Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

Στατιστικά μοντέλα ταξινόμησης με εφαρμογή στην ανίχνευση επιπλοκών στη φυσιολογική ανάπτυξη και την καλή υγεία εμβρύων

Φίλιππος Παρθενόπουλος

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς

Σεπτέμβριος 2023

Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

Στατιστικά μοντέλα ταξινόμησης με εφαρμογή στην ανίχνευση επιπλοκών στη φυσιολογική ανάπτυξη και την καλή υγεία εμβρύων

Φίλιππος Παρθενόπουλος

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς

Σεπτέμβριος 2023

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN

APPLIED STATISTICS

Statistical classification models and an application of them in the detection of complications in the normal development and good health of fetus.

By

Filippos Parthenopoulos

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus Greece

September 2023

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. …….. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική.

Τα μέλη της Επιτροπής ήταν:

- Καθηγητής Μάρκος Κούτρας (Επιβλέπων)

- Καθηγήτρια Γεωργία Βερροπούλου

- Αναπληρωτής Καθηγητής Γεώργιος Τζαβελάς

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και

Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των

γνωμών του συγγραφέα.

Ευχαριστίες

Με την παρούσα διπλωματική εργασία ολοκληρώνεται ένας μεγάλος κύκλος της ζωής μου. Μέσα από το μεταπτυχιακό πρόγραμμα στην Εφαρμοσμένη Στατιστική στο τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς κατάφερα να εμβαθύνω περισσότερο στον τομέα της στατιστικής και των στοχαστικών μαθηματικών τους οποίους είχα γνωρίσει και αγαπήσει στο προπτυχιακό μου τμήμα των Μαθηματικών του Πανεπιστημίου Κρήτης. Στο διάστημα αυτό γνώρισα πολύ σημαντικά και αξιόλογα άτομα τα οποία είτε ήταν καθηγητές οι οποίοι με βοήθησαν πολύ με την κατανόηση και την ένταξη στο αντικείμενο, είτε φοιτητές οι οποίοι με βοήθησαν σημαντικά και πλέον μπορούμε επίσημα να αλληλοχαρακτηριστούμε ως συνάδελφοι. Ένα ιδιαίτερο ευχαριστώ θα ήθελα να πω στον επιβλέποντα καθηγητή μου κ. Μάρκο Κούτρα, για την εμπιστοσύνη που μου έδειξε να ασχοληθώ με το θέμα και για την πολύτιμη βοήθεια που μου παρείχε η οποία σε πολλά σημεία της εργασίας έπαιξε κομβικό ρόλο.

Περίληψη

Στην παρούσα διπλωματική εργασία παρουσιάζεται μια οικογένεια προβλημάτων που είναι γνωστή ως προβλήματα ταξινόμησης και ο τρόπος αντιμετώπισής τους μέσω των στατιστικών μοντέλων ταξινόμησης. Επικεντρωνόμαστε σε τρία διαφορετικά μοντέλα τα οποία θεμελιώνονται τόσο σε θεωρητικό όσο και σε πρακτικό επίπεδο. Μέσω της ενίσχυσής τους με ένα σύνολο τεχνικών και μεθοδολογιών της μηχανικής μάθησης τα μοντέλα αυτά είναι ικανά να αντιμετωπίσουν με επιτυχία προβλήματα από πολλούς διαφορετικούς τομείς όπως αυτοί των χρηματοοικονομικών, της ιατρικής, της επιστήμης δεδομένων και άλλων. Στο τελευταίο κομμάτι της διπλωματικής εργασίας παρουσιάζεται ένα πραγματικό πρόβλημα από τον τομέα της ιατρικής το οποίο αφορά δεδομένα που προκύπτουν από ιατρικές μετρήσεις σε έμβρυα μέσω μίας κλινικής εξέτασης που ονομάζεται καρδιοτοκογράφημα. Σκοπός της διπλωματικής εργασίας είναι η εφαρμογή, η σύγκριση, η αξιολόγηση και η βελτιστοποίηση των μοντέλων που παρουσιάστηκαν αρχικά σε θεωρητικό επίπεδο πάνω σε αυτά τα ιατρικά δεδομένα, ώστε να δούμε ποια από αυτά είναι σε θέση να ανιχνεύουν τυχόν επιπλοκές στην φυσιολογική ανάπτυξη και καλή υγεία του εμβρύου αξιόπιστα.

Abstract

In this thesis, a family of problems known as classification problems and the techniques for addressing them through statistical classification models are presented. We focus on three different models that are based on both theoretical and practical levels. By enhancing them with a set of techniques and methodologies from machine learning, these models are capable of successfully dealing with problems from various domains such as finance, medical, data science, and others. The last part of the thesis presents a real-life problem from field of health sciences, which involves data derived from medical measurements on embryos through a clinical examination called cardiotocography. The purpose of the thesis is to apply, compare, evaluate, and optimize the models that were initially presented at a theoretical level on the medical data, in order to determine which of them are able to detect effectively any complications in the normal development and good health of the fetus.

Περιεχόμενα

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ ΚΑΙ ΣΧΗΜΑΤΩΝ

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή στα προβλήματα και μοντέλα ταξινόμησης.

Ο τομέας της στατιστικής και της επιστήμης δεδομένων αποτελείται από ένα σύνολο θεωρητικά θεμελιωμένων τεχνικών μέσω των οποίων μπορούμε είτε να μάθουμε μέσα από τα δεδομένα είτε να λύσουμε σημαντικά προβλήματα της καθημερινότητας. Μία μεγάλη οικογένεια προβλημάτων είναι τα λεγόμενα προβλήματα ταξινόμησης ή κατηγοριοποίησης. Στα προβλήματα αυτά έχουμε παρατηρήσεις με διάφορα χαρακτηριστικά (*features)* και μία ετικέτα (*label)* αυτών που τα χαρακτηρίζει. Σε τέτοιες περιπτώσεις καλούμαστε να αναπτύξουμε μία τεχνική ,μία μέθοδο ή ένα κανόνα ο οποίος, όταν τροφοδοτείται από τα αντίστοιχα *features* είναι σε θέση να προβλέψει το *label* μέσα από ένα σύνολο προκαθορισμένων τιμών. Προβλήματα τέτοιας φύσης ποικίλουν στην καθημερινότητά μας, τόσο ως προς το είδος του όσο και ως προς τον κλάδο από τον οποίο προέρχονται. Παρακάτω, δίνονται μερικά παραδείγματα.

a) Τομέας χρηματοοικονομικών και τραπεζικής

Ένα πρόβλημα σε μία τράπεζα μπορεί να είναι η έγκριση ενός δανείου ή μίας κάρτας δοθέντος ενός προφίλ πελάτη, που περιλαμβάνει ένα πλήθος χαρακτηριστικών ενός νέου πελάτη ταξινομώντας τον σε καλοπληρωτή ή κακοπληρωτή αντίστοιχα.

b) Τομέας Ιατρικής και Βιοστατιστικής

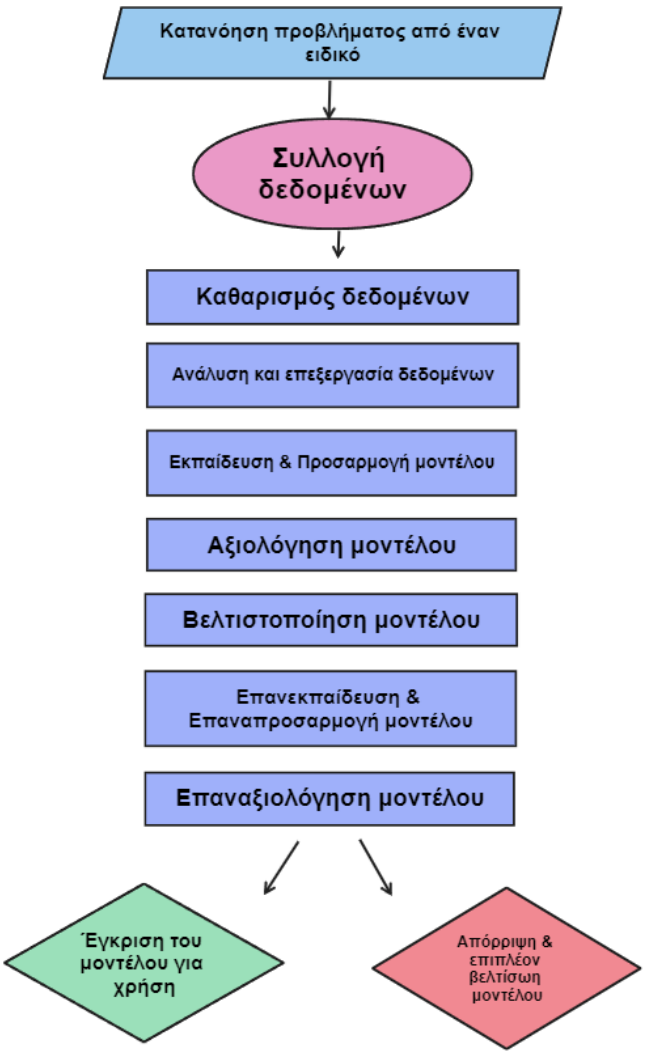
Ένα πρόβλημα σε μία κλινική ή ένα ιατρείο μπορεί να είναι ή διάγνωση μίας ασθένειας δοθέντος μίας σειράς ιατρικών μετρήσεων του εκάστοτε υποψήφιου ασθενή.

c) Τομέας επιστήμης δεδομένων και επεξεργασίας φυσικής γλώσσας.

Μπορούμε να διακρίνουμε μέσω επεξεργασίας φυσικής γλώσσας και μοτίβων που υπάρχουν σε ένα κείμενο που αφορά κριτική για μία παράσταση, για το αν αυτή αποτελεί μία καλή ή κακή κριτική.

Τα μοντέλα ταξινόμησης είναι εκείνα που προβλέπουν την ετικέτα και προσδιορίζουν την απόφαση μας. Το όνομα τους προέρχεται από το γεγονός ότι καλούνται να ταξινομήσουν παρατηρήσεις σε μία από τις διαθέσιμες ετικέτες. Έτσι εμείς με τα δεδομένα που διαθέτουμε αφού τα καθαρίσουμε και τα αναλύσουμε, εκπαιδεύουμε τα μοντέλα όπου με τον όρο αυτό εννοούμε ότι προβαίνουμε στην εκτίμηση των παραμέτρων τους, και στην συνέχεια μπορούμε να τα τροφοδοτήσουμε με χαρακτηριστικά νέων παρατηρήσεων για να πάρουμε τις αντίστοιχες προβλέψεις. Τα μοντέλα ταξινόμησης ανήκουν σε μία ευρύτερη οικογένεια μοντέλων που ονομάζεται μοντέλα επιτηρούμενης μάθησης (*supervised learning*) καθώς γνωρίζουμε εκ των προτέρων τις διαθέσιμες ετικέτες στις οποίες μπορεί να ταξινομηθεί μία νέα παρατήρηση και η ετικέτα αυτή χρησιμοποιείται στο κομμάτι της εκπαίδευσης. Λόγω αυτού εμείς μετά την δημιουργία και την προσαρμογή ενός μοντέλου ταξινόμησης μπορούμε να το αξιολογήσουμε και να το βελτιστοποιήσουμε ευκολότερα μέσα από ένα πλήθος τεχνικών της στατιστικής μηχανικής μάθησης πριν κρίνουμε το μοντέλο ότι είναι ικανό να αντιμετωπίσει επιτυχώς το εκάστοτε

πρόβλημα. Παρακάτω φαίνεται ένα σχήμα που δείχνει τα βήματα κατασκευής ενός μοντέλου συνοπτικά.



Σχήμα 1.1. Βήματα κατασκευής μοντέλου

Στο επόμενο κεφάλαιο θα παρουσιάσουμε κάποιες μαθηματικές έννοιες και ορισμούς τους οποίους θα τους χρειαστούμε για την μαθηματική θεμελίωση των μοντέλων που παρουσιάζονται και μία σειρά από μεθόδους στατιστικής μηχανικής μάθησης οι οποίες χρειάζονται είτε στο πλαίσιο της επεξεργασίας των δεδομένων είτε στο πλαίσιο της βελτίωσης και αξιολόγησης του εκάστοτε μοντέλου. Ακολουθούν τρία κεφάλαια τα οποία παρουσιάζουν τρία διαφορετικά και από τα πιο δημοφιλή μοντέλα ταξινόμησης τέλος παρουσιάζεται η εφαρμογή αυτών σε ένα πραγματικό πρόβλημα από τον τομέα της ιατρικής καθώς και μία μέθοδος που συνδυάζει και τα τρία αυτά μοντέλα για καλύτερη πρόβλεψη.

ΚΕΦΑΛΑΙΟ 2

Μαθηματικές έννοιες και μέθοδοι στατιστικής μηχανικής μάθησης.

2.1 Έννοιες μαθηματικών και στατιστικής

2.1.1 Ορισμός κυρτής συνάρτησης

Έστω $\boldsymbol{S} \subseteq \boldsymbol{R^n}$ ένα κυρτό και μη κενό σύνολο. Μία συνάρτηση ονομάζεται κυρτή στο S , $f: S \to R$ αν για κάθε $\boldsymbol{x_1}, \boldsymbol{x_2} \in \boldsymbol{S}$ και για κάθε $\theta \in [0,1]$ ισχύει, ([20])

$$f(\theta x_1 + (1-\theta)x_2) \leq \theta f(x_1) + (1-\theta) f(x_2).$$

Αντίστοιχα ως αυστηρά κυρτή συνάρτηση στο S θεωρείται κάθε συνάρτηση $f: S \to R$ για την οποία, αν $x_1, x_2 \in S$ με $x_1 \neq x_2$ και για κάθε $\theta \in [0,1]$ ισχύει ότι :

$$f(\theta x_1 + (1-\theta)x_2) < \theta f(x_1) + (1-\theta) f(x_2).$$

2.1.2 Χρήση πολλαπλασιαστών Lagrange για επίλυση προβλήματος βελτιστοποίησης

Έστω ότι έχουμε ένα πρόβλημα βελτιστοποίησης με σκοπό την ελαχιστοποίηση μίας κυρτής συνάρτησης f υπό συγκεκριμένους γραμμικούς περιορισμούς $g_i(\boldsymbol{x}) \geq 0$ με $\boldsymbol{x} \in R^p$ και $i = 1,2, \dots n$

Για κάθε ένα τέτοιο πρόβλημα γραμμικού προγραμματισμού υπάρχει και ένα άλλο πρόβλημα γραμμικού προγραμματισμού το οποίο ονομάζεται το δυικό (dual) του. Το αρχικό πρόβλημα ονομάζεται πρωταρχικό (primal) πρόβλημα.

Για την επίλυση τώρα του πρωταρχικού προβλήματος, δημιουργούμε το πρόβλημα βελτιστοποίησης Lagrange, το οποίο βασίζεται σε αυτό το πρωταρχικό πρόβλημα.

$$\max_{\boldsymbol{a}} \min_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{a})$$

όπου

$$L(\boldsymbol{x}, \boldsymbol{a}) = f(\boldsymbol{x}) - \sum_{i=1}^{n} a_i g_i(\boldsymbol{x}) \text{ με } \boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \dots \alpha_n)', a_i \geq 0 \; \forall \, i$$

Η $L(\boldsymbol{x}, \boldsymbol{a})$ ονομάζεται συνάρτηση Lagrange και αποτελείται από την αρχική συνάρτηση f και από ένα γραμμικό συνδυασμό των γραμμικών περιορισμών $g_i(\boldsymbol{x})$ και των όρων a_i οι οποίοι ονομάζονται πολλαπλασιαστές Lagrange και είναι μη αρνητικές σταθερές. Βλέπουμε ότι το νέο πρόβλημα αυτό αποτελεί ένα πρόβλημα ελαχιστοποίησης ως προς $\boldsymbol{x}$ της συνάρτησης Lagrange για σταθερό $\boldsymbol{a}$, ενώ αποτελεί ένα πρόβλημα μεγιστοποίησης ως προς $\boldsymbol{a}$ για σταθερό $\boldsymbol{x}$. Ως λύση του προβλήματος έχουμε 2 διανύσματα, έστω τα $\boldsymbol{a}^*$ και $\boldsymbol{x}^*$ για τα οποία υπάρχει μοναδική λύση δοθέντος ότι έχουμε υποθέσει ότι η f αποτελεί μία κυρτή συνάρτηση και οι περιορισμοί $g_i(\boldsymbol{x})$ είναι γραμμικοί. Επομένως για το $\boldsymbol{x}^*$ για το οποίο η L θα ελαχιστοποιείται, θα ισχύει ότι

$$\frac{\partial L(\boldsymbol{\alpha}, \boldsymbol{x}^*)}{\partial \boldsymbol{x}} = 0$$

Έστω λοιπόν τώρα ότι η λύση του προβλήματος Lagrange ως προς $\boldsymbol{\alpha}$ να είναι η $\boldsymbol{\alpha}^*$. Έχουμε ότι:

$$\max_{a} \min_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{a}) = L(\boldsymbol{x}^*, \boldsymbol{a}^*) = f(\boldsymbol{x}^*) - \sum_{i=1}^{n} a_i^* g_i(\boldsymbol{x}^*)$$

Η λύση $\boldsymbol{x}^*$ είναι λύση της αρχικής συνάρτησης αν και μόνο αν ισχύουν οι συνθήκες Karush-Kuhn-Tucker (KKT - conditions)

$$\frac{\partial L(\boldsymbol{\alpha}^*, \boldsymbol{x}^*)}{\partial \boldsymbol{x}} = 0 \qquad (2.1.2.1)$$

$$a_i g_i(\boldsymbol{x}^*) = 0 \qquad (2.1.2.2)$$

$$g_i(\boldsymbol{x}^*) \geq 0 \qquad (2.1.2.3)$$

$$a_i \geq 0\,. \qquad (2.1.2.4)$$

Η πρώτη συνθήκη (2.1.2.1) μας διασφαλίζει την στασιμότητα του σημείου $\boldsymbol{x}^*$,η δεύτερη συνθήκη (2.1.2.2) μας διασφαλίζει τον μηδενισμό των γινομένων των γραμμικών περιορισμών και των πολλαπλασιαστών Lagrange. Ακόμα, η τρίτη συνθήκη (2.1.2.3) μας διασφαλίζει ότι οι περιορισμοί του αρχικού προβλήματος ικανοποιούνται ενώ από την τέταρτη (2.1.2.4) έχουμε την αντίστοιχη περίπτωση για την ικανοποίηση των συνθηκών του δυικού προβλήματος ([10][20]). Εν κατακλείδι, με την χρήση πολλαπλασιαστών Lagrange έχουμε την αναγωγή του αρχικού προβλήματος το οποίο ήταν το

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \quad \text{υπό περιορισμούς } g_i(\boldsymbol{x}) \geq 0 \qquad \text{με} \quad \boldsymbol{x} \in R^p \text{ και } i = 1,2, \dots n,$$

στο

$$\max_{\boldsymbol{a}} f'(\boldsymbol{a}) \quad \text{όπου} \quad f'(\boldsymbol{a}) = L(\boldsymbol{a}, \boldsymbol{x}^*) \text{ με } a_i \geq 0 \text{ για } i = 1,2,3, \dots n.$$

2.1.3 Εκθετική οικογένεια κατανομών

Μία κατανομή πιθανότητας λέμε ορίζει μια εκθετική οικογένεια κατανομών όταν η συνάρτηση πιθανότητας (ή πυκνότητας για συνεχής τ.μ.) της κατανομής μπορεί να γραφθεί στη μορφή

$$f(x; \theta, \varphi) = \exp\left[\frac{x\theta - b(\theta)}{\alpha(\varphi)} + c(x, \varphi)\right] \qquad (2.1.3.1)$$

όπου a, b, c να είναι τρείς γνωστές συναρτήσεις ενώ οι θ, φ είναι παράμετροι. Αν το φ είναι γνωστό τότε έχουμε την εκθετική οικογένεια με μία παράμετρο και το θ αναφέρεται ως η κανονική παράμετρος (canonical parameter) της κατανομής. Αν το φ δεν είναι γνωστό, τότε μπορούμε σε

πολλές περιπτώσεις να το θεωρήσουμε σαν μία παράμετρο κλίμακας για την κατανομή, οπότε αποκαλείτε παράγοντας όχλησης (nuisance factor) της κατανομής, ([7]).

Όπου το θ καλείται ως κανονική παράμετρος (canonical parameter) και το φ είναι μία οχληρή παράμετρος γνωστή και ως παράγοντας όχλησης (nuisance factor).

Για παράδειγμα ας θεωρήσουμε ότι $X \sim B(n,p)$. Τότε η συνάρτηση πιθανότητας γράφεται για γνωστό n ως,

$$f_X(x;p) = \binom{n}{x} p^x (1-p)^{n-x} \; \mu\varepsilon \; x = 0,1,2,\ldots\ldots n, 0 < p < 1$$

Και μπορεί εναλλακτικά να γραφτεί ως

$$f_x(x;p) = \exp\left[xlog\left(\frac{p}{1-p}\right) + nlog(1-p) + log\binom{n}{x}\right]$$

Άρα για $\theta = \log\left(\frac{p}{1-p}\right)$, $b(\theta) = nlog(1+e^{\theta})$, $a(\varphi) = 1$, $c(x,\varphi) = log\binom{n}{x}$

Επομένως βλέπουμε ότι η Διωνυμική κατανομή άρα και η κατανομή Bernoulli η οποία είναι μία ειδική περίπτωση αυτής για n=1, μπορούν να αναπαρασταθούν όπως η (2.1.3.1) άρα ανήκουν στην εκθετική οικογένεια κατανομών.

2.1.4 Εκτιμητές μέγιστης πιθανοφάνειας

Έστω $\boldsymbol{x} = (x_1, x_2, \ldots\ldots x_n)$ ένα τυχαίο δείγμα από ένα πληθυσμό με συνάρτηση/πυκνότητα πιθανότητας $f(\boldsymbol{x}\,;\theta)$ όπου θ η παράμετρος ή το διάνυσμα παραμέτρων που καλούμαστε να εκτιμήσουμε και $\theta \in \Theta$ με Θ να είναι ο παραμετρικός χώρος. Η από κοινού συνάρτηση πυκνότητας πιθανότητας του δείγματος

$$L(\theta) = L(\theta;\boldsymbol{x})$$

η οποία θεωρείται συνάρτηση της παραμέτρου θ ονομάζεται συνάρτηση πιθανοφάνειας. Επομένως η συνάρτηση πιθανοφάνειας θα ισούται με

$$L(\theta) = L(\theta;\boldsymbol{x}) = L(\theta; x_1, x_2 \ldots\ldots x_n) = \prod_{i=1}^{n} f(x_i;\theta)$$

Ως εκτιμητής μεγίστης πιθανοφάνειας θεωρείται εκείνη η στατιστική συνάρτηση $\hat{\theta}(\boldsymbol{x})$ για την οποία η συνάρτηση πιθανοφάνειας μεγιστοποιείται δηλαδή

$$L(\hat{\theta}) = \sup_{\theta \in \Theta} L(\theta)$$

Οι εκτιμητές μεγίστης πιθανοφάνειας αποτελούν αν όχι την πιο δημοφιλή μία από τις πιο δημοφιλείς μεθόδους εκτίμησης. Ο λόγος είναι ότι εκτός της εύκολης εύρεσης αυτών, έχουν και μία σειρά από καλές ασυμπτωτικές ιδιότητες τις οποίες εκμεταλλευόμαστε.

Μερικές από τις ιδιότητες είναι ότι καθώς το δείγμα τείνει στο άπειρο οι εκτιμητές μεγίστης πιθανοφάνειας ακολουθούν ασυμπτωτικά την κανονική κατανομή με μέση τιμή την παράμετρο που καλούνται να εκτιμήσουν και διακύμανση το κάτω φράγμα της ανισότητας Cramer-Rao. Αυτό τους κάνει ασυμπτωτικά κανονικούς, ασυμπτωτικά αμερόληπτους και ακόμα, καθώς το κάτω φράγμα της ανισότητας Cramer-Rao είναι η ελάχιστη τιμή που μπορούμε να επιτύχουμε, εξασφαλίζουμε για μεγάλο δείγμα και την βέλτιστη διακύμανση. Ακόμα είναι και συνεπείς, το οποίο σημαίνει ότι καθώς το δείγμα αυξάνει η εκτιμήτρια συνάρτηση συγκλίνει στην πραγματική τιμή με πιθανότητα 1, ([3][4][7]).

Δηλαδή έχουμε ότι

$$\hat{\theta} \sim N(\theta, I(\theta)^{-1})$$

όπου

$$I(\theta) = E[-l''(\theta)] \quad \text{πληροφορία κατά Fisher}$$

και επιπλέον ισχύει

$$\lim_{n\to\infty} P(\left|\hat{\theta} - \theta\right| < \varepsilon) = 1\,, \quad \forall\, \varepsilon > 0.$$

2.1.5 Συναρτήσεις σύνδεσης

Έστω μία γραμμική συνάρτηση πρόβλεψης

$$\eta_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

Ως συνάρτηση σύνδεσης θεωρείται εκείνη η συνάρτηση g η οποία συνδέει την αναμενόμενη τιμή της απόκρισης με την παραπάνω συνάρτηση πρόβλεψης, δηλαδή

$$\eta_i = g(\mu_i) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

Η g πρέπει να είναι μία συνάρτηση μονότονη και διαφορίσιμη. Όταν ισχύει ότι

$$g = (b')^{-1}$$

όπου b είναι η συνάρτηση που αντιστοιχεί στην παράγραφο 2.1.3 τότε έχουμε την περίπτωση της κανονικής συνάρτησης σύνδεσης *(*canonical link function*)*. Για περισσότερες λεπτομέρειες βλ., ([7]).

2.2 Μέθοδοι στατιστικής μηχανικής μάθησης

2.2.1 Πίνακας συγχύσεως

Ο πίνακας συγχύσεως αποτελεί την πρώτη εικόνα που έχει κάποιος μετά την εφαρμογή ενός μοντέλου ταξινόμησης. Είναι ένας πίνακας όπου ως γραμμές έχει τις διάφορες πραγματικές

ετικέτες μίας παρατήρησης ενώ στις στήλες έχει τις αντίστοιχες προβλεφθείσες ετικέτες που εκτιμούνται από το μοντέλο που χρησιμοποιήσαμε για ταξινόμηση. Καθώς η απόκριση μας αποτελεί μία ποιοτική τυχαία μεταβλητή πρέπει να την αντιμετωπίσουμε ως παρόντα. Οι παράγοντες λαμβάνουν ως τιμές τις διάφορες κατηγορίες ή επίπεδα μίας ποιοτικής τυχαίας μεταβλητής. Για περιπτώσεις που έχουμε δίτιμα δεδομένα επιλέγουμε το ένα επίπεδο της ετικέτας και το θεωρούμε ως επιτυχία του παράγοντα, Στην συνέχεια είτε όταν προβλέπουμε είτε όταν γνωρίζουμε ότι μία παρατήρηση έχει αυτόν τον παράγοντα το ορίζουμε ως θετικό και διαφορετικά αρνητικό. Έτσι ανάλογα με το αν συμφωνεί η πρόβλεψη με την πραγματικότητα ή όχι έχουμε τις περιπτώσεις: αληθές θετικό – true positive (TP) , αληθές αρνητικό – true negative (TN) , ψευδές θετικό – false positive (FP) και ψευδές αρνητικό – false negative (FN).

Σε περιπτώσεις όπου έχουμε παραπάνω από 2 επίπεδα και θέλουμε να κάνουμε μία ανάλυση ως προς ένα συγκεκριμένο επίπεδο, το ορίζουμε αυτό ως θετικό, τα υπόλοιπα ως αρνητικά και συμπεριφερόμαστε με παρόμοιο τρόπο μόνο που μπορούμε να παρατηρούμε στις περιπτώσεις όπου έχουμε ασυμφωνία πρόβλεψης με την πραγματικότητα πόση διαφορά είχε η ασυμφωνία μας. Παρακάτω φαίνεται μία κλασσική εικόνα ενός πίνακα συγχύσεως για την περίπτωση της δίτιμης απόκρισης, βλ.([21][22])

	Predicted Positive	**Predicted Negative**
True Positive	TP	FN
True Negative	FP	TN

Πίνακας 2.2.1.1 Παράδειγμα Πίνακα συγχύσεως

2.2.2 Μέτρα αξιολόγησης μοντέλων ταξινόμησης – Εξωτερικά μέτρα

Μετά την δημιουργία ενός μοντέλου ταξινόμησης καλούμαστε να το αξιολογήσουμε, με σκοπό να έχουμε μία εικόνα της απόδοσης αυτού σε δεδομένα που εκπαιδεύτηκε αλλά κυρίως σε ξένα για το μοντέλο δεδομένα. Τα εξωτερικά μέτρα ή αλλιώς μέτρα αξιολόγησης του μοντέλου, είναι ένα σύνολο μετρικών οι οποίες καλούνται να αξιολογήσουν μοντέλα και τεχνικές ταξινόμησης. Τα μέτρα αυτά προκύπτουν μέσα από τον πίνακα συγχύσεως. Παρακάτω λοιπόν ορίζονται μερικά από τα πιο δημοφιλή εξωτερικά μέτρα, βλ. ([21][22]).

- $\boldsymbol{Precision} = \frac{TP}{TP+FP}$
- $\boldsymbol{Recall} = Sensitivity = \frac{TP}{TP+FN}$
- $\boldsymbol{F1-score} = \frac{2Precision*Recall}{Precision+Recall}$
- $\boldsymbol{Accuracy} = \frac{TP+TN}{TP+FN+TN+FP}$

Το Precision μας δείχνει το ποσοστό των σωστών θετικών ταξινομήσεων ανάμεσα σε όλες τις θετικές προβλέψεις. Το Recall ή αλλιώς ευαισθησία ($Sensitivity$) μας επιστρέφει το ποσοστό των θετικών προβλέψεων ανάμεσα σε όλες τις πραγματικά θετικές περιπτώσεις. Το F1-score αποτελεί των αρμονικό μέσο των Recall & Precision με σκοπό να μας δείχνει μια πιο συνολική

εικόνα της απόδοσης του μοντέλου. Τέλος accuracy αποτελεί το ποσοστό των συνολικών σωστών ταξινομήσεων.

2.2.3 Τεχνικές διαχωρισμού του συνόλου των δεδομένων

2.2.3.1 Train-test split

Όπως αναφέραμε και στην Παράγραφο 2.2.2, όταν καλούμαστε να αξιολογήσουμε ένα μοντέλο καταλαβαίνουμε ότι αν έχουμε ένα σύνολο δεδομένων και το χρησιμοποιήσουμε ολόκληρο και την εκπαίδευση του μοντέλου, και για την αξιολόγηση του φυσικό είναι να υπάρχει υπερεκτίμηση της ακρίβειας. Αυτό συμβαίνει καθώς οι διάφοροι παράμετροι του μοντέλου θα έχουν εκτιμηθεί χρησιμοποιώντας ως δείγμα τα δεδομένα αυτά. Έτσι μία φαινομενικά καλή απόδοση με βάση τα εξωτερικά μέτρα που θα χρησιμοποιήσουμε ίσως να μην είναι αντιπροσωπευτική.

Για τον λόγο αυτό μία συνήθης τεχνική στην στατιστική μηχανική μάθηση είναι ο διαχωρισμός του συνόλου δεδομένων σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου. Στην συνέχεια, χρησιμοποιούμε ένα μέρος από τα δεδομένα μας για να εκπαιδεύσουμε το μοντέλο και αφήνουμε ένα μέρος από το δείγμα μας για να δοκιμάσουμε και να ελέγξουμε το μοντέλο σε δεδομένα που δεν έχει δει. Με την τεχνική αυτή αποφεύγουμε και την λανθασμένη εικόνα που μπορεί να έχουμε για το μοντέλο αναφορικά με την απόδοσή του αλλά και προβλήματα υπερπροσαρμογής.

2.2.3.2 Cross-Validation

Σε πολλά από τα μοντέλα της μηχανικής μάθησης και της στατιστικής, υπάρχουν παράμετροι οι οποίοι καθορίζονται από τον αναλυτή ανάλογα με την περίπτωση που εξετάζεται. Για τον λόγο αυτό ο εκάστοτε αναλυτής πρέπει να επιλέξει μία τιμή για την εκάστοτε παράμετρο που θα επηρεάσει το σύνολο της απόδοσης του μοντέλου. Μία συνήθης τεχνική επιλογής τιμής τέτοιων παραμέτρων είναι αυτή της διασταυρωμένης επικύρωσης γνωστή και ως Cross-Validation (CV).

Ο τρόπος που λειτουργεί αυτή η τεχνική είναι ο διαχωρισμός του συνόλου εκπαίδευσης σε k ξένα ανά δύο υποσύνολα. Στην συνέχεια αν εμείς επιθυμούμε να κάνουμε tunning σε μία παράμετρο, δηλαδή να βρούμε την τιμή αυτή που βελτιστοποιεί την απόδοση του μοντέλου, διαλέγουμε ως προς ποια μετρική από τα εξωτερικά μέτρα θα το κάνουμε.

Έστω λοιπόν ότι θέλουμε να κουρδίσουμε (tunning) την παράμετρο β ως προς το accuracy. Διαλέγουμε ένα πλήθος διαφορετικών τιμών για την παράμετρο β και προσαρμόζουμε k μοντέλα εκπαιδευμένα από τα $k-1$ υποσύνολα και ελεγμένα από τα αντίστοιχα υποσύνολα που δεν έχουν συμπεριληφθεί στην εκπαίδευση, για την κάθε διαφορετική τιμή της β. Στην συνέχεια υπολογίζουμε τον μέσο όρο του accuracy όλων των k μοντέλων για κάθε τιμή του β και επιλέγουμε εκείνο το οποίο αντιστοιχεί στο μεγαλύτερο accuracy κατά μέσο όρο.

2.2.4 Μη ισορροπημένα δεδομένα και η μέθοδος του Under-sampling

Πολλές φορές όταν αντιμετωπίζουμε προβλήματα ταξινόμησης ερχόμαστε αντιμέτωποι με το πρόβλημα των μη ισορροπημένων ομάδων της ετικέτας μας. Σε περιπτώσεις όπου ένα επίπεδο της ετικέτας μας υπερεκπροσωπείται είτε υποεκπροσωπείται μπορεί να δημιουργηθούν πολλά προβλήματα. Για παράδειγμα ας θεωρήσουμε ένα πρόβλημα όπου έχουμε 10.000 παρατηρήσεις με μία ετικέτα που παίρνει τις τιμές {0,1}. Ακόμα υποθέτουμε ότι οι 1500 παρατηρήσεις παίρνουν την ετικέτα 0 ενώ οι 8500 την ετικέτα 1 .Αν εμείς δημιουργήσουμε ένα μοντέλο το οποίο ανεξαρτήτως των χαρακτηριστικών μίας παρατήρησης αυτό ταξινομεί την κάθε παρατήρηση στην κατηγορία 1 θα είχαμε ως αποτέλεσμα ένα μοντέλο με accuracy=85% το οποίο ενώ φαινομενικά από το νούμερο φαίνεται αποδοτικό δεν είναι. Υπάρχουν διάφορες τεχνικές διόρθωσης αυτού του προβλήματος, μία από αυτές είναι η μέθοδος under-sampling. Η τεχνική αυτή από την ομάδα η οποία αποτελεί την πλειοψηφία στο dataset, αφαιρεί ένα μέρος των παρατηρήσεων με σκοπό είτε να φέρει την ισορροπία στα δεδομένα μας είτε να μειώσει την μεροληψία ως προς αυτήν την ομάδα. Το καινούργιο πλήθος της ισορροπημένης ομάδας καθορίζεται από τον αναλυτή. Συχνά χρησιμοποιείται είτε ο μέσος όρος των υπόλοιπων ομάδων, είτε το άθροισμα των υπόλοιπων ομάδων είτε γενικότερα μία συνάρτηση των διαφορετικών πληθών των ομάδων.

2.2.5 Ανάλυση σε κύριες συνιστώσες

Ένα σύνηθες πρόβλημα στον κλάδο της στατιστικής είναι η οπτικοποίηση των δεδομένων μας. Στην πλειονότητα των περιπτώσεων το σύνολο των μεταβλητών που αναλύουμε είναι μεγαλύτερο του 3 με αποτέλεσμα να μην μπορούμε να προβούμε σε ακριβή οπτικοποίηση για το σύνολο των δεδομένων μας. Για τον λόγο αυτό έχουν προταθεί πολλές τεχνικές οι οποίες είναι γνωστές και ως μέθοδοι μείωσης διαστάσεων.

Η ανάλυση των κύριων συνιστωσών ή αλλιώς PCA αποτελεί την πιο δημοφιλή μέθοδο μείωσης διαστάσεων η οποία έχει σκοπό να εκμεταλλευτεί τυχόν συσχετίσεις ανάμεσα στα δεδομένα μας και να μπορέσει να δημιουργήσει λίγες νέες μεταβλητές οι οποίες αντιπροσωπεύουν το σύνολο των δεδομένων μας σε ένα ικανοποιητικό βαθμό. Ο τρόπος που γίνεται αυτό είναι με την εύρεση των ιδιοτιμών του πίνακα $Z^T Z$, όπου Z είναι ο πίνακας των κεντρικοποιημένων δεδομένων μας, και παίρνοντας με φθίνουσα σειρά αυτές τις ιδιοτιμές βρίσκουμε τις αντίστοιχες κύριες συνιστώσες μέσω του τύπου

$$y_i = u_{i1}X_1 + u_{i2}X_2 \ldots\ldots + u_{ip}X_p$$

όπου τα u_{ij} αποτελούν τις συντεταγμένες του αντίστοιχου μοναδιαίου ιδιοδιανύσματος. Το ποσοστό της μεταβλητότητας που εξηγεί η κάθε κύρια συνιστώσα δίνεται από το λόγο της αντίστοιχης ιδιοτιμής προς το συνολικό άθροισμα αυτών με αποτέλεσμα να μπορούμε να συγκεντρώσουμε μεγάλο μέρος της μεταβλητότητας στις πρώτες μόνο κύριες συνιστώσες. Σε αρκετές περιπτώσεις μέχρι και τις πρώτες τρεις, οπότε μπορούμε να επιτύχουμε την οπτικοποίηση των δεδομένων μας έχοντας ερμηνεύσει ένα ικανοποιητικό ποσοστό της μεταβλητότητας τους, για περισσότερες λεπτομέρειες βλ., ([2]).

2.2.6 Μετασχηματισμός δεδομένων

Πολλές φορές τα δεδομένα μας πριν τα χρησιμοποιήσουμε σε κάποιο μοντέλο ή μία άλλη στατιστική μέθοδο είναι καλό να τα επεξεργαζόμαστε πρώτα. Για τα ποσοτικά δεδομένα το πιο σύνηθες είναι η τυποποίηση τους. Ο λόγος που γίνεται αυτό είναι ότι πολλές φορές η κλίμακα στην οποία κινείται μία μεταβλητή διαφέρει σημαντικά από τις άλλες. Μέσω της τυποποίησης καταφέρνουμε να ξεπεράσουμε αυτό το εμπόδιο και πλεόν η σημαντικότητα της κάθε μεταβλητής σε ένα μοντέλο να μην επηρεάζεται από αυτόν τον παράγοντα. Η τυποποίηση γίνεται μετά από την εύρεση των αμερόληπτων εκτιμητών της αναμενόμενης τιμής και διακύμανσης της κάθε μεταβλητής.

Έστω ότι για μια τυχαία μεταβλητή X έχουμε συγκεντρώσει ένα δείγμα $x_1, x_2, \dots . x_n$. Θεωρούμε την αμερόληπτη εκτιμήτρια της αναμενόμενης τιμής

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

και την αμερόληπτη εκτιμήτρια της διακύμανσης

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}.$$

Στην συνέχεια προβαίνουμε στην τυποποίηση μέσω του τύπου

$$z_i = (x_\iota - \bar{x})/s.$$

Αυτό το κάνουμε για κάθε ποσοτική μεταβλητή που έχουμε στα δεδομένα μας.

Υπάρχει όμως κάτι που πρέπει να διευκρινιστεί. Το δείγμα που χρησιμοποιούμε για την εκτίμηση των παραμέτρων είναι αυτό του συνόλου εκπαίδευσης. Στην συνέχεια, έχοντας εκτιμήσει της παραμέτρους αυτές, τυποποιούμε τα δεδομένα ελέγχου με τις εκτιμηθήσες παραμέτρους από τα δεδομένα εκπαίδευσης και δεν προβαίνουμε σε νέα εκτίμηση.

Ότι αναφέρθηκε παραπάνω αφορά ποσοτικές μεταβλητές. Με διαφορετικό τρόπο στην περίπτωση ποιοτικών μεταβλητών χρησιμοποιούμε τις λεγόμενες δείκτριες συναρτήσεις που είναι γνωστές ως ψευδομεταβλητές ή dummy variables. Πιο συγκεκριμένα το πρώτο βήμα είναι να ορίσουμε ένα επίπεδο της αντίστοιχης ποιοτικής μεταβλητής ως επίπεδο αναφοράς (*baseline*) ενώ για τα υπόλοιπα επίπεδα αυτής ορίζουμε μία αντίστοιχη ψευδομεταβλητή. Αν μία παρατήρηση ανήκει σε ένα από τα επίπεδα εκτός αυτού του επιπέδου αναφοράς, θέτουμε όλες τις ψευδομεταβλητές ίσες με το μηδέν εκτός την αντίστοιχης του επιπέδου που ανήκει.

Για παράδειγμα αν έχουμε μία μεταβλητή X που περιγράφει το φύλο σε μορφή {άνδρας, γυναίκα}. Επιλέγουμε μία κατηγορία ως επίπεδο αναφοράς (π.χ. τους άνδρες) και ορίζουμε την ψευδομεταβλητή

$$X' = \begin{cases} 1 \; \alpha\nu \; \epsilon\acute{\iota}\nu\alpha\iota \; \gamma\upsilon\nu\alpha\acute{\iota}\kappa\alpha \\ 0 \; \alpha\nu \; \epsilon\acute{\iota}\nu\alpha\iota \; \acute{\alpha}\nu\delta\rho\alpha\varsigma. \end{cases}$$

Σε περιπτώσεις που έχουμε παραπάνω από 2 επίπεδα για παράδειγμα μεταβλητή X που παίρνει τις τιμές {Α,Β,Γ}, επιλέγουμε ένα επίπεδο αναφοράς π.χ. το Α και χρησιμοποιούμε 2 ψευδομεταβλητές για τα άλλα δύο επίπεδα, πιο συγκεκριμένα

$$X_{1B} = \begin{cases} 1 \text{ αν είναι } B \\ 0 \text{ αν όχι } B \end{cases}, X_{1\Gamma} = \begin{cases} 1 \text{ αν είναι } \Gamma \\ 0 \text{ αν όχι } \Gamma. \end{cases}$$

Γενικά για μία ποιοτική μεταβλητή με k επίπεδα χρειαζόμαστε k-1 ψευδομεταβλητές.

2.2.7 Μέθοδοι επιλογής χαρακτηριστικών

Πολλές φορές στα πλαίσια δημιουργίας ενός για τα διαθέσιμα δεδομένα, χρειάζεται το υποσύνολο εκείνων των χαρακτηριστικών που είτε δίνουν την καλύτερη απόδοση ως προς μία μετρική είτε δίνουν την καλύτερη απόδοση σε σχέση με το πλήθος των μεταβλητών που χρειάζεται εμείς να το τροφοδοτήσουμε. Σε προβλήματα που έχουμε πολλές επεξηγηματικές μεταβλητές καλούμαστε να επιλέξουμε τα χαρακτηριστικά εκείνα τα οποία είτε βελτιώνουν την απόδοση του μοντέλου είτε την διατηρούν σε ένα ικανοποιητικό επίπεδο, κρατώντας παράλληλα το πλήθος των επεξηγηματικών μεταβλητών εντός ενός ορίου. Παρακάτω παρουσιάζονται 2 κλασσικές τεχνικές επιλογής χαρακτηριστικών.

a. Forward Selection

Με την μέθοδο forward selection ή αλλιώς προς τα εμπρός επιλογή ο αναλυτής επιλέγει μία μετρική η οποία καθορίζει την απόδοση του μοντέλου. Ο αναλυτής αρχίζει με το κενό μοντέλο και στην αρχή προσθέτει την μεταβλητή εκείνη που μεγιστοποιεί από μόνη της την μετρική αυτή. Στην συνέχεια προσθέτει την αμέσως επόμενη και είτε σταματάει αφού ξεπεράσει ένα όριο μεταβλητών που έχει ορίσει για το μοντέλο είτε όταν καθώς προσθέτοντας καινούργιες μεταβλητές η απόδοση του μοντέλου αυξάνεται πολύ λίγο.

b. Backward Elimination

Με την μέθοδο backward elimination ή προς τα πίσω απαλοιφή ο αναλυτής και πάλι επιλέγει έναν δείκτη όπου ως προς τον οποίο θέλει να βελτιστοποιήσει το μοντέλο του. Ο αναλυτής σε αυτήν την περίπτωση ξεκινάει με το πλήρες μοντέλο που περιλαμβάνει όλο το σύνολο των μεταβλητών. Στο πρώτο βήμα αρχίζει και αφαιρεί είτε μεταβλητές που δεν ικανοποιούν κάποια στατιστική ή μαθηματική συνθήκη είτε μεταβλητές για τις οποίες κατά την αφαίρεσή τους η απόδοση του μοντέλου η μειώνεται αμελητέα. Συνεχίζει την διαδικασία μέχρι είτε να φτάσει στο πλήθος των μεταβλητών που επιθυμεί είτε όταν αρχίζει να χάνει σημαντική πληροφορία ως προς την απόδοση του μοντέλου.

ΚΕΦΑΛΑΙΟ 3

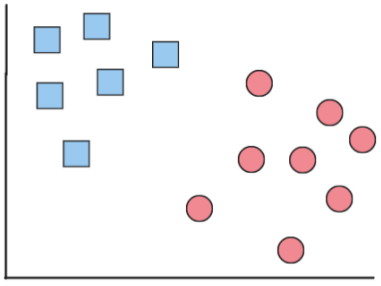
Μηχανές διανυσμάτων υποστήριξης

3.1 Εισαγωγή

Οι μηχανές διανυσμάτων υποστήριξης ή αλλιώς SVM (*support vectors machines*) αποτελούν μία μέθοδο μηχανικής μάθησης η οποία εντάσσεται στην κατηγορία της επιτηρούμενης μάθησης (*supervised learning*) και έχει εφαρμογές σε προβλήματα παλινδρόμησης και ταξινόμησης. Στο πλαίσιο αυτής της μελέτης θα επικεντρωθούμε στο κομμάτι της ταξινόμησης. Η κεντρική ιδέα της μεθόδου είναι η εύρεση ενός υπερεπιπέδου το οποίο διαχωρίζει τα δεδομένα μας με την μέθοδο του μέγιστου περιθωρίου. Η μέθοδος εφαρμόζεται για την περίπτωση των δύο κλάσεων και εύκολα γενικεύεται για πλήθος κλάσεων μεγαλύτερου του δύο. Η μέθοδος αυτή αρχικά ξεκίνησε με την περίπτωση των γραμμικά διαχωρίσιμων δεδομένων (*Hard Margin SVM*) όμως η φήμη και η διάδοση της μεθόδου επεκτάθηκε σημαντικά όταν κατάφερε να εντάξει και τεχνικές που μπορούσαν να αντιμετωπίσουν και περιπτώσεις που είτε η απλή τεχνική των γραμμικά διαχωρίσιμων δεδομένων δεν απέδιδε ικανοποιητικά και έπρεπε να καταφύγουμε σε τεχνικές όπως αυτή του χαλαρού περιθωρίου SVM (*Soft margin SVM),* είτε σε τεχνικές όπου καταφεύγαμε σε μεγαλύτερες διαστάσεις για να ξεπεράσουμε τον μη γραμμικό διαχωρισμό, το γνωστό και ως τέχνασμα του πυρήνα (*Kernel Trick*). Για περισσότερες λεπτομέρειες σχετικά με όσα παρουσιάζονται ο ενδιαφερόμενος αναγνώστης παραπέμπεται στα ([9][10][11][17][20][24]).

