**Εργαστηριακή Άσκηση**

**για το μάθημα Θεωρία Αποφάσεων**

**2022-2023**

**Ονοματεπώνυμο: ……………………………………………………. ΑΜ:………….**

**ΑΝΤΙΚΕΙΜΕΝΟ:**

"**Σχεδιασμός και Αξιολόγηση Συστήματος, από Δημογραφικά και Βιοϊατρικά Δεδομένα, για Πρόβλεψη Ασθένειας** "

Σε αυτή την εργαστηριακή άσκηση καλείστε να χρησιμοποιήσετε ένα σύνολο δεδομένων για να εκπαιδεύσετε και αξιολογήσετε ένα σύνολο ταξινομητών για την πρόβλεψη ασθένειας στο συκώτι.

Το σύνολο δεδομένων που σας δίνεται στο Indian Liver Patient Dataset (ILPD).csv παρέχει ένα σύνολο βιοϊατρικών μετρήσεων για κάθε ασθενή καθώς και την πληροφορία για το αν αυτός πάσχει ή όχι από ασθένεια στο συκώτι. Κάθε γραμμή στο αρχείο αυτό περιέχει πληροφορία για διαφορετικό ασθενή. Η πρώτη στήλη αφορά τον την ηλικία και η δεύτερη το φύλλο, ενώ η τελευταία δείχνει το αν η καταγραφή αφορά μέτρηση ασθενούς ατόμου στο συκώτι (τιμή 2) ή υγιούς ατόμου (τιμή 1). Όλες οι άλλες στήλες αντιστοιχούν σε βιοϊατρικές μετρήσεις που θα πρέπει επίσης να χρησιμοποιήσετε σαν εισόδους στους ταξινομητές που θα δημιουργήσετε. Για περισσότερες πληροφορίες για τα δεδομένα αυτά μπορείτε να δείτε την αναλυτική περιγραφή τους στην βάση δεδομένων μηχανικής μάθησης UCI ([http://archive.ics.uci.edu/ml/datasets/ILPD+%28Indian+Liver+Patient+Dataset%29#](http://archive.ics.uci.edu/ml/datasets/ILPD+%28Indian+Liver+Patient+Dataset%29) ) από την οποία προέρχονται.

**Ερώτημα 1.** Προεπεξεργασία δεδομένων

Να αναφέρετε, πόσα είναι τα χαρακτηριστικά κάθε δείγματος και πόσα δείγματα εκπαίδευσης περιέχει το αρχείο.

Η δεύτερη στήλη περιέχει το φύλο του ανθρώπου που συμμετείχε στο δείγμα. Στο αρχείο όμως είναι σημειωμένη με Male για αρσενικό και Female για θηλυκό. Προκειμένου να την χρησιμοποιήσουμε σαν είσοδο θα πρέπει να αντιστοιχίσετε το Male με την τιμή 0 και το Female με την τιμή 1.

Το εύρος τιμών των δεδομένων που σας έχουν δοθεί διαφέρει σημαντικά ανά χαρακτηριστικό. Για αυτό τον λόγο, για να μην υπερεκτιμηθεί η συνεισφορά κάποιου χαρακτηριστικού έναντι άλλων, θα πρέπει πριν την επεξεργασία των χαρακτηριστικών εισόδου να κανονικοποιηθούν στο εύρος [-1,1]. Χρησιμοποιήστε το matlab (ή όποια άλλη εφαρμογή θέλετε) τόσο για το διάβασμα του αρχείου που σας δίνεται όσο και για την κανονικοποίηση των δεδομένων εισόδου στο εύρος τιμών [-1,1].

**Ερώτημα 2.** Στο μάθημα συζητήθηκε εκτεταμένα ο ταξινομητής Bayes. Στη βιβλιογραφία, υπάρχει μια παραλλαγή του που λέγεται Αφελής Ταξινομητής Bayes (Naïve Bayes), με την υπόθεση ότι τα χαρακτηριστικά είναι στατιστικά ανεξάρτητα. Αναζητήστε τη σχετική βιβλιογραφία στο Internet, και να κάνετε μια σύντομη παρουσίαση του αλγορίθμου. Στη συνέχεια να κάνετε μια σύγκριση με τον Ταξινομητή Bayes.

