ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ Μ/Υ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΣΧΟΛΗ ΝΑΥΤΙΛΙΑΣ ΚΑΙ ΒΙΟΜΗΧΑΝΙΑΣ
ΤΜΗΜΑΤΟΣ ΒΙΟΜΗΧΑΝΙΚΗΣ ΔΙΟΙΚΗΣΗΣ & ΤΕΧΝΟΛΟΓΙΑΣ
ΔΙΑΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΤΕΧΝΟ-ΟΙΚΟΝΟΜΙΚΑ ΣΥΣΤΗΜΑΤΑ»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΚΑΡΚΙΝΟΥ ΤΟΥ ΠΝΕΥΜΟΝΑ ΜΕ ΤΗΝ ΧΡΗΣΗ ΑΛΓΟΡΙΘΜΩΝ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΣΥΓΚΡΙΤΙΚΗ ΑΝΑΣΚΟΠΗΣΗ ΤΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

ΕΥΘΥΜΙΟΣ
ΑΛΕΞΑΝΔΡΙΔΗΣ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ :
ΣΥΜΕΩΝ ΠΑΠΑΒΑΣΙΛΕΙΟΥ, ΚΑΘΗΓΗΤΗΣ ΤΗΣ ΣΧΟΛΗΣ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ ΤΟΥ ΕΘΝΙΚΟΥ ΜΕΤΣΟΒΙΟΥ ΠΟΛΥΤΕΧΝΕΙΟΥ

ΑΠΡΙΛΙΟΣ 2019

ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΚΑΡΚΙΝΟΥ ΤΟΥ ΠΝΕΥΜΟΝΑ ΜΕ ΤΗΝ ΧΡΗΣΗ ΑΛΓΟΡΙΘΜΩΝ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΣΥΓΚΡΙΤΙΚΗ ΑΝΑΣΚΟΠΗΣΗ ΤΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

ΜΕΛΗ ΤΡΙΜΕΛΟΥΣ ΕΠΙΤΡΟΠΗΣ:
ΓΕΩΡΓΙΟΣ ΜΑΤΣΟΠΟΥΛΟΣ, ΚΑΘΗΓΗΤΗΣ ΤΗΣ ΣΧΟΛΗΣ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ ΤΟΥ ΕΘΝΙΚΟΥ ΜΕΤΣΟΒΙΟΥ ΠΟΛΥΤΕΧΝΕΙΟΥ

ΙΩΑΝΝΑ ΡΟΥΣΣΑΚΗ, ΕΠΙΚΟΥΡΗ ΚΑΘΗΓΗΤΡΙΑ ΤΗΣ ΣΧΟΛΗΣ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ ΤΟΥ ΕΘΝΙΚΟΥ ΜΕΤΣΟΒΙΟΥ ΠΟΛΥΤΕΧΝΕΙΟΥ

Ευχαριστίες

Αναμφισβήτητα οφείλω ένα μεγάλο ευχαριστώ στην οικογένεια μου που στάθηκε αρωγός στις σπουδές μου, μέσω καθημερινών θυσιών. Με την πολύτιμη βοήθεια και στήριξη τους έχω καταφέρει και καταφέρνω συνεχώς να υλοποιήσω κάθε επιθυμία - στόχο που θέτω.

Τέλος επιθυμώ να ευχαριστήσω θερμά και τους φίλους μου, που στάθηκαν δίπλα μου όλο αυτό το διάστημα και μου έδωσαν κουράγιο σε κάθε δυσκολία που αντιμετώπισα.

Πίνακας περιεχομένων

Περίληψη

Στις μέρες η τεχνητή νοημοσύνη αποτελεί έναν από τους πιο αναπτυσσόμενους τομείς και εισβάλει ολοένα στην καθημερινότητα μας, για την λήψη αυτοματοποιημένων αποφάσεων στην σύγχρονη τεχνολογία. Στην παρούσα εργασία αναλύονται οι μέθοδοι και οι αλγόριθμοι που μπορεί κανείς να χρησιμοποιήσει προκειμένου να εκπαιδεύσει ένα σύστημα και τελικά να εξάγει κάποια συμπεράσματα για τα αποτελέσματα και την επιτυχία της μεθόδου που ακολούθησε. Αναλύεται το λογισμικό weka μέσω του οποίου πραγματοποιείται η επεξεργασία ανοιχτού περιεχομένου δεδομένων που αφορούν τον καρκίνο του πνεύμονα. Ο ιατρικός τομέας αποτελεί έναν από τους πιο απαιτητικούς τομείς που η ακρίβεια και η μείωση της πιθανότητας λάθους μπορεί να οδηγήσει στην σωτηρία μια ζωής. Οι περισσότερες αποφάσεις που παίρνονται από τους ιατρούς λαμβάνονται με βάση κάποια στατιστικά στοιχεία προηγούμενων παρόμοιων περιπτώσεων. Για παράδειγμα, για να χορηγηθεί ένα φάρμακο ή μια διαδικασία όπως η αντιμετώπιση καρκινικών κυττάρων πρέπει να ληφθεί υπόψιν η επιτυχία του συγκεκριμένου φαρμάκου ή της διαδικασίας σε προηγούμενα ανάλογα περιστατικά – ασθένειες. Ωστόσο με τόσο μεγάλο λοιπόν όγκο πληροφοριών ήταν αδύνατο να παραχθούν ασφαλή αποτελέσματα χωρίς κάποια συγκεκριμένη διαδικασία και αργά ή γρήγορα θα δημιουργούνταν η ανάγκη για εφεύρεση συγκεκριμένων διαδικασιών που θα βοηθούσαν σε αυτό το σκοπό. Η διαδικασία αυτή ονομάζεται Εξόρυξη Δεδομένων (ΕΔ) και έχει εξελιχθεί σε μια από τις μεγαλύτερες επιστήμες τον τελευταίο αιώνα. Διάφοροι κλάδοι επιστημονικοί ή μη εξαρτώνται από την ΕΔ. Γι’ αυτό λοιπόν τον λόγο θα εστιάσουμε και θα εφαρμόσουμε στην πράξη όσα θα δούμε πρώτα σε θεωρητικό υπόβαθρο, προκειμένου να καταφέρουμε να εκπαιδεύσουμε το σύστημα μας με το μικρότερο δυνατό ποσοστό λάθους και την μέγιστη δυνατή ακρίβεια και αξιοπιστία.

Εισαγωγή

Όπως προαναφέραμε ο ρόλος της Εξόρυξης Δεδομένων (ΕΔ) είναι αναμφισβήτητα καθοριστικός στις μέρες μας συμβάλλοντας ριζικά στην εκπαίδευση ενός συστήματος και τελικά στην λήψη μιας απόφασης. Στόχος της παρούσας εργασίας είναι αναλυθούν τόσο θεωρικά όσο και πρακτικά οι μέθοδοι / αλγόριθμοι, να παρουσιαστούν οι προηγμένες τεχνικές εξόρυξης δεδομένων (συλλογή, προεπεξεργασία, επεξεργασία δεδομένων, κατηγοριοποίησης ή ομαδοποίησης) καθώς και οι βασικές ιατρικές-βιολογικές έννοιες που θα χρησιμοποιηθούν για την παραγωγή ενός ευφυούς συστήματος αυτοματοποιημένης κατηγοριοποίησης ασθενών ή υγιών ατόμων με βάση τα συγκεκριμένα χαρακτηριστικά που θα έχουν εισαχθεί στο σύστημα μας, όπως για παράδειγμα συγκεκριμένο χαρακτηριστικό των καρκινικών κυττάρων του ατόμου. Ουσιαστικά τα δεδομένα μας αφορούν τον καρκίνο του πνεύμονα και περιλαμβάνουν κάποια χαρακτηριστικά των κυττάρων όπου με βάση αυτά τα χαρακτηριστικά το άτομο κατατάσσεται σε υγιή ή ασθενή. Εκπαιδεύοντας λοιπόν το σύστημα μας με τον δυνατότερο αμερόληπτο τρόπο γίνεται προσπάθεια να εντοπιστεί ο αλγόριθμος και τα χαρακτηριστικά αυτού που θα επιφέρουν το μικρότερο ποσοστό λάθους και την μεγαλύτερη ακρίβεια και αξιοπιστία, δηλαδή τον αλγόριθμο που για το συγκεκριμένο dataset θα οδηγήσει στην καλύτερη εκπαίδευση του συστήματος. Με αυτό λοιπόν τον τρόπο δίνοντας κάποια χαρακτηριστικά των κυττάρων θα πραγματοποιείται ταξινόμηση με το συγκεκριμένο ποσοστό ακρίβειας το οποίο έχει εξαχθεί από την εκπαίδευση του συστήματος, καθορίζοντας αν το άτομο αποτελεί ασθενή ή υγιές άτομο ή είναι πιθανό να εμφανίσει καρκίνο του πνεύμονα. Τέλος συγκρίνονται τα αποτελέσματα της εκπαίδευσης του συστήματος μας με άλλες σχετικές έρευνες που έχουν πραγματοποιηθεί για το συγκεκριμένο dataset με την χρήση άλλων αλγορίθμων. Οι παραπάνω διαδικασίες για το παραπάνω ιατρικό dataset θα πραγματοποιηθούν με την χρήση του ανοιχτού λογισμικού, weka.

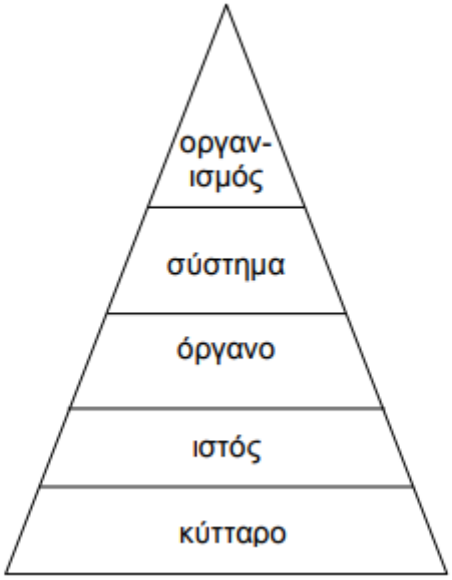
1ο Κεφάλαιο - Ιατρικές / Βιολογικές Έννοιες

1.1 Η έννοια του κυττάρου

Το κύτταρο αποτελεί τη βάση της δομικής και λειτουργικής οργάνωσης ενός οργανισμού. Συγκεκριμένα, όμοια κύτταρα συνδέονται μεταξύ τους για να σχηματίσουν ιστό, διαφορετικοί ιστοί συνδυάζονται για να σχηματίσουν ένα

όργανο, διαφορετικά όργανα συνεργάζονται για να αποτελέσουν ένα σύστημα που επιτελεί μια συγκεκριμένη λειτουργία του οργανισμού, και τέλος όλα τα συστήματα μαζί δημιουργούν έναν οργανισμό.

Ουσιαστικά λοιπόν ως κύτταρο ορίζουμε την δομική και λειτουργική μονάδα της ζωής όπως φαίνεται και στην ακόλουθη εικόνα.

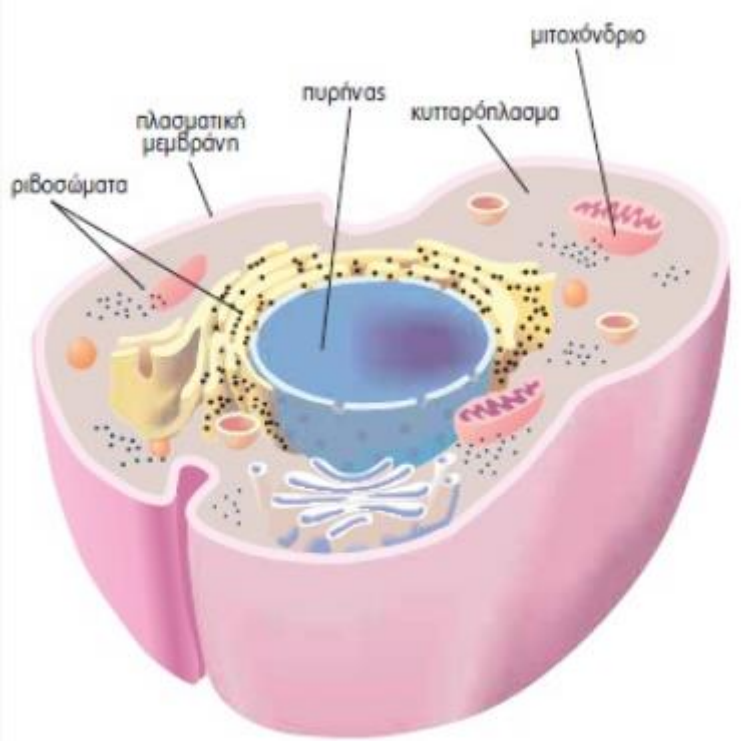


Εικόνα 1.1 : Απεικόνιση δομικής και λειτουργικής μονάδας.

Ο αριθμός των κυττάρων σε έναν οργανισμό είναι ένα χαρακτηριστικό της κατηγοριοποίησής τους σε μονοκύτταρους ή πολυκύτταρους οργανισμούς αντίστοιχα. Ο χώρος εντός του οποίου βιώνουν τα κύτταρα των πολυκύτταρων οργανισμών ονομάζεται μεσοκυττάριο υγρό. Μεγάλες ομάδες ομοειδών κυττάρων, κατά σύσταση και συγκεκριμένη φυσιολογική λειτουργία, ονομάζονται ιστοί (μυϊκός ιστός, επιθηλιακός ιστός κλπ.) οι οποίοι αποτελούν την μονάδα δεύτερης τάξης στον ανθρώπινο οργανισμό μετά τα κύτταρα. Όλα τα κύτταρα αποτελούνται από μία ουσία που ονομάζεται πρωτόπλασμα η οποία μοιάζει με γέλη (ζελε), είναι θολή και άχρωμη και αποτελείται κυρίως από νερό και άλλες διαλυμένες ουσίες. Η λέξη κυτταρόπλασμα χρησιμοποιείται συχνά για να περιγράψει το πρωτόπλασμα που σχηματίζει την κύρια μάζα του κυττάρου. Χωρίζεται από τον πυρήνα, που βρίσκεται συνήθως στο κέντρο του κυττάρου και περιέχει το γενετικό υλικό, με την πυρηνική μεμβράνη, και από το εξωτερικό περιβάλλον την κυτταρική μεμβράνη. Το κυτταρόπλασμα περιέχει μόρια ριβονουκλεινικού οξέος (RNA), τα οποία λειτουργούν ως αγγελιοφόροι που μεταφέρουν πληροφορίες από τον πυρήνα στο κυτταρόπλασμα. Η κυτταρική μεμβράνη παίζει σπουδαίο ρόλο στην συγκράτηση των περιεχομένων του κυττάρου στο εσωτερικό του και είναι αρκετά εύκαμπτη ώστε να επιτρέπει στα κύτταρα να παίρνουν διάφορα σχήματα. Η κυτταρική μεμβράνη ρυθμίζει επίσης την μεταφορά ουσιών, όπως είναι τα θρεπτικά

συστατικά, οι ηλεκτρολύτες και τα άχρηστα προϊόντα του μεταβολισμού , από προς το εσωτερικό του κυττάρου. Για το λόγο αυτό, η κυτταρική μεμβράνη περιγράφεται ως εκλεκτικά διαπερατή ή ημιδιαπερατή. (Χριστοδουλάκης, 1994)

Η κυτταρική μεμβράνη αποτελείται από πρωτεΐνες και μόρια που ονομάζονται φωσφολιπίδια. Η δομή των μορίων των φωσφολιπιδίων, η οποία τα καθιστά υδρόφοβα (απωθούνται από το νερό) ή υδρόφιλα (έλκονται από το νερό) στο άλλο, οδηγεί στην μοναδική δομή της κυτταρικής μεμβράνης που είναι ένα διπλό στρώμα, γνωστό ως φωσφολιπιδική διπλοστιβάδα. (Μαργαρίτης, 1985)



Εικόνα 1.2 : Δομή του κυττάρου, πηγή: (Μαργαρίτης, 1985)

1.2 Καρκινικά κύτταρα

Με τον όρο «καρκίνος» περιγράφεται το σύνολο των ασθενειών ή διαταραχών που χαρακτηρίζονται κυρίως από ανεξέλεγκτο κυτταρικό πολλαπλασιασμό. Τα καρκινικά κύτταρα εξαπλώνονται είτε άμεσα στον παρακείμενο ιστό με «διήθηση», ή μεταφέρονται σε άλλες θέσεις του οργανισμού μέσω των αιμοφόρων αγγείων και

των λεμφαγγείων (μετάσταση). Ο καρκίνος αντιπροσωπεύει μια από τις πιο ενδιαφέρουσες προκλήσεις της μοντέρνας ιατρικής, καθώς κατέχει την δεύτερη θέση σε σειρά θνησιμότητας, παγκοσμίως, μετά τις καρδιαγγειακές παθήσεις (Καρδαμάκης, 2004).

Η επίπτωση του καρκίνου παρουσιάζει αύξηση στην Ευρώπη. Σήμερα ο καρκίνος διαγιγνώσκεται σε περισσότερους από 1,2 εκατομμύρια ανθρώπους το χρόνο εντός των χωρών της Ευρωπαϊκής Ένωσης (Παντελάκος, 2005). Στις ανεπτυγμένες χώρες, οι πιο συχνά εμφανιζόμενες μορφές καρκίνου είναι πνεύμονα, μαστού, δέρματος, εντέρου και προστάτη.

Κάθε πολυκύτταρος οργανισμός όπως προαναφέραμε λειτουργεί ως οικοσύστημα του οποίου τα μέλη είναι τα κύτταρα. Για τη διατήρηση του εν λόγω οικοσυστήματος, τα κύτταρα αναπαράγονται μέσω της κυτταροδιαίρεσης και οργανώνονται σε στενά και αυστηρά συνεργαζόμενες δομές, τους λεγόμενους ιστούς. Όλες οι σωματικές κυτταρικές σειρές ανεξαρτήτου καταγωγής είναι καταδικασμένες σε θάνατο και αφιερώνουν την ύπαρξή τους στην υποστήριξη της διατήρησης των σπερματικών βλαστικών κυττάρων (germ cells) τα οποία είναι τα μοναδικά που διαθέτουν πιθανότητα επιβίωσης. Τα κύτταρα ενός πολυκύτταρου οργανισμού είναι υποχρεωμένα να συνεργάζονται. Για να συντονίσουν τη συμπεριφορά τους στέλνουν λαμβάνουν και ερμηνεύουν ένα πολύπλοκο σύνολο σημάτων. Το σήμα αυτό περιέχει πληροφορία που μεταφράζεται ως εντολής καθοδήγησης και υποδεικνύει την ενέργεια που πρέπει να γίνει. Κατά συνέπεια, κάθε κύτταρο που συμπεριφέρεται με τον κοινωνικά αρμόζοντα τρόπο, συνεχίζοντας δηλαδή τη λειτουργία του με διαίρεση, διαφοροποίηση ή θάνατο, ανάλογα με το τι απαιτείται κάθε φορά προς όφελος του οργανισμού. Οι μοριακές διαταραχές (π.χ. μεταλλάξεις) που ανατρέπουν αυτή την αρμονία σημαίνουν πρόβλημα για την πολυκυτταρική κοινωνία. (Kleinsmith, 2006) Μεγαλύτερο κίνδυνο όμως επιφέρουν οι μοριακές διαταραχές που επιτρέπουν σε ένα κύτταρο να διαιρείται και να πολλαπλασιάζεται ανεξέλεγκτα με αποτέλεσμα το κύτταρο αυτό να μετατρέπεται σε έναν αυξανόμενο κλώνο μεταλλάξεων. Μια τέτοια μετάλλαξη που δίνει αφορμή για τέτοια «εγωιστική» συμπεριφορά από κάποια ξεχωριστά μέλη της «κοινωνίας» μπορεί να διακινδυνεύσει το μέλλον ολόκληρου του οργανισμού. (Κρεμιώτης, 2016)«Οι επαναλαμβανόμενοι κύκλοι μεταλλάξεων, ανταγωνισμού και φυσικής επιλογής που λειτουργούν μέσα στον πληθυσμό των σωματικών κυττάρων μπορούν να ωθήσουν τα πράγματα σε δυσμενείς καταστάσεις. Αυτές οι μεταλλάξεις (καρκινικά κύτταρα) είναι μια ασθένεια στην οποία μεμονωμένοι κλώνοι μεταλλάξεων των κυττάρων αρχίζουν να ευημερούν εις βάρος των γειτόνων τους αλλά στο τέλος καταστρέφουν ολόκληρη τη κυτταρική κοινωνία » (Κρεμιώτης, 2016)

«Ένα απομονωμένο παθολογικό κύτταρο που δεν πολλαπλασιάζεται περισσότερο από τους φυσιολογικούς γείτονές του, δεν προξενεί σημαντική ζημιά με τις όποιες μπορεί να έχει δυσάρεστες ιδιότητες, ενώ αν ο πολλαπλασιασμός του γίνει εκτός ελέγχου θα δώσει αφορμή για το σχηματισμό ενός όγκου ή μιας νεοπλασματικής μάζας με υπερβολική εξάπλωση των παθολογικών κυττάρων. Όταν τα νεοπλασματικά κύτταρα παραμείνουν συγκεντρωμένα ως μια ενιαία μάζα, ο όγκος λέγεται ότι είναι καλοήθης. Σε αυτή τη φάση, μια πλήρης θεραπεία μπορεί να επιτευχθεί με τη χειρουργική αφαίρεση της μάζας. Ένα όγκος θεωρείται καρκινικός μόνο εάν είναι κακοήθης, δηλαδή μόνο όταν τα κύτταρά του έχουν αποκτήσει τη δυνατότητα να εισβάλουν στους περιβάλλοντες ιστούς. Η διηθητικότητα αυτών των κυττάρων υπονοεί τη δυνατότητα ανωμάλων κυττάρων να ξεφύγουν, να εισαχθούν στη κυκλοφορία του αίματος ή των λεμφαγγείων και να διαμορφώσουν τους δευτερογενείς όγκους σε άλλες περιοχές του σώματος που λέγονται μεταστάσεις. Όσο πιο εκτεταμένος είναι ο καρκίνος, τόσο λιγοστεύουν οι πιθανότητες θεραπείας» (Κρεμιώτης, 2016)

Οι διάφοροι τύποι καρκίνου ταξινομούνται με βάση τον τύπο του και των κυττάρων από τα οποία προέρχονται. Τα καρκινικά κύτταρα συνήθως προέρχονται από τα επιθηλιακά κύτταρα και ονομάζονται καρκινώματα, ενώ εκείνοι που προκύπτουν από τα κύτταρα του συνδετικού ιστού ή τα μυϊκά κύτταρα ονομάζονται σαρκώματα. Τα καρκινικά κύτταρα που δεν εντάσσονται σε καμία από τις παραπάνω ευρείες κατηγορίες περιλαμβάνουν τις διάφορες λευχαιμίες, που προέρχονται από τα αιμοποιητικά κύτταρα και τα κύτταρα του νευρικού συστήματος. (Κρεμιώτης, 2016)

1.3 Κύριες αιτίες καρκίνου

Στη διερεύνηση των αιτιών του καρκίνου, δύο κύριες οδοί μπορούν να ακολουθηθούν: Η πρώτη αφορά στο γενετικό επίπεδο και η άλλη στις αιτιολογικές συσχετίσεις. Οι αιτίες καρκίνου που περιγράφονται μέχρι σήμερα είναι αρκετές, και οι μηχανισμοί της δημιουργίας του είναι αρκετά πολύπλοκοι (J. Larry Jameson, 1997) . Οι αιτίες αυτές μπορεί να είναι χημικές (χημικά καρκινογόνα, ιοντίζουσες και υπεριώδεις ακτινοβολίες), ιογενείς (ιοί και ογκογονίδια), και άλλες, όπως η κληρονομικότητα, τα τραύματα, και οι προκαρκινικές κακώσεις. Το 25% του πληθυσμού καταλήγει κάποια στιγμή με μια μορφή καρκίνου, ενώ μόνο το 40% αυτών των ασθενών επιβιώνει για μια πενταετία μετά τη διάγνωση της νόσου. Η βιολογική μονάδα του καρκίνου είναι το καρκινικό κύτταρο, το οποίο είναι παρόμοιο με ένα φυσιολογικό κύτταρο από πολλές απόψεις. Οι θεμελιώδεις ιδιότητες όμως του καρκινικού κυττάρου, είναι αυτές που το διακρίνουν από τα φυσιολογικά. Κυρίαρχο χαρακτηριστικό αποτελεί ο μη ελεγχόμενος πολλαπλασιασμός και η ικανότητα να κάνει μεταστάσεις (J. Larry Jameson, 1997).

