ΜΠΣ ΒΙΟΣΤΑΤΙΣΤΙΚΗ

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΙΑΤΡΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

Δ Ι Π Λ Ω Μ Α Τ Ι Κ Η Ε Ρ Γ Α Σ Ι Α

ΒΑΦΕΙΑΔΗ ΑΝΔΡΟΜΑΧΗ

Ομαδοποίηση και Διαχωρισμός Καρκινοπαθών με τη Χρήση Μετρήσεων από το Τεστ Παπανικολάου (Pap-smear test)

ΑΘΗΝΑ, ΕΤΟΣ 2009

ΜΠΣ ΒΙΟΣΤΑΤΙΣΤΙΚΗ

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΙΑΤΡΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

Δ Ι Π Λ Ω Μ Α Τ Ι Κ Η Ε Ρ Γ Α Σ Ι Α

ΒΑΦΕΙΑΔΗ ΑΝΔΡΟΜΑΧΗ

Ομαδοποίηση και Διαχωρισμός Καρκινοπαθών με τη Χρήση Μετρήσεων από το Τεστ Παπανικολάου (Pap-smear test)

ΑΘΗΝΑ, ΕΤΟΣ 2009

Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο των σπουδών για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στη

ΒΙΟΣΤΑΤΙΣΤΙΚΗ

που απονέμει η Ιατρική Σχολή και το Τμήμα Μαθηματικών του Εθνικού & Καποδιστριακού Πανεπιστημίου Αθηνών και το Τμήμα Μαθηματικών του Πανεπιστημίου Ιωαννίνων.

Εγκρίθηκε την……. από την εξεταστική επιτροπή:

ΟΝΟΜΑΤΕΠΩΝΥΜΟ	***ΒΑΘΜΙΔΑ***	***ΥΠΟΓΡΑΦΗ***
Ι. ΝΤΖΟΥΦΡΑΣ (Επιβλέπων)	**ΕΠ. ΚΑΘΗΓΗΤΗΣ**	**……**
Γ. ΔΟΥΝΙΑΣ	**ΑΝ. ΚΑΘΗΓΗΤΗΣ**	**…….**
Γ.ΤΟΥΛΟΥΜΗ	**ΕΠ. ΚΑΘΗΓΗΤΡΙΑ**	**…….**

Στον αγαπημένο μου πατέρα…

Ευχαριστίες

Ευχαριστώ θερμά τους καθηγητές που ασχολήθηκαν με τη διπλωματική αυτή εργασία τόσο για τις γνώσεις που μου προσέφεραν απλόχερα όσο και για την υπομονή και κατανόηση που έδειξαν ως το τέλος.

Ευχαριστώ επίσης όλους τους καθηγητές που με δίδαξαν στο μεταπτυχιακό της βιοστατιστικής

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1:Επίχρισμα Παπανικολάου (smear data)……...9

1.1 Εισαγωγή ……..... 10

1.2 Σκοπός ……..... 10

1.3 Ο διαχωρισμός των επιχρισμάτων Παπανικολάου (pap – smear)…… 10

1.4 Η παλαιά βάση δεδομένων επιχρίσματος Παπανικολάου …….. 13

1.5 Η νέα βάση δεδομένων επιχρίσματος Παπανικολάου …… 18

ΚΕΦΑΛΑΙΟ 2:Αλγόριθμοι Πληροφορικής……19

2.1 Νευρωνικά δίκτυα, ασαφής λογική και γενετικοί αλγόριθμοι……...19

2.2 Μετρήσεις σφάλματος (error measurements) ……. 21

2.3 Επαναξιολόγηση κ δειγμάτων (k – fold cross – validation) …… 22

2.4 Επανάληψη της μεθόδου επαναξιολόγησης (cross – validation rerunning)……. 22

2.5 Προσομειωμένη ανόπτηση (simulated annealing)…… 22

2.6 Επιτηρούμενη δημιουργία συστάδων (supervised clustering) …… 24

2.7 Ομαδοποίηση με την μέθοδο των C-μέσων ……..… 25

2.7.1 Hard C – μέσων …… 25

2.7.2 Αλγόριθμος fuzzy C – μέσων (Bezdek 1981) …….. 26

2.8 Δημιουργία συστάδων Gustafson – Kessel (Gustafson – Kessel clustering)…… 27

2.9 Μέθοδος ελαχίστων τετραγώνων (Least Square Method) ……... 28

2.10 Αλγόριθμοι Κ κοντινότερου γείτονα (KNN και WKNN) …… 29

2.11 Κέντρο βαρύτητας κοντινότερης ομάδας (Nearest Class Gravity Center-NCC) …….. 29

2.12 Νευροασαφής μέθοδος εξαγωγής συμπεράσματος (Neurofuzzy Inference Method- NFI) 31

2.13 Μέθοδοι επιλογής χαρακτηριστικών …….. 32

2.13.1 Γενετικός Αλγόριθμος (Genetic Algorithm-GA) …… 32

2.13.2 Tabu search (TS)(Yannis Marinakis $ George Dounias 2006) …….. 33

2.13.3 Βελτιστοποίηση Αποικίας Μυρμηγκιών (Ant Colony Optimization-ACO)……..... 34

ΚΕΦΑΛΑΙΟ 3: Αλγόριθμοι Στατιστικής……..….37

3.1 συντελεστής συσχέτισης ……... 37

3.2 t έλεγχος για την διαφορά των μέσων τιμών (t-test) …… 38

3.3 Ανάλυση διασποράς (ANOVA) …… 39

3.4 Λογαριθμιστική παλινδρόμηση ……. 39

3.5 διαβαθμισμένη λογαριθμιστική παλινδρόμηση ……. 40

3.6 Δημιουργία συστάδων Κ μέσων …… 41

3.7 Διαχωριστική ανάλυση …….. 42

ΚΕΦΑΛΑΙΟ 4: Μονομεταβλητή στατιστική ανάλυση……..43

4.1 Παλαιά βάση δεδομένων επιχρίσματος Παπανικολάου …… 43

4.1.1 περιγραφικά στοιχεία ……. 43

4.1.2 Συσχετίσεις μεταξύ όλων των χαρακτηριστικών της βάσης δεδομένων (correlation)….. 46

4.1.3 έλεγχος διαφοράς μέσων τιμών των χαρακτηριστικών ανάμεσα στις κατηγορίες υγιών και μη υγιών κυττάρων (t-test) ……. 48

4.1.4 ανάλυση διασποράς κάθε χαρακτηριστικού ανάμεσα στις 7 κατηγορίες των κυττάρων (oneway ANOVA) ……. 49

4. 2 Νέα βάση δεδομένων επιχρίσματος Παπανικολάου ……. 50

4. 2. 1 περιγραφικά στοιχεία …… 50

4.2.2 Συσχετίσεις μεταξύ όλων των χαρακτηριστικών της βάσης δεδομένων (correlation) … 53

4.2.3 έλεγχος διαφοράς μέσων τιμών των χαρακτηριστικών ανάμεσα στις κατηγορίες υγιών και μη υγιών κυττάρων (t-test) ……..... 55

4.2.4 Ανάλυση διασποράς κάθε χαρακτηριστικού ανάμεσα στις 7 κατηγορίες των κυττάρων (oneway ANOVA) …….. 56

ΚΕΦΑΛΑΙΟ 5: Αποτελέσματα αλγορίθμων Στατιστικής…….58

5.1 Λογιστική παλινδρόμηση (logistic regression) …… 58

5.2 Διατάξιμη λογιστική παλινδρόμηση (ordinal logistic regression) ……..… 62

5.3 Ανάλυση κατά συστάδες (cluster analysis) …… 67

5.4 Διαχωριστική ανάλυση(discriminant analysis) …….. 70

5.5 Σύγκριση αποτελεσμάτων μεταξύ των δύο βάσεων δεδομένων……. 75

ΚΕΦΑΛΑΙΟ 6:Αποτελέσματα αλγορίθμων πληροφορικής……..77

6.1 Αποτελέσματα αλγορίθμου hard C μέσων …… 79

6.2 Αποτελέσματα αλγορίθμου FCM …….. 82

6.3 Αποτελέσματα αλγορίθμου Gustafson – Kessel …… 85

6.4 Αποτελέσματα μεθόδου ελαχίστων τετραγώνων ……... 88

6.5 Αποτελέσματα αλγορίθμων KNN και WKNN (Norup2005) …… 88

6.6 Αποτελέσματα αλγορίθμου NCC(Norup2005) ……. 89

6.7 Αποτελέσματα μεθόδου NFI(Norup2005) …… 89

6.8 Αποτελέσματα αλγορίθμου GA(Yannis Marinakis $ George Dounias 2006) …….. 90

6.9 Αποτελέσματα μεθόδου έρευνας tabu(Yannis Marinakis $ George Dounias 2006) …… 93

6.10 Αποτελέσματα μεθόδου ACO(Yannis Marinakis $ George Dounias 2006) ……. 96

6.11 Συγκριτικός πίνακας επιλεγμένων χαρακτηριστικών για τους αλγόριθμους GA, TS, ACO… 99

ΚΕΦΑΛΑΙΟ 7:Σύγκριση αποτελεσμάτων……..100

ΚΕΦΑΛΑΙΟ 8: Συζήτ6ηση……..103

ΠΑΡΑΡΤΗΜΑ……...107

ΚΕΦΑΛΑΙΟ 1

Επίχρισμα Παπανικολάου (smear data)

1.1 ΕΙΣΑΓΩΓΗ

Ο καρκίνος του τραχήλου της μήτρας είναι ο τρίτος κατά σειρά συχνότητας καρκίνος στις γυναίκες. Η θνησιμότητα από τον καρκίνο του τραχήλου έχει ελαττωθεί κατά 50% στα τελευταία 40 χρόνια και η συχνότητα της νόσου σε προχωρημένο στάδιο κατά τη διάγνωση έχει ελαττωθεί κατά 70% στο ίδιο χρονικό διάστημα (en.wikipedia.org). Αυτό είναι αποτέλεσμα της έγκαιρης διάγνωσης και θεραπείας, στην οποία έχει συμβάλλει κατά πολύ το τεστ Παπανικολάου. Ο Γεώργιος Παπανικολάου από το 1928 ανακάλυψε ότι ο καρκίνος του τραχήλου μπορεί να διαγνωσθεί έγκαιρα, μια θεωρία που αναπτύχθηκε περαιτέρω τα επόμενα χρόνια και βρίσκει εφαρμογή παγκόσμια, έχοντας οδηγήσει στην σημαντική μείωση της θνησιμότητας από τον καρκίνο του τραχήλου. Χρησιμοποιώντας την τεχνική που ονομάζεται επίχρισμα Παπανικολάου (pap-smear), κατέστη δυνατόν να χρωματίσουμε τα κύτταρα. Με ένα μικροσκόπιο οι κυτταροτεχνικοί μπορούν να ανιχνεύσουν προκαρκινικά κύτταρα στο μητριαίο τράχηλο. Μεταξύ άλλων, η μέθοδος χρησιμοποιήθηκε και στο Πανεπιστημιακό Νοσοκομείο Herlev της Δανίας από το οποίο είναι και οι βάσεις δεδομένων που χρησιμοποιούνται σε αυτήν την εργασία. Στο τμήμα της παθολογίας η ταξινόμηση γίνεται από καλά εκπαιδευμένους κυτταροτεχνικούς, χρησιμοποιώντας μικροσκόπιο. Αυτή η μέθοδος δεν είναι τόσο γρήγορη και απαιτεί τη διαθεσιμότητα καλά καταρτισμένων κυτταροτεχνικών. Όμως, χρησιμοποιώντας έναν υπολογιστή για την ταξινόμηση, θα μπορούσε να επιτευχθεί μια καλύτερη λύση, θεωρώντας φυσικά δεδομένο ότι δεν θα αυξάνεται το σφάλμα. Μερικά συστήματα είναι ήδη διαθέσιμα αλλά και πολύ ακριβά. Τα τελευταία 10 χρόνια έχουν γραφτεί αρκετές διπλωματικές εργασίες σχετικές με την ταξινόμηση των επιχρισμάτων Παπανικολάου, όπως του Byriel G.το 1999 με τίτλο "Neuro - fuzzy classification of cells in cervical smears", του Martin E.το 2003 με τίτλο "Pap smear classification" και του Norup J.το 2005 με τίτλο "Classification of pap smear data by transductive neuro – fuzzy methods", σε συνεργασία με το Πανεπιστημιακό Νοσοκομείο Herlev, τόσο στο Orsted της Δανίας, όσο και στο Πανεπιστήμιο του Αιγαίου στη Χίο. Η συνεργασία αυτή διεύρυνε τη γνώση πάνω σε μεθόδους ταξινόμησης και επέφερε την ανάπτυξη μιας μεγάλης βάσης δεδομένων που αποτελείται από δεδομένα επιχρίσματος (smear data), προσεκτικά εξετασμένα από κυτταροτεχνικούς και γιατρούς. Αυτή η βάση δεδομένων καθώς και μια παλιότερη εκδοχή της, διατίθενται ελεύθερα στο διαδίκτυο(fuzzy.iau.dtu.dk/download/smear2005 και fuzzy.iau.dtu.dk/smear) με σκοπό να παρέχουν τη δυνατότητα πρόσβασης σε ερευνητές και να τους βοηθούν να αναπτύξουν νέες μεθόδους ταξινόμησης των κυττάρων.

1.2 Σκοπός

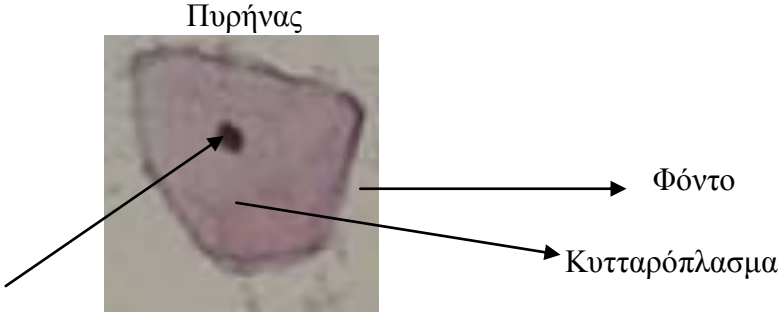
Σκοπός της διπλωματικής αυτής εργασίας είναι η εφαρμογή γνωστών στατιστικών μεθόδων στα δεδομένα επιχρίσματος Παπανικολάου για την διεξαγωγή συμπερασμάτων σχετικών με την επίδραση των χαρακτηριστικών των κυττάρων που περιέχουν οι βάσεις στη διάγνωση της κατάστασης του κυττάρου αλλά και η προσπάθεια ταξινόμησης των κυττάρων στις ήδη υπάρχουσες κατηγορίες με όσο το δυνατόν μεγαλύτερα ποσοστά επιτυχούς ταξινόμησης. Επίσης γίνεται μια ανασκόπηση των μεθόδων που έχουν ήδη εφαρμοστεί στα δεδομένα καθώς και παράθεση των αποτελεσμάτων τους με σκοπό την σύγκρισή τους με τα αποτελέσματα αυτής της εργασίας.

1.3 Ο διαχωρισμός των επιχρισμάτων Παπανικολάου (pap – smear)

Η μέθοδος επιχρίσματος Παπανικολάου (Papanicolaou Smear) είναι μια ιατρική διαδικασία που έχει σκοπό την ανίχνευση προκαρκινικών κυττάρων στον τράχηλο της μήτρας. Η ιατρική περιγραφή που ακολουθεί είναι βασισμένη στον Martin (2003) και στον Dr. Indman (2005).

Χρησιμοποιώντας ένα μικρό βουρτσάκι, βαμβακοφόρο στυλεό ή ξύλινο στυλεό, συλλέγεται ένα κυτταρολογικό δείγμα από τον τράχηλο και αλείφεται πάνω σε ένα λεπτό γυάλινο πλακίδιο. Για να αποσαφηνιστούν τα χαρακτηριστικά των κυττάρων, το επίχρισμα χρωματίζεται σύμφωνα με τη μέθοδο Παπανικολάου, έτσι ώστε να δίνεται έμφαση στα διαφορετικά συστατικά των κυττάρων με συγκεκριμένα χρώματα.

Γενικά, η εικόνα ενός χρωματισμένου κυττάρου περιέχει τον πυρήνα, ο οποίος περικλείεται από το κυτταρόπλασμα, πάνω σε ένα φόντο, όπως φαίνεται στο Διάγραμμα 1.1



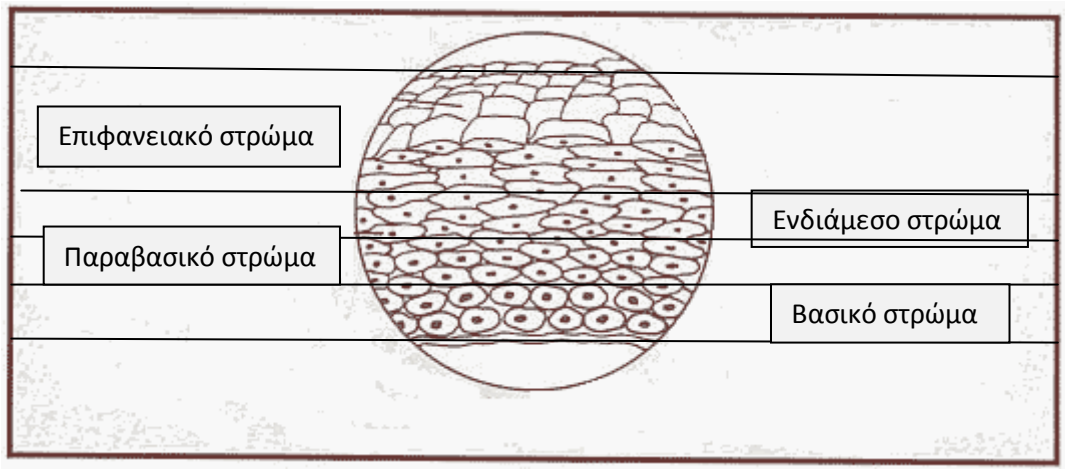
Διάγραμμα 1.1 : Κύτταρο επιχρίσματος Παπανικολάου (Norup2005)

Από την επισκόπηση των χαρακτηριστικών των κυττάρων όπως μέγεθος, χρώμα, σχήμα και υφή του πυρήνα και του κυτταροπλάσματος, οι κυτταροτεχνικοί μπορούν να διαγνώσουν τα καρκινικά κύτταρα. Η εργασία αυτή είναι πολύ απαιτητική και προϋποθέτει την ύπαρξη πολύ καλά εκπαιδευμένων κυτταροτεχνικών. Κάθε διαφάνεια στο μικροσκόπιο περιέχει περίπου 300.000 κύτταρα με διαφορετικό προσανατολισμό και αλληλοοεπικάλυψη.

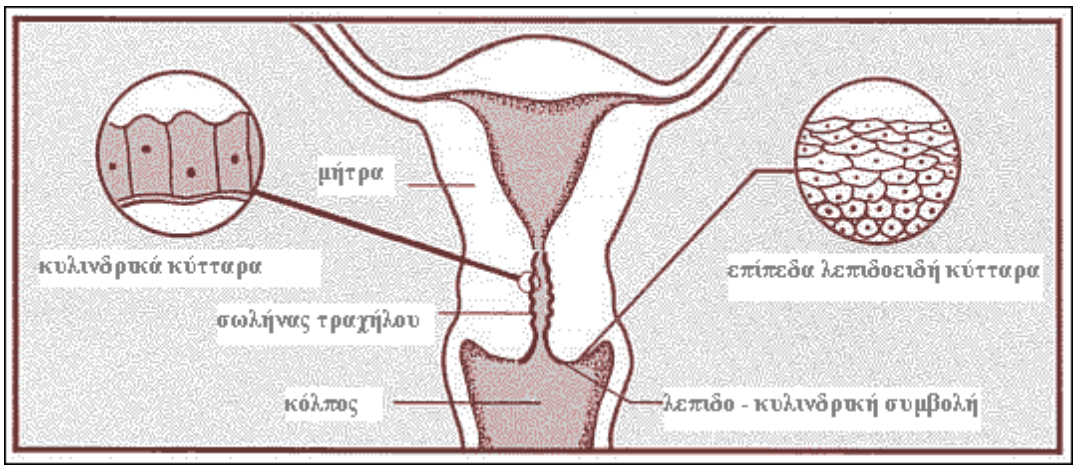
Στον τράχηλο υπάρχουν διάφορα είδη κυττάρων και βρίσκονται σε ξεχωριστές περιοχές: α) λεπιδοειδής περιοχή (squamous area) και β) κυλινδρική περιοχή (columnar area).

Οι λεπιδοειδείς περιοχές βρίσκονται στο κάτω μέρος, στον σωλήνα του τραχήλου, ακριβώς έξω από τον κόλπο. Τα κύτταρα εδώ χωρίζονται σε 4 στρώματα: στο βασικό (basal), στο παραβασικό (parabasal), στο ενδιάμεσο (intermediate) και στο επιφανειακό στρώμα (superficial). Όταν τα κύτταρα ωριμάζουν μετακινούνται μεταξύ των στρωμάτων και τελικά εκτινάσσονται από την επιφάνεια στο επιφανειακό στρώμα. Καθώς μετακινούνται μεταξύ των στρωμάτων, τα κύτταρα αλλάζουν σχήμα, χρώμα και άλλα χαρακτηριστικά. Τα κύτταρα που βρίσκονται στο βασικό στρώμα είναι μικρά και σφαιρικά, με μεγάλο πυρήνα και μικρό κυτταρόπλασμα. Μετακινούμενα μεταξύ των στρωμάτων το κυτταρόπλασμα γίνεται μεγαλύτερο και ο πυρήνας μικρότερος. Το γενικό σχήμα του κυττάρου παίρνει οβάλ μορφή για αυτό και τα κύτταρα στο επιφανειακό στρώμα αναφέρονται και ως επίπεδα λεπιδοειδή κύτταρα (flat squamous cells). Το σχήμα των κυττάρων και τα τέσσερα στρώματα φαίνονται στο Διάγραμμα 1.2

Η κυλινδρική περιοχή βρίσκεται στο επάνω μέρος και ειδικότερα στον σωλήνα του τραχήλου. Τα κυλινδρικά κύτταρα υπάρχουν μόνο σε ένα στρώμα, το βασικό. Χαρακτηριστικό αυτών των κυττάρων είναι το κυλινδρικό σχήμα με επιμήκες κυτταρόπλασμα και μεγάλο πυρήνα που βρίσκεται στη μια άκρη. Κάπου ανάμεσα σε αυτές τις δύο περιοχές τα κύτταρα συναντώνται στην λεπιδο – κυλινδρική συμβολή (squamo – columnar junction). Η συμβολή αυτή μπορεί να βρίσκεται είτε εντός είτε εκτός του τραχήλου. Η κυλινδρική περιοχή και άλλες λεπτομέρειες της μήτρας φαίνονται στο Διάγραμμα 1.3



Διάγραμμα 1.2: Διαγραμματική απεικόνιση των λεπιδοειδών κυττάρων μεταξύ των τεσσάρων στρωμάτων. (Norup2005)`



Διάγραμμα 1.3 : Απεικόνιση της μήτρας καθώς και της θέσης των λεπιδοειδών και των κυλινδρικών κυττάρων. (Dr. Indman, 2005)

Όταν η γενετική πληροφορία σε ένα κύτταρο αλλάξει με κάποιον τρόπο, το κύτταρο δεν θα διαιρεθεί όπως έπρεπε και θα μετατραπεί σε προκαρκινικό κύτταρο. Χρησιμοποιώντας ιατρική ορολογία, τα κύτταρα χωρίζονται σε 2 βασικά είδη διάγνωσης:

1) **Δυσπλασία**. Ο όρος δυσπλασία σημαίνει διαταραγμένη ανάπτυξη. Η τραχηλική δυσπλασία χωρίζεται σε 3 τύπους: ελαφριά (mild), μεσαία (moderate) και βαριά (severe), περιγράφοντας έτσι τον κίνδυνο να εξελιχθεί ένα κύτταρο σε κακοήθες καρκινικό. Τα χαρακτηριστικά των κυττάρων στην δυσπλασία εξαρτώνται από το είδος. Στην ελαφριά δυσπλασία έχουν μεγεθυμένο και φωτεινό πυρήνα. Στην μεσαία δυσπλασία ο πυρήνας είναι μεγαλύτερος και πιο σκούρος. Ο πυρήνας έχει αρχίσει να εκφυλίζεται, γεγονός που φαίνεται ως κοκκίωση του πυρήνα. Στην βαριά δυσπλασία ο

πυρήνας είναι μεγάλος, σκούρος και συχνά παραμορφωμένος. Το κυτταρόπλασμα είναι σκούρο και μικρό σε σύγκριση με τον πυρήνα.

2) **Μη διηθητικός καρκίνος (in – situ)**: τα κύτταρα αυτού του τύπου χαρακτηρίζονται από πολύ μεγάλο πυρήνα.

Σύμφωνα με τις ιατρικές περιγραφές που αναφέρθηκαν, ιδιότητες όπως το μέγεθος, η περιοχή, το σχήμα και η φωτεινότητα είναι καλά περιγραφικά γνωρίσματα. Επιπρόσθετα, το σχετικό μέγεθος του πυρήνα ως προς το κυτταρόπλασμα είναι αρκετά περιγραφικό γιατί μεγαλώνει στα προκαρκινικά κύτταρα. Έτσι ορίζουμε την αναλογία πυρήνα/κυτταροπλάσματος ως:

$$\text{Αναλογία πυρήνα / κυτταροπλάσματος} = \frac{\text{Εμβαδόν πυρήνα}}{\text{Εμβαδόν πυρήνα} + \text{Εμβαδόν κυτταροπλάσματος}} \quad (1.1)$$

1.4 Η παλαιά βάση δεδομένων επιχρίσματος Παπανικολάου

Η πρώτη βάση με δεδομένα κυττάρων επιχρίσματος Παπανικολάου δημιουργήθηκε στο τμήμα παθολογίας, στο Πανεπιστημιακό Νοσοκομείο του Herlev. Αναφέρεται ως παλαιά γιατί αργότερα δημιουργήθηκε και μια νέα βάση δεδομένων παρόμοιου τύπου. Η συλλογή των δεδομένων έγινε με τη βοήθεια κυτταροτεχνικών του Πανεπιστημίου και συλλέχθησαν συνολικά 500 εικόνες κυττάρων με ψηφιακή κάμερα και μικροσκόπιο. Ειδικοί κυτταροτεχνικοί πήραν ψηφιακές εικόνες των κυττάρων χρησιμοποιώντας μικροσκόπιο με ανάλυση 0.201 μm/pixel. Κάθε εικόνα ενός κυττάρου ταξινομείται χειροκίνητα σε μια από τις 7 διαφορετικές κατηγορίες κυττάρων που περιγράφονται στον Πίνακα 1.1. Για την αξιολόγηση της διαδικασίας, η ταξινόμηση γίνεται δύο φορές από διαφορετικούς κυτταροτεχνικούς. Αν οι απόψεις τους είναι διαφορετικές τότε η εικόνα δεν συμπεριλαμβάνεται στην βάση. Τα κύτταρα χωρίστηκαν σε 7 κατηγορίες, 4 κατηγορίες υγιών κυττάρων και 3 κατηγορίες μη υγιών κυττάρων όπως φαίνεται στον Πίνακα 1.1.

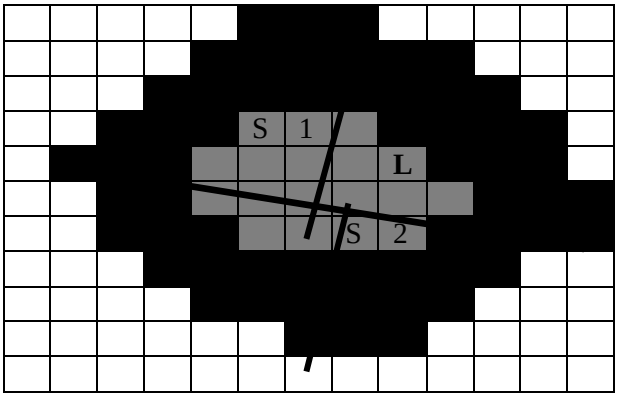
Φυσιολογικά κύτταρα	**Αριθμός κυττάρων**	**Μη φυσιολογικά κύτταρα**	**Αριθμός κυττάρων**
Κυλινδρικά επιθήλια	50	**Ελαφριά λεπιδοειδής μη κερατινώδης δυσπλασία**	100
Παραβασικά λεπιδοειδή επιθήλια	50	**Μεσαία λεπιδοειδής μη κερατινώδης δυσπλασία**	100
Ενδιάμεσα λεπιδοειδή επιθήλια	50	**Βαριά λεπιδοειδής μη κερατινώδη δυσπλασία**	100
Επιφανειακά λεπιδοειδή επιθήλια	50		
Σύνολο	200		300

Πίνακας 1.1 : Τύποι κυττάρων επιχρίσματος Παπανικολάου

Από τις εικόνες των κυττάρων που συλλέχθησαν, αντλήθηκαν χαρακτηριστικά που περιγράφουν τα κύτταρα. Η άντληση αυτών των χαρακτηριστικών έγινε με το εμπορικό πρόγραμμα CHAMP (*www.*dimac-imaging.*com*). Το πρόγραμμα αυτό αναλύει εικόνες και δημιουργεί μια βάση δεδομένων από αντικείμενα που υπάρχουν σ' αυτές. Το CHAMP αναγνωρίζει 3 είδη αντικειμένων στα κύτταρα: το φόντο, το κυτταρόπλασμα και τον πυρήνα. Το πρόγραμμα είναι εκπαιδευμένο να τραβάει γραμμές ανάμεσα στα αντιπροσωπευτικά αντικείμενα κάθε κατηγορίας. Αν θέλουμε να εκπαιδεύσουμε το πρόγραμμα να αναγνωρίζει το φόντο, τότε σχηματίζονται γραμμές στο φόντο της εικόνας. Τα pixel που βρίσκονται κάτω από τις γραμμές θεωρούνται αντιπροσωπευτικά της συγκεκριμένης κατηγορίας. Αυτό γίνεται για όλες τις κατηγορίες αντικειμένων που πρέπει να αναγνωρίζει το πρόγραμμα.

Η αναγνώριση βασίζεται σε πληροφορίες για το χρώμα. Τα τυπικά χρώματα κάθε κατηγορίας αντικατοπρίζονται σε έναν τρισδιάστατο χρωματικό χώρο. Τα χρωματικά σημεία που δημιουργούνται καθορίζουν ομάδες στο χρωματικό χώρο, μια ομάδα για κάθε κατηγορία. Το πρόγραμμα χρησιμοποιεί πληροφορίες από τις ομάδες για να διαχωρίσει τις εικόνες σε κατηγορίες. Όταν το πρόγραμμα εκπαιδευτεί και με τις 3 κατηγορίες, μπορεί να διαχωρίσει τις εικόνες σε αντικείμενα φόντου, κυτταροπλάσματος και πυρήνα. Το πώς γίνεται αυτός ο διαχωρισμός εξαρτάται από τα δεδομένα εκπαίδευσης.

Από κάθε εικόνα μετριέται ένας αριθμός χαρακτηριστικών και εισάγεται σε μια βάση δεδομένων. Κάθε σειρά στη βάση δεδομένων περιέχει μετρήσεις για μία εικόνα. Οι μετρήσεις περιγράφουν χαρακτηριστικά του πυρήνα και του κυτταροπλάσματος. Στο Διάγραμμα 1.4 φαίνεται η δυαδική εικόνα ενός κυττάρου, όπως την αναλύει το πρόγραμμα CHAMP.



Διάγραμμα 1.4 : Δυαδική εικόνα ενός κυττάρου με το φόντο (λευκό τμήμα), το κυτταρόπλασμα (μαύρο τμήμα) και τον πυρήνα (γκρι τμήμα). Για το κυτταρόπλασμα φαίνονται η μεγαλύτερη διάμετρος (L) και η μικρότερη διάμετρος (S1 και S2). (Martin2003).

	Χαρακτηριστικά των δεδομένων επιχρίσματος Παπανικολάου		
1	N area	Εμβαδόν πυρήνα	Kerne_a
2	C area	Εμβαδόν κυτταροπλάσματος	Cyto_a
3	N/C ratio	Αναλογία πυρήνα/κυτταροπλάσματος	NC
4	N brightness	Φωτεινότητα πυρήνα	Kerneycol
5	C brightness	Φωτεινότητα κυτταροπλάσματος	Cytoycol
6	N shortest diameter	Μικρότερη διάμετρος πυρήνα	Kerneshort
7	N longest diameter	Μεγαλύτερη διάμετρος πυρήνα	Kernelong
8	N elongation	Επιμήκυνση πυρήνα	Kerneelong
9	N roundness	Σφαιρικότητα πυρήνα	Kernerund
10	C shortest diameter	Μικρότερη διάμετρος κυτταροπλάσματος	Cytoshort
11	C longest diameter	Μεγαλύτερη διάμετρος κυτταροπλάσματος	Cytolong
12	C elongation	Επιμήκυνση κυτταροπλάσματος	Cytoelong
13	C roundness	Σφαιρικότητα κυτταροπλάσματος	Cytorund
14	N perimeter	Περίμετρος πυρήνα	Kerneperi
15	C perimeter	Περίμετρος κυτταροπλάσματος	Cytoperi
16	N relative position	Θέση πυρήνα	Kernepos
17	Maxima in N	Μέγιστο πυρήνα	Kernemax
18	Minima in N	Ελάχιστο πυρήνα	Kernemin
19	Maxima in C	Μέγιστο κυτταροπλάσματος	Cytomax
20	Minima in C	Ελάχιστο κυτταροπλάσματος	Cytomin

Πίνακας 1.2: Χαρακτηριστικά των δεδομένων επιχρίσματος Παπανικολάου

Έτσι συλλέχθηκαν 20 χαρακτηριστικά που περιγράφουν τα κύτταρα και φαίνονται στον Πίνακα 1.2

Ακολουθούν επεξηγήσεις για τα χαρακτηριστικά του Πίνακα 1.2 :

1) Εμβαδόν του πυρήνα και 2) εμβαδόν κυτταροπλάσματος

Έχει υπολογιστεί μετρώντας τα αντίστοιχα pixels της τμηματοποιημένης εικόνας. Η περιοχή κάθε pixel ισούται με $(0.201\ \mu m)^2$ όπου $\mu m = 10^{-6}$ μέτρα. Μετριέται σε μm.

3) Αναλογία πυρήνα / κυτταροπλάσματος

Μας δείχνει πόσο μικρή είναι η περιοχή του πυρήνα σε σχέση με την περιοχή του κυτταροπλάσματος. Δίνεται από τον τύπο 1.1

4) Φωτεινότητα πυρήνα και 5) φωτεινότητα κυτταροπλάσματος

Η φωτεινότητα υπολογίζεται από τη μέση διακριτή φωτεινότητα, μέσω μιας συνάρτησης του μήκους κύματος των χρωμάτων. Στην συγκεκριμένη περίπτωση υπολογίζεται ως:

$$Y = 0.299 \cdot Red_{\mu} + 0.587 \cdot Green_{\mu} + 0.114 \cdot Blue_{\mu} .$$

Όπου Red_{μ}, $Green_{\mu}$ και $Blue_{\mu}$ είναι η μέση ένταση για καθένα από τα χρώματα. Είναι σταθμισμένα με την διακριτή φωτεινότητα του ανθρώπινου οφθαλμού.

6) Μικρότερη διάμετρος πυρήνα και 10) μικρότερη διάμετρος κυτταροπλάσματος

Αυτή είναι η μεγαλύτερη διάμετρος που μπορεί να έχει ένας κύκλος όταν είναι πλήρως εγγεγραμένος στο αντικείμενο. Όπως φαίνεται στο Διάγραμμα 1.4 η μικρότερη διάμετρος είναι το άθροισμα των S1 και S2. Μετριέται σε μm.

7) Μεγαλύτερη διάμετρος πυρήνα και 11) μεγαλυτερη διάμετρος κυτταροπλάσματος

Αυτή είναι η μικρότερη διάμετρος που μπορεί να έχει ένας κύκλος όταν περιγράφεται γύρω από ολόκληρο το αντικείμενο. Υπολογίζεται ως η μεγαλύτερη απόσταση μεταξύ των pixels στο όριο του αντικειμένου και είναι το L στο διάγραμμα 1.4.

8) Επιμήκυνση πυρήνα και 12) επιμήκυνση κυτταροπλάσματος

Η επιμήκυνση υπολογίζεται από τον λόγο της μικρότερης διαμέτρου προς τη μεγαλύτερη διάμετρο ενός αντικειμένου. Μετριέται σε μm.

9) Σφαιρικότητα του πυρήνα και 13) σφαιρικότητα του κυτταροπλάσματος

Η σφαιρικότητα υπολογίζεται από τον λόγο του πραγματικού εμβαδού προς το εμβαδόν του κύκλου με τη μεγαλύτερη διάμετρο που δίνεται από τον τύπο :

$$N_{εμβαδόν\ κύκλου} = \frac{\pi}{4} \cdot N^2_{μεγαλύτερη\ διάμετρος}\ \ άρα\ \ N_{σφαιρικότητα} = \frac{εμβαδόν}{εμβαδόν\ κύκλου}$$

14) Περίμετρος πυρήνα και 15) περίμετρος κυτταροπλάσματος

Το μήκος της περιμέτρου του πυρήνα και του κυτταροπλάσματος αντίστοιχα. Μετριέται σε μm.

16) Θέση του πυρήνα

Αυτό είναι ένα μέτρο του πόσο καλά είναι κεντραρισμένος ο πυρήνας μέσα στο κυτταρόπλασμα. Υπολογίζεται από την απόσταση μεταξύ του κέντρου του πυρήνα και του κέντρου του κυτταροπλάσματος.

$$θέση\ πυρήνα = \frac{\sqrt{(XN - XC)^2 + (YN - YC)^2}}{μεγαλύτερη\ διάμετρος\ κυτταροπλάσματος}$$

Εδώ η θέση του πυρήνα και του κυτταροπλάσματος δίνεται αντίστοιχα από τα σημεία (XN, YN) και (XC, YC). Μετριέται σε μm.

17) Μέγιστο πυρήνα και 19) μέγιστο κυτταροπλάσματος

Αυτή είναι μια μέτρηση του πόσα pixels είναι η μέγιστη τιμή στο εσωτερικό μιας ακτίνας 3 pixel. Μετριέται σε μm.

18) Ελάχιστο πυρήνα και 20) ελάχιστο κυτταροπλάσματος

Αυτή είναι μια μέτρηση του πόσα pixels είναι η ελάχιστη τιμή στο εσωτερικό μιας ακτίνας 3 pixel. Μετριέται σε μm.

1.5 Η νέα βάση δεδομένων επιχρίσματος Παπανικολάου

Η νέα βάση δεδομένων είναι η τελευταία εκ των δύο που δημιουργήθηκαν στο Πανεπιστημιακό Νοσοκομείο Herlev, στο τμήμα της Παθολογίας και στο τμήμα Αυτοματισμού του Τεχνικού

Πανεπιστημίου της Δανίας. Τα χαρακτηριστικά των κυττάρων που περιγράφονται είναι ίδια με αυτά της παλαιάς βάσης, αλλά οι κατηγορίες είναι ελαφρώς διαφορετικές. Στην νέα βάση δεδομένων τα χαρακτηριστικά αντλήθηκαν από τον Martin (2003) χρησιμοποιώντας το πρόγραμμα Matlab ως γλώσσα προγραμματισμού. Η βάση αποτελείται από 917 παρατηρήσεις η κατανομή των οποίων στις 7 κατηγορίες φαίνεται στον Πίνακα 1.3

Φυσιολογικά κύτταρα	**Αριθμός κυττάρων**	**Μη φυσιολογικά κύτταρα**	**Αριθμός κυττάρων**
Επιφανειακά λεπιδοειδή επιθήλια	74	**Ελαφριά λεπιδοειδής μη κερατινώδης δυσπλασία**	182
Ενδιάμεσα λεπιδοειδή επιθήλια	70	**Μεσαία λεπιδοειδής μη κερατινώδης δυσπλασία**	146
Κυλινδρικά επιθήλια	98	**Βαριά λεπιδοειδής μη κερατινώδη δυσπλασία**	197
		Ενδιάμεσο λεπιδοειδές κύτταρο με καρκίνωμα in – situ	150
Σύνολο	242		675

Πίνακας 1.3 : Τύποι κυττάρων επιχρίσματος Παπανικολάου.

ΚΕΦΑΛΑΙΟ 2

Αλγόριθμοι Πληροφορικής

Αυτό το κεφάλαιο έχει σκοπό να παραθέσει τους πιο γνωστούς τρόπους ταξινόμησης των pap – smear δεδομένων που χρησιμοποιούνται στην πληροφορική. Τα τελευταία 10 χρόνια πολλοί ερευνητές έχουν ασχοληθεί με τις 2 υπάρχουσες βάσεις δεδομένων, δημιουργώντας ταξινομητές, ικανούς να διαχωρίσουν τα κύτταρα σε 7 κατηγορίες ή έστω στην απλούστερη μορφή τους, στις 2 κατηγορίες, φυσιολογικών και μη φυσιολογικών κυττάρων. Πριν όμως από αυτό, αναφέρονται βασικοί τρόποι επιλογής χαρακτηριστικών καθώς και μέτρα υπολογισμού των σφαλμάτων.στις παραγράφους 2.2-2.6.

2.1 Νευρωνικά δίκτυα, ασαφής λογική και γενετικοί αλγόριθμοι

Για την καλύτερη κατανόηση των μεθόδων της πληροφορικής που περιγράφονται στο κεφάλαιο, παρατείθενται σε αυτήν την παράγραφο οι έννοιες των νευρωνικών δικτύων και της ασαφούς λογικής καθώς και η γενική λειτουργία των γενετικών αλγορίθμων αφού σε αυτές τις έννοιες βασίζονται οι συγκεκριμένοι αλγόριθμοι.