3.2 Hard margin SVM

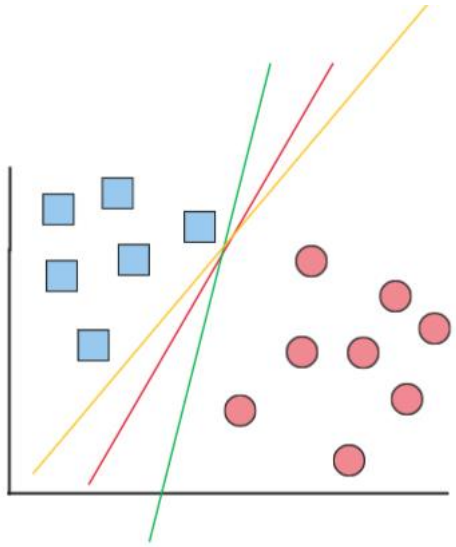
Έστω ότι έχουμε ένα σύνολο δεδομένων της μορφής $\{\boldsymbol{x_i}, y_i\}$ με $\boldsymbol{x_i} \in R^2$ και $y \in \{-1,1\}$. Αρχικά υποθέτουμε ότι τα δεδομένα μας είναι γραμμικά διαχωρίσιμα και αναπαρίστανται από το παρακάτω γράφημα.



Σχήμα 3.2.1.1 Αρχικά δεδομένα

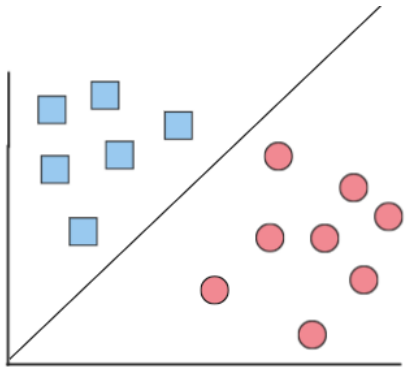
Αναζητούμε έναν κανόνα ή μία μεθοδολογία σύμφωνα με την οποία θα διαχωρίσουμε τα δεδομένα μας και θα κατηγοριοποιούμε νέες παρατηρήσεις. Καθώς βρισκόμαστε στον R^2 ο πιο λογικός

τρόπος διαχωρισμού των δεδομένων είναι μία ευθεία. Το ερώτημα είναι ποια από τις πολλές ευθείες που μπορούν να επιτύχουν αυτόν τον διαχωρισμό είναι η βέλτιστη.

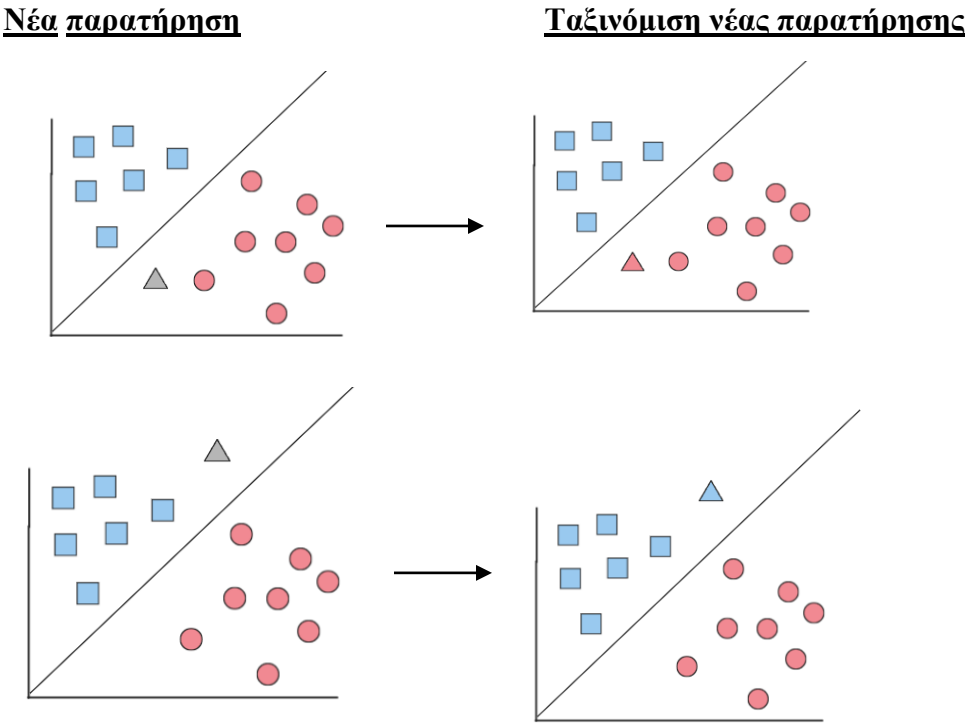


Σχήμα 3.2.1.2 Δεδομένα με πιθανές ευθείες διαχωρισμού

Όπως βλέπουμε στο Σχήμα 3.2.1.2 τα δεδομένα μας μπορούν διαχωριστούν με πολλές ευθείες, εμείς θέλουμε ένα κριτήριο σύμφωνα με το οποίο μία ευθεία θεωρείται καλύτερη από μία άλλη ως προς την κατηγοριοποίηση που προσφέρει ώστε να μπορέσουμε τελικά να βρούμε την βέλτιστη. Ακόμα για την επιλογή της ευθείας θα πρέπει να προσδιορίσουμε και τον τρόπο χρήσης αυτής ως κατηγοριοποιητή. Εμείς καθώς η ευθεία διαχωρίζει τις ομάδες μας, επιθυμούμε όταν έρχεται μία νέα παρατήρηση αναλόγως με την θέση του καινούργιου αυτού σημείου ως προς την ευθεία να το κατατάσσουμε σε μία από τις 2 ομάδες αντίστοιχα. Για παράδειγμα, αν επιλέξουμε την ευθεία που φαίνεται στο παρακάτω Σχήμα 3.2.1.3 και έχουμε μία καινούργια παρατήρηση αναλόγως με την θέση της ως προς την ευθεία θα τα ταξινομηθεί αντίστοιχα όπως φαίνεται στην Σχήμα 3.2.1.4.

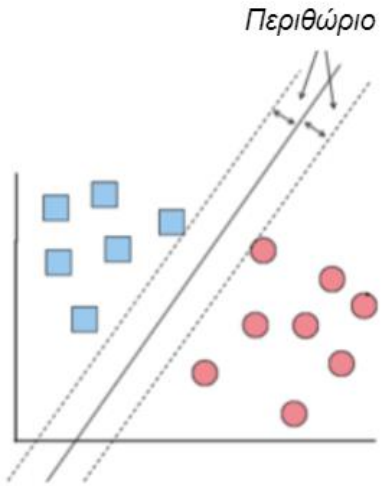


Σχήμα 3.2.1.3 Δεδομένα με πιθανές ευθείες διαχωρισμού



Σχήμα 3.2.1.4 Ταξινόμηση νέας παρατήρησης ανάλογα με την θέση ως προς την ευθεία

Ως βέλτιστη ευθεία θεωρείται εκείνη του μέγιστου περιθωρίου. Η ευθεία αυτή ονομάζεται έτσι καθώς μεγιστοποιεί το περιθώριο που έχει από τις δύο κατηγορίες όπως φαίνεται παρακάτω στο Σχήμα 3.2.1.5

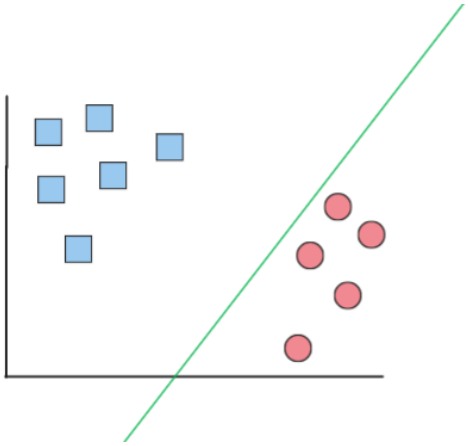


Σχήμα 3.2.1.5 Βέλτιστη ευθεία

Γενικότερα όταν προσδιορίζουμε μία ευθεία καταλαβαίνουμε ότι όσο πιο μακριά είμαστε από την μία ομάδα τόσο πιο εύκολα θα κατηγοριοποιούμε νέες παρατηρήσεις σε αυτήν ομάδα

καθώς καταλαμβάνει μεγαλύτερο συνολικό χώρο στο επίπεδο συγκριτικά με την άλλη. Για παράδειγμα στο παρακάτω Σχήμα 3.2.1.6 φαίνεται μία περίπτωση μεροληψίας ως προς την μία ομάδα.

Ευκολότερη ταξινόμηση για νέες μπλέ παρατηρήσεις



Σχήμα 3.2.1.6 Μεροληπτική ευθεία ως προς μία ομάδα

Επομένως βλέπουμε ότι όσο μεγαλύτερο περιθώριο υπάρχει από την μία ομάδα τόσο πιο ασφαλές είναι να κατηγοριοποιήσουμε σωστά μία νέα παρατήρηση ως προς αυτήν την ομάδα. Άρα η ευθεία του μέγιστου περιθωρίου έχει την ιδιότητα και το πλεονέκτημα σε σχέση με όλες τις άλλες ευθείες, ότι στον ερχομό μίας νέα παρατήρησης μειώνεται η πιθανότητα να κατηγοριοποιηθεί λάθος η καινούργια παρατήρηση. Στο παρακάτω σχήμα φαίνεται η χρήση της βέλτιστης ευθείας ως κατηγοριοποιητής.

Σχήμα 3.2.1.7 Ταξινόμηση μέσω της βέλτιστης ευθείας

Είναι σημαντικό σημειώσουμε εδώ ότι η προσδιορισμός της ευθείας επηρεάζεται άμεσα από τις πιο κοντινές παρατηρήσεις, ενώ οι απομακρισμένες παρατηρήσεις δεν επηρεάζουν την θέση και την κλίση της ευθείας.

3.3 Μαθηματικός προσδιορισμός βέλτιστου υπερεπιπέδου.

Αρχικά αφού ψάχνουμε τον προσδιορισμό ενός υπερεπιπέδου, γνωρίζουμε ότι θα περιγράφεται από την ακόλουθη εξίσωση:

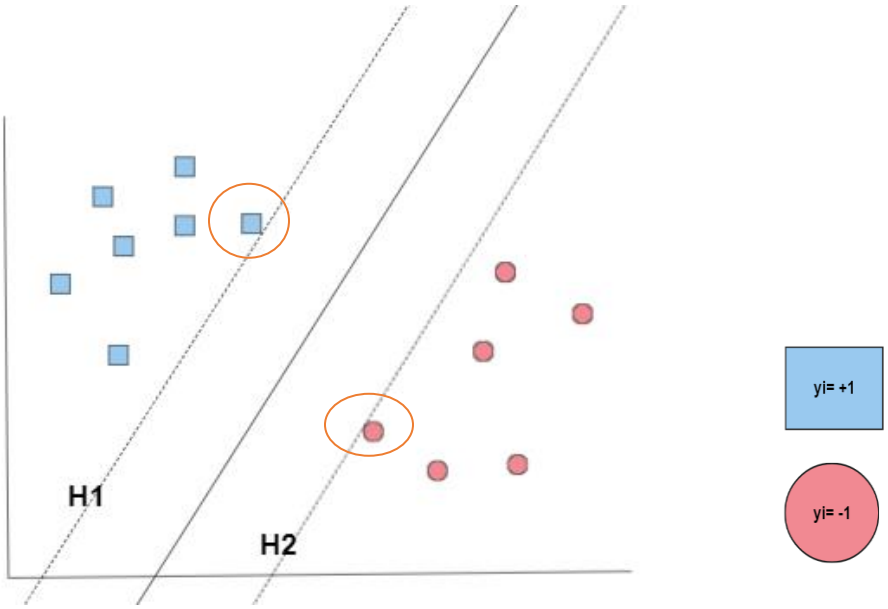
$$w^T x + b = 0$$

όπου $\boldsymbol{w} = (w_1, w_2, w_3, \dots \dots w_n)$ είναι το κάθετο διάνυσμα στο επίπεδο, $\boldsymbol{b}$ να είναι το κατώφλι και $\boldsymbol{x} = (x_1, x_2, x_3, \dots \dots x_n)$ είναι οι παρατηρήσεις μας που χρησιμοποιούμε για εκπαίδευση και μελλοντική ταξινόμηση (στο παράδειγμα μας κινούμαστε στον R^2).

Στην συνέχεια ορίζουμε τα υπερεπίπεδα τα οποία προσδιορίζονται από τα γειτονικά σημεία της κάθε ομάδας τα οποία φαίνονται στο Σχήμα 3.3.1.8. Έστω λοιπόν , H_1 & H_2 τέτοια ώστε

$$H_1: w^T x_i + b = +1$$

$$H_2: w^T x_i + b = -1.$$



Σχήμα 3.3.1.8 Γράφημα με τα H_1 & H_2

Άρα ως κανόνα ταξινόμησης για τον οποίο έχουμε μηδενικές λάθος ταξινομήσεις έχουμε ότι

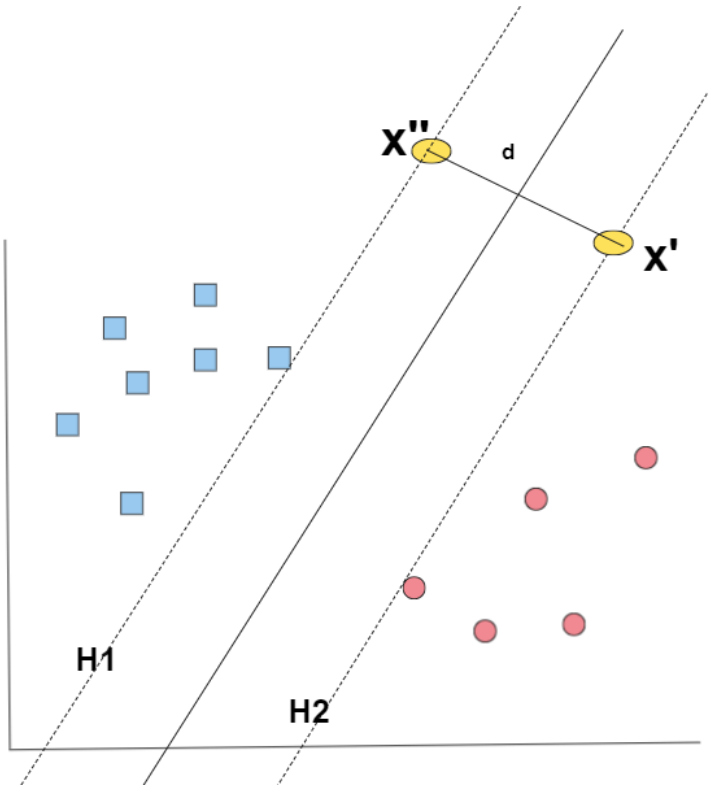
Αν $y_i = 1$ τότε $w^T x_i + b \geq 1$

Αν $y_i = -1$ τότε $w^T x_i + b \leq -1$

Το ενδιάμεσο υπερεπίπεδο από τα H_1 & H_2 είναι αυτό που θέλουμε να προσδιορίσουμε και αναφερθήκαμε στην αρχή και θα το συμβολίζουμε με H_0. Ακόμα παρατηρούμε τα πιο κοντινά σημεία της κάθε κλάσης να είναι κυκλωμένα. Ο λόγος είναι ότι αυτά ακριβώς τα σημεία προσδιορίζουν άμεσα την θέση των H_1 & H_2 και επομένως και του H_0 .Τα σημεία αυτά ονομάζονται διανύσματα υποστήριξης. Οπότε για τον προσδιορισμό του βέλτιστου υπερεπιπέδου χρειαζόμαστε τον υπολογισμό του περιθωρίου ανάμεσα στα H_1 & H_2 ώστε να το μεγιστοποιήσουμε.

Έστω ένα x' όπου κινείται πάνω στο H_2 και d η κάθετη απόσταση του x' από το H_1. Παίρνοντας το διάνυσμα $u = d * \frac{w}{||w||}$ θα ισχύει ότι το μέτρο του u θα ισούται με d και ότι $u \perp H_1$.

Έστω τώρα x'' το σημείο που προκύπτει φέρνοντας την κάθετη ευθεία από το x'ως προς H_2 που τέμνει το H_1.



Σχήμα 3.3.1.9 Απόσταση των H_1 & H_2

Αφού το x''ανήκει στο H_1 θα ικανοποιεί και την εξίσωσή του άρα έχουμε ότι

$$w^T x'' + b = 1 => w^T(x' + u) + b = 1 => w^T\left(x' + d * \frac{w}{||w||}\right) + b = 1 =>$$

$$w^T x' + \frac{d w^T w}{||w||} + b = 1 => w^T x' + \frac{d||w||^2}{||w||} + b = 1 => w^T x' + d||w|| + b = 1 =>$$

$$w^T x' + b = 1 - d||w||$$

Γνωρίζουμε όμως και ότι

$w^T x' + b = -1$ αφού x' $ανήκει$ $στο$ H_2 άρα έχουμε ότι

$w^T\ x' + b = 1 - d||w||$και $\ w^T\ x' + b = -1$

Άρα καταλήγουμε ότι

$$-1 = 1 - d||w|| => d = \frac{2}{||w||} \qquad (3.3.1)$$

Εμείς για την εύρεση του βέλτιστου υπερεπιπέδου θέλουμε την μεγιστοποίηση της απόστασης των H_1& H_2 άρα την μεγιστοποίηση της (3.3.1) η οποία ονομάζεται περιθώριο.

Συνεπώς θέλουμε την μεγιστοποίηση της $\frac{2}{||w||}$ με την προϋπόθεση ότι ισχύουν οι γραμμικοί περιορισμοί:

$$w^T x_i + b \geq\ 1\ αν\ y_i = 1$$

$$w^T x_i + b \leq\ -1\ αν\ y_i = -1,$$

ή ισοδύναμα

$$y_i(w^T x_i + b) \geq\ 1.$$

Ακόμα το πρόβλημα της μεγιστοποίησης της εξίσωσης (3.3.1) για την εύρεση του βέλτιστου υπερεπιπέδου μπορεί να αντικατασταθεί από την ελαχιστοποίηση της $\frac{2}{||w||}$ ή ισοδύναμα της $\frac{||w||^2}{2}$.

Εν κατακλείδι θέλουμε να προβούμε στην λύση του προβλήματος βελτιστοποίησης $\min_{w,b} \frac{||w||^2}{2}$ υπό τους γραμμικούς περιορισμούς

$$y_i(w^T x_i + b) \geq\ 1\ με\ i = 1,2, \ldots . n$$

Το πρόβλημα αυτό ανήκει στη οικογένεια προβλημάτων βελτιστοποίησης που λύνεται με την χρήση πολλαπλασιαστών Lagrange που είδαμε στην παράγραφο 2.1.2. Η επιλογή της συγκεκριμένης τεχνικής για την επίλυση του προβλήματος γίνεται καθώς οι γραμμικοί περιορισμοί που έχουμε την στιγμή θα αντικατασταθούν με των πολλαπλασιαστών Lagrange που τους κάνει πιο εύκολα διαχειρίσιμους από μαθηματική σκοπιά. Άρα εφόσον προβαίνουμε στην επίλυση του προβλήματος μέσω των πολλαπλασιαστών Lagrange, αρχικά ορίζουμε τη συνάρτηση

$$L_p = \frac{||w||^2}{2} - \sum_{i=1}^{n} a_i(y_i(wx_i + b) - 1)) \qquad (3.3.2)$$

$$με\ a_i \geq 0$$

Στην συνέχεια βρίσκουμε τις μερικές παραγώγους ως προς w και b και τις εξισώνουμε με το 0 αντίστοιχα, οπότε προκύπτουν οι εξισώσεις

$$w - \sum_{\iota=1}^{n} a_i y_i x_i \ = 0 => w = \sum_{\iota=1}^{n} a_i y_i x_i \tag{3.3.3}$$

$$\sum_{\iota=1}^{n} a_i y_i = 0 \tag{3.3.4}$$

Αντικαθιστώντας στην εξίσωση (3.3.2) του πρωταρχικού προβλήματος τις (3.3.3) και (3.3.4) έχουμε την δυική μορφή του προβλήματος

$$L_D = \frac{1}{2} \sum_{\iota=1}^{n} (a_i y_i x_i) \sum_{j=1}^{n} (a_j y_j x_j) \ - \sum_{\iota=1}^{n} (a_i y_i x_i \sum_{j=1}^{n} (a_i y_i x_i)) + \sum_{\iota=1}^{n} a_i y_i b + \sum_{\iota=1}^{n} a_i$$

όπου έπειτα από πράξεις

$$L_D = \frac{1}{2} \sum_{\iota=1}^{n} \sum_{j=1}^{n} a_i a_j y_i y_j x_i x_j \ - \sum_{\iota=1}^{n} \sum_{j=1}^{n} a_i a_j y_i y_j x_i x_j \ + \sum_{\iota=1}^{n} a_i y_i b + \sum_{\iota=1}^{n} a_i$$

0

καταλήγουμε στην εξίσωση

$$L_D = \sum_{\iota=1}^{n} a_i \ - \frac{1}{2} \sum_{\iota=1}^{n} \sum_{j=1}^{n} a_i a_j y_i y_j x_i x_j \ . \tag{3.3.5}$$

Μέσω της (3.2.5) βλέπουμε ότι κάθε πολλαπλασιαστής Lagrange a_i αντιστοιχεί σε μία παρατήρηση x_i.Για κάθε x_i το οποίο δεν ανήκει σε ένα από τα H_1 & H_2 ο αντίστοιχος πολλαπλασιαστής a_i ισούται με μηδέν, ενώ για τις περιπτώσεις που έχουμε θετικό a_i αυτός αντιστοιχεί σε ένα διάνυσμα υποστήριξης.

Για τον τελικό καθορισμό των w, b χρησιμοποιούνται οι συνθήκες KKT (Karush-Kuhn-Tucker)(2.1.2.1),(2.1.2.2),(2.1.2.3),(2.1.2.4) που είδαμε στην παράγραφο 2.1.2, οι οποίες στην περίπτωση μας παίρνουν την μορφή (βλ. [10])

$\frac{\partial L_p}{\partial w_j} = w_j - \sum_{i=1}^{n} (a_i y_i x_i) = 0, \frac{\partial L_p}{\partial b} = -\sum_{i=1}^{n} (a_i y_i) = 0, y_i (w^T x_i + b) - 1 \geq \ 0, a_i \geq 0.$

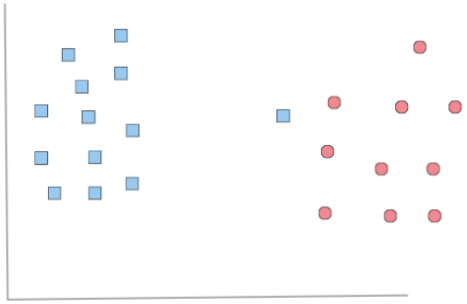
Ακόμα παίρνοντας την τιμή για ένα x_i το οποίο είναι διάνυσμα υποστήριξης άρα ο αντίστοιχος πολλαπλασιαστής Lagrange a_i είναι διάφορος από το μηδέν μπορούμε να καθορίσουμε το b μέσω της συνθήκης

$$\alpha_i (y_i (w^T x_i + b) - 1) = 0$$

Αφού είπαμε ότι η ισότητα ισχύει για τα διανύσματα υποστήριξης.

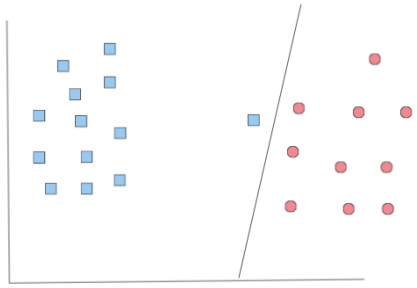
3.4 Soft margin SVM

Στην προηγούμενη περίπτωση είχαμε να αντιμετωπίσουμε ένα σχετικά απλό πρόβλημα εύρεσης ενός κανόνα διαχωρισμού των δεδομένων. Ο λόγος ήταν επειδή η φύση των δεδομένων μας ήταν τέτοια ώστε οι δύο ομάδες ήταν αρκετά μακριά μεταξύ τους. Συχνά όμως τα δεδομένα στα οποία θα εκπαιδευτεί το εκάστοτε μοντέλο καθορίζουν άμεσα την μορφή του χωρίς απαραίτητα αυτή να είναι η πιο λογική και σωστή, καθώς μπορεί τα δεδομένα αυτά να έχουν ορισμένες μη αντιπροσωπευτικές παρατηρήσεις. Βλέπουμε για παράδειγμα το Σχήμα 3.4.1.1



Σχήμα 3.4.1.1 Παράδειγμα κοντινών ομάδων

Αν εμείς ακολουθούσαμε αυστηρά την λογική των Hard margin SVM θα καταλήγαμε στον παρακάτω κανόνα ο οποίος λογικό είναι ότι σε μελλοντικές κόκκινες παρατηρήσεις θα έκανε αρκετές λάθος ταξινομήσεις.

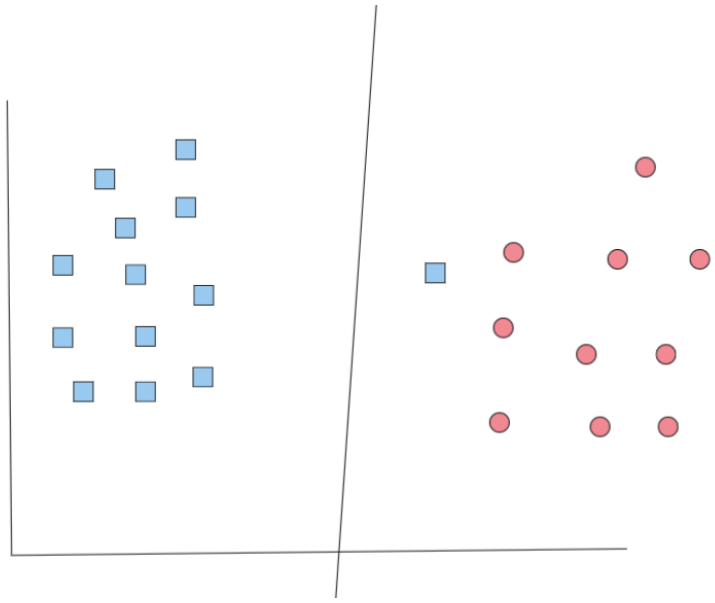


Σχήμα 3.4.1.2 Προσαρμογή βέλτιστης ευθείας σε παράδειγμα κοντινών ομάδων κοντινών ομάδων

Σχήμα 3.4.1.3 Ταξινόμηση σε κοντινές ομάδες

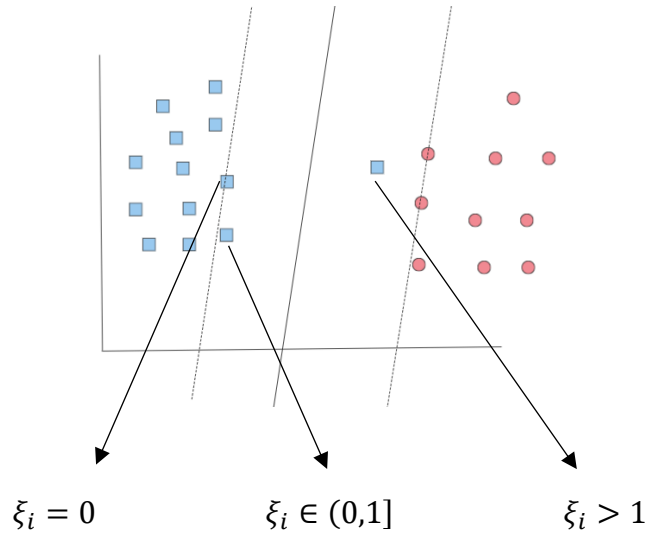
Για παράδειγμα βλέπουμε ότι, η νέα παρατήρηση στην Σχήμα 3.4.1.3, παρόλο που εμφανώς είναι πιο κοντά στην ομάδα των κόκκινων ταξινομείται ως μπλε επειδή έτσι ορίζει η ευθεία του μέγιστου περιθωρίου η οποία εμφανώς επηρεάζεται από την ακραία μπλε παρατήρηση.

Μια ευθεία η οποία θα αγνοούσε την ακραία αυτή παρατήρηση και θα επέτρεπε και τις λάθος ταξινομήσεις με σκοπό να πετύχουμε καλύτερη απόδοση σε μελλοντικές νέες παρατηρήσεις φαίνεται στο Σχήμα 3.4.1.4



Σχήμα 3.4.1.4 Soft margin ευθεία

Για την επίτευξη αυτού λοιπόν, του καθορισμού ενός υπερεπιπέδου το οποίο επιτρέπει τις λάθος ταξινομήσεις εισάχθηκαν κάποιες νέες μεταβλητές ξ_i οι οποίες είναι γνωστές ως μεταβλητές χαλαρότητας (slack variables). Σε κάθε x_i αντιστοιχεί ένα ξ_i θα ισχύει ότι αν το x_i έχει ταξινομηθεί ορθά τότε το αντίστοιχο $\xi_i = 0$,αν το x_i είναι ορθά ταξινομημένο αλλά μέσα στο περιθώριο το αντίστοιχο $\xi_i \in (0,1]$, ενώ αν το x_i είναι λάθος ταξινομημένο $\xi_i > 1$.Οι αντίστοιχες περιπτώσεις φαίνονται στο παρακάτω Σχήμα 3.4.1.5



Σχήμα 3.4.1.5 Εισαγωγή των όρων χαλαρότητας

Στην αντικειμενική συνάρτηση θα εισαχθεί και ένας νέος όρος της μορφής

$$C\left(\sum_{i=1}^{n}\xi_i\right)^k.$$

Ο όρος αυτός αποτελείται από το άθροισμα των όρων χαλαρότητας και από μία θετική σταθερά C η οποία καλείται ως trade off παράμετρος. Ο ρόλος αυτής είναι η εξισορρόπηση ανάμεσα στο πλήθος των λανθασμένων ταξινομήσεων και της μεγιστοποίησης του περιθωρίου. Ονομάζεται trade off παράμετρος καθώς μέσω αυτής θυσιάζουμε ενός μέρος της μεροληψίας μας ώστε να έχουμε καλύτερη απόδοση σε μελλοντικά δεδομένα το οποίο είναι γνωστό στην ορολογία της μηχανικής μάθησης ως bias-variance trade off.

Γενικά ισχύει ότι καθώς το C τείνει στο μηδέν τότε αγνοούνται οι μεταβλητές χαλαρότητας, άρα επιτρέπονται αρκετές λάθος ταξινομήσεις ενώ για μεγάλο C το περιθώριό μας γίνεται πιο αυστηρό προς αυτές. Για C να τείνει στο άπειρο έχουμε την περίπτωση του hard margin SVM. Ο καθορισμός της σταθεράς αυτής γίνεται συνήθως μετά από χρήση της τεχνικής Cross-Validation που αναφέραμε στην Παράγραφο 2.2.3.2.

Επομένως το πρόβλημα μας για k=1, γίνεται πλέον ένα πρόβλημα βελτιστοποίησης της ποσότητας

$$\min_{w,b}\frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i$$

υπό τους περιορισμούς

$$y_i(w^T x_i + b) \geq 1 - \xi_i \text{ και } \xi_i \geq 0 \qquad \gamma\iota\alpha\ i = 1,2,..n.$$

Η συνάρτηση Lagrange του πρωτεύοντος προβλήματος έχει τη μορφή

$$L_p = \frac{||w||^2}{2} - \sum_{i=1}^{n} a_i(y_i(wx_i + b) - 1 + \xi_i)) + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \mu_i \xi_i. \qquad (3.4.1)$$

Ενώ οι περιορισμοί μας είναι οι παρακάτω

$$\mu_i \geq 0 \;\&\; a_i \geq 0 \quad \text{(Πολλαπλασιαστές Lagrange)}$$

$$y_i(w^T x_i + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0.$$

Στην συνέχεια βρίσκουμε και πάλι τις μερικές παραγώγους ως προς τα w, b και ξ_i και εξισώνοντας τες με το μηδέν προκύπτουν οι συνθήκες

$$w - \sum_{\iota=1}^{n} a_i y_i x_i = 0 \Rightarrow w = \sum_{\iota=1}^{n} a_i y_i x_i \qquad (3.4.2)$$

$$\sum_{\iota=1}^{n} a_i y_i = 0 \qquad (3.4.3)$$

$$C - a_i = \mu_i. \qquad (3.4.4)$$

Η συνάρτηση Lagrange για το δυϊκό πρόβλημα είναι η

$$L_D = \frac{1}{2}\sum_{\iota=1}^{n}(a_i y_i x_i)\sum_{j=1}^{n}(a_j y_j x_j) + C\sum_{i=1}^{n}\xi_i - \sum_{\iota=1}^{n}(a_i y_i x_i \sum_{j=1}^{n}(a_i y_i x_i)) + \sum_{\iota=1}^{n} a_i y_i b + \sum_{\iota=1}^{n} a_i(1 - \xi_i) + \sum_{i=1}^{n}\mu_i\xi_i$$

ή ισοδύναμα

$$L_D = \frac{1}{2}\sum_{\iota=1}^{n}\sum_{j=1}^{n} a_i a_j y_i y_j x_i x_j - \sum_{\iota=1}^{n}\sum_{j=1}^{n} a_i a_j y_i y_j x_i x_j + C\sum_{i=1}^{n}\xi_i - \sum_{\iota=1}^{n} a_i y_i b + \sum_{\iota=1}^{n} a_i - \sum_{\iota=1}^{n} a_i\xi_i + \sum_{i=1}^{n}\mu_i\xi_i. \qquad (3.4.5)$$

Με την (3.4.5) και σε συνδυασμό με τις KKT οι οποίες στην περίπτωση του Soft Margin SVM είναι οι

$$w = \sum_{\iota=1}^{n} a_i y_i x_i,\; -\sum_{i=1}^{n}(a_i y_i) = 0,\; C - a_i = \mu_i, y_i(w^T x_i + b) - 1 + \xi_i \geq 0,$$

$$\xi_i \geq 0, \mu_i \geq 0, a_i \geq 0,\; a_i(y_i(w^T x_i + b) - 1 + \xi_i) = 0,\; \mu_i\xi_i = 0$$

η (3.4.5) γίνεται

$$L_D = \sum_{\iota=1}^{n} a_i - \frac{1}{2}\sum_{\iota=1}^{n}\sum_{j=1}^{n} a_i a_j y_i y_j x_i x_j \qquad (3.4.6)$$

και το δυικό πρόβλημά μας σε αυτήν την περίπτωση είναι η μεγιστοποίηση τις (3.3.6). Η λύση του προβλήματος δηλαδή ο καθορισμός των παραμέτρων γίνεται μέσω των KKT όπως και στην περίπτωση του Hard margin SVM.

Αξίζει να κάνουμε ένα σχόλιο για την εξίσωση (3.4.4) που είναι η

$$C - a_i = \mu_i.$$

Σε αυτήν καθώς όλοι οι όροι μας είναι μη αρνητικοί βλέπουμε να προκύπτει μία συνθήκη ότι οι πολλαπλασιαστές Lagrange δεν μπορούν να ξεπεράσουν το C, επομένως

$$0 \leq a_i \leq C.$$

3.5 SVM σε μη γραμμικά δεδομένα

3.5.1 Εισαγωγή

Οι παράγραφοι 3.2, 3.3 και 3.3 που είδαμε πιο πριν καλύπτουν όλο το εύρος των περιπτώσεων όπου τα δεδομένα μας διαχωρίζονται με γραμμικό τρόπο. Η πλειοψηφία των περιπτώσεων όμως δεν κατατάσσονται σε αυτήν την κατηγορία. Το πιο σύνηθες φαινόμενο είναι οι περιπτώσεις όπου τα δεδομένα είναι μπλεγμένα μεταξύ τους και η επίτευξη του διαχωρισμού αυτών μέσω ενός γραμμικού ταξινομητή, είναι ανέφικτη. Ο τρόπος μέσω του οποίου αυτό το εμπόδιο ξεπερνιέται είναι η διαφυγή σε έναν χώρο μεγαλύτερης διάστασης (πολλές φορές ακόμα και άπειρης διάστασης), με σκοπό στον καινούργιο χώρο να μπορεί να βρεθεί ένα επίπεδο το οποίο θα μπορέσει να κατηγοριοποιήσει τα μετασχηματισμένα δεδομένα.

Παρακάτω φαίνεται ένα απλό παράδειγμα με το οποίο ο αναγνώστης μπορεί να πάρει μία ιδέα για τον τρόπο λειτουργίας της τεχνικής και πως οι περισσότερες διαστάσεις μέσω ενός μετασχηματισμού μπορούν να μας δώσουν την λύση. Το παράδειγμα θα γίνει για παρατηρήσεις που κινούνται στον R^1 οπότε μιλάμε για σημεία πάνω στην ευθεία και ο γραμμικός τρόπος διαχωρισμού είναι ένα σημείο. Αρχικά είχαμε τις απλές περιπτώσεις των *Hard margin SVM* και *Soft margin SVM.*

Στην περίπτωση του *Hard margin SVM* είχαμε απλά ένα σημείο το οποίο ισαπέχει από τις δύο ομάδες.

Σχήμα 3.5.1.1 Περίπτωση Hard margin SVM στον R^1

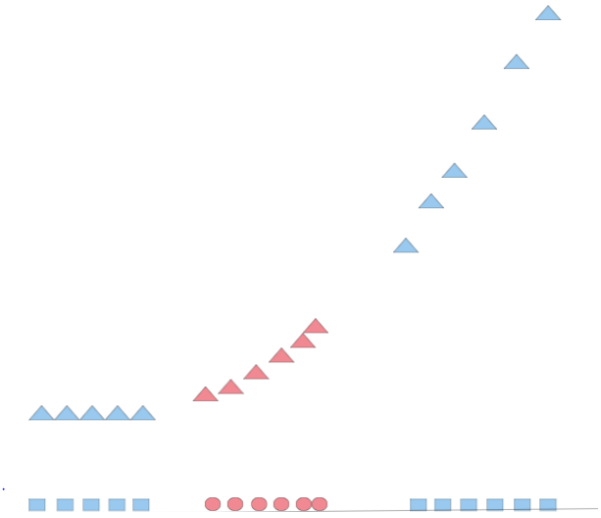
Στην περίπτωση των Soft margin SVM χρησιμοποιήσαμε ως διαχωριστικό σημείο πάλι εκείνο που ισαπέχει από τις δύο ομάδες αφήνοντας μεταξύ τους το μέγιστο περιθώριο αγνοώντας κάποιες παρατηρήσεις οι οποίες ίσως δεν είναι τόσο αντιπροσωπευτικές και μας δίνουν ένα πολύ κακό περιθώριο για ταξινόμηση νέων παρατηρήσεων.

Σχήμα 3.5.1.2 Περίπτωση Soft margin SVM στον R^1

Έστω τώρα ότι έχουμε την περίπτωση όπου δεν υπάρχει θέση στην οποία μπορεί να τοποθετηθεί ένα σημείο ώστε να διαχωρίζει τα δεδομένα μας σε δύο ομάδες, δηλαδή δεν είναι γραμμικά διαχωρίσιμες.

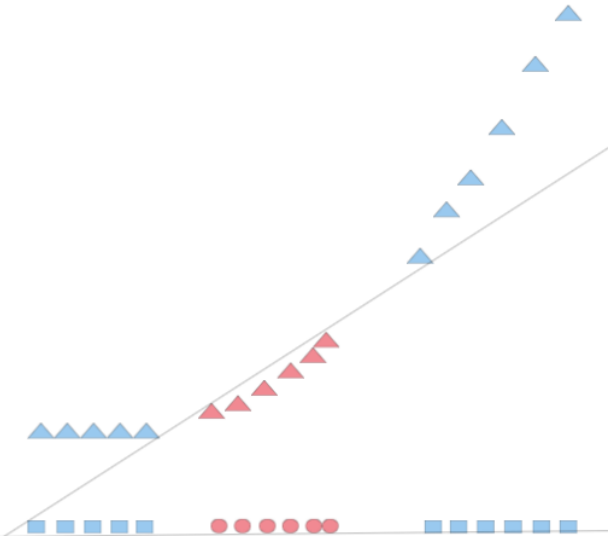
Σχήμα 3.5.1.3 Περίπτωση μη γραμμικών διαχωρίσιμων δεδομένων στον R^1

Αυτό που μπορούμε να κάνουμε είναι να πάρουμε ένα μετασχηματισμό των δεδομένων μας, για παράδειγμα να τα υψώσουμε στο τετράγωνο ,στον κύβο ή σε κάποια άλλη δύναμη. Με αυτόν τον τρόπο παίρνουμε τις νέες παρατηρήσεις μας οι οποίες είναι μετασχηματισμένες και συμβολίζονται με τρίγωνο όπως φαίνεται στην Σχήμα 3.5.1.4..



Σχήμα 3.5.1.4 Μετασχηματισμός δεδομένων για διαφυγή σε μεγαλύτερη διάσταση.

Τώρα πλέον μπορούμε να βρούμε έναν γραμμικό τρόπο ο οποίος μπορεί να διαχωρίσει ικανοποιητικά τα δεδομένα μας στο χώρο που κινούνται οι μετασχηματισμένες παρατηρήσεις μας. Αφού πλέον έχουμε διαφύγει μέσω του μετασχηματισμού στον R^2 ο γραμμικός τρόπος διαχωρισμού θα είναι μία ευθεία η οποία φαίνεται στο Σχήμα 3.5.1.5.



Σχήμα 3.5.1.5 Διαχωρισμός δεδομένων με γραμμικό τρόπο σε μεγαλύτερη διάσταση

Βλέπουμε ότι όλα τα δεδομένα πάνω από την ευθεία ανήκουν στην μπλε ομάδα ενώ κάτω από αυτήν έχουμε τα σημεία της άλλης ομάδας. Με τον τρόπο αυτό είδαμε γραφικά πως μπορούμε μέσω ενός μετασχηματισμού και διαφυγής σε έναν χώρο μεγαλύτερης διάστασης, να ξεπεράσουμε το πρόβλημα των μη γραμμικά διαχωρίσιμων δεδομένων. Η λογική λοιπόν αυτής της μεθόδου είναι ότι μέσω ενός μετασχηματισμού διαφεύγουμε σε έναν χώρο μεγαλύτερης διάστασης, στον οποίο τα δεδομένα μας μπορούν να διαχωριστούν με γραμμικό τρόπο. Στην συνέχεια θα δούμε πιο αναλυτικά σε ποιες περιπτώσεις και υπό ποιες συνθήκες είναι αυτό εφικτό.

3.5.2 Μηχανές διανυσμάτων υποστήριξης μη γραμμικά διαχωρίσιμων δεδομένων και το τέχνασμα του πυρήνα.

Όπως είδαμε και στην εισαγωγή μας, για να ξεπεράσουμε τον μη γραμμικό διαχωρισμό των δεδομένων μας θα πρέπει να μετασχηματίσουμε τις παρατηρήσεις μας σε διανύσματα ενός άλλου χώρου μεγαλύτερης διάστασης χρησιμοποιώντας μία απεικόνιση Φ.[9][10]

$$\Phi: R^p => H.$$

Ο R^P ονομάζεται χώρος εισόδου καθώς σε αυτόν ανήκουν είτε οι παρατηρήσεις οι οποίες θα εκπαιδεύσουν τον αλγόριθμο είτε οι μελλοντικές τις οποίες καλούμαστε να ταξινομήσουμε, και ο H είναι ο χώρος των μετασχηματισμένων δεδομένων. Η συνάρτηση Lagrange του δυϊκού προβλήματος για την περίπτωση των μη γραμμικά διαχωρίσιμων δεδομένων είναι η

$$L_D = \sum_{\iota=1}^{n} a_i \ - \frac{1}{2} \sum_{\iota=1}^{n} \sum_{j=1}^{n} a_i a_j y_i y_j \Phi(x_i) \Phi(x_j) \qquad (3.5.1)$$

Στον τύπο αυτό βλέπουμε να υπάρχει πλήρης ομοιότητα με τις προηγούμενες περιπτώσεις με μόνη διαφοροποίηση ότι οι παρατηρήσεις μας εισέρχονται με την μορφή μετασχηματισμού αλλά και πάλι σε μορφή εσωτερικού γινομένου. Αντίστοιχα χρησιμοποιώντας τις παρακάτω εξισώσεις για

τα support vectors, μπορούν να καθοριστούν οι παράμετροι του προβλήματος όπως και στις προηγούμενες περιπτώσεις.

$$\boldsymbol{w} = \sum_{\iota=1}^{N} a_i y_i \Phi(x_i)$$

όπου N είναι το πλήθος των support vectors.

Η συνάρτηση απόφασης πλέον η οποία θα καθορίζει το πως ταξινομούνται οι καινούργιες παρατηρήσεις $\boldsymbol{x}$ είναι η (3.4.2)

$$f(\boldsymbol{x}) = \sum_{\iota=1}^{N} a_i y_i \Phi(x_i)^T \Phi(\boldsymbol{x}) + b \qquad (3.5.2)$$

Στην (3.5.2) βλέπουμε ότι υπάρχει ένα εσωτερικό γινόμενο το οποίο είναι στην μορφή $\Phi(x_i)^T \Phi(\boldsymbol{x})$. Το εσωτερικό γινόμενο αυτό συμβολίζεται με $K(\boldsymbol{x}, x_i)$ και ονομάζεται εσωτερικό γινόμενο πυρήνα ή πυρήνας ή συνάρτηση πυρήνα. Αυτό που μετρά ο πυρήνας είναι το εσωτερικό γινόμενο των μετασχηματισμένων διανυσμάτων σε μία μεγαλύτερη διάσταση. Επομένως εμείς χρειαζόμαστε να βρούμε ένα μετασχηματισμό Φ και ένα χώρο H τέτοια ώστε από των χώρο των δεδομένων μας να πάμε στον H μέσω της Φ και να ισχύει ότι:

$$K(x_j, x_i) = \Phi(x_i)^T \Phi\left(x_j\right)$$

Για παράδειγμα, έστω ότι τα δεδομένα μας ανήκουν στο R^2και επιλέγουμε ως πυρήνα τον $K\left(x_j, x_i\right) = \left(x_i^T x_j\right)^2$, $H = R^3$ και για

$$\Phi(x_i) = \begin{pmatrix} x_{i,1}^2 \\ \sqrt{2}\, x_{i,1} x_{i,2} \\ x_{i,2}^2 \end{pmatrix}.$$

Τότε

$$\Phi(x_1)^T \Phi(x_2) = \begin{pmatrix} x_{1,1}^2 \\ \sqrt{2}\, x_{1,1} x_{1,2} \\ x_{1,2}^2 \end{pmatrix}^T \begin{pmatrix} x_{2,1}^2 \\ \sqrt{2}\, x_{2,1} x_{2,2} \\ x_{2,2}^2 \end{pmatrix} = x_{11}^2 x_{21}^2 + 2x_{11} x_{12} x_{21} x_{22} + x_{12}^2 x_{22}^2$$
$$= (x_{11} x_{21} + x_{12} x_{22})^2 = (x_1^T x_2)^2 = K\left(x_j, x_i\right)$$

Για να μπορεί να υπάρξει ένας πυρήνα $K\left(x_j, x_i\right)$,μία απεικόνιση Φ και ένας χώρος H, πρέπει να ικανοποιούνται κάποιες συνθήκες οι οποίες διατυπώνονται στο θεώρημα Mercer. Σύμφωνα με αυτό, υπάρχει απεικόνιση Φ και ο πυρήνας $K(x, x')$ μπορεί να αναπτυχθεί στην μορφή, βλ. ([10])

$$K(x, x') = \sum_{\iota=1}^{\infty} \varphi_i(x)\varphi_i(x')$$

αν και μόνο αν για κάθε *g(x)* τέτοια ώστε

$$\int_{\alpha}^{b} g^2(x)\mathrm{d}x$$

να είναι πεπερασμένο, τότε

$$\int_{\alpha}^{b}\int_{a}^{b} K(x, x')g(x)g(x')dxdx' \geq 0$$

Οι συναρτήσεις πυρήνα δεν είναι μοναδικές και ακόμα δεν καθορίζονται και μοναδικά από τις αντίστοιχες απεικονίσεις. Ανάλογα με τη φύση των δεδομένων μας επιλέγουμε κάθε φορά τον πυρήνα μας. Μερικοί από τους πιο διαδεδομένους πυρήνες είναι οι ακόλουθοι.

