ΑΛΕΞΑΝΔΡΕΙΟ ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΘΕΣΣΑΛΟΝΙΚΗΣ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΕΥΦΥΕΙΣ ΤΕΧΝΟΛΟΓΙΕΣ ΔΙΑΔΙΚΤΥΟΥ - WEB INTELLIGENCE

Ανάλυση βιολογικών δεδομένων με χρήση αλγορίθμων μηχανικής μάθησης με εφαρμογή στη διάγνωση του γαστρεντερικού καρκίνου

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΑΛΕΞΑΝΔΡΟΥ ΠΕΡΗΦΑΝΟΥ

Επιβλέπων: **Κωνσταντίνος Διαμαντάρας**
Καθηγητής, ΑΤΕΙ Θεσσαλονίκης

Θεσσαλονίκη, Φεβρουάριος 2019

Ευχαριστίες

Θέλω να ευχαριστήσω ιδιαιτέρως θερμά τον επιβλέποντα Καθηγητή μου, κ. Διαμαντάρα Κωνσταντίνο, καταρχάς για την δυνατότητα που μου έδωσε να ασχοληθώ με ένα τόσο σημαντικό και ενδιαφέρον θέμα. Κατά δεύτερον, για την εμπιστοσύνη που έδειξε στο πρόσωπό μου, την πολυτιμότατη καθοδήγηση, την υπομονή και φυσικά τον επαγγελματισμό του κατά τη διάρκεια εκπόνησης της διπλωματικής μου εργασίας. Επιπροσθέτως, θέλω να ευχαριστήσω την Επίκουρη Καθηγήτρια του Τμήματος Βιολογίας του Α.Π.Θ., κα. Ντάφου Δήμητρα και κ. Μώκο Παναγιώτη για την παροχή του δείγματος που χρησιμοποιήσαμε και φυσικά για τη συνολική βοήθεια τους. Ευχαριστώ επίσης το τμήμα Η/Υ του Α.τ.ε.ι.θ για τη διάθεση του απαραίτητου εξοπλισμού για την εκτέλεση των πειραμάτων.

Τέλος, θα ήθελα να ευχαριστήσω βαθύτατα την οικογένεια μου, πρωτίστως τους γονείς μου Χρήστο και Σοφία, και δευτερευόντως τα αδέρφια μου, Ιωάννη και Ιάσωνα για την τεράστια υποστήριξη και συνεισφορά τους κατά τη διάρκεια του κύκλου σπουδών μου.

Περίληψη

Η στόχευση της παρούσας διπλωματικής εργασίας είναι να εντοπιστεί ένας όσο το δυνατόν μικρότερος αριθμός γονιδίων, με χρήση Μεθόδων Μηχανικής Μάθησης, τα οποία περιέχουν σημαντική πληροφορία για την κατασκευή ενός ταξινομητή που θα έχει εξαιρετική απόδοση και φυσικά θα μπορεί να γενικεύει. Το σύνολο δεδομένων που εξετάσαμε αφορά στο γαστρεντερικό καρκίνο και προέρχεται από τη βάση δεδομένων του TCGA (The Cancer Genome Atlas). Η συγκεκριμένη μορφή καρκίνου που μελετιέται, έχει 5 είδη, αλλά εμείς θα ασχοληθούμε συγκεκριμένα με 4, τον οισοφαγικό, τον στομαχικό, τον παγκρεατικό και της χοληδόχου κύστης. Οι μεταβλητές-στόχοι είναι η ύπαρξη ή μη της ασθένειας. Οι τιμές των γονιδίων αποτελούν τιμές έκφρασης μετασχηματισμένες από αλληλλούχιση RNA (RNA-seq). Αρχικά, βρέθηκαν τα κοινά γονίδια για όλους τους τύπους καρκίνου προς ανάλυση. Μετέπειτα, με τη χρήση των μεθόδων Μείωσης Διαστάσεων/επιλογής χαρακτηριστικών Αμοιβαίας Πληροφορίας (Mutual Information), του κριτηρίου Kolmogorov-Smirnov για 2 δείγματα (KS 2Samples Test) και τέλος της Επαναλαμβανόμενης Εξάλειψης Χαρακτηριστικών με Διασταυρούμενη Επικύρωση (Recursive Feature Elimination with Cross Validation) αξιολογήσαμε τα γονίδια και τα ταξινομήσαμε με βάση τη σημαντικότητα τους. Για τις 2 πρώτες (σ.σ. Mutual Information, Kolmogorov Smirnov 2 Samples), επιλέξαμε τα πρώτα σημαντικά γονίδια της κάθε μεθόδου Επιλογής Χαρακτηριστικών ξεκινώντας από 10 μέχρι 5000. Η τελευταία (σ.σ. rfecv) εφαρμόζοντας βαρύτητες στα χαρακτηριστικά με βάση το εκάστοτε μοντέλο που χρησιμοποιεί, εξαλείφει κάθε φορά τα χειρότερα χαρακτηριστικά σύμφωνα με κάποιο βήμα. Αφού εντοπίστηκε το βέλτιστο υποσύνολο χαρακτηριστικών, κάναμε σύγκριση ταξινομητών και έπειτα εκτελέσαμε Εξαντλητική αναζήτηση (Grid Search) για την εύρεση των παραμέτρων που οι αποδοτικότεροι ταξινομητές πετυχαίνουν την καλύτερη τιμή μέσης ακρίβειας (k-Fold Cross Validation). Επιπροσθέτως, άλλες μέθοδοι χρησιμοποιήθηκαν όπως κανονικοποίηση των δεδομένων και δημιουργία συνθετικών δεδομένων για τη μειοψηφούσα κλάση(υγιείς) καθώς το δείγμα μας ήταν κατά πολύ μη ισορροπημένο. Τα αποτελέσματα των πειραμάτων έδειξαν ότι το κριτήριο rfecv υπερέχει των υπολοίπων κριτηρίων αξιολόγησης που εξετάσαμε αφού κατάφερε να εντοπίσει το μικρότερο αριθμό σημαντικών γονιδίων (χαρακτηριστικών) τα οποία περιέχουν σημαντική πληροφορία για την κατασκευή ενός ταξινομητή SVM RBF ο οποίος διαθέτει καλύτερη ικανότητα γενίκευσης έναντι άλλων υποσυνόλων σημαντικών γονιδίων που προήλθαν από τα υπόλοιπα κριτήρια αξιολόγησης που εξετάσαμε(η απόδοση του ταξινομητή δεν μειώθηκε παρόλο που χρησιμοποιήσαμε και μεθόδους δημιουργίας συνθετικών τιμών για τη μειοψηφούσα κλάση). Επιπλέον παρατηρήθηκε ότι η με κανονικοποίηση των γονιδιακών τιμών, πέτυχαμε τα καλύτερα αποτελέσματα.

Λέξεις-Κλειδιά: Μηχανική Μάθηση, Μέθοδοι Μείωσης Διαστάσεων, Μέθοδοι Επιλογής Χαρακτηριστικών, Mutual Information, Kolmogorov-Smirnov 2 Samples, rfecv, SVM, Γαστρεντερικός Καρκίνος, Οισοφαγικός Καρκίνος, Στομαχικός Καρκίνος, Παγκρεατικός Καρκίνος, Καρκίνος της Χοληδόχου Κύστης, Γονίδια, Αλληλλούχιση RNA (RNA-seq)

Abstract

The purpose of this thesis is to identify as few genes as possible, using Machine Learning methods, which contain important information for the construction of a classifier that will perform extremely well and will, of course, be able to generalize. The dataset that we examined, concerns gastrointestinal cancer and was taken from the TCGA database (The Cancer Genome Atlas). The specific form of cancer studied, has 5 species, but we will deal specifically with 4, esophageal, stomach, pancreatic and gallbladder. The target variables are the existence or absence of the disease. Gene expression data were transformed by RNA sequencing (RNA-Seq). Initially, common genes were found for all types of cancer to be analyzed. Subsequently, using the Dimension Reduction/Feature Selection Methods, Kolmogorov Smirnov 2 Samples Test (KS 2Samples Test), Mutual Information (MI) and Recursive Feature Elimination with Cross Validation (RFE-CV), we evaluated the genes and ranked them according to their significance. For the first two (Mutual Information, Kolmogorov Smirnov 2 Samples), we chose the first more important genes of each feature selection method starting from 10 to 5000. The latter (RFE-CV) refers to the classification of features according to the weights given by the respective model used, with repeated deletion of features in regards to a specific step and then selecting the best number of features through cross-validation. We compared classifiers and then we performed Grid Search to find the parameters at which the most efficient classifiers achieve the best results (k-Fold Cross Validation). In addition, other methods were used such as data standardization and generation of synthetic data for the minority class (healthy) as our sample was very unbalanced. The results of the experiments showed that the rfecv criterion is superior to the other evaluation criteria we tested as it was able to find the smaller number of significant genes containing important information for the construction of an SVM RBF classifier that has a better generalization capability than other subsets of important genes derived from the other evaluation criteria we tested (the classifier's performance did not decrease even after we used methods for synthetic data generation). In addition, it was observed that with standardization of gene values, we achieved the best results.

Keywords: Machine Learning, Dimension Reduction Methods, Feature Selection Methods, Mutual Information, Kolmogorov-Smirnov 2 Samples, rfecv, SVM, Gastrointestinal Cancer, Esophageal Cancer, Stomach Cancer, Pancreatic Cancer, Gallbladder Cancer, Genes, RNA Sequencing (RNA-Seq)

Πίνακας Περιεχομένων

1. Εισαγωγή

1.1 Μηχανική Μάθηση – Ορισμός και Βασικές Έννοιες

Η **Μηχανική Μάθηση (Machine Learning)** είναι ένας κλάδος της Τεχνητής Νοημοσύνης. Μπορεί να οριστεί ως οι υπολογιστικές μέθοδοι που χρησιμοποιούν εμπειρία για να βελτιώσουν την απόδοση ενός συστήματος ή να πραγματοποιήσουν ακριβείς προβλέψεις [1]. Η έννοια της **εμπειρίας (experience)** άπτεται στην πληροφορία του παρελθόντος η οποία διατίθεται στο σύστημα μάθησης και συνήθως παίρνει τη μορφή ηλεκτρονικών δεδομένων που συλλέχθηκαν και είναι διαθέσιμα για ανάλυση. Τα δεδομένα αυτά συνήθως έχουν τη μορφή συνόλων εκπαίδευσης που έχουν χαρακτηριστεί και ταξινομηθεί χειροκίνητα (από ανθρώπους). Σε κάθε περίπτωση, η ποιότητα και το μέγεθος του συνόλου αποτελούν έναν κρίσιμο παράγοντα για την επιτυχία των προβλέψεων του συστήματος.

Περιλαμβάνει τη σχεδίαση αποδοτικών αλγορίθμων για τη δημιουργία ακριβών προβλέψεων. Όπως και σε άλλα επιστημονικά πεδία, δύο σημαντικές μετρικές ποιότητας των αλγορίθμων αυτών αποτελούν η χρονική και χωρική πολυπλοκότητά τους. Αναφορικά με τη Μηχανική Μάθηση, εισάγεται μια νέα έννοια που αφορά στην **πολυπλοκότητα δείγματος (sample complexity)** και αναφέρεται στην αξιολόγηση του μεγέθους δείγματος που χρειάζεται ένας αλγόριθμος για να αποδώσει έχοντας «κατανοήσει» μια οικογένεια κλάσεων. Η αποτελεσματικότητά του εξαρτάται από την πολυπλοκότητα των κλάσεων και το μέγεθος του συνόλου εκπαίδευσης.

Καθώς η επιτυχία ενός αλγορίθμου Μηχανικής Μάθησης σχετίζεται άρρηκτα με τα δεδομένα που χρησιμοποιούμε, είναι ασφαλές να ειπωθεί πως, η ανάλυση δεδομένων και η στατιστική γενικότερα, είναι οι δυο σημαντικότεροι πυλώνες της. Ως εκ τούτου, οι τεχνικές εκμάθησης βασίζονται στα δεδομένα και συνδυάζουν ιδέες από στατιστική, πιθανότητες και τεχνικές βελτιστοποίησης [1].

1.1.1 Ορολογία

Θεωρώντας το ζήτημα της αυτόματης ταξινόμησης ανθρώπων σε υγιείς και ασθενείς ως προς κάποια ασθένεια, όπως στην περίπτωσή μας, του καρκίνου, γίνεται μια πρώτη αναφορά σε βασικές έννοιες [1].

- Δείγματα (Samples): Στιγμιότυπα δεδομένων που χρησιμοποιούνται για εκμάθηση ή αξιολόγηση. Για παράδειγμα, σε ένα πρόβλημα ταξινόμησης ως προς κάποια ασθένεια, τα δείγματα αντιστοιχούν σε ένα σύνολο ανθρώπων και τα δεδομένα τους, συνήθως γονιδιακές εκφράσεις που θα χρησιμοποιηθούν για εκμάθηση και φυσικά τον έλεγχο αποτελεσματικότητας του αλγορίθμου Ταξινόμησης Μηχανικής Μάθησης.

- Χαρακτηριστικά (Features): Είναι το σύνολο γνωρισμάτων που σχετίζεται με ένα δείγμα. Τα χαρακτηριστικά συνήθως σχηματίζουν ένα διάνυσμα. Στο παράδειγμα μας, τα χαρακτηριστικά είναι αποτέλεσμα εργαστηριακών εξετάσεων και πιο συγκεκριμένα εκφράσεις των γονιδίων τους.

- Κλάσεις (Classes): Αφορά στις κατηγορίες που έχουν ανατεθεί στα δεδομένα μας. Στην περίπτωση μας, έχουμε δύο κλάσεις, υγιείς και ασθενείς.

- Σύνολο Εκπαίδευσης (Training Set): Αφορά στα δείγματα που χρησιμοποιούνται για την εκπαίδευση ενός αλγορίθμου Μηχανικής Μάθησης. Οι κλάσεις των στιγμιότυπων αυτού του υποσυνόλου είναι γνωστές στον αλγόριθμο. Το μέγεθος του συνόλου μπορεί να διαφέρει, συνήθως αποτελεί το 70% του συνόλου.

- Σύνολο Ελέγχου (Test Set): Αφορά στα δείγματα που χρησιμοποιούνται για την αξιολόγηση ενός αλγορίθμου Μηχανικής Μάθησης. Αποτελεί το υποσύνολο για το οποίο ο αλγόριθμος πρέπει να προβλέψει τις κλάσεις για κάθε στιγμιότυπο. Στη συνέχεια, οι προβλέψεις συγκρίνονται με τις πραγματικές κλάσεις κάθε στιγμιότυπου για τη μέτρηση της απόδοσης του αλγορίθμου.

1.2 Μηχανική Μάθηση στην Ιατρική και τη Βιοπληροφορική

Οι Τεχνολογικές εξελίξεις των τελευταίων χρόνων όπως η Μηχανική Μάθηση, έχουν τεράστιες επιδράσεις σε κλάδους όπως η Ιατρική. Με τη βοήθεια της Μηχανικής Μάθησης, μπορούν να επιλυθούν διαγνωστικά και προγνωστικά προβλήματα που αφορούν σε ένα ευρύ φάσμα Ιατρικών πεδίων. Οι μέθοδοι Μηχανικής Μάθησης, χρησιμοποιούνται πλέον για την ανάλυση σημαντικότητας κλινικών παραμέτρων αναφορικά με τις προγνώσεις, π.χ. πρόγνωση της πορείας μιας ασθένειας, περισσότερη ιατρική γνώση και σχεδιασμό θεραπειών. Επιπροσθέτως, γίνεται χρήση για ανάλυση δεδομένων όπως ανίχνευση ανωμαλιών και καλύτερη κατανόηση δεδομένων. Θεωρείται δε, ότι η εφαρμογή μεθόδων Μηχανικής Μάθησης με επιτυχία, μπορεί να παρέχει ευκαιρίες βελτίωσης της αποτελεσματικότητας και ποιότητας στον τομέα της υγείας [2].

Όσον αφορά στη Βιολογία-Βιοπληροφορική, με την επιτυχή διεξαγωγή του Προγράμματος Αποκρυπτογράφησης του Ανθρώπινου Γονιδιώματος (The Human Genome Project, HGP), ενός από τα σπουδαιότερα της σύγχρονης επιστήμης, που αναφέρεται στον καθορισμό της ακολουθίας των ζευγών από αποτελούν το ανθρώπινο DNA αλλά και την εύρεση των γονιδίων αυτών καθ’ αυτών, η σύμπλευση με τεχνολογικές εξελίξεις όπως η Μηχανική Μάθηση, θεωρούνταν δεδομένες. Αν το δούμε σε ένα γενικότερο πλαίσιο, η εξέλιξη στον Πληροφοριακό τομέα, ευνοεί και προσφέρει σε αυτόν της Βιολογίας, τα μέσα για την ανάλυση αλλά και διαχείριση βιολογικών δεδομένων.

Όπως είναι απόλυτα κατανοητό, η Μηχανική Μάθηση χρησιμοποιείται σε πολλές εφαρμογές της Βιοπληροφορικής, αφού αποτελεί ένα από τα σημαντικότερα, αν όχι το σημαντικότερο εργαλείο στην ανάλυση των δεδομένων που προέρχονται από βιολογικές αναλύσεις. Ήδη, μέθοδοι Μηχανικής Μάθησης χρησιμοποιούνται για διάγνωση και πρόγνωση ασθενειών, υποβοήθηση ιατρών στη λήψη αποφάσεων κ.α.

Βέβαια, ένα ζήτημα που προκύπτει είναι ο τεράστιος αριθμός μεταβλητών σε αντιδιαστολή πάντα με τον αριθμό των διαθέσιμων δειγμάτων προς ανάλυση. Όπως προαναφέρθηκε, τούτο μπορεί να επιλυθεί μερικώς, με επιλογή χαρακτηριστικών όπου επιλέγουμε τα περισσότερα σημαντικά σε σχέση με το εκάστοτε πρόβλημα.

Ό εντοπισμός των χαρακτηριστικών που διαθέτουν σημαντική πληροφορία για την κατασκευή ενός μοντέλου Μηχανικής Μάθησης ο οποίος φυσικά διαθέτει την ικανότητα γενίκευσης, αποτελεί αναμφισβήτητα ένα από τα πιο σημαντικά ερευνητικά πεδία της Βιοπληροφορικής και όχι μόνο.

1.3 Αντικείμενο Διπλωματικής Εργασίας

Η παρούσα διπλωματική έχει ως αντικείμενο την ανάλυση βιολογικών δεδομένων με χρήση αλγορίθμων μηχανικής μάθησης με εφαρμογή στη διάγνωση του καρκίνου. Για τις ανάγκες της διπλωματικής χρησιμοποιήθηκαν 1065 δείγματα ανθρώπων από τους οποίους οι 1004 ήταν ασθενείς ενώ οι 61 υγιείς. Η συγκεκριμένη μορφή καρκίνου που μελετάται αφορά στον γαστρεντερικό καρκίνο (Gastrointestinal Cancer) που έχει 5 είδη, τον οισοφαγικό (esophageal), τον στομαχικό (stomach), τον παγκρεατικό(pancreatic), του συκωτιού(liver) και της χοληδόχου κύστης (gallbladder). Ασχοληθήκαμε με όλα τα είδη εκτός του συκωτιού. Για τον κάθε άνθρωπο είχαν συλλεχθεί οι εκφράσεις των γονιδίων τους σε αριθμητικές τιμές. Τα κοινά γονίδια των ανθρώπων για όλους τους τύπους καρκίνου προς ανάλυση ήταν 20501.

Ο σκοπός φυσικά είναι να εντοπιστεί και να χρησιμοποιηθεί ένα κατάλληλο υποσύνολο αυτών των γονιδίων-χαρακτηριστικών για την κατασκευή ενός αποδοτικού ταξινομητή. Με τον όρο κατάλληλο εννοούμε γονίδια-χαρακτηριστικά που περιέχουν σημαντική πληροφορία. Επιπλέον, η απόδοση του ταξινομητή σχετίζεται άμεσα με την ικανότητα του να γενικεύει.

Οι τιμές των γονιδίων των δεδομένων προς εξέταση, είναι εκφρασμένες μετά από μετασχηματισμό αλληλούχισης RNA (RNA-Seq). Η αλληλούχιση RNA είναι ουσιαστικά με μέθοδος ποσοτικοποίησης, συνήθως εφαρμοσμένη σε βιολογικά δεδομένα ώστε να γίνει περισσότερο αντιληπτή η διαφορετική τους έκφραση.

Οι δυσκολίες που προέκυψαν κατά την εκτέλεση των πειραμάτων αφορούν κατά κύριο λόγο στην ανισορροπία του δείγματος(κλάση 0 (υγιείς): 61, κλάση 1 (ασθενείς): 1004). Τούτο το ζήτημα ξεπεράστηκε μερικώς με τη χρήση κατάλληλων μεθόδων συνθετικής δημιουργίας στιγμιότυπων, όπως θα δούμε και στη συνεχεία. Επιπλέον, οι απαιτήσεις σε υπολογιστική δύναμη ήταν μεγάλες. Τόσο ο μεγάλος αριθμός των γονιδίων-χαρακτηριστικών, τόσο οι διάφορες τεχνικές όπως η αναζήτηση πλέγματος (Grid Search) κατέστησαν τα πειράματα πολύωρα και δύσκολα. Σε αυτό το σημείο, θα ήθελα να ευχαριστήσω εκ νέου τον καθηγητή μου, κ. Διαμαντάρα για την παροχή ενός συστήματος Η/Υ από το Α.τ.ε.ι.Θ.

1.3.1 Συνεισφορά

Ή συνεισφορά της διπλωματικής εργασίας είναι η εξής:

- Μελετήσαμε 20501 κοινά γονίδια σε σχέση με τα 4 είδη γαστρεντερικού καρκίνου που δουλέψαμε. Η μοναδική μεταβλητή-στόχος αφορούσε στην ύπαρξη ή μη της ασθένειας (κλάση 0=υγιείς, κλάση 1=ασθενείς). Στα κοινά γονίδια εφαρμόστηκαν μέθοδοι επιλογής χαρακτηριστικών όπως Kolmogorov-Smirnov 2 Samples, Mutual Information kai rfecv.

- Ως εκ τούτου, τα κοινά γονίδια (χαρακτηριστικά) ταξινομήθηκαν βάσει σημαντικότητας σε σχέση με τη μεταβλητή-στόχο μας με τις προαναφερθείσες μεθόδους.

- Τα δεδομένα μας επέστησαν περαιτέρω μετασχηματισμό καθώς χρησιμοποιήθηκαν επίσης μέθοδοι όπως κανονικοποίηση (Standardization) και εμπλουτισμός δείγματος με χρήση μεθόδων δημιουργίας συνθετικών στιγμιότυπων για τη μειοψηφούσα κλάση (σ.σ. κλάση 0=υγιείς).

- Μετά τη χρήση του καλύτερου δυνατού υποσυνόλου χαρακτηριστικών, εφαρμόστηκε αναζήτηση πλέγματος(Grid Search) για τους αποδοτικότερους αλγορίθμους ώστε να βρεθούν οι παράμετροι που βοηθούν τους ταξινομητές μας να γενικεύσουν καλύτερα.

- Μετά τη σύγκριση των ταξινομητών, βρέθηκε ο καλύτερος εξ αυτών (SVM with RBF Kernel) σε σχέση με μετρικές ακρίβειας, που εφαρμόστηκε στα μετασχηματισμένα (Standardized) δεδομένα μας, και συγκεκριμένα στο υποσύνολο χαρακτηριστικών που προέκυψε από τις επιλογές της μεθόδου rfecv. Τα αποτελέσματα ήταν ίδια μετά και την εφαρμογή στο εμπλουτισμένο δείγμα όπου εφαρμόστηκε η μέθοδος SMOTEEN (Δημιουργία συνθετικών (όχι ρέπλικες) στιγμιότυπων για τη μειοψηφούσα κλάση με εξάλειψη των περιθωριακών στιγμιότυπων της μειοψηφούσας κλάσης). Εδώ να τονιστεί ότι τα συνθετικά δείγματα χρησιμοποιήθηκαν μόνο για την εκπαίδευση των αλγορίθμων.

1.4 Οργάνωση κειμένου

Η οργάνωση κειμένου της διπλωματικής εργασίας που ακολουθείται είναι η εξής.

Στο **κεφάλαιο 2**, παρουσιάζονται σχετικές εργασίες και βιβλιογραφία σχετικά με το θέμα της διπλωματικής, περιληπτικά.

Στο **κεφάλαιο 3**, εμβαθύνουμε στις βασικές έννοιες της Μηχανικής Μάθησης, την πολυτιμότατη συμβολή της στον ιατρικό τομέα, και φυσικά περιγράφονται οι μέθοδοι και τεχνολογίες που χρησιμοποιούνται στην παρούσα διπλωματική εργασία, πιο συγκεκριμένα αναφερόμαστε στις μεθόδους επιλογής χαρακτηριστικών και μετασχηματισμού του δείγματος, στους διάφορους αλγορίθμους μηχανικής μάθησης που χρησιμοποιήθηκαν για την εκτέλεση των πειραμάτων και τέλος μια σύντομη περιγραφή της τεχνικής σχετικά με τον εμπλουτισμό του δείγματος μας.

Στο **κεφάλαιο 4**, προχωράμε στη λεπτομερή περιγραφή του συνόλου δεδομένων μας, των μεθόδων Επιλογής Χαρακτηριστικών που χρησιμοποιήσαμε, των διαφόρων τεχνικών που ερευνήθηκαν ώστε το αποτέλεσμα να είναι αποδοτικότερο, όπως κανονικοποίηση αρχικά και εμπλουτισμός του δείγματος λόγω ανισορροπίας μετέπειτα.

