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

ΣΧΟΛΗ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗΣ ΚΑΙ ΣΤΑΤΙΣΤΙΚΗΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΣΥΓΚΡΙΣΗ ΤΕΧΝΙΚΩΝ ΔΙΑΧΩΡΙΣΤΙΚΗΣ ΑΝΑΛΥΣΗΣ

Στέφανος Β. Μυλωνίδης

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
Μάιος 2018

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος 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 Κούτρας Μάρκος (Επιβλέπων)
- Αναπληρωτής Καθηγητής Πολίτης Κωνσταντίνος
- Επίκουρος Καθηγητής Ευαγγελάρας Χαράλαμπος

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

U N I V E R S I T Y O F P I R A E U S

SCHOOL OF FINANCE AND STATISTICS

DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

COMPARISON OF DISCRIMINATION TECHNIQUES

By

Stefanos V. Mylonidis

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
May 2018

Σε όσους με στήριξαν ...

Ευχαριστίες

Θα ήθελα να ευχαριστήσω ιδιαιτέρως τον επιβλέποντα Καθηγητή κ. Μάρκο Κούτρα, ο οποίος επέδειξε εξαιρετική υπομονή και επιμονή, κατά τη διάρκεια συγγραφής της παρούσας εργασίας, μεταδίδοντας ανεκτίμητες γνώσεις και πράττοντας σημαντικές υποδείξεις, αλλά το σημαντικότερο, στηρίζοντας συναισθηματικά και ψυχολογικά, ώστε η όλη προσπάθεια να ολοκληρωθεί επιτυχώς, εντός των προβλεπόμενων χρονικών ορίων.

Επιπρόσθετα, θα επιθυμούσα να ευχαριστήσω θερμά τα υπόλοιπα μέλη της Τριμελούς Επιτροπής, τον Αναπληρωτή Καθηγητή κ. Κωνσταντίνο Πολίτη, καθώς και τον Επίκουρο Καθηγητή κ. Χαράλαμπο Ευαγγελάρα, για τον πολύτιμο χρόνο που αφιέρωσαν στη μελέτη και διόρθωση της εργασίας αυτής.

Τέλος, θα ήταν σημαντική παράλειψη να μην ευχαριστήσω, όλους εκείνους τους πολύ κοντινούς μου ανθρώπους, οι οποίοι με στήριξαν, προκειμένου να ολοκληρώσω εργαζόμενος γαρ, ένα τόσο ποιοτικό και συνάμα τόσο απαιτητικό Μεταπτυχιακό Πρόγραμμα, επιστέγασμα του οποίου αποτελεί η παρούσα διπλωματική εργασία.

Περίληψη

Το πεδίο της Διαχωριστικής Ανάλυσης παρουσιάζει ιδιαίτερο ενδιαφέρον, τόσο σε ερευνητικό επίπεδο, όσο και σε επίπεδο πρακτικής εφαρμογής στο χώρο των επιχειρήσεων, σε ένα ευρύ πεδίο δραστηριοτήτων. Στον Χρηματοπιστωτικό τομέα, τα τραπεζικά ιδρύματα ενδιαφέρονται για τον έγκαιρο εντοπισμό των πελατών με υψηλή πιστοληπτική διαβάθμιση, προκειμένου να αποφανθούν για την χορήγηση ή μη πίστωσης, στον τομέα της Ιατρικής Επιστήμης, προκειμένου να εντοπιστούν και προληπτικά αντιμετωπιστούν σοβαρές ασθένειες, με βάση τα επίπεδα τιμών συγκεκριμένων ιατρικών δεικτών και συναφών συμπτωμάτων, στον τομέα των Κοινωνικών Επιστημών και της Εγκληματολογίας, προκειμένου να εντοπιστούν κοινωνικές ομάδες συγκεκριμένου ενδιαφέροντος ή κινδύνου, με κριτήριο μία σειρά από χαρακτηριστικά (δημογραφικά, ιατρικά, οικονομικά κλπ.), καθώς και στον τομέα του Marketing, προκειμένου να αποφασιστεί το κατά πόσον ένα συγκεκριμένο τμήμα της Αγοράς, κρίνεται κατάλληλο για την τοποθέτηση και ανάπτυξη νέων προϊοντικών κατηγοριών ή υπηρεσιών. Βασικός σκοπός της διαχωριστικής διαδικασίας είναι να κατατάξει τις υπό μελέτη πειραματικές μονάδες, σε έναν από πολλούς γνωστούς πληθυσμούς, με βάση τις τιμές επιλεχθέντων παρατηρούμενων χαρακτηριστικών. Το παραπάνω καθίσταται εφικτό, μέσω της διαμόρφωσης κατάλληλου διαχωριστικού κανόνα, βάση του οποίου κατατάσσεται κάθε πειραματική μονάδα, σε έναν από τους διαθέσιμους πληθυσμούς.

Στην παρούσα διπλωματική εργασία, αναπτύσσεται λεπτομερώς το θεωρητικό πλαίσιο της μεθόδου Πολυωνυμικής Λογιστικής Παλινδρόμησης, καθώς και των Διαχωριστικών Αλγορίθμων *ID3, C4_5* και *CART*. Ακολούθως, δίδονται συγκεκριμένα παραδείγματα εφαρμογής εκάστης μεθόδου, προκειμένου να καταστεί ευκολότερη η κατανόηση των σχετικών εννοιών θεμελίωσής τους. Καταληκτικά, πραγματοποιείται εκτενής εφαρμογή της μεθόδου Πολυωνυμικής Λογιστικής Παλινδρόμησης, καθώς και του διαχωριστικού αλγορίθμου *C 4_5*, σε πρωτογενές σύνολο δεδομένων, προκειμένου να διαχωριστούν οι φορολογούμενοι συγκεκριμένης χώρας, ως προς την οικογενειακή τους κατάσταση, πληροφορία η οποία είναι χρήσιμη στις φορολογικές αρχές, για την αποτελεσματικότερη φορολογική διαχείριση των μη μονίμων κατοίκων αλλοδαπής εθνικότητας. Η απόδοση και διαχωριστική ακρίβεια των δύο μεθόδων η οποία επιτεύχθηκε, κρίνεται απολύτως συγκρίσιμη σε σύνολα δεδομένων, τα οποία είναι πλήρη επί των χρησιμοποιούμενων χαρακτηριστικών, δηλαδή χωρίς την εμφάνιση ελλειπουσών τιμών, ενώ στην περίπτωση ενσωμάτωσης χαρακτηριστικών με υψηλό ποσοστό ελλειπουσών τιμών και μετέπειτα διαχείρισης αυτών, ο διαχωριστικός αλγόριθμος *C 4_5*, εμφανίστηκε ανθεκτικότερος, σημειώνοντας καλύτερη ακρίβεια διαχωρισμού, σε σχέση με εκείνον της Πολυωνυμικής Λογιστικής Παλινδρόμησης.

Abstract

The field of Discriminant Analysis is of particular interest, both at the research scientific area, as well as, the direction of practical application in the field of business, in a wide range of activities. The main purpose of the discrimination process is to classify the experimental units under study, to one of many known populations, based on the values of specific observed characteristics. The above is made possible through the formation of an appropriate discrimination rule, based on which each experimental unit is classified, in one of the above mentioned available populations.

Along the specific diploma thesis, the theoretical framework of the Polynomial Logistic Regression method, as well as the ID3, C4_5 and CART Algorithms, is thoroughly presented. Next, specific examples are given of each method, in order to make it easier to understand their concepts of foundation. Finally, extensive use of the Polynomial Logistic Regression method and the C 4_5 algorithm is performed, in a primary set of data, so as to separate taxpayers of a given country, in terms of their marital status, which is particular useful to the tax authorities, for tax administration of non-residents of a foreign nationality. The efficiency and separation accuracy of the two methods achieved, is judged to be entirely comparable to data sets, which are complete on the characteristics under consideration, without the occurrence of missing values, whereas in the case of incorporation of features with a high percentage of missing values and subsequent management of them, the C 4_5 algorithm appeared to be more robust, showing better separation accuracy than that of the Polynomial Logistic Regression.

.

Πίνακας Περιεχομένων

1 Εισαγωγή

1.1 Ορισμός του προβλήματος της Διαχωριστικής Ανάλυσης

Το πρόβλημα της Διαχωριστικής Ανάλυσης (*Discriminant Analysis*) , έγκειται στην εκμάθηση και μετέπειτα αξιολόγηση της αποτελεσματικότητας αλγοριθμικής διαδικασίας, μέσω της οποίας προβλέπεται η κλάση πειραματικής μονάδας, εκμεταλλευόμενοι τη γνώση των τιμών ενός συνόλου χαρακτηριστικών γνωρισμάτων.