Επιπρόσθετα ένα ακόμη χαρακτηριστικό είναι το γεγονός ότι μεταξύ των καρκινικών κυττάρων παρατηρούνται διαφορές στη λειτουργία ή/και τη δομή, όταν προέρχονται από καρκίνους διαφορετικής προέλευσης, πράγμα που δεν παρατηρείται ανάμεσα στα φυσιολογικά κύτταρα. Παρόλο, όμως, που ο καρκίνος είναι μια διαταραχή των κυττάρων, συνήθως κατά τη διάγνωση, εμφανίζεται σαν ένας ορατός όγκος που είναι το τελικό αποτέλεσμα μιας ολόκληρης σειράς γενετικών αλλαγών στα κύτταρα που προάγουν την εμφάνιση νέων χαρακτηριστικών στο ανθρώπινο σώμα, που μπορεί να έχει πάρει χρόνια για να αναπτυχθεί. Ο καρκίνος, λοιπόν, είναι μια νόσος που προκαλείται όταν τα φυσιολογικά κύτταρα διαταραχθούν από κάποιον αιτιολογικό παράγοντα και τότε οδηγούνται σε μη ελεγχόμενη λειτουργία και ανάπτυξη. Αυτή η μη ελεγχόμενη ανάπτυξη προκαλεί το σχηματισμό του όγκου. Αν δεν αντιμετωπισθεί, ο όγκος μπορεί να προκαλέσει προβλήματα με την εισβολή του σε φυσιολογικούς γειτονικούς ιστούς ή ασκώντας πίεση σε άλλες γειτονικές δομές του σώματος, όπως επίσης και να αποβεί θανατηφόρος για τη ζωή του ασθενούς (Eugene Braunwald, 1997)

1.4 Καρκίνος του πνεύμονα

Ο καρκίνος του πνεύμονα αποτελεί ένα σημαντικότατο πρόβλημα υγείας εξαιτίας της επίπτωσης και της θνητότητάς του. Είναι μια ασθένεια που χαρακτηρίζεται από ανεξέλεγκτη ανάπτυξη των κυττάρων στους ιστούς του πνεύμονα. Εάν δεν θεραπευτεί , όπως προαναφέραμε, ο ανεξέλεγκτος πολλαπλασιασμός μπορεί να εξαπλωθεί και πέραν του πνεύμονα με μια διαδικασία που ονομάζεται μετάσταση σε κοντινό ιστό και, τελικά, σε άλλα μέρη του σώματος . Οι περισσότεροι καρκίνοι που ξεκινούν στον πνεύμονα, γνωστοί ως πρωτογενείς καρκίνοι του πνεύμονα, είναι καρκινώματα που προέρχονται από επιθηλιακά κύτταρα.

Οι κυριότεροι τύποι καρκίνου του πνεύμονα είναι το καρκίνωμα των μικρών κυττάρων του πνεύμονα (SCLC), που ονομάζεται επίσης καρκίνος κυττάρων βρώμης, και το καρκίνωμα των μη μικρών κυττάρων του πνεύμονα (NSCLC). Η πιο κοινή αιτία του καρκίνου του πνεύμονα είναι η μακροχρόνια έκθεση στον καπνό η οποία προκαλεί το 80-90% των καρκίνων του πνεύμονα. Οι μη καπνιστές αντιστοιχούν στο 10-15% των περιπτώσεων καρκίνου του πνεύμονα, που συχνά αποδίδεται σε ένα συνδυασμό γενετικών παραγόντων στο αέριο ραδόνιο, στον αμίαντο, την ατμοσφαιρική ρύπανση και το παθητικό κάπνισμα. Τα πιο συχνά συμπτώματα είναι ο βήχας (συμπεριλαμβανομένης της αιμόπτυσης), η απώλεια βάρους και η δυσκολία στην αναπνοή.

Ο καρκίνος του πνεύμονα μπορεί να παρατηρηθεί σε ακτινογραφία θώρακος και αξονική τομογραφία (CT scan). Η διάγνωση επιβεβαιώνεται με βιοψία . Αυτό γίνεται συνήθως με βρογχοσκόπηση ή CT καθοδηγούμενη βιοψία. Η θεραπεία και η πρόγνωση εξαρτώνται από τον ιστολογικό τύπο του καρκίνου, το στάδιο (βαθμό εξάπλωσης), και γενικά την ευημερία του ασθενούς, που μετράται με την κατάσταση απόδοσης. Οι κοινές θεραπείες περιλαμβάνουν χειρουργική επέμβαση, χημειοθεραπεία και ακτινοθεραπεία. Το καρκίνωμα NSCLC μερικές φορές αντιμετωπίζεται με χειρουργική επέμβαση, ενώ το SCLC συνήθως ανταποκρίνεται καλύτερα στη χημειοθεραπεία και την ακτινοθεραπεία.

Η σταδιοποίηση επιτρέπει στο γιατρό να κατανοήσει πλήρως την έκταση της νόσου, προκειμένου να αποφασιστεί το καλύτερο θεραπευτικό πλάνο για τον ασθενή. Οι γιατροί χρησιμοποιούν πολύ εξιδεικευμένους όρους, προκειμένου να περιγράψουν τα στάδια του καρκίνου, αλλά μια απλοποίηση των πραγμάτων μας δίνει 3 τύπους:

- *Τοπικός:* ο καρκίνος περιορίζεται στον πνεύμονα.
- *Περιοχικός:* ο καρκίνος έχει προχωρήσει στους λεμφαδένες του θώρακος.
- *Απομακρυσμένος:* υπάρχουν μεταστάσεις σε άλλα σημεία του σώματος.

Περίπου 90% των καρκίνων του πνεύμονος ξεκινούν από τους βρόγχους, παρόλα αυτά καρκίνος μπορεί να αναπτυχθεί και σε αδένες κάτω από την γραμμή των βρόγχων, συχνότερα στο εξωτερικό τμήμα του πνεύμονος. Οι καρκίνοι του πνεύμονας (όγκοι πνεύμονος) χωρίζονται σε δύο τύπους. Στο μικροκυτταρικό και στο μη μικροκυτταρικό καρκίνο πνεύμονος, οι οποίοι αναπτύσσονται διαφορετικά και έχουν διαφορετική πρόγνωση και θεραπεία.

- *Μη μικροκυτταρικός καρκίνος πνεύμονος*: είναι πολύ πιο συχνός και συνήθως αναπτύσσεται πολύ πιο αργά από τον άλλο τύπο. Χωρίζεται σε 3 υποκατηγορίες:
  1. Αδενοκαρκίνωμα
  2. Πλακώδες καρκίνωμα
  3. Μεγαλοκυτταρικό καρκίνωμα

- *Μικροκυτταρικός καρκίνος πνεύμονος*: είναι λιγότερο συχνός, και αντιπροσωπεύει το 15% όλων των καρκίνων του πνεύμονος. Παρόλα αυτά η ανάπτυξη του είναι ταχύτατη, και πολύ συχνά προχωρημένη τη στιγμή της διάγνωσης, αφού εξαπλώνεται γρήγορα σε όλο το σώμα.

- *Σπάνιοι καρκίνοι θώρακος*: Υπάρχουν περισσότεροι από 12 τύποι καρκίνου που μπορούν να αναπτυχθούν στο θώρακα (όγκοι πνεύμονος), οι οποίοι είτε ξεκινούν από τον πνεύμονα, είτε όχι. Κάποιοι από τους λιγότερο συχνούς

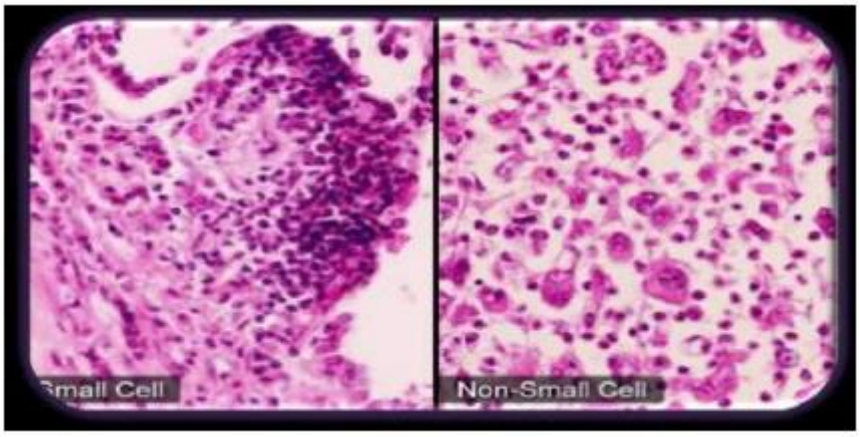
τύπους καρκίνου περιλαμβάνουν το καρκινοειδές, το κακοήθες μεσοθηλίωμα και τους νευροενδοκρινείς όγκους. (Devereux & Taylor JA, 1996)

- *Μεσοθηλίωμα:* το μεσοθηλίωμα είναι ένας καρκίνος που προσβάλει το μεσοθήλιο, την προστατευτική μεμβράνη, που καλύπτει τα περισσότερα από τα εσωτερικά μας όργανα. Στον πνεύμονα αναπτύσσεται στη μεμβράνη που ονομάζεται υπεζωκότας και σπάνια στο περικάρδιο (τη μεμβράνη που καλύπτει την καρδιά). Είναι άμεσα συνδεδεμένο με πολυετή έκθεση σε αμίαντο.

Εικόνα 1.3 : Απεικόνιση όγκων στον πνεύμονα. (πηγή: wikipedia)

Όλες οι πληροφορίες, ο απεικονιστικός έλεγχος, οι βιοψίες, ο τύπος του καρκίνου, καθώς και η γενικότερη κατάσταση του ασθενούς συνεκτιμώνται, προκειμένου να επιλεγεί η κατάλληλη θεραπεία. Ο πιο κριτικός παράγοντας που καθορίζει την επιβίωση είναι το στάδιο τη στιγμή της διάγνωσης. Δυστυχώς μόνο το 25% των ασθενών βρίσκονται σε χειρουργήσιμο στάδιο.

Παρόλα αυτά υπάρχει λύση και για τους ασθενείς σε προχωρημένο στάδιο, όπως προηγμένες χημειοθεραπείες, εκλεκτική ακτινοβολία, και τα τελευταία χρόνια με μεγάλη επιτυχία, η ανοσοθεραπεία. Είναι απαραίτητο σε αυτά τα στάδια να υπάρχει ολοκληρωμένη ομάδα ιατρών, περιλαμβανομένων ογκολόγων, ακτινολόγων, ακτινοθεραπευτών και πνευμονολόγων.



Εικόνα 1.4 : Τύποι καρκίνων του πνεύμονα, μικροκυτταρικός καρκίνος και μη μικροκυτταρικός (πηγή: Iatropedia 2015)

Παγκοσμίως, ο καρκίνος του πνεύμονα είναι η πιο κοινή αιτία καρκίνου που σχετίζεται με το θάνατο σε άνδρες και γυναίκες και είναι υπεύθυνη για 1.380.000 θανάτους ετησίως από το 2008. (Ferlay, 2010)

2ο Κεφάλαιο - Τεχνητή νοημοσύνη & data Mining

Στις μέρες μας η τεχνητή νοημοσύνη αποτελεί έναν από τους πιο αναπτυσσόμενους τομείς, σχετίζεται με τη σχεδίαση και την υλοποίηση υπολογιστικών συστημάτων που μιμούνται στοιχεία της ανθρώπινης συμπεριφοράς όπως την μάθηση, την προσαρμοστικότητα, την εξαγωγή συμπερασμάτων, καθώς και την επίλυση προβλημάτων. (Ι. Βλαχάβας, 2011)

Συστήματα που σκέπτονται και ενεργούν σαν τον άνθρωπο, ορθολογικά όπως αναφέρει χαρακτηριστικά και ο Kurzweil, ”Η τέχνη της δημιουργίας μηχανών που πραγματοποιούν λειτουργίες οι οποίες απαιτούν νοημοσύνη όταν πραγματοποιούνται από τους ανθρώπους” (Kurzweil,1990)

Συγκεκριμένα ο όρος της τεχνητής νοημοσύνης επινοήθηκε το 1956, αλλά έγινε πιο δημοφιλές τα τελευταία χρόνια λόγω του αυξημένου όγκου δεδομένων, των προηγμένων αλγορίθμων και των βελτιώσεων στην ισχύ των υπολογιστών και την αποθήκευση των δεδομένων.

Η τεχνητή νοημοσύνη αυτοματοποιεί την επαναληπτική μάθηση, εκτελεί συχνά, μεγάλου όγκου μηχανογραφημένα έργα, αξιόπιστα και χωρίς κόπο. Χρησιμοποιώντας αλγορίθμους στους οποίους θα αναφερθούμε παρακάτω καταφέρνει να ταξινομήσει ή να κατηγοριοποιήσει τα δεδομένα καθώς και να εκπαιδεύσει το συστήματα προκειμένου να επιτύχει καλύτερα αποτελέσματα. Επιτυγχάνει απίστευτη ακρίβεια μέσω deep neural networks – κάτι που προηγουμένως ήταν αδύνατο. Χαρακτηριστικό παράδειγμα αποτελούν οι αλληλεπιδράσεις με το Alexa, το Google Search που βασίζονται στο deep learning – και συνεχίζονται να γίνονται πιο ακριβείς όσο περισσότερο τα χρησιμοποιεί κανείς. Στον ιατρικό τομέα, τεχνικές τεχνητής νοημοσύνης όπως το deep learning, η ταξινόμηση εικόνων και η αναγνώριση αντικειμένων μπορούν πλέον να χρησιμοποιηθούν για την ανίχνευση καρκίνου σε απεικονίσεις με μαγνητική τομογραφία, με ακρίβεια παρόμοια με αυτή καταρτισμένων ακτινολόγων ή την κατηγοριοποίηση ενός ασθενούς με βάση τις ιατρικές εξετάσεις του. Με αυτό λοιπόν τον τρόπο μπορεί άμεσα να γίνει η διάγνωση της ασθένειας χωρίς την παρέμβαση ανθρώπινου νου, με την χρήση απλά της τεχνητής νοημοσύνης, αφού πρωτύτερα θα έχει εκπαιδευτεί ορθά το σύστημα μας, στην αναγνώριση συγκεκριμένων χαρακτηριστικών και μορφών που θα οδηγήσουν στο τελικό συμπέρασμα. (Kantardzic, 2003)

Αντίστοιχα και στην δική μας περίπτωση που τα δεδομένα μας αποτελούν ιατρικά δεδομένα καρκίνου του πνεύμονα, θα πραγματοποιηθεί προσπάθεια εκπαίδευσης του συστήματος ώστε με τα συγκεκριμένα εισαχθέντα χαρακτηριστικά να μπορεί να ολοκληρωθεί η κατάταξη με τον καλύτερο δυνατό τρόπο σε υγιή ή σε ασθενή άτομο, με το μικρότερο δυνατό ποσοστό λάθους.

Πλέον εισβάλει σε όλους τους τομείς όλο ένα και περισσότερο δίνοντας άμεσα λύσεις και βελτιώνοντας τις υπηρεσίες. Ικανότητες αγορών που προσφέρουν εξατομικευμένες προτάσεις και τη δυνατότητα συζήτησης των επιλογών αγοράς με τον πελάτη, καλύτερη διαχείριση των Logistics, βέλτιστη ανάλυση δεδομένων μέσα από εφαρμογές διαδικτύου των πραγμάτων (IoT) καθώς και ψυχαγωγία μέσω ηλεκτρονικών παιχνιδιών που παρέχουν άμεση αλληλεπίδραση με τον χρήστη αποτελούν μερικά από τα κύρια χαρακτηριστικά της τεχνητής νοημοσύνης που εισβάλει όλο ένα στην καθημερινότητα μας. Ουσιαστικά η τεχνητή νοημοσύνη λειτουργεί σε συνδυασμό μεγάλων ποσοτικών δεδομένων με γρήγορους, επαναληπτικής διαδικασίας και ευφυείς αλγορίθμους, επιτρέποντας στο λογισμικό να μαθαίνει αυτόματα από μορφές ή χαρακτηριστικά των δεδομένων και εν τέλει να εξάγει αποτελέσματα.

Εικόνα 2.1: Απεικόνιση τεχνητής νοημοσύνης, πηγή: (Sepe, 2018)

Μερικά πεδία της τεχνητής νοημοσύνης είναι τα ακόλουθα:

- ***Machine learning (μηχανική μάθηση):*** αυτοματοποιεί την κατασκευή αναλυτικών μοντέλων. Χρησιμοποιεί μεθόδους από τα νευρωνικά δίκτυα (neural networks), τη στατιστική, την επιχειρησιακή έρευνα (operational research) και τη φυσική για την εύρεση κρυφών γνώσεων εντός των δεδομένων χωρίς να έχει προγραμματιστεί εμφανώς για το πού να εξετάσει τι να συμπεράνει. (Sas - machine- learning, 2019)
- ***Neural network (νευρωνικό δίκτυο):*** είναι ένας τύπος μηχανικής μάθησης που αποτελείται από αλληλοσυνδεόμενες μονάδες (όπως οι νευρώνες) που επεξεργάζονται τις πληροφορίες ανταποκρινόμενο σε εξωτερικές εισαγωγές δεδομένων, προωθώντας πληροφορίες μεταξύ κάθε μονάδας. Η διαδικασία απαιτεί πολλαπλές διελεύσεις στα δεδομένα προκειμένου να βρεθούν συνδέσεις και να γίνει εξαγωγή νοήματος από ακαθόριστα δεδομένα.
- ***Deep learning (σε βάθος μάθηση):*** Χρησιμοποιεί τεράστια neural networks με πολλά επίπεδα μονάδων επεξεργασίας, αξιοποιώντας τις εξελίξεις στην υπολογιστική ισχύ και τις βελτιωμένες τεχνικές εκπαίδευσης για την μάθηση πολύπλοκων μορφών σε μεγάλες ποσότητες δεδομένων. Οι κοινές εφαρμογές της περιλαμβάνουν την αναγνώριση εικόνας και ομιλίας. (Sas - deep-learning, 2019)
- ***Cognitive computing (γνωστική υπολογιστική):*** Κατά τη χρήση τεχνικών AI και cognitive computing, ο απώτατος στόχος είναι η προσομοίωση ανθρώπινων αλληλεπιδράσεων από μια μηχανή μέσω της ικανότητας να ερμηνευτούν εικόνες και ομιλία – και να υπάρξει κανονική απάντηση από την μηχανή. (Sas -Big Data, 2019)
- ***Computer vision*** **:** βασίζεται στην αναγνώριση μορφών (pattern recognition) και στο deep learning ώστε να αναγνωρίζεται τι υπάρχει σε μια εικόνα ή ένα βίντεο. Όταν οι μηχανές μπορούν να επεξεργαστούν, να αναλύσουν και να

κατανοήσουν εικόνες μπορούν να συλλάβουν εικόνες ή βίντεο σε πραγματικό χρόνο και να ερμηνεύσουν τα περιβάλλοντά τους.

- ***Natural language processing (NLP)* - (επεξεργασία φυσικής γλώσσας ή ΕΦΓ):** είναι η ικανότητα των υπολογιστών να αναλύουν, να κατανοούν και να παράγουν ομιλούμενη γλώσσα, συμπεριλαμβανομένης της ομιλίας. Το επόμενο στάδιο στην ΕΦΓ είναι η φυσική γλωσσική αλληλεπίδραση, η οποία επιτρέπει στους ανθρώπους να επικοινωνούν με υπολογιστές χρησιμοποιώντας την κανονική, καθημερινή γλώσσα για την εκτέλεση καθηκόντων.

2.1 Εξόρυξη δεδομένων (Data mining)

Εξόρυξη δεδομένων είναι η εξεύρεση μιας πληροφορίας ή προτύπων από μεγάλες βάσεις δεδομένων με χρήση αλγορίθμων ομαδοποίησης ή κατηγοριοποίησης και των αρχών της στατιστικής, της τεχνητής νοημοσύνης, της μηχανικής μάθησης και των συστημάτων βάσεων δεδομένων. Στόχος της εξόρυξης δεδομένων είναι η πληροφορία που θα εξαχθεί και τα πρότυπα που θα προκύψουν να έχουν δομή κατανοητή προς τον άνθρωπο έτσι ώστε να τον βοηθήσουν να πάρει τις κατάλληλες αποφάσεις. (Wikipedia, 2019)

Ιδιαίτερη συμβολή στην ανάπτυξη της εξόρυξης δεδομένων είχε πρωταρχικά η βελτίωση της υπολογιστικής ισχύος, τα νευρωνικά δίκτυα, η συσταδοποίηση, οι γενετικοί αλγόριθμοι (1950), τα δέντρα απόφασης (1960) και η μηχανή υποστήριξης διανυσμάτων (1990), με σκοπό την αποκάλυψη άγνωστων προτύπων. (Mehmed, 2003)

Η διαδικασία ανακάλυψης γνώσης από βάσεις δεδομένων(KDD) συνήθως ορίζεται από τα εξής στάδια: (Usama, 1996)

1. Συλλογή
2. Προεπεξεργασία
3. Μετασχηματισμός
4. Εξόρυξη δεδομένων
5. Ερμηνεία/Αξιολόγηση

2.1.1 Απαιτήσεις της εξόρυξης δεδομένων

Προκειμένου να επιτευχθεί ένα ολοκληρωμένο αποτέλεσμα από μία διαδικασία Εξόρυξης Δεδομένων (ΕΔ) προαπαιτείται έλεγχος των χαρακτηριστικών του συστήματος καθώς και των απαιτήσεων για την εφαρμογή των τεχνικών.

