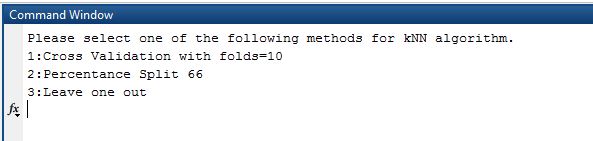
Πανεπιστήμιο Δ. Μακεδονίας Εξόρυξη Δεδομένων Τμήμα Μηχ. Πληροφορικής & ΤηλεπικοινωνιώνΝτάγιου Άννα 432

**Υλοποίηση αλγόριθμου k-NN και αξιολόγηση απόδοσης του σε γλώσσα προγραμματισμού Matlab**

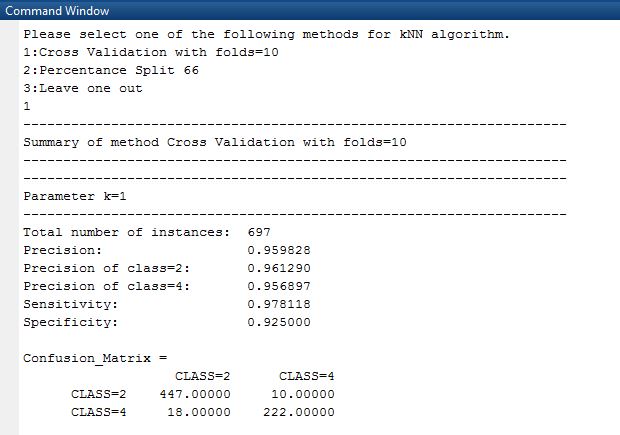
Η κατηγοριοποίηση που βασίζεται στην εύρεση και στην εξέταση κοντινότερων γειτόνων (κατηγοριοποίηση KNN) είναι μια από τις πιο γνωστές μεθόδους κατηγοριοποίησης. Η παράμετρος Κ της τεχνικής, ορίζει τον αριθμό των κοντινότερων γειτόνων που έχουν λόγο στην κατηγοριοποίηση. Η υπό εξέταση πλειάδα τοποθετείται στην κατηγορία στην οποία ανήκουν οι περισσότεροι από τους Κ κοντινότερους γείτονες. Το πόσοι κοντινότεροι γείτονες θα ληφθούν υπόψη για να γίνει η κατηγοριοποίηση είναι ένα σημαντικό ζήτημα αφού η ακρίβεια της κατηγοριοποίησης εξαρτάται από αυτόν τον αριθμό. Σε αυτήν την εργασία η παράμετρος Κ έχει εύρος από 1 έως 15 και λαμβάνει υπ’ όψιν μόνο τους περιττούς αριθμούς γιατί οι κατηγορίες στο σύνολο δεδομένων είναι δύο.

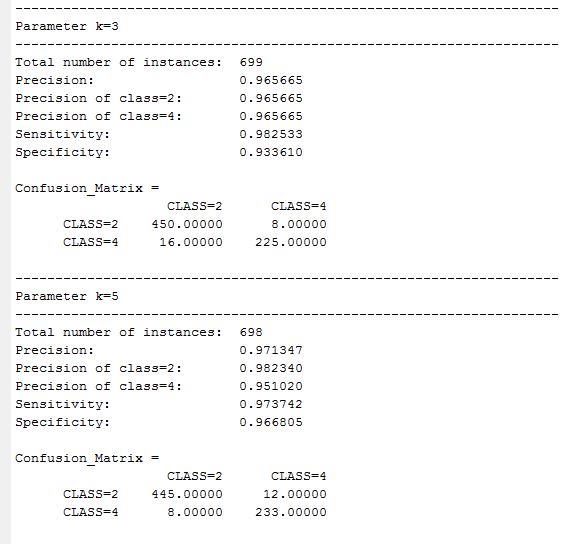
Το σύνολο δεδομένων που μας δόθηκε αποτελείται από 699 περιπτώσεις γυναικών στις οποίες παρουσιάστηκε καρκίνος του μαστού. Τα χαρακτηριστικά της κάθε περίπτωσης αναφέρονται σε βαθμονομημένα χαρακτηριστικά των κυττάρων των γυναικών μετά από ιστολογικές – κυτταρολογικές εξετάσεις. Τα χαρακτηριστικά βρίσκονται από την 2η έως την 10η στήλη του συνόλου δεδομένων και οι τιμές τους κυμαίνονται από το 1 έως το 10. Ο καρκίνος ο οποίος εντοπίστηκε στις γυναίκες κατηγοριοποιήθηκε σε 2 κλάσεις Καλοήθης (Benign) και Κακοήθης (Malignant). Εκ των 699 περιπτώσεων, 16 δεν έχουν τιμές σε ορισμένα χαρακτηριστικά. Αυτά τα χαρακτηριστικά επιλέχτηκε να αντικατασταθούν με τον μέσο όρο της κατηγορίας των περιπτώσεων που ανήκουν στην ίδια κλάση.