Τα τεχνητά νευρωνικά δίκτυα είναι προγράμματα για υπολογιστές που προσομοιώνουν τη βιολογική οργάνωση και τη λειτουργία των βιολογικών νευρώνων. Βασικό τους πλεονέκτημα είναι η ευπλαστότητα, όπως συμβαίνει με τα εγκεφαλικά μας κύτταρα, έτσι τα τεχνητά νευρωνικά δίκτυα δε χρειάζεται να επαναπρογραμματιστούν αν αλλάξει το περιβάλλον**.** Επιπλέον μπορούν να "μαθαίνουν" από μόνα τους αυτό που πρέπει να υπολογίσουν, χάρη σε ειδικά προγράμματα που σταδιακά διορθώνουν τα λάθη τους καθώς μεταβάλλεται η κατάσταση. Πρόκειται δηλαδή για την προσέγγιση της περιγραφής της λειτουργίας του νευρικού συστήματος μέσω μαθηματικών συναρτήσεων.

Οι αλγόριθμοι του κεφαλαίου αυτού χρησιμοποιούν την λειτουργία της ασαφούς λογικής. Με τον όρο αυτό εννοείται η αναπαράσταση ανακριβούς ή αδιευκρίνιστης γνώσης χρησιμοποιώντας την έννοια του βαθμού συμμετοχής και όχι τον απόλυτο διαχωρισμό αλήθειας-ψεύδους. Ένα κλασσικό (crisp) σύνολο Α ορίζεται μέσω της χαρακτηριστικής συνάρτησης

$$f_A(x):X\rightarrow(0,1) \qquad (2.1)$$

όπου

$$f_A(x) = \begin{cases} 1, \alpha\nu\ x\in A \\ 0, \alpha\nu\ x\notin A \end{cases} \qquad (2.2)$$

Ένα ασαφές σύνολο Α ορίζεται μέσω της συνάρτησης συμμετοχής (membership function)

$$\mu_A(x):X\rightarrow(0,1) \qquad (2.3)$$

όπου

$$\mu_{,A}(x) = \begin{cases} 1, \text{αν } x \text{ ολικά στο } A \\ 0, \text{αν } x \text{ καθόλου στο } A \\ (0,1) \text{αν } x \text{ μερικώς στο } A \end{cases} \qquad (2.4)$$

Το $\mu_A(x)$ είναι ένας πραγματικός αριθμός που παριστάνει τον βαθμό στον οποίο το x είναι στοιχείο του Α και ονομάζεται τιμή συμμετοχής (membersahip value).

Στις τελευταίες παραγράφους του κεφαλαίου αυτού παρουσιάζονται τρεις μέθοδοι βελτιστοποίησης που ανήκουν σε μια κατηγορία αλγορίθμων, τους γενετικούς αλγόριθμους. Ο τρόπος λειτουργίας των γενετικών αλγορίθμων είναι εμπνευσμένος απο την βιολογία. Χρησιμοποιεί την ιδέα της εξέλιξης μέσω γενετικής μετάλλαξης, φυσικής επιλογής και διασταύρωσης. Οι τιμές για τις παραμέτρους του συστήματος πρέπει να κωδικοποιούνται με τρόπο ώστε να αναπαρασταθούν από μια μεταβλητή που περιέχει σειρά χαρακτήρων ή δυαδικών ψηφίων (0/1). Αυτη η μεταβλητή μιμείται το γενετικό κώδικα που υπάρχει στους ζωντανούς οργανισμούς. Αρχικά, ο γενετικός αλγόριθμος παράγει πολλαπλά αντίγραφα της μεταβλητής/γεννητικού κώδικα, συνήθως με τυχαίες τιμές, δημιουργώντας έναν πληθυσμό λύσεων. Κάθε λύση (τιμές για τις παραμέτρους του συστήματος) δοκιμάζεται για το πόσο κοντά φέρνει την αντίδραση του συστήματος στην επιθυμητή, μέσω μιας συνάρτησης που δίνει το μέτρο ικανότητας της λύσης και η οποία ονομάζεται συνάρτηση ικανότητας (Σ.Ι). Οι λύσεις που βρίσκονται πιο κοντά στην επιθυμητή, σε σχέση με τις άλλες, σύμφωνα με το μέτρο που μας δίνει η Σ.Ι, αναπαράγονται στην επόμενη γενιά λύσεων και λάμβανουν μια τυχαία μετάλλαξη. Επαναλαμβάνοντας αυτή τη διαδικασία για αρκετές γενιές, οι τυχαίες μεταλλάξεις σε συνδυασμό με την επιβίωση και αναπαραγωγή των γονιδίων/λύσεων που πλησιάζουν καλύτερα το επιθυμητό αποτέλεσμα θα παράγουν ένα γονίδιο/λύση που θα περιέχει τις τιμές για τις παραμέτρους που ικανοποιούν όσο καλύτερα γίνεται την Σ.Ι.(en.wikipedia.org)

Οι γενετικοί αλγόριθμοι δεν επιλύουν το πρόβλημα με αναλυτικό/μαθηματικό τρόπο αλλά με βιολογικό. Συνεπώς έχουν μεγαλύτερη ενδογενή ευελιξία και ελευθερία να επιλέγουν μια επιθυμητή βέλτιστη λύση σύμφωνα με τις προδιαγραφές του προβλήματος. Ουσιαστικά οι γενετικό αλγόριθμοι

είναι αλγόριθμοι αναζήτησης (heuristics) που προσπαθούν να αναζητήσουν την λύση του προβλήματος που τους αναθέτουμε.

2.2 Μετρήσεις σφάλματος (error measurements)

Για την αξιολόγηση της απόδοσης των ταξινομητών ορίζονται 4 διαφορετικά μέτρα απόδοσης: ψευδώς αρνητικό σφάλμα FN (false – negative error), ψευδώς θετικό σφάλμα FP (false – positive error), συνολικό σφάλμα OE (overall error) και ρίζα του μέσου τετραγωνικού σφάλματος RMSE (root – mean – square error).

		Εκτιμώμενη τιμή (T)	
		-	+
Πραγματική τιμή (D)	-	TN	FP
	+	FN	TP

Πίνακας 2.1 : Ορισμός ψευδώς θετικών (FP) και ψευδώς αρνητικών (FN) σφαλμάτων

Στους ακόλουθους ορισμούς, ως θετικό (positive) αποτέλεσμα του τεστ εννοούμε μη φυσιολογικά κύτταρα, ενώ ως αρνητικό (negative) αποτέλεσμα του τεστ εννοούμε φυσιολογικά κύτταρα. Έτσι ορίζουμε :

$$FN\% = P\left(T^{-} \mid D^{+}\right) \times 100\% \; = \frac{FN}{TP+FN} \times 100\% \tag{2.5}$$

$$FP\% = P\left(T^{+} \mid D^{-}\right) \times 100\% \; = \frac{FP}{TN+FP} \times 100\% \tag{2.6}$$

Οι ποσότητες FN και FP δείχνουν τον αριθμό των κυττάρων που έχουν λανθασμένα ταξινομηθεί ως αρνητικά και θετικά κύτταρα αντίστοιχα. Οι ποσότητες TN και TP δείχνουν τον αριθμό των κυττάρων που έχουν ταξινομηθεί σωστά. Το σφάλμα FN είναι πιο καταστροφικό από το σφάλμα FP επειδή οι ασθενείς με προκαρκινικές ενδείξεις, λανθασμένα ταξινομούνται ως υγιείς. Η εκτίμηση του συνολικού σφάλματος περιγράφεται στον τύπο 2.7 ως ο αριθμός των σφαλμάτων σε σχέση με όλα τα κύτταρα.

$$OE\% = \frac{FN+FP}{TP+FN+TN+FP} \times 100\% \quad (2.7)$$

Σαν εναλλακτική των σφαλμάτων ταξινόμησης που περιγράφησαν, ορίζεται η ρίζα του μέσου τετραγωνικού σφάλματος (RMSE).

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(y_i - T_i)^2}{N}} \quad (2.8)$$

Το μέσο τετραγωνικό σφάλμα RMSE δείχνει τη μέση απόσταση ανάμεσα στο αποτέλεσμα του μοντέλου ταξινόμησης y_i και στην πραγματική κατηγορία T_i που ανήκει κάθε μία από τις N παρατηρήσεις. Για το πρόβλημα των 2 κατηγοριών, το μέσο τετραγωνικό σφάλμα ισούται με τη ρίζα του συνολικού σφάλματος, δηλαδή :

$$RMSE = \sqrt{OE}$$

2.3 Επαναξιολόγηση κ δειγμάτων (k – fold cross – validation)

Η δημιουργία και ο έλεγχος των ταξινομητών, απαιτεί την διαίρεση των δεδομένων σε ξεχωριστά σετ δεδομένων εκμάθησης (training data) και δεδομένων ελέγχου (test data). Όμως, η βέλτιστη αναλογία του αριθμού των δεδομένων εκμάθησης και των δεδομένων ελέγχου εξαρτάται από τη μέθοδο.

Για να χρησιμοποιηθούν στο έπακρο τα διαθέσιμα δεδομένα γίνεται επαναξιολόγηση. Η μέθοδος αυτή περιγράφεται από τους Wabba & Wold στο Bishop (1995). Ολόκληρο το σετ δεδομένων N παρατηρήσεων χωρίζεται τυχαία σε κ υποδείγματα. Έπειτα, χρησιμοποιούνται κ -1 δείγματα ως δεδομένα εκμάθησης, για να δημιουργηθεί το μοντέλο και να εκτιμηθούν οι παράμετροί του. Η απόδοση του ταξινομητή υπολογίζεται μέσω του δείγματος που δεν συμπεριλήφθηκε στην αρχική ανάλυση. Η διαδικασία αυτή μπορεί να επαναληφθεί κ φορές επειδή υπάρχουν κ διαφορετικές δυνατότητες αποκλεισμού ενός από τα κ δείγματα. Το συνολικό σφάλμα του ελέγχου δίδεται από τη μέση τιμή των κ μοντέλων που κατασκευάστηκαν.

2.4 Επανάληψη της μεθόδου επαναξιολόγησης (cross – validation rerunning)

Στην μέθοδο επαναξιολόγησης κ δειγμάτων, το δείγμα χωρίζεται σε κ υποδείγματα αλλά η επιλογή των παρατηρήσεων που θα ανήκουν σε καθένα από τα υποδείγματα είναι τυχαία. Κι όταν τα δεδομένα εκμάθησης και ελέγχου επιλέγονται τυχαία και χρησιμοποιούνται για την δημιουργία και τον έλεγχο του μοντέλου, η εκτίμηση των σφαλμάτων μπορεί να διαφέρει. Για αυτό το λόγο, η

επαναξιολόγηση κ δειγμάτων επαναλαμβάνεται R φορές και υπολογίζεται το μέσο σφάλμα των R αυτών επαναλήψεων, ως μια πιο αξιόπιστη εκτίμηση σφάλματος.

2.5 Προσομοιωμένη ανόπτηση (simulated annealing)

Η προσομοιωμένη ανόπτηση είναι μια στοχαστική υπολογιστική τεχνική η οποία αναπτύχθηκε για την επίλυση προβλημάτων βελτιστοποίησης και ο στόχος της είναι η ελαχιστοποίηση μιας συνάρτησης κόστους, η οποία καλείται και ενέργεια. Η βασική ιδέα της προσομοιωμένης ανόπτησης προήλθε από τη θερμοδυναμική, όπου η κατάσταση χαμηλής ενέργειας ενός μετάλλου παράγεται τήκοντάς το και στη συνέχεια μειώνοντας τη θερμοκρασία. Αν το μέταλλο ψυχθεί επαρκώς αργά, τα άτομά του σχηματίζουν έναν καθαρό κρύσταλλο, που αποτελεί τη δομή με τη χαμηλότερη ενεργειακή στάθμη, που για ένα μαθηματικό πρόβλημα βελτιστοποίησης αντιστοιχεί στη βέλτιστη λύση. Αντίθετα, αν το μέταλλο ψυχθεί γρήγορα, φτάνει σε μια κατάσταση με μεγαλύτερη ενέργεια από την ελάχιστη, κάτι που αντιστοιχεί σε μια υπο-βέλτιστη λύση ενός μαθηματικού προβλήματος

Το πιο κρίσιμο κομμάτι της προσομοιωμένης ανόπτησης είναι αυτό που καθορίζει πόσο γρήγορα θα πέσει η «θερμοκρασία» από ψηλές σε χαμηλές τιμές. Αυτό συνήθως χρειάζεται αρκετό πειραματισμό καθότι εξαρτάται από τη φύση του προβλήματος. Η σημασία μάλιστα αυτής της διαδικασίας πιστοποιείται και από το γεγονός ότι μπορεί να αποδειχθεί ότι αν η θερμοκρασία T μειώνεται επαρκώς αργά, ο αλγόριθμος θα βρει το ολικό βέλτιστο με πιθανότητα που συγκλίνει στο 1 (A.W. Johnsona, S.H. Jacobson). Το πρόβλημα όμως είναι ότι μπορεί να χρειαστεί χρόνος μεγαλύτερος ακόμα κι από αυτόν που θα χρειαζόταν ένας εξαντλητικός αλγόριθμος. Επομένως, αν και δεν είναι ιδιαίτερα πρακτικό το να περιμένουμε την εύρεση της βέλτιστης λύσης, οι σχεδόν-βέλτιστες λύσεις που δίνει ο αλγόριθμος αρκετά γρήγορα είναι συνήθως ικανοποιητικές, ενώ παράλληλα η αποδεδειγμένη σύγκλιση αποτελεί ένα από τα πλεονεκτήματα της προσομοιωμένης ανόπτησης σε σχέση με άλλους αλγόριθμους. (Παναγιώτης Γεωργαλλής 2006)

Ένας τρόπος επιλογής των μεταβλητών που χρησιμοποιούνται για περαιτέρω ανάλυση στην παρούσα διπλωματική εργασία είναι η χρήση του αλγορίθμου προσομοιωμένης ανόπτησης (Jang J.S.R. 1997).Ο αλγόριθμος ξεκινά με μια τυχαία τιμή $\mathbf{x_1}$. Ύστερα δημιουργεί μια μικρή μετατόπιση στο $\mathbf{x_1}$, την $\Delta\mathbf{x_1}$ την οποία και προσθέτει στο αρχικό σημείο. Έτσι δημιουργείται ένα νέο σημείο $\mathbf{x}_{new} = \mathbf{x_1} + \Delta\mathbf{x_1}$. Το νέο αυτό σημείο αποθηκεύεται μαζί με μια πιθανότητα υπολογισμένη από την συνάρτηση πυκνότητας πιθανότητας Boltzmann ($P \propto e^{-\frac{E}{kT}}$ όπου E η ενέργεια του συστήματος, T η «θερμοκρασία» και k η σταθερά Boltzmann). Η διαδικασία αυτή επαναλαμβάνεται αρκετές φορές μέχρι να βρεθεί η βέλτιστη λύση. Στις μεθόδους που ακολουθούν σε αυτήν τη διπλωματική εργασία, χρησιμοποιείται μία

απλουστευμένη εκδοχή του αλγόριθμου για το λόγο ότι είναι πιο γρήγορος στην εφαρμογή του. Ο απλουστευμένος αλγόριθμος προσομοιωμένης ανόπτησης αποθηκεύει την νέα τιμή αν αυτή είναι καλύτερη από την προηγούμενη ($f(\mathbf{x}_{new}) < f(\mathbf{x_1})$). Έτσι, παραλείπεται η διαδικασία εύρεσης πιθανοτήτων για κάθε νέα τιμή. Ο αλγόριθμος έχει 6 βήματα:

1.Διάλεξε ένα τυχαίο αρχικό σημείο $\mathbf{x_1}$. Όρισε τον μετρητή επαναλήψεων k=1.

2.Εκτίμησε την αντικειμενική συνάρτηση:

$$E = f(\mathbf{x_1})$$

3.Όρισε $\mathbf{x}_{new} = \mathbf{x_1} + \Delta\mathbf{x_1}$ όπου $\Delta\mathbf{x_1}$ μια μικρή μετατόπιση του $\mathbf{x_1}$.

4.Υπολόγισε την νέα τιμή της αντικειμενικής συνάρτησης:

$$E_{new} = f(\mathbf{x}_{new})$$

5. Αν $E_{new} < E$ θέσε $\mathbf{x_1}$ στο $\mathbf{x}_{new}$ και E στο E_{new}

6. Αύξησε τον αριθμό επαναλήψεων k και επανέλαβε. Αν το k φτάσει στην μέγιστη τιμή του K τότε ο αλγόριθμος σταματά. Αλλιώς, πήγαινε στο βήμα 3.

2.6 Επιτηρούμενη δημιουργία συστάδων (supervised clustering)

Αυτή η μέθοδος βοηθάει στην εύρεση κέντρων των συστάδων όταν οι συστάδες δεν είναι εύκολα διαχωρίσιμες μεταξύ τους λόγω αλληλοεπικάλυψης των κατηγοριών. Η επιτηρούμενη δημιουργία συστάδων ορίζεται σαν μια διαδικασία στην οποία χρησιμοποιείται η διεγνωσμένη κατηγορία των δεδομένων εκμάθησης στην διαδικασία δημιουργίας συστάδων (Duda et al.2000).

Για παράδειγμα, οι μέθοδοι δημιουργίας συστάδων C – means και Gustafson – kessel συνήθως δουλεύουν με την εύρεση φυσικών συστάδων στα δεδομένα. Οι αλγόριθμοι όμως, δεν κάνουν χρήση της διεγνωσμένης κατηγορίας που ανήκουν οι παρατηρήσεις των δεδομένων εκμάθησης. Οι συστάδες υπολογίζονται υποθέτοντας ότι είναι φυσικά διαχωρίσιμες. Αν παρατηρείται αλληλοεπικάλυψη, τότε δεν βρίσκονται απαραίτητα συστάδες με καλά κέντρα κι έτσι οδηγούμαστε σε φτωχή ταξινόμηση. Η προτεινόμενη διαδικασία θα μπορούσε να βρει καλύτερα κέντρα, χρησιμοποιώντας τη διεγνωσμένη κατηγορία που ανήκουν οι παρατηρήσεις των δεδομένων εκμάθησης.

Η συγκεκριμένη διαδικασία δεν αλλάζει τον τρόπο που δουλεύουν οι αλγόριθμοι. Είναι ένας απλός τρόπος που χρησιμοποιεί τους αλγόριθμους και τα αποτελέσματα τους για να βρει καλύτερα κέντρα συστάδων. Η διαδικασία έχει ως εξής:

1. Τα δεδομένα εκμάθησης χωρίζονται σύμφωνα με την διεγνωσμένη κατηγορία που ανήκει κάθε παρατήρηση.
2. Βρίσκονται συστάδες για κάθε διεγνωσμένη κατηγορία ξεχωριστά, σύμφωνα με τον αλγόριθμο δημιουργίας συστάδων που έχει επιλεχθεί.

3. Τα κέντρα των συστάδων που έχουν βρεθεί για κάθε διεγνωσμένη κατηγορία συγχωνεύονται και φτιάχνουν ένα μοντέλο.

Η επιτηρούμενη δημιουργία συστάδων είναι εφαρμόσιμη μόνο όταν η κατηγορία των παρατηρήσεων των δεδομένων εκμάθησης είναι γνωστή εξ αρχής.

2.7 Ομαδοποίηση με την μέθοδο των C-μέσων (C – means clustering) (J.MacQuenn 1967)

Ο αλγόριθμος C – means είναι σχεδιασμένος για να διαχωρίζει τα δεδομένα σε συστάδες λαμβάνοντας υπόψη την συνεισφορά (membership) της κάθε παρατήρησης στην δημιουργία της συστάδας . Ο αλγόριθμος απαιτεί την εκτίμηση λίγων παραμέτρων και γι'αυτό είναι μια προτιμητέα μέθοδος. Επιπλέον, τα δεδομένα δεν έχουν περιορισμό διαστάσεων, δηλαδή μπορούν να χρησιμοποιηθούν όσα χαρακτηριστικά είναι διαθέσιμα. Υπάρχουν 2 είδη αλγόριθμων C – μέσων : ο αυστηρά(hard) C – μέσων και ο ασαφής(fuzzy) C – μέσων.

2.7.1 Αυστηρά C – μέσων (Hard C-means-HCM)

Ο αλγόριθμος αυστηρά C – μέσων (HCM) ξεκινά με τυχαία επιλογή των θέσεων των κέντρων για όλες τις συστάδες. Έπειτα, σε όλα τα δεδομένα προσδίδεται το κοντινότερο κέντρο συστάδας με μια τιμή συμμετοχής (membership value) που ισούται με ένα για αυτό το κέντρο και με μηδέν για όλα τα υπόλοιπα. Χρησιμοποιείται η Ευκλείδια απόσταση. Βρίσκονται καινούρια κέντρα των συστάδων με τον υπολογισμό της μέσης τιμής της θέσης των παρατηρήσεων που ανήκουν σε κάθε συστάδα, όπως φαίνεται στον τύπο 2.5

$$\boldsymbol{c_i} = \frac{1}{\mathrm{C}} \cdot \sum_{k,u_k \in C} u_k \qquad (2.9)$$

Στον τύπο 2.9 ως $\boldsymbol{c_i}$ δίδεται το κέντρο της συστάδας i, με C συμβολίζεται το σύνολο όλων των συστάδων και με u_k η συνεισφορά της παρατήρησης k. Έπειτα, τα δεδομένα καταχωρούνται ξανά στο κοντινότερο κέντρο συστάδας και καινούρια κέντρα συστάδων υπολογίζονται. Αυτή η διαδικασία είναι επαναληπτική, μέχρι μια αντικειμενική συνάρτηση να βρεθεί κάτω από ένα όριο που ονομάζεται κατώφλι ή ο αριθμός των επαναλήψεων να υπερβεί ένα συγκεκριμένο αριθμό. Η αντικειμενική συνάρτηση, δίνεται συνήθως από το άθροισμα των αποστάσεων όλων των κέντρων των συστάδων από τις αντίστοιχες παρατηρήσεις, όπως φαίνεται στον τύπο 2.10

$$J = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} (\sum_{k,u_k \in C} \|u_k - c_i\|) \quad (2.10)$$

Στον τύπο 2.10 το J συμβολίζει την αντικειμενική τιμή και το c συμβολίζει τον αριθμό των συστάδων. Αυτή η αντικειμενική συνάρτηση μειώνεται καθώς βρίσκουμε τα κέντρα των συστάδων επαναληπτικά. Η τελική τιμή εξαρτάται από την αρχική τοποθέτηση των κέντρων. Κάθε φορά που υπολογίζονται καινούρια κέντρα συστάδων, μια παρατήρηση μπορεί να καταχωρηθεί σε άλλη συστάδα. Αυτός είναι ένας από τους λόγους που ο αλγόριθμος HCM δεν εντοπίζει πάντα τα βέλτιστα κέντρα συστάδων. Αυτό το πρόβλημα το χειρίζεται καλύτερα ο αλγόριθμος ασαφής C – μέσων..

2.7.2 Ασαφής C – μέσων (fuzzy C-means-FCM) (Bezdek 1981)

Ο αλγόριθμος FCM τροποποιεί τον αλγόριθμο HCM με το να επιτρέπει στα δεδομένα να ανήκουν σε όλες τις συστάδες, με συνεισφορές που παίρνουν τιμές στο διάστημα [0,1]. Το άθροισμα των συμμετοχών κάθε παρατήρησης πρέπει να ισούται με ένα. Οι συμμετοχές υπολογίζονται από τον τύπο 2.11

$$m_{ik} = [\sum_{j=1}^{c} (\frac{d_{ik}}{d_{jk}})^{\frac{2}{q-1}}]^{-1} \quad (2.11)$$

Στον τύπο 2.11 το m_{ik} συμβολίζει την συνεισφορά της παρατήρησης k στο κέντρο της συστάδας i, το K είναι ο αριθμός των παρατηρήσεων, το d_{jk} είναι η απόσταση του κέντρου συστάδας j από την παρατήρηση k. Το q∈[1,∞] και είναι ο εκθέτης που καθορίζει πόσο ξερή (crisp) πρέπει να είναι η συμμετοχή. Αν q≈1 τότε η συνεισφορά είναι ξερή, ενώ όταν το q→ ∞ τότε είναι ίση για όλες τις συστάδες ασχέτως από τη θέση τους. Καινούρια κέντρα συστάδων υπολογίζονται με τις ασαφείς (fuzzy) συμμετοχές σύμφωνα με τον τύπο 2.12

$$c_i = \frac{\sum_{k=1}^{K} m_{ik}^{q} u_k}{\sum_{k=1}^{K} m_{ik}^{q}} \quad (2.12)$$

Ο αλγόριθμος FCM εξαρτάται πολύ λιγότερο από την αρχική θέση των κέντρων των συστάδων, αν έχει επιλεχθεί μια κατάλληλη τιμή για το q. Αυτό συμβαίνει επειδή η θέση των συστάδων καθορίζεται από όλα τα δεδομένα αλλά δέχεται και μια επιρροή που καθορίζεται από την ασαφή συμμετοχή τους. Για αυτό, τα δεδομένα δεν θα αλλάξουν απότομα συστάδα. Παρόλα αυτά, η τυχαία αρχική θέση των κέντρων των συστάδων, μπορεί να επηρεάσει τις τελικές θέσεις των κέντρων των συστάδων.

Η αντικειμενική τιμή του πόσο καλά τοποθετημένα είναι τα κέντρα των συστάδων βρίσκεται από τον τύπο 2.13

$$J = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{k=1}^{K} m_{ik}^{q} \, d_{ik}^{2} \quad (2.13)$$

Στον τύπο 2.9 το J_i συμβολίζει την αντικειμενική τιμή για την συστάδα i. Ο αλγόριθμος FCM εκτιμάει τα κέντρα των συστάδων επαναληπτικά μέχρι η αντικειμενική συνάρτηση να βρεθεί κάτω από ένα κατώφλι ή μέχρι ο αριθμός των επαναλήψεων να υπερβεί ένα όριο, όπως και στον αλγόριθμο HCM.

Παράμετροι αλγορίθμων C –μέσων

Ο αλγόριθμος FCM, όπως φαίνεται στους τύπους (2.11)-(2.13) περιλαμβάνει την παράμετρο q που αποφασίζει την ξηρότητα (crispness) των συμμετοχών. Όταν q≈1, ο αλγόριθμος FCM μοιάζει με τον αλγόριθμο HCM. Ο αλγόριθμος HCM μπορεί να έχει τόσες συστάδες όσες και παρατηρήσεις. Αυτό το όριο δεν υπάρχει στον αλγόριθμο FCM. Επίσης, κατά τον υπολογισμό των συμμετοχών, στον μεν αλγόριθμο HCM, μια παρατήρηση μπορεί να ανήκει μόνο σε μια συστάδα, στο δε αλγόριθμο FCM, υποδεικνύονται παραπάνω από μία συστάδες ως πιθανές.

Και οι δύο αλγόριθμοι πρέπει να αποφασίσουν εξ αρχής πόσες ομάδες θα χρησιμοποιήσουν. Αυτή είναι μια σημαντική παράμετρος γιατί πάρα πολλές συστάδες μπορεί να έχουν ως αποτέλεσμα την υπερπροσαρμογή του μοντέλου στον θόρυβο των δεδομένων. Η εκμάθηση των κέντρων των συστάδων σταματά όταν η αντικειμενική συνάρτηση βρεθεί κάτω από το κατώφλι, ή όταν ξεπεραστεί το όριο των επαναλήψεων με τα οποία ο αλγόριθμος τερματίζεται. Αυτά πρέπει να επιλεχθούν εξαρχής κι η επιλογή τους να είναι τέτοια έτσι ώστε τα κέντρα των συστάδων να συγκλίνουν σε μοναδική λύση.

2.8 Δημιουργία συστάδων Gustafson – Kessel (Gustafson – Kessel clustering) (Gustafson – Kessel 1979)

Αυτή η παράγραφος ασχολείται με την περιγραφή του αλγόριθμου Gustafson – Kessel (GK), ο οποίος έχει εφαρμοστεί στα δεδομένα επιχρίσματος Παπανικολάου (pap-smear). Ο αλγόριθμος FCM παράγει σφαιρικές συστάδες και δεν επιτρέπει σε μια συστάδα να αλλάξει σχήμα ανάλογα με τα δεδομένα. Η μέθοδος GK δίνει την δυνατότητα σε μια συστάδα να πάρει υπερελλειψοειδές σχήμα.

Ο αλγόριθμος GK μοιάζει αρκετά με τον αλγόριθμο FCM. Η μόνη διαφορά είναι στον τρόπο που υπολογίζονται οι αποστάσεις. Ο FCM χρησιμοποιεί την Ευκλείδεια απόσταση ενώ ο GK την mahalanobis όπως φαίνεται στον τύπο 2.10.

$$D_{ik}^{2} = (x_k - c_i) A_i (x_k - c_i)^{T} \quad (2.14)$$

Στον τύπο 2.14, D_{ik} συμβολίζει την απόσταση, c_i το κέντρο της συστάδας i και x_k την παρατήρηση k. Το A_i είναι ο πίνακας των αποστάσεων mahalanobis. Αν ο πίνακας A_i είναι μοναδιαίος τότε χρησιμοποιείται η Ευκλείδεια απόσταση. Ο πίνακας A_i συνήθως ορίζεται από τον τύπο 2.15

$$A_i = p_i \cdot det(F_i)^{(1/N)} F_i^{-1} \tag{2.15}$$

όπου, N είναι ο αριθμός των χαρακτηριστικών, p_i ο βαθμός που θα έχει η συστάδα i και F_i^{-1} ο αντίστροφος του πίνακα F_i. Ο F_i είναι ο πίνακας διακυμάνσεων-συνδιακυμάνσεων της συστάδας i και υπολογίζεται από τον τύπο (2.16)

$$F_i = \frac{\sum_{k=1}^{K} (m_{ik})^q (x_k - c_i)^T (x_c - c_i)}{\sum_{k=1}^{K} (m_{ik})^q} \tag{2.16}$$

όπου, K συμβολίζει τον αριθμό των ομάδων, q τον επιλεγμένο ασαφή εκθέτη και m_{ik} τη συνεισφορά όπως αυτή ορίστηκε στη σχέση (2.15)

Η αντικειμενική συνάρτηση του GK έχει σκοπό να ελαχιστοποιήσει τις αποστάσεις των κέντρων των συστάδων από τα δεδομένα. Οι αποστάσεις δίνονται από την (2.14) και ελέγχονται από τον πίνακα A_i

Τα ιδιοδιανύσματα και οι ιδιοτιμές του A_i καθορίζουν το σχήμα, την θέση και το βαθμό της συστάδας i. Οι υπερελλειψοειδείς συστάδες σχηματίζονται από διανύσματα με φορά προς την κατεύθυνση των ιδιοδιανυσμάτων και έχουν μήκη που δίνονται από την ν-οστή ρίζα των αντίστοιχων ιδιοτιμών. Από τη στιγμή που όλα τα ιδιοδιανύσματα είναι κάθετα μεταξύ τους, μια προσεγγιστική έκφραση για τον βαθμό της συστάδας θα μπορούσε να είναι το γινόμενο των Νιοστών ριζών των ιδιοτιμών της.

2.9 Μέθοδος ελαχίστων τετραγώνων για ευθεία παλινδρόμησης (Least Square Method)

Με τη μέθοδο των ελαχίστων τετραγώνων μπορούμε εύκολα να προσαρμόσουμε μια ευθεία $Y = \alpha \cdot x + \beta$ (ή γενικότερα ένα πολυώνυμο m βαθμού) στα πειραματικά δεδομένα (x_1, y_1), (x_2, y_2)..., (x_n, y_n), έτσι ώστε να ελαχιστοποιηθεί το άθροισμα των τετραγώνων των αποκλίσεων (sum of squared residuals) S:

$$S = \Sigma[Y(x_i) - y_i]^2 \tag{2.17}$$

Λαμβάνοντας τις μερικές παραγώγους του S ως προς τα α, β και εξισώνοντας αυτές με το μηδέν, προκύπτει το ακόλουθο σύστημα 2-εξισώσεων και 2-αγνώστων (α, β):

$$\Sigma(y_i) = \alpha \cdot \Sigma(x_i) + N \cdot \beta \quad (2.18)$$

$$\Sigma(x_i \cdot y_i) = \alpha \cdot \Sigma(x_i^2) + \beta \cdot \Sigma(x_i) \quad (2.19)$$

όπου N είναι ο αριθμός των παρατηρήσεων. Το σύστημα αυτό είναι γνωστό ως σύστημα κανονικών εξισώσεων και οι ζητούμενοι συντελεστές α και β αποτελούν τη μοναδική λύση αυτού του συστήματος.Με την επίλυση του συστήματος προκύπτουν οι εξισώσεις

$$\alpha = [N \cdot \Sigma(x_i \cdot y_i) - \Sigma(x_i) \cdot \Sigma(y_i)] / [N \cdot \Sigma(x_i^2) - \Sigma(x_i) \cdot \Sigma(x_i)] \quad (2.20)$$

$$\beta = [\Sigma(y_i) \cdot \Sigma(x_i^2) - \Sigma(x_i) \cdot \Sigma(y_i \cdot x)] / [N \cdot \Sigma(x_i^2) - \Sigma(x_i) \cdot \Sigma(x_i)] \quad (2.21)$$

Παρατήρηση: Η μέθοδος ελαχίστων τετραγώνων δεν μπορεί να εφαρμοστεί εφόσον η μεταβλητή y είναι κατηγορική . Παρόλ'αυτά στην πληροφορική έχει εφαρμοστεί ο αλγόριθμος και τα αποτελέσματά του θα παρατεθούν στην εργασία αυτή.

2.10 Αλγόριθμοι Κ εγγύτερου γείτονα (KNN και WKNN)

Οι μέθοδοι που περιγράφονται σε αυτή την παράγραφο είναι 2 απλοί μέθοδοι ταξινόμησης. Έχοντας ένα σύνολο δεδομένων εκμάθησης με γνωστή την κατηγορία στην οποία ανήκει κάθε παρατήρηση, ο αλγόριθμος KNN επιλέγει τα K κοντινότερα σημεία από ένα σημείο ελέγχου. Το σημείο ελέγχου μπαίνει στην κατηγορία που απαντάται πιο συχνά ανάμεσα σ' αυτά τα K σημεία.

Ο αλγόριθμος WKNN είναι ίδιος με τον KNN με μόνη διαφορά ότι η κατηγορία των K κοντινότερων παρατηρήσεων είναι σταθμισμένη. Η στάθμιση της κάθε παρατήρησης είναι αντιστρόφως ανάλογη της απόστασης μεταξύ των παρατηρήσεων εκμάθησης και ελέγχου.

Για την επιλογή των K κοντινότερων σημείων από τα δεδομένα εκμάθησης χρησιμοποιείται η κανονικοποιημένη Ευκλείδεια απόσταση (Normalized Euclidean Distance) d_i που είναι η απόσταση μεταξύ μιας παρατήρησης εκμάθησης $\boldsymbol{x}_i = \{x_{i1}, x_{i2}, \dots, x_{iP}\}$ και του σημείου ελέγχου $\boldsymbol{x}_q = \{x_{q1}, x_{q2}, \dots, x_{qP}\}$, όπου P είναι ο αριθμός των χαρακτηριστικών που περιγράφουν το δείγμα.

$$d_i = \|\boldsymbol{x}_q - \boldsymbol{x}_i\| = \left(\sum_{j=1}^{P} \frac{|x_{qj} - x_{ij}|^2}{P}\right)^{\frac{1}{2}} \quad (2.22)$$

2.11 Κέντρο βαρύτητας κοντινότερης ομάδας (Nearest Class Gravity Center-NCC)

Η σύνθεση των δεδομένων εκμάθησης είναι πολύ σημαντική για την απόδοση των μεθόδων WKNN και KNN. Μιας και δεν εκτελείται δημιουργία συστάδων για τις K επιλεγμένες παρατηρήσεις εκμάθησης, το μεγάλο μέγεθος μιας ομάδας θα την ευνοεί στο αποτέλεσμα του μοντέλου. Για τη βελτίωση αυτού του σφάλματος, χρησιμοποιείται ο παρακάτω NCC αλγόριθμος που περιγράφεται σε 6 βήματα.

1. Υπολογισμός όλων των αποστάσεων $D_{global} = \{d_1, d_2, \dots, d_N\}$ των σημείων ελέγχου $\boldsymbol{x}_q$ από όλα τα σημεία εκμάθησης $\boldsymbol{x}_i$ για i=1, 2, …, N

$$d_i = \|\boldsymbol{x}_q - \boldsymbol{x}_i\| = \{\sum_{j=1}^{P} \frac{|x_{qj} - x_{ij}|^2}{P}\}^{\frac{1}{2}}$$

2. Τοποθέτηση όλων των αποστάσεων του D σε αύξουσα σειρά. Επιλογή των K κοντινότερων γειτόνων. Τα S και S_k περιέχουν και τα σημεία εκμάθησης $\boldsymbol{x}_i$ για να μπορούν να υπολογιστούν τα κέντρα βαρύτητας (centers of gravity) στο βήμα 3.

$$S = \begin{matrix} d_1 & C_1 & \boldsymbol{x}_1 \\ \vdots & \vdots & \vdots \\ d_K & C_K & \boldsymbol{x}_K \\ \vdots & \vdots & \vdots \\ d_N & C_N & \boldsymbol{x}_N \end{matrix}, d_1 < d_2 < \cdots < d_K < \cdots < d_N, S_K = \begin{matrix} d_1 & C_1 & \boldsymbol{x}_1 \\ \vdots & \vdots & \vdots \\ d_K & C_K & \boldsymbol{x}_K \end{matrix}$$

3. Υπολογισμός των κέντρων των κατηγοριών $\boldsymbol{\Phi}_l = \{\Phi_1\ \Phi_2 \dots \Phi_P\}$ για τις κατηγορίες l = 1, 2, …, M των K κοντινότερων γειτόνων. Ο αριθμός των σημείων εκμάθησης με κατηγορία C_i είναι K_l και $\sum_{L=1}^{M} K_l = K$

$$\Phi_{lj} = \frac{1}{K_l} \sum_{i=1}^{K_l} x_{ij} \ \ για\ C_i = l \quad i = 1, 2, \dots, K$$

4. Υπολογισμός αποστάσεων $\sigma = \{\sigma_1, \sigma_2, \dots, \sigma_M\}$ του σημείου ελέγχου x_q από τα M κέντρα βαρύτητας.

$$\sigma_l = \|\boldsymbol{x}_q - \boldsymbol{\Phi}_i\| = \{\sum_{j=1}^{P} \frac{|x_{qj} - \boldsymbol{\Phi}_{ij}|^2}{P}\}^{\frac{1}{2}}$$

5. Τοποθέτηση των αποστάσεων $\sigma = \{\sigma_1, \sigma_2, \dots, \sigma_M\}$ σε αύξουσα σειρά.

$$S_\sigma = \begin{matrix} \sigma_1 \\ \vdots \\ \sigma_M \end{matrix} \begin{matrix} C_1 \\ \vdots \\ C_M \end{matrix}, \qquad \sigma_1 < \sigma_2 < \cdots < \sigma_M$$

5. Επιλογή των κοντινότερων κέντρων ως προς το S(1) και τοποθέτηση της κατηγορίας C_q του αποτελέσματος στην κατηγορία του κέντρου αυτού.

2.12 Νευροασαφής μέθοδος εξαγωγής συμπεράσματος (Neurofuzzy Inference Method- NFI)

Ο αλγόριθμος NFI είναι τοπικό νευρο-ασαφές (neural-fuzzy) σύστημα εξαγωγής συμπερασμάτων που χρησιμοποιεί έναν αλγόριθμο ανάστροφης μετάδοσης (back-propagation-BP) για την βελτιστοποίηση των παραμέτρων του. Ο μηχανισμός που χρησιμοποιείται στην συγκεκριμένη αναφορά είναι τύπου Takagi-Sugeno. Για να διατηρηθούν οι συμβολισμοί που χρησιμοποιήθηκαν από τους Kasabov&Song , στην παράγραφο αυτή, οι K κοντινότεροι γείτονες θα αναγράφονται ως N_q. Για τον διαχωρισμό των υπόχωρων που θα χρησιμοποιηθούν έπειτα για την δημιουργία των ασαφών κανόνων, χρησιμοποιήθηκε ο αλγόριθμος ECM (Evolving Clustering Method). Ο αλγόριθμος αυτός διατηρεί μικρό αριθμό συστάδων μέσα στους υπόχωρους (Kasabov&Song 2002). Τα κέντρα και οι ακτίνες των κατηγοριών που βρίσκονται από τον ECM χρησιμοποιούνται μετέπειτα για την δημιουργία κανόνων των συναρτήσεων ασαφών συνεισφορών. Ακολουθεί ο αλγόριθμος NFI σε 7 βήματα.

1. Υπολογισμός των αποστάσεων d_i , $i = 1,2, \dots, N$ του διανύσματος των δεδομένων $\boldsymbol{x}_q$ από καθένα από τα N δεδομένα εκμάθησης σε όλο το σύνολο των δεδομένων εκμάθησης.

2. Επιλογή των N_q κοντινότερων παρατηρήσεων εκμάθησης, με την μικρότερη απόσταση d_i,με σκοπό τον σχηματισμό του υποσυνόλου D_q . Η τιμή του N_q μπορεί να προκαθοριστεί με βάση την εμπειρία ή να βρεθεί με διαδικασία βελτιστοποίησης.