- $K(x_j, x_i) = x_j^T x_i \quad \gamma\rho\alpha\mu\mu\iota\kappa\acute{o}\varsigma\ \pi\upsilon\rho\acute{\eta}\nu\alpha\varsigma$
- $K(x_j, x_i) = \left(x_j^T x_i + r\right)^p\ \pi o\lambda\upsilon\omega\nu\upsilon\mu\iota\kappa\acute{o}\varsigma\ \pi\upsilon\rho\acute{\eta}\nu\alpha\varsigma$
- $K(x_j, x_i) = tahn\left(ax_j^T x_i + r\right)\ \sigma\iota\gamma\mu o\epsilon\iota\delta\acute{\eta}\varsigma\ \pi\upsilon\rho\acute{\eta}\nu\alpha\varsigma$
- $K(x_j, x_i) = \exp\left\{-\gamma\left|\left|x_i - x_j\right|\right|^2\right\}\ Radial\ Basis\ Function\ (RBF)$

Σε όλες τις περιπτώσεις η επιλογή των παραμέτρων όπως και η επιλογή του C που είδαμε πιο πάνω γίνεται μετά από δοκιμές ή μέσω cross validation. Ο RBF πυρήνας ο οποίος είναι ο πιο διαδεδομένος, με αντικατάσταση του γ με μία συνάρτηση της διακύμανσης παίρνει την παρακάτω μορφή που είναι γνωστή και ως Γκαουσιανός πυρήνας.

$$K(x_j, x_i) = \exp\left\{\frac{\left|\left|x_i - x_j\right|\right|^2}{2\sigma^2}\right\}\ Gaussian\ kernel$$

3.6 Γενίκευση των SVM μέσω της one-against-rest μεθόδου

Όπως είδαμε πριν όλες οι εφαρμογές και η θεωρία των SVM αναπτύχθηκε για περιπτώσεις όπου η ετικέτα μας (το y_i) είναι μία δίτιμη τυχαία μεταβλητή. Η γενίκευση σε προβλήματα όπου οι κλάσεις μας είναι περισσότερες από δύο, τα γνωστά και ως *multiclass-classification problems* μπορεί να επιτευχθεί μέσω διάφορων μεθόδων μία εκ των οποίων είναι η *one-against-rest(*μία ομάδα εναντίον υπολοίπων) βλ., ([16]). Η μέθοδος αυτό που κάνει είναι να ανάγει το αρχικό πρόβλημα σε πολλαπλά binary (δίτιμα) προβλήματα. Πρώτο βήμα της μεθόδου είναι η επιλογή μίας κλάσης και η θεώρηση ότι όλα τα στοιχεία της κλάσης αυτής έχουν την ετικέτα +1, ενώ όλα τα υπόλοιπα στοιχεία έχουν την ετικέτα -1.

Έστω λοιπόν τα δεδομένα μας $\{ x_i \ , y_i\}$ με $x_i \in R^p$ και $y_i \in \{1,2, \dots . k\}$. Επιλέγουμε μία κλάση την φορά και αναθέτουμε σε αυτήν την τιμή +1 ενώ στις υπόλοιπες $k-1$ την τιμή -1. Μέσω της τεχνικής των SVM καταλήγουμε σε μία συνάρτηση απόφασης και αφού έχουμε k ομάδες καταλήγουμε να έχουμε k συναρτήσεις απόφασης που θα είναι της μορφής

$$\boldsymbol{w^T \Phi(x)} + b$$

Άρα ως κλάση του $\boldsymbol{x}$ ορίζεται η

$$\boldsymbol{x} = \arg \max_{i=1,2,\dots k} (\boldsymbol{w_i^T \Phi(x)} + b_i)$$

όπου η συνάρτηση argmax επιστρέφει τον δείκτη εκείνο ο οποίος μεγιστοποιεί την μαθηματική αναπαράσταση που περιέχει.

ΚΕΦΑΛΑΙΟ 4

Πολυωνυμική λογιστική παλινδρόμηση

4.1 Εισαγωγή

Η λογιστική παλινδρόμηση ίσως είναι ένα από τα πιο χαρακτηριστικά μοντέλα τα οποία μπορούν να αντιμετωπίσουν με επιτυχία προβλήματα κατηγοριοποίησης. Η λογιστική παλινδρόμηση ανήκει στην οικογένεια των γενικευμένων γραμμικών μοντέλων. Η χρήση αυτών των μοντέλων κρίθηκε αναγκαία καθώς η κλασσική γραμμική παλινδρόμηση είτε δεν μπορούσε να αντιμετωπίσει ένα μεγάλο εύρος προβλημάτων, είτε παραβιάζονταν οι απαραίτητες προϋποθέσεις της γραμμικής παλινδρόμησης ώστε αυτή να μπορέσει να εφαρμοστεί ορθά. Για παράδειγμα, όταν καλούμαστε να αντιμετωπίσουμε ένα πρόβλημα ταξινόμησης του οποίου η αντίστοιχη ετικέτα δεν ακολουθεί την κανονική κατανομή όπως είναι αναμενόμενο αλλά ακολουθεί πιθανότατα μία διακριτή κατανομή όπως η κατανομή Bernoulli.Παρακάτω παρουσιάζονται τα μοντέλα της λογιστικής παλινδρόμησης και της πολυωνυμικής λογιστικής παλινδρόμησης τα οποία ξεπερνάνε τέτοιου είδους εμπόδια με επιτυχία. Βιβλιογραφία κεφαλαίου 4.[1][2][7][17]

4.2 Λογιστική παλινδρόμηση

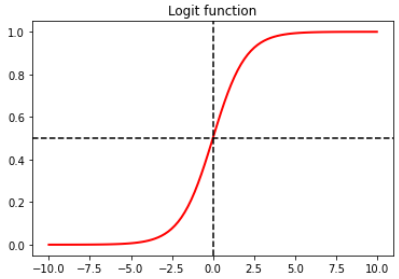
Η λογιστική παλινδρόμηση χρησιμοποιείται για την πρόβλεψη της πιθανότητας εμφάνισης ή πραγματοποίησης ενός γεγονότος. Η κατανομή της απόκρισής μας είναι η κατανομή Bernoulli η οποία όπως είδαμε στην παράγραφο 2.1.3 ανήκει στην εκθετική οικογένεια κατανομών. Άρα στην περίπτωση της Bernoulli έχουμε ότι

$$\text{Για } X \sim B(1,p) \Rightarrow E(X) = p$$

ως συνάρτηση σύνδεσης χρησιμοποιείται κυρίως η συνάρτηση $logit$ με

$$logit(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$$

Στο Σχήμα 4.2.1.1 φαίνεται και η γραφική παράσταση αυτής



Σχήμα 4.2.1.1 Γραφική αναπαράσταση συνάρτηση logit

Άλλες γνωστές συναρτήσεις σύνδεσης είναι η probit και η Complementary log-log.

Το στατιστικό μοντέλο λογιστικής παλινδρόμησης είναι το

$$logit(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \cdots \beta_\kappa x_{ni} \qquad (4.2.1)$$

Η αναγκαιότητα της συνάρτησης σύνδεσης προκύπτει από το γεγονός ότι σε ένα μοντέλο κλασσικής παλινδρόμησης, ανάλογα το είδος και τις τιμές των επεξηγηματικών μεταβλητών, η μεταβλητή απόκρισης μπορεί να πάρει τιμές σε όλο το *R*. Στην περίπτωση μας όμως έχοντας ως απόκριση μας μια πιθανότητα γίνεται αντιληπτό ότι θα παραβιάζεται αρκετές φορές η θεμελιώδες συνθήκη ότι η τιμή αυτή λαμβάνει τιμές στο διάστημα [0,1]. Έτσι μέσω της συνάρτησης σύνδεσης καταφέρουμε να ξεπεράσουμε αυτό το πρόβλημα.

Η εκτίμηση των παραμέτρων γίνεται με την μέθοδο μεγίστης πιθανοφάνειας βλ. ([7]). Καθώς εμείς επιθυμούμε να βρούμε εκτίμηση για το διάνυσμα

$$\boldsymbol{\beta^T} = (\beta_0, \ldots . \beta_\kappa)$$

χρειάζεται να λύσουμε το σύστημα που προκύπτει από τον μηδενισμό του λογαρίθμου της παραγώγου

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_i} = 0$$

για κάθε i.

Για μία τ.μ. Y η οποία ακολουθεί την διωνυμική κατανομή με παραμέτρους *n,p* έχουμε ότι

$$f_Y(y;p) = \exp\left[ylog\left(\frac{p}{1-p}\right) + nlog(1-p) + log\binom{n}{y}\right]$$

Αφού είμαστε στην περίπτωση της Bernoulli έχουμε ότι n=1 άρα

$$f_Y(y;p) = \exp\left[ylog\left(\frac{p}{1-p}\right) + log(1-p)\right]c(y)$$

όπου η συνάρτηση $c(y)$ δεν εξαρτάται από την παράμετρο *p*.

Άρα ως πιθανοφάνεια του δείγματος έχουμε την

$$L(\boldsymbol{p};y) = \exp\left[\sum_{i=1}^{n} y_i\, log\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{n} log(1-p)\right]c_n(y)$$

με $\boldsymbol{p^T} = (p_1, \ldots . p_n)$, $\boldsymbol{y^T} = (y_1, \ldots . y_n)$

Αγνοώντας την c_n η οποία δεν συνδέεται με τις παραμέτρους έχουμε ότι ο λογάριθμος της συνάρτησης πιθανοφάνειας γράφεται ως

$$l(\boldsymbol{p}; y) = \sum_{i=1}^{n} y_i \log\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{n} y_i \log(1-p_i) \qquad (4.2.2)$$

Αντικαθιστώντας τώρα την αρχική γραμμική σχέση (4.2.1) στην σχέση (4.2.2) έχουμε ότι

$$l(\boldsymbol{\beta}; y) = \sum_{i=1}^{n}\sum_{j=0}^{k} y_i \beta_j x_{ij} - \sum_{i=1}^{n} \log\left(1 + exp \sum_{j=0}^{k} \beta_j x_{ij}\right)$$

Ακόμα λύνοντας την (4.2.1) ως προς p_i έχουμε ότι

$$p_i = \frac{\exp\left(\sum_{j=0}^{k} \beta_j x_{ij}\right)}{1 + \exp\left(\sum_{j=0}^{k} \beta_j x_{ij}\right)} \qquad (4.2.3)$$

Στην συνέχεια αντικαθιστώντας την (4.2.3) στην (4.2.2) έχουμε ότι

$$\frac{\partial l}{\partial \beta_r} = \sum_{i=1}^{n} y_i \, x_{ir} - \sum_{i=1}^{n} p_i x_{ir} = \sum_{i=1}^{n} (y_i - p_i) x_{ir}$$

Το οποίο αν το θέσουμε ίσο με το μηδέν για κάθε r έχουμε ένα σύστημα κ+1 εξισώσεων ως προς τις παραμέτρους β_r οι οποίες εμπεριέχονται μέσα στα p_i από την (4.2.3). Το σύστημα αυτό λύνετε μόνο με μεθόδους αριθμητικής αναλύσεως μέσω κάποιας γλώσσας προγραμματισμού όπως η Python ή η R ή μέσω κάποιου στατιστικού πακέτου όπως το SPSS. Για την εύρεση των τιμών χρησιμοποιείται μία μέθοδος που ονομάζεται επαναληπτική μέθοδος των Newton-Raphson.

Όπως είδαμε και στην σχέση (4.2.3) αφού έχουμε πλέον τις εκτιμήσεις των β_i μπορούμε να βρούμε και τις αντίστοιχες εκτιμήσεις των p_i μέσω της

$$\widehat{p_i} = \frac{e^{\sum_{j=0}^{k} \widehat{\beta}_j x_{ij}}}{1 + e^{\sum_{j=0}^{k} \widehat{\beta}_j x_{ij}}}$$

Στο πλαίσιο της στατιστικής συμπερασματολογίας μπορούμε να προβούμε σε ελέγχους πάνω στο μοντέλο και των παραμέτρων αυτού. Οι πιο βασικοί έλεγχοι είναι για τις παραμέτρους β_i. Πολλές φορές καλούμαστε να ελέγξουμε αν η πραγματική τιμή κάποιας παραμέτρου είναι ίση με το μηδέν καθώς αυτό συνεπάγεται ότι η απόκριση μας δεν εξαρτάται από την αντίστοιχη επεξηγηματική μεταβλητή ή δοθέντος ότι υπάρχουν οι υπόλοιπες επεξηγηματικές μεταβλητές στο μοντέλο η συγκεκριμένη δεν κρίνεται στατιστικά σημαντική .

Έτσι για τον έλεγχο

$$H_0: \beta_i = 0$$

$$H_1: \beta_i \neq 0$$

εκμεταλλευόμαστε την ασυμπτωτική ιδιότητα της κανονικότητας των εκτιμητών μεγίστης πιθανοφάνειας και χρησιμοποιούμε την στατιστική συνάρτηση του ελέγχου Wald

$$W = \frac{\hat{\beta}_i}{s(\hat{\beta}_i)}$$

η οποία υπό την μηδενική υπόθεση ακολουθεί προσεγγιστικά την τυπική κανονική κατανομή N(0,1).

Επομένως απορρίπτουμε την μηδενική υπόθεση όταν

$$\left|\frac{\hat{\beta}_i}{s(\hat{\beta}_i)}\right| > z_{a/2}$$

όπου α είναι το επίπεδο σημαντικότητας του ορίζει ο αναλυτής και $z_{a/2}$ είναι το άνω ποσοστιαίο σημείο της τυπικής κανονικής κατανομής.

Μία σημαντική έννοια στο κομμάτι του μοντέλου λογιστικής παλινδρόμησης είναι το *Deviance* ή αλλιώς η απόκλιση.

$$D = -2Loglikelihood(model)$$

Με την ποσότητα αυτή και μέσω του ελέγχου γενικευμένων λόγων πιθανοφανειών μπορούμε για εμφωλευμένα μοντέλα (nested models) να προβούμε σε ελέγχους για την συνεισφορά νέων εισερχομένων μεταβλητών. Ως nested models θεωρούνται δύο μοντέλα για τα οποία το σύνολο των επεξηγηματικών μεταβλητών του ενός μοντέλου είναι υποσύνολο των επεξηγηματικών μεταβλητών του άλλου μοντέλου.

Επομένως έχοντας ένα μοντέλο m_1 και ένα εμφωλευμένο του μοντέλο m_2 μπορούμε

$$D_2 - D_1 = -2(Loglikelihood(m_2) - Loglikelihood(m_1))$$

όπου η διαφορά των αποκλίσεων ακολουθεί προσεγγιστικά την κατανομή X_p^2 όπου p είναι η διαφορά των βαθμών ελευθερίας της παλινδρόμησης των 2 μοντέλων, συγκεκριμένα

$$H_0: m_1 - m_2 \text{ η διαφορά των μοντέλων } m_1 \text{ \& } m_2 \text{ δεν είναι στατιστικά σημαντική}$$

$$H_1: \text{τα μοντέλα διαφέρουν σημαντικά}$$

Όπως είδαμε η λογιστική παλινδρόμηση μπορεί να εκτιμήσει της πιθανότητες p_i μέσω της

$$\widehat{p_i} = \frac{e^{\sum_{j=0}^{k} \widehat{\beta}_j x_{ij}}}{1 + e^{\sum_{j=0}^{k} \widehat{\beta}_j x_{ij}}}$$

Επομένως ο τρόπος που μπορούμε εμείς να χρησιμοποιήσουμε το μοντέλο της λογιστικής παλινδρόμησης ως εργαλείο ταξινόμησης είναι ο ακόλουθος. Διαλέγουμε ένα όριο για το p_i το οποίο να είναι ανάμεσα στο 0 και στο 1 , και αν το διάνυσμα των επεξηγηματικών μεταβλητών της καινούργιας παρατήρησης προσαρμοσμένο στις εκτιμηθείσες παραμέτρους $\hat{\beta}_i$ μας δώσει μία εκτίμηση πάνω από το όριο που το έχουμε θέσει τότε εμείς ταξινομούμε αυτήν την παρατήρηση ως επιτυχία δηλαδή ως 1. Αλλιώς την θεωρούμε ως αποτυχία και της βάζουμε ως label την τιμή 0.Το πιο σύνηθες όριο προφανώς είναι το 0.5.

4.3 Πολυωνυμική λογιστική παλινδρόμηση

Η πολυωνυμική λογιστική παλινδρόμηση έρχεται να γενικεύσει την ιδέα και την τεχνική της λογιστικής παλινδρόμησης όταν η απόκριση μας έχει παραπάνω από δύο κατηγορίες.

Έστω λοιπόν ότι έχουμε την απόκριση y_i η οποία έχει k κατηγορίες με $k > 2$, και έστω ότι η κάθε κατηγορία έχει πιθανότητα εμφάνισης $p_1, p_2, \dots . p_k$. Τότε θα ισχύει ότι

$$\sum_{i=1}^{\kappa} p_i = 1$$

όπου

$$P(y = i) = p_i \quad , \forall\, i = 1, \dots . k$$

Ο στόχος μας και πάλι είναι η μοντελοποίηση των πιθανοτήτων p_i συναρτήσει των επεξηγηματικών μεταβλητών x_i. Για την επίτευξη αυτού πρέπει να καθορίσουμε μία κατηγορία y_i η οποία θα γίνει η κατηγορία αναφοράς (baseline category). Η λογική είναι ότι όπως πριν μοντελοποιούσαμε την $\left(\frac{p}{1-p}\right)$ η οποία μπορεί να γραφτεί και ως $\left(\frac{p1}{p2}\right)$ όπου p_1, p_2 οι πιθανότητες μία παρατήρηση να ανήκει σε μία από τις 2 πιθανές κατηγορίες, έτσι και τώρα θα ως παρονομαστή θα βάλουμε την κατηγορία αναφοράς. Πρακτικά δεν αλλάζει κάτι στην πρόβλεψη και στην εφαρμογή του μοντέλου με την επιλογή μας, απλώς πρέπει να είμαστε λίγο προσεκτικοί στον τρόπο ερμηνείας. Χωρίς βλάβη της γενικότητας έστω ότι διαλέγουμε την πρώτη κατηγορία ως κατηγορία αναφοράς. Ακόμα έστω ότι έχουμε ένα σύνολο επεξηγηματικών μεταβλητών $x_i \, με \, i = 1,2 \dots p$ και y_i η οποία έχει k κατηγορίες $με\, k > 2$ και ορίζουμε ως κατηγορία αναφοράς της y_i την πρώτη. Τότε θα έχουμε k-1 υπομοντέλα της μορφής

$$log\left(\frac{p_2}{p_1}\right) = \beta_{20} + \beta_{21}x_1 + \ \dots\dots\ \beta_{2p}x_p$$

$$log\left(\frac{p_3}{p_1}\right) = \beta_{30} + \beta_{31}x_1 + \ \dots\dots\ \beta_{3p}x_p$$

.

.

$$log\left(\frac{p_k}{p_1}\right) = \beta_{k0} + \beta_{k1}x_1 + \ \dots\dots\ \beta_{kp}x_p$$

Οι εκτιμήσεις γίνονται και πάλι με την μέθοδο της μεγίστης πιθανοφάνειας, και ως αποτέλεσμα για τις πιθανότητες p_i έχουμε :

- Για την κατηγορία αναφοράς

$$p_1 = \frac{1}{1 + e^{x^T\beta_2} + e^{x^T\beta_3} + e^{x^T\beta_\kappa}}$$

- Ενώ για τις άλλες κατηγορίες

$$p_i = \frac{e^{x^T \beta_i}}{\sum_{j=1}^{k} e^{x^T \beta_j}}$$

Οι τρόποι αξιολόγησης και εξαγωγής στατιστικών συμπερασμάτων για την λογιστική παλινδρόμηση ισχύουν και για την πολυωνυμική. Ας σταθούμε λίγο στις ερμηνείες των β_i.

Έστω ότι είχαμε ένα απλό μοντέλο με δύο επεξηγηματικές μεταβλητές και p την πιθανότητα να ανήκει μία παρατήρηση στην κατηγορία 1 με $y_i \in \{0,1\}$

$$logit(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

Η ερμηνεία του β_1 για παράδειγμα είναι ότι με αύξηση της ερμηνευτικής μεταβλητής x_1κατά μία μονάδα δοθέντος ότι η άλλη επεξηγηματική μεταβλητή παραμένει σταθερή αναμένουμε ότι ο λογάριθμος της σχετικής πιθανότητας να ανήκει στην κατηγορία 1 θα αυξηθεί κατά β_1 ή ισοδύναμα ότι η σχετική πιθανότητα θα αυξηθεί πολλαπλασιαστικά κατά e^{β_1}

Στην πολυωνυμική παλινδρόμηση έστω ότι έχουμε $y_i \in \{1,2,3\}$ με επίπεδο αναφοράς την 1 και όπως και πριν 2 επεξηγηματικές μεταβλητές.'Ετσι θα έχουμε την εξής μοντελοποίηση:

$$log\left(\frac{p_2}{p_1}\right) = \beta_{20} + \beta_{21} x_1 + \beta_{22} x_2$$

$$log\left(\frac{p_3}{p_1}\right) = \beta_{30} + \beta_{31} x_1 + \beta_{32} x_2$$

Για την πρώτη περίπτωση θα έχουμε ότι για αύξηση της επεξηγηματικής μεταβλητής x_1 κατά μία μονάδα ενώ η άλλη επεξηγηματική μεταβλητή παραμένει σταθερή αναμένουμε η λογαριθμημένη σχετική πιθανότητα της πραγματοποίησης του γεγονότος 2 έναντι του γεγονότος 1 να αυξηθεί κατά β_{21},αντίστοιχα για το γεγονός 3 έναντι του γεγονότος 1 κατά $\beta_{31.}$ Αντίστοιχα ερμηνεύονται και οι σχετικές πιθανότητες.

Η ταξινόμηση με το μοντέλο της πολυωνυμικής παλινδρόμησης γίνεται με αντίστοιχο τρόπο όπως και στην λογιστική παλινδρόμηση. Ειδικότερα αφού πλέον έχουμε τις εκτιμήσεις με την μέθοδο της μεγίστης πιθανοφάνειας για τα p_i από τους τύπους

$$\widehat{p_\iota} = \frac{1}{1 + e^{x^T \widehat{\beta}_2} + e^{x^T \widehat{\beta}_3} + \cdots . e^{x^T \widehat{\beta}_\kappa}}$$

$$\widehat{p_\iota} = \frac{e^{\sum_{j=0}^{k} \widehat{\beta}_j x_{ij}}}{1 + e^{\sum_{j=0}^{k} \widehat{\beta}_j x_{ij}}},$$

προβαίνουμε στην ταξινόμηση της εκάστοτε παρατήρησης στην κατηγορία της οποίας η πιθανότητα μεγιστοποιείται.

4.4 Ειδικά μέτρα αξιολόγησης του μοντέλου λογιστικής και πολυωνυμικής παλινδρόμησης.

Όπως έχουμε αναφέρει, καθώς η πολυωνυμική και η λογιστική παλινδρόμηση αποτελούν μοντέλα κατηγοριοποίησης ο τρόπος αξιολόγησής τους γίνεται κυρίως μέσω της χρήσης των εξωτερικών μέτρων και του πίνακα συνάφειας. Όμως υπάρχουν και άλλες στατιστικές μετρικές οι οποίες μπορούν να μας δείξουν μερικές πληροφορίες για την αποδοτικότητα του μοντέλου μας, πιο συγκεκριμένα

- Ψευδό-R^2(Pseudo-R^2)
- Κριτήριο του Akaike / Akaike Information Criterion (AIC)
- Κριτήριο του Bayes / Bayesian Information Criterion (BIC)

Όπως και στην γραμμική παλινδρόμηση το R^2 (ο συντελεστής προσαρμογής) μας δίνει το ποσοστό της μεταβλητότητας της απόκρισης μας που ερμηνεύεται από το μοντέλο. Επειδή κάτι τέτοιο όμως στην λογιστική παλινδρόμηση δεν μπορεί να βρεθεί έχουν προταθεί διάφορες εκδοχές για τα λεγόμενα ψευδό-R^2 τα οποία δίνουν μία αντίστοιχη εικόνα σε μοντέλα λογιστικής παλινδρόμησης. Το πιο δημοφιλές είναι αυτό του McFadden , το οποίο ορίζεται από τον παρακάτω τύπο:

$$R_{ps}^2 = 1 - \frac{\log(L_m)}{\log\,(L_0)}$$

Όπου L_m είναι η λογαριθμημένη πιθανοφάνεια του προσαρμοσμένου μοντέλου και L_0 η λογαριθμημένη πιθανοφάνεια του κενού μοντέλου χωρίς επεξηγηματικές μεταβλητές.

Για μοντέλα λογιστικής παλινδρόμησης και πολυωνυμικής, τιμές που θεωρούνται ικανοποιητικές είναι κοντά στα 0.2-0.4.

Τα κριτήρια AIC – BIC περιέχουν μία μορφή ποινής και επιλέγουμε τα μοντέλα τα οποία έχουν όσο το δυνατόν μικρότερη τιμή στα πληροφοριακά κριτήρια. Αυτά τα πληροφοριακά κριτήρια βοηθούν πολύ στην αποφυγή της υπερπροσαρμογής (*overfitting)* στα δεδομένα. Τα δύο κριτήρια υπολογίζονται από τους τύπους.

$$AIC = 2k - \log\,(L)$$

$$BIC = k\log\,(n) - 2\log\,(L)$$

Βλέπουμε ότι στο πληροφοριακό κριτήριο του Bayes υπάρχει μεγαλύτερη ποινή για περισσότερες παραμέτρους έχουμε άρα και αντίστοιχες μεταβλητές. Επίσης η σύγκριση μοντέλων μέσω των πληροφοριακών τους κριτηρίων δεν προϋποθέτει απαραίτητα αυτά τα μοντέλα να είναι εμφωλευμένα μεταξύ τους.

ΚΕΦΑΛΑΙΟ 5

Δέντρα Απόφασης

5.1 Εισαγωγή

Το τελευταίο μοντέλο που θα εξετάσουμε είναι γνωστό ως δέντρα απόφασης. Τα δέντρα απόφασης μπορούν να αντιμετωπίσουν με επιτυχία και προβλήματα παλινδρόμησης αλλά και προβλήματα ταξινόμησης στα οποία θα επικεντρωθούμε. Το σκεπτικό των δέντρων απόφασης είναι να δημιουργήσουμε μία ροή από την ρίζα ως τα φύλλα η οποία θα καθορίζεται από την μορφή των δεδομένων εκπαίδευσης. Τα δέντρα απόφασης αποτελούνται από ένα σύνολο κόμβων οι οποίοι εμπεριέχουν συνθήκες που αφορούν τα χαρακτηριστικά των παρατηρήσεών μας και αναλόγως αν ικανοποιούνται οι συνθήκες αυτές ή όχι αποφασίζεται η συνέχεια της διαδικασίας. Η βιβλιογραφία για το κεφάλαιο 5 είναι η ([2][6][17]).

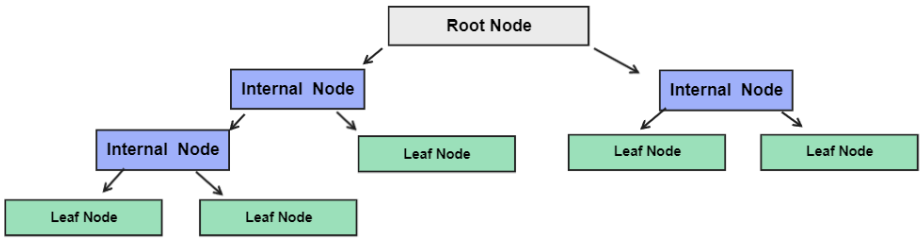
5.2 Το μοντέλο των δέντρων απόφασης σε προβλήματα ταξινόμησης.

Ένα δέντρο απόφασης αποτελείται από κόμβους ή αλλιώς nodes οι οποίοι ανήκουν σε μία από τις παρακάτω κατηγορίες:

- Ρίζα (Root Node)
- Εσωτερικοί κόμβοι (Internal Nodes)
- Φύλλα (Leave Nodes)

Ως ρίζα (root node) ορίζεται το πρώτο node το οποίο είναι το μοναδικό και τροφοδοτεί άλλα nodes ενώ το ίδιο δεν τροφοδοτείται από κάποιο άλλο. Στην συνέχεια έχουμε τα εσωτερικά nodes τα οποία τροφοδοτούν αλλά και τροφοδοτούνται τα ίδια. Τέλος τα φύλλα είναι εκείνα τα nodes τα οποία μόνο τροφοδοτούνται από άλλα nodes. Οι ακμές που συνδέουν τους κόμβους ονομάζονται κλαδιά.

Τα δέντρα απόφασης ξεκινούν από την ρίζα και φιλτράροντας τα χαρακτηριστικά της παρατήρησης που επιθυμούμε να ταξινομήσουμε μέσα στα εσωτερικά nodes καταλήγουμε σε ένα φύλλο το οποίο μας καθορίζει την τελική ομάδα που θα ταξινομηθεί η αντίστοιχη παρατήρηση. Ο τρόπος που γίνεται αυτό είναι μέσω μιας συνθήκης που σχετίζεται με το node της οποίας το αποτέλεσμα αναλόγως αν είναι ψευδές ή αληθές αποφασίζεται η συνέχεια της διαδικασίας. Αν η συνθήκη ισχύει τότε συνεχίζουμε στο αριστερό Internal node ή leave node, αλλιώς επιλέγουμε το δεξί όπως φαίνεται στο Σχήμα 5.2.1.



Σχήμα 5.2.1 Μορφή ενός δέντρου ταξινόμησης

Για την κατασκευή του μοντέλου, έστω ότι έχουμε ένα σύνολο δεδομένων $D = \{\boldsymbol{x}, y_i\}$ με $\boldsymbol{x}$ να αποτελεί το διάνυσμα των χαρακτηριστικών της παρατήρησης i η οποία αποτελείται από p μεταβλητές (ποιοτικές ή ποσοτικές) και y_i να αποτελεί την αντίστοιχη ετικέτα της παρατήρησης. Ο σκοπός μας είναι να δημιουργήσουμε μία αλληλουχία συνθηκών ο οποίες εφαρμόζονται στα x_i και η παραβίαση τους ή μη μας οδηγεί σε αντίστοιχα φύλλα τα οποία είναι η πρόβλεψη για την μεταβλητή y_i. Όπως καταλαβαίνουμε όμως μεγάλη σημασία έχουν τα πιο ψηλά nodes καθώς επιλέγοντας μία πορεία δεν μπορούμε να την ανακαλέσουμε. Επομένως η σειρά με την οποία θα μπουν οι συνθήκες μέσα στα nodes παίζει κομβικό ρόλο.

Ο τρόπος δημιουργίας του μοντέλου και ο καθορισμός της σειράς των nodes μπορεί να γίνει με διάφορες μεθόδους μία από τις πιο δημοφιλείς είναι μέθοδος του δείκτη *Gini.*

Ο δείκτης *Gini* υπολογίζεται μέσω της

$$Gini_{index}(D) = 1 - \sum_{j=1}^{k} p_j^2$$

όπου p_j είναι η πιθανότητα εμφάνισης της κλάσης j από ένα σύνολο κλάσεων k σε ένα σύνολο δεδομένων D με n παρατηρήσεις. Όταν το σύνολο D διαχωρίζεται σε δύο υποσύνολα D_1, D_2 μεγέθους n_1, n_2 αντίστοιχα τότε θα ισχύει ότι

$$Gini_{index}(D) = \frac{n_1}{n} Gini_{index}(D_1) + \frac{n_2}{n} Gini_{index}(D_2)$$

δηλαδή σταθμίζεται ο δείκτης *Gini* αναλόγως με το πλήθος των παρατηρήσεων του κάθε υποσυνόλου. Ο δείκτης *Gini* αποτελεί έναν δείκτη καθαρότητας ή αλλιώς τον βαθμό του impurity ενός node.

Ο τρόπος για την επιλογή ενός node λοιπόν είναι ο υπολογισμός όλων των δεικτών *Gini* για κάθε μεταβλητή ο οποίος προκύπτει μέσα από την στάθμιση των επιπέδων του και επιλέγοντας τον μικρότερο. Στην συνέχεια για να προσθέσουμε κλαδιά βλέπουμε την πορεία του δείκτη *Gini* μέσα στο υποσύνολο που πλέον κινούμαστε και στο τέλος καταλήγουμε σε απόφαση όταν καταφέρουμε να μειώσουμε τον δείκτη αυτόν. Παρακάτω φαίνεται ένα παράδειγμα για να γίνει περισσότερο κατανοητή η μέθοδος.

Έστω ότι έχουμε 3 χαρακτηριστικά X_1, X_2, X_3 και την αντίστοιχη ετικέτα Y όπως φαίνεται στον επόμενο πίνακα δεδομένων.

X_1	X_2	X_3	Y
a	a	a	0
a	a	a	0
b	a	a	1
c	b	a	1
c	c	b	0
b	c	b	1

Πίνακας 5.1 Πίνακας ποιοτικών δεδομένων παραδείγματος

Υπολογίζουμε τους δείκτες *Gini* για κάθε μεταβλητή. Για την πρώτη έχουμε

$$g(x_1) = \frac{2}{6}g(a) + \frac{2}{6}g(b) + \frac{2}{6}g(c)$$

$$g(a) = 1 - \left(p^2_{(Y=0\,|X_1=a)} - p^2_{(Y=1\,|X_1=a)}\right) = 1 - 1 = 0$$

$$g(b) = 1 - \left(p^2_{(Y=0\,|X_1=b)} - p^2_{(y=1\,|x_1=b)}\right) = 1 - 1 = 0$$

$$g(c) = 1 - \left(p^2_{(Y=0\,|X_1=c)} - p^2_{(Y=1\,|X_1=c)}\right) = 1 - (\frac{1}{4} + \frac{1}{4}) = 1 - \frac{1}{2} = \frac{1}{2}$$

Επομένως

$$g(x_1) = 0 + 0 + \frac{1}{3} * \frac{1}{2} = \frac{1}{6} = 0.16$$

Ομοίως έχουμε

$$g(x_2) = \frac{3}{6}g(a) + \frac{1}{6}g(b) + \frac{2}{6}g(c)$$

όπου

$$g(a) = 1 - \left(p^2_{(y=0\,|x_2=a)} - p^2_{(y=1\,|x_2=a)}\right) = \frac{4}{9} = 0$$

$$g(b) = 1 - \left(p^2_{(y=0\,|x_2=b)} - p^2_{(y=1\,|x_2=b)}\right) = 0$$

$$g(c) = 1 - \left(p^2_{(y=0\,|x_2=c)} - p^2_{(y=1\,|x_2=c)}\right) = 1 - \frac{1}{2} = \frac{1}{2}$$

οπότε

$$g(x_2) = \frac{3}{6}g(a) + \frac{1}{6}g(b) + \frac{2}{6}g(c) = 0.35$$

Τέλος

$$g(x_3) = \frac{4}{6}g(a) + \frac{2}{6}g(b)$$

με

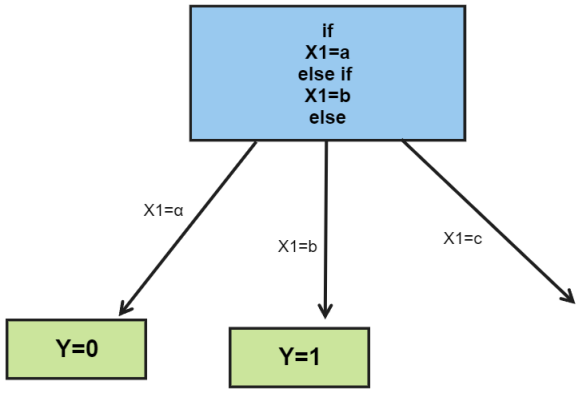
$$g(a) = 1 - \left(p^2_{(y=0\,|x_3=a)} - p^2_{(y=1\,|x_3=a)}\right) = 1 - \frac{1}{2} = \frac{1}{2}$$

$$g(b) = 1 - \left(p^2_{(y=0\,|x_3=b)} - p^2_{(y=1\,|x_3=b)}\right) = \frac{1}{2}$$

οπότε

$$g(x_3) = 0.5$$

Αφού ο μικρότερος δείκτης *Gini* αντιστοιχεί στην μεταβλητή X_1 δημιουργούμε το πρώτο node. Στην συνέχεια παρατηρούμε ότι για τις περιπτώσεις όπου $X_1 = a$ ή $X_1 = b$ όλα τα Y ταξινομούνται με τον ίδιο τρόπο. Επομένως αυτόματα δημιουργούμε επιπλέον leaf nodes απο κάτω ενώ όταν το $X_1 = c$ πρέπει να υπάρχει περαιτέρω διερεύνηση.



Σχήμα 5.2.2 Δέντρο απόφασης παραδείγματος πρώτου επιπέδου

Συνεχίζουμε την διαδικασία για όταν έχουμε $X_1 = c$. Βλέπουμε τώρα τον υποπίνακα

X_1	X_2	X_3	Y
c	b	a	1
c	c	b	0

Πίνακας 5.2 Πίνακας δεδομένων παραδείγματος πρώτου επιπέδου

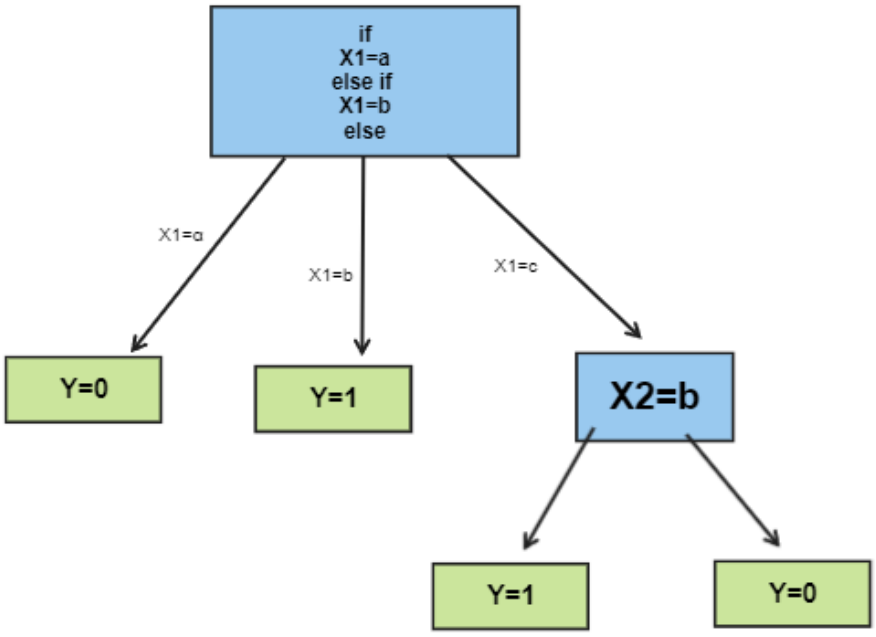
και παρατηρούμε ότι για τις υπόλοιπές μεταβλητές x_2, x_3 έχουμε αντιστοιχία διαφορετικών τιμών του χαρακτηριστικού με την απόκριση, οπότε επιλέγουμε μία από τις δύο και υπολογίζουμε τον αντίστοιχο δείκτη *Gini*. Αν επιλέξουμε την X_2 θα έχουμε

$$g(X_2|X_1 = c) = \frac{1}{2}g(b) + \frac{1}{2}g(c)$$

$$g(b) = 1 - \left(p^2_{(Y=0\,|X_2=b, X_1=c)} - p^2_{(Y=1\,|X_2=b, X_1=c)}\right) = 1 - 1 = 0$$

$$g(c) = 1 - \left(p^2_{(Y=0\,|X_2=c, X_1=c)} - p^2_{(Y=1\,|X_2=c, X_1=c)}\right) = 1 - 1 = 0$$

‘Αρα καταλήγουμε πάλι σε προσθήκη leave nodes.



Σχήμα 5.2.3 Δέντρο απόφασης παραδείγματος δεύτερου επιπέδου

Με το ίδιο σκεπτικό συνεχίζουμε την διαδικασία δημιουργίας κλαδιών και nodes αντίστοιχα. Αυτό το παράδειγμα που μόλις είδαμε είναι ένα σχετικά απλό παράδειγμα στο οποίο όντως καταφέρνουμε τον μηδενισμό του δείκτη Gini και την δημιουργία των φύλλων δηλαδή της απόφασης γρήγορα. Στην πράξη όμως όταν έχουμε πολλά δεδομένα με πολλές παρατηρήσεις και πολλές μεταβλητές δεν φτάνουμε τόσο εύκολα στα φύλλα. Αντιθέτως υπάρχουν πολλά internal nodes τα οποία φιλτράρουν ξανά και ξανά τις παρατηρήσεις μέσω πολλών συνθηκών για να καταλήξουν σε ένα φύλλο του οποίου η απόφαση πολλές φορές φαίνεται να έχει καθοριστεί στα πιο ψηλά επίπεδα. Για τον λόγο αυτό πολλές φορές, όταν δεν θέλουμε να έχουμε μία μεγάλη πολυπλοκότητα στο δέντρο μας, επιλέγουμε να κόψουμε ορισμένα από τα κλαδιά και να μετατρέψουμε το τελευταίο επίπεδο των internal nodes σε leaves.Το βάθος και το σημείο του

κλαδέματος ορίζεται από τον αναλυτή εάν κρίνεται ότι δεν μειώνεται η απόδοση του μοντέλου σημαντικά.

Επίσης στο παράδειγμα που μόλις είδαμε, είχαμε να κάνουμε μόνο με ποιοτικές μεταβλητές. Τα δέντρα ταξινόμησης όμως δεν περιορίζονται σε αυτό καθώς μπορούν εκπαιδευτούν με επιτυχία και για ποσοτικές μεταβλητές. Για την επίτευξη αυτού αρκεί να διατάξουμε το διάνυσμα της ποσοτικής μεταβλητής σε αύξουσα σειρά και να υπολογίσουμε τους ανά 2 μέσους όρους των διαδοχικών όρων και να κάνουμε χρήση αυτών για την δημιουργία των απαραίτητων συνθηκών που χρειαζόμαστε.

Για παράδειγμα, έστω ότι στο προηγούμενο παράδειγμα είχαμε μία επιπλέον μεταβλητή X_4 η οποία μπορούσε να κινηθεί σε όλο το R, και ότι οι παρατηρηθείσες τιμές για αυτήν είναι αυτές που φαίνονται στον Πίνακα 5.3.

X_4	Y
2	0
-9	0
5	1
-4	1
19	0
45	1

Πίνακας 5.3 Πίνακας ποσοτικών δεδομένων παραδείγματος

Αρχικά διατάσσουμε τις τιμές σε αύξουσα σειρά,

X_4	Y
-9	0
-4	1
2	0
5	1
19	0
45	1

Πίνακας 5.4 Πίνακας ποσοτικών δεδομένων παραδείγματος σε αύξουσα σειρά

και στην συνέχεια βρίσκουμε τους ανά 2 μέσους όρους διαδοχικών τιμών.

averages
-6.5
-1
3.5
12
32

Πίνακας 5.5 Πίνακας μέσον όρων διαδοχικών τιμών

Ορίζουμε αυτούς ως κατώφλι (*threshold)* για τον υπολογισμό των δεικτών *Gini*. Άρα δοθέντος έχουμε το εκάστοτε threshold δηλαδή τους μέσους όρους που βρήκαμε προηγουμένως υπολογίζουμε

$$g(x_4) = \frac{1}{6} g(X_4 < -6.5) + \frac{5}{6} g(X_4 \geq -6.5)$$

$$g(X_4 < -6.5) = 1 - \left(p^2_{(Y = 0\ |X_4 < -6.5)} - p^2_{(Y = 1\ |X_4 < -6.5)}\right) = 0$$

$$g(X_4 \geq -6.5) = 1 - \left(p^2_{(Y = 0\ |X_4 \geq -6.5)} - p^2_{(Y = 1\ |X_4 \geq -6.5)}\right) = 1 - \left(\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right)$$
$$= 1 - 0.52 = 0.48$$

Κάνουμε την ίδια διαδικασία για όλες τις πιθανές συνθήκες που προκύπτουν από τους μέσους όρους. Βρίσκουμε την συνθήκη με το χαμηλότερο δείκτη *Gini* και χρησιμοποιούμε αυτήν.

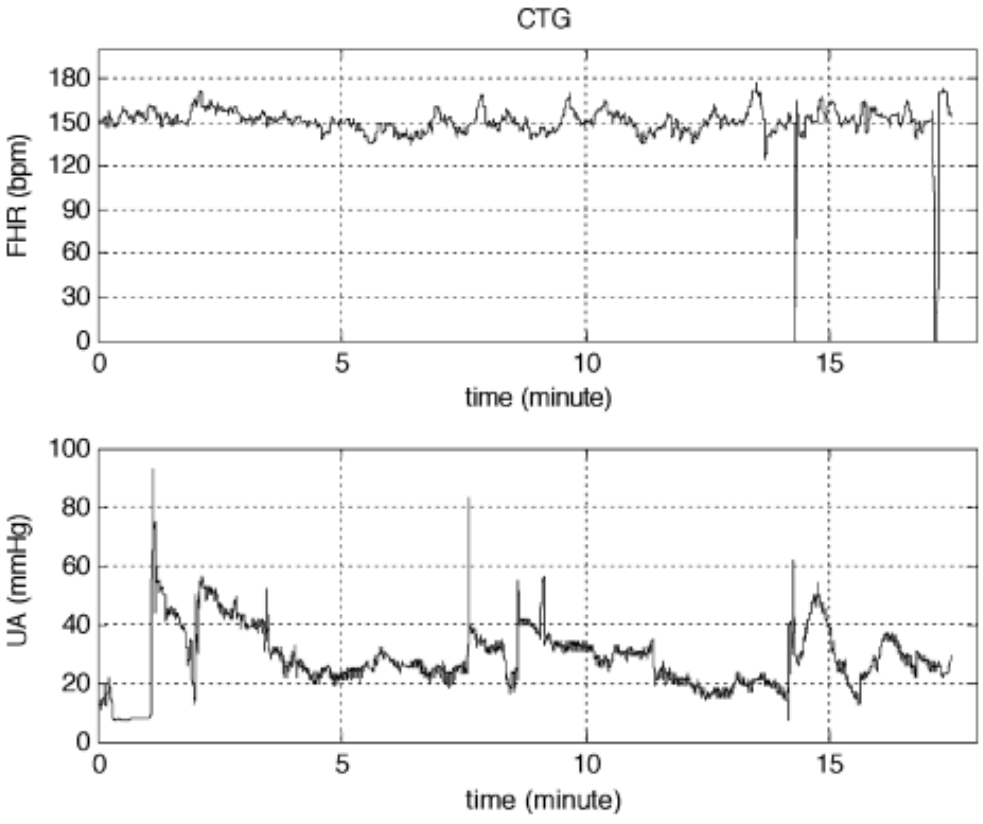
Τώρα αφού πλέον είδαμε τον τρόπο υπολογισμού του δείκτη *Gini* και για ποσοτικές καταλαβαίνουμε εύκολα ότι η δημιουργία των nodes και των leaves γίνεται με αντίστοιχο τρόπο όπως και στις ποιοτικές μεταβλητές.

ΚΕΦΑΛΑΙΟ 6

Εφαρμογή μοντέλων ταξινόμησης για την αξιολόγηση της καλής κατάστασης και υγείας του εμβρύου.

6.1 Εισαγωγή

Η καρδιοτοκογραφία είναι η κύρια μέθοδος που χρησιμοποιείται για την αξιολόγησης της κατάστασης και υγείας ενός εμβρύου. Για τον λόγο αυτό έχουν γίνει πολλές μελέτες πάνω σε αυτήν και προσπάθειες για την αύξηση της αποτελεσματικότητάς της, αξιοποιώντας καινούργια εργαλεία και μεθόδους της τεχνολογίας. Η καρδιοτοκογραφία με λίγα λόγια είναι ένας διαγνωστικός έλεγχος που συνήθως πραγματοποιείται κατά το τρίτο τρίμηνο της εγκυμοσύνης και σκοπό έχει να ανιχνεύσει τυχόν δυσλειτουργίες στην ανάπτυξη και καλή υγεία του εμβρύου. Η μέθοδος της καρδιοτοκογραφίας γίνεται με την τοποθέτηση δύο αισθητήρων στην κοιλιά της γυναίκας. Ο πρώτος παρακολουθεί τον ρυθμό της καρδιάς του εμβρύου (FHR) και ο άλλος τις συστολικές συσπάσεις της μήτρας (UA). Το αποτέλεσμα από τους αισθητήρες καταγράφεται πάνω σε ένα γράφημα το οποίο έχει την μορφή που απεικονίζεται στο παρακάτω Σχήμα, βλ. ([8] [12][15][23]).



