαδικασία αυτή επαναλαμβάνεται k φορές για κάθε test set και η ακρίβεια του ταξινομητή υπολογίζεται από την σωστή ταξινόμηση όλων των δεδομένων. Κατά την μέθοδο αξιολόγησης Bootstrap τα δεδομένα χωρίζονται τυχαία σε training set και test set, για παράδειγμα τα 3/4 των δεδομένων χρησιμοποιούνται για την εκπαίδευση και το 1/4 για την αξιολόγηση. Σε αυτή την μέθοδο τα δεδομένα εκπαίδευσης επιλέγονται με επανατοποθέτηση (re-substitution), το οποίο σημαίνει ότι τα ίδια πρότυπα μπορούν να επιλεχθούν παραπάνω από μια φορές. Αν τα δεδομένα είναι λίγα τότε η διαδικασία επαναλαμβάνεται για έναν συγκεκριμένο αριθμό επαναλήψεων, ο οποίος αναφέρεται ως εποχή (epoch) και η τελική επίδοση του συστήματος προκύπτει από την μέση επίδοση και την τυπική απόκλιση των επιδόσεων [60]. Στην συγκεκριμένη εργασία η μέθοδος Bootstrap επαναλήφθηκε για 50 epochs, όπου κάθε φορά επιλεγόταν τυχαία ένα 80% των δεδομένων για εκπαίδευση και μέσω της μεθόδου k-fold χωριζόταν σε 3-folds. Για κάθε αλγόριθμο η διαδικασία της ταξινόμησης επαναλήφθηκε 10 φορές και κάθε φορά σημειωνόταν η ακρίβεια του ταξινομητή. Η τελική ακρίβεια ταξινόμησης προέκυψε από την μέση τιμή των τιμών ταξινόμησης μετά και από τις 10 επαναλήψεις. Οι τιμές των παραμέτρων των ταξινομητών που επιλέχθηκαν ήταν οι προκαθορισμένες από το πακέτο του λογισμικού. Οι αλγόριθμοι ταξινόμησης που χρησιμοποιηθήκαν ήταν οι εξής:

- Ταξινομητής ελάχιστης απόστασης (MDC) με την συνάρτηση sklearn.neighbors.NearestCentroid.
- Ταξινομητής πλησιέστερων γειτόνων (KNN) με την συνάρτηση sklearn.neighbors.KNeighborsClassifier για k=3.
- Απλοϊκός Bayes ταξινομητής (Naive Bayessian) με την συνάρτηση sklearn.naive_bayes.GaussianNB.
- Ταξινομητής ανάλυσης γραμμικής διάκρισης (LDA) με την συνάρτηση sklearn.discriminant_analysis.LinearDiscriminantAnalysis.
- Ταξινομητής λογιστικής παλινδρόμησης (Logistic Regression) με την συνάρτηση sklearn.linear_model.LogisticRegression.
- Ταξινομητής Perceptron με την συνάρτηση sklearn.linear_model.Perceptron και με τις παραμέτρους tol=1e-3, random_state=0.
- Ταξινομητής Perceptron πολλαπλών επιπέδων (MLP) με την συνάρτηση sklearn.neural_network.MLPClassifier και με τις παραμέτρους solver='lbfgs', alpha=1e-5, max_iter=10**30, hidden_layer_sizes=(nFeats*2, 2), random_state=1, όπου nFeats είναι ο αριθμός των χαρακτηριστικών.
- Ταξινομητής μηχανών στήριξης διανυσμάτων (SVM) με την συνάρτηση sklearn.svm.LinearSVC.

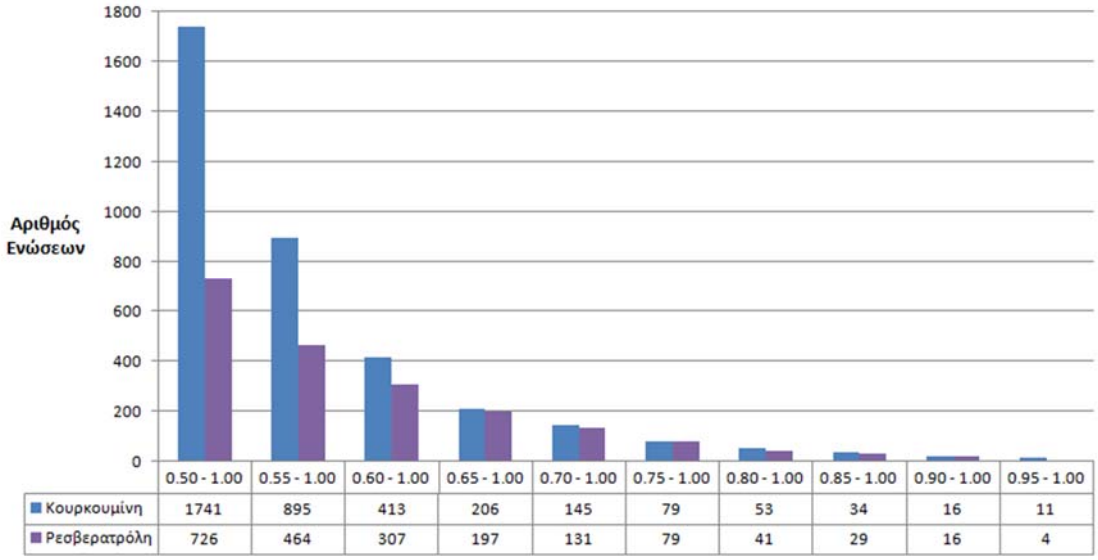
- Ταξινομητής τυχαίου δάσους (Random Forest) με την συνάρτηση sklearn.ensemble.RandomForestClassifier και με την παράμετρο n_estimators=10.

- Ταξινομητής δένδρων απόφασης (Decision Trees) με την συνάρτηση sklearn.tree.DecisionTreeClassifier.