3. Υπολογισμός των βαρών $\boldsymbol{w_i}$ των N_q κοντινότερων παρατηρήσεων $w_i = 1 - (d_i - \min(\boldsymbol{d})),\ i = 1,2,\dots,N_q$,min $(\boldsymbol{d})$ είναι η ελάχιστη από όλες τις αποστάσεις $\boldsymbol{d} = [d_1, d_2, \dots, d_{N_q}]$ στον υπόχωρο D_q .

4. Χρήση του αλγόριθμου ECM για την δημιουργία συστάδων και τον διαχωρισμό των υπόχωρων D_q των N_q επιλεγμένων παρατηρήσεων εκμάθησης.

5. Δημιουργία ασαφών κανόνων και ρύθμιση των αρχικών παραμέτρων τους σύμφωνα με τα αποτελέσματα του αλγόριθμου ECM .Κάθε κατηγορία αντιστοιχεί σε έναν κανόνα. Τα κέντρα και οι ακτίνες των συστάδων χρησιμοποιούνται για την δημιουργία του κέντρου και του πλάτους των συναρτήσεων συμμετοχής.

6. Εφαρμογή της διαδικασίας back-propagation για την βελτιστοποίηση των ασαφών παραμέτρων στο τοπικό μοντέλο M_q, με την ελαχιστοποίηση της συνάρτησης του σφάλματος.

7. Υπολογισμός του αποτελέσματος y_q του μοντέλου για το σημείο ελέγχου $\boldsymbol{x_q}$, με τη χρήση του βέλτιστου τοπικού μοντέλου M_q, δεδομένων των ασαφών κανόνων .

2.13 Μέθοδοι επιλογής χαρακτηριστικών

Σε αυτή την παράγραφο, παραθέτονται 3 σημαντικές μέθοδοι επιλογής χαρακτηριστικών. Οι μέθοδοι συνδυάζονται με τις μεθόδους ταξινόμησης 1NN, KNN και WKNN. Οι μέθοδοι KNN και WKNN έχουν αναλυθεί στην παράγραφο 2.10. Η μέθοδος 1NN είναι η ίδια,με την KNN με τη διαφορά οτι για κάθε παρατήρηση του συνόλου ελέγχου, υπολογίζεται η Ευκλείδεια απόσταση της από κάθε μία από τις παρατηρήσεις του συνόλου εκμάθησης. Με αυτή τη διαδικασία, υπολογίζεται η κοντινότερη παρατήρηση από το σύνολο των δεδομένων εκμάθησης. Έτσι, κάθε σημείο ελέγχου ταξινομείται στην ίδια κατηγορία που ανήκει η κοντινότερη παρατήρηση σε αυτό, από το σύνολο των δεδομένων εκμάθησης.

2.13.1 Γενετικός Αλγόριθμος (Genetic Algorithm-GA) (Yannis Marinakis $ George Dounias 2006)

Οι γενετικοί αλγόριθμοι είναι διαδικασίες αναζήτησης που βασίζονται στο μηχανισμό της φυσικής επιλογής των ικανοτέρων και στην γενετική. Μιμούνται την εξελικτική διαδικασία της φύσης, μιμούνται

δηλαδή την βιολογική διαδικασία κατά την οποία αναπτύσσονται νέοι και καλύτεροι πληθυσμοί από διαφορετικά είδη κατά την εξέλιξη. Οι γενετικοί αλγόριθμοι χρησιμοποιούν πληροφορίες από έναν πληθυσμό από λύσεις που ονομάζονται υποκείμενα (individuals), όταν ψάχνουν για καλύτερες λύσεις. Ένας γενετικός αλγόριθμος είναι μια στοχαστική επαναληπτική διαδικασία κατά την οποία το μέγεθος του πληθυσμού διατηρείται σταθερό σε κάθε επανάληψη. Κάθε επανάληψη ονομάζεται γενιά (generation). Η βασική λειτουργία των γενεών είναι ο συνδυασμός δύο λύσεων με σκοπό τη δημιουργία μιας νέας λύσης. Για να δημιουργηθεί ένας καινούριος πληθυσμός εφαρμόζονται δύο τελεστές (operators), ο δυαδικός που ονομάζεται τελεστής διασταύρωσης (crossover) και ο μοναδιαίος (unary) που ονομάζεται τελεστής μετάλλαξης (mutation). Ο τελεστής διασταύρωσης παίρνει δύο υποκείμενα που ονομάζονται γονείς και παράγει δύο νέα υποκείμενα που ονομάζονται γόνοι, συνδυάζοντας μέρη από τους γονείς.

Ακολουθούν τα βήματα του γενετικού αλγορίθμου που χρησιμοποιείται σε αυτή τη διπλωματική: Κάθε υποκείμενο στον πληθυσμό είναι μια πιθανή λύση στο πρόβλημα της επιλογής των χαρακτηριστικών. Έστω m ο συνολικός αριθμός των χαρακτηριστικών. Καθένα από αυτά αντιπροσωπεύεται από ένα δυαδικό διάνυσμα m διαστάσεων. Αν πάρει την τιμή 1 σημαίνει ότι το αντίστοιχο χαρακτηριστικό επιλέγεται, διαφορετικά δεν επιλέγεται. Ο αρχικός πληθυσμός δημιουργείται τυχαία. Για να εξερευνηθούν τα υποσύνολα των διαφόρων χαρακτηριστικών, ο αριθμός των μονάδων (1) για κάθε υποκείμενο δημιουργείται τυχαία. Επιτρέπονται μόνο διαφορετικά υποκείμενα. Έτσι, στον αρχικό πληθυσμό, δεν υπάρχουν υποκείμενα με τα ίδια χαρακτηριστικά. Με αυτόν τον τρόπο διατηρείται η πολυπλοκότητα του αρχικού πληθυσμού.

Ο μηχανισμός επιλογής είναι υπεύθυνος για την επιλογή του γονέα από τον πληθυσμό και τη δημιουργία του συνδυασμού. Ο μηχανισμός αυτός μιμείται το μηχανισμό της φύσης που βασίζεται στην επιβίωση του καλύτερου (survival of the fittest). Όπως είναι αναμενόμενο ένα καλύτερο χρωμόσωμα έχει μεγαλύτερη πιθανότητα να επιβιώσει κατά την εξέλιξη. Σε αυτήν την εργασία, χρησιμοποιείται η επιλογή ρουλέτα (roulette wheel selection) που είναι ένας εύκολα εφαρμόσιμος μηχανισμός επιλογής και δουλεύει ως εξής: κάθε υποκείμενο του πληθυσμού καταλαμβάνει έναν τομέα στην εικονική ρουλέτα. Ανάλογα με την τιμή προσαρμογής (fitness value) του υποκειμένου, ο τομέας καταλαμβάνει μεγαλύτερη περιοχή όταν το αντίστοιχο υποκείμενο έχει καλύτερη τιμή προσαρμογής.

Χρησιμοποιείται ο τελεστής διασταύρωσης ενός σημείου (1- point crossover). Σύμφωνα με αυτόν το τελεστή, οι δύο γονείς χωρίζονται σε δύο μέρη σε ένα συγκεκριμένο σημείο. Οι δύο γόνοι παίρνουν το ένα μέρος από τον ένα γονιό και το δεύτερο μέρος από τον άλλο. Έπειτα υπολογίζεται η συνάρτηση προσαρμογής για τον κάθε γόνο. Για παράδειγμα:

Γονέας 1: 1 0 0 1| 1 0 1 0 1 Γονέας 2: 1 1 1 1| 0 0 0 0 0

Οι δύο γονείς χωρίζονται ανάμεσα στο τέταρτο και στο πέμπτο χαρακτηριστικό. Ακολουθούν οι παραγόμενοι γόνοι:

Γόνος 1: 1 0 0 1 0 0 0 0 0 Γόνος 2: 1 1 1 1 1 0 1 0 1

Σε ένα συγκεκριμένο ποσοστό των γόνων εφαρμόζεται μετάλλαξη. Η μετάλλαξη λειτουργεί σε μία μόνο τιμή. Έπειτα υπολογίζεται η συνάρτηση προσαρμογής για κάθε γόνο. Για παράδειγμα:

Πριν την μετάλλαξη: 1 1 1 **1** 1 0 1 0 1 Μετά την μετάλλαξη: 1 1 1 **0** 1 0 1 0 1

Στην επόμενη γενιά επιβιώνει ο καλύτερος από όλο τον πληθυσμό. Με τον όρο όλο τον πληθυσμό εννοείται ο αρχικός πληθυσμός μαζί με τους γόνους που παράχθηκαν από τις φάσεις μετάλλαξης και διασταύρωσης. Έτσι, ο πληθυσμός ταξινομείται με βάση την συνάρτηση προσαρμογής των υποκειμένων και στην επόμενη γενιά, επιβιώνουν τα καλύτερα υποκείμενα.

Ο αλγόριθμος σταματά όταν φτάσει στο μέγιστο αριθμό γενεών ή συγκλίνει σε μοναδική λύση.

2.13.2 Διερεύνηση ταμπού (Tabu search-TS) (Y. Marinakis & G. Dounias 2006)

Ο αλγόριθμος διερεύνησης ταμπού εισήχθη από τον Glover και είναι μια επαναληπτική διαδικασία για την επίλυση προβλημάτων βελτιστοποίησης. Η εμπειρία δείχνει ότι ο αλγόριθμος TS είναι μια καλά καθιερωμένη τεχνική προσέγγισης που μπορεί να συναγωνιστεί κάθε άλλη γνωστή τεχνική. Έχει την μορφή της αναζήτησης τοπικού γείτονα. Κάθε λύση S έχει ένα σύνολο από γείτονες N(S). Μια λύση S’$\in$ N(S) μπορεί να προσεγγιστεί από το S, μέσω μιας λειτουργίας που ονομάζεται κίνηση (move). Ο TS κινείται από μια λύση προς τον καλύτερα αποδεκτό γείτονα, ακόμα κι αν αυτό χειροτερέψει την αντικειμενική συνάρτηση. Για να αποφευχθούν οι άσκοποι κύκλοι του αλγορίθμου, οι λύσεις που έχουν πρόσφατα εξερευνηθεί, θεωρούνται απαγορευμένες ή Tabu, για έναν αριθμό επαναλήψεων. Η Tabu κατάσταση μιας λύσης διαγράφεται όταν ικανοποιηθούν κάποια συγκεκριμένα κριτήρια. Ο αλγόριθμος ξεκινά με μια αρχική λύση, δηλαδή μια αρχική επιλογή χαρακτηριστικών. Το διάνυσμα της επιλογής χαρακτηριστικών αντιπροσωπεύεται από δυαδικές μεταβλητές, όπου το 0 δείχνει ότι το χαρακτηριστικό δεν επιλέγεται, ενώ το 1 ότι επιλέγεται. Αρχικά, μόνο δύο χαρακτηριστικά επιλέγονται και υπολογίζεται το μέσο τετραγωνικό σφάλμα της λύσης με τον ταξινομητή 1 – nn. Έπειτα δημιουργείται μια γειτονική λύση. Οι γείτονες δημιουργούνται από την τυχαία ενεργοποίηση ή απενεργοποίηση ενός χαρακτηριστικού στο διάνυσμα των χαρακτηριστικών. Από τους γείτονες, επιλέγεται αυτός με το καλύτερο μέσο τετραγωνικό σφάλμα και θεωρείται η καινούρια τρέχουσα λύση για την επόμενη επανάληψη. Υπάρχουν δύο περιορισμοί κατά τη χρήση αυτού του αλγορίθμου. Ο πρώτος είναι ότι το διάνυσμα των χαρακτηριστικών δεν επιτρέπεται να έχει λιγότερα από δύο ενεργοποιημένα χαρακτηριστικά και ο δεύτερος είναι οι κινήσεις Tabu (Tabu Moves). Με τον όρο Tabu Moves εννοείται μια λίστα Tabu που πρέπει να διατηρείται ώστε να αποφεύγεται η επιστροφή σε προηγούμενες λύσεις. Έτσι, ένα χαρακτηριστικό που έχει προστεθεί ή διαγραφεί από τη λύση, σε μία επανάληψη, δεν επιτρέπεται να ξαναμπεί στη λύση για έναν αριθμό επαναλήψεων που ισούται με το μέγεθος της λίστας Tabu. Αν μια κίνηση οδηγήσει σε μια επαρκή λύση

ακόμα και αν είναι απαγορευμένη, ο περιορισμός της λίστας Tabu δεν ενεργοποιείται και η κίνηση επιτρέπεται.

2.13.3 Βελτιστοποίηση αποικίας μυρμηγκιών (Ant Colony Optimization – ACO) (Y. Marinakis & G. Dounias 2006)

Η μέθοδος βελτιστοποίησης αποικίας μυρμηγκιών είναι μια σχετικά καινούρια τεχνική για την επίλυση προβλημάτων βελτιστοποίησης και βασίζεται στο σύστημα μυρμηγκιών (Ant System-AS) που αναπτύχθηκε από τους Dorigo, Maniezzo και Colorni. Ο αλγόριθμος βελτιστοποίησης αποικίας μυρμηγκιών είναι ένα σύστημα που βασίζεται σε πράκτορες που προσομειώνουν την φυσική συμπεριφορά των μυρμηγκιών, συμπεριλαμβανομένων των μηχανισμών συνεργασίας και προσαρμογής. Ο αλγόριθμος ACO μιμείται τις τεχνικές που έχουν τα μυρμήγκια για να βρίσκουν την συντομότερη πορεία από την πηγή τροφής έως την φωλιά τους και αντίστροφα, χωρίς τη χρήση οπτικών πληροφοριών. Τα μυρμήγκια ψάχνουν την περιοχή που περιβάλλει την φωλιά τους με τυχαίο τρόπο και καθώς κινούνται, μια συγκεκριμένη ποσότητα φερομόνης αποβάλλεται από τον οργανισμό τους στο έδαφος, σημαδεύοντας το μονοπάτι με ίχνη της ουσίας. Η ποσότητα της φερομόνης εξαρτάται από την απόσταση, την ποσότητα και την ποιότητα της πηγής τροφής. Όταν ένα μυρμήγκι ανιχνεύσει την φερομόνη, είναι πολύ πιθανό να αποφασίσει να ακολουθήσει το συγκεκριμένο μονοπάτι. Αυτό το μυρμήγκι θα αφήσει επίσης μια συγκεκριμένη ποσότητα φερομόνης κι έτσι θα ενισχύσει τα ίχνη φερομόνης για το συγκεκριμένο μονοπάτι. Όμως, όσο περνάει η ώρα η φερομόνη αρχίζει να εξατμίζεται. Συνεπώς, ένα κοντινότερο μονοπάτι θα το επισκεφθούν περισσότερα μυρμήγκια. Τελικά όλα τα μυρμήγκια θα συγκλίνουν στο κοντινότερο μονοπάτι. Ακολουθούν τα βήματα του αλγόριθμου:

Κάθε πιθανό χαρακτηριστικό στον αλγόριθμο ACO, αντιπροσωπεύεται από ένα δυαδικό μυρμήγκι που παίρνει την τιμή 1 αν το χαρακτηριστικό επιλεχθεί και την τιμή 0 αν δεν επιλεχθεί. Υπολογίζεται ένας αρχικός πληθυσμός λύσεων r με σκοπό την εύρεση μιας αρχικής βέλτιστης λύσης η οποία θα χρησιμοποιηθεί στον υπολογισμό της συνάρτησης n_i του χαρακτηριστικού i. Συνήθως στον αλγόριθμο ACO η τιμή n_i χρησιμοποιείται σε συνδυασμό με την τιμή φερομόνης για να αποφασιστούν οι αλλαγές που θα γίνουν. Αυτή η τιμή δίνει μια εκτίμηση της ποιότητας κάθε χαρακτηριστικού σε σχέση με τη δυνατότητά του να βελτιώνει την ακρίβεια πρόβλεψης. Το n_i υπολογίζεται από τις r_j καλύτερες λύσεις του αρχικού πληθυσμού. Επιθυμητή είναι μια αρχική εκτίμηση των πιο σημαντικών χαρακτηριστικών. Για αυτό, υπολογίζονται τα χαρακτηριστικά που ανήκουν στις r_j καλύτερες λύσεις και όλα τα χαρακτηριστικά σταθμίζονται με βάση τις φορές που εμφανίζεται κάθε χαρακτηριστικό στις r_j καλύτερες λύσεις. Αυτά τα χαρακτηριστικά έχουν μεγαλύτερη τιμή στον πίνακα n.

Κάθε μυρμήγκι που χρησιμοποιείται στον αλγόριθμο ξεκινά από διαφορετικό μέρος του διανύσματος χαρακτηριστικών και ακολουθεί την δική του πορεία. Όλα τα μυρμήγκια αρχίζουν να

χτίζουν λύσεις ταυτόχρονα. Κάθε μυρμήγκι έχει τη δυνατότητα να επισκεφτεί όλα τα χαρακτηριστικά και να χτίσει λύσεις. Κάθε μυρμήγκι χρησιμοποιείται για έναν αριθμό γενεών, ξεκινώντας πάντα από το ίδιο χαρακτηριστικό και επιλέγοντας σε κάθε γενιά διαφορετικά χαρακτηριστικά με βάση την ποσότητα της φερομόνης που υπάρχει σε καθένα από αυτά.

Η αρχική ποσότητα φερομόνης τ_i για το χαρακτηριστικό i υπολογίζεται από τον τύπο:

$$\tau_i = \frac{ant\_size}{init\_opt} \tag{2.23}$$

Όπου ant_size είναι ο αρχικός πληθυσμός μυρμηγκιών και $init\_opt$ είναι η ακρίβεια της βέλτιστης λύσης του αρχικού πληθυσμού.

Ένα μυρμήγκι που βρίσκεται στο χαρακτηριστικό j αποφασίζει αν το χαρακτηριστικό i επιλέγεται ή όχι σύμφωνα με τον τύπο:

$$p_i = \frac{[\tau_i]^{\alpha}[n_i]^{\beta}}{\sum_{l=1}^{M}[\tau_i]^{\alpha}[n_i]^{\beta}} \tag{2.24}$$

Όπου M είναι ο αριθμός των χαρακτηριστικών και α, β είναι δύο παράμετροι που έχουν επιλεχθεί εμπειρικά. Αν α = 0 τα χαρακτηριστικά που επιλέγονται στις αρχικές λύσεις είναι πιο πιθανό να επιλεχθούν ξανά και αν β = 0 χρησιμοποιείται μόνο η φερομόνη. Έπειτα, υπολογίζεται η προσαρμοστικότητα κάθε μυρμηγκιού και καθένα από αυτά επιλέγει το επόμενο χαρακτηριστικό που θα επισκεφθεί. Στον αλγόριθμο αυτόν υπάρχει ο περιορισμός που δεν επιτρέπει στα μυρμήγκια να δημιουργήσουν το καθένα ένα μονοπάτι όπου όλα τα χαρακτηριστικά θα είναι ενεργοποιημένα. Όταν όλα τα μυρμήγκια συμπληρώσουν τα μονοπάτια τους, εφαρμόζεται μια απλή τοπική έρευνα σε καθένα από αυτά, με σκοπό την βελτιστοποίηση των λύσεων. Τα χαρακτηριστικά που δεν έχουν ενεργοποιηθεί στην τρέχουσα λύση ενεργοποιούνται τώρα κι αντίστροφα με σκοπό την εύρεση καλύτερης λύσης.

ΚΕΦΑΛΑΙΟ 3

Αλγόριθμοι Στατιστικής

Σε αυτό το κεφάλαιο παρουσιάζονται θεωρητικά οι στατιστικοί μέθοδοι ανάλυσης και ταξινόμησης των παρατηρήσεων που εφαρμόζονται σε αυτήν την εργασία

3.1 Συντελεστής συσχέτισης

Ο συντελεστής συσχέτισης μεταξύ δύο μεταβλητών X και Y ορίζεται ως το πηλίκο

$$\rho_{XY} = \frac{cov(X,Y)}{S_X S_Y}$$

όπου cov(X,Y) είναι η συνμεταβλητότητα των δύο μεταβλητών και S_X , S_Y οι τυπικές αποκλίσεις των μεταβλητών X και y αντίστοιχα .

Ο δειγματικός συντελεστής συσχέτισης ορίζεται ως

$$r_{XY} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

Ο συντελεστής συσχέτισης παίρνει τιμές στο διάστημα [-1,1] . Τιμές κοντά στα άκρα δηλώνουν ισχυρές συσχετίσεις μεταξύ των μεταβλητών . Θετικές τιμές υποδεικνύουν αναλογική σχέση ενώ αρνητικές δηλώνουν αντιστρόφως ανάλογη σχέση μεταξύ των μεταβλητών

Γ ια αν ελεγχθεί η σημαντικότητα του συντελεστή συσχέτισης χρησιμοποιείται το στατιστικό Τ όπου

$$T = r\sqrt{\frac{n-2}{1-r^2}}$$

Που ακολουθεί την κατανομή student (t_{n-2}) στους n-2 βαθμούς ελευθερίας. Η μηδενική υπόθεση του ελέγχου είναι H_0 : r=0 έναντι της εναλλακτικής H_α : r ≠ 0. Η μηδενική υπόθεση απορρίπτεται αν $T>t_{n-2;1-a/2}$ ή $T<t_{n-2;a/2}$

3.2 t έλεγχος για την διαφορά των μέσων τιμών (t-test)

Έστω ότι συγκρίνονται δύο μέσες τιμές $\overline{x_1}$, $\overline{x_2}$ που προέρχονται από 2 δείγματα μεγέθους n_1 και n_2 αντίστοιχα. Η μηδενική υπόθεση είναι H_0: οι μέσες τιμές δεν διαφέρουν σε στατιστικά σημαντικό βαθμό έναντι της $H_{\alpha:}$ οι μέσες τιμές διαφέρουν σε στατιστικά σημαντικό βαθμό. Για την σύγκριση αυτή υπολογίζεται η τιμή του κριτηρίου t με βάση τον τύπο

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{SE_1{}^2 + SE_2{}^2}}$$ όπου

SE_1 και SE_2 τα τυπικά σφάλματα των μέσων τιμών $\overline{x_1}$ και $\overline{x_2}$ αντίστοιχα .Η τιμή αυτή συγκρίνεται με την αντίστοιχη θεωρητική τιμή της κατανομής t που εξαρτάται από το προκαθορισμένο επίπεδο σημαντικότητας και από τα μεγέθη n_1 και n_2 με τη μορφή $n_1 + n_2 - 1$ η οποί ορίζει τους βαθμούς ελευθερίας.

3.3 Ανάλυση διασποράς (ANOVA)

Ενώ ο t έλεγχος για την διαφορά των μέσων τιμών εφαρμόζεται μόνο ανάμεσα σε δύο κατηγορίες, η ανάλυση διασποράς μπορεί να συγκρίνει την επίδραση ενός χαρακτηριστικού ανάμεσα σε παραπάνω από δύο κατηγορίες μιας ποιοτικής μεταβλητής.

Έστω y το χαρακτηριστικό του οποίου την επίδραση θέλουμε να εξετάσουμε κι έστω x η μεταβλητή που δηλώνει τις κατηγορίες ανάμεσα στις οποίες θέλουμε να συγκρίνουμε την επίδραση. Το μοντέλο που δείχνει την εκτίμηση της γραμμικής συσχέτισης μεταξύ τους είναι $\hat{y} = b_0 + b_i x_i$ όπου b_0 και b_i η

εκτιμώμενη σταθερά και οι συντελεστές του μοντέλου αντίστοιχα. x_i είναι η μεταβλητή που δηλώνει την κατηγορία και παίρνει την τιμή 0 αν η παρατήρηση δεν ανήκει στην κατηγορία και 1 αν ανήκει.

Ορίζουμε ως άθροισμα τετραγώνων του μοντέλου $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$, ως άθροισμα τετραγώνων των υπολοίπων $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ και ως συνολικό άθροισμα τετραγώνων $SST = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$

Στον Πίνακα 3.1 φαίνονται τα αθροίσματα των τετραγώνων , τα μέσα των τετραγώνων των αθροισμάτων ώστε να γίνει αντιληπτή η διαδικασία ελέγχου με ανάλυση διασποράς

πηγή	**Άθροισμα τετραγώνων**	**Βαθμοί ελευθερίας**	**Μέσο άθροισμα τετραγώνων**	**Στατιστικό F**
Μοντέλο	SSR	p	MSR=SSR/p	F=MSR/MSE
Σφάλμα	SSE	n-(p+1)	MSE=SSE/[n-(p+1)]	
συνολικό	SST	n-1	MST=SST/(n-1)	

Πίνακας 3.1 : Ανάλυση διασποράς και ορισμός στατιστικού F

Η ποσότητα F ακολουθεί την κατανομή $F_{a,p,n-(p+1)}$ όπου a το επίπεδο σημαντικότητας και p, n-(p+1) οι βαθμοί ελευθερίας του MSR και MSE αντίστοιχα .Το F δείχνει αν υπάρχει στατιστικά σημαντική διαφορά του χαρακτηριστικού που εξετάζεται ανάμεσα στις κατηγορίες της ποιοτικής μεταβλητής.

Επίσης ορίζεται η ποσότητα R^2 =SSR/SST που δείχνει το ποσοστό της μεταβλητότητας που εξηγεί το μοντέλο.

Για τον έλεγχο της διαφοράς μεταξύ μιας κατηγορίας και της κατηγορίας αναφοράς υπολογίζεται το στατιστικό $t = \frac{b_i}{SE_{b_i}}$ που ακολουθεί την κατανομή student. Η μηδενική υπόθεση του ελέγχου είναι H_0: $b_i = 0$ έναντι της εναλλακτικής H_{α}: $b_i \neq 0$. Η μηδενική υπόθεση απορρίπτεται αν $|t| > t_{\frac{\alpha}{2},n-2}$

3.4 Λογιστική παλινδρόμηση

Έστω μία μεταβλητή y που παίρνει τιμές 0,1. Η πιθανότητα να πάρει την τιμή 0 έστω ότι είναι $1-\pi_1$ και η πιθανότητα να πάρει την τιμή 1 είναι άρα π_1. Το μοντέλο που δημιουργείται θα έχει και ένα σύνολο από επεξηγηματικές μεταβλητές x_i των οποίων θέλουμε να μελετήσουμε τη σχέση με την y. Επειδή η πιθανότητα παίρνει τιμές στο διάστημα (0,1) πρέπει να χρησιμοποιηθεί ένας μετασχηματισμός που θα αντικατοπτρίζει το σύνολο (0,1) στο (-∞ , +∞). Στην συγκεκριμένη εργασία ο μετασχηματισμός αυτός είναι η λογαριθμιστική συνάρτηση log(π/1-π). Έτσι το μοντέλο παίρνει τη μορφή

$$\log[\pi/(1-\pi)] = b_0 + b_i x_i \qquad (3.1)$$

Υπολογίζοντας τον αντιλογάριθμο των συντελεστών b_i βρίσκεται ο λόγος συμπληρωματικών πιθανοτήτων(OR) και ερμηνεύονται ανάλογα.

3.5 Διατάξιμη λογιστική παλινδρόμηση

Έστω μία μεταβλητή y που αντιπροσωπεύει ένα διαβαθμισμένο χαρακτηριστικό όπως για παράδειγμα τα προκαρκινικά στάδια των κυττάρων του τραχήλου της μήτρας που εξετάζονται στην εργασία αυτή .Η διαβαθμισμένη λογιστική παλινδρόμηση χρησιμοποιεί αθροιστικές συναρτήσεις πιθανοτήτων $\gamma_j=pr(y<j)$ και το μοντέλο που δημιουργείται έχει την μορφή

$$\log[\gamma_j(x)/(1-\gamma_j(x)] = \kappa_j - \beta^T x \quad , j=1, \dots ,l-1$$

όπου l ο αριθμός των κατηγοριών. Αυτό το μοντέλο είναι γνωστό ως μοντέλο αναλογικών συμπληρωματικών πιθανοτήτων (proportional – odds model) επειδή ο λόγος των odds του $y<j$ για x_1 και x_2 είναι ανεξάρτητο του j.

$$\frac{\gamma j(x_1)/(1-\gamma j(x_1)}{\gamma j(x_2)/(1-\gamma j(x_2)} = \exp\left[-\beta^T(x_1-x_2)\right]$$

Έστω μια συνεχής μεταβλητή U που συνδέεται με το αποτέλεσμα y έτσι ώστε $\pi=pr(y=1)=pr(U>\kappa)$ και $1-p=pr(y=0)=pr(U\leq\kappa)$. Η συνεχής μεταβλητή U συνδέεται με ένα σύνολο επεξηγηματικών μεταβλητών σύμφωνα με το μοντέλο

$U=\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon$ όπου το ε ακολουθεί την κατανομή F(ε) . Έτσι:

$\pi = pr(U>\kappa) =$

$pr(U-\beta^*X>\kappa- \beta^*X) =$

$pr(\varepsilon> \kappa- \beta^*X) =$

$1 - F(\kappa- \beta^*X)$ και

$1-\pi= F(\kappa- \beta^*X)$

$F^{-1}(1-\pi)=\kappa- \beta^*X=n^*$ όπου $n^*=\kappa - \beta_1 x_1 - \beta_2 x_2 - \dots - \beta_p x_p$.

Αυτό είναι ένα γενικευμένο γραμμικό μοντέλο με σύνδεσμο (link) $n=g(\pi) = F^{-1}(1-\pi)$ για κάποια κατανομή F.

Αυτή η προσέγγιση του μοντέλου με την μεταβλητή U έχει τα πλεονεκτήματα ότι προκύπτει πολύ συχνά στα προβλήματα που συναντώνται, βοηθά στην κατανόηση άλλων περιπτώσεων κατανομής για το ε και οδηγεί σε μια πιο ομαλή γενίκευση προς το αποτέλεσμα των διαβαθμισμένων μεταβλητών.

3.6 Δημιουργία συστάδων Κ μέσων

Η ανάλυση κατά συστάδες είναι η μέθοδος που έχει σκοπό να κατατάξει σε ομάδες τις παρατηρήσεις χρησιμοποιώντας τις υπάρχουσες μεταβλητές. Η βασική αρχή της είναι ότι οι παρατηρήσεις μέσα σε κάθε ομάδα πρέπει να είναι όσο γίνεται πιο ομοιογενείς και οι παρατηρήσεις διαφορετικών ομάδων να διαφέρουν όσο γίνεται περισσότερο.

Συγκεκριμένα στην ανάλυση Κ μέσων ο αριθμός των ομάδων είναι γνωστός από την αρχή. Με έναν επαναληπτικό αλγόριθμο οι παρατηρήσεις κατανέμονται στις ομάδες ανάλογα με το ποια ομάδα είναι πιο κοντά στην παρατήρηση . Για συνεχή δεδομένα συνήθως χρησιμοποιείται η Ευκλείδεια απόσταση

$$d(x,y) = \sqrt{\sum_{i=1}^{\rho} (x_i - y_i)^2}$$

Η Ευκλείδεια απόσταση δεν προτιμάται όταν κάποια από τις μεταβλητές έχει τεράστια διακύμανση σε σχέση με τις υπόλοιπες γιατί αυτή θα παίζει σπουδαιότερο ρόλο και θα κατευθύνει τα αποτελέσματα. Σε αυτήν την περίπτωση προτείνεται η απόσταση Mahalanobis

$$d^2(x,y) = (x-y)^t S^{-1}(x-y)$$

Η οποία λαμβάνει υπόψη διακυμάνσεις και συνδιακυμάνσεις

Ο αλγόριθμος της μεθόδου Κ μέσων έχει τα εξής βήματα:

1.Βρίσκονται τα αρχικά κέντρα, αυθαίρετα ή με συγκεκριμένο τρόπο

2. Κατάταξη της κάθε παρατήρησης στην ομάδα της οποίας το κέντρο έχει τη μικρότερη απόσταση

3. Από τις παρατηρήσεις που είναι μέσα στη νέα ομάδα υπολογίζονται τα νέα κέντρα. Τα κέντρα υπολογίζονται από τη μέση τιμή για κάθε μεταβλητή όλων των παρατηρήσεων της ομάδας .

4. Τα νέα κέντρα διαφέρουν από τα παλιά; Αν όχι η συσταδική ανάλυση σταματά. Αν ναι τότε πήπηγαίνειε στο βήμα 3

3.7 Διαχωριστική ανάλυση

Σκοπός της διαχωριστικής ανάλυσης είναι να κατανείμει κάθε παρατήρηση του δείγματος στις κ ήδη γνωστές ομάδες. Για να γίνει αυτό πρέπει να βρεθεί ένας διαχωριστικός κανόνας που μπορεί να καταχωρίσει όσο το δυνατόν πιο σωστά περισσότερες παρατηρήσεις. Ο διαχωριστικός κανόνας μπορεί να είναι μια συνάρτηση (ευθεία, ή υπερ-επίπεδο) που διαχωρίζει κατά τον καλύτερο τρόπο τα δεδομένα.

Σκοπός λοιπόν είναι να βρεθεί μια συνάρτηση της μορφής:

$$y = w^T x$$

Έστω ότι οι αριθμητικοί μέσοι:

$$\mu_i = \frac{1}{N_i}\sum_{x\in w_i} x \text{ και } \tilde{\mu}_i = \frac{1}{N_i}\sum_{y\in w_i} y = \frac{1}{N_i}\sum_{y\in w_i} w^T x = w^T \mu_i$$

Αν υποθέσουμε ότι έχουμε 2 ομάδες με μέσους μ_1 και μ_2 τότε η απόσταση των μέσων ορίζεται από τη σχέση:

$$J(w) = |\tilde{\mu_1} - \tilde{\mu_2}| = |w^T(\mu_1 - \mu_2)|$$

Το παραπάνω μέτρο απόστασης δεν λαμβάνει όμως υπόψη την μεταβλητότητα εντός των ομάδων. Η ιδανική διαχωριστική συνάρτηση είναι αυτή που τα στοιχεία της κάθε ομάδας προβάλλονται όσο πιο κοντά γίνεται μεταξύ τους, ενώ οι αντίστοιχοι μέσοι έχουν την μεγαλύτερη δυνατή απόσταση.

Για να λάβει υπόψη την ενδο-μεταβλητότητα ο Fisher πρότεινε ως διαχωριστική συνάρτηση την $Y = w^T x$, η οποία μεγιστοποιεί την μετρική (απόσταση):

$$J(w) = \frac{|\tilde{\mu_1} - \tilde{\mu_2}|^2}{\tilde{S_1}^2 + \tilde{S_2}^2} \qquad \text{όπου}$$

$$\tilde{S_i}^2 = \sum_{y\in w_i}(y - \tilde{\mu_i})^2$$

Μεγάλες ιδιοτιμές της συνάρτησης υποδηλώνουν ότι η συγκεκριμένη διάσταση έχει καλή διαχωριστική ικανότητα και ερμηνεύει ικανοποιητικά την μεταβλητότητα μεταξύ των ομάδων.

3.8 Δείκτης Κάππα (Kappa)

Η στατιστική τιμή K είναι ένα μέτρο της διαφοράς μεταξύ της πραγματικής συμφωνίας μεταξύ των δοσμένων δεδομένων ενός προβλήματος και ενός αυτόματου αλγόριθμου ταξινόμησης, και της τυχαίας συμφωνίας μεταξύ των δοσμένων δεδομένων και του τυχαίου αλγόριθμου ταξινόμησης. Θεωρητικά, ο K μπορεί να οριστεί ως:

K= (παρατηρηθείσα ακρίβεια - τυχαία συμφωνία) / (1-τυχαία συμφωνία)

Η στατιστική παράμετρος **K** υπολογίζεται ως:

$$\hat{k} = \frac{N\cdot\sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r}(x_{i+}\cdot x_{+i})}{N^2 - \sum_{i=1}^{r}(x_{i+}\cdot x_{+i})}$$

όπου:

r = ο αριθμός των γραμμών του πίνακα σύγχυσης,

xii = ο αριθμός των παρατηρήσεων στη γραμμή i και τη στήλη i (στην κύρια διαγώνιο),

xi+ = το σύνολο των παρατηρήσεων στη γραμμή i (στο δεξί μέρος του πίνακα),

x+i= το σύνολο των παρατηρήσεων στη στήλη i (στο κάτω μέρος του πίνακα), και

N = ο συνολικός αριθμός των παρατηρήσεων που περιλαμβάνονται στον πίνακα.

Ως πίνακας σύγχισης ορίζεται ο πίνακας που έχει τη μορφή:

	Προβλεπόμενες κατηγορίες κατάταξης παρατήρησης				
Πραγματικές Κατηγορίες κατάταξης παρατήρησης	**Κατηγορία 1 (αρνητικά)**	**Κατηγορία 2 (θετικά)**	**…**	**Κατηγορία r**	**Σύνολο**
Κατηγορία 1 (αρνητικά)	Αληθώς κατηγορία 1	Ψευδώς κατηγορία 2	**…**	Ψευδώς κατηγορία r	x_{1+}
Κατηγορία 2 (θετικά)	Ψευδώς κατηγορία 1	Αληθώς κατηγορία 2	**…**	Ψευδώς κατηγορία r	χ_{2+}
…	**…**	**…**	**…**		x_{i+}
Κατηγορία r	Ψευδώς κατηγορία 1	Ψευδώς κατηγορία 2	**…**	Αληθώς κατηγορία r	x_{r+}
Σύνολο	x_{+1}	x_{+2}	x_{+i}	x_{+r}	N

ΚΕΦΑΛΑΙΟ 4

Μονομεταβλητή στατιστική ανάλυση

Σε αυτό το κεφάλαιο, παρατίθενται περιγραφικά στατιστικά στοιχεία και μονοπαραγοντική ανάλυση των δεδομένων των βάσεων επιχρίσματος-Παπανικολάου.

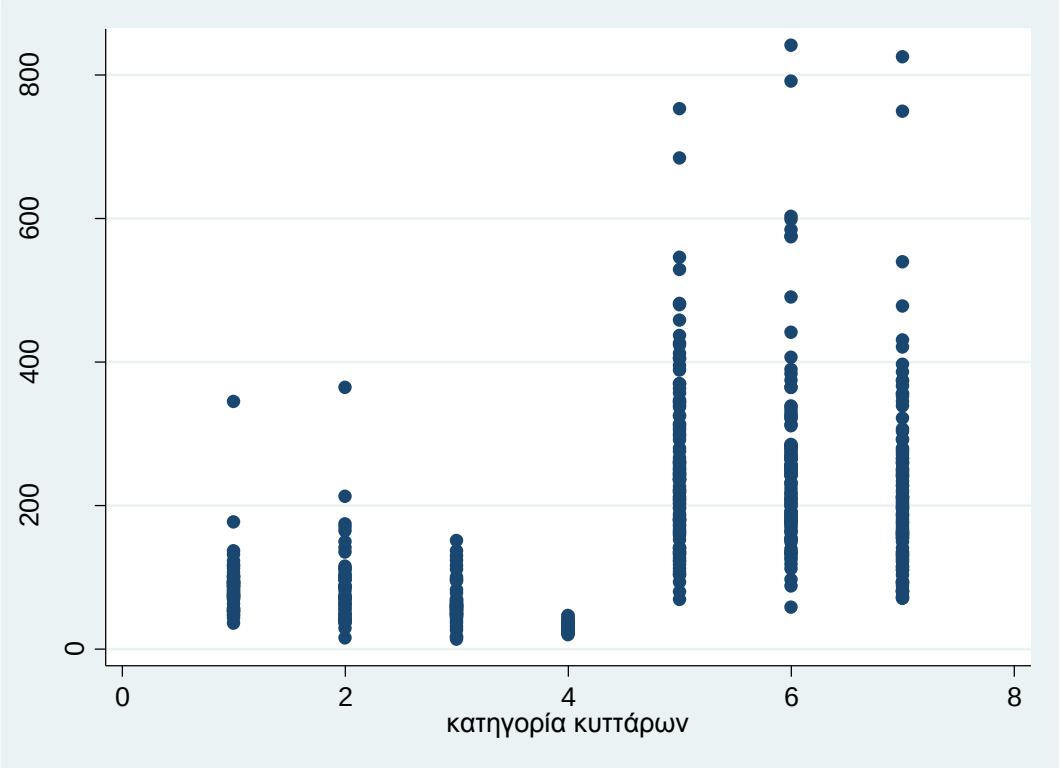
4.1 Παλαιά βάση δεδομένων επιχρίσματος Παπανικολάου

4.1.1 Περιγραφικά στοιχεία

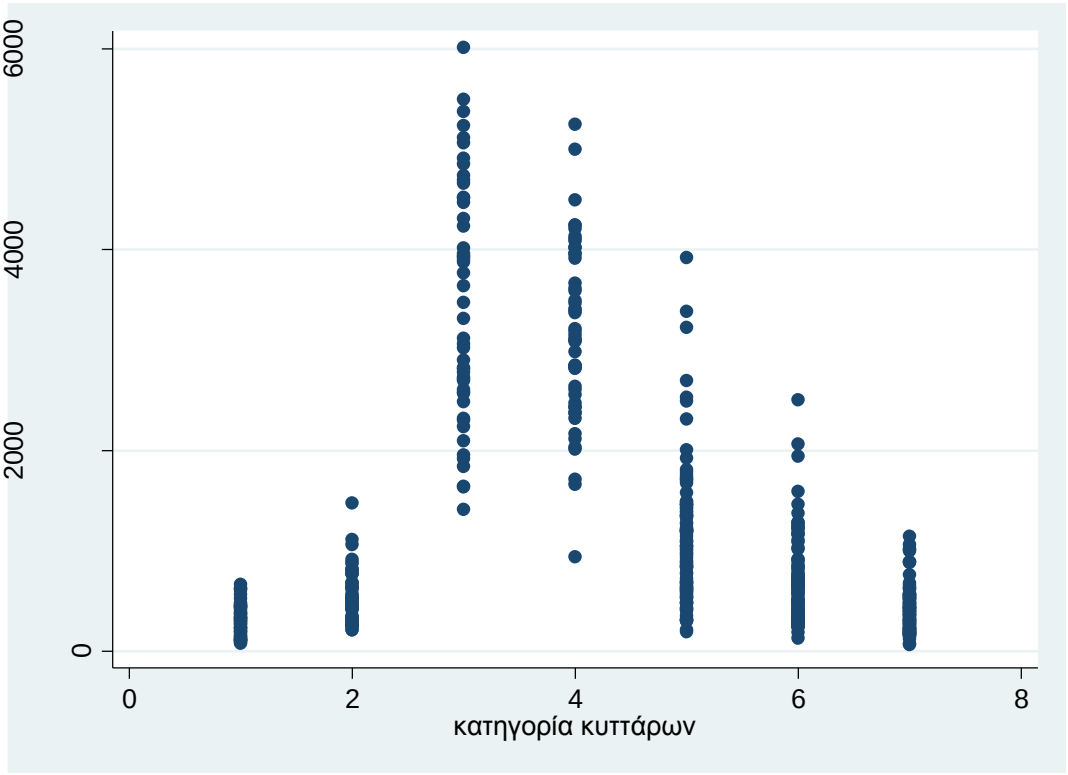
Σε αυτήν την παράγραφο παρατείθενται βασικοί περιγραφικοί δείκτες για τα δεδομένα της παλαιάς βάσης έτσι ώστε να επιτευχθεί μια γενική περιγραφή των κυττάρων των διαφόρων κατηγοριών για μια

πρώτη εκτίμηση. Η μέση τιμή, η τυπική απόκλιση, το μέγιστο και το ελάχιστο υπολογίστηκαν για όλες τις παρατηρήσεις μαζί αλλά και ξεχωριστά για κάθε κατηγορία κατάταξης των παρατηρήσεων.