Παρ όλα αυτά πολλούς από τους στόχους που θέτουμε παρακάτω για την υλοποίηση μπορεί να είναι αντικρουόμενοι. Για παράδειγμα, ο στόχος της προστασίας της ασφάλειας δεδομένων μπορεί να αντικρούει στην απαίτηση για διαλογική εξόρυξη πολυεπίπεδης γνώσης από διαφορετικές σκοπιές. Με βάση τους (Chen, 1996), τα κυριότερα ζητήματα (ομαδοποιήσεις απαιτήσεων) που οφείλουμε κάθε φορά να λαμβάνουμε υπόψη είναι:

i) ***Χειρισμός διαφορετικών τύπων δεδομένων.***
Ένα σύστημα ΕΔ πρέπει να μπορεί να εφαρμόζεται σε διαφορετικούς τύπους δεδομένων, καθώς συχνά χρησιμοποιούνται διαφορετικοί τύποι και ΒΔ σε διαφορετικές εφαρμογές. Επίσης, παρατηρείται συχνά η ύπαρξη συγγενών (relational) βάσεων δεδομένων. Επομένως, πρέπει ένα σύστημα ΕΔ να είναι σε θέση να υποστηρίζει τεχνικές για αποδοτική και αποτελεσματική ανάλυση συγγενικών δεδομένων. Επιπρόσθετα ένα τέτοιο σύστημα θα έπρεπε να λειτουργεί ανεξάρτητα από τύπους δεδομένων, καθώς πολλά σύγχρονα συστήματα βάσεων δεδομένων περιέχουν σύνθετους τύπους δεδομένων (δομές δεδομένων και σύνθετα αντικείμενα, υπερκείμενο και στοιχεία πολυμέσων, χωροχρονικά στοιχεία κ.λπ.).
Η ποικιλία των τύπων δεδομένων και οι διαφορετικοί στόχοι της ΕΔ κάνουν πιο απίθανη την ύπαρξη ενός συστήματος ΕΔ που να μπορεί να χειριστεί όλα αυτά τα είδη δεδομένων.

ii) ***Αποδοτικότητα των αλγορίθμων ΕΔ***
Για να έχουμε αποτελεσματική εξόρυξη γνώσης από μεγάλα σύνολα δεδομένων, πρέπει να έχουμε αλγορίθμους κατάλληλα προσαρμοσμένους σε αυτά. Επομένως, ο χρόνος εκτέλεσης των αλγορίθμων πρέπει να είναι αποδεκτός και αναμενόμενος για μεγάλες βάσεις δεδομένων. Συνεπώς οι αλγόριθμοι πρέπει να είναι αποδοτικοί ικανοί να επεξεργαστούν μεγάλο όγκο δεδομένων σε μικρό χρονικό διάστημα.

iii) ***Χρησιμότητα, βεβαιότητα, εκφραστικότητα των αποτελεσμάτων της ΕΔ***

Η εξορυγμένη γνώση πρέπει να παρουσιάζει με ακριβή τρόπο τα περιεχόμενα των βάσεων δεδομένων και να είναι χρήσιμη για συγκεκριμένες εφαρμογές. Η ακρίβεια των αποτελεσμάτων θα μπορούσε να εκφραστεί μέσω κάποιων μέτρων βεβαιότητας, προσεγγιστικά ή ποσοτικά. Εξαιρέσεις όπως θόρυβος και outliers πρέπει να αντιμετωπιστούν από τα συστήματα ΕΔ. Το γεγονός αυτό δίνει το κίνητρο για μια συστηματική μελέτη της ποιότητας της εξορυγμένης γνώσης, κατασκευάζοντας στατιστικά ή αναλυτικά μοντέλα, μοντέλα προσομοίωσης, καθώς και τα εργαλεία αυτών.

iv) ***Εκφράσεις διαφορετικού τύπου για τα αποτελέσματα***

Όπως μπορούμε να φανταστούμε, από μεγάλα σύνολα δεδομένων μπορούν να προκύψουν διαφορετικοί τύποι γνώσεων. Συνεπώς, θα ήταν πολύ χρήσιμο να μπορούμε να ελέγξουμε τη γνώμη από διαφορετικές απόψεις και να την εκφράσουμε σε διάφορες μορφές. Θεωρείται ότι θα ήταν πολύ καλό να μπορούν να εκφραστούν τα ερωτήματα της ΕΔ και η εξορυμένη γνώση σε γλώσσες υψηλού επιπέδου ή μέσω γραφικών διεπαφών των χρηστών. Έτσι, η ΕΔ θα μπορούσε να είναι εφαρμόσιμη και από μη ειδικούς και η εξορυμένη γνώση θα χρησιμοποιούταν άμεσα από όλους. Τέλος, απαιτείται το σύστημα να υιοθετήσει εκφραστικές τεχνικές αναπαράστασης της γνώσης, έτσι ώστε να επιτευχθεί η αποτελεσματική παρουσίαση της γνώσης.

v) ***Διαλογική ανακάλυψη γνώσης στα πολλαπλά εννοιολογικά επίπεδα***

Είναι δύσκολο να προβλεφθεί αυτό που θα μπορούσε να ανακαλυφθεί επακριβώς από μια βάση δεδομένων. Γι’ αυτό, θα μπορούσε να καθοριστεί μια σειρά ερωτήσεων της ΕΔ προκειμένου να διαμορφωθεί η εστίαση στα δεδομένα, να δημιουργηθεί ένα λεπτομερέστερο επίπεδο ΕΔ και να παρατηρηθούν τα αποτελέσματα της ΕΔ σε πολλαπλά επίπεδα και από διαφορετικές πτυχές. Όλα αυτά μπορούν να επιτευχθούν μέσω της διαλογικής ανακάλυψης της γνώσης.

vi) ***Εξόρυξη πληροφορίας από διαφορετικές πηγές δεδομένων***

Σε σχέση με τη σύνδεση των διάφορων πηγών δεδομένων, υπάρχει προβάδισμα της ευρέως διαθέσιμης σύνδεσης υπολογιστών σε τοπικό και ευρύτερο δίκτυο, συμπεριλαμβανομένου του διαδικτύου. Αυτό οδηγεί στη δημιουργία μεγάλων κατανεμημένων και ετερογενών βάσεων δεδομένων. Επιπλέον, το τεράστιο μέγεθος των βάσεων δεδομένων, η υψηλή κατανομή των δεδομένων και η υπολογιστική

πολυπλοκότητα ορισμένων μεθόδων ΕΔ οδηγούν στην ανάπτυξη παράλληλων και κατανεμημένων αλγορίθμων ΕΔ.

vii) ***Προστασία ιδιωτικότητας και ασφάλεια δεδομένων***
Η προστασία και αποκλειστικότητα των δεδομένων απειλείται στην περίπτωση που αυτά μπορούν να παρατηρηθούν από πολλές διαφορετικές σκοπιές. Είναι σημαντικό να μελετηθεί πότε μπορεί να οδηγηθεί ένα σύστημα σε μια εισβολή στην ιδιωτικότητα μέσω της KDD και τι μέτρα ασφαλείας μπορούν να αναπτυχθούν για να εμποδιστεί η αποκάλυψη των ευαίσθητων πληροφοριών. (Σταυλιώτης, 2008)

3ο Κεφάλαιο - Μέθοδοι εξόρυξης δεδομένων

Στην εξόρυξη δεδομένων χρησιμοποιούνται πολλά είδη μεθόδων όπως οι ακόλουθοι:

- Ταξινόμηση (Classification) / Δυαδική Ταξινόμηση(Binary Classification)
- Συσχέτιση (Association)
- Ομαδοποίηση (Clustering)

3.1 Ταξινόμηση (Classification)

Η ταξινόμηση ή αλλιώς κατηγοριοποίηση είναι μία τεχνική εξόρυξης δεδομένων κατά την οποία ένα στοιχείο ανατίθεται σε ένα προκαθορισμένο σύνολο κατηγοριών. (Βαζιργιάννης, 2005)

Η Δυαδική ταξινόμηση έχει ως στόχο την ταξινόμηση των στοιχείων ενός συνόλου σε δύο ομάδες, με βάση έναν συγκεκριμένο κανόνα ταξινόμησης.

Πιο συγκεκριμένα οι αλγόριθμοι ταξινόμησης εφαρμόζονται σε δεδομένα τα οποία έχουν πρώτα ταξινομηθεί σε συγκεκριμένες κλάσεις με σκοπό την εξαγωγή κανόνων οι οποίοι χρησιμοποιούνται μετέπειτα για την ταξινόμηση νέων δειγμάτων στις ίδιες κλάσεις. Το κάθε σύνολο των εξαγόμενων κανόνων ονομάζεται ταξινομητής (classifier).

Η κατηγοριοποίηση αποτελεί μια διαδικασία δύο σταδίων. Το πρώτο βήμα αποτελεί η κατασκευή ενός μοντέλου πρόβλεψης βασιζόμενο στην εκπαίδευση του συνόλου δεδομένων. Το δεύτερο βήμα αναφέρεται στη δοκιμή και αξιολόγηση του μοντέλου. Εάν η ακρίβεια ταξινόμησης των δοκιμαστικών συνόλων δεδομένων είναι αποδεκτή, το μοντέλο μπορεί να χρησιμοποιηθεί για την ταξινόμηση νέων–άγνωστων δεδομένων σε νέες εγγραφές (Olson & Shi, 2005).

Για την κατασκευή του μοντέλου πρόβλεψης καθιστάται απαραίτητη η αρχική εκπαίδευση του συστήματος από ένα σύνολο δεδομένων μέσω του οποίου θα προκύψουν οι κανόνες κατηγοριοποίησης της εξαρτημένης μεταβλητής.

Οι γνωστότερες μέθοδοι κατηγοριοποίησης είναι :

- *Η μάθηση των εννοιών*
  Το σύστημα τροφοδοτείται με θετικά παραδείγματα (ανήκουν σε κάποια κατηγορία) ή αρνητικά (δεν ανήκουν σε κάποια κατηγορία), και παράγει ένα γενικευμένο κανόνα ώστε να είναι σε θέση στην συνέχεια να αποφασίσει για άγνωστες περιπτώσεις.
- *Τα δέντρα απόφασης*
  Δημιουργούνται για ένα συγκεκριμένο σύνολο δεδομένων μέσω μιας διαδικασίας εκπαίδευσης. Ο γνωστότερος αλγόριθμος είναι ο ID3 μέσω του οποίου κατασκευάζεται ένα δέντρο για το συγκεκριμένο σύνολο δεδομένων ορίζοντας το κριτήριο διαχωρισμού των δεδομένων (για παράδειγμα την ανεξάρτητη μεταβλητή)
- *Η μάθηση με βάση τα παραδείγματα*
  Τα δεδομένα σύμφωνα με τον αλγόριθμο k-Nearest Neighbors αναπαρίστανται γραφικά σε κάποιο Ευκλείδειο χώρο με τόσες διαστάσεις όσα και τα πεδία των εγγραφών που υφίσταται .
- *Η μάθηση με βάση την θεωρία του Bayes.*
  Κατά την μάθηση με βάση την θεωρία του Bayes κάθε παράδειγμα εκπαίδευσης μπορεί να οδηγήσει σε μείωση ή σε αύξηση της πιθανότητας Μια υπόθεση δεν απορρίπτεται αλλά μειώνεται η πιθανότατα της.

3.2 Ομαδοποίηση (Clustering)

Στόχος είναι η εύρεση ενός συνόλου από ομάδες με κοινά χαρακτηριστικά. Η μέθοδος της ομαδοποίησης μοιάζει με την μέθοδο της ταξινόμησης ωστόσο η κυρίαρχη διαφορά είναι ότι τα δεδομένα του συνόλου εκπαίδευσης δεν είναι προταξινομημένα. Το σύνολο των στοιχείων με κοινά χαρακτηριστικά-ομοιότητες

χωρίζεται σε ομάδες , σε ορισμένους αλγορίθμους μια εγγραφή μπορεί να ανήκει σε περισσότερες από μια ομάδες ταυτόχρονα (διάγραμμα Venn).

Πιο συγκεκριμένα δημιουργούνται ομάδες όμοιων αντικειμένων , «ομάδες αποτελούν το σύνολο των σημείων για τα οποία ισχύει ότι η απόσταση ανάμεσα στα σημεία που ανήκουν στην ίδια ομάδα είναι μικρότερη από την απόσταση μεταξύ ενός σημείου της ομάδας και ενός από οποιαδήποτε άλλη ομάδα» (Dunham, 2004)

Γενικότερα υπάρχουν οι εξής τρείς κατηγορίες αλγορίθμων ομαδοποίησης :

- Οι αλγόριθμοι οι οποίοι προσπαθούν να βρουν τον καλύτερο διαχωρισμό ενός συνόλου δεδομένων σε κάποιο συγκεκριμένο αριθμό ομάδων. Χαρακτηριστικό παράδειγμα αποτελεί ο K-Means όπου ως κ ορίζουμε τον αριθμό των ομάδων.
- Οι ιεραρχικοί αλγόριθμοι οι οποίοι με ιεραρχικό τρόπο προσπαθούν να ανακαλύψουν τον αριθμό των ομάδων. Χαρακτηριστικό παράδειγμα τέτοιων αλγορίθμων είναι οι αλγόριθμοι της συγχώνευσης.
- Οι πιθανοκρατικοί αλγόριθμοι οι οποίοι βασίζονται σε μοντέλα πιθανοτήτων.

Η μέθοδος της ομαδοποίησης μπορεί να είναι είτε στατιστική είτε αριθμητική οπότε χρησιμοποιούνται διάφορα κριτήρια ομοιότητας.

Η τεχνική της ομαδοποίησης υπάγεται στην ευρύτερη κατηγορία των τεχνικών μάθησης χωρίς επίβλεψη. (Pourrajabi, Moulavi, & Campello, 2014)

3.3 Συσχέτιση (Association)

Αποτελεί μία από τις πιο διαδεδομένες τεχνικές εξόρυξης δεδομένων για την δημιουργία αλληλεξαρτήσεων. Ουσιαστικά προσπαθούμε να βρούμε συσχετίσεις στα δεδομένα μας, γίνεται εύρεση σημαντικών/συχνών εξαρτήσεων μεταξύ των γνωρισμάτων μας. Η Ανάλυση Κανόνων Συσχέτισης θεωρείται το πιο γνήσιο τέκνο της Εξόρυξης Δεδομένων, καθώς οι άλλες μέθοδοι προέρχονται από τη Μηχανική Μάθηση, τη Στατιστική κλπ. Στόχος των Κανόνων Συσχέτισης είναι η ανακάλυψη σχέσεων μεταξύ τιμών των γνωρισμάτων, οι οποίες εμφανίζονται συχνά μαζί. Για την καλύτερη κατανόηση του αντικειμένου παραθέτουμε το παρακάτω παράδειγμα. Θεωρήστε τις πωλήσεις ενός σούπερ μάρκετ. Για κάθε συναλλαγή πώλησης (απόδειξη λιανικής) καταγράφονται τα προϊόντα που αγόρασε ο καταναλωτής. Το ερώτημα είναι εάν υπάρχουν προϊόντα τα οποία πωλούνται συχνά μαζί, εάν υπάρχουν δηλαδή ομάδες καταναλωτών που επιλέγουν να αγοράσουν

κοινά προϊόντα. Οι Κανόνες Συσχέτισης ανακαλύπτουν τέτοιες σχέσεις και τις ποσοτικοποιούν, καταγράφοντας ποσοστά εμφάνισης τους. Για παράδειγμα, εξορύσσονται κανόνες που αναφέρουν ότι όταν αγοράζεται το προϊόν Α, τότε αγοράζεται ταυτόχρονα και το προϊόν Β, και παρατίθενται οι πιθανότητες εμφάνισης αυτού του γεγονότος. Οι Κανόνες Συσχέτισης μπορούν να χρησιμοποιηθούν για τη διαρρύθμιση των ραφιών ενός σούπερ μάρκετ, ώστε παροτρύνοντας τον καταναλωτή να αυξηθούν οι πωλήσεις. Γενικώς, η Ανάλυση Κανόνων Συσχέτισης εφαρμόζεται σε μεγάλο βαθμό για την επίτευξη διασταυρούμενων πωλήσεων. Ειδικά στο ηλεκτρονικό εμπόριο, όπου υπάρχει άμεση τροφοδότηση δεδομένων από τον πελάτη, καθώς και δυνατότητα άμεσης ανάλυσης αυτών των δεδομένων και αντιπαραβολής με ιστορικά στοιχεία, η προώθηση πρόσθετων προϊόντων στον πελάτη γίνεται άμεσα, την ώρα της επίσκεψης στην ιστοθέση. Ο πελάτης πραγματοποιεί αγορές και ταυτόχρονα οι αλγόριθμοι εξόρυξης δεδομένων εντοπίζουν άλλα προϊόντα, τα οποία πωλούνται συχνά μαζί με τα προϊόντα που επέλεξε ο συγκεκριμένος πελάτης. Άμεσα παρουσιάζεται στον πελάτη ένα μήνυμα της μορφής «Οι πελάτες που αγόρασαν αυτά τα προϊόντα αγόρασαν επίσης τα παρακάτω προϊόντα». Η εξόρυξη Κανόνων Συσχέτισης επιτρέπει τον εντοπισμό καταναλωτικών προτύπων και την καλύτερη κατανόηση των πραγματικών αναγκών των πελατών. Οι πληροφορίες αυτές χρησιμοποιούνται για την προσωποποιημένη κατεύθυνση στον πελάτη και τη διεξαγωγή μάρκετινγκ ένα-προς-ένα. Επιπρόσθετα σχεδόν σε κάθε βιβλίο Εξόρυξης Δεδομένων αναφέρεται το παράδειγμα, όπου βρέθηκε ότι πωλούνται συχνά μαζί μπύρες και πάνες υγιεινής για βρέφη. Η εξήγηση του φαινομένου είναι αρκετά απλή. Άρρενες γονείς, όταν επισκέπτονται το σούπερ μάρκετ για τις οικογενειακές αγορές, αγοράζουν ταυτόχρονα και τις αγαπημένες τους μπύρες. Με αυτό λοιπόν τον τρόπο τα δεδομένα αυτά μπορεί να χρησιμοποιηθούν για να καθοριστεί το μίγμα και η θέση των προϊόντων που διατίθενται στο υποκατάστημα.

4ο Κεφάλαιο - Αλγόριθμοι

4.1 Αλγόριθμος Κ-Μέσων

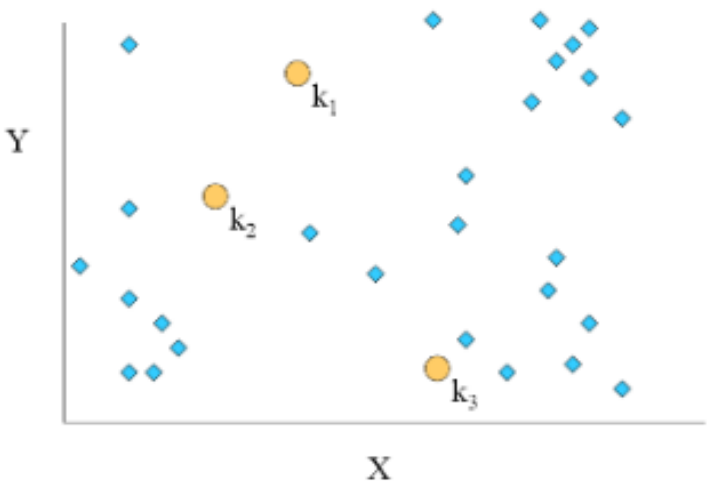
Ο αλγόριθμος Κ- Μέσων αποτελεί έναν από τους δημοφιλέστερους αλγορίθμους ομαδοποίησης που ανήκουν στις τεχνικές μάθησης χωρίς επίβλεψη. Τα βήματα που ακολουθεί ο αλγόριθμος είναι τα εξής:

Αρχικά καθορίζεται ο αριθμός των ομάδων (Κ) που θα προκύψουν. Η κύρια ιδέα είναι να προσδιοριστούν αρχικά k centroids (κεντροειδή), ένα για κάθε cluster. Αυτά τα αρχικά centroids πρέπει να επιλεγούν με επιδέξιο τρόπο, γιατί διαφορετικές αρχικές θέσεις για τα centroids δίνουν διαφορετικά αποτελέσματα. Δηλαδή, η αρχική θέση των centroids επηρεάζει το αποτέλεσμα που θα δώσει ο αλγόριθμος. Έτσι, συχνά θεωρείται καλύτερη η επιλογή εκείνων των centroids ώστε να απέχουν μεταξύ τους όσο περισσότερο γίνεται.

Το επόμενο βήμα είναι επιλογή κάθε στοιχείου από το σύνολο δεδομένων και συσχέτιση του με το κοντινότερο σε αυτό centroid. Όταν αυτό γίνει για όλα τα στοιχεία του συνόλου δεδομένων, το πρώτο βήμα έχει ολοκληρωθεί και μία πρώτη και «πρόχειρη» ομαδοποίηση έχει ήδη προκύψει.

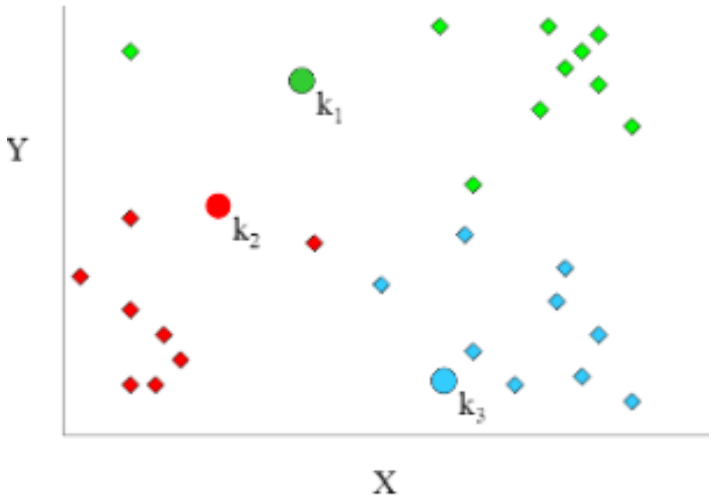
Στη συνέχεια, απαιτείται να υπολογιστούν ξανά k νέα centroids, τα οποία θα αποτελούν το κέντρο βάρους για κάθε ένα cluster που προέκυψε από το προηγούμενο βήμα. Αφού λοιπόν οριστούν τα νέα k centroids, ακολουθεί και πάλι η ίδια διαδικασία ανάθεσης καθενός από τα στοιχεία του συνόλου δεδομένων στο κοντινότερο με αυτό, νέο πλέον, centroid. Αποτέλεσμα αυτής της διαδικασίας είναι να αλλάζουν τα centroid σε κάθε εκτέλεση και τα στοιχεία αντίστοιχα ανατίθενται στο κατάλληλο cluster με βάση το κοντινότερο centroid που ορίστηκε νωρίτερα. Όταν μετά από κάποιο αριθμό επαναλήψεων δεν αλλάξουν θέσεις τα κεντρικά στοιχεία τότε ο αλγόριθμος τερματίζει και το αποτέλεσμα είναι η ομαδοποίηση σε κ clusters.

Παρακάτω αναλύεται γραφικά ένα παράδειγμα Κ μέσων ορίζοντας εξαρχής 3 clusters. Στην ακόλουθη εικόνα διαφαίνεται το αρχικό στάδιο όπου ορίζουμε τα αρχικά σημεία k1,k2,k3.



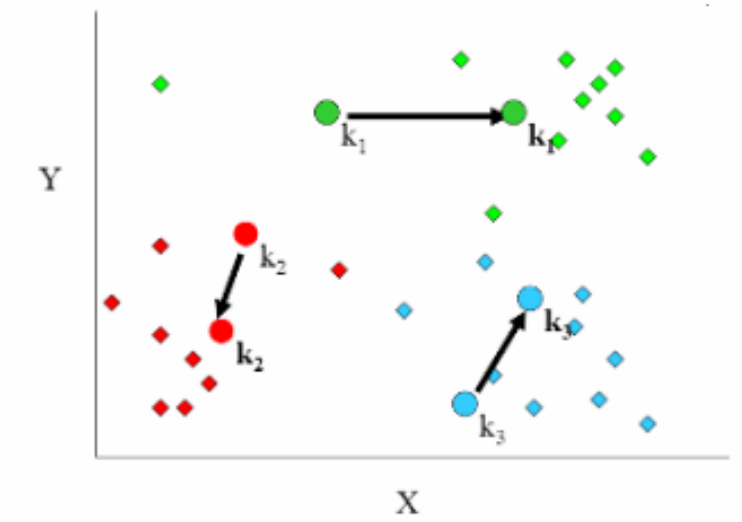
Εικόνα 4.1 : Ορισμός κεντροειδών

Στο επόμενο στάδιο τα στοιχεία μας ανατίθενται στα πιο γειτονικά σημεία από τα σημεία που έχουν ορισθεί k1,k2,k3.