Λόγω του όγκου των δεδομένων που πρέπει να παρουσιαστούν, ο χρήστης επιλέγει με ποια από τις τρεις μεθόδους θέλει να κατηγοριοποιήσει τα δεδομένα του.

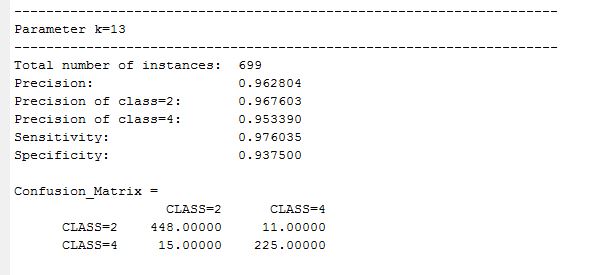


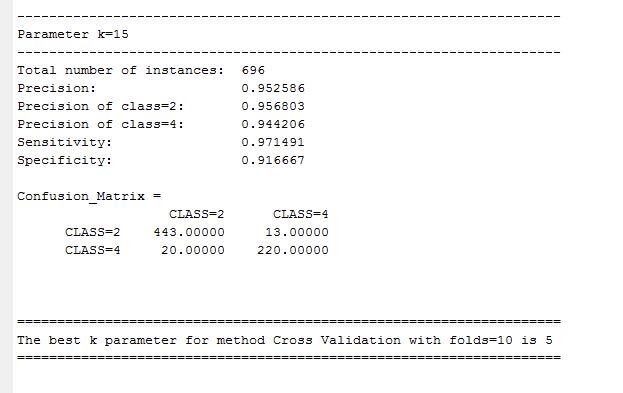
Μόλις ο χρήστης επιλέξει μέθοδο εμφανίζονται σε αυτόν τα αποτελέσματα για κάθε τιμή της παραμέτρου Κ, όπως Total number of instances, Precision, Precision of every class, Sensitivity, Specificity και Confusion\_Matrix.



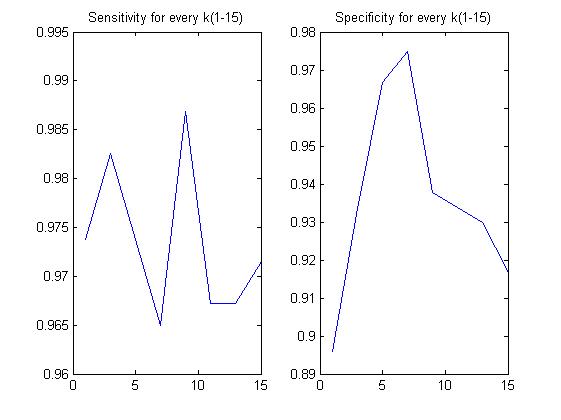


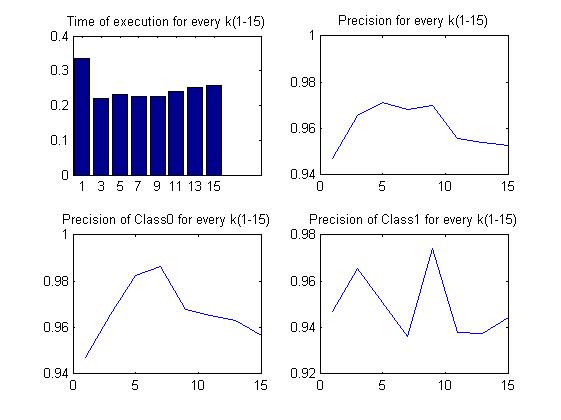
Μετά και την τελευταία κατηγοριοποίηση με Κ=15, υπολογίζεται και εμφανίζεται η βέλτιστη τιμή της παραμέτρου Κ, δηλαδή η τιμή με την οποία ο αλγόριθμος είχε την βέλτιστη ακρίβεια.