Οι αναλυτικοί Πίνακες 4.1-4.8 με τους περιγραφικούς δείκτες βρίσκονται στο παράρτημα. Σύμφωνα με τους δείκτες, τα υγιή κύτταρα έχουν σημαντικά μικρότερο πυρήνα σε σχέση με τα μη υγιή. Ιδιαίτερα η τιμή του εμβαδού του πυρήνα έχει απότομη αύξηση αφού για τις τέσσερις κατηγορίες των υγιών κυττάρων έχει μέση τιμή 89,34 88,44 64,73 και 30,90 μm^2 αντίστοιχα και για τις τρεις κατηγορίες των μη υγιών κυττάρων αυξάνει σε 258,39 263,03 και 223,48 μm^2 αντίστοιχα. Ο διαχωρισμός των κυττάρων στις δύο μεγάλες κατηγορίες υγιών και μη υγιών είναι ξεκάθαρος μέσω του μεγέθους του πυρήνα αλλά είναι αρκετά δύσκολος να επιτευχθεί ανάμεσα στις τρεις κατηγορίες δυσπλασίας.

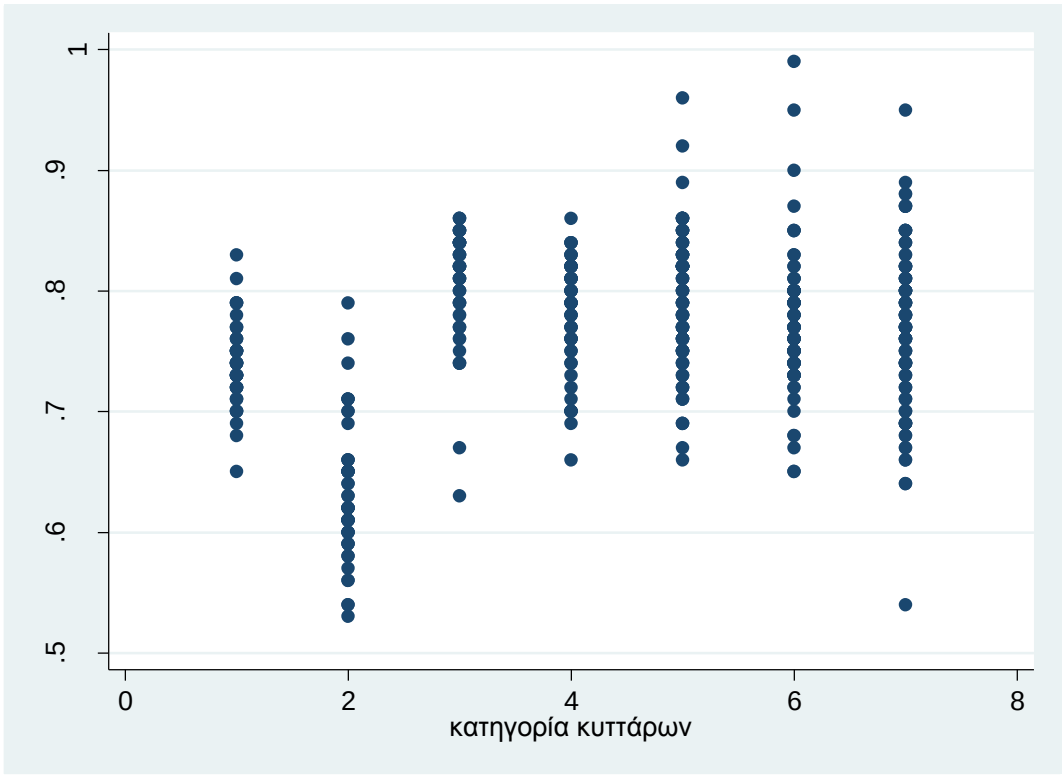


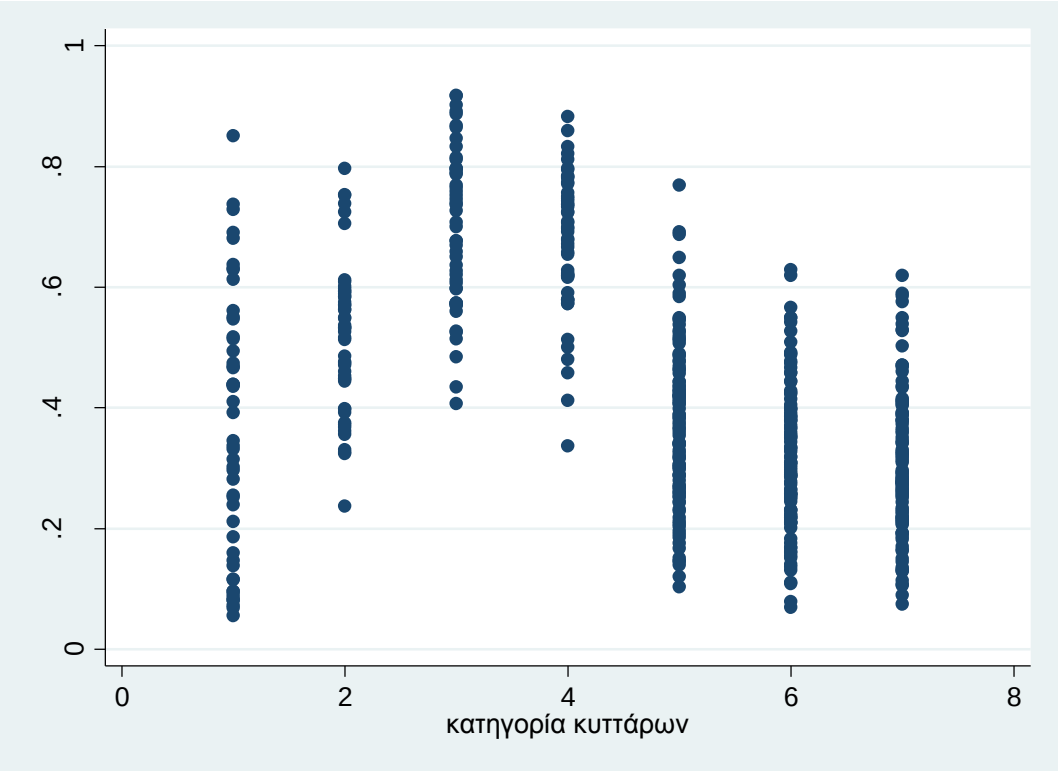
Στις κατηγορίες των υγιών κυττάρων παρατηρείται μεγάλη διαφορά στην τιμή του εμβαδού του κυτταροπλάσματος. Οι δύο πρώτες κατηγορίες έχουν εντυπωσιακά μικρότερο εμβαδόν κυτταροπλάσματος από την τρίτη και τέταρτη κατηγορία με μέσες τιμές 309,25 553,67 και 3482,24 3166,41 μm^2 αντίστοιχα. Στις κατηγορίες των μη υγιών κυττάρων παρατηρείται μείωση της τιμής της μεταβλητής σε 1110,94 664,30 και 405,10 μm^2 αλλά και πάλι δεν βρίσκεται κάτω από τα επίπεδα που βρίσκεται στις δύο πρώτες κατηγορίες των υγιών κυττάρων. Οι δύο πρώτες κατηγορίες των υγιών κυττάρων φαίνεται να αλληλοκαλύπτονται με τις δύο τελευταίες των μη υγιών.



Οι τιμές της φωτεινότητας του πυρήνα και του κυτταροπλάσματος επίσης φαίνεται να αυξάνουν ελαφγρά στα κύτταρα με κάποια μορφή δυσπλασίας σε σχέση με τα υγιή κύτταρα κι έτσι παρατηρείται αλληλοεπικάλυψη των κατηγοριών.

Τέλος, καθώς περνάμε από τα υγιή στα μη υγιή κύτταρα μειώνεται η τιμή της σφαιρικότητας του κυτταροπλάσματος καθώς το κυτταρόπλασμα των μη υγιών κυττάρων τείνει να χάσει το σφαρικό του σχήμα.





4.1.2 Συσχετίσεις μεταξύ όλων των χαρακτηριστικών της βάσης δεδομένων (correlation)

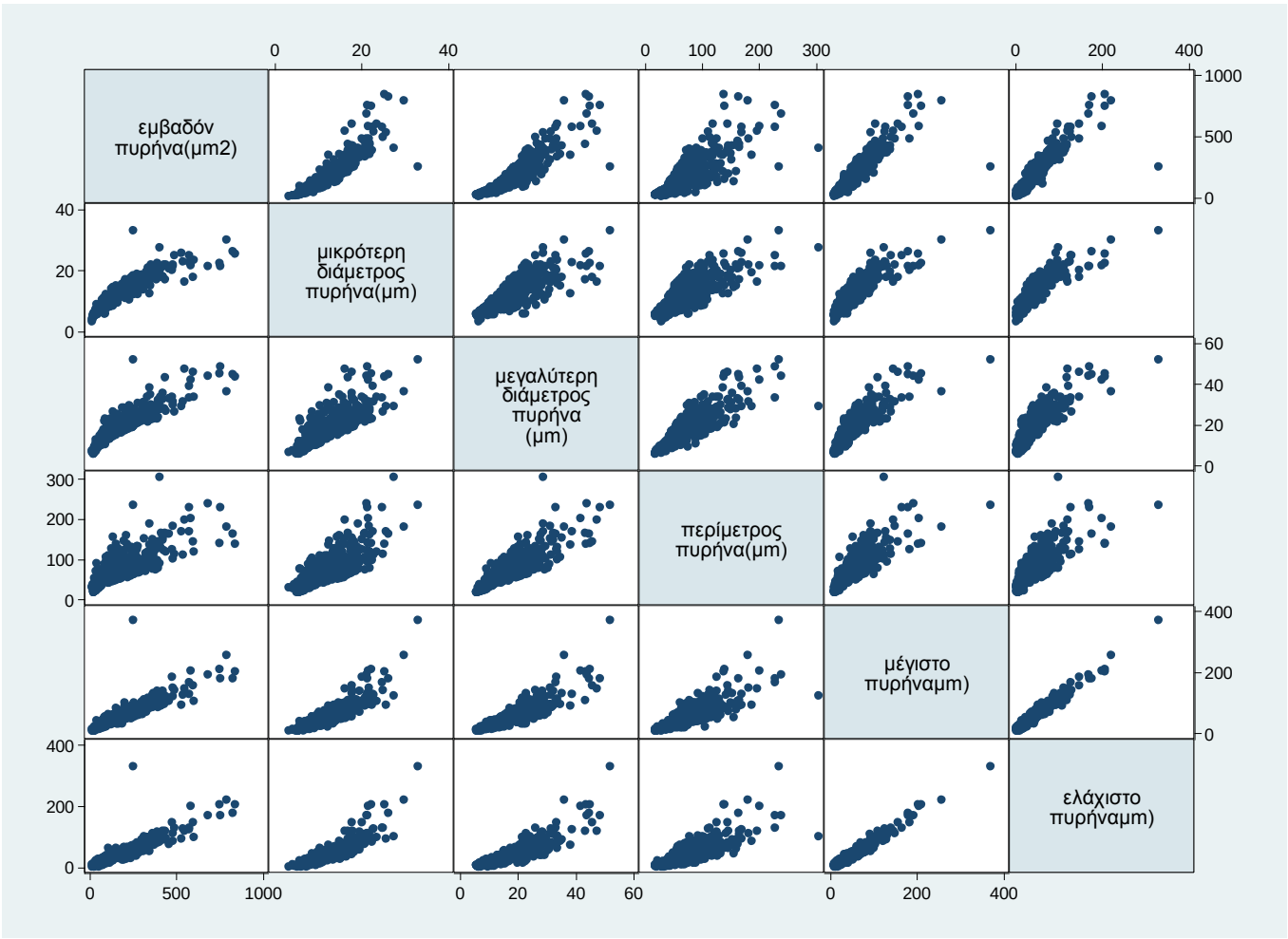
Στον Πίνακα 4.9 του παραρτήματος υπάρχουν όλες οι συσχετίσεις μεταξύ των είκοσι μεταβλητών της παλαιάς βάσης δεδομένων. Ανάμεσα σε αυτές αναφέρονται οι συσχετίσεις με συντελεστή συσχέτισης πάνω από 0,7 που είναι και οι πιο ισχυρές.

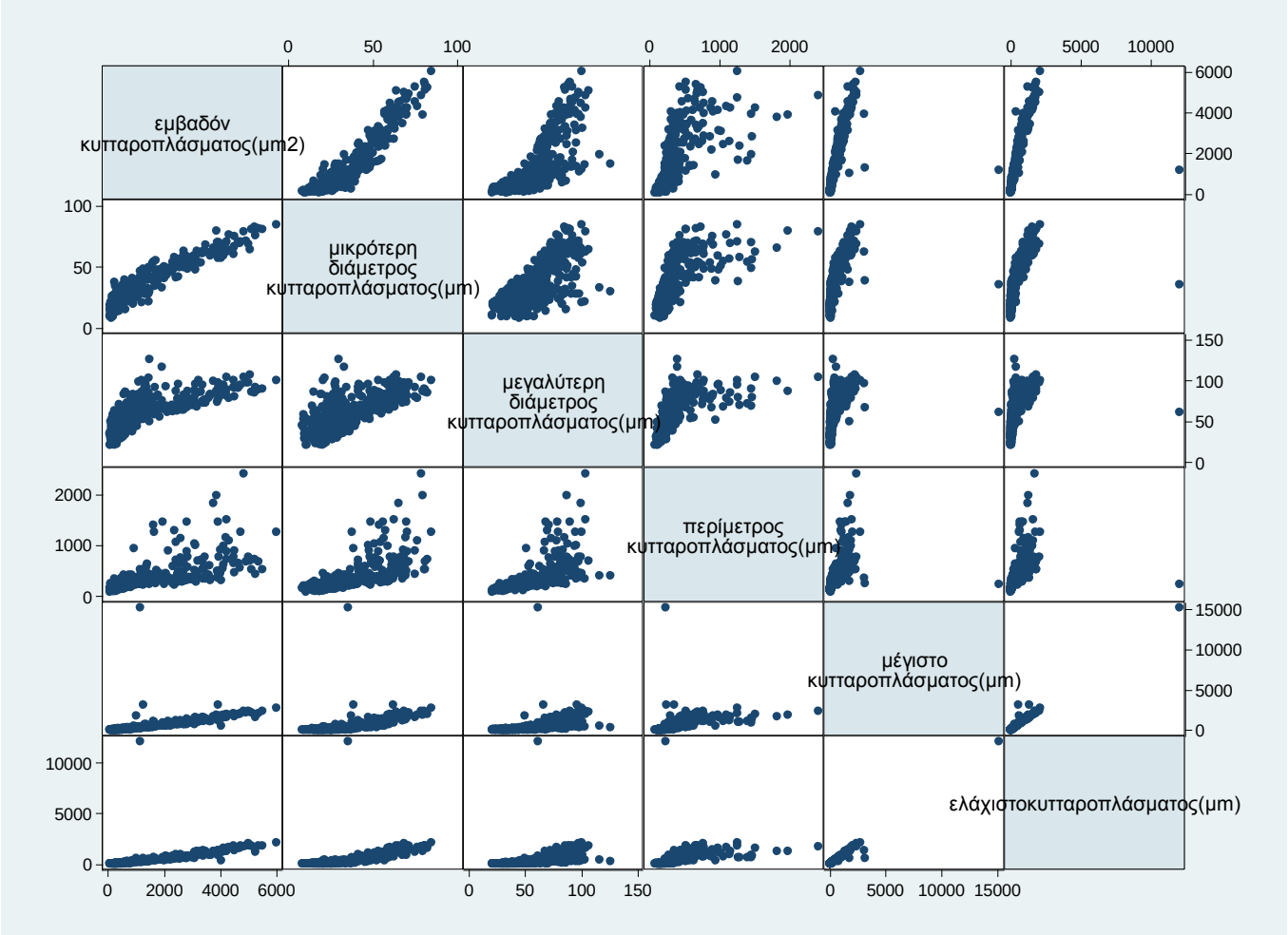
Έτσι, φαίνεται ότι το εμβαδόν του πυρήνα είναι ισχυρά συσχετισμένο με την μικρότερη και μεγαλύτερη διάμετρο πυρήνα, με την περίμετρο πυρήνα και με το ελάχιστο και μέγιστο πυρήνα , γεγονός λογικό αφού τα μεγέθη είναι εξαρτώμενα.

Με τα αντίστοιχα μεγέθη είναι συσχετισμένο και το εμβαδόν κυτταροπλάσματος ,με λίγο πιο μικρές τιμές του συντελεστή συσχέτισης.

Η μικρότερη και η μεγαλύτερη διάμετρος του πυρήνα και του κυτταροπλάσματος είναι ισχυρά συσχετισμένες με την περίμετρο,το ελάχιστο και το μέγιστο πυρήνα και κυτταροπλάσματος αντίστοιχα αφού είναι εξαρτώμενα μεταξύ τους.

Οι ισχυρές συσχετίσεις που αναφέροναται, παρουσιάζονται και στα παρακάτω διαγράμματα:


εμβαδόν πυρήνα(μm2)
μικρότερη διάμετρος πυρήνα(μm)
μεγαλύτερη διάμετρος πυρήνα (μm)
περίμετρος πυρήνα(μm)
μέγιστο πυρήναμm)
ελάχιστο πυρήναμm)
0
20
40
0
100
200
300
0
200
400
1000
500
0
40
20
0
60
40
20
0
300
200
100
0
400
200
0
400
200
0
0
500
1000
0
20
40
60
0
200
400


εμβαδόν κυτταροπλάσματος(μm2)
μικρότερη διάμετρος κυτταροπλάσματος(μm)
μεγαλύτερη διάμετρος κυτταροπλάσματος(μm)
περίμετρος κυτταροπλάσματος(μm)
μέγιστο κυτταροπλάσματος(μm)
ελάχιστοκυτταροπλάσματος(μm)
0
50
100
0
1000
2000
0
5000
10000
6000
4000
2000
0
100
50
0
150
100
50
0
2000
1000
0
15000
10000
5000
0
10000
5000
0
0
2000
4000
6000
0
50
100
150
0
5000
10000
15000

4.1.3 έλεγχος διαφοράς μέσων τιμών των χαρακτηριστικών ανάμεσα στις κατηγορίες υγιών και μη υγιών κυττάρων (t-test)

μεταβλητή	Τιμή ελέγχου	p-value
Εμβαδόν πυρήνα	-18.807	<0.001
Εμβαδόν κυτταροπλάσματος	11.212	<0.001
Αναλογία πυρήνα/κυτταροπλάσματος	-14.454	<0.001
Φωτεινότητα πυρήνα	-14.242	<0.001
Φωτεινότητα κυτταροπλάσματος	-6.167	<0.001
Μικρότερη διάμετρος πυρήνα	-23.459	<0.001
Μεγαλύτερη διάμετρος πυρήνα	-20.990	<0.001
Επιμήκυνση πυρήνα	3.453	<0.001
Σφαιρικότητα πυρήνα	3.096	0.0021
Μικρότερη διάμετρος κυτταροπλάσματος	8.088	<0.001
Μεγαλύτερη διάμετρος κυτταροπλάσματος	3.019	0.0027
Επιμήκυνση κυτταροπλάσματος	5.603	<0.001
Σφαιρικότητα κυτταροπλάσματος	15.351	<0.001
Περίμετρος πυρήνα	-15.192	<0.001
Περίμετρος κυτταροπλάσματος	7.459	<0.001
Θέση πυρήνα	0.530	0.5961
Μέγιστο πυρήνα	-16.437	<0.001
Ελάχιστο πυρήνα	-15.724	<0.001
Μέγιστο κυτταροπλάσματος	5.250	<0.001
Ελάχιστο κυτταροπλάσματος	5.324	<0.001

Πίνακας 4.10: Αποτελέσματα στατιστικού ελέγχου για την διαφορά των μέσων τιμών των χαρακτηριστικών ανάμεσα στις 2 κατηγορίες, υγιών και μη υγιών κυττάρων

Όπως φαίνεται στον Πίνακα 4.10 οι διαφορές των μέσων τιμών των χαρακτηριστικών ανάμεσα στις κατηγορίες υγιών και μη υγιών κυττάρων είναι στατιστικά σημαντικές σε επίπεδο σημαντικότητας 95%,εκτός από την μεταβλητή που αντιπροσωπεύει την θέση του πυρήνα. Το χαρακτηριστικό αυτό δεν φαίνεται να διαφοροποιείται σημαντικά ανάμεσα στις 2 κατηγορίες .

4.1.4 ανάλυση διασποράς κάθε χαρακτηριστικού ανάμεσα στις 7 κατηγορίες των κυττάρων (oneway ANOVA)

μεταβλητή	R^2-προσαρμοσμένο (R^2-adjusted)	Στατιστικό F και p-value	Σύγκριση κατηγοριών με την κατηγορία 7(σε 95% επίπεδο εμπιστοσύνης)
Εμβαδόν πυρήνα	0.430	63.88(<0.001)	Σ.σ.
Εμβαδόν κυτταροπλάσματος	0.766	274.45(<0.001)	Σ.σ. εκτός από τις κατηγορίες 1,2
Αναλογία πυρήνα/κυτταροπλάσματος	0.512	88.40(<0.001)	Σ.σ.
Φωτεινότητα πυρήνα	0.475	76.27(<0.001)	Σ.σ. εκτός από τις κατηγορίες 1,5,6
Φωτεινότητα κυτταροπλάσματος	0.432	64.33(<0.001)	Σ.σ. εκτός από τις κατηγορίες 4,6
Μικρότερη διάμετρος πυρήνα	0.559	106.49(<0.001)	Σ.σ.
Μεγαλύτερη διάμετρος πυρήνα	0.518	90.69(<0.001)	Σ.σ. εκτός από τις κατηγορίες 5,6
Επιμήκυνση πυρήνα	0.129	13.38(<0.001)	Σ.σ. εκτός από τις κατηγορίες 2,5,6
Σφαιρικότητα πυρήνα	0.118	12.19(<0.001)	Σ.σ. εκτός από τις κατηγορίες 2,5,6
Μικρότερη διάμετρος κυτταροπλάσματος	0.723	218.78(<0.001)	Σ.σ. εκτός από τις κατηγορίες 2
Μεγαλύτερη διάμετρος κυτταροπλάσματος	0.532	95.72(<0.001)	Σ.σ.
Επιμήκυνση κυτταροπλάσματος	0.178	19.10(<0.001)	Σ.σ. εκτός από τις κατηγορίες 5,6
Σφαιρικότητα κυτταροπλάσματος	0.512	88.42(<0.001)	Σ.σ. εκτός από τις κατηγορίες 6
Περίμετρος πυρήνα	0.408	58.45(<0.001)	Σ.σ.
Περίμετρος κυτταροπλάσματος	0.452	69.62(<0.001)	Σ.σ. εκτός από τις κατηγορίες 1,2
Θέση πυρήνα	0.217	24.10(<0.001)	Σ.σ. εκτός από τις κατηγορίες 6
Μέγιστο πυρήνα	0.371	50.15(<0.001)	Σ.σ. εκτός από τις κατηγορίες 5
Ελάχιστο πυρήνα	0.341	44.17(<0.001)	Σ.σ. εκτός από τις κατηγορίες 5
Μέγιστο κυτταροπλάσματος	0.243	27.75(<0.001)	Σ.σ. εκτός από τις κατηγορίες 1,2
Ελάχιστο κυτταροπλάσματος	0.245	28.06(<0.001)	Σ.σ. εκτός από τις κατηγορίες 1,2

Πίνακας 4.11: Αποτελέσματα ανάλυσης διασποράς**.** Η ποσότητα R^2-προσαρμοσμένο αν εκφραστεί σε ποσοστό % δείχνει το ποσοστό της μεταβλητότητας των κατηγοριών που εξηγείται από την εκάστοτε μεταβλητή . Η συντομογραφία Σ.σ. σημαίνει στατιστικά σημαντική

Από τον Πίνακα 4.11 προκύπτει ότι το εμβαδόν του κυτταροπλάσματος και η μικρότερη διάμετρος του κυτταροπλάσματος εξηγούν αρκετά μεγάλο ποσοστό της μεταβλητότητας που οφείλεται στις 7 διαφορετικές κατηγορίες των κυττάρων, η κάθε μία από μόνη της, χωρίς να έχει συμπεριληφθεί καμία

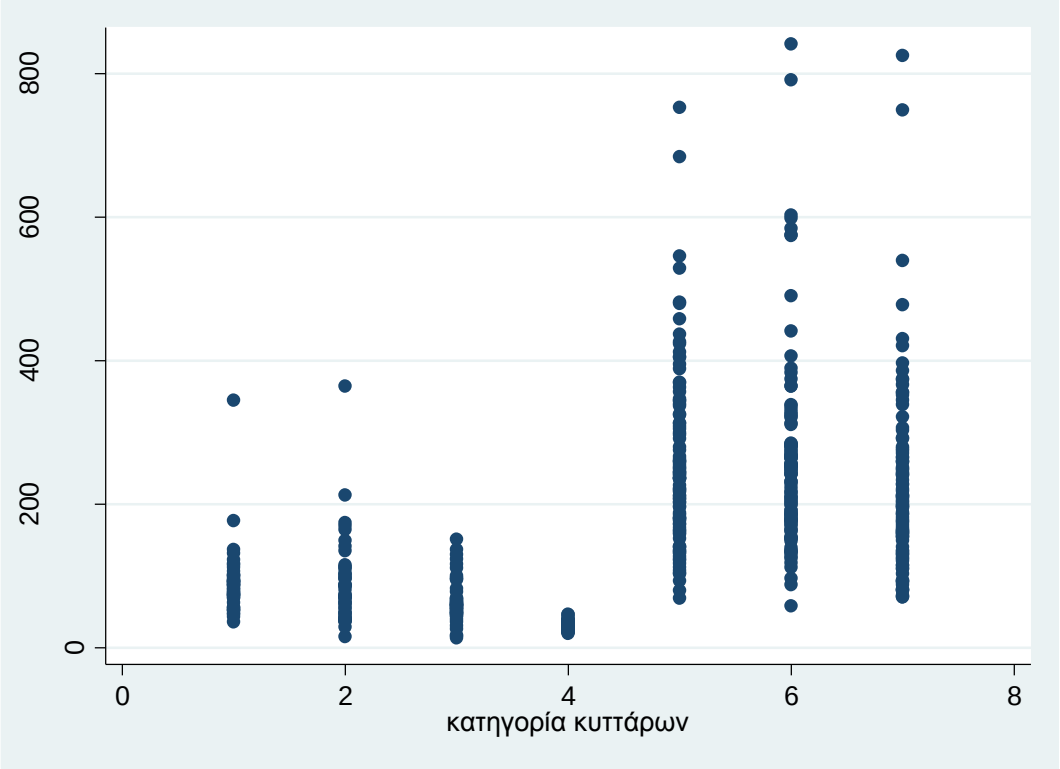
άλλη μεταβλητή στην ανάλυση . Είναι επίσης φανερό ότι οι κατηγορίες των κυττάρων δεν είναι ικανοποιητικά διαχωρίσιμες μεταξύ τους. Παρατηρούνται συγχύσεις, κυρίως μεταξύ των κατηγοριών 2,5,6 με την κατηγορία 7 . Βέβαια αυτό συνάγεται μόνο από την χρήση κάθε μεταβλητής ξεχωριστά στην προσπάθεια διαχωρισμού των κυττάρων στις 7 κατηγορίες και δεν αποτελεί γενικό συμπέρασμα .

4. 2 Νέα βάση δεδομένων επιχρίσματος Παπανικολάου

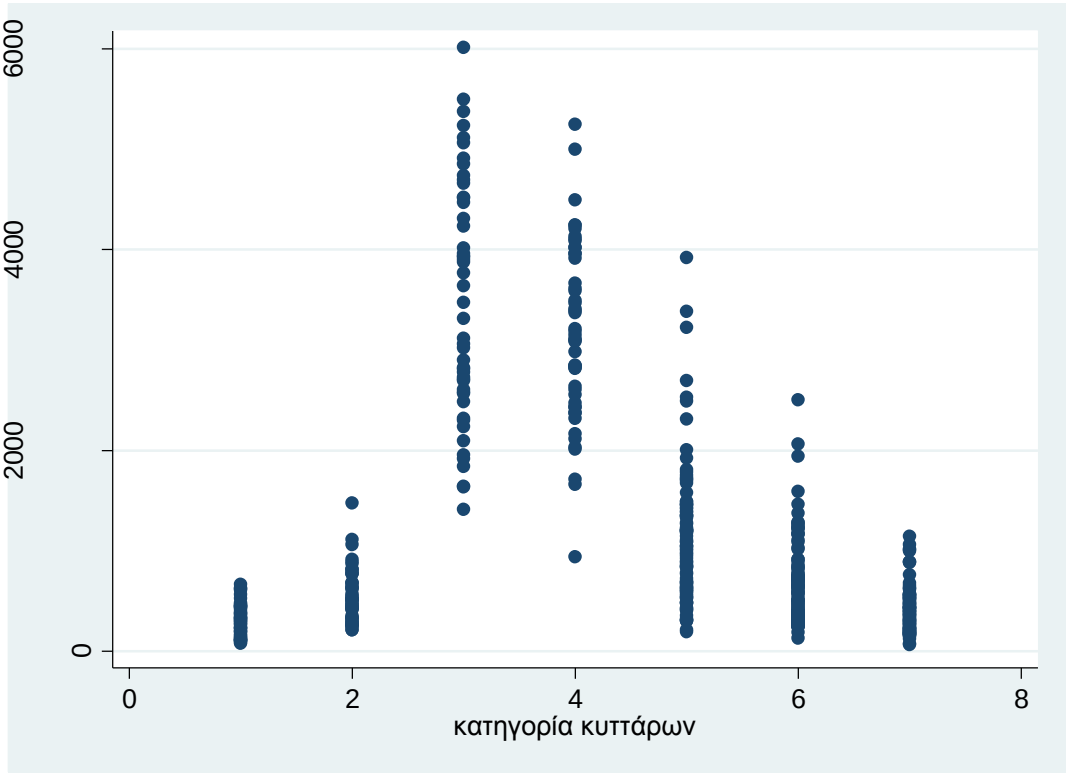
4. 2. 1 Περιγραφικά στοιχεία

Σε αυτήν την παράγραφο παρατείθενται βασικοί περιγραφικοί δείκτες για τα δεδομένα της νέας βάσης έτσι ώστε να επιτευχθεί μια γενική περιγραφή των κυττάρων των διαφόρων κατηγοριών για μια πρώτη εκτίμηση. Η μέση τιμή, η τυπική απόκλιση, το μέγιστο και το ελάχιστο υπολογίστηκαν για όλες τις παρατηρήσεις μαζί αλλά και ξεχωριστά για κάθε κατηγορία κατάταξης των παρατηρήσεων.

Οι αναλυτικοί Πίνακες 4.12-4.19 με τους περιγραφικούς δείκτες βρίσκονται στο παράρτημα. Σύμφωνα με τους δείκτες, τα υγιή κύτταρα έχουν σημαντικά μικρότερο πυρήνα σε σχέση με τα μη υγιή. Η τιμή του εμβαδού του πυρήνα έχει απότομη αύξηση αφού για τις τρεις κατηγορίες των υγιών κυττάρων έχει μέση τιμή 2990,83 630,87 και 1315,33 μm^2 αντίστοιχα και για τις τέσσερις κατηγορίες των μη υγιών κυττάρων αυξάνει σε 1591,43 4690,10 3872,80 και 2948,99 μm^2 αντίστοιχα.Ο διαχωρισμός των κυττάρων στις δύο μεγάλες κατηγορίες υγιών και μη υγιών είναι ξεκάθαρος μέσω του μεγέθους του πυρήνα αλλά είναι αρκετά δύσκολος να επιτευχθεί ανάμεσα στις τρεις κατηγορίες δυσπλασίας.

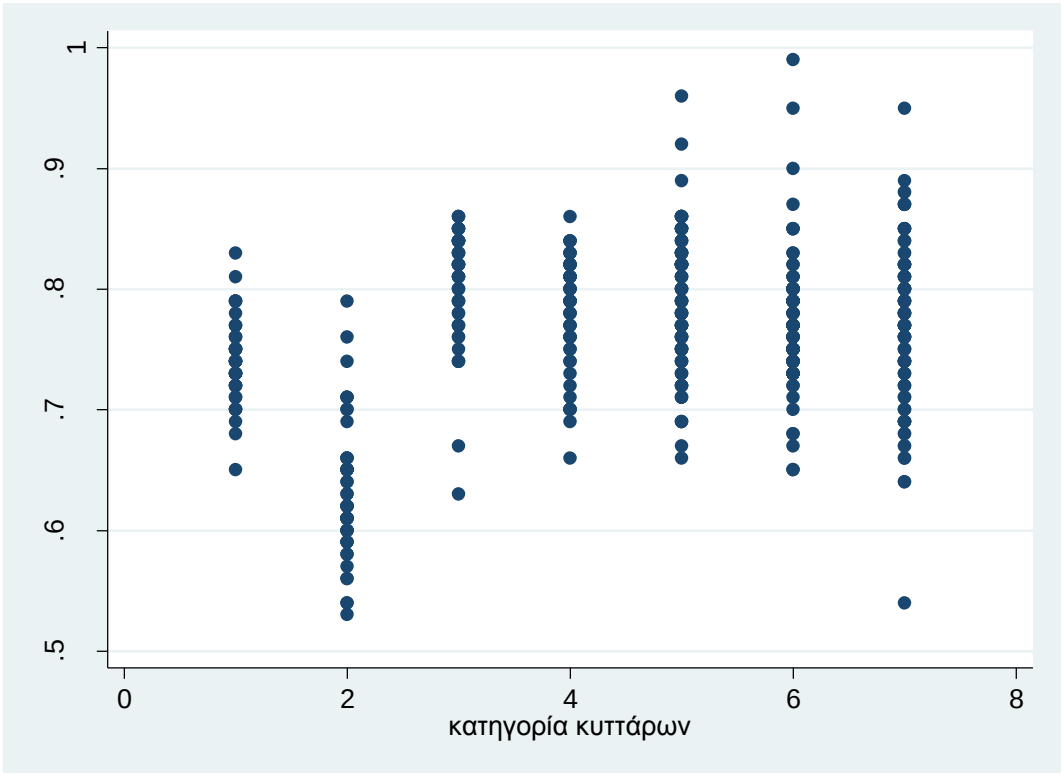


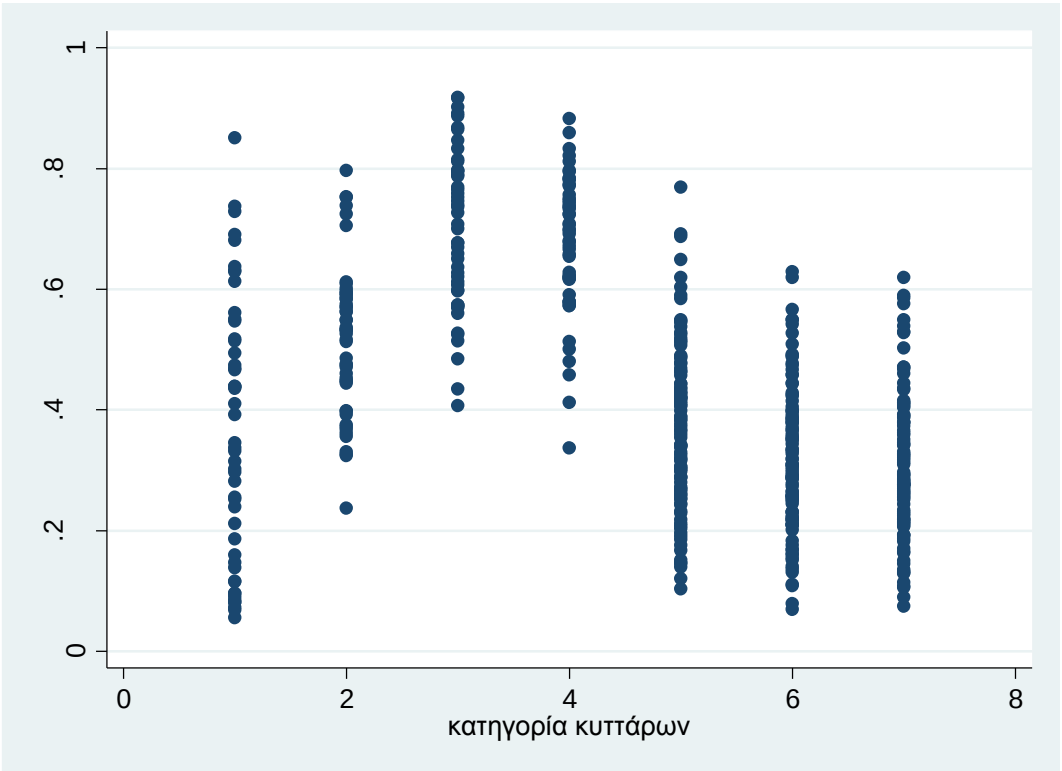
Στις κατηγορίες των υγιών κυττάρων παρατηρείται μεγάλη διαφορά στην τιμή του εμβαδού του κυτταροπλάσματος. Οι δύο πρώτες κατηγορίες έχουν κατά μέσο όρο μεγαλύτερο εμβαδόν κυτταροπλάσματος από την τρίτη κατηγορία υγιών κυττάρων καθώς και από τις τρεις από τις τέσσερις κατηγορίες μη υγιών κυττάρων με τιμές 14053,90 61487,28 και 44961,50 μm^2 για τα υγιή κύτταρα και 3289,86 15458,71 7288,19 και 3415,32 μm^2 για τα μη υγιή αντίστοιχα.



Οι τιμές της φωτεινότητας του πυρήνα και του κυτταροπλάσματος επίσης φαίνεται να αυξάνουν ελαφγρά στα κύτταρα με κάποια μορφή δυσπλασίας σε σχέση με τα υγιή κύτταρα κι έτσι παρατηρείται αλληλοεπικάλυψη των κατηγοριών.

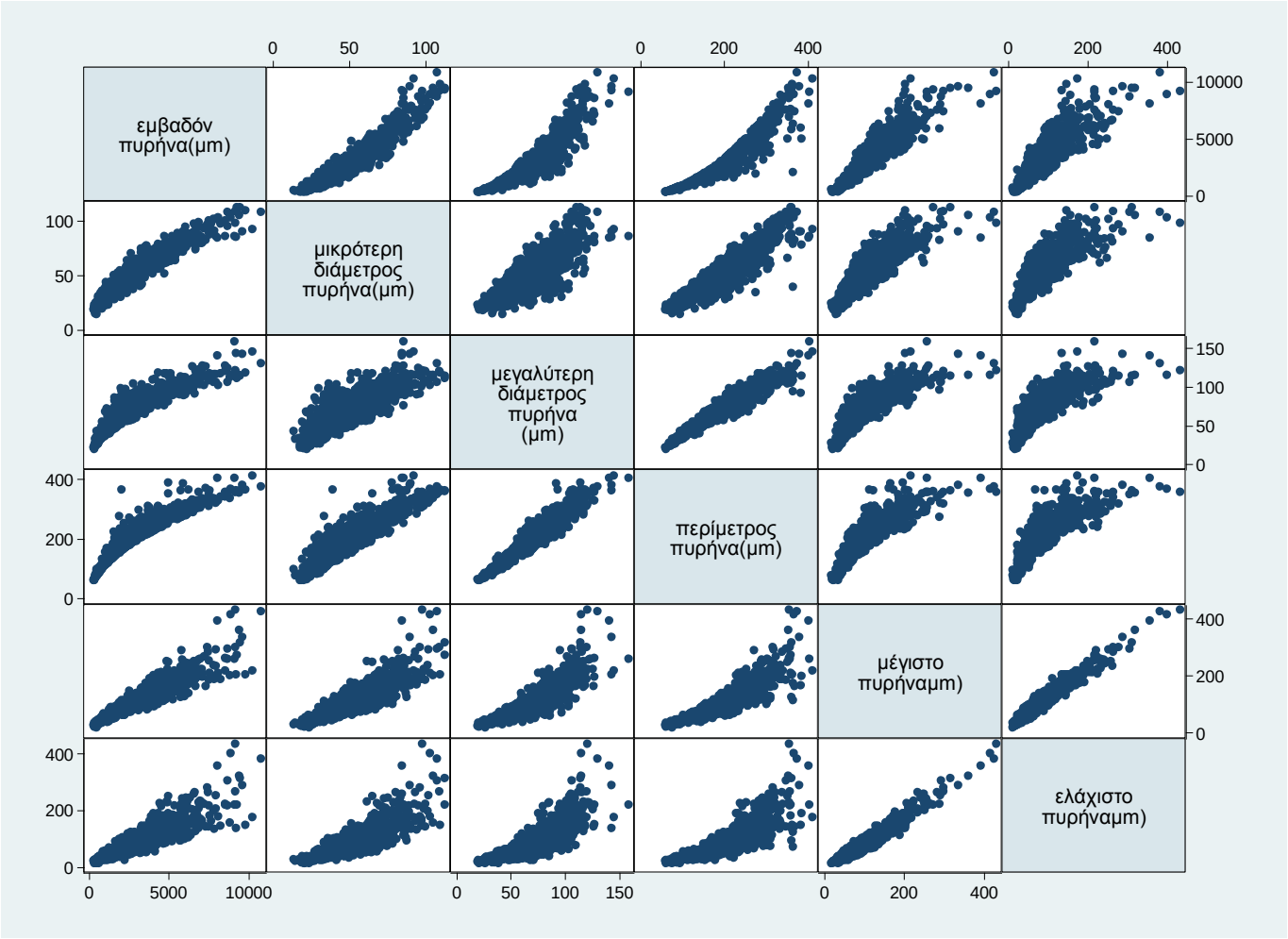
Τέλος, καθώς περνάμε από τα υγιή στα μη υγιή κύτταρα μειώνεται η τιμή της σφαιρικότητας του κυτταροπλάσματος καθώς το κυτταρόπλασμα των μη υγιών κυττάρων τείνει να χάσει το σφαρικό του σχήμα.

