Ανάλυση Κατηγορικών Δεδομένων

Εργασία 2

Ονοματεπώνυμο : Ιωάννα Τσέτσι

A.M.: s6160095

Διδάσκων : Αντωνία Κορρέ

June 30, 2022

Το Ινδικό Ινστιτούτο του Διαβήτη, πραγματοποίησε μία ερεύνα με σκοπό τη διερεύνηση της σημασίας κάποιων προγνωστικών παραγόντων στην πρόβλεψη της πιθανότητας για διαβήτη σε γυναίκες ηλικίας από 21 ετών και πάνω που προέρχονται απο την φυλή Pima.Οι προγνωστικοί παράγοντες που μετρήθηκαν παρουσίαζονται στον Πίνακα 1 [1].

Για την υλοποιήση της ανάλυσης θα χρησιμοποιήσουμε το στατιστικό πακέτο R. Τα δεδομένα βρίσκονται στην βιβιοθήκη mlbench και το σετ δεδομένων ονομάζεται Pima Indians Diabetes.

Εφόσον εχουμε εγκαταστήσει τις απαραίτητες βιβλιοθηκες μπορούμε να προχώρησουμε στην στατιστική ανάλυση.Αρχίκα βλέπουμε οτί στα δεδομένα έχουμε αρκετές ελλιπεις τιμές (missing values, NA).Για την αναλυσή των δεδομένων θα τις αφαιρέσουμε απο το αρχείο[2] .

Αριθμός Μεταβλήτης	Όνομα	Τύπος	Σήμασια
1	Pregnant	Αριθμητική	Πόσες φορές έμεινε έγκυος
2	Glocose	Αριθμητική	Συγκέντρωση γλυκόζης στο πλάσμα
3	Pressure	Αριθμητική	Αρτηριακή πίεση (mm Hg)
4	Triceps	Αριθμητική	Πάχος δερματικής πτυχής (mm)
5	Insulin	Αριθμητική	Ινσουλίνη ορού (mu U/ml)
6	Mass	Αριθμητική	Δείκτης μάζας σώματος (βάρος σε kg/(ύψος σε m)^2)
7	Pedigree	Αριθμητική	Score για την πιθανότητα διαβήτη από κληρονομικά αίτια
8	Age	Αριθμητική	Ηλικία (σε έτη)
9	Diabetes	Κατηγορίκη	Νόσησης από διαβήτη

Table 1: Μεταβλήτες Αρχείου Pima Indians Diabetes

[1] Για τις εντολές του έλεγχου του τυπού καθε μεταβλήτης βλεπε το Παράρτημα Listing 3

[2] Για τις εντολές για την αφαιρεση NA βλέπε στο Παράρτημα Listing 2

Στην συνέχεια θα προχωρήσουμε στη ομαδοποιήση δυο μεταβλήτων[3].Η πρώτη είναι η ομαδοποιήση της μεταβλήτης ηλικίας (Age) στις ομάδες 20-30, 31-40, 41-50 και 50+.
Ενώ η αλλή μεταβλήτη είναι ο αριθμός των κυήσεων (Pregnant) ομαδοποιήμενες στις ομάδες 0-5,6-10 και 10+. Τα δεδομένα παρουσιάζονται αναλυτικά στον Πίνακα 2[4].

Αριθμός Μεταβλήτης	Όνομα	Τύπος	Σήμασια	Τιμές
9	Diabetes	Κατηγορίκη	Νόσησης από διαβήτη	negative positive
8	age	Κατηγορίκη	Ομαδοποιήση της μεταβλήτης Age	"20-30","31-40" "41-50 " ,"50+"
1	pregnant	Κατηγορίκη	Ομαδοποιήση της μεταβλήτης Pregnant	"0-5",6-10" "10+"

Table 2: Μεταβλήτες Κατγορίκων Δεδόμενων .

Σκοπός της μελέτης είναι να φτιάξουμε ένα μοντέλο ωστέ να ερμηνευσούμε την πιθανότητα για διαβήτη σε γυναίκες ηλικίας 21 ετών και πάνω σε σχέση με τους προγνωστικούς παραγόντες που έχουν συλλεχθεί.Για την ανάλυση αυτή θα χρησιμοποιήσουμε το λογιστικό μοντέλο παλινδρομησης (logistic regression) για την διτιμή μεταβλήτη diabetes που εκφράζει το ενδεχόμενο νοσησής απο διαβήτη. Έστω Y η μεταβλήτη diabetes έχουμε οτί

$$Y_i = \begin{cases} 1 & \text{να έχει καποία διαβήτη με πιθανοτητα p} \\ 0 & \text{να μην έχει καποία διαβήτη με πιθανοτητα 1-p} \end{cases} \quad (1)$$

Η λογιστική σύναρτηση είναι η συνάρτηση συνδέσης συνδεσή σε αυτό το γενικεύμενο γραμμικό μοντέλο και για τα δεδομένα μας εχούμε

$$logit(p) = logit(\frac{p}{1-p}) = \boldsymbol{\beta}'\mathbf{X} = \beta_0 + \beta_1\chi_1 + ... + \beta_{11}\chi_{11}$$

[3]Για τις εντολές για την ομαδοποίηση των μεταβλήτων βλέπε Παράρτημα Listing 4
[4]Για τις εντολές του τύπου κάθε μεταβλητης βλέπε στο Παράρτημα Listing 5

Τα μοντέλο που προκύπτει με τα δεδομένα μας είναι το : [5]