Κεφάλαιο 3: Αποτελέσματα

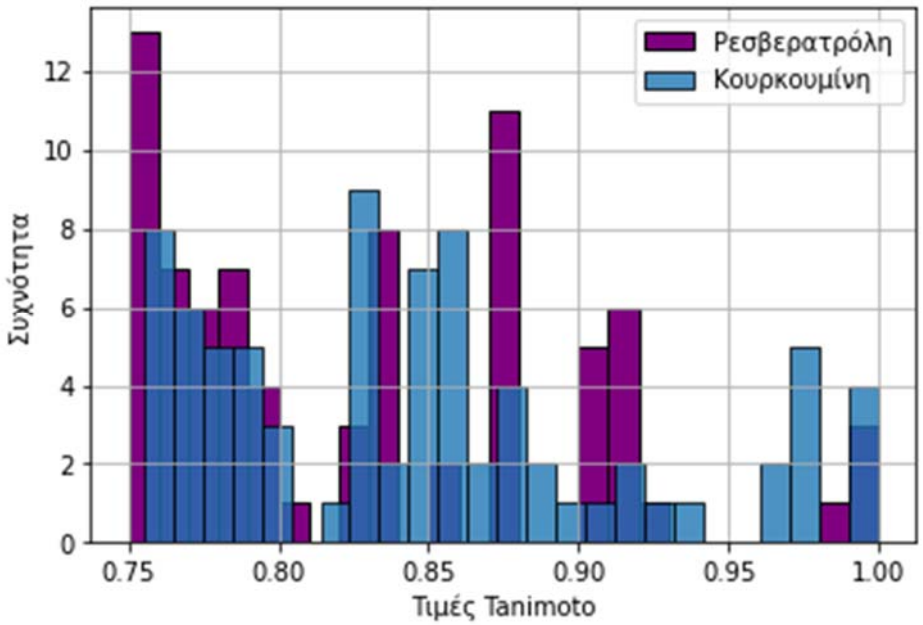
3.1 Αποτελέσματα επιλογής δεδομένων

Παρακάτω παραθέτονται τα αποτελέσματα από την αναζήτηση ομοιότητας Tanimoto η οποία πραγματοποιήθηκε μεταξύ της κουρκουμίνης και όλων των ενώσεων που βρίσκονται στη βάση δεδομένων Cnatural, καθώς και μεταξύ της ρεσβερατρόλης και όλων των ενώσεων της βάσης δεδομένων. Ο έλεγχος πραγματοποιήθηκε για τιμές από Tc ≥ 0,5 και για μεγαλύτερες τιμές. Στην Εικόνα 3.1 φαίνεται το πλήθος των ενώσεων που η χημική τους δομή έμοιαζε με την κουρκουμίνη και την ρεσβερατρόλη για διάφορες τιμές του συντελεστή Tanimoto. Πιο συγκεκριμένα, ο οριζόντιος άξονας περιέχει τα διαστήματα των τιμών Tanimoto για τα οποία έγινε η αναζήτηση ομοιότητας στην βάση δεδομένων Cnatural, ξεκινώντας από την τιμή 0,5. Κάτω από τις τιμές των διαστημάτων των τιμών Tanimoto, παρουσιάζεται το πλήθος των ενώσεων που βρέθηκαν ότι είναι όμοιες με την κουρκουμίνη και την ρεσβερατρόλη στο εκάστοτε διάστημα. Για παράδειγμα, στην αναζήτηση ομοιότητας όπου η τιμή του συντελεστή ομοιότητας είναι μεγαλύτερη ή ίση με 0,85, δηλαδή στο διάστημα 0,85 – 1,00, βρέθηκαν 34 ενώσεις με χημική δομή παρόμοια με την δομή της κουρκουμίνης και 29 ενώσεις με χημική δομή παρόμοια με την δομή της ρεσβερατρόλης. Στην Εικόνα 3.2 παρουσιάζεται το ιστόγραμμα συχνοτήτων των ενώσεων με συντελεστή Tanimoto μεγαλύτερο ή ίσο από 0,75, οι οποίες ήταν τελικά αυτές που επιλέχθηκαν για την δημιουργία των συνόλων δεδομένων. Στην Εικόνα 3.3 απεικονίζεται ο τριγωνικός χάρτης συσχέτισης (triangular correlation heatmap), ο οποίος δείχνει την συσχέτιση που παρουσιάζουν όλοι οι μοριακοί περιγραφείς μεταξύ τους σύμφωνα με τον συντελεστή συσχέτισης Pearson (Pearson’s correlation coefficient). Η σχέση (1.11) στην ενότητα 1.3.4 περιγράφει το πώς υπολογίζεται η τιμή του συντελεστή συσχέτισης Pearson.