Η αρχική ιδέα ξεκίνησε από τον Άγγλο Μαθηματικό Karl Pearson (1857–1936), ο οποίος εισήγαγε ένα δείκτη απόστασης μεταξύ ομάδων (*intergroup distance index*), τον οποίο ονόμασε *CRL: 'Coefficient of Racial Likeness*'. Ο στατιστικός G. M. Morant (1899–1964) το 1920, μελέτησε και εξέλιξε την αρχική ιδέα του Pearson, όπως και ο P. C. Mahalanobis (1893–1972), ο οποίος το 1930, εισήγαγε στην Ινδία έναν αντίστοιχο δείκτη απόστασης μεταξύ ομάδων. Το 1930 ο στατιστικός R. A. Fischer (1890–1962), εισήγαγε το βασικό μοντέλο Διαχωριστικής Ανάλυσης πολυδιάστατων παρατηρήσεων μεταξύ δύο ομάδων, μέσω της κατασκευής ενός γραμμικού διαχωριστικού κανόνα, ο οποίος αξιοποιεί ένα πλήθος αρχικά συσχετισμένων μεταξύ τους ερμηνευτικών μεταβλητών. Η επέκταση του αρχικού μοντέλου του Fischer, σε ένα μοντέλο Διαχωριστικής Ανάλυσης πολλών ομάδων, δόθηκε τελικώς από το στατιστικό C. R. Rao το 1948. Απαραίτητες προϋποθέσεις ισχύος των παραπάνω, είναι η εκ των προτέρων γνώση των κλάσεων ισοδυναμίας του προβλήματος κατηγοριοποίησης, καθώς και η ύπαρξη ενός συνόλου αρχικών δεδομένων εκπαίδευσης, τα οποία αποτελούνται από εγγραφές για τις οποίες γνωρίζουμε εκ των προτέρων την κλάση ισοδυναμίας στην οποία ανήκουν. Με διαφορετικά αλλά ισοδύναμα λόγια και σύμφωνα με μία πιο αυστηρή διατύπωση του προβλήματος κατηγοριοποίησης, η όλη αλγοριθμική διαδικασία επί της ουσίας επιδιώκει το καθορισμό της απεικόνισης f :

$$f : D \to C$$

όπου

$$D = \{t_1, t_2, \dots, t_n\}$$

μία βάση δεδομένων, η οποία αποτελείται από n στο πλήθος, p – διάστατες εγγραφές t_i , $i = 1, \dots, n$ και

$$C = \{C_1, C_2, \dots, C_m\}$$

$C_j, j = 1, \dots, m$ είναι οι m στο πλήθος προκαθορισμένες κλάσεις ισοδυναμίας του προβλήματος.

Πρόκειται για μία στατιστική διαδικασία ανάθεσης, μίας πολυδιάστατης παρατήρησης, σε μία εκ των m προκαθορισμένων ομάδων ή κλάσεων/κατηγοριών ταξινόμησης, με απώτερο σκοπό, την κατασκευή ενός μοντέλου Διαχωριστικής Ανάλυσης, το οποίο βελτιστοποιεί την ακρίβεια διάκρισης και ταξινόμησης των παρατηρήσεων, μεταξύ των m αυστηρά προκαθορισμένων κλάσεων (Wehrens, 2010). Η προαναφερθείσα διαδικασία της εκχώρησης των παρατηρήσεων / πειραματικών αντικειμένων, σε m προκαθορισμένες και ει δυνατόν πιο ομοιογενείς ομάδες, πραγματοποιείται με τη χρήση κατάλληλων στατιστικών τεχνικών.

Οι σημαντικότερες μέθοδοι ταξινόμησης στο παραπάνω πρόβλημα μείζονος πρακτικού και ερευνητικού ενδιαφέροντος, είναι η Γραμμική Διαχωριστική Ανάλυση (*LDA:Linear Discriminant Analysis*), η Πολυωνυμική Λογιστική Παλινδρόμηση *(MLR:Multinomial Logistic Regression*), καθώς και μέθοδοι κατασκευής Δέντρων Απόφασης, μέσω αξιοποίησης σύγχρονων αλγορίθμων βασισμένων στην εντροπία της πληροφορίας (*ID3 , C 4.5 , CART*), ή αλγορίθμων βασισμένων στη χρήση κατάλληλα ορισμένης απόστασης (*KNN*).

Οι βασικοί στόχοι ενός προβλήματος Διαχωριστικής Ανάλυσης είναι οι ακόλουθοι:

- ✓ Να αναγνωρίσει το μικρότερο δυνατό πλήθος, στατιστικά σημαντικών ερμηνευτικών μεταβλητών, οι οποίες επιτυγχάνουν το καλύτερο δυνατό διαχωριστικό αποτέλεσμα.

- ✓ Να εκτιμήσει τους συντελεστές του γραμμικού συνδυασμού, μεταξύ της εξαρτημένης κατηγορικής μεταβλητής απόκρισης και των προαναφερθέντων στατιστικά σημαντικών ερμηνευτικών μεταβλητών, μέσω της αξιοποίησης των δειγματικών παρατηρήσεων που συνιστούν το σύνολο των Δεδομένων Εκπαίδευσης (*Training dataset*).

- ✓ Να αξιολογήσει τη διαχωριστική ισχύ του προκύπτοντος μοντέλου, μέσω της ταξινόμησης των παρατηρήσεων του συνόλου των Δεδομένων Ελέγχου (*Test dataset*), σε μία εκ των προκαθορισμένων ομάδων ταξινόμησης.

- ✓ Να αξιολογήσει κατά πόσον, οι *m* προκαθορισμένες ομάδες του προβλήματος ταξινόμησης εμφανίζουν μεταξύ των, στατιστικώς σημαντικές διαφορές.

- ✓ Η εκ των υστέρων ταξινόμηση, της οιασδήποτε μετέπειτα συλλεχθείσας δειγματικής παρατηρήσεως, σε μία εκ των *m* προκαθορισμένων ομάδων του προβλήματος.

Στο σχετικό του βιβλίο ο Ogum (2002), διατείνεται πως το πρόβλημα της Εποπτευόμενης Διαχωριστικής Ανάλυσης (*Supervised Classification*), συνίσταται στην αξιόπιστη ταξινόμηση πειραματικών μονάδων, σε *m* προκαθορισμένες κλάσεις. Ως εκ τούτου, ένα διαχωριστικό μοντέλο φέρει τόσο περιγραφική, όσο και προβλεπτική χρησιμότητα.

Σύμφωνα δε, με τη μελέτη του Prempeh (2009), το πρόβλημα της Διαχωριστικής Ανάλυσης δύναται να λάβει τρεις διαφορετικές μορφές:

- ✓ Ταξινόμηση *(Classification / Discrimination)*
  Με δεδομένη την ύπαρξη *m* διακεκριμένων πληθυσμών, λαμβάνεται δείγμα από έκαστο πληθυσμό, με απώτερο σκοπό την κατασκευή κανόνα ταξινόμησης, ο οποίος θα μας επιτρέψει σε επόμενο χρόνο, να ταξινομήσουμε σε έναν από αυτούς, μια νέα πειραματική μονάδα άγνωστης πληθυσμιακής προελεύσεως.

- ✓ Διάκριση *(Desertion)*
  Πρόκειται κατ' ουσία για ένα πρόβλημα ταξινόμησης, στο οποίο όμως τα όρια διάκρισης μεταξύ των κλάσεων, ενδέχεται να μην είναι φυσικώς υπαρκτά. Το εν λόγω χαρακτηριστικό της μεθόδου, καθιστά την όλη προσέγγιση ικανή να εφαρμοστεί και εντός ομοιογενών κλάσεων, δηλαδή ακόμη και επί εκείνων των κλάσεων που έχουν ήδη προκύψει, κατόπιν της κλασσικής εφαρμογής της μεθόδου Διαχωριστικής Ανάλυσης.

Σε αντίστοιχη μελέτη του ο Onyeagu (2003), υποστήριξε πως το πρόβλημα της Διαχωριστικής Ανάλυσης, έγκειται στη διαδικασία ταξινόμησης τυχούσας πειραματικής μονάδας άγνωστης προελεύσεως, σε μία από *m* το πλήθος προκαθορισμένες και ξεκάθαρα διακεκριμένες κλάσεις, για την οποία πειραματική μονάδα έχουν ληφθεί μετρήσεις επί συγκεκριμένων ερμηνευτικών χαρακτηριστικών.

Σε ανάλογο πνεύμα κινήθηκαν και οι μελέτες των Lachenbruch (1975) και Cooley & Lohnes (1962), οι οποίοι υποστήριξαν πως η ενέργεια ταξινομήσεως αγνώστου προελεύσεως πειραματικής μονάδας, σε μία από *m* προκαθορισμένες και ξεκάθαρα διακεκριμένες κλάσεις, πρέπει να ελαχιστοποιεί συγκεκριμένο κριτήριο εσφαλμένης ταξινόμησης. Η συνάρτηση ή οι συναρτήσεις ταξινόμησης που θα χρησιμοποιηθούν για το σκοπό αυτό, έχουν κατασκευαστεί με ανάλογο τρόπο και ακολουθούν ανάλογες υποθέσεις εφαρμογής, με εκείνες που ισχύουν στην κλασσική μεθοδολογία της Πολυμεταβλητής Ανάλυσης της Διακύμανσης (*MANOVA: Multivariate Analysis of Variance*).