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-8.990364	1.170461	-7.681	1.58e-14	***
glucose	0.039232	0.005865	6.689	2.24e-11	***
pressure	-0.002747	0.011892	-0.231	0.8173	
triceps	0.015717	0.017320	0.907	0.3642	
insulin	-0.000692	0.001351	-0.512	0.6085	
mass	0.063485	0.027425	2.315	0.0206	*
pedigree	1.051124	0.434989	2.416	0.0157	*
age31-40	0.802485	0.395535	2.029	0.0425	*
age41-50	1.377709	0.536629	2.567	0.0102	*
age50+	1.236659	0.633326	1.953	0.0509	.
pregnant6-10	-0.150064	0.424600	-0.353	0.7238	
pregnant10+	0.930171	0.779488	1.193	0.2327	

Table 3: Μοντέλο λογιστικής παλινδρόμησης με όλες τις μεταβλητές

$$\boldsymbol{\beta}' = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \\ \hat{\beta}_5 \\ \hat{\beta}_6 \\ \hat{\beta}_7 \\ \hat{\beta}_8 \\ \hat{\beta}_9 \\ \hat{\beta}_{10} \\ \hat{\beta}_{11} \end{bmatrix} \approx \begin{bmatrix} -8.990364 \\ 0.039232 \\ -0.002747 \\ 0.015717 \\ -0.000692 \\ 0.063485 \\ 1.051124 \\ 0.802485 \\ 1.377709 \\ 1.236659 \\ -0.150064 \\ 0.930171 \end{bmatrix}$$

[5] Για τις εντολές για την λογιστική παλινδρομηση ολών των μεταβλήτων βλέπε στο Παράρτημα Listing 6

Απο τον Πίνακα 3 βλέπουμε οτι η σταθέρα $\hat{\beta}_0$ εχεί την τιμή -8.90364.Αυτο σημαινει οτι το log odds για να νοσήσει κάποια απο διαβήτη η οποία ανήκει στην ηλικιάκη ομάδα $20-30$ και ο αριθμός των κύησεων της είναι απο $0-5$ είναι -8.90364.

Για την πάραμετρο glucose $\hat{\beta}_1$ βλέπουμε οτί η 1 μοναδιαία αυξήση της συγκέντρωσης γλύκοζης στο πλάσμα αυξάνει τα log odd κατα 0.039232 , οταν ολές οι υπολοιπές μεταβλήτες παραμένουν σταθερές.

Ενώ για την παράμετρο pressure $\hat{\beta}_2$ έχουμε οτι 1 μοναδιαία αυξηση της αρτηριακής πίεσης(mm Hg) μείωνει τα log odds κάτα -0.002747 , όταν ολές οι υπολοιπές μεταβήτες είναι παραμενουν σταθερές.

Ακόμα για την παράμετρο triceps $\hat{\beta}_3$ έχουμε οτι 1 μοναδιαία αυξηση του πάχους δερματικής πτύχης(mm) αυξάνει τα log odds κατα 0.015717, όταν ολές οι υπολοιπές μεταβήτες είναι παραμενουν σταθερές.

Επιπλέον για την παράμετρο insulin $\hat{\beta}_4$ παρατηρούμε οτι 1 μοναδιαία αυξήση του ινσουλινη ορού (mu U/ml) μείωνει τα log odds κατα-0.000692, όταν ολές οι υπολοιπές μεταβήτες είναι παραμενουν σταθερές.

Αντιστοιχα για την παράμετρο mass $\hat{\beta}_5$ παρατηρούμε οτι 1 μοναδιαία αυξήση του δείκτη μάζας σώματος (βάρος σε kg/(ύψος σε m)^2) αυξάνει τα log odds κατα 0.063485, όταν ολές οι υπολοιπές μεταβήτες είναι παραμενουν σταθερές.

Ενώ για την παράμετρο pedigree $\hat{\beta}_6$ παρατηρούμε οτι 1 μοναδιαία αυξήση του Score για την πιθανότητα διαβήτη από κληρονομικά αίτια αυξάνει τα log odds κατα 1.051124 οταν ολές οι υπολοιπές μεταβήτες είναι παραμενουν σταθερές.

Ακόμα για τις γυναίκες που ανήκουν στην ηλικιάκη ομάδα 31-40 απο την παραμετρο age31-40, $\hat{\beta}_7$, εχούμε

$$\hat{\beta}_0 + \hat{\beta}_7 = -8.90364 + 0.802485 = -8.187879$$

Δηλάδη τα log odds για τις γυναικές που νοσούν απο διαβήτη και ανήκουν στην ηλικιάκη κατηγορία 31-40 μέιωνονται κατα -8.187879 απο τις γυναίκες που ανήκουν στην ηλικιακή κατηγορία 20-30 και νοσουν απο διαβήτη.

Ενώ για τις γυναίκες που ανήκουν στην ηλικιάκη ομάδα 41-50 απο την παραμετρο age41-50 , $\hat{\beta}_8$, εχούμε

$$\hat{\beta}_0 + \hat{\beta}_8 = -8.90364 + 1.377709 = -7.525931$$

Δηλάδη τα log odds για τις γυναικές που νοσούν απο διαβήτη και ανήκουν στην ηλικιάκη κατηγορία 41-50 μέιωνονται κατα -7.525931 απο τις γυναίκες που ανήκουν στην ηλικιακή κατηγορία 20-30 και νοσουν απο διαβήτη.

Για τις γυναίκες που ανήκουν στην ηλικιάκη ομάδα 50 και πανω απο την παραμετρο age50+ , $\hat{\beta}_9$, εχούμε

$$\hat{\beta}_0 + \hat{\beta}_9 = -8.90364 + 1.236659 = -7.666981$$

Δηλάδη τα log odds για τις γυναικές που νοσούν απο διαβήτη και ανήκουν στην ηλικιάκη κατηγορία 50 και πάνω μέιωνονται κατα -7.666981 απο τις γυναίκες που ανήκουν στην ηλικιακή κατηγορία 20-30 και νοσουν απο διαβήτη.