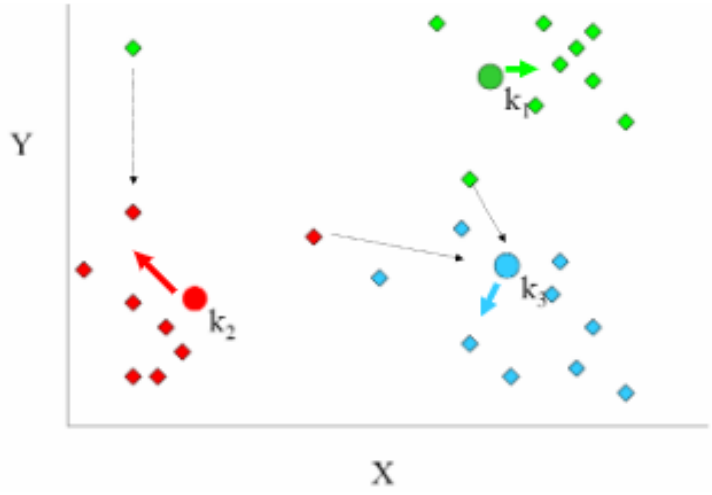
Εικόνα 4.2 : Συσχέτιση κάθε στοιχείου στο κοντινότερο κεντροειδή

Στο μετέπειτα στάδιο γίνεται επανα-υπολογισμός του κέντρου βάρους κάθε σημείου όπως διαφαίνεται στην ακόλουθη εικόνα. (N. Tan, 2006)



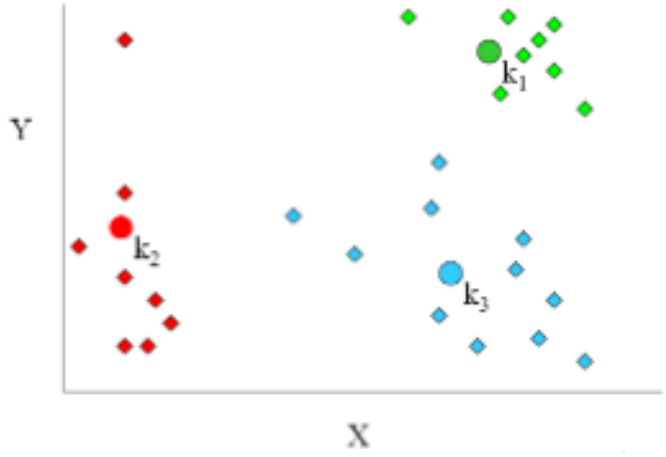
Εικόνα 4.3 : Επαναυπολογισμός κεντροειδή

Συνεπώς έχουμε πλέον νέα ανάθεση σημείων και εκ νέου υπολογισμός του κέντρου βάρους των στοιχείων.



Εικόνα 4.4 : Υπολογισμός νέου κεντρου βάρους

Ο αλγόριθμος θα ολοκληρωθεί μόλις τα κεντρικά σημεία πλέον δεν αλλάζουν θέση μετά από κάποιο αριθμό επαναλήψεων, οπότε και έχουν δημιουργηθεί τα τρία clusters.



Εικόνα 4.5 : Τελική δημιουργία ομάδων μετά από κάποιο αριθμό επαναλήψεων

Ο αλγόριθμος στοχεύει να ελαχιστοποιήσει μία αντικειμενική συνάρτηση, την λεγόμενη συνάρτηση τετραγωνικού λάθους που ορίζεται ως εξής:

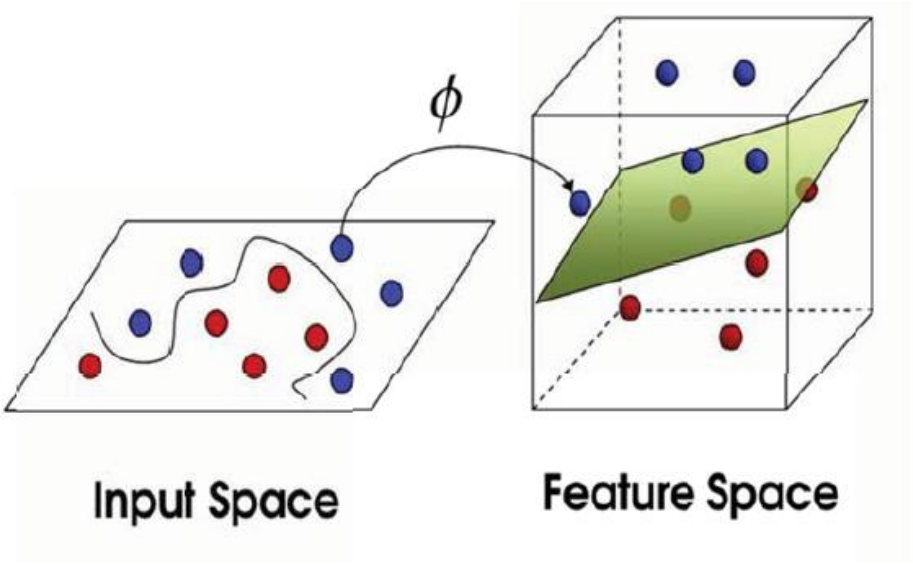
$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

$\left\| x_i^{(j)} - c_j \right\|^2$: Αντιστοιχεί στο μέτρο απόστασης που χρησιμοποιείται για να μετρά την απόσταση κάθε στοιχείου ()jix από το centroid jc του κάθε cluster. Όπου n το σύνολο των στοιχείων του συνόλου των δεδομένων.

4.2 Μηχανές διανυσμάτων υποστήριξης (SVM)

Οι μηχανές διανυσμάτων υποστήριξης ή αλλιώς Support Vector Machines (SVM) αποτελούν πιθανότατα την ακριβέστερη μέθοδο μεταξύ των παραπάνω αλγορίθμων. Ο SVM αλγόριθμος λειτουργεί ως εξής: χαρτογραφεί το δοθέν σύνολο σε ένα πιθανό πολυδιάστατο χώρο διανυσμάτων και προσπαθεί να εντοπίσει σε αυτό το χώρο ένα πεδίο το οποίο να διαχωρίζει τα θετικά από τα αρνητικά παραδείγματα. Έχοντας βρει ένα τέτοιο πεδίο ο αλγόριθμος μπορεί να προβλέψει την κατηγοριοποίηση ενός αχαρακτήριστου παραδείγματος χαρτογραφώντας το στον χώρο που περιέχει τα χαρακτηριστικά και ψάχνοντας σε ποια πλευρά του διαχωριστικού πεδίου βρίσκεται. Πιο συγκεκριμένα έστω ότι έχουμε ένα πρόβλημα δύο τάξεων και προσπαθούμε να βρούμε την καλύτερη συνάρτηση ταξινόμησης που να ξεχωρίζει τα μέλη των δύο τάξεων. Αυτό μπορεί να γίνει αντιληπτό

γεωμετρικά ως εξής. Για το σετ λοιπόν αυτό που θεωρούμε ότι είναι γραμμικά διαχωρίσιμο αντιστοιχεί μία γραμμική συνάρτηση στο υπερεπίπεδο διαχωρισμού f(x), η οποία περνά μεταξύ των δύο τάξεων με αποτέλεσμα να χωρίζονται σε δύο διαφορετικά μέρη. Αφού λοιπόν προσδιορίσουμε αυτή την συνάρτηση κάθε νέο στοιχείο που έχουμε μπορεί να ταξινομηθεί απλά βρίσκοντας το πρόσημο της συνάρτησης αυτής καθώς αν f(x)<0 τότε θα ανήκει στην αρνητική τάξη.



Εικόνα 4.6 : Μετατροπή σε υπερεπίπεδο διαχωρισμού μέσω svm, πηγή: (Crdd, 2018)

Προκειμένου να δούμε αν το υπερεπίπεδο είναι μέγιστου περιθωρίου ο ταξινομητής προσπαθεί να μεγιστοποιήσει την ακόλουθη συνάρτηση ως προς τα w και b :

$$L_p = \frac{1}{2}\|\vec{w}\| - \sum_{i=1}^{t} a_i y_i (\vec{w} \cdot \vec{x}_i + b) + \sum_{i=1}^{t} a_i$$

Όπου t είναι ο αριθμός των υποδειγμάτων εκπαίδευσης και a είναι οι Λανγκραντιανοί πολλαπλασιαστές και η L η Λανγκραζιανή συνάρτηση. Τέλος το διάνυσμα w και η σταθερά b ορίζουν το υπερεπίπεδο. (Jordan, 2019)

4.3 Κατηγοριοποιητής k- πλησιέστερων γειτόνων (KNN)

Ο αλγόριθμος k πλησιέστερων γειτόνων είναι ένας σειριακός αλγόριθμος, όπου τα στοιχεία συγχωνεύονται επαναληπτικά στις πλησιέστερες μεταξύ των κλάσεων που υπάρχουν, σε κάθε νέα επανάληψη. Ουσιαστικά κατά τον KNN εισάγουμε τα δεδομένα εκπαίδευσης καθώς και το σύνολο των δεδομένων για test. Ο k-NN εκπαιδεύεται, αποθηκεύοντας τα διανύσματα που αντιστοιχούν στα παραδείγματα εκπαίδευσης, και μαζί και τις εξόδους των στοιχείων αυτών. Αποθηκεύει δηλαδή τα σημεία ενός πολυδιάστατου χώρου, στον οποίο μπορούν να αναπαρασταθούν τα στοιχεία αυτά. Κατά την φάση της ταξινόμησης, δηλαδή κατά τη χρήση του εκπαιδευμένου k-NN, το σύστημα λαμβάνει τις νέες εισόδους, για τις οποίες δεν γνωρίζει την έξοδο και υπολογίζει για κάθε μία τη διανυσματική της αναπαράσταση, δηλαδή το αντίστοιχο σημείο στον πολυδιάστατο χώρο. Έπειτα, υπολογίζεται η απόσταση του σημείου του στοιχείου εισόδου από κάθε σημείο που αντιστοιχεί σε ένα αποθηκευμένο παράδειγμα εκπαίδευσης. Η απόσταση ορίζεται χρησιμοποιώντας Ευκλείδεια μετρική, ενώ στην περίπτωση μη ετερόκλητων χαρακτηριστικών, απαιτείται κανονικοποίηση των τιμών τους σε κοινή ακτίνα τιμών.

Αφού υπολογιστούν οι αποστάσεις αυτές, είναι εύκολο να βρεθούν τα k στοιχεία εκπαίδευσης με τη μικρότερη απόσταση από το σημείο της εισόδου και έτσι η είσοδος κατατάσσεται στην κατηγορία που είναι πιο συχνή μεταξύ των k κοντινότερων παραδειγμάτων εκπαίδευσης. Το k είναι ένας φυσικός αριθμός ο οποίος αποτελεί παράμετρο του αλγορίθμου. Συνήθως είναι ένας περιττός φυσικός αριθμός για να μπορεί να υπάρξει πλειοψηφία. Ο αλγόριθμος απαιτεί περισσότερους υπολογισμούς κατά την κατάταξη νέων στοιχείων, όσο αυξάνει το πλήθος των παραδειγμάτων εκπαίδευσης, αφού υπολογίζεται κάθε φορά η απόσταση του από όλα τα παραδείγματα εκπαίδευσης. Έχει επίσης, μεγάλες απαιτήσεις μνήμης, αφού πρέπει να αποθηκεύονται όλα τα παραδείγματα εκπαίδευσης ωστόσο αποτελεί έναν ταχύτατο αλγόριθμο εκπαίδευσης. (Dunham, 2004), (Χρονάκης, 2006)

Αφού γίνει η κατάταξη ο αλγόριθμος ελέγχει την απόδοση του με τρείς τρόπους. Αρχικά με τα δεδομένα ελέγχου (dataset for test), που λαμβάνει σαν παράμετρο μαζί με τα δεδομένα εκπαίδευσης. Έπειτα ο αλγόριθμός μας δίνει σαν έξοδο την μήτρα αληθείας (confusion matrix), η οποία μας δίνει το πλήθος των αληθών θετικών και αρνητικών σε σχέση με το σύνολο των δεδομένων, όπως και το σύνολο των ψευδών θετικών και αρνητικών. Και τέλος μπορούμε να ελέγξουμε την απόδοση του αλγορίθμου μέσω της καμπύλης OC (operating characteristic).

Στην ακόλουθη εικόνα φαίνεται ο αλγόριθμος KNN, όπου N Θεωρούμε μια δομή οργανωμένη με βάση την ομοιότητα.

```
Input:
  T            //training data
  K            //Number of neighbors
  t            //Input tuple to classify
Output:
  c            //Class to which t is assigned
KNN algorithm:  //Algorithm to classify tuple using KNN
begin
  N = ∅;       //Find set of neighbors, N, for t
  for each d ∈ T do
       if |N| ≤ K, then
              N = N ∪ {d};
       else
              if ∃ u ∈ N such that sim(t,u) ≤ sim(t,d), then
                     begin  N = N - {u}; N = N ∪ {d}; end
              //Find class for classification
  c = class to which the most u ∈ N are classified
end
```

Εικόνα 4.7 : Αλγόριθμος Knn, πηγή : (Dunham, 2004)

4.4 Αλγόριθμος C4.5

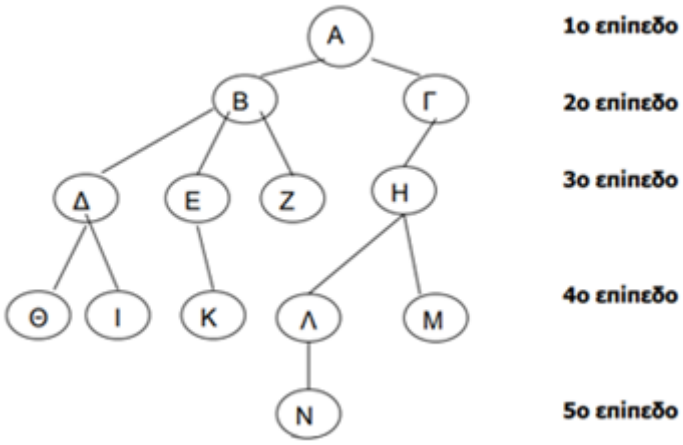
Μια από τις ποιο γνωστές μεθόδους μηχανικής μάθησης για τον σχηματισμό κανόνων απόφασης μέσω δέντρου είναι ο αλγόριθμος C4.5 ο οποίος είναι μια εξέλιξη του αλγορίθμου ID3.

Τα δέντρα απόφασης αποτελούν μια από τις ποιο βασικές μεθόδους πρόβλεψης και ταξινόμησης και αποτελούν μια επαγωγική διαδικασία. Αναπαριστούν κανόνες και είναι αρκετά διαδεδομένα καθώς ο μηχανισμός της διαδικασίας απόφασης είναι αρκετά εμφανής και επιτρέπει αρκετά καλή ανάλυση της γνώσης την οποία λαμβάνουμε. Τα κύρια χαρακτηριστικά του είναι οι κόμβοι, τα κλαδιά και τα φύλλα. Σύμφωνα με έναν ορισμό, ένα δέντρο απόφασης είναι ένα δέντρο με τις ακόλουθες ιδιότητες:

- Κάθε εσωτερικός κόμβος ονοματίζεται με το όνομα ενός χαρακτηριστικού x.
- Κάθε κλαδί / σύνδεση ονοματίζεται με ένα κατηγόρημα που μπορεί να εφαρμοστεί στο χαρακτηριστικό που αποτελεί το όνομα του κόμβου.
- Κάθε φύλλο ονοματίζεται με το όνομα μιας κλάσης

Οι κόμβοι του δέντρου αφορούν τον έλεγχο ενός συγκεκριμένου χαρακτηριστικού. Τα κλαδιά περιέχουν συνθήκες σύγκρισης (κυρίως ανισότητες) της τιμής που λαμβάνει το συγκεκριμένο χαρακτηριστικό με μια άλλη τιμή η οποία και θα

προσδιορίσει σε ποιο κόμβο «παιδί» θα συνεχίσουμε την αναζήτηση ώστε να φτάσουμε στην κλάση την οποία και θέλουμε να προβλέψουμε και αναπαρίσταται στα φύλλα του δέντρου.



Εικόνα 4.8 : Κάθετη ανάπτυξη δέντρου

Ο αλγόριθμος j48 είναι η έκδοση του C4.5 για την πλατφόρμα του WEKA. Ο ID3 υπήρξε ο κυριότερος εκπρόσωπος των TDIDT δέντρων μέχρι την έλευση του C4.5. Ήταν ο πρώτος αλγόριθμος που χρησιμοποίησε για το κριτήριο καταλληλόλητας τεμαχισμού το κέρδος Gain από τη θεωρία πληροφορίας.

Το αποτέλεσμα είναι μια δενδροειδής δομή που με γραφικό τρόπο αναπαριστά τις συσχετίσεις στα δεδομένα εκπαίδευσης ή διαφορετικά, περιγράφει τα δεδομένα. Αρχικά, μια από τις παραμέτρους του συνόλου εκπαίδευσης ορίζεται ως παράμετρος στόχος. Οι υπόλοιπες παράμετροι θεωρούνται παράμετροι εισόδου. Τα βήματα τα οποία ακολουθεί ο αλγόριθμος είναι τα παρακάτω:

1) Βρίσκει την ανεξάρτητη μεταβλητή η οποία αν χρησιμοποιηθεί ως κριτήριο διαχωρισμού των δεδομένων εκπαίδευσης θα οδηγήσει σε κόμβους κατά το δυνατό διαφορετικούς σε σχέση με την εξαρτημένη μμεταβλητή

2) Κάνει το διαχωρισμό

3) Επαναλαμβάνει τη διαδικασία για κάθε έναν από τους κόμβους που προέκυψαν μέχρι να μην είναι δυνατός περαιτέρω διαχωρισμός.

Ένας από τους πιο διαδεδομένους μηχανισμούς διαχωρισμού είναι αυτός της εντροπίας της πληροφορίας (information entropy) ο οποίος επιλέγει εκείνη την ανεξάρτητη μεταβλητή που οδηγεί σε περισσότερο συμπαγές δένδρο. Η τιμή της εντροπίας της πληροφορίας δίνεται από τη σχέση:

$$E(S) = -p_+.\log_2(p_+) - p_-.\log_2(p\_)$$

όπου S είναι το σύνολο των δεδομένων εκπαίδευσης στο στάδιο (κόμβο) του διαχωρισμού, p+ είναι το κλάσμα των θετικών παραδειγμάτων του S και p- είναι το κλάσμα των αρνητικών παραδειγμάτων του S. (Quinlan, 1996)

Γενικότερα, για c διαφορετικές κατηγορίες, η εντροπία ορίζεται από τη σχέση:

$$E(S) = -\sum_{i=1}^{c} p_i.\log_2\ p(i)$$

όπου pi το ποσοστό των παραδειγμάτων του S που ανήκουν στην κατηγορία i.

Η εντροπία της πληροφορίας μετρά ουσιαστικά την ανομοιογένεια που υπάρχει στο S αναφορικά με την υπό εξέταση εξαρτημένη μεταβλητή και έχει τις ρίζες της θεωρίας των πληροφοριών (information theory). Στην περίπτωση που έχουμε δυο κατηγορίες, η τιμή της είναι 0 αν όλα τα μέλη του S ανήκουν στην ίδια κατηγορία και 1 αν τα μισά μέλη ανήκουν στην μια και τα άλλα μισά στην άλλη κατηγορία. Σε όλους δε τους υπολογισμούς, θεωρούμε την ποσότητα 0-log2(0) ίση με μηδέν.

Στην πράξη, χρησιμοποιείται το κέρδος πληροφορίας (information gain), Gain(S, A) ή G(S,A) που αναπαριστά τη μείωση της εντροπίας του συνόλου εκπαίδευσης S αν επιλεγεί ως παράμετρος διαχωρισμού η μεταβλητή A. Όταν μειώνεται η πληροφοριακή εντροπία, αυξάνεται η πυκνότητα πληροφορίας και άρα η περιγραφή γίνεται περισσότερο συμπαγής. Το κέρδος πληροφορίας δίνεται από τη σχέση:

$$G(S,A) = E(S) - \sum_{u \in Values(A)} \frac{|Su|}{S}.E(Su)$$

Όπου E(S) είναι η εντροπία πληροφορίας του υπό εξέταση κόμβου, A είναι η ανεξάρτητη μεταβλητή, με τιμές Values(A), βάσει της οποίας επιχειρείται ο επόμενος διαχωρισμός, u είναι μία από τις δυνατές τιμές του A, Su είναι το πλήθος των εγγραφών με A=u και E(Su) η εντροπία πληροφορίας του υπό εξέταση κόμβου ως προς την τιμή A=u. Ουσιαστικά, ο δεύτερος όρος είναι η εντροπία των παραδειγμάτων μετά το διαχωρισμό τους σύμφωνα με την τιμή του χαρακτηριστικού A και αποτελείται από το άθροισα της εντροπίας για το κάθε σύνολο που προκύπτει μετά το διαχωρισμό. (Kotsiantis, 2007)

Όπως αναφέραμε και προηγουμένως ο C4.5 είναι μια εκλέπτυνση του ID3. Η μέθοδος του C4.5 είναι βασισμένη στην διαδοχή Bernouli και έχει να κάνει με την εύρεση διαστήματος εμπιστοσύνης από τα δεδομένα εκπαίδευσης και ευρετική επιλογή ορίου για κλάδεμά. Η εκτίμηση σφάλματος του υποδέντρου αποτελεί σταθμισμένο άθροισα των εκτιμήσεων σφάλματος όλων των φύλλων του. Η εκτίμηση σφάλματος κόμβου υπολογίζεται από την συνάρτηση:

$$e = \left(f + \frac{z^2}{2N} + z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N}} \right) / \left(1 + \frac{z^2}{N} \right)$$

όπου f είναι το σφάλμα στα δεδομένα εκπαίδευσης ,και N ο αριθμός των υποδειγμάτων που καλύπτονται από το φύλλο.

4.5 Bayes

Η Μάθηση κατά Μπαιεζ (Bayes) αποτελεί μια άλλη ιδιαίτερα δημοφιλή προσέγγιση για την επαγωγική κατασκευή ταξινομητών, αφενός διότι εκπορεύεται από τον χώρο των πιθανοτήτων και αφετέρου διότι έχει επιδείξει σημαντικά αποτελέσματα σε ένα ευρύτατο φάσμα εφαρμογών. Η λειτουργία αυτής της κατηγορίας αλγορίθμων στηρίζεται στην υπόθεση ότι η υπό εκμάθηση έννοια σχετίζεται άμεσα με την κατανομή των πιθανοτήτων που παρουσιάζουν τα στιγμιότυπα του προβλήματος αναφορικά με την κλάση στην οποία ανήκουν. Συνεπώς, υπάρχει μια τελείως διαφορετική αντιμετώπιση του χώρου υποθέσεων. Δεν κατασκευάζεται ένας ταξινομητής ο οποίος βελτιώνεται σύμφωνα με τις επιδόσεις του στο σύνολο υποθέσεων, αλλά αναζητείται η υπόθεση με την μεγαλύτερη πιθανότητα να ταξινομεί σωστά τα στιγμιότυπα του συνόλου εκπαίδευσης. Εν συνεχεία δίνονται κάποιες από τις βασικότερες έννοιες στο χώρο της μηχανικής μάθησης που βασίζονται στην στατιστική. Το σημαντικότερο από αυτά είναι το θεώρημα του Μπαιεζ.

Έστω μια υπόθεση h ενός χώρου υποθέσεων H και D το σύνολο δεδομένων που χρησιμοποιείται στην εκπαίδευση. Η πιθανότητα η υπόθεση h να ταξινομεί σωστά (ή τουλάχιστον με την επιζητούμενη ακρίβεια) τα στιγμιότυπα συμβολίζεται με P(h) ενώ η πιθανότητα η υπόθεση να ταξινομεί σωστά τα στιγμιότυπα του D συμβολίζεται με P(h|D) και καλείται δεσμευμένη πιθανότητα. Ο Bayes διατύπωσε

το θεμελιώδες για την στατιστική θεώρημα που είναι γνωστό με το όνομα του και συνοψίζεται στον τύπο:

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

Η εξίσωση αυτή είναι γνωστή ως κανόνας του Bayes (καθώς και ως νόμος του Bayes ή θεώρημα του Bayes). Αυτή η απλή εξίσωση αποτελεί το θεμέλιο για όλα τα μοντέρνα συστήματα για πιθανοτικό συμπερασμό.