Η μελέτη του Anderson (2003), προσεγγίζει το πρόβλημα της Διαχωριστικής Ανάλυσης, υπό του πρίσματος των κατανομών πιθανότητας (*Statistical Decision Functions*), των αντιστοίχων κλάσεων. Πιο συγκεκριμένα, διατείνεται πως τυχούσα πειραματική μονάδα θα πρέπει να καταταχθεί σε εκείνη τη κλάση, στην οποία μεγιστοποιείται η αντίστοιχη συνάρτηση πιθανοφάνειας.

1.2 Η κλασσική προσέγγιση της Γραμμικής Διαχωριστικής Ανάλυσης

Η Μέθοδος της Γραμμικής Διαχωριστικής Ανάλυσης, είναι μια στατιστική τεχνική ταξινόμησης ενός συνόλου πολυδιάστατων δειγματικών παρατηρήσεων, σε *m* το πλήθος προκαθορισμένες κλάσεις. Επιπλέον, χρησιμοποιείται για να προσδιορίσει εκείνες τις ερμηνευτικές μεταβλητές, οι οποίες επιτυγχάνουν την καλύτερη δυνατή διάκριση των παρατηρήσεων, μεταξύ δύο ή και περισσοτέρων κλάσεων ή συναφών διακριτικών ομάδων.

Οι βασικές υποθέσεις οι οποίες πρέπει να πληρούνται, προκειμένου να καθίσταται δυνατή η εφαρμογή της μεθόδου της Γραμμικής Διαχωριστικής Ανάλυσης, έχουν ως ακολούθως:

α. Μέγεθος Δείγματος

Διαφορετικά μεγέθη δειγμάτων είναι γενικώς αποδεκτά, αλλά ο συνήθης εμπειρικός κανόνας είναι πως, το μικρότερο δυνατό μέγεθος δείγματος πρέπει να είναι 4 με 5 φορές μεγαλύτερο, του πλήθους των ερμηνευτικών μεταβλητών του μοντέλου (Pai, 2009).

β. Κανονικότητα Κατανομής Δεδομένων

Προϋποτίθεται πως τα πειραματικά δεδομένα ακολουθούν πολυδιάστατη Κανονική Κατανομή. Η εν λόγω υπόθεση αποτελεί και τον κυριότερο περιοριστικό παράγοντα ασφαλούς εφαρμογής της μεθόδου. Παρόλα αυτά, πιθανή παραβίαση της υποθέσεως της Κανονικότητας, δεν είναι εκ των προτέρων επιζήμια για την ακρίβεια και την αξιοπιστία του μοντέλου Διαχωριστικής Ανάλυσης. Το παραπάνω συμπέρασμα ισχύει υπό την προϋπόθεση πως, η απώλεια της Κανονικότητας των δεδομένων, είναι αποτέλεσμα της λοξότητας (*skewness)* της πραγματικής κατανομής των πειραματικών δεδομένων και όχι εξαιτίας της ύπαρξης έκτροπων παρατηρήσεων (*outliers*), (Tabachnick & Fidell, 1996, Wang, 2008, Huberty & Olejnik, 2006).

Στην ειδική περίπτωση κατά την οποία, οι ερμηνευτικές μεταβλητές του μοντέλου είναι διχοτομικές κατηγορικές μεταβλητές, η παραβίαση της Κανονικότητας των πειραματικών δεδομένων, δεν έχει καμία επίπτωση στην ακρίβεια των αποτελεσμάτων του εφαρμοζόμενου Μοντέλου Διαχωριστικής Ανάλυσης (Klecka, 1980).

γ. Ομοιογένεια Πινάκων Διακυμάνσεων – Συνδιακυμάνσεων

Η ακρίβεια των αποτελεσμάτων ενός μοντέλου Γραμμικής Διαχωριστικής Ανάλυσης, είναι ιδιαιτέρως ευαίσθητη στην παραβίαση της υποθέσεως περί ετερογένειας των αντιστοίχων Πινάκων Διακυμάνσεων – Συνδιακυμάνσεων, των κλάσεων ισοδυναμίας του προβλήματος. Η ετερογένεια των Πινάκων Διακυμάνσεων – Συνδιακυμάνσεων των κλάσεων του προβλήματος, είναι αναγκαίο και απολύτως επιβεβλημένο να ελεγχθεί και να εξασφαλιστεί, προτού αποδεχθούμε ως αξιόπιστα τα όποια αποτελέσματα προκύπτουν, από την εφαρμογή του Διαχωριστικού μοντέλου. Πιθανή διάγνωση ομοιογένειας των πινάκων αυτών, πρέπει άμεσα να διορθωθεί, με κατάλληλη επέμβαση στη δομή του μοντέλου (Barfield et al., 2004). Ο συνήθης έλεγχος ομοιογένειας των Πινάκων Διακυμάνσεων – Συνδιακυμάνσεων, πραγματοποιείται μέσω του *"Box's M" Test*, όπου η σχετική του ελέγχου μηδενική υπόθεση διατείνεται πως, οι πίνακες Διακυμάνσεων – Συνδιακυμάνσεων των αντιστοίχων ομάδων, είναι μεταξύ των ομοιογενείς. Ισοδύναμα, ο παραπάνω έλεγχος μπορεί να πραγματοποιηθεί, μέσω της σύγκρισης του λογαρίθμου των οριζουσών των εν λόγω Πινάκων (*log determinants*), όπου η ουσιαστική διαφοροποίηση του εν λόγω στατιστικού, υποδηλοί ανομοιογένεια των αντιστοίχων Πινάκων Διακυμάνσεων – Συνδιακυμάνσεων, των ομάδων του Διαχωριστικού Μοντέλου (Geoffry, 1992).

δ. Έκτροπες Παρατηρήσεις (*Outliers*)

Τα μοντέλα Γραμμικής Διαχωριστικής Ανάλυσης, είναι εξαιρετικά ευαίσθητα στην ενσωμάτωση έκτροπων ή ακραίων παρατηρήσεων. Μία παρατήρηση νοείται ως ακραία, στη περίπτωση κατά την οποία απέχει εξαιρετικά μεγάλη απόσταση από το κέντρο βάρους (*centroid*) της αντιστοίχου ομάδας, σε σχέση με τις υπόλοιπες παρατηρήσεις οι οποίες συνιστούν την ομάδα αυτή. Η πιθανή διατήρηση των παρατηρήσεων αυτών στην ανάλυση, θα επηρεάσει τόσο το μέσο, όσο και τη διακύμανση της αντιστοίχου ομάδας, με άμεση συνέπεια τον επηρεασμό της αξιοπιστίας του διαχωριστικού μοντέλου. Επομένως, πιθανή ύπαρξη τέτοιων έκτροπων παρατηρήσεων πρέπει να αντιμετωπίζεται, είτε με την άμεση εξαίρεσή τους, είτε με κατάλληλο μετασχηματισμό (Khattree & Naik ,1995, Riemann et. Al, 2008).

ε. Πολυσυγγραμικότητα Ερμηνευτικών Μεταβλητών (*Multi - Collinearity*)

Το πρόβλημα της Πολυσυγγραμικότητας μεταξύ των ερμηνευτικών μεταβλητών, είναι κρίσιμο για την αξιοπιστία του προκύπτοντος μοντέλου Γραμμικής Διαχωριστικής Ανάλυσης. Οι ερμηνευτικές μεταβλητές του μοντέλου, ιδεατά πρέπει να είναι μεταξύ των καθολικά ασυσχέτιστες. Σε περίπτωση κατά την οποία, ο δειγματικός συντελεστής συσχέτισης μεταξύ δύο οιοδήποτε ερμηνευτικών μεταβλητών υπερβαίνει το 0.80, τότε επιβάλλεται η εξαίρεση εκ του μοντέλου, κάποιων εκ των υψηλά συσχετισμένων μεταβλητών. Κατά αυτόν τον τρόπο καθίσταται ακριβής η εκτίμηση των αντιστοίχων συντελεστών του μοντέλου, περιορίζοντας το μέγεθος των τυπικών σφαλμάτων εκτίμησης.

Στο σημείο αυτό να σημειώσουμε πως, μας ενδιαφέρει η εξασφάλιση της μη - Πολυσυγγραμικότητας των ερμηνευτικών μεταβλητών, στο εσωτερικό των ομάδων του προβλήματος Διαχωριστικής Ανάλυσης (*within groups Multi Collinearity*), παρά η αντίστοιχη επί του συνόλου των πειραματικών δεδομένων (*pooled Multi Collinearity*) (Brace et al. , 2009, Poulsen & French, 2004).

Η μέθοδος της Γραμμικής Διαχωριστικής Ανάλυσης, διακρίνεται σε δύο βασικές κατηγορίες, (Simar and Hardle, 2000), αναλόγως του πλήθους των ομάδων κατηγοριοποίησης:

Α. Γραμμική Διαχωριστική Ανάλυση Δύο Ομάδων Ταξινόμησης (*LDA: Linear Discriminant Analysis*)

Πρόκειται για την απλούστερη μορφή μοντέλου Διαχωριστικής Ανάλυσης, στην οποία η κατηγορική εξαρτημένη μεταβλητή, λαμβάνει τιμές μεταξύ δύο συγκεκριμένων και ανεξάρτητων επιπέδων, τα οποία αντιστοιχούν στις κλάσεις, ή ομάδες ταξινομήσεως του Διαχωριστικού Μοντέλου.