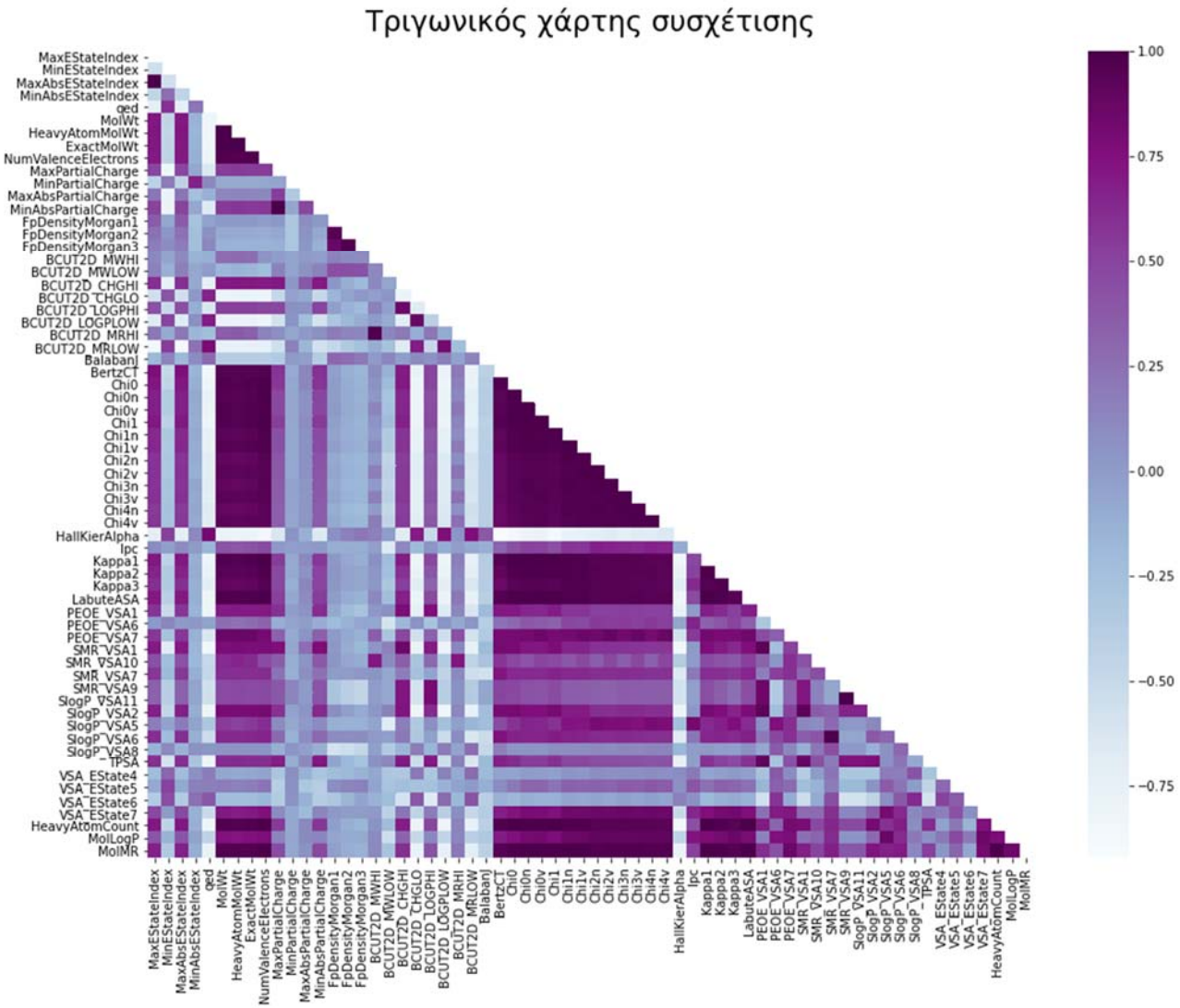


	0.50 - 1.00	0.55 - 1.00	0.60 - 1.00	0.65 - 1.00	0.70 - 1.00	0.75 - 1.00	0.80 - 1.00	0.85 - 1.00	0.90 - 1.00	0.95 - 1.00
Κουρκουμίνη	1741	895	413	206	145	79	53	34	16	11
Ρεσβερατρόλη	726	464	307	197	131	79	41	29	16	4

Εικόνα 3.1: *Πλήθος όμοιων ενώσεων με την κουρκουμίνη και την ρεσβερατρόλη σύμφωνα με τον συντελεστή Tanimoto.*



Εικόνα 3.2: *Ιστόγραμμα συχνοτήτων για τις ενώσεις με Tc ≥ 0,75.*



Εικόνα 3.3: *Τριγωνικός χάρτης συσχέτισης των μοριακών περιγραφέων σύμφωνα με τον συντελεστή συσχέτισης Pearson.*

3.2 Αποτελέσματα Μηχανικής Μάθησης

Στους πίνακες που παρουσιάζονται στην συνέχεια αναφέρεται η επί τοις εκατό ακρίβεια ταξινόμησης του κάθε ταξινομητή και η τυπική απόκλιση της ακρίβειάς του, για την περίπτωση στην οποία χρησιμοποιήθηκε ο συνδυασμός των 64 μοριακών περιγραφέων (Πίνακας 3.1) αλλά και για τον συνδυασμό των 26 βέλτιστων μοριακών περιγραφέων (Πίνακας 3.2). Να σημειωθεί εδώ ότι η διαδικασία της ταξινόμησης πραγματοποιήθηκε 10 φορές για τον κάθε ταξινομητή, οπότε οι τιμές που παρουσιάζονται αφορούν την μέση τιμή της ακρίβειας του κάθε ταξινομητή και για αυτό τον λόγο παρατίθεται και η τυπική απόκλιση της ακρίβειας. Επιπλέον, στον Πίνακα 3.3 και στον Πίνακα 3.4 παρουσιάζονται η μέση τιμή, το εύρος των τιμών αλλά και η τυπική απόκλιση της ακρίβειας του ταξινομητή Random Forest για όλους τους συνδυασμούς περιγραφέων που αναφέρθηκαν στην ενότητα 2.5.

Πίνακας 3.1: *Επί τοις εκατό ακρίβεια και τυπική για τον συνδυασμό όλων των μοριακών περιγραφέων για κάθε ταξινομητή.*

Ταξινομητές	% Ακρίβεια για 64 Περιγραφείς	Τυπική Απόκλιση
MDC	97.187	0.220
KNN	98.110	0.189
Bayesian	97.899	0.339
LDA	98.453	0.227
LogReg	88.429	1.756
PERCEPTRON	99.096	0.137
MLP	69.397	2.676
SVM	99.475	0.072
RF	**99.864**	**0.064**
CART	99.638	0.103

Πίνακας 3.2: *Επί τοις εκατό ακρίβεια και τυπική απόκλιση για τον συνδυασμό όλων των βέλτιστων μοριακών περιγραφέων για κάθε ταξινομητή.*

Ταξινομητές	% Ακρίβεια για 26 Περιγραφείς	Τυπική Απόκλιση
MDC	98.635	0.221
KNN	98.728	0.115
Bayesian	98.242	0.191
LDA	99.534	0.105
LogReg	87.819	1.018
PERCEPTRON	99.261	0.121
MLP	49.885	0.696
SVM	99.581	0.096
RF	**99.915**	**0.049**
CART	99.590	0.078

Πίνακας 3.3: *Μέση τιμή, εύρος και τυπική απόκλιση της ακρίβειας για συνδυασμούς όλων των μοριακών περιγραφέων με βάση τον ταξινομητή Random Forest.*

Συνδυασμοί Περιγραφέων (πλήθος τιμών)	Μέση τιμή Ακρίβειας	Εύρος Ακρίβειας	Τυπική Απόκλιση
Ανά 63 (64)	99.869	99.712 – 99.984	0.063
Ανά 62 (2016)	99.864	99.568 – 100	0.070
Ανά 61 (41664)	99.863	99.408 – 100	0.071
Ανά 60 (635376)	99.860	98.747 – 100	0.092

Πίνακας 3.4: *Μέση τιμή, εύρος και τυπική απόκλιση της ακρίβειας για συνδυασμούς των βέλτιστων μοριακών περιγραφέων με βάση τον ταξινομητή Random Forest.*