Στο **κεφάλαιο 5**, αναφέρονται τα αποτελέσματα των πειραμάτων που έγιναν όπως επίσης και η μεθοδολογία η οποία ακολουθήθηκε. Έπειτα προχωρούμε στο σχολιασμό των αποτελεσμάτων και των μεθόδων αυτών κλείνοντας με μία τελική σύνοψη των αποτελεσμάτων-συμπερασμάτων.

Το **κεφάλαιο 6** αναφέρεται στις προγραμματιστικές πλατφόρμες-εργαλεία, λογισμικά, βιβλιοθήκες κ.α. που χρησιμοποιήσαμε για την διενέργηση των πειραμάτων μας.

Τέλος, στο **κεφάλαιο 7**, καταθέτουμε τα ευρήματά μας και παρατείνουμε προτάσεις και ιδέες για μελλοντική έρευνα και τυχόν επεκτάσεις της παρούσας διπλωματικής.

Το **κεφάλαιο 8** περιέχει τη βιβλιογραφία πάνω στην οποία βασιστήκαμε για την ολοκλήρωση του βιβλιογραφικού μέρους αυτής της διπλωματικής εργασίας.

2. Σχετική Βιβλιογραφία

Στα πλαίσια της έρευνας μας στο Διαδίκτυο για σχετικές εργασίες που εφάπτονται στο θέμα της διπλωματικής εργασίας μας, διαπιστώσαμε το προφανές. Η βιβλιογραφία που υπάρχει είναι τεράστια. Είναι τόσο μεγάλης σημαντικότητας το πεδίο της Μηχανικής Μάθησης όχι μόνον για τον Ιατρικό τομέα που κάθε μέρα ξεπετάγονται καινούριες εργασίες και έρευνες επί του θέματος.

Φυσικά, το πρόβλημα που αντιμετωπίσαμε κατά την συγγραφή της διπλωματικής και ενόσω διενεργούσαμε τα πειράματά μας αποτελεί ένα κοινό στοιχείο όλων των πρότερων εργασιών/ερευνών. Και αυτό είναι η ύπαρξη μικρού δείγματος ανθρώπων από τη μια, ενώ από την άλλη ο αριθμός των γονιδίων είναι πολύ μεγάλος. Το φαινόμενο αυτό είναι ευρύτερα γνωστό ως η «κατάρα των πολλών διαστάσεων» ή αλλιώς η «κατάρα της Διαστασιμότητας» και είναι υπεύθυνο για τη δυσκολία δημιουργίας ενός πραγματικά αποδοτικού αλγορίθμου Μηχανικής Μάθησης με την ικανότητα να γενικεύει. Αυτό συμβαίνει διότι οι πολλές διαστάσεις/χαρακτηριστικά εισάγουν θόρυβο και μειώνουν την απόδοση του εκάστοτε συστήματος. Κάποια χαρακτηριστικά μπορεί να σχετίζονται μεταξύ τους και με το πρόβλημα ενώ άλλα ενδεχομένων να παρέχουν εντελώς ανούσια πληροφορία ως προς το πρόβλημα. Επιπλέον, λόγω των πολλών διαστάσεων, ο χώρος αυξάνεται σημαντικά, τα δεδομένα είναι αραιά κατανεμημένα και ως εκ τούτου μέθοδοι που προσπαθούν να μελετήσουν στατιστική σημαντικότητα, αποτυγχάνουν.

Στις περισσότερες εργασίες, όπως φυσικά και στη δική μας, γίνεται μια προσπάθεια μείωσης τούτων των διαστάσεων με κατάλληλες μεθόδους Επιλογής Χαρακτηριστικών/Μείωσης Διαστάσεων. Με την ανίχνευση λοιπόν και φυσικά τη χρήση των σημαντικών χαρακτηριστικών (στην περίπτωσή μας, γονιδίων), μειώνουμε τον αριθμό των διαστάσεων, και συνολικά τον όγκο των δεδομένων μας, η πολυπλοκότητα μειώνεται και μπορούμε ένα επιτύχουμε έναν δυνητικά αποδοτικότερο αλγόριθμο με τη δυνατότητα να γενικεύει.

Αναφορικά με τις μεθόδους Μείωσης Διαστάσεων/Επιλογής Χαρακτηριστικών που προαναφέρθηκαν, οι περισσότερες από αυτές, αν όχι όλες βασίζονται στη στατιστική και βάσει διαφόρων συσχετίσεων κατηγοριοποιούν τα χαρακτηριστικά ανά σημαντικότητα.

Ο κατάλληλος συνδυασμός μεθόδου Επιλογής Χαρακτηριστικών και μοντέλου Μηχανικής Μάθησης, μαζί με κάποιες άλλες ίσως παραμέτρους, θα καθορίσουν την επιτυχία ή όχι του πειράματος και άρα την αποδοτικότητα ή μη του τελικού αλγορίθμου μας.

Στα επόμενα υποκεφάλαια, παρουσιάζονται ενδεικτικά και περιληπτικά μερικές σχετικές πρόσφατες έρευνες επί του θέματος της διάγνωσης καρκίνου με μεθόδους Μηχανικής Μάθησης μετά από βιολογική ανάλυση δεδομένων.

2.1 Ανίχνευση καρκίνου του μαστού: Εφαρμογή Αλγορίθμων Μηχανικής Μάθησης στο διαγνωστικό σύνολο δεδομένων του Wisconsin

Ο Abien Fred M. Agarap στην έρευνά του, παρουσιάζει και συγκρίνει μια πληθώρα αλγορίθμων Μηχανικής Μάθησης για την ανίχνευση του καρκίνου του Μαστού. Οι αλγόριθμοι Μηχανικής Μάθησης εκπαιδεύτηκαν ώστε να ανιχνεύσουν τον καρκίνο του μαστού με τη χρήση του διαγνωστικού συνόλου δεδομένων καρκίνου του μαστού του Wisconsin (WDBC) [3]. Σύμφωνα με το [3], το σύνολο δεδομένων αποτελείται από χαρακτηριστικά που υπολογίστηκαν από μια ψηφιοποιημένη εικόνα παρακέντησης με λεπτή βελόνα (FNA) μιας μάζας στήθους. Τα εν λόγω χαρακτηριστικά περιγράφουν πληροφορίες των πυρήνων των κυττάρων που βρίσκονται στην εικόνα [3].

Υπάρχουν 569 στιγμιότυπα στο σύνολο δεδομένων: 212 - Κακοήθη, 357 - καλοήθη. Συνεπώς, τα χαρακτηριστικά των δεδομένων είναι τα εξής: (1) ακτίνα, (2) υφή, (3) περίμετρος, (4) περιοχή, (5) ομαλότητα, (6) συμπάγεια, (9) συμμετρία, και (10) διάσταση φράκταλ. Κάθε χαρακτηριστικό έχει τρεις πληροφορίες [20]: (1) μέση τιμή, (2) τυπικό σφάλμα και (3) "χειρότερη" ή μεγαλύτερη (μέση τιμή από τις τρεις μεγαλύτερες τιμές). Έτσι, δουλεύει ουσιαστικά με συνολικά 30 χαρακτηριστικά δεδομένων.

Το άρθρο παρουσιάζει τη σύγκριση μεταξύ έξι αλγορίθμων Μηχανικής Μάθησης, και συγκεκριμένα των GRU-SVM[4], Linear Regression, Multilayer Perceptron (MLP), Nearest Neighbor (NN) Search, Softmax Regression και τέλος SVM. Μετρήθηκαν οι τιμές ακρίβειας, ευαισθησίας (Sensitivity) και εξειδίκευσης (Specificity). Το δείγμα χωρίστηκε με τον εξής τρόπο, το 70% χρησιμοποιήθηκε στη φάση της εκπαίδευσης και το εναπομείναν 30% για τη φάση ελέγχου. Οι τιμές των παραμέτρων που χρησιμοποιήθηκαν για όλους τους αλγορίθμους/μοντέλα, ανατέθηκαν χειροκίνητα.

Όπως είναι προφανές, καθώς η πολυπλοκότητα του δείγματος ήταν μικρή, λόγω του λιγοστών χαρακτηριστικών, όλοι οι αλγόριθμοι τα πήγαν εξαιρετικά. Συγκεκριμένα, τα αποτελέσματα έδειξαν ότι όλοι τους είχαν ποσοστό ακρίβειας μεγαλύτερο από 90% στην επιτυχή ταξινόμηση των στιγμιότυπων του δείγματος. Ιδιαίτερα, ο αλγόριθμος MLP ξεχώρισε έναντι των υπολοίπων με ποσοστό ακρίβειας περίπου ίσο με 99,04 %.

Αυτό είναι ιδιαίτερα σημαντικό αν αναλογιστούμε πως ο καρκίνος του Μαστού είναι ένας από τους συνηθέστερους μαζί αυτούς του πνεύμονα, του προστάτη, του παχέος εντέρου και φυσικά του παγκρεατικού ο οποίος μάλιστα αποτελεί υποαντικείμενο της δικής μας διπλωματικής. Ο καρκίνος του μαστού, αντιπροσωπεύοντας το 15% όλων των νέων περιπτώσεων στις ΗΠΑ μόνο [5], αποτελεί αντικείμενο έρευνας με μεγάλη αξία. Η αξιοποίηση των προσεγγίσεων της Επιστήμης των δεδομένων (Data Science) και της Μηχανικής Μάθησης (Machine Learning) στα Ιατρικά πεδία αποδεικνύεται καθ' όλα παραγωγική αφού τέτοιες προσεγγίσεις θεωρούνται ότι προσφέρουν σημαντικότατη βοήθεια στη διαδικασία λήψης αποφάσεων των Ιατρών.

Καλώς ή κακώς, με την ατυχώς αυξανόμενη τάση κρουσμάτων καρκίνου του μαστού [5], γίνονται διαθέσιμα πολλά δεδομένα που είναι ύψιστης σημασίας στη συνέχιση κλινικών ερευνών και πειραμάτων και τελικά στην εφαρμογή τέτοιων μεθόδων μετέπειτα στον προαναφερθέντα τομέα.

Σε σχέση με την διενέργηση των πειραμάτων και των αποτελεσμάτων της συγκεκριμένης εργασίας, καταρχάς να τονίσουμε ότι οι αριθμητικές τιμές κανονικοποιήθηκαν, όπως ακριβώς και στην παρούσα διπλωματική εργασία. Θεωρούμε πως η κανονικοποίηση των δεδομένων αποτελεί ένα σημαντικό κομμάτι της προεπεξεργασίας δεδομένων.

Ο παρακάτω πίνακας δείχνει τις τιμές των παραμέτρων που χρησιμοποιήθηκαν ανά αλγόριθμο.

Table 1: Hyper-parameters used for the ML algorithms.

Hyper-parameters	GRU-SVM	Linear Regression	MLP	Nearest Neighbor	Softmax Regression	SVM
Batch Size	128	128	128	N/A	128	128
Cell Size	128	N/A	[500, 500, 500]	N/A	N/A	N/A
Dropout Rate	0.5	N/A	None	N/A	N/A	N/A
Epochs	3000	3000	3000	1	3000	3000
Learning Rate	1e-3	1e-3	1e-2	N/A	1e-3	1e-3
Norm	L2	N/A	N/A	L1, L2	N/A	L2
SVM C	5	N/A	N/A	N/A	N/A	5

Ο παρακάτω πίνακας συνοψίζει τα αποτελέσματα των πειραμάτων.

Table 2: Summary of experiment results on the ML algorithms.

Parameter	GRU-SVM	Linear Regression	MLP	L1-NN	L2-NN	Softmax Regression	SVM
Accuracy	93.75%	96.09375%	99.038449585420729%	93.567252%	94.736844%	97.65625%	96.09375%
Data points	384000	384000	512896	171	171	384000	384000
Epochs	3000	3000	3000	1	1	3000	3000
FPR	16.666667%	10.204082%	1.267042%	6.25%	9.375%	5.769231%	6.382979%
FNR	0	0	0.786157%	6.542056%	2.803738%	0	2.469136%
TPR	100%	100%	99.213843%	93.457944%	97.196262%	100%	97.530864%
TNR	83.333333%	89.795918%	98.732958%	93.75%	90.625%	94.230769%	93.617021%

Όπως ήταν αναμενόμενο, οι γραμμικοί ταξινομητές (Linear Regression και SVM) είχαν το πλεονέκτημα καθώς το χρησιμοποιούμενο σύνολο δεδομένων ήταν γραμμικά διαχωρίσιμο.

Συνοψίζοντας, η συγκεκριμένη εργασία παρουσιάζει την εφαρμογή διαφορετικών αλγορίθμων μηχανικής μάθησης για τη διάγνωση του καρκίνου του μαστού. Όλοι οι αλγόριθμοι Μηχανικής Μάθησης που παρουσιάστηκαν παρουσίασαν υψηλές επιδόσεις στη δυαδική ταξινόμηση του καρκίνου του μαστού, δηλ. προσδιορίζοντας αν το κάθε στιγμιότυπο αφορά καλοήθεις ή κακοήθεις όγκους. Συνεπώς, οι μέθοδοι που χρησιμοποιήθηκαν σχετικά με το πρόβλημα ταξινόμησης ήταν επίσης ικανοποιητικοί. Για να τεκμηριωθούν περαιτέρω τα αποτελέσματα αυτής της μελέτης, πρέπει να χρησιμοποιηθούν επιπλέον μέθοδοι όπως η διασταυρούμενη επικύρωση (Cross Validation). Η εφαρμογή μιας τέτοιας τεχνικής όχι μόνο θα παρέχει μια πιο ακριβή μέτρηση της απόδοσης πρόβλεψης μοντέλου, αλλά θα βοηθήσει επίσης στον προσδιορισμό των βέλτιστων τιμών παραμέτρων για τους αλγορίθμους Μηχανικής Μάθησης.

2.2 Επιλογή βιοδεικτών με χρήση επαναλαμβανόμενων Μηχανών Υποστήριξης Διανυσμάτων (Recursive SVM) για έγκαιρη ανίχνευση καρκίνου του μαστού στο αίμα

Οι Fan Zhang, Howard L Kaufman, Youping Deng, και Renee Drabier, παρουσιάζουν στην εργασία τους έναν αλγόριθμο Μηχανών Υποστήριξης Διανυσμάτων βασισμένο σε επαναλαμβανόμενη εξάλειψη χαρακτηριστικών και διασταυρούμενη επικύρωση(SVM-RFECV) για την έγκαιρη ανίχνευση καρκίνου του μαστού στο αίμα και δείχνουν πως να χρησιμοποιηθεί ο εν λόγω αλγόριθμος ώστε να μοντελοποιηθεί το πρόβλημα ταξινόμησης και πρόγνωσης του καρκίνου προς έρευνα.

Το σύνολο εκπαίδευσης αποτελείται από 32 υγιή και 33 ασθενή στιγμιότυπα ενώ το σύνολο ελέγχου αποτελείται από 31 υγιή και 34 ασθενή στιγμιότυπα τα οποία διαχωρίστηκαν τυχαία από ένα σύνολο δεδομένων το οποίο μεταφορτώθηκε από το Gene Express Omnibus (GEO).

Αρχικά, αναγνωρίστηκαν οι 43 διαφορετικά εκφρασμένοι Βιοδείκτες μεταξύ κανονικότητας και καρκίνου. Στη συνέχεια, με τη χρήση του αλγορίθμου SVM-RFECV εξήχθησαν 15 Βιοδείκτες. Τέλος, έγινε σύγκριση της επίδοσης ταξινόμησης και πρόγνωσης των SVM-RFECV, SVM και SVM-RFE.

Η προσέγγιση του Guyon [6] για χρήση του SVM-RFE προς επιλογή γονιδίων, αποτελεί μία από τις αποδοτικότερες μεθόδους επιλογής χαρακτηριστικών που έχουν χρησιμοποιηθεί επιτυχώς στην επιλογή σημαντικών γονιδίων για ταξινόμηση του καρκίνου. Πρόκειται για μια προσέγγιση επιλογής προς τα πίσω η οποία επιλέγει γονίδια βασισμένη στην επίδρασή τους (βάρος) σε σχέση με μια Μηχανή Υποστήριξης Διανυσμάτων. Πρώτα υπολογίζει την κατάταξη βάσει των βαρών. Έπειτα, εξαλείφει χαρακτηριστικά με την μικρότερη θέση στην κατάταξη. Τέλος, επαναλαμβάνει τη διαδικασία έως ότου επιτευχθεί η υψηλότερη τιμή ακρίβειας ταξινόμησης.

Ο SVM-RFE χρησιμοποιείται για την εύρεση διακριτών σχέσεων εντός κλινικών συνόλων δεδομένων και εντός συνόλων γονιδιακής έκφρασης. Ωστόσο, η επαναλαμβανόμενη εξάλειψη χαρακτηριστικών (RFE) είναι ευαίσθητη σε μικρές διαταραχές/διαφοροποιήσεις του τεστ εκπαίδευσης. Τα χαρακτηριστικά τα οποία εξάγονται από το σετ εκπαίδευσης μπορεί να μην έχουν καλή απόδοση σε ένα ανεξάρτητο σετ ελέγχου. Αυτό πιθανόν οφείλεται στην υπερπροσαρμογή (Overfitting) της μεθόδου και μπορεί να προκύψει όταν ο αριθμός των χαρακτηριστικών είναι μεγάλος ενώ αντίστοιχα ο αριθμός του υποσυνόλου προς εκπαίδευση μικρός. Επίσης, όταν εμφανίζονται κάποιες κανονικότητες στο σύνολο εκπαίδευσης οι οποίες απουσιάζουν από το σύνολο ελέγχου. Προκειμένου να αποφευχθεί η υπερπροσαρμογή και να αποκτήσουμε την καλύτερη ακρίβεια πρόβλεψης για το σετ ελέγχου, προτείνεται μια Μηχανή Υποστήριξης Διανυσμάτων (SVM) βασισμένη στην Επαναλαμβανόμενη Εξάλειψη Χαρακτηριστικών (RFE)και στην Διασταυρούμενη Επικύρωση (CV) για την εξαγωγή των βέλτιστων χαρακτηριστικών (SVM-RFE-CV).

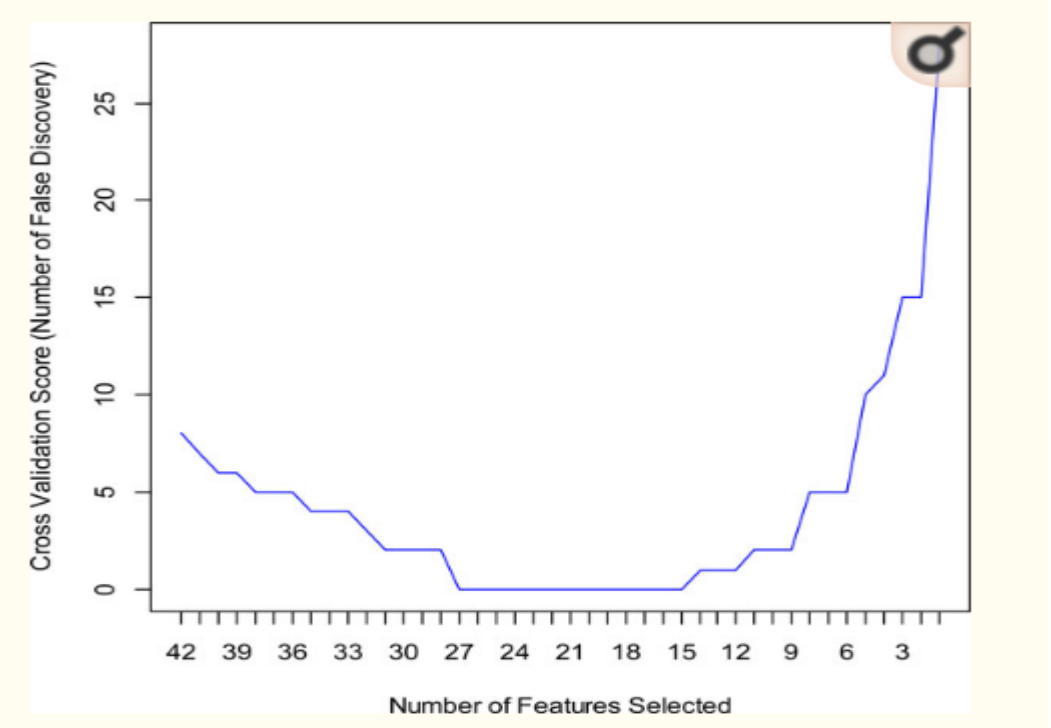
Επιστρέφοντας στο πείραμα, να πούμε εδώ πως κάποια μορφή κανονικοποίησης χρησιμοποιήθηκε για την καλύτερη κατανομή διαφοροποίησης ανά γονίδιο.

Στην παρακάτω εικόνα βλέπουμε τη σύγκριση των διαφορετικών αλγορίθμων-μεθόδων για διαφορετικούς αριθμούς γονιδίων. Είναι ξεκάθαρο πως με τη χρήση του Rfecv και τη μείωση των χαρακτηριστικών/γονιδίων, οι επιδόσεις αυξάνονται θεαματικά.

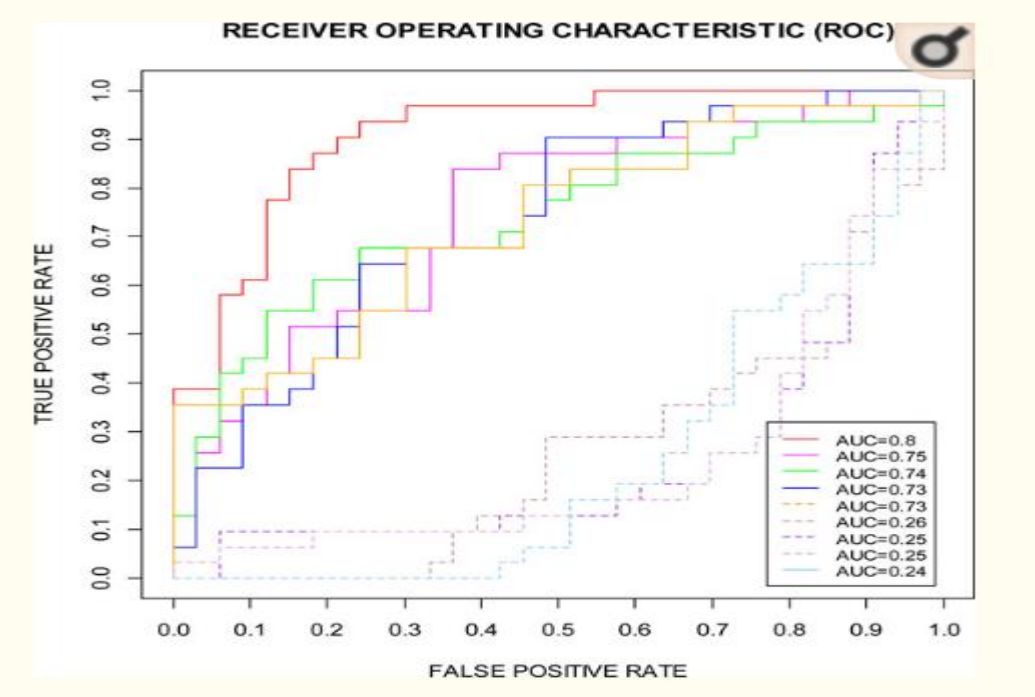
performance comparison of SVM, SVM-RFE, and SVM-RFE-CV

Measure	**SVM**		**SVM-RFE**		**SVM-RFE-CV**	
#genes	**42**		**18**		**15**	
	Training set	**Testing set**	**Training set**	**Testing set**	**Training set**	**Testing set**
Precision	97.0%	58.8%	100%	71.4%	100%	74.29%
Accuracy	98.4%	56.9%	100%	70.8%	100%	73.85%
Sensitivity	100.0%	58.8%	100%	73.5%	100%	76.47%
Specificity	96.9%	54.8%	100%	67.7%	100%	70.97%
AUC	0.98	0.56	1.0	0.75	1.0	0.80

Στην παρακάτω εικόνα βλέπουμε τον αυτόματο συντονισμό του αριθμού των χαρακτηριστικών με τη βοήθεια του RFE και CV.



Αναφορικά με τα συμπεράσματα, είναι προφανές ότι ο αλγόριθμος SVM-RFE-CV είναι κατάλληλος για ανάλυση δεδομένων μεγάλου όγκου με θόρυβο. Επιπλέον, αποδίδει καλύτερα σε σχέση με τον απλό SVM-RFE σε σχέση με την ανθεκτικότητα στο θόρυβο και την ικανότητα να βρίσκει σημαντικά χαρακτηριστικά και τέλος μπορεί να βελτιώσει την απόδοση πρόβλεψης (Area Under Curve) στο σετ ελέγχου από 0.58 σε 0.78.



Συνοψίζοντας, αναπτύχθηκε μια ολοκληρωμένη υπολογιστική προσέγγιση που αφορούσε ένα σημαντικό πρόβλημα ανάπτυξης βιοδεικτών στην έγκαιρη ανίχνευση καρκίνου του Μαστού χρησιμοποιώντας περιφερειακό αίμα. Η προσέγγιση αυτή, συνδύασε Επαναλαμβανόμενη Εξάλειψη Χαρακτηριστικών (RFE) με τη Χρήση του Αλγορίθμου Μηχανών Υποστήριξης Διανυσμάτων (SVM) και διασταυρούμενη Επικύρωση (CV). Διέγνωσε αυτόματα μη-γραμμικές συσχετίσεις μεταξύ των χαρακτηριστικών (Features) και αποτελέσματος (τιμές κλάσης ανά στιγμιότυπο) για τη δημιουργία του βέλτιστου μοντέλου πρόγνωσης με τον ελάχιστο αριθμό χαρακτηριστικών, το οποίο επιτύγχανε επιδόσεις AUC=0.80 με ευαισθησία (Sensitivity) 0.76 και Εξειδίκευση (Specificity) 0.70 στα δεδομένα ελέγχου.