Σε αυτή τη περίπτωση, η Γραμμική Διαχωριστική Συνάρτηση διέρχεται από τα κέντρα βάρους των ομάδων αυτών, επιτρέποντας εν συνεχεία την κατάταξη των πολυδιάστατων δειγματικών παρατηρήσεων, εντός των δύο εναλλακτικών κατηγοριών της μεταβλητής απόκρισης (Antonogeorgos et al., 2009).

Η Γραμμική Διαχωριστική Συνάρτηση (*LDF: Linear Discriminant Function)*, προκύπτει από την εκτίμηση των αντιστοίχων συντελεστών – σταθμών βαρύτητας, των στατιστικά σημαντικών ερμηνευτικών μεταβλητών, με καταληκτικό στόχο, την ελαχιστοποίηση της μεταβλητότητας εντός των ομάδων (*within clusters variability*), μεγιστοποιώντας συνάμα τη μεταβλητότητα μεταξύ των ομάδων (*between clusters variability*), (Albayrak, A. S. 2009, Rencher, 2002).

Η γενική μορφή της Γραμμικής Διαχωριστικής Συναρτήσεως, είναι η ακόλουθη:

$$(LDF):\ D = a + \beta_1 X_1 + \beta_2 X_2 + \ \ldots + \beta_p X_p$$

όπου

a, η σταθερά του διαχωριστικού μοντέλου

X_i , i= 1, … , p , το score της i στατιστικά σημαντικής ερμηνευτικής μεταβλητής

β_i , i= 1, … , p , ο αντίστοιχος συντελεστής βαρύτητας της i ερμηνευτικής μεταβλητής.

Η εφαρμογή της παραπάνω Γραμμικής Διαχωριστικής Συναρτήσεως, επί των πειραματικών Δεδομένων Εκπαίδευσης, θα οδηγήσει στην εκτίμηση των συντελεστών β_i του μοντέλου, κατά τρόπον ώστε, να ελαχιστοποιηθεί η μεταβλητότητα στο εσωτερικό εκάστης ομάδας, ενώ συνάμα μεγιστοποιείται η μεταβλητότητα μεταξύ των ομάδων της Ανάλυσης.

Σε δεύτερο χρόνο, κάθε επόμενη παρατήρηση του συνόλου των Δεδομένων Αξιολόγησης (*test data set*), ταξινομείται σε μία από τις δύο εναλλακτικές κλάσεις ταξινομήσεως, αναλόγως της τιμής που λαμβάνει επί της εκτιμημένης Γραμμικής Διαχωριστικής Συναρτήσεως, ώστε να αποτυπωθεί η διαχωριστική αποτελεσματικότητα του μοντέλου, καθώς και το φαινομενικό σφάλμα ταξινόμησης (*APER:Apparent Error Rate/*), (Press & Wilson , 1978). Βασική προϋπόθεση ισχύος των παραπάνω είναι, η συνάρτηση κατανομής πυκνότητας πιθανότητας κάθε p - διάστατου πειραματικού σημείου, να είναι η p - διάστατη Κανονική κατανομή. Αυτό εξασφαλίζει πως, η από κοινού συνάρτηση πυκνότητας πιθανότητας, των δειγματικών παρατηρήσεων που συνιστούν εκάστη ομάδα ταξινόμησης, προς τη κατεύθυνση εκείνου του ίδιο - διανύσματος που μεγιστοποιεί τη μεταβλητότητα μεταξύ των ομάδων, είναι η μονοδιάστατη Κανονική κατανομή. Πρόσθετα των παραπάνω να σημειωθεί πως πρέπει κατάλληλα να οριστεί το διαχωριστικό σημείο *c* (*cut off point / cutting discriminant score*). Σε περίπτωση κατά την οποία, η τιμή της διαχωριστικής συναρτήσεως *LDF*, λάβει τιμή ανώτερη του *c*, τότε η πειραματική μονάδα ταξινομείται στο ένα cluster, διαφορετικά ταξινομείται στο άλλο (Rencher, 2002).

Β. Γραμμική Διαχωριστική Ανάλυση Πολλαπλών Ομάδων Ταξινόμησης (MDA: Multiple Discriminant Analysis)

Η φυσική επέκταση της μεθόδου *LDA*, είναι η μέθοδος της Πολλαπλής Γραμμικής Διαχωριστικής Ανάλυσης (*MDA*), στην οποία η κατηγορική μεταβλητή απόκρισης συνίσταται από *m >2* στο πλήθος, ομάδες – κατηγορίες ταξινόμησης. Στην εν λόγω περίπτωση των *m* διαχωριστικών ομάδων ταξινόμησης, θα απαιτηθεί η κατασκευή (m – 1) Γραμμικών Διαχωριστικών Συναρτήσεων (*LDF)*, οι οποίες είναι ανά δύο ορθογώνιες μεταξύ των, της γενικής μορφής

$$D_j = \alpha_j + \beta_{1j}X_{1j} + \beta_{2j}X_{2j} + \cdots + \beta_{pj}X_{pj}\,, \qquad j = 1, \dots, (m-1)$$

όπου

D_j, $j = 1, \dots, m$, το διαχωριστικό αποτέλεσμα της *MDA*, προκειμένου για την j ομάδα ταξινόμησης.

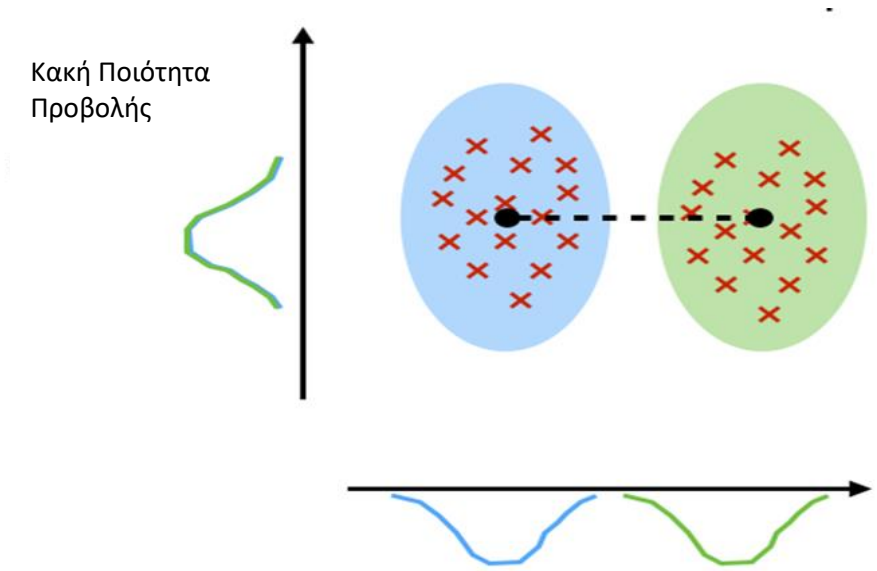
α_j, η σταθερά του διαχωριστικού μοντέλου, προκειμένου για την j ομάδα , $j = 1, \dots, m$.

X_{ij} ,το score της i στατιστικά σημαντικής ερμηνευτικής μεταβλητής, προκειμένου για την j ομάδα, $i = 1, \dots, p$, $j = 1, \dots, m$.

β_{ij}, ο αντίστοιχος συντελεστής βαρύτητας της i ερμηνευτικής μεταβλητής, προκειμένου για την j ομάδα, $i = 1, \dots, p$, $j = 1, \dots, m$.

Η πρώτη Γραμμική Διαχωριστική Συνάρτηση (*LDF*), κατασκευάζεται κατά τρόπον ώστε, να διαχωριστούν τα πειραματικά δεδομένα σε δύο ομάδες, τα κέντρα βαρών των οποίων απέχουν όσο το δυνατόν περισσότερο. Διαφορετικά αλλά ισοδύναμα, θα πρέπει να ελαχιστοποιείται η μεταβλητότητα στο εσωτερικό των ομάδων ταξινόμησης, ενώ την ίδια στιγμή, να μεγιστοποιείται η μεταβλητότητα μεταξύ των ομάδων ταξινόμησης.

Η βασική ιδέα του Γραμμικού Διαχωριστικού Μοντέλου, είναι η προβολή των αρχικών πολυδιάστατων δεδομένων, σε ένα χώρο λιγότερων διαστάσεων, στον οποίο είναι εφικτή η πραγματοποίηση ευκρινούς και άρα αποτελεσματικής κατηγοριοποίησης των δεδομένων. Ενδεικτικά αποτυπώνεται η προβολή δύο κλάσεων ισοδυναμίας, δισδιάστατων πειραματικών δεδομένων, σε δύο ευθείες (μονοδιάστατοι χώροι προβολής), όπου η προβολή στον άξονα xx΄, κρίνεται ως ευκρινέστερη και άρα ως καταλληλότερη της προβολής στον άξονα των yy΄.