Παρατηρείται ότι η πιθανότητα της υπόθεσης που μελετάται να ταξινομεί σωστά το σύνολο εκπαίδευσης αυξάνει όταν αυξάνεται η πιθανότητα να ταξινομεί σωστά όλα τα πιθανά στιγμιότυπα. Στη συνέχεια θα δοθούν κάποιοι ορισμοί εννοιών που συναντώνται συχνά στις μεθόδους μηχανικής μάθησης που έχουν ως βάση το θεώρημα του Μπέυζ. Μια υπόθεση λέγεται MAP (μέγιστη a posteriori) και συμβολίζεται h_{MAP} αν και μόνο αν

$$h_{MAP} = \arg_{h \in H} \max P(h \mid D) = \arg_{h \in H} \max \frac{P(D \mid h)P(h)}{P(D)} = \arg_{h \in H} \max P(D \mid h)P(h)$$

Το P(D) παραλήφθηκε καθώς είναι σταθερά ως προς τις υποθέσεις. Επίσης μερικές φορές δεν έχουμε γνώση για τις υποθέσεις h. Τότε μπορούμε να θεωρήσουμε πως και ο όρος P(h) είναι σταθερός για όλες τις υποθέσεις και να τον απαλείψουμε από τον τύπο .Έτσι προκύπτει η μέγιστη πιθανοφάνεια (maximum likelihood).

$$h_{ML} = \arg_{h \in H} \max P(D \mid h)$$

Προηγουμένως δόθηκε απάντηση στο ερώτημα της πιο πιθανής υπόθεσης. Πώς όμως κρίνουμε την καταλληλότερη ταξινόμηση. Μια υπόθεση πρέπει να καλύπτει το σύνολο των στιγμιότυπων όμως κατά την αναζήτηση της κατάλληλης υπόθεσης είναι πολύ πιο απλό να κατασκευαστεί μια υπόθεση που να μπορεί να ταξινομεί τα περισσότερα (και όχι όλα) στιγμιότυπα σωστά. Κατά την ταξινόμηση ενός στιγμιότυπου, προτεραιότητα έχει η ορθή ταξινόμηση του συγκεκριμένου στιγμιότυπου, χωρίς να έχει κάποια σημασία η επίδοση του ταξινομητή στο σύνολο των στιγμιότυπων. Επιφανειακά ο κανόνας του Bayes δεν φαίνεται και πολύ χρήσιμος. Απαιτεί τρείς όρους- μία υπο συνθήκη πιθανότητα και δύο χωρίς συνθήκη πιθανότητες- απλώς και μόνο για τον υπολογισμό μιας υπο συνθήκη πιθανότητας.

Ωστόσο ο κανόνας του Bayes είναι χρήσιμος στην πράξη επειδή υπάρχουν τόσες πολλές περιπτώσεις όπου έχουμε καλές πιθανοτικές εκτιμήσεις για αυτές τις τρείς τιμές και χρειάζεται να υπολογίσουμε την τέταρτη τιμή.

Σε μία ιατρική διάγνωση έχουμε συχνά υπό συνθήκη πιθανότητες για τις αιτιολογικές συσχετίσεις και θέλουμε να παράγουμε μια διάγνωση. Για παράδειγμα ένας γιατρός γνωρίζει ότι η μηνιγγίτιδα κάνει το 50% των φορών τον ασθενή να έχει δύσκαμπτο λαιμό. Ο γιατρός επίσης γνωρίζει κάποια γεγονότα χωρίς συνθήκη: η εκ των προτέρων πιθανότητα να έχει ένας ασθενής μηνιγγίτιδα είναι 1/50.000 και η εκ των προτέρων πιθανότητα να έχει ένας ασθενής δύσκαμπτο λαιμό είναι 1/20. Αν λοιπόν S είναι η πρόταση να έχει ο ασθενής δύσκαμπτο λαιμό και m είναι η πρόταση να έχει μηνιγγίτιδα, τότε έχουμε:

$P(s|m)=0.5$

$P(m)=1/50.000$

$P(s)=1/20$

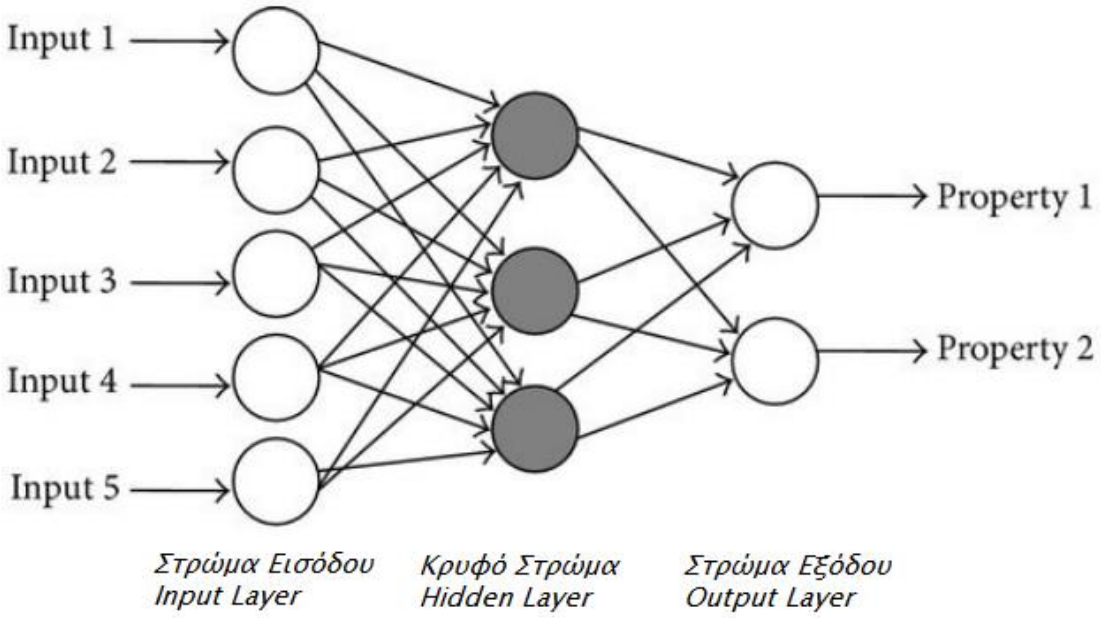
Συνεπώς : $P(m|s)= P(s|m) *P(m) / P(s) = (0.5 *1/50.000)/(1/20)=0.0002$

Άρα αυτό σημαίνει ότι αναμένουμε μόνο 1 στους 5000 ασθενείς με δύσκαμπτο λαιμό να έχει μηνιγγίτιδα. Παρ όλο λοιπόν που ο δύσκαμπτος λαιμός είναι ένα αρκετά συνηθισμένο σύμπτωμα μηνιγγίτιδας (με πιθανότητα όπως προαναφέραμε 0.5), η πιθανότητα της μηνιγγίτιδας στον ασθενή παραμένει χαμηλή, καθώς η εκ των προτέρων πιθανότητα του δύσκαμπτου λαιμού είναι πολύ υψηλότερη από εκείνη της μηνιγγίτιδας. (Russell, 2004)

4.6 Τεχνική πολλαπλών στρωμάτων (Multilayer Perceptron)

Η τεχνική πολλαπλών στρωμάτων (MLP) είναι ένα Νευρωνικό Δίκτυο που έχει πολύ καλές επιδόσεις και επιλύει προβλήματα που ένα απλό Perceptron δεν μπορεί.Δεν χρησιμοποιεί ως συνάρτηση ενεργοποίησης την βηματική (0/1 ή -1/1) αλλά συνήθως την Σιγμοειδή. Αυτό γίνεται γιατί, καθώς κατά την εκμάθηση, χρησιμοποιούνται αλγόριθμοι βελτιστοποίησης που βασίζονται στην κατάβαση δυναμικού (δηλαδή στην παραγώγιση), η βηματική δεν είναι παραγωγίσιμη. Η Σιγμοειδής και είναι πολύ κοντά στην βηματική και είναι παραγωγίσιμη. Η γενική τοπολογία ενός MLP δικτύου είναι η εξής:

Όπως φαίνεται και από την ακόλουθη εικόνα έχουμε ένα αρχικό στρώμα, αποκαλούμενο ως στρώμα εισόδου, ακολουθεί το κρυφό στρώμα που δεν είναι απαραίτητο να υφίσταται ενώ μπορεί να υπάρχουν και πολλαπλά κρυφά στρώματα. Τέλος έχουμε το τελικό στρώμα χαρακτιρισμένο ως στρώμα εξόδου στην τοπολογία του MLP.



Εικόνα 4.9 : Τοπολογία ενός MLP δικτύου.

Η λειτουργία ενός δικτύου πολλαπλών στρωμάτων έχει ως εξής. Αρχικά το επίπεδο με τους κόμβους εισόδου τροφοδοτεί το πρώτο κρυφό στρώμα με το διάνυσμα εισόδου, ενεργοποιώντας το. Έπειτα πραγματοποιούνται οι υπολογισμοί στους νευρώνες του πρώτου κρυφού στρώματος και παράγονται σήματα εξόδου. Αυτά τα σήματα εξόδου χρησιμοποιούνται ως είσοδοι στο δεύτερο κρυφό στρώμα νευρώνων που με τη σειρά του θα τροφοδοτήσει το επόμενο στρώμα. Η διαδικασία συνεχίζεται με τον ίδιο τρόπο μέχρι το σήμα να φτάσει στο στρώμα εξόδου, από το οποίο εξάγεται η απόκριση του δικτύου στο σήμα εισόδου που του δόθηκε μέσω των κόμβων στο επίπεδο εισόδου.

Η εκπαίδευση ενός δικτύου MLP έχει ιδιαίτερο ενδιαφέρον λόγω της ικανότητας του MLP να συμπεριφέρεται ως «Καθολικός Προσεγγιστής» (Universal Approximator). Αποδεικνύεται πως εάν έχουμε το κατάλληλο μέγεθος δικτύου, τότε μπορούμε να το εκπαιδεύσουμε να μάθει όποια συνάρτηση θέλουμε και με οποιαδήποτε ακρίβεια θέλουμε. Αυτό αιτιολογεί και την μεγάλη δημοτικότητα των

αλγορίθμων εκπαίδευσης MLP. Ο πιο γνωστός αλγόριθμος εκπαίδευσης είναι ο Back-Propagation.

Η εκπαίδευση είναι μία πολύ σημαντική διαδικασία για τα νευρωνικά δίκτυα, καθώς δηλώνει την ικανότητα του δικτύου να μαθαίνει από το περιβάλλον του και του δίνει τη δυνατότητα σταδιακά να βελτιώσει την απόδοσή του. Η διαδικασία της εκπαίδευσης σχετίζεται άμεσα με ένα μετρούμενο μέγεθος του δικτύου το οποίο μεταβάλλεται με το χρόνο και το οποίο επηρεάζει τις μεταβολές που πραγματοποιούνται στα συναπτικά βάρη καθώς και στις πολώσεις του δικτύου. Ιδανικά η απόδοση του δικτύου θα βελτιώνεται και εκείνο θα αποκτά μεγαλύτερη γνώση του περιβάλλοντος μετά από κάθε επανάληψη στην διαδικασία εκμάθησης. Ένας ορισμός της εκπαίδευσης δίνεται από το Simon Haykin στο βιβλίο του (S, 1999), ο οποίος περιγράφει την εκπαίδευση ως « μια διαδικασία με την οποία οι ελεύθερες παράμετροι του δικτύου μεταβάλλονται μέσω μιας διαδικασίας διέγερσης από το περιβάλλον στο οποίο το δίκτυο είναι ενσωματωμένο. Το είδος της εκμάθησης καθορίζεται από τον τρόπο με τον οποίο γίνεται η μεταβολή των παραμέτρων.»

Για την πραγματοποίηση της εκπαίδευσης απαραίτητος είναι ένας αλγόριθμος εκπαίδευσης ο οποίος αποτελείται από ένα σετ οδηγιών που υλοποιούν τον τρόπο με τον οποίο θα εκπαιδευτεί το δίκτυο. Υπάρχει ποικιλία τέτοιων αλγορίθμων οι οποίοι διαφέρουν ο ένας από τον άλλο στον τρόπο με τον οποίο πραγματοποιούνται οι προσαρμογές των συναπτικών βαρών του δικτύου. Επίσης επισημαίνεται ότι οι περισσότεροι από τους αλγορίθμους εκπαίδευσης που χρησιμοποιούνται σήμερα έχουν καλή απόδοση σε πληθώρα προβλημάτων. Παρόλα αυτά σε διάφορα προβλήματα υπάρχουν κάποιοι αλγόριθμοι που παρουσιάζουν καλύτερα αποτελέσματα. Αυτός είναι και ο λόγος για τον οποίο δεν μπορεί να προταθεί ένας μοναδικός αλγόριθμος εκπαίδευσης ο οποίος θα έχει καθολική ισχύ.

Συνοψίζοντας τα βασικά χαρακτηριστικά των MLP είναι:

- Οι συναρτήσεις των νευρώνων τους είναι μη γραμμικές και κυρίως συνεχείς και παραγωγίσιμες, σε αντίθεση με το αρχικά προτεινόμενο Perceptron. Συνήθως χρησιμοποιούνται σιγμοειδείς συναρτήσεις όπως η λογιστική και η συνάρτηση tanh(x).
- Το δίκτυο αποτελείται από ένα ή περισσότερα κρυμμένα στρώματα, τα οποία του προσδίδουν χαρακτηριστική ευελιξία καθώς έχουν την ικανότητα να αποσπούν σταδιακά, σημαντικές πληροφορίες για τις ιδιότητες της εισόδου.
- Παρουσιάζουν μεγάλο βαθμό συνεκτικότητας εξαιτίας των συνάψεων.
- Χαρακτηρίζονται από αιτιατές και χωρίς ανάδραση εξισώσεις. Η έξοδός τους δηλαδή, είναι συνάρτηση αποκλειστικά και μόνο της παρούσας εισόδου.

5ο Κεφάλαιο - Εισαγωγή στο Weka & Επεξεργασία Δεδομένων

Προκειμένου να αναλύσουμε και να επεξεργαστούμε τα δεδομένα μας με στόχο την εύρεση του καλύτερα αξιόπιστου αλγορίθμου ο οποίος θα οδηγήσει στην ορθή ταξινόμηση θα χρησιμοποιηθεί το λογισμικό weka. Το συγκεκριμένο λογισμικό δημιουργήθηκε από το πανεπιστήμιο του Waikato στην Νέα Ζηλανδία και μπορεί κανείς να το κατεβάσει δωρεάν καθώς αποτελεί open source και ελεύθερης διανομής λογισμικό από τον επίσημο ιστότοπο του πανεπιστημίου. Η πλήρη ονομασία του είναι Waikato Environment for Knowledge Analysis και αποτελεί μια συλλογή από τους αλγορίθμους μηχανικής μάθησης. Μερικές από τις δυνατότητες που παρέχει το συγκεκριμένο λογισμικό είναι οι ακόλουθες:

-Προεπεξεργασία Δεδομένων

-Ταξινόμηση

-Ομαδοποίηση

-Εύρεση κανόνων συσχέτισης

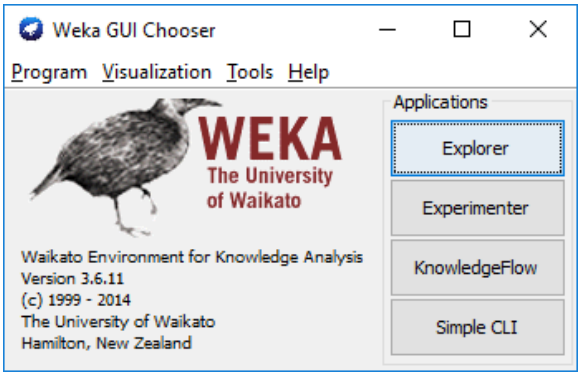
-Παλινδρόμηση

5.1 Περιβάλλον λογισμικού Weka

Το πρόγραμμα weka είναι όπως προαναφέραμε λογισμικό ανοιχτού κώδικα το οποίο δημιουργήθηκε το 1992 αποτελώντας το πιο αναγνωρισμένο σύστημα για μεθόδους εξόρυξης δεδομένων και μηχανικής μάθησης. Κατά κύριο λόγο χρησιμοποιούνται αρχεία μορφής .arff (όπως το dataset που χρησιμοποιήθηκε στην παρούσα εργασία) καθώς και αρχεία μορφής .csv.

Όπως διαφαίνεται και στο ακόλουθο dataset που επακολούθησε η επεξεργασία διαφαίνονται οι εγγραφές, οι τιμές ενώ σχηματίζεται ένα ιστόγραμμα για την απεικόνιση αυτών. Το weka διαθέτει τεχνικές κατηγοριοποίησης (classification), συσταδοποίησης (clustering), παλινδρόμησης(regression), καθώς και κανόνων συσχέτισης (association rules).

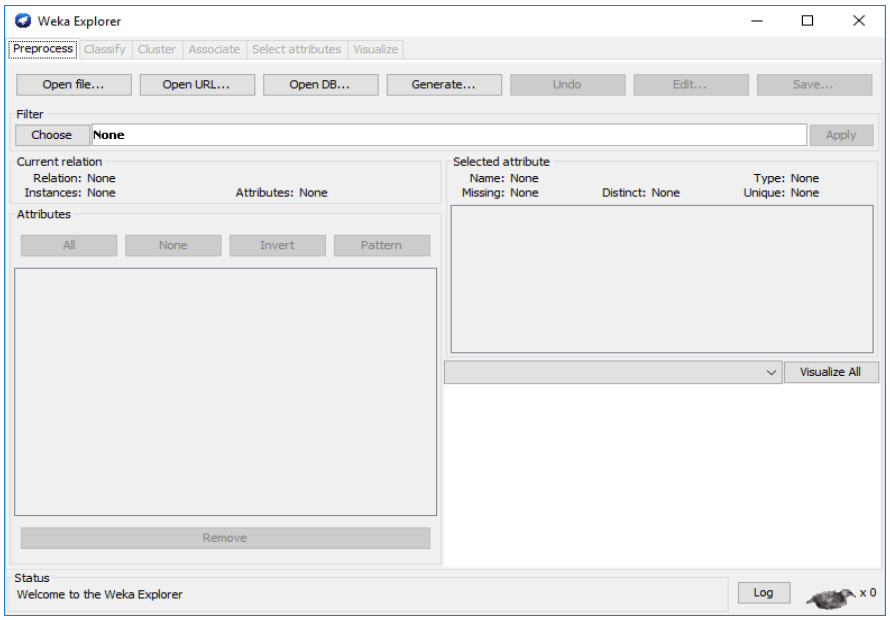
Μόλις εγκαταστήσουμε το λογισμικό μας ανοίγει το ακόλουθο μενού δίνοντας τις εξής επιλογές στο χρήστη: Explorer, Experimenter, KnowledgeFlow, Simple Cli. Όπως μπορεί να διακρίνει κανείς στην ακόλουθη εικόνα. (G. Holmes, 1994)



Εικόνα 5.1 : Application Λογισμικού weka

Στην παρούσα εργασία θα χρησιμοποιηθεί το γραφικό περιβάλλον “explorer”, το οποίο χρησιμοποιείται κυρίως για την επεξεργασία δεδομένων, τα οποία δεν έχουν υποστεί άλλη επεξεργασία. Στο περιβάλλον “experimenter” υλοποιούνται κυρίως εργασίες που αφορούν την στατιστική χρήση, ενώ στο περιβάλλον “knowledgeflow” μπορεί να αντλήσει κανείς γνώσεις από προηγούμενες εργασίες που είχαν υλοποιηθεί, με παρόμοια μορφή. Τέλος το στο γραφικό “simple CLI” δίνει την δυνατότητα εξαγωγής αποτελεσμάτων μέσα από την γραμμή εντολών.

Επιλέγοντας συνεπώς το πεδίο με όνομα Explorer εμφανίζεται το παρακάτω παράθυρο του περιβάλλοντος.



Εικόνα 5.2 : Γραφικό περιβάλλον “explorer”

Σε αυτό το παράθυρο ο χρήστης έχεις τις επιλογές:

- ***Open file:*** όπου ο χρήστης έχει τη δυνατότητα να εισάγει δεδομένα από αρχείο (τύπου .arff)

- ***Open Url:*** όπου ο χρήστης έχει τη δυνατότητα να εισάγει δεδομένα από διαδικτυακό ιστότοπο

- ***Open DB***: όπου ο χρήστης έχει τη δυνατότητα να εισάγει δεδομένα από βάση δεδομένων εγκατεστημένο όμως στο συγκεκριμένο υπολογιστή.

- ***Generate:*** όπου το Weka παράγει τυχαία δεδομένα (χρησιμοποιείται σε στάδιο πειραματισμού).

Το weka όπως αναφέρθηκε, έχει μια ειδική μορφή αρχείου με κατάληξη .arff . Για την κατασκευή ενός αρχείο με κατάληξη .arff στο επίπεδο που είναι απαραίτητο για την παρούσα εργασία, ο χρήστης θα πρέπει να ανοίξει ένα απλό αρχείο κειμένου (.txt) και να το αποθηκεύσει ως αρχείου τύπου με κατάληξη .arff κάνοντας τις ανάλογες τροποποιήσεις όπως απαιτείται για την ορθή αναγνώριση των μεταβλητών, όπως διαφαίνεται παρακάτω. Ουσιαστικά θα πρέπει με τις τροποποιήσεις αυτές να ορίσουμε τι είναι το κάθε στοιχείο (είδος) καθώς και σε ποια ομάδα ανήκει (area , perimeter κλπ)

Για παράδειγμα:

@cancer dataset

@attribute Radius numeric

@attribute Perimeter numeric

@attribute Smoothness numeric

@attribute concavity numeric

@data

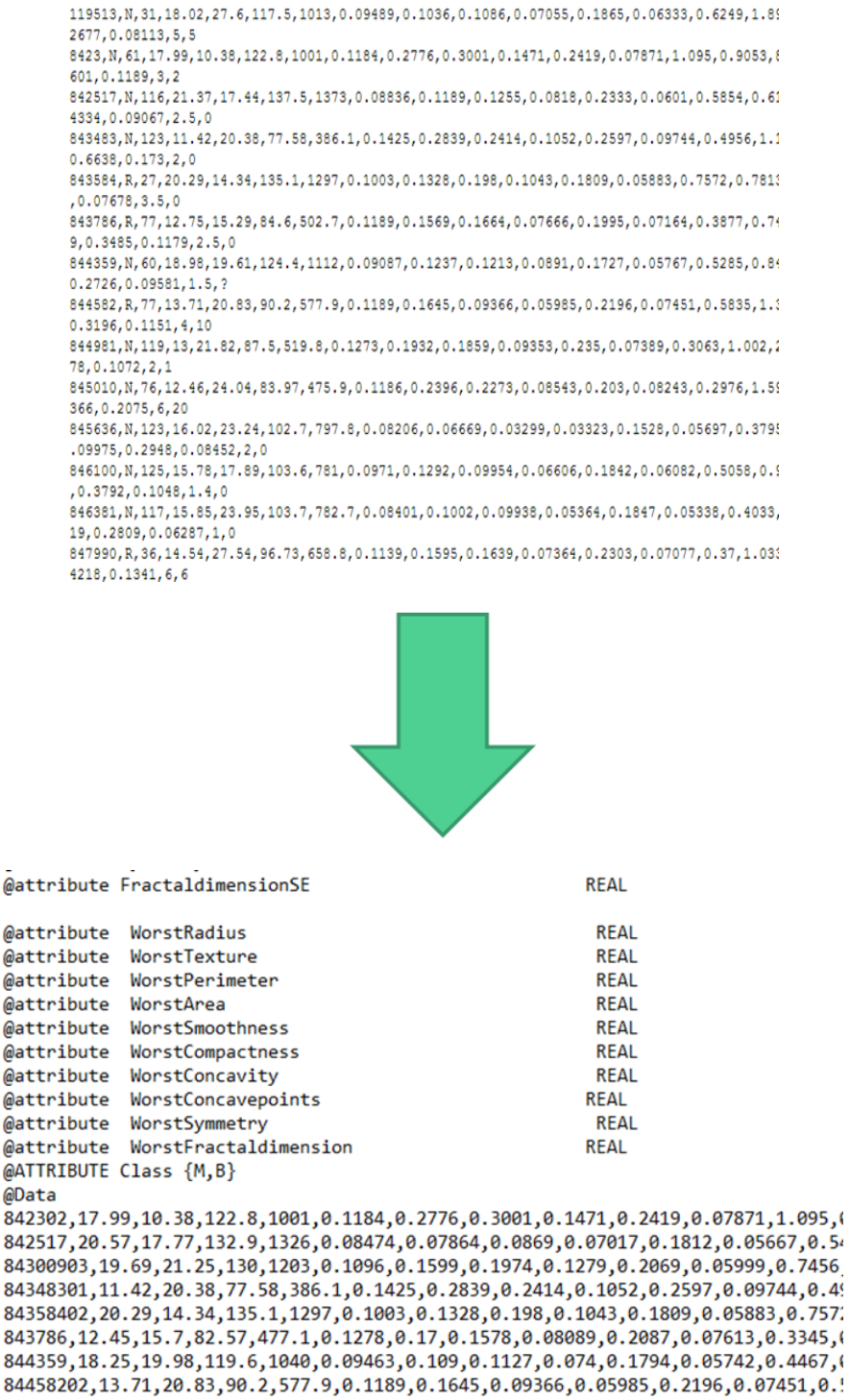
Radius, perimeter, smoothness, concavity

Συνεπώς σύμφωνα με το παραπάνω παράδειγμα έχουμε τα εξής κατηγορήματα:

-Το κατηγόρημα @relation: είναι μια δεσμευμένη εντολή η οποία ονομάζει το σύνολο των δεδομένων που εισάγονται.