Σχήμα 6.1.1 Παράδειγμα μορφής καρδιοτοκογραφήματος

Στο πάνω γράφημα βλέπουμε την καταγραφή του ρυθμού καρδιάς του εμβρύου ενώ στο κάτω τις συστολικές συσπάσεις της μήτρας. Εκτιμάται ότι οι τυχόν διακυμάνσεις που υπάρχουν στο πάνω γράφημα που αφορούν τον ρυθμό της καρδιάς συνδέονται με αλληλεπιδράσεις του συμπαθητικού νευρικού συστήματος και του παρασυμπαθητικού νευρικού συστήματος γνωστά και ως SNS και PSNS αντίστοιχα. Το SNS είναι το μέρος του νευρικού συστήματος του ανθρώπου το οποίο είναι υπεύθυνο για την ανταπόκριση του σώματος σε καταστάσεις άμεσου κινδύνου ενώ το PSNS είναι υπεύθυνο για την χαλάρωση και αποκατάσταση του σώματος σε καταστάσεις ηρεμίας. Η διέγερση του SNS παραπέμπει σε αύξηση του ρυθμού της καρδιάς ενώ του PSNS σε μείωση. Ακόμα οι μητρικές συστολικές συσπάσεις της μήτρας οι οποίες φαίνονται στο δεύτερο γράφημα, αποτελούν μία κατάσταση πίεσης για το έμβρυο έχοντας μία άμεση επιρροή στο SNS άρα και στον ρυθμό της καρδιάς του εμβρύου. Στο κεφάλαιο αυτό θα προβούμε σε μία μελέτη των χαρακτηριστικών αυτών και όχι μόνο, τα οποία χρησιμοποιώντας τα ως ένα διάνυσμα εισόδου στο αντίστοιχο μοντέλο που θα προσαρμόσουμε θα μας οδηγήσουν σε μία πρόβλεψη για την κατάσταση του εμβρύου, πιο συγκεκριμένα θα το κατατάξουν σε μία από τρεις διακριτές κατηγορίες.

6.2 Παρουσίαση του συνόλου δεδομένων.

Παραπάνω όπως είδαμε παρουσιάστηκε ένα πραγματικό πρόβλημα από τον κλάδο της Ιατρικής το οποίο συνδέεται με την κατάσταση του εμβρύου και κατά πόσο αυτή είναι ανησυχητική ή όχι. Ο κύριος στόχος λοιπόν για τον οποίο παρουσιάσαμε τα αντίστοιχα μοντέλα στα προηγούμενα κεφάλαια είναι για να μπορούν να κληθούν και να δοκιμαστούν σε πραγματικά προβλήματα όπως ένας διαγνωστικός έλεγχος που προβλέπει αν η κατάσταση του εμβρύου είναι φυσιολογική ή όχι. Το σύνολο δεδομένων που θα χρησιμοποιήσουμε για την εκπαίδευση, την προσαρμογή αλλά και την αξιολόγηση των στατιστικών μοντέλων ταξινόμησης που αναφέραμε είναι το Cardiotocography Data Set από το UCI Machine Learning Repository [8].Το σύνολο δεδομένων αυτό αποτελείται από 2126 καρδιοτοκογραφήματα (CTGs) τα οποία μετά από επεξεργασία που υποβλήθηκαν, κατέληξαν σε 21 χαρακτηριστικά τα οποία συνδέονται με τις δύο μετρικές ενός καρδιοτοκογραφήματος δηλαδή τα FHR και UA. Στην συνέχεια, τρεις ειδικοί ανέλαβαν την ταξινόμηση του κάθε καρδιοτοκογραφήματος, με αποτέλεσμα στην κάθε παρατήρηση να αντιστοιχούν δύο ετικέτες. Η πρώτη ετικέτα αντιστοιχεί ένα μοτίβο του ρυθμού της καρδιάς λαμβάνοντας 10 πιθανές τιμές και είναι γνωστή ως CLASS. Η δεύτερη ετικέτα στην οποία εμείς θα επικεντρωθούμε είναι ένας δείκτης ο οποίος αντιπροσωπεύει την κατάσταση του εμβρύου. Ο δείκτης αυτός, γνωστός και ως NSP ή fetal-health μπορεί να λάβει 3 διαφορετικές τιμές οι οποίες αντιστοιχούν σε φυσιολογική κατάσταση (Normal), ύποπτη κατάσταση (Suspect) και τέλος παθολογική κατάσταση (Pathological), τα αρχικά των οποίων έδωσαν και την αντίστοιχη ονομασία στον δείκτη. Όπως είναι προφανές, η πρώτη κλάση μας δείχνει ότι η κατάσταση του εμβρύου είναι φυσιολογική και δεν έχουμε επιπλοκές, η δεύτερη κλάση μας καλεί να προβούμε σε περαιτέρω εξετάσεις αλλά δεν σημαίνει ότι υπάρχει αναγκαστικά κάποιο παθολογικό πρόβλημα καθώς μπορεί να οφείλεται σε κάποια κίνηση του παιδιού ή σε σύσπαση της μήτρας. Τέλος η παθολογική κατάσταση είναι ανησυχητική καθώς μπορεί το έμβρυο να κινδυνεύει. Μία τέτοια κατάσταση μπορεί να οφείλεται σε αίτια όπως ότι δεν λαμβάνει επαρκές οξυγόνο, ότι υπάρχει ανεπάρκεια του πλακούντα ή υπέρταση της μητέρας. Σκοπός μας λοιπόν είναι η δημιουργία ενός μοντέλου το οποίο τροφοδοτώντας το με τα 21 χαρακτηριστικά ή ένα

υποσύνολο αυτών, μίας καινούργιας παρατήρησης να μπορούμε επιτυχώς και με αρκετή ακρίβεια να προβλέψουμε την κατάσταση του εμβρύου ή αντίστοιχα την σωστή κατηγορία στην οποία ανήκει ο NSP. Για την επίτευξη αυτού πρέπει πρώτα να μελετήσουμε τα 21 αυτά χαρακτηριστικά τα οποία θα παρατηρούμε και με βάση αυτά να προβούμε σε μία εκτίμηση για την τιμή του NSP. Παρακάτω φαίνονται τα χαρακτηριστικά αυτά η αλλιώς τα *features* με τα οποία θα δουλέψουμε.

1. LB - FHR baseline (beats per minute)	Σφυγμοί ανά λεπτό
2. AC - # of accelerations per second	Αριθμός επιταχύνσεων ανά δευτερόλεπτο
3. FM - # of fetal movements per second	Αριθμός κινήσεων εμβρύου ανά δευτερόλεπτο
4. UC - # of uterine contractions per second	Συσπάσεις μήτρας ανά δευτερόλεπτο
5. DL - # of light decelerations per second	Αριθμός ελαφρών επιβραδύνσεων ανά δευτερόλεπτο
6. DS - # of severe decelerations per second	Αριθμός έντονων επιβραδύνσεων ανά δευτερόλεπτο
7. DP - # of prolongued decelerations per second	Παρατεταμένη επιβράδυνση ανά δευτερόλεπτο
8. ASTV - percentage of time with abnormal short term variability	Ποσοστό χρόνου με μη φυσιολογική βραχυπρόθεσμή μεταβλητότητα
9. MSTV - mean value of short term variability	Μέση τιμή βραχύχρονης μεταβλητότητας
10. ALTV - percentage of time with abnormal long term variability	Ποσοστό χρόνου με μη φυσιολογική μακροπρόθεσμη μεταβλητότητα
11. MLTV - mean value of long term variability	Μέση τιμή μακροπρόθεσμης μεταβλητότητας
12. Width - width of FHR histogram	Πλάτος ιστογράμματος του FHR
13. Min - minimum of FHR histogram	Μικρότερη τιμή ιστογράμματος του FHR
14. Max - Maximum of FHR histogram	Μεγαλύτερη τιμή ιστογράμματος του FHR
15. Nmax - # of histogram peaks	Αριθμός κορυφών ιστογράμματος του FHR
16. Nzeros - # of histogram zeros	Αριθμός μηδενικών ιστογράμματος του FHR
17. Mode - histogram mode	Επικρατούσα τιμή ιστογράμματος του FHR
18. Mean - histogram mean	Μέση τιμή ιστογράμματος του FHR
19. Median - histogram median	Διάμεσος ιστογράμματος του FHR
20. Variance - histogram variance	Διακύμανση ιστογράμματος του FHR
21. Tendency - histogram tendency	Συμμετρία ιστογράμματος του FHR (αριστερή ασσυμετρία=-1,δεξιά ασσυμετρία=1,συμμετρία=0)
22. NSP - fetal state class code based on expert obstetricians' classification (1=Normal, 2=Suspect, 3=Pathological)	NSP -Fetal Health 1=Κανονική,2=Ύποπτη,3=Παθολογική

Πίνακας 6.1 Χαρακτηριστικά του συνόλου δεδομένων

Όπως βλέπουμε έχουμε 21 χαρακτηριστικά και στο τέλος η εικοστή δεύτερη μεταβλητή είναι το label μας. Παρατηρούμε, ότι όλα τα χαρακτηριστικά είναι αλληλένδετα με τα δύο

χαρακτηριστικά που μετρήθηκαν από του αισθητήρες μέσω του καρδιοτοκογράφου, πιο συγκεκριμένα τις μεταβλητές 1 (FHR-baseline) και 4 (UC uterine contractions).

6.3 Ανάλυση και επεξεργασία του συνόλου δεδομένων.

Στο σύνολο δεδομένων μας ξεχωρίζει η μεταβλητή 21 Tendency, η οποία είναι μία ποιοτική μεταβλητή ή αλλιώς όπως ορίζεται στον κλάδο της στατιστικής ένας παράγοντας που αποτελείται από 3 επίπεδα. Για τον λόγο αυτό θα πρέπει να λάβει ιδιαίτερη μεταχείριση κατά την επεξεργασία. Η μεταβλητή αυτή όμως δεν είναι απαραίτητο να αποτελέσει τον μοναδικό παράγοντα του συνόλου δεδομένων, καθώς είναι πολύ πιθανόν, μετά την επεξεργασία και την ανάλυση που θα κάνουμε στα δεδομένα αυτό να αλλάξει και μια φαινομενικά ποσοτική μεταβλητή δηλαδή μία μεταβλητή που αντιπροσωπεύει κάτι μετρήσιμο καθώς αυτή λαμβάνει πολύ μικρό εύρος διαφορετικών τιμών να πρέπει να αντιμετωπιστεί ως ποιοτική.

Η γλώσσα προγραμματισμού με την οποία θα δουλέψουμε για την στατιστική ανάλυση των μοντέλων και την επεξεργασία των δεδομένων είναι η Python, μία γλώσσα προγραμματισμού αλληλένδετη με τομέα του data science. Καθώς κάποιες από τις τεχνικές που θα χρησιμοποιήσουμε περιλαμβάνουν μέσα μία τυχαιοποίηση ως κατάσταση τυχαίοτητας ή αλλιώς όπως αναφέρεται στην Python ως *random state* θα χρησιμοποιήσουμε τον αριθμό 42 ο οποίος αποτελεί και την προεπιλεγμένη επιλογή στις περισσότερες μεθόδους και συναρτήσεις που αφορούν το κομμάτι της επιστήμης δεδομένων που υπάρχουν στο περιβάλλον της Python. Αρχικά βλέπουμε μία πρώτη εικόνα των δεδομένων.

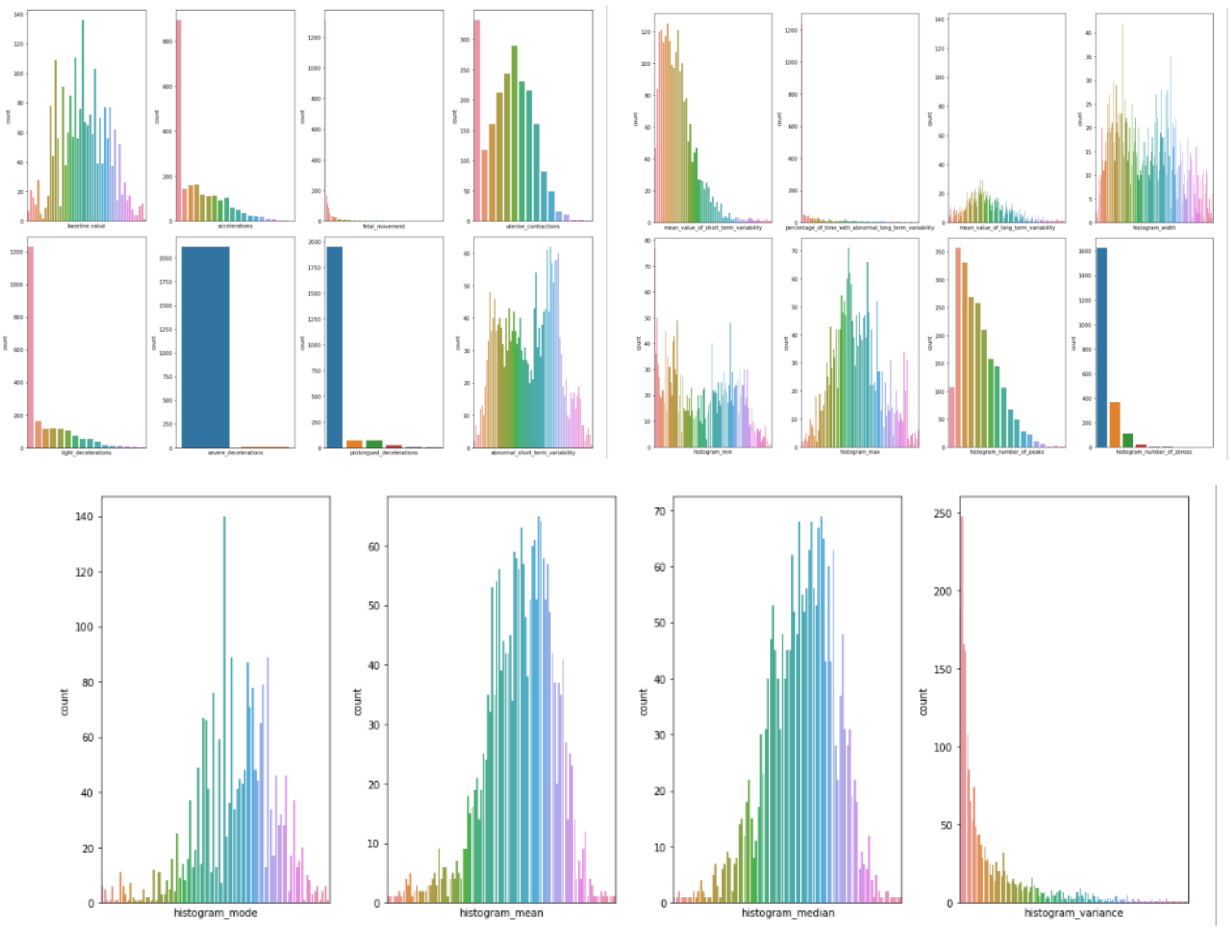
:

	baseline.value	accelerations	fetal_movement	uterine_contractions	light_decelerations	severe_decelerations	prolongued_decelerations	abnormal_short_ter
0	120	0.000	0.000	0.000	0.000	0.0	0.0	
1	132	0.006	0.000	0.006	0.003	0.0	0.0	
2	133	0.003	0.000	0.008	0.003	0.0	0.0	
3	134	0.003	0.000	0.008	0.003	0.0	0.0	
4	132	0.007	0.000	0.008	0.000	0.0	0.0	
...	...	...	...	...	...	...	...	
2121	140	0.000	0.000	0.007	0.000	0.0	0.0	
2122	140	0.001	0.000	0.007	0.000	0.0	0.0	
2123	140	0.001	0.000	0.007	0.000	0.0	0.0	
2124	140	0.001	0.000	0.006	0.000	0.0	0.0	
2125	142	0.002	0.002	0.008	0.000	0.0	0.0	

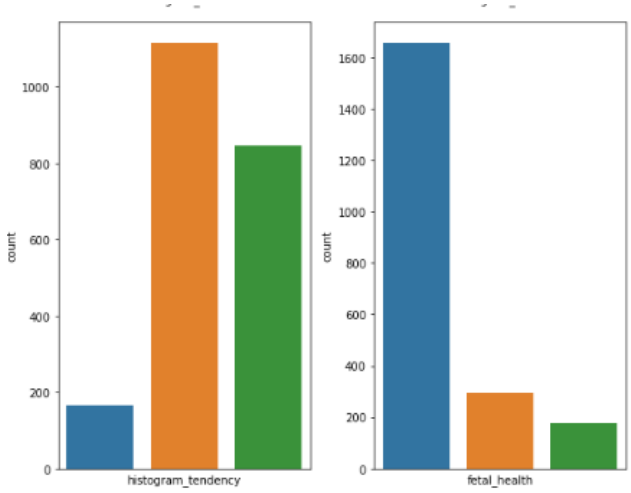
2126 rows × 22 columns

Σχήμα 6.3.1 Πρώτη εικόνα του συνόλου δεδομένων

Το πρώτο που θέλουμε να κάνουμε είναι να ξεχωρίσουμε το είδος των χαρακτηριστικών. Όπως είδαμε πριν η μεταβλητή Tendency, αποτελεί ένα παράγοντα με 3 επίπεδα, όμως υπάρχουν και άλλες μεταβλητές η οποίες είτε παίρνουν διακριτές τιμές είτε το πλήθος των διαφορετικών που παίρνουν είναι μικρό. Στο Σχήμα 6.3.2 λοιπόν βλέπουμε για όλες τις μεταβλητές τα ιστογράμματα τα οποία μας δίνουν μία πρώτη εικόνα τόσο για την φύση των δεδομένων όσο και για το πλήθος των διαφορετικών τιμών που μπορεί να λάβει μία μεταβλητή.



Σχήμα 6.3.2 Ιστογράμματα ποσοτικών μεταβλητών



Σχήμα 6.3.3 Ιστογράμματα ποιοτικών μεταβλητών

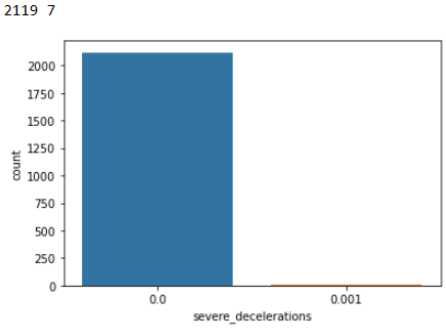
Παρατηρούμε στις περισσότερες μεταβλητές, να υπάρχουν διακυμάνσεις μέσα σε ένα μεγάλο εύρος τιμών, όμως υπάρχουν και ορισμένες από αυτές που λαμβάνουν ένα μικρό. Τις μεταβλητές αυτές θα τις διαχειριστούμε ως παράγοντες, με επίπεδα το μικρό εύρος τιμών που έχουν. Ο τρόπος της εκλογής των μεταβλητών που θα αντιμετωπισθούν ως παράγοντες είναι ο εξής. Θα ορίσουμε ένα κατώφλι (*threshold)* πλήθους διαφορετικών παρατηρήσεων ανά μεταβλητή. Μετά θα πάμε και θα ελέγξουμε σε κάθε μεταβλητή αν το πλήθος των διαφορετικών τιμών το οποίο περιέχει δεν ξεπερνάει το *threshold* αυτό, το οποίο θα σημαίνει ότι η μεταβλητή αυτή πλέον θα οριστεί ως παράγοντας.

Για τον ορισμό του *threshold* θα επιλέξουμε μία μικρή τιμή καθώς μετά από μεγάλο αριθμό επιπέδων δημιουργούνται προβλήματα στο κομμάτι της ερμηνείας. Επομένως για *threshold=8* έχουμε ότι οι μεταβλητές που πρέπει να αντιμετωπιστούν ως παράγοντες είναι οι ακόλουθες.

severe_decelerations (DS)
prolongued_decelerations (DP)
histogram_tendency (Tendency)

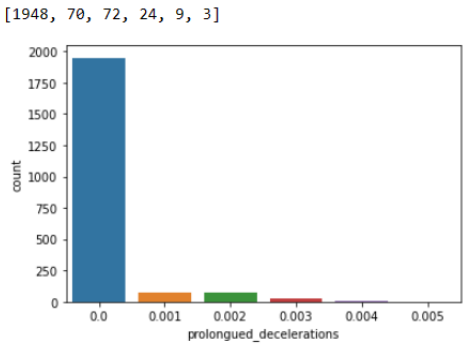
Πίνακας 6.2 Μεταβλητές που πρέπει να αντιμετωπιστούν ως παράγοντες του συνόλου δεδομένων

Για την μεταβλητή **histogram_tendency** είδαμε πριν ότι ήταν από την φύση της ποιοτική μεταβλητή οπότε η εμφάνιση της στους παράγοντες ήταν αναμενόμενη. Ενδιαφέρον έχουν οι άλλες δύο μεταβλητές. Ξεκινώντας με την **DS** βλέπουμε και πάλι το αντίστοιχο γράφημα της.



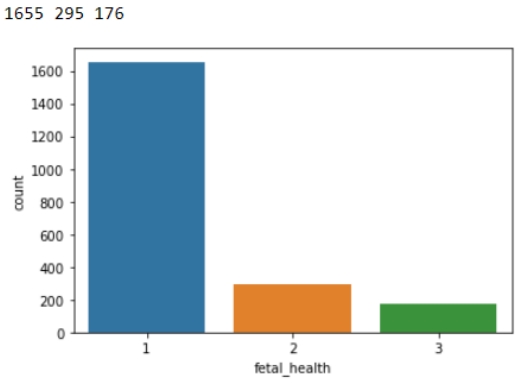
Σχήμα 6.3.4 Ιστόγραμμα DS

Βλέπουμε ότι η μεταβλητή DS αποτελείται από 2 επίπεδα τα οποία είναι τα 0 και 0.001. Πιο συγκεκριμένα βλέπουμε ότι από τις 2126 παρατηρήσεις οι 2119 λαμβάνουν την τιμή 0 ενώ οι υπόλοιπες λαμβάνουν την τιμή 0.001. Αυτή η μη ύπαρξη ισορροπίας ανάμεσα στα επίπεδα αυτού του παράγοντα οδηγεί σε ελάχιστη διακύμανση ανάμεσα στα επίπεδα του, οπότε είναι πολύ πιθανόν καθώς μη προσφέροντας αρκετή πληροφορία, στην διαδικασία της επιλογής χαρακτηριστικών (*feature selection*) του μοντέλου να εξαιρεθεί. Προς το παρών θα την κρατήσουμε με επιφύλαξη αλλά δεν αναμένουμε να μπορέσει να περάσει το στάδιο του *feature selection.*



Σχήμα 6.3.5 Ιστόγραμμα DP

Ως προς την μεταβλητή **DP**, βλέπουμε να αποτελείται από 6 διαφορετικά επίπεδα τα οποία είναι τα 0,0.001,0.002,0.003,0.004,0.005 αντίστοιχα. Βλέποντας και πάλι μέσω του γραφήματος την μεγάλη διαφορά στο πλήθος των παρατηρήσεων σε κάθε κλάση. Ειδικότερα παρατηρώντας το πλήθος της κάθε κλάσης βλέπουμε να έχουμε 1948 για την πρώτη κλάση ,70 για την δεύτερη, 72 για την τρίτη, 24 για την τέταρτη, 9 για την πέμπτη και 3 για την έκτη. Έχουμε πάλι μία αντίστοιχη περίπτωση με την DS,όχι στον ίδιο βαθμό βέβαια για όλα τα επίπεδα. Αναμένουμε λοιπόν και σε αυτήν να δημιουργηθούν αντίστοιχα προβλήματα αλλά αυτό θα αφήσουμε να φανεί στο κομμάτι του *feature selection* του κάθε μοντέλου. Πριν προβούμε στην περαιτέρω μετατροπή των παραπάνω μεταβλητών σε ποιοτικές αφού τις εντοπίσαμε αξίζει να δούμε και λίγο περισσότερο την απόκριση μας. Η απόκριση μας η οποία είναι ο NSP ή αλλιώς fetal_health λαμβάνει τρεις διαφορετικές τιμές, αλλά βλέπουμε να υπάρχει και στην απόκριση μία έλλειψη ισορροπίας.



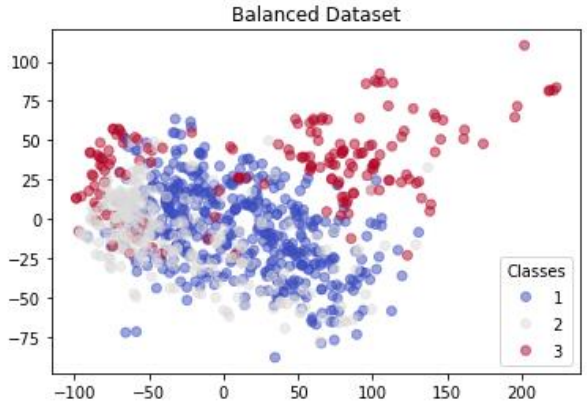
Σχήμα 6.3.6 Ιστόγραμμα fetal_health

Στο Σχήμα 6.3.6 βλέπουμε ότι οι περιπτώσεις όπου το έμβρυο είναι σε φυσιολογική κατάσταση είναι 1655 από τις 2126, οι ύποπτες είναι 295 ενώ οι παθολογικές είναι 176. Αυτό δεν σημαίνει ότι η προσαρμογή των μοντέλων μας δεν θα ικανοποιητική αλλά πολύ πιθανόν να περιέχει μία δόση μεροληψίας υπέρ της κλάσης 1. Γενικότερα αν εμείς απλά δημιουργήσουμε ένα μοντέλο το οποίο ανεξαρτήτως του καρδιοτοκογραφήματος και των υπολοίπων χαρακτηριστικών προβλέπει συνέχεια ότι η παρατήρηση ανήκει στην κατηγορία Normal (1), αυτό θα έμοιαζε πρόχειρο αλλά παράλληλα θα είχε $\frac{1655}{1655+295+176} = 77\%$ σωστές προβλέψεις. Για τον λόγο αυτό

μέσω της Python θα προβούμε σε ισορρόπηση του συνόλου των δεδομένων. Η μέθοδος εξισορρόπησης που θα χρησιμοποιήσουμε είναι το *under-sampling* μέσω του οποίου θα μειώσουμε το πλήθος των παρατηρήσεων τα οποία αποτελούν την πλειοψηφία του dataset και πιο συγκεκριμένα της κατάστασης 1. Ως νέο αριθμό πλήθους της κατηγορίας 1 θα ορίσουμε να είναι το άθροισμα των άλλων δύο κλάσεων με σκοπό να έχουμε ακόμα ένα αρκετά μεγάλο δείγμα στο σύνολο και να διατηρείται η εμφανής πλειοψηφία της κατάστασης 1 χωρίς όμως να υποσκιάζει σε τέτοιο βαθμό τις άλλες κλάσεις.

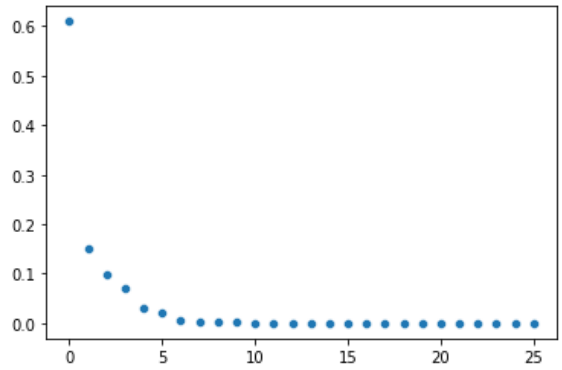
Επόμενο βήμα για την επεξεργασία των δεδομένων μας είναι ο διαχωρισμός σε σύνολο επεξηγηματικών μεταβλητών και απόκρισης ή αλλιώς σε *features* και *label*. Ακόμα μιας και καθορίσαμε πλέον το ποιοι θα είναι οι παράγοντες μας θα προβούμε στην δημιουργία *dummy variables* για αυτές κρατώντας το μικρότερο επίπεδο ως επίπεδο αναφοράς.

Στην συνέχεια θα προβούμε σε μία οπτικοποίηση των δεδομένων μας. Ο τρόπος που θα το επιτύχουμε αυτό είναι μέσω της *principal component analysis* (μέθοδος κυρίων συνιστωσών) γνωστή και ως *PCA* που είδαμε στην Παράγραφο 2.2.5. Έτσι μέσω της *PCA* και της python παίρνουμε την παρακάτω εικόνα για το πως κινούνται τα δεδομένα μας ανά κατηγορία.



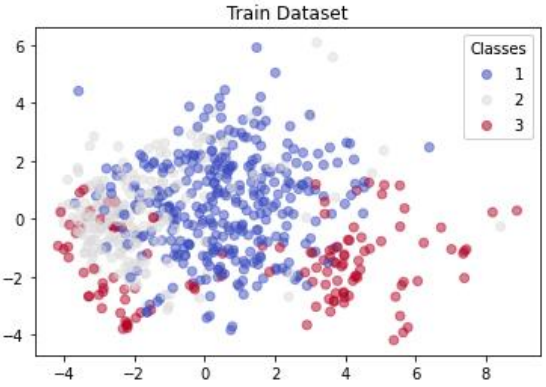
Σχήμα 6.3.7 Οπτικοποίηση του συνόλου δεδομένων μέσω της *PCA*)

Η εικόνα αυτή καταφέρνει μέσα από 2 μόνο κύριες συνιστώσες να μας εξηγεί το 76% περίπου της ολικής εικόνας των δεδομένων μας.



Σχήμα 6.3.8 Ποσοστό επεξηγηματικότητας της κάθε κύριας συνιστώσας

Στην συνέχεια θα προβούμε σε διαχωρισμό του συνόλου των δεδομένων μας, σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου γνωστό και ως *train-test split.* Ο λόγος που το κάνουμε αυτό είναι για εκμηδενίσουμε τυχόν μεροληψία που μπορεί να υπάρξει στην αξιολόγηση του εκάστοτε μοντέλου. Ο διαχωρισμός στον οποίο θα προβούμε είναι η χρήση του 70% των δεδομένων για εκπαίδευση και η χρήση του 30% για αξιολόγηση. Ακόμα εκτός από τον διαχωρισμό των δεδομένων μας θα προβούμε στην τυποποίηση των δεδομένων εκπαίδευσης και με τις παραμέτρους αυτών θα τυποποιήσουμε αντίστοιχα και τα δεδομένα ελέγχου. Η εφαρμογή της τυποποίησης προφανώς θα γίνει μόνο για τις ποσοτικές μεταβλητές και όχι για τα επίπεδα των παραγόντων. Στην συνέχεια προβαίνουμε και στην οπτικοποίηση του κάθε συνόλου δεδομένων (*train-test)* για να έχουμε μία εικόνα σύγκρισης για όταν γνωρίζομαι το *label* και όταν μας είναι άγνωστο.



Σχήμα 6.3.9 Οπτικοποίηση του train dataset

Σχήμα 6.3.10 Οπτικοποίηση του test dataset

Σχήμα 6.3.11 Οπτικοποίηση του test dataset με άγνωστο label

6.4 Χρήση των μοντέλων Support Vector Machines για εκτίμηση του δείκτη υγείας

a) Προσαρμογή Linear SVM

Έχοντας κάνει τις κατάλληλες προσαρμογές στα δεδομένα μας μπορούμε πλέον να προβούμε στην προσαρμογή του μοντέλου SVM. Αρχικά θα προσαρμόσουμε το γραμμικό SVM και θα ορίσουμε την trade-off παράμετρο *C=1*. Όπως αναφέρθηκε στο Κεφάλαιο 3, όσο μεγαλύτερη τιμή έχει η παράμετρος αυτή τόσο πιο αυστηρό γίνεται το μοντέλο μας ως προς τις λανθασμένες ταξινομήσεις. Μέσω της Python λοιπόν και της βιβλιοθήκης scikit-learn προσαρμόζουμε το μοντέλο και στην συνέχεια δημιουργούμε το classification report το οποίο βασίζεται πάνω στο test dataset και όχι σε αυτό που το μοντέλο έχει εκπαιδευτεί. Το classification report αποτελεί ένα πίνακα που περιέχει ένα σύνολο εξωτερικών μέτρων που είδαμε στην Παράγραφο 2.2.2 για να λάβουμε μία γενική εικόνα απόδοσης του μοντέλου μας.

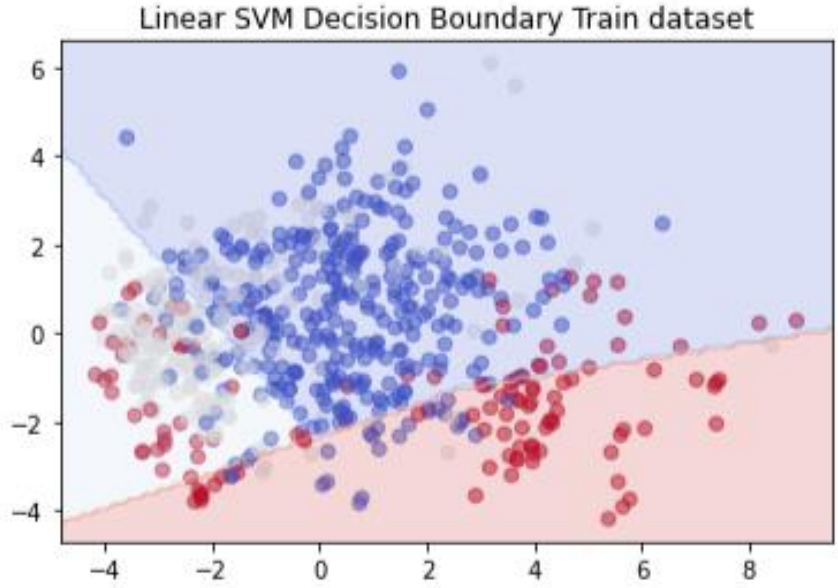
Classification report for Linear SVM:

	precision	recall	f1-score	support
1	0.93	0.90	0.92	145
2	0.78	0.84	0.81	87
3	0.85	0.80	0.83	51
accuracy			0.87	283
macro avg	0.85	0.85	0.85	283
weighted avg	0.87	0.87	0.87	283

Σχήμα 6.4.1 Classification Report Linear SVM

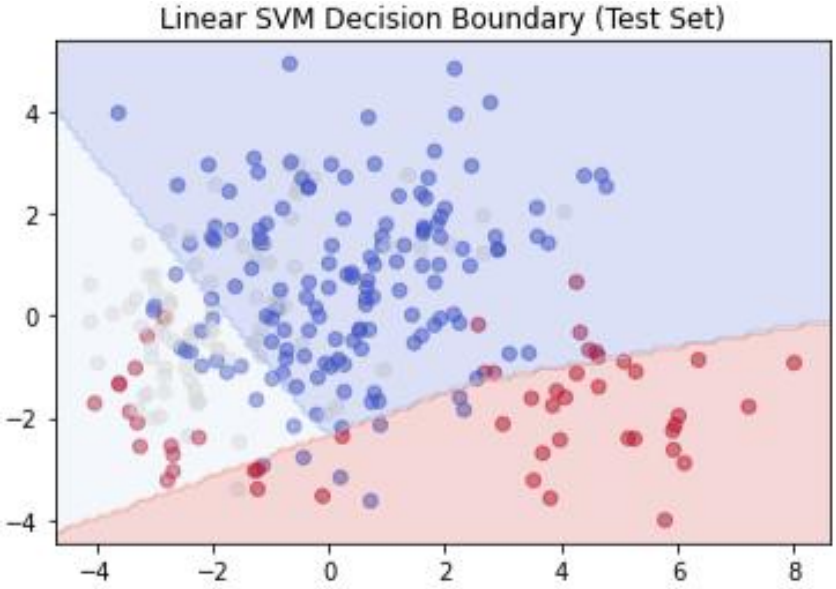
Μπορούμε μέσω του Classification Report να δούμε τις τιμές των εξωτερικών μέτρων. Στις τρείς πρώτες στήλες έχουμε τις τιμές των Precision,Recall και F1-Score ανά κατηγορία του label μας και ως τελευταία στήλη έχουμε το πλήθος των παρατηρήσεων για κάθε κατηγορία που υπάρχουν στο test dataset.Ακόμα από κάτω μπορούμε να δούμε την τιμή του accuracy και στις 2 τελευταίες γραμμές τις τιμές των Precision,Recall και F1-Score κατά μέσο όρο και κατά σταθμισμένο μέσο όρο. Σε αυτό το μοντέλο βλέπουμε να έχουμε ως πρώτη εικόνα ικανοποιητικά αποτελέσματα καθώς για δεδομένα εκπαίδευσης έχουμε περίπου 85-87% σε όλα τα εξωτερικά μέτρα.

Στην συνέχεια προβαίνουμε στον γραφική αναπαράσταση των δεδομένων μας στον R^2 μέσω της τεχνικής *PCA* για να δούμε πως δημιουργούνται οι βέλτιστες ευθείες που χωρίζουν τα δεδομένα μας.



Σχήμα 6.4.2 Decision Boundary Linear SVM train dataset

Μπορούμε να διακρίνουμε μέσα από το γράφημα αυτό ότι οι γραμμές διαχωρίζουν σε ένα ικανοποιητικό βαθμό τα δεδομένα παρόλο που αυτά είναι μπλεγμένα μεταξύ. Παρακάτω φαίνεται η εικόνα του συνόλου δεδομένων ελέγχου πάνω στα σύνορα που καθόρισαν τα δεδομένα εκπαίδευσης.



Σχήμα 6.4.3 Decision Boundary Linear SVM test dataset

Παρατηρούμε ότι τα δεδομένα εκπαίδευσης έχουν ταξινομηθεί σε έναν ικανοποιητικό βαθμό δοθέντος ότι μιλάμε για γραμμικό διαχωρισμό των δεδομένων μας.

b) Προσαρμογή Gaussian SVM

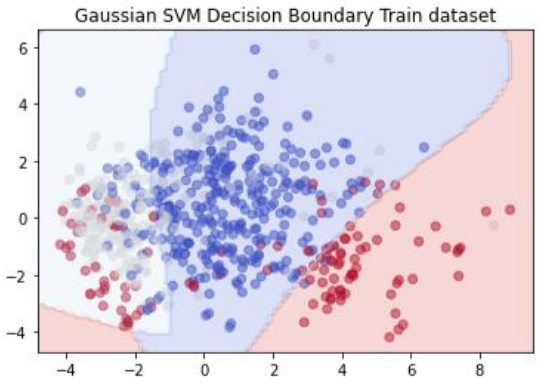
Στην συνέχεια προβαίνουμε και στην προσαρμογή του Gaussian *SVM* με *C*=1.

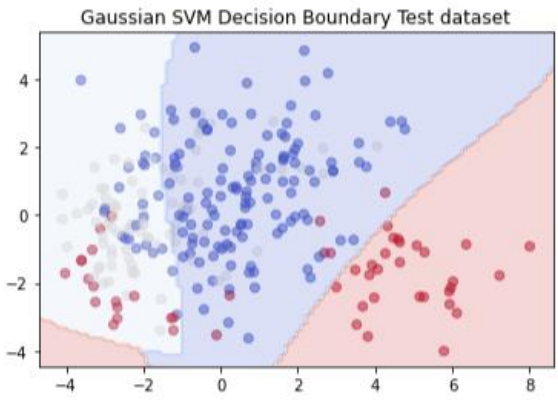
Classification report for Gaussian SVM:

	precision	recall	f1-score	support
1	0.92	0.90	0.91	145
2	0.76	0.85	0.80	87
3	0.95	0.80	0.87	51
accuracy			0.87	283
macro avg	0.88	0.85	0.86	283
weighted avg	0.88	0.87	0.87	283

Σχήμα 6.4.4 Classification Report Gaussian SVM

Με την χρήση του Gaussian kernel βλέπουμε να έχουμε μία ελαφρά βελτίωση των αποτελεσμάτων μας ιδιαίτερη ως προς την κλάση 3.Η συνολική εικόνα των εξωτερικών μέτρων πάλι βλέπουμε να κινείται κοντά στο 85-87%.Αξίζει να δούμε την εικόνα με βάση την οποία διαχωρίζονται τα δεδομένα μας και στο κομμάτι της εκπαίδευσης και στο κομμάτι του ελέγχου.





Σχήμα 6.4.5 Gaussian SVM για train dataset Σχήμα 6.4.6 Gaussian SVM για test dataset

Τονίζουμε ότι στην περίπτωση του Gaussian SVM τα σύνορα ανάμεσα στις ομάδες δεν είναι γραμμικά. Ειδικότερα βλέπουμε στην κάτω αριστερά γωνία να υπάρχει ένα μικρό χωρίο το οποίο ταξινομεί παρατηρήσεις στην κλάση 2 παρόλο που υπάρχει και ένα μεγαλύτερο στην δεξιά μεριά. Εκ του αποτελέσματος βλέποντας το πως κινούνται τα δεδομένα μας αν αυτό το χωρίο ήταν ελαφρά πιο ψηλό και περιλάμβανε και μερικές από τις παρατηρήσεις της κατηγορίας 2 που βρίσκονται εκεί, καταλαβαίνουμε ότι θα είχαμε καλύτερα αποτελέσματα και ειδικά ως προς την ύποπτη κατάσταση (2) η οποία είναι αυτή που υστερεί ως προς τα εξωτερικά μέτρα συγκριτικά με τις άλλες.

c) Προσαρμογή SVM με πολυωνυμικό πυρήνα τρίτου βαθμού

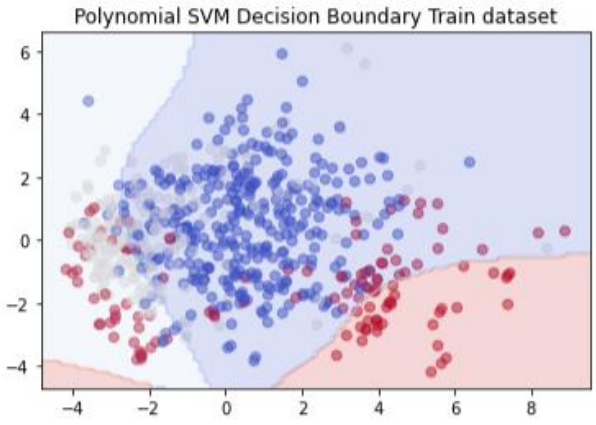
Το επόμενο μοντέλο είναι το πολυωνυμικό SVM τρίτου βαθμού με τιμή της trade-off παραμέτρου *C=1*.

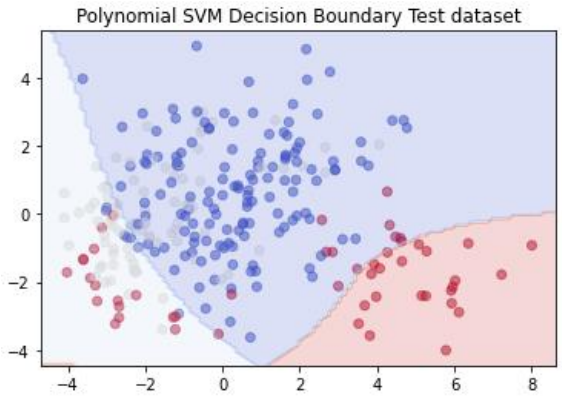
Classification report for Polynomial SVM:

	precision	recall	f1-score	support
1	0.83	0.94	0.88	145
2	0.78	0.69	0.73	87
3	1.00	0.80	0.89	51
accuracy			0.84	283
macro avg	0.87	0.81	0.84	283
weighted avg	0.85	0.84	0.84	283

Σχήμα 6.4.7 Classification Report Polynomial SVM

Παρατηρούμε ότι σαν συνολική εικόνα απόδοσης του μοντέλου έχουμε εμφανώς πιο μικρά νούμερα το οποίο μας προϊδεάζει ότι ίσως ο πολυωνυμικός πυρήνας να μην είναι ο καταλληλότερος, καθώς βλέπουμε μια γενική απόδοση κοντά στο 81-84%.Ενδιαφέρον έχει να παρατηρήσουμε ότι παρόλο που συνολικά έχουμε την χειρότερη απόδοση σε σύγκριση με τα άλλα δύο μοντέλα ως προς την pathological κατηγορία έχουμε τα υψηλότερα ποσοστά και με το ποσοστό των θετικών προβλέψεων ανάμεσα σε όλες τις θετικές προβλέψεις για την κατηγορία αυτή να αγγίζει το 100%.Παρακάτω βλέπουμε και τα αντίστοιχα γραφήματα για train και test dataset.





Σχήμα 6.4.8 Polynomial SVM για train dataset Σχήμα 6.4.9 Polynomial SVM για test dataset

Βλέπουμε και σε αυτήν την περίπτωση μία εικόνα ικανοποιητική σε γενικό βαθμό, και ότι το κύριο πρόβλημα της ταξινόμησης με την χρήση αυτού του SVM αφορά την κατηγορία 2 που ακόμα και εκεί έχουμε υψηλά ποσοστά απλώς είναι χαμηλότερα από τις 2 προηγούμενες επιλογές.

d) Βελτιστοποίηση του μοντέλου

Στα προηγούμενα βήματα είδαμε τις αποδόσεις ως προς 3 πυρήνες για συγκεκριμένη τιμή της trade-off παραμέτρου και συγκεκριμένες τιμές των παραμέτρων του κάθε πυρήνα. Στο επόμενο βήμα θα προβούμε στην βελτιστοποίηση του μοντέλου. Ο τρόπος που θα το κάνουμε αυτό είναι μέσω μίας συνάρτησης της Scikit-learn η οποία λέγεται GridSearchCV. Με την συνάρτηση αυτή θα δώσουμε διάφορες τιμές της παραμέτρου *C*, διαφορετικά Kernels και διαφορετικές τιμές για τις παραμέτρους του κάθε Kernel. Η συνάρτηση αρχίζει μία επαναληπτική διαδικασία και χρησιμοποιώντας Cross-Validation θα μας επιστρέψει ως αποτέλεσμα ποιο μοντέλο και με ποιες παραμέτρους πετυχαίνει την καλύτερη απόδοση κατά μέσο όρο με βάση ένα μέτρο αξιολόγησης της επιθυμίας μας.

Θα βελτιστοποιήσουμε ως προς το f1-score που παρέχει μία πιο γενική εικόνα και ως τιμή του k στο Cross-Validation (πλήθος ομάδων στις οποίες θα χωρίσουμε τα δεδομένα μας) θα χρησιμοποιήσουμε την τιμή 3. Ακόμα θα ελέγξουμε για τα 3 πιθανά kernels που είδαμε και πριν, στην περίπτωση της πολυωνυμικής θα ελέγξουμε από πολυωνυμικά kernels πρώτου βαθμού μέχρι τετάρτου ενώ ως προς για την γ παράμετρο του RBF kernel θα αφήσουμε να κινηθεί στις τιμές $\gamma \in \{0.1,1,10, scale\}$.Διευκρινίζεται ότι η τιμή scale στην Python για την τιμή γ σημαίνει ότι έχουμε την περίπτωση του Gaussian kernel. Τέλος για την trade-off παράμετρο *C* θα βάλουμε ένα μεγάλο εύρος από πολύ χαλαρή ως προς πολύ αυστηρή ώστε να δούμε την κατάλληλη τιμή της, πιο συγκεκριμένα

$$C \in \{0.1,1,10,1000\}.$$

```
from sklearn.model_selection import GridSearchCV
# Define the parameter grid for GridSearchCV
param_grid = {'kernel': ['linear', 'rbf', 'poly'], 'C': [0.1, 1, 10, 100, 1000],"degree":[1,2,3,4],"gamma":["scale",1,0.1,10]}

# Create a GridSearchCV object with the SVC classifier, parameter grid, and f1_score as the scoring metric
grid_search = GridSearchCV(estimator=SVC(), param_grid=param_grid, scoring='f1_weighted',cv=3)

# Fit the GridSearchCV object to the training data
grid_search.fit(X_train_scaled, y_train)

# Get the best parameters and best score
best_params = grid_search.best_params_
best_score = grid_search.best_score_

# Print the best parameters and best score
print("Best parameters: ", best_params)
print("Best score: ", best_score)
```

```
Best parameters:  {'C': 10, 'degree': 1, 'gamma': 0.1, 'kernel': 'rbf'}
Best score:  0.8766129853754173
```

Σχήμα 6.4.10 Αποτέλεσμα του GridSearchCV αλγορίθμου

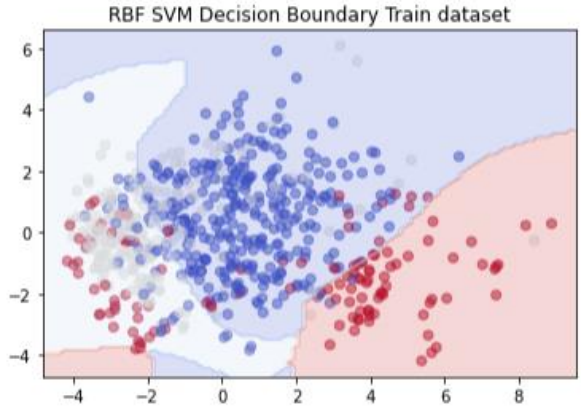
Βλέπουμε ότι το καλύτερο κατά μέσο όρο f1-score στο σύνολο δεδομένων εκπαίδευσης ισούται με 87.6% και το πετυχαίνουμε με χρήση ενός RBF kernel με τιμή της αντίστοιχης παραμέτρου *γ=0.1* και παράμετρο χαλαρότητας *C=10*. Μετά την προσαρμογή αυτού του SVM έχουμε ότι.