Οσό αναφορα τον αριθμό κυήσεων έχουμε που είχε η κάθε γυναίκα βλέπουμε απο υην μεταβλήτη pregnant6-10, $\hat{\beta}_{10}$,οτί

$$\hat{\beta}_0 + \hat{\beta}_{10} = -8.90364 - 0.150064 = -9.053704$$

Δηλάδη τα log odds για τις γυναικές που νοσούν απο διαβήτη και ο αριθμός των κυήσεων τους είανι απο 6-10 μέιωνονται κατα -7.666981 απο τις γυναίκες που ο αριθμός των κυήσεων τους είναι απο 0-5 και νοσουν απο διαβήτη.

Τέλος βλέπουμε οτι για τις γυναίκες με αριθμό κυήσεων πάνω απο 10 στην παράμετρο pregnant10+, $\hat{\beta}_{11}$,οτί

$$\hat{\beta}_0 + \hat{\beta}_{11} = -8.90364 + 0.930171 = -7.973469$$

Δηλάδη τα log odds για τις γυναικές που νοσούν απο διαβήτη και ο αριθμός των κυήσεων τους είανι απο 10 και πάνω μέιωνονται κατα -7.973469 απο τις γυναίκες που ο αριθμός των κυήσεων τους είναι απο 0-5 και νοσουν απο διαβήτη.

Στην συνέχεια θα προχώρησουμε στον έλεγχο αν το μοντέλο μας είναι προσαρμοζεί καλύτερα απο ένα μοντέλο με μονο ορό την σταθερα. Ο ελέγχος που κάνουμε για την διαφορα των αποκλίσεων των δυο μοντέλων είναι

$$D_n(\hat{\beta}) - D_m(\hat{\beta}) \to x^2_{df_{null}-df_m}$$

[6]Απο οπού και συμπαιρένουμε οτί υπαρχεί σημαντική διαφορα ανάμεσα στο μοντέλο μας απο το μοντέλο απο την σταθέρα.

Στον Πίνακα 3 παρουσίαζονται και τα αποτελέσματα για τους συντελέστες του μοντέλου. Απο την δευτερη στηλή εχούμε το τυπικό σφάλμα για κάθε παραμέτρο (Std.Error), ενω στην τρίτη στηλή εχούμε τον στατιστικό έλεγχου Z για τον ελέγχο

$$H_0 : \beta_i = 0 \text{vs} H_1 : \beta_i \neq 0$$

και στην τέλευταια στήλη παρουσίαζονται οι αντίστοιχες p-τιμές για κάθε έλεγχο αντίστοιχα. Αυτο που παρατηρούμε είναι οτι αρκέτες απο τις παραμέτρους έχουμε ισχύρες στατιστικές ενδείξεις για να αποδεχτουμε την μηδενική υποθεση ,δηλαδη οτί $\beta_i = 0$,οπως για παράδειγμα η παραμτρεος pressure.Ενώ για αλλές παραμέτρους βλέπουμε οτί ορίακα μπορούμε να τις θεωρησουμε στατιστικά σημαντικες για το μοντέλο μας($p - value \approx a(= 0.05)$).

Για την επιλογη των μεταβλήτων που θα χρησιμοποιήσουμε στο μοντέλο είναι το κριτήριο Bayesian Information Criterion (BIC).Ακόμα η μέθοδος επιλογής που θα χρήσιμοποιήσουμε είναι σε κάθε βήμα να γίνεται και η εφαρμογή και των δυο τεχνικών,δηλάδη με την μεθοδο επιλογης προς τα εμπρός και πρός τα πίσω.Συμφώνα με το BIC το επίλεγμενο μοντέλο περιέχει τις εξής παραμετρους [7]:

- glucose : Συγκέντρωση γλυκόζης στο πλάσμα
- mass :Δείκτης μάζας σώματος
- pedigree :Score για την πιθανότητα διαβήτη από κληρονομικά αίτια
- age : Ηλικία

[6]Για τις εντολές για τον ελεγχος προσαρμογής για το μοντέλο βλέπε στο Παράρτημα Listing 7

[7]Για τις εντολές για την επιλογή του μοντέλου βλέπε στο Παράρτημα Listing 8

Τα μοντέλο που προκύπτει συμώνα με το κρίτηριο BIC είναι το [8]:

$$logit(\frac{\hat{P}(\text{νοσει απο διαβήτη})}{\hat{P}(\text{δεν νοσει απο διαβήτη})}) = -9.023004 + 0.037482 \times (glucose) + 0.074965 \times (mass)$$

$$+1.088329 \times (pedigree) + 0.798086 \times (group\_age31 - 40)$$

$$+1.553980 \times (group\_age41 - 50) + 1.291923 \times (group\_age50+)$$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-9.023004	1.009742	-8.936	< 2e-16 ***
glucose	0.037482	0.005036	7.443	9.82e-14 ***
mass	0.074965	0.020504	3.656	0.000256 ***
pedigree	1.088329	0.427625	2.545	0.010926 *
age31-40	0.798086	0.342608	2.329	0.019836 *
age41-50	1.553980	0.416508	3.731	0.000191 ***
age50+	1.291923	0.526703	2.453	0.014173 *

Table 4: Μοντέλο λογιστικής παλινδρόμησης συμφωνα με το κρίτηριο BIC.

Απο τον Πίνακα 4 και την τελευταια στήλη βλέπουμε οτί ολές οι παραμετροι είναι στατιστικά σημαντικές σε επιπέδο σημαντικότητας 5%.

Επιπλέον εχούμε οτι η σταθέρα $\hat{\beta}_0$ εχεί την τιμή -9.023004.Αυτο σημαινει οτι το log odds για να νοσήσει κάποια απο διαβήτη η οποία ανήκει στην ηλικιάκη ομάδα $20 - 30$ είναι -9.023004.