Συνδυασμοί Περιγραφέων (πλήθος τιμών)	Μέση τιμή Ακρίβειας	Εύρος Ακρίβειας	Τυπική Απόκλιση
Ανά 25 (26)	99.885	99.696 – 99.984	0.069
Ανά 24 (325)	99.883	99.648 – 100	0.069
Ανά 23 (2600)	99.873	99.312 – 100	0.088
Ανά 22 (14950)	99.864	98.928 – 100	0.099
Ανά 21 (65780)	99.853	98.912 – 100	0.113
Ανά 20 (230230)	99.841	98.240 – 100	0.139
Ανά 19 (657800)	99.827	98.320 – 100	0.145

Κεφάλαιο 4: Συζήτηση – Μελλοντικές Προοπτικές

4.1 Σχολιασμός Αποτελεσμάτων

Κατά την διαδικασία επιλογής δεδομένων από την βάση δεδομένων Cnatural με σκοπό την δημιουργία των 2 συνόλων δεδομένων, ήταν επιθυμητό να επιλεχθούν ενώσεις που θα είχαν όσο το δυνατόν υψηλότερο συντελεστή ομοιότητας Tanimoto με τις ενώσεις αναφοράς, την κουρκουμίνη και την ρεσβερατρόλη. Επιπλέον, ήταν ζητούμενο το κατώφλι της τιμής του συντελεστή Tanimoto να είναι κοινό και για τις 2 ομάδες ενώσεων, όπως επίσης και ο αριθμός των ενώσεων που θα περιέχει το κάθε σύνολο δεδομένων να είναι ίδιος ή παραπλήσιος, έτσι ώστε να αποφευχθεί η ανομοιογένεια. Όπως φαίνεται στην Εικόνα 3.1 οι παραπάνω προϋποθέσεις ικανοποιούνται όταν η τιμή του συντελεστή Tanimoto είναι ίση ή μεγαλύτερη από 0,75. Σε αυτή την περίπτωση προέκυψαν 79 ενώσεις με χημική δομή παρόμοια με την δομή της κουρκουμίνης και 79 ενώσεις με χημική δομή παρόμοια με την δομή της ρεσβερατρόλης. Στην Εικόνα 3.2 παρουσιάζεται η κατανομή των ενώσεων που χρησιμοποιήθηκαν για την δημιουργία των συνόλων δεδομένων αξιοποιώντας το ιστόγραμμα συχνοτήτων με κριτήριο την τιμή Tanimoto. Παρατηρείται ότι οι περισσότερες ενώσεις και για τις δύο ομάδες βρίσκονται στη περιοχή όπου η τιμή Tanimoto κυμαίνεται από 0,75 έως 0,90. Στο διάστημα 0,90 με 0,95 φαίνεται οι ενώσεις που ανήκουν στην ομάδα της ρεσβερατρόλης να είναι παραπάνω από διπλάσιες από τις ενώσεις που βρίσκονται στην ομάδα της κουρκουμίνης, ενώ για το διάστημα 0,95 με 1,00 συμβαίνει ακριβώς το αντίθετο. Στην Εικόνα 3.3 φαίνεται η συσχέτιση που έχουν οι μοριακοί περιγραφείς μεταξύ τους. Σε αυτό το γράφημα το μπλε χρώμα υποδηλώνει ότι δεν υπάρχει συσχέτιση μεταξύ των συγκεκριμένων περιγραφέων. Το σκούρο μωβ και το λευκό χρώμα αντιστοιχούν στις τιμές 1 και -1 του συντελεστή Pearson και δηλώνουν την ύπαρξη θετικής και αρνητικής συσχέτισης αντίστοιχα. Η θετική συσχέτιση δείχνει ότι όσο αυξάνονται οι τιμές των μεταβλητών στον x άξονα, αυξάνονται αναλογικά και στον y άξονα. Αντίστροφα η αρνητική συσχέτιση δείχνει ότι όσο αυξάνονται οι τιμές των μεταβλητών στον x άξονα, μειώνονται αντίστοιχα στον y άξονα.

Το μέρος της εργασίας που αφορά την Μηχανική Μάθηση, όπως αναφέρθηκε και στην ενότητα 2.5, χωρίστηκε σε 2 περιπτώσεις. Στη πρώτη περίπτωση χρησιμοποιήθηκαν όλοι οι 64 μοριακοί περιγραφείς για να γίνει η εκπαίδευση του συστήματος και η ταξινόμηση και στην συνέχεια όλοι οι πιθανοί συνδυασμοί ανά 63, 62, 61 και 60 περιγραφείς. Ήταν σημαντικό να χρησιμοποιηθούν όσο το δυνατόν περισσότεροι μοριακοί περιγραφείς, διότι με αυτόν τον τρόπο θα λαμβανόταν υπόψη περισσότερη πληροφορία σχετικά με τις ιδιότητες της κάθε ένωσης για την ταξινόμηση. Ο χρόνος υπολογισμού των αποτελεσμάτων του κάθε ταξινομητή όταν αξιοποιήθηκε ο συνδυασμός όλων των μοριακών περιγραφέων ήταν περίπου 15-30 λεπτά. Ενώ ο χρόνος υπολογισμού των αποτελεσμάτων για τους συνδυασμούς των μοριακών περιγραφέων διέφερε ανάλογα με το πλήθος των συνδυασμών από 1 έως 3 μέρες. Η δεύτερη στήλη του Πίνακα 3.1 δείχνει το ποσοστό της ακρίβειας του κάθε ταξινομητή, όταν για είσοδο του συστήματος χρησιμοποιήθηκε ο συνδυασμός όλων

των μοριακών περιγραφέων. Ο ταξινομητής τυχαίου δάσους (Random Forest) φαίνεται να έχει την καλύτερη επίδοση με ποσοστό ακρίβειας 99,86%. Αμέσως μετά ακολουθεί ο ταξινομητής CART με ποσοστό ακρίβειας 99.63%, ενώ ο ταξινομητής Perceptron πολλαπλών επιπέδων (MLP) φαίνεται να παρουσιάζει την χαμηλότερη επίδοση με ποσοστό ακρίβειας 69,39%. Στην τρίτη στήλη του Πίνακα 3.1 φαίνεται και η τυπική απόκλιση της ακρίβειας του κάθε ταξινομητή, καθώς η διαδικασία της ταξινόμησης πραγματοποιήθηκες 10 φορές για τον κάθε ταξινομητή. Όπως φαίνεται ο ταξινομητής Random Forest έχει την μικρότερη τυπική απόκλιση από όλους τους υπόλοιπους ταξινομητές. Στον Πίνακα 3.3 παρουσιάζονται οι τιμές της μέσης τιμής, του εύρους και της τυπικής απόκλισης της ακρίβειας για του συνδυασμούς των μοριακών περιγραφέων, όπως αναφέρθηκαν παραπάνω, που αφορούν τον ταξινομητή τυχαίου δάσους (Random Forest), ο οποίος σημείωσε την υψηλότερη επίδοση. Παρατηρείται ότι καθώς ελαττώνεται ο αριθμός των μοριακών περιγραφέων που χρησιμοποιούνται στο σύστημα, μειώνεται και η τελική ακρίβεια του συστήματος. Επίσης με την μείωση των μοριακών περιγραφέων και κατά συνέπεια την αύξηση των πιθανών συνδυασμών, παρατηρείται ότι αυξάνει το εύρος των τιμών της ακρίβειας και η τυπική απόκλιση.