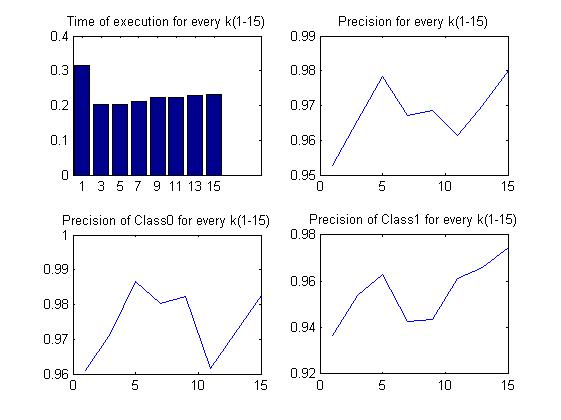
Τελικά εμφανίζονται στον χρήστη γραφικές παραστάσεις που απεικονίζουν την απόδοση του αλγόριθμου k-NN.

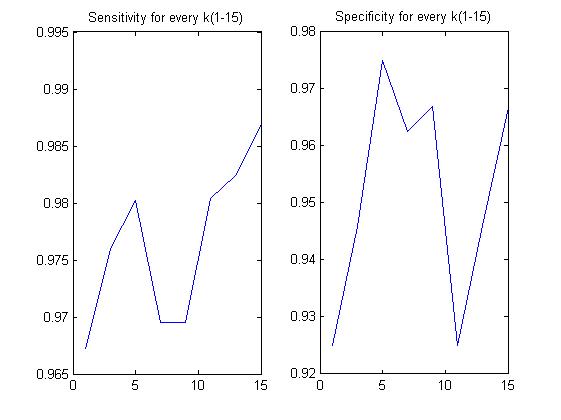




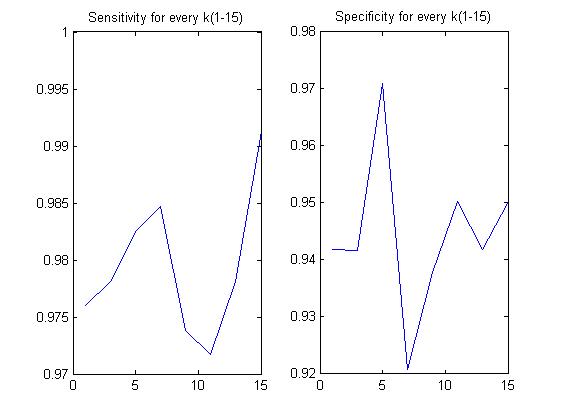
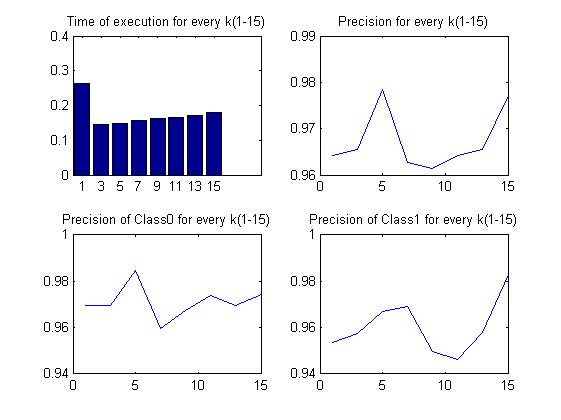
**Σύγκριση των αποτελεσμάτων του αλγόριθμου k-NN**