Καλή Ποιότητα Προβολής με αποτελεσματικό Διαχωρισμό Ομάδων

Εάν $\boldsymbol{\ell_j}$ είναι το p – διάστατο διάνυσμα συντελεστών του γραμμικού συνδυασμού προβολής, προκειμένου για τη j κλάση ταξινόμησης, προκύπτει το j - score της k πειραματικής μονάδας

$$Y_{kj} = \boldsymbol{\ell_j}^{\mathbf{T}} \mathbf{X_k}\,,\ k = 1, \dots, n\,,\ \forall\, j = 1\,, \dots, m$$

όπου

$\mathbf{X_k}$ είναι το p διάστατο διάνυσμα τιμών, των χαρακτηριστικών της k πειραματικής μονάδας,

Επομένως

$$E(\mathbf{Y_j}) = \frac{1}{n}\sum_{k=1}^{n} Y_{kj}$$

Όπου

$$\mathbf{Y_j}^{\mathbf{T}} = \left(Y_{1j}\,, Y_{2j}\,, \dots\ Y_{nj}\right), \qquad \forall j = 1, \dots m$$

Επιπλέον έχουμε

$$Var\left(\mathbf{Y_j}^{\mathbf{T}}\right) = \boldsymbol{\ell_j}^{\mathbf{T}} \Sigma\, \boldsymbol{\ell_j}^{\mathbf{T}}$$

όπου $\mathbf{\Sigma}$ είναι ο p – διάστατος πίνακας Διακυμάνσεων – Συνδιακυμάνσεων, των πολυδιάστατων χαρακτηριστικών, $X = (\mathbf{X_1}, \mathbf{X_2}, \dots, \mathbf{X_p})$, του συνόλου των αρχικών πειραματικών δεδομένων της μελέτης , ανεξαρτήτου κλάσεως ισοδυναμίας.

Η επιλογή των συντελεστών $\boldsymbol{\ell_j}$ του γραμμικού συνδυασμού προβολής, για κάθε κλάση ταξινόμησης, θα πρέπει να είναι τέτοια ώστε, η προβολή στο χώρο των λιγότερων διαστάσεων, να ελαχιστοποιεί τη διακύμανση (άρα να μεγιστοποιεί την ομοιογένεια) στο εσωτερικό των ομάδων, ενώ συνάμα θα πρέπει να μεγιστοποιεί τη διακύμανση (να ελαχιστοποιεί την ομοιογένεια) μεταξύ των ομάδων αυτών. Με διαφορετικά αλλά ισοδύναμα λόγια, θα πρέπει να μεγιστοποιείται ο λόγος του αθροίσματος τετραγώνων μεταξύ των κλάσεων (*B: between classes Sum of Squares*), προς το άθροισμα τετραγώνων στο εσωτερικό των κλάσεων (*W: within classes Sum of Squares*), έτσι ώστε:

$$B \to \Sigma \ \ \& \ \ W \to 0$$

Συγκεκριμένα, η επιλογή των συντελεστών προβολής του διανύσματος $\boldsymbol{\ell_j}$, για κάθε κλάση ταξινόμησης *j = 1, … m*, θα πρέπει να τέτοια ώστε, να μεγιστοποιείται ο λόγος

$$F_j = \frac{\boldsymbol{\ell_j}^{\mathbf{T}} * B * \boldsymbol{\ell_j}}{\boldsymbol{\ell_j}^{\mathbf{T}} * \Sigma * \boldsymbol{\ell_j}}$$

Υπό πραγματικές συνθήκες εφαρμογής, οι πίνακες *B* και *Σ* δεν είναι γνωστοί και πρέπει να εκτιμηθούν από τα διαθέσιμα δειγματικά δεδομένα, του συνόλου των δεδομένων εκπαίδευσης. Ας υποθέσουμε πως η κατηγοριοποίηση αφορά *m* στο πλήθος ομάδες ταξινόμησης και για κάθε μία εξ' αυτών, διαθέτουμε n_j πειραματικά δεδομένα, *j = 1, … , m*. Ακολούθως προκύπτουν οι ακόλουθες δειγματικές εκτιμήσεις:

Η εκτίμηση του πίνακα *B* είναι η ακόλουθη:

$$\hat{B} = \sum_{j=1}^{m} [\ (\hat{\boldsymbol{\mu}}_{\mathbf{j}} - \hat{\boldsymbol{\mu}})\ (\hat{\boldsymbol{\mu}}_{\mathbf{j}} - \hat{\boldsymbol{\mu}})^{\mathrm{T}}]$$

όπου:

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{j}} = \frac{1}{n_j} \sum_{k=1}^{n_j} \mathbf{X_{kj}}\ , \qquad j = 1, \dots, m$$

και

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{j=1}^{m} \sum_{k=1}^{n_j} \mathbf{X_{kj}}, \quad n = \ (n_1 + n_2 + \cdots + n_m)$$

ενώ η δειγματική εκτίμηση του πίνακα *Σ* προκύπτει ως εξής

$$\hat{\Sigma} = S_{pooled} = \frac{1}{n}\,\widehat{W}$$

όπου

$$\widehat{W} = \sum_{j=1}^{m}\left(n_j - 1\right) S_j \; =$$

$$= \sum_{j=1}^{m}\sum_{k=1}^{n_j}(\mathbf{X_{jk}} - \mathbf{\hat{\mu}_j})\,(\mathbf{X_{jk}} - \mathbf{\hat{\mu}_j})^T$$

Επομένως, ο στόχος μεγιστοποίησης του προαναφερθέντος λόγου F_j , λαμβάνει τη μορφή:

$$\max_{\ell_j} \widehat{F_j} = \max_{\ell_j} \frac{\boldsymbol{\ell_j}^{\mathbf{T}} * \hat{B} * \boldsymbol{\ell_j}}{\boldsymbol{\ell_j}^{\mathbf{T}} * \hat{\Sigma} * \boldsymbol{\ell_j}} = \max_{\ell_j} \frac{\boldsymbol{\ell_j}^{\mathbf{T}} * \hat{B} * \boldsymbol{\ell_j}}{\boldsymbol{\ell_j}^{\mathbf{T}} * \widehat{W} * \boldsymbol{\ell_j}}$$

Προκειμένου να προσδιοριστεί ο βέλτιστος πολυδιάστατος χώρος προβολής των αρχικών πρωτογενών δεδομένων, ο οποίος μεγιστοποιηθεί την προαναφερθείσα στατιστική συνάρτηση, θα πρέπει να προσδιοριστεί εκείνο το βέλτιστο πλήθος ιδιοδιανυσμάτων, τα οποία αντιστοιχούν στις μεγαλύτερες ιδιοτιμές, του ακόλουθου προβλήματος εύρεσης ιδιοτιμών:

$$\hat{B}\ \boldsymbol{\ell_j} = \widehat{W}\ \Lambda\ \boldsymbol{\ell_j} \quad j = 1, \dots, m$$

Εκείνο το οποίο πρέπει εν συνεχεία να γίνει, είναι η εύρεση και ταξινόμηση των, το πολύ p στο πλήθος, μη μηδενικών πραγματικών ιδιοτιμών του πίνακα

$$\hat{\Lambda} = \widehat{W}^{(-1)}\,\hat{B}$$

ώστε να διατηρηθούν εκείνα τα ιδιοδιανύσματα, τα οποία αντιστοιχούν στις ιδιοτιμές αυτές. Τα ιδιοδιανύσματα αυτά, είναι ανά δύο ορθογώνια μεταξύ των και ορίζουν εκείνον τον βέλτιστο πολυδιάστατο χώρο επί του οποίου θα πρέπει να πραγματοποιηθεί η προβολή των αρχικών πρωτογενών δεδομένων.

Έστω ότι οι μη μηδενικές και πραγματικές ιδιοτιμές του παραπάνω δειγματικά εκτιμώμενου πίνακα $\hat{\Lambda}$ διατεταγμένες κατά φθίνουσα τάξη, είναι οι

$$\lambda_1, \lambda_2, \dots\ \lambda_s, \qquad s \le min\{m-1, p\}$$

Ενώ τα αντίστοιχα ιδιοδιανύσματα, που αντιστοιχούν σε αυτές, έστω πως είναι τα

$$\mathbf{e_1}, \mathbf{e_2}, \dots\ \mathbf{e_s}, \qquad s \le min\{m-1, p\}$$

Επομένως, το πρώτο διάνυσμα συντελεστών, που μεγιστοποιεί το λόγο F_1, στη κατεύθυνση της πρώτης κύριας συνιστώσας, είναι το

$$\hat{\boldsymbol{\ell}}_{\mathbf{1}} = \hat{\boldsymbol{e}}_{\mathbf{1}}$$

Το παραπάνω μέγιστο του στατιστικού F_1, αντιστοιχεί στη μεγαλύτερη την τάξη ιδιοτιμή λ_1, ενώ η πρώτη διαχωριστική συνάρτηση η οποία διατηρεί το μέγιστο της πληροφορίας που φέρουν τα αρχικά δεδομένα, είναι η

$$\hat{\boldsymbol{\ell}}_{\mathbf{1}}\, X$$

Ακολούθως, το αμέσως επόμενο διάνυσμα συντελεστών που μεγιστοποιεί το λόγο F_2, αντιστοιχεί στο δεύτερο ιδιοδυάνυσμα

$$\hat{\boldsymbol{\ell}}_{\mathbf{2}} = \hat{\boldsymbol{e}}_{\mathbf{2}}$$

Αντιστοιχεί στη δεύτερη μεγαλύτερη ιδιοτιμή λ_2, ενώ η δεύτερη διαχωριστική συνάρτηση, η οποία διατηρεί το αμέσως μικρότερο ποσοστό της αρχικής μεταβλητότητας των δεδομένων, είναι η:

$$\hat{\boldsymbol{\ell}}_{\mathbf{2}}\, X$$

Συνεχίζοντας με τον ίδιο τρόπο, δημιουργούμε s στο πλήθος διαχωριστικές συναρτήσεις, όπου $s \leq min\ \{m-1,\ p\}$.

