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)

Η ατομική μου εργασία αφορά μια μελέτη για εμφάνιση καρκίνου στο στήθος κάποιων εξεταζόμενων γυναικών, παραθέτοντας και επιπλέον δεδομένα της εξέτασης τους, πέραν από την κατάσταση τους (όγκος με κακοήθεια ή καλοήθεια).

## 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 (τις μεταβλητές):

* "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.

Στην προκειμένη περίπτωση έχουμε ένα πρόβλημα πρόβλεψης, για το οποίο θα επιδιωχθεί η εξαγωγή ενός κανόνα, στον οποίο θα λαμβάνονται υπόψιν όλες οι μεταβλητές εκτός αυτής της κλάσης, μεταβλητές εισόδου: ["Clump Thickness", "Uniformity of Cell Size", "Uniformity of Cell Shape" , "Marginal Adhesion" , "Single Epithelial Cell Size" , "Bare Nuclei" , "Bland Chromatin" , "Normal Nucleoli" , "Mitoses "], η οποία κλάση επί τοις ουσίας είναι μια μεταβλητή κατηγοριοποίησης, που κατατάσσει την έκαστη περίπτωση εξεταζομένου σε δυο κατηγορίες benigh (2) or malignant (4), δηλαδή αποτελεί την μεταβλητή εξόδου στο πρόβλημά μας. Οπότε το προς κατασκευή μοντέλο θα έχει ως στόχο την πρόβλεψη την έκαστη περίπτωση ασθενούς, αν δηλαδή ανήκει στην κατηγορία benigh (2) - καλοήθειας ή malignant (4) – κακοήθειας.

|  |  |
| --- | --- |
| **Όνονα μεταβλητής** | **Περιγραφή** |
| Sampldate code number | Κωδικοποίηση του δείγματος εξεταζομένου. |
| Clump Thickness | Μεταβλητές εισόδου που μετρώνται σε καθένα από τα δείγματα - δηλαδή τους εξεταζόμενους. |
| Uniformity of Cell Size |
| Uniformity of Cell Shape |
| Marginal Adhesion |
| Single Epithelial Cell Size |
| Bare Nuclei |
| Bland Chromatin |
| Normal Nucleoli |
| Mitoses |
| Class | Μεταβλητή εξόδου, κατηγοριοποιημένη σε δυο καταστάσεις benigh (2) ή malignant (4) - ορισμένη η κατάσταση του εξεταζομένου από τον επιστήμονα γιατρό που παρακολουθεί τον έκαστο εξεταζόμενο. |

Πίνακας 1: Περιγραφή μεταβλητών.

Το αρχείο των δεδομένων που διαθέτουμε εμπεριέχει 699 εγγραφές, δηλαδή 699 δειγματοληψίες που αντιστοιχεί σε 699 εξεταζόμενες από τον γιατρό με σκοπό το να αποφανθούμε αν η έκαστη περίπτωση φέρει καλοήθεια ή κακοήθεια.

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

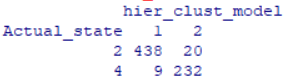
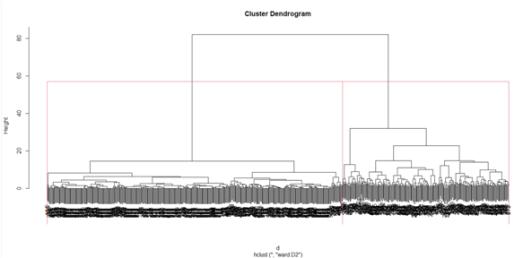
Το πρόβλημα το οποίο έχει επιλεγεί είναι η χρήση όλων των μεθόδων (Ιεραρχικής Συσταδοποίησης, K-means Συσταδοποίησης, Δέντρου Αποφάσεων και Γραμμικής Παλινδρόμησης), οι οποίες έχουν διδαχτεί στα πλαίσια του παρόντος μαθήματος, πάνω στα δεδομένα. Το σημείο αναφοράς όλων το προβλημάτων είναι η κακοήθεια (benigh) ή καλοήθεια (malignant) κάθε περιστατικού ξεχωριστά, λαμβάνοντας υπόψιν όλα τα υπόλοιπα αποτελέσματα. Αυτό συνεπάγεται ότι όλες οι υπόλοιπες μεταβλητές, αποτελούν μεταβλητές εισόδου, και το έκαστο μοντέλο βασισμένο σε αυτές τις μεταβλητές εισόδου επιδιώκει να αποδώσει έναν κανόνα εξαγωγής συμπεράσματος για την μεταβλητή εξόδου που είναι η Class [ benigh / malignant ]. 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). Παρακάτω περιγράφονται τα αποτελέσματα της ανάλυσης δεδομένων και παρουσιάζονται τα αποτελέσματα.

## ΠΑΡΟΥΣΙΑΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ (l2036-c1-u3)

### Hierarchical Clustering

Στόχος του συγκεκριμένου μοντέλου είναι η δημιουργία ενός δέντρου που θα χωρίζει τα δείγματα μας σε 2 παρακλάδια, δηλαδή τους εξεταζόμενους σε δυο ομάδες - κλάσεις, βασισμένοι στην αρχική μας λογική όπως δηλαδή ομαδοποιούν οι γιατροί τις διάφορες περιπτώσεις σε καλοήθεις (benigh- 2) και κακοήθεις (malignant - 4). Το μοντέλο λαμβάνει υπόψιν όλες τις μεταβλητές εισόδου.

Αρχικά εκτελούμε την εντολή κανονικοποίησης των μεταβλητών εισόδου (scale). Στην συνέχεια εξάγουμε την απόσταση μεταξύ των σημείων του συνόλου δεδομένων (dist), για να το εισάγουμε στην εντολή δημιουργίας δέντρου (hclust). Παρατηρούμε λοιπόν ότι ο κώδικας της ιεραρχικής ομαδοποίησης μας δίνει το παρακάτω δέντρο. Επίσης πρέπει να τονιστεί ότι για την διεκπεραίωση αυτού του προβλήματος γίνεται χρήση της μεθόδου ward.D2.

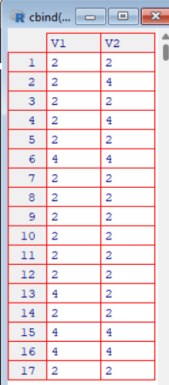
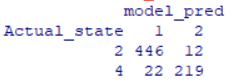


Εικόνα 1: Αρχική μοντελοποίηση ιεραρχικής ομαδοποίησης, για 2 ομάδες.

Μπορούμε να δούμε λοιπόν αριστερά το δέντρο που δημιουργείται από τον κώδικα, και παρατηρούμε πως τα χωρίζουμε με κόκκινο περίγραμμα - τετράγωνο σε δυο μεγάλες ομάδες. Στον πίνακα δεξιά γίνεται σύγκριση μεταξύ των πραγματικών δεδομένων Class και αυτών που προέκυψαν από την μέθοδο ιεραρχικής ομαδοποίηση (τα ίδια απεικονίζονται αριστερά στην 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

Στόχος του συγκεκριμένου μοντέλου είναι η δημιουργία ομαδοποίησης με την χρήση της μεθόδου k-means. Το μοντέλο αυτό για να κάνει την κανονικοποίηση λαμβάνει υπόψιν μόνο τις μεταβλητές εισόδου. Αρχικά κανονικοποιούμε τις μεταβλητές εισόδου (scale) και στην συνέχεια εισάγουμε τα κανονικοποιημένα δεδομένα στην εντολή kmeans για την δημιουργία επιλεγμένου αριθμού ομάδων των εξεταζομένων γυναικών. Στην δική μας περίπτωση επιθυμούμε να έχουμε 2 ομάδες. Προφανώς για τις περιπτώσεις κακοήθειας – καλοήθειας.

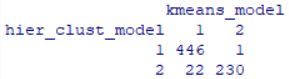


Εικόνα 2: Ομαδοποίηση k-means για 2 ομάδες.

Στον παραπάνω αριστερά πίνακα φαίνονται οι αποκλίσεις που έχει αυτή η μέθοδος σε σχέση με τις πραγματικές τιμές του πεδίου Class [2,4]. Οι γραμμές [2,4] αναπαριστούν τις πραγματικές περιπτώσεις καλοήθειας και κακοήθειας, ενώ οι στήλες [1,2] αποδίδουν τις προβλέψεις του μοντέλου σε σχέση με αυτές τις δυο καταστάσεις. Παρατηρούμε ότι σχετικά με τις περιπτώσεις αστοχίας του μοντέλου, έχουμε 22 τιμές κακοήθειας που εμφανίζονται στο μοντέλο σαν καλοήθεια και 12 τιμές καλοήθειας που τις αποδίδει το μοντέλο σαν κακοήθεια.

Στον δεξιά πίνακα βλέπουμε μια αναπαράσταση όλων και των 699 εξεταζομένων περιπτώσεων.