Ο Αλγόριθμος SVM-RFE με βάση την διασταυρούμενη επικύρωση είναι σε θέση να προσδιορίσει τον βέλτιστο πίνακα πολλαπλών δεικτών με τον μικρότερο αριθμό γονιδίων. Μπορεί να φιλτράρει άσχετα γονίδια συγκεκριμένου ιστού από εκείνα που σχετίζονται με κακοήθεια. Είναι επίσης σε θέση να προσδιορίσει μοτίβα γονιδιακής έκφρασης που σχετίζονται με τη σοβαρότητα της νόσου. Είναι μια αποτελεσματική μέθοδος για την εύρεση δεικτών που εμπλέκονται σε καρκίνους.

2.3 Επιλογή Χαρακτηριστικών για τη Διάγνωση του Καρκίνου του Πνεύμονα με τη χρήση SVM βασισμένου σε RFE

Οι Kesav Kancherla και Srinivas Mukkamala στην προηγούμενη εργασία τους, μελέτησαν την απόδοση γνωστών μεθόδων Μηχανικής Μάθησης σε σχέση με την ικανότητα τους να ταξινομούν πάνω στην εσωτερική μελέτη Biomoda. Χρησιμοποίησαν 79 χαρακτηριστικά σχετιζόμενα με το σχήμα, την ένταση και την υφή. Πέτυχαν ποσοστό ακρίβειας 80% χρησιμοποιώντας το τρέχον σύνολο χαρακτηριστικών. Προκειμένου να βελτιώσουν τις επιδόσεις των μεθόδων τους, πραγματοποίησαν επιλογή χαρακτηριστικών πάνω στα 79 υπάρχοντα. Χρησιμοποίησαν τον αλγόριθμο Μηχανής Υποστήριξης Διανυσμάτων (SVM) βασισμένο στην Επαναλαμβανόμενη Εξάλειψη Χαρακτηριστικών (RFE) για την διεξαγωγή των πειραμάτων τους. Πέτυχαν ποσοστό ακρίβειας 87.5% με τη χρήση ενός υποσυνόλου 19 χαρακτηριστικών.

Διάφορες μέθοδοι όπως σάρωση CT (Computed Tomography), ακτινογραφία θώρακα, ανάλυση πτυέλων, ανάλυση δεδομένων μικροσυστοιχιών χρησιμοποιούνται για την ανίχνευση του καρκίνου του θώρακα [7]. Η μαζική ανίχνευση με σάρωση CT είναι ελπιδοφόρα, ωστόσο, δεν συνίσταται καθώς είναι δαπανηρή και δεν μπορεί να τεκμηριωθεί η μακροπρόθεσμη ασφάλεια της μεθόδου αυτής λόγω έκθεσης σε ακτινοβολία [8] . Από την άλλη, η χρήση δεδομένων μικροσυστοιχιών είναι δαπανηρή επίσης. Σε αυτή την εργασία εξετάζεται η χρήση της Τετρακικής Καρβοξυφαινυλικής Πορφίνης (Tetrakis Carboxy Phenyl Porphine or TCPP) ως μια εναλλακτική μέθοδος έγκαιρης ανίχνευσης του καρκίνου του Πνεύμονα.

Η χρήση Μηχανικής Μάθησης για την υποβοήθηση της ανίχνευσης και πρόβλεψης καρκίνου ερευνήθηκε στο [9]. Τεχνικές Μηχανικής Μάθησης όπως Τεχνητά Νευρωνικά Δίκτυα (ANN) και Δέντρα Αποφάσεων (DT) χρησιμοποιούνται για τη διάγνωση του καρκίνου εδώ και περίπου 20 χρόνια [10, 11 και 12]. Η δυναμική της χρήσης μεθόδων Μηχανικής Μάθησης για διάγνωση καρκινικών κυττάρων η όγκων μέσω ακτινών Χ, σαρώσεις CT αναφέρεται στο [13, 14]. Μέθοδοι Μηχανικής Μάθησης που χρησιμοποιούνται για ταξινόμηση όγκων ή ανίχνευση καρκίνου με χρήση δεδομένα μικροσυστοιχίας ή γονιδιακές εκφράσεις είναι οι, Γραμμική Διακριτή Ανάλυση Φίσερ (Fisher Linear Discriminant Analysis) [15], Κ-Κοντινότεροι Γείτονες (K-Nearest Neighbor) [16], Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines) [17], Ενίσχυση (Boosting) και Χάρτες Αυτό-Οργάνωσης (SOM) [18], Ιεραρχική Στοίχιση [19] και τέλος Θεωρητικές Προσεγγίσεις Γραφημάτων [20].

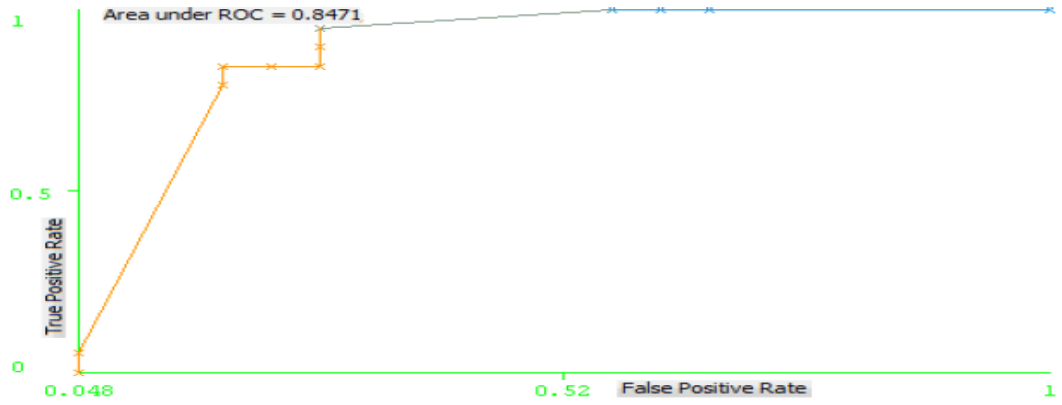
Η προσέγγιση του αλγορίθμου SVM-RFE [6] είναι μια δημοφιλής τεχνική για επιλογή χαρακτηριστικών, ειδικά στον τομέα της Βιοπληροφορικής. Η λειτουργία της επαναλαμβανόμενης Εξάλειψης χαρακτηριστικών (RFE) περιγράφηκε παραπάνω. Μετά την εκτέλεση τεχνικών επεξεργασίας εικόνας εξήχθησαν 79 χαρακτηριστικά. Το σύνολο δεδομένων αποτελείται από 119 στιγμιότυπα, από τα οποία τα 60 προέρχονται από δείγματα καρκίνου ενώ τα υπόλοιπα 59 από δείγματα φυσιολογικών ασθενών. Το 66% του συνολικού δείγματος χρησιμοποιήθηκε για την εκπαίδευση και το εναπομείναν 34% για τον

έλεγχο. Προκειμένου να αποφευχθεί τυχόν υπερπροσαρμογή (Overfitting), χρησιμοποιήθηκε Διασταυρούμενη Επικύρωση 5 υποσυνόλων (5-Fold Cross Validation) κατά τη διάρκεια της εκπαίδευσης. Ο μη γραμμικός πυρήνας RBF (Radial Basis Function) χρησιμοποιήθηκε ώστε να βρεθούν οι βέλτιστες παράμετροι.

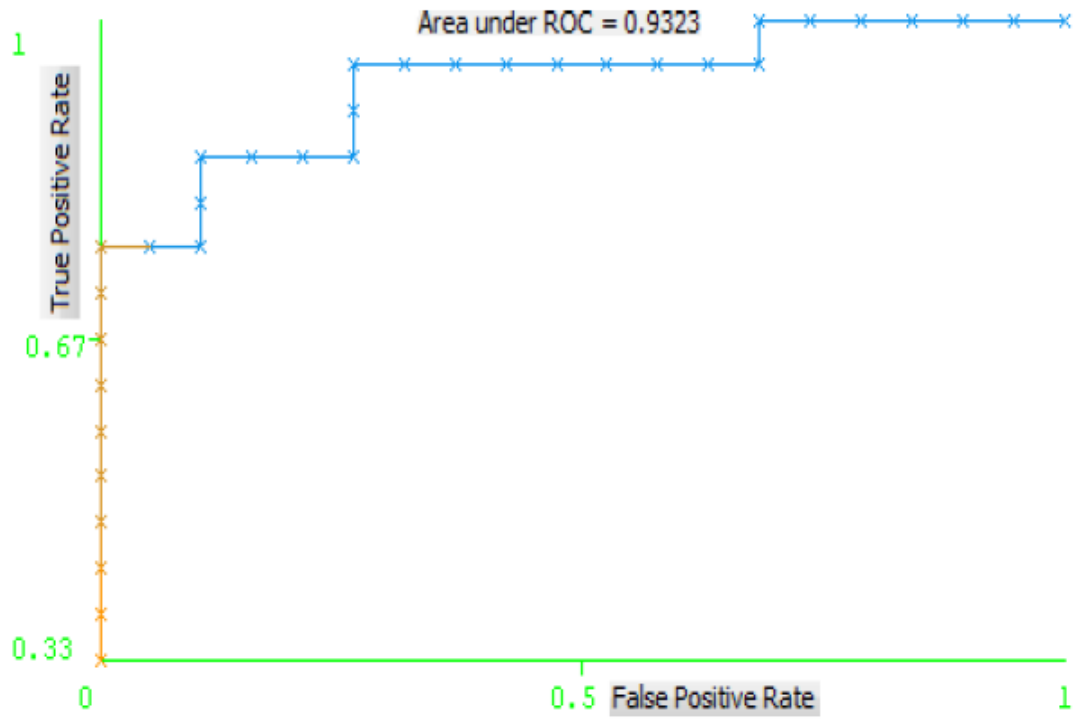
Αναφορικά με τα πειράματα που διενεργήθηκαν, σε αυτό το σύνολο δεδομένων, εκτελέστηκε ο αλγόριθμος SVM-RFE. Τα χαρακτηριστικά ταξινομήθηκαν βάσει των βαρών που απέδωσε ο αλγόριθμος SVM. Αφού αφαιρέθηκαν τα λιγότερο σημαντική χαρακτηριστικά με βήμα 5, έγινε ανάλυση των εναπομεινάντων χαρακτηριστικών. Η ακρίβεια που επετεύχθη για κάθε βήμα φαίνεται στον πίνακα 1.

Number of features used	Accuracy obtained
79	80
74	80
69	82.5
64	82.5
59	85
54	85
49	85
44	85
39	85
34	82.5
29	82.5
24	85
19	87.5

Στη συνέχεια, δείχνεται η ακρίβεια της μεθόδου με τη χρήση καμπυλών ROC (Receiver Operating Characteristic). Οι εν λόγω καμπύλες παράγονται για τους αλγορίθμους SVM εξετάζοντας τον ρυθμό συσσώρευσης αληθινών-θετικών (True Positives) έναντι του ρυθμού των ψευδών-θετικών (False Positives) οι οποίοι αντιστοιχούν στον κάθετο και οριζόντιο άξονα του σχήματος 1. Είναι σαφές ότι το σημείο (0,1) αντιστοιχεί στον ιδεατό τέλειο ταξινομητή. Το παρακάτω σχήμα δείχνει την καμπύλη ROC για 79 χαρακτηριστικά.



Αντίστοιχα, το παρακάτω σχήμα δείχνει την καμπύλη ROC για 19 χαρακτηριστικά.



Συνοψίζοντας, σε αυτήν την εργασία, πραγματοποιήθηκε Επιλογή Χαρακτηριστικών με χρήση του αλγορίθμου SVM βασισμένου στην Επαναλαμβανόμενη Εξάλειψη Χαρακτηριστικών για 79 χαρακτηριστικά (RFE). Χαρακτηριστικά εξαλείφθηκαν βάσει της κατάταξης τους και επετεύχθη ένα ποσοστό ακρίβειας της τάξης του 87.5% χρησιμοποιώντας τα 19 πιο σημαντικά από αυτά. Εκτός από τη χρήση του ως ενός δυναμικού εργαλείο διαλογής του καρκίνου του πνεύμονα, αυτή η μέθοδος μπορεί να χρησιμοποιηθεί για την παρακολούθηση της αποτελεσματικότητας μιας θεραπείας , να ανιχνεύσει την επανεμφάνιση του καρκίνου του πνεύμονα, αλλά και να εντοπίσει τους ασθενείς που ενδεχομένων να χρειαστούν μια επεμβατική διαγνωστική διαδικασία. Τα αποτελέσματά της εργασίας δείχνουν τη δυναμική της χρήσης επιλογής χαρακτηριστικών για τη βελτίωση της ακρίβειας και της αποτελεσματικότητας της ανίχνευσης καρκίνου του πνεύμονα.

3 Θεωρητικό Υπόβαθρο

Όπως προαναφέρθηκε, στόχος της παρούσας διπλωματικής εργασίας είναι η ανάλυση βιολογικών δεδομένων με απώτερο σκοπό την κατασκευή ενός αποδοτικού (με την ικανότητα να γενικεύει) ταξινομητή ώστε να επιτευχθεί στον μέγιστο δυνατό βαθμό ακριβής διάγνωση του γαστρεντερικού καρκίνου, πάντα με μεθόδους Μηχανικής Μάθησης.

Αρχικά, εντοπίστηκαν τα κοινά γονίδια καθώς η ερευνά μας αφορούσε 4 τύπους του γαστρεντερικού καρκίνου (οισοφαγικός, στομαχικός, παγκρεατικός και της χοληδόχου κύστης). Έπειτα, με τη χρήση τεχνικών Μηχανικής Μάθησης, αξιολογήθηκαν οι συσχετίσεις μεταξύ των διαφόρων γονιδίων και της ύπαρξης ή μη της ασθένειας. Μετέπειτα, τα σημαντικότερα γονίδια, δηλαδή εκείνα που κουβαλούσαν ικανή πληροφορία ώστε να μας βοηθήσουν να κάνουμε μια ασφαλή πρόβλεψη, επιλέχθηκαν. Αυτό έγινε με διάφορες τεχνικές και ως εκ τούτου προέκυψαν διαφορετικά υποσύνολα σημαντικών υποσυνόλων. Τέλος, χρησιμοποιώντας, τα διάφορα υποσύνολα αυτά, διενεργήσαμε τα πειράματα μας κάνοντας χρήση διαφόρων αλγορίθμων Μηχανικής Μάθησης. Στους αποδοτικότερους από αυτούς χρησιμοποιήθηκε επίσης η μέθοδος αναζήτησης πλέγματος ή αλλιώς εξαντλητική αναζήτηση (Grid Search) για την εύρεση των βέλτιστων παραμέτρων για κάθε ταξινομητή ώστε τελικά να καταλήξουμε στον αποδοτικότερο όλων.

Στο κεφάλαιο αυτό, αρχικά, αναφέρονται οι βασικές έννοιες της Μηχανικής Μάθησης, τίθεται επί τάπητος το ζήτημα της ταξινόμησης το οποίο μας απασχολεί στην παρούσα διπλωματική εργασία, ενώ φυσικά αναφερόμαστε τα κριτήρια επίδοσης ταξινομητών. Στη συνέχεια προχωρούμε σε μια ανάλυση κάποιων από τους αλγορίθμους που χρησιμοποιήθηκαν και σημαντικότερα στις Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machine) που ήταν και ο νικητής της παρούσας διπλωματικής στη μάχη των ταξινομητών. Επιπλέον, αναλύονται οι διάφορες μέθοδοι επιλογής χαρακτηριστικών που βοήθησαν στη διεξαγωγή των πειραμάτων αναφορικά με την εύρεση του καταλληλότερου υποσυνόλου χαρακτηριστικών για να δουλέψουμε, πιο συγκεκριμένα, θα αναφερθούμε στην τεχνική της Αμοιβαίας Πληροφορίας (Mutual Information), του Kolmogorov-Smirnov για δυο δείγματα (Kolmogorov-Smirnov 2 Samples) και τέλος της Επαναλαμβανόμενης Εξάλειψης Χαρακτηριστικών με Διασταυρούμενη Επικύρωση (Recursive Feature Elimination with Cross Validation).

3.1 Μηχανική Μάθηση – Εφαρμογές και Σενάρια Μάθησης

Μια συνοπτική αναφορά στη Μηχανική Μάθηση, πραγματοποιήθηκε στην εισαγωγή. Εκεί, δώσαμε ένα σύντομο ορισμό της Μηχανικής Μάθησης, ενώ αναφερθήκαμε σε κάποιες έννοιες όπως αυτές της Εμπειρίας (Experience), της Πολυπλοκότητας Δείγματος. Τέλος δόθηκαν κάποιες πληροφορίες σχετικά με την ορολογία που σχετίζεται με το πρόβλημα της Ταξινόμησης (Classification) όπως Δείγμα (Sample), Χαρακτηριστικά (Features), Κλάσεις (Classes), Σύνολο Εκπαίδευσης (Training Set) και Σύνολο Ελέγχου (Test Set).

Σε αυτό το σημείο θα ήταν καλό να τονίσουμε τις διαφορετικές δυνατότητες των μεθόδων Μηχανικής Μάθησης. Μιλήσαμε πολύ για την ταξινόμηση και φυσικά την επιλογή χαρακτηριστικών. Ωστόσο, οι διάφορες μέθοδοι και αλγόριθμοι Μηχανικής Μάθησης χρησιμοποιούνται επίσης για Αναγνώριση Προτύπων (Pattern Recognition), παλινδρόμηση (Regression) που συνήθως αφορά προβλέψεις (Forecast) και Ομαδοποίηση (Clustering). Όπως είναι αντιληπτό, στα πλαίσια της δικής μας διπλωματικής εργασίας, εμβαθύνουμε στο πρόβλημα της ταξινόμησης με ταυτόχρονη χρήση μεθόδων Επιλογής Χαρακτηριστικών.

Οι Ταξινομητές είναι ουσιαστικά αλγόριθμοι που χωρίζουν διάφορα στιγμιότυπα σε κλάσεις, είτε αυτά αφορούν αντικείμενων, είτε εικόνες είτε ακόμα και χρονοσειρές. Φυσικά, το σύνολο των δεδομένων που δέχεται ένας ταξινομητής, αποτελείται πάντοτε από αριθμητικές τιμές.

Στο σημείο αυτό πρέπει να αναφερθούμε στα διαφορετικά σενάρια Μάθησης. Τα σενάρια αυτά διαφοροποιούνται ως προς τον τύπο των διαθέσιμων δεδομένων και τη μέθοδο με την οποία πραγματοποιείται η φάση της εκπαίδευσης και ελέγχου [1]. Γενικά, υπάρχουν διάφοροι τρόποι διαχωρισμού των σεναρίων αυτών. Θα αναφερθούμε περιληπτικά σε μερικά από αυτά:

- **Μάθηση υπό Επίβλεψη (Supervised Learning).** Το σύστημα Μάθησης δέχεται ένα σύνολο δειγμάτος κατηγοριοποιημένο σε κλάσεις και το χρησιμοποιεί για εκπαίδευση. Έπειτα, πραγματοποιεί προβλέψεις για νέα δεδομένα. Αποτελεί ίσως το πιο κοινό σενάριο και σχετίζεται με ζητήματα όπως η ταξινόμηση και η πρόβλεψη τιμής συνάρτησης.

- **Μάθηση χωρίς Επίβλεψη (Unsupervised Learning).** Το σύστημα Μάθησης δέχεται ένα σύνολο δείγματος για το οποίο δεν είναι οι γνωστές οι κλάσεις για κάθε στιγμιότυπο για εκπαίδευση. Έπειτα, πραγματοποιεί προβλέψεις για νέα δεδομένα. Όπως γίνεται εύκολα αντιληπτό, είναι δύσκολο να ποσοτικοποιηθεί η απόδοση ενός τέτοιου συστήματος. Παραδείγματα συστημάτων Μάθησης χωρίς επίβλεψη αποτελούν η Ομαδοποίηση και η Μείωση Διαστάσεων.

- **Μάθηση με ενίσχυση (Reinforcement Learning).** Το σύστημα Μάθησης εναλλάσσει τις φάσεις εκπαίδευσης και ελέγχου. Για να συλλέξει πληροφορίες, το σύστημα μάθησης αλληλοεπιδρά ενεργά με το περιβάλλον ενώ μερικές φορές το επηρεάζει κιόλας. Παρόλα αυτά, το σύστημα συνήθως αντιμετωπίζει το δίλλημα της εξερεύνησης ή εκμετάλλευσης (exploration versus exploitation) διότι θα πρέπει να επιλέξει μεταξύ της εξερεύνησης άγνωστων ενεργειών για να λάβει περισσότερη πληροφορία και της εκμετάλλευσης της ήδη υπάρχουσας πληροφορίας.

Παρακάτω ακολουθούν περισσότερα σενάρια μάθησης, ίσως όχι τόσα κοινώς χρησιμοποιούμενα:

- **Μάθηση υπό μερική Επίβλεψη (Semi-supervised Learning).** Όπως σωστά μπορεί να υποτεθεί, το εν λόγω σενάριο συνδυάζει τα σενάρια με και χωρίς επίβλεψη. Δηλαδή, το σύστημα Μάθησης δέχεται ένα σύνολο δείγματος με γνώστες τις κλάσεις για κάποια στιγμιότυπα ενώ για άλλα, όχι, για εκπαίδευση. Έπειτα, πραγματοποιεί προβλέψεις για νέα δεδομένα. Το συγκεκριμένο σενάριο εφαρμόζεται συχνά σε περιπτώσεις όπου η συλλογή δεδομένων χωρίς γνωστές κλάσεις είναι εύκολή, ενώ από την άλλη, δεδομένα με γνωστές τις κλάσεις είναι δύσκολο να αποκτηθούν, είτε λόγω κόστους είτε λόγω άλλων παραγόντων. Σχετίζεται με ζητήματα όπως η ταξινόμηση και η πρόβλεψη τιμής συνάρτησης. Ένα παράδοξο της χρήσης τέτοιων σεναρίων αποτελεί το γεγονός ότι η κατανομή των δεδομένων με άγνωστες κλάσεις μπορεί να βοηθήσει στην επίτευξη καλύτερης απόδοσης από την Μάθηση με Επίβλεψη. Αυτό είναι κάτι που αποτελεί αντικείμενο θεωρητικής και εφαρμοσμένης έρευνας στον τομέα της Μηχανικής Μάθησης.

- **Μεταβιβαστικός συμπερασμός (Transductive Inference).** Το σύστημα μάθησης δέχεται ένα σύνολο εκπαίδευσης για το οποίο είναι γνωστές οι κλάσεις, μαζί με ένα σύνολο ελέγχου χωρίς κλάσεις. Αυτό το είδος μάθησης παράγει προβλέψεις μόνον για τα συγκεκριμένα δεδομένα ελέγχου. Τούτος ο τρόπος εκμάθησης είναι ευκολότερος και συναντάται σε διάφορες σύγχρονες εφαρμογές. Όπως είναι αναμενόμενο, οι συνθήκες κάτω από τις οποίες παράγει υψηλή απόδοση είναι αντικείμενο έρευνας.

- **Μάθηση με απευθείας σύνδεση (Online Learning).** Σε αντίθεση με τα προηγούμενα σενάρια μάθησης, το σενάριο τούτο, περιλαμβάνει πολλές επαναλήψεις κατά τις οποίες, οι φάσεις εκπαίδευσης και ελέγχου εναλλάσσονται. Για κάθε επανάληψη, το σύστημα μάθησης δέχεται ένα σύνολο εκπαίδευσης χωρίς τιμές χαρακτηριστικών, παράγει μια πρόβλεψη για αυτό και τελικώς λαμβάνει τις πραγματικές τιμές και υπολογίζει το σφάλμα. Ο στόχος του σεναρίου αυτού είναι η ελαχιστοποίηση του συσωρευτικού σφάλματος για όλες τις επαναλήψεις. Στη μάθηση με απευθείας σύνδεση δεν γίνεται κάποια υπόθεση ως προς την κατανομή των δεδομένων.

- **Ενεργή μάθηση (Active Learning).** Το σύστημα μάθησης συλλέγει με δυναμικά, ανάλογα με τις εκάστοτε ανάγκες, δείγματα εκπαίδευσης με ερωτήσεις προς κάποιο άλλο σύστημα για τις κλάσεις των στιγμιότυπων των δειγμάτων και τις τιμές των χαρακτηριστικών τους. Ο στόχος αυτού του σεναρίου Μάθησης είναι να επιτευχθεί απόδοση η οποία θα είναι συγκρίσιμη με ένα σενάριο μάθησης υπό επίβλεψη, αλλά με λιγότερα κατηγοριοποιημένα δείγματα. Η ενεργή μάθηση χρησιμοποιείται σε περιπτώσεις όπου η κατηγοριοποίηση των δειγμάτων κοστίζει υπολογιστικά, όπως σε εφαρμογές υπολογιστικής βιολογίας.

- **Μάθηση βασισμένη σε δείγματα (Instance-based Learning).** Για το σενάριο αυτό, το οποίο ονομάζεται και μάθηση βασισμένη σε μνήμη (memory-based Learning), αντί να παράγονται άμεσα προβλέψεις, τα νέα δεδομένα συγκρίνονται με τα δεδομένα που έχει επεξεργαστεί το σύστημα κατά την εκπαίδευση και τα οποία έχουν αποθηκευτεί στην μνήμη. Το σενάριο αυτό ανήκει στην κατηγορία της οκνηρής μάθησης (Lazy Learning), στην οποία η διαδικασία μάθησης γίνεται κατά το χρόνο εκτέλεσης του συστήματος (δεν προηγείται κάποια διαδικασία εκπαίδευσης). Το σενάριο αυτό ονομάζεται μάθηση βασισμένη σε δείγματα διότι δημιουργεί υποθέσεις απευθείας από τα δείγματα εκπαίδευσης που διαθέτει.