Στη δεύτερη περίπτωση, μέσω κατάλληλου αλγορίθμου επιλέχθηκαν 26 βέλτιστοι μοριακοί περιγραφείς. Η επιλογή του πλήθους των 26 βέλτιστων μοριακών περιγραφέων έγινε με βάση έναν εμπειρικό κανόνα ελάττωσης χαρακτηριστικών. Ο οποίος αναφέρει ότι ο μέγιστος αριθμός χαρακτηριστικών σε έναν συνδυασμό πρέπει να είναι μικρότερος ή ίσος από το 1/3 του αριθμού των προτύπων της μικρότερης αριθμητικά κατηγορίας. Έτσι αποφεύγεται το φαινόμενο της υπερπροσαρμογής [60]. Στην συγκεκριμένη περίπτωση η μικρότερη αριθμητικά κατηγορία ήταν το σύνολο δεδομένων της ρεσβερατρόλης, που περιείχε 78 ενώσεις. Ο χρόνος υπολογισμού των αποτελεσμάτων του κάθε ταξινομητή όταν αξιοποιήθηκε ο συνδυασμός των βέλτιστων μοριακών περιγραφέων ήταν περίπου 10-20 λεπτά. Ενώ ο χρόνος υπολογισμού των αποτελεσμάτων για τους συνδυασμούς των μοριακών περιγραφέων διέφερε ανάλογα με το πλήθος των συνδυασμών από 1 έως 3 μέρες. Επίσης όπως φαίνεται και από το γράφημα που απεικονίζεται στην Εικόνα 3.3 οι βέλτιστοι περιγραφείς που επιλέχθηκαν από τον αλγόριθμο RFE έχουν αρκετά μικρή συσχέτιση μεταξύ τους. Ο σκοπός ήταν να χρησιμοποιηθούν όσο το δυνατόν περισσότεροι μοριακοί περιγραφείς, για αυτόν το λόγο επιλέχθηκαν 26 περιγραφείς στην αρχή και στην συνέχεια συνδυασμοί αυτών. Αρχικά χρησιμοποιήθηκαν για την εκπαίδευση και αξιολόγηση του συστήματος όλοι οι βέλτιστοι περιγραφείς και στην συνέχεια για τον ταξινομητή που σημείωσε την καλύτερη επίδοση αξιοποιήθηκαν όλοι οι πιθανοί συνδυασμοί ανά 25, 24, 23, 22, 21, 20 και 19 περιγραφείς. Στην δεύτερη στήλη του Πίνακα 3.2 φαίνεται το ποσοστό της ακρίβειας για τον κάθε ταξινομητή, όταν σαν είσοδος του συστήματος χρησιμοποιείται ο συνδυασμός των 26 βέλτιστων χαρακτηριστικών. Όπως και στην πρώτη περίπτωση, ο ταξινομητής με την μεγαλύτερη ακρίβεια ήταν ο ταξινομητής τυχαίου δάσους (Random Forest) με ποσοστό 99,91%. Την αμέσως καλύτερη επίδοση μετά από τον ταξινομητή Random Forest σημείωσε ο ταξινομητής CART με ποσοστό ακρίβειας 99.59%, ενώ την χειρότερη επίδοση σημείωσε πάλι ο ταξινομητής Perceptron πολλαπλών επιπέδων (MLP) με ποσοστό ακρίβειας 49,88%. Η τρίτη στήλη του Πίνακα 3.2 παρουσιάζει την τυπική απόκλιση της ακρίβειας του κάθε ταξινομητή, καθώς η διαδικασία της ταξινόμησης πραγματοποιήθηκες 10 φορές για τον κάθε ταξινομητή, όπως και στην

πρώτη περίπτωση. Όπως φαίνεται ο ταξινομητής Random Forest έχει την μικρότερη τιμή τυπικής απόκλισης από τους υπόλοιπους ταξινομητές. Στον Πίνακα 3.4 παρουσιάζονται η μέση τιμή, το εύρος και η τυπική απόκλιση της ακρίβειας του κάθε συνδυασμού για τον ταξινομητή Random Forest, ο οποίος σημείωσε την καλύτερη επίδοση. Παρόμοια με το προηγούμενο στάδιο, παρατηρείται μείωση της ακρίβειας με την μείωση του αριθμού των μοριακών περιγραφέων και αύξηση του εύρους και της τυπικής απόκλισης.

Συγκρίνοντας τα δύο στάδια μεταξύ τους και τα αποτελέσματα που φαίνονται στον Πίνακα 3.1 και στον Πίνακα 3.2, παρατηρείται ότι οι περισσότεροι ταξινομητές έχουν καλύτερη επίδοση όταν έχουν ως είσοδο τον συνδυασμό με τους 26 βέλτιστους μοριακούς περιγραφείς. Όμως οι διαφορές είναι αρκετά μικρές, εκτός από την περίπτωση του MLP όπου η διαφορά είναι κοντά στο 20%, γεγονός που πιθανώς οφείλεται στην λανθασμένη ρύθμιση των παραμέτρων του συγκεκριμένου ταξινομητή για τα συγκεκριμένα δεδομένα. Συνεπώς, κρίνεται απαραίτητο να γίνει ελάττωση χαρακτηριστικών και επιλογή βέλτιστων χαρακτηριστικών ώστε το σύστημα ταξινόμησης να επιτύχει όσο το δυνατόν καλύτερη ακρίβεια. Επιπλέον με την μείωση των χαρακτηριστικών αποφεύγεται και το φαινόμενο της υπερπροσαρμογής (overfitting), κατά το οποίο το σύστημα ταξινόμησης κάνει υπερεκτίμηση της ταξινομικής του ακρίβειας. Από τα παραπάνω μπορούν να εξαχθούν τα εξής συμπεράσματα:

- Η αξιοποίηση όσο το δυνατόν περισσότερων μοριακών περιγραφέων παρέχει αρκετά ικανοποιητικά αποτελέσματα και εμπεριέχει περισσότερη φυσικοχημική πληροφορία.