Για την πάραμετρο glucose $\hat{\beta}_1$ βλέπουμε οτί η 1 μοναδιαία αυξήση της συγκέντρωσης γλύκοζης στο πλάσμα αυξάνει τα log odd κατα 0.037482 , οταν ολές οι υπολοιπές μεταβλήτες παραμένουν σταθερές.

Αντιστοιχα για την παράμετρο mass $\hat{\beta}_2$ παρατηρούμε οτι 1 μοναδιαία αυξήση του δείκτη μάζας σώματος (βάρος σε kg/(ύψος σε m)^2) αυξάνει τα log odds κατα 0.074965, όταν ολές οι υπολοιπές μεταβήτες είναι παραμενουν σταθερές.

[8] Για τις εντολές για την λογιστική παλινδρομηση με τις μεταβλήτες απο το BIC βλέπε στο Παράρτημα Listing 9

Ενώ για την παράμετρο pedigree $\hat{\beta}_3$ παρατηρούμε οτι 1 μοναδιαία αυξήση του Score για την πιθανότητα διαβήτη από κληρονομικά αίτια αυξάνει τα log odds κατα 1.088329 οταν ολές οι υπολοιπές μεταβήτες είναι παραμενουν σταθερές.

Ακόμα για τις γυναίκες που ανήκουν στην ηλικιάκη ομάδα 31-40 απο την παραμετρο age31-40 , $\hat{\beta}_4$, εχούμε

$$\hat{\beta}_0 + \hat{\beta}_4 = -9.023004 + 0.798086 = -8.224918$$

Δηλάδη τα log odds για τις γυναικές νοσούν απο διαβήτη που ανήκουν στην ηλικιάκη κατηγορία 31-40 μέιωνονται κατα -8.187879 απο τις γυναίκες που ανήκουν στην ηλικιακή κατηγορία 20-30 και νοσουν απο διαβήτη.

Ενώ για τις γυναίκες που ανήκουν στην ηλικιάκη ομάδα 41-50 απο την παραμετρο age41-50 , $\hat{\beta}_5$, εχούμε

$$\hat{\beta}_0 + \hat{\beta}_5 = -9.023004 + 1.553980 = -7.469024$$

Δηλάδη τα log odds για τις γυναικές νοσούν απο διαβήτη και ανήκουν στην ηλικιάκη κατηγορία 41-50 μέιωνονται κατα -8.224918 απο τις γυναίκες που ανήκουν στην ηλικιακή κατηγορία 20-30 και νοσουν απο διαβήτη.

Για τις γυναίκες που ανήκουν στην ηλικιάκη ομάδα 50 και πανω απο την παραμετρο age50+ ,$\hat{\beta}_6$, εχούμε

$$\hat{\beta}_0 + \hat{\beta}_6 = -9.023004 + 1.291923 = -7.731081$$

Δηλάδη τα log odds για τις γυναικές νοσούν απο διαβήτη και ανήκουν στην ηλικιάκη κατηγορία 50 και πάνω μέιωνονται κατα -7.731081 απο τις γυναίκες που ανήκουν στην ηλικιακή κατηγορία 20-30 και νοσουν απο διαβήτη.

Για μια γυναίκα ηλικίας 35 χρονών έχουμε οτί η πιθανοτητα να νοσήσει μια γυναικα απο διάβητη με ολές τις μεταβλήτες να είναι σταθερές
(στις μέσες τιμές τους) είναι [9]

$$logit\Big(\frac{\hat{P}(\text{νοσει απο διαβήτη}|\text{age }=35)}{\hat{P}(\text{δεν νοσει απο διαβήτη}|\text{age }=35)}\Big) = 0.3591437$$

Ενώ τα odds ratio για μια γυναίκα να νοσήσει απο διαβήτη στην ηλικία των 35 είναι [10]

$$exp(0.3591437) = 1.432103$$

Ενώ για μια γυναίκα ηλικίας 45 χρονών έχουμε οτί η πιθανοτητα να νοσήσει μια γυναικα απο διάβητη με ολές τις μεταβλήτες να είναι σταθερές
(στις μέσες τιμές τους) είναι [11]

$$logit\Big(\frac{\hat{P}(\text{νοσει απο διαβήτη}|\text{age }=45)}{\hat{P}(\text{δεν νοσει απο διαβήτη}|\text{age }=45)}\Big) = 0.544088$$

Ενώ τα odds ratio για μια γυναίκα να νοσήσει απο διαβήτη στην ηλικία των 45 είναι [12]

$$exp(0.544088) = 1.723036$$

Ακόμα τα odds ratio για την εμφάνιση διαβήτη μεταξύ της γυναίκας 35 με την εμφάνιση του διαβήτη για την γυναίκα 45 ετών είναι [13]

$$\frac{exp(0.3591437)}{exp(0.544088)} = \frac{1.432103}{1.723036} = 0.8311509$$

Δηλάδη οι γυναικές που ανήκουν στην ηλικιάκη ομάδα 45 ετών αυξανονται τα odds ratio κατα 8.3% για την εμφάνσιση του διαβήτη σε σχέση τις γυναίκες που ανήκουν στην ηλικιακή ομαδα 35 ετών.