Μέθοδος Cross Validation with folds=10

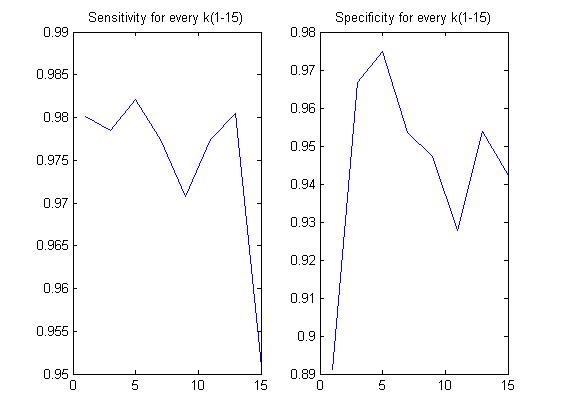
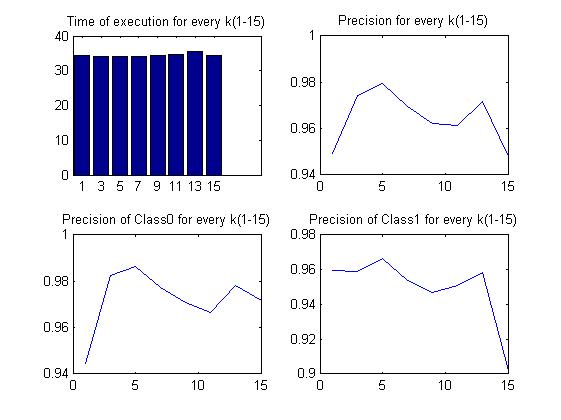




Μέθοδος Percentance Split 66 ( 33% Test Data και 66% Train Data)



Μέθοδος Leave one out



**Συμπεράσματα**

Η επίδοση των αλγορίθμων εξετάζεται με την εκτίμηση της ακρίβειας της κατηγοριοποίησης, δηλαδή την ικανότητα του μοντέλου να προβλέπει την κατηγορία μιας νέας περίπτωσης. Η εκτίμηση της ακρίβειας είναι ένα πολύ σημαντικό ζήτημα στο χώρο της κατηγοριοποίησης αφού κάτι τέτοιο μας δείχνει το πόσο καλά ανταποκρίνεται ο αλγόριθμος μας για δεδομένα με τα οποία δεν έχει εκπαιδευτεί.

Στην υλοποίηση αυτήν βλέπουμε ότι η μέθοδος Percentance Split 66 έχει μεγαλύτερη ακρίβεια (Precision) (εύρος 0.961 - 0.982) από ότι οι άλλες δύο μέθοδοι (Cross Validation with folds=10 εύρος 0.953 - 0.980 και Leave one out εύρος 0.948 - 0.980). Ειδικότερα παρατηρούμε καλύτερη ακρίβεια στην κλάση 0 (δηλαδή class=2 ) και αυτό γίνεται λόγω τις πλειοψηφίας των περιπτώσεων αυτής της κλάσης.

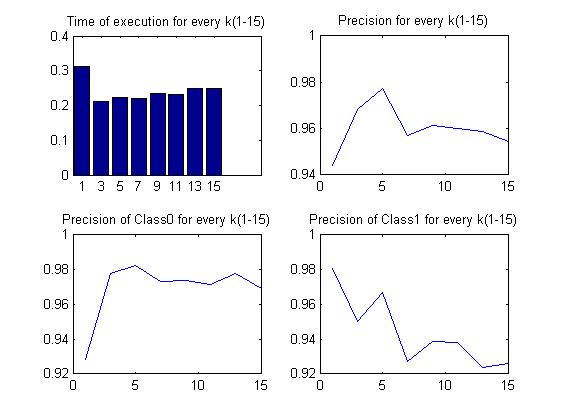
Στην συνέχεια παρατηρούμε ότι η μέθοδος Percentance Split 66 έχει πολύ μικρότερο χρόνο εκτέλεσης (εύρος 0.144 - 0.262 sec) από ότι οι άλλες δύο μέθοδοι (Cross Validation with folds=10 εύρος 0.203 - 0.315 sec και Leave one out εύρος 34.096 - 35.616 sec) και αυτό συμβαίνει διότι στην μέθοδο Percentance Split 66 το σύνολο δεδομένων χωρίζεται σε μεγαλύτερα μέρη απ’ ότι στις άλλες μεθόδους.