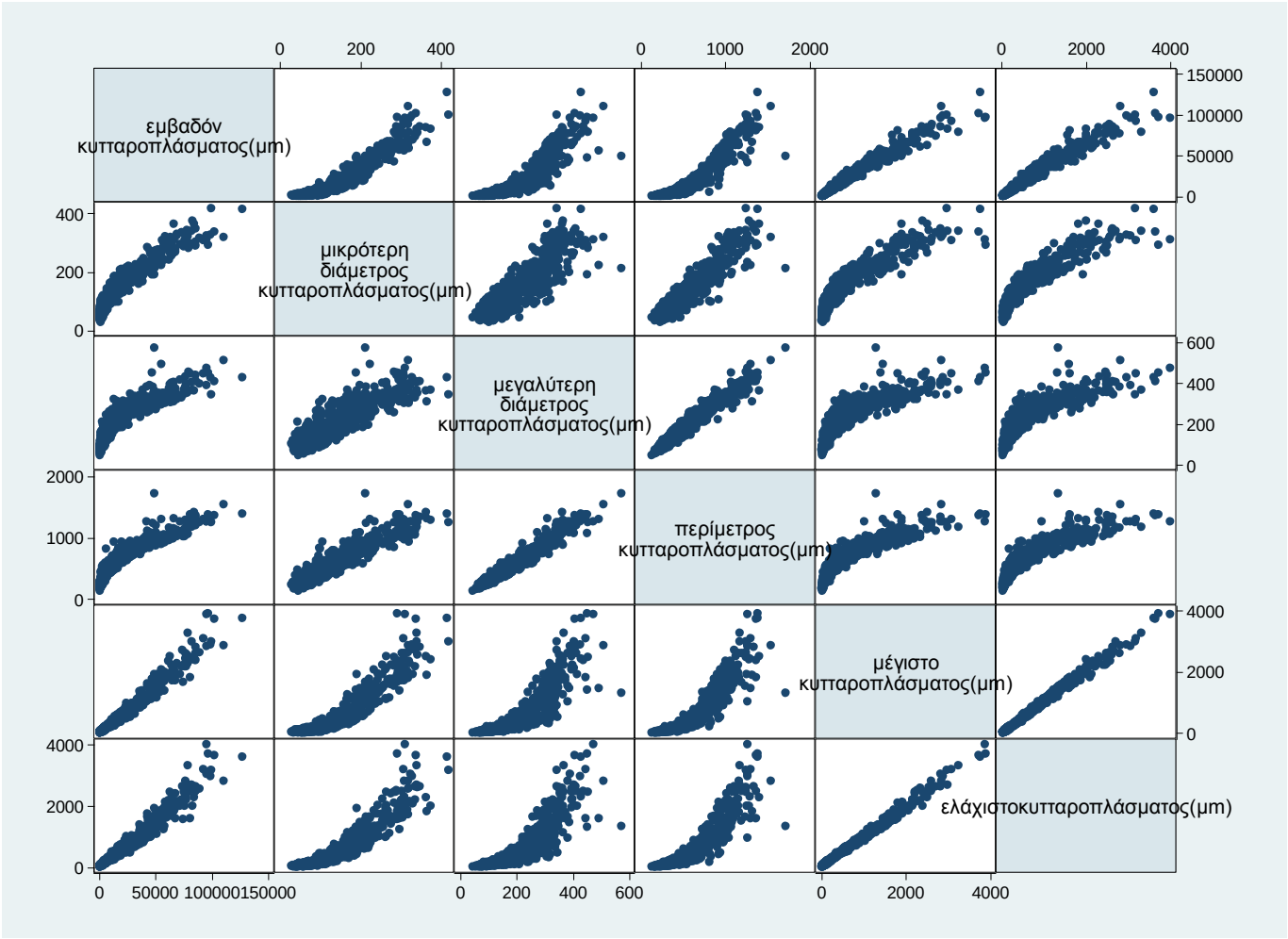

1
.9
.8
.7
.6
.5
0
2
4
6
8
κατηγορία κυττάρων


1
.8
.6
.4
.2
0
0
2
4
6
8
κατηγορία κυττάρων

4.2.2 Συσχετίσεις μεταξύ όλων των χαρακτηριστικών της βάσης δεδομένων (correlation)

Όπως και στην παράγραφο 4.1.2, οι ισχυρές συσχετίσεις και στη νέα βάση δεδομένων είναι μεταξύ εξαρτώμενων χαρακτηριστικών και φαίνονται στα διαγράμματα που ακολουθούν. Ο Πίνακας 4.20 είναι ο αναλυτικός πίνακας όλων των συσχετίσεων των μεταβλητών και βρίσκεται στο παράρτημα.




εμβαδόν κυτταροπλάσματος(μm)
μικρότερη διάμετρος κυτταροπλάσματος(μm)
μεγαλύτερη διάμετρος κυτταροπλάσματος(μm)
περίμετρος κυτταροπλάσματος(μm)
μέγιστο κυτταροπλάσματος(μm)
ελάχιστοκυτταροπλάσματος(μm)
0
200
400
0
1000
2000
0
2000
4000
150000
100000
50000
0
400
200
0
600
400
200
0
2000
1000
0
4000
2000
0
4000
2000
0
0
50000
100000
150000
0
200
400
600
0
2000
4000

4.2.3 Έλεγχος διαφοράς μέσων τιμών των χαρακτηριστικών ανάμεσα στις κατηγορίες υγιών και μη υγιών κυττάρων (t-test)

μεταβλητή	Τιμή ελέγχου	p-value
Εμβαδόν πυρήνα	-20.742	<0.001
Εμβαδόν κυτταροπλάσματος	20.534	<0.001
Αναλογία πυρήνα/κυτταροπλάσματος	-21.114	<0.001
Φωτεινότητα πυρήνα	-11.429	<0.001
Φωτεινότητα κυτταροπλάσματος	-3.310	<0.001
Μικρότερη διάμετρος πυρήνα	-22.046	<0.001
Μεγαλύτερη διάμετρος πυρήνα	-27.126	<0.001
Επιμήκυνση πυρήνα	3.156	0.0017
Σφαιρικότητα πυρήνα	5.153	<0.001
Μικρότερη διάμετρος κυτταροπλάσματος	15.917	<0.001
Μεγαλύτερη διάμετρος κυτταροπλάσματος	14.976	<0.001
Επιμήκυνση κυτταροπλάσματος	0.480	0.6310
Σφαιρικότητα κυτταροπλάσματος	15.090	<0.001
Περίμετρος πυρήνα	-26.555	<0.001
Περίμετρος κυτταροπλάσματος	16.099	<0.001
Θέση πυρήνα	-2.277	0.0230
Μέγιστο πυρήνα	-16.667	<0.001
Ελάχιστο πυρήνα	-18.526	<0.001
Μέγιστο κυτταροπλάσματος	20.114	<0.001
Ελάχιστο κυτταροπλάσματος	19.231	<0.001

Πίνακας 4.21: Αποτελέσματα στατιστικού ελέγχου για την διαφορά των μέσων τιμών των χαρακτηριστικών ανάμεσα στις 2 κατηγορίες, υγιών και μη υγιών κυττάρων

Όπως φαίνεται από τον Πίνακα 4.21, υπάρχει στατιστικά σημαντική διαφορά στις μέσες τιμές όλων των μεταβλητών μεταξύ των κατηγοριών υγιών και μη υγιών κυττάρων.

4.2.4 Ανάλυση διασποράς κάθε χαρακτηριστικού ανάμεσα στις 7 κατηγορίες των κυττάρων (oneway ANOVA)

μεταβλητή	R^2-προσαρμοσμένο (R^2-adjusted)	Στατιστικό F και p-value	Σύγκριση κατηγοριών με την κατηγορία 7(σε 95% επίπεδο εμπιστοσύνης)
Εμβαδόν πυρήνα	0.442	121.93(<0.001)	Σ.σ. εκτός από τις κατηγορίες 6
Εμβαδόν κυτταροπλάσματος	0.776	532.26(<0.001)	Σ.σ. εκτός από τις κατηγορίες 3,6
Αναλογία πυρήνα/κυτταροπλάσματος	0.718	390.90(<0.001)	Σ.σ.
Φωτεινότητα πυρήνα	0.230	46.78(<0.001)	Σ.σ. εκτός από τις κατηγορίες 3,4,6
Φωτεινότητα κυτταροπλάσματος	0.022	4.50(<0.001)	Σ.σ. εκτός από τις κατηγορίες 6
Μικρότερη διάμετρος πυρήνα	0.485	145.18(<0.001)	Σ.σ. εκτός από τις κατηγορίες 6
Μεγαλύτερη διάμετρος πυρήνα	0.534	176.38(<0.001)	Σ.σ. εκτός από τις κατηγορίες 6
Επιμήκυνση πυρήνα	0.068	12.30(<0.001)	Σ.σ. εκτός από τις κατηγορίες 6
Σφαιρικότητα πυρήνα	0.111	20.21(<0.001)	Σ.σ. εκτός από τις κατηγορίες 3,6
Μικρότερη διάμετρος κυτταροπλάσματος	0.770	514.01(<0.001)	Σ.σ.
Μεγαλύτερη διάμετρος κυτταροπλάσματος	0.689	339.54(<0.001)	Σ.σ. εκτός από τις κατηγορίες 4,5
Επιμήκυνση κυτταροπλάσματος	0.098	17.67(<0.001)	Σ.σ.
Σφαιρικότητα κυτταροπλάσματος	0.533	175.55(<0.001)	Σ.σ. εκτός από τις κατηγορίες 6
Περίμετρος πυρήνα	0.538	178.78(<0.001)	Σ.σ. εκτός από τις κατηγορίες 3
Περίμετρος κυτταροπλάσματος	0.717	389.51(<0.001)	Σ.σ. εκτός από τις κατηγορίες 4,5,6
Θέση πυρήνα	0.106	19.19(<0.001)	Σ.σ. εκτός από τις κατηγορίες 6
Μέγιστο πυρήνα	0.354	84.67(<0.001)	Σ.σ. εκτός από τις κατηγορίες 6
Ελάχιστο πυρήνα	0.298	65.82(<0.001)	Σ.σ. εκτός από τις κατηγορίες 3,6
Μέγιστο κυτταροπλάσματος	0.739	435.37(<0.001)	Σ.σ. εκτός από τις κατηγορίες 1,2
Ελάχιστο κυτταροπλάσματος	0.729	412.97(<0.001)	Σ.σ. εκτός από τις κατηγορίες 1,2

Πίνακας 4.22: Αποτελέσματα ανάλυσης διασποράς**.** Η ποσότητα R^2-προσαρμοσμένο αν εκφραστεί σε ποσοστό % δείχνει το ποσοστό της μεταβλητότητας των κατηγοριών που εξηγείται από την εκάστοτε μεταβλητή .Η συντομογραφία Σ.σ. σημαίνει στατιστικά σημαντική

Από τον Πίνακα 4.22 φαίνεται ότι το εμβαδόν του κυτταροπλάσματος, η αναλογία πυρήνα/κυτταροπλάσματος, η μικρότερη και η μεγαλύτερη διάμετρος κυτταροπλάσματος, η περίμετρος κυτταροπλάσματος, το μέγιστο και το ελάχιστο κυτταροπλάσματος έχουν αρκετά σημαντικό ποσοστό διαχωριστικής ικανότητας των κυττάρων στις 7 κατηγορίες. Παρατηρείται σύγχιση μεταξύ των

κατηγοριών 6 και 7 που είναι λογικό αλλά και μεταξύ των κατηγοριών των φυσιολογικών κυττάρων με την κατηγορία 7 των μη φυσιολογικών .

ΚΕΦΑΛΑΙΟ 5

Αποτελέσματα αλγορίθμων Στατιστικής

Οι στατιστικές μέθοδοι για ανάλυση με πολλούς παράγοντες που αναπτύχθηκαν στο κεφάλαιο 3, εφαρμόστηκαν στα δεδομένα των 2 βάσεων επιχρίσματος Παπανικολάου. Τα αποτελέσματα των μεθόδων αυτών παρουσιάζονται αναλυτικά σε αυτό το κεφάλαιο

5.1 Λογιστική παλινδρόμηση (logistic regression)

Η λογιστική παλινδρόμηση δημιουργεί ένα μοντέλο που η εξαρτημένη μεταβλητή του είναι δίτιμη. Στο συγκεκριμένο πρόβλημα η μεταβλητή αυτή δείχνει τα υγιή όταν παίρνει την τιμή μηδέν και τα μη υγιή κύτταρα όταν παίρνει την τιμή ένα. Στην ανάλυση έγινε επιλογή χαρακτηριστικών με μέθοδο stepwise και διατηρήθηκαν τελικά 7 χαρακτηριστικά για την παλαιά βάση και 11 για τη νέα. Στον Πίνακα 5.1 φαίνονται τα αποτελέσματα της παλινδρόμησης.

	Παλαιά βάση δεδομένων			**Νέα βάση δεδομένων**		
μεταβλητή	**Συντελεστής**	**p-value**	**Λόγος συμπληρωματικών πιθανοτήτων**	**Συντελεστή**	**p-value**	**Λόγος συμπληρωματικών πιθανοτήτων**
Μικρότερη διάμετρος πυρήνα	4.853	<0.001	128.215			
Σφαιρικότητα πυρήνα	-0.369	<0.001	0.691			
Περίμετρος πυρήνα	-0.254	<0.001	0.775			
Φωτεινότητα πυρήνα	0.352	0.005	1.421	-0.066	0.005	0.935
Αναλογία πυρήνα/κυτταροπλάσματος	-0.318	<0.001	0.727			
Μέγιστο πυρήνα	-0.228	<0.001	0.795	-0.060	<0.001	0.941
Εμβαδόν κυτταροπλάσματος	-0.007	<0.001	0.992			
Θέση πυρήνα	-0.131	<0.001	0.877	- 3.443	<0.001	0.031
Φωτεινότητα κυτταροπλάσματος	0.240	0.049	1.272			
Μεγαλύτερη διάμετρος πυρήνα	1.019	0.007	2.771	0.083	<0.001	1.087
Επιμήκυνση κυτταροπλάσματος	0.080	0.021	1.084	0.029	<0.001	1.029
Περίμετρος κυτταροπλάσματος				-0.009	<0.001	0.990
Εμβαδόν πυρήνα				0.004	<0.001	1.004

Πίνακας 5.1: Συντελεστές και λόγοι συμπληρωματικών πιθανοτήτων του μοντέλου λογαριθμιστικής παλινδρόμησης

Για την ερμηνεία των συντελεστών του μοντέλο χρησιμοποιούνται οι αντιλογάριθμοι επειδή ερμηνεύονται απευθείας ως λόγος συμπληρωματικών πιθανοτήτων (odds ratio) και η ερμηνεία απλουστεύεται. Κάποια χαρακτηριστικά λαμβάνουν τιμές στο διάστημα (0,1). Καθώς η ερμηνεία των αντιλογαριθμισμένων συντελεστών γίνεται για μια μονάδα αύξησης της τιμής του χαρακτηριστικού, οι τιμές των συγκεκριμένων μεταβλητών πολλαπλασιάστηκαν με 100 ώστε η ερμηνεία των συντελεστών να γίνεται για αύξηση ενός εκατοστού της μονάδας. Έτσι, για την παλαιά βάση δεδομένων προκύπτει ότι:

- Για 1 μm αύξηση της μικρότερης διαμέτρου του πυρήνα του κυττάρου, η συμπληρωματική πιθανότητα (odds) να είναι το κύτταρο μη υγιές είναι ίση με 128 φορές την πιθανότητα να είναι υγιές ,έχοντας λάβει υπόψη τις άλλες μεταβλητές .

Στις υπόλοιπες ερμηνείες των μεταβλητών θα χρησιμοποιείται ο όρος odds αντί για συμπληρωματικές πιθανότητες για ευκολία .

- Για ένα εκατοστό του μm αύξηση της σφαιρικότητας του πυρήνα, μειώνονται τα odds να μην είναι υγιές το κύτταρο κατά 30% περίπου σε σχέση με το να είναι υγιές, έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για 1 μm αύξηση της περιμέτρου του πυρήνα, μειώνονται τα odds να μην είναι υγιές το κύτταρο κατά 23% περίπου σε σχέση με το να είναι υγιές, έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για ένα εκατοστό του μm αύξηση στην φωτεινότητα του πυρήνα, αυξάνονται κατά περίπου 42% τα odds να μην είναι υγιές σε σχέση με το να είναι υγιές, έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για ένα εκατοστό του μm αύξηση στην αναλογία πυρήνα κυτταροπλάσματος, μειώνονται τα odds να μην είναι υγιές το κύτταρο κατά 28% περίπου, σε σχέση με το να είναι υγιές , έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για 1 μm αύξηση στο μέγιστο του πυρήνα, τα odds να μην είναι υγιές το κύτταρο μειώνονται περίπου κατά 20% σε σχέση με το να είναι υγιές, έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για 1 μm^2 αύξηση του εμβαδού κυτταροπλάσματος , μειώνονται τα odds το κύτταρο να μην είναι υγιές κατα περίπου 1% σε σχέση με το να είναι υγιές , έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για ένα εκατοστό του μm αύξηση της θέσης του πυρήνα, τα odds να μην είναι το κύτταρο υγιές μειώνονται κατά περίπου 13% σε σχέση με το ναείναι υγιές , έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για ένα εκατοστό του μm αύξηση στην φωτεινότητα του κυτταροπλάσματος, αυξάνονται τα odds να μην είναι υγιές το κύτταρο κατά περίπου 27% σε σχέση με το να είναι υγιές , έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για 1 μm αύξηση της μεγαλύτερης διαμέτρου του πυρήνα, τα odds του να είναι το κύτταρο μη υγιές είναι ίσα με 2,77 φορές τα odds να είναι υγιές, έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για ένα εκατοστό του μm αύξηση στην επιμήκυνση του κυτταροπλάσματος, το κύτταρο έχει κατά περίπου 8% αυξημένα odds να μην είναι υγιές σε σχέση με το να είναι υγιές, έχοντας λάβει υπόψη τις άλλες μεταβλητές.

Αντίστοιχα, για τη νέα βάση δεδομένων προκύπτει ότι:

- Για 1 μm αύξηση της περιμέτρου του κυτταροπλάσματος του κυττάρου, η συμπληρωματική πιθανότητα (odds) να είναι το κύτταρο μη υγιές είναι 1% μειωμένη σε σχέση με την πιθανότητα να είναι υγιές ,έχοντας λάβει υπόψη τις άλλες μεταβλητές .

- Για 1 μm αύξηση της μεγαλύτερης διαμέτρου του πυρήνα, αυξάνονται τα odds να μην είναι υγιές το κύτταρο κατά 8% περίπου σε σχέση με το να είναι υγιές, έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για ένα εκατοστό του μm αύξηση επιμήκυνσης του κυτταροπλάσματος αυξάνονται τα odds να μην είναι υγιές κατά 3% περίπου σε σχέση με το να είναι υγιές, έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για μία μονάδα αύξηση της φωτεινότητας του πυρήνα, μειώνονται τα odds να μην είναι υγιές το κύτταρο κατά 7% περίπου σε σχέση με το να είναι υγιές, έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για 1 μm^2 αύξηση του εμβαδού του πυρήνα , αυξάνονται τα odds το κύτταρο να μην είναι υγιές κατά 5‰ περίπου σε σχέση με το να είναι υγιές , έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για 1 μm αύξηση του μέγιστου του πυρήνα, μειώνονται τα odds να μην είναι υγιές το κύτταρο κατά 6% περίπου σε σχέση με το να είναι υγιές, έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για 1 μm αύξηση της θέσης του πυρήνα, μειώνονται τα odds να μην είναι υγιές το κύτταρο κατά 68% περίπου σε σχέση με το να είναι υγιές, έχοντας λάβει υπόψη τις άλλες μεταβλητές

	Παλαιά βάση δεδομένων			Νέα βάση δεδομένων		
	Κατηγορία πρόβλεψης μοντέλου			Κατηγορία πρόβλεψης μοντέλου		
Κατηγορία κατάταξης	**Υγιή κύτταρα**	**Μη υγιή κύτταρα**	**Σύνολο**	**Υγιή κύτταρα**	**Μη υγιή κύτταρα**	**Σύνολο**
Υγιή κύτταρα	195	5	200	195	5	200
Μη υγιή κύτταρα	2	298	300	2	298	300
Σύνολο	197	303	500	197	303	500

Πίνακας 5.2: Κατάταξη τωνπαρατηρήσεων σύμφωνα με την κατηγορία πρόβλεψης του μοντέλου και την πραγματική κατηγορία που ανήκουν

Τα σφάλματα υπολογίζονται και προκύπτει ότι

Σφάλμα	**Παλαιά βάση δεδομένων**	**Νέα βάση δεδομένων**
FN%	0,7%	2,6%
FP%	2.5%	14,9%
OE%	1.4%	5.8%

Πίνακας 5.3:Σφάλματα ταξινόμησης

Τα σφάλματα της μεθόδου λογιστικής παλινδρόμησης είναι αρκετά κάτω από το 5% για την παλιά βάση, γεγονός που την καθιστά μια αρκετά σίγουρη μέθοδο για την ταξινόμηση των κυττάρων. Τα σφάλματα ταξινόμησης είναι πάνω από το 5% για τη νέα βάση δεδομένων αν και τα ψευδώς αρνητικά που είναι πιο κρίσιμα στην ιατρική βρίσκονται κάτω από το 5%

Ο δείκτης κάππα υπολογίζεται σύμφωνα με τον τύπο

$$\kappa = \frac{N \cdot \sum_i^r x_{ii} - \sum_i^r x_{i+} \cdot x_{+i}}{N^2 - \sum_i^r x_{i+} \cdot x_{+i}}$$

και στην περίπτωση αυτή ισούται με κ=97,1% για την παλαιά βάση δεδομένων και κ=84,7% για τη νέα που είναι ιδιαίτερα ικανοποιητικά ποσοστά συμφωνίας μεταξύ της πραγματικής κατηγορίας που ανήκουν οι παρατηρήσεις και της κατηγορίας που προβλέπει το μοντέλο.Έτσι, επαληθεύεται ο ισχυρισμός ότι το μοντέλο παλινδρόμησης είναι επιτυχές ιδιαίτερα για την παλαιά βάση.

5.2 Διατάξιμη λογιστική παλινδρόμηση (ordinal logistic regression)

Οι κατηγορίες των μη υγιών κυττάρων είναι διαβαθμισμένες, κάτι που δεν συμβαίνει με τις κατηγορίες των υγιών κυττάρων, Στην διατάξιμη λογιστική παλινδρόμηση η εξαρτημένη μεταβλητή πρέπει να έχει διαβαθμίσεις. Για αυτό το λόγο η κατηγορία όλων των υγιών κυττάρων θα θεωρείται ως ενιαία σε αυτή την ανάλυση και των μη υγιών διαβαθμισμένη. Με τη μέθοδο stepwise επιλέχθηκαν τελικά 10 χαρακτηριστικά για την παλαιά βάση δεδομένων και 12 για τη νέα. Τα αποτελέσματα του αλγόριθμου φαίνονται στον Πίνακα 5.2

	Παλαιά βάση δεδομένων			Νέα βάση δεδομένων		
μεταβλητή	**Συντελεστής**	**p-value**	**Λόγος συμπληρωματικών πιθανοτήτων**	**Συντελε-στής**	**p-value**	**Λόγος συμπληρωματικών πιθανοτήτων**
Εμβαδόν κυτταροπλάσματος	-0.001	<0.001	0.998	0.001	<0.001	1.001
Σφαιρικότητα πυρήνα	-6.077	<0.001	0 .002	0.105	<0.001	1.111
Μικρότερη διάμετρος πυρήνα	0.786	<0.001	2.195	0.207	<0.001	1.231
Περίμετρος κυτταροπλάσματος	-0.005	0.006	0.994			
Θέση πυρήνα	-2.058	<0.001	0.127	-1.901	<0.001	0.149
Φωτεινότητα κυτταροπλάσματος	0.098	<0.001	1.103	0.016	<0.001	1.016
εμβαδόν πυρήνα	-0.023	<0.001	0.977	-0.002	<0.001	0.997
Περίμετρος πυρήνα	-0.031	<0.001	0.968			
ελάχιστο πυρήνα	0.027	0.003	1.028			
Επιμήκυνση πυρήνα	0.191	0.003	1.211	-15.438	<0.001	0.856
Αναλογία πυρήνα/κυτταροπλάσματος				0.067	<0.001	1.069
Μεγαλύτερη διάμετρος πυρήνα				0.075	<0.001	1.078
Επιμήκυνση κυτταροπλάσματος				0.041	<0.001	1.042
Μικρότερη διάμετρος κυτταροπλάσματος				-0.038	<0.001	0.962
φωτεινότητα πυρήνα				-0.026	<0.001	0.974

Πίνακας 5.4: Συντελεστές και λόγοι συμπληρωματικών πιθανοτήτων διαβαθμισμένης λογιστικής παλινδρόμησης

Το μοντέλο είναι αναλογικών συμπληρωματικών πιθανοτήτων (proportional odds model). Αυτό πρακτικά σημαίνει ότι ο λόγος συμπληρωματικών πιθανοτήτων είναι σταθερός για κάθε διχοτόμηση των κατηγοριών της εξαρτημένης μεταβλητής. Έτσι, για την παλαιά βάση δεδομένων προκύπτει ότι:

- Για 1 μm^2 αύξηση του εμβαδού κυτταροπλάσματος, μειώνονται κατά περίπου 1‰ τα odds να ανήκει το κύτταρο στην κατηγορία βαριάς δυσπλασίας σε σχέση με το να βρίσκεται από μεσαία μέχρι υγιή , έχοντας λάβει υπόψη τις άλλες μεταβλητές. Ή για 1 μm^2 αύξηση του εμβαδού κυτταροπλάσματος, μειώνονται κατά περίπου 1‰ τα odds να ανήκει το κύτταρο στην κατηγορία τουλάχιστον μεσαία δυσπλασία σε σχέση με το να βρίσκεται στην κατηγορία το πολύ ελαφριά δυσπλασία .

Στις υπόλοιπες ερμηνείες των μεταβλητών θα καταγράφεται μία ενδεικτική ερμηνεία, δηλαδή μία διχοτόμηση του αποτελέσματος .

- Για 1 μm αύξηση της σφαιρικότητας του πυρήνα, μειώνονται κατά περίπου 98% τα odds το κύτταρο να βρίσκεται στην κατηγορία βαριάς δυσπλασίας σε σχέση με το να βρίσκεται το πολύ στην κατηγορία μεσαίας δυσπλασίας , έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για 1 μm αύξηση της μικρότερης διαμέτρου του πυρήνα, το κύτταρο έχει διπλάσια odds να βρίσκεται στην κατηγορία βαριάς δυσπλασίας σε σχέση με το να βρίσκεται το πολύ στην κατηγορία μεσαίας δυσπλασίας , έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για 1 μm αύξηση στην περίμετρο του κυτταροπλάσματος, μειώνονται κατά περίπου 1% τα odds το κύτταρο να βρίσκεται στην κατηγορία βαριάς δυσπλασίας σε σχέση με το να βρίσκεται το πολύ στην κατηγορία μεσαίας δυσπλασίας , έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για 1 μm αύξηση στην θέση του πυρήνα, μειώνονται κατά περίπου 83% τα odds το κύτταρο να βρίσκεται στην κατηγορία βαριάς δυσπλασίας σε σχέση με το να βρίσκεται το πολύ στην κατηγορία μεσαίας δυσπλασίας, έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για ένα εκατοστό μm αύξηση της φωτεινότητας του κυτταροπλάσματος αυξάνονται κατά περίπου 10% τα odds το κύτταρο να βρίσκεται στην κατηγορία βαριάς δυσπλασίας σε σχέση με το να βρίσκεται το πολύ στην κατηγορία μεσαίας δυσπλασίας, έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για 1 μm^2 αύξηση του εμβαδού πυρήνα, μειώνονται κατά περίπου 2% τα odds να ανήκει το κύτταρο στην κατηγορία βαριάς δυσπλασίας σε σχέση με το να βρίσκεται από μεσαία μέχρι υγιή , έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για 1 μm αύξηση της περιμέτρου του πυρήνα, μειώνονται κατά περίπου 3% τα odds το κύτταρο να βρίσκεται στην κατηγορία βαριάς δυσπλασίας σε σχέση με το να βρίσκεται το πολύ στην κατηγορία μεσαίας δυσπλασίας , έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για 1 μm αύξηση της φωτεινότητας του κυτταροπλάσματος αυξάνονται κατά περίπου 2% τα odds το κύτταρο να βρίσκεται στην κατηγορία βαριάς δυσπλασίας σε σχέση με το να βρίσκεται το πολύ στην κατηγορία μεσαίας δυσπλασίας , έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για 1 μm αύξηση της επιμήκυνσης του πυρήνα αυξάνονται κατά περίπου 21% τα odds το κύτταρο να βρίσκεται στην κατηγορία βαριάς δυσπλασίας σε σχέση με το να βρίσκεται το πολύ στην κατηγορία μεσαίας δυσπλασίας , έχοντας λάβει υπόψη τις άλλες μεταβλητές

Για τη νέα βάση δεδομένων προκύπτει ότι:

- Για 1/100 αύξηση της αναλογίας πυρήνα/ κυτταροπλάσματος αυξάνονται κατά περίπου 7% τα odds το κύτταρο να βρίσκεται στην κατηγορία βαριάς δυσπλασίας σε σχέση με το να βρίσκεται το πολύ στην κατηγορία μεσαίας δυσπλασίας , έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για ένα μm αύξηση της θέσης του πυρήνα, μειώνονται κατά περίπου 85% τα odds το κύτταρο να βρίσκεται στην κατηγορία βαριάς δυσπλασίας σε σχέση με το να βρίσκεται το πολύ στην κατηγορία μεσαίας δυσπλασίας , έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για ένα μm αύξηση της μεγαλύτερης διαμέτρου του πυρήνα, αυξάνονται κατά περίπου 8% τα odds το κύτταρο να βρίσκεται στην κατηγορία βαριάς δυσπλασίας σε σχέση με το να βρίσκεται το πολύ στην κατηγορία μεσαίας δυσπλασίας , έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για ένα μm^2 αύξηση του εμβαδού του πυρήνα , μειώνονται τα odds κατά 2‰ περίπου το κύτταρο να βρίσκεται στην κατηγορία βαριάς δυσπλασίας σε σχέση με το να βρίσκεται το πολύ στην κατηγορία μεσαίας δυσπλασίας , έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για ένα μm αύξηση της μικρότερης διαμέτρου του πυρήνα, αυξάνονται κατά περίπου 23% τα odds το κύτταρο να βρίσκεται στην κατηγορία βαριάς δυσπλασίας σε σχέση με το να βρίσκεται το πολύ στην κατηγορία μεσαίας δυσπλασίας , έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για ένα μm αύξηση της επιμήκυνσης του κυτταροπλάσματος, αυξάνονται κατά περίπου 4% τα odds το κύτταρο να βρίσκεται στην κατηγορία βαριάς δυσπλασίας σε σχέση με το να βρίσκεται το πολύ στην κατηγορία μεσαίας δυσπλασίας , έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για ένα μm αύξηση της μικρότερης διαμέτρου του κυτταροπλάσματος, μειώνονται κατά περίπου 4% τα odds το κύτταρο να βρίσκεται στην κατηγορία βαριάς δυσπλασίας σε σχέση με το να βρίσκεται το πολύ στην κατηγορία μεσαίας δυσπλασίας , έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για ένα μm^2 αύξηση του εμβαδού του κυτταροπλάσματος , αυξάνονται τα odds το κύτταρο να βρίσκεται στην κατηγορία βαριάς δυσπλασίας σε σχέση με το να βρίσκεται το πολύ στην κατηγορία μεσαίας δυσπλασίας κατά 1%, έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για 1/100 μm αύξηση της επιμήκυνσης του πυρήνα, μειώνονται κατά περίπου 15% τα odds το κύτταρο να βρίσκεται στην κατηγορία βαριάς δυσπλασίας σε σχέση με το να βρίσκεται το πολύ στην κατηγορία μεσαίας δυσπλασίας , έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για 1/100 μm αύξηση της φωτεινότητας του πυρήνα, μειώνονται κατά περίπου 3% τα odds το κύτταρο να βρίσκεται στην κατηγορία βαριάς δυσπλασίας σε σχέση με το να βρίσκεται το πολύ στην κατηγορία μεσαίας δυσπλασίας , έχοντας λάβει υπόψη τις άλλες μεταβλητές

- Για ένα μm αύξηση της φωτεινότητας του κυτταροπλάσματος, αυξάνονται κατά περίπου 2% τα odds το κύτταρο να βρίσκεται στην κατηγορία βαριάς δυσπλασίας σε σχέση με το να βρίσκεται το πολύ στην κατηγορία μεσαίας δυσπλασίας , έχοντας λάβει υπόψη τις άλλες μεταβλητές

Ο δείκτης κάππα υπολογίζεται σύμφωνα με τον τύπο

$$\kappa = \frac{N \cdot \sum_i^r x_{ii} - \sum_i^r x_{i+} \cdot x_{+i}}{N^2 - \sum_i^r x_{i+} \cdot x_{+i}}$$

Και στην περίπτωση αυτή ισούται με κ=49% για την παλαιά βάση καιπ κ=39% για τη νέα, που δείχνει ότι η συμφωνία μεταξύ της πραγματικής και της προβλεπόμενης από το μοντέλο κατηγορίας κατάταξης των κυττάρων δεν είναι ιδιαίτερα ικανοποιητική.

5.3 Ανάλυση κατά συστάδες (cluster analysis)

κατηγορία	Συστάδα 1	Συστάδα 2	Συστάδα 3	Συστάδα 4	Συστάδα 5	Συστάδα 6	Συστάδα 7
Κυλινδρικά επιθήλια	0	44	0	6	0	0	0
Παραβασικά λεπιδοειδή επιθήλια	0	28	1	21	0	0	0
Ενδιάμεσα λεπιδοειδή επιθήλια	19	0	4	0	10	17	0
Επιφανειακά λεπιδοειδή επιθήλια	20	0	3	0	19	8	0
Ελαφριά λεπιδοειδής μη κερατινώδης δυσπλασία	5	18	40	34	2	1	0
Μεσαία λεπιδοειδής μη κερατινώδης δυσπλασία	3	47	15	34	0	0	1
Βαριά λεπιδοειδής μη κερατινώδης δυσπλασία	0	71	2	27	0	0	0
σύνολο	47	208	65	122	31	26	1

Σε αυτήν την παράγραφο παρουσιάζονται τα αποτελέσματα της ανάλυσης κατά συστάδες κ-μέσων

Πίνακας 5.5: Κατάταξη κυττάρων στις 7 συστάδες ανά κατηγορία που ανήκουν πραγματικά, παλαιά βάση δεδομένων

κατηγορία	Συστάδα 1	Συστάδα 2	Συστάδα 3	Συστάδα 4	Συστάδα 5	Συστάδα 6	Συστάδα 7
Κυλινδρικά επιθήλια	26	0	7	17	0	23	1
Παραβασικά λεπιδοειδή επιθήλια	4	0	10	32	0	22	2
Ενδιάμεσα λεπιδοειδή επιθήλια	0	22	0	0	73	0	3
Επιφανειακά λεπιδοειδή επιθήλια	0	50	43	13	0	2	68
Ελαφριά λεπιδοειδής μη κερατινώδης δυσπλασία	0	75	9	1	33	0	28
Μεσαία λεπιδοειδής μη κερατινώδης δυσπλασία	0	55	0	0	134	0	8
Βαριά λεπιδοειδής μη κερατινώδης δυσπλασία	0	18	0	0	131	0	1
σύνολο	30	220	69	63	371	47	111

Πίνακας 5.6: Κατάταξη κυττάρων στις 7 συστάδες ανά κατηγορία που ανήκουν πραγματικά, νέα βάση δεδομένων

Όπως φαίνεται από τους Πίνακες 5.5 και 5.6 οι συστάδες που δημιουργήθηκαν δεν αντιπροσωπεύουν ικανοποιητικά τις πραγματικές κατηγορίες στις οποίες ανήκουν τα κύτταρα . Το αξιοσημείωτο είναι ότι η συσταδική ανάλυση κατηγοριοποιεί μία συγκεκριμένη παρατήρηση της παλαιάς βάσης δεδομένων μόνη της σε μία συστάδα,την συστάδα 7 . Αυτή η παρατήρηση προφανώς έχει κάποια διαφορετικά χαρακτηριστικά και πρέπει να ερευνηθεί .Πρόκειται για την παρατήρηση 323 της βάσης η οποία φαίνεται να έχει αρκετά μεγάλη απόλιση από τις μέσες τιμές των χαρακτηριστικών μέγιστο και ελάχιστο κυτταροπλάσματος στην κατηγορία στην οποία ανήκει,δηλαδή στη μεσαία λεπιδοειδή μη κερατινώδη δυσπλασία .(μέγιστο κυτταροπλάσματος=15155 ενώ η μέση τιμή της κατηγορίας είναι 382,24 και ελάχιστο κυτταροπλάσματος=12049 ενώ η μέση τιμή της κατηγορίας είναι 299,17) Ίσως οι τιμές των χαρακτηριστικών να είναι λανθασμένες ή να πρόκειται για μία περίπτωση που πρέπει να εξεταστεί ξεχωριστά .

Παρακάτω παρουσιάζονται τα αποτελέσματα της συσταδικής ανάλυσης για το πρόβλημα της κατανομής των παρατηρήσεων σε δύο συστάδες, των υγιών και των μη υγιών κυττάρων

	Παλαιά βάση δεδομένων			**Νέα βάση δεδομένων**		
κατηγορία	**Συστάδα 1**	**Συστάδα 2**	**Σύνολο**	**Συστάδα 1**	**Συστάδα 2**	**Σύνολο**
Υγιή κύτταρα	87	113	200	118	124	242
Μη υγιή κύτταρα	11	289	300	659	16	675
σύνολο	98	402	500	777	140	917

Πίνακας 5.7: Κατανομή των κυττάρων ανα συστάδα και πραγματική κατηγορία

Από τον Πίνακα 5.7 μπορούμε να υπολογίσουμε τα σφάλματα $FN\% = \frac{FN}{TP+FN} \times 100\%$, $FP\% = \frac{FP}{TN+FP} \times 100\%$ και $OE\% = \frac{FN+FP}{TP+FN+TN+FP} \times 100\%$. Έτσι προκύπτει:

Σφάλμα	**Παλαιά βάση δεδομένων**	**Νέα βάση δεδομένων**
FN%	11.2%	11.4%
FP%	28.1%	15.2%
OE%	24.8%	14.6%

Πίνακας 5.8: Σφάλματα ταξινόμησης

Τα αποτελέσματα της ανάλυσης δεν είναι ικανοποιητικά αφού παρατηρείται αρκετά μεγάλο ποσοστό ψευδώς αρνητικών και ψευδώς θετικών κυττάρων . Αυτό που έχει μεγαλύτερη σημασία στην ανάλυση είναι το ποσοστό των ψευδώς αρνητικών που είναι πιο χαμηλό αλλά και πάλι δεν είναι αρκετά ικανοποιητικό

Υπολογίζεται επίσης ο δείκτης κάππα ως κ=43,5% για την παλαιά βάση και κ=-28,6% για τη νέα που δείχνει ότι το μέγεθος της συμφωνίας μεταξύ των συστάδων και της κατηγορίας των κυττάρων δεν είναι ιδιαίτερα ικανοποιητικό για την πρώτη βάση και καθόλου ικανοποιητικό για τη δεύτερη. Το αρνητικό πρόσημο στο ποσοστό του δείκτη Κάππα υποδηλώνει χειρότερη συμφωνία και από την τυχαία συμφωνία μεταξύ των κατηγοριών.