[9] Για τις εντολές για την πρόβλεψη για την εμφάνιση διαβήτη για μια γυναίκα ηλικίας 35,αποτέλεσμα σε log odd βλέπε στο Παράρτημα Listing 10

[10] Για τις εντολές για τα Odds Ratio για την εμφάνιση διαβήτη για μια γυναίκα ηλικίας 35 βλέπε στο Παράρτημα Listing 11

[11] Για τις εντολές για την Πρόβλεψη για την εμφάνιση διαβήτη για μια γυναίκα ηλικίας 45,αποτέλεσμα σε log odds βλέπε στο Παράρτημα Listing 12

[12] Για τις εντολές για Odds Ratio για την εμφάνιση διαβήτη για μια γυναίκα ηλικίας 45 βλέπε στο Παράρτημα Listing 13

[13] Για τις εντολές για τα Odds Ratio εμφάνισης διαβήτη μεταξύ της γυναίκας 35 με την εμφάνιση του διαβήτη για την γυναίκα 45 ετών βλέπε στο Παράρτημα Listing 14

Στην συνεχεία θα προχωρήσουμε στην δημιουργεία ενός Confusion matrix. Ο πίνακας αυτός στην ουσία μας επιτρεπεί να οπτικοποιήσουμε την αποδόση ενος αλγοριθμου,στην περιπρωση μας του λογιστικου μοντέλου που χρήσιμοποιήσαμε παραπάνω.

[14]Αυτό που πρόσπαθουμέ να κάνουμε είναι να πρόβλεψουμε πότε ο αλγορίθμος για την δίτιμη παραμετρο του διαβήτη θα δώσει ως αποτέλεσμα y=0 ή y=1. Για την προβλέψη οτί η γυναίκα θα νοσήσει απο διαβήτη (y=1) υποθετουμε οτί η πιθανοτητα θα είναι μεγαλήτερη απο ένα όριο $\hat{\pi}_0$ που θα έχουμε ορίσει εμείς δηλαδή $\hat{\pi}_i > \hat{\pi}_0$ και για την πρόβλεψή οτι η γυναίκα δεν θα νοσήσει (y=0)υποθέτουμε οτί $\hat{\pi}_i < \hat{\pi}_0$ Εμείς για $\hat{\pi}_0$ θα το ορίσουμε ισό με $\hat{\pi}_0 = 0.5$

[15]Τα αποτέλεσματα που παίρνουμε παρουσιάζονται στον παρακάτω πίνακα:

Prediction, π_0 =0.5		
Actual	neg	pos
neg	111	30
pos	15	40

Table 5: Confusion matrix.

[16]Απο οπου και συμπαιρένουμε οτί το ποσοστο ορθής ταξονομησης εκείνων που νοσήσαν απο διαβήτη είναι 57.15%.

$$\text{Sensitivity(Ευαισθησία)} = P(\hat{y} = 1|y = 1) = \frac{40}{40 + 30} = 57.14286\%$$

Ενώ το ποσοστό ορθής ταξινομησης εκείνων που δεν νοσήσαν είναι 88.1%.

$$\text{Specificity(Ειδικότητα)} = P(\hat{y} = 0|y = 0) = 1 - \frac{15}{111 + 15} = 88.09524\%$$

Αυτο που βλέπουμε είναι οτί η ειδικότητα είναι αρκέτα μεγαλή.

[14]Για τις εντολές Λογιστικη παλινδρομηση για την πρόβλεψη της μεταβλητής diabete βλέπε στο Παράρτημα Listing 15

[15]Για τις εντολές Δημιρουργεία Confusion Matric βλέπε στο Παράρτημα Listing 16

[16]ΥπολογισμοςΕυαισθησία (Sensitivity) και Ειδικότητα (Specificity) βλέπε στο Παράρτημα Listing 17

Ακόμα υπάρχουν τα δεδόμενα για τρείς γυναικές με συγκεκριμένα χαρακτηριστικά οπως παρουσιάζονται στον Πίνακα 5.

Γυναίκες	1	2	3
Pregnant	3	2	0
Glocose	95	55	80
Pressure	50	62	70
Triceps	21	17	30
Insulin	70	100	110
Mass	30	33	35
Pedigree	0.5	0.7	0.9
Age	50	40	35

Table 6: Δεδόμενα για τα 95% διαστηματα εμπιστοσύνης

Για τα παραπάνω δεδομένα θα χρείαστει να δημιουργήσουμε σημειακά 95% διαστήματα πρόβλεψης για την πιθανότητα εμφάνισης διαβήτη στις τρείς αυτές γυναίκες [17].

Γυναίκες	Πίθανότητα Πρόβλεψης	Ανώ όριο δ.ε	Κάτω όριο δ.ε
1	0.2469113	0.4225588	0.1280815
2	0.1369591	0.2534081	0.06907139
3	0.06238096	0.1475301	0.02493914

Table 7: 95% σημείακα διαστηματα πρόβλεψης

Για την Γυναίκα 1 με τα χαρακτηριστηκα οπως δινονται στον πίνακα 6 έχουμε ότι η πιθανοτητα πρόβλψης να νοσήσει είναι 0.24 και με σημειακα διαστήματα (0.12,0.42). Ενώ για την δεύτερη γυναίκα η πιθανοτητα πρόβλψης να νοσήσει είναι 0.13 και με σημειακα διαστήματα (0.07,0.25).Τέλος για την τρίτη γυναίκα η πιθανοτητα πρόβλψης να νοσήσει είναι 0.06 και με σημειακα διαστήματα (0.02,0.14).