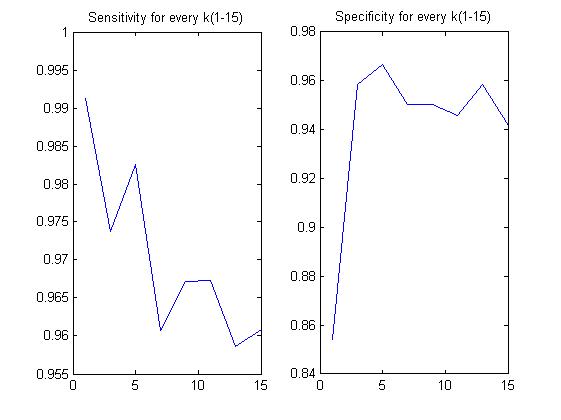
Έπειτα η ευαισθησία (sensitivity) δηλαδή το ποσοστό των θετικών παραδειγμάτων που κατηγοριοποιούνται σωστά είναι μεγαλύτερη στην μέθοδο Percentance Split 66 (εύρος 0.972 - 0.991) και μικρότερη στην μέθοδο Leave one out (εύρος 0.951 - 0.982). Η ιδιαιτερότητα (specificity) δηλαδή το ποσοστό των αρνητικών παραδειγμάτων που κατηγοριοποιούνται σωστά είναι καλύτερη στην μέθοδο Cross Validation with folds=10 (εύρος 0.925 - 0.975) και χειρότερη στην μέθοδο Leave one out (εύρος 0.891 - 0.975).

Γενικότερα τα ποσοστά που μας δείχνουν εάν ο αλγόριθμος κατηγοριοποίησε σωστά τις περιπτώσεις που δεν έχει εκπαιδευτεί είναι υψηλά αλλά εάν πρέπει να επιλέξουμε μία μέθοδο από τις τρεις θα επιλέγαμε την Percentance Split λόγω της ακρίβειας και του χρόνου εκτέλεσης.

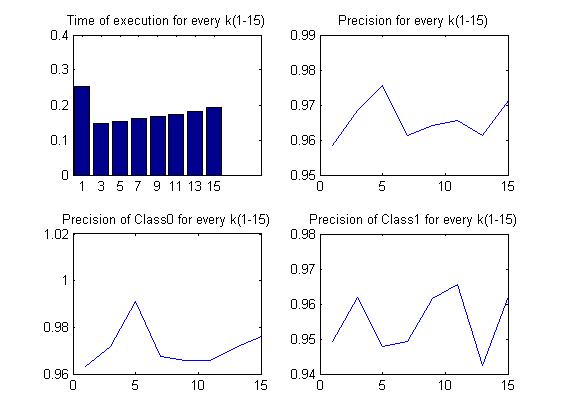
**Σύγκριση των αποτελεσμάτων του αλγόριθμου Weighted k-NN**