1.3 Η προσέγγιση Bayes

Η εν λόγω τεχνική κατηγοριοποίησης, στηρίζεται στη στατιστική προσέγγιση του Thomas Bayes (1702 – 1762), απαιτεί δε την εξασφάλιση των ακόλουθων δύο βασικών παραδοχών

- ✓ Τα χαρακτηριστικά γνωρίσματα της μελέτης, πρέπει να είναι στατιστικώς σημαντικά, εισφέροντας ουσιαστική πληροφορία στο πρόβλημα κατηγοριοποίησης.
- ✓ Τα χαρακτηριστικά γνωρίσματα της μελέτης, πρέπει να είναι μεταξύ των ανεξάρτητα.

Η πιθανότητα να συμβεί ένα ενδεχόμενο C, δοθέντος ότι έχει παρατηρηθεί το ενδεχόμενο A, δίνεται από τον τύπο

$$P(\mathrm{C}|\mathrm{A}) = \frac{P(\mathrm{A}|\mathrm{C})\, P(\mathrm{C})}{P(\mathrm{A})}$$

Όπου

- ✓ $P(\mathrm{C})$: "*A – priori*" πιθανότητα πραγματοποίησης του ενδεχομένου C, χωρίς την επίκληση της μαρτυρίας του παρατηρούμενου ενδεχομένου A.

- ✓ $P(\mathrm{C}|\mathrm{A})$: "*A – posteriori*" πιθανότητα πραγματοποίησης του ενδεχομένου C, μετά την επίκληση της μαρτυρίας του παρατηρούμενου ενδεχομένου Α.

Στην περίπτωση του υπό εξέταση προβλήματος κατηγοριοποίησης, επιδιώκουμε την εκτίμηση της πιθανότητας πραγματοποίησης του ενδεχομένου C, δοθέντος του γεγονότος πραγματοποίησης του ενδεχομένου Α, λαμβάνοντας υπόψιν συγκεκριμένο σύνολο, ανεξάρτητων μεταξύ τους χαρακτηριστικών.

Επομένως

- ✓ Το ενδεχόμενο C ορίζεται ως το ενδεχόμενο, η εξεταζόμενη πειραματική μονάδα, να ανήκει σε μία εκ των δύο κλάσεων ισοδυναμίας του προβλήματος.
- ✓ Το ενδεχόμενο Α, αντιστοιχεί σε μία καταγεγραμμένη εγγραφή της βάσεως των πειραματικών δεδομένων εκπαίδευσης, η οποία αφορά συγκεκριμένη πειραματική μονάδα, με καταγεγραμμένες τιμές επί των προεπιλεγμένων ανεξάρτητων χαρακτηριστικών της μελέτης.

Η επέκταση του θεωρήματος του Bayes μπορεί να γίνει, αν θεωρήσουμε τα ενδεχόμενα $C_1, \dots, C_m$, τα οποία διαμερίζουν το δειγματικό χώρο $\boldsymbol{\Omega}$, έτσι ώστε $C_i \cap C_j = \emptyset$, για κάθε $i \neq j$ και $C_1 \cup \dots \cup C_m = \Omega$. Σε αυτήν την περίπτωση θα έχουμε

$$P\left(C_j \,|A\right) = \frac{P(A \mid C_j)P(C_j)}{\sum_{j=1}^{m} P(A \mid C_j)P(C_j)}, j = 1, \dots, m$$

Απαραίτητη προϋπόθεση προκειμένου να υπολογιστούν οι παραπάνω πιθανότητες είναι, καμία από τις παρατηρούμενες συχνότητες (*observed frequencies*), να μην εμφανίζεται ως μηδενική. Σε περίπτωση που κάτι τέτοιο συμβεί, στο πίνακα των παρατηρούμενων συχνοτήτων, προσθέτουμε παντού μία μονάδα και υπολογίζουμε εκ νέου τις παραπάνω πιθανότητες.

Στη πράξη, η παραπάνω μέθοδος κατά Bayes δουλεύει εξαιρετικά αποτελεσματικά, ακόμη και σε περιπτώσεις κατά τις οποίες δεν εξασφαλίζεται απολύτως η ανεξαρτησία, μεταξύ των χρησιμοποιούμενων γνωρισμάτων. Αυτό συμβαίνει διότι η μέθοδος, δεν απαιτεί ακριβείς εκτιμήσεις των επιμέρους πιθανοτήτων.

2 Λογιστική Παλινδρόμηση

Η κατηγοριοποίηση μέσω της μεθόδου της Λογιστικής Παλινδρόμησης, χρησιμοποιείται στη περίπτωση κατά την οποία, επιχειρείται η μοντελοποίηση μίας κατηγορικής μεταβλητής Απόκρισης, ενάντια ενός συνόλου ερμηνευτικών ανεξαρτήτων μεταβλητών (είτε κατηγορικών, είτε συνεχών). Μέσω της εν λόγω μεθόδου εκτιμάται η πιθανότητα, η υπό εξέταση πειραματική μονάδα να ανήκει σε ένα εκ των (m-1) διακεκριμένων επιπέδων ($m \geq 2$) της μεταβλητής Απόκρισης, δεδομένου ενός επιπέδου αναφοράς της (Hosmer & Lemeshow, 2007, Agresti, 1990).

Τα είδη της μοντελοποίησης τα οποία υλοποιούνται μέσω της τεχνικής της Λογιστικής Παλινδρόμησης είναι δύο, αναλόγως του πλήθους των επιπέδων της μεταβλητής Απόκρισης. Πιο συγκεκριμένα, σε περίπτωση κατά την οποία η εξαρτημένη μεταβλητή εμφανίζει δύο επίπεδα, υλοποιείται το μοντέλο της Διωνυμικής Λογιστικής Παλινδρόμησης (με απλά ή ομαδοποιημένα Δίτιμα δεδομένα), ενώ στη περίπτωση κατά την οποία τα επίπεδα είναι περισσότερα των δύο, υλοποιείται η μοντελοποίηση της Πολυωνυμικής Λογιστικής Παλινδρόμησης (με απλά ή ομαδοποιημένα Πολυωνυμικά δεδομένα).Μέσω της εν λόγω τεχνικής, είναι δυνατή η μελέτη της συνολικής επάρκειας του μοντέλου μετά τη προσαρμογή στα πειραματικά δεδομένα, ο προσδιορισμός του ποσοστού της συνολικής μεταβλητότητας των δεδομένων το οποίο επεξηγείται από το μοντέλο, η ταξινόμηση των ερμηνευτικών χαρακτηριστικών αναλόγως της ερμηνευτικής των ικανότητας, η εκτίμηση των αντιστοίχων συντελεστών μέσω της μεθόδου μεγιστοποίησης της Πιθανοφάνειας, καθώς και η μελέτη της στατιστικής σημαντικότητας τυχόν αλληλεπιδράσεων, μεταξύ των ερμηνευτικών μεταβλητών (Kutner , 2004).

Στο απλό Διχοτομικό μοντέλο, εκτιμάται η πιθανότητα η πειραματική μονάδα να ανήκει σε ένα εκ των δύο επιπέδων της μεταβλητής Απόκρισης, αναλόγως της τιμής της εκτιμώμενης πιθανότητας. Εάν η τιμή της εκτιμώμενης πιθανότητας, υπερβαίνει το τιθέμενο διαχωριστικό σημείο αποκοπής (*cut off point / threshold*), τότε η πειραματική μονάδα κατατάσσεται στη δεύτερη κατηγορία της Διχοτομικής Μεταβλητής Απόκρισης, άλλως στη κατηγορία αναφοράς.

Στο Μοντέλο της Πολυωνυμικής Λογιστικής Παλινδρόμησης, η μεταβλητή Απόκρισης διαθέτει περισσότερες των δύο κατηγορίες και μοντελοποιείται ενάντια ενός συνόλου ερμηνευτικών μεταβλητών, είτε κατηγορικών, είτε συνεχών. Μέσω της διαδικασίας αυτής, εκτιμάται η πιθανότητα μία πειραματική μονάδα, να ανήκει σε μία εκ των εναλλακτικών επιπέδων της μεταβλητής Απόκρισης, δεδομένου του επιπέδου αναφοράς της. Κατά αντιστοιχία με το απλό Διχοτομικό Μοντέλο, η εκτίμηση των παραμέτρων του μοντέλου Πολλαπλής Λογιστικής Παλινδρόμησης, υλοποιείται μέσω της μεθόδου Μεγιστοποίησης της Πιθανοφάνειας.