3.2 Ταξινόμηση (Classification)

Η ταξινόμηση ή κατηγοριοποίηση (Classification) ορίζεται ως το πρόβλημα ανάθεσης μιας προκαθορισμένης κλάσης (Class) σε ένα στιγμιότυπο δείγματος ή πιο γενικά, πρότυπο (pattern). Όπως προείπαμε, θέμα της παρούσης διπλωματικής εργασίας αποτελεί η ταξινόμηση ενός συνόλου προτύπων ανάμεσα σε δύο κλάσεις. Ως κλάση ορίζεται ένα σύνολο ομοειδών αντικειμένων, ενώ ως πρότυπο, ένα διάνυσμα που περιέχει αριθμητικά χαρακτηριστικά ενός αντικειμένου από μία κλάση.

Η δυαδική ταξινόμηση αφορά στην διαδικασία κατά την οποία δοσμένου ενός συνόλου προτύπων Xi = {x1, x2, χ3..., χν} και ενός συνόλου δύο κλάσεων C = {C0, C1}, πρέπει να καθοριστεί σε ποια από τις δύο κλάσεις ανήκει κάθε ένα από τα πρότυπα X. Η επίτευξη της παραπάνω διαδικασίας είναι βασισμένη στην εύρεση μίας συνάρτησης στόχου ή διαχωρισμού (Target / Discriminant Function) f, που απεικονίζει το κάθε σύνολο τιμών ενός αντικειμένου X σε μία από τις δύο προκαθορισμένες κλάσεις, ώστε να είναι δυνατή η ταξινόμηση μελλοντικών προτύπων.

$$y=f(x;w)$$

Η διαδικασία που ακολουθείται με σκοπό την ταξινόμηση προτύπων στην μηχανική μάθηση αποτελείται, αρχικά, από το στάδιο κατασκευής του μοντέλου ταξινόμησης (εκμάθηση), κατά δεύτερον, τον έλεγχο του μοντέλου ταξινόμησης (επικύρωση) και τέλος την εφαρμογή του μοντέλου. Αρχικά, τα δεδομένα (πρότυπα) χωρίζονται σε δύο σύνολα, στο σύνολο εκπαίδευσης (Training Set) και στο σύνολο ελέγχου (Test Set).

Αναφορικά με το πρώτο στάδιο στη διαδικασία της ταξινόμησης, καταρχάς να πούμε ότι συνήθως αναφερόμαστε σε Μάθηση υπό Επίβλεψη, όπως και στην περίπτωση της παρούσας διπλωματικής εργασίας. Στο πρώτο στάδιο λοιπόν, εισάγουμε τα σύνολο δεδομένων μας μαζί με την αντίστοιχη κλάση που αφορά το κάθε στιγμιότυπο. Έπειτα, αναλύουμε τα δεδομένα μας ώστε να αναγνωρίσουμε πιθανές συσχετίσεις μεταξύ χαρακτηριστικών και κλάσεων. Τέλος, κατασκευάζουμε το αρχικό μοντέλο ταξινόμησης μας. Σε αυτό το σημείο, είναι σημαντικό να σημειωθεί ότι είναι καίριας σημασίας η επιλογή των υποσυνόλων για εκπαίδευση και έλεγχο, καθώς μπορεί να έχουμε ως αποτέλεσμα ένα μεροληπτικό μοντέλο.

Αυτό θα φανεί στο δεύτερο στάδιο ταξινόμησης που αφορά στον έλεγχο του μοντέλου μας κάνοντας προβλέψεις για το υποσύνολο ελέγχου. Ο έλεγχος αυτός θα δείξει την απόδοση του ταξινομητή μας στο να προβλέπει τις κλάσεις για στιγμιότυπα για τα οποία οι κλάσεις θεωρούνται άγνωστες. Έπειτα, συγκρίνουμε τις προβλέψεις του μοντέλου με τις πραγματικές κλάσεις στις οποίες ανήκει το υποσύνολο ελέγχου και υπολογίζουμε τις τιμές ακρίβειας πρόβλεψης του μοντέλου μας. Όπως προαναφέρθηκε πολλάκις μέχρι στιγμής στην διπλωματική μας, πρωταρχικός μας στόχος είναι η δημιουργία ενός μοντέλου με την ικανότητα να γενικεύει. Και αυτό σημαίνει να προβλέπει σωστά τις κλάσεις για στιγμιότυπα που δεν έχει αναλύσει κατά την εκπαίδευση.

Ένας σημαντικός παράγοντας στην δημιουργία ενός αποδοτικού μοντέλου είναι φυσικά η αναλογία στιγμιότυπων ανά κλάση και επίσης ο αριθμός των χαρακτηριστικών γιατί όπως τονίσαμε σε προγενέστερο κεφάλαιο, η πολυπλοκότητα του δείγματος μπορεί να μειώσει δραστικά την απόδοση του αλγορίθμου καθώς στατιστικά θα υπάρχουν χαρακτηριστικά μη συσχετιζόμενα με τις κλάσεις και άρα απλά εισάγουν θόρυβο στο δείγμα μας. Για τον λόγο αυτό, ένα από τα σημαντικότερα βήματα στην ταξινόμηση, ειδικά όταν έχουν πολλά χαρακτηριστικά για λίγα στιγμιότυπα είναι η χρήση τεχνικών Επιλογής Χαρακτηριστικών και άρα Μείωσης Διαστάσεων. Αυτές οι τεχνικές αναλύονται παρακάτω.

Αν δεν είμαστε προσεκτικοί κατά την κατασκευή του μοντέλου μας σε σχέση με όλα τα παραπάνω που προαναφέρθηκαν, είναι πολύ πιθανό να οδηγηθούμε στη λεγόμενη υπερπροσαρμογή (Overfitting). Αυτό αφορά μοντέλα ταξινόμησης που δημιουργούν πολύπλοκες συσχετίσεις μεταξύ προτύπων και χαρακτηριστικών που δεν γενικεύουν καλά όταν πρόκειται να προβλέψουν πρότυπα με περιθωριακές τιμές χαρακτηριστικών. Το αντίθετο πρόβλημα που αφορά στην κατασκευή ενός μοντέλου με πολύ απλές συσχετίσεις μεταξύ προτύπων και χαρακτηριστικών (Underfitting) μπορεί να οδηγήσει στο ίδιο μη επιθυμητό αποτέλεσμα.

Για την αποφυγή τέτοιων αποτελεσμάτων, δηλαδή τη δημιουργία ενός είτε απλού είτε πολύπλοκου μοντέλου και φυσικά για την αποφυγή κακής επιλογής υποσυνόλου για εκπαίδευση, χρησιμοποιούνται τεχνικές που κάνουν διασταυρούμενη επικύρωση (Cross Validation). Γενικά η λογική είναι ότι, το σύνολο χωρίζεται σε κ υποσύνολα, τα κ-1 χρησιμοποιούνται για εκπαίδευση και το εναπομείναν για έλεγχο, αυτό επαναλαμβάνεται κ φορές αφήνοντας κάθε φορά ένα διαφορετικό υποσύνολο κ για έλεγχο. Αυτό σημαίνει ότι όλα τα πρότυπα τους δείγματος μας χρησιμοποιούνται τόσο για εκπαίδευση όσο και για έλεγχο. Έτσι υπολογίζεται μια συνολική τιμή ακρίβειας που αφορά όλο το σύνολο μας και θεωρείται πολύ αντικειμενική και αντιπροσωπευτική της ικανότητας του μοντέλου μας. Η τεχνική που περιγράφηκε παραπάνω αφορά στην διασταυρούμενη επικύρωση κ-σετς (K-fold Cross Validation). Αποτελεί μάλιστα την τεχνική διασταυρούμενης επικύρωσης που χρησιμοποιήθηκε στην παρούσα εργασία. Φυσικά υπάρχουν και άλλες που μπορεί να διαφοροποιούνται λίγο σε κάποια βήματα αλλά το αποτέλεσμα και η ουσία είναι λίγο πολύ το ίδιο.

3.3 Κριτήρια Επίδοσης Μοντέλων Μηχανικής Μάθησης (Ταξινόμηση)

Υπάρχουν πολλές μέθοδοι και κριτήρια επίδοσης της ταξινόμησης ταξινομητών στη Μηχανική Μάθηση. Αν θεωρήσουμε ένα πρόβλημα ταξινόμησης δύο κλάσεων (C0 και C1), και συνήθως χρησιμοποιούμε ένα δυαδικό διαχωρισμό άρα θεωρούμε τη μια κλάση αρνητική και την άλλη θετική, ένας ταξινομητής δημιουργεί μια συνάρτηση διαχωρισμού και σε ένα ιδανικό σενάριο όλα τα πρότυπα της μίας κλάσης θα ήταν στην μία πλευρά ενώ τα υπόλοιπα στην άλλη πλευρά. Αυτό φυσικά δεν συμβαίνει συνήθως, τουναντίον διακρίνουμε τέσσερις περιπτώσεις. Η πρώτη αφορά τα πραγματικά αρνητικά (True Negatives) δηλαδή ο ταξινομητής ανέθεσε σε ένα πρότυπο την αρνητική κλάση (C0) και όντως το πρότυπο ανήκε εκεί. Έπειτα έχουμε τα εσφαλμένα αρνητικά (False Negatives) δηλαδή ο ταξινομητής ανέθεσε σε ένα πρότυπο την αρνητική κλάση (C0) ενώ ανήκουν στη θετική κλάση. Αντίστοιχα, έχουμε την περίπτωση των πραγματικών θετικών (True Positives) όπου ο ταξινομητής ανέθεσε ένα πρότυπο στη θετική κλάση (C1) το οποίο ανήκει στη θετική κλάση. Τέλος, έχουμε την περίπτωση των εσφαλμένων θετικών (False Positives) όπου ο ταξινομητής ανέθεσε ένα πρότυπο στην θετική κλάση (C1) ενώ αυτό ανήκει στην αρνητική.

Ουσιαστικά, ο τρόπος αυτός αξιολόγησης της επίδοσης ενός ταξινομητή συνοψίζεται στον παρακάτω πίνακα που ονομάζεται Πίνακας Σύγχυσης (Confusion Matrix).

	Ταξινομήθηκαν στην Κλάση C0	Ταξινομήθηκαν στην Κλάση C1
Ανήκουν στην Κλάση C0	True Negatives (TN)	False Positives (FP)
Ανήκουν στην Κλάση C1	False Negatives (FN)	True Positives (TP)

Όπως είναι λογικό, στην ιδανική περίπτωση θα θέλαμε, ο αριθμός των εσφαλμένων αρνητικών και θετικών να είναι 0 (False Negatives (FN) = False Positives (FN) = 0). Η ακρίβεια λοιπό, που αποτελεί μία από τις συνηθέστερες και συνάμα σημαντικότερες μετρικές επίδοσης δίνεται από τον παρακάτω τύπο:

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP}$$

Η ακρίβεια παίρνει τιμές μεταξύ 0,1 δηλαδή 0 <= ακρίβεια <= 1. Όταν FN και FP ισούνται με μηδέν, τότε η ακρίβεια ισούται με ένα. Το πρόβλημα της παραπάνω μετρικής είναι το γεγονός ότι είναι πολύ γενικευμένη. Όταν όμως για παράδειγμα, ο αριθμός των προτύπων ανά κλάση δεν είναι ισορροπημένος, η τιμή ακρίβειας μπορεί να είναι παραπλανητική. Επειδή λοιπόν και το δείγμα μας δεν είναι καθόλου ισορροπημένο, δεν θα μπορούσαμε να

στηριχθούμε στα ευρήματα της ακρίβειας για την αξιολόγηση της επίδοσης του μοντέλου μας. Για το λόγο αυτό, έχουν οριστεί ακόμα δυο τύποι που κάνουν χρήση των TP, FP, FN και είναι περισσότερα αντικειμενικά, συγκεκριμένα:

$$Precision = \frac{TP}{CLASSIFIED\ AS\ POSITIVE} = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{ALL\ POSITIVE} = \frac{TP}{TP + FN}$$

Το κριτήριο Precision είναι ο λόγος του αριθμού των προτύπων που ταξινομήθηκαν ως θετικά και ανήκουν όντως στην κλάση 1, προς όλα όσα ταξινομήθηκαν θετικά. Αντίστοιχα, το κριτήριο Recall είναι ο λόγος του αριθμού των προτύπων που ταξινομήθηκαν ως θετικά και ανήκουν όντως στην κλάση 1, προς όλα τα θετικά (ανήκουν στην κλάση 1). Οι τιμές των δύο παραπάνω μετρικών επίσης κυμαίνονται μεταξύ 0 και 1. Στην ιδανική περίπτωση, θα είχαμε Precision = Recall =1. Και πάλι όμως, ακόμα και με τη χρήση των Precision, Recall, δεν είμαστε σε θέση να εκτιμήσουμε επαρκώς έναν ταξινομητή. Μία άλλη μετρική αξιολόγησης είναι το κριτήριο F-measure ή αλλιώς F1-Score και δίνεται από τον παρακάτω τύπο:

$$F - measure = 2\ x\ \frac{Precision\ x\ Recall}{(Precision + Recall)}$$

Το εύρος των τιμών του F-measure είναι κυμαίνεται επίσης μεταξύ 0 και 1. Σε μια ιδανική περίπτωση θα είχαμε F-measure = 1.

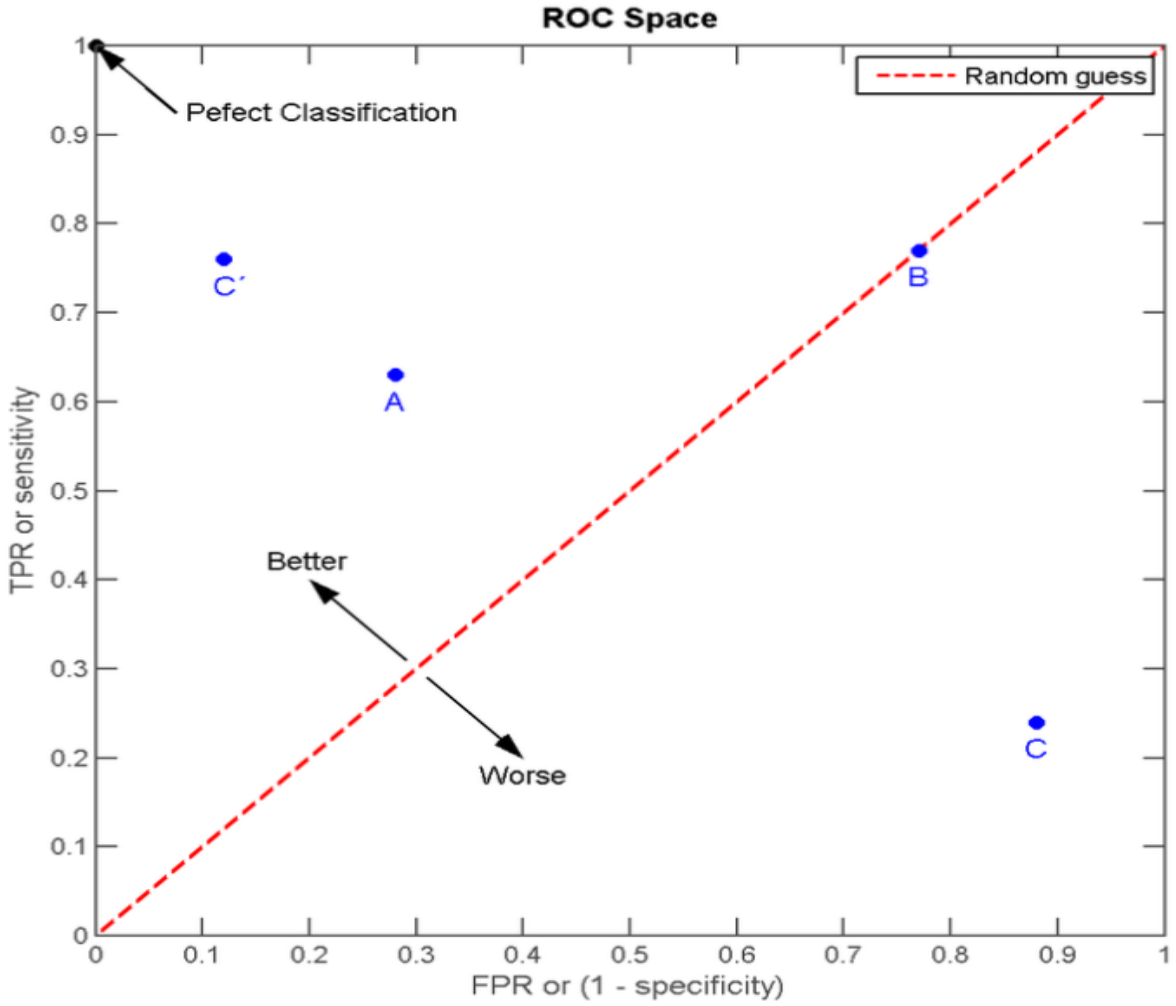
Σε περίπτωση που δεν το διαπιστώσατε, να πούμε εδώ πως ένα μεγάλο μειονέκτημα των μετρικών Precision και Recall είναι πως εστιάζουν αποκλειστικά στη θετική κλάση (C1). Για το λόγο αυτό, έχουν οριστεί δύο ακόμα κριτήρια επίδοσης που αφορούν και τις δύο κλάσεις:

$$Sensitivity = \frac{TP}{ALL\ POSITIVE} = \frac{TP}{TP + FN}\ (= Recall = True\ Positive\ Rate)$$

$$Specificity = \frac{TN}{ALL\ NEGATIVE} = \frac{TN}{TN + FP} (= True\ Negative\ Rate)$$

Το κριτήριο Sensitivity ή αλλιώς True Positive Rate (TPR) είναι ουσιαστικά το γνωστό μας Recall. Όσο για τη μετρική Specificity, είναι αλλιώς γνωστή ως True Negative Rate (TNR) και είναι ο λόγος του αριθμού των προτύπων που ταξινομήθηκαν ως αρνητικά και ανήκουν όντως στην κλάση 0, προς όλα τα αρνητικά (ανήκουν στην κλάση 0). Ουσιαστικά οι δύο αυτές μετρικές είναι το ίδιο με μοναδική εξαίρεση την κλάση στην οποία εστιάζουν (Sensitivity για την κλάση 1 και Specificity για την κλάση 0).

Και πάλι όμως, θεωρείται ότι τα προαναφερθέντα κριτήρια, από μόνα τους δεν είναι ικανά να εκτιμήσουν πλήρως την απόδοση ενός ταξινομητή. Ένα ακόμα κριτήριο, το Receiver Operating Characteristic (ROC), συνδυάζει τα ευρήματα των Sensitivity και Specificity σε ένα γράφημα. Πιο συγκεκριμένα κάνει χρήση του Sensitivity σε σχέση με το 1 – Specificity ή αλλιώς False Positive Rate.

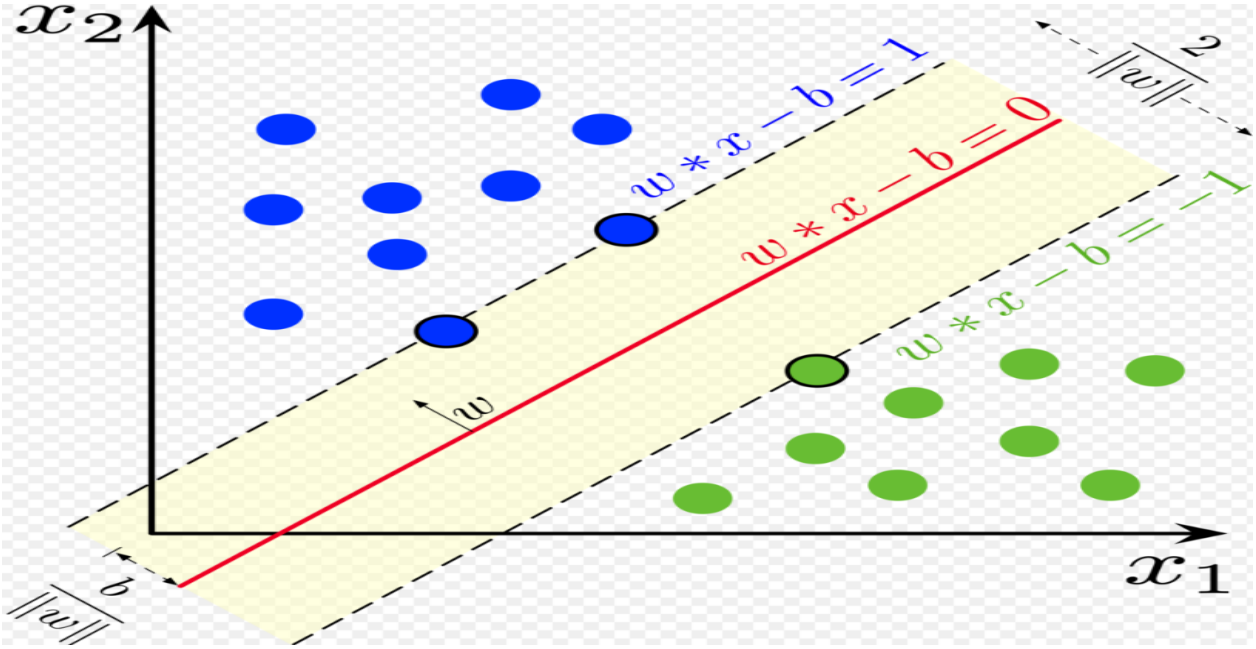


Στην παραπάνω εικόνα βλέπουμε στην κόκκινη διακεκομμένη διαγώνιο την επίδοση ενός τυχαίου ταξινομητή. Κάτω από τη διαγώνιο έχουμε τις χειρότερες τιμές ενώ από πάνω έχουμε τις καλύτερες. Ένας ιδανικός ταξινομητής θα ήταν εντελώς πάνω και αριστερά ώστε Sensitivity =1 και 1 – Specificity = 0.

3.4 Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines)

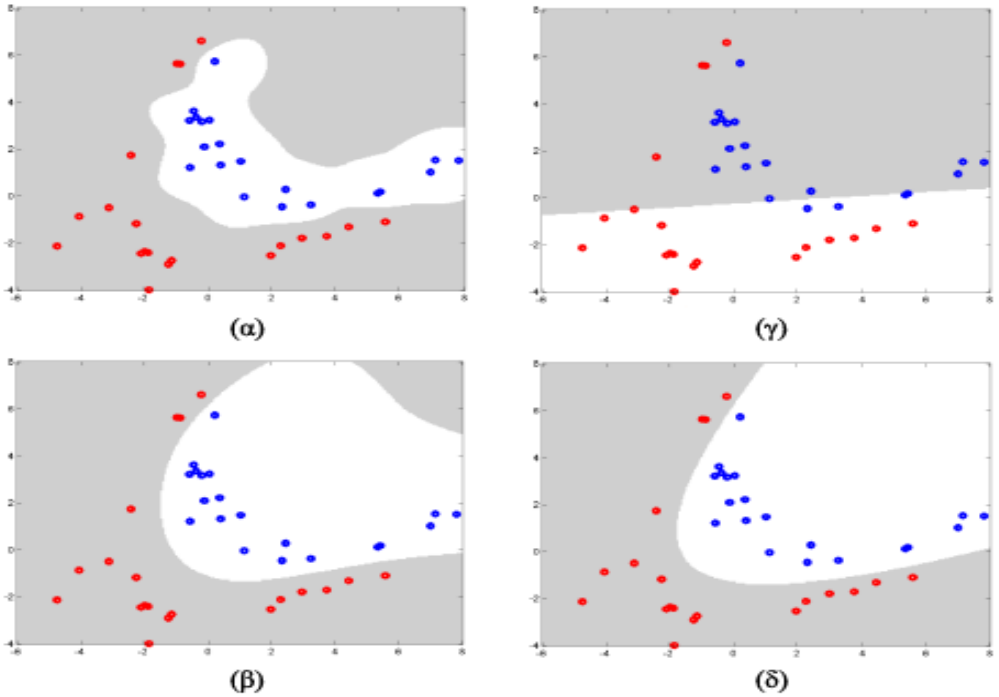
Ο Vladimir Vapnik, το μακρινό 1995, ήταν ο πρώτος που πρότεινε μία μέθοδο Μηχανικής Μάθησης (με επίβλεψη) για προβλήματα ταξινόμησης (Classification) και παλινδρόμησης (Regression), τις Μηχανές Διανυσμάτων Υποστήριξης [21]. Η πρόταση του έλαβε ερεθίσματα φυσικά από τη Στατιστική αλλά και από νευρωνικά δίκτυα τύπου Perceptron.

Ο συγκεκριμένος αλγόριθμος προσπαθεί να εντοπίσει το βέλτιστο υπέρ-επίπεδο του χώρου των χαρακτηριστικών (Features) το οποίο να διαχωρίζει καλύτερα τα αρνητικά από τα θετικά παραδείγματα προτύπων. Μάλιστα, ο διαχωρισμός γίνεται με γνώμονα το όσο το δυνατόν μεγαλύτερο περιθώριο (margin) μεταξύ των περιθωριακών παραδειγμάτων για όλες τις κλάσεις. Το τελευταίο είναι περισσότερο γνωστό με τον όρο Maximum Margin Hyperplane [22]. Τα παραδείγματα που απέχουν λιγότερο από το διαχωρισμού του υπέρ-επιπέδου, είναι και εκείνα που δίνουν το όνομα τους στον αλγόριθμο, δηλαδή ονομάζονται διανύσματα υποστήριξης (Support Vectors).