- Οι ταξινομητές που βασίζονται σε δένδρα απόφασης, όπως ο Random Forest και ο CART, σημείωσαν τις καλύτερες επιδόσεις, συνεπώς μπορούν να θεωρηθούν ως οι καταλληλότεροι ταξινομητές για τον διαχωρισμό των ενώσεων.

- Κρίνεται απαραίτητη η επιλογή βέλτιστων μοριακών περιγραφέων, καθώς με αυτόν τον τρόπο επιτυγχάνεται μεγαλύτερη ακρίβεια στην ταξινόμηση και αποφεύγεται το φαινόμενο της υπερπροσαρμογής.

Γενικά τα αποτελέσματα της ταξινόμησης που προέκυψαν ήταν πολύ υψηλά, γεγονός που οφείλεται αφενός στην αρκετά διαφορετική χημική δομή που έχουν μεταξύ τους οι δύο ενώσεις αναφοράς, κουρκουμίνη και ρεσβερατρόλη, και αφετέρου στον μικρό αριθμό δειγμάτων και μοριακών περιγραφέων. Για αυτό τον λόγο δεν διερευνήθηκε περαιτέρω η βελτιστοποίηση των τιμών των παραμέτρων του κάθε ταξινομητή. Για πιο αντιπροσωπευτικά αποτελέσματα κρίνεται αναγκαίο να χρησιμοποιηθούν περισσότερες χημικές ενώσεις και μεγαλύτερος αριθμός μοριακών περιγραφέων.

4.2 Μελλοντικές Προοπτικές

Σκοπός της παρούσας εργασίας ήταν να μελετηθεί σε ένα αρχικό στάδιο το κατά πόσο διάφορες μέθοδοι Χημειοπληροφορικής και ειδικότερα Μηχανικής Μάθησης μπορούν να διαχωρίσουν χημικές ενώσεις και να τις κατηγοριοποιήσουν ανάμεσα σε δύο κλάσεις, αξιοποιώντας τις φυσικοχημικές τους ιδιότητες. Μια σημαντική επέκταση αυτής της εργασίας είναι η ανάπτυξη ενός συστήματος Μηχανικής Μάθησης, το οποίο θα μπορεί να ταξινομεί χημικές ενώσεις με άγνωστες ή ελλιπείς φυσικοχημικές ιδιότητες σε περισσότερες από δύο κατηγορίες. Επίσης, πέραν των φυσικοχημικών ιδιοτήτων, θα μπορούσε να συμπεριληφθεί ποσοτική πληροφορία σχετικά με την βιολογική δράση ενώσεων έναντι κυτταρικών σειρών ή βιολογικών στόχων. Με αυτόν τον τρόπο θα μπορούσε να δημιουργηθεί ένας αλγόριθμος κατηγοριοποίησης ή και πρόβλεψης βιολογικής δράσης για ενώσεις οι οποίες δεν έχουν δοκιμαστεί σε *in vitro* πειράματα. Βέβαια, αυτό προϋποθέτει την αξιοποίηση περισσότερων ενώσεων, για την δημιουργία συνόλων δεδομένων με καταγεγραμμένες βιολογικές ιδιότητες και την χρήση μεγαλύτερου αριθμού μοριακών περιγραφέων.

Αναφορές - Πηγές

[1] (2007), "All natural", *Nat Cem Biol*, **3**, pp 351. https://doi.org/10.1038/nchembio0707-351

[2] Cragg GM, Newman DJ, (2013), "Natural products: a continuing source of novel drug leads", *Biochim Biophys Acta,* **1830**, **(6)**, pp 3670-3695. https://doi.org/10.1016/j.bbagen.2013.02.008

[3] Atanasov AG, Waltenberger B, Pferschy-Wenzig EM, et al., (2015), "Discovery and resupply of pharmacologically active plant-derived natural products: A review", *Biotechnol Adv*, **33**, **(8)**, pp 1582-1614. https://doi.org/10.1016/j.biotechadv.2015.08.001

[4] Rossiter SE, Fletcher MH, Wuest WM, (2017), "Natural Products as Platforms To Overcome Antibiotic Resistance", *Chem Rev*, **117**, **(19)**, pp 12415–12474. https://doi.org/10.1021/acs.chemrev.7b00283.

[5] Sorokina M, Steinbeck C, (2020), "Review on natural products databases: where to find data in 2020", *J Cheminform*, **12**, **(1)**, pp 20. https://doi.org/10.1186/s13321-020-00424-9

[6] Newman DJ, Cragg GM, (2016), "Natural Products as Sources of New Drugs from 1981 to 2014", *J Nat Prod*, **79**, **(3)**, pp 629–661. https://doi.org/10.1021/acs.jnatprod.5b01055

[7] Kunnumakkara AB, Bordoloi D, Padmavathi G, et al., (2017), "Curcumin, the golden nutraceutical: multitargeting for multiple chronic diseases", *Br J Pharmacol*, **174**, **(11)**, pp 1325–1348 https://doi.org/10.1111/bph.13621

[8] Hewlings SJ, Kalman DS, (2017), "Curcumin: A Review of Its Effects on Human Health", *Foods*, **6**, **(10)**, pp 92 https://doi.org/10.3390/foods6100092

[9] Aggarwal BB, Harikumar KB, (2009), "Potential therapeutic effects of curcumin, the anti-inflammatory agent, against neurodegenerative, cardiovascular, pulmonary, metabolic, autoimmune and neoplastic diseases", I*nt J Biochem Cell Biol*, **41**, **(1)**, pp 40–59 https://doi.org/10.1016/j.biocel.2008.06.010

[10] Vallianou NG, Evangelopoulos A, Schizas N, Kazazis C, (2015), "Potential anticancer properties and mechanisms of action of curcumin", *Anticancer Res*, **35**, **(2)**, pp 645–651

[11] Mbese Z, Khwaza V, Aderibigbe BA, (2019), "Curcumin and Its Derivatives as Potential Therapeutic Agents in Prostate, Colon and Breast Cancers", *Molecules*, **24**, **(23)**, pp 4386. https://doi.org/10.3390/molecules24234386

[12] Alrawaiq NS, Abdullah A, (2014), "A review of antioxidant polyphenol curcumin and its role in detoxification", *Int J Pharmtech Res*, **6**, pp 280-289.