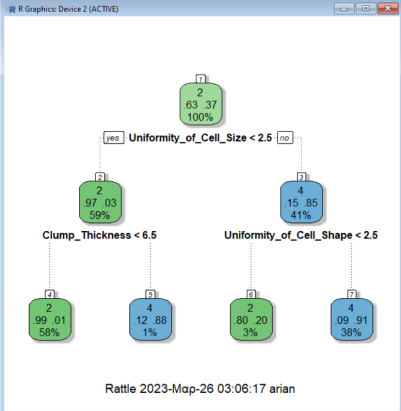
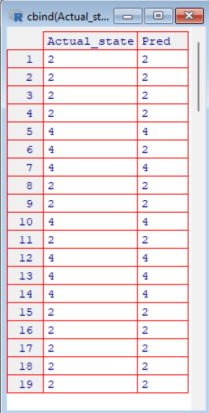
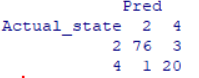
Συμπερασματικά λοιπόν μπορούμε να πούμε ότι το μοντέλο της ιεραρχικής ομαδοποίησης δούλεψε πολύ καλύτερα από αυτό της k-means καθώς αστοχεί για 9 περιπτώσεις κακοήθειας όγκου, ενώ αυτό της k-means αστοχεί για 22 περιπτώσεις κακοήθειας. Συνεπώς από την σκοπιά της κρισιμότητας της πρόβλεψης το μοντέλο ιεραρχικής ομαδοποίησης συμπεριφέρεται πολύ πιο ωφέλιμα. Καθώς επίσης μπορούμε να δούμε και τον παρακάτω πίνακα σύγκρισης προβλέψεων μοντέλων.



Εικόνα 3: Σύγκριση ιεραρχικής ομαδοποίησης και μοντέλου k-means.

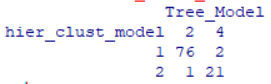
### Predictive Decision Tree

Γίνεται χρήση όλων των πρώτων 599 πρώτων δεδομένων (από όλα τα πεδία) για την δημιουργία του μοντέλου. Πιο συγκεκριμένα κάνουμε χρήση της rpart για την δημιουργία ενός μοντέλου το οποίο με βάση όλες τις υπόλοιπες μεταβλητές εκτιμάει την κατάσταση της ασθενούς. Στην συνέχεια γίνεται πρόβλεψη για τις 100 τελευταίες περιπτώσεις κακοήθειας ή καλοήθειας. Είναι πολύ σημαντικό να τονιστεί ότι πριν χτίσουμε το μοντέλο δέντρου αποφάσεων πρέπει να μετατρέψουμε το διάνυσμα Class των δεδομένων μας σε μεταβλητή κατηγορική, που θα έχει 2 level, θα μπορεί να πάρει μόνο τις τιμές 2 και 4, οι οποίες όπως προαναφέρθηκε αντιστοιχούν σε benigh – καλοήθη και malignant – κακοήθη αντίστοιχα.



Εικόνα 4: Μοντέλο δέντρου αποφάσεων.

Βλέπουμε στον πίνακα αριστερά ότι έχουμε 1 τιμή κακοήθειας που εμφανίζεται ως καλοήθεια και 3 τιμές καλοήθειας που εμφανίζονται σαν κακοήθεια. Σε αυτές τις περιπτώσεις το μοντέλο μας αστοχεί. Βλέπουμε και την σύγκριση των 100 πραγματικών τιμών (αριστερά στήλη) και των τιμών που προβλέπει το μοντέλο (δεξιά στήλη). Συμπεραίνουμε επίσης ότι αυτό το μοντέλο για υποεπταπλάσιο αριθμό συγκρινόμενων δεδομένων – προβλέψεων, έχει περίπου υποεπταπλάσιο αριθμό αστοχιών, συνεπώς εκτιμάται ότι έχει περίπου την ίδια συμπεριφορά με το μοντέλο της ιεραρχικής ομαδοποίησης. Παρατηρούμε λοιπόν στην παρακάτω εικόνα ότι οι προβλέψεις των μοντέλων αυτών διαφέρουν ελάχιστα – πήραμε τις 100 τελευταίες τιμές και για τα δυο μοντέλα.

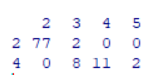
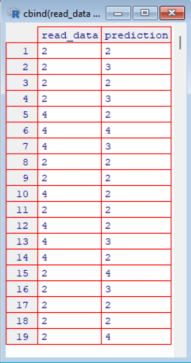


Εικόνα 5: Σύγκριση αποτελεσμάτων ιεραρχικής ομαδοποίησης και δέντρου αποφάσεων.

### Linear Regression Model

Δεν μπορεί να γίνει χρήση αυτής της μεθόδου στο πρόβλημά μας επειδή η μεταβλητή εξόδου είναι μια μεταβλητή 2 καταστάσεων ενώ οι μεταβλητές εισόδου είναι αριθμητικές. Αυτό σημαίνει ότι δεν μπορούμε να δημιουργήσουμε μοντέλο με αυτήν την μεταβλητή, ως κύρια μεταβλητή εξόδου ενώ ένα τέτοιο μοντέλο θα μπορούσε να χτιστεί αποκλειστικά και μόνο με αριθμητικές μεταβλητές.

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 – factor 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 για να χτιστεί το μοντέλο έγινε χρήση των 599 αρχικών δειγμάτων.



Εικόνα 6: Αποτελέσματα γραμμικής παλινδρόμισης.

Παρατηρούμε ότι δεν μας εξυπηρετεί.