[17]Για τις εντολές για τα σημειακα δ.ε. βλέπε στο Παράρτημα για την πρωτη γυναικα Listing 18 ,για την 2 γυναικα Listing 19 και για την τριτη γυναικα Listing 20

ΠΑΡΑΡΤΗΜΑ

```r

rm(list=ls(all=TRUE))

install.packages("mlbench")
library("mlbench")

data(PimaIndiansDiabetes2)
data<-PimaIndiansDiabetes2
```

Listing 1: Εγκατάσταση βιβλιοθηκών και δεδομένων

ΚΩΔΙΚΑΣ ΕΡΩΤΗΜΑΤΟΣ 1

```r
sum(is.na(data))
# [1] 652

new.data<-na.omit(data)
sum(is.na(new.data))
#[1] 0
```

Listing 2: Ελέγχος ελλειπών τιμών και αφαιρεσή τους

```r
str(new.data)
#'data.frame':   392 obs. of  9 variables:
# $ pregnant: num  1 0 3 2 1 5 0 1 1 3 ...
# $ glucose : num  89 137 78 197 189 166 118 103 115 126 ...
# $ pressure: num  66 40 50 70 60 72 84 30 70 88 ...
# $ triceps : num  23 35 32 45 23 19 47 38 30 41 ...
# $ insulin : num  94 168 88 543 846 175 230 83 96 235 ...
# $ mass    : num  28.1 43.1 31 30.5 30.1 25.8 45.8 43.3 34.6
   39.3 ...
# $ pedigree: num  0.167 2.288 0.248 0.158 0.398 ...
# $ age     : num  21 33 26 53 59 51 31 33 32 27 ...
# $ diabetes: Factor w/ 2 levels "neg","pos": 1 2 2 2 2 2 2 1 2
    1 ...
# - attr(*, "na.action")= 'omit' Named int [1:376] 1 2 3 6 8 10
    11 12 13 16 ...
#  ..- attr(*, "names")= chr [1:376] "1" "2" "3" "6" ...
```

Listing 3: Ελέγχος του τύπου των μεταβλήτων

```
1
2 new.data$age<-cut(new.data$age, breaks=c(20,31,41,50,100),
    labels=c("20-30","31-40","41-50","50+"))
3
4  levels(new.data$age)
5 #[1] "20-30" "31-40" "41-50" "50+"
6
7 new.data$pregnant<-cut(new.data$pregnant, breaks=c(-1,5,10,17),
    labels=c("0-5","6-10","10+"))
8
9 levels(new.data$pregnant)
10 > levels(new.data$pregnant)
11 # [1] "0-5"  "6-10" "10+"
```

Listing 4: Ομαδοποιήση των μεταβλήτης Age και Pregnant

```
1  str(new.data)
2 #'data.frame':   392 obs. of  9 variables:
3 # $ pregnant: Factor w/ 3 levels "0-5","6-10","10+": 1 1 1 1 1
    1 1 1 1 1 ...
4 # $ glucose : num  89 137 78 197 189 166 118 103 115 126 ...
5 # $ pressure: num  66 40 50 70 60 72 84 30 70 88 ...
6 # $ triceps : num  23 35 32 45 23 19 47 38 30 41 ...
7 # $ insulin : num  94 168 88 543 846 175 230 83 96 235 ...
8 # $ mass    : num  28.1 43.1 31 30.5 30.1 25.8 45.8 43.3 34.6
    39.3 ...
9 # $ pedigree: num  0.167 2.288 0.248 0.158 0.398 ...
10 # $ age     : Factor w/ 4 levels "20-30","31-40",..: 1 2 1 4 4
    4 1 2 2 1 ...
11 # $ diabetes: Factor w/ 2 levels "neg","pos": 1 2 2 2 2 2 2 1 2
     1 ...
12 # - attr(*, "na.action")= 'omit' Named int [1:376] 1 2 3 6 8 10
     11 12 13 16 ...
13 #  ..- attr(*, "names")= chr [1:376] "1" "2" "3" "6" ...
```

Listing 5: Ελέγχος του τύπου των μεταβλήτων για το τελικό αρχείο.

```
1
2 mylogit <- glm(formula = diabetes~glucose+ pressure+ triceps+
    insulin+mass+pedigree+age+pregnant,
3 +                 family = "binomial", data = new.data)
4  summary(mylogit)
5
6 #Call:
7 #glm(formula = diabetes ~ glucose + pressure + triceps +
    insulin +
8 #    mass + pedigree + age + pregnant, family = "binomial",
    data = new.data)
9 #
10 #Deviance Residuals:
11 #    Min       1Q   Median       3Q      Max
12 #-2.6284  -0.6525  -0.3584   0.5869   2.6191
13 #
14 #Coefficients:
15 #              Estimate Std. Error z value Pr(>|z|)
16 #(Intercept)  -8.990364   1.170461  -7.681 1.58e-14 ***
17 #glucose       0.039232   0.005865   6.689 2.24e-11 ***
18 #pressure     -0.002747   0.011892  -0.231   0.8173
19 #triceps       0.015717   0.017320   0.907   0.3642
20 #insulin      -0.000692   0.001351  -0.512   0.6085
21 #mass          0.063485   0.027425   2.315   0.0206 *
22 #pedigree      1.051124   0.434989   2.416   0.0157 *
23 #age31-40      0.802485   0.395535   2.029   0.0425 *
24 #age41-50      1.377709   0.536629   2.567   0.0102 *
25 #age50+        1.236659   0.633326   1.953   0.0509 .
26 #pregnant6-10 -0.150064   0.424600  -0.353   0.7238
27 #pregnant10+   0.930171   0.779488   1.193   0.2327
28 #---
29 #Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
30 #
31 #(Dispersion parameter for binomial family taken to be 1)
32 #
33 #    Null deviance: 498.10  on 391  degrees of freedom
34 #Residual deviance: 340.65  on 380  degrees of freedom
35 #AIC: 364.65
36 #
37 #Number of Fisher Scoring iterations: 5
```

Listing 6: Λογιστική Παλινδρομήση με ολες τις μεταβλήτες.

```
1  with(mylogit, pchisq(null.deviance - deviance,
2 df.null - df.residual, lower.tail = FALSE))
3
4 #[1] 4.462115e-28
```

Listing 7: Έλεγχος προσαρμογής για το μοντέλο.