Classification report for RBF SVM:

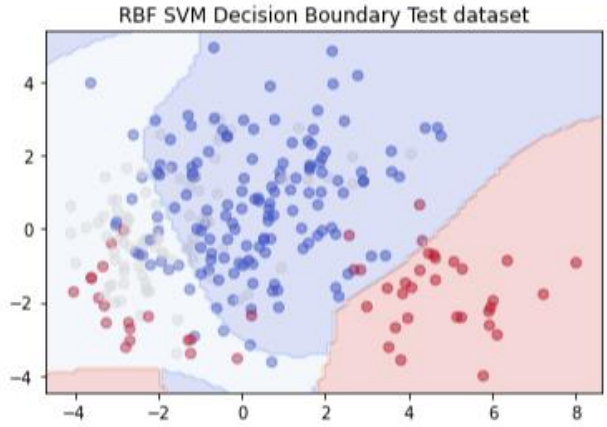
	precision	recall	f1-score	support
1	0.93	0.92	0.93	145
2	0.85	0.86	0.86	87
3	0.88	0.88	0.88	51
accuracy			0.90	283
macro avg	0.89	0.89	0.89	283
weighted avg	0.90	0.90	0.90	283

Σχήμα 6.4.11 Classification Report RBF SVM μετά το GridSearchCV

Βλέπουμε μία γενική αύξηση σε όλα τα μέτρα τα οποία πλέον κινούνται στο 89-90% και για την ομάδα 2 που υπήρχε ένα μικρό πρόβλημα πριν, έχουμε πλέον ανεβασμένα ποσοστά. Στην συνέχεια προβαίνουμε στην γραφική αναπαράσταση των συνόρων διαχωρισμού των ομάδων.



Σχήμα 6.4.12 Decision Boundary RBF SVM για Train dataset μετά το GridSearchCV



Σχήμα 6.4.13 Decision Boundary RBF SVM για Test dataset μετά το GridSearchCV

Παρατηρούμε ότι πλέον το μοντέλο μας περιγράφει σε πολύ καλύτερο βαθμό τα δεδομένα μας και ότι έχουμε λίγα σημεία τα οποία έχουν ταξινομηθεί λανθασμένα.

e) Feature Selection

Ως επόμενο βήμα θα είναι η επιλογή των τελικών χαρακτηριστικών που θα κρατήσουμε στο μοντέλο. Στόχος μας είναι είτε να μειώσουμε τις μεταβλητές που χρειαζόμαστε για να κάνουμε μία ικανοποιητική και αξιόπιστη πρόβλεψη όπως τώρα, είτε μειώνοντας τις μεταβλητές να πετύχουμε βελτίωση της απόδοσης του μοντέλου. Ο τρόπος που θα το κάνουμε αυτό είναι μέσω της μεθόδου forward-selection και χρήση Cross-Validation με διαχωρισμό του συνόλου δεδομένων στα 2.Ως κριτήριο βελτίωσης του μοντέλου θα έχουμε και πάλι το f1-score. Έτσι μέσω της Python παίρνουμε τα εξής αποτελέσματα.

```
from mlxtend.feature_selection import SequentialFeatureSelector
sfs_selector_f1 = SequentialFeatureSelector(rbf_svm, k_features=len(X_train.columns), cv=2,scoring="f1_weighted")
sfs1_f1=sfs_selector_f1.fit(X_train_scaled,y_train)
sfs1pdf_f1 = pd.DataFrame(sfs1_f1.subsets_)
sfs1pdf_f1 = sfs1pdf_f1.T
sfs1pdf_f1
```

	feature_idx	cv_scores	avg_score	feature_names
1	(7,)	[0.6939410971061555, 0.6898745759029331]	0.691908	(percentage_of_time_with_abnormal_long_term_va...
2	(7, 15)	[0.8171487957124948, 0.7798893998995272]	0.798519	(percentage_of_time_with_abnormal_long_term_va...
3	(1, 7, 15)	[0.852614895009798, 0.8196934102241058]	0.836154	(accelerations, percentage_of_time_with_abnorm...
4	(1, 5, 7, 15)	[0.8668271940201766, 0.8375134162133681]	0.85217	(accelerations, abnormal_short_term_variabilit...
5	(0, 1, 5, 7, 15)	[0.888917014229659, 0.8563468361294609]	0.872632	(baseline.value, accelerations, abnormal_short...
6	(0, 1, 3, 5, 7, 15)	[0.8917817198574296, 0.8690490043304427]	0.880415	(baseline.value, accelerations, uterine_contra...
7	(0, 1, 3, 5, 7, 15, 20)	[0.8945500705984578, 0.8693445496038021]	0.881947	(baseline.value, accelerations, uterine_contra...
8	(0, 1, 3, 5, 7, 15, 20, 21)	[0.8949156528046193, 0.8724675666926661]	0.883692	(baseline.value, accelerations, uterine_contra...
9	(0, 1, 3, 5, 7, 15, 20, 21, 22)	[0.8979283597580127, 0.8724675666926661]	0.885198	(baseline.value, accelerations, uterine_contra...
10	(0, 1, 3, 5, 7, 15, 20, 21, 22, 23)	[0.8979283597580127, 0.8724675666926661]	0.885198	(baseline.value, accelerations, uterine_contra...
11	(0, 1, 3, 5, 7, 15, 18, 20, 21, 22, 23)	[0.8949156528046193, 0.8724675666926661]	0.883692	(baseline.value, accelerations, uterine_contra...
12	(0, 1, 3, 5, 7, 15, 18, 20, 21, 22, 23, 25)	[0.8975559262400331, 0.8661289005961568]	0.881842	(baseline.value, accelerations, uterine_contra...
13	(0, 1, 3, 5, 7, 9, 15, 18, 20, 21, 22, 23, 25)	[0.9065525098656819, 0.875457252946123]	0.891005	(baseline.value, accelerations, uterine_contra...
14	(0, 1, 3, 5, 7, 9, 15, 18, 20, 21, 22, 23, 24,...	[0.9028476563644761, 0.8816201722871558]	0.892234	(baseline.value, accelerations, uterine_contra...
15	(0, 1, 3, 5, 7, 9, 14, 15, 18, 20, 21, 22, 23,...	[0.9001371628938276, 0.8755384400950207]	0.887838	(baseline.value, accelerations, uterine_contra...
16	(0, 1, 3, 5, 7, 9, 12, 14, 15, 18, 20, 21, 22,...	[0.8941267956410714, 0.878576912696868]	0.886352	(baseline.value, accelerations, uterine_contra...
17	(0, 1, 3, 5, 7, 9, 10, 12, 14, 15, 18, 20, 21,...	[0.8910801133023354, 0.8815857635113088]	0.886333	(baseline.value, accelerations, uterine_contra...
18	(0, 1, 3, 5, 7, 9, 10, 12, 14, 15, 16, 18, 20,...	[0.8881029072540393, 0.8726032506282879]	0.880353	(baseline.value, accelerations, uterine_contra...
19	(0, 1, 3, 5, 7, 9, 10, 12, 14, 15, 16, 17, 18,...	[0.8851794247257397, 0.8668094972165704]	0.875994	(baseline.value, accelerations, uterine_contra...
20	(0, 1, 3, 4, 5, 7, 9, 10, 12, 14, 15, 16, 17, ...	[0.8971250597203454, 0.854474071655193]	0.8758	(baseline.value, accelerations, uterine_contra...
21	(0, 1, 3, 4, 5, 6, 7, 9, 10, 12, 14, 15, 16, 1...	[0.8999043268471546, 0.85733404734327]	0.878619	(baseline.value, accelerations, uterine_contra...
22	(0, 1, 3, 4, 5, 6, 7, 9, 10, 12, 14, 15, 16, 1...	[0.8969103310135919, 0.8601946641768027]	0.878552	(baseline.value, accelerations, uterine_contra...
23	(0, 1, 2, 3, 4, 5, 6, 7, 9, 10, 12, 14, 15, 16...	[0.8908877494285103, 0.8691634010812036]	0.880026	(baseline.value, accelerations, fetal_movement...
24	(0, 1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 14, 15...	[0.884892241616651, 0.8633308344786328]	0.874112	(baseline.value, accelerations, fetal_movement...
25	(0, 1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14...	[0.8786753319661178, 0.8478851560125514]	0.86328	(baseline.value, accelerations, fetal_movement...
26	(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,...	[0.8755236475577466, 0.8475612460815564]	0.861542	(baseline.value, accelerations, fetal_movement...

Σχήμα 6.4.14 Feature Selection για SVM

Στην πρώτη στήλη παρατηρούμε τους δείκτες των μεταβλητών που έχουν επιλεγεί ανά βήμα και στην τελευταία τα ονόματα τους, στην δεύτερη στήλη τις τιμές που κυμαίνεται το εξωτερικό μέτρο που επιλέξαμε. Στην τρίτη στήλη την τιμή του CV- score κατά μέσο όρο.

Στην συνέχεια βρίσκουμε την μέγιστη τιμή του f1-score κατά μέσο όρο και βλέπουμε τα αντίστοιχα features

```
sfs1pdf_f1[sfs1pdf_f1.iloc[:,2]==max(sfs1pdf_f1.iloc[:,2])]
```

	feature_idx	cv_scores	avg_score	feature_names
14	(0, 1, 3, 5, 7, 9, 15, 18, 20, 21, 22, 23, 24,...	[0.9028476563644761, 0.8816201722871558]	0.892234	(baseline.value, accelerations, uterine_contra...

```
sfs1pdf_f1.iloc[13,3]#Features που αντιστοιχούν στην μέγιστη τιμή του f1-score
```

```
('baseline.value',
 'accelerations',
 'uterine_contractions',
 'abnormal_short_term_variability',
 'percentage_of_time_with_abnormal_long_term_variability',
 'histogram_width',
 'histogram_mean',
 'severe_decelerations_0.001',
 'prolongued_decelerations_0.002',
 'prolongued_decelerations_0.003',
 'prolongued_decelerations_0.004',
 'prolongued_decelerations_0.005',
 'histogram_tendency_0',
 'histogram_tendency_1')
```

Σχήμα 6.4.15 Επιλεγμένα features για SVM

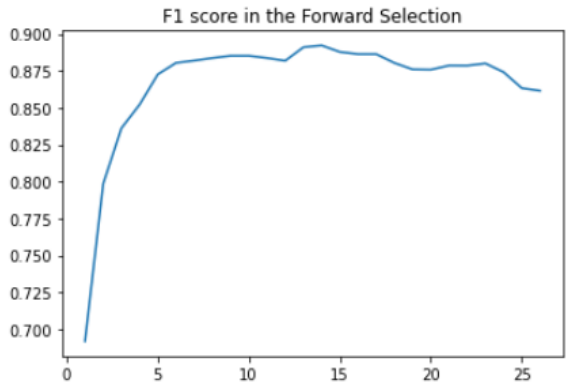
Παρατηρούμε ότι οι μεταβλητές που μας εξασφαλίζουν την μέγιστη απόδοση είναι 10 καθώς υπάρχουν και οι τρεις μεταβλητές που αντιμετωπίζονται ως ποιοτικές. Βλέπουμε το επίπεδο 0.001 της μεταβλητής prolongued_decelarations να μην συμπεριλαμβάνεται. Παρόλα αυτά εμείς θα το συμπεριλάβουμε στο τελικό μοντέλο αφού τουλάχιστον ένα επίπεδο από τον αντίστοιχο παράγοντα κρίνεται στατιστικά σημαντικό ως προς την βελτίωση του f1-score.

Άρα τα τελικά χαρακτηριστικά τα οποία θα συμπεριληφθούν για την εκπαίδευση και προσαρμογή του ελέγχου θα είναι τα

Baseline.value
accelerations
Uterine_contractions
Abnormal_short_term_variability
Percentage_of_time_with_abnormal_long term variability
Histogram witdth
Histogram mean
Sever decelerations
Prolongued decelerations
Histogram tendency

Πίνακας 6.3 Επιλεγμένα features

Παρακάτω βλέπουμε την πορεία του F1-score καθώς εκτελείται το forward-selection.



Σχήμα 6.4.16 Πορεία του f1-score μέσα στο Forward selection

Στην εικόνα αυτήν παρατηρούμε στην αρχή αύξηση του μέτρου μέχρι που μεγιστοποιείται κοντά στο πλήθος 14 και μετά παραμένει σταθερό ή μειώνεται.

f) Προσαρμογή βελτιστοποιημένου Radial Basis Function Support Vector Machines Model

Προχωράμε τώρα στην τελική προσαρμογή του μοντέλου μας. Προσαρμόζουμε το τελικό βελτιστοποιημένο SVM με τον RBF ως πυρήνα , τιμή $\gamma=0.1$ και τιμή της trade-off παραμέτρου C=10. Παρακάτω φαίνεται και ο κώδικας με τον οποίο προσαρμόζεται το τελικό μοντέλο. Με παρόμοιο τρόπο δημιουργήθηκαν και τα προηγούμενα SVM με διαφορετικές παραμέτρους.

```
# Δημιουργία σκελετού του SVM και προσαρμογή του
rbf_svm = SVC(kernel='rbf',C=10,gamma=0.1)
rbf_svm.fit(X_train_scaled, y_train)

# Δημιουργία classification report για βελτιστοποιημένο RBF SVM
y_pred_rbf = rbf_svm.predict(X_test_scaled)
print("Classification report for optimized RBF SVM:\n", classification_report(y_test, y_pred_rbf))

# Χρήση της Principal Component Analysis για την οπτικοποίηση των δεδομένων
pca = PCA(n_components=2)
X_train_pca = pca.fit_transform(X_train_scaled)

# Visualization για δεδομένα συνόλου εκπαίδευσης
fig, ax = plt.subplots()
ax.scatter(X_train_pca[:, 0], X_train_pca[:, 1], c=y_train, cmap='coolwarm', alpha=0.5)
xlim = ax.get_xlim()
ylim = ax.get_ylim()
xx, yy = np.meshgrid(np.linspace(xlim[0], xlim[1], 100), np.linspace(ylim[0], ylim[1], 100))
Z = rbf_svm.predict(pca.inverse_transform(np.c_[xx.ravel(), yy.ravel()]))
Z = Z.reshape(xx.shape)
ax.contourf(xx, yy, Z, cmap='coolwarm', alpha=0.2)
ax.set_title('Optimized RBF SVM Decision Boundary Train dataset')
plt.show()



# Visualization για δεδομένα συνόλου ελέγχου
pca = PCA(n_components=2)
X_test_pca = pca.fit_transform(X_test_scaled)
fig, ax = plt.subplots()
ax.scatter(X_test_pca[:, 0], X_test_pca[:, 1], c=y_test, cmap='coolwarm', alpha=0.5)
xlim = ax.get_xlim()
ylim = ax.get_ylim()
xx, yy = np.meshgrid(np.linspace(xlim[0], xlim[1], 100), np.linspace(ylim[0], ylim[1], 100))
Z = rbf_svm.predict(pca.inverse_transform(np.c_[xx.ravel(), yy.ravel()]))
Z = Z.reshape(xx.shape)
ax.contourf(xx, yy, Z, cmap='coolwarm', alpha=0.2)
ax.set_title('RBF SVM Decision Boundary Test dataset')
plt.show()
#Δημιουργία πίνακα συγχίσεως σε μορφή heatmap
conf_mat = confusion_matrix(y_test, y_pred_rbf)
labels = ['1', '2', '3']
conf_mat = pd.DataFrame(conf_mat, index=labels, columns=labels)
ax=sns.heatmap(conf_mat,annot=True,cmap="Purples",fmt='')
ax.set_xlabel('Predicted Values')
ax.set_ylabel('Actual Values ')
```

Σχήμα 6.4.17 Κώδικας δημιουργίας optimized SVM

Αρχικά βλέπουμε τον πίνακα συγχύσεως του μοντέλου ως προς τα δεδομένα του συνόλου εκπαίδευσης και το αντίστοιχο classification report.

Σχήμα 6.4.18 Πίνακας Συγχύσεως optimized SVM

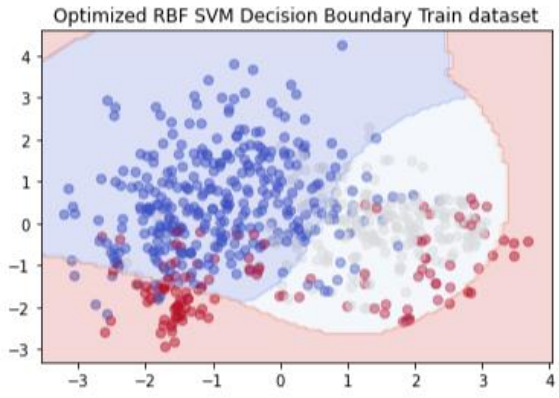
Classification report for optimized RBF SVM:

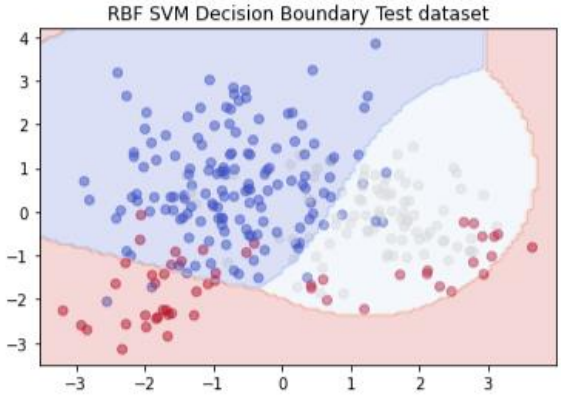
	precision	recall	f1-score	support
1	0.95	0.90	0.92	145
2	0.83	0.92	0.87	87
3	0.96	0.94	0.95	51
accuracy			0.91	283
macro avg	0.91	0.92	0.92	283
weighted avg	0.92	0.91	0.91	283

Σχήμα 6.4.19 Classification Report optimized SVM

Από τον πίνακα συγχύσεως του Σχήματος 6.4.19 βλέπουμε ότι για τις φυσιολογικές (N) περιπτώσεις από τις 145 που έχουμε, συνολικά 130 ταξινομήθηκαν σωστά ενώ 14 ταξινομήθηκαν ως ύποπτες και 1 ως παθολογική. Ως προς την δεύτερη ομάδα των ύποπτων (S) περιπτώσεων, από τις 87 που έχουμε οι 80 ταξινομήθηκαν σωστά ενώ 6 ταξινομήθηκαν ως φυσιολογικές και 1 ως παθολογική. Τέλος ως προς την τρίτη ομάδα των παθολογικών (P) περιπτώσεων, έχουμε ότι 48 από τις 51 περιπτώσεις διαγνώστηκαν από το μοντέλο, ενώ είχαμε 1 περίπτωση που κρίθηκε ως φυσιολογική και 2 περιπτώσεις που κρίθηκαν ως ύποπτες. Από το Σχήμα 6.4.19 βλέπουμε ότι ως προς όλες τις ομάδες και όλες τις μετρικές έχουμε καταφέρει να πετύχουμε τις υψηλότερες επιδόσεις και αξίζει να σημειωθεί ότι το πετύχαμε με τις 10 μόνο μεταβλητές από τις 21 που αρχίσαμε. Άρα μέσω της μείωσης του κόστους της συλλογής δεδομένων καταφέραμε σημαντική πρόοδο στην σωστή ανίχνευση του μοντέλου. Μία γενική εικόνα για το μοντέλο μας από τα ποσοστά του θα λέγαμε ότι είναι κοντά στο 92% για το σύνολο τους, και με πολύ καλύτερα ποσοστά ως προς την ομάδα 3 που αποτελεί την πιο κρίσιμη καθώς έχουμε ένα πρόβλημα ιατρικής φύσεως οπότε η έγκαιρη και ακριβής πρόβλεψη σε αυτό το κομμάτι παίζει κομβικό ρόλο.

Τέλος θα προβούμε στον σχεδιασμό του συνόλου δεδομένων μας. Σημαντικό είναι να διευκρινίσουμε ότι καθώς μειώσαμε το σύνολο των features δεν αναμένουμε να είχαμε την αρχική εικόνα του αντίστοιχου scatterplot.Έτσι για δεδομένα του συνόλου εκπαίδευσης και ελέγχου έχουμε την παρακάτω μορφή.





Σχήμα 6.4.20 Optimized RBF SVM για train dataset Σχήμα 6.4.21 Optimized RBF SVM για test dataset

Στα παραπάνω γραφήματα βλέπουμε μία πολύ πιο ξεκάθαρη εικόνα ως προς τα δεδομένα μας. Βλέπουμε το σύνολο των παθολογικών καταστάσεων να είναι συγκεντρωμένες στην δεξία μεριά, οι ύποπτες περιπτώσεις κυρίως στο κάτω μέρος ενώ το μεγαλύτερο ποσοστό των φυσιολογικών περιπτώσεων να υπάρχει στο πάνω αριστερό μέρος. Η πιο καθαρή διαχώριση των δεδομένων μας που εμφανίζεται στα γραφήματα είναι αυτή που μας δίνει τα υψηλά ποσοστά στα μέτρα αξιολόγησης που είδαμε πριν. Είναι σημαντικό τέλος να διευκρινίσουμε και πάλι ότι, η απεικόνιση αυτήν είναι μία προσέγγιση της πραγματικής εικόνας καθώς έχει γίνει μέσω της τεχνικής principal component analysis, αλλά μας δίνει μία γενική εικόνα συμπεριφοράς των δεδομένων μας.

Ακόμα μπορούμε να δούμε και ποιες από τις παρατηρήσεις ορίζονται ως support vectors δηλαδή κινούνται πάνω στα σύνορα που δημιουργήθηκαν και ορίζουν άμεσα την σημασία αυτών.

	baseline.value	accelerations	uterine_contractions	abnormal_short_term_variability	percentage_of_time_with_abnormal_long_term_variability	histogram_mean
0						
2	0.704604	-0.631753	0.424027	-0.694770	-0.709164	0.679757
16	1.003147	0.555204	1.080744	0.106975	-0.371369	0.902373
52	0.505575	-0.631753	0.424027	0.565114	-0.582491	-0.155054
57	0.903633	0.258465	1.080744	-0.866572	-0.498042	1.069335
70	0.107518	-0.631753	1.409103	0.794184	-0.244696	-0.043746
...	...	...	...	...	...	...
571	-1.285684	-0.631753	-0.232691	1.882266	-0.455818	-0.655941
606	-1.186169	-0.631753	1.080744	-1.553782	-0.709164	-1.602060
615	1.301690	-0.631753	0.424027	1.366859	-0.540267	0.902373
620	-1.285684	-0.631753	-1.217768	1.710463	-0.455818	-0.655941
656	-0.091511	0.555204	-0.889409	0.393312	-0.709164	-0.878557

238 rows × 11 columns

Σχήμα 6.4.22 Classification Report optimized SVM

Βλέπουμε για παράδειγμα ότι η δεύτερη παρατήρηση του συνόλου εκπαίδευσης αποτελεί ένα τέτοιο κομβικό σημείο.

6.5 Χρήση της Πολυωνυμικής παλινδρόμησης για εκτίμηση του δείκτη υγείας

Αφού είδαμε πως το μοντέλο των SVM μπορεί να αντιμετωπίσει με επιτυχία το πρόβλημα μας, θα προβούμε και σε μία εναλλακτική μορφή αντιμετώπισης μέσω της εφαρμογής του μοντέλου της πολυωνυμικής παλινδρόμησης που είδαμε στο Κεφάλαιο 4.

Καθώς έχουμε ένα μεγάλο πλήθος μεταβλητών θα ξεκινήσουμε με το πλήρες μοντέλο και θα προσπαθήσουμε να βελτιώσουμε το μοντέλο αυτό αναλόγως με την απόδοσή του βασισμένοι σε στατιστικούς ελέγχους και άλλα κριτήρια αξιολόγησης. Προσαρμόζουμε αρχικά το πλήρες μοντέλο με όλα τα features του συνόλου εκπαίδευσης. Μέσω της Python έχουμε την εξής εικόνα.

```
model = sm.MNLogit(y_train,X_train)
results = model.fit()
print(results.summary())
```

```
Warning: Maximum number of iterations has been exceeded.
         Current function value: 0.317279
         Iterations: 35
                          MNLogit Regression Results
==============================================================================
Dep. Variable:           fetal_health   No. Observations:                  659
Model:                        MNLogit   Df Residuals:                      605
Method:                           MLE   Df Model:                           52
Date:                Mon, 17 Apr 2023   Pseudo R-squ.:                 0.6912
Time:                        11:22:28   Log-Likelihood:                -209.09
converged:                      False   LL-Null:                       -677.11
Covariance Type:            nonrobust   LLR p-value:                2.135e-162
==============================================================================
```

Σχήμα 6.5.1 Πληροφορίες προσαρμογής του πλήρες μοντέλου

Πρώτα βλέπουμε κάποιες πληροφορίες σχετικά με την προσαρμογή του μοντέλου. Όπως είπαμε και στο Κεφάλαιο 4, η μέθοδος εκτίμησης που γίνεται για τον υπολογισμό των παραμέτρων του μοντέλο είναι η μέθοδος μεγίστης πιθανοφάνειας *(maximum likelihood estimation)* η οποία μας οδηγεί σε ένα σύστημα το οποίο επιλύεται μέσω μίας επαναληπτικής μεθόδου. Οπότε ως κύριο κριτήριο για την καλή προσαρμογή του μοντέλου είναι η σύγκλιση ή όχι της μεθόδου. Βλέπουμε στο Σχήμα 6.5.1 ένα αποτέλεσμα που αναφέρεται ως converged το οποίο μας λέει ακριβώς αυτό, το αν έχει συγκλίνει δηλαδή ο αλγόριθμος. Στην περίπτωση μας βλέπουμε να μας επιστρέφει False κάτι το οποίο μας δείχνει ότι το πλήρες μοντέλο δεν ενδείκνυται για χρήση. Ακόμα βλέπουμε ότι η επαναληπτική μέθοδος έκανε 35 επαναλήψεις κάτι το οποίο είναι υψηλό και οδηγεί στην μη σύγκλιση. Στα δεξιά βλέπουμε κάποιες γενικές πληροφορίες για το μοντέλο μας. Παρακάτω φαίνεται το αποτέλεσμα των εκτιμήσεων των παραμέτρων του μοντέλου και οι στατιστικοί έλεγχοι με τις ελεγχοσυναρτήσεις και το αντίστοιχο p-value.Ακόμα στο μοντέλο πολυωνυμικής λογιστικής παλινδρόμησης έχουμε 3 labels άρα πρέπει ένα εξ αυτών να καθοριστεί ως επίπεδο αναφοράς. Ως επίπεδο αναφοράς παίρνουμε την φυσιολογική ομάδα (N) άρα θα έχουμε 2 μοντέλα ένα για την ύποπτη ομάδα (S) και ένα για την παθολογική (P) συγκριτικά με το επίπεδο αναφοράς που θέσαμε.

fetal_health=2	coef	std err	z	P>|z|	[0.025	0.975]
const	-17.6182	3.171	-5.556	0.000	-23.833	-11.403
baseline.value	-0.1529	0.061	-2.513	0.012	-0.272	-0.034
accelerations	-1053.4087	168.329	-6.258	0.000	-1383.328	-723.490
fetal_movement	9.2918	12.240	0.759	0.448	-14.698	33.282
uterine_contractions	-232.9400	71.131	-3.275	0.001	-372.354	-93.526
light_decelerations	183.1254	139.807	1.310	0.190	-90.891	457.141
abnormal_short_term_variability	0.0960	0.019	5.047	0.000	0.059	0.133
mean_value_of_short_term_variability	-0.2970	0.479	-0.620	0.535	-1.236	0.641
percentage_of_time_with_abnormal_long_term_variability	0.0124	0.011	1.176	0.240	-0.008	0.033
mean_value_of_long_term_variability	-0.0950	0.062	-1.533	0.125	-0.216	0.026
histogram_width	0.0007	nan	nan	nan	nan	nan
histogram_min	0.0078	nan	nan	nan	nan	nan
histogram_max	0.0099	nan	nan	nan	nan	nan
histogram_number_of_peaks	0.1398	0.080	1.745	0.081	-0.017	0.297
histogram_number_of_zeroes	0.0542	0.224	0.242	0.809	-0.385	0.493
histogram_mode	-0.0629	0.050	-1.247	0.212	-0.162	0.036
histogram_mean	0.4336	0.107	4.066	0.000	0.225	0.643
histogram_median	-0.1359	0.126	-1.082	0.279	-0.382	0.110
histogram_variance	0.0386	0.015	2.602	0.009	0.010	0.068
severe_decelerations_0.001	-75.3887	5.6e+20	-1.35e-19	1.000	-1.1e+21	1.1e+21
prolongued_decelerations_0.001	3.8483	1.180	3.262	0.001	1.536	6.161
prolongued_decelerations_0.002	6.8456	1.446	4.734	0.000	4.011	9.680
prolongued_decelerations_0.003	8.2768	1.889	4.381	0.000	4.574	11.980
prolongued_decelerations_0.004	-0.2557	3.88e+04	-6.59e-06	1.000	-7.61e+04	7.61e+04
prolongued_decelerations_0.005	1.1109	3.26e+04	3.4e-05	1.000	-6.4e+04	6.4e+04
histogram_tendency_0	-0.2354	0.859	-0.274	0.784	-1.919	1.449
histogram_tendency_1	-0.4152	1.060	-0.392	0.695	-2.493	1.663

fetal_health=3	coef	std err	z	P>|z|	[0.025	0.975]
const	-22.7280	4.871	-4.666	0.000	-32.275	-13.181
baseline.value	0.1399	0.082	1.699	0.089	-0.022	0.301
accelerations	-930.2924	502.742	-1.850	0.064	-1915.649	55.064
fetal_movement	22.4654	13.440	1.672	0.095	-3.876	48.807
uterine_contractions	-191.9355	103.917	-1.847	0.065	-395.608	11.737
light_decelerations	326.3406	153.604	2.125	0.034	25.282	627.400
abnormal_short_term_variability	0.2065	0.032	6.556	0.000	0.145	0.268
mean_value_of_short_term_variability	-0.7386	0.651	-1.135	0.256	-2.014	0.537
percentage_of_time_with_abnormal_long_term_variability	0.0733	0.016	4.670	0.000	0.043	0.104
mean_value_of_long_term_variability	0.0311	0.101	0.309	0.758	-0.166	0.228
histogram_width	0.0127	1.47e+04	8.67e-07	1.000	-2.87e+04	2.87e+04
histogram_min	0.0099	1.47e+04	6.73e-07	1.000	-2.87e+04	2.87e+04
histogram_max	0.0253	1.47e+04	1.73e-06	1.000	-2.87e+04	2.87e+04
histogram_number_of_peaks	-0.4186	0.162	-2.584	0.010	-0.736	-0.101
histogram_number_of_zeroes	0.3372	0.404	0.835	0.404	-0.454	1.128
histogram_mode	-0.0378	0.062	-0.607	0.544	-0.160	0.084
histogram_mean	0.0464	0.072	0.647	0.518	-0.094	0.187
histogram_median	-0.1290	0.103	-1.252	0.210	-0.331	0.073
histogram_variance	0.0531	0.017	3.207	0.001	0.021	0.086
severe_decelerations_0.001	20.4738	7577.995	0.003	0.998	-1.48e+04	1.49e+04
prolongued_decelerations_0.001	2.7767	1.437	1.932	0.053	-0.041	5.594
prolongued_decelerations_0.002	8.1875	1.763	4.643	0.000	4.731	11.644
prolongued_decelerations_0.003	9.7665	2.304	4.239	0.000	5.251	14.282
prolongued_decelerations_0.004	16.6779	6482.429	0.003	0.998	-1.27e+04	1.27e+04
prolongued_decelerations_0.005	15.8682	5427.132	0.003	0.998	-1.06e+04	1.07e+04
histogram_tendency_0	-0.7963	1.087	-0.733	0.464	-2.927	1.334
histogram_tendency_1	-0.8858	1.348	-0.657	0.511	-3.527	1.755

Σχήμα 6.5.2 Output πλήρες μοντέλου πολυωνυμικής παλινδρόμησης

Όπως αναγράφεται έχουμε στο πάνω μέρος το μοντέλο που αφορά την ομάδα όπου

$$y_i = fetal\ health = 2\ ,$$

ενώ από κάτω

$$y_i = fetal\ health = 3.$$

Επομένως εμείς είτε μέσω του p-value, είτε μέσω της ελεγχοσυνάρτησης και της συνθήκης του ελέγχου του Wald είτε μέσω του διαστήματος εμπιστοσύνης βλέποντας αν περιέχεται το 0 μέσα σε αυτό μπορούμε να διακρίνουμε την σημαντικότητα των μεταβλητών. Όπως είπαμε εμείς

θα εργαστούμε με την τεχνική του Backward-elimination,και καθώς το κύριο πρόβλημα μας σε αυτό το στάδιο είναι η μη σύγκλιση του αλγορίθμου, πρώτο μας μέλημα είναι η λύση αυτού πετώντας μεταβλητές που ίσως δημιουργούν αυτό το πρόβλημα. Κατευθείαν λοιπόν μέσω του Σχήματος 6.5.2 βλέπουμε αρκετές μη στατιστικά σημαντικές μεταβλητές στο πλήρες μοντέλο καθώς υπάρχουν πολλά p-value μεγαλύτερα του 0.05 τα οποία δεν μας αφήνουν να απορρίψουμε τις αντίστοιχες υποθέσεις ότι

$$H_0: \beta_i = 0$$

$$H_1: \beta_i \neq 0$$

για το αντίστοιχο i.

Εκτός όμως από τα υψηλά p-value διακρίνουμε και μερικές τιμές για τις οποίες δεν ορίζεται αντίστοιχο p-value. Πιο συγκεκριμένα βλέπουμε ότι για το μοντέλο που αντιστοιχεί σε fetal_health=2 και για τις μεταβλητές, Histogram_width, histogram_min, histogram_max έχουμε να μην ορίζεται το τυπικό σφάλμα του εκτιμητή και επομένως να μην μπορούν να οριστούν οι αντίστοιχοι έλεγχοι καθώς η ελεγχοσυνάρτηση του Wald χρειάζεται το τυπικό σφάλμα του εκτιμητή στον παρονομαστή. Πάμε να δούμε την επίδοση του μοντέλου στο κομμάτι της ταξινόμησης καθώς μας ενδιαφέρει να δούμε αν αυτή θα αυξηθεί μετά την αφαίρεση μη στατιστικά σημαντικών μεταβλητών στο πλήρες μοντέλο. Παρακάτω φαίνεται ο αντίστοιχος κώδικας που θα μας παρέχει τα επιθυμητά αποτελέσματα.

```
#δημιουργούμε της προβλέψεις μας
probabilities = results.predict(X_test)
predictions = probabilities.idxmax(axis=1)+1
print(predictions)
#Κατασκευάζουμε το classification report
print(classification_report(y_test,predictions ))
#κατασκευάζουμε το confusion matrix
conf_mat = confusion_matrix(y_test, predictions)
labels = ['1', '2', '3']
conf_mat = pd.DataFrame(conf_mat, index=labels, columns=labels)
ax=sns.heatmap(conf_mat,annot=True,cmap="Blues",fmt='')
ax.set_xlabel('Predicted Values')
ax.set_ylabel('Actual Values ')
#επιστρέφουμε τα κριτηρια αξιολόγησης
print("AIC",results.aic)
print("BIC",results.bic)
print("pseudo R-squared",results.prsquared)
print("Deviance",-2*results.llf)
#αποθηκεύουμε κριτήρια αξιολόγησης για μελλοντική σύγκριση
list_acc.append(accuracy_score(y_test,predictions ))
list_rsq.append(results.prsquared)
list_aic.append(results.aic)
list_bic.append(results.bic)
list_dev.append(-2*results.llf)
list_rec.append(recall_score(y_test,predictions,average="weighted"))
list_f1.append(f1_score(y_test,predictions,average="weighted" ))
```

Σχήμα 6.5.3 Κώδικας δημιουργίας Πίνακα Συγχύσεως και Classification Report

```
Length: 283, dtype: int64
              precision    recall  f1-score   support

           1       0.93      0.92      0.93       145
           2       0.80      0.79      0.80        87
           3       0.81      0.84      0.83        51

    accuracy                           0.87       283
   macro avg       0.85      0.85      0.85       283
weighted avg       0.87      0.87      0.87       283

AIC 526.173699979155
BIC 768.6727708422904
pseudo R-squared 0.6912074810202109
Deviance 418.173699979155
```

Σχήμα 6.5.4 Classification Report στο πλήρες μοντέλο Σχήμα 6.5.5 Πίνακας Συγχύσεως στο πλήρες μοντέλο

Μπορούμε να παρατηρήσουμε ότι παρά τη μη σύγκλισης του αλγορίθμου τα αποτελέσματα στο κομμάτι της ταξινόμησης είναι ικανοποιητικά. Γενικότερα βλέπουμε να έχουμε στα εξωτερικά μέτρα τιμές περίπου στο 85% και μία καλή τιμή ως προς το ψευδό-Rsquared.

Επειδή έχουμε 3 μεταβλητές οι οποίες το αντίστοιχο τυπικό σφάλμα τους δεν ορίζεται ίσως πρέπει να εξαιρεθούν, καθώς αυτό πιθανόν να ευθύνεται για την μη σύγκλιση της επαναληπτικής διαδικασίας, αλλά αυτό θα το δούμε σταδιακά. Παρακάτω φαίνεται η διαδικασία του Backward Elimination με κριτήριο τα *p–values* μέχρι να επιτύχουμε την σύγκλιση του αλγορίθμου. Σε κάθε βήμα βλέπουμε αν η αντίστοιχη μεταβλητή δεν κρίνεται στατιστικά σημαντική και στα 2 υπό μοντέλα της πολυωνυμικής παλινδρόμησης. Στην αρχή αν δεν κριθούν στατιστικά σημαντικές θα εστιάσουμε στις 3 μεταβλητές όπου το αντίστοιχο *p–value* δεν ορίζεται και στην συνέχεια θα προχωρήσουμε και στις υπόλοιπες μεταβλητές. Τα αντίστοιχα πλήρη Outputs περιλαμβάνονται στο Παράρτημα.

Μεταβλητή η οποία εξαιρέθηκε από το μοντέλο του προηγούμενου βήματος	**Converged**
Πλήρες μοντέλο	**FALSE**
Histogram width	**FALSE**
Histogram max	**FALSE**
Histogram min	**FALSE**
Sever decelerations	**FALSE**
Histogram Number of zeros	**FALSE**
Histogram Median	**FALSE**
Histogram Tendency	**FALSE**
Prolongues decelerations	**TRUE**
Mean value of short term variability	**TRUE**
Light Decelerations	**TRUE**
Histogram number of peaks	**TRUE**

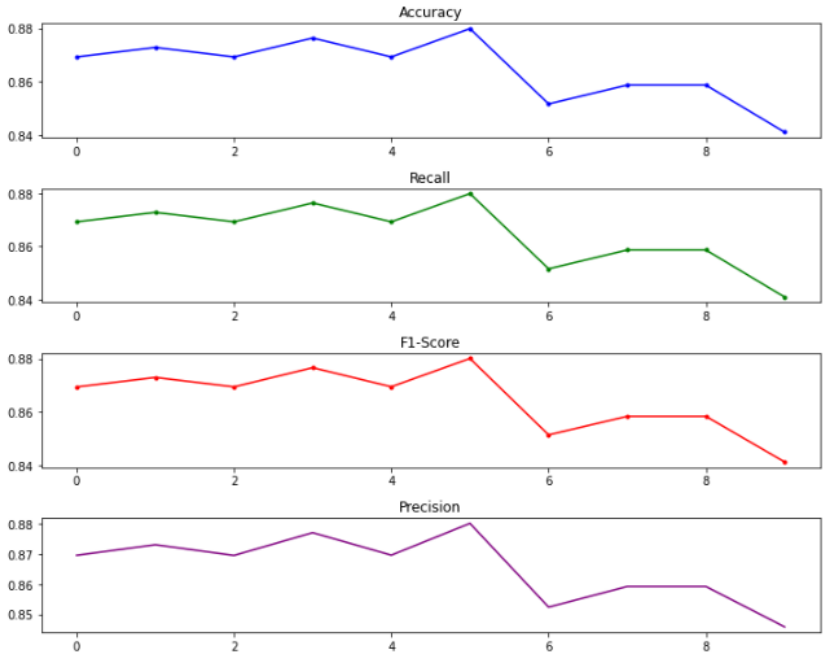
Πίνακας 6.4 Αφαίρεση χαρακτηριστικών για την σύγκλιση του αλγορίθμου

Στον Πίνακα 6.4 φαίνεται η διαδικασία του Backward Elimination. Σε κάθε βήμα αντίστοιχα αφαιρούνταν η αντίστοιχη μεταβλητή από το μοντέλο του προηγούμενου βήματος μέχρι να πετύχουμε την σύγκλιση και την δημιουργία ενός μοντέλου το οποίο όλες οι μεταβλητές είναι στατιστικά σημαντικές για τουλάχιστον ένα από τα 2 υπό μοντέλα της πολυωνυμικής παλινδρόμησης. Παρακάτω φαίνονται οι πληροφορίες για το μοντέλο αυτό.

```
Optimization terminated successfully.
         Current function value: 0.375283
         Iterations 12
                          MNLogit Regression Results
==============================================================================
Dep. Variable:           fetal_health   No. Observations:                  659
Model:                        MNLogit   Df Residuals:                      637
Method:                           MLE   Df Model:                           20
Date:                Mon, 17 Apr 2023   Pseudo R-squ.:                  0.6348
Time:                        11:22:37   Log-Likelihood:                -247.31
converged:                       True   LL-Null:                       -677.11
Covariance Type:            nonrobust   LLR p-value:                3.084e-169
=========================================================================================================================
                                     fetal_health=2       coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------------------
const                                                -11.7872      2.454     -4.804      0.000     -16.596      -6.978
baseline.value                                        -0.0770      0.041     -1.872      0.061      -0.158       0.004
accelerations                                       -921.9140    133.632     -6.899      0.000   -1183.829    -659.999
fetal_movement                                         7.8516      6.764      1.161      0.246      -5.406      21.109
uterine_contractions                                -231.4114     57.821     -4.002      0.000    -344.739    -118.084
abnormal_short_term_variability                        0.0644      0.013      4.809      0.000       0.038       0.091
percentage_of_time_with_abnormal_long_term_variability 0.0091      0.009      0.996      0.319      -0.009       0.027
mean_value_of_long_term_variability                   -0.1170      0.044     -2.672      0.008      -0.203      -0.031
histogram_mode                                        -0.0726      0.033     -2.221      0.026      -0.137      -0.009
histogram_mean                                         0.2221      0.044      5.101      0.000       0.137       0.307
histogram_variance                                     0.0565      0.011      5.338      0.000       0.036       0.077
-------------------------------------------------------------------------------------------------------------------------
                                     fetal_health=3       coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------------------
const                                                -13.1445      3.793     -3.465      0.001     -20.579      -5.710
baseline.value                                         0.1698      0.048      3.507      0.000       0.075       0.265
accelerations                                      -1222.4714    325.092     -3.760      0.000   -1859.640    -585.303
fetal_movement                                        17.5169      6.799      2.576      0.010       4.190      30.843
uterine_contractions                                -190.6177     88.702     -2.149      0.032    -364.471     -16.764
abnormal_short_term_variability                        0.1645      0.027      6.155      0.000       0.112       0.217
percentage_of_time_with_abnormal_long_term_variability 0.0526      0.012      4.380      0.000       0.029       0.076
mean_value_of_long_term_variability                   -0.0129      0.067     -0.193      0.847      -0.145       0.119
histogram_mode                                        -0.1234      0.036     -3.422      0.001      -0.194      -0.053
histogram_mean                                        -0.0409      0.039     -1.045      0.296      -0.117       0.036
histogram_variance                                     0.0612      0.013      4.799      0.000       0.036       0.086
=========================================================================================================================
```

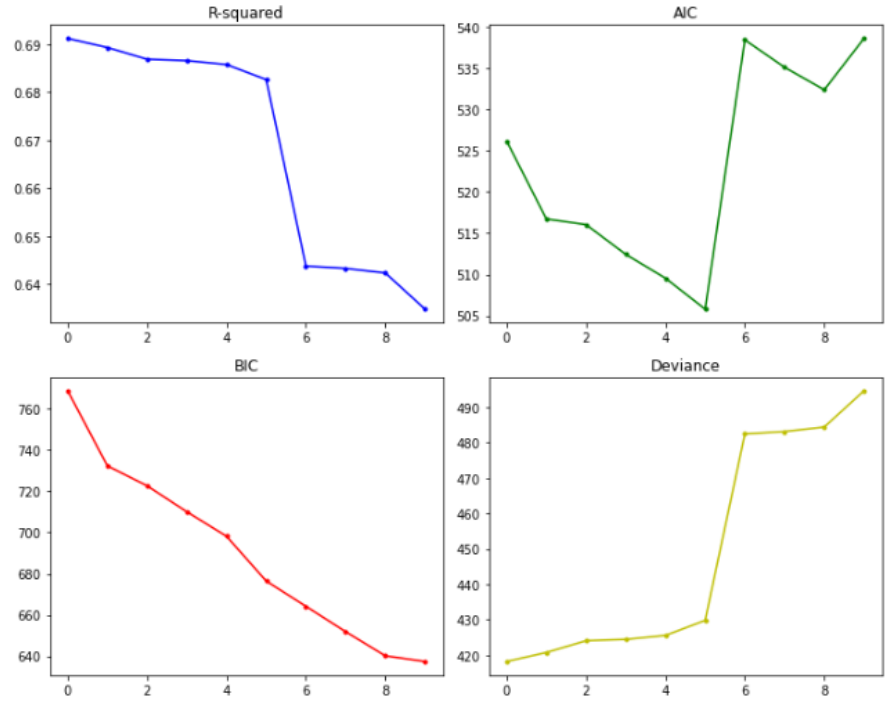
Σχήμα 6.5.6 Μοντέλο που προέκυψε μετά το Backward Elimination

Μπορούμε να διακρίνουμε ορισμένα *p-values* να είναι μεγαλύτερα του 0.05 αλλά τα αντίστοιχα *p-values* για την ίδια μεταβλητή για το άλλο μοντέλο είναι σημαντικά οπότε πρέπει να κρατήσουμε της μεταβλητές αυτές. Ακόμα αξίζει να δούμε στο τελικό μοντέλο αυτό μετά την διαδικασία του Backward-Elimination την απόδοση του στο σύνολο ελέγχου. Παρακάτω βλέπουμε την πορεία των εξωτερικών μέτρων κατά την διάρκεια του feature selection.