-Το κατηγόρημα @attribute : περιγράφει ένα χαρακτηριστικό και τον τύπου του.

- Το κατηγόρημα @data: καθορίζει στο αρχείο ότι από εκείνο το σημείο και μετά ακολουθούν αυστηρά μόνο δεδομένα. Τα δεδομένα της ίδια γραμμής και διαφορετικών χαρακτηριστικών χωρίζονται μεταξύ τους με κόμμα.



Εικόνα 5.3 : Απεικόνιση μετατροπής δεδομένων σε αρχείο μορφής .arff , επεξεργάσιμο από το λογισμικό weka.

Αφότου λοιπόν εισαχθούν τα δεδομένα στην κατάλληλη μορφή υλοποιούμε στην πράξη πλέον του προαναφερθέντες αλγορίθμους προκειμένου να εξάγουμε αποτελέσματα και να εκπαιδεύσουμε ορθά το σύστημα μας.

5.2 Επεξεργασία δεδομένων με την χρήση weka

Στην συγκεκριμένη εργασία αναλύουμε δεδομένα που αφορούν τον καρκίνο του πνεύμονα. Τα δεδομένα πάρθηκαν από το (UCI) και αναλύονται με την χρήση του weka. Χρησιμοποιούνται διάφορες μέθοδοι clussify όπως Ibk (Knn), Smo (Svm) Multilayer Perceptron, και εξάγονται αποτελέσματα προκειμένου να πραγματοποιηθεί σύγκριση μεταξύ αυτών.

Το συγκεκριμένο λοιπόν data set περιλάμβανε 32 attributes συνολικά εκ των οποίων τα 30 περιείχαν τιμές για διάφορα δεδομένα σχετικά με κάποια χαρακτηριστικά της συγκεκριμένης ασθένειας (π.χ χαρακτηριστικά των κυττάρων), ενώ ένα όριζε σε ποιά κλάση ανήκει το άτομο (υγιής ή ασθενής) και άλλο ένα το id. Πιο συγκεκριμένα τα δεδομένα αφορούν τον καρκίνο του πνεύμονα και έχουν συλλεγεί από τον Dr. William H. Wolberg (General Surgery Dept, University of Wisconsin).

Περιλαμβάνουν 10 χαρακτηριστικά γνωρίσματα για κάθε πυρήνα κυττάρου τα οποία είναι :

a) Radius

b) Texture

c) Perimeter

d) Area

e) Smoothness

f) Compactness

g) concavity

h) Concave points

i) Symmetry

j) Fractal dimension

Αυτά τα χαρακτηριστικά διαφοροποιούνται σε κατηγορίες standard error , worst , largest υπολογίστηκαν για κάθε μια εικόνα και τελικά προέκυψαν τα 32 attributes που χρησιμοποιήθηκαν. Αναλυτικότερα με βάση αυτά τα χαρακτηριστικά των κυττάρων το άτομο κατατάσσεται σε ασθενή ή υγιές άτομο, συνεπώς έχουμε δύο ομάδες κατηγοριοποίησης με βάση τα χαρακτηριστικά αυτά.

Αφού λοιπόν κατεβάσουμε το συγκεκριμένο dataset πραγματοποιείται μετατροπή σε μορφή κατάλληλη προς επεξεργασία από το weka, arff. Μετά λοιπόν από την μετατροπή του αρχείου σε κατάλληλη μορφή τρέχοντας τις διάφορες μεθόδους συγκρίνονται τα αποτελέσματα και το ποσοστό επιτυχίας του κάθε αλγορίθμου. Ουσιαστικά αναζητείται ο καλύτερος αλγόριθμος οποίος αφού εκπαιδευτεί ορθά το σύστημα μας θα μας επιφέρει το καλύτερο ποσοστό επιτυχίας στην κατηγοριοποίηση ασθενή ή υγιές άτομο με βάση τα συγκεκριμένα χαρακτηριστικά.

Ύστερα από την συγκριτική ανασκόπηση των αποτελεσμάτων των μεθόδων εξάγονται κάποια συμπεράσματα προκειμένου να βρεθεί ποιά μέθοδος αποδίδει καλύτερα αποτελέσματα για το συγκεκριμένο data set, καθώς και πραγματοποιείται σύγκριση του συγκεκριμένου dataset σε άλλες σχετικές έρευνες που έχουν γίνει χρησιμοποιώντας άλλους αλγορίθμους. Αναλυτικότερα έχοντας τα συγκεκριμένα δεδομένα που αφορούν τον καρκίνο του πνεύμονα γίνεται προσπάθεια να αναζητήσουμε την καλύτερη μέθοδο ταξινόμησης, συγκρίνοντας τις μεθόδους προκειμένου να διαπιστώσουμε την καλύτερη μέθοδο και αλλάζοντας το μεγαλύτερο σύνολο των παραμέτρων όπως για παράδειγμα το κ στον knn ή τα folds κλπ. Με αυτό λοιπόν τον τρόπο εντοπίζεται η καλύτερη μέθοδος που θα καταφέρει να ταξινομήσει καλύτερα άλλα δεδομένα κατηγοριοποιώντας σε ασθενή ή μη ασθενή με βάση τα χαρακτηριστικά που θα έχουν εισαχθεί. Παράλληλα αρκετά papers χρησιμοποιούν άλλες ή ίδιες (με άλλες παραμέτρους) μεθόδους, γίνεται συνεπώς προσπάθεια σύγκρισης αυτών των μεθόδων για τον εντοπισμό των αποτελεσμάτων της καλύτερης ταξινόμησης για τα συγκεκριμένα δεδομένα.

Data Set Characteristics:	Multivariate	**Number of Instances:**	569
Attribute Characteristics:	Real	**Number of Attributes:**	32
Associated Tasks:	Classification	**Missing Values?**	No

Εικόνα 5.4 : Χαρακτηριστικά του dataset που χρησιμοποιήθηκε

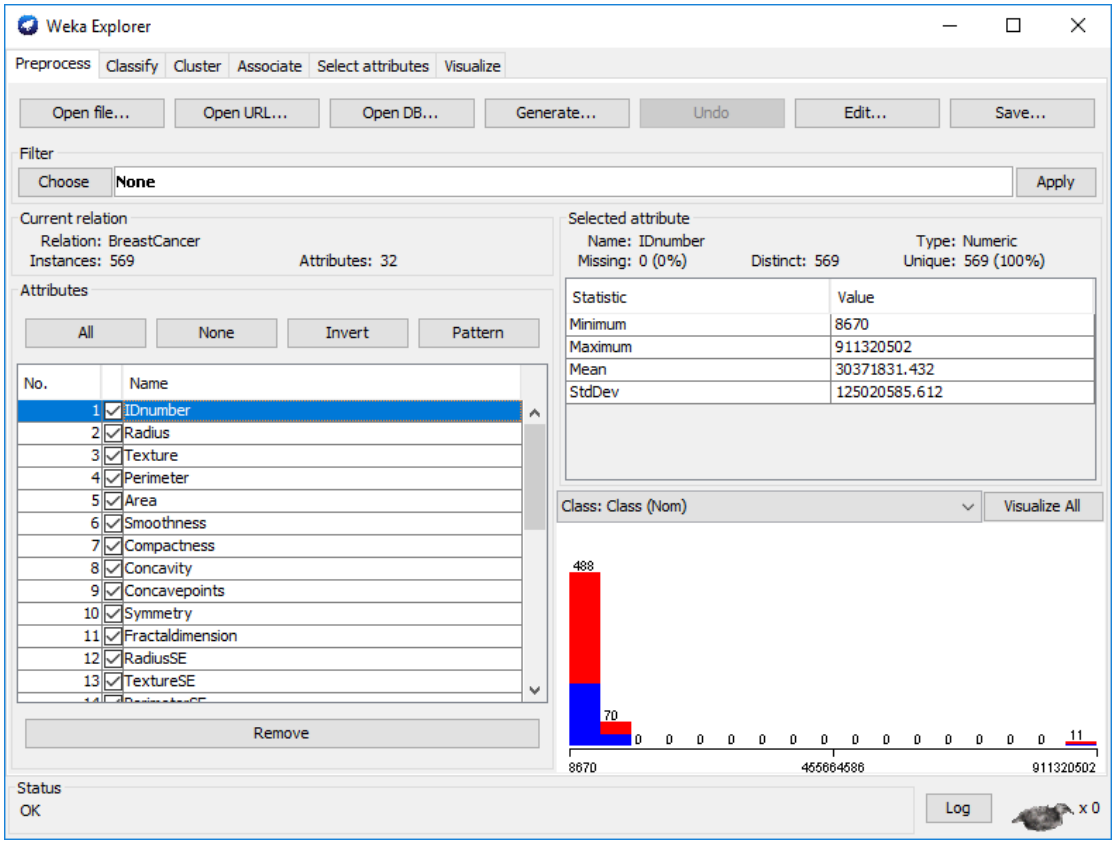
5.2.1 Μεθοδολογία

Στην συγκεκριμένη υλοποίηση που πραγματοποιήθηκε δεν χρησιμοποιείται κάποιο test set για την υλοποίηση σε άγνωστα δεδομένα και την μετέπειτα κατηγοριοποίηση τους σε κλάσεις ,αλλά χρησιμοποιείται το συγκεκριμένο training set.

Μέσω του cross validation αξιολογείται η ταξινόμηση μας με τον κατά δυνατότερο αμερόληπτο τρόπο. Σύμφωνα με το cross validation χωρίζονται τα δεδομένα μας σε κομμάτια ανάλογα με τις διπλες (folds) που έχουν οριστεί πρωτύτερα. Αν για παράδειγμα έχουμε ορίσει folds=10 (το οποίο προτείνεται από την βιβλιογραφία (Geoffrey, 2004)) , τότε παίρνουμε τα 9 για training και το 1 για test , επαναλαμβάνοντας την ίδια διαδικασία για όλες τις διαφορετικές περιπτώσεις training και test (μέχρι όλα τα folds να έχουν χρησιμοποιηθεί έστω και μία φορά για test).

Ένας άλλος τρόπος αξιολόγησης είναι το percentage split όπου μέσω αυτού κατά κερματίζουμε τα δεδομένα σε training set και test ανάλογα με το ποσοστό που έχουμε ορίσει. Για παράδειγμα όταν οριστεί 66% percentage split θα έχουμε το 66% των δεδομένων που εισάγαμε ως training set , και το υπόλοιπο 34% ως test.

Ως έξοδο επιλέγεται ο πίνακας ταξινόμησης (confusion matrix) , το ποσοστό ακρίβειας και λάθους καθώς και διάφορα σφάλματα. Συγκεκριμένα πραγματοποιείται εστίαση κυρίως στον πίνακα ταξινόμησης καθώς και στο ποσοστό λάθους προκειμένου να δημιουργηθούν γραφικές και μέσω των οποίων τελικά θα γίνει η τελική σύγκριση των μεθόδων. Στο συγκεκριμένο πρόβλημα όπου στα δεδομένα υπάρχουν δύο κλάσεις μπορεί να χρησιμοποιηθεί το ποσοστό λάθους σε αντίθεση με την περίπτωση που υπήρχαν περισσότερες εκ των δύο κλάσεων.



Εικόνα 5.5 : Απεικόνιση των δεδομένων που έχουν εισαχθεί στο weka μετά από μετατροπή σε μορφή .Arff, πριν την την χρήση τεχνικών κατηγοριοποίησης

5.2.2 Τεχνικές Κατηγοριοποίησης στην παρούσα επεξεργασία

- *Τεχνική Κατηγοριοποίησης k-nn*

Η τεχνική των κοντινότερων γειτόνων (Nearest Neighbor (NN)) είναι μια απλή προσέγγιση του προβλήματος της κατηγοριοποίησης. Έτσι ένα νέο στοιχείο κατηγοριοποιείται χρησιμοποιώντας την πλειοψηφία μεταξύ των κατηγοριών από k παραδείγματα που είναι τα πιο κοντινά σε αυτό που δίνεται. Για να κατηγοριοποιηθεί ένα στοιχείο x προσδιορίζουμε τους k πλησιέστερους γείτονες. Για παράδειγμα για κ=3 εννοούμε 3 κοντινότερους γείτονες ενώ μπορεί να δοθεί βάρος στον πιο κοντινότερο γείτονα. Για την εφαρμογή του αλγορίθμου, χρειάζεται εκ των προτέρων να γνωρίζουμε τη τιμή του k και τη μετρική απόστασης. (Hall, 2008) Εισάγεται λοιπόν διαφορετικό αριθμό κ , διαφορετικά folds και διαφορετικό ποσοστό training data , και αποθηκεύονται τα αποτελέσματα της μεθόδου προκείμενου να συγκριθούν μετέπειτα. (Everitt, 2011)

- *Μηχανές Υποστήριξης Διανυσμάτων (SVM)*

Στην κατηγορία των function βρίσκεται η μέθοδος Smo (svm) που αφορά τα νευρωνικά δίκτυα. Τα νευρωνικά παράγουν μία συνάρτηση διαχωρισμού.

Συγκεκριμένα η SVM είναι μια μέθοδος μηχανικής μάθησης για δυαδικά προβλήματα ταξινόμησης. Ουσιαστικά αποτελεί την προβολή των σημείων εκπαίδευσης σε χώρο περισσοτέρων διαστάσεων και βρίσκουν το υπερεπίπεδο το οποίο διαχωρίζει βέλτιστα τα σημεία των δύο τάξεων.

Τα άγνωστα σημεία ταξινομούνται σύμφωνα με την πλευρά του υπερεπίπεδου στην οποία βρίσκονται, ενώ τα διανύσματα τα οποία ορίζουν το υπερεπίπεδο σε δύο τάξεις, χαρακτηρίζονται ως support vectors. (Nello, 2000)

- *Τεχνική πολλαπλών στρωμάτων, πρόσθιας διάδοσης (multilayer perceptron)*

Είναι ένα feed forward τεχνητό νευρικό μοντέλο δικτύων αποτελείται από τα πολλαπλάσια στρώματα των κόμβων σε μια κατευθυνόμενη γραφική παράσταση, με κάθε στρώμα που συνδέεται πλήρως με το επόμενο.

Ουσιαστικά αποτελέι ένα τεχνητό νευρωνικό δίκτυο όπου σε αυτό το δίκτυο η πληροφορία μετακινείται προς μια κατεύθυνση την πρόσθια , από τους κόμβους εισόδου προς τους ενδιάμεσους νευρώνες (αν υπάρχουν) και τελικά στους κόμβους εξόδου. (Frank, 1961) , (Rumelhart, 1986)

5.3 Περιγραφή αποτελεσματικότερης μεθόδου

Υλοποιώντας την τεχνική knn και συγκρίνοντας τα τελικά αποτελέσματα όπως φαίνονται παρακάτω τα βέλτιστα ποσοστά βγαίνουν στην περίπτωση που έχει οριστεί fold=9 καθώς και για k=11 όπου το ποσοστό της σωστής ταξινόμησης ανέρχεται στο 97,7153 %.

Αντίστοιχα για τις μηχανές υποστήριξης διανυσμάτων (Svm) η καλύτερη ταξινόμηση επέρχεται όταν οριστεί c=0.9 όπου η πιθανότητα σωστής ταξινόμησης ανέρχεται στο 98,2425%. Παράλληλα αξίζει να αναφερθεί πως στο ίδιο ποσοστό με το παραπάνω (98,2425%) οδηγεί η μέθοδος αυτή ορίζοντας ως ποσοστό για training 80% και το 20% test.

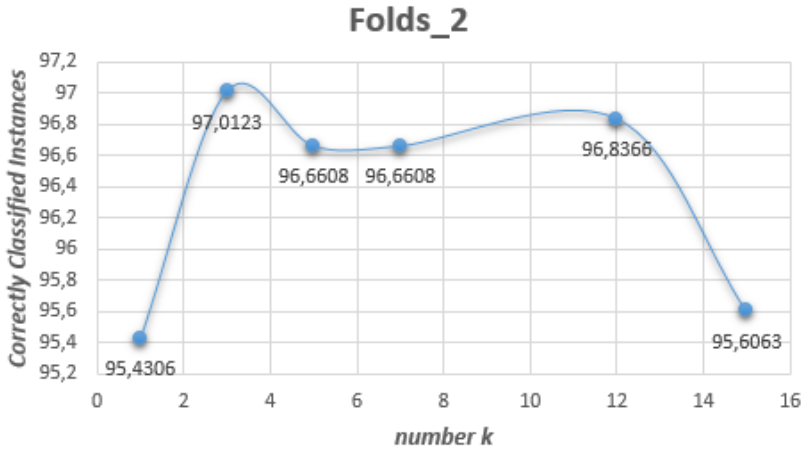
Όσο αναφορά την τεχνική πολλαπλών στρωμάτων συναντάται η βέλτιστη λύση για την τιμή folds=14 όπου το ποσοστό ανέρχεται στο 96,8366%.

Συνεπώς χρησιμοποιώντας τις παραπάνω μεθόδους οδηγούμαστε στην βέλτιστη λύση ταξινόμησης με την μικρότερη πιθανότητα λάθους μέσω των μηχανών υποστήριξης διανυσμάτων (svm).

5.3.1 Αποτελέσματα Μεθόδων Ταξινόμησης

- ***KNN***

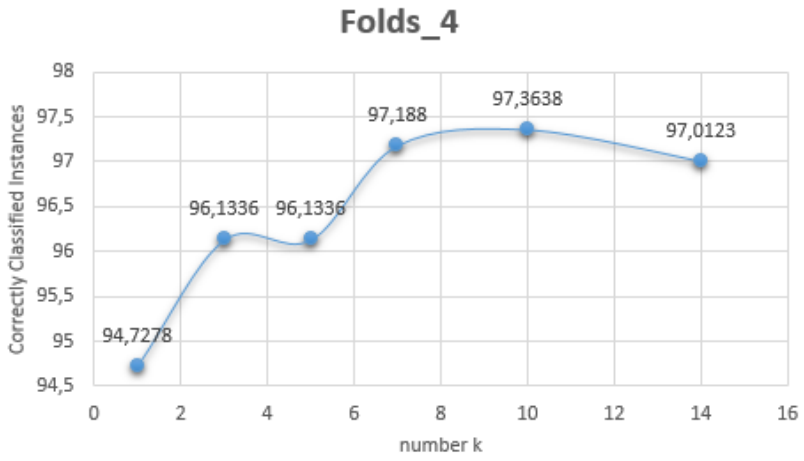
Στον knn αλγόριθμο επιλέγοντας το cross-validation, χωρίζονται τα δεδομένα μας σε δίπλες για να προκύψουν τα δεδομένα εκπαίδευσης και τα δεδομένα test. Αρχικά επιλέγεται η τιμή folds=2, ενώ μεταβάλλονται διάφορες τιμές του k προκειμένου να παρατηρηθούν τα καλύτερα αποτελέσματα. Πιο συγκεκριμένα στην παρακάτω γραφική φαίνεται για folds=2 η αλλαγή των k πως επηρεάζει το ποσοστό επιτυχίας. Όπως προαναφέραμε ένα νέο στοιχείο κατηγοριοποιείται χρησιμοποιώντας την πλειοψηφία μεταξύ των κατηγοριών από k παραδείγματα που είναι τα πιο κοντινά σε αυτό που δίνεται. Για παράδειγμα για κ=3 εννοούμε 3 κοντινότερους γείτονες ενώ μπορούμε να δώσουμε βάρος στον πιο κοντινότερο γείτονα.



Εικόνα 5.6 :Γραφική απεικόνιση αποτελεσμάτων αλγορίθμου KNN για folds= 2

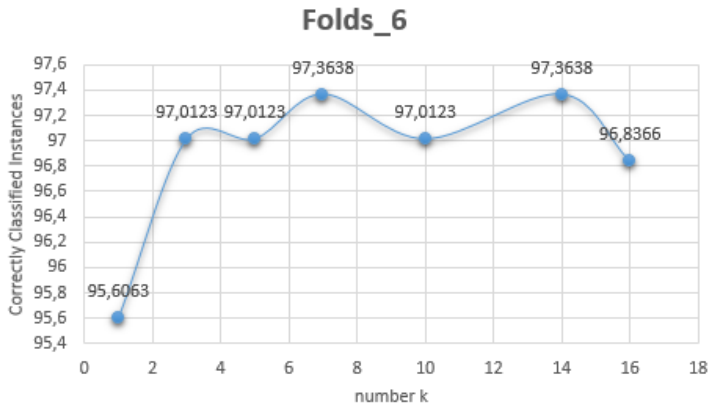
Παρατηρείται λοιπόν ότι η ελάχιστη πιθανότητα λάθους είναι για κ=3, δηλαδή επιλέγοντας τους 3 κοντινότερους γείτονες. Αντίστοιχα υλοποιούμε με τον ίδιο τρόπο για διαφορετικά folds και κ στην cross-validation παίρνοντας τα ακόλουθα διαγράμματα. Στα διαγράμματα που ακολουθούν παρατηρείται για διαφορετικά

folds και κ ,το ποσοστό επιτυχίας που επιτευχθεί από την χρήση του συγκεκριμένου αλγορίθμου.