ΚΩΔΙΚΑΣ ΕΡΩΤΗΜΑΤΟΣ 2

```
1
2  n <- dim(new.data)[1]
3  model_bic <- step(mylogit, trace=TRUE, direction = 'both', k =
       log(n))
4
5  #Start:  AIC=412.31
6  #diabetes ~ glucose + pressure + triceps + insulin + mass +
       pedigree +
7  #    age + pregnant
8  #
9  #           Df Deviance    AIC
10 #- pregnant  2   342.90 402.61
11 #- age       3   349.12 402.86
12 #- pressure  1   340.70 406.39
13 #- insulin   1   340.91 406.60
14 #- triceps   1   341.47 407.16
15 #- mass      1   346.14 411.83
16 #<none>          340.65 412.31
17 #- pedigree  1   346.80 412.49
18 #- glucose   1   395.09 460.78
19 #
20 #Step:  AIC=402.61
21 #diabetes ~ glucose + pressure + triceps + insulin + mass +
       pedigree +
22 #    age
23 #
24 #           Df Deviance    AIC
25 #- pressure  1   342.92 396.66
26 #- insulin   1   343.29 397.03
27 #- triceps   1   343.62 397.36
28 #- age       3   360.41 402.20
29 #- mass      1   348.62 402.36
30 #<none>          342.90 402.61
31 #- pedigree  1   349.52 403.26
32 #+ pregnant  2   340.65 412.31
33 #- glucose   1   397.69 451.43
34 #
35 #
36 #
37 #
38 #
39 #
40 #
41 #
42 #
```

```
#
#
#Step:  AIC=396.66
#diabetes ~ glucose + triceps + insulin + mass + pedigree + age
#
#           Df Deviance    AIC
#- insulin   1   343.30 391.07
#- triceps   1   343.64 391.41
#<none>          342.92 396.66
#- mass      1   348.91 396.68
#- age       3   361.34 397.17
#- pedigree  1   349.60 397.37
#+ pressure  1   342.90 402.61
#+ pregnant  2   340.70 406.39
#- glucose   1   398.23 446.00
#
#Step:  AIC=391.07
#diabetes ~ glucose + triceps + mass + pedigree + age
#
#           Df Deviance    AIC
#- triceps   1   344.05 385.85
#- mass      1   348.96 390.76
#<none>          343.30 391.07
#- pedigree  1   349.80 391.59
#- age       3   361.77 391.63
#+ insulin   1   342.92 396.66
#+ pressure  1   343.29 397.03
#+ pregnant  2   340.95 400.66
#- glucose   1   412.61 454.40
#
#Step:  AIC=385.85
#diabetes ~ glucose + mass + pedigree + age
#
#           Df Deviance    AIC
#<none>          344.05 385.85
#- pedigree  1   350.93 386.76
#- age       3   363.70 387.59
#+ triceps   1   343.30 391.07
#+ insulin   1   343.64 391.41
#+ pressure  1   344.04 391.81
#- mass      1   358.36 394.19
#+ pregnant  2   341.80 395.54
#- glucose   1   413.84 449.67
```

Listing 8: Επιλογή μοντέλου με το κρίτηριο BIC

```
1
2  mylogit1 <- glm(formula =diabetes ~ glucose + mass + pedigree +
       age,
3                  family = "binomial", data = new.data)
4  summary(mylogit1)
5
6  #Call:
7  #glm(formula = diabetes ~ glucose + mass + pedigree + age,
       family = "binomial",
8  #    data = new.data)
9  #
10 #Deviance Residuals:
11 #    Min       1Q   Median       3Q      Max
12 #-2.7102  -0.6399  -0.3706   0.6409   2.6590
13 #
14 #Coefficients:
15 #             Estimate Std. Error z value Pr(>|z|)
16 #(Intercept) -9.023004   1.009742  -8.936  < 2e-16 ***
17 #glucose      0.037482   0.005036   7.443 9.82e-14 ***
18 #mass         0.074965   0.020504   3.656 0.000256 ***
19 #pedigree     1.088329   0.427625   2.545 0.010926 *
20 #age31-40     0.798086   0.342608   2.329 0.019836 *
21 #age41-50     1.553980   0.416508   3.731 0.000191 ***
22 #age50+       1.291923   0.526703   2.453 0.014173 *
23 #---
24 #Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
25 #
26 #(Dispersion parameter for binomial family taken to be 1)
27 #
28 #    Null deviance: 498.10  on 391  degrees of freedom
29 #Residual deviance: 344.05  on 385  degrees of freedom
30 #AIC: 358.05
31 #
32 #Number of Fisher Scoring iterations: 5
```

Listing 9: Λογιστική Παλινδρόμηση για τις μεταβλήτες συμφωνα με το κρίτηριο BIC .

ΚΩΔΙΚΑΣ ΕΡΩΤΗΜΑΤΟΣ 3

```
as.numeric(predict(mylogit1 ,
newdata=data.frame(
age=factor(factor('31-40',levels=c('20-30','31-40','41-50','
   50+'))),
glucose =mean(new.data$glucose ),
pressure=mean(new.data$pressure),
triceps =mean(new.data$triceps ),
insulin =mean(new.data$insulin ),
mass=mean(new.data$mass),
pedigree =mean(new.data$pedigree )),type="response"))
# [1] 0.3591437
```

Listing 10: Πρόβλεψη για την εμφάνιση διαβήτη για μια γυναίκα ηλικίας 35, αποτέλεσμα σε log odds

```
as.numeric(predict(mylogit1 ,
newdata=data.frame(
age=factor(factor('41-50',levels=c('20-30','31-40','41-50','
   50+'))),
glucose =mean(new.data$glucose ),
pressure=mean(new.data$pressure),
triceps =mean(new.data$triceps ),
insulin =mean(new.data$insulin ),
mass=mean(new.data$mass),
pedigree =mean(new.data$pedigree )),type="response"))
# [1] 0.544088
```