5.4 Διαχωριστική ανάλυση (discriminant analysis)

Πραγματοποιήθηκε διαχωριστική ανάλυση για τα δεδομένα της βάσης χρησιμοποιώντας όλα τα χαρακτηριστικά των κυττάρων. Σκοπός της ανάλυσης είναι ο διαχωρισμός των κυττάρων στις δύο κατηγορίες υγιών και μη υγιών κυττάρων.

Στον Πίνακα 5.9 φαίνονται οι τυποποιημένοι συντελεστές των μεταβλητών μέσω των οποίων μπορούν να καταταχθούν οι μεταβλητές σε σειρά σύμφωνα με τη διαχωριστική τους ικανότητα. Όσο πιο μεγάλη είναι η απόλυτη τιμή ενός συντελεστή τόσο μεγαλύτερη είναι η διαχωριστική ικανότητα της μεταβλητής, δηλαδή τόσο περισσότερο συνεισφέρει στο διαχωρισμό των παρατηρήσεων. Οι τυποποιημένοι συντελεστές λαμβάνουν υπόψη τις συσχετίσεις της εκάστοτε μεταβλητής με τις υπόλοιπες.

μεταβλητή	Τυποποιημένος συντελεστής διαχωριστικής συνάρτησης	Κατάταξη μεταβλητής	Τυποποιημένος συντελεστής διαχωριστικής συνάρτησης	Κατάταξη μεταβλητής
Εμβαδόν πυρήνα	0,145	17	-2,150	4
Εμβαδόν κυτταροπλάσματος	0,160	15	0,429	11
Αναλογία πυρήνα/κυτταροπλάσματος	0,579	10	0,000	20
Φωτεινότητα πυρήνα	-0,141	18	-0,436	10
Φωτεινότητα κυτταροπλάσματος	-0,503	11	0,135	16
Μικρότερη διάμετρος πυρήνα	-1,227	1	2,201	3
Μεγαλύτερη διάμετρος πυρήνα	-0,987	3	1,756	5
Επιμήκυνση πυρήνα	-0,657	7	-1,681	6
Σφαιρικότητα πυρήνα	0,930	4	1,544	7
Μικρότερη διάμετρος κυτταροπλάσματος	0,164	14	-1,135	8
Μεγαλύτερη διάμετρος κυτταροπλάσματος	-0,255	13	-0,079	19
Επιμήκυνση κυτταροπλάσματος	-0,930	4	0,585	9
Σφαιρικότητα κυτταροπλάσματος	1,072	2	-0,141	15
Περίμετρος πυρήνα	0,599	9	-0,197	14
Περίμετρος κυτταροπλάσματος	0,356	12	0,122	17
Θέση πυρήνα	0,127	19	-0,198	13
Μέγιστο πυρήνα	0,904	5	-0,286	12
Ελάχιστο πυρήνα	-0,156	16	-0,081	18
Μέγιστο κυτταροπλάσματος	0,746	6	-2,562	2
Ελάχιστο κυτταροπλάσματος	-0,638	8	2,590	1

Πίνακας 5.9 : Τυποποιημένοι συντελεστές διαχωριστικής συνάρτησης

Η κατάταξη των μεταβλητών από αυτή που συνεισφέρει περισσότερο στον διαχωρισμό προς αυτή που συνεισφέρει λιγότερο φαίνεται στον Πίνακα 5.9. Οι 5 μεταβλητές που συνεισφέρουν περισσότερο για την παλαιά βάση δεδομένων είναι: **Μικρότερη διάμετρος πυρήνα, Σφαιρικότητα κυτταροπλάσματος, Μεγαλύτερη διάμετρος πυρήνα, Σφαιρικότητα πυρήνα και Επιμήκυνση κυτταροπλάσματος** και οι 5 μεταβλητές που συνεισφέρουν λιγότερο είναι**: Εμβαδόν κυτταροπλάσματος, Ελάχιστο πυρήνα, Εμβαδόν πυρήνα, Φωτεινότητα πυρήνα, Θέση πυρήνα.**

Οι 5 μεταβλητές που συνεισφέρουν περισσότερο για τη νέα βάση δεδομένων είναι: **Ελάχιστο κυτταροπλάσματος, Μέγιστο κυτταροπλάσματος, Μικρότερη διάμετρος πυρήνα, Εμβαδόν πυρήνα,**

Μεγαλύτερη διάμετρος πυρήνα και οι 5 μεταβλητές που συνεισφέρουν λιγότερο είναι**: Φωτεινότητα κυτταροπλάσματος, Περίμετρος κυτταροπλάσματος, Ελάχιστο πυρήνα, Μεγαλύτερη διάμετρος κυτταροπλάσματος, Αναλογία πυρήνα/κυτταροπλάσματος**

Η διαχωριστική ανάλυση που πραγματοποιήθηκε απέφερε πολύ ικανοποιητικά αποτελέσματα. Στον Πίνακα 5.10 φαίνεται ότι το Λ του Wilks είναι στατιστικά σημαντικό. Αυτό σημαίνει ότι η διαχωριστική συνάρτηση έχει σημαντική διαχωριστική ικανότητα κάτι που φαίνεται και από τον Πίνακα 5.11 Ο συντελεστής συσχέτισης της συνάρτησης μεταξύ των ομάδων είναι 0.871. Όταν πρόκειται για 2 ομάδες όπως στη συγκεκριμένη περίπτωση,το τεράγωνο του συντελεστή αυτού ισούται με το συντελεστή συσχέτισης του Pearson , δηλαδή ο συντελεστής παλινδρόμησης του σκορ της συνάρτησης με ανεξάρτητη μεταβλητή αυτή που δηλώνει τις 2 ομάδες και εκφράζει το ποσοστό της διακύμανσης της διαχωριστικής συνάρτησης που ερμηνεύεται από τη διαφορά μεταξύ των 2 ομάδων.

	Παλαιά βάση δεδομένων				**Νέα βάση δεδομένων**			
Έλεγχος συνάρτησης	**Λ του Wilks**	**X^2 έλεγχος**	**Βαθμοί ελευθερίας**	**σημαντικότητα**	**Λ του Wilks**	**X^2 έλεγχος**	**Βαθμοί ελευθερίας**	**σημαντικότητα**
1	0,242	692,088	20	<0.001	0,320	1031,029	20	<0.001

Πίνακας 5.10 : |Το Λ του Wilks και ο έλεγχος για τη σημαντικότητά του

	Παλαιά βάση δεδομένων		**Νέα βάση δεδομένων**	
Συνάρτηση	**Ιδιοτιμή**	**Συντελεστής συσχέτισης**	**Ιδιοτιμή**	**Συντελεστής συσχέτισης**
1	3,130	0.871	2,124	0.825

Πίνακας 5.11: Συντελεστής συσχέτισης της συνάρτησης με τις ομάδες

Η διαχωριστική ανάλυση καταχωρεί σωστά το 96% των υγιών κυττάρων και το 97,3% των μη υγιών για την παλαιά βάση δεδομένων και το 86,8% των υγιών κυττάρων και το 96,9% των μη υγιών για τη νέα βάση. Συνολικά καταχωρεί σωστά το 96,8% των παρατηρήσεων της παλαιάς βάσης και το 94,2% της νέας, δηλαδή ένα πολύ ικανοποιητικό ποσοστό όπως φαίνεται και στον πίνακα 5.12

κατηγορία	Ποσοστό % σωστής ταξινόμησης	Συνολικό ποσοστό σωστής ταξινόμησης	Ποσοστό % σωστής ταξινόμησης	Συνολικό ποσοστό σωστής ταξινόμησης
Υγιή κύτταρα	96,0	96,8	86,8	94,2
Μη υγιή κύτταρα	97,3		96,9	

Πίνακας 5.12: Ποσοστά σωστής ταξινόμησης των παρατηρήσεων του δείγματος

Στον Πίνακα 5.13 φαίνονται οι συντελεστές ταξινόμησης της συνάρτησης (classification function coefficients) μέσω των οποίων δημιουργούνται συναρτήσεις που καταχωρούν κάθε νέα παρατήρηση σε μια ομάδα

μεταβλητή	Συντελεστής για την κατηγορία υγιών κυττάρων	Συντελεστής για την κατηγορία μη υγιών κυττάρων	Συντελεστής για την κατηγορία υγιών κυττάρων	Συντελεστής για την κατηγορία μη υγιών κυττάρων
Εμβαδόν πυρήνα	-0,232	**-0,237**	-0,034	**-0,038**
Εμβαδόν κυτταροπλάσματος	-0,026	**-0,026**	-0,003	**-0,003**
Αναλογία πυρήνα/κυτταροπλάσματος	152,879	**138,437**	275,004	**275,011**
Φωτεινότητα πυρήνα	9,695	**16,143**	0,065	**-0,006**
Φωτεινότητα κυτταροπλάσματος	238,403	**265,068**	0,165	**0,183**
Μικρότερη διάμετρος πυρήνα	-4,315	**-2,990**	-2,029	**-1,550**
Μεγαλύτερη διάμετρος πυρήνα	8,049	**8,668**	3,944	**4,279**
Επιμήκυνση πυρήνα	198,307	**214,658**	177,176	**141,527**
Σφαιρικότητα πυρήνα	1,410	**-20,677**	103,524	**136,958**
Μικρότερη διάμετρος κυτταροπλάσματος	-1,828	**-1,867**	-0,229	**-0,288**
Μεγαλύτερη διάμετρος κυτταροπλάσματος	2,519	**2,564**	1,009	**1,006**
Επιμήκυνση κυτταροπλάσματος	80,462	**99,325**	-18,075	**-7,176**
Σφαιρικότητα κυτταροπλάσματος	171,642	**148,105**	325,562	**322,556**
Περίμετρος πυρήνα	0,053	**-0,015**	-0,097	**-0,110**
Περίμετρος κυτταροπλάσματος	-0,006	**-0,011**	-0,016	**-0,014**
Θέση πυρήνα	-16,169	**-18,873**	9,346	**5,365**
Μέγιστο πυρήνα	-0,236	**-0,342**	-0,046	**-0,065**
Ελάχιστο πυρήνα	-0,079	**-0,060**	0,107	**0,102**
Μέγιστο κυτταροπλάσματος	-0,016	**-0,019**	-0,024	**-0,040**
Ελάχιστο κυτταροπλάσματος	0,015	**0,018**	0,036	**0,052**

Πίνακας 5.13: Συντελεστές ταξινόμησης της συνάρτησης

Οι συναρτήσεις που προκύπτουν από τον Πίνακα 5.13 είναι:

$\mathbf{Z_1}$ **=-273,338 -0,232*** Εμβαδόν πυρήνα**-0,026*** Εμβαδόν κυτταροπλάσματος **+ 152,879*** Αναλογία πυρήνα/κυτταροπλάσματος **+ 9,695*** Φωτεινότητα πυρήνα **+ 238,403*** Φωτεινότητα κυτταροπλάσματος**-4,315*** Μικρότερη διάμετρος πυρήνα **+ 8,049*** Μεγαλύτερη διάμετρος πυρήνα **+ 198,307*** Επιμήκυνση πυρήνα **+ 1,410*** Σφαιρικότητα πυρήνα**-1,828*** Μικρότερη διάμετρος κυτταροπλάσματος **+ 2,519*** Μεγαλύτερη διάμετρος κυτταροπλάσματος **+ 80,462*** Επιμήκυνση κυτταροπλάσματος **+ 171,642*** Σφαιρικότητα κυτταροπλάσματος **+ 0,053*** Περίμετρος πυρήνα**-0,006*** Περίμετρος κυτταροπλάσματος**-16,169*** Θέση πυρήνα**-0,236*** Μέγιστο πυρήνα**-0,079*** Ελάχιστο πυρήνα**-0,016*** Μέγιστο κυτταροπλάσματος **+ 0,015*** Ελάχιστο κυτταροπλάσματος

και

$\mathbf{Z_1}$ = **-297,063 -0,034*** Εμβαδόν πυρήνα-**0,003*** Εμβαδόν κυτταροπλάσματος + **275,004*** Αναλογία πυρήνα/κυτταροπλάσματος + **0,065*** Φωτεινότητα πυρήνα + **0,165*** Φωτεινότητα κυτταροπλάσματος-**2,029*** Μικρότερη διάμετρος πυρήνα + **3,944*** Μεγαλύτερη διάμετρος πυρήνα + **177,176*** Επιμήκυνση πυρήνα + **103,524*** Σφαιρικότητα πυρήνα-**0,229*** Μικρότερη διάμετρος κυτταροπλάσματος + **1,009*** Μεγαλύτερη διάμετρος κυτταροπλάσματος **-18,075*** Επιμήκυνση κυτταροπλάσματος + **325,562*** Σφαιρικότητα κυτταροπλάσματος **-0,097*** Περίμετρος πυρήνα-**0,016*** Περίμετρος κυτταροπλάσματος + **9,346*** Θέση πυρήνα-**0,046*** Μέγιστο πυρήνα + **0,107*** Ελάχιστο πυρήνα-**0,024*** Μέγιστο κυτταροπλάσματος + **0,036*** Ελάχιστο κυτταροπλάσματος

καθώς και

$\mathbf{Z_2}$ =**-305,897 -0,237*** Εμβαδόν πυρήνα-**0,026*** Εμβαδόν κυτταροπλάσματος + **138,437*** Αναλογία πυρήνα/κυτταροπλάσματος + **16,143*** Φωτεινότητα πυρήνα + **265,068*** Φωτεινότητα κυτταροπλάσματος-**2,990*** Μικρότερη διάμετρος πυρήνα + **8,668*** Μεγαλύτερη διάμετρος πυρήνα + **214,658*** Επιμήκυνση πυρήνα **-20,677*** Σφαιρικότητα πυρήνα-**1,867*** Μικρότερη διάμετρος κυτταροπλάσματος + **2,564*** Μεγαλύτερη διάμετρος κυτταροπλάσματος + **99,325*** Επιμήκυνση κυτταροπλάσματος + **148,105*** Σφαιρικότητα κυτταροπλάσματος **-0,015*** Περίμετρος πυρήνα-**0,011*** Περίμετρος κυτταροπλάσματος-**18,873*** Θέση πυρήνα-**0,342*** Μέγιστο πυρήνα-**0,060*** Ελάχιστο πυρήνα-**0,019*** Μέγιστο κυτταροπλάσματος + **0,018*** Ελάχιστο κυτταροπλάσματος

και

$\mathbf{Z_2}$ = **-317,624 -0,038*** Εμβαδόν πυρήνα-**0,003*** Εμβαδόν κυτταροπλάσματος + **275,011*** Αναλογία πυρήνα/κυτταροπλάσματος **-0,006*** Φωτεινότητα πυρήνα + **0,183*** Φωτεινότητα κυτταροπλάσματος-**1,550*** Μικρότερη διάμετρος πυρήνα + **4,279*** Μεγαλύτερη διάμετρος πυρήνα + **141,527*** Επιμήκυνση πυρήνα + **136,958*** Σφαιρικότητα πυρήνα-**0,288*** Μικρότερη διάμετρος κυτταροπλάσματος + **1,006*** Μεγαλύτερη διάμετρος κυτταροπλάσματος **-7,176*** Επιμήκυνση κυτταροπλάσματος + **322,556*** Σφαιρικότητα κυτταροπλάσματος **-0,110*** Περίμετρος πυρήνα-**0,014*** Περίμετρος κυτταροπλάσματος + **5,365*** Θέση πυρήνα-**0,065*** Μέγιστο πυρήνα + **0,102*** Ελάχιστο πυρήνα-**0,040*** Μέγιστο κυτταροπλάσματος + **0,052*** Ελάχιστο κυτταροπλάσματος

για τις δύο βάσεις δεδομένων αντίστοιχα.

Για μία νέα καταχώρηση υπολογίζονται οι τιμές για τις συναρτήσεις $\mathbf{Z_1}$ και $\mathbf{Z_2}$. Αν $\mathbf{Z_1} > \mathbf{Z_2}$ τότε η παρατήρηση καταχωρείται στην ομάδα των υγιών ατόμων, διαφορετικά καταχωρείται στην ομάδα των μη υγιών ατόμων .

5.5 Σύγκριση αποτελεσμάτων μεταξύ των δύο βάσεων δεδομένων

Στις παραγράφους του κεφαλαίου 5 παρουσιάζονται τα αποτελέσματα των στατιστικών μεθόδων για την παλαιά και τη νέα βάση δεδομένων. Θεωρείται σκόπιμη μια σύγκριση των αποτελεσμάτων των δύο βάσεων για όλες τις μεθόδους. Η σύγκριση γίνεται μέσω του Πίνακα 5.14

	Παλαιά βάση δεδομένων			Νέα βάση δεδομένων		
Μέθοδος	**FN%**	**FP%**	**K%**	**FN%**	**FP%**	**K%**
Λογαριθμιστική παλινδρόμηση (logistic regression)	**2.70**	**2.50**	**97.10**	2.60	14.90	84.70
Διατάξιμη λογαριθμιστική παλινδρόμηση (ordinal logistic regression)			**49.00**			39.00
Ανάλυση σε συστάδες(cluster analysis)	11.20	28.10	**43.50**	11.40	15.20	28.60
Διαχωριστική ανάλυση (discriminant analysis)	**2.70**	**4.00**		5.80	13.20	

Πίνακας 5.14: Σύγκριση παλαιάς και νέας βάσης σύμφωνα με τα αποτελέσματα των στατιστικών μεθόδων

Όπως είναι φανερό, οι μέθοδοι που εφαρμόζονται έχουν σαφώς καλύτερα αποτελέσματα στην παλαιά βάση δεδομένων, κάτι που συμβαίνει και με τα αποτελέσματα των αλγορίθμων της πληροφορικής στο κεφάλαιο 6. Η λογαριθμιστική παλινδρόμηση επέφερε αρκετά ικανοποιητικά αποτελέσματα και στις δύο βάσεις δεδομένων με φαερή υπεροχή όμως της παλαιάς βάσης. Η διατάξιμη λογαριθμιστική παλινδρόμηση αδυνατεί να ταξινομήσει τα κύτταρα στις κατηγορίες μη υγιών κυττάρων σε ικανοποιητικό βαθμό και στις δύο βάσεις δεδομένων με ελαφρώς καλύτερα αποτελέσματα στην παλαιά βάση όπως φαίνεται από το δείκτη κάππα. Η συσταδική ανάλυση δεν ενδείκνυται στα επιχρίσματα Παπανικολάου αφού δεν καταφέρνει να ταξινομήσει τα κύτταρα στις δύο μεγάλες κατηγορίες υγιών και μη υγιών κυττάρων σε ενθαρρυντικό βαθμό σε καμία από τις δύο βάσεις δεδομένων. Τέλος, η διαχωριστική ανάλυση είναι μια ενδεδειγμένη μέθοδος, ιδιαίτερα για την παλαιά βάση αφού δίνει τη δυανατότητα διαχωρισμού των κυττάρων στις δύο κατηγορίες υγιών και μη υγιών διατηρώντας το σφάλμα σε χαμηλά επίπεδα.

ΚΕΦΑΛΑΙΟ 6

Αποτελέσματα αλγορίθμων Πληροφορικής

Στο κεφάλαιο αυτό παρατίθενται τα αποτελέσματα της εφαρμογής των αλγορίθμων του κεφαλαίου 2 για τις δύο βάσεις δεδομένων επιχρίσματος Παπανικολάου.

6.1 Αποτελέσματα αλγορίθμου αυστηρά C μέσων

Σε αυτή την παράγραφο παρουσιάζονται τα αποτελέσματα του αλγορίθμου HCM για τα παλιά και για τα νέα δεδομένα. Τα αποτελέσματα εξάγονται με τον αλγόριθμο FCM με q≈1 αφού πρακτικά δίνει τα ίδια αποτελέσματα με τον HCM.

Αποτελέσματα HCM						
Παλιά βάση δεδομένων				**Νέα βάση δεδομένων**		
Μέθοδος ταξινόμησης	**OE %**	**FP%**	**FN%**	**OE %**	**FP%**	**FN%**
Μη επιτηρούμενη	5.99	6.90	5.38	9.71	25.67	4.00
επιτηρούμενη	4.03	4.95	3.42	8.77	18.81	5.24

Πίνακας 6.1 : Αποτελέσματα αλγορίθμου HCM με όλα τα χαρακτηριστικά. (Martin2003)

Ο αριθμός των χρησιμοποιούμενων συστάδων καθορίστηκε χρησιμοποιώντας επαναξιολόγηση 2 δειγμάτων με 20 επαναλήψεις. Χρησιμοποιήθηκαν όλα τα χαρακτηριστικά. Αφού ο αριθμός των χρησιμοποιούμενων συστάδων είναι πλέον γνωστός, υπολογίζεται το σφάλμα του αλγορίθμου με τη χρήση επαναξιολόγησης 10 συστάδων με 20 επαναλήψεις. 50 και 100 συστάδες χρησιμοποιήθηκαν για επιτηρούμενη και μη επιτηρούμενη δημιουργία συστάδων αντίστοιχα. Από τον Πίνακα 6.1 φαίνεται ότι η επιτηρουμένη δημιουργία συστάδων δίνει αρκετά καλύτερα αποτελέσματα σε σχέση με τη μη επιτηρούμενη αλλά για την νέα βάση δεδομένωντα σφάλματα δεν πλησιάζουν το επιθυμητό 5%. Για την καλυτέρευση των σφαλμάτων έγινε επιλογή χαρακτηριστικών με την μέθοδο της προσομειωμένης

ανόπτησης. Τα αποτελέσματα της προσομειωμένης ανόπτησης καθώς και τα καινούρια σφάλματα φαίνονται αντίστοιχα στους πίνακες 6.2 και 6.3. Και πάλι, 50 και 100 συστάδες χρησιμοποιήθηκαν για την επιτηρούμενη και μη επιτηρούμενη δημιουργία συστάδων αντίστοιχα. Χρησιμοποιήθηκε επαναξιολόγηση 10 δειγμάτων με 20 επαναλήψεις για την εκτίμηση του σφάλματος.

Αποτελέσματα επιλογής χαρακτηριστικών		
Παλιά βάση δεδομένων		**Νέα βάση δεδομένων**
Μέθοδος ταξινόμησης	Επιλεγμένα χαρακτηριστικά	
Μη επιτηρούμενη	1,3,4,5,6,7,10,11,12,16,18,20	1,3,6,7,11,14,15,18,20
επιτηρούμενη	1,3,4,5,6,7,11,12,13,16,18	1,2,3,4,5,6,7,8,11,12,14,15,16,17,18,19,20

Πίνακας 6.2 : Επιλεγμένα χαρακτηριστικά για μη επιτηρούμενη και επιτηρούμενη δημιουργία συστάδων. Οι αριθμοί αντιστοιχούν στα χαρακτηριστικά που περιγράφησαν στον πίνακα 1.2

Αποτελέσματα HCM μετά την επιλογή χαρακτηριστικών						
Παλιά βάση δεδομένων				**Νέα βάση δεδομένων**		
Μέθοδος ταξινόμησης	OE %	FP%	FN%	OE %	FP%	FN%
Μη επιτηρούμενη	3.88	4.17	3.68	8.28	21.81	3.41
επιτηρούμενη	2.62	2.80	2.50	7.86	17.12	4.55

(Martin2003)

Πίνακας 6.3 : Αποτελέσματα για μη επιτηρούμενο και επιτηρούμενο αλγόριθμο HCM μετά την επιλογή χαρακτηριστικών. (Martin2003)

Μετά την επιλογή χαρακτηριστικών τα σφάλματα για την παλιά βάση δεδομένων και το ποσοστό των ψευδώς αρνητικών για την νέα βάση δεδομένων είναι κάτω από το επιθυμητό 5%. Και πάλι το αποτέλεσμα είναι καλύτερο για την επιτηρούμενη δημιουργία συστάδων.

Συμπέρασμα

Είναι φανερό ότι αποδεκτά αποτελέσματα εξήχθησαν από τον αλγόριθμο HCM για τα παλιά δεδομένα. Το σφάλμα για τα καινούρια δεδομένα είναι πολύ υψηλό και αυτό καθιστά τον αλγόριθμο HCM ακατάλληλο. Η επιτηρούμενη διαδικασία δημιουργίας συστάδων επιφέρει καλύτερα αποτελέσματα αλλά και πάλι μη αποδεκτά. Αυτό υποδεικνύει ότι οι συστάδες δεν είναι καλά διαχωρίσιμες μεταξύ τους. Ενδιαφέρον παρουσιάζει το γεγονός ότι όσο περισσότερες συστάδες χρησιμοποιούνται, τόσο μειώνεται το σφάλμα. Αυτό μπορεί να συμβαίνει επειδή κάποιες συστάδες συμπληρώνουν η μια την άλλη και σχηματίζουν πιο σύνθετες συστάδες στα δεδομένα.

6.2 Αποτελέσματα αλγορίθμου FCM

Σε αυτή την παράγραφο παρουσιάζονται τα αποτελέσματα του αλγορίθμου FCM για την παλιά και την καινούρια βάση δεδομένων. Εξάγονται αποτελέσματα ταξινόμησης για μη επιτηρούμενη και επιτηρούμενη δημιουργία συστάδων. Επίσης, ορίζονται παράμετροι όπως ο αριθμός των συστάδων και ο ασαφής εκθέτης. Έπειτα, παρουσιάζονται αποτελέσματα ταξινόμησης μετά από διαδικασία επιλογής χαρακτηριστικών.

Στον Πίνακα 6.4 φαίνονται τα αποτελέσματα του αλγορίθμου FCM χωρίς επιλογή χαρακτηριστικών.

Αποτελέσματα FCM						
Παλιά βάση δεδομένων				**Νέα βάση δεδομένων**		
Μέθοδος ταξινόμησης	OE %	FP%	FN%	OE %	FP%	FN%
Μη επιτηρούμενη	3.31	5.60	1.78	8.72	25.36	2.75
επιτηρούμενη	3.06	4.17	2.32	7.56	20.38	2.96

Πίνακας 6.4 : Αποτελέσματα μη επιτηρούμενου και επιτηρούμενου αλγορίθμου FCM, με όλα τα χαρακτηριστικά. (Martin2003)

Τα σφάλματα υπολογίστηκαν με επαναξιολόγηση 10 πτυχών με 20 επαναλήψεις. Ο Πίνακας 6.4 δείχνει ότι τα σφάλματα είναι ικανοποιητικά για την παλιά βάση δεδομένων, εκτός από τα ψευδώς θετικά στην μη επιτηρούμενη ταξινόμηση. Για την νέα βάση δεδομένων τα αποτελέσματα είναι ενθαρρυντικά μόνο για το ποσοστό των ψευδώς αρνητικών.

Μετά από επιλογή χαρακτηριστικών με προσομειωμένη ανόπτηση τα αποτελέσματα βελτιώνονται. Τα επιλεγμένα χαρακτηριστικά καθώς και τα καινούρια σφάλματα φαίνονται στους Πίνακες 6.5 και 6.6.

Αποτελέσματα επιλογής χαρακτηριστικών		
Παλιά βάση δεδομένων		**Νέα βάση δεδομένων**
Μέθοδος ταξινόμησης	Επιλεγμένα χαρακτηριστικά	
Μη επιτηρούμενη	1,2,3,4,5,6,7,10,11,12,15,16,17,18,20	1,2,3,4,5,6,7,10,12,13,14
επιτηρούμενη	2,3,4,5,6,7,10,11,12,15,16,18,20	1,3,4,5,6,7,10,14,15,17,18,19,20

Πίνακας 6.5 : Επιλεγμένα χαρακτηριστικά για μη επιτηρούμενη και επιτηρούμενη δημιουργία συστάδων. Οι αριθμοί αντιστοιχούν στα χαρακτηριστικά που περιγράφησαν στον πίνακα 1.2 (Martin2003)

Αποτελέσματα FCM μετά την επιλογή χαρακτηριστικών						
Παλιά βάση δεδομένων				**Νέα βάση δεδομένων**		
Μέθοδος ταξινόμησης	OE %	FP%	FN%	OE %	FP%	FN%
Μη επιτηρούμενη	1.81	2.78	1.17	8.08	20.97	3.45
επιτηρούμενη	1.64	2.02	1.38	6.10	13.89	3.29

Πίνακας 6.6 : Αποτελέσματα για μη επιτηρούμενο και επιτηρούμενο αλγόριθμο FCM μετά την επιλογή χαρακτηριστικών. (Martin2003)

Μετά την επιλογή χαρακτηριστικών, μειώθηκε το συνολικό σφάλμα και τα ψευδώς θετικά αλλά αυξήθηκαν ελαφρώς τα ψευδώς αρνητικά στην νέα βάση δεδομένων. Και πάλι τα σφάλματα είναι πανω από 5%. Αντίθετα για την παλιά βάση δεδομένων τα αποτελέσματα είναι πολύ ικανοποιητικά. Όλα τα σφάλματα βρίσκονται κάτω από 5%.

Παράμετροι

Έγινε η βέλτιστη επιλογή του εκθέτη q. Για αυτό το σκοπό το σφάλμα της επαναξιολόγησης εκτιμήθηκε για μια σειρά από διαφορετικές τιμές του q. Τα σφάλματα βρέθηκαν με επαναξιολόγηση 2 δειγμάτων με 20 επαναλήψεις. Τελικά επιλέχθηκε η τιμή q=1.25 για τη μη επιτηρούμενη δημιουργία συστάδων και q=1.20 για την επιτηρούμενηγια την παλιά βάση δεδομένων και q=1.13 για την μη επιτηρούμενη δημιουργία συστάδων και q=1.2 για την επιτηρούμενη για την νέα βάση δεδομένων

Η επιλογή του αριθμού των συστάδων έγινε με ανάλυση του αριθμού των συστάδων με το σφάλμα ελέγχου επαναξιολόγησης. Τα σφάλματα βρέθηκαν με επαναξιολόγηση 2 συστάδων με 20 επαναλήψεις.

Συμπέρασμα

Τα αποτελέσματα είναι αποδεκτά για τα παλιά δεδομένα αφού το σφάλμα βρέθηκε κάτω από το 5%. Για τα καινούρια δεδομένα, τα αποτελέσματα δεν είναι τόσο καλά. Ο αλγόριθμος FCM δεν μπόρεσε να μειώσει το σφάλμα κάτω από 5% που ήταν και ο στόχος. Παρόλα αυτά, τα σφάλματα του αλγορίθμου FCM είναι καλύτερα από του αλγορίθμου HCM. Φαίνεται ότι η επιλογή του fuzzy εκθέτη είναι πολύ σημαντική στον αλγόριθμο FCM. Αν επιλεχθεί κακός εκθέτης τότε ο αλγόριθμος FCM μπορεί να δώσει χειρότερα αποτελέσματα από τον HCM.

Η επιλογή χαρακτηριστικών βελτιώνει το σφάλμα και στις 2 περιπτώσεις, της μη επιτηρούμενης και επιτηρούμενης δημιουργίας συστάδων. Είναι επίσης φανερό, ότι η επιτηρούμενη δημιουργία συστάδων δίνει καλύτερα αποτελέσματα από την μη επιτηρούμενη και αυτό είναι μια ένδειξη ότι οι συστάδες δεν είναι καλά διαχωρίσιμες.

6.3 Αποτελέσματα αλγορίθμου Gustafson – Kessel

Ο fuzzy εκθέτης και ο αριθμός των συστάδων πρέπει να προκαθοριστούν για την μη επιτηρούμενη και επιτηρούμενη δημιουργία συστάδων. Για την μη επιτηρούμενη δημιουργία συστάδων επιλέχθηκε q=1.91 για την παλιά βάση δεδομένων και q=1.5 για τη νέα. Ο αριθμός των συστάδων που παράγει καλύτερα αποτελέσματα είναι 7 για την παλιά και 4 για τη νέα βάση δεδομένων. Για την επιτηρούμενη δημιουργία συστάδων, το σφάλμα ελέγχου φαίνεται να είναι μικρότερο όταν q≈1. Ο βέλτιστος αριθμός συστάδων για κάθε μια από τις κατηγορίες φυσιολογικών και μη φυσιολογικών κυττάρων είναι ο αριθμός 1.

.

Τα σφάλματα ελέγχου υπολογίστηκαν και φαίνονται στον Πίνακα 6.7.

.

Αποτελέσματα GK						
Παλιά βάση δεδομένων				**Νέα βάση δεδομένων**		
Μέθοδος ταξινόμησης	OE %	FP%	FN%	OE %	FP%	FN%
Μη επιτηρούμενη	14.34	33.71	1.31	10.90	41.32	0.04
επιτηρούμενη	4.44	2.13	5.97	7.04	17.99	3.11

Πίνακας 6.7 : Αποτελέσματα μη επιτηρούμενου και επιτηρούμενου αλγορίθμου GK, με όλα τα χαρακτηριστικά. .(Jens Byriel 1999)

Αν και η επιτηρούμενη ταξινόμηση είναι καλύτερη από τη μη επιτηρούμενη, η πλειοψηφία των σφαλμάτων είναι πάνω από 5% και ο αλγόριθμος δεν είναι αποδεκτός Είναι επίσης φανερό ότι στην μη επιτηρούμενη ταξινόμηση υπάρχει μεροληπτικό σφάλμα.

Για να εξάγουμε καλύτερα αποτελέσματα γίνεται επιλογή χαρακτηριστικών με προσομειωμένη ανόπτηση. Τα επιλεγμένα χαρακτηριστικά φαίνονται στον Πίνακα 6.8.

Αποτελέσματα επιλογής χαρακτηριστικών		
Παλιά βάση δεδομένων		**Νέα βάση δεδομένων**
Μέθοδος ταξινόμησης	Επιλεγμένα χαρακτηριστικά	
Μη επιτηρούμενη	1,3,4,6,7,8,10,14,16,17	1,6,7,8,9,10,13,14,17,18,20
επιτηρούμενη	1,2,3,4,5,7,8,9,10,11,12,14,15,16,17,18	1,2,4,5,6,7,8,9,10,12,13,14,15,16,17,18,19,20

Πίνακας 6.8 : Επιλεγμένα χαρακτηριστικά για μη επιτηρούμενη και επιτηρούμενη δημιουργία συστάδων. Οι αριθμοί αντιστοιχούν στα χαρακτηριστικά που περιγράφησαν στον πίνακα 1.2 .(Jens Byriel 1999)

Με αυτά τα χαρακτηριστικά, τα σφάλματα υπολογίζονται με επαναξιολόγηση 10 συστάδων με 20 επαναλήψεις. Τα αποτελέσματα φαίνονται στον Πίνακα 6.9

Αποτελέσματα GK μετά την επιλογή χαρακτηριστικών						
Παλιά βάση δεδομένων				**Νέα βάση δεδομένων**		
Μέθοδος ταξινόμησης	**OE %**	**FP%**	**FN%**	**OE %**	**FP%**	**FN%**
Μη επιτηρούμενη	9.92	23.83	0.86	7.68	22.66	2.31
επιτηρούμενη	2.89	4.46	1.77	6.06	13.88	3.26

Πίνακας 6.9 : Αποτελέσματα για μη επιτηρούμενο και επιτηρούμενο αλγόριθμο GK μετά την επιλογή χαρακτηριστικών. (Jens Byriel 1999)

Για την παλιά βάση δεδομένων, η μη επιτηρούμενη δημιουργία συστάδων παρουσιάζει καλύτερα αποτελέσματα σε σύγκριση με πριν την επιλογή χαρακτηριστικών αλλά και πάλι το σφάλμα είναι πάνω από 5%. Η επιτηρούμενη όμως δημιουργία συστάδων βελτιώθηκε ως προς το σφάλμα της και αποτελεί μια αποδεκτή διαδικασία, αφού το σφάλμα είναι κάτω από 5%. Στη νέα βάση δεδομένων, αν και τα σφάλματα βελτιώθηκαν, απέχουν αρκετά από το 5%. Άρα, ο αλγόριθμος δεν είναι αποδεκτός για τα νέα δεδομένα.

Συμπέρασμα

Η μη επιτηρούμενη δημιουργία συστάδων δίνει πολύ χειρότερα αποτελέσματα από την επιτηρούμενη. Τα αποτελέσματα της επιτηρούμενης δημιουργίας συστάδων πριν την επιλογή χαρακτηριστικών, είναι καλύτερα από τα αποτελέσματα της μη επιτηρούμενης δημιουργίας συστάδων μετά την επιλογή χαρακτηριστικών. Αυτό δείχνει ότι είναι πιο σημαντική η χρήση επιτηρούμενης διαδικασίας, από την διαδικασία επιλογής χαρακτηριστικών. Ο αλγόριθμος GK είναι αποδεκτός για την παλιά βάση δεδομένων αλλά όχι για την καινούρια

6.4 Αποτελέσματα μεθόδου ελαχίστων τετραγώνων

Αποτελέσματα μεθόδου ελαχίστων τετραγώνων (καινούρια βάση δεδομένων) (Norup2005)

Πρόβλημα 2 κατηγοριών

	Μέση τιμή	**Τυπική απόκλιση**	**Ελάχιστο**	**μέγιστο**
FN%	1.23	1.34	0.00	7.25
FP%	20.66	6.26	4.00	36.00
OE%	6.40	1.93	1.06	11.70
RMSE	0.25	0.02	0.21	0.30

Πίνακας 6.10 : Σφάλματα ελέγχου για το πρόβλημα των 2 κατηγοριών με κατώφλι = 1.5, k = 10 και επαναλήψεις = 50.

Παρατηρείται μεγάλη διαφορά ανάμεσα στα ψευδώς θετικά και ψευδώς αρνητικά. Αυτό είναι λογικό επειδή τα μη φυσιολογικά κύτταρα είναι πολύ περισσότερα από τα φυσιολογικά. Χρησιμοποιώντας διαφορετικό κατώφλι τα αποτελέσματα αλλάζουν σημαντικά. Για κατώφλι = 1.65 τα αποτελέσματα φαίνονται στον παρακάτω πίνακα. Παρόλα αυτά, το RMSE φαίνεται να μην εξαρτάται από το κατώφλι.

Κατώφλι	**FN%**	**FP%**	**OE%**
1.50	1.23	20.66	6.40
1.65	6.04	7.82	6.52

Πρόβλημα 7 κατηγοριών

Για να γίνει η ταξινόμηση στις 7 κατηγορίες θα χρησιμοποιηθούν 6 τιμές για το κατώφλι : κατώφλι = {1.5 2.5 3.5 4.5 5.5 6.5}. Τα αποτελέσματα φαίνονται στον Πίνακα 6.11.

	Μέση τιμή	**Τυπική απόκλιση**	**Ελάχιστο**	**μέγιστο**
FN%				
FP%				
OE%	42.86	4.67	26.09	56.04
RMSE	0.91	0.06	0.69	1.11

Πίνακας 6.11 : Σφάλματα ελέγχου για τις 7 κατηγορίες, k = 10, επαναλήψεις = 50.

Στον Πίνακα 6.12 φαίνονται τα αποτελέσματα του ελέγχου για το πρόβλημα των 2 κατηγοριών χρησιμοποιώντας όμως και τις 7 κατηγορίες για την δημιουργία του μοντέλου.

	Μέση τιμή	**Τυπική απόκλιση**	**Ελάχιστο**	**μέγιστο**
FN%	3.31	2.10	0.00	10.45
FP%	34.99	4.47	20.83	41.67
OE%	11.67	1.85	6.52	17.39
RMSE	0.91	0.06	0.69	1.11

Πίνακας 6.12 : Σφάλματα ελέγχου για το πρόβλημα των 2 κατηγοριών, χρησιμοποιώντας και τις 7 κατηγορίες, k = 10, επαναλήψεις = 50.

Στον Πίνακα 6.13 φαίνονται τα σφάλματα του ελέγχου για το πρόβλημα των 7 κατηγοριών, χρησιμοποιώντας δυαδικό δείκτη.

	Μέση τιμή	**Τυπική απόκλιση**	**Ελάχιστο**	**μέγιστο**
FN%	6.46	2.87	0.00	16.18
FP%	11.86	5.99	0.00	37.50
OE%	7.88	2.67	1.10	18.48
RMSE	1.13	0.13	0.76	1.57

Πίνακας 6.13 : Σφάλματα ελέγχου για το πρόβλημα των 7 κατηγοριών χρησιμοποιώντας δυαδικό δείκτη, k = 10, επαναλήψεις = 50.

6.5 Αποτελέσματα αλγορίθμων KNN και WKNN (Norup2005)

	OE%	**FN%**	**FP%**	**RMSE**	**N**	**κλιμακοποίηση**
WKNN	7.09	1.78	21.90	0.229	8	εύρος{0.1}
KNN	6.76	1.78	22.64	0.226	8	εύρος{0.1}

Πίνακας 6.14 : Αποτελέσματα WKNN και KNN με κλιμακοποίηση στο διάστημα {0.1}

Για βελτίωση των αποτελεσμάτων, εφαρμόστηκε μια διαφορετική κλιμακοποίηση, η κανονικοποίηση των δεδομένων. Με αυτόν τον τρόπο επιτεύχθηκε μια μικρή βελτίωση των σφαλμάτων, όπως φαίνεται στον Πίνακα 6.14.

	OE%	**FN%**	**FP%**	**RMSE**	**N**	**κλιμακοποίηση**
WKNN	6.54	2.52	17.77	0.232	3	Μέση τιμή=0/τυπική απόκλιση=1
KNN	6.65	2.52	18.18	0.234	3	Μέση τιμή=0/τυπική απόκλιση=1

Πίνακας 6.15 : Αποτελέσματα WKNN και KNN με κανονικοποίηση.