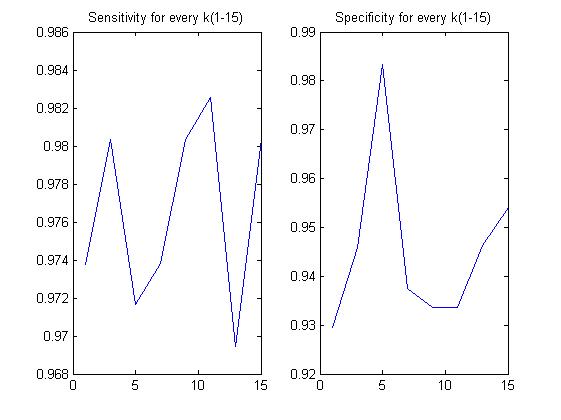
Μέθοδος Cross Validation with folds=10



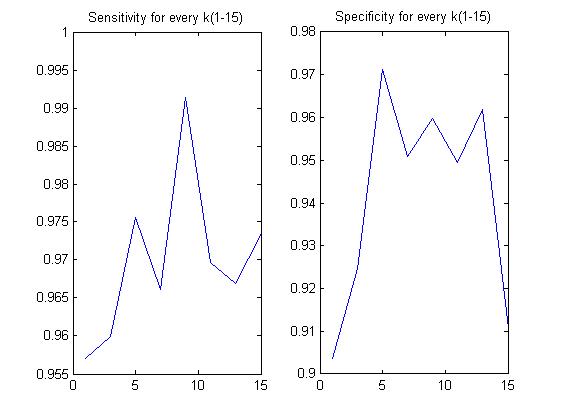
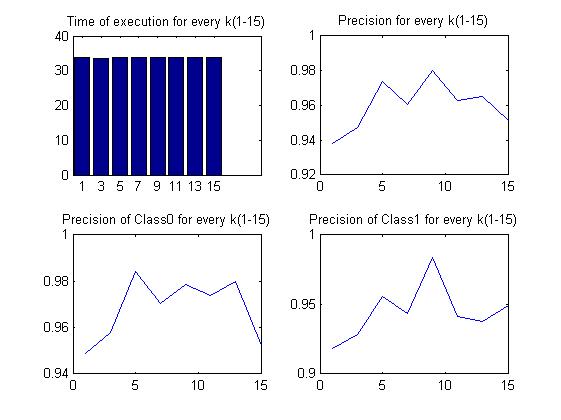


Μέθοδος Percentance Split 66 ( 33% Test Data και 66% Train Data)





Μέθοδος Leave one out



**Σύγκριση των αλγορίθμων k-NN και Weighted k-NN**

Τελικά μετά το πέρας όλων των εκτελέσεων όλων των μεθόδων για κάθε αλγόριθμο παρατηρούμε ότι ο αλγόριθμος Weighted k-NN είναι ο βέλτιστος διότι η ακρίβεια, η ευαισθησία και η ιδιαιτερότητα ξεπερνούν σε πολλές περιπτώσεις το ποσοστό των 0.99.

Ειδικότερα στην υλοποίηση του αλγόριθμου Weighted k-NN βλέπουμε ότι η μέθοδος και Leave one out έχει μεγαλύτερη ακρίβεια (Precision) (εύρος 0.94 - 0.982) από ότι οι άλλες δύο μέθοδοι (Cross Validation with folds=10 εύρος 0.938 - 0.979 και Percentance Split 66 εύρος 0.955 - 0.978 ). Ειδικότερα παρατηρούμε καλύτερη ακρίβεια στην κλάση 0 (δηλαδή class=2 ) και αυτό γίνεται λόγω τις πλειοψηφίας των περιπτώσεων αυτής της κλάσης.

Στην συνέχεια παρατηρούμε ότι οι χρόνοι εκτέλεσης κυμαίνονται στα ίδια επίπεδα με τον αλγόριθμο k-NN και αυτό συμβαίνει διότι ο χρόνος υπολογισμού των γειτόνων με κριτήριο κάποιο βάρος είναι αμελητέος.

Έπειτα η ευαισθησία (sensitivity) δηλαδή το ποσοστό των θετικών παραδειγμάτων που κατηγοριοποιούνται σωστά είναι μεγαλύτερη στην μέθοδο Leave one out (εύρος 0.955 - 0.995) και μικρότερη στην μέθοδο Percentance Split 66 (εύρος 0.969 - 0.982). Η ιδιαιτερότητα (specificity) δηλαδή το ποσοστό των αρνητικών παραδειγμάτων που κατηγοριοποιούνται σωστά είναι καλύτερη στην μέθοδο Percentance Split 66 (εύρος 0.930 - 0.984) και χειρότερη στην μέθοδο Cross Validation with folds=10 (εύρος 0.855 - 0.969).

Γενικότερα τα ποσοστά που μας δείχνουν εάν ο αλγόριθμος κατηγοριοποίησε σωστά τις περιπτώσεις που δεν έχει εκπαιδευτεί είναι υψηλότερα από τον προηγούμενο αλγόριθμο αλλά εάν πρέπει να επιλέξουμε μία μέθοδο από τις τρεις θα επιλέγαμε την Leave one out λόγω της ακρίβειας αλλά όχι του χρόνου εκτέλεσης. Εάν θέλαμε έναν συνδυασμό χρόνου και ακρίβειας η μέθοδος που θα επιλεγόταν θα ήταν η Cross Validation with folds=10.