2.1 Υποθέσεις Μοντέλων Λογιστικής Παλινδρόμησης

Το βασικό πλεονέκτημα της Λογιστικής Παλινδρόμησης είναι το γεγονός πως, δεν απαιτεί την ισχύ καμίας περιοριστικής υπόθεσης, σε αντίθεση με τα κλασσικά μοντέλα Γραμμικής Διαχωριστικής Ανάλυσης. Επομένως, δεν υφίστανται υποχρεωτικά οι περιοριστικές υποθέσεις της γραμμικής σχέσεως μεταξύ της μεταβλητής απόκρισης και των ερμηνευτικών μεταβλητών της μελέτης, της κανονικότητας της απόκρισης, ή ισοδύναμα των καταλοίπων του μοντέλου, καθώς και της ομοσκεδαστικότητας των καταλοίπων.

Οι μόνες περιοριστικές υποθέσεις οι οποίες απαιτείται να ισχύουν, σε ένα μοντέλο Λογιστικής Παλινδρόμησης, είναι η κατανομή της απόκρισης να προέρχεται από την Εκθετική Οικογένεια Κατανομών, καθώς και της ανεξαρτησίας μεταξύ των προβλεπουσών ερμηνευτικών μεταβλητών της ανάλυσης. Ειδικότερα, η φυσική ερμηνεία της υπόθεσης της ανεξαρτησίας των ερμηνευτικών μεταβλητών, έχει να κάνει με το γεγονός πως, η πιθανότητα μία πειραματική μονάδα να ανήκει σε μία εκ των κατηγοριών της μεταβλητής απόκρισης, δεν επηρεάζει, ούτε επηρεάζεται, από τη πιθανότητα η εν λόγω πειραματική μονάδα να ανήκει σε μία διαφορετική, εκ των εναλλακτικών κατηγοριών της μεταβλητής απόκρισης.

Ο βασικότερος έλεγχος μέσω του οποίου ελέγχεται η μηδενική υπόθεση ανεξαρτησίας των ερμηνευτικών μεταβλητών, είναι ο έλεγχος των Hausman – Mc Fadden, (Mertler & Vannatta, 2002). Ανακεφαλαιώνοντας, οι βασικές υποθέσεις ενός μοντέλου Λογιστικής Παλινδρόμησης, είναι οι ακόλουθες:

- ✓ Η σχέση της μεταβλητής απόκρισης και των ανεξάρτητων ερμηνευτικών μεταβλητών της ανάλυσης, δεν είναι απαραιτήτως γραμμική.
- ✓ Οι ερμηνευτικές μεταβλητές του μοντέλου πρέπει να είναι ανεξάρτητες μεταξύ των
- ✓ Δεν είναι υποχρεωτική η ισχύς των περιοριστικών υποθέσεων της ομοσκεδαστικότητας, καθώς και της κανονικότητας της κατανομής των καταλοίπων, ανά κατηγορία της μεταβλητής απόκρισης, μετά τη προσαρμογή του μοντέλου στα πειραματικά δεδομένα.
- ✓ Η κατανομή της μεταβλητής απόκρισης, πρέπει να ανήκει στην Εκθετική Οικογένεια Κατανομών.
- ✓ Η μεταβλητή απόκρισης πρέπει να είναι κατηγορική, με τουλάχιστον δύο διαφορετικά επίπεδα.
- ✓ Τα επίπεδα της κατηγορικής μεταβλητής απόκρισης, πρέπει να είναι αυστηρώς καθορισμένα και αμοιβαίως αποκλειόμενα.

Εξαιτίας του γεγονότος πως η εκτίμηση των παραμέτρων του μοντέλου γίνεται αποκλειστικά με τη μέθοδο Μεγιστοποίησης της Πιθανοφάνειας, απαιτείται μεγάλο μέγεθος δείγματος πειραματικών δεδομένων εκπαίδευσης, τουλάχιστον 50 παρατηρήσεων ανά ερμηνευτική μεταβλητή.

2.2 Διχοτομική Λογιστική Παλινδρόμηση

Το βασικό πρότυπο του μοντέλου Απλής (Διχοτομικής) Λογιστικής Παλινδρόμησης, είναι το ακόλουθο

$$Logit\ (\pi_i) = Log\left(\frac{\pi_i}{1-\pi_i}\right) = Log\left[odds_{(i)}\right] = \ b_o + \sum_{j=1}^{p} b_j\, X_{ji}\,, \quad i = 1, \dots, n$$

το οποίο ισοδύναμα γράφεται ως εξής

$$\pi_i = \frac{Exp[b_0 \ + \ \sum_{j=1}^{p} b_j\, X_{ji}]}{1 + Exp[b_0 \ + \ \sum_{j=1}^{p} b_j\, X_{ji}]}, \qquad i = 1, \dots, n$$

όπου p το πλήθος των ερμηνευτικών μεταβλητών και n το πλήθος των παρατηρήσεων του δείγματος.

Η πιθανότητα $\pi_i\ ,\ i = 1, \dots, n$, εκφράζει την πιθανότητα η Διχοτομική (Δίτιμη) μεταβλητή Απόκρισης, να λάβει τιμή ίση με 1, επομένως κατ' ουσία εκφράζει τη πιθανότητα υλοποίησης του ενδεχόμενου επιτυχίας. Σε επίπεδο εκτιμητικής, η εκτιμώμενη πιθανότητα πραγματοποίησης του ενδεχομένου επιτυχίας, θα πρέπει να λαμβάνει τιμή ανώτερη του τιθέμενου διαχωριστικού ορίου αποκοπής (*cut off point*), (Tabachnick , 1996).

2.3 Πολυωνυμική Λογιστική Παλινδρόμηση

Το βασικό πρότυπο του μοντέλου Πολλαπλής Λογιστικής Παλινδρόμησης, είναι το ακόλουθο:

$$Logit[\pi_k(\mathbf{X_i})] = Log\left[\frac{\pi_k(\mathbf{X_i})}{\pi_1(\mathbf{X_i})}\right] = Log[odds_{(ik)}] =$$

$$= b_{0k} + \sum_{j=1}^{p} b_{jk}\, X_{ji}\ , \qquad i = 1, \dots, n\ , \qquad k = 2, \dots, m$$

Ισοδύναμα

$$\pi_k(\mathbf{X_i}) = \frac{Exp[b_{0k} + \sum_{j=1}^{p} b_{jk} * X_{ji}]}{1 + Exp[b_{0k} + \sum_{j=1}^{p} b_{jk} * X_{ji}]}\ ,\ i = 1, \dots, n \quad k = 2, \dots, m$$

Η $\pi_k(\mathbf{X_i})$ εκφράζει τη πιθανότητα, η πειραματική μονάδα *i* να ανήκει στο k επίπεδο της μεταβλητής απόκρισης, δεδομένου του γεγονότος ότι, το πρώτο επίπεδο λαμβάνεται ως το επίπεδο αναφοράς της ανάλυσης, ενάντια του οποίου πραγματοποιούνται όλες οι δυνατές ανά δύο συγκρίσεις (Chatterjee & Hadi, 2006).

2.4 Πλεονεκτήματα & Περιορισμοί Μοντέλων Λογιστικής Παλινδρόμησης, έναντι Μοντέλων Γραμμικής Διαχωριστικής Ανάλυσης

Τα βασικότερα πλεονεκτήματα ενός Μοντέλου Λογιστικής Παλινδρόμησης, έναντι ενός Μοντέλου Γραμμικής Διαχωριστικής Ανάλυσης, είναι τα ακόλουθα:

- ✓ Δεν απαιτείται τα κατάλοιπα του μοντέλου να κατανέμονται κανονικά.
- ✓ Δεν απαιτείται ομοσκεδαστικότητα των καταλοίπων, στις διάφορες κατηγορίες της μεταβλητής απόκρισης.
- ✓ Δεν απαιτείται η διασφάλιση της γραμμικής σχέσης, μεταξύ μεταβλητής απόκρισης και συνόλου ερμηνευτικών μεταβλητών.
- ✓ Σε ένα μοντέλο Λογιστικής Παλινδρόμησης, η εισαγωγή και διαχείριση μη γραμμικών όρων αλληλεπίδρασης, είναι άμεση και πολύ πιο απλή, σε σχέση με ένα μοντέλο Διαχωριστικής Ανάλυσης.