Με την κόκκινη διαχωριστική γραμμή βλέπουμε τον εντοπισμό του βέλτιστου υπέρ-επιπέδου που έχει το μεγαλύτερο περιθώριο για όλα τα παραδείγματα ανά κλάση.

Μέχρι τώρα μιλούσαμε κυρίως για γραμμικώς διαχωρίσιμα προβλήματα. Οι Μηχανές Διανυσμάτων Υποστήριξης, ωστόσο, μπορούν να ανταπεξέρθουν και σε περιπτώσεις μη-γραμμικώς διαχωρίσιμων προβλημάτων. Τότε, αναλαμβάνουν δράση οι λεγόμενες συναρτήσεις πυρήνα (Kernel Functions) προς εντοπισμού ενός βέλτιστου υπέρ-επιπέδου σε ένα μετασχηματισμένο αυτή τη φορά, χώρου χαρακτηριστικών.



Στην παραπάνω εικόνα βλέπουμε τρόπους επίλυσης του ίδιου προβλήματος με χρήση διαφορετικών συναρτήσεων πυρήνα. (α)Γκαουσιανός σ^2=1 (β) Γκαουσιανός σ^2=10 (γ)Γραμμικός (δ)Πολυωνυμικός (τετραγωνικός)

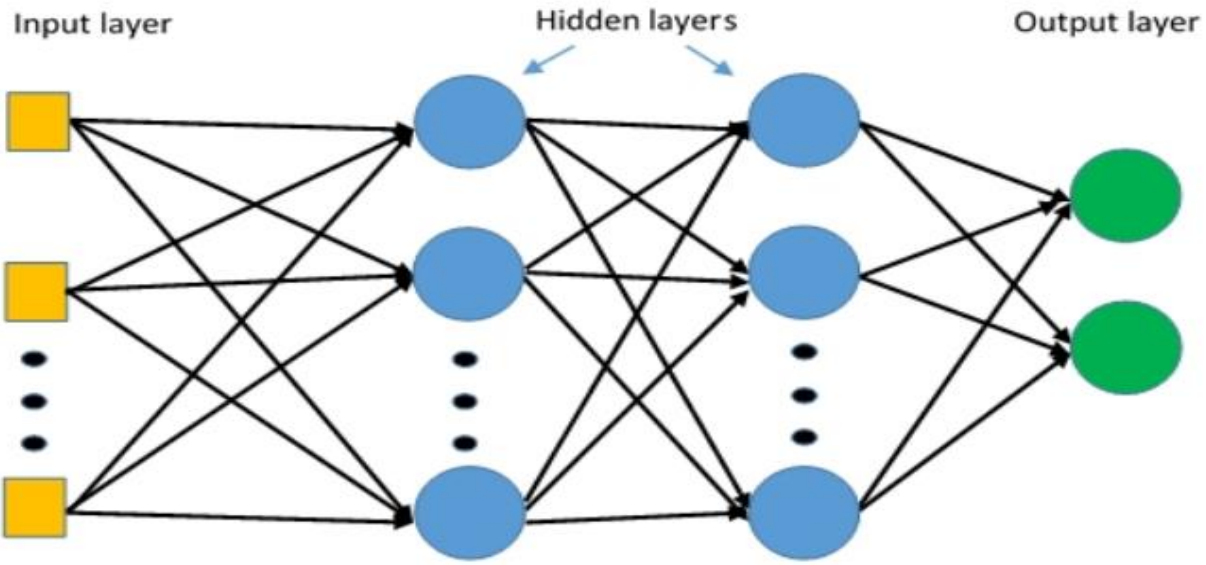
Γενικά, κάθε συνάρτηση πυρήνα χρησιμοποιεί τον δικό της τύπο:

$e^{-\|x-y\|^2/(2\sigma^2)}$	Γκαουσιανή (RBF)
$[\boldsymbol{x}^T\boldsymbol{y}+\theta]^p$	Πολυωνυμική
$\tanh(\alpha\,\boldsymbol{x}^T\mathbf{y}+\theta)$	Σιγμοειδής

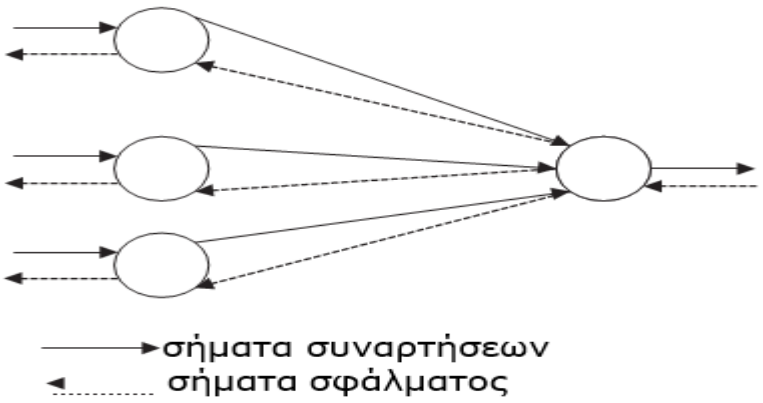
Οι Μηχανές Υποστήριξης Διανυσμάτων αποτελούν αδιαμφησβήτητα σήμερα μία από τις περισσότερο γνωστές και περισσότερο χρησιμοποιούμενες μεθόδους ταξινόμησης γραμμικών ή μη προβλημάτων. Αυτό συμβαίνει διότι διακρίνονται για την αποτελεσματικότητα και συνάμα ταχύτητα τους, όπως επίσης και γιατί είναι ιδιαίτερα αποδοτικά στο να εντοπίζουν μη γραμμικά υπέρ-επίπεδα.

3.5 Πολύ-επίπεδα Perceptron (Multi-Layer Perceptron)

Τα Πολύ-επίπεδα Perceptron είναι ουσιαστικά, δίκτυα τα οποία αποτελούνται από ένα σύνολο κόμβων εισόδου, το λεγόμενο επίπεδο εισόδου (input layer), ένα ή περισσότερα κρυφά επίπεδα νευρώνων (hidden layers) οι οποίοι είναι υπεύθυνοι για τους υπολογισμούς, και τέλος το επίπεδο εξόδου (output layer) που αποτελούν την έξοδο [23].



Σε αυτό το δίκτυο, συναντάμε δύο είδη σημάτων, τα σήματα συναρτήσεων και τα σήματα σφάλματος. Το πρώτο, εφαρμόζεται στην είσοδο, προχωράει προς τα κρυφά επίπεδα νευρώνων και τελικά εμφανίζεται στην έξοδο του δικτύου ως σήμα εξόδου. Το όνομα του έχει να κάνει με την υπόθεση ότι αποτελεί μια χρήσιμη συνάρτηση στην έξοδο του δικτύου. Αντίθετα, τα σήματα σφάλματος, παράγονται στην έξοδο και διαδίδονται προς τα πίσω. Το όνομα του έχει να κάνει με το γεγονός ότι ο υπολογισμός του περιλαμβάνει τη διαφορά της πραγματικής εξόδου σε σχέση με το επιθυμητό αποτέλεσμα (σφάλμα).



Κάθε Νευρώνας ενός τέτοιου δικτύου που βρίσκεται στο κρυφό στρώμα ή στο στρώμα εξόδου κάνει δύο υπολογισμούς, το σήμα συνάρτησης στην έξοδο το οποίο εκφράζεται σαν μια συνεχής γραμμική συνάρτηση του σήματος εισόδου και των βαρών, και τέλος την εκτίμηση της κλίσης.

Τα πολύ-επίπεδα Perceptron χρησιμοποιούνται συνήθως στη Μηχανική Μάθηση με Επίβλεψη και κάνουν χρήση του αλγορίθμου Ανάστροφης Μετάδοσης Σφάλματος (Error Back-Propagation). Ο αλγόριθμος με την διάδοση του σφάλματος προς τα πίσω πετυχαίνει την εύρεση των κατάλληλων συντελεστών βαρών. Η εκμάθηση μπορεί να γίνει αντίστοιχα με τη ροή των σημάτων, είτε σε ευθεία κατεύθυνση είτε σε ανάστροφη [23].

Στην ευθεία κατεύθυνση, από ένα διάνυσμα εισόδου, υπολογίζονται οι έξοδοι των νευρώνων. Οι τιμές των βαρών διατηρούνται σταθερές.

Στην ανάστροφη κατεύθυνση, οι τιμές των βαρών ρυθμίζονται βάσει κάποιου κανόνα με στόχο την εξάλειψη του σφάλματος.

Τέλος, να πούμε εδώ ότι μια επανάληψη του αλγορίθμου για ολόκληρο το σύνολο δεδομένων εκπαίδευσης ονομάζεται εποχή (epoch). Τις πιο πολλές φορές, χρειάζονται πολλές επαναλήψεις (εποχές) για την ολοκλήρωση εκπαίδευσης ενός νευρωνικού δικτύου.

3.6 Μείωση Διαστάσεων (Dimension Reduction)

Οι αλγόριθμοι Μηχανικής Μάθησης δημιουργήθηκαν σταδιακά τις περασμένες δεκαετίες με σκοπό την επίλυση πολλών πολύπλοκων προβλημάτων όχι μόνο στον τομέα της Βιοπληροφορικής και της υγείας αλλά και σε διάφορα άλλα πεδία. Η απαρχή και αιτία της δημιουργίας των αλγορίθμων μηχανικής μάθησης λοιπόν αποτελεί μια από τις μεγαλύτερες προκλήσεις στην βελτιστοποίηση της αποτελεσματικότητας τους. Η πολυπλοκότητα των προβλημάτων που αντιμετωπίζουν οι αλγόριθμοι μηχανικής μάθησης, αφορά στην ύπαρξη μεγάλου αριθμού διαστάσεων στα διάφορα σύνολα δεδομένων. Όσο λοιπόν αυξάνεται ο αριθμός των χαρακτηριστικών τόσο δυσκολότερο είναι για τους αλγορίθμους να αποδίδουν και να γενικεύουν. Όπως προαναφέρθηκε σε προηγούμενο κεφάλαιο, το πρόβλημα αυτό είναι ευρέως γνωστό με τον όρο «κατάρα των πολλών διαστάσεων» ή αλλιώς «κατάρα της Διαστασιμότητας». Είναι αναγκαία λοιπόν η προσπάθεια μείωσης των διαστάσεων με τη δημιουργία υποσυνόλων χαρακτηριστικών. Χρησιμοποιείται συνήθως σε προβλήματα ταξινόμησης (Classification) ή ομαδοποίησης (Clustering).

Πολλές φορές, ένα σύνολο δεδομένων περιέχει χαρακτηριστικά, των οποίων η πληροφορία δεν βοηθάει στο να καθοριστεί η κλάση ενός στιγμιότυπου. Αυτό μπορεί να συμβαίνει γιατί η εν λόγω πληροφορία είναι είτε άσχετη με το αποτέλεσμα είτε υπάρχει ήδη σε κάποιο άλλο χαρακτηριστικό. Όπως προείπαμε, στόχος μιας μεθόδου μείωσης διαστάσεων είναι η εύρεση του βέλτιστου υποσυνόλου χαρακτηριστικών που αποτελείται από εκείνα τα χαρακτηριστικά που περιέχουν την πιο σημαντική πληροφορία με την οποία ο αλγόριθμος μπορεί να ταξινομήσει ένα πρότυπο σε μία κλάση [24].

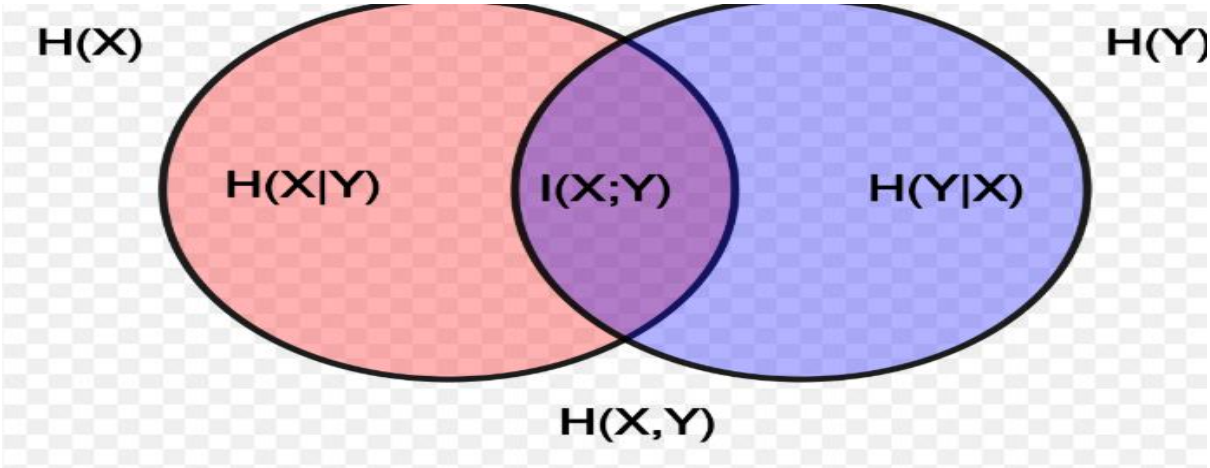
Με την εξάλειψη των περιττών και άσχετων χαρακτηριστικών, κάνουμε ένα πολύ σημαντικό βήμα στη δημιουργία ενός αποδοτικού ταξινομητή. Αρχικά, μειώνουμε την πιθανότητα υπερπροσαρμογής (overfitting) του μοντέλου μας. Κατά δεύτερον, με μικρότερο αριθμό χαρακτηριστικών το πρόβλημα είναι περισσότερο κατανοητό και ερμηνεύσιμο ακόμα και από το μάτι του ανθρώπου. Τέλος, πρέπει πάντα να συνυπολογίζουμε και το υπολογιστικό κόστος το οποίο στην προκειμένη περίπτωση μπορεί να μειωθεί δραματικά.

Οι κύριοι τρόποι Μείωσης Διαστάσεων των χαρακτηριστικών είναι πρώτον η Εξαγωγή Χαρακτηριστικών (Feature Extraction, FE) κατά την οποία, το σύνολο των χαρακτηριστικών αναλύεται, μετασχηματίζεται και προκύπτουν νέα χαρακτηριστικά, ασυσχέτιστα μεταξύ τους και σαφώς λιγότερα από πριν, και δεύτερον η Επιλογή ενός Υποσυνόλου Χαρακτηριστικών (Subset Feature Selection, SFS) κατά την οποία, εξαλείφονται τα χαρακτηριστικά που περιέχουν άσχετη πληροφορία (σε σχέση με τη μεταβλητή-στόχο) ή παρόμοια με άλλα, και διατηρούνται μόνον όσα περιέχουν σημαντική πληροφορία. Μια συνηθισμένη μέθοδος Εξαγωγής Χαρακτηριστικών είναι η Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis, PCA). Για την επιλογή υποσυνόλου Χαρακτηριστικών οι δυο βασικές προσεγγίσεις είναι οι wrapper approach και filter approach. Για την πρώτη, ένας ταξινομητής χρησιμοποιείται για τη διαδικασία επιλογής, ενώ για τη δεύτερη χρησιμοποιείται το εκάστοτε κριτήριο αξιολόγησης [24].

Παρακάτω θα αναφερθούμε στις τρείς μεθόδους επιλογής χαρακτηριστικών που χρησιμοποιήθηκαν στην παρούσα διπλωματική και συγκεκριμένα στην Αμοιβαία Πληροφορία (Mutual Information), στο κριτήριο Kolmogorov-Smirnov 2 δειγμάτων(KS 2 Samples Test) και τέλος στην Επαναλαμβανόμενη Εξάλειψη Χαρακτηριστικών με χρήση Διασταυρούμενης Επικύρωσης (Recursive Feature Elimination with Cross Validation, RFE-CV)

3.6.1 Αμοιβαία Πληροφορία (Mutual Information, MI)

Ή Αμοιβαία Πληροφορία (Mutual Information) είναι από τις βασικότερες μεθόδους Επιλογής Χαρακτηριστικών. Σχετίζεται με την εντροπία (entropy) της πληροφορίας η οποία αποτελεί μέτρο αβεβαιότητας της τιμής μιας τυχαίας μεταβλητής και βασίζεται ολοκληρωτικά στην κατανομή πιθανότητας της μεταβλητής. Όσο λοιπόν μεγαλώνει η εντροπία για μια μεταβλητή, τόσο μεγαλώνει και η αβεβαιότητα γύρω από αυτή, άρα και λιγότερες οι πιθανότητες μας για σωστή πρόβλεψη με βάση αυτή την μεταβλητή [25]. Η ποσότητα της πληροφορίας μετριέται σε bits. Η αμοιβαία πληροφορία βέβαια σχετίζεται περισσότερο με τη λεγόμενη δεσμευμένη εντροπία (Conditional Entropy) που είναι η τιμή μιας τυχαίας μεταβλητής με βάση την τιμή μιας άλλης τυχαίας μεταβλητής. Από την άλλη, η διαφορά της αβεβαιότητας για μια τυχαία μεταβλητή μείον τη διαφορά της αβεβαιότητας για μια τυχαία μεταβλητή με δεδομένη μια άλλη τυχαία μεταβλητή είναι το λεγόμενο Κέρδος Πληροφορίας (Information Gain).



Στην παραπάνω εικόνα βλέπουμε μια γραφική απεικόνιση μεταξύ των σχέσεων της (δεσμευμένης) εντροπίας (H(X|Y) και (H(Y|X)) και της αμοιβαίας πληροφορίας (I(X,Y)) για δύο τυχαίες μεταβλητές X,Y.

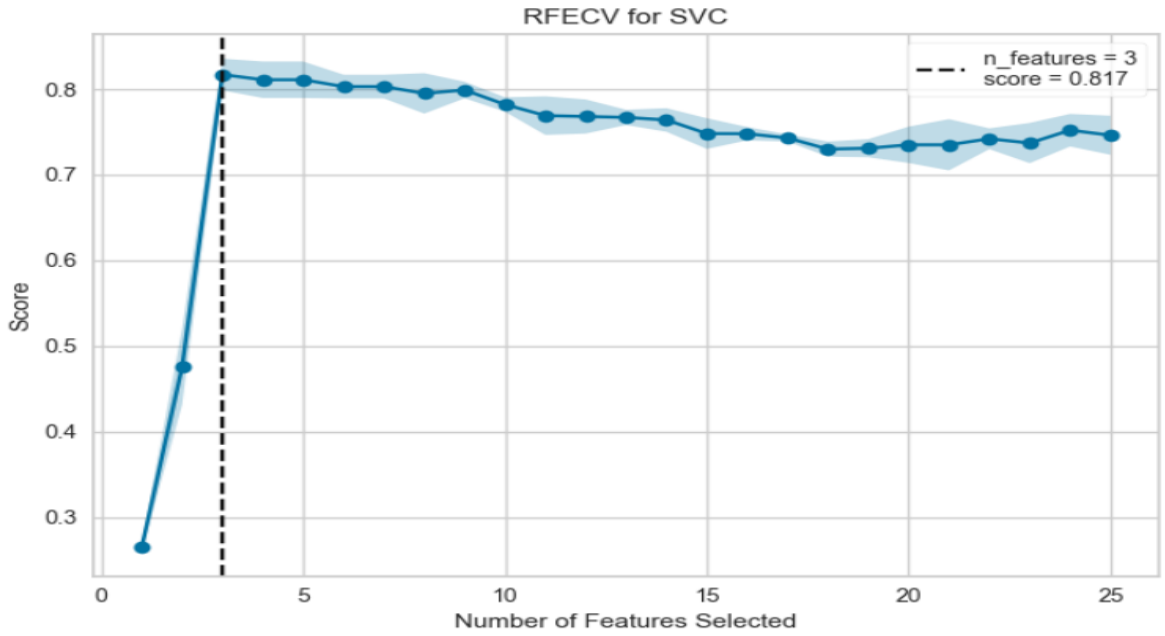
3.6.2 Κριτήριο Kolmogorov-Smirnov 2 δειγμάτων (KS 2 Samples Test)

Το κριτήριο Kolmogorov-Smirnov 2 δειγμάτων συγκρίνει τις κατανομές δύο συνόλων δειγμάτων και ελέγχει αν δύο δείγματα αποτελούμενα από συνεχείς τιμές ανήκουν στην ίδια κατανομή [26]. Πρακτικά λοιπόν, χωρίζουμε το δείγμα μας στα δύο, με την κλάση 0 να είναι στο πρώτο υποσύνολο και την κλάση 1 στο δεύτερο. Έπειτα, για κάθε χαρακτηριστικό, συγκρίνονται οι κατανομές που προέρχονται από το πρώτο δείγμα, με τις κατανομές που προέρχονται από το δεύτερο. Αν δοθέντος ενός χαρακτηριστικού, οι δυο αυτές κατανομές είναι ίδιες ή παρόμοιες, σημαίνει ότι το συγκεκριμένο χαρακτηριστικό συμπεριφέρεται το ίδιο ανεξάρτητα κλάσης, άρα δεν μας προσφέρει κάποια πληροφορία που μπορούμε να εκμεταλλευτούμε. Αντίστοιχα, αν για ένα χαρακτηριστικό, οι δύο κατανομές που προκύπτουν είναι διαφορετικές, σημαίνει ότι το χαρακτηριστικό αυτό συμπεριφέρεται διαφορετικά ανάλογα με την κλάση στην οποία ανήκει το πρότυπο στο οποίο εκφράζεται, και άρα μπορούμε να το χρησιμοποιήσουμε για άντληση πληροφοριών για αποτελεσματικότερη ταξινόμηση.

3.6.3 Επαναλαμβανόμενη Εξάλειψη Χαρακτηριστικών με χρήση Διασταυρούμενης Επικύρωσης (Recursive Feature Elimination with Cross Validation, RFE-CV)

Η Επαναλαμβανόμενη Εξάλειψη Χαρακτηριστικών (RFE) είναι μια μέθοδος επιλογής χαρακτηριστικών που χρησιμοποιεί ένα μοντέλο και αφαιρεί τα πιο αδύναμα χαρακτηριστικά (αυτά δηλαδή που παρέχουν την λιγότερο χρήσιμη πληροφορία) έως ότου επιτευχθεί ο καθορισμένος αριθμός χαρακτηριστικών. Τα χαρακτηριστικά κατατάσσονται από τις τιμές coefficient ή feature_importances_attributes του εκάστοτε μοντέλου και εξαλείφοντας επαναλαμβανόμενα έναν μικρό αριθμό χαρακτηριστικών ανά βρόχο, η RFE επιχειρεί να εξαλείψει τις εξαρτήσεις και την συγγραμικότητα που μπορεί να υπάρχουν στο μοντέλο.

Η RFE απαιτεί να διατηρηθεί ένας καθορισμένος αριθμός χαρακτηριστικών, ωστόσο συχνά δεν είναι γνωστό εκ των προτέρων πόσα χαρακτηριστικά πρέπει να διατηρηθούν. Για να βρεθεί ο βέλτιστος αριθμός χαρακτηριστικών, η διασταυρούμενη επικύρωση χρησιμοποιείται σε συνδυασμό με την RFE για να βαθμολογήσει διαφορετικά υποσύνολα χαρακτηριστικών και να επιλέξει την καλύτερη συλλογή χαρακτηριστικών. Το γράφημα της RFECV καταγράφει τον αριθμό των χαρακτηριστικών του μοντέλου μαζί με την τιμή ακρίβειας ελέγχου και τη μεταβλητότητα με χρήση της διασταυρούμενης επικύρωσης και απεικονίζει τον επιλεγμένο αριθμό χαρακτηριστικών.



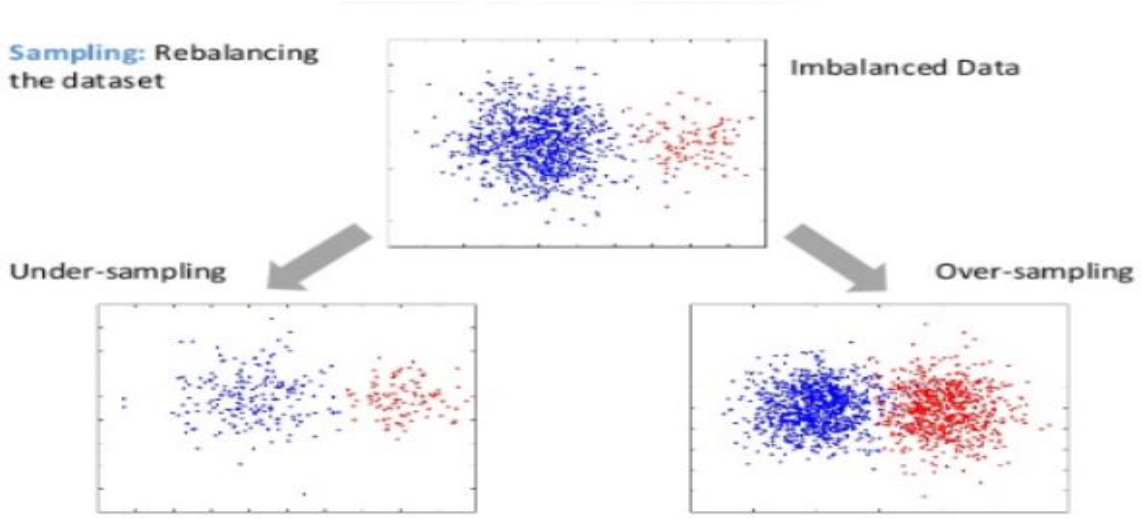
Το παραπάνω σχήμα δείχνει μια ιδανική καμπύλη RFE-CV, η καμπύλη έχει βέλτιστη ακρίβεια όταν χρησιμοποιούνται τα τρία σημαντικά χαρακτηριστικά, στη συνέχεια μειώνεται σταδιακά η ακρίβεια καθώς τα μη σχετικά χαρακτηριστικά προστίθενται στο μοντέλο. Η σκιασμένη περιοχή αντιπροσωπεύει τη μεταβλητότητα της διασταυρούμενης επικύρωσης, μία τυπική απόκλιση πάνω και κάτω από τη μέση τιμή ακρίβειας που έχει σχεδιαστεί από την καμπύλη.