[13] Galiniak S, Aebisher D, Bartusik-Aebisher D, (2019), "Health benefits of resveratrol administration", *Acta Biochim Pol*, **66**, **(1)**, pp 13–21 https://doi.org/10.18388/abp.2018_2749

[14] Catalgol B, Batirel S, Taga Y, Ozer NK, (2012), "Resveratrol: French paradox revisited", *Front Pharmacol*, **3**, pp 141 https://doi.org/10.3389/fphar.2012.00141

[15] Keylor MH, Matsuura BS, Stephenson CR, (2015), "Chemistry and Biology of Resveratrol-Derived Natural Products", *Chem Rev*, **115**, **(17)**, pp 8976–9027 https://doi.org/10.1021/cr500689b

[16] Breuss JM, Atanasov AG, Uhrin P, (2019), "Resveratrol and Its Effects on the Vascular System", *Int J Mol Sci*, **20**, **(7)**, pp 1523 https://doi.org/10.3390/ijms20071523

[17] Bonnefont-Rousselot D, (2016), "Resveratrol and Cardiovascular Diseases", *Nutrients*, **8**, **(5)**, pp 250 https://doi.org/10.3390/nu8050250

[18] Li H, Xia N, Hasselwander S, Daiber A, (2019), "Resveratrol and Vascular Function", *Int J Mol Sci*, **20**, **(9)**, pp 2155 https://doi.org/10.3390/ijms20092155

[19] Shaito A, Posadino AM, Younes N, et al., (2020), "Potential Adverse Effects of Resveratrol: A Literature Review", *Int J Mol Sci*, **21**, **(6)**, pp 2084 https://doi.org/10.3390/ijms21062084

[20] Ramírez-Garza SL, Laveriano-Santos EP, Marhuenda-Muñoz M, et al., (2018), "Health Effects of Resveratrol: Results from Human Intervention Trials", *Nutrients*, **10**, **(12)**, pp 1892 https://doi.org/10.3390/nu10121892

[21] Ko JH, Sethi G, Um JY, et al., (2017), "The Role of Resveratrol in Cancer Therapy", *Int J Mol Sci*, **18**, **(12)**, pp 2589 https://doi.org/10.3390/ijms18122589

[22] Aggarwal BB, Bhardwaj A, Aggarwal RS, et al., (2004), "Role of resveratrol in prevention and therapy of cancer: preclinical and clinical studies", *Anticancer Res*, **24**, **(5A)**, pp 2783–2840

[23] Khan OS, Bhat AA, Krishnankutty R, et al., (2016), "Therapeutic Potential of Resveratrol in Lymphoid Malignancies", *Nutr Cancer*, **68**, **(3)**, pp 365–373 https://doi.org/10.1080/01635581.2016.1152386

[24] Gasteiger J, (2016), "Chemoinformatics: Achievements and Challenges, a Personal View", *Molecules*, **21**, **(2)**, pp 151 https://doi.org/10.3390/molecules21020151

[25] Gasteiger J, Engel T, editors, (2003), "Chemoinformatics–A Textbook", *Wiley-VCH Verlag GmbH & Co. KGaA.*, **ISBN 3-527-30681-1**

[26] Lo YC, Rensi SE, Torng W, Altman RB, (2018), "Machine learning in chemoinformatics and drug discovery", *Drug Discov Today*, **23**, **(8)**, pp 1538–1546 https://doi.org/10.1016/j.drudis.2018.05.010

[27] Grisoni F, Consonni V, Todeschini R, (2018), "Impact of Molecular Descriptors on Computational Models", *Methods Mol Biol*, **1825**, pp 171–209 https://doi.org/10.1007/978-1-4939-8639-2_5

[28] Schneider M, Pons JL, Labesse G, Bourguet W, (2019), "In Silico Predictions of Endocrine Disruptors Properties", *Endocrinology*, **160**, **(11)**, pp 2709–2716 https://doi.org/10.1210/en.2019-00382

[29] Bajusz D, Rácz A, Héberger K, (2017), "Chemical Data Formats, Fingerprints, and Other Molecular Descriptions for Database Analysis and Searching", In: Chackalamannil S, Rotella DP, Ward SE (eds) Comprehensive medicinal chemistry III, Elsevier, Oxford, pp 329–378 https://doi.org/10.1016/B978-0-12-409547-2.12345-5

[30] Cereto-Massagué A, Ojeda MJ, Valls C, et al., (2015), "Molecular fingerprint similarity search in virtual screening", *Methods*, **71**, pp 58–63 https://doi.org/10.1016/j.ymeth.2014.08.005

[31] Gao K, Nguyen DD, Sresht V, et al., (2020), "Are 2D fingerprints still valuable for drug discovery?", Phys *Chem Chem Phys*, **22**, **(16)**, pp 8373–8390 https://doi.org/10.1039/d0cp00305k

[32] Durant JL, Leland BA, Henry DR, Nourse JG, (2002), "Reoptimization of MDL keys for use in drug discovery", *J Chem Inf Comput Sci*, **42, (6)**, pp 1273–1280 https://doi.org/10.1021/ci010132r

[33] Bolton EE, Wang Y, Thiessen PA, Bryant SH, (2008), "PubChem: Integrated Platform of Small Molecules and Biological Activities", *Annu Rep Comput Chem*, **4**, pp 217–241 http://dx.doi.org/10.1016/S1574-1400(08)00012-1

[34] Barnard JM, Downs GM, (1997), "Chemical Fragment Generation and Clustering Software§", *J Chem Inf Comput Sci*, **37**, **(1)**, pp 141–142 https://doi.org/10.1021/ci960090k

[35] O'Boyle NM, Banck M, James CA, et al., (2011), "Open Babel: An open chemical toolbox", *J Cheminform*, **3**, pp 33 https://doi.org/10.1186/1758-2946-3-33

[36] I. Daylight Chemical Information Systems, *Daylight*, http://www.daylight.com/ (ανακτήθηκε 14/5/2021)

[37] Hall LH, Kier LB, (1995), "Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information", *J Chem Inf Comput Sci*, **35**, **(6)**, pp 1039–1045 https://doi.org/10.1021/ci00028a014