Εικόνα 5.7 :Γραφική απεικόνιση αποτελεσμάτων αλγορίθμου KNN για folds= 4

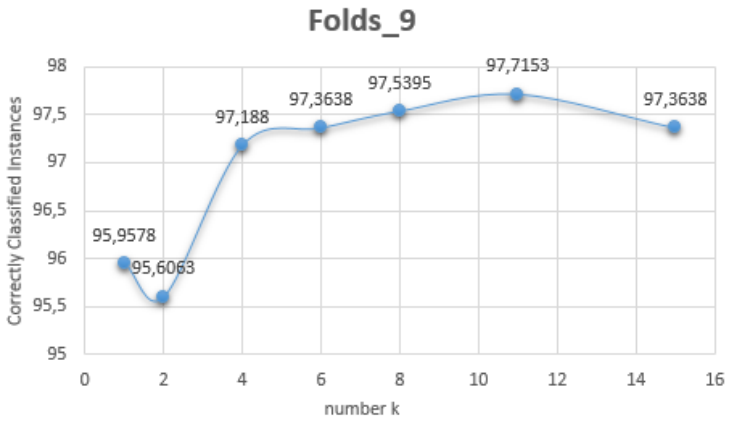
Όταν χωρίστηκαν τα δεδομένα σε 4 δίπλες παρατηρείται ότι το σύστημα βγάζει καλύτερα ποσοστά επιτυχίας με μικρό ποσοστό λάθους για K=10 κοντινότερους γείτονες όπου το ποσοστό επιτυχίας ανέρχεται σε 97,3638 % . Ουσιαστικά δηλαδή μας δείχνει ότι εκπαιδεύοντας το σύστημα με αυτά τα χαρακτηριστικά το ποσοστό λάθους να ταξινομήσει λάθος ένα άτομο είναι μόλις 2,6362%.



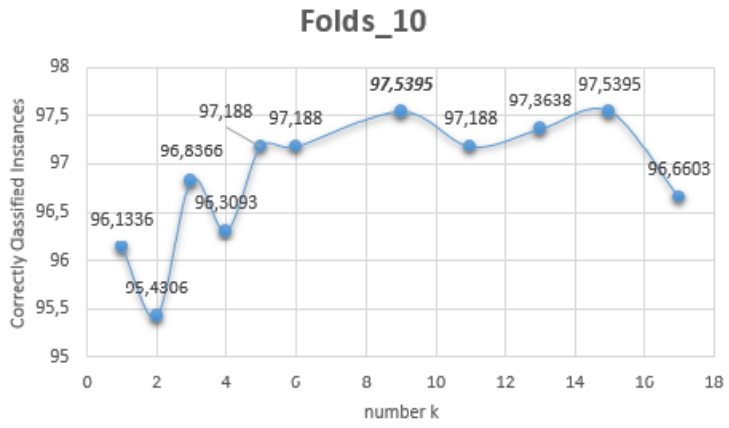
Εικόνα 5.8 :Γραφική απεικόνιση αποτελεσμάτων αλγορίθμου KNN για folds= 6

Αντίστοιχα λοιπόν μεταβάλλεται ουσιαστικά ο αριθμός των κοντινότερων γειτόνων καθώς και ο αριθμός από τις δίπλες που διαχωρίζονται τα δεδομένα μας προκειμένου να παρατηρηθούν εκ νέου τα ποσοστά επιτυχίας. Να υπενθυμίσουμε

πως αν για παράδειγμα ορισθούν 10 δίπλες (folds) τότε ουσιαστικά χρησιμοποιούνται τα 9 folds για την εκπαίδευση του συστηματος και το 1 fold ως data test. Με αυτό λοιπόν τον τρόπο εργαζόμαστε ορίζοντας folds=9 καθώς και folds=10 όπως φαίνεται στα ακόλουθα διαγράμματα.



Εικόνα 5.9 :Γραφική απεικόνιση αποτελεσμάτων αλγορίθμου KNN για folds= 9



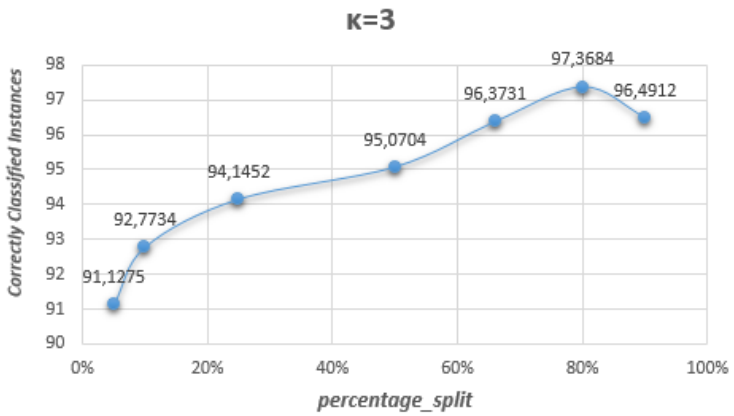
Εικόνα 5.10 :Γραφική απεικόνιση αποτελεσμάτων αλγορίθμου KNN για folds= 10

Έτσι λοιπόν τα καλύτερα αποτελέσματα τα εξάγονται χωρίζοντας το dataset σε folds=9 και συγκεκριμένα επιλέγοντας κ=11 κοντινότερους γείτονες.

Με την ίδια λογική αλλάζοντας την παράμετρο που ορίζει το ποσοστό για training και για test δίνοντας διάφορα ποσοστά ,για κ=2 (τα ποσοστά επιτυχίας της ταξινόμησης κυμαίνονται από 86,5% έως 94,7%) ,κ=3 (91,1% έως 97,3%) , κ=5 (90% έως 94,7%), κ=7 (90,4% έως 96,5%) . Δηλαδή αναφέροντας μέγιστο 96.5% για κ=7 εννοείται ότι το μέγιστο ποσοστό επιτυχίας της ταξινόμησης με βάση των χαρακτηριστικών που έχει εισαχθεί είναι με 96,5% επιτυχία, κατηγοριοποιώντας το

άτομο σε υγιή ή ασθενή με βάση την εκπαίδευση που έχει κάνει στο σύστημα για το συγκεκριμένο data set. Το ποσοστό αυτό προκύπτει καθώς μετά από την εκπαίδευση του συστήματος γίνεται ουσιαστικά έλεγχος σε τι ποσοστό ταξινομήθηκαν σωστά τα δεδομένα που είχαμε ορίσει ως training test data set.

Τα καλύτερα λοιπόν αποτελέσματα εμφανίζονται στο κ=3, όπως φαίνεται αναλυτικότερα στην εικόνα με μέγιστη πιθανότητα σωστής ταξινόμησης το 97,37%.



Εικόνα 5.11 :Γραφική απεικόνιση αποτελεσμάτων αλγορίθμου ορίζοντας κ=3

Παρατηρείται επίσης σε μικρά ποσοστά percentage_split η ύπαρξη χαμηλής απόδοσης του αλγορίθμου, πράγμα φυσιολογικό καθώς όταν για παράδειγμα χρησιμοποιείται το 25% τον δεδομένων για εκπαίδευση και το 75% για test είναι λογικό το σύστημα να αυξάνει την πιθανότητα λάθους (100 - 94.1452 = 5.8548 %) καθώς τα δεδομένα με τα οποία πραγματοποιείται εκπαίδευση του συστήματος είναι πολύ λίγα. Αντίστοιχα όσο αυξάνουμε τα ποσοστά των δεδομένων εκπαίδευσης παρατηρείται και αντίστοιχη αύξηση στο ποσοστό επιτυχίας, με μέγιστο ποσοστό στο 80% όπως ορίζει και η βιβλιογραφία μας.

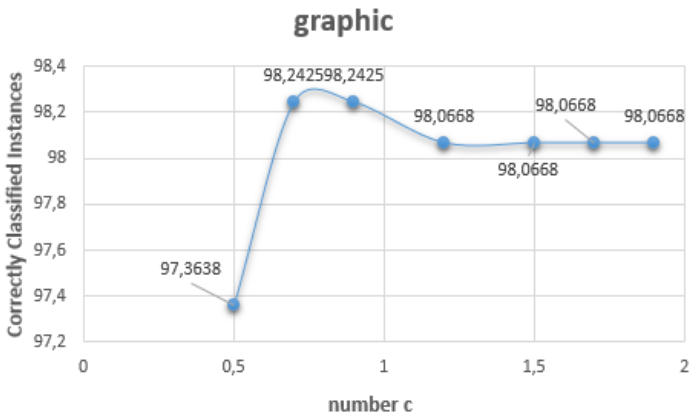
Ουσιαστικά λοιπόν, αυξάνοντας το ποσοστό του training σε σχέση με το ποσοστό του test παρατηρείται να εξάγονται καλύτερα αποτελέσματα , πράγμα που είναι απόλυτα λογικό καθώς το σύστημα μας εκπαιδεύεται καλύτερα καθώς χρησιμοποιεί περισσότερα δεδομένα για την εκπαίδευση του. Αυξάνοντας λοιπόν το percentage split σωστά παίρνουμε τα ανάλογα αποτελέσματα καλύτερης ταξινόμησης.

Παρατηρώντας λοιπόν όλα τα παραπάνω δεδομένα, η καλύτερη επίλυση για τον knn δίνεται για cross-validation με folds=9 , και k=11.

Επιπρόσθετα παρατηρείται ότι μεταβάλλοντας την τιμή των folds για διάφορες τιμές (από 2 – 10 folds) ενώ ταυτόχρονα μεταβάλλουμε το κ , το εύρος της μέγιστης και της ελάχιστης πιθανότητας σωστής ταξινόμησης είναι σχεδόν ίδιο καθώς μεταβάλλεται με ακρίβεια δεκαδικού. Πιο συγκεκριμένα κυμαίνεται μεταξύ ~ 95.5% έως ~ 97.5% .

- **SVM**

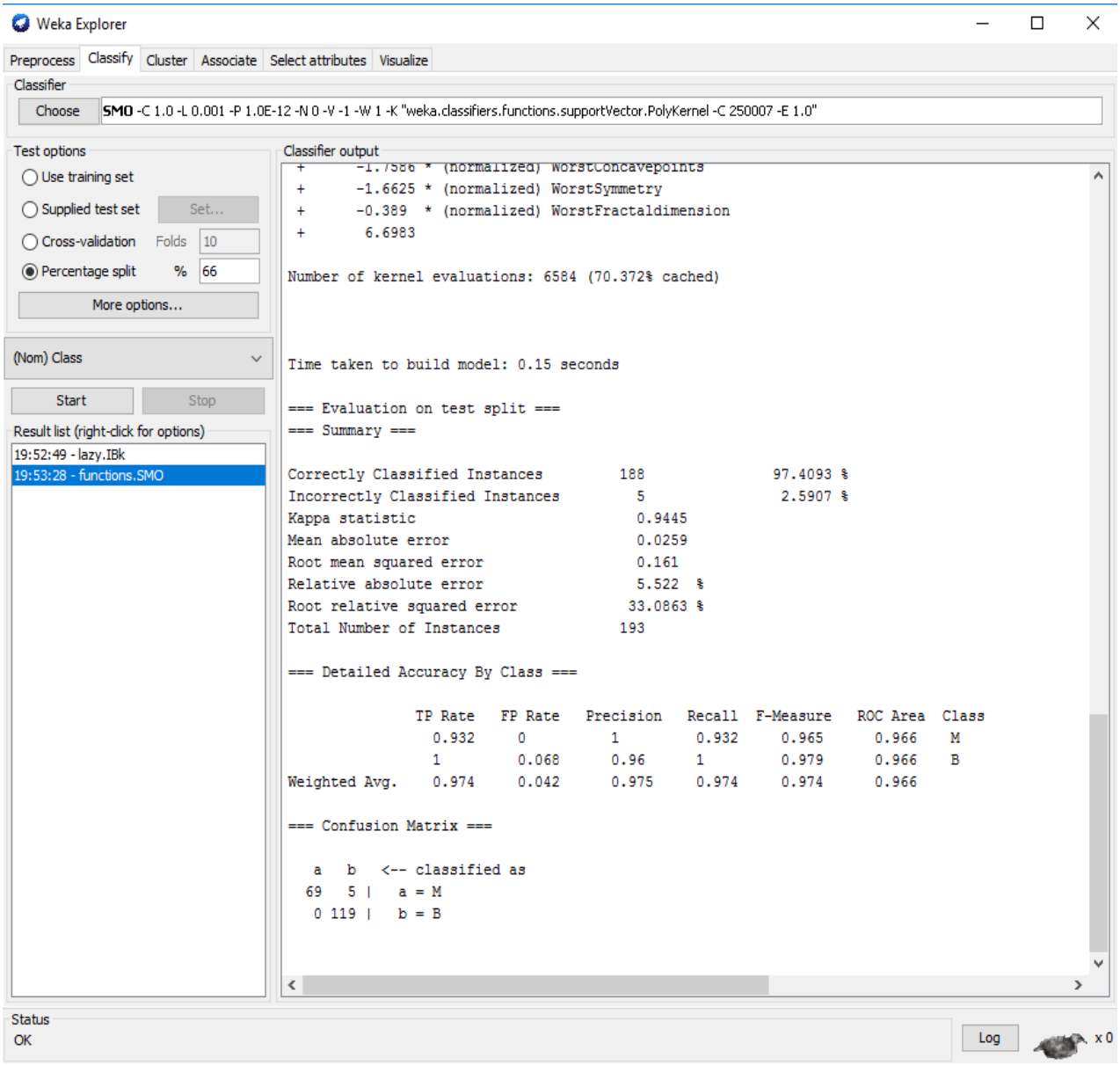
Όσο αναφορά τον svm αλλάζοντας την παράμετρο c εμφανίζονται οι ακόλουθες αλλαγές στην σωστή ταξινόμηση των δεδομένων.



Εικόνα 5.12 :Γραφική απεικόνιση αποτελεσμάτων αλγορίθμου μεταβάλλοντας την τιμή c

Συνεπώς μεταβάλλοντας συνεχώς την παράμετρο c η πιθανότητα σωστής ταξινόμησης μεταβάλλεται, ενώ τα καλύτερα αποτελέσματα εμφανίζονται για c=0.9.

Αντίστοιχα μεταβάλλοντας το ποσοστό των δεδομένων training και set έχουμε μέγιστη για ποσοστό training 80% και 20% test με ποσοστό επιτυχίας της ταξινόμησης στα 98,2456 %, όπως φαίνεται παρακάτω.



Εικόνα 5.13 : εκτέλεση αλγορίθμου ορίζοντας c=1.0 καθώς και percentage split 66% με επιτυχία ταξινόμησης στα 97.4093 %

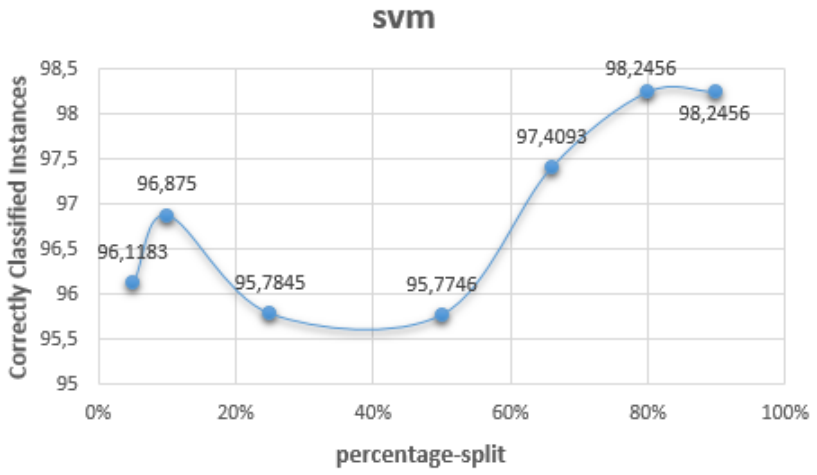
```
=== Confusion Matrix ===

   a   b   <-- classified as
  69   5 |   a = M
   0 119 |   b = B
```

Εικόνα 5.14: Απεικόνιση confusion Matrix από την εκτέλεση του παραπάνω αλγορίθμου.

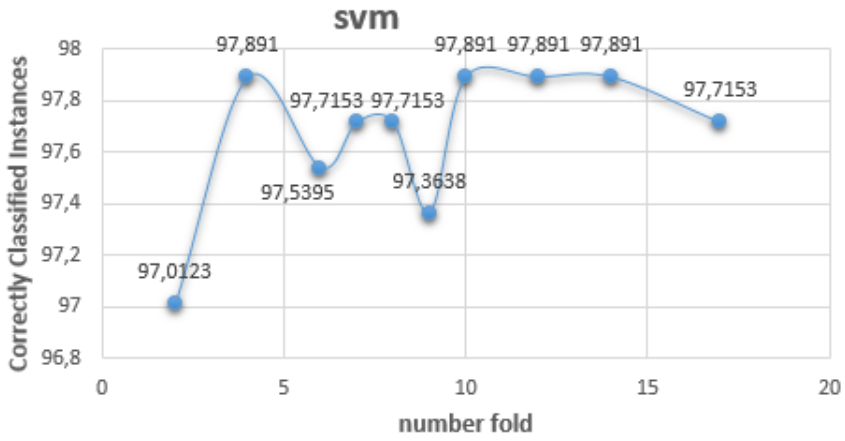
Παρατηρούμε λοιπόν μέσα από τον confusion Matrix ότι 69 από τα άτομα μας τα ταξινόμησε ο αλγόριθμος στην κλάση A (ασθενής) ήταν όντως στην κλάση A ενώ αντίστοιχα ταξινόμησε 5 άτομα στην κλάση B (υγιές άτομο) ενώ στην πραγματικότητα άνηκαν στην κλάση A (ασθενής). Με τον ίδιο τρόπο κατ' αντιστοιχία με τα παραπάνω παρατηρούμε ότι ταξινομεί ορθά 119 στην κλάση B που όντως στην πραγματικότητα ανήκουν στην κλάση B.

Γενικότερα επεξεργάζοντας τα παραπάνω και μεταβάλλοντας τους συντελεστές που προαναφέραμε έχουμε την εξής καμπύλη σε σχέση με το ποσοστό επιτυχίας (correctly classified instances) και του ποσοστού που ορίζουμε για εκπαίδευση και test set (percentage- split):



Εικόνα 5.15: Γραφική απεικόνιση του svm μεταβάλοντας το percentage split

Τέλος όσο αναφορά την svm μέθοδο μεταβάλλοντας όπως και στον knn τα folds παρατηρούνται τα εξής αποτελέσματα.



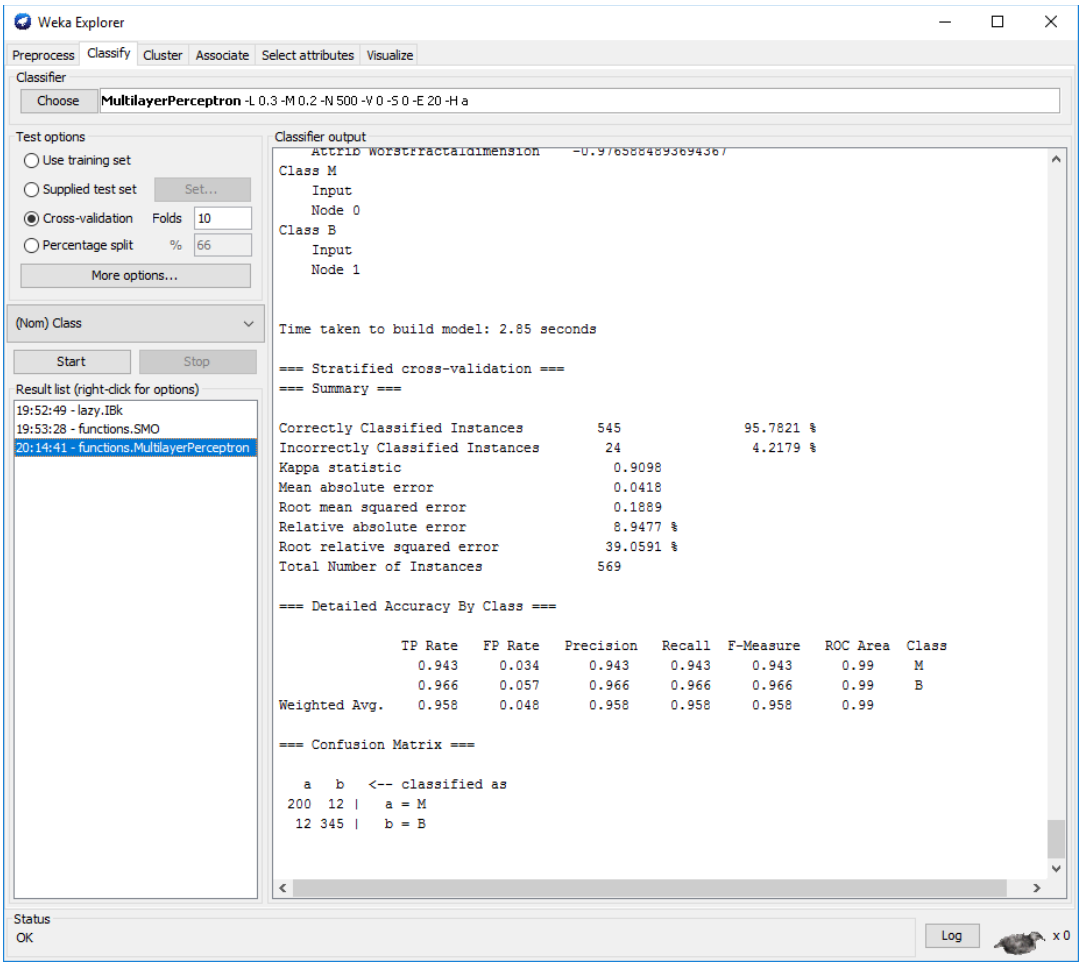
Εικόνα 5.16 : Γραφική απεικόνιση του svm μεταβάλοντας τον αριθμό των folds

Τα καλύτερα λοιπόν αποτελέσματα (μεγαλύτερο ποσοστό επιτυχίας) εμφανίζονται χωρίζοντας το σύστημα σε 10 έως και 13 δίπλες, ενώ μετέπειτα από αν χωριστεί το σύστημα σε 14 δίπλες, εμφανίζεται αύξηση του ποσοστού λάθους.

Έτσι λοιπόν μέσα από τις αλλαγές αυτές των παραμέτρων του svm εντοπίζεται η καλύτερη λύση για το svm ορίζοντας τιμή c=0.9 με ποσοστό 98,2425% πιθανότητα σωστής ταξινόμησης.

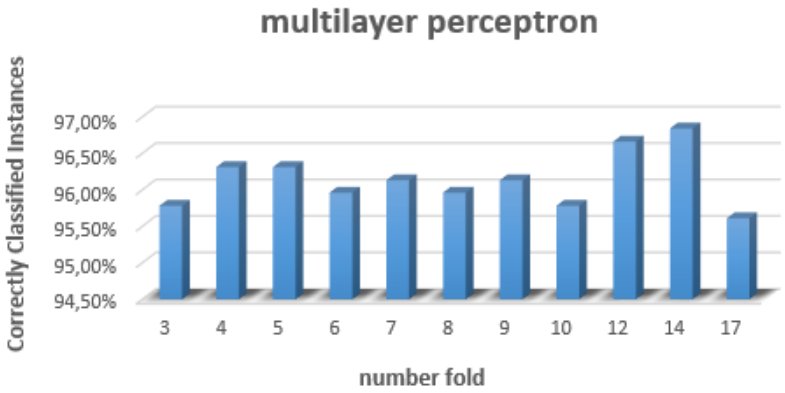
- **Multilayer Perceptron**

Μετέπειτα εξετάζεται η Τεχνική πολλαπλών στρωμάτων, πρόσθιας διάδοσης (multilayer perceptron) μεταβάλλοντας και εδώ τα folds και το ποσοστό του training set αντίστοιχα με τις παραπάνω μεθόδους



Εικόνα 5.17 : Εκτέλεση multilayer Perceptron ορίζοντας την τιμή fold=10

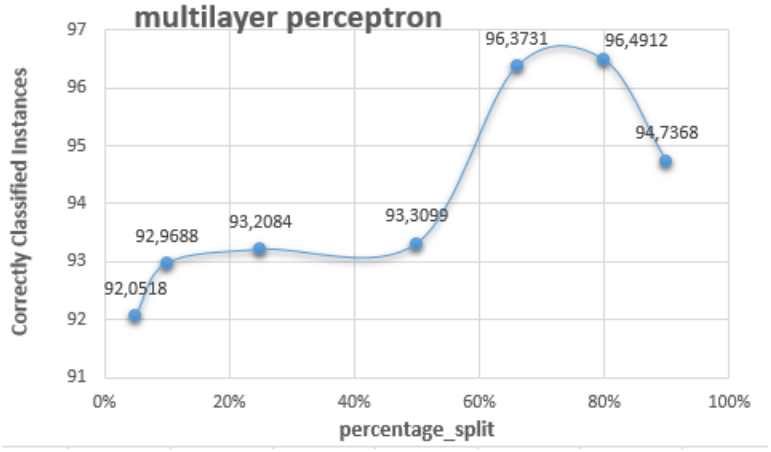
Συνεπώς εμφανίζονται τα ακόλουθα αποτελέσματα από την εκτέλεση των παραπάνω.