3.7 Εξισορρόπηση του δείγματος

Τα μη ισορροπημένα δείγματα υπάρχουν παντού. Πρόσφατα η εξισορρόπηση των δειγμάτων χρησιμοποιώντας τεχνικές υπερδειγματοληψίας για τη μειοψηφούσα κλάση έχει γίνει μια συνήθης προσέγγιση για τη βελτίωση της ποιότητας των προβλέψεων. Με την υπερδειγματοληψία, τα μοντέλα Μηχανικής Μάθησης είναι μερικές φορές πιο ικανά να εντοπίσουν μοτίβα που διαφοροποιούν κλάσεις μεταξύ τους. Η υπερδειγματοληψία είναι ένας γνωστός τρόπος βελτίωσης μοντέλων εκπαιδευμένων σε μη ισορροπημένα δεδομένα.

3.7.1 Συνθετική Υπερδειγματοληψία Μειοψηφούσας κλάσης με ταυτόχρονη Υποδειγματοληψία για περιθωριακά στιγμιότυπα με χρήση Προσαρμοσμένων Κοντινότερων Γειτόνων

Ο αλγόριθμος SMOTE εφαρμόζει την τεχνική KNN (K Nearest Neighbors) όπου επιλέγει κ κοντινότερους γείτονες, τους συνδέει και δημιουργεί συνθετικά στιγμιότυπα στο χώρο. Ο αλγόριθμος λαμβάνει τα διανύσματα των χαρακτηριστικών και των κοντινότερων γειτόνων και υπολογίζει την απόσταση μεταξύ των διανυσμάτων αυτών. Η διαφορά πολλαπλασιάζεται με έναν τυχαίο αριθμό μεταξύ 0,1 και προστίθεται στο χαρακτηριστικό. Αποτελεί έναν πρωτοποριακό αλγόριθμο και πολλοί άλλοι αντίστοιχοι του προέρχονται από αυτόν.



Η επέκταση του αλγορίθμου SMOTE, ο SMOTEENN, είναι πανομοιότυπος, με τη μόνη διαφορά ότι αφαιρεί στιγμιότυπα των οποίων η κλάση είναι διαφορετική σε σχέση με δύο από τους τρεις κοντινότερους γείτονες του, δηλαδή δεν λαμβάνει υπόψιν του περιθωριακά στιγμιότυπα. Η ENN εξαλείφει στιγμιότυπα της πλειοψηφούσας κλάσης, για τα οποία, η πρόβλεψη τους από τον KNN είναι διαφορετική από αυτήν της πλειοψηφούσας κλάσης. Ο ENN μπορεί να αφαιρέσει στιγμιότυπα που αποτελούν θόρυβο παρέχοντας καλύτερα πιθανότητα σωστών αποφάσεων.

4. Ανάλυση βιολογικών δεδομένων με χρήση αλγορίθμων Μηχανικής Μάθησης με εφαρμογή στη διάγνωση του γαστρεντερικού καρκίνου

Σε αυτό το κεφάλαιο, αρχικά περιγράφεται το σύνολο δεδομένων που χρησιμοποιήθηκε στην παρούσα εργασία και αφορά στον γαστρεντερικό καρκίνο. Στη συνέχεια περιγράφονται οι μέθοδοι που χρησιμοποιήθηκαν για τον καθορισμό των συσχετίσεων των χαρακτηριστικών και των μεταβλητών-στόχων ώστε να προκύψει η κατάλληλη επιλογή χαρακτηριστικών. Επιπλέον, γίνεται μία περιγραφή κάποιων από τους αλγορίθμους που χρησιμοποιήσαμε και κάποιων άλλων μεθόδων όπως εξαντλητική αναζήτηση κ.λπ.

Η διενέργηση των πειραμάτων έγινε χρησιμοποιώντας την γλώσσα προγραμματισμού Python, και συγκεκριμένα της έκδοσης 3. Η επιλογή αυτή έγινε καθώς υπάρχει πληθώρα βιβλιοθηκών αναφορικά με αλγορίθμους Μηχανικής Μάθησης, στατιστικές αναλύσεις, μεθόδους επιλογής χαρακτηριστικών και οπτικοποίησης δεδομένων όπως Scikit-learn, Numpy, Matplotlib, imbalanced-learn και άλλες.

4.1 Περιγραφή Συνόλου Δεδομένων

Το σύνολο δεδομένων που χρησιμοποιήσαμε στην παρούσα εργασία αφορά τον γαστρεντερικό καρκίνο και συγκεκριμένα τέσσερα από τα πέντε είδη του (τον οισοφαγικό(esophageal), τον στομαχικό(stomach), τον παγκρεατικό(pancreatic) και της χοληδόχου κύστης(gallbladder). Το δείγμα προήρθε από τη βάση δεδομένων TCGA (The Cancer Genome Atlas).

Λάβαμε το δείγμα από το Ινστιτούτο Εφαρμοσμένων Βιοεπιστημών (INAB) του Εθνικού Κέντρου Έρευνας και Τεχνολογικής Ανάπτυξης (ΕΚΕΤΑ). Το δείγμα, αφού πρώτα επεξεργάστηκε καταλλήλως (filtering, transformation) ήταν έτοιμο για την διεξαγωγή των πειραμάτων μας.

Τα δεδομένα μας περιέχονται σε δύο αρχεία. Το ένα αρχείο είναι ένας πίνακας του οποίου οι στήλες αποτελούν τα γονίδια (χαρακτηριστικά) και οι γραμμές αφορούν τα πρότυπα (στιγμιότυπα) ανθρώπων. Οι τιμές των γονιδίων είναι ουσιαστικά τιμές έκφρασης μετασχηματισμένες από αλληλλούχιση RNA (RNA – Seq). Η αλληλλούχιση RNA αποτελεί μια μέθοδο ποσοτικοποίησης βιολογικών δεδομένων ώστε να μπορεί να ερευνηθεί η διαφοροποίηση στη συμπεριφορά τους. Ο αριθμός των προτύπων (στιγμιότυπων) είναι 1065 και ο αριθμός των κοινών γονιδίων για τους τέσσερις τύπους γαστρεντερικού καρκίνου είναι 20501.

Το δεύτερο αρχείο είναι ένας πίνακας μίας γραμμής και 1065 στηλών. Η γραμμή αφορά την ύπαρξη ή μη της ασθένειας του καρκίνου (cancer) ενώ οι στήλες αφορούν τον αριθμό των προτύπων (στιγμιότυπων). Όπως είναι προφανές, στην παρούσα διπλωματική θα ασχοληθούμε με την στήλη της ύπαρξης ή μη του καρκίνου σε σχέση με τα γονίδια (χαρακτηριστικά) για κάθε στιγμιότυπο.

Οι τιμές της γραμμής cancer παίρνουν την τιμή 0 και 1. Η κλάση 0 αφορά τους υγιείς ανθρώπους ενώ η κλάση 1 τους ασθενείς.

Αριθμός Γονιδίων (Χαρακτηριστικά)	**Αριθμός Προτύπων (Στιγμιότυπα)**	**Πρόβλημα Ταξινόμησης**	**Κλάση**	**Αριθμός Προτύπων ανά Κλάση**
20501	1065	**Cancer**	Υγιείς = 0	61
			Ασθενείς = 1	1004

Στην παραπάνω εικόνα βλέπουμε το σύνολο δεδομένων που αφορά στα τέσσερα είδη γαστρεντερικού καρκίνου που δουλέψαμε, τον αριθμό των χαρακτηριστικών, τον συνολικό αριθμό προτύπων αλλά και ανά κλάση.

4.2 Μέθοδοι Αξιολόγησης Χαρακτηριστικών για τη Μείωση Διαστάσεων του Συνόλου Δεδομένων

4.2.1 Mutual Information (MI)

Για να εφαρμόσουμε τη μέθοδο της Αμοιβαίας Πληροφορίας για την αξιολόγηση των χαρακτηριστικών, χρησιμοποιήθηκε η συνάρτηση sklearn.feature_selection.mutual_info_classif η οποία ανήκει στη βιβλιοθήκη Scikit-learn. Η συνάρτηση βαθμολογεί το κάθε χαρακτηριστικό με τιμές από 0 (κανένα κέρδος στην πληροφορία) μέχρι 1 (μέγιστο κέρδος στην πληροφορία) με τους τρόπους που αναλύθηκαν στο προηγούμενο κεφάλαιο.

Οι παράμετροι που χρησιμοποιήθηκαν για τον υπολογισμό του κέρδους πληροφορίας από τη μέθοδο της Αμοιβαίας Πληροφορίας είναι οι εξής:

- Discrete_features=False: θεώρηση των δεδομένων ως διακριτά ή συνεχή.

- n_neighbors=3: Ο αριθμός των κοντινότερων γειτόνων προς χρήση για υπολογισμό της Αμοιβαίας Πληροφορίας για συνεχείς μεταβλητές.

- random_state=None: Η δημιουργία θορύβου στις συνεχείς μεταβλητές για την αποφυγή επαναλαμβανόμενων τιμών.

Τέλος με τη χρήση της συνάρτησης sklearn.feature_selection.SelectKBest αποθηκεύσαμε σε ένα καινούριο πίνακα τα χ καλύτερα χαρακτηριστικά σε φθίνουσα σειρά. Όπως προείπαμε, όσο πιο κοντά στο 1 είναι η τιμή του κριτηρίου Αμοιβαίας Πληροφορίας, τόσο μεγαλύτερο είναι το κέρδος της πληροφορίας.

4.2.2 Kolmogorov Smirnov 2 Samples Test (KS 2Samples Test)

Αναφορικά με το κριτήριο Kolmogorov-Smirnov, χρησιμοποιήθηκε η συνάρτηση scipy.stats.ks_2samp της βιβλιοθήκης Scipy. Για τις ανάγκες αυτής της μεθόδου, χωρίσαμε το δείγμα μας σε δύο υποσύνολα, όπου το πρώτο αφορούσε ένα πίνακα με τις τιμές ενός χαρακτηριστικού που αντιστοιχούν στην κλάση 0 (=υγιείς) ενώ το δεύτερο που αντιστοιχούν στην κλάση 1 (=ασθενείς), αντίστοιχα. Αυτό έγινε για κάθε χαρακτηριστικό ξεχωριστά, σε ένα βρόγχο.

Η συνάρτηση για κάθε επανάληψη μας έδινε δύο τιμές, το ks_statistic_grade και το p-value. Η πρώτη αφορά στην μεγαλύτερη απόλυτη απόκλιση μεταξύ δύο κατανομών, ενώ η δεύτερη μας δίνει την πιθανότητα οι 2 κατανομές να είναι παρόμοιες. Πρακτικά, για μικρό ks_statistic_grade ή μεγάλο p-value δεν μπορούμε να απορρίψουμε την υπόθεση ότι οι κατανομές των δειγμάτων είναι ίδιες.

Τέλος με τη χρήση της συνάρτησης sklearn.feature_selection.SelectKBest αποθηκεύσαμε σε ένα καινούριο πίνακα τα χ καλύτερα χαρακτηριστικά σε φθίνουσα σειρά. Όπως προείπαμε, όσο πιο κοντά στο 0 είναι η τιμή του p-value, τόσο πιθανότερο είναι οι δύο κατανομές για κάθε χαρακτηριστικό να είναι διαφορετικές άρα να έχουμε ένα χαρακτηριστικό με σημαντική πληροφορία.

4.2.3 Recursive Feature Elimination with Cross Validation (RFE-CV)

Για την εφαρμογή της μεθόδου Recursive Feature Elimination with Cross Validation, χρησιμοποιήσαμε τη συνάρτηση sklearn.feature_selection.RFECV της βιβλιοθήκης Scikit-learn. Ο αλγόριθμος που χρησιμοποιήθηκε ως κριτήριο στις παραμέτρους ήταν οι Μηχανές Διανυσμάτων Υποστήριξης με Γραμμικό πυρήνα (SVM with Linear kernel) με τη χρήση της συνάρτησης sklearn.svm.SVC της βιβλιοθήκης Scikit-learn. Οι υπόλοιπες παράμετροι ήταν οι εξής:

- step=1: Η μέθοδος αφαιρεί σε κάθε επανάληψη το χειρότερο (1) χαρακτηριστικό.

- cv=StratifiedKFold (5): Για την αξιολόγηση της επίδοσης για κάθε υποσύνολο χαρακτηριστικών χρησιμοποιείται η συγκεκριμένη διασταυρούμενη επικύρωση με χρήση της συνάρτησης sklearn.model_selection.StratifiedKFold. Είναι μια παραλλαγή της μεθόδου KFold η οποία απλά διατηρεί το ποσοστό των δειγμάτων ανά κλάση για κάθε υποσύνολο που δημιουργεί.

- scoring=accuracy: Το κριτήριο επίδοσης που χρησιμοποιήθηκε για την κατάταξη και εξάλειψη χαρακτηριστικών.

Τέλος, αφού δημιουργήσαμε ένα πίνακα που περιείχε τον βέλτιστο αριθμό χαρακτηριστικών, κάναμε μια γραφική παράσταση της επίδοσης για κάθε στιγμιότυπο εφαρμογής της μεθόδου.

4.3 Μέθοδοι Μετασχηματισμού των Δεδομένων

Για τη βέλτιστη απόδοση των μοντέλων στα πειράματα που κάναμε, διερευνήσαμε τη πιθανή χρήση μεθόδων μετασχηματισμού των δεδομένων.

4.3.1 Κανονικοποίηση

Μία από τίς μεθόδους που εξετάστηκε ήταν η κανονικοποίηση των δεδομένων. Αυτό επετεύχθη με τη χρήση της συνάρτησης sklearn.preprocessing.StandardScaler(). Η μέθοδος αυτή κανονικοποιεί δεδομένα αφαιρώντας τη μέση τιμή και μετασχηματίζοντας την τυπική απόκλιση στην τιμή 1. Το κεντράρισμα και ο μετασχηματισμός γίνεται ανεξάρτητα για κάθε χαρακτηριστικό υπολογίζοντας σχετικά στατιστικά στοιχεία των δειγμάτων. Η μέση τιμή και η τυπική απόκλιση αποθηκεύονται ώστε να χρησιμοποιηθούν μετά κατά το στάδιο του μετασχηματισμού. Η κανονικοποίηση ενός δείγματος είναι ένα συνηθισμένο προαπαιτούμενο πολλών μοντέλων Μηχανικής Μάθησης. Η απόδοσή τους ενδεχομένως να μειώνεται αν τα μεμονωμένα χαρακτηριστικά δεν μοιάζουν λίγο πολύ με τυπικά κανονικά κατανεμημένα δεδομένα.

4.4 Αναζήτηση Πλέγματος με Διασταυρούμενη Επικύρωση(Grid Search with Cross Validation)

Για την εφαρμογή Αναζήτησης Πλέγματος, χρησιμοποιήθηκε η συνάρτηση sklearn.model_selection.GridSearchCV() της βιβλιοθήκης Scikit-learn με σκοπό τον εντοπισμό των βέλτιστων παραμέτρων για κάθε μοντέλο στο οποίο χρησιμοποιήθηκε. Οι παράμετροι που χρησιμοποιήθηκαν ήταν οι εξής:

- param: ο πίνακας των παραμέτρων που θα εξεταστεί με όλους τους πιθανούς συνδυασμούς.
- classifier: το μοντέλο του αλγορίθμου που θα χρησιμοποιηθεί κάθε φορά.
- cv: το κριτήριο διασταυρούμενης επικύρωσης που θα χρησιμοποιηθεί. Στα πλαίσια των πειραμάτων μας, χρησιμοποιήσαμε το StratifiedKFold (5).

Τα κριτήρια αξιολόγησης της επίδοσης του εκάστοτε μοντέλου Μηχανικής Μάθησης ήταν η ακρίβεια πρόβλεψης, όπως επίσης και τα κριτήρια precision, recall και f1-score. Οι αλγόριθμοι που εισήχθησαν στην αναζήτηση πλέγματος ήταν εκείνοι που είχαν την καλύτερη επίδοση αφού κανονικοποιήσαμε τα δεδομένα μας, επιλέξαμε το υποσύνολο χαρακτηριστικών που είχε την καλύτερη απόδοση και τέλος κάναμε μια σύγκριση μεταξύ διαφόρων εκτιμητών. Οι παράμετροι που ελέγχθηκαν αποτέλεσαν μέρος μιας εξαντλητικής αναζήτησης και εξερεύνησης πολλών συνδυασμών παραμέτρων.

4.5 Μέθοδοι Υπερδειγματοληψίας με Υποδειγματοληψία για ακραίες παρατηρήσεις (Methods for Oversampling with Undersampling for outliers)

Μετά σκέψη και συνδιάλεξη με τον επιβλέποντα Καθηγητή κ. Διαμαντάρα, αποφασίστηκε να χρησιμοποιηθεί κάποια μέθοδος με σκοπό να εξισορροπηθεί το δείγμα μας ώστε κατά πρώτον να σιγουρευτούμε για την αντικειμενικότητα και την ικανότητα γενίκευσης των αποτελεσμάτων των μοντέλων μας. Για το σκοπό αυτό χρησιμοποιήθηκε ο αλγόριθμος SMOTEENN (Synthetic Minority Over-sampling Technique and Under-sampling with Edited Nearest Neighbors) [27].

4.5.1 Synthetic Minority Oversampling and Undersampling using Edited Nearest Neighbors Technique (SMOTEENN)

Ο εν λόγω αλγόριθμος πραγματοποιεί τεχνικές υπερδειγματοληψίας για την μειοψηφούσα κλάση δημιουργώντας συνθετικά στιγμιότυπα (όχι ρέπλικες) με σκοπό την εξισορρόπηση του δείγματος ενώ πραγματοποιεί επίσης τεχνικές υποδειγματοληψίας με τη χρήση των Προσαρμοσμένων Κοντινότερων Γειτόνων (ENN) για να μειώσει το θόρυβο των περιθωριακών στιγμιότυπων πριν τη δημιουργία συνθετικών. Εδώ, πρέπει να τονιστεί το γεγονός πως τα καινούρια συνθετικά αυτά στιγμιότυπα χρησιμοποιήθηκαν μόνο κατά τη φάση εκμάθησης των αλγορίθμων. Η συνάρτηση που χρησιμοποιήθηκε είναι η imblearn.combine.SMOTEENN της βιβλιοθήκης imbalanced-learn.

5. Αξιολόγηση

Αυτό το κεφάλαιο περιέχει αρχικά τις μετρικές που χρησιμοποιήθηκαν για την αξιολόγηση της επίδοσης των μοντέλων Μηχανικής Μάθησης που χρησιμοποιήθηκαν. Έπειτα, αναλύονται τεχνικές που χρησιμοποιήθηκαν για επιλογή χαρακτηριστικών μαζί με τα αντίστοιχα αποτελέσματα για κάθε τεχνική (βήμα 1). Μετέπειτα, αναλύεται η σύγκριση των ταξινομητών που έγινε μετά τη χρήση του βέλτιστου υποσυνόλου που προέκυψε από το βήμα 1 και εντοπίζονται οι καλύτεροι εξ αυτών (βήμα 2). Μετέπειτα, καταγράφονται τα αποτελέσματα της αναζήτησης πλέγματος για τους ταξινομητές που προέκυψαν από το βήμα 2 και εντοπίζεται ο βέλτιστος αλγόριθμος(βήμα 3). Τέλος, με τη χρήση μεθόδων για υπέρ- και υπό-δειγματοληψία, αξιολογείται εκ νέου ο βέλτιστος αλγόριθμος που προέκυψε από το βήμα 3.

5.1 Κριτήρια Αξιολόγησης

Τα κριτήρια αξιολόγησης που χρησιμοποιήθηκαν στα πειράματα μας για την αξιολόγηση της επίδοσης των αλγορίθμων Μηχανικής Μάθησης είναι τα εξής:

- Accuracy Score: Αποτελεί τη τιμή ακρίβειας πρόβλεψης των δεδομένων.

- Precision: Είναι ουσιαστικά η ικανότητα του μοντέλου να μην ταξινομεί ως θετικό, ένα αρνητικό στιγμιότυπο.

- Recall: Είναι ουσιαστικά η ικανότητα του μοντέλου να εντοπίσει όλα τα θετικά στιγμιότυπα.

- F1-score: Το συγκεκριμένο κριτήριο είναι ένας σταθμισμένος μέσος του precision, recall.

- Mean Squared Error: Η μέση τυπική απόκλιση εις το τετράγωνο.

5.2 Οργάνωση Πειραμάτων και Αξιολόγηση

Η μεθοδολογία που ακολουθήσαμε κατά τη διεξαγωγή των πειραμάτων μας μπορεί να συνοψιστεί σε 6 βήματα.

- **Εισαγωγή** του συνόλου δεδομένων (τιμές γονιδίων για κάθε στιγμιότυπο) και των αντίστοιχων κλάσεων ανά στιγμιότυπο.

- **Προεπεξεργασία** του συνόλου δεδομένων ανάλογα με την τεχνική Επιλογής Χαρακτηριστικών που θέλαμε να εφαρμόσουμε. (Ουσιαστικά η μόνη προεπεξεργασία που χρειάστηκε αφορούσε τη μέθοδο Kolmogorov-Smirnov 2 Samples Test).

- Εφαρμογή των μεθόδων **Επιλογής Χαρακτηριστικών** και εντοπισμός του βέλτιστου δυνατού υποσυνόλου Χαρακτηριστικών. Ωστόσο, κατά τη διάρκεια εφαρμογής της Επιλογής Χαρακτηριστικών, επιλέχθηκε να προχωρήσουμε σε κανονικοποίηση των δεδομένων, πράγμα που βελτίωσε τις επιδόσεις των μοντέλων Μηχανικής Μάθησης μας.

- **Σύγκριση Ταξινομητών** με βάση το βέλτιστο υποσύνολο χαρακτηριστικών και εντοπισμός των επιλαχόντων αλγορίθμων.

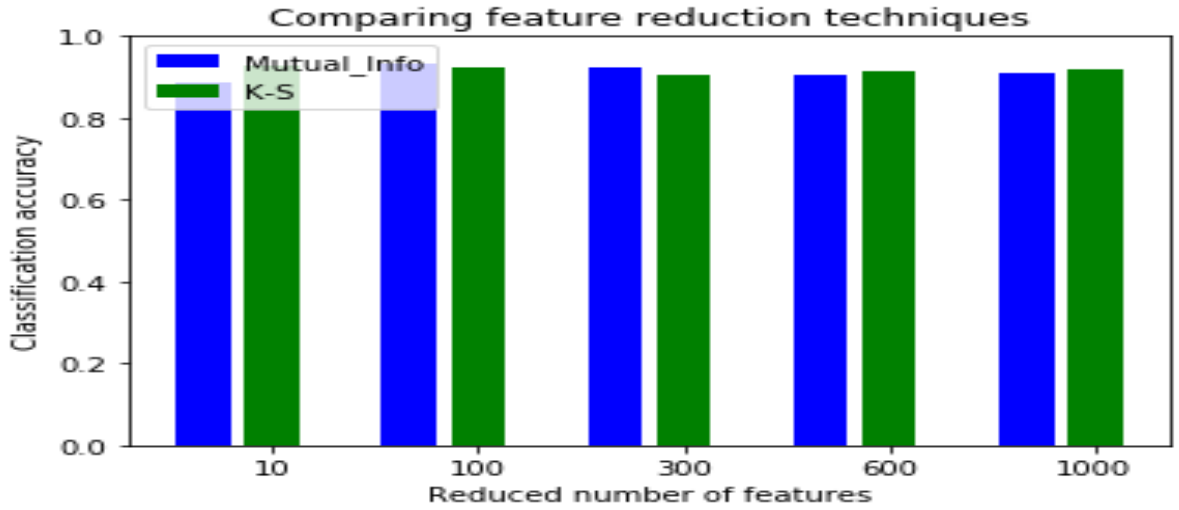
- Εφαρμογή **Εξαντλητικής Αναζήτησης** για τους καλύτερους αλγορίθμους του προηγούμενου βήματος και εντοπισμός του βέλτιστου ταξινομητή.

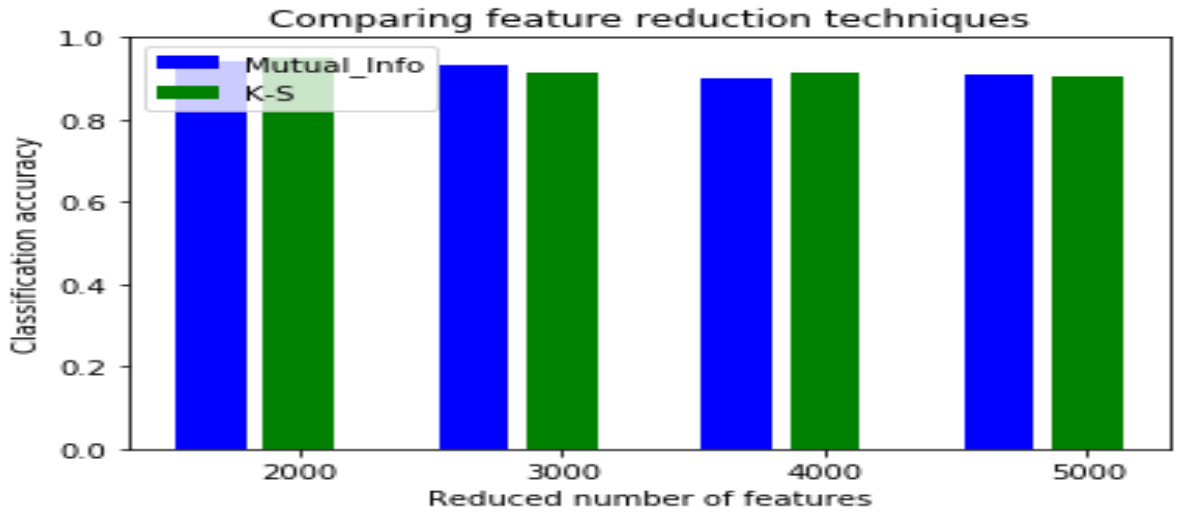
- Εφαρμογή **τεχνικών Εμπλουτισμού δείγματος** για να σιγουρευτούμε για την ικανότητα του αλγορίθμου μας να γενικεύει.