Όπως φαίνεται, τα αποτελέσματα δεν είναι εντυπωσιακά για καμία από τις 2 μεθόδους. Πολλά δείγματα αλλάζουν κατηγορία όταν το K αλλάζει. Το μεγάλο σφάλμα υποδεικνύει την μεγάλη αλληλεπίβαση μεταξύ κάποιων κατηγοριών. Όλες οι μέθοδοι είναι ευαίσθητες στην κατανομή των δεδομένων και για αυτό κάποιες κατηγορίες ευνοούνται στο αποτέλεσμα του μοντέλου. Για την βελτίωση αυτού του σφάλματος, χρησιμοποιήθηκε η μέθοδος NCC.

6.6 Αποτελέσματα αλγορίθμου NCC(Norup2005)

Ο ταξινομητής NCC υιοθετεί την κατηγορία του κοντινότερου κέντρου βαρύτητας. Κατά συνέπεια, το αποτέλεσμα του μοντέλου είναι ένα μονόμετρο μέγεθος που δείχνει τις κατηγορίες. Στον πίνακα 6.16 φαίνονται τα αποτελέσματα του αλγορίθμου NCC για 2 διαφορετικές περιπτώσεις κλιμακοποίησης.

	OE%	**FN%**	**FP%**	**RMSE**	**N**	**Κλιμακοποίηση**
NCC	6.22	5.04	9.50	0.249	21	εύρος{0.1}
NCC	5.13	4.30	7.44	0.226	21	Μέση τιμή=0/τυπική απόκλιση=1

Πίνακας 6.16 : Αποτελέσματα NCC για το πρόβλημα των 2 κατηγοριών, με 2 διαφορετικές μεθόδους κλιμακοποίησης.

Τα αποτελέσματα του πίνακα 6.16 είναι καλύτερα από των αλγορίθμων KNN και WKNN αλλά και πάλι δεν βρίσκονται κάτω από το 5%.

6.7 Αποτελέσματα μεθόδου NFI (Norup2005)

νέα βάση δεδομένων

Ο αριθμός των χρησιμοποιούμενων συστάδων εξαρτάται από την ακτίνα κατώφλι D_{thr} που υπολογίζεται με τον αλγόριθμο ECM και από τον αριθμό των κοντινότερων γειτόνων N_q . Αυτές οι 2 παράμετροι είναι κρίσιμες για την απόδοση του αλγόριθμου NFI .Αν το κατώφλι πάρει μικρές τιμές τότε ο αριθμός των συστάδων αυξάνεται με το N_q . Αλλά αν το κατώφλι πάρει μεγάλες τιμές τότε ο αριθμός των συστάδων είναι ανεξάρτητος του N .

Για την καλύτερη απόδοση του NFI, ο αριθμός των κοντινότερων γειτόνων πρέπει να είναι πολύ μεγαλύτερος από τον αριθμό των συστάδων. Η βέλτιστη ταξινόμηση επιτεύχθηκε για $D_{thr} = 0.300$.

	OE%	FN%	FP%	RMSE	N_q	D_{thr}
NFI	5.67	3.85	10.74	0.221	35	0.300

Πίνακας 6.17 : Αποτελέσματα NFI στα δεδομένα επιχρίσματος Παπανικολάου

Συμπέρασμα

Η μέθοδος NFI αποδίδει χειρότερα από την μέθοδο NCC και είναι και πιο αργή. Άρα η μέθοδος NCC είναι προτιμότερη αν και οι 2 δεν κατάφεραν να αγγίξουν το επιθυμητό 5% .

6.8 Αποτελέσματα αλγορίθμου GA (Y. Marinakis & G. Dounias 2006)

Στον πίνακα 6.18 φαίνονται τα αποτελέσματα του αλγορίθμου GA σε συνδυασμό με τους αλγορίθμους ταξινόμησης 1NN, KNN, WKNN, για τα παλιά και τα νέα δεδομένα και για το πρόβλημα των 2 κατηγοριών. Έγινε επαναξιολόγηση κ δειγμάτων για κ = 2,3,4,5,10,20. Εδώ, παρουσιάζονται τα αποτελέσματα για κ = 2 και κ = 20, που είναι και τα βέλτιστα. Παρατηρείται ότι, ο αλγόριθμος GA λειτουργεί βέλτιστα με τον 1NN. Ομοίως, στον πίνακα 6.19 φαίνονται τα αποτελέσματα για το πρόβλημα των 7 κατηγοριών. (Y. Marinakis, G. Dounias)

	Νέα βάση δεδομένων				**Παλιά βάση δεδομένων**			
	RMSE	**FN%**	**FP%**	**OE%**	**RMSE**	**FN%**	**FP%**	**OE%**
	Επαναξιολόγηση 2 δειγμάτων							
GA – 1NN	0.218	2.66	10.74	4.79	0.107	1.00	1.50	1.20
GA – KNN	0.235	1.33	17.35	5.56	0.089	0.00	2.00	0.80
GA - WKNN	0.242	1.92	16.94	5.88	0.094	0.66	1.50	1.00
	Επαναξιολόγηση 20 δειγμάτων							
GA – 1NN	0.022	0.00	1.25	0.32	0.000	0.00	0.00	0.00
GA – KNN	0.073	0.00	4.10	1.08	0.000	0.00	0.00	0.00
GA - WKNN	0.076	0.00	4.51	1.19	0.000	0.00	0.00	0.00

Πίνακας 6.18 : Αποτελέσματα GA αλγόριθμου για το πρόβλημα των 2 κατηγοριών.

	Νέα βάση δεδομένων		Παλιά βάση δεδομένων	
	RMSE	**OE%**	**RMSE**	**OE%**
	Επαναξιολόγηση 2 δειγμάτων			
GA – 1NN	1.019	55.12	0.883	1.40
GA – KNN	0.998	6.54	0.910	2.00
GA - WKNN	1.011	6.10	0.860	1.40
	Επαναξιολόγηση 20 δειγμάτων			
GA – 1NN	0.624	1.85	0.341	0.00
GA – KNN	0.660	1.63	0.432	0.00
GA - WKNN	0.678	1.95	0.372	0.00

Πίνακας 6.19 : Αποτελέσματα GA αλγόριθμου για το πρόβλημα των 7 κατηγοριών.

Για τα νέα δεδομένα ο μέσος ελάχιστος αριθμός χαρακτηριστικών είναι 8.5 και παρατηρήθηκε στο GA – KNN με Επαναξιολόγηση 2 δειγμάτων, ενώ ο μέσος μέγιστος αριθμός χαρακτηριστικών είναι 11.33 και παρατηρήθηκε στο GA – WKNN με επαναξιολόγηση 3 δειγμάτων, για το πρόβλημα των 2 κατηγοριών. Για το πρόβλημα των 7 κατηγοριών ο ελάχιστος μέσος αριθμός χαρακτηριστικών είναι 8 και παρατηρήθηκε στο GA – KNN με επαναξιολόγηση 2 δειγμάτων, ενώ ο μέσος μέγιστος αριθμός χαρακτηριστικών είναι 12.6 και παρατηρήθηκε στο GA – WKNN με επαναξιολόγηση 5 δειγμάτων. Για τα παλιά δεδομένα ο μέσος ελάχιστος αριθμός χαρακτηριστικών είναι 8.33 και παρατηρήθηκε στο GA – 1NN με επαναξιολόγηση 3 δειγμάτων, ενώ ο μέσος μέγιστος αριθμός χαρακτηριστικών είναι 13.5 και παρατηρήθηκε στο GA – KNN με 2 επαναξιολόγηση 2 δειγμάτων, για το πρόβλημα των 2 κατηγοριών. Για το πρόβλημα των 7 κατηγοριών ο μέσος ελάχιστος αριθμός χαρακτηριστικών είναι 8.75 και παρατηρήθηκε στο GA – 1NN επαναξιολόγηση 4 δειγμάτων, ενώ ο μέσος μέγιστος αριθμός χαρακτηριστικών είναι 13.6 και παρατηρήθηκε στο GA – KNN με επαναξιολόγηση 5 δειγμάτων.

6.9 Αποτελέσματα μεθόδου έρευνας tabu(Yannis Marinakis & George Dounias 2006)

Στον πίνακα 6.20 φαίνονται τα αποτελέσματα του αλγορίθμου TS, σε συνδυασμό με τους αλγορίθμους ταξινόμησης 1NN KNN και WKNN, για τα παλιά και νέα δεδομένα και για το πρόβλημα των 2 κατηγοριών. Παρουσιάζονται μόνο τα βέλτιστα αποτελέσματα. (Yiannis Marinakis, George Dounias)

	Νέα βάση δεδομένων				**Παλιά βάση δεδομένων**			
	RMSE	**FN%**	**FP%**	**OE%**	**RMSE**	**FN%**	**FP%**	**OE%**
	Επαναξιολόγηση 2 δειγμάτων							
Tabu – 1NN	**0.260**	**4.74**	**12.39**	**6.76**	0.140	1.33	3.00	2.00
Tabu 5NN	0.264	2.07	20.66	6.97	**0.128**	**1.66**	**2.00**	**1.80**
	Επαναξιολόγηση 3 δειγμάτων							
Tabu – 1NN	**0.213**	**2.22**	**11.14**	**4.57**	0.132	1.66	1.97	1.79
Tabu – W10NN	0.223	1.92	13.62	**5.01**	**0.089**	**0.66**	**1.99**	**1.19**
	Επαναξιολόγηση 3 δειγμάτων							
Tabu – 1NN	**0.208**	**2.37**	**9.91**	**4.36**	0.107	0.66	2.00	1.20

Tabu 5NN	0.230	2.22	14.05	5.34	**0.092**	**1.00**	**1.50**	**1.20**
Tabu – 10NN	0.263	1.33	22.71	6.97	**0.092**	**0.66**	**2.00**	**1.20**
	Επαναξιολόγηση 5 δειγμάτων							
Tabu – 1NN	**0.203**	**2.51**	**8.69**	**4.14**	0.074	0.66	1.50	1.00
Tabu – W5NN	0.216	1.92	12.38	4.68	**0.060**	**1.00**	**1.00**	**1.00**
	Επαναξιολόγηση 10 δειγμάτων							
Tabu – 1NN	**0.164**	**1.18**	**7.45**	**2.83**	0.042	0.66	0.50	0.60
Tabu – 3NN	0.197	0.44	13.66	3.92	**0.020**	**0.33**	**0.50**	**0.40**
	Επαναξιολόγηση 20 δειγμάτων							
Tabu – 1NN	**0.090**	**0.74**	**3.71**	**1.52**	**0.000**	**0.00**	**0.00**	**0.00**

Πίνακας 6.20 : Αποτελέσματα αλγορίθμου TS για το πρόβλημα των 2 κατηγοριών.

Στον πίνακα 6.21 φαίνονται τα αποτελέσματα του αλγορίθμου TS για το πρόβλημα των 7 κατηγοριών.

	Νέα βάση δεδομένων		**Παλιά βάση δεδομένων**	
	RMSE	**OE%**	**RMSE**	**OE%**
	Επαναξιολόγηση 2 δειγμάτων			
Tabu – 1NN	1.097	6.97	**0.979**	**2.40**
Tabu W10NN	**1.041**	**6.76**	1.027	2.80
	Επαναξιολόγηση 3 δειγμάτων			
Tabu – 1NN	**0.998**	**5.23**	0.901	2.39
Tabu – W8NN	1.103	6.43	**0.851**	**1.99**
	Επαναξιολόγηση 4 δειγμάτων			
Tabu – 1NN	**0.962**	**5.23**	0.818	1.40
Tabu 10NN	1.070	7.08	**0.794**	**1.20**
	Επαναξιολόγηση 5 δειγμάτων			
Tabu – 8NN	0.975	5.88	**0.727**	**0.80**
Tabu – W10NN	**0.966**	**5.66**	0.731	1.20

	Επαναξιολόγηση 10 δειγμάτων			
Tabu – 1NN	**0.905**	**4.25**	0.669	0.60
Tabu – W10NN	0.910	4.14	**0.659**	**1.00**
	Επαναξιολόγηση 20 δειγμάτων			
Tabu – 1NN	**0.776**	**3.04**	**0.487**	**0.00**

Πίνακας 6.21 : Αποτελέσματα του αλγορίθμου TS για το πρόβλημα των 7 κατηγοριών.

Για τα νέα δεδομένα ο μέσος ελάχιστος αριθμός χαρακτηριστικών είναι 6.5 και παρατηρήθηκε στο TS – 1NN με επαναξιολόγηση 2 δειγμάτων, ενώ ο μέσος μέγιστος αριθμός χαρακτηριστικών είναι 12.33 και παρατηρήθηκε στο TS – W3NN και στο TS - W10NN με επαναξιολόγηση 3 δειγμάτων, για το πρόβλημα των 2 κατηγοριών. Για το πρόβλημα των 7 κατηγοριών ο ελάχιστος μέσος αριθμός χαρακτηριστικών είναι 7.67 και παρατηρήθηκε στο TS – W10NN με επαναξιολόγηση 3 δειγμάτων, ενώ ο μέσος μέγιστος αριθμός χαρακτηριστικών είναι 12.67 και παρατηρήθηκε στο TS – 1NN με επαναξιολόγηση 3 δειγμάτων. Για τα παλιά δεδομένα ο μέσος ελάχιστος αριθμός χαρακτηριστικών είναι 7.4 και παρατηρήθηκε στο TS– 1NN με επαναξιολόγηση 3 δειγμάτων, ενώ ο μέσος μέγιστος αριθμός χαρακτηριστικών είναι 14.67 και παρατηρήθηκε στο TS – 10NN Επαναξιολόγηση 3 δειγμάτων, για το πρόβλημα των 2 κατηγοριών. Για το πρόβλημα των 7 κατηγοριών ο μέσος ελάχιστος αριθμός χαρακτηριστικών είναι 8.67 και παρατηρήθηκε στο TS – 3NN με επαναξιολόγηση 3 δειγμάτων, ενώ ο μέσος μέγιστος αριθμός χαρακτηριστικών είναι 14 και παρατηρήθηκε στο TS – 5NN, TS – W8NN με επαναξιολόγηση 2 δειγμάτων και TS – W8NN με επαναξιολόγηση 4 δειγμάτων.

6.10 Αποτελέσματα μεθόδου ACO(Y. Marinakis & G. Dounias 2006)

. Τα αποτελέσματα για τα παλιά και τα νέα δεδομένα, για το πρόβλημα των 2 και 7 κατηγοριών αντίστοιχα φαίνονται στους Πίνακες 6.22 και 6.23. (Y. Marinakis, G. Dounias)

	Νέα βάση δεδομένων				Παλιά βάση δεδομένων			
	RMSE	**FN%**	**FP%**	**OE%**	**RMSE**	**FN%**	**FP%**	**OE%**
	Επαναξιολόγηση 2 δειγμάτων							
ACO – 1NN	0.218	2.66	10.74	4.79	0.107	1.00	1.50	1.20
ACO - WKNN	0.249	1.77	18.59	6.21	0.107	0.66	2.00	1.20
	Επαναξιολόγηση 20 δειγμάτων							
ACO – 1NN	0.014	0.14	0.41	0.21	0.000	0.00	0.00	0.00
ACO - WKNN	0.069	0.00	4.10	1.08	0.000	0.00	0.00	0.00

Πίνακας 6.22 : Αποτελέσματα του αλγορίθμου ACO για το πρόβλημα των 2 κατηγοριών.

	Νέα βάση δεδομένων		Παλιά βάση δεδομένων	
	RMSE	**OE%**	**RMSE**	**OE%**
	Επαναξιολόγηση 2 δειγμάτων			
ACO – 1NN	1.030	5.78	0.898	1.60
ACO - WKNN	1.020	6.65	0.914	1.80
	Επαναξιολόγηση 20 δειγμάτων			
ACO – 1NN	0.638	2.07	0.366	0.00
ACO - WKNN	0.702	2.18	0.393	0.00

Πίνακας 6.23 : αποτελέσματα του αλγορίθμου ACO για το πρόβλημα των 7 κατηγοριών.

Για τα νέα δεδομένα ο μέσος ελάχιστος αριθμός χαρακτηριστικών είναι 8.5 και παρατηρήθηκε στο ACO – 1NN επαναξιολόγηση 4 δειγμάτων, ενώ ο μέσος μέγιστος αριθμός χαρακτηριστικών είναι 11 και παρατηρήθηκε στο ACO – WKNN με επαναξιολόγηση 4 δειγμάτων, και στο ACO 1NN με επαναξιολόγηση 3 δειγμάτων για το πρόβλημα των 2 κατηγοριών. Για το πρόβλημα των 7 κατηγοριών ο ελάχιστος μέσος αριθμός χαρακτηριστικών είναι 9 και παρατηρήθηκε στο ACO – 1NN με επαναξιολόγηση 2 δειγμάτων και στο ACO – WKNN με επαναξιολόγηση 5 δειγμάτων, ενώ ο μέσος μέγιστος αριθμός χαρακτηριστικών είναι 10.9 και παρατηρήθηκε στο ACO – WKNN με 10 επαναξιολόγηση 10 δειγμάτων. Για τα παλιά δεδομένα ο μέσος ελάχιστος αριθμός χαρακτηριστικών είναι 7.75 και παρατηρήθηκε στο ACO – 1NN με επαναξιολόγηση 4 δειγμάτων, ενώ ο μέσος μέγιστος αριθμός χαρακτηριστικών είναι 14.5 και παρατηρήθηκε στο ACO – WKNN με επαναξιολόγηση 2 δειγμάτων, για το πρόβλημα των 2 κατηγοριών. Για το πρόβλημα των 7 κατηγοριών ο μέσος ελάχιστος αριθμός χαρακτηριστικών είναι 8.25 και παρατηρήθηκε στο ACO – 1NN με επαναξιολόγηση 20 δειγμάτων, ενώ ο μέσος μέγιστος αριθμός χαρακτηριστικών είναι 11.5 και παρατηρήθηκε στο ACO – 1NN με επαναξιολόγηση 2 δειγμάτων.

Συμπέρασμα

Και οι 3 αλγόριθμοι που περιγράφησαν στην παράγραφο 2.12 δίνουν πολύ καλά αποτελέσματα τόσο στην ταξινόμηση, όσο και στην επιλογή χαρακτηριστικών. Σχεδόν σε όλες τις περιπτώσεις, χρησιμοποιήθηκαν λιγότερα από τα μισά χαρακτηριστικά.

6.11 Συγκριτικός πίνακας επιλεγμένων χαρακτηριστικών για τους αλγόριθμους GA, TS, ACO (Y. Marinakis & G. Dounias 2006)

Πρόβλημα 2 κατηγοριών				
	Νέα δεδομένα (τα 5 καλύτερα χαρακτηριστικά)	**Νέα δεδομένα (τα 5 χειρότερα χαρακτηριστικά)**	**Παλιά δεδομένα (τα 5 καλύτερα χαρακτηριστικά**	**Παλιά δεδομένα (τα 5 χειρότερα χαρακτηριστικά)**
ACO	5,4,3,7,1	12,17,13,8,9	16,15,5,18,1	11,12,6,14,9
TS	3,7,4,1,5	15,16,11,8,9	5,16,1,3,20	9,13,15,10,14
GA	4,3,5,7,1	19,12,13,8,9	5,16,7,15,19	11,13,17,9,14
Πρόβλημα 7 κατηγοριών				
	Νέα δεδομένα (τα 5 καλύτερα χαρακτηριστικά)	**Νέα δεδομένα (τα 5 χειρότερα χαρακτηριστικά)**	**Παλιά δεδομένα (τα 5 καλύτερα χαρακτηριστικά**	**Παλιά δεδομένα (τα 5 χειρότερα χαρακτηριστικά)**
ACO	3,5,1,4,7	16,19,12,8,9	16,5,4,3,11	14,18,13,9,8
TS	3,5,4,13,7	6,12,16,8,9	16,5,3,4,13	14,17,15,8,9
GA	3,5,14,4,7	6,12,9,16,8	5,16,4,3,11	12,15,17,9,8

Στον πίνακα, οι αριθμοί εμφανίζονται κατά φθίνουσα σειρά με βάση την συχνότητα προτίμησης, από τα πιο συχνά χρησιμοποιούμενα χαρακτηριστικά προς τα πιο σπάνια.

ΚΕΦΑΛΑΙΟ 7

Σύγκριση αποτελεσμάτων

Σε αυτήν την εργασία παρουσιάστηκαν διάφορες μέθοδοι ταξινόμησης και επιλογής χαρακτηριστικών τόσο στην πληροφορική όσο και στη στατιστική. Η προσπάθεια που έχει γίνει μέσω μεθόδων νευρωνικών δικτύων και γενετικών αλγορίθμων είναι μακρόχρονη και έτσι υπάρχουν πολυάριθμες μέθοδοι που έχουν χρησιμοποιηθεί ως σήμερα και γι'αυτό δεν παρουσιάζονται όλες. Επιλέχθηκαν να παρουσιαστούν οι μέθοδοι που επέφεραν τα πιο ενθαρρυντικά αποτελέσματα και μπορούν έως κάποιο βαθμό να συγκριθούν με μεθόδους στατιστικής ώστε να έχει νόημα η σύγκριση.

Καταρχάς να τονιστεί το γεγονός ότι τα αποτελέσματα των εφαρμογών των αλγορίθμων διαφέρουν αρκετά ανάμεσα στις δύο βάσεις δεδομένων ενώ και οι δύο βάσεις περιέχουν τα ίδια ακριβώς χαρακτηριστικά απλά έχουν μια διαφορά στην κατηγοριοποίηση των κυττάρων.

Ξεκινώντας λοιπόν με τους αλγορίθμους δημιουργίας συστάδων κ μέσων, του κεφαλαίου 6 και για τις δύο βάσεις δεδομένων τα καλύτερα αποτελέσματα επιτεύχθηκαν από τον αλγόριθμο ασαφών c-μέσων (FCM) με επιτηρούμενη δημιουργία συστάδων, μετά από επιλογή 13 εκ των 20 χρακτηριστικών με προσομοιωμένη ανόπτηση. Έτσι, με ποσοστό ψευδώς αρνητικών 1,17% και ψευδώς θετικών 2,02% είναι ένας αρκετά αξιόπιστος αλγόριθμος ταξινόμησης των κυττάρων επιχρίσματος Παπανικολάου της παλαιάς βάσης σε υγιή και μη υγιή. Για τη νέα βάση δεδομένων το ποσοστό των ψευδώς αρνητικών κυττάρων είναι 3,29% και των ψευδώς θετικών 13,89%. Η διαφορά στα δύο ποσοστά είναι αρκετά μεγάλη, γεγονός που υποδεικνύει την ύπαρξη μεροληψίας στα δεδομένα υπέρ των μη υγιών κυττάρων. Και στους δύο αλγόριθμους, αυστηρών c-μέσων (HCM) και ασαφών c-μέσων (FCM) η επιτηρούμενη δημιουργία συστάδων επέφερε καλύτερα αποτελέσματα, απόδειξη ότι οι συστάδες δεν είναι καλά διαχωρίσιμες μεταξύ τους.

Με την μέθοδο δημιουργίας συστάδων Gustafson- Kessel, τα καλύτερα αποτελέσματα και για τις δύο βασεις δεδομένων επιτεύχθηκαν με επιτηρούμενη δημιουργία συστάδων μετά από την επιλογή 18 εκ των 20 χρακτηριστικών. Ετσι για την παλαιά βάση δεδομένων το ποσοστό των ψευδώς αρνητικών ταξινομήσεων είναι 1,77% και των ψευδών θετικών 4,46% και για την νέα βάση δεδομένων τα αντίστοιχα ποσοστά είναι 3,26% και 13,88%. Και πάλι για τη νέα βάση δεδομένων υπάρχει αρκετά μεγάλη διαφορά στα ποσοστά ψευδώς αρνητικών και ψευδώς θετικών ταξινομήσεων. Για τις μεθόδους εγγύτερων γειτόνων KNN και WKNN τα αποτελέσματα δεν είναι εντυπωσιακά. Η διαφορά μεταξύ των ποσοστών ψευδώς θετικών και ψευδώς αρνητικών κυττάρων είναι πολύ μεγάλη. Οι μέθοδοι αυτές είναι ευαίσθητες στην κατανομή των δεδομένων και δίνουν μεροληπτικά αποτελέσματα επειδή κάποιες κατηγορίες ευνοούνται έναντι άλλων.

Τα πρόβλημα της εξάρτησης των αποτελεσμάτων από την κατανομή των δεδομένων το χειρίζεται ο αλγόριθμος κέντρου βαρύτητας κοντινότερης ομάδας (NCC) που συγκρινόμενος με τις μεθόδους KNN και WKNN δίνει καλύτερα αποτελέσματα. Εφαρμόστηκε στην νέα βάση δεδομένων και ταξινόμησε τις παρατηρήσεις με 4,3% ψευδώς αρνητικά κύτταρα και 7,44% ψευδώς θετικά.

Ο τελευταίος από τους αλγόριθμους ταξινόμησης με δημιουργία συστάδων που περιγράφεται σε αυτήν την εργασία είναι η νευροασαφής μέθοδος εξαγωγής συμπερασμάτων (NFI). Ο αλγόριθμος αυτός εφαρμόστηκε στη νέα βάση δεδομένων και τα αποτελέσματα ήταν ικανοποιητικά με 3,85% ψευδώς αρνητικά κύτταρα και 10,74% ψευδώς θετικά. Παρόλαυτά δεν απέδωσε καλύτερα από τον NCC και είναι και πιο αργός στην εφαρμογή του.

Ο αλγόριθμος κ μέσων που εφαρμόστηκε, βάση των αποτελεσμάτων που παρουσιάζονται στο κεφάλαιο 5, δεν απέδωσε ικανοποιητικά αποτελέσματα. Η αρχική επιλογή των συστάδων ήταν 7 αλλά η κατανομή των παρατηρήσεων σε αυτές δεν προσέγγισε την πραγματική κατανομή των δεδομένων στις 7 κατηγορίες για καμία πό τις δύο βάσεις δεδομένων. Σε επίπεδο δύο κατηγοριών, υγιών και μη υγιών κυττάρων, οι δύο συστάδες που δημιουργήθηκαν απέδωσαν ένα σφάλμα της τάξης του 11,2% για τα ψευδώς αρνητικά κύτταρα και 28,1% για τα ψευδώς θετικά για την παλαιά βάση δεδομένων και αντίστοιχα 11,4 % για τα ψευδώς αρνητικά και 15,2% για τα ψευδώς θετικά στην νέα βάση δεδομένων. Σαφώς τα αποτελέσματα δεν είναι ικανοποιητικά για καμία από τις δύο βάσεις δεδομένων κι έτσι εξάγεται το συμπέρασμα ότι μεταξύ όλων των προαναφερθείσων μεθόδων ταξινόμησης με δημιουργία συστάδων, ο αλγόριθμος που απέδωσε καλύτερα είναι ο NCC με μεγάλη διαφορά από τον αλγόριθμο κ μέσων της στατιστικής ανάλυσης.

Στο κεφάλαιο 6 παρουσιάζονται τα αποτελέσματα της μεθόδου ελαχίστων τετραγώνων που καταχρηστικά εφαρμόζετασι στην πληροφορική. Τα καλύτερα αποτελέσματα απέδωσε ο αλγόριθμος για το πρόβλημα των 7 κατηγοριών χρησιμοποιώντας δυαδικό δείκτη, αφού επέφερε ποσοστό ψευδώς αρνητικών 6,46% και ποσοστό ψευδώς θετικών 11,86% και πάλι όμως δεν κράτησε το σφάλμα κάτω από το επιθυμητό 5%. Στατιστικά, με τις μεθόδους που ίσως θα μπορούσε να συγκριθεί ο αλγόριθμος αυτός είναι με τα μοντέλα λογαριθμιστικής παλινδρόμησης μιας και δεν εφαρμόστηκε απλή παλινδρόμηση εφόσον το αποτέλεσμα δεν είναι συνεχής μεταβλητή. Έτσι λοιπόν, με τα μοντέλα λογαριθμιστικής παλινδρόμησης μπορούν να υπολογιστούν οι πιθανότητες για οποιοδήποτε κύτταρο να ανήκει στην κατηγορία των υγιών ή των μη υγιών κυττάρων καθώς και την πιθανότητα να ανήκει σε κάποια από τις συγκεκριμένες κατηγορίες δυσπλασίας. Το μοντέλο βρέθηκε ότι ερμηνεύει στατιστικά σημαντικό κομμάτι της μεταβλητότητας των χαρακτηριστικών με λιγότερα από τα 20 αρχικά χαρακτηριστικά. Για την παλαιά βάση δεδομένψν η λογαριθμιστική παλινδρόμηση απέδωσε σφάλμα 0,7% ψευδώς αρνητικών κυττάρων και 2.5% ψευδώς θετικών με τον δείκτη κάππα να παίρνει την τιμή 97,1% που είναι εξαιρετικά ικανοποιητικό ποσοστό έναντι του 49% που υπολογίστηκε στη διατάξιμη λογαριθμιστική παλινδρόμηση.

Για τη νέα βάση δεδομένων η λογαριθμιστική παλινδρόμηση απέδωσε σφάλμα 2,6% ψευδώς αρνητικών κυττάρων και 14.9% ψευδώς θετικών με υπολογισμένο δείκτη κάππα ίσο με 84,7% έναντι του 39% που απέδωσε η διατάξιμη λογαριθμιστική παλινδρόμηση. Το εξίσου σημαντικό εύρημα των μοντέλων λογαριθμιστικής παλινδρόμησης στη στατιστική είναι το ότι υπολογίζεται ξεχωριστά η επίδραση του κάθε χαρακτηριστικού στη διαμόρφωση του αποτελέσματος, δηλαδή υπολογίζεται η αύξηση ή αντίστοιχα η μείωση του κινδύνου όταν αυξάνει η τιμή κάποιου χαρακτηριστικού.

Με τον έλεγχο διαφοράς των μέσων τιμών για κάθε χρακτηριστικό διαπιστώθηκε επίσης η διαφοροποίηση όλων των χαρακτηριστικών, πλην της θέσης του πυρήνα, ανάμεσα στις κατηγορίες υγιών και μη υγιών κυττάρων και με την ανάλυση διασποράς διαπιστώθηκε η διαφοροποίηση των χαρακτηριστικών ανάμεσα στις 7 κατηγορίες αλλα και η αδυναμία των χαρακτηριστικών από μονα τους το καθένα να διαχωρίσει τα κύτταρα στις 7 κατηγορίες.

Στο θέμα της ταξινόμησης των κυττάρων στις κατηγορίες υγιών και μη υγιών κυττάρων, επέφερε πολύ ικανοποιητικά αποτελέσματα η διαχωριστική ανάλυση και για τις δύο βάσεις δεδομένων αλλά κυρίως για την παλαιά βάση αφού υπολόγισε ποσοστό ψευδώς αρνητικών κυττάρων 2,7% και ψευδώς αρνητικών κυττάρων 4% έναντι 3,1% και 13,2% αντίστοιχα για τη νέα βάση. Η διαχωριστική ανάλυση αξιοποίησε και τα 20 χαρακτηριστικα. Δεδομένων των ικανοποιητικών αποτελεσμάτων της και της ευκολίας εφαρμογής της, είναι μια μέθοδος που μπορεί να χρησιμοποιείται στα δεδομένα επιχρίσματος Παπανικολάου.

Η εφαρμογή της διαχωριστικής ανάλυσης επέφερε κι ένα άλλο πολύ σημαντικό αποτέλεσμα, την αξιολόγηση των χαρακτηριστικών των βάσεων με βάση την διαχωριστική τους ικανότητα. Κατέταξε δηλαδή τα 20 χαρακτηριστικά των βάσεων δεδομένων σε σειρά σύμφωνα με την ικανότητά τους να διαχωρίζουν τα κύτταρα στις 2 κατηγορίες των υγιών και των μη υγιών . Αυτή η κατάταξη έρχεται σε απόλυτη συμφωνία με τα αποτελέσματα των συγκρίσεων μέσων τιμών στα οποία η θέση του πυρήνα φαίνεται να μην διαφοροποιείται στατιστικά σημαντικά ανάμεσα στις 2 κατηγορίες υγιών και μη υγιών κυττάρων.

Αντίστοιχα, οι αλγόριθμοι επιλογής χρακατηριστικών γενετικός αλγόριθμος (GA), διερεύνηση ταμπού (TS) και βελτιστοποίηση αποικίας μηρμηγκιών (ACO) του κεφαλαίου 6, έχουν πολύ ικανοποιητικά αποτελέσματα τόσο στην επιλογή των καλύτερων χαρακτηριστικών όσο και στην κατάταξη των παρατηρήσεων των βάσεων στις κατηγορίες όταν συνδυάστηκαν με τους αλγόριθμους 1NN, KNN,WKNN. Παρατηρείται επίσης ότι στους αλγόριθμους λογαριθμιστικής παλινδρόμησης η επιλογή των χαρακτηριστικών με stepwise selection δίνει παρόμοια αποτελέσματα με αυτά των γενετικών αλγορίθμων αλλά όχι ίδια. Δηλαδή δεν επιλέγει τις ίδιες ακριβώς μεταβλητές για να χρησιμοποιηθούν στα

μοντέλα. Ομοίως και η διαχωριστική ανάλυση. Τα 5 καλύτερα χαρακτηριστικά της διαχωριστικής ανάλυσης δεν είναι τα ίδια 5 χαρακτηριστικά που επιλέγονται από τους αλγόριθμους GA, TS, ACO.

Στον Πίνακα 7.1 παρουσιάζονται συνοπτικά τα αποτελέσματα των αλγόριθμων στατιστικής και πληροφορικής για άμεση σύγκριση αποτελεσμάτων.

	ΠΛΗΡΟΦΟΡΙΚΗ					ΣΤΑΤΙΣΤΙΚΗ					
	Παλαιά βάση δεδομένων		Νέα βάση δεδομένων			Παλαιά βάση δεδομένων			Νέα βάση δεδομένων		
ΜΕΘΟΔΟΣ	**FN%**	**FP%**	**FN%**	**FP%**	**ΜΕΘΟΔΟΣ**	**FN%**	**FP%**	**K%**	**FN%**	**FP%**	**K%**
Αυστηρά c-μέσων (HCM-επιτηρούμενη,με επιλεγμένα χαρακτηριστικά)	**2.50**	**2.80**	4.55	17.12	**Ανάλυση σε συστάδες (cluster analysis)**	11.20	28.10	43.50	11.40	15.20	28.60
Ασαφών c-μέσων (FCM - επιτηρούμενη,με επιλεγμένα χαρακτηριστικά)	**1.38**	**2.02**	3.29	13.89	**Διαχωριστική ανάλυση (discriminan analysis)**	**2.70**	**4.00**		5.80	13.20	
GUSTAFSON-KESSEL	**1.77**	**4.46**	3.26	13.88	**Λογαριθμιστική παλινδρόμηση (logisti regression)**	**2.70**	**2.50**	**97.10**	**2.60**	**14.90**	**84.70**
Ελαχίστων τετραγώνω (LS- με δυαδικό δείκτη)			6.46	11.86	**Διατάξιμη λογαριθμιστική παλινδρόμηση (ordina logistic regression)**			49.00			39.00
Εγγύτερου γείτονα (KNN)			**2.52**	18.18							
Εγγύτερου γείτονα με στάθμες (WKNN)			**2.52**	17.77							
Κέντρου βαρύτητας κοντινότερης ομάδας (NCC)			**4.30**	**7.44**							
Νευροασαφής μέθοδος (NFI)			3.85	10.74							

Πίνακας 7.1: Σύγκριτικά αποτελέσματα μεθόδων στατιστικής και πληροφορικής (αναφέρονται μόνο τα καλύτερα αποτελέσματα από κάθε μέθοδο)

ΚΕΦΑΛΑΙΟ 8

Συζήτηση

Ο καρκίνος του τραχήλου είναι μια μορφή καρκίνου που απασχολεί πολύ έντονα τις γυναίκες σε όλον τον κόσμο. Η ανακάλυψη της μεθόδου χρωματισμού των κυττάρων από τον Γεώργιο Παπανικολάου έδωσε τη δυνατότητα ανίχνευσης των στοιχείων εκείνων μέσω των οποίων μπορεί να χαρακτηριστεί ένα κύτταρο υγιές ή προκαρκινικό και και να βρεθεί σε ποιο ακριβώς στάδιο βρίσκεται. Η εξέλιξη αυτή ήταν και είναι επαναστατική. Παρόλαυτά τα ποσοστά θνητότητας από την συγκεκριμένη μορφή καρκίνου ελαττώθηκαν πολύ σημαντικά αλλά όχι στο ελάχιστο ποσοστό που θα αναμενόταν δεδομένου του τεστ Παπανικολάου. Αυτό οφείλεται από την μία πλευρά στην άγνοια και στην αμέλεια των γυναικών οι οποίες σε πολύ σημαντικό ποσοστό δεν κάνουν την εξέταση Παπανικολάου στα χρονικά διαστήματα που επιβάλλεται και κατά δεύτερον στα σφάλματα κατά τη διάγνωση που είναι και αυτό που μας απασχολεί σε αυτήν την εργασία.

Η διάγνωση των κυττάρων γίνεται από έμπειρους και εξειδικευμένους κυτταροτεχνικούς. Όμως ο τεράστιος όγκος των κυττάρων προς εξέταση καθώς και οι πολύ λεπτομερείς διαφορές στα χαρακτηριστικά των κυττάρων δεν μπορούν να εξασφαλίσουν την ύπαρξη ενός πολύ χαμηλού σφάλματος. Η διάγνωση μέσω υπολογιστών εξελίχθηκε τα τελευταία χρόνια αλλά το κόστος είναι πολύ μεγάλο. Ετσι, με την δημιουργία των δύο βάσεων δεδομένων και μέσω διαφόρων αλγορίθμων επιτεύχθηκε η ταξινόμηση των κυττάρων με πολύ ικανοποιητικά αποτελέσματα για κάποιους από αυτούς. Οποιαδήποτε γνώση όμως πάνω στη διάγνωση των κυττάρων καθώς και στην επίδραση των χαρακτηριστικών των βάσεων είναι πολύ σημαντική για την εξέλιξη της διαδικασίας αυτόματης ταξινόμησης των κυττάρων. Μέσω των μεθόδων στατιστικής ανάλυσης βρέθηκαν συσχετίσεις των μεταβλητών με τις κατηγορίες των κυττάρων αλλά και επιδράσεις αυτών στο αποτέλεσμα. Παρουσιά στηκαν και άλλοι μέθοδοι ταξινόμησης που άλλοι με καλά και άλλοι με όχι και τόσο ικανοποιητικά αποτελέσματα, χειρίστηκαν τις δύο βάσεις και άνοιξαν των δρόμο για μελλοντικές στατιστικές αναλύσεις.

Δεν χρειάζεται να τονιστεί ξανά η σημασία οποιουδήποτε ευρήματος από την ανάλυση των δεδομένων των βάσεων επιχρίσματος Παπανικολάου. Όπως διαπιστώθηκε, οι κατηγορίες των κυττάρων δεν είναι πολύ καλά φυσικά διαχωρίσιμες και αυτό δυσκολεύει την ταξινόμηση άλλων κυττάρων σε μία από τις κατηγορίες. Ένα θέμα λοιπόν που προκύπτει είναι η επανεξέταση των βάσεων δεδομένων και των τιμών των χαρακτηριστικών αυτών για τη δημιουργία μιας καλύτερης εκδοχής τους στο μέλλον για να επανεξεταστούν οι ήδη εφαρμοσμένες μέθοδοι και να αξιολογηθούν. Ίσως κάποιες παρατηρήσεις δεν πρέπει να συμπεριλαμβάνονται στις βάσεις.

Από άποψη ανάλυσης των δεδομένων, στην πληροφορική έχουν εφαρμοστεί ήδη πολλοί ταξινομητές και αρκετές μέθοδοι επιλογής χαρακτηριστικών. Ίσως θα μπορούσε να εξεταστεί ο συνδυασμός των μεθόδων επιλογής χαρακτηριστικών GA, TS, ACO με άλλους ταξινομητές εκτός από τους 1NN, KNN, WKNN. Στην στατιστική, μιας και δεν υπάρχει αρκετό παρελθόν ανάλυσης, μπορούν να εφαρμοστούν αρκετές ακόμα μέθοδοι ταξινόμησης αλλά και δημιουργίας μοντέλων. Συγκεκριμένα για τις μεθόδους που εφαρμόστηκαν στην διπλωματική αυτή, μπορούν να εξεταστούν και διάφορες παραλλαγές τους στο μέλλον. Έτσι, στην ανάλυση κατά συστάδες κ μέσων, δεν εξετάστηκε η περίπτωση δημιουργίας παραπάνω από 10 συστάδων. Στην διαχωριστική ανάλυση, χρησιμοποιήθηκαν ίσες εκ των προτέρων πιθανότητες υγιών και μη υγιών κυττάρων . Μια ανάλυση με εκ των προτέρων πιθανότητες υπέρ των υγιών κυττάρων, όπως άλλωστε είνσι στην πραγματικότητα θα επέφερε ακόμα πιο ικανοποιητικά αποτελέσματα. Απλά η συγκεκριμένη εφαρμογή της ανάλυσης έγινε με περισσότερη αυστηρότητα. Επίσης, μπορεί να πραγματοποιηθεί ανάλυση με επιλογή χαρακτηριστικών αφού στην εργασία αυτή χρησιμοποιήθηκαν και τα 20 χαρακτηριστικά των βάσεων.

Στα μοντέλα λογαριθμιστικής παλινδρόμησης χρησιμοποιήθηκε επιλογή χαρακτηριστικών με stepwise selection. Η μελλοντική δημιουργία των μοντέλων με διαφορετικό τρόπο επιλογής των χαρακτηριστικών είναι ένα θέμα προς εξέταση.

Τέλος, ίσως να είχε νόημα η ένωση των δύο βάσεων δεδομένων με κάποιους περιορισμούς και η ανάλυση των δεδομένων συνολικά, κυρίως για την ταξινόμηση των κυττάρων στις δύο κατηγορίες υγιών και μη υγιών κυττάρων μιας και οι 7 κατηγορίες είναι διαφορετικές στις δύο βάσεις.