Εικόνα 5.18 : Απεικόνιση αποτελεσμάτων multilayer perceptron μεταβαλλόντας τον αριθμό των folds.

Έτσι χαρακτηριστικά εντοπίζεται για κ=14 να εμφανίζεται μεγαλύτερο ποσοστό σε αντίθεση με τα υπόλοιπα κ που βάζουμε.

Τέλος μεταβάλοντας το ποσοστό training και test και πετυχαίνονται τα εξής αποτελέσματα, με καλύτερο στο 80% training.



Εικόνα 5.19 : Απεικόνιση αποτελεσμάτων multilayer perceptron μεταβαλλόντας τον αριθμό το ποσοστό του percentage split.

Συνεπώς όπως προκύπτει και από το παραπάνω γράφημα το ποσοστό αυξάνεται όλο ένα μέχρι και την στιγμή που ορίζουμε ποσοστό 80%, μετέπειτα το ποσοστό επιτυχίας μειώνεται δραματικά(όπως δηλαδή όριζε και η βιβλιογραφία πρωτύτερα).

5.4 Σύγκριση Μεθόδων Ταξινόμησης

Προσπαθούμε λοιπόν να κάνουμε σύγκριση των καλύτερων αποτελεσμάτων των μηχανών ταξινόμησης προκειμένου να βρούμε την καλύτερη μέθοδο υπολογίζοντας κάποιες νέες παραμέτρους στο σύστημα μας, την ευαισθησία, την ακρίβεια και την ειδικότητα.

Συνοψίζοντας τα καλύτερα ποσοστά επιτυχίας για κάθε μέθοδο, προκύπτουν επιγραμματικά τα εξής:

Μέθοδος	*Τρόπος*	*Πιθανότητα Σωστής Ταξινόμησης*
knn	cross-validation με folds=9 και k=11	97,71%
svm	αλλάζοντας το c=0.9	98,24%
multilayer perceptron	cross-validation με folds=14	96,83%

Εικόνα 5.19 : Απεικόνιση καλύτερων αποτελεσμάτων επιτυχούς ταξινόμησης

Συνεπώς πλέον υπολογίζεται η ευαισθησία, η ακρίβεια και η ειδικότητα για τις παραπάνω μεθόδους μέσω του confusion matrix που έχουμε παράξει από τους μηχανισμούς ταξινόμησης με την χρήση των ακόλουθων τύπων:

Πίνακας ταξινόμησης (confusion matrix)

		Ταξινομήθηκαν: Θετικά	Ταξινομήθηκαν: Αρνητικά
Πραγματικά	Θετικά	*a true positive*	*b false negative*
	Αρνητικά	*c false positive*	*d true negative*

Ακρίβεια (Accuracy) $= \frac{a+d}{a+b+c+d} =$ 1 - Σφάλμα (Error)

Ευαισθησία (Sensitivity) $= TPR = \frac{a}{a+b}$

$$\text{Ειδικότητα (Specificity)} \quad = 1 - FPR = \left(1 - \frac{c}{c+d}\right)$$

Με βάση λοιπόν τους παραπάνω τύπους θα εξάγουμε τα αποτελέσματα από τους υπολογισμούς στον confusion matrix, μετά από την υλοποίηση κάθε αλγορίθμου που έχουμε τρέξει και έχουμε λάβει τα αντίστοιχα αποτελέσματα.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         556               97.7153 %
Incorrectly Classified Instances        13                2.2847 %
Kappa statistic                          0.9506
Mean absolute error                      0.0623
Root mean squared error                  0.1622
Relative absolute error                 13.323  %
Root relative squared error             33.5457 %
Total Number of Instances              569

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.943     0.003       0.995     0.943     0.969      0.994    M
                 0.997     0.057       0.967     0.997     0.982      0.994    B
Weighted Avg.    0.977     0.037       0.978     0.977     0.977      0.994

=== Confusion Matrix ===

   a   b   <-- classified as
 200  12 |   a = M
   1 356 |   b = B
```

Εικόνα 5.20 : Απεικόνιση Αποτελεσμάτων αλγορίθμου KNN για folds=9 και k=11

Συνεπώς για τον παραπάνω αλγόριθμο με βάση τους γενικούς τύπους που προαναφέραμε έχουμε:

Ακρίβεια: 556/569 =0.977152

(TPR) Ευαισθησία : 200/212=0.9433

(1-FRP) Ειδικότητα : 1- (1/357)=0.9972

Γενικότερα λοιπόν με βάση τα παραπάνω έχουμε τα εξής αποτελέσματα που αφορούν τις κύριες μεθόδους ταξινόμησης στις οποίες εξαγάγαμε τα καλύτερα αποτελέσματα.

KNN	***FOLD=9***	***number k***	***T R P***	**1-FPR**	***F R P***
		K=1	0,9433	0,97	0,03
		K=2	0,9622	0,953	0,047
		K=4	0,957	0,984	0,0196
		K=6	0,9528	0,986	0,014
		K=8	0,9575	0,986	0,014
		K=11	0,9433	0,9972	0,0028
		K=15	0,9433	0,9916	0,0084
	SVM	**number c**	**T P R**	**1 - F P R**	
		c=0,5	0,9339	0,9972	
		c=0,9	0,957	0,9972	
		c=1,2	0,9528	0,9972	
		c=1,5	0,9575	0,9944	
		c=1,7	0,9575	0,9972	
	MULTILAYER	**number folds**	**T P R**	**1- F P R**	**F P R**
		folds=3	0,9292	0,975	0,025
		folds=5	0,9433	0,975	0,025
		folds=7	0,9481	0,97	0,03
		folds=9	0,9433	0,972	0,028
		folds=12	0,9433	0,981	0,019

Εικόνα 5.21 : Υπολογισμός ευαισθησίας, ακρίβειας και ειδικότητας για τα καλύτερα αποτελέσματα ταξινόμησης

Ειδικότερα παρατηρούνται καλά ποσοστά ακρίβειας και ευαισθησίας στις μεθόδους με τα καλύτερα αποτελέσματα για τις συγκεκριμένες παραμέτρους που έχουμε ορίσει:

		Accuracy	**Sensitivity**
KNN	k=11	0,9771	0,95
SVM	c=0,9	0,982	0,96
Multilayer Perceptron	f=14	0,9683	0,94

Καταλήγοντας, παρατηρούνται ότι τα μεγαλύτερα ποσοστά επιτυχίας από τις 3 μεθόδους που χρησιμοποιήθηκαν με όλες τις δυνατές αλλαγές στις παραμέτρους όπως προαναφέραμε εμφανίζει ο svm με ποσοστό επιτυχίας ταξινόμησης 98,2425% ορίζοντας ως παράμετρο την c ίση με 0.9. Επιπρόσθετα παρατηρείται στον svm πολύ καλά ποσοστά ακρίβειας και ευαισθησίας που σημαίνει ότι είναι αξιόπιστη μέθοδος.

5.4.1 Σύγκριση αποτελεσμάτων με άλλες έρευνες για το συγκεκριμένο dataset.

Συγκρίνεται λοιπόν τώρα το αξιόπιστο 98,2425% ποσοστό επιτυχίας, με αντίστοιχα αποτελέσματα της μεθόδου svm καθώς και με άλλες πιθανές μεθόδους που έχουν γίνει σε άλλες έρευνες προκειμένου να βρεθεί η αποτελεσματικότερη μέθοδος επίλυσης για το συγκεκριμένο dataset.

Παρατηρούμε λοιπόν ότι επιτυγχάνεται σχεδόν ίδιο αποτέλεσμα (αναφερόμενοι πάντα για επεξεργασία και χρήση αλγορίθμων ίδιο data set σε αυτές τις έρευνες) μέσω του svm με ποσοστό ~98,25% (98,2425%) και με πολύ καλή ακρίβεια και ευαισθησία, σε αντιπαραβολή με το (Setiono, 2000) που το αποτέλεσμα της έρευνας ανέρχεται σε ποσοστό 98,24%, ενώ το αποτέλεσμα αυτό φαίνεται να είναι χαμηλότερο έναντι του αποτελέσματος από το (Sarkar, 2000) στο οποίο με την χρήση του αλγορίθμου KNN και την εκπαίδευση του συστήματος το ποσοστό ανέρχεται σε 99.12%. Στην παρούσα φάση αξίζει να διευκρινιστεί ότι αντιπαραβάλαμε τα στοιχεία στις παρούσες έρευνες με την δική μας στο στάδιο κατά το οποίο με το συγκεκριμένο data set πραγματοποιήθηκε training και test και χωρίς την χρήση νέου data set για test και έλεγχο τον αποτελεσμάτων. Επιπρόσθετα σε αντίθεση με τις δύο παραπάνω έρευνες το μέγιστο καλύτερο αποτέλεσμα ταξινόμησης που εξαγάγαμε εμείς με την χρήση του αλγορίθμου KNN ανέρχεται στο ποσοστό 97,71% ορίζοντας folds=9, κ=8 (Βλ. Εικόνα 5.9 :Γραφική απεικόνιση αποτελεσμάτων αλγορίθμου KNN για folds= 9), ενώ παρατηρείται ότι για training and test set το (Setiono, 2000), βγάζει 98,24% ενώ εμείς μέσω του svm με ποσοστό 80% για training και 20% για test βγάζουμε 98,25%.

Επίσης για το training set με την χρήση του αλγορίθμου knn και κ=1 βγάζουμε το μεγαλύτερο ποσοστό επιτυχίας ταξινόμησης όπως και αντίστοιχα και σε αντίστοιχη έρευνα για το συγκεκριμένο dataset. (Sarkar, 2000).

Αντίστοιχα παρατηρούμε πως σε άλλες σχετικές έρευνες σε σύγκριση με διαφορετικούς μηχανισμούς ο svm και ο KNN βγάζει καλύτερα αποτελέσματα έναντι του Bayes rule (Kapouleas, 1990) που βγάζει 97.03% σε training set ωστόσο δεν βγάζει καλύτερα αποτελέσματα από τον Pvm στην αντίστοιχη έρευνα. Τέλος αξίζει να αναφερθεί πως σύμφωνα με την έρευνα του (Wl odzisl) το ποσοστό αυξάνεται ριζικά έναντι των υπολογισμών μας καθώς ανέρχεται σε ποσοστό 99.89%, χρησιμοποιώντας ένα υβριδικό μοντέλο, στο οποίο έχει αναπτυχθεί μια υβριδική μέθοδος για την εξαγωγή των λογικών κανόνων από δεδομένα. Η υβριδική μέθοδος βασίζεται σε μια περιορισμένη multilayer perceptron (C-MLP2LN). (Wl odzisl)

6ο Κεφάλαιο - Συμπεράσματα

Από την μελέτη των δεδομένων που αφορούσαν τον καρκίνο του πνεύμονα και την ανάλυση των δεδομένων αυτών, με την μέθοδο Knn , svm , και multilayer perceptron μέσω του ανοιχτού λογισμικού weka εξάγουμε τα ακόλουθα αποτελέσματα:

Παρατηρούμε λοιπόν ότι η καλύτερη μέθοδος που επιφέρει τα πιο αξιόπιστα αποτελέσματα ταξινόμησης στην κλάση του ασθενή ή υγιή ατόμου είναι η svm, ύστερα από την σύγκριση των 3 καλύτερων μεθόδων, οι οποίες απέδωσαν τα καλύτερα αποτελέσματα. Ορίζοντας για c=0.9 έχουμε την μέγιστη πιθανότητα ορθής ταξινόμησης στα 98,24% με ακρίβεια 0,982 και ευαισθησία 0,96 έναντι των άλλων δυο μεθόδων που χρησιμοποιήθηκαν. Επιπρόσθετα παρατηρούμε ίδια αποτελέσματα με κάποιες άλλες έρευνες, και σε ορισμένες περιπτώσεις βλέπουμε ένα μικρό ποσοστό βελτίωσης από κάποιες άλλες. Έχοντας λοιπόν το συγκεκριμένο data set αξιολογήσαμε το σύστημα μας, καταλήγοντας στο καλύτερο σύστημα, το οποίο προκύπτει με την χρήση του αλγορίθμου svm.

Θα μπορούσαμε να συνεχίσουμε μετέπειτα υλοποιώντας το τελικό σύστημα χρησιμοποιώντας το training set για την εκπαίδευση αυτού και για την αξιολόγηση μπορούμε να χρησιμοποιήσουμε ένα άλλο ξένο data set για να κάνουμε πραγματικό test, οδηγώντας έτσι στην πραγματική έξοδο στην οποία θα καταφέρνει το σύστημα μας να κατατάξει με βάση τα συγκεκριμένα στοιχεία εισόδου αν το εξετάζον άτομο με τα συγκεκριμένα χαρακτηριστικά ανήκει στην κατηγορία του υγιούς ατόμου ή του ασθενούς εμφανίζοντας καρκίνο του πνεύμονα.

Η επιστήμη του data mining αναμφισβήτητα εισβάλει ολοένα και περισσότερο στις μέρες μας σε όλους τους τομείς της σύγχρονης τεχνολογίας όσο και στην ιατρική επιστήμη δίνοντας καίριες λύσεις σε σπουδαία αναπάντητα ερωτήματα, οδηγώντας όλο και πιο κοντά τους επιστήμονες στην αντιμετώπιση σημαντικών ασθενειών

Η ποικιλία των δεδομένων, οι εργασίες εξόρυξης δεδομένων, τα δεδομένα και οι προσεγγίσεις της εξόρυξης δεδομένων δημιουργούν ένα πεδίο που παρέχει πολύ τροφή για το μέλλον της τεχνολογίας και της λήψης αποφάσεων. Η ανάπτυξη αποτελεσματικών δεδομένων μεθόδων εξόρυξης και των συστημάτων, ο σχεδιασμός διάφορων γλωσσών, καθώς και η εφαρμογή των τεχνικών της εξόρυξης δεδομένων για την επίλυση των μεγάλων προβλημάτων εφαρμογής είναι σημαντικά καθήκοντα για τους ερευνητές των δεδομένων του συστήματος εξόρυξης και ανάπτυξης εφαρμογών. Ενέργειες που θέτουν τα θεμέλια και δημιουργούν τις

τάσεις της εξόρυξης δεδομένων που επιδιώκουν σε νέες αποτελεσματικές προκλήσεις.

Δεδομένου του τεράστιου όγκου των διαθέσιμων πληροφοριών στο διαδίκτυο και τον όλο και πιο σημαντικό ρόλο που διαδραματίζει το Web στη σημερινή κοινωνία, το web mining κατακτά όλο ένα και περισσότερο τις σύγχρονες απαιτήσεις τις κοινωνίας. Η Weblog εξόρυξη, και η εξόρυξης δεδομένων υπηρεσιών στο Διαδίκτυο θα αποτελέσουν ένα από τα πιο σημαντικά για άνθηση υποπεδία στην εξόρυξη δεδομένων

Επιπρόσθετα πολλές από τις εφαρμογές που αφορούν δεδομένα ροής (όπως το ηλεκτρονικό εμπόριο, εξόρυξη στο Web, την ανάλυση των αποθεμάτων, live ιαρικά δεδομένα από ιατρικά μηχανήματα κλπ) απαιτούν δυναμικά μοντέλα εξόρυξης δεδομένων που θα κατασκευαστούν σε πραγματικό χρόνο. Πρόσθετη ανάπτυξη είναι απαραίτητη σε αυτόν τον τομέα για Real-time or time-critical data mining.

Η Visual εξόρυξη δεδομένων είναι ένας αποτελεσματικός τρόπος για να ανακαλύψει κανείς τη γνώση από τις τεράστιες ποσότητες δεδομένων. Η συστηματική μελέτη και η ανάπτυξη των οπτικών δεδομένων τεχνικών εξόρυξης θα διευκολύνουν την προώθηση και χρήση του Data Mining ως ένα κύριο εργαλείο για ανάλυση των δεδομένων.

Το αυξανόμενο ωστόσο ενδιαφέρον για την εξόρυξη δεδομένων δημιουργεί και κάποιες απειλές, καθώς με αφθονία λοιπόν καταγράφονται όλο και περισσότερες προσωπικές πληροφορίες και διατίθενται σε ηλεκτρονική μορφή στο διαδίκτυο, σε συνδυασμό, με όλο και πιο ισχυρά εργαλεία εξόρυξης δεδομένων. Συνεπώς ελλοχεύει η απειλή για την προστασία της ιδιωτικής ζωής μας και την ασφάλεια των προσωπικών μας δεδομένων. Συνεπώς κρίνεται η οριοθέτηση ιδιωτικής ζωής, της ασφάλειας των δεδομένων και της χρήσης αυτών η οποία δεν θα διαταραχτεί από την χρήση των τεχνικών του Data Mining, αλλά αντιθέτως θα προσφέρει μόνο θετικά στοιχεία στην σύγχρονα τεχνολογικά εξελιγμένη κοινωνίας μας.

Βιβλιογραφία

Chen, M. H. (1996). Data Mining: An Overview from Database Perspective. *IEEE Transactions on Knowledge and Data Engineering*, pp. 866-883.

Crdd. (2018). Retrieved from http://crdd.osdd.net/raghava/rbpred/svm.jpeg

Devereux, T., & Taylor JA, B. J. (1996). Molecular mechanisms of lung cancer. Interaction of environmental and genetic factors. American College of Chest Physicians) .

Dunham, M. H. (2004). Data Mining: Εισαγωγικά και Προηγμένα Θέματα Εξόρυξης Γνώσης από Δεδομένα. In Γ. Θ. Βασίλης Βερύκιος. Εκδόσεις Νέων Τεχνολογιών.

Eugene Braunwald, K. J. (1997). *Harrison's Principles of Internal Medicine. 14th edition* .

Everitt, B. (2011). Miscellaneous Clustering Methods, in Cluster Analysis. In L. L. Everitt. UK: Wiley.

Ferlay, J. S. (2010). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. . *International Journal of Cancer* .

Frank, R. (1961). Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. . Spartan Books.

G. Holmes, A. D. (1994). *Weka: A machine learning workbench.*

Geoffrey, M. (2004). Analyzing microarray gene expression data. Wiley.

Hall, P. (2008). Choice of neighbor order in nearest-neighbor classification. 2135-2152.

J. Larry Jameson, A. S. (1997). In H. P. Medicine.

Jordan, J. (2019). *Jeremy Jordan support vector machines*. Retrieved from https://www.jeremyjordan.me/support-vector-machines

Kantardzic, K. (2003). *Data Mining : Concepts, Models, Methods, and Algorithms.* John Wiley & Sons.

Kapouleas, I. (1990). An empirical comparison of pattern recognition, neural nets and machine learning classification methods. *Sm Weiss*.

Kleinsmith, L. J. (2006). Principles of cancer biology. . Pearson Benjamin Cummings.

Kotsiantis, S. (2007). Supervised Machine Learning: A Review of Classification Techniques . *Informatica* , 249-268.

Mehmed, K. (2003). Data Mining:Concepts,Models,Methods, and Algorithms. In j. W. sons.

N. Tan, M. (2006). *Introduction to Data Mining.* Addison Wesley.

Nello, C. (2000). An Introduction to Support Vector Machines and other kernel-based learning methods. *Cambridge University Press*.

Pourrajabi, M., Moulavi, D., & Campello. (2014). Model Selection for Semi-Supervised Clustering . *Proceedings of the 17th International Conference on Extending Database Technology (EDBT)*.

Quinlan, J. R. (1996). Improved use of continuous attributes in c4.5. *Journal of Artificial Intelligence Research*, 4:77-90.

Rumelhart. (1986). Learning Internal Representations by Error Propagation. In G. E. David E. Mit Press.

Russell, S. (2004). Τεχνητή Νοημοσύνη, Μια σύγχρονη προσέγγιση. Εκδόσεις Κλειδάριθμος.

S, H. (1999). Neural Networks : A Comprehensive Foundation. Prentice Hall International Inc.

Sarkar, M. (2000). Application of K-Nearest Neigthbors Algorithm on Breast Cancer Diagnosis Problem. *Proc AMIA Symp*, 759–763.

Sas - machine- learning. (2019). Retrieved from www.sas.com/el_gr/insights/analytics/machine-learning.html

Sas -Big Data. (2019). Retrieved from https://www.sas.com/el_gr/insights/articles/big-data/executives-guide-to-cognitive-computing.html

Sas -deep-learning. (2019). Retrieved from https://www.sas.com/en_us/insights/analytics/deep-learning.html

Sepe. (2018). Retrieved from http://www.sepe.gr/gr/information/news/article/10041211/protaseis-gia-tin-tehniti-noimosuni-fernei-i-ee-stis-arhes-tou-2018-/

Setiono, R. (2000). Generating concise and accurate classification rules for breast cancer diagnosis. 205-219.

UCI. (n.d.). Retrieved from http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/.

Usama, F. (1996). From Data Mining Knowledge Discovery in Databases.

Wikipedia. (2019). *Wikipedia*. Retrieved from https://el.wikipedia.org/wiki/Εξόρυξη_δεδομένων

Wl odzisl, R. A. (n.d.). A hybrid method for extraction of logical rules from data. . *Department of Computer Methods, Nicholas Copernicus University*.

Βαζιργιάννης, Μ. (2005). Εξόρυξη Γνώσης απο Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό. In Μ. Χ. Μ. Βαζιργιάννης. Gutenberg.

Ι. Βλαχάβας, Π. Κ. (2011). Τεχνητή Νοημοσύνη Γ' Έκδοση .

Καρδαμάκης, Δ. (2004). Πανεπιστημιακές Παραδόσεις Ακτινοβιολογίας, Ακτινοπροστασίας και Ακτινοθεραπείας. Πατρα.

Κρεμιώτης, Θ. (2016). *Kremiotis*. Retrieved from http://kremiotis.mysch.gr/CancerCells.pdf

Μαργαρίτης, Λ. (1985). Κυτταρική βιολογία. Εκδ. Επτάλοφος.

Παντελάκος, Π. (2005). Διασφάλιση Ποιότητας Ακτινοθεραπείας. *2ο Διεταιρικό Αντικαρκινικό Συνέδριο* , (pp. 364-375).

Σταυλιώτης, Γ. Ε. (2008). ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ (DATA MINING)ΚΑΙ ΚΑΤΗΓΟΡΙΚΑ ΔΕΔΟΜΕΝΑ. pp. 5-8.

Χριστοδουλάκης, Ν. (1994). Σύγχρονη Βιολογία. Εκδ. Πατάκη .

Χρονάκης, Ι. (2006). Επεκτάσεις και Περαιτέρω Αξιολόγηση Συστήματος Αναγνώρισης Μερών του Λόγου για Ελληνικά Κείμενα.