5.3 Αποτελέσματα

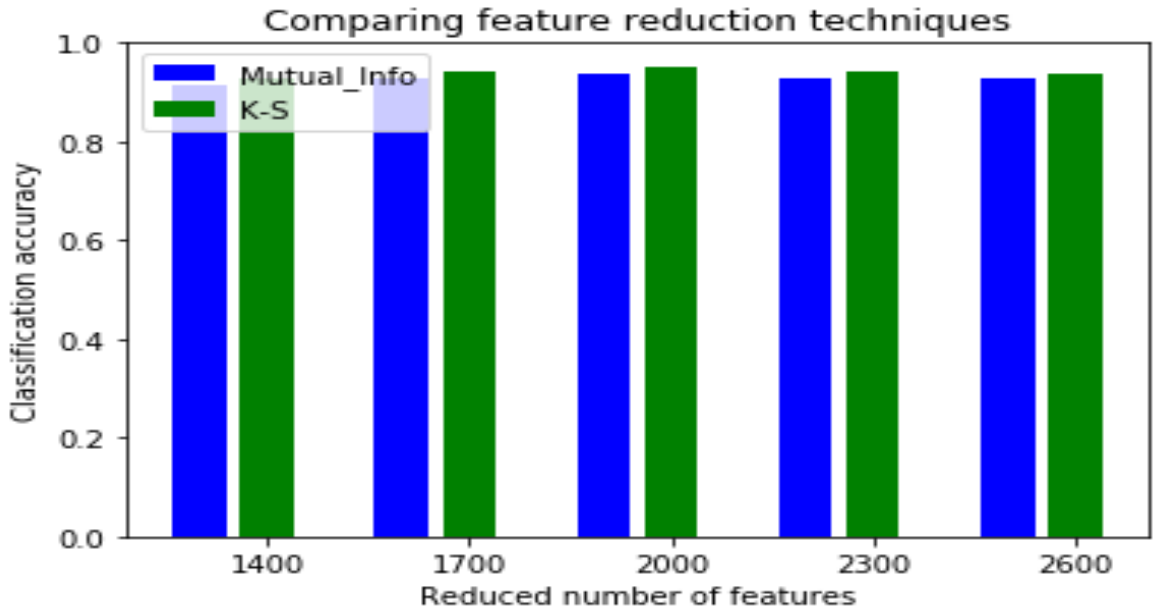
5.3.1 Επιλογή Χαρακτηριστικών

Για την επιλογή χαρακτηριστικών χρησιμοποιήθηκαν αρχικά οι εξής 2 αλγόριθμοι, ο mutual information kai o Kolmogorov- Smirnov. Ο mutual information βρίσκει την ποσότητα της πληροφορίας που εξάγεται από κάθε χαρακτηριστικό σε σχέση με το στόχο. Για τον Kolmogorov-Smirnov, χωρίσαμε το δείγμα σε ασθενείς και υγιείς και αντιδιαστείλαμε τις κατανομές των χαρακτηριστικών για ασθενείς και υγιείς. Αν για ένα χαρακτηριστικό, οι 2 κατανομές ήταν παρόμοιες θεωρούμε ότι το συγκεκριμένο γονίδιο-χαρακτηριστικό δεν αλλάζει τη συμπεριφορά του αν κάποιος είναι υγιής ή ασθενής άρα μας είναι άχρηστο. Αν υπήρχε διαφορά στην κατανομή για κάποιο χαρακτηριστικό, το θεωρούσαμε σημαντικό. Με αυτόν τον τρόπο, εξήχθησαν πίνακες που δίνουν την ακρίβεια του αλγορίθμου μηχανικής μάθησης (στην προκειμένη περίπτωση χρησιμοποιήθηκε ο Linear SVC) για συγκεκριμένο αριθμό χαρακτηριστικών. Παρατηρήθηκε ότι ο αλγόριθμος είχε την καλύτερη του επίδοση με τη χρήση των καλύτερων γονιδίων για μια τιμή κοντά στα 2000 από τα 20501 κοινά.

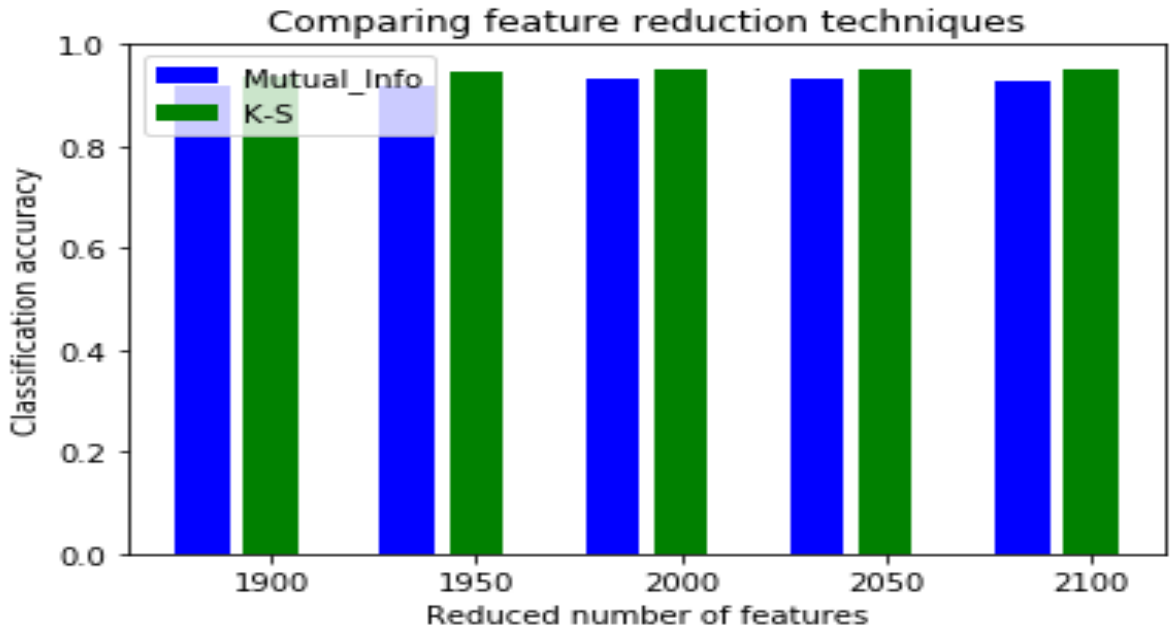




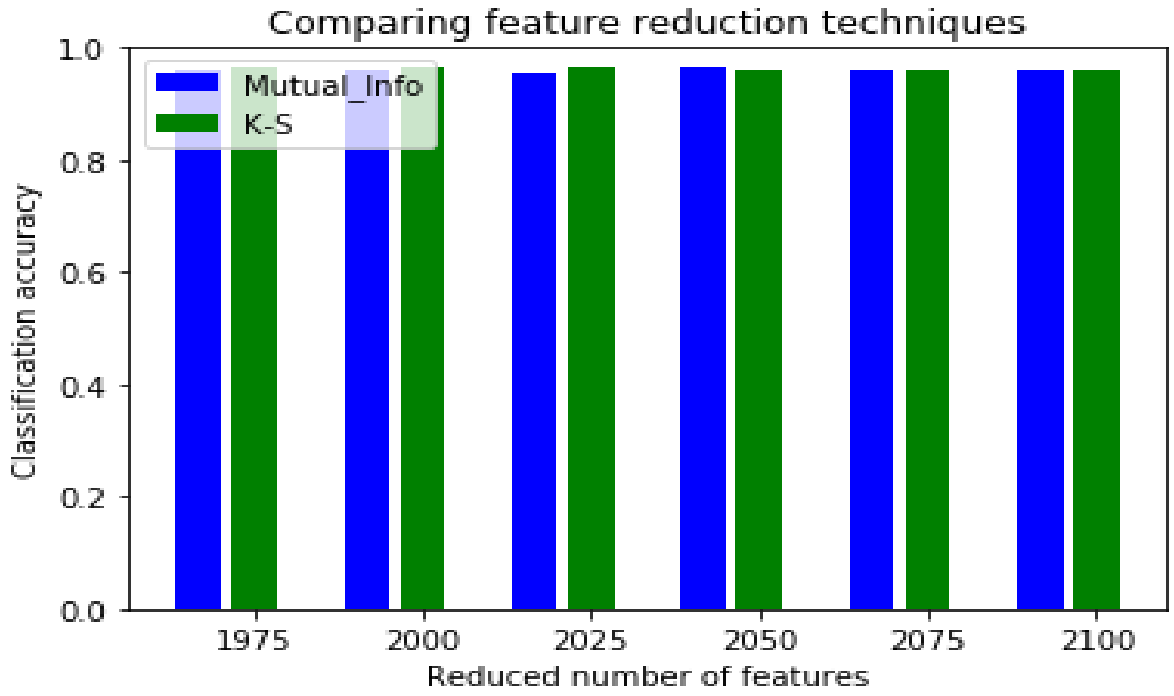
Για το λόγο αυτό, επαναλάβαμε τα πειράματα μας εστιάζοντας στα 1400 μέχρι 2600 γονίδια με βήμα 300.



Έπειτα καθώς περίπου στα 2000 γονίδια είχαμε ξανά τις καλύτερες προβλέψεις επαναλάβαμε για 1900 μέχρι 2100 γονίδια με βήμα 50.



Η τελευταία επανάληψη έγινε για αριθμό γονιδίων από 1975 μέχρι 2100 με βήμα 25.



Δυστυχώς, δεν φαίνεται καλά στα προηγούμενα σχήματα, αλλά πετύχαμε την καλύτερη μας επίδοση με χρήση 2025 ακριβώς γονιδίων.

```
v_score: 0.97183

---- Confusion Matrix ----
[[ 18   6]
 [  6 396]]

             precision    recall  f1-score   support

          0       0.75      0.75      0.75        24
          1       0.99      0.99      0.99       402

avg / total       0.97      0.97      0.97       426


Mean Squared Error 0.0281690140845
```

Στην παραπάνω εικόνα φαίνεται η απόδοση του ταξινομητή **SVM Polynomial** στο δείγμα μας με **2025 γονίδια**. Το 60% του δείγματος χρησιμοποιήθηκε για εκπαίδευση και το 40% για έλεγχο.

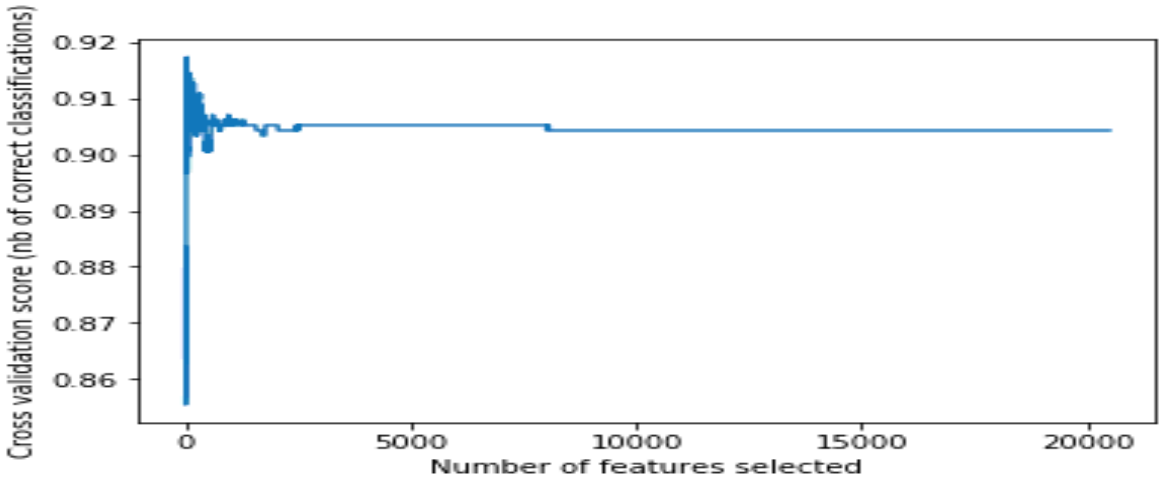
Όπως προαναφέρθηκε, κατά τη διάρκεια εφαρμογής μεθόδων Επιλογής Χαρακτηριστικών έγιναν σκέψεις μετασχηματισμού του δείγματος για να ερευνηθεί η πιθανότητα βελτίωσης των αποτελεσμάτων. Η μέθοδος που ανακαλύφθηκε ότι όντως βελτιώνει τα αποτελέσματα μας ήταν η κανονικοποίηση (Standardization).

```
v_score: 0.98592

---- Confusion Matrix ----
[[ 19   5]
 [  1 401]]

             precision    recall  f1-score   support

          0       0.95      0.79      0.86        24
          1       0.99      1.00      0.99       402

avg / total       0.99      0.99      0.99       426


Mean Squared Error 0.0140845070423
```

Στην παραπάνω εικόνα φαίνεται η απόδοση του ταξινομητή **SVM Polynomial** στο δείγμα μας με **2025 γονίδια** μετά την **κανονικοποίηση** που κάναμε. Το 60% του δείγματος χρησιμοποιήθηκε για εκπαίδευση και το 40% για έλεγχο.

Καθότι δεν είχαμε επαρκώς ικανοποιητικά αποτελέσματα με τις μεθόδους που εξετάστηκαν προηγουμένως, χρησιμοποιήθηκε επίσης ο αλγόριθμος rfecv που αφορά στην κατάταξη χαρακτηριστικών με επαναλαμβανόμενη εξάλειψη χαρακτηριστικών και επιλογή του καλύτερου αριθμού χαρακτηριστικών μέσω Cross Validation.

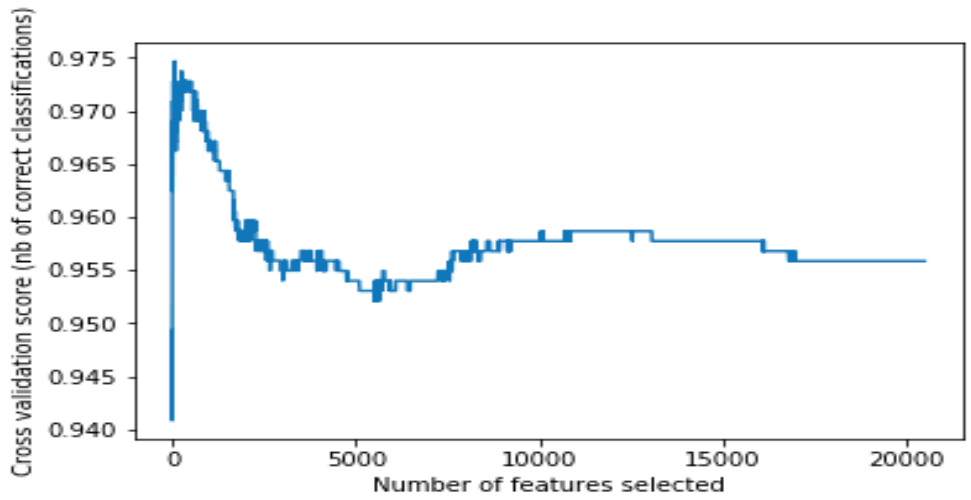
```
Optimal number of features : 37
```



Στην παραπάνω εικόνα βλέπουμε τη χρήση του rfecv στα αρχικά δεδομένα μας.

Έπειτα, κάναμε το ίδιο αφού πρώτα κανονικοποιήσαμε τα δεδομένα μας.

```
Optimal number of features : 76
```



Στην παραπάνω εικόνα βλέπουμε τη χρήση του rfecv στα **κανονικοποιημένα** δεδομένα μας.

```
v_score: 0.99061

---- Confusion Matrix ----
[[ 20   4]
 [  0 402]]

              precision    recall  f1-score   support

           0       1.00      0.83      0.91        24
           1       0.99      1.00      1.00       402

   micro avg       0.99      0.99      0.99       426
   macro avg       1.00      0.92      0.95       426
weighted avg       0.99      0.99      0.99       426


Mean Squared Error 0.009389671361502348
```

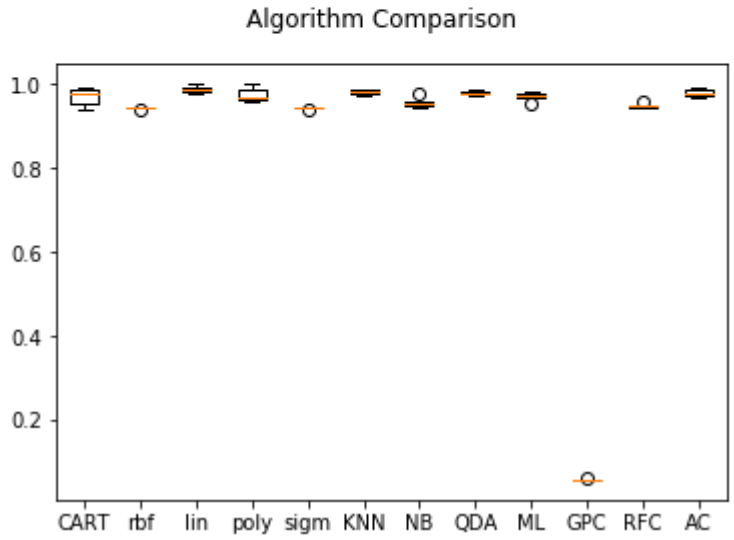
Στην παραπάνω εικόνα φαίνεται η απόδοση του ταξινομητή **SVM Polynomial** στο δείγμα μας μετά τη χρήση του **rfecv** στα **κανονικοποιημένα** δεδομένα μας (76).

Όπως είναι αντιληπτό, η καλύτερη μέθοδος επιλογής χαρακτηριστικών που μας παρείχε με το βέλτιστο υποσύνολο χαρακτηριστικών ήταν η **Επαναλαμβανόμενη Εξάλειψη Χαρακτηριστικών με Διασταυρούμενη Επικύρωση (RFE-CV).**

5.3.2 Σύγκριση Ταξινομητών

Όπως προαναφέραμε, μετά τον **επιτυχή εντοπισμό του βέλτιστου υποσυνόλου χαρακτηριστικών με τη χρήση της μεθόδου rfecv**, προχωρήσαμε σε μια σύγκριση πολλών ταξινομητών.

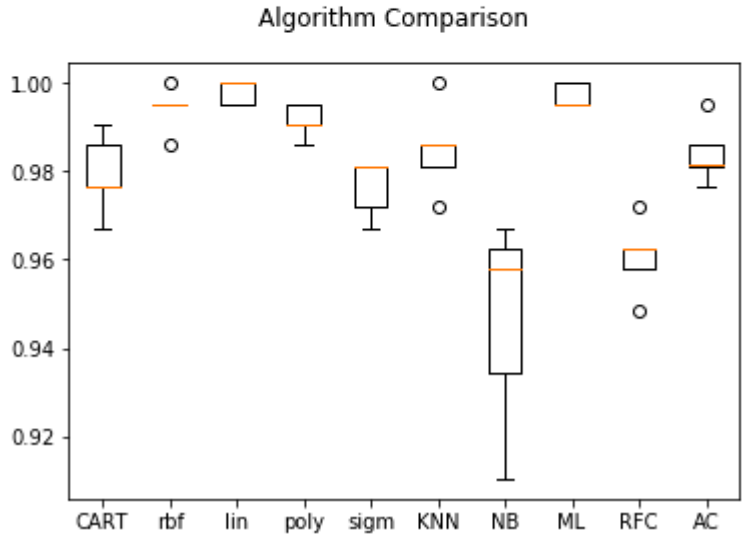
```
CART: 0.969049 (0.019735)
rbf: 0.942727 (0.001740)
lin: 0.986868 (0.008045)
poly: 0.974674 (0.015828)
sigm: 0.942727 (0.001740)
KNN: 0.980290 (0.005447)
NB: 0.955855 (0.011410)
QDA: 0.978408 (0.004778)
ML: 0.969935 (0.009750)
GPC: 0.057273 (0.001740)
RFC: 0.948352 (0.005194)
AC: 0.978395 (0.008743)
```



Στην παραπάνω εικόνα φαίνεται η σύγκριση ταξινομητών που διενεργήσαμε στα αρχικά μας δεδομένα μετά τη χρήση του rfecv. Οι ταξινομητές που συγκρίθηκαν ήταν οι εξής:

- Decision Tree Classifier (CART)
- SVM with kernels, (rbf), (lin), (poly) και (sigm)
- K Neighbors Classifier (KNN)
- Gaussian NB (NB)
- Quadratic Discriminant Analysis (QDA)
- MLP Classifier (ML)
- Gaussian Process Classifier (GPC)
- Random Forest Classifier (RFC)
- Adaboost Classifier (AC)

```
CART: 0.979334 (0.008204)
rbf: 0.994366 (0.004600)
lin: 0.998118 (0.002305)
poly: 0.991549 (0.003513)
sigm: 0.976534 (0.005904)
KNN: 0.984990 (0.009064)
NB: 0.946456 (0.021229)
ML: 0.997179 (0.002304)
RFC: 0.960567 (0.007606)
AC: 0.984033 (0.006387)
```



Στην παραπάνω εικόνα φαίνεται η δεύτερη σύγκριση ταξινομητών την οποία πραγματοποιήσαμε στα **κανονικοποιημένα** δεδομένα μας με τη χρήση του rfecv.

Παρατηρούμε ότι τα δεδομένα συμπεριφέρονται καλύτερα αφού κανονικοποιηθούν. Για το λόγο αυτό, στα επόμενα βήμα χρησιμοποιούμε μόνο τα **κανονικοποιημένα δεδομένα που προέκυψαν από τη χρήση rfecv** (76 βέλτιστα γονίδια).

5.3.3 Εξαντλητική Αναζήτηση Επιλεγμένων Αλγορίθμων

Αφού κάναμε μια ενδελεχή σύγκριση μεταξύ διαφόρων αλγορίθμων, επιλέξαμε τους καλύτερους από αυτούς ώστε να προχωρήσουμε σε μια εξαντλητική αναζήτηση για να βρούμε τις βέλτιστες τιμές παραμέτρων ανά ταξινομητή και τελικά να εντοπίσουμε τον καλύτερο όλων. Να θυμίσουμε σε αυτό το σημείο ότι το βέλτιστο υποσύνολο αποτελείται από 76 γονίδια και δημιουργήθηκε με χρήση του rfecv σε κανονικοποιημένα δεδομένα. Η επιλογή των καλύτερων ταξινομητών έγινε με βάση αυτό το υποσύνολο.

```
GRID SEARCH:
Best accuracy_score: 0.993
Best parameters set:
        C: 0.005
        kernel: 'linear'
        tol: 0.1

CROSS VALIDATION:
Best accuracy_score: 0.99 (+/- 0.00)
 ------ Confusion Matrix ------
[[  55    6]
 [   1 1003]]

             precision    recall  f1-score   support

          0       0.98      0.90      0.94        61
          1       0.99      1.00      1.00      1004

avg / total       0.99      0.99      0.99      1065
```

Στην παραπάνω εικόνα βλέπουμε τη χρήση Gridsearch για τον αλγόριθμο **SVM Linear**.

```
GRID SEARCH:
Best accuracy_score: 0.997
Best parameters set:
        activation: 'identity'
        alpha: 0.01
        learning_rate: 'invscaling'
        max_iter: 5000
        momentum: 0.5
        solver: 'lbfgs'
        tol: 0.0001

CROSS VALIDATION:
Best accuracy_score: 0.99 (+/- 0.01)
 ------ Confusion Matrix ------
[[ 57   4]
 [  5 999]]

             precision    recall  f1-score   support

          0       0.92      0.93      0.93        61
          1       1.00      1.00      1.00      1004

avg / total       0.99      0.99      0.99      1065
```

Στην παραπάνω εικόνα βλέπουμε τη χρήση Gridsearch για τον αλγόριθμο **MLP (Multi-layer Perceptron)**.

```
GRID SEARCH:
Best accuracy_score: 0.995
Best parameters set:
        C: 0.0005
        coef0: 5
        degree: 7
        gamma: 5e-05
        kernel: 'poly'
        tol: 1

CROSS VALIDATION:
Best accuracy_score: 1.00 (+/- 0.00)
 ------ Confusion Matrix ------
[[  56    5]
 [   0 1004]]

             precision    recall  f1-score   support

          0       1.00      0.92      0.96        61
          1       1.00      1.00      1.00      1004

avg / total       1.00      1.00      1.00      1065
```

Στην παραπάνω εικόνα βλέπουμε τη χρήση Gridsearch για τον αλγόριθμο **SVM Polynomial**.

```
GRID SEARCH:
Best accuracy_score: 0.997
Best parameters set:
        C: 1.6
        gamma: 0.001
        kernel: 'rbf'
        tol: 0.0001

CROSS VALIDATION:
Best accuracy_score: 1.00 (+/- 0.00)
 ------ Confusion Matrix ------
[[  59    2]
 [   1 1003]]

             precision    recall  f1-score   support

          0       0.98      0.97      0.98        61
          1       1.00      1.00      1.00      1004

avg / total       1.00      1.00      1.00      1065
```

Στην παραπάνω εικόνα βλέπουμε τη χρήση Gridsearch για τον αλγόριθμο **SVM Rbf**.