Σχήμα 6.5.7 Πορεία των εξωτερικών μέτρων κατά το Backward Elimination

Ως προς τα εξωτερικά μέτρα παρατηρούμε ότι καθώς προχωρούσαμε σε αφαίρεση μεταβλητών υπήρχε μία ελαφρά αύξηση αυτών ενώ σε ένα από τα βήματα του backward-elimination είχαμε μία πτώση σε όλα τα μέτρα η οποία αυτήν δεν ανατράπηκε μετά. Η πτώση αυτή δεν είναι μεγάλη καθώς κινούμαστε πάντα ανάμεσα στο 85-88%.



Σχήμα 6.5.8 Πορεία των ειδικών μέτρων αξιολόγησης κατά το Backward Elimination

Στα ειδικά μέτρα αξιολόγησης της πολυωνυμικής παλινδρόμησης βλέπουμε για το ψευδό R τετράγωνο να έχει μία αμελητέα πτώση η οποία έχει μία απότομη μεταβολή και αυτή στο ίδιο βήμα με την πτώση των εξωτερικών μέτρων. Για τα πληροφοριακά κριτήρια βλέπουμε να μειώνονται καθώς αφαιρούμε μεταβλητές το οποίο είναι καλό αλλά για το πληροφοριακό κριτήριο το Akaike βλέπουμε να έχουμε μια ραγδαία αύξηση πάλι στο ίδιο βήμα ενώ το πληροφοριακό κριτήριο του Bayes που δίνει μεγαλύτερη έμφαση στο πλήθος των παραμέτρων δεν επηρεάστηκε και τόσο. Στο κομμάτι του Deviance βλέπουμε μία παρόμοια συμπεριφορά όπως του AIC. Αυτό σημαίνει ότι η μεταβλητή που αφαιρέθηκε στο βήμα αυτό μας έδινε μία σημαντική πληροφορία που δεν μπορέσαμε να την καλύψουμε. Η μεταβλητή αυτή είναι η prolongued decelerations. Καθώς όμως η μεταβλητή αυτή όταν την αφαιρέσαμε λύθηκε το πρόβλημα της σύγκλισης δεν μπορούμε απλά να την προσθέσουμε ξανά, αλλά πρέπει να την μελετήσουμε λίγο παραπάνω. Όπως είδαμε και πριν στο κομμάτι της ανάλυσης του συνόλου δεδομένων είδαμε ότι η συγκεκριμένη μεταβλητή ίσως μας δημιουργήσει πρόβλημα πράγμα που όντως έγινε καθώς μερικά από τα επίπεδα της υπό εκπροσωπούνταν. Για τον λόγο αυτό θα δούμε αν κάποια από τα επίπεδα έχουν μία κοινή συμπεριφορά ώστε να μπορέσουμε να προβούμε στην ένωσή τους. Η συμπεριφορά που θα ελέγξουμε θα είναι ως προς την απόκριση μας. Παρακάτω φαίνεται ένας πίνακας που μας δείχνει το πλήθος τιμών του κάθε επιπέδου της απόκρισης ανάμεσα στα επίπεδα του παράγοντα prolonogued decelerations.

prolonogued decelerations (DP)	**Fetal health – y**		
	1	**2**	**3**
DP 0.000	452	280	82
DP 0.001	14	5	11
DP 0.002	4	7	51
DP 0.003	1	3	20
DP 0.004	0	0	9
DP 0.005	0	0	3

Πίνακας 6.5 Πλήθος τιμών κάθε επιπέδου απόκρισης ανάμεσα στα επίπεδα της DP

Αυτό που παρατηρούμε είναι ότι τα επίπεδα 0.002,0.003,0.004,0.005 της DP έχουν μία παρόμοια συμπεριφορά καθώς πλειονότητα των περιπτώσεων των συγκεκριμένων επιπέδων αποτελεί μία παθολογική κατάσταση. Για τον λόγο αυτό καθώς αυτά είναι και τα επίπεδα που δεν εκπροσωπούνται επαρκώς αυτό που θα κάνουμε είναι να προβούμε στην δημιουργία ενός καινούργιου στρώματος το οποίο θα λέγεται prolonogued decelerations 0.002-0.005.Η αντίστοιχη δείκτρια του επιπέδου θα παίρνει την τιμή 1 όταν η αντίστοιχη μεταβλητή prolonogued decelerations λαμβάνει την τιμή ενός από τα 0.002,0.003,0.004,0.005 και κρατάμε πάλι το 0.000 ως επίπεδο αναφοράς.

Δημιουργώντας πλέον αυτό το επίπεδο και ξανά εντάσσοντας την μεταβλητή prolongued στο μοντέλο προβαίνουμε στην προσαρμογή αυτού.

```
Optimization terminated successfully.
         Current function value: 0.341600
         Iterations 12
                          MNLogit Regression Results
==============================================================================
Dep. Variable:           fetal_health   No. Observations:                  659
Model:                        MNLogit   Df Residuals:                      633
Method:                          MLE   Df Model:                           24
Date:                Mon, 17 Apr 2023   Pseudo R-squ.:                 0.6675
Time:                        17:39:35   Log-Likelihood:                -225.11
converged:                       True   LL-Null:                       -677.11
Covariance Type:            nonrobust   LLR p-value:                2.071e-175
=========================================================================================================================
                                     fetal_health=2       coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------------------
const                                                -15.2604      2.788     -5.474      0.000     -20.724      -9.797
baseline.value                                        -0.1441      0.048     -3.029      0.002      -0.237      -0.051
accelerations                                       -953.3964    142.246     -6.702      0.000   -1232.194    -674.599
fetal_movement                                         7.2949     11.411      0.639      0.523     -15.071      29.660
uterine_contractions                                -233.6255     63.063     -3.705      0.000   -357.226    -110.025
abnormal_short_term_variability                        0.0911      0.016      5.599      0.000      0.059       0.123
percentage_of_time_with_abnormal_long_term_variability     0.0121      0.010      1.249      0.212     -0.007       0.031
mean_value_of_long_term_variability                   -0.0750      0.045     -1.663      0.096     -0.163       0.013
histogram_mode                                        -0.0845      0.036     -2.344      0.019     -0.155      -0.014
histogram_mean                                         0.3103      0.054      5.723      0.000      0.204       0.417
histogram_variance                                     0.0482      0.011      4.258      0.000      0.026       0.070
prolongued_decelerations_0.001                         2.9010      1.150      2.524      0.012      0.648       5.154
prolongued_decelerations_0.002-0.005                   6.3963      1.173      5.452      0.000      4.097       8.696
-------------------------------------------------------------------------------------------------------------------------
                                     fetal_health=3       coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------------------
const                                                -17.7027      4.085     -4.333      0.000     -25.710      -9.695
baseline.value                                         0.1331      0.064      2.086      0.037       0.008       0.258
accelerations                                       -973.0454    354.485     -2.745      0.006   -1667.822    -278.269
fetal_movement                                        15.8907     12.376      1.284      0.199      -8.366      40.148
uterine_contractions                                -171.6868     98.072     -1.751      0.080   -363.904      20.530
abnormal_short_term_variability                        0.1955      0.028      6.889      0.000      0.140       0.251
percentage_of_time_with_abnormal_long_term_variability     0.0567      0.013      4.386      0.000      0.031       0.082
mean_value_of_long_term_variability                    0.0331      0.072      0.462      0.644      -0.107      0.173
histogram_mode                                        -0.1435      0.039     -3.640      0.000      -0.221      -0.066
histogram_mean                                         0.0304      0.053      0.574      0.566      -0.073      0.134
histogram_variance                                     0.0510      0.015      3.460      0.001      0.022       0.080
prolongued_decelerations_0.001                         1.4937      1.309      1.141      0.254      -1.072      4.059
prolongued_decelerations_0.002-0.005                   6.4245      1.400      4.589      0.000      3.681       9.168
=========================================================================================================================
```

Σχήμα 6.5.9 Μοντέλο μετά την Backward elimination

Το πρώτο πράγμα που βλέπουμε είναι η έχουμε επιτυχημένη σύγκλιση του αλγορίθμου ακόμα όλα τα επίπεδα του παράγοντα κρίνονται στατιστικά σημαντικά για τα μοντέλα μας πράγμα που δεν ήταν αναμενόμενο αν σκεφτούμε ότι ξεκινήσαμε από ορισμένα p-values που κινούντουσαν κοντά στο 1. Ακόμα βλέπουμε ότι πλέον μετά την είσοδο αυτής της μεταβλητής ορισμένες από τις μεταβλητές που θωρούντουσαν στατιστικά σημαντικές πλέον κρίνονται ως μη στατιστικά σημαντικές, καθώς πολύ πιθανόν η επεξηγηματικότητα που προσφέρεται από αυτές να καλύπτεται από την DP. Για τον λόγο αυτό θα αφού έχουμε συμπεριλάβει πλέον και αυτήν την καινούργια μεταβλητή θα συνεχίσουμε την διαδικασία του Back ward Elimination.

Μεταβλητή η οποία εξαιρέθηκε από το μοντέλο του προηγούμενου βήματος	Converged
Fetal Movement	TRUE
mean value of long term variability	TRUE

Πίνακας 6.6 Συνέχιση του feature selection μετά την επανένταξη της μετασχηματισμένης DP

Αφού τώρα τελειώσαμε με την διαδικασία του feature selection και συμπεριλαμβανομένης της DP πάμε να δούμε τα χαρακτηριστικά του μοντέλου αυτού.

```
Optimization terminated successfully.
         Current function value: 0.349766
         Iterations 12
                          MNLogit Regression Results
==============================================================================
Dep. Variable:           fetal_health   No. Observations:                  659
Model:                        MNLogit   Df Residuals:                      637
Method:                           MLE   Df Model:                           20
Date:                Mon, 17 Apr 2023   Pseudo R-squ.:                 0.6596
Time:                        17:53:13   Log-Likelihood:                -230.50
converged:                       True   LL-Null:                       -677.11
Covariance Type:            nonrobust   LLR p-value:                2.167e-176
=================================================================================================================================
                                    fetal_health=2       coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------------------------------
const                                                -16.0734      2.748     -5.850      0.000     -21.459     -10.688
baseline.value                                        -0.1372      0.047     -2.899      0.004      -0.230      -0.044
accelerations                                       -893.1440    136.705     -6.533      0.000   -1161.080    -625.208
uterine_contractions                                -218.4945     58.455     -3.738      0.000    -333.065    -103.924
abnormal_short_term_variability                       0.0954      0.016      5.989      0.000       0.064       0.127
percentage_of_time_with_abnormal_long_term_variability    0.0186      0.009      2.084      0.037       0.001       0.036
histogram_mode                                       -0.0680      0.036     -1.913      0.056      -0.138       0.002
histogram_mean                                        0.2850      0.053      5.426      0.000       0.182       0.388
histogram_variance                                    0.0440      0.011      3.898      0.000       0.022       0.066
prolongued_decelerations_0.001                        3.4084      1.089      3.129      0.002       1.274       5.543
prolongued_decelerations_0.002-0.005                  6.7224      1.151      5.842      0.000       4.467       8.978
---------------------------------------------------------------------------------------------------------------------------------
                                    fetal_health=3       coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------------------------------
const                                                -15.9984      3.827     -4.181      0.000     -23.499      -8.498
baseline.value                                        0.1164      0.060      1.940      0.052      -0.001       0.234
accelerations                                      -1007.2283    337.838     -2.981      0.003   -1669.378    -345.078
uterine_contractions                                -221.3461     94.178     -2.350      0.019    -405.932     -36.760
abnormal_short_term_variability                       0.1803      0.026      6.995      0.000       0.130       0.231
percentage_of_time_with_abnormal_long_term_variability    0.0536      0.011      4.713      0.000       0.031       0.076
histogram_mode                                       -0.1350      0.040     -3.380      0.001      -0.213      -0.057
histogram_mean                                        0.0375      0.047      0.794      0.427      -0.055       0.130
histogram_variance                                    0.0520      0.014      3.739      0.000       0.025       0.079
prolongued_decelerations_0.001                        1.7010      1.330      1.279      0.201      -0.905       4.307
prolongued_decelerations_0.002-0.005                  6.9857      1.341      5.209      0.000       4.357       9.614
=================================================================================================================================
```

Σχήμα 6.5.10 Τελικό μοντέλο πολυωνυμικής παλινδρόμησης

Στην εικόνα αυτή παρουσιάζεται το τελικό μοντέλο της πολυωνυμικής λογιστικής παλινδρόμησης. Βλέπουμε ότι όλες οι μεταβλητές κρίνονται στατιστικά σημαντικές βάση του ελέγχου το Wald για τουλάχιστον ένα από τα δύο υπομοντέλα.

Το τελικό μοντέλο της πολυωνυμικής λογιστικής παλινδρόμησης είδαμε στο κεφάλαιο 4 ότι είναι της μορφής

$$log\left(\frac{p_i}{p_1}\right) = \boldsymbol{x'\beta_i} = \beta_{i0} + \beta_{i1}x_1 + \ \ \ldots\ldots\ \beta_{ip}x_p\ , \mu\varepsilon\ i = 2{,}3\ \ldots . k$$

Στην εφαρμογή μας έχουμε 3 κλάσεις στην απόκριση άρα θα πάρουμε *k=3-1=2* μοντέλα που έχουν ως baseline την φυσιολογική κατάσταση του εμβρύου.

Άρα έχουμε

$$log\left(\frac{p_{Suspect}}{p_{Normal}}\right) = -16.0734 - 0.1372LB - 893.1440AC - 218.4945UC + 0.0954ASTV + 0.0186ALTV - 0.0680Mode + 0.285Mean + 0.0440Variance + 3.4084DP_{0.001} + 6.7224DP_{0.002-0.005} \quad (6.5.1.1.)$$

$$log\left(\frac{p_{pathological}}{p_{Normal}}\right) = -15.9984 + 0.1164LB - 1007.22830AC - 221.3461UC + 0.18034ASTV + 0.0536ALTV - 0.135Mode + 0.0375Mean + 0.052Variance + 1.701DP_{0.001} + 6.9657DP_{0.002-0.005} \quad (6.5.1.2.)$$

Από το πρόσημο της εκτιμηθείσας τιμής της κάθε παραμέτρου μπορούμε να διακρίνουμε για τις ποσοτικές μεταβλητές αν επρόκειτο για αύξηση της σχετικής πιθανότητας ή μείωση αυτής το να είναι έμβρυο σε ύποπτη κατάσταση ή παθολογική σε σχέση με φυσιολογική. Θα ερμηνεύσουμε την ποσοτική μεταβλητή LB που αντιστοιχεί στην baseline.value που είναι οι σφυγμοί της καρδιάς του εμβρύου όπου υπάρχει εναλλαγή στα πρόσημα έτσι ώστε να έχουμε μία εικόνα για την ερμηνεία της κάθε μεταβλητής.

Αν λοιπόν εμείς παρατηρήσουμε μία αύξηση στους σφυγμούς της καρδιάς κατά μία μονάδα ενώ όλες οι άλλες μεταβλητές του τελικού μοντέλου παραμένουν σταθερές θα έχουμε ότι η λογαριθμημένη σχετική πιθανότητα του να είναι αυτή η παρατήρηση ύποπτη έναντι του να είναι φυσιολογική θα μειωθεί κατά 0.1372. Αυτό ισοδύναμα σημαίνει ότι η σχετική πιθανότητα πολλαπλασιάζεται επί $e^{-0.1372} = 0.8717958$ άρα έχουμε μείωση κατά 12.8%. Ενώ όσον αφορά το δεύτερο μοντέλο αναμένουμε η αντίστοιχη λογαριθμημένη σχετική πιθανότητα του να είναι η αντίστοιχη παρατήρηση μία παθολογική περίπτωση να αυξηθεί κατά 0.1164. Αυτό ισοδύναμα σημαίνει ότι η σχετική πιθανότητα πολλαπλασιάζεται επί $e^{0.1164} = 1.123445$ άρα έχουμε αύξηση κατά 12.3%.

Αυτά αναφορικά με τις ποσοτικές μεταβλητές. Για την ποιοτική μας μεταβλητή η οποία είναι η DP (prolongued decelerations), έστω ότι όλες οι ποσοτικές μεταβλητές παραμένουν σταθερές. Αν δούμε μία παρατεταμένη επιβράδυνση της τάξεως του 0.001 τότε αναμένουμε να έχουμε μία αύξηση της λογαριθμημένης σχετικής πιθανότητας στο να θεωρηθεί αυτή η περίπτωση ως ύποπτη σε σχέση με φυσιολογική κατά 3.4084 σε σχέση με μία παρατήρηση που δεν έχει παρατεταμένη επιβράδυνση η οποία αντιστοιχεί σε 2921.686% αύξηση. Σε περίπτωση που έχουμε ένδειξη παρατεταμένης επιβράδυνσης παραπάνω από 0.001 τότε η αντίστοιχη λογαριθμημένη σχετική πιθανότητα αυξάνεται σημαντικά κατά 6.7224 σε σύγκριση με κάποιον που δεν έχει παρατεταμένη επιβράδυνση άρα σε 82980.91% αύξηση. Όσον αφορά την παθολογική κατάσταση αναμένουμε ένα έμβρυο με παρατεταμένη επιβράδυνση να έχει αυξημένη λογαριθμημένη σχετική πιθανότητα κατά 1.701 στο να είναι όντως παθολογική η κατάσταση έναντι του να είναι φυσιολογική συγκριτικά με περίπτωση μηδενικής παρατεταμένης επιβράδυνσης το οποίο αντιστοιχεί σε 447.9424% αύξηση. Σε περίπτωση που είναι ακόμα υψηλότερη, η λογαριθμημένη σχετική πιθανότητα αυξάνεται παραπάνω στο 6.9657 που αντιστοιχεί σε 105865.6% αύξηση.

Αφού τελειώσαμε με τις ερμηνείες μπορούμε να προβούμε στην γενική αξιολόγηση του μοντέλου μας. Το μοντέλο αυτό έχει 9 χαρακτηριστικά εισόδου. Οι βαθμοί ελευθερίας των καταλοίπων είναι 637 από τις 659 παρατηρήσεις που έχουμε, ενώ οι βαθμοί ελευθερίας του μοντέλου είναι 20.

Σχήμα 6.5.11 Τελικό μοντέλο πολυωνυμικής παλινδρόμησης

Από το Σχήμα 6.5.11 παρατηρούμε ότι από τις 145 παρατηρήσεις του συνόλου ελέγχου που ανήκουν στην ομάδα 1, οι 134 έχουν ταξινομηθεί σωστά ενώ 10 έχουν ταξινομηθεί ως ύποπτες και μία ως παθολογική. Ακόμα 77 από τις 87 ύποπτες παρατηρήσεις έχουν προβλεφθεί σωστά, έχοντας κάνει λάθος σε 6 προβλέψεις ως φυσιολογικές και σε 4 ως παθολογικές. Τέλος από τις 51 παθολογικές καταστάσεις έχουμε ταξινομήσει 2 λανθασμένα ως φυσιολογικές ενώ 9 ως ύποπτες.

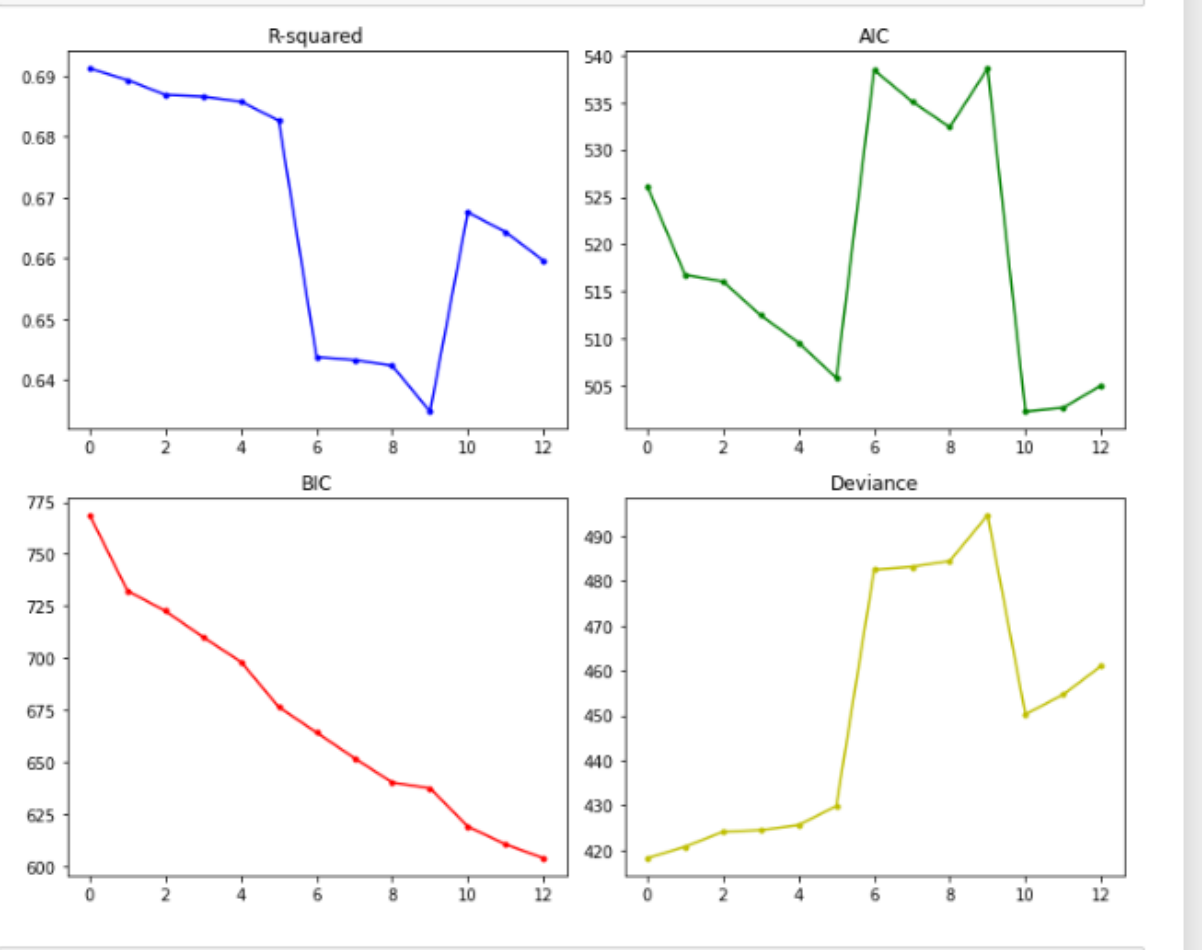
Στην συνέχεια θα δημιουργήσουμε το αντίστοιχο classification report που παράγεται από τον παραπάνω πίνακα καθώς και τα ειδικά μέτρα αξιολόγησης της πολυωνυμικής παλινδρόμησης.

	precision	recall	f1-score	support
1	0.94	0.92	0.93	145
2	0.80	0.89	0.84	87
3	0.89	0.78	0.83	51
accuracy			0.89	283
macro avg	0.88	0.86	0.87	283
weighted avg	0.89	0.89	0.89	283

AIC 504.9916525759451
BIC 603.7875703350003
pseudo R-squared 0.6595893676845824
Deviance 460.9916525759451

Σχήμα 6.5.12 Classification Report τελικού μοντέλου πολυωνυμικής παλινδρόμησης

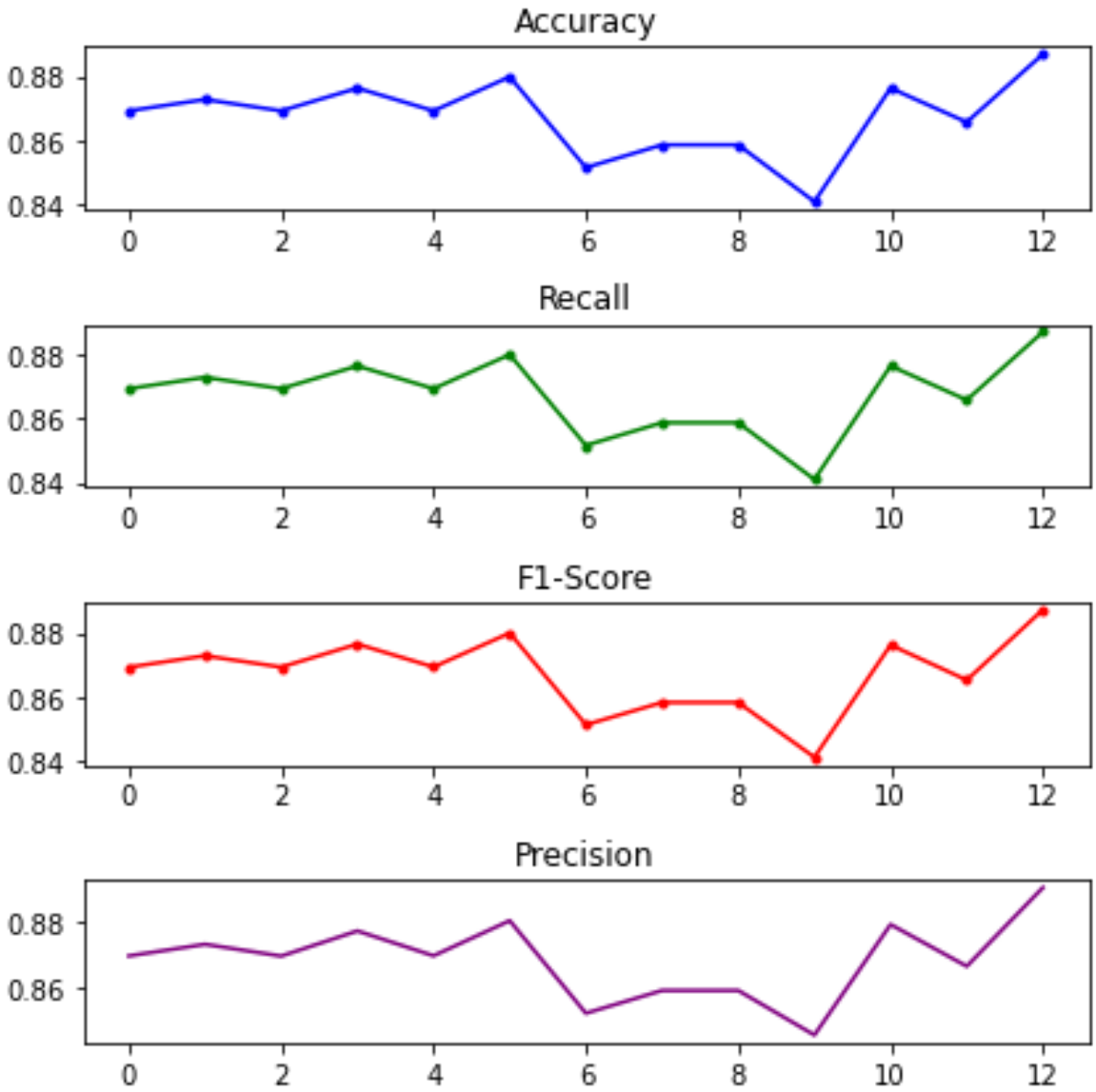
Στο κομμάτι των εξωτερικών μέτρων βλέπουμε υψηλά ποσοστά επιτυγχάνοντας το με την αφαίρεση πάνω από των μισών χαρακτηριστικών μας, το οποίο μας ωφελεί σημαντικά στο θέμα του κόστους. Γενικότερα βλέπουμε μία εικόνα για την πρώτη ομάδα απόδοσης κοντά στο 93%, στην δεύτερη ομάδα κοντά στο 86% ενώ στην τρίτη κοντά στο 85%. Το ποσοστό των συνολικά σωστών ταξινομήσεων που προέβλεψε το μοντέλο είναι 89%. Μέσω του ψευδό- R^2 μπορούμε να πούμε ότι περίπου το 65% της μεταβλητότητας της λογαριθμημένης σχετικής πιθανότητας για την κατηγορία που ανήκει μία παρατήρηση ερμηνεύεται από το μοντέλο κάτι το οποίο είναι παραπάνω από ικανοποιητικό καθώς όπως είπαμε και στο κεφάλαιο 4 η καλή προσαρμογή του μοντέλου κυμαίνεται στις τιμές 20%-40%. Τέλος για τα πληροφοριακά κριτήρια και το Deviance είναι καλύτερα να προβούμε σε σύγκριση με άλλα μοντέλα ώστε να μπορέσουμε να διακρίνουμε την βελτίωση τους η μη κατά την πορεία του feature selection.



Σχήμα 6.5.13 Πορεία των ειδικών μέτρων αξιολόγησης μέχρι το τελικό μοντέλο παλινδρόμησης

Μέσω του παραπάνω σχήματος βλέπουμε την πορεία των μέτρων κατά την ολική διάρκεια της ανάλυσης. Ως προς το ψευδό- R^2 βλέπουμε ότι είχαμε μία πτώση στο πέμπτο βήμα αλλά με την σωστή εισαγωγή της DP (παρατεταμένης επιβράδυνσης) ως μεταβλητή στο μοντέλο, διορθώνοντας τα προβλήματα που δημιουργούσε, βλέπουμε να έχουμε με πολύ λίγες μεταβλητές ένα μεγάλο ποσοστό κοντά στο αρχικό με 12 παραπάνω μεταβλητές. Το πληροφοριακό κριτήριο

του Bayes έχει μία γενική σταθερή φθίνουσα πορεία καθώς φεύγαν οι μεταβλητές από το μοντέλο πράγμα το οποίο είναι καλή ένδειξη. Για το πληροφοριακό κριτήριο του Akaike είδαμε και πριν ότι μόλις αφαιρέθηκε η DP είχαμε μεγάλη αύξηση που μας προβλημάτισε. Έτσι καθώς μετά προβήκαμε στην επανένταξή της με τον σωστό τρόπο αυτήν την φορά καταφέραμε στο AIC το καλύτερο αποτέλεσμα. Τέλος αξίζει να δούμε και την πορεία των εξωτερικών μέτρων κατά την ολική ανάλυση του μοντέλου.



Σχήμα 6.5.14 Πορεία των εξωτερικών μέτρων μέχρι το τελικό μοντέλο παλινδρόμησης

Βλέπουμε την βελτίωση των εξωτερικών μέτρων κατά την διάρκεια της διαδικασίας και την διόρθωση του απώλειας πληροφορίας στο βήμα 10 μέχρι που όλα τα μέτρα φτάνουν την μέγιστη τους τιμή στο τελικό μας μοντέλο.

6.6 Χρήση των Δένδρων απόφασης για εκτίμηση του δείκτη υγείας

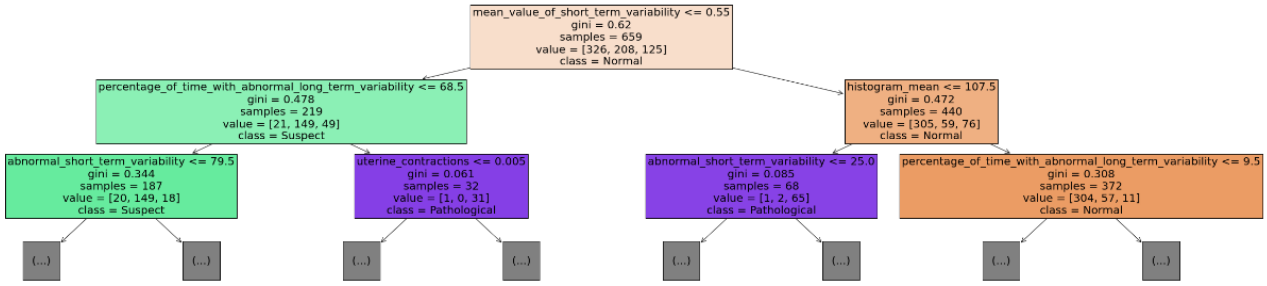
Στην παράγραφο αυτή θα δούμε την εφαρμογή του μοντέλου των Decision Trees για την αντιμετώπιση του προβλήματος. Σκοπός μας είναι η δημιουργία nodes που εμπεριέχουν συνθήκες ως προς της μεταβλητές μας και η ιεράρχησή τους βασισμένοι πάνω στο αντίστοιχο δείκτη *Gini*.

Αρχικά προσαρμόζουμε το μοντέλο μας για το σύνολο το μεταβλητών μας μέσω της Python.

```
#Δημιουργία μοντέλου δέντρου απόφασης
dt = DecisionTreeClassifier(random_state=42)
#Προσαρμογή μοντέλου
dt.fit(X_train, y_train)


#Πρόβλεψη μοντέλου στο σύνολο ελέγχου
y_pred = dt.predict(X_test)



#Visualization του δένδρου ταξινόμησης σε βάθος τριών επιπέδων
fig, ax = plt.subplots(figsize=(42, 10))
plot_tree(dt, ax=ax, feature_names=X.columns, class_names=["Normal", "Suspect", "Pathological"], filled=True,
          fontsize=20,max_depth=2)
plt.show()
```

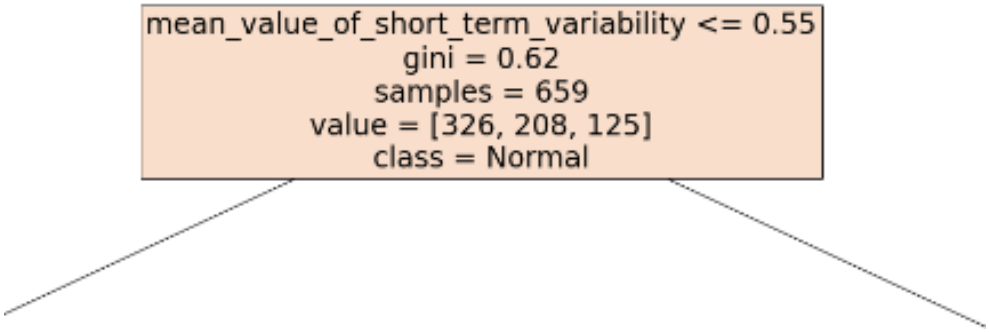
Σχήμα 6.6.1 Κώδικας δημιουργίας δέντρου ταξινόμησης

Στην συνέχεια εμφανίζουμε τα πρώτα επίπεδα του δέντρου αυτού για να δούμε την δομή του και πως παρουσιάζεται το δέντρο μας στην Python.



Σχήμα 6.6.2 Οπτικοποίηση πρώτων επιπέδων του δέντρου

Στο Σχήμα 6.6.2 βλέπουμε την ρίζα του δέντρου καθώς και τα 2 πρώτα επίπεδα των internal nodes.Σε κάθε node μέσα εμφανίζονται ορισμένες πληροφορίες.Ας δούμε λίγο ένα από αυτά λίγο πιο αναλυτικά για να εξηγήσουμε την πληροφορία που υπάρχει μέσα σε αυτό.



Σχήμα 6.6.3 Rood node του πλήρες μοντέλου

Στην πρώτη γραμμή του node βλέπουμε την συνθήκη που έχει δημιουργηθεί από την μεταβλητή με τον μικρότερο αντίστοιχο δείκτη *Gini*. Από κάτω βλέπουμε την τιμή του δείκτη *Gini* για το συγκεκριμένο δείγμα του οποίου το πλήθος φαίνεται από κάτω. Στην τέταρτη γραμμή βλέπουμε 3 αριθμούς. Αυτοί οι αριθμοί αντιπροσωπεύουν την εκπροσώπηση της κάθε ομάδας στο δείγμα που έχουμε. Συγκεκριμένα για αυτό το node του Σχήματος 6.6.3 που αποτελεί και την ρίζα του δέντρου μας, φαίνεται ότι έχουμε στις 659 παρατηρήσεις, από αυτές 326 κανονικές περιπτώσεις, 208 ύποπτες και 125 παθολογικές. Τέλος βλέπουμε στο κάθε node από κάτω να αποδίδεται μία ετικέτα από αυτές που έχουμε ως επιλογή και ακόμα διαφορετικά χρώματα για κάθε node. Ο λόγος για τον οποίο γίνετε αυτό είναι ότι όπως είπαμε και στο κεφάλαιο 5 πολλές φορές προβαίνουμε σε κλάδεμα των κλαδιών του δέντρου και μετατρέποντας κάποια από τα ενδιάμεσα nodes σε φύλλα. Σε περίπτωση που μετατρέψουμε ένα node ως φύλλο πρέπει να μας ορίζει μία πρόβλεψη. Η αντίστοιχη πρόβλεψη για την παρατήρηση είναι το επίπεδο της ετικέτας το οποίο κατέχει την πλειοψηφία στο αντίστοιχο node.Έτσι στο κάθε node μπορούμε να δούμε την πλειοψηφία αυτή στην τελευταία γραμμή που μας ορίζει και την πρόβλεψη. Για το συγκεκριμένο node αν επιθυμούσαμε να το ορίσουμε ως φύλλο θα προέβλεπε ότι η παρατήρηση είναι φυσιολογική. Ακόμα για την κάθε κλάση αντιστοιχεί ένα χρώμα.

- Φυσιολογική 1 (N) – πορτοκαλί
- Ύποπτη 2 (S) – πράσινο
- Παθολογική (P) – μωβ

Όσο πιο καθαρό είναι ένα node δηλαδή όσο μεγαλύτερη διαφορά υπάρχει της πλειοψηφίας από της άλλες κλάσεις τόσο πιο έντονη απόχρωση παίρνει το αντίστοιχο node, ενώ σε μικρές διαφορές το αντίστοιχο node παίρνει ανοιχτές αποχρώσεις.

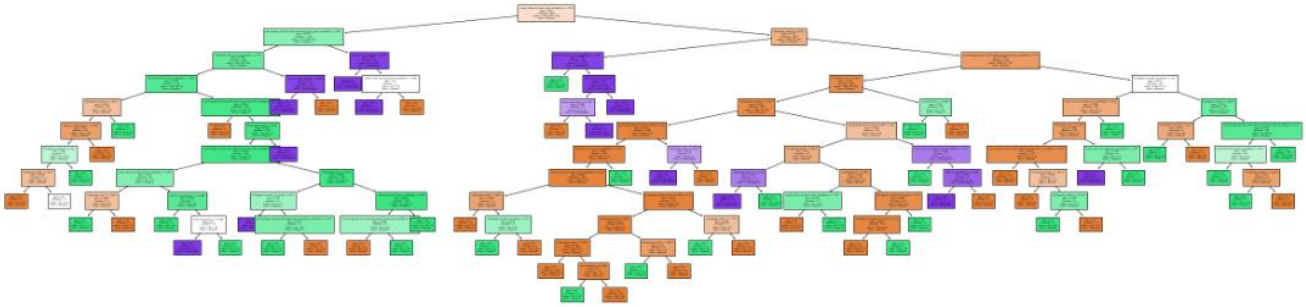
Ακόμα στο σχήμα αυτό βλέπουμε την αντίστοιχη συνθήκη που δημιουργείται από την μεταβλητή μας και πιο συγκεκριμένα την mean_value_of_short_term_variability. Η μεταβλητή αυτή ήρθε στην κορυφή καθώς θεωρείται από το μοντέλο ως η πιο σημαντική με βάση τον αντίστοιχο δείκτη *Gini* για την διαδικασία της πρόβλεψης. Για το συγκεκριμένο μοντέλο το οποίο περιέχει όλες τις μεταβλητές μπορούμε να πάρουμε την σημαντικότητα που κατέχει η κάθε μία μεταβλητή μέσω της Python.

```
u=pd.DataFrame([dt.feature_importances_,X_train.columns])
u.T
```

0	0.001562	baseline.value
1	0.022316	accelerations
2	0.010349	fetal_movement
3	0.014549	uterine_contractions
4	0.004091	light_decelerations
5	0.110836	abnormal_short_term_variability
6	0.249086	mean_value_of_short_term_variability
7	0.175747	percentage_of_time_with_abnormal_long_term_var...
8	0.018355	mean_value_of_long_term_variability
9	0.005182	histogram_width
10	0.031915	histogram_min
11	0.009281	histogram_max
12	0.038094	histogram_number_of_peaks
13	0.004686	histogram_number_of_zeroes
14	0.021278	histogram_mode
15	0.2211	histogram_mean
16	0.021602	histogram_median
17	0.022315	histogram_variance
18	0.0	severe_decelerations_0.001
19	0.0	prolongued_decelerations_0.001
20	0.011418	prolongued_decelerations_0.002
21	0.006238	prolongued_decelerations_0.003
22	0.0	prolongued_decelerations_0.004
23	0.0	prolongued_decelerations_0.005
24	0.0	histogram_tendency_0
25	0.0	histogram_tendency_1

Σχήμα 6.6.4 Σημαντικότητα μεταβλητών στο πλήρες δέντρο απόφασης

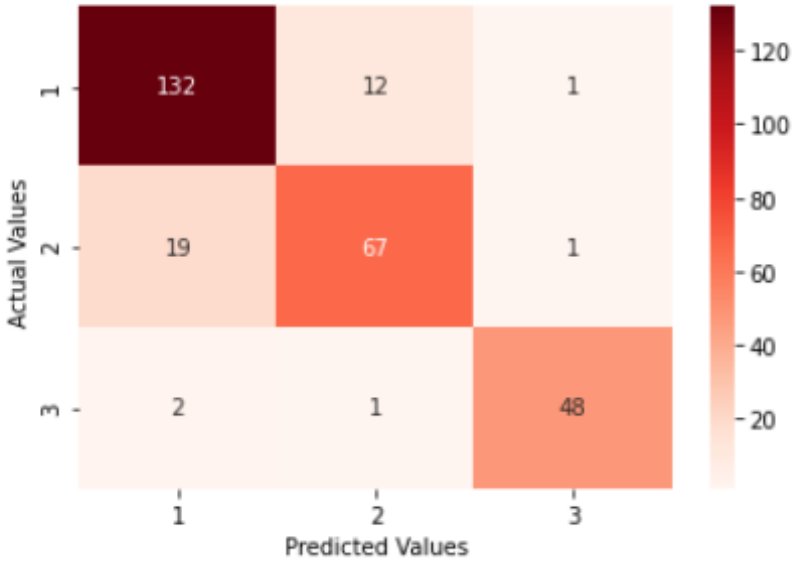
Στο Σχήμα 6.6.4 μπορούμε να διακρίνουμε την σημαντικότητα των μεταβλητών μας για το μοντέλο μας. Παρατηρούμε ότι η μέγιστη τιμή αντιστοιχεί στην μεταβλητή που ορίστηκε ως η ρίζα την mean_value_of_short_term_variability. Στο Σχήμα 6.6.2 βλέπουμε ότι και τα επόμενα nodes έχουν δημιουργηθεί με βάση τις percentage_of time_with_abnormal_long_term_variability και histogram mean που καταλαμβάνουν υψηλές θέσεις επίσης. Στο Σχήμα 6.6.5 φαίνεται η ολική δομή του δέντρου.



Σχήμα 6.6.5 Ολική δομή πλήρους δέντρου απόφασης

Όπως καταλαβαίνουμε σε μία τέτοια εικόνα η οποία είναι πολύπλοκη, η ύπαρξη των χρωμάτων στα διάφορα nodes είναι επιβεβλημένη ώστε ο αναλυτής να μπορέσει να πάρει μία πρόχειρη εικόνα για την αρχιτεκτονική του δέντρου.

Πάμε να δούμε όμως πως αυτό το δέντρο αποδίδει στα δεδομένα ελέγχου.



Σχήμα 6.6.6 Πίνακας συγχύσεως πλήρους μοντέλου δέντρου απόφασης

Από τον πίνακα συγχύσεως του Σχήματος 6.6.6 παρατηρούμε ότι από τις 145 παρατηρήσεις που είναι φυσιολογικές περιπτώσεις έχουν διαγνωστεί σωστά ως φυσιολογικές οι 132, 12 έχουν διαγνωστεί λανθασμένα ως ύποπτες και 1 λανθασμένα ως παθολογική. Στην συνέχεια από τις 87 ύποπτες περιπτώσεις οι 67 έχουν όντως ταξινομηθεί σωστά οι 19 λανθασμένα ως φυσιολογικές και 1 περίπτωση ως παθολογική. Τέλος από τις 51 παθολογικές περιπτώσεις έχουμε 48 σωστές διαγνώσεις 2 λανθασμένα τοποθετημένες ως φυσιολογικές και 1 ως ύποπτη. Βλέπουμε γενικότερα ότι η μεγαλύτερη σύγχυση υπάρχει ανάμεσα στις ομάδες 1 και 2.

Στην συνέχεια κατασκευάζουμε το classification report για να δούμε την απόδοση του.

	precision	recall	f1-score	support
1	0.86	0.91	0.89	145
2	0.84	0.77	0.80	87
3	0.96	0.94	0.95	51
accuracy			0.87	283
macro avg	0.89	0.87	0.88	283
weighted avg	0.87	0.87	0.87	283

Σχήμα 6.6.7 Classification Report πλήρους μοντέλου δέντρου απόφασης

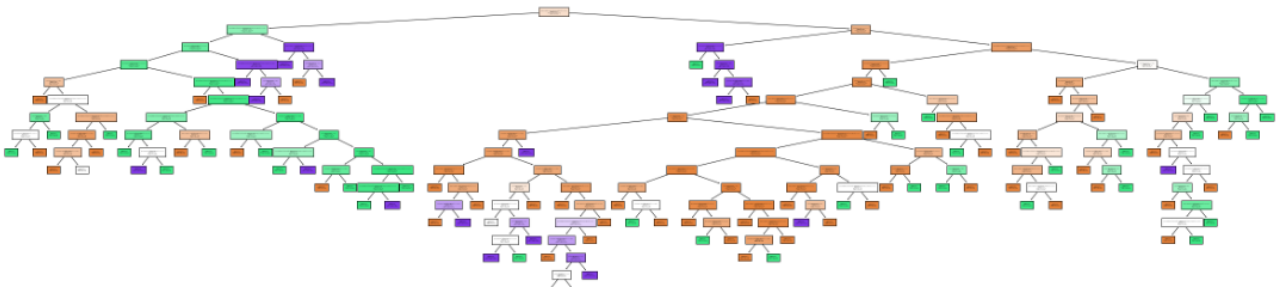
Από το Σχήμα 6.6.7 βλέπουμε μια πάρα πολύ καλή απόδοση του μοντέλου αν σκεφτούμε κιόλας ότι δεν έχουμε παρέμβει ακόμα αρκετά επάνω του. Βλέπουμε μία αδυναμία στην δεύτερη ομάδα και σε αυτό το report συγκριτικά με τις άλλες ομάδες.

Όπως είδαμε και πριν κάποιες από τις μεταβλητές μας είναι οι πιο σημαντικές από άλλες. Οπότε ίσως καταφέρουμε να πετύχουμε με λιγότερες μεταβλητές καλύτερη απόδοση ή ίδια απόδοση στο οποίο θα ωφεληθούμε και στο θέμα κόστους. Ας προσαρμόσουμε για παράδειγμα ένα δέντρο απόφασης μόνο για τις 4 πιο σημαντικές μεταβλητές του Σχήματος 6.6.4 που είδαμε πριν. Αυτές είναι οι

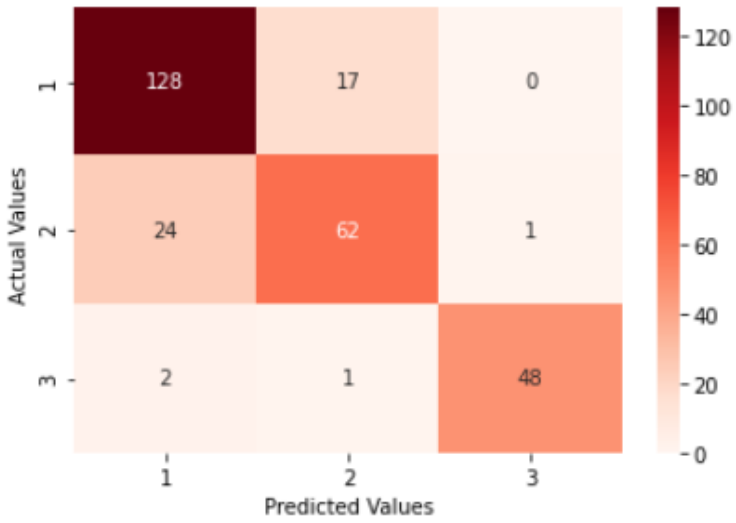
mean_value_of_short_term_variability
histogram_mean
percentage_of_time_with_abnormal_long_term_variability
abnormal_short_term_variability

Πίνακας 6.7 Τα τέσσερα πιο σημαντικά χαρακτηριστικά του πλήρες μοντέλου δέντρου απόφασης

Προσαρμόζοντας λοιπόν το αντίστοιχο δέντρο απόφασης έχουμε ότι η αρχιτεκτονική του και η απόδοση του είναι:



Σχήμα 6.6.8 Ολική δομή μοντέλου δέντρου απόφασης με 4 χαρακτηριστικά



Σχήμα 6.6.9 Πίνακας συγχύσεως μοντέλου δέντρου απόφασης με 4 χαρακτηριστικά

	precision	recall	f1-score	support
1	0.83	0.88	0.86	145
2	0.78	0.71	0.74	87
3	0.98	0.94	0.96	51
accuracy			0.84	283
macro avg	0.86	0.85	0.85	283
weighted avg	0.84	0.84	0.84	283

Σχήμα 6.6.10 Classification Report μοντέλου δέντρου απόφασης με 4 χαρακτηριστικά

Βλέπουμε ότι μόλις με 4 μόνο μεταβλητές καταφέραμε να πετύχουμε μία αρκετά καλή απόδοση. Συγκεκριμένα βλέπουμε ότι τα νούμερα στα πιο πολλά μέτρα έχουν πέσει ελαφρά αλλά δοθέντος ότι από τις 21 μεταβλητές πέσαμε στις 4 και διατηρούμε τόσο καλά ποσοστά και μάλιστα στις μετρικές της ομάδας 3 πετυχαίνουμε και ελαφρώς καλύτερα νούμερα είναι αξιόλογο.

