Βασίλειος Δημητρίου - Capstone

**Εγώ ο Βασίλειος Δημητρίου βεβαιώνω ότι η παρούσα εργασία βασίζεται σε δεδομένα για τα οποία έχω άδεια μεταβίβασης και επεξεργασίας. Βεβαιώνω επίσης ότι τα δεδομένα αυτά τηρούν όλους τους ισχύοντες νόμους και διατάξεις περί προστασίας πνευματικής ιδιοκτησίας και προσωπικών δεδομένων.**

Η ατομική εργασία είναι αναρτημένη στον ακόλουθο σύνδεσμο, όπου περιλαμβάνονται και τα δεδομένα που χρησιμοποίησα αλλά και η ανάλυση και τα αποτελέσματα.

[Vasilisdi/Data-Analyst-NTUA-Course---Vasileios-Dimitriou: Assignment in the context of the NTUA course (github.com)](https://github.com/Vasilisdi/Data-Analyst-NTUA-Course---Vasileios-Dimitriou)

## ΕΠΙΛΟΓΗ / ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ-(l2036-c1-u1)

Στα πλαίσια της εξέτασης στο μάθημα επιλογή/επεξεργασία δεδομένων (i2036-c1-u1), γίνεται χρήση του repository Breast Cancer Wisconsin (Diagnostic) Data Set. Πιο συγκεκριμένα από το Data Folder του, θα γίνει αξιοποιηθούν τα αρχεία **«breast-cancer-wisconsin.data»** και **«breast-cancer-wisconsin.names»,** τα οποία περιλαμβάνουν τα πεδία:

* "Sampldate code number",
* "Clump Thickness",
* "Uniformity of Cell Size",
* "Uniformity of Cell Shape" ,
* "Marginal Adhesion" ,
* "Single Epithelial Cell Size" ,
* "Bare Nuclei" ,
* "Bland Chromatin" ,
* "Normal Nucleoli" ,
* "Mitoses " ,
* "Class"

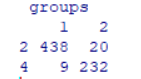
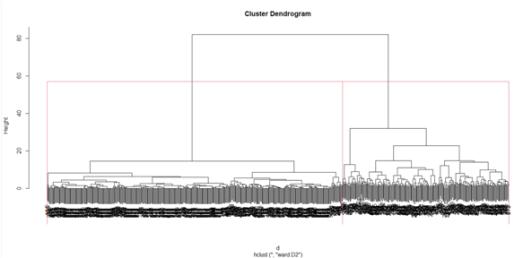
Παρατηρούμε λοιπόν ότι το πεδίο "Bare Nuclei" , περιλαμβάνει κάποια «?»-ερωτηματικά, για τα οποία έχει κρίθηκε ωφέλιμο το να απαλειφθούν. Μη έχοντας κάποια περεταίρω δεδομένα συμπληρώθηκαν στοχαστικά.

## ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ-ΚΑΙ ΣΥΜΠΕΡΑΣΜΑΤΑ(l2036-c1-u2)

Το πρόβλημα το οποίο έχει επιλεγεί είναι η χρήση όλων των μεθόδων (Ιεραρχικής Συσταδοποίησης, K-means Συσταδοποίησης, Δέντρου Αποφάσεων και Γραμμικής Παλινδρόμησης), οι οποίες έχουν διδαχτεί στα πλαίσια του παρόντος μαθήματος, πάνω στα δεδομένα. Το σημείο αναφοράς όλων το προβλημάτων είναι η κακοήθεια ή καλοήθεια κάθε περιστατικού ξεχωριστά, λαμβάνοντας υπόψιν όλα τα υπόλοιπα αποτελέσματα. Συνεπώς επιλέγω να έχω ένα πρόβλημα της μορφής [“Class” ~ “Clump Thickness”, "Uniformity of Cell Size", "Uniformity of Cell Shape", "Marginal Adhesion", "Single Epithelial Cell Size", "Bare Nuclei", "Bland Chromatin", "Normal Nucleoli", "Mitoses "]. Επίσης σε αυτό το σημείο κρίνεται απαραίτητο να τονιστεί ότι το πεδίο Class, είναι categorical variable, καθώς παίρνει τιμές 2 και 4, επομένως ανα περίπτωση θα πρέπει να προσεγγίσουμε το πρόβλημα έτσι ώστε να μπορέσουμε να εκφράσουμε μοντέλα που μας δίνουν υπολογισμό αποτελέσματος για Class με δυο εκδοχές - benigh (2) or malignant (4).

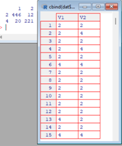
### Hierarchical Clustering

Παρατηρούμε λοιπόν ότι ο κώδικας της ιεραρχικής ομαδοποίησης μας δίνει, έχοντας επιλέξει να μας αποδώσει 2 κλάσσεις - ομάδες: Επίσης γίνεται χρήση της μεθόδου ward.D2.



Στον πίνακα δεξιά γίνεται σύγκριση μεταξύ των δεδομένων Class και αυτών που προέκυψαν από την μέθοδο ιεραρχικής ομαδοποίηση. Συμπεραίνουμε ότι εδώ η προκειμένη μέθοδος κάνει μια αρκετά καλή προσέγγιση χάνοντας μόνο λίγες τιμές [20 τιμές καλοήθειας που τις δίνει το μοντέλο σαν κακοήθεις και 9 τιμές κακοήθειας που αστοχεί το μοντέλο και τις εμφανίζει σας καλοήθεις]. Οι υπόλοιπες φαίνεται να έχουν συμπέσει με τις τιμές του πεδίου Class.

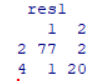
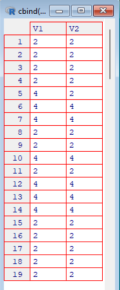
### K-means Clustering



Στον παραπάνω πίνακα φαίνονται οι αποκλίσεις που έχει αυτή η μέθοδος σε σχέση με τις πραγματικές τιμές του πεδίου Class. Παρατηρούμε ότι έχουμε 20 τιμές κακοήθειας που εμφανίζονται στο μοντέλο σαν καλοήθεια και 12 τιμές καλοήθειας που τις αποδίδει το μοντέλο σαν κακοήθεια.

### Predictive Decision Tree

Γίνεται χρήση όλων των δεδομένων 599 πρώτων δεδομένων (από όλα τα πεδία) για την δημιουργία του μοντέλου. Στην συνέχεια γίνεται πρόβλεψη για τις 100 τελευταίες περιπτώσεις κακοήθειας ή καλοήθειας. Βλέπουμε ότι έχουμε 1 τιμή κακοήθειας που εμφανίζεται ως καλοήθεια και 2 τιμές καλοήθειας που εμφανίζονται σαν κακοήθεια.



### Linear Regression Model

Έγινε προσπάθεια προσέγγισης του προβλήματος με γραμμικό μοντέλο, παρατηρούμε ότι η συγκεκριμένη εφαρμογή είναι αρκετά δύσχρηστη και δεν μας δίνει την δυνατότητα εύκολα να κάνουμε αναγωγή σε μικρότερες τιμές έτσι ώστε να μπορέσουμε να εμφανίσουμε categorical variables όπως έγινε και στα προηγούμενα μοντέλα. Έγινε και πάλι προσπάθεια πρόβλεψης των 100 τελευταίων τιμών, όμως ήταν δύσκολο στην προκειμένη περίπτωση, καθώς όπως φαίνεται έχουμε μεγάλες αποκλίσεις. Επίσης χρησιμοποιήθηκαν όλα τα δεδομένα και τα 699 για την προκειμένη ανάλυση.