Όπως φαίνεται στα παραπάνω σχήματα, Ο αλγόριθμος **SVM με πυρήνα RBF** παρέχει τα καλύτερα αποτελέσματα. Τα επόμενα μας πειράματα θα συνεχιστούν για τον επιλεγμένο αλγόριθμο.

5.3.4 Εξακρίβωση Επίδοσης του Βέλτιστου Ταξινομητή (SVM with Rbf Kernel)

Predictions

Αρχικά, χωρίσαμε τα δεδομένα μας σε δεδομένα εκπαίδευσης (70%) και ελέγχου (30%). Μετά από πολλά πειράματα, στα οποία χρησιμοποιήθηκε η τεχνική ανακατέματος (shuffling) στην οποία τα υποσύνολα εκπαίδευσης-ελέγχου αλλάζουν συνεχώς, καταθέτουμε τα χειρότερα / καλύτερα αποτελέσματα που πετύχαμε.

```
training accuracy: 0.99866

testing accuracy: 1.00000

---- Confusion Matrix ----
[[ 18   0]
 [  0 302]]

             precision    recall  f1-score   support

          0       1.00      1.00      1.00        18
          1       1.00      1.00      1.00       302

avg / total       1.00      1.00      1.00       320


Mean Squared Error 0.0
```

Στην παραπάνω εικόνα βλέπουμε τα **καλύτερα** αποτελέσματα του αλγορίθμου **SVM Rbf**.

```
training accuracy: 1.00000

testing accuracy: 0.99062

---- Confusion Matrix ----
[[ 16   2]
 [  1 301]]

             precision    recall  f1-score   support

          0       0.94      0.89      0.91        18
          1       0.99      1.00      1.00       302

avg / total       0.99      0.99      0.99       320


Mean Squared Error 0.009375
```

Στην παραπάνω εικόνα βλέπουμε τα **χειρότερα** αποτελέσματα του αλγορίθμου **SVM Rbf**.

Cross Validation

Έπειτα, διενεργήθηκαν πειράματα cross-validation με 5-folds. Ξανά, επειδή τα δεδομένα γίνονται shuffle οπότε κάθε φορά τα υποσύνολα είναι διαφορετικά, τα πειράματα έτρεξαν πολλές φορές και παρουσιάζονται τα καλύτερα / χειρότερα αποτελέσματα.

```
The selected number of Features is: 76

cvd_predict: 1.00000

 ------ Confusion Matrix ------
[[  61    0]
 [   0 1004]]

             precision    recall  f1-score   support

          0       1.00      1.00      1.00        61
          1       1.00      1.00      1.00      1004

avg / total       1.00      1.00      1.00      1065


Mean Squared Error: 0.000000
```

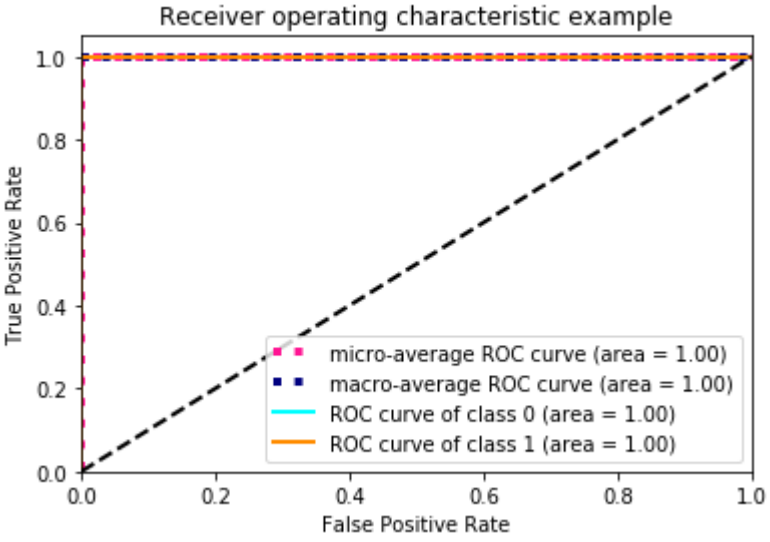
Στην παραπάνω εικόνα βλέπουμε τα **καλύτερα** αποτελέσματα του αλγορίθμου **SVM Rbf** μετά τη διασταυρούμενη επικύρωση.

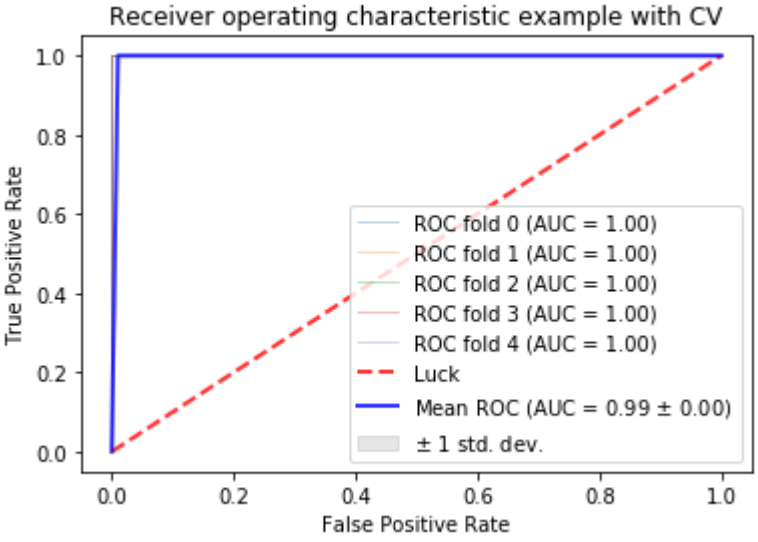
```
The selected number of Features is: 76

cvd_predict: 0.99624

 ------ Confusion Matrix ------
[[  59    2]
 [   2 1002]]

             precision    recall  f1-score   support

          0       0.97      0.97      0.97        61
          1       1.00      1.00      1.00      1004

avg / total       1.00      1.00      1.00      1065


Mean Squared Error: 0.003756
```

Στην παραπάνω εικόνα βλέπουμε τα **χειρότερα** αποτελέσματα του αλγορίθμου **SVM Rbf** μετά τη διασταυρούμενη επικύρωση.

Roc Curves

Οι καμπύλες ROC τυπικά διαθέτουν τον αληθώς θετικό ρυθμό στον άξονα Y και τον ψευδώς θετικό ρυθμό στον άξονα X. Αυτό σημαίνει ότι η επάνω αριστερή γωνία της γραφικής παράστασης είναι το "ιδανικό" σημείο - ένα ψευδώς θετικό ποσοστό μηδέν και ένα αληθώς θετικό ποσοστό ενός. Αυτό δεν είναι πολύ ρεαλιστικό, αλλά σημαίνει ότι μια ευρύτερη περιοχή κάτω από την καμπύλη (AUC) είναι συνήθως καλύτερη.





Στα παραπάνω σχήματα, με τη βοήθεια των γραφικών παραστάσεων καμπυλών ROC (Receiver Operating Characteristic) είναι προφανές πόσο καλά αποδίδει ο αλγόριθμος μας, και μάλιστα χωρίς να αποκλίνει καθόλου για τα διαφορετικά υποσύνολα που δημιουργήθηκαν για την αντικειμενική αξιολόγηση του.

5.3.5 Τεχνικές Εμπλουτισμού Δείγματος

Τέλος, διενεργήθηκαν πειράματα μετά τη χρήση του SMOTEENN (Εξισορρόπηση του δείγματος με χρήση over-sampling methods για τη μειοψηφούσα κλάση και under-sampling methods για την καρατόμηση των περιθωριακών στιγμιότυπων). Να σημειωθεί εδώ ότι αφού χωρίστηκε το δείγμα σε δεδομένα εκπαίδευσης (70%) και δεδομένα ελέγχου (30%), έγινε χρήση του SMOTEENN μόνο στα δεδομένα εκπαίδευσης δημιουργώντας περίπου 40 συνθετικά στιγμιότυπα για τη μειοψηφούσα κλάση.

Original training dataset shape Counter ({1: 702, 0: 43})

Resampled training dataset shape Counter ({1: 702, 0: 78})

Test Dataset shape Counter ({1: 302, 0: 18})

Τα πειράματα έγιναν πολλές φορές με shuffling και παρουσιάζονται τα καλύτερα / χειρότερα αποτελέσματα.

```
Original training dataset shape Counter({1: 702, 0: 43})
Resampled training dataset shape Counter({1: 701, 0: 87})
Test dataset shape Counter({1: 302, 0: 18})

The selected number of Features is: 76

v_score: 1.00000

---- Confusion Matrix ----
[[ 18   0]
 [  0 302]]

             precision    recall  f1-score   support

          0       1.00      1.00      1.00        18
          1       1.00      1.00      1.00       302

avg / total       1.00      1.00      1.00       320


Mean Squared Error 0.0
```

Στην παραπάνω εικόνα βλέπουμε τα **καλύτερα** αποτελέσματα του αλγορίθμου **SVM Rbf** μετά τη χρήση του **SMOTEENN**.

```
Original training dataset shape Counter({1: 702, 0: 43})
Resampled training dataset shape Counter({1: 702, 0: 85})
Test dataset shape Counter({1: 302, 0: 18})

The selected number of Features is: 76

v_score: 0.98750

---- Confusion Matrix ----
[[ 14   4]
 [  0 302]]

             precision    recall  f1-score   support

          0       1.00      0.78      0.88        18
          1       0.99      1.00      0.99       302

avg / total       0.99      0.99      0.99       320


Mean Squared Error 0.0125
```

Στην παραπάνω εικόνα βλέπουμε τα **χειρότερα** αποτελέσματα του αλγορίθμου **SVM Rbf** μετά τη χρήση του **SMOTEENN**.

Κατόπιν, έγινε το ίδιο με τη χρήση διασταυρούμενης επικύρωσης πέντε υποσυνόλων.

```
The selected number of Features is: 76

Classifier Accuracy Score: 0.998

 ------ Confusion Matrix ------
[[  59    2]
 [   0 1004]]

             precision    recall  f1-score   support

          0       1.00      0.97      0.98        61
          1       1.00      1.00      1.00      1004

avg / total       1.00      1.00      1.00      1065


Mean Squared Error: 0.002
```

Στην παραπάνω εικόνα βλέπουμε τα **καλύτερα** αποτελέσματα του αλγορίθμου **SVM Rbf** μετά τη χρήση του **SMOTEENN** και **Cross Validation**.

```
The selected number of Features is: 76

Classifier Accuracy Score: 0.996

 ------ Confusion Matrix ------
[[  58    3]
 [   1 1003]]

             precision    recall  f1-score   support

          0       0.98      0.95      0.97        61
          1       1.00      1.00      1.00      1004

avg / total       1.00      1.00      1.00      1065


Mean Squared Error: 0.004
```

Στην παραπάνω εικόνα βλέπουμε τα **χειρότερα** αποτελέσματα του αλγορίθμου **SVM Rbf** μετά τη χρήση του **SMOTEENN** και **Cross Validation**.

<u>Παρατηρήσεις</u>

Παρατηρούμε ότι τα χειρότερα αποτελέσματα με χρήση SMOTEENN είναι ελαφρώς χειρότερα από ότι χωρίς τη χρήση SMOTEENN. Αυτό σημαίνει ότι ο αλγόριθμος μας γενικεύει καλά ακόμα και μετά τις προσπάθειες μας για εμπλουτισμό του δείγματος ώστε να εξισορροπηθεί.

Τα αποτελέσματα κρίνονται εξαιρετικά και άκρως ικανοποιητικά ασχέτως του ποιες μετρικές χρησιμοποιήθηκαν στην αξιολόγηση των αποτελεσμάτων. Κατά τη γνώμη μας, οι κινήσεις που ήταν καίριας σημασίας ήταν η κανονικοποίηση των δεδομένων αρχικά, και η χρήση του rfecv κατά δεύτερον για την Επιλογή Χαρακτηριστικών.

6 Τεχνικές Λεπτομέρειες

Στο παρόν κεφάλαιο, καταγράφουμε τα προγραμματιστικά εργαλεία, λογισμικά και βιβλιοθήκες που χρησιμοποιήσαμε για τη διεξαγωγή των πειραμάτων μας και την ολοκλήρωση αυτής της διπλωματικής εργασίας

6.1 Προγραμματιστικά εργαλεία και Βιβλιοθήκες

Το μεγαλύτερο μέρος των πειραμάτων έγινε σε ένα μηχάνημα το οποίο μας διατέθηκε από το τμήμα Πληροφορικής του Αλεξάνδρειου Τεχνολογικού Εκπαιδευτικού Ιδρύματος Θεσσαλονίκης. Για αυτό, θέλω να ευχαριστήσω ξανά τον επιβλέποντα καθηγητή μου κ. Διαμαντάρα. Για τις ανάγκες των πειραμάτων, χρησιμοποιήθηκε επίσης ένας προσωπικός Ηλεκτρονικός Υπολογιστής αναβαθμισμένος ώστε να έχει την απαραίτητη υπολογιστική δύναμη για την ολοκλήρωση των εργασιών μας.

Η γλώσσα προγραμματισμού που χρησιμοποιήθηκε ήταν η Python 3. Η πλατφόρμα στην οποία στηριχθήκαμε ήταν η Anaconda. Οι βασικότερες βιβλιοθήκες που χρησιμοποιήθηκαν ήταν οι εξής:

- Spyder: Περιβάλλον ανάπτυξης προγραμμάτων σε κώδικα Python

- Scikit-learn: Βιβλιοθήκη σχετική με τη Μηχανική Μάθηση. Οι συναρτήσεις που μας παρείχε ήταν πολυάριθμες.

- Scipy: Βιβλιοθήκη για στατιστική ανάλυση. Συγκεκριμένα, χρησιμοποιήσαμε τη μέθοδο Kolmogorov-Smirnov.

- Matplotlib: Βιβλιοθήκη για δημιουργία διαγραμμάτων και γραφημάτων για καλύτερη και ευκολότερη κατανόηση των αποτελεσμάτων.

- Imbalanced-learn: Βιβλιοθήκη για την εξισορρόπηση μη ισορροπημένων δειγμάτων με χρήση μεθόδων υπερδειγματοληψίας και υποδειγματοληψίας.

7 Επίλογος

Αυτό το κεφάλαιο περιέχει συγκεντρωμένα όλα τα συμπεράσματα που βγάλαμε κατά τη διενέργηση των πειραμάτων μας και την συγγραφή της διπλωματικής αυτής εργασίας γενικότερα. Τέλος, καταθέτονται κάποιες ιδέες για μελλοντικές επεκτάσεις της έρευνας μας.

7.1 Σύνοψη και Συμπεράσματα

Όπως προαναφέραμε, αυτοσκοπός της παρούσας διπλωματικής ήταν ο εντοπισμός ενός μικρού αλλά και συνάμα ικανού αριθμού γονιδίων (χαρακτηριστικών) από τα συνολικά 20501 με τα οποία δουλέψαμε, ο οποίος θα περιείχε τα περισσότερο σημαντικά γονίδια, αυτά δηλαδή που περιείχαν την περισσότερο σημαντική πληροφορία για την κατασκευή ενός ταξινομητή για το πρόβλημα της ταξινόμησης του δοθέντος δείγματος μας στις κλάσεις υγιείς, ασθενείς που θα είχε μάλιστα την ικανότητα γενίκευσης. Τα πρότυπα που εξετάσαμε αφορούσαν στον γαστρεντερικό καρκίνο και συγκεκριμένα σε τέσσερα από τα συνολικά πέντε είδη του. Οι μεταβλητές-στόχοι αφορούσαν στην ύπαρξη ή μη της ασθένειας. Οι τιμές των γονιδίων που εξετάστηκαν αποτελούσαν τιμές έκφρασης μετασχηματισμένες από αλληλλούχιση RNA (RNA – Seq).

Τα αποτελέσματα της διπλωματικής μας εργασίας μπορούν να συνοψιστούν ως εξής:

- Το κριτήριο επιλογής χαρακτηριστικών της Επαναλαμβανόμενης Εξάλειψης Χαρακτηριστικών με Διασταυρούμενη Επικύρωση (RFE-CV) υπερέχει έναντι των υπολοίπων κριτηρίων Επιλογής Χαρακτηριστικών που εξετάστηκαν με σκοπό τον εντοπισμό ενός βέλτιστου αριθμού γονιδίων που περιέχουν σημαντική πληροφορία για την κατασκευή ενός αποδοτικού ταξινομητή (Αριθμός Επιλεγμένων Γονιδίων=76). Κατά την διερεύνηση των κριτηρίων αυτών, συνειδητοποιήσαμε πως εφαρμόζοντας επίσης κανονικοποίηση στο δείγμα μας καταφέραμε να βελτιώσουμε εκ νέου την απόδοση όλων των κριτηρίων επιλογής χαρακτηριστικών.

- Στη συνέχεια, προχωρήσαμε στη σύγκριση αρκετών μοντέλων Μηχανικής Μάθησης για τον εντοπισμό των καλύτερων από αυτών. Η σύγκριση έγινε στα κανονικοποιημένα δεδομένα μας μετά τη χρήση της μεθόδου RFE-CV για την επιλογή χαρακτηριστικών. Οι επιλαχόντες αλγόριθμοι που προέκυψαν ήταν οι Μηχανές Διανυσμάτων Υποστήριξης (SVM) για όλους τους πιθανούς πυρήνες και ο Perceptron Πολλαπλών Επιπέδων (Multi-layer Perceptron).

- Μετέπειτα, οι παραπάνω αλγόριθμοι υπέστησαν εξαντλητική αναζήτηση για την εύρεση των βέλτιστων παραμέτρων που μεγιστοποιούν τις επιδόσεις του κάθε ταξινομητή. Ο αποδοτικότερος όλων αναδείχθηκε ο SVM with Rbf Kernel. Η απόδοσή του εξετάστηκε ενδελεχώς με απλό διαχωρισμό του δείγματος σε υποσύνολα εκπαίδευσης και ελέγχου αλλά και φυσικά με τη χρήση διασταυρούμενης επικύρωσης.

- Τέλος, πραγματοποιήθηκε εμπλουτισμός του δείγματος με μεθόδους υπερδειγματοληψίας της μειοψηφούσας κλάσης με υποδειγματοληψία για τα περιθωριακά στιγμιότυπα (SMOTEENN) λόγω της μεγάλης ανισορροπίας του δείγματος, ώστε να εξακριβωθεί η ικανότητα του ταξινομητή μας να γενικεύει σε ένα δείγμα διαφορετικό από αυτό που ελέγχθηκε.

7.2 Ιδέες για Μελλοντικές Επεκτάσεις

Θεωρούμε ότι κάναμε μια αρκετά ενδελεχή έρευνα κατά τη διεξαγωγή των πειραμάτων μας, ωστόσο πάντα υπάρχει χώρος για περαιτέρω επέκταση. Τα σημεία τα οποία μπορούν να επεκτείνουν την συνεισφορά της παρούσας διπλωματικής εργασίας συνοψίζονται παρακάτω:

- Πρώτον και κυριότερο, ο εμπλουτισμός του δείγματος με αληθινά δεδομένα, συγκεκριμένα μας ενδιαφέρει περισσότερο να αυξηθούν τα δείγματα που αφορούν τη μειοψηφούσα κλάση (Υγιείς). Με αυτόν τον τρόπο τυχόν μελλοντικά αποτελέσματα και καινούριες αξιολογήσεις μεθόδων θα ήταν πιο αντικειμενικές και περισσότερο συμπαγείς.

- Επιπλέον, περισσότερες μέθοδοι Μείωσης Διαστάσεων μπορούν να χρησιμοποιηθούν ακόμα και να συνδυαστούν. Με αυτόν τον τρόπο ίσως βρεθεί ένα ακόμα πιο βέλτιστο υποσύνολο Γονιδίων (Χαρακτηριστικών).

- Όπως είδαμε, η κανονικοποίηση των δεδομένων είχε ευεργετικές ιδιότητες στη βελτίωση της αποτελεσματικότητας των μεθόδων επιλογής χαρακτηριστικών και κατ' επέκταση στην απόδοση των ταξινομητών. Ως εκ τούτου, θεωρούμε πως περισσότερες τεχνικές προεπεξεργασίας μπορούν να διερευνηθούν.

- Τέλος, αναφορικά με τους ταξινομητές που δοκιμάσαμε, τα διαθέσιμα μοντέλα που υπάρχουν είναι πολλά και φυσικά χρησιμοποιήσαμε τους περισσότερους από αυτούς που παραδοσιακά αποδίδουν καλά σε αντίστοιχα προβλήματα, αλλά καινούρια μοντέλα δημιουργούνται συνέχεια οπότε είναι πιθανόν στο εγγύς μέλλον να υπάρξει κάποιο νέο μοντέλο με ακόμα καλύτερες επιδόσεις.

8. Βιβλιογραφία

[1] Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar, *Foundations of Machine Learning*, MIT Press, 2012.

[2] George D. Magoulas, Andriana Prentza, *Machine Learning in Medical Applications, Proceeding Machine Learning and Its Applications, Advanced Lectures*, Pages 300- 307, Springer-Verlag London,2001.

[3] William H Wolberg, W Nick Street, and Olvi L Mangasarian. 1992. Breast cancerWisconsin (diagnostic) data set.UCI Machine Learning Repository [http://archive.ics. uci. edu/ml/]1992).

[4] Abien Fred Agarap. 2017. A Neural Network Architecture Combining GatedRecurrent Unit (GRU) and Support Vector Machine (SVM) for Intrusion Detection in Network Traffic Data.arXiv preprint arXiv:1709.03082(2017).

[5] [n. d.]. ([n. d.]). https://seer.cancer.gov/statfacts/html/breast.html

[6] Guyon I, Weston J, Barnhill S, Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines. Mach Learn. 2002;46(1-3):389–422.

[7] Nanda, K., McCory, P., Myers, E., et al.: Accuracy of the PAP Test in Screening for and Follow up of Cervical Cytologic Abnormalities a Systematic Review. Annals of Internal Medicine,16 132(10), 810–819 (2000)

[8] Taxdal, S.R., Ward, G.E., Figge, F.H.J.: Fluorescence of Human Lymphatic and Cancer Tissues Following High Doses of Intravenous Hematoporphyrin. Surg. Forum. 5, 619–624 (1955)

[9] Lipson, R.L., Baldes, E.J., Olsen, A.M.: Hematoporphyrin Derivative: A New Aid for Endoscopic Detection of Malignant Disease. J. Thorac. Cardiovasc. Surg. 42, 623–629 (1961)

[10] Galeotti, T., Borrello, S., Minotti, G., Masotti, L.: Membrane Alterations in Cancer Cells: The Role of Oxy Radicals. Ann. NY Acad. Sci. 488, 468–480 (1986)

[11] Campanella, R.: Membrane Lipids Modifications in Human Gliomas of Different Degree of Malignancy. J. Neurosurg. Sci. 36, 11–25 (1992)

[12] Liu, H., Kho, A.T., Kohane, I.S., Sun, Y.: Predicting Survival within the Lung Cancer Histopathological Hierarchy Using a Multi-Scale Genomic Model of Development. PLoS Medicine 3(7), 1090–1102 (2006)

[13] Ruth, S.V., Baas, P., Zoetmulder, F.A.N.: Surgical Treatment of Malignant Pleural Mesothelioma. Chest Journal 123(2), 551–561 (2003)

[14] http://www.mesotheliomahelp.net/default.asp

[15] Brown, P., Botstein, D.: Exploring the New World of the Genome with DNA Microarrays. Nature Genetics Supplement 21, 33–37 (1999)

[16] Dudoit, S., Fridlyand, J., Speed, T.: Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. J. Am. Statistical Assoc. 97, 77–87 (2002)

[17] Peterson, Ringner, M.: Analysis Tumor Gene Expression Profiles. Artificial Intelligence in Medicine 28(1), 59–74 (2002)

[18] Eisen, M., Spellman, P., Brown, P., Botstein, D.: Cluster Analysis and Display of Genome-Wide Expression Patterns. Proc. Nat'l Acad. Sci. USA 95, 14863–14868 (1998)

[19] Tamyo, P., et al.: Interpreting Patterns of Gene Expression with Self-Organizing Maps: Methods and Application to Hematopoietic Differentiation. Proc. Nat'l Acad. Sci. USA 96, 2907–2912 (1999)

[20] Liu, D., Fei, S., Hou, Z., Zhang, H., Sun, C.: Advances in Neural Networks, 4th International Symposium on Neural Networks, ISNN 2007, Nanjing (2007)

[21] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

[22] Brereton, R. G., & Lloyd, G. R. (2010). Support vector machines for classification and regression. *Analyst*, *135*(2), 230-267.

[23] Simon Haykin, Neural Networks, A Comprehensive Foundation, 2nd Edition, Prentice Hall International, 1999

[24] Liu, H., & Motoda, H. (2008). Less is more. *Computational Methods of Feature Selection. Chapman & Hall/CRC*, 3-17

[25] Jakulin, A. (2005). *Machine learning based on attribute interactions* (Doctoral dissertation, Univerza v Ljubljani).

[26] Berger, V. W., & Zhou, Y. (2014). Kolmogorov–smirnov test: Overview. *Wiley StatsRef: Statistics Reference Online*.

[27] *(1, 2)* G. Batista, R. C. Prati, M. C. Monard. "A study of the behavior of several methods for balancing machine learning training data," ACM Sigkdd Explorations Newsletter 6 (1), 20-29, 2004.