Ως επόμενο βήμα θα προβούμε σε feature selection.Η τεχνική που θα χρησιμοποιήσουμε είναι η forward selection και ως μέτρο το οποίο μετρά την απόδοση του μοντέλου θα ορίσουμε το accuracy. Τέλος το forward selection θα γίνει με cross-validation με cv=3.

Έτσι έχουμε ότι

```
from mlxtend.feature_selection import SequentialFeatureSelector
sfs_selector = SequentialFeatureSelector(dt, k_features=len(X_train.columns), cv=3,scoring="accuracy")
sfs1=sfs_selector.fit(X_train,y_train)
sfs1pdf = pd.DataFrame(sfs1.subsets_)
sfs1pdf= sfs1pdf.T
sfs1pdf
```

	feature_idx	cv_scores	avg_score	feature_names
1	(7,)	[0.6681818181818182, 0.6818181818181818, 0.662...	0.6707	(percentage_of_time_with_abnormal_long_term_va...
2	(7, 15)	[0.7681818181818182, 0.759090909090909, 0.7488...	0.75871	(percentage_of_time_with_abnormal_long_term_va...
3	(5, 7, 15)	[0.8090909090909091, 0.8272727272727273, 0.835...	0.823993	(abnormal_short_term_variability, percentage_o...
4	(5, 7, 15, 17)	[0.8363636363636363, 0.8636363636363636, 0.826...	0.842161	(abnormal_short_term_variability, percentage_o...
5	(3, 5, 7, 15, 17)	[0.8454545454545455, 0.8636363636363636, 0.858...	0.855846	(uterine_contractions, abnormal_short_term_var...
6	(1, 3, 5, 7, 15, 17)	[0.8590909090909091, 0.8909090909090909, 0.835...	0.861872	(accelerations, uterine_contractions, abnormal...
7	(1, 3, 5, 7, 15, 17, 25)	[0.8590909090909091, 0.8909090909090909, 0.849...	0.866438	(accelerations, uterine_contractions, abnormal...
8	(1, 3, 5, 7, 15, 17, 20, 25)	[0.8727272727272727, 0.8954545454545455, 0.835...	0.867933	(accelerations, uterine_contractions, abnormal...
9	(1, 3, 5, 7, 15, 17, 20, 22, 25)	[0.8727272727272727, 0.8954545454545455, 0.844...	0.870977	(accelerations, uterine_contractions, abnormal...
10	(1, 3, 5, 7, 15, 17, 18, 20, 22, 25)	[0.8727272727272727, 0.8909090909090909, 0.844...	0.869462	(accelerations, uterine_contractions, abnormal...
11	(1, 3, 5, 7, 14, 15, 17, 18, 20, 22, 25)	[0.8590909090909091, 0.9, 0.867579908675799]	0.875557	(accelerations, uterine_contractions, abnormal...
12	(1, 3, 5, 7, 14, 15, 17, 18, 20, 22, 24, 25)	[0.8636363636363636, 0.8954545454545455, 0.863...	0.874035	(accelerations, uterine_contractions, abnormal...
13	(1, 3, 5, 7, 14, 15, 17, 18, 20, 21, 22, 24, 25)	[0.8727272727272727, 0.9045454545454545, 0.849...	0.875529	(accelerations, uterine_contractions, abnormal...
14	(1, 3, 5, 7, 14, 15, 17, 18, 20, 21, 22, 23, 2...	[0.8681818181818182, 0.9045454545454545, 0.863...	0.87858	(accelerations, uterine_contractions, abnormal...
15	(1, 3, 5, 7, 14, 15, 17, 18, 19, 20, 21, 22, 2...	[0.8636363636363636, 0.9045454545454545, 0.863...	0.877065	(accelerations, uterine_contractions, abnormal...
16	(1, 3, 5, 7, 13, 14, 15, 17, 18, 19, 20, 21, 2...	[0.8545454545454545, 0.9045454545454545, 0.853...	0.870991	(accelerations, uterine_contractions, abnormal...
17	(0, 1, 3, 5, 7, 13, 14, 15, 17, 18, 19, 20, 21...	[0.8318181818181818, 0.9045454545454545, 0.863...	0.866459	(baseline.value, accelerations, uterine_contra...
18	(0, 1, 3, 5, 7, 10, 13, 14, 15, 17, 18, 19, 20...	[0.8636363636363636, 0.9, 0.863013698630137]	0.866459	(baseline.value, accelerations, uterine_contra...
19	(0, 1, 2, 3, 5, 7, 10, 13, 14, 15, 17, 18, 19,...	[0.85, 0.8863636363636364, 0.8538812785388128]	0.863415	(baseline.value, accelerations, fetal_movement...
20	(0, 1, 2, 3, 4, 5, 7, 10, 13, 14, 15, 17, 18, ...	[0.8636363636363636, 0.8954545454545455, 0.858...	0.863422	(baseline.value, accelerations, fetal_movement...
21	(0, 1, 2, 3, 4, 5, 6, 7, 10, 13, 14, 15, 17, 1...	[0.8545454545454545, 0.8954545454545455, 0.849...	0.866438	(baseline.value, accelerations, fetal_movement...
22	(0, 1, 2, 3, 4, 5, 6, 7, 10, 11, 13, 14, 15, 1...	[0.8454545454545455, 0.8954545454545455, 0.844...	0.861886	(baseline.value, accelerations, fetal_movement...
23	(0, 1, 2, 3, 4, 5, 6, 7, 10, 11, 12, 13, 14, 1...	[0.8545454545454545, 0.8954545454545455, 0.853...	0.86796	(baseline.value, accelerations, fetal_movement...
24	(0, 1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14...	[0.85, 0.8818181818181818, 0.8538812785388128]	0.8619	(baseline.value, accelerations, fetal_movement...
25	(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,...	[0.85, 0.8863636363636364, 0.821917808219178]	0.85276	(baseline.value, accelerations, fetal_movement...
26	(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,...	[0.8590909090909091, 0.8818181818181818, 0.835...	0.858842	(baseline.value, accelerations, fetal_movement...

Σχήμα 6.6.11 Feature Selection για το μοντέλο των δέντρων απόφασης

Στην συνέχεια βρίσκουμε τα features τα οποία μας δίνουν την καλύτερη απόδοση.

```
sfs1pdf[sfs1pdf.iloc[:,2]==max(sfs1pdf.iloc[:,2])]
```

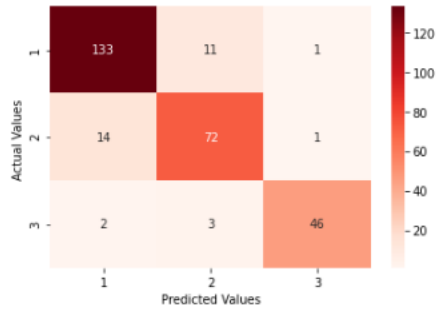
	feature_idx	cv_scores	avg_score	feature_names
14	(1, 3, 5, 7, 14, 15, 17, 18, 20, 21, 22, 23, 2...	[0.8681818181818182, 0.9045454545454545, 0.863...	0.87858	(accelerations, uterine_contractions, abnormal...

```
sfs1pdf.iloc[13,3]
```

```
('accelerations',
 'uterine_contractions',
 'abnormal_short_term_variability',
 'percentage_of_time_with_abnormal_long_term_variability',
 'histogram_mode',
 'histogram_mean',
 'histogram_variance',
 'severe_decelerations_0.001',
 'prolongued_decelerations_0.002',
 'prolongued_decelerations_0.003',
 'prolongued_decelerations_0.004',
 'prolongued_decelerations_0.005',
 'histogram_tendency_0',
 'histogram_tendency_1')
```

Σχήμα 6.6.12 Χαρακτηριστικά που επιλέχθηκαν από το Feature Selection

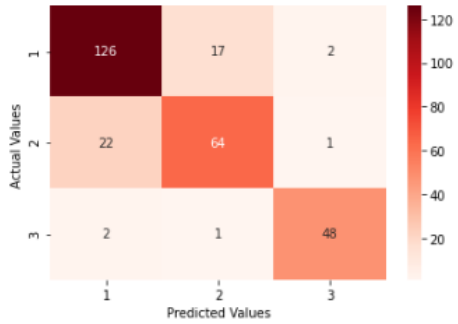
Παρατηρούμε ότι έχουν συμπεριληφθεί όλες οι ποιοτικές μεταβλητές ενώ η mean_value_of_short_term_variability η οποία σε σημαντικότητα αρχικά ήταν η πρώτη τώρα δεν έχει καν επιλεχθεί. Αυτό πολύ πιθανόν να συμβαίνει καθώς η επεξήγηση που μας δίνει να δίνεται μέσω άλλων μεταβλητών. Επιπλέον στην μεταβλητή prolongued decelerations παρατηρούμε ότι το επίπεδο 0.001 έχει εξαιρεθεί. Εμείς παρόλο που δεν έχει επιλεχθεί θα το προσθέσουμε για τεχνικούς λόγους. Έτσι προσαρμόζουμε το δέντρο το οποίο δημιουργείται μέσω του feature selection. Ως αποτελέσματα έχουμε.



	precision	recall	f1-score	support
1	0.89	0.92	0.90	145
2	0.84	0.83	0.83	87
3	0.96	0.90	0.93	51
accuracy			0.89	283
macro avg	0.90	0.88	0.89	283
weighted avg	0.89	0.89	0.89	283

Σχήμα 6.6.13 Πίνακας συγχύσεως τελικού μοντέλου Σχήμα 6.6.14 Classification Report τελικού μοντέλου

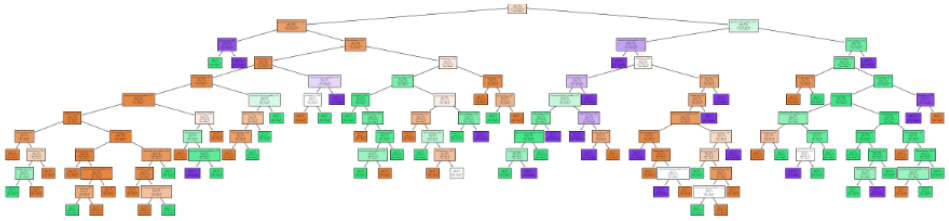
Στο μοντέλο αυτό βλέπουμε ελαφρώς βελτιωμένα νούμερα. Αξίζει όμως να δούμε λίγο παραπάνω την μεταβλητή mean_value_of_short_term_variability, η οποία εξαιρέθηκε στο μοντέλο μας ενώ πριν θεωρούταν η πιο σημαντική. Για αυτόν τον λόγο θα προσαρμόσουμε ένα μοντέλο και με αυτήν για να κάνουμε μία σύγκριση.



	precision	recall	f1-score	support
1	0.84	0.87	0.85	145
2	0.78	0.74	0.76	87
3	0.94	0.94	0.94	51
accuracy			0.84	283
macro avg	0.85	0.85	0.85	283
weighted avg	0.84	0.84	0.84	283

Σχήμα 6.6.15 Πίνακας συγχύσεως μοντέλου με MSTV Σχήμα 6.6.16 Classification Report μοντέλου με MSTV

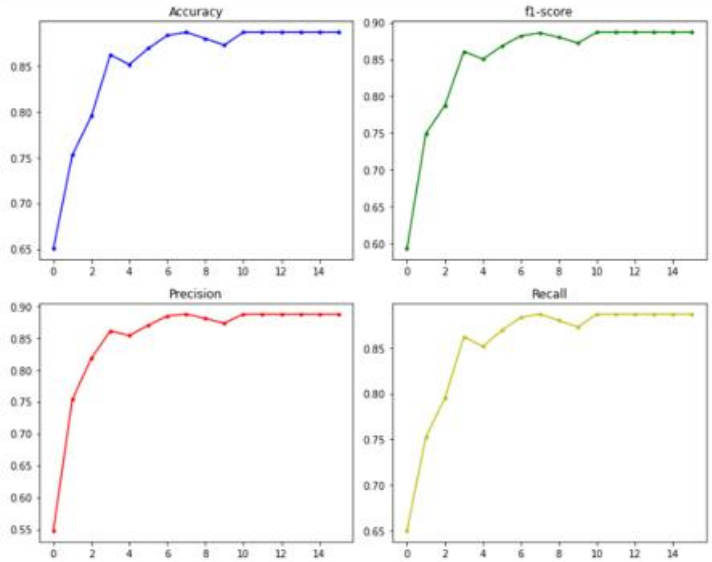
Βλέπουμε ότι με την εισόδου αυτής της μεταβλητής, ενώ υπάρχουν ήδη οι μεταβλητές που επιλέχθηκαν μέσω του forward selection, <<καταστρέφει>> τα αποτελέσματα μας, για τον λόγο αυτό θα την διατηρήσουμε εκτός της ανάλυσής μας και θα συνεχίσουμε με το προηγούμενο μοντέλο. Στην συνέχεια βλέπουμε την αρχιτεκτονική του δέντρου μας.



Σχήμα 6.6.17 Αρχιτεκτονική τελικού μοντέλου δέντρου

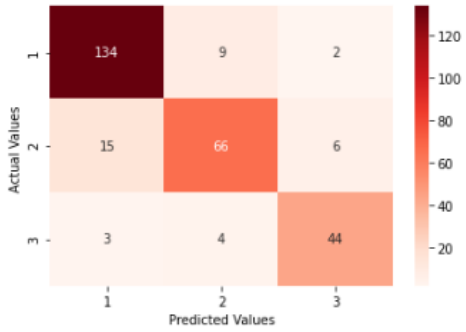
Όπως είπαμε και στο κεφάλαιο 5 που πρωτοπαρουσιάστηκαν θεωρητικά τα δέντρα απόφασης πολλές φορές προβαίνουμε σε κλάδεμα μερικών κλαδιών τω δέντρων με σκοπό να μειώσουμε την πολυπλοκότητα του μοντέλου. Για τον λόγο αυτό θα δημιουργήσουμε μία επανάληψη στην οποία κάθε φορά θα προσαρμόζεται ένα δέντρο ταξινόμησης με συγκεκριμένο βάθος στα φύλλα του και θα μετράμε την απόδοση του. Στην συνέχεια θα επιλέξουμε εκείνο του οποίου το κλάδεμα μας δίνει την καλύτερη απόδοση.

```
#Δημιουργούμε τα πιθανά μέγιστα επίπεδα που μπορεί να έχει το δέντρο μας
depths = range(1, 17)

# Δημιουργούμε τις λίστες που θα αποθηκεύσουμε της μετρικές απόδοσης του μοντέλου
accuracy_scores = []
f1_scores=[]
recall_scores=[]
precision_scores=[]

#Επαναλαμβάνουμε την διαδικασία της προσαρμογής του μοντέλου για όλα τα βάθη
for depth in depths:
    dt = DecisionTreeClassifier(max_depth=depth, random_state=42)
    dt.fit(X1_train, y_train)
    y_pred = dt.predict(X1_test)

    # Αποθηκεύουμε τα μέτρια αξιολόγησης του μοντέλου
    accuracy = accuracy_score(y_test, y_pred)
    precision_scores.append(precision_score(y_test, y_pred,average="weighted"))
    recall_scores.append(recall_score(y_test, y_pred,average="weighted"))
    f1_scores.append(f1_score(y_test, y_pred,average="weighted"))
    accuracy_scores.append(accuracy)
```



Σχήμα 6.6.18 Πορεία εξωτερικών μέτρων καθώς προβαίνουμε σε κλάδεμα

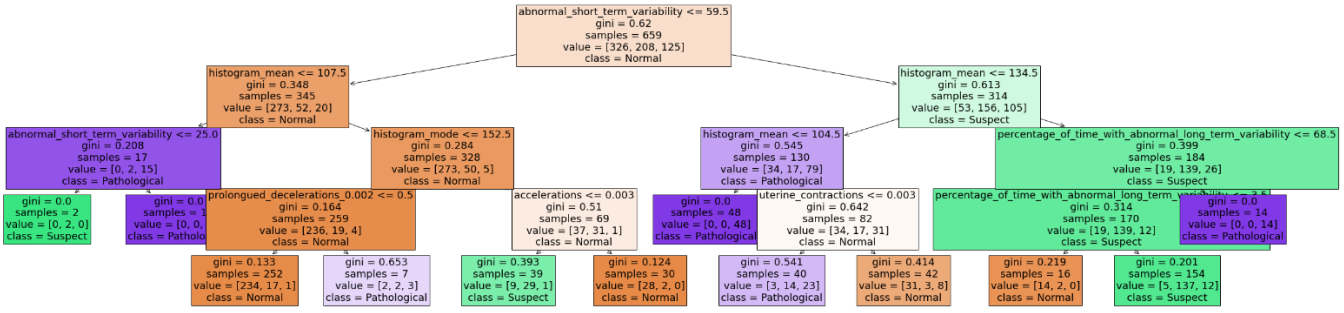
Παρατηρούμε να υπάρχει μία σχετική συμφωνία ανάμεσα στα γραφήματα. Δηλαδή το μέγιστο σημείο για όλες τις μετρικές να αντιστοιχεί στο σημείο 10 του άξονα x άρα σε βάθος 11 υπολογίζοντας ότι η Python ξεκινάει από το 0 την μέτρηση, ενώ μετά να μην έχουν ιδιαίτερες διακυμάνσεις. Όμως ο λόγος που γίνεται αυτό είναι ότι το δέντρο μας έχει μέγιστο βάθος 11.Αν θέλουμε επομένως να κλαδέψουμε για να μειώσουμε την πολυπλοκότητα του μοντέλου ένα καλό σημείο είναι το βάθος 4 στο οποίο γίνεται μία απότομη βελτίωση στις μετρικές μας. Επομένως θα προσαρμόσουμε το μοντέλο με βάθος 4 μόνο και μόνο για να δούμε λίγο την αρχιτεκτονική του και την απόδοση του συγκριτικά με το προηγούμενο μοντέλο το οποίο θα το κρατήσουμε ως τελικό καθώς αποδίδει καλύτερα στα δεδομένα μας.



	precision	recall	f1-score	support
1	0.88	0.92	0.90	145
2	0.84	0.76	0.80	87
3	0.85	0.86	0.85	51
accuracy			0.86	283
macro avg	0.85	0.85	0.85	283
weighted avg	0.86	0.86	0.86	283

Σχήμα 6.6.19 Πίνακας συγχύσεως μοντέλου βάθους 4 Σχήμα 6.6.20 Classification Report μοντέλου βάθους 4

Βλέπουμε ότι έχουμε εμφανώς μειωμένη απόδοση αλλά όχι σε πολύ χαμηλό βαθμό. Τέλος αξίζει να δούμε τα nodes που παρέμειναν μετά το κλάδεμα.



Σχήμα 6.6.21 Αρχιτεκτονική μοντέλου βάθους 4

Ενδιαφέρον έχει να προσπαθήσουμε να δούμε την διαδικασία για μία παρατήρηση του συνόλου ελέγχου. Έστω λοιπόν ότι παίρνουμε την πρώτη παρατήρηση του συνόλου ελέγχου. Έχουμε ότι.

accelerations	0.010
uterine_contractions	0.003
abnormal_short_term_variability	41.000
percentage_of_time_with_abnormal_long_term_variability	0.000
histogram_mode	186.000
histogram_mean	180.000
histogram_variance	11.000
severe_decelerations_0.001	0.000
prolongued_decelerations_0.001	0.000
prolongued_decelerations_0.002	0.000
prolongued_decelerations_0.003	0.000
prolongued_decelerations_0.004	0.000
prolongued_decelerations_0.005	0.000
histogram_tendency_0	0.000
histogram_tendency_1	1.000

Σχήμα 6.6.22 Χαρακτηριστικά πρώτης παρατήρησης συνόλου ελέγχου

Ξεκινάμε από την ρίζα ελέγχουμε την πρώτη συνθήκη

$$ASTV \leq 59.5 \,|True \;=>$$

το οποίο ισχύει οπότε συνεχίζουμε στα αριστερά

$$histogram\ mean \leq 107.5 \mid False =>$$

το οποίο δεν ισχύει οπότε συνεχίζουμε στα δεξιά

$$histogram\ mode \leq 152.5 \mid False =>$$

το οποίο δεν ισχύει οπότε συνεχίζουμε στα δεξιά

$$accelerations \leq 0.003 \,|False =>$$

το οποίο δεν ισχύει οπότε ως εκτίμηση για την απόκριση μας βάζουμε το δεξί φύλλο που προβλέπει την τιμή.

$$Estimated\ value\ \hat{Y} = 1\ (Normal)$$

Στην συνέχεια βλέπουμε την πραγματική τιμή της παρατήρησης η οποία είναι όντως ένα άρα το μοντέλο μας προέβλεψε σωστά. Με αυτόν τον τρόπο βλέπουμε την διαδικασία που γίνετε μέσω του δέντρου απόφασης για την εκτίμηση της ετικέτας μίας παρατήρησης. Θυμίζουμε ότι εμείς έχουμε καταλήξει στο δέντρο που αντιστοιχεί στα Σχήματα (6.6.13, 6.6.14, 6.6.17.) απλώς τώρα είδαμε την διαδικασία για το ίδιο μοντέλο στο οποίο έχουμε κλαδέψει ορισμένα από τα κλαδιά του.

6.7 Σύγκριση των τελικών τριών μοντέλων και χρήση της Ensemble Modeling μεθόδου.

Στις προηγούμενες παραγράφους του Κεφαλαίου 6 είδαμε την προσαρμογή και των τριών μοντέλων που παρουσιάστηκαν στα πλαίσια αυτής της εργασίας. Αξίζει να κάνουμε μία σύγκριση των μοντέλων καθώς εμείς πρέπει να προτείνουμε ένα μοντέλο για χρήση και να δούμε τις δυνατότητες και τις αδυναμίες του κάθε μοντέλου.

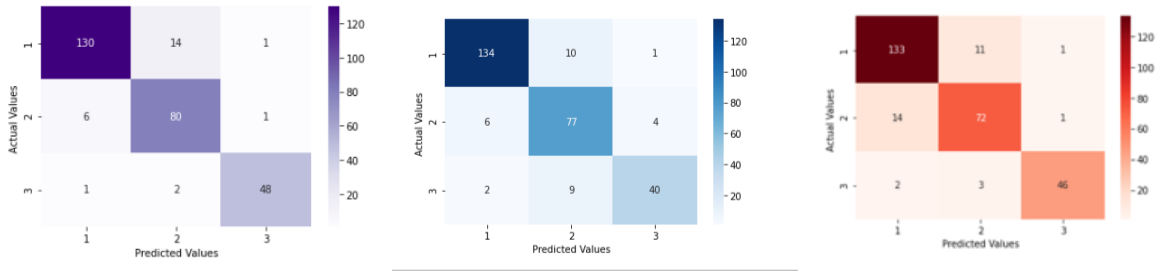
Αρχικά ας δούμε για το κάθε μοντέλο πόσες και ποιες από τις μεταβλητές απαιτούνται για την χρήση τους.

Features		
SVM	**Polynomial Regression**	**Classification Trees**
AC	AC	AC
UC	UC	UC
ASTV	ASTV	ASTV
ALTV	ALTV	ALTV
Prolongued Decelerations	Prolongued Decelerations	Prolongued Decelerations
Mean	Mean	Mean
Variance	Variance	Variance
LB-FHR	LB-FHR	Severe deceleration
	Mode	Mode
		Histogram tendency

Πίνακας 6.8 Μεταβλητές που επιλέχθηκαν για το κάθε μοντέλο

Βλέπουμε ότι το μεγαλύτερο κόστος ως προς τον αριθμό των μεταβλητών που απαιτούνται για την πρόβλεψη είναι αυτό των classification Trees με 10 μεταβλητές από τις 21.Στην συνέχεια έχουμε την πολυωνυμική παλινδρόμηση με 9 μεταβλητές ενώ τέλος το μοντέλο των SVM 8 το οποίο είναι το πιο οικονομικό. Αν δούμε ακόμα την ένωση όλων αυτών το μεταβλητών μιλάμε για 11 διαφορετικές μεταβλητές όπου οι 7 είναι κοινές για τα 3 μοντέλα. Ας ξαναθυμηθούμε τώρα τους πίνακες συγχύσεως των τριών μοντέλων.

SVM	Polynomial Regression	Classification Trees



Σχήμα 6.7.1 Πίνακες συγχύσεως των τριών τελικών μοντέλων

Παρατηρούμε ότι ως προς την πρώτη ομάδα την καλύτερη απόδοση την έχει η πολυωνυμική παλινδρόμηση κάνοντας 134 από τις 145 σωστές προβλέψεις με 10 λανθασμένα ταξινομημένες παρατηρήσεις στην κατηγορία 2 και μόλις 1 στην παθολογική κατάσταση. Ακολουθούν τα δέντρα ταξινόμησης με μόνη διαφορά σε μία παραπάνω παρατήρηση στις ύποπτες και τέλος τα SVM με 130 σωστές προβλέψεις 14 λανθασμένα ταξινομημένες στις ύποπτες και 1 λανθασμένα ταξινομημένη στην παθολογική. Ως προς την κατηγορία 2 έχουμε την καλύτερη επίδοση να την έχουν τα SVM με 80 από τις 87 σωστές προβλέψεις ενώ 6 να έχουν τοποθετηθεί ως φυσιολογικές και 1 ως παθολογική. Στην συνέχεια ακολουθεί η πολυωνυμική παλινδρόμηση με 77 σωστές προβλέψεις 6 λανθασμένα τοποθετημένες στην κατηγορία 1 και 4 λανθασμένα τοποθετημένες στην κατηγορία 3.Τελευταίο σε επίδοση ως προς την κατηογρία 2 έρχεται το μοντέλο των δέντρων απόφασης πετυχαίνοντας 72 σωστές προβλέψεις 14 λανθασμένα τοποθετημένες στην κατηγορία 1 και 1 λανθασμένα τοποθετημένη στην κατηγορία 3.Αναφορικά με την κατηγορία 3 βλέπουμε και πάλι την υψηλότερη επίδοση να την έχουν τα SVM με 48 από τις 51 σωστές ταξινομήσεις ενώ 1 να έχει τοποθετηθεί λανθασμένα στην κατηγορία 1 και 2 λανθασμένα στην κατηγορία 2.Ακολουθούν τα δέντρα απόφασης με 46 σωστές προβλέψεις, 2 λανθασμένα τοποθετημένες παρατηρήσεις στην κατηγορία 1 και 3 λανθασμένα τοποθετημένες στην κατηγορία 2.

Γενικότερα βλέπουμε ότι το μοντέλο των SVM έχει καλύτερη επίδοση στο να ανιχνεύει όταν υπάρχει κάτι ύποπτο σε σχέση με τα άλλα δύο μοντέλα. Το μοντέλο των δέντρων ταξινόμησης έχει μία γενικά καλή επίδοση με εξαίρεση την ομάδα 2 που υστερεί συγκριτικά με τα άλλα μοντέλα και το μοντέλο της πολυωνυμικής παλινδρόμησης έχει την καλύτερη επίδοση στο να εντοπίζει τα υγιείς περιπτώσεις ενώ υστερεί ελαφρά σε σύγκριση με τα άλλα δύο να ξεχωρίσει τις ομάδες των 2 και 3.Εμάς ως αναλυτές πρέπει να δούμε τι μας καλύπτει και τι επιθυμούμε γενικότερα να κάνει το μοντέλο μας. Θέλουμε να μπορούμε να διασφαλίζουμε την υγεία πιο εύκολα και να μην αναγκάζουμε τους ασθενείς όπου επρόκειτο για έγγυες γυναίκες στους τελευταίους μήνες της εγκυμοσύνης να κάνουν παραπάνω αχρείαστες εξετάσεις καθώς υπάρχει ταλαιπωρία για τις ίδιες και το έμβρυο, και σε συνδυασμό με την δημιουργία ανησυχίας στην μητέρα στο άκουσμα της ανίχνευσης κάποιου ανησυχητικού παράγοντα, η οποία μπορεί και αυτή με την σειρά της να φέρει επιπλέον επιπλοκές λόγο άγχους; Ή προτιμάμε την πιο σίγουρη διασφάλιση της υγείας με σκοπό τον περιορισμό πλήρως της πιθανότητας ότι δεν έχει ανιχνευθεί σωστά κάτι ανησυχητικό. Συνεχίζουμε την ανασκόπηση μας για την μελέτη της απόδοσης των μοντέλων βλέποντας τα αντίστοιχα classification report τους.

Classification report for optimized RBF SVM:

	precision	recall	f1-score	support
1	0.95	0.90	0.92	145
2	0.83	0.92	0.87	87
3	0.96	0.94	0.95	51
accuracy			0.91	283
macro avg	0.91	0.92	0.92	283
weighted avg	0.92	0.91	0.91	283

Σχήμα 6.7.2 Classification Report SVM

Classification report for Multinomial Regression

	precision	recall	f1-score	support
1	0.94	0.92	0.93	145
2	0.80	0.89	0.84	87
3	0.89	0.78	0.83	51
accuracy			0.89	283
macro avg	0.88	0.86	0.87	283
weighted avg	0.89	0.89	0.89	283

Σχήμα 6.7.3 Classification Report Πολυωνυμικής παλινδρόμησης

Classification report for Classification Trees

	precision	recall	f1-score	support
1	0.89	0.92	0.90	145
2	0.84	0.83	0.83	87
3	0.96	0.90	0.93	51
accuracy			0.89	283
macro avg	0.90	0.88	0.89	283
weighted avg	0.89	0.89	0.89	283

Σχήμα 6.7.4 Classification Report Δέντρων απόφασης

Metrics	SVM	Polynomial Regression	Classification Trees
Accuracy	0.91	0.89	0.89
Precision	0.91	0.88	0.9
Precision-weighted	0.92	0.89	0.89
Recall	0.92	0.86	0.88
Recall-weighted	0.91	0.89	0.89
F1-score	0.92	0.87	0.89
F1-score -weighted	0.91	0.89	0.89

Πίνακας 6.9 Μέτρα αξιολόγησης κάθε μοντέλου

Βλέπουμε ότι το μοντέλο των SVM είναι εκείνο το οποίο έχει τις καλύτερες αποδόσεις συνολικά στα δεδομένα ελέγχου αλλά οι διαφορές είναι αρκετά μικρές. Αν λάβουμε υπόψιν μας όσα έχουμε πει μέχρι τώρα δηλαδή το κόστος ως προς το σύνολο το μεταβλητών, την φύση του προβλήματος ως πρόβλημα ιατρικής επομένως η διασφάλιση της υγείας να παίζει κομβικό ρόλο και την επίδοση στα δεδομένα ελέγχου κάποιος θα μπορούσε να πει ότι το μοντέλο των Support Vector Machines αποτελεί την καλύτερη επιλογή από τις διαθέσιμες. Όμως επειδή οι διαφορές

είναι αρκετά μικρές και καθώς με όλα τα μοντέλα μπορούμε να αντιμετωπίσουμε ικανοποιητικά το πρόβλημα, μπορούμε να δούμε εκτός από τα νούμερα τι άλλο μας προσφέρει το κάθε μοντέλο.

Αρχικά τα support vector machines εκτός από την υψηλή απόδοση μας προσφέρουν την δυνατότητα την γραφικής αναπαράστασης των δεδομένων και των συνόρων τα οποία τα χωρίζουν, πράγμα το οποίο για την περιγραφή του προβλήματος είναι σημαντικό πλεονέκτημα. Η πολυωνυμική παλινδρόμηση μας προσφέρει την δυνατότητα της ερμηνείας της κάθε μεταβλητής που έχει κριθεί στατιστικά σημαντική και περιλαμβάνεται στο μοντέλο. Μέσω αυτού εμείς μπορούμε να δούμε τις επιρροές που έχει η κάθε μεταβλητή σε τυχόν διακυμάνσεις της. Τέλος τα δέντρα απόφασης μας δίνουν γραφικά μία εύκολη χρήση του μοντέλου και παρακολούθηση της πορείας του φιλτραρίσματος της εκάστοτε παρατήρησης.

Συνοψίζοντας θα συστήναμε και τα τρία μοντέλα για χρήση και θα παροτρύναμε τον εκάστοτε επιβλέποντα γιατρό αναλόγως με τις προτεραιότητες που θεωρεί εκείνος ως ειδικός και τις ανάγκες της κάθε περίπτωσης να χρησιμοποιήσει το αντίστοιχο μοντέλο. Θα μπορούσαμε όμως να αξιοποιήσουμε και τα τρία μοντέλα μαζί.

Κλείνοντας αυτήν την ενότητα θα παρουσιάσουμε την ensemble modeling method είναι μία τεχνική η οποία μπορεί να χρησιμοποιηθεί όταν έχουμε παραπάνω από ένα μοντέλο ταξινόμησης [23]. Αυτό που κάνουμε ουσιαστικά είναι να συνδυάζουμε τα μοντέλα αυτά μέσω των προβλέψεων τους για να επιτύχουμε καλύτερη πρόβλεψη στο αποτέλεσμα. Έτσι προβαίνουμε σε εκτίμηση για την ετικέτα Y μέσω και των τριών μοντέλων ξεχωριστά και στο τέλος ως πρόβλεψη για την απόκριση ορίζουμε την επικρατούσα τιμή των εκτιμήσεων. Με αυτό τον τρόπο καταφέρνουμε να φιλτράρουμε το αποτέλεσμα μέσα από διαφορετικά μοντέλα με σκοπό να έχουμε περισσότερη αξιοπιστία και ακρίβεια στα αποτελέσματα μας.

Στην περίπτωση μας καθώς όπως είπαμε και στην αρχή της Παραγράφου 6.7 η ένωση του συνόλου των features που χρησιμοποιούμε για κάθε μοντέλο είναι 11, καταλαβαίνουμε ότι δεν θα αυξηθεί και σημαντικά το κόστος για την εφαρμογή του ensemble model. Έτσι λοιπόν δημιουργούμε ένα dataframe το οποίο περιέχει όλες τις προβλέψεις πάνω στο σύνολο ελέγχου και προσθέτουμε και μία τέταρτη στήλη η οποία περιέχει την επικρατούσα τιμή της κάθε γραμμής. Επειδή έχουμε 3 στήλες και τρεις πιθανές ετικέτες θα προβούμε σε έλεγχο της κάθε γραμμής για να ελέγξουμε αν υπάρχει έστω και μία γραμμή όπου έχουμε ολική ασυμφωνία των μοντέλων με την έννοια ότι το κάθε μοντέλο προβλέπει διαφορετική τιμή. Τότε θα το κατατάσσουμε σε ύποπτη κατάσταση ώστε να θορυβήσουμε τον γιατρό για να το διερευνήσει παραπάνω μέσω των πλεονεκτημάτων του κάθε μοντέλου.

```
ensemble_model=pd.DataFrame([y_pred_rbf,y_pred,predictions]).T
u= []
#Δημιουργία επανάληψης για έλεγχο πλήρης ασυμφωνίας προβλέψεων
for i in range(len(ensemble_model)):
    #έλεγχος αν το πλήθος των διαφορετικών προβλέψεων ισούται με 3
    if len(set(ensemble_model.iloc[i, :])) == 3:
        #αν ναι κρατάμε τον δείκτη της παρατήρησης
        u.append(i)
print(u)
```

```
[]
```

Σχήμα 6.7.5 Κώδικας δημιουργίας πρόβλεψης μέσω της Ensemble modeling μεθόδου

Βλέπουμε ότι δεν υπάρχει παρατήρηση στην οποία να έχουμε πλήρη ασυμφωνία. Στο Σχήμα 6.7.6 βλέπουμε την μορφή που έχει το dataframe μας, στην τελευταία στήλη του οποίου περιέχονται οι προβλέψεις για την απόκριση μας.

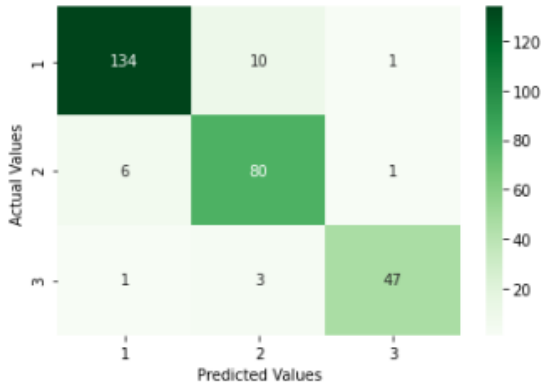
```
ensemble_model['Prediction'] = cb_pred.mode(axis=1)
ensemble_model.columns=["SVM","Decision Trees","Multinomial Regression","Prediction"]
ensemble_model
```

	SVM	Decision Trees	Multinomial Regression	Prediction
0	1	1	1	1
1	3	3	3	3
2	3	3	3	3
3	2	1	2	2
4	2	2	2	2
...	...	...	...	...
278	1	1	1	1
279	1	1	1	1
280	1	1	1	1
281	1	1	1	1
282	1	1	1	1

Σχήμα 6.7.6 Πρόβλεψη μέσω της ensemble modeling μεθόδου

Σε κάθε γραμμή βλέπουμε πως διαμορφώνονται οι προβλέψεις μας σύμφωνα με τις προβλέψεις των τριών μοντέλων. Για παράδειγμα στην γραμμή 3 βλέπουμε να έχει προβλεφθεί από τα SVM και την πολυωνυμική παλινδρόμηση ότι η παρατήρηση είναι ύποπτη ενώ από τα δέντρα απόφασης έχει προβλεφθεί ως φυσιολογική. Επομένως η παρατήρηση ταξινομείται ως ύποπτη.

Ας πάμε όμως να δούμε την απόδοση του ensemble μοντέλου συνολικά στο σύνολο ελέγχου.



Σχήμα 6.7.7 Πίνακας συγχύσεως ensemble modeling μεθόδου

Μέσω του πίνακα συγχύσεως βλέπουμε να έχουμε την καλύτερη γενική εικόνα που είδαμε μέχρι τώρα. Συγκεκριμένα βλέπουμε ότι 134 από τις 145 φυσιολογικές παρατηρήσεις μας έχουν

ταξινομηθεί ορθά ενώ 10 έχουν ταξινομηθεί ως ύποπτες και μόλις μία ως παθολογική. Από τις 87 ύποπτες παρατηρήσεις βλέπουμε να έχουμε 80 σωστά ταξινομημένες παρατηρήσεις ενώ 6 έχουν ταξινομηθεί ως φυσιολογικές και 1 ως παθολογική. Τέλος από τις 51 παθολογικές περιπτώσεις έχουμε 47 να έχουν προβλεφθεί σωστά ενώ έχουμε 3 λανθασμένα ταξινομημένες ως ύποπτες και 1 ταξινομημένη ως κανονική. Στην συνέχεια βλέπουμε και το αντίστοιχο classification report.

Classification report for Ensemble modeling:

	precision	recall	f1-score	support
1	0.95	0.92	0.94	145
2	0.86	0.92	0.89	87
3	0.96	0.92	0.94	51
accuracy			0.92	283
macro avg	0.92	0.92	0.92	283
weighted avg	0.92	0.92	0.92	283

Σχήμα 6.7.8 Classification Report ensemble modeling μεθόδου

Στην εικόνα αυτή φαίνεται ξεκάθαρα ότι μοντέλο αυτό που συνδυάζει όλα τα προηγούμενα τρία υπερέχει με διαφορά αφού οι τιμές που λαμβάνουμε ξεπερνούν το 92% (με εξαίρεση δύο).

6.8 Συμπεράσματα

Στα πλαίσια αυτής της διπλωματικής αρχικά παρουσιάστηκαν 3 μοντέλα ταξινόμησης τόσο στο πλαίσιο της στατιστικής θεωρίας όσο και μέσα από παραδείγματα. Στην συνέχεια παρουσιάσαμε ένα πραγματικό πρόβλημα από τον κλάδο της Ιατρικής και πως η εφαρμογή αυτών των μοντέλων μπορεί με επιτυχία να αντιμετωπίσει τέτοιου είδους προβλήματα.

Συγκεκριμένα παρουσιάζοντας την ιατρική εξέταση μέσω των καρδιοτοκογραφημάτων και την καταγραφή διάφορων χαρακτηριστικών της μητέρας ή του εμβρύου, εμείς καλούμασταν να δημιουργήσουμε ένα μοντέλο το οποίο θα έπρεπε να είναι σε θέση να προβλέψει αξιόπιστα αν επρόκειτο για μία φυσιολογική, μία ύποπτη ή μια παθολογική περίπτωση ως προς την κατάσταση και υγεία του εμβρύου με σκοπό είτε την περαιτέρω μελέτη της ύποπτης περίπτωσης είτε της ιατρικής παρέμβασης σε παθολογική περίπτωση. Επεξεργαζόμενοι τα δεδομένα και κάνοντας κάποιες προσαρμογές και διορθώσεις σε αυτά, στην συνέχεια προχωρήσαμε στην προσαρμογή των τριών αυτών μοντέλων και στην βελτιστοποίηση τους. Καταλήξαμε ότι όλα τα μοντέλα απέδιδαν πολύ καλά σε καινούργια δεδομένα που δεν είχαν ξανά αντιμετωπίσει με μικρές διαφορές.

Πιο συγκεκριμένα είχαμε το μοντέλο των SVM να ανιχνεύει καλύτερα τις παθολογικές και τις ύποπτες περιπτώσεις διασφαλίζοντας έτσι περισσότερο την έγκυρη διάγνωση στις κακές περιπτώσεις, το μοντέλο πολυωνυμικής παλινδρόμησης ανίχνευε καλύτερα τις υγιείς περιπτώσεις ενώ το μοντέλο των δέντρων απόφασης αποτελούσε μία ενδιάμεση κατάσταση με μία μικρή αδυναμία ως προς την ύποπτη ομάδα. Ακόμα είδαμε τα οφέλη του κάθε μοντέλου πέραν των επιδόσεων και του κόστους τους. Πιο συγκεκριμένα είδαμε ότι μέσω των SVM στον αναλυτή ή στον επιβλέποντα γιατρό δίνετε το πλεονέκτημα της γραφικής αναπαράστασης των δεδομένων

μας και η πληροφορία για το πόσο κοντά ή όχι είναι μία παρατήρηση στα σύνορα που χωρίζουν την κάθε ομάδα. Στην πολυωνυμική παλινδρόμηση από την άλλη δίνεται επιπλέον η δυνατότητα της ερμηνείας των επεξηγηματικών μεταβλητών και η γρήγορη πρόβλεψη του αποτελέσματος σε μεμονωμένες αλλαγές αυτών, ενώ για τα δέντρα απόφασης δίνεται η επιλογή της εποπτείας της μεθόδου καθώς αυτή εκτελείται.

Αναφορικά με το κόστος είδαμε ότι το λιγότερο κοστοβόρο μοντέλο ήταν αυτό των SVM, ακολουθούσε εκείνο της πολυωνυμικής παλινδρόμησης ενώ τρίτο ερχόταν το μοντέλο των δέντρων απόφασης με διαφορά μόλις μίας μεταβλητής κάθε φορά. Τέλος παρουσιάστηκε η μέθοδος του ensemble modeling η οποία συνδυάζει και τα τρία αυτά μοντέλα πετυχαίνοντας έτσι την μέγιστη απόδοση ως προς όλα τα μέτρα αξιολόγησης χωρίς να ανεβάζει σημαντικά το κόστος ως προς το πλήθος των μεταβλητών που απαιτούνται για αυτήν, γεγονός που την κάνει να προτείνεται για ασφαλέστερη πρόβλεψη.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνική

[1] Ηλιόπουλος Γ. (2022). *Γενικευμένα γραμμικά μοντέλα*, Σημειώσεις Μεταπτυχιακού Μαθήματος του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης Πανεπιστήμιο Πειραιώς.

[2] Κούτρας Μ. (2022). *Εφαρμοσμένη Πολυμεταβλητή Ανάλυση,* Σημειώσεις Μεταπτυχιακού Μαθήματος του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης Πανεπιστήμιο Πειραιώς.

[3] Κούτρας Μ., Μπούτσικας Μ. (2011). *Στατιστική ΙΙ,* Σημειώσεις Προπτυχιακού Μαθήματος του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης.

[4] Κρητικού Δ. (2019). *Παραμετρική Στατιστική*, Σημειώσεις Προπτυχιακού Μαθήματος του τμήματος Μαθηματικών του Πανεπιστημίου Κρήτης.

[5] Μπερσίμης Σ. (2021). *Στατιστική μηχανική μάθηση,* Σημειώσεις Μεταπτυχιακού Μαθήματος του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης Πανεπιστήμιο Πειραιώς.

[6] Πελέκης Ν. (2023). *Στατιστικές μέθοδοι εξόρυξης δεδομένων,* Σημειώσεις Μεταπτυχιακού Μαθήματος του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης.

[7] Πολίτης Κ. (2022), *Γενικευμένα γραμμικά μοντέλα*, Σημειώσεις Μεταπτυχιακού Μαθήματος του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης Πανεπιστήμιο Πειραιώς.

Ξένη

[8] Ayres de Campos, D., Sisio dos Santos, M. E., Bernardes, J., & Costa-Pereira, A. (2000). Cardiotocography dataset, UCI Machine Learning Repository.

[9] Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory (COLT '92) (pp. 144-152). Association for Computing Machinery. doi:10.1145/130385.130401

[10] Burges, C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2), 121-167. doi:10.1023/A:1009715923555

[11] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297. doi:10.1007/BF00994018

[12] Deressa, T. D., & Kadam, K. (2018). Prediction of Fetal Health State during Pregnancy: A Survey. *International Journal of Computer Science Trends and Technology (IJCST)*, 6(1), 29.

[13] Everitt, B. S. and Dunn, G. (1991). *Applied Multivariate Data Analysis*, Arnold, New York.

[14] Flury, B., & Riedwyl, H. (1988). *Multivariate Statistics: A Practical Approach*. Prentice Hall, New York.

[15] Georgoulas, G., Stylios, D., & Groumpos, P. (2006). Predicting the risk of metabolic acidosis for newborns based on fetal heart rate signal classification using support vector machines. *IEEE Transactions on Biomedical Engineering*, 53(5), 875-884. doi:10.1109/TBME.2006.872814

[16] Hsu, C.-W., & Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2), 415-425. doi:10.1109/72.991427

[17] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. Springer, New York. doi:10.1007/978-1-4614-7138-7

[18] Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). Springer Series in Statistics. Springer-Verlag New York.

[19] Li, J., & Liu, X. (2021). Fetal Health Classification Based on Machine Learning. In *2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)* (pp. 899-902). Nanchang, China. doi:10.1109/ICBAIE52039.2021.9389902

[20] Nocedal, J., & Wright, S. J. (1999). Sequential Quadratic Programming. In *Numerical Optimization* (pp. 526-573). Springer New York. doi:10.1007/0-387-22742-3_18

[21] Powers, D. M. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies. 2:1 pp.37-63

[22] Sasaki, Y. (2007). *The truth of the F-measure.* Version: 26th October, 2007. Retrieved from https://nicolasshu.com/assets/pdf/Sasaki_2007_The%20Truth%20of%20the%20F-measure.pdf

[23] Zhou, Z. (2006). Ensemble methods in machine learning: A survey. *ACM Computing Surveys*, 38(1), Article 13.