[38] Rogers D, Hahn M, (2010), "Extended-Connectivity Fingerprints", *J Chem Inf Model*, **50**, **(5)**, pp 742–754 https://doi.org/10.1021/ci100050t

[39] McGregor MJ, Muskal SM, (1999), "Pharmacophore fingerprinting. 1. Application to QSAR and focused library design", *J Chem Inf Comput Sci*, **39**, **(3)**, pp 569–574 https://doi.org/10.1021/ci980159j

[40] Maggiora GM, Shanmugasundaram V, (2011), "Molecular similarity measures", *Methods Mol Biol*, **672**, pp 39–100 https://doi.org/10.1007/978-1-60761-839-3_2

[41] Vogt M, Bajorath J, (2020), "ccbmlib - a Python package for modeling Tanimoto similarity value distributions", *F1000Research*, **9**, Chem Inf Sci-100. https://doi.org/10.12688/f1000research.22292.2

[42] Tovar A, Eckert H, Bajorath J, (2007), "Comparison of 2D fingerprint methods for multiple-template similarity searching on compound activity classes of increasing structural diversity", *ChemMedChem*, **2**, **(2)**, pp 208–217 https://doi.org/10.1002/cmdc.200600225

[43] Maggiora G, Vogt M, Stumpfe D, Bajorath J, (2014), "Molecular similarity in medicinal chemistry", *J Med Chem*, **57**, **(8)**, pp 3186–3204 https://doi.org/10.1021/jm401411z

[44] GSI Technology, https://www.gsitechnology.com/Hardware-Accelerated-Search-for-Drug-Discovery (ανακτήθηκε 17/5/2021)

[45] Flower DR, (1998), "On the Properties of Bit String-Based Measures of Chemical Similarity", *J Chem Inf Comput Sci*, **38**, **(3)**, pp 379–386 https://doi.org/10.1021/ci970437z

[46] Weininger D, (1988), "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules", *J Chem Inf Comput Sci*, **28**, **(1)**, pp 31–36 https://doi.org/10.1021/ci00057a005

[47] Saldívar-González FI, Huerta-García CS, Medina-Franco JL, (2020), "Chemoinformatics-based enumeration of chemical libraries: a tutorial", *J Cheminform*, **12**, **(1)**, pp 64 https://doi.org/10.1186/s13321-020-00466-z

[48] M Karthikeyan, R Vyas, (2014), "Practical Chemoinformatics", *Springer India*, **ISBN 978-81-322-1780-0**

[49] PubChem, https://pubchem.ncbi.nlm.nih.gov/ (ανακτήθηκε 21/5/2021)

[50] ChEMBL, https://www.ebi.ac.uk/chembl/ (ανακτήθηκε 21/5/2021)

[51] ChemSpider, http://www.chemspider.com/ (ανακτήθηκε 21/5/2021)

[52] Deo RC, (2015), "Machine Learning in Medicine", *Circulation*, **132**, **(20)**, pp 1920–1930 https://doi.org/10.1161/CIRCULATIONAHA.115.001593

[53] Jayatilake S, Ganegoda GU, (2021), "Involvement of Machine Learning Tools in Healthcare Decision Making", *J Healthc Eng*, **2021**, 6679512 https://doi.org/10.1155/2021/6679512

[54] Sarker IH, (2021), "Machine Learning: Algorithms, Real-World Applications and Research Directions", *SN Comput Sci*, **2**, **(3)**, pp 160 https://doi.org/10.1007/s42979-021-00592-x

[55] Henkel T, Brunne RM, Muller H, Reichel F, (1999), "Statistical investigation into the structural complementarity of natural products and synthetic compounds", *Angew Chem Int Ed Engl*, **38**, pp 643-647

[56] Stahura FL, Godden JW, Xue L, Bajorath J,m(2000), "Distinguishing between Natural Products and Synthetic Molecules by Descriptor Shannon Entropy Analysis and Binary QSAR Calculations", *J Chem Inf Comput Sci*, **40**, **(5)**, pp 1245–1252 https://doi.org/10.1021/ci0003303

[57] Ertl P, Roggo S, Schuffenhauer A, (2008), "Natural Product-likeness Score and Its Application for Prioritization of Compound Libraries", *J Chem Inf Model*, **48**, **(1)**, pp 68–74 https://doi.org/10.1021/ci700286x

[58] Καλατζής Ι, (2021), "Αναγνώριση προτύπων με εφαρμογές σε ιατρικά απεικονιστικά συστήματα", Σημειώσεις Μαθήματος

[59] Uddin S, Khan A, Hossain ME, Moni MA, (2019), "Comparing different supervised machine learning algorithms for disease prediction", *BMC Med Inform Decis Mak*, **19**, **(1)**, pp 281 https://doi.org/10.1186/s12911-019-1004-8

[60] Κάβουρας Δ, (2019), "Μηχανική Μάθηση", Σημειώσεις Μαθήματος

[61] Shi W, Xue B, Guo S, et al., (2018), "Obstructive Sleep Apnea Detection Using Difference in Feature and Modified Minimum Distance Classifier", *Annu Int Conf IEEE Eng Med Biol Soc*, Annual International Conference, 2018, pp 1–4 https://doi.org/10.1109/EMBC.2018.8513093

[62] Santucci E, (2017), "Quantum Minimum Distance Classifier", *Entropy*, **19**, **(12)**, pp 659 doi:10.3390/e19120659

[63] Μήκος ΙΕ, (2016), "Συστήματα Υποστήριξης Διάγνωσης της Νόσου του Parkinson με Χρήση Φωνητικών Καταγραφών", Διπλωματική εργασία, https://pergamos.lib.uoa.gr/uoa/dl/frontend/file/lib/default/data/1320201/theFile

[64] Ghanou Y, Bencheikh G, (2016), "Architecture Optimization and Training for the Multilayer Perceptron using Ant System", *Int J Comput Sci*,**43**, **(1)**, pp 20–26

[65] Theodoridis S, Koutroubas K (2012), "Αναγνώριση Προτύπων, 4ήΕκδοση", *Εκδόσεις Π. Χ. Πασχαλίδης*, **ISBN 978-960-489-145-0**

[66] Cnatural, http://cnatural.gr

[67] https://www.psycopg.org/docs/

[68] RDKit, https://rdkit.org/docs/source/rdkit.html

[69] http://www.cadaster.eu/sites/cadaster.eu/files/challenge/descr.htm

[70] Sklearn, https://scikit-learn.org/stable/