ΠΑΡΑΡΤΗΜΑ

μεταβλητή	Μέση τιμή	Τυπική απόκλιση	Ελάχιστη τιμή	Μέγιστη τιμή
Εμβαδόν πυρήνα	176.33	136.920	12.90	840.92
Εμβαδόν κυτταροπλάσματος	1187.40	1257.001	63.56	6008.90
Αναλογία πυρήνα/κυτταροπλάσματος	0.2248262	0.1719713	0.0026522	0.8669367
Φωτεινότητα πυρήνα	0.45	0.094	0.27	0.79
Φωτεινότητα κυτταροπλάσματος	0.76	0.07	0.53	0.99
Μικρότερη διάμετρος πυρήνα	12.25	4.84	3.28	32.96
Μεγαλύτερη διάμετρος πυρήνα	17.85	7.886	5.59	51.95
Επιμήκυνση πυρήνα	0.72	0.146	0.26	0.99
Σφαιρικότητα πυρήνα	0.65	0.153	0.12	1.03
Μικρότερη διάμετρος κυτταροπλάσματος	33.31	16.086	8.13	84.73
Μεγαλύτερη διάμετρος κυτταροπλάσματος	53.65	20.548	20.58	125.7
Επιμήκυνση κυτταροπλάσματος	0.63	0.183	0.16	0.98
Σφαιρικότητα κυτταροπλάσματος	0.42	0.199	0.05	0.91
Περίμετρος πυρήνα	72.49	38.412	17.48	304.48
Περίμετρος κυτταροπλάσματος	318.96	265.977	78.38	2406.31
Θέση πυρήνα	0.21	0.169	0.01	1.14
Μέγιστο πυρήνα	54.41	38.056	9,00	369,00
Ελάχιστο πυρήνα	42.18	35.353	3,00	331,00
Μέγιστο κυτταροπλάσματος	454.45	857.885	21,00	15155,00
Ελάχιστο κυτταροπλάσματος	361.31	670.140	25,00	12049,00

Πίνακας 4.1: Περιγραφικοί δείκτες των 500 παρατηρήσεων της βάσης δεδομένων.

Κυλινδρικά επιθήλια				
Μεταβλητή	**Μέση τιμή**	**Τυπική απόκλιση**	**Ελάχιστη τιμή**	**Μέγιστη τιμ**
Εμβαδόν πυρήνα	89.38	45.347	35.82	344.75
Εμβαδόν κυτταροπλάσματος	309.24	161.708	74.59	663.07
Αναλογία πυρήνα/κυτταροπλάσματος	0.26	0.160	0.08	0.76
Φωτεινότητα πυρήνα	0.47	0.054	0.35	0.60
Φωτεινότητα κυτταροπλάσματος	0.73	0.033	0.65	0.83
Μικρότερη διάμετρος πυρήνα	8.61	1.578	5.55	12.33
Μεγαλύτερη διάμετρος πυρήνα	14.66	4.906	9.83	38.25
Επιμήκυνση πυρήνα	0.63	0.182	0.25	0.93
Σφαιρικότητα πυρήνα	0.55	0.169	0.22	0.87
Μικρότερη διάμετρος κυτταροπλάσματος	16.86	5.095	8.13	28.15
Μεγαλύτερη διάμετρος κυτταροπλάσματος	35.43	8.665	20.58	57.81
Επιμήκυνση κυτταροπλάσματος	0.51	0.211	0.16	0.96
Σφαιρικότητα κυτταροπλάσματος	0.36	0.217	0.05	0.85
Περίμετρος πυρήνα	64.23	17.986	34.86	136.07
Περίμετρος κυτταροπλάσματος	184.59	42.440	109.34	294.29
Θέση πυρήνα	0.43	0.237	0.01	1.14
Μέγιστο πυρήνα	31.70	12.825	13,00	90,00
Ελάχιστο πυρήνα	18.38	11.248	3,00	72,00
Μέγιστο κυτταροπλάσματος	103.44	49.223	29,00	231,00
Ελάχιστο κυτταροπλάσματος	82,00	38.964	25,00	198,00

Πίνακας 4.2 : Περιγραφικοί δείκτες των 500 παρατηρήσεων της βάσης δεδομένων για την κατηγορία κυλινδρικά επιθήλια

Παραβασικά λεπιδοειδή επιθήλια				
Μεταβλητή	**Μέση τιμή**	**Τυπική απόκλιση**	**Ελάχιστη τιμ**	**Μέγιστη τιμή**
Εμβαδόν πυρήνα	88.44	57.715	15.59	364.04
Εμβαδόν κυτταροπλάσματος	553.67	262.093	207.54	1475.94
Αναλογία πυρήνα/κυτταροπλάσματος	0.14	0.087	0.01	0.56
Φωτεινότητα πυρήνα	0.34	0.022	0.27	0.38
Φωτεινότητα κυτταροπλάσματος	0 .63	0.055	0.53	0.79
Μικρότερη διάμετρος πυρήνα	9.13	2.835	4.11	21.4
Μεγαλύτερη διάμετρος πυρήνα	12.62	3.491	7.12	23.40
Επιμήκυνση πυρήνα	0.72	0.115	0.48	0.97
Σφαιρικότητα πυρήνα	0.65	0.136	0.38	1.02
Μικρότερη διάμετρος κυτταροπλάσματος	25.24	25.248	15.74	39.87
Μεγαλύτερη διάμετρος κυτταροπλάσματος	36.54	8.483	21.28	56.32
Επιμήκυνση κυτταροπλάσματος	0.70	0.149	0.41	0.96
Σφαιρικότητα κυτταροπλάσματος	0.51	0.124	0.23	0.79
Περίμετρος πυρήνα	56.87	25.932	29.21	159.29
Περίμετρος κυτταροπλάσματος	174.31	54.107	101.81	331.68
Θέση πυρήνα	0.14	0.116	0.01	0.75
Μέγιστο πυρήνα	37.24	20.611	14,00	126,00
Ελάχιστο πυρήνα	26.18	18.327	5,00	112,00
Μέγιστο κυτταροπλάσματος	116.08	59.961	51,00	354,00
Ελάχιστο κυτταροπλάσματος	101.68	51.155	34,00	311,00

Πίνακας 4.3 : Περιγραφικοί δείκτες των 50 παρατηρήσεων της βάσης δεδομένων για την κατηγορία παραβασικά λεπιδοειδή επιθήλια

Ενδιάμεσα λεπιδοειδή επιθήλια				
μεταβλητή	**Μέση τιμή**	**Τυπική απόκλιση**	**Ελάχιστη τιμή**	**Μέγιστη τιμή**
Εμβαδόν πυρήνα	64.72	30.216	12.90	150.49
Εμβαδόν κυτταροπλάσματος	3482.23	1183.166	1410.12	6008.90
Αναλογία πυρήνα/κυτταροπλάσματος	0.02	0.009	0.02	0.04
Φωτεινότητα πυρήνα	0.40	0.060	0.32	0.56
Φωτεινότητα κυτταροπλάσματος	0.80	0.046	0.63	0.86
Μικρότερη διάμετρος πυρήνα	8.09	1.951	3.28	12.16
Μεγαλύτερη διάμετρος πυρήνα	10.56	2.256	6.27	16.89
Επιμήκυνση πυρήνα	0.76	0.106	0.50	0.92
Σφαιρικότητα πυρήνα	0.70	0.118	0.39	0.88
Μικρότερη διάμετρος κυτταροπλάσματος	61.45	12.197	37.73	84.73
Μεγαλύτερη διάμετρος κυτταροπλάσματος	79.18	13.475	54.78	106.41
Επιμήκυνση κυτταροπλάσματος	0.78	0.118	0.49	0.98
Σφαιρικότητα κυτταροπλάσματος	0.69	0.131	0.40	0.91
Περίμετρος πυρήνα	41.08	12.224	28.03	98.53
Περίμετρος κυτταροπλάσματος	705.29	483.232	279.42	2406.31
Θέση πυρήνα	0.13	0.094	0.01	0.54
Μέγιστο πυρήνα	23.70	9.802	10,00	51,00
Ελάχιστο πυρήνα	15.08	8.677	3,00	39,00
Μέγιστο κυτταροπλάσματος	1340.38	609.545	474,00	2735,00
Ελάχιστο κυτταροπλάσματος	1066.64	510.918	350,00	2130,00

Πίνακας 4.4 : Περιγραφικοί δείκτες των 50 παρατηρήσεων της βάσης δεδομένων για την κατηγορία ενδιάμεσα λεπιδοειδή επιθήλια

Επιφανειακά λεπιδοειδή επιθήλια				
μεταβλητή	**Μέση τιμή**	**Τυπική απόκλιση**	**Ελάχιστη τιμή**	**Μέγιστη τιμή**
Εμβαδόν πυρήνα	30.90	5.978	19.23	46.09
Εμβαδόν κυτταροπλάσματος	3166.41	891.637	939.75	5245.81
Αναλογία πυρήνα/κυτταροπλάσματος	0.01	0.004	0.01	0.03
Φωτεινότητα πυρήνα	0.31	0.027	0.28	0.41
Φωτεινότητα κυτταροπλάσματος	0.77	0.046	0.66	0.86
Μικρότερη διάμετρος πυρήνα	5.98	0.623	4.13	7.66
Μεγαλύτερη διάμετρος πυρήνα	7.11	0.778	5.59	9.41
Επιμήκυνση πυρήνα	0.84	0.090	0.55	0.98
Σφαιρικότητα πυρήνα	0.77	0.087	0.44	0.90
Μικρότερη διάμετρος κυτταροπλάσματος	56.50	8.789	37.84	75.26
Μεγαλύτερη διάμετρος κυτταροπλάσματος	76.93	10.478	51.12	103.96
Επιμήκυνση κυτταροπλάσματος	0.73	0.102	0.47	0.92
Σφαιρικότητα κυτταροπλάσματος	0.67	0.116	0.33	0.88
Περίμετρος πυρήνα	21.98	2.790	17.48	31.27
Περίμετρος κυτταροπλάσματος	624.32	335.573	221.22	1509.79
Θέση πυρήνα	0.17	0.081	0.03	0.34
Μέγιστο πυρήνα	14.54	3.459	9,00	23,00
Ελάχιστο πυρήνα	9.42	4.005	3,00	21,00
Μέγιστο κυτταροπλάσματος	1219.48	448.748	451,00	2374,00
Ελάχιστο κυτταροπλάσματος	956.08	398.187	335,00	2053,00

Πίνακας 4.5 : Περιγραφικοί δείκτες των 50 παρατηρήσεων της βάσης δεδομένων για την κατηγορία επιφανειακά λεπιδοειδή επιθήλια

Ελαφριά λεπιδοειδής μη κερατινώδης δυσπλασία				
μεταβλητή	**Μέση τιμή**	**Τυπική απόκλιση**	**Ελάχιστη τιμή**	**Μέγιστη τιμή**
Εμβαδόν πυρήνα	258.38	124.309	68.64	753.19
Εμβαδόν κυτταροπλάσματος	1110.93	679.822	188.63	3916.86
Αναλογία πυρήνα/κυτταροπλάσματος	0.22	0.120	0.04	0.68
Φωτεινότητα πυρήνα	0.49	0.069	0.36	0.73
Φωτεινότητα κυτταροπλάσματος	0.79	0.053	0.66	0.96
Μικρότερη διάμετρος πυρήνα	15.49	3.646	8.02	27.51
Μεγαλύτερη διάμετρος πυρήνα	22.62	6.738	13.31	48.40
Επιμήκυνση πυρήνα	0.70	0.135	0.34	0.97
Σφαιρικότητα πυρήνα	0.63	0.141	0.29	0.93
Μικρότερη διάμετρος κυτταροπλάσματος	33.79	10.106	16.01	62.72
Μεγαλύτερη διάμετρος κυτταροπλάσματος	61.82	19.724	23.88	125.70
Επιμήκυνση κυτταροπλάσματος	0.58	0.181	0.20	0.93
Σφαιρικότητα κυτταροπλάσματος	0.37	0.138	0.10	0.76
Περίμετρος πυρήνα	98.03	43.756	47.44	304.48
Περίμετρος κυτταροπλάσματος	309.36	102.996	103.16	670.13
Θέση πυρήνα	0.19	0.171	0.01	1.06
Μέγιστο πυρήνα	71.95	30.555	28.00	193.00
Ελάχιστο πυρήνα	57.67	28.882	13.00	170.00
Μέγιστο κυτταροπλάσματος	381.72	383.118	61.00	3100.00
Ελάχιστο κυτταροπλάσματος	295.93	215.106	42.00	1337.00

Πίνακας 4.6 : Περιγραφικοί δείκτες των 100 παρατηρήσεων της βάσης δεδομένων για την κατηγορία ελαφριά λεπιδοειδής μη κερατινώδης δυσπλασία

Μεσαία λεπιδοειδής μη κερατινώδης δυσπλασία				
μεταβλητή	**Μέση τιμ**	**Τυπική απόκλιση**	**Ελάχιστη τιμή**	**Μέγιστη τιμή**
Εμβαδόν πυρήνα	263.03	135.969	57.87	840.92
Εμβαδόν κυτταροπλάσματος	664.29	432.272	129.07	2502.35
Αναλογία πυρήνα/κυτταροπλάσματος	0.31	0.130	0.10	0.86
Φωτεινότητα πυρήνα	0.47	0.071	0.33	0.71
Φωτεινότητα κυτταροπλάσματος	0.77	0.051	0.65	0.99
Μικρότερη διάμετρος πυρήνα	15.77	4.124	7.19	32.96
Μεγαλύτερη διάμετρος πυρήνα	22.90	6.722	14.39	51.95
Επιμήκυνση πυρήνα	0.70	0.140	0.38	0 .97
Σφαιρικότητα πυρήνα	0.63	0.155	0.11	0.92
Μικρότερη διάμετρος κυτταροπλάσματος	28.53	8.558	10.32	52.38
Μεγαλύτερη διάμετρος κυτταροπλάσματος	51.02	14.995	29.89	116.30
Επιμήκυνση κυτταροπλάσματος	0.58	0.173	0.20	0.91
Σφαιρικότητα κυτταροπλάσματος	0.31	0.123	0.06	0.62
Περίμετρος πυρήνα	93.56	34.911	50.85	234.97
Περίμετρος κυτταροπλάσματος	250.00	74.333	111.79	588.61
Θέση πυρήνα	0.21	0.115	0.02	0.52
Μέγιστο πυρήνα	78.23	47.617	21.00	369.00
Ελάχιστο πυρήνα	63.59	45.036	9.00	331.00
Μέγιστο κυτταροπλάσματος	382.24	1527.871	28.00	15155.00
Ελάχιστο κυτταροπλάσματος	299.17	1194.608	36.00	12049.00

Πίνακας 4.7 : Περιγραφικοί δείκτες των 100 παρατηρήσεων της βάσης δεδομένων για την κατηγορία μεσαία λεπιδοειδής μη κερατινώδης δυσπλασία

Βαριά λεπιδοειδής μη κερατινώδης δυσπλασία				
μεταβλητή	**Μέση τιμ**	**Τυπική απόκλιση**	**Ελάχιστη τιμή**	**Μέγιστη τιμή**
Εμβαδόν πυρήνα	223.48	126.842	70.70	825.14
Εμβαδόν κυτταροπλάσματος	405.97	229.533	63.56	1144.72
Αναλογία πυρήνα/κυτταροπλάσματος	0.37	0.153	0.07	0.83
Φωτεινότητα πυρήνα	0.48	0.095	0.31	0.79
Φωτεινότητα κυτταροπλάσματος	0.76	0.065	0.54	0.95
Μικρότερη διάμετρος πυρήνα	14.03	3.711	7.23	26.22
Μεγαλύτερη διάμετρος πυρήνα	21.24	6.135	11.76	44.97
Επιμήκυνση πυρήνα	0.68	0.148	0.32	0.97
Σφαιρικότητα πυρήνα	0.61	0 .155	0.24	0.91
Μικρότερη διάμετρος κυτταροπλάσματος	24.19	6.235	10.38	41.94
Μεγαλύτερη διάμετρος κυτταροπλάσματος	41.37	12.332	20.64	86.32
Επιμήκυνση κυτταροπλάσματος	0.61	0.168	0.17	0.97
Σφαιρικότητα κυτταροπλάσματος	0.30	0.123	0.07	0.61
Περίμετρος πυρήνα	78.78	25.333	42.05	163.79
Περίμετρος κυτταροπλάσματος	191.19	56.968	78.38	358.29
Θέση πυρήνα	0.23	0.161	0.03	1.13
Μέγιστο πυρήνα	68.28	31.416	27.00	211.00
Ελάχιστο πυρήνα	55.11	30.992	18.00	206.00
Μέγιστο κυτταροπλάσματος	118.61	72.643	21.00	394.00
Ελάχιστο κυτταροπλάσματος	108.26	65.104	29.00	341.00

Πίνακας 4.8 : Περιγραφικοί δείκτες των 100 παρατηρήσεων της βάσης δεδομένων για την κατηγορία βαριά λεπιδοειδής μη κερατινώδης δυσπλασία

Πίνακας 4.9: Συσχετίσεις όλων των μεταβλητών ανά δύο

μεταβλητή	Μέση τιμή	Τυπική απόκλιση	Ελάχιστη τιμή	Μέγιστη τιμή
Εμβαδόν πυρήνα	2990.82	1878.177	316.87	10794.63
Εμβαδόν κυτταροπλάσματος	14053.90	20357.670	467.87	127313.80
Αναλογία πυρήνα/κυτταροπλάσματος	0.35	0.213	0.00	0.88
Φωτεινότητα πυρήνα	90.49	21.780	17.91	174.99
Φωτεινότητα κυτταροπλάσματος	139.35	25.012	69.88	230.48
Μικρότερη διάμετρος πυρήνα	52.49	18.754	13.97	112.72
Μεγαλύτερη διάμετρος πυρήνα	71.19	23.242	19.41	158.82
Επιμήκυνση πυρήνα	0.75	0.156	0.30	1.19
Σφαιρικότητα πυρήνα	0.70	0.154	0.27	1.13
Μικρότερη διάμετρος κυτταροπλάσματος	118.13	71.904	29.45	418.49
Μεγαλύτερη διάμετρος κυτταροπλάσματος	171.59	90.183	44.64	571.90
Επιμήκυνση κυτταροπλάσματος	0.69	0.177	0.21	1.21
Σφαιρικότητα κυτταροπλάσματος	0.38	0.172	0.06	1.07
Περίμετρος πυρήνα	201.16	67.798	60.75	411.37
Περίμετρος κυτταροπλάσματος	490.03	274.857	136.12	1718.00
Θέση πυρήνα	0.27	0.165	0.00	0.90
Μέγιστο πυρήνα	107.89	57.507	19.00	430.00
Ελάχιστο πυρήνα	86.65	54.833	13.00	432.00
Μέγιστο κυτταροπλάσματος	416.86	633.817	14.00	3906.00
Ελάχιστο κυτταροπλάσματος	423.70	621.690	21.00	4000.00

Πίνακας 4.12: Περιγραφικοί δείκτες των 917 παρατηρήσεων της βάσης δεδομένων

Επιφανειακά λεπιδοειδή επιθήλια				
μεταβλητή	**Μέση τιμή**	**Τυπική απόκλιση**	**Ελάχιστη τιμή**	**Μέγιστη τιμή**
Εμβαδόν πυρήνα	630.87	207.529	316.87	1409.50
Εμβαδόν κυτταροπλάσματος	61487.28	23942.290	12631.25	127313.80
Αναλογία πυρήνα/κυτταροπλάσματος	0.01	0.006	0.00	0.03
Φωτεινότητα πυρήνα	66.30	16.650	17.91	97.64
Φωτεινότητα κυτταροπλάσματος	134.01	23.044	94.56	207.72
Μικρότερη διάμετρος πυρήνα	25.42	5.218	13.97	40.54
Μεγαλύτερη διάμετρος πυρήνα	31.41	5.684	19.41	47.16
Επιμήκυνση πυρήνα	0.82	0.154	0.32	1.19
Σφαιρικότητα πυρήνα	0.81	0.152	0.33	1.13
Μικρότερη διάμετρο κυτταροπλάσματος	270.60	62.250	123.83	418.49
Μεγαλύτερη διάμετρο κυτταροπλάσματος	337.55	68.945	162.79	509.51
Επιμήκυνση κυτταροπλάσματος	0.80	0.147	0.44	1.21
Σφαιρικότητα κυτταροπλάσματος	0.66	0.127	0.37	1.07
Περίμετρος πυρήνα	87.58	15.026	60.75	131.37
Περίμετρος κυτταροπλάσματος	1034.44	222.618	458.50	1541.75
Θέση πυρήνα	0.17	0.096	0.02	0.49
Μέγιστο πυρήνα	38.89	10.365	19.00	64.00
Ελάχιστο πυρήνα	27.47	8.131	13.00	54.00
Μέγιστο κυτταροπλάσματος	1881.10	813.209	398.00	3906.00
Ελάχιστο κυτταροπλάσματος	1882.74	820.540	444.00	4000.00

Πίνακας 4.13 : Περιγραφικοί δείκτες των 74 παρατηρήσεων της βάσης δεδομένων για την κατηγορία επιφανειακά λεπιδοειδή επιθήλια

Ενδιάμεσα λεπιδοειδή επιθήλια				
μεταβλητή	**Μέση τιμή**	**Τυπική απόκλιση**	**Ελάχιστη τιμή**	**Μέγιστη τιμή**
Εμβαδόν πυρήνα	1315.32	392.313	568.62	2943.25
Εμβαδόν κυτταροπλάσματος	44961.49	15455.750	12507.88	84196.00
Αναλογία πυρήνα/κυτταροπλάσματος	0.03	0.014	0.01	0.09
Φωτεινότητα πυρήνα	66.60	19.432	34.08	150.61
Φωτεινότητα κυτταροπλάσματος	131.13	22.296	85.50	217.31
Μικρότερη διάμετρος πυρήνα	36.52	7.122	20.16	56.69
Μεγαλύτερη διάμετρος πυρήνα	46.68	7.384	28.28	67.41
Επιμήκυνση πυρήνα	0.79	0.153	0.45	1.07
Σφαιρικότητα πυρήνα	0.77	0.154	0.43	1.06
Μικρότερη διάμετρο κυτταροπλάσματος	227.52	52.917	98.75	362.88
Μεγαλύτερη διάμετρο κυτταροπλάσματος	303.11	53.009	180.24	452.04
Επιμήκυνση κυτταροπλάσματος	0.75	0.144	0.42	1.12
Σφαιρικότητα κυτταροπλάσματος	0.61	0.122	0.29	0.87
Περίμετρος πυρήνα	130.38	19.436	83.25	196.62
Περίμετρος κυτταροπλάσματος	893.69	167.092	461.50	1408.62
Θέση πυρήνα	0.17	0.090	0.01	0.42
Μέγιστο πυρήνα	62.22	13.652	34.00	99.00
Ελάχιστο πυρήνα	44.50	10.656	25.00	80.00
Μέγιστο κυτταροπλάσματος	1338.84	560.369	340.00	3255.00
Ελάχιστο κυτταροπλάσματος	1257.51	540.550	307.00	3311.00

Πίνακας 4.14 : Περιγραφικοί δείκτες των 70 παρατηρήσεων της βάσης δεδομένων για την κατηγορία ενδιάμεσα λεπιδοειδή επιθήλια

Κυλινδρικά επιθήλια				
μεταβλητή	**Μέση τιμή**	**Τυπική απόκλιση**	**Ελάχιστη τιμή**	**Μέγιστη τιμή**
Εμβαδόν πυρήνα	1591.43	702.238	673.62	4418.75
Εμβαδόν κυτταροπλάσματος	3289.85	1837.972	563.62	9610.37
Αναλογία πυρήνα/κυτταροπλάσματος	0.34	0.103	0.15	0.66
Φωτεινότητα πυρήνα	94.08	24.974	61.42	156.47
Φωτεινότητα κυτταροπλάσματος	138.04	36.220	83.82	225.75
Μικρότερη διάμετρος πυρήνα	38.80	9.745	19.25	63.72
Μεγαλύτερη διάμετρος πυρήνα	54.77	13.558	31.57	105.23
Επιμήκυνση πυρήνα	0.73	0.188	0.32	1.17
Σφαιρικότητα πυρήνα	0.68	0.175	0.28	1.05
Μικρότερη διάμετρο κυτταροπλάσματος	62.87	18.406	29.45	131.90
Μεγαλύτερη διάμετρο κυτταροπλάσματος	117.27	36.178	58.18	211.02
Επιμήκυνση κυτταροπλάσματος	0.56	0.186	0.25	1.07
Σφαιρικότητα κυτταροπλάσματος	0.31	0.119	0.09	0.62
Περίμετρος πυρήνα	153.10	34.883	96.87	280.00
Περίμετρος κυτταροπλάσματος	322.89	103.356	166.37	614.37
Θέση πυρήνα	0.37	0.216	0.03	0.87
Μέγιστο πυρήνα	68.81	24.211	33.00	171.00
Ελάχιστο πυρήνα	52.36	21.483	24.00	131.00
Μέγιστο κυτταροπλάσματος	99.28	51.374	14.00	275.00
Ελάχιστο κυτταροπλάσματος	101.24	50.663	26.00	296.00

Πίνακας 4.15 : Περιγραφικοί δείκτες των 98 παρατηρήσεων της βάσης δεδομένων για την κατηγορία κυλινδρικά επιθήλια

Ελαφριά λεπιδοειδής μη κερατινώδης δυσπλασία				
μεταβλητή	**Μέση τιμή**	**Τυπική απόκλιση**	**Ελάχιστη τιμή**	**Μέγιστη τιμή**
Εμβαδόν πυρήνα	4690.10	1906.201	1121.37	9602.62
Εμβαδόν κυτταροπλάσματος	15458.71	10568.330	2612.87	62517.50
Αναλογία πυρήνα/κυτταροπλάσματος	0.26	0.102	0.09	0.62
Φωτεινότητα πυρήνα	97.55	16.925	60.17	158.80
Φωτεινότητα κυτταροπλάσματος	142.47	19.286	92.72	216.48
Μικρότερη διάμετρος πυρήνα	69.67	16.407	33.57	112.68
Μεγαλύτερη διάμετρος πυρήνα	88.89	18.833	37.94	158.82
Επιμήκυνση πυρήνα	0.79	0.138	0.37	1.17
Σφαιρικότητα πυρήνα	0.74	0.133	0.30	1.07
Μικρότερη διάμετρο κυτταροπλάσματος	138.83	40.317	60.28	264.97
Μεγαλύτερη διάμετρο κυτταροπλάσματος	211.51	73.113	82.97	571.90
Επιμήκυνση κυτταροπλάσματος	0.68	0.168	0.32	1.19
Σφαιρικότητα κυτταροπλάσματος	0.41	0.116	0.19	0.68
Περίμετρος πυρήνα	257.39	54.829	126.37	404.37
Περίμετρος κυτταροπλάσματος	588.78	203.142	239.12	1718.00
Θέση πυρήνα	0.25	0.130	0.01	0.61
Μέγιστο πυρήνα	151.57	60.909	58.00	415.00
Ελάχιστο πυρήνα	124.32	63.665	25.00	399.00
Μέγιστο κυτταροπλάσματος	442.05	340.578	79.00	2511.00
Ελάχιστο κυτταροπλάσματος	463.79	344.437	81.00	2472.00

Πίνακας 4.16 : Περιγραφικοί δείκτες των 182 παρατηρήσεων της βάσης δεδομένων για την κατηγορία ελαφριά λεπιδοειδής μη κερατινώδης δυσπλασία

Μεσαία λεπιδοειδής μη κερατινώδης δυσπλασία				
μεταβλητή	**Μέση τιμή**	**Τυπική απόκλιση**	**Ελάχιστη τιμή**	**Μέγιστη τιμή**
Εμβαδόν πυρήνα	3872.79	1657.032	1312.12	9806.00
Εμβαδόν κυτταροπλάσματος	7288.18	5225.137	1030.37	32476.88
Αναλογία πυρήνα/κυτταροπλάσματος	0.37	0.119	0.14	0.75
Φωτεινότητα πυρήνα	91.89	15.438	35.51	135.42
Φωτεινότητα κυτταροπλάσματος	134.62	18.373	87.23	209.40
Μικρότερη διάμετρος πυρήνα	62.08	14.977	31.07	112.72
Μεγαλύτερη διάμετρος πυρήνα	81.83	17.007	46.32	140.73
Επιμήκυνση πυρήνα	0.76	0.130	0.43	1.12
Σφαιρικότητα πυρήνα	0.71	0.128	0.41	1.07
Μικρότερη διάμετρο κυτταροπλάσματος	106.35	31.910	54.45	226.57
Μεγαλύτερη διάμετρο κυτταροπλάσματος	153.88	46.357	71.56	326.74
Επιμήκυνση κυτταροπλάσματος	0.71	0.174	0.33	1.17
Σφαιρικότητα κυτταροπλάσματος	0.37	0.123	0.12	0.80
Περίμετρος πυρήνα	231.28	49.078	131.37	402.37
Περίμετρος κυτταροπλάσματος	442.75	141.348	219.37	935.00
Θέση πυρήνα	0.30	0.168	0.00	0.82
Μέγιστο πυρήνα	132.52	58.952	59.00	430.00
Ελάχιστο πυρήνα	109.39	56.910	40.00	432.00
Μέγιστο κυτταροπλάσματος	196.96	139.295	43.00	783.00
Ελάχιστο κυτταροπλάσματος	213.43	141.874	51.00	878.00

Πίνακας 4.17 : Περιγραφικοί δείκτες των 146 παρατηρήσεων της βάσης δεδομένων για την κατηγορία μεσαία λεπιδοειδής μη κερατινώδης δυσπλασία

Ενδιάμεσο λεπιδοειδές κύτταρο με μη διηθητικό καρκίνο				
Βαριά λεπιδοειδής μη κερατινώδης δυσπλασία				
μεταβλητή	**Μέση τιμή**	**Τυπική απόκλιση**	**Ελάχιστη τιμή**	**Μέγιστη τιμή**
Εμβαδόν πυρήνα	2948.98	1478.143	757.50	10794.63
Εμβαδόν κυτταροπλάσματος	3415.31	2282.241	594.25	14108.75
Αναλογία πυρήνα/κυτταροπλάσματος	0.48	0.142	0.17	0.84
Φωτεινότητα πυρήνα	93.66	21.734	53.35	174.99
Φωτεινότητα κυτταροπλάσματος	143.46	29.090	69.88	230.48
Μικρότερη διάμετρος πυρήνα	52.23	14.436	23.76	107.78
Μεγαλύτερη διάμετρος πυρήνα	74.44	18.234	31.95	130.24
Επιμήκυνση πυρήνα	0.71	0.151	0.30	1.05
Σφαιρικότητα πυρήνα	0.66	0.143	0.27	0.96
Μικρότερη διάμετρος κυτταροπλάσματος	75.51	20.233	33.47	155.08
Μεγαλύτερη διάμετρος κυτταροπλάσματος	118.74	36.554	44.64	263.30
Επιμήκυνση κυτταροπλάσματος	0.66	0.171	0.29	1.10
Σφαιρικότητα κυτταροπλάσματος	0.30	0.120	0.07	0.74
Περίμετρος πυρήνα	207.88	52.131	99.25	384.25
Περίμετρος κυτταροπλάσματος	323.04	94.926	136.12	666.62
Θέση πυρήνα	0.29	0.169	0.00	0.90
Μέγιστο πυρήνα	110.58	48.765	37.00	425.00
Ελάχιστο πυρήνα	88.91	47.899	23.00	381.00
Μέγιστο κυτταροπλάσματος	100.40	63.131	20.00	357.00
Ελάχιστο κυτταροπλάσματος	116.20	65.173	21.00	357.00

Πίνακας 4.18 : Περιγραφικοί δείκτες των 197 παρατηρήσεων της βάσης δεδομένων για την κατηγορία βαριά λεπιδοειδής μη κερατινώδης δυσπλασία

μεταβλητή	Μέση τιμή	Τυπική απόκλιση	Ελάχιστη τιμή	Μέγιστη τιμή
Εμβαδόν πυρήνα	2985.95	1283.793	1160.50	10271.50
Εμβαδόν κυτταροπλάσματος	2115.22	1495.384	467.87	10863.38
Αναλογία πυρήνα/κυτταροπλάσματος	0.60	0.132	0.23	0.88
Φωτεινότητα πυρήνα	97.10	17.626	61.98	141.56
Φωτεινότητα κυτταροπλάσματος	142.09	22.285	77.57	210.21
Μικρότερη διάμετρος πυρήνα	52.40	13.662	28.51	101.54
Μεγαλύτερη διάμετρος πυρήνα	76.86	14.994	42.01	145.27
Επιμήκυνση πυρήνα	0.68	0.152	0.34	1.12
Σφαιρικότητα πυρήνα	0.63	0.149	0.29	0.94
Μικρότερη διάμετρος κυτταροπλάσματος	70.27	17.924	33.20	129.89
Μεγαλύτερη διάμετρος κυτταροπλάσματος	102.06	24.761	57.00	209.34
Επιμήκυνση κυτταροπλάσματος	0.70	0.165	0.21	1.16
Σφαιρικότητα κυτταροπλάσματος	0.24	0.101	0.06	0.63
Περίμετρος πυρήνα	215.24	48.176	121.12	411.37
Περίμετρος κυτταροπλάσματος	287.78	67.595	160.25	517.50
Θέση πυρήνα	0.27	0.154	0.00	0.65
Μέγιστο πυρήνα	108.26	38.846	41.00	282.00
Ελάχιστο πυρήνα	87.09	37.560	25.00	264.00
Μέγιστο κυτταροπλάσματος	70.83	43.54469	22.00	311.00
Ελάχιστο κυτταροπλάσματος	85.34	45.78632	23.00	330.00

Πίνακας 4.19 : Περιγραφικοί δείκτες των 150 παρατηρήσεων της βάσης δεδομένων για την κατηγορία ενδιάμεσο λεπιδοειδές κύτταρο με μη διηθητικό καρκίνο

ΠΕΡΙΛΗΨΗ

Ο Γεώργιος Παπανικολάου ανακάλυψε το 1928 μια μέθοδο χρωματισμού των κυττάρων μέσω της οποίας κατέστη δυνατή η διάγνωση προκαρκινικών σταδίων του καρκίνου του τραχήλου της μήτρας. Με αυτή τη μέθοδο, ο καρκίνος προλαμβάνεται και θεραπεύεται, αν το άτομο υποβάλλεται στο λεγόμενο τεστ παπ σε τακτά χρονικά διαστήματα. Μετά την ανακάλυψη του τεστ παπ, τα ποσοστά θνησιμότητας από τη συγκεκριμένη μορφή καρκίνου μειώθηκαν σημαντικά σε όλο τον κόσμο. Η διάγνωση των κυττάρων, που αρχικά γινόταν από κυτταροτεχνικούς, είναι μια εργασία αρκετά λεπτομερής και χρονοβόρα και παρουσιάζει σφάλμα κυρίως λόγω ψευδώς αρνητικών ταξινομήσεων. Τα τελευταία δέκα χρόνια έχουν γίνει αρκετές προσπάθειες αυτόματης ταξινόμησης των κυττάρων μέσω υπολογιστών. Για τον σκοπό αυτό, δημιουργήθηκαν δύο βάσεις δεδομένων επιχρίσματος Παπανικολάου και αντλήθηκαν 20 χαρακτηριστικά για τα κύτταρα της κάθε βάσης.

Στην πληροφορική, έχουν ήδη εφαρμοστεί πολλοί αλγόριθμοι ταξινόμησης και επιλογής χαρακτηριστικών. Κάποιοι από αυτούς επέφεραν πολύ ικανοποιητικά αποτελέσματα, όπως ο αλγόριθμος ταξινόμησης του κοντινότερου κέντρου βαρύτητας και οι γενετικοί αλγόριθμοι επιλογής χαρακτηριστικών GA, TS, ACO που κατάφεραν να μειώσουν το σφάλμα κάτω από το 5%.

Σε αυτή τη διπλωματική, περιγράφονται αρχικά οι μέθοδοι αυτοί και τα αποτελέσματα τους. Ύστερα αναλύονται στατιστικές μέθοδοι ταξινόμησης και δημιουργίας μοντέλων. Τα μοντέλα λογαριθμιστικής παλινδρόμησης προσφέρουν σημαντική γνώση για την επίδραση των χαρακτηριστικών στην πιθανότητα ένα κύτταρο να ανήκει στα υγιή ή στα μη υγιή καθώς και στην πιθανότητα να ανήκει σε κάθε ένα από τα προκαρκινικά στάδια. Εφαρμόζεται επίσης διαχωριστική ανάλυση και δημιουργία συστάδων. Η διαχωριστική ανάλυση, με σφάλμα μικρότερο του 5%, διαχωρίζει ικανοποιητικά τα κύτταρα σε υγιή και μη υγιή. Η ταξινόμηση σε συστάδες δεν επέφερε σημαντικά αποτελέσματα.

Τέλος, γίνεται σύγκριση των αποτελεσμάτων της στατιστικής ανάλυσης με τα αποτελέσματα της πληροφορικής.

ABSTRACT

George Papanicolaou discovered in 1928 a method of colouring cells through which, pre – malignant cell diagnosis in the uterine cervix became possible. With this method, cancer can be diagnosed in early stages and it can be treated, if a person submit her self in a, so called, pap test in regular time periods. After test pap's discovery, the mortality rate had an important decrease all over the world. Cell diagnosis, which was initially handled by cyto – technicians, is a detailed and time consuming task and thus, it produces an important error, especially a false – negative one. Over the last ten years, many attempts of automatic classification have taken place. For that purpose, two pap – smear data bases have been created and 20 features have been extracted for each cell.

In the information technology field, numerous algorithms have been applied on both data bases, for cell classification and feature selection. Same of them, like nearest center of gravity and genetic algorithms GA, TS, ACO for feature selection, derived very satisfying results as they managed to reduce error below 5%.

In this master's thesis, those methods are described, along with their results. Then, statistical methods for classification and model building are analysed and results are presented. Logistic regression models offer important knowledge on the effect of features on probabilities of a cell belonging to healthy or unhealthy category or to one of the pre – cancerous stages. Discriminant and cluster analysis are applied. Discriminant analysis, with error rates below 5%, can classify cells in healthy or unhealthy cell categories. Cluster analysis does not give satisfy results.

Finally, results from both statistics and information technology are compared.

ΑΝΑΦΟΡΕΣ

1) Bezdek, J.C., (1981) Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York.

2) Byriel G., (1999) Neuro - fuzzy classification of cells in cervical smears. Master's thesis, Technical University of Denmark: Oersted – DTU, Automation

3)Dorigo M., Maniezzo V., Colorni A. (1996) The Ant System: Optimization by a Colony of Cooperating Agents

4)Duda, R.O. Hart, P.E.. & Stork P.G. (2000), Pattern Classification, 2 endition, A Wiley-Interscience Publication

5) Gustafson E. and Kessel W. (1979) Fuzzy clustering with a fuzzy covariance matrix. In Proc. of IEEE CDC.

6) Hosmer D.W. and Lemeshow S. (2000). Applied Logistic Regression, 2nd ed. New York, Chichester, Wiley

7) Indman, D. P. (2005). M.d. http://www.gynalternatives.com/abnpap.htm.

8) Jang J.S.R., Sun C.T. Mizutani E. (1997) Neuro-Fuzzy and Soft Computing, A computational Approach to Learning and Machine Intelligence. Upper Saddle River, N.J. Prentice Hall

9) Jantzen, J. (1998). Design of fuzzy controllers. Technical report, Technical University of Denmark, Dept. of Automation, 98-E-864.

10) Johnsona A.W., Jacobson S.H. (2002) On the convergence of generalized hill climbing algorithms.Discrete Applied Mathematics 119 pg.37–57

11) Kasabov, N. and Song, Q. (2002). Denfis : Dynamic, evolving neaural fuzzy inference system and its application for time-series prediction. *IEEE Transaction on Fuzzy System*, 10(2):144–154.

12) Kasabov, N. and Song, Q. (2005). Nfi: A neuro-fuzzy inference method for transductive reasoning. *IEEE Transactions on Fuzzy System*,

13) MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297

14) Marinakis Y. and Dounias G. (2006) Nature inspired intelligent techniques for pap smear diagnosis: ant colony optimization for cell classification

15) Marinakis Y. and Dounias G. (2006) Nearest neighbor based pap smear cell classification using tabu search for feature selection

16) Marinakis Y. and Dounias G. (2006) Pap smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbor classification

17) Martin E. (2003) Pap smear classification, Master's thesis, Technical University of Denmark: Oersted – DTU, Automation

18) McLachlan G.J. (2004) Discriminant Analysis and Statistical Pattern Recognition Wiley-Interscience, New Ed edition

19) Norup J., (2005) Classification of pap smear data by transductive neuro – fuzzy methods, Master's thesis, Technical University of Denmark: Oersted – DTU, Automation

20) Shakhnarovish, Darrell, and Indy (2005) Nearest-Neighbor Methods in Learning and Vision. The MIT Press

21)Zadeh L.A. (1965) Fuzzy Sets, Information and Control, Vol. 8

22)Γεωργαλλής Π. (2006) Υλοποίηση, μελέτη και ανάλυση επίδοσης αλγορίθμων κατανομής φάσματος για dvb-T allotments

ΔΙΑΔΙΚΤΥΑΚΕΣ ΠΗΓΕΣ

1) en.wikipedia.org
2) fuzzy.iau.dtu.dk/smear
3) fuzzy.iau.dtu.dk/download/smear 2005
4) www.dimac-imaging.com