Listing 11: Πρόβλεψη για την εμφάνιση διαβήτη για μια γυναίκα ηλικίας 45, αποτέλεσμα σε log odds

```
exp(0.3591437)
#[1] 1.432103
```

Listing 12: Odds Ratio για την εμφάνιση διαβήτη για μια γυναίκα ηλικίας 35

```
exp(0.544088)
#[1] 1.723036
```

Listing 13: Odds Ratio για την εμφάνιση διαβήτη για μια γυναίκα ηλικίας 45

```
exp(0.3591437)/exp(0.544088)
#[1] 0.8311506
```

Listing 14: Odds Ratio εμφάνισης διαβήτη μεταξύ της γυναίκας 35 με την εμφάνιση του διαβήτη για την γυναίκα 45 ετών

ΚΩΔΙΚΑΣ ΕΡΩΤΗΜΑΤΟΣ 4

```
n <- nrow(new.data)
train <- sample(n, n/2)
glm.fit <- glm(formula =diabetes ~ glucose + mass + pedigree + age,
               data = new.data,subset = train, family = binomial)

glm.probs <- predict(glm.fit, new.data[-train, ], type = "response")
glm.pred <- rep("neg", n/2)
glm.pred[glm.probs>0.5] <- "pos"
```

Listing 15: Λογιστικη παλινδρομηση για την πρόβλεψη της μεταβλητής diabetes

```

table(glm.pred)
#  glm.pred
#  neg pos
#  141  55

confTab <- table(glm.pred, diabetes [-train])
confTab

#glm.pred  neg  pos
#     neg  111   30
#     pos   15   40
```

Listing 16: Δημιουργεία Confusion Matric

```

# Sensitivity
((confTab[2,2])/(confTab[1,2]+confTab[2,2]))*100
#[1] 57.14286

#Specificity
(1-(confTab[2,1])/(confTab[1,1]+confTab[2,1]))*100
#[1] 88.09524
```

Listing 17: ΥπολογισμοςΕυαισθησία (Sensitivity) και Ειδικότητα (Specificity)

ΚΩΔΙΚΑΣ ΕΡΩΤΗΜΑΤΟΣ 5

```
1
2 newdata1<- with(new.data,
3    data.frame(pregnant=factor("0-5",levels=c( "0-5" ,"6-10", "
     10+")),glucose=95,pressure=50,
4  triceps=21,insulin=70,mass=30,pedigree=0.5,
5  age=factor('41-50',levels=c('20-30','31-40','41-50','50+'))))
6
7
8  data1<- cbind(newdata1,predict(mylogit1, newdata = newdata1,
     type="link", se=TRUE))
9
10 newdata4<- within(data1 ,{
11   PredictedProb <- plogis(fit)
12   LL <- plogis(fit - (1.96 * se.fit))
13   UL <- plogis(fit + (1.96 * se.fit))
14 })
15 (newdata4)
16 #   pregnant glucose pressure triceps insulin mass pedigree
17 #1      0-5      95       50      21      70   30      0.5
18
19 #    age       fit     se.fit   residual.scale          UL
       LL
20 # 1  41-50  -1.115154  0.4096304              1   0.4225588
     0.1280815
21
22 #   PredictedProb
23 #1     0.2469113
```

Listing 18: Προβλέψη για την Γυναικα 1 και δ.ε 95%

```
1
2  newdata2<- with(new.data,
3     data.frame(pregnant=factor("0-5",levels=c( "0-5" ,"6-10", "
       10+")),glucose=  80,pressure=62,
4   triceps=30,insulin=100,mass=35,pedigree=0.7,
5   age=factor('31-40',levels=c('20-30','31-40','41-50','50+'))))
6
7  data2<- cbind(newdata2,predict(mylogit1, newdata = newdata2,
       type="link", se=TRUE))
8
9   newdata5<- within(data2,{
10    PredictedProb <- plogis(fit)
11    LL <- plogis(fit - (1.96 * se.fit))
12    UL <- plogis(fit + (1.96 * se.fit))
13  })
14  (newdata5)
15
16 #     pregnant  glucose  pressure  triceps  insulin  mass
       pedigree
17 #        0-5       80        62       30         100     35       0.7
18
19 #     age       fit       se.fit     residual.scale
20 #   31-40  -1.84078   0.387889               1
21
22 #   UL          LL   PredictedProb
23 #  0.2534081 0.06907139      0.1369591
```

Listing 19: Προβλέψη για την Γυναικα 2 και 95% δ.ε

```
1
2 newdata3<- with(new.data,
3    data.frame(pregnant=factor("0-5",levels=c( "0-5" ,"6-10", "
     10+")),glucose= 55,pressure=70,
4  triceps=17,insulin= 110,mass=33,pedigree=0.9,
5  age=factor('31-40',levels=c('20-30','31-40','41-50','50+'))))
6
7  data3<- cbind(newdata3,predict(mylogit1, newdata = newdata3,
     type="link", se=TRUE))
8
9  newdata6<- within(data3,{
10   PredictedProb <- plogis(fit)
11   LL <- plogis(fit - (1.96 * se.fit))
12   UL <- plogis(fit + (1.96 * se.fit))
13 })
14 (newdata6)
15
16 #    pregnant  glucose  pressure  triceps  insulin  mass
     pedigree
17 # 1      0-5       55        70       17      110       33       0.9
18
19 # age        fit     se.fit  residual.scale
20 # 31-40 -2.710084 0.4877437              1
21
22 # UL          LL       PredictedProb
23 #0.1475301 0.02493914     0.06238096
```

Listing 20: Προβλέψη για την Γυναικα 3 και 95% δ.ε