**Ερώτημα 3.** Με χρήση της μεθόδου 5-fold cross validation, να εκπαιδεύσετε τον Naïve Bayes ταξινομητή, να παρουσιάστε και να σχολιάσετε την απόδοσή του. Μπορείτε να χρησιμοποιήσετε κατάλληλες συναρτήσεις του matlab ή οποιαδήποτε εφαρμογή επιθυμείτε (ή να υλοποιήσετε δικό σας κώδικα). Για την αξιολόγηση της απόδοσης του ταξινομητή να χρησιμοποιήσετε τις μετρικές του ερωτήματος 4, παρακάτω.

**Ερώτημα 4.** Με χρήση της μεθόδου 5-fold cross validation και του matlab, εκπαιδεύσετε τους παρακάτω ταξινομητές, να παρουσιάστε και να σχολιάσετε την απόδοσή τους:

* Support Vector Machines (με Radial Basis Function kernel function):
  + Ρυθμίστε την παράμετρο C με διαδοχική αναζήτηση του βέλτιστου C στο διάστημα 1-200 με βήμα 5 και χρήση γραμμικών SVM. Στη συνέχεια, ρυθμίστε την παράμετρο γ με χρήση του βέλτιστου C που βρέθηκε από πριν, και διαδοχική αναζήτηση του βέλτιστου γ στο διάστημα 0-10 με βήμα 0.5 και χρήση RBFSVM.
* Ταξινομητής K-Κοντινότερου Γείτονα
  + Ρυθμίστε την παράμετρο Κ με διαδοχική αναζήτηση της βέλτιστης τιμής στο διάστημα 3-15.

Ποιο συγκεκριμένα ζητείται να παρουσιάσετε για κάθε ταξινομητή την μέση απόδοση του με χρήση 5 fold cross validation σε σχέση με την μετρική του γεωμετρικού μέσου (Geometric Mean) της ευαισθησίας (Sensitivity) και της ειδίκευσής (Specificity) του:

Geometric Mean = sqrt (Sensitivity \* Specificity)

Η μετρική αυτή χρησιμοποιείται για προβλήματα ταξινόμησης όπου παραδείγματα εκπαίδευσης της μίας κλάσης είναι περισσότερα από τα παραδείγματα εκπαίδευσης της άλλης κλάσης.

Στη συνέχεια, παρουσιάστε τα ενδιάμεσα αποτελέσματα που πήρατε από τα πειράματα για την ρύθμιση των παραμέτρων των αλγορίθμων. Γιατί η κάθε μέθοδος ταξινόμησης δίνει διαφορετικά αποτελέσματα; Με βάση τα παραπάνω αποτελέσματα ποιά από τις μεθόδους προτείνετε εσείς να χρησιμοποιηθεί για το παραπάνω πρόβλημα και γιατί;

**Ερώτημα 5.** Για τα δεδομένα που σας δίνονται διατάξτε τα χαρακτηριστικά εισόδου με χρήση της μεθόδου student t-test με βάση την σημαντικότητά τους στην πρόβλεψη της ασθένειας στο συκώτι. Στη συνέχεια, κρατείστε τα 5 πιο σημαντικά χαρακτηριστικά, και επαναλάβετε την εκπαίδευση του βέλτιστου ταξινομητή που βρήκατε από προηγούμενο ερώτημα. Σχολιάστε την τελική απόδοση.

**Παρατηρήσεις:**

* Η αξιολόγηση της εργασίας θα έχει βαρύτητα 40% του τελικού βαθμού.
* Η εργασία σας θα παραδοθεί τμηματικά. Το πρώτο μέρος (Ερωτήματα 1-3) πρέπει να παραδοθεί (με ανάρτηση στο e-class) μέχρι τις 21/12/2018 στις 23.55. Το 2ο μέρος της εργασίας (τα υπόλοιπα ερωτήματα), πρέπει να αναρτηθούν την παραμονή της εξέτασης του μαθήματος στις 23.55.
* Για κάθε μέρος της εργασία, πρέπει να παραδώσετε μία αναφορά, η οποία θα περιέχει τις απαντήσεις σας, σύμφωνα με τις εκφωνήσεις. Καλό είναι να χρησιμοποιήσετε αυτό το κείμενο, και μετά από κάθε ερώτηση να έχετε την απάντηση. Το όνομα του αρχείου θα έχει τη μορφή: ΘΑ\_ΜέροςΑ (ή Β΄)\_ΑΜxxxxxxx.
* Για απορίες σχετικές με την εργασία σας μπορείτε να επικοινωνείτε με email με τον κο Λυκοθανάση (Α’ Μέρος) και τον κο Κοσμόπουλο (Β΄ Μέρος).
* Σύντομα θα ανακοινωθούν και ώρες γραφείου.