[24] Zisserman, A. (2015). *The SVM Classifier, Lecture 2, C19 Machine learning. Hilary.* Retrieved from https://www.robots.ox.ac.uk/~az/lectures/ml/lect2.pdf

ΠΑΡΑΡΤΗΜΑ

Παράγραφος 6.5

converged:	False	LL-Null:	-677.11
Covariance Type:	nonrobust	LLR p-value:	1.138e-163

fetal_health=2	coef	std err	z	P>|z|	[0.025	0.975]
const	-17.6182	3.171	-5.556	0.000	-23.833	-11.403
baseline value	-0.1529	0.061	-2.513	0.012	-0.272	-0.034
accelerations	-1053.4087	168.329	-6.258	0.000	-1383.328	-723.490
fetal_movement	9.2918	12.240	0.759	0.448	-14.698	33.282
uterine_contractions	-232.9400	71.131	-3.275	0.001	-372.354	-93.526
light_decelerations	183.1254	139.807	1.310	0.190	-90.891	457.141
abnormal_short_term_variability	0.0960	0.019	5.047	0.000	0.059	0.133
mean_value_of_short_term_variability	-0.2970	0.479	-0.620	0.535	-1.236	0.641
percentage_of_time_with_abnormal_long_term_variability	0.0124	0.011	1.176	0.240	-0.008	0.033
mean_value_of_long_term_variability	-0.0950	0.062	-1.533	0.125	-0.216	0.026
histogram_min	0.0071	0.012	0.603	0.546	-0.016	0.030
histogram_max	0.0106	0.021	0.517	0.605	-0.030	0.051
histogram_number_of_peaks	0.1398	0.080	1.745	0.081	-0.017	0.297
histogram_number_of_zeroes	0.0542	0.224	0.242	0.809	-0.385	0.493
histogram_mode	-0.0629	0.050	-1.247	0.212	-0.162	0.036
histogram_mean	0.4336	0.107	4.066	0.000	0.225	0.643
histogram_median	-0.1359	0.126	-1.082	0.279	-0.382	0.110
histogram_variance	0.0386	0.015	2.602	0.009	0.010	0.068
severe_decelerations_0.001	-0.4826	3.1e+05	-1.56e-06	1.000	-6.08e+05	6.08e+05
prolongued_decelerations_0.001	3.8483	1.180	3.262	0.001	1.536	6.161
prolongued_decelerations_0.002	6.8456	1.446	4.734	0.000	4.011	9.680
prolongued_decelerations_0.003	8.2768	1.889	4.381	0.000	4.574	11.980
prolongued_decelerations_0.004	2.4847	2.22e+04	0.000	1.000	-4.34e+04	4.34e+04
prolongued_decelerations_0.005	-2.6560	4.81e+05	-5.52e-06	1.000	-9.43e+05	9.43e+05
histogram_tendency_0	-0.2354	0.859	-0.274	0.784	-1.919	1.449
histogram_tendency_1	-0.4152	1.060	-0.392	0.695	-2.493	1.663

fetal_health=3	coef	std err	z	P>|z|	[0.025	0.975]
const	-22.7280	4.871	-4.666	0.000	-32.275	-13.181
baseline value	0.1399	0.082	1.699	0.089	-0.022	0.301
accelerations	-930.2924	502.742	-1.850	0.064	-1915.649	55.064
fetal_movement	22.4654	13.440	1.672	0.095	-3.876	48.807
uterine_contractions	-191.9355	103.917	-1.847	0.065	-395.608	11.737
light_decelerations	326.3405	153.604	2.125	0.034	25.282	627.400
abnormal_short_term_variability	0.2065	0.032	6.556	0.000	0.145	0.268
mean_value_of_short_term_variability	-0.7386	0.651	-1.135	0.256	-2.014	0.537
percentage_of_time_with_abnormal_long_term_variability	0.0733	0.016	4.670	0.000	0.043	0.104
mean_value_of_long_term_variability	0.0311	0.101	0.309	0.758	-0.166	0.228
histogram_min	-0.0028	0.017	-0.173	0.863	-0.035	0.029
histogram_max	0.0380	0.028	1.350	0.177	-0.017	0.093
histogram_number_of_peaks	-0.4186	0.162	-2.584	0.010	-0.736	-0.101
histogram_number_of_zeroes	0.3372	0.404	0.835	0.404	-0.454	1.128
histogram_mode	-0.0378	0.062	-0.607	0.544	-0.160	0.084
histogram_mean	0.0464	0.072	0.647	0.518	-0.094	0.187
histogram_median	-0.1290	0.103	-1.252	0.210	-0.331	0.073
histogram_variance	0.0531	0.017	3.207	0.001	0.021	0.086
severe_decelerations_0.001	25.0578	7.5e+04	0.000	1.000	-1.47e+05	1.47e+05
prolongued_decelerations_0.001	2.7767	1.437	1.932	0.053	-0.041	5.594
prolongued_decelerations_0.002	8.1875	1.763	4.643	0.000	4.731	11.644
prolongued_decelerations_0.003	9.7665	2.304	4.239	0.000	5.251	14.282
prolongued_decelerations_0.004	17.9572	1.23e+04	0.001	0.999	-2.41e+04	2.41e+04
prolongued_decelerations_0.005	17.5107	1.23e+04	0.001	0.999	-2.42e+04	2.42e+04
histogram_tendency_0	-0.7963	1.087	-0.733	0.464	-2.927	1.334
histogram_tendency_1	-0.8858	1.348	-0.657	0.511	-3.527	1.755

Π.1 Output μοντέλου πολυωνυμικής παλινδρόμησης χωρίς την **Histogram width.**

converged:	False	LL-Null:	-677.11
Covariance Type:	nonrobust	LLR p-value:	1.357e-164

fetal_health=2	coef	std err	z	P>|z|	[0.025	0.975]
const	-17.2729	3.128	-5.522	0.000	-23.404	-11.142
baseline value	-0.1524	0.061	-2.509	0.012	-0.271	-0.033
accelerations	-1037.3198	165.561	-6.265	0.000	-1361.813	-712.827
fetal_movement	8.9563	12.049	0.743	0.457	-14.659	32.571
uterine_contractions	-228.1740	70.485	-3.237	0.001	-366.321	-90.027
light_decelerations	193.3764	138.682	1.394	0.163	-78.435	465.188
abnormal_short_term_variability	0.0964	0.019	5.062	0.000	0.059	0.134
mean_value_of_short_term_variability	-0.2642	0.474	-0.557	0.578	-1.194	0.665
percentage_of_time_with_abnormal_long_term_variability	0.0124	0.011	1.175	0.240	-0.008	0.033
mean_value_of_long_term_variability	-0.0874	0.061	-1.432	0.152	-0.207	0.032
histogram_min	0.0043	0.011	0.395	0.693	-0.017	0.025
histogram_number_of_peaks	0.1493	0.077	1.938	0.053	-0.002	0.300
histogram_number_of_zeroes	0.0502	0.226	0.222	0.824	-0.393	0.493
histogram_mode	-0.0598	0.050	-1.199	0.231	-0.158	0.038
histogram_mean	0.4382	0.106	4.152	0.000	0.231	0.645
histogram_median	-0.1321	0.125	-1.058	0.290	-0.377	0.113
histogram_variance	0.0393	0.014	2.722	0.006	0.011	0.068
severe_decelerations_0.001	0.3886	6.53e+04	5.95e-06	1.000	-1.28e+05	1.28e+05
prolongued_decelerations_0.001	3.8784	1.166	3.325	0.001	1.593	6.164
prolongued_decelerations_0.002	6.8810	1.432	4.804	0.000	4.074	9.688
prolongued_decelerations_0.003	8.2742	1.882	4.397	0.000	4.586	11.963
prolongued_decelerations_0.004	0.4967	2.82e+05	1.76e-06	1.000	-5.53e+05	5.53e+05
prolongued_decelerations_0.005	-1.8354	3.41e+06	-5.39e-07	1.000	-6.68e+06	6.68e+06
histogram_tendency_0	-0.4032	0.800	-0.504	0.614	-1.972	1.165
histogram_tendency_1	-0.7440	0.889	-0.837	0.403	-2.486	0.998

fetal_health=3	coef	std err	z	P>|z|	[0.025	0.975]
const	-22.0863	4.788	-4.613	0.000	-31.471	-12.701
baseline value	0.1366	0.082	1.663	0.096	-0.024	0.298
accelerations	-754.8842	430.542	-1.753	0.080	-1598.731	88.963
fetal_movement	20.2886	12.880	1.575	0.115	-4.955	45.533
uterine_contractions	-195.8561	103.177	-1.898	0.058	-398.080	6.367
light_decelerations	372.9284	150.808	2.473	0.013	77.350	668.507
abnormal_short_term_variability	0.2097	0.031	6.720	0.000	0.149	0.271
mean_value_of_short_term_variability	-0.5020	0.625	-0.803	0.422	-1.726	0.723
percentage_of_time_with_abnormal_long_term_variability	0.0737	0.016	4.703	0.000	0.043	0.104
mean_value_of_long_term_variability	0.0665	0.096	0.690	0.490	-0.122	0.255
histogram_min	-0.0079	0.016	-0.492	0.623	-0.039	0.024
histogram_number_of_peaks	-0.3394	0.148	-2.291	0.022	-0.630	-0.049
histogram_number_of_zeroes	0.2686	0.375	0.715	0.474	-0.467	1.004
histogram_mode	-0.0300	0.059	-0.510	0.610	-0.145	0.085
histogram_mean	0.0568	0.073	0.774	0.439	-0.087	0.201
histogram_median	-0.1032	0.099	-1.047	0.295	-0.296	0.090
histogram_variance	0.0567	0.016	3.448	0.001	0.024	0.089
severe_decelerations_0.001	23.1986	2.25e+04	0.001	0.999	-4.41e+04	4.41e+04
prolongued_decelerations_0.001	2.7694	1.417	1.955	0.051	-0.007	5.546
prolongued_decelerations_0.002	8.4370	1.747	4.830	0.000	5.013	11.861
prolongued_decelerations_0.003	9.7038	2.296	4.227	0.000	5.204	14.203
prolongued_decelerations_0.004	21.8274	6.35e+04	0.000	1.000	-1.24e+05	1.24e+05
prolongued_decelerations_0.005	22.6180	1.13e+05	0.000	1.000	-2.22e+05	2.22e+05
histogram_tendency_0	-1.3984	0.962	-1.453	0.146	-3.285	0.488
histogram_tendency_1	-1.9220	1.080	-1.780	0.075	-4.038	0.194

Π.2 Output μοντέλου πολυωνυμικής παλινδρόμησης χωρίς την **Histogram max.**

```
Warning: Maximum number of iterations has been exceeded.
         Current function value: 0.319218
         Iterations: 35
                          MNLogit Regression Results
==============================================================================
Dep. Variable:           fetal_health   No. Observations:                  659
Model:                        MNLogit   Df Residuals:                      611
Method:                           MLE   Df Model:                           46
Date:                Mon, 10 Apr 2023   Pseudo R-squ.:                 0.6893
Time:                        21:26:20   Log-Likelihood:                -210.36
converged:                      False   LL-Null:                       -677.11
Covariance Type:            nonrobust   LLR p-value:                9.652e-166
=================================================================================================================================
                                    fetal_health=2       coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------------------------------
const                                                -17.1004      3.108     -5.501      0.000     -23.193     -11.008
baseline.value                                        -0.1520      0.060     -2.525      0.012      -0.270      -0.034
accelerations                                      -1041.0219    164.957     -6.311      0.000   -1364.331    -717.713
fetal_movement                                         9.3301     11.943      0.781      0.435     -14.078      32.739
uterine_contractions                                -221.6915     69.073     -3.210      0.001    -357.071     -86.312
light_decelerations                                  177.1485    139.441      1.270      0.204     -96.150     450.447
abnormal_short_term_variability                        0.0958      0.019      5.033      0.000       0.059       0.133
mean_value_of_short_term_variability                  -0.3050      0.473     -0.645      0.519      -1.231       0.621
percentage_of_time_with_abnormal_long_term_variability     0.0122      0.011      1.162      0.245      -0.008       0.033
mean_value_of_long_term_variability                   -0.0980      0.059     -1.659      0.097      -0.214       0.018
histogram_number_of_peaks                              0.1354      0.068      1.991      0.047       0.002       0.269
histogram_number_of_zeroes                             0.0700      0.225      0.311      0.755      -0.371       0.511
histogram_mode                                        -0.0575      0.050     -1.149      0.251      -0.156       0.041
histogram_mean                                         0.4427      0.106      4.194      0.000       0.236       0.650
histogram_median                                      -0.1354      0.125     -1.081      0.280      -0.381       0.110
histogram_variance                                     0.0386      0.014      2.693      0.007       0.010       0.067
severe_decelerations_0.001                            -8.4838   6076.214     -0.001      0.999   -1.19e+04    1.19e+04
prolongued_decelerations_0.001                         3.9267      1.158      3.390      0.001       1.656       6.197
prolongued_decelerations_0.002                         6.7854      1.429      4.747      0.000       3.984       9.587
prolongued_decelerations_0.003                         8.3592      1.852      4.513      0.000       4.729      11.989
prolongued_decelerations_0.004                         0.6853   1.85e+05   3.71e-06      1.000   -3.62e+05    3.62e+05
prolongued_decelerations_0.005                        -1.5227      7e+05  -2.18e-06      1.000   -1.37e+06    1.37e+06
histogram_tendency_0                                  -0.4489      0.790     -0.568      0.570      -1.997       1.099
histogram_tendency_1                                  -0.8458      0.831     -1.017      0.309      -2.475       0.784
---------------------------------------------------------------------------------------------------------------------------------
                                    fetal_health=3       coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------------------------------
const                                                -22.2066      4.777     -4.648      0.000     -31.570     -12.843
baseline.value                                         0.1335      0.082      1.621      0.105      -0.028       0.295
accelerations                                       -745.8648    433.476     -1.721      0.085   -1595.462     103.733
fetal_movement                                        19.3086     12.733      1.516      0.129      -5.648      44.265
uterine_contractions                                -187.2438    103.304     -1.813      0.070    -389.716      15.229
light_decelerations                                  388.3782    151.163      2.569      0.010      92.105     684.652
abnormal_short_term_variability                        0.2089      0.031      6.704      0.000       0.148       0.270
mean_value_of_short_term_variability                  -0.5063      0.628     -0.806      0.420      -1.737       0.724
percentage_of_time_with_abnormal_long_term_variability     0.0736      0.016      4.731      0.000       0.043       0.104
mean_value_of_long_term_variability                    0.0915      0.092      0.998      0.318      -0.088       0.271
histogram_number_of_peaks                             -0.3161      0.141     -2.243      0.025      -0.592      -0.040
histogram_number_of_zeroes                             0.2483      0.372      0.667      0.505      -0.481       0.978
histogram_mode                                        -0.0302      0.059     -0.508      0.611      -0.146       0.086
histogram_mean                                         0.0458      0.073      0.629      0.530      -0.097       0.189
histogram_median                                      -0.0966      0.099     -0.974      0.330      -0.291       0.098
histogram_variance                                     0.0564      0.016      3.451      0.001       0.024       0.088
severe_decelerations_0.001                            43.5146        nan        nan        nan         nan         nan
prolongued_decelerations_0.001                         2.8237      1.436      1.966      0.049       0.009       5.638
prolongued_decelerations_0.002                         8.5328      1.764      4.838      0.000       5.076      11.990
prolongued_decelerations_0.003                         9.8135      2.286      4.293      0.000       5.334      14.293
prolongued_decelerations_0.004                        21.5481   4.57e+04      0.000      1.000   -8.96e+04    8.97e+04
prolongued_decelerations_0.005                        20.1482   2.73e+04      0.001      0.999   -5.36e+04    5.36e+04
histogram_tendency_0                                  -1.3476      0.952     -1.415      0.157      -3.214       0.519
histogram_tendency_1                                  -1.7582      0.991     -1.775      0.076      -3.700       0.183
=================================================================================================================================
```

Π.3 Output μοντέλου πολυωνυμικής παλινδρόμησης χωρίς την **Histogram min.**

```
Warning: Maximum number of iterations has been exceeded.
         Current function value: 0.321731
         Iterations: 35
                           MNLogit Regression Results
==============================================================================
Dep. Variable:           fetal_health   No. Observations:                  659
Model:                        MNLogit   Df Residuals:                      613
Method:                           MLE   Df Model:                           44
Date:                Mon, 10 Apr 2023   Pseudo R-squ.:                0.6869
Time:                        21:26:20   Log-Likelihood:               -212.02
converged:                      False   LL-Null:                      -677.11
Covariance Type:            nonrobust   LLR p-value:                2.206e-166
=========================================================================================================================
                                        fetal_health=2       coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------------------
const                                                    -16.8516      3.092     -5.449      0.000     -22.913     -10.791
baseline.value                                            -0.1457      0.060     -2.441      0.015      -0.263      -0.029
accelerations                                          -1036.8082    164.770     -6.292      0.000   -1359.751    -713.865
fetal_movement                                             9.4229     12.229      0.771      0.441     -14.545      33.391
uterine_contractions                                    -218.2361     69.086     -3.159      0.002    -353.642     -82.830
light_decelerations                                      161.5035    140.860      1.147      0.252    -114.577     437.584
abnormal_short_term_variability                            0.0928      0.019      4.966      0.000       0.056       0.129
mean_value_of_short_term_variability                      -0.3772      0.487     -0.775      0.438      -1.331       0.577
percentage_of_time_with_abnormal_long_term_variability     0.0116      0.010      1.101      0.271      -0.009       0.032
mean_value_of_long_term_variability                       -0.0944      0.059     -1.598      0.110      -0.210       0.021
histogram_number_of_peaks                                  0.1297      0.068      1.917      0.055      -0.003       0.262
histogram_number_of_zeroes                                 0.0699      0.225      0.311      0.756      -0.371       0.511
histogram_mode                                            -0.0655      0.050     -1.297      0.195      -0.164       0.033
histogram_mean                                             0.4398      0.104      4.222      0.000       0.236       0.644
histogram_median                                          -0.1312      0.123     -1.066      0.286      -0.372       0.110
histogram_variance                                         0.0419      0.014      2.898      0.004       0.014       0.070
prolongued_decelerations_0.001                             3.8953      1.172      3.325      0.001       1.599       6.191
prolongued_decelerations_0.002                             6.6664      1.426      4.675      0.000       3.872       9.461
prolongued_decelerations_0.003                             8.1315      1.845      4.407      0.000       4.515      11.748
prolongued_decelerations_0.004                             2.9745   4.45e+04   6.69e-05      1.000   -8.71e+04    8.71e+04
prolongued_decelerations_0.005                            -2.7705   2.19e+06  -1.26e-06      1.000    -4.3e+06     4.3e+06
histogram_tendency_0                                      -0.4226      0.785     -0.538      0.591      -1.962       1.117
histogram_tendency_1                                      -0.7909      0.828     -0.955      0.339      -2.413       0.832
-------------------------------------------------------------------------------------------------------------------------
                                        fetal_health=3       coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------------------
const                                                    -21.1611      4.686     -4.516      0.000     -30.345     -11.977
baseline.value                                            0.1548      0.082      1.878      0.060      -0.007       0.316
accelerations                                          -899.4500    446.142     -2.016      0.044   -1773.872     -25.028
fetal_movement                                            18.3515     13.128      1.398      0.162      -7.379      44.082
uterine_contractions                                    -190.0654    104.064     -1.826      0.068    -394.027      13.896
light_decelerations                                      362.6239    148.889      2.436      0.015      70.807     654.441
abnormal_short_term_variability                            0.1940      0.029      6.637      0.000       0.137       0.251
mean_value_of_short_term_variability                      -0.8033      0.596     -1.348      0.178      -1.971       0.364
percentage_of_time_with_abnormal_long_term_variability     0.0729      0.016      4.666      0.000       0.042       0.104
mean_value_of_long_term_variability                        0.1166      0.093      1.249      0.212      -0.066       0.300
histogram_number_of_peaks                                 -0.3559      0.141     -2.533      0.011      -0.631      -0.080
histogram_number_of_zeroes                                 0.2399      0.374      0.641      0.521      -0.494       0.973
histogram_mode                                            -0.0768      0.052     -1.475      0.140      -0.179       0.025
histogram_mean                                             0.0246      0.071      0.345      0.730      -0.115       0.164
histogram_median                                          -0.0498      0.096     -0.519      0.604      -0.238       0.138
histogram_variance                                         0.0626      0.016      3.865      0.000       0.031       0.094
prolongued_decelerations_0.001                             3.2111      1.427      2.250      0.024       0.413       6.009
prolongued_decelerations_0.002                             8.5029      1.735      4.901      0.000       5.103      11.903
prolongued_decelerations_0.003                             9.7520      2.273      4.290      0.000       5.296      14.208
prolongued_decelerations_0.004                            20.0082   3.14e+04      0.001      0.999   -6.16e+04    6.16e+04
prolongued_decelerations_0.005                            20.9666   5.74e+04      0.000      1.000   -1.12e+05    1.12e+05
histogram_tendency_0                                      -1.3529      0.948     -1.428      0.153      -3.210       0.504
histogram_tendency_1                                      -1.6738      0.985     -1.700      0.089      -3.604       0.256
=========================================================================================================================
```

Π.4 Output μοντέλου πολυωνυμικής παλινδρόμησης χωρίς την **Severe decelerations.**

```
Warning: Maximum number of iterations has been exceeded.
         Current function value: 0.322028
         Iterations: 35
                           MNLogit Regression Results
==============================================================================
Dep. Variable:           fetal_health   No. Observations:                  659
Model:                        MNLogit   Df Residuals:                      615
Method:                          MLE   Df Model:                           42
Date:               Mon, 10 Apr 2023   Pseudo R-squ.:                 0.6866
Time:                       21:26:20   Log-Likelihood:                -212.22
converged:                     False   LL-Null:                       -677.11
Covariance Type:           nonrobust   LLR p-value:                1.199e-167
=========================================================================================================================
                                        fetal_health=2       coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------------------
const                                                    -17.0392      3.042     -5.601      0.000     -23.001     -11.077
baseline.value                                            -0.1446      0.059     -2.433      0.015      -0.261      -0.028
accelerations                                          -1038.6531    164.702     -6.306      0.000   -1361.463    -715.843
fetal_movement                                             8.5494     12.053      0.709      0.478     -15.074      32.173
uterine_contractions                                    -218.6776     69.138     -3.163      0.002    -354.186     -83.169
light_decelerations                                      167.2689    138.910      1.204      0.229    -104.990     439.528
abnormal_short_term_variability                            0.0930      0.019      5.027      0.000       0.057       0.129
mean_value_of_short_term_variability                      -0.3615      0.481     -0.751      0.452      -1.304       0.581
percentage_of_time_with_abnormal_long_term_variability     0.0118      0.010      1.133      0.257      -0.009       0.032
mean_value_of_long_term_variability                       -0.0897      0.056     -1.601      0.109      -0.200       0.020
histogram_number_of_peaks                                  0.1345      0.067      2.015      0.044       0.004       0.265
histogram_mode                                            -0.0659      0.050     -1.309      0.191      -0.165       0.033
histogram_mean                                             0.4340      0.103      4.209      0.000       0.232       0.636
histogram_median                                          -0.1253      0.121     -1.034      0.301      -0.363       0.112
histogram_variance                                         0.0416      0.014      2.904      0.004       0.014       0.070
prolongued_decelerations_0.001                             3.8976      1.161      3.357      0.001       1.622       6.173
prolongued_decelerations_0.002                             6.6977      1.405      4.769      0.000       3.945       9.451
prolongued_decelerations_0.003                             8.0635      1.839      4.385      0.000       4.459      11.668
prolongued_decelerations_0.004                            -1.2270   2.8e+06  -4.38e-07      1.000   -5.49e+06   5.49e+06
prolongued_decelerations_0.005                            -2.4482  1.66e+05  -1.47e-05      1.000   -3.26e+05   3.26e+05
histogram_tendency_0                                      -0.4285      0.786     -0.545      0.586      -1.969       1.112
histogram_tendency_1                                      -0.7982      0.827     -0.965      0.335      -2.420       0.823
-------------------------------------------------------------------------------------------------------------------------
                                        fetal_health=3       coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------------------
const                                                    -21.5846      4.658     -4.634      0.000     -30.714     -12.455
baseline.value                                            0.1585      0.083      1.910      0.056      -0.004       0.321
accelerations                                          -930.1510    442.811     -2.101      0.036   -1798.045     -62.257
fetal_movement                                            16.8085     12.939      1.299      0.194      -8.552      42.169
uterine_contractions                                    -182.6130    104.359     -1.750      0.080    -387.154      21.928
light_decelerations                                      380.0059    147.266      2.580      0.010      91.370     668.642
abnormal_short_term_variability                            0.1926      0.029      6.605      0.000       0.135       0.250
mean_value_of_short_term_variability                      -0.7952      0.597     -1.333      0.183      -1.964       0.374
percentage_of_time_with_abnormal_long_term_variability     0.0736      0.016      4.712      0.000       0.043       0.104
mean_value_of_long_term_variability                        0.1322      0.090      1.469      0.142      -0.044       0.308
histogram_number_of_peaks                                 -0.3349      0.135     -2.475      0.013      -0.600      -0.070
histogram_mode                                            -0.0824      0.051     -1.599      0.110      -0.183       0.019
histogram_mean                                             0.0202      0.072      0.283      0.777      -0.120       0.161
histogram_median                                          -0.0412      0.096     -0.429      0.668      -0.229       0.147
histogram_variance                                         0.0613      0.016      3.837      0.000       0.030       0.093
prolongued_decelerations_0.001                             3.1977      1.448      2.208      0.027       0.359       6.036
prolongued_decelerations_0.002                             8.6181      1.750      4.924      0.000       5.188      12.049
prolongued_decelerations_0.003                             9.6783      2.272      4.260      0.000       5.226      14.131
prolongued_decelerations_0.004                            24.9773   3.29e+05   7.59e-05      1.000   -6.45e+05   6.45e+05
prolongued_decelerations_0.005                            16.3795   4921.514      0.003      0.997   -9629.611   9662.370
histogram_tendency_0                                      -1.3479      0.949     -1.421      0.155      -3.208       0.512
histogram_tendency_1                                      -1.6629      0.984     -1.690      0.091      -3.591       0.266
=========================================================================================================================
```

Π.5 Output μοντέλου πολυωνυμικής παλινδρόμησης χωρίς την **Histogram Number of zeros.**

```
Warning: Maximum number of iterations has been exceeded.
         Current function value: 0.322870
         Iterations: 35
                          MNLogit Regression Results
==============================================================================
Dep. Variable:          fetal_health   No. Observations:                  659
Model:                       MNLogit   Df Residuals:                      617
Method:                         MLE   Df Model:                           40
Date:               Mon, 10 Apr 2023   Pseudo R-squ.:                 0.6858
Time:                       21:26:21   Log-Likelihood:                -212.77
converged:                     False   LL-Null:                       -677.11
Covariance Type:           nonrobust   LLR p-value:                8.764e-169
==========================================================================================================================
                                    fetal_health=2       coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------------------------
const                                                -16.5285      2.977     -5.552      0.000     -22.363     -10.694
baseline.value                                        -0.1635      0.056     -2.896      0.004      -0.274      -0.053
accelerations                                      -1045.7536    163.830     -6.383      0.000   -1366.855    -724.653
fetal_movement                                         6.9947     11.822      0.592      0.554     -16.175      30.165
uterine_contractions                                -231.2568     67.903     -3.406      0.001    -364.344     -98.170
light_decelerations                                  186.9528    136.276      1.372      0.170     -80.143     454.048
abnormal_short_term_variability                        0.0928      0.018      5.040      0.000       0.057       0.129
mean_value_of_short_term_variability                  -0.2741      0.466     -0.588      0.557      -1.188       0.639
percentage_of_time_with_abnormal_long_term_variability 0.0129      0.010      1.254      0.210      -0.007       0.033
mean_value_of_long_term_variability                   -0.0829      0.055     -1.501      0.133      -0.191       0.025
histogram_number_of_peaks                              0.1179      0.064      1.834      0.067      -0.008       0.244
histogram_mode                                        -0.1007      0.039     -2.595      0.009      -0.177      -0.025
histogram_mean                                         0.3579      0.069      5.220      0.000       0.223       0.492
histogram_variance                                     0.0345      0.013      2.720      0.007       0.010       0.059
prolongued_decelerations_0.001                         3.7300      1.167      3.196      0.001       1.443       6.017
prolongued_decelerations_0.002                         6.9554      1.386      5.020      0.000       4.240       9.671
prolongued_decelerations_0.003                         8.1335      1.850      4.396      0.000       4.507      11.760
prolongued_decelerations_0.004                         2.3986   2.27e+04      0.000      1.000   -4.45e+04    4.45e+04
prolongued_decelerations_0.005                        -2.0415   1.63e+06  -1.25e-06      1.000   -3.2e+06     3.2e+06
histogram_tendency_0                                  -0.5429      0.793     -0.685      0.493      -2.096       1.011
histogram_tendency_1                                  -0.9653      0.826     -1.169      0.242      -2.584       0.653
--------------------------------------------------------------------------------------------------------------------------
                                    fetal_health=3       coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------------------------
const                                                -20.9811      4.502     -4.660      0.000     -29.805     -12.157
baseline.value                                         0.1384      0.072      1.912      0.056      -0.003       0.280
accelerations                                      -1010.5932    383.479     -2.635      0.008   -1762.198    -258.989
fetal_movement                                        15.7645     12.724      1.239      0.215      -9.175      40.704
uterine_contractions                                -176.2327    104.391     -1.688      0.091    -380.835      28.369
light_decelerations                                  386.2517    142.858      2.704      0.007     106.256     666.248
abnormal_short_term_variability                        0.1932      0.029      6.580      0.000       0.136       0.251
mean_value_of_short_term_variability                  -0.8178      0.589     -1.388      0.165      -1.973       0.337
percentage_of_time_with_abnormal_long_term_variability 0.0739      0.016      4.739      0.000       0.043       0.104
mean_value_of_long_term_variability                    0.1336      0.090      1.481      0.139      -0.043       0.310
histogram_number_of_peaks                             -0.3408      0.132     -2.588      0.010      -0.599      -0.083
histogram_mode                                        -0.0932      0.041     -2.265      0.024      -0.174      -0.013
histogram_mean                                         0.0052      0.065      0.080      0.936      -0.121       0.132
histogram_variance                                     0.0612      0.016      3.772      0.000       0.029       0.093
prolongued_decelerations_0.001                         3.2095      1.423      2.256      0.024       0.421       5.998
prolongued_decelerations_0.002                         8.8278      1.722      5.127      0.000       5.453      12.202
prolongued_decelerations_0.003                         9.7385      2.266      4.297      0.000       5.297      14.180
prolongued_decelerations_0.004                        18.3876   8102.409      0.002      0.998   -1.59e+04    1.59e+04
prolongued_decelerations_0.005                        21.3983    4.3e+04      0.000      1.000   -8.43e+04    8.44e+04
histogram_tendency_0                                  -1.4165      0.934     -1.517      0.129      -3.247       0.414
histogram_tendency_1                                  -1.7519      0.972     -1.802      0.072      -3.658       0.154
==========================================================================================================================
```

Π.6 Output μοντέλου πολυωνυμικής παλινδρόμησης χωρίς την **Histogram median.**

```
Warning: Maximum number of iterations has been exceeded.
         Current function value: 0.326092
         Iterations: 35
                          MNLogit Regression Results
==============================================================================
Dep. Variable:           fetal_health   No. Observations:                  659
Model:                        MNLogit   Df Residuals:                      621
Method:                           MLE   Df Model:                           36
Date:                Mon, 10 Apr 2023   Pseudo R-squ.:                 0.6826
Time:                        21:26:21   Log-Likelihood:                -214.89
converged:                      False   LL-Null:                       -677.11
Covariance Type:            nonrobust   LLR p-value:                1.070e-170
=========================================================================================================================
                                     fetal_health=2       coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------------------
const                                                -16.0959      2.904     -5.542      0.000     -21.789     -10.403
baseline.value                                        -0.1751      0.056     -3.108      0.002      -0.286      -0.065
accelerations                                      -1030.1524    157.620     -6.536      0.000   -1339.083    -721.222
fetal_movement                                         7.9377     11.915      0.666      0.505     -15.416      31.291
uterine_contractions                                -230.2318     66.764     -3.448      0.001    -361.087     -99.376
light_decelerations                                  171.5281    133.951      1.281      0.200     -91.011     434.068
abnormal_short_term_variability                        0.0936      0.018      5.167      0.000       0.058       0.129
mean_value_of_short_term_variability                  -0.1835      0.454     -0.404      0.686      -1.074       0.707
percentage_of_time_with_abnormal_long_term_variability     0.0129      0.010      1.252      0.211      -0.007       0.033
mean_value_of_long_term_variability                   -0.0958      0.054     -1.767      0.077      -0.202       0.010
histogram_number_of_peaks                              0.1111      0.064      1.743      0.081      -0.014       0.236
histogram_mode                                        -0.1119      0.039     -2.888      0.004      -0.188      -0.036
histogram_mean                                         0.3727      0.068      5.486      0.000       0.240       0.506
histogram_variance                                     0.0351      0.013      2.742      0.006       0.010       0.060
prolongued_decelerations_0.001                         3.6221      1.153      3.141      0.002       1.362       5.882
prolongued_decelerations_0.002                         6.9757      1.375      5.073      0.000       4.281       9.671
prolongued_decelerations_0.003                         7.9921      1.826      4.377      0.000       4.413      11.571
prolongued_decelerations_0.004                         2.2113   8.71e+04   2.54e-05      1.000   -1.71e+05    1.71e+05
prolongued_decelerations_0.005                        -1.3286   7.21e+05  -1.84e-06      1.000   -1.41e+06    1.41e+06
-------------------------------------------------------------------------------------------------------------------------
                                     fetal_health=3       coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------------------
const                                                -20.7813      4.471     -4.648      0.000     -29.544     -12.019
baseline.value                                         0.1329      0.072      1.840      0.066      -0.009       0.274
accelerations                                       -881.8889    361.819     -2.437      0.015   -1591.042    -172.736
fetal_movement                                        16.1521     12.807      1.261      0.207      -8.949      41.253
uterine_contractions                                -174.9445    102.726     -1.703      0.089    -376.284      26.395
light_decelerations                                  354.2232    140.568      2.520      0.012      78.715     629.731
abnormal_short_term_variability                        0.1934      0.029      6.656      0.000       0.136       0.250
mean_value_of_short_term_variability                  -0.6567      0.560     -1.173      0.241      -1.754       0.441
percentage_of_time_with_abnormal_long_term_variability     0.0704      0.015      4.644      0.000       0.041       0.100
mean_value_of_long_term_variability                    0.1095      0.087      1.264      0.206      -0.060       0.279
histogram_number_of_peaks                             -0.3076      0.126     -2.440      0.015      -0.555      -0.061
histogram_mode                                        -0.1122      0.040     -2.786      0.005      -0.191      -0.033
histogram_mean                                         0.0195      0.063      0.307      0.759      -0.105       0.144
histogram_variance                                     0.0577      0.016      3.603      0.000       0.026       0.089
prolongued_decelerations_0.001                         2.8523      1.358      2.101      0.036       0.191       5.513
prolongued_decelerations_0.002                         8.5576      1.672      5.120      0.000       5.281      11.834
prolongued_decelerations_0.003                         9.2173      2.204      4.183      0.000       4.898      13.536
prolongued_decelerations_0.004                        20.8516   2.31e+04      0.001      0.999   -4.53e+04    4.54e+04
prolongued_decelerations_0.005                        20.2654   2.26e+04      0.001      0.999   -4.43e+04    4.44e+04
=========================================================================================================================
```

Π.7 Output μοντέλου πολυωνυμικής παλινδρόμησης χωρίς την **Histogram Tendency.**

```
Optimization terminated successfully.
         Current function value: 0.366046
         Iterations 12
                          MNLogit Regression Results
==============================================================================
Dep. Variable:          fetal_health   No. Observations:                  659
Model:                       MNLogit   Df Residuals:                      631
Method:                         MLE   Df Model:                           26
Date:              Mon, 10 Apr 2023   Pseudo R-squ.:                  0.6437
Time:                      21:26:21   Log-Likelihood:                -241.22
converged:                     True   LL-Null:                       -677.11
Covariance Type:           nonrobust   LLR p-value:                5.024e-167
=========================================================================================================================
                                    fetal_health=2       coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------------------
const                                                -11.7795      2.521     -4.672      0.000     -16.722      -6.838
baseline.value                                        -0.0545      0.046     -1.181      0.237      -0.145       0.036
accelerations                                       -970.0850    143.920     -6.740      0.000   -1252.163    -688.007
fetal_movement                                        8.3806      7.452      1.125      0.261      -6.225      22.986
uterine_contractions                                -219.7730     59.981     -3.664      0.000    -337.333    -102.213
light_decelerations                                 -113.5708    115.996     -0.979      0.328    -340.919     113.778
abnormal_short_term_variability                       0.0583      0.015      3.945      0.000       0.029       0.087
mean_value_of_short_term_variability                 -0.2136      0.407     -0.525      0.600      -1.011       0.584
percentage_of_time_with_abnormal_long_term_variability    0.0034      0.009      0.362      0.717      -0.015       0.022
mean_value_of_long_term_variability                  -0.1679      0.050     -3.347      0.001      -0.266      -0.070
histogram_number_of_peaks                             0.1246      0.059      2.097      0.036       0.008       0.241
histogram_mode                                       -0.0902      0.037     -2.465      0.014      -0.162      -0.018
histogram_mean                                        0.2228      0.053      4.198      0.000       0.119       0.327
histogram_variance                                    0.0551      0.012      4.664      0.000       0.032       0.078
-------------------------------------------------------------------------------------------------------------------------
                                    fetal_health=3       coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------------------
const                                                -11.7038      3.998     -2.928      0.003     -19.539      -3.869
baseline.value                                        0.2013      0.055      3.638      0.000       0.093       0.310
accelerations                                      -1210.2305    343.691     -3.521      0.000   -1883.853    -536.608
fetal_movement                                       17.7867      7.399      2.404      0.016       3.285      32.288
uterine_contractions                                -208.2210     92.968     -2.240      0.025    -390.435     -26.007
light_decelerations                                  -28.4307    109.221     -0.260      0.795    -242.501     185.639
abnormal_short_term_variability                       0.1461      0.028      5.153      0.000       0.091       0.202
mean_value_of_short_term_variability                 -0.3534      0.450     -0.786      0.432      -1.235       0.528
percentage_of_time_with_abnormal_long_term_variability    0.0484      0.013      3.759      0.000       0.023       0.074
mean_value_of_long_term_variability                  -0.0238      0.077     -0.311      0.756      -0.174       0.126
histogram_number_of_peaks                            -0.1567      0.111     -1.416      0.157      -0.374       0.060
histogram_mode                                       -0.1043      0.039     -2.684      0.007      -0.181      -0.028
histogram_mean                                       -0.0875      0.048     -1.839      0.066      -0.181       0.006
histogram_variance                                    0.0710      0.014      4.895      0.000       0.043       0.099
=========================================================================================================================
```

Π.8 Output μοντέλου πολυωνυμικής παλινδρόμησης χωρίς την **Prolongues decelerations.**

```
Optimization terminated successfully.
         Current function value: 0.366558
         Iterations 12
                          MNLogit Regression Results
==============================================================================
Dep. Variable:          fetal_health   No. Observations:                  659
Model:                       MNLogit   Df Residuals:                      633
Method:                         MLE   Df Model:                           24
Date:              Mon, 10 Apr 2023   Pseudo R-squ.:                  0.6432
Time:                      21:26:21   Log-Likelihood:                -241.56
converged:                     True   LL-Null:                       -677.11
Covariance Type:           nonrobust   LLR p-value:                1.917e-168
=========================================================================================================================
                                    fetal_health=2       coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------------------
const                                                -12.0321      2.491     -4.831      0.000     -16.914      -7.151
baseline.value                                        -0.0619      0.045     -1.390      0.165      -0.149       0.025
accelerations                                       -976.4158    143.560     -6.801      0.000   -1257.789    -695.043
fetal_movement                                        7.4570      7.292      1.023      0.307      -6.836      21.750
uterine_contractions                                -223.8877     59.066     -3.790      0.000    -339.655    -108.120
light_decelerations                                 -119.1681    108.233     -1.101      0.271    -331.300      92.964
abnormal_short_term_variability                       0.0614      0.014      4.429      0.000       0.034       0.089
percentage_of_time_with_abnormal_long_term_variability    0.0045      0.009      0.484      0.629      -0.014       0.023
mean_value_of_long_term_variability                  -0.1678      0.050     -3.346      0.001      -0.266      -0.070
histogram_number_of_peaks                             0.1175      0.058      2.038      0.042       0.004       0.231
histogram_mode                                       -0.0900      0.036     -2.467      0.014      -0.162      -0.018
histogram_mean                                        0.2295      0.052      4.400      0.000       0.127       0.332
histogram_variance                                    0.0531      0.011      4.734      0.000       0.031       0.075
-------------------------------------------------------------------------------------------------------------------------
                                    fetal_health=3       coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------------------
const                                                -12.2560      3.942     -3.109      0.002     -19.982      -4.530
baseline.value                                        0.1855      0.050      3.691      0.000       0.087       0.284
accelerations                                      -1224.0251    344.726     -3.551      0.000   -1899.676    -548.374
fetal_movement                                       17.1127      7.051      2.427      0.015       3.293      30.932
uterine_contractions                                -210.0834     92.966     -2.260      0.024    -392.294     -27.873
light_decelerations                                  -40.4784    105.347     -0.384      0.701    -246.954     165.997
abnormal_short_term_variability                       0.1537      0.027      5.700      0.000       0.101       0.207
percentage_of_time_with_abnormal_long_term_variability    0.0499      0.013      3.941      0.000       0.025       0.075
mean_value_of_long_term_variability                  -0.0203      0.075     -0.272      0.786      -0.167       0.126
histogram_number_of_peaks                            -0.1666      0.109     -1.527      0.127      -0.380       0.047
histogram_mode                                       -0.1026      0.039     -2.658      0.008      -0.178      -0.027
histogram_mean                                       -0.0748      0.046     -1.635      0.102      -0.164       0.015
histogram_variance                                    0.0688      0.014      4.949      0.000       0.042       0.096
=========================================================================================================================
```

Π.9 Output μοντέλου πολυωνυμικής παλινδρόμησης χωρίς την **Mean value of short term variability.**

```
Optimization terminated successfully.
         Current function value: 0.367524
         Iterations 12
                          MNLogit Regression Results
==============================================================================
Dep. Variable:           fetal_health   No. Observations:                  659
Model:                        MNLogit   Df Residuals:                      635
Method:                           MLE   Df Model:                           22
Date:                Mon, 10 Apr 2023   Pseudo R-squ.:                 0.6423
Time:                        21:26:22   Log-Likelihood:                -242.20
converged:                      True   LL-Null:                       -677.11
Covariance Type:            nonrobust   LLR p-value:                8.997e-170
=================================================================================================================
                                    fetal_health=2       coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------------------
const                                                -12.0473      2.497     -4.824      0.000     -16.942      -7.153
baseline.value                                        -0.0805      0.042     -1.932      0.053      -0.162       0.001
accelerations                                       -982.9584    142.635     -6.891      0.000   -1262.518    -703.399
fetal_movement                                        7.6727      7.056      1.087      0.277      -6.157      21.503
uterine_contractions                                -229.1664     58.659     -3.907      0.000    -344.136    -114.197
abnormal_short_term_variability                       0.0634      0.014      4.629      0.000       0.037       0.090
percentage_of_time_with_abnormal_long_term_variability     0.0063      0.009      0.684      0.494      -0.012       0.024
mean_value_of_long_term_variability                  -0.1577      0.049     -3.207      0.001      -0.254      -0.061
histogram_number_of_peaks                             0.1069      0.057      1.888      0.059      -0.004       0.218
histogram_mode                                       -0.0979      0.035     -2.780      0.005      -0.167      -0.029
histogram_mean                                        0.2541      0.048      5.337      0.000       0.161       0.347
histogram_variance                                    0.0501      0.011      4.679      0.000       0.029       0.071
-----------------------------------------------------------------------------------------------------------------
                                    fetal_health=3       coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------------------
const                                                -12.2848      3.900     -3.150      0.002     -19.930      -4.640
baseline.value                                        0.1826      0.049      3.692      0.000       0.086       0.280
accelerations                                     -1235.3464    344.186     -3.589      0.000   -1909.939    -560.753
fetal_movement                                       16.7828      6.730      2.494      0.013       3.592      29.973
uterine_contractions                                -212.7349     92.304     -2.305      0.021    -393.648     -31.821
abnormal_short_term_variability                       0.1540      0.027      5.716      0.000       0.101       0.207
percentage_of_time_with_abnormal_long_term_variability     0.0513      0.012      4.232      0.000       0.028       0.075
mean_value_of_long_term_variability                  -0.0135      0.072     -0.187      0.852      -0.155       0.128
histogram_number_of_peaks                            -0.1789      0.105     -1.702      0.089      -0.385       0.027
histogram_mode                                       -0.1032      0.038     -2.719      0.007      -0.178      -0.029
histogram_mean                                       -0.0718      0.045     -1.606      0.108      -0.159       0.016
histogram_variance                                    0.0673      0.014      4.955      0.000       0.041       0.094
=================================================================================================================
```

Π.10 Output μοντέλου πολυωνυμικής παλινδρόμησης χωρίς την **Light decelerations.**

```
Optimization terminated successfully.
         Current function value: 0.344944
         Iterations 12
                          MNLogit Regression Results
==============================================================================
Dep. Variable:           fetal_health   No. Observations:                  659
Model:                        MNLogit   Df Residuals:                      635
Method:                           MLE   Df Model:                           22
Date:                Mon, 10 Apr 2023   Pseudo R-squ.:                 0.6643
Time:                        21:26:24   Log-Likelihood:                -227.32
converged:                      True   LL-Null:                       -677.11
Covariance Type:            nonrobust   LLR p-value:                4.340e-176
=================================================================================================================
                                    fetal_health=2       coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------------------
const                                                -15.2926      2.785     -5.491      0.000     -20.751      -9.834
baseline.value                                        -0.1430      0.048     -2.987      0.003      -0.237      -0.049
accelerations                                       -951.6009    142.415     -6.682      0.000   -1230.729    -672.473
uterine_contractions                                -243.1663     61.081     -3.981      0.000    -362.883    -123.450
abnormal_short_term_variability                       0.0903      0.016      5.579      0.000       0.059       0.122
percentage_of_time_with_abnormal_long_term_variability     0.0120      0.010      1.236      0.216      -0.007       0.031
mean_value_of_long_term_variability                  -0.0749      0.045     -1.666      0.096      -0.163       0.013
histogram_mode                                       -0.0810      0.036     -2.227      0.026      -0.152      -0.010
histogram_mean                                        0.3067      0.054      5.637      0.000       0.200       0.413
histogram_variance                                    0.0483      0.012      4.191      0.000       0.026       0.071
prolongued_decelerations_0.001                        3.1086      1.104      2.816      0.005       0.945       5.272
prolongued_decelerations_0.002-0.005                  6.3521      1.147      5.537      0.000       4.104       8.600
-----------------------------------------------------------------------------------------------------------------
                                    fetal_health=3       coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------------------
const                                                -17.7073      4.041     -4.381      0.000     -25.628      -9.786
baseline.value                                        0.1467      0.064      2.280      0.023       0.021       0.273
accelerations                                       -914.8186    336.808     -2.716      0.007   -1574.951    -254.686
uterine_contractions                                -217.3518     96.142     -2.261      0.024    -405.786     -28.918
abnormal_short_term_variability                       0.1868      0.027      6.992      0.000       0.134       0.239
percentage_of_time_with_abnormal_long_term_variability     0.0583      0.013      4.523      0.000       0.033       0.084
mean_value_of_long_term_variability                   0.0578      0.067      0.867      0.386      -0.073       0.188
histogram_mode                                       -0.1417      0.040     -3.549      0.000      -0.220      -0.063
histogram_mean                                        0.0187      0.053      0.349      0.727      -0.086       0.123
histogram_variance                                    0.0500      0.014      3.503      0.000       0.022       0.078
prolongued_decelerations_0.001                        1.6168      1.330      1.215      0.224      -0.991       4.224
prolongued_decelerations_0.002-0.005                  6.9535      1.339      5.194      0.000       4.329       9.578
=================================================================================================================
```

Π.11 Output μοντέλου πολυωνυμικής παλινδρόμησης χωρίς την **Fetal movement.**