Παρόλα αυτά, υφίστανται οι ακόλουθοι περιορισμοί:

- ✓ Η μεταβλητή απόκρισης ενός μοντέλου Λογιστικής Παλινδρόμησης, πρέπει υποχρεωτικά να είναι κατηγορική.
- ✓ Η μοντελοποίηση μέσω της μεθόδου της Λογιστικής Παλινδρόμησης και ειδικότερα της Πολυωνυμικής, προκειμένου να εξασφαλίσει ακριβείς εκτιμήσεις των παραμέτρων του μοντέλου, απαιτεί μεγάλα μεγέθη δειγμάτων.
- ✓ Ο αυστηρός περιορισμός της γραμμικής σχέσης μεταξύ της μεταβλητής απόκρισης και του συνόλου των ερμηνευτικών μεταβλητών του μοντέλου, δεν είναι απαραίτητο να ισχύει. Παρόλα αυτά, η γραμμική σχέση η οποία κρίνεται ως υποχρεωτική, προκειμένου να λειτουργήσει αποτελεσματικά και αξιόπιστα η μοντελοποίηση, είναι εκείνη μεταξύ του λογαρίθμου της σχετικής πιθανότητας του ενδεχομένου επιτυχίας (*Logit function*) και των ερμηνευτικών μεταβλητών του μοντέλου (Hilbe , 2009).
- ✓ Η κατανομή πιθανότητας της μεταβλητής Απόκρισης, θα πρέπει να προέρχεται από την Εκθετική Οικογένεια Κατανομών. Μία κατανομή πιθανότητας ανήκει στην εκθετική οικογένεια κατανομών, όταν η συνάρτηση πιθανότητας της κατανομής (ή πυκνότητας πιθανότητας αν η κατανομή είναι συνεχής), μπορεί να γραφεί στη γενική μορφή

$$f_Y(y\,;\,\theta\,,\varphi) = exp[\frac{y\theta - b(\theta)}{\alpha(\varphi)} + c(y,\varphi)]$$

όπου a, b, c είναι τρεις γνωστές συναρτήσεις, ενώ οι θ, φ είναι παράμετροι. Αν η παράμετρος φ είναι γνωστή, τότε έχουμε την εκθετική οικογένεια κατανομών με μία παράμετρο και το θ αναφέρεται ως η 'κανονική παράμετρος' (*canonical parameter*) της κατανομής. Αν η παράμετρος φ δεν είναι γνωστή, τότε μπορούμε σε πολλές περιπτώσεις να τη θεωρήσουμε ως μία παράμετρο κλίμακας για την κατανομή, οπότε αποκαλείται 'παράγοντας όχλησης' (*nuisance factor*) της κατανομής.

3 Διαχωριστικοί Αλγόριθμοι κατασκευής Δέντρων Απόφασης

Η κατηγοριοποίηση μίας βάσης δεδομένων μέσω της χρήσης Δέντρων Απόφασης, στηρίζεται στην αρχή της διαμέρισης του χώρου σε ξένες ανά δύο περιοχές. Η πατρότητα της συγκεκριμένης μεθοδολογίας ανήκει στους Sonquist, Morgan (1964), οι οποίοι αρχικά αξιοποίησαν τεχνικές κλασικής Παλινδρόμησης, προκειμένου να κατασκευάσουν Δέντρα Απόφασης με προβλεπτική και επεξηγηματική ικανότητα (*AID: Automatic Interaction Detection*). Οι βασικοί αλγόριθμοι παραγωγής τέτοιων δέντρων κατηγοριοποίησης, διαφέρουν στον τρόπο με τον οποίο επιλέγεται η ρίζα του δέντρου (*root*), καθώς και στον τρόπο με τον οποίο δημιουργείται η μετέπειτα δομή του. Ως ρίζα του δέντρου, επιλέγεται το κατά περίπτωση σημαντικότερο χαρακτηριστικό γνώρισμα της ανάλυσης, ενώ τα υπόλοιπα γνωρίσματα της κατηγοριοποίησης, αντιστοιχούν στους εσωτερικούς κόμβους (*nodes*) του δέντρου. Τα τόξα κάθε κόμβου, αντιστοιχούν στα επίπεδα εκάστου υπό μελέτη χαρακτηριστικού. Καταληκτικά, τα φύλλα (*leafs*) του δέντρου, αντιστοιχούν στις κλάσεις ισοδυναμίας του προβλήματος κατηγοριοποίησης.

Τα βασικότερα ζητήματα τα οποία πρέπει κανείς να διαχειριστεί, προκειμένου να κατασκευαστεί ένα δέντρο απόφασης, είναι τα ακόλουθα:

3.1 Επιλογή Χαρακτηριστικών Γνωρισμάτων

Καθορισμός των σημαντικότερων χαρακτηριστικών μεταξύ των διαθεσίμων μεταβλητών της ανάλυσης, τα οποία απαιτείται να διατηρηθούν και χρησιμοποιηθούν, για τη κατηγοριοποίηση των δεδομένων και τη μετέπειτα διαμέριση του Δειγματικού χώρου σε ξένες ανά δύο περιοχές. Το σημαντικότερο εξ' αυτών, θα χρησιμοποιηθεί ως ρίζα του δέντρου, ενώ τα υπόλοιπα θα αποτελέσουν τους εσωτερικούς του κόμβους

α. Επιλογή Επιπέδων ανά Χαρακτηριστικό Γνώρισμα

Για κάθε χαρακτηριστικό διάσπασης, θα πρέπει να προσδιοριστεί το βέλτιστο πλήθος επιπέδων. Τα επίπεδα αυτά μπορεί να είναι, είτε κάποια εκ των αρχικών επιπέδων, είτε κάποια νέα που προέκυψαν ως συγχώνευση των αρχικών. Επιπλέον των παραπάνω, σε περίπτωση κατά την οποία κάποιο εκ των χαρακτηριστικών της ανάλυσης είναι συνεχής μεταβλητή, απαιτείται η κατηγοριοποίηση της, δηλαδή η δημιουργία κλάσεων, από τις οποίες θα προκύψουν τα επίπεδα που θα χρησιμοποιηθούν.

β. Δομή Δέντρου Απόφασης

Το παραγόμενο Δέντρο Απόφασης, ενδέχεται να είναι είτε Δυαδικό, όπου σε κάθε κόμβο απόφασης έχουμε δύο διαφορετικά αμοιβαίως αποκλειόμενα επίπεδα κατάταξης, είτε πολλαπλά επίπεδα, τα οποία συνιστούν τα αντίστοιχα τόξα ανά κόμβο λήψης απόφασης. Σε κάθε περίπτωση, το παραγόμενο Δέντρο Απόφασης πρέπει να είναι ει δυνατόν ισοζυγισμένο, δηλαδή όλες οι εναλλακτικές διαδρομές, από τη ρίζα προς τα φύλλα του δέντρου, να έχουν συγκρίσιμο μήκος. Μετά την ολοκλήρωση ενός Δέντρου Απόφασης, ενδέχεται να είναι απαραίτητη η εφαρμογή κανόνων 'κλαδέματος' (*Pruning*), ώστε να παραχθεί η πλέον ισοζυγισμένη μορφή.

γ. Κριτήρια Τερματισμού

Καθορισμός κατάλληλων κριτηρίων τερματισμού της αλγοριθμικής διαδικασίας, ώστε να ικανοποιηθούν, οι ακόλουθες αντικρουόμενες απαιτήσεις

- ✓ Ακρίβεια κατηγοριοποίησης
- ✓ Ταχύτητα αλγορίθμου
- ✓ Πρόβλημα υπερπροσαρμογής (*Over fitting*)

3.2 Κριτήρια επιλογής Χαρακτηριστικών διάσπασης

Προκειμένου να επιλεγεί το κατά περίπτωση καταλληλότερο χαρακτηριστικό διάσπασης, υιοθετείται Ευριστικός Κανόνας (*Heuristic Rule*) μεγιστοποίησης, κατάλληλα επιλεγμένης συνάρτησης καταλληλότητας (*Fitness Function*). Η σημαντικότερη συνάρτηση καταλληλότητας, είναι εκείνη που μεγιστοποιεί το 'Κέρδος Πληροφορίας' (*Information Gain*), προκειμένου για τον Αλγόριθμο *ID3* και η συνάρτηση καταλληλότητας 'Λόγος Κέρδους Πληροφορίας' (*Information Gain Ratio*), προκειμένου για τον Αλγόριθμο *C_4.5*. Και οι δύο προαναφερθείσες συναρτήσεις καταλληλότητας, βασίζονται στην έννοια της Εντροπίας της Πληροφορίας των Δεδομένων (*Information Entropy*).

Έστω $C = \{C_1, C_2, \dots, C_m\}$ μία διαμέριση του χώρου των πειραματικών δεδομένων και $p_i,\ i = 1, \dots, m$ η πιθανότητα μία πειραματική μονάδα του χώρου των δεδομένων, να ανήκει στη κλάση i. Προφανώς ισχύει ότι:

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

Η εντροπία της πληροφορίας κατά (Benjamin Devéze , Matthieu Fouquin, 2005), ορίζεται ως ακολούθως:

$$H(C) = \sum_{i=1}^{m} [\, p_i \ log_\beta(1/p_i) \,] \in [0,1]$$

Στον παραπάνω τύπο υπολογισμού της εντροπίας, απαραίτητη κρίνεται η χρήση λογάριθμου βάσεως 2. Διαισθητικά, η έννοια της εντροπίας της πληροφορίας, συνδέεται άμεσα με την έννοια της αβεβαιότητας. Σε καθεστώς πλήρους βεβαιότητας, η εντροπία μηδενίζεται, ενώ σε καθεστώς πλήρους αβεβαιότητας, η εντροπία μεγιστοποιείται λαμβάνοντας τιμή ίση με 1. Πράγματι, εάν παραδειγματικά θεωρήσουμε μία διαμέριση δύο περιοχών, τότε θα έχουμε:

$$H(p, 1-p) = p \ log_2(1/p) + (1-p)\, log_2\left[1/(1-p)\right] = -[\, p \ log_2(p) + (1-p)\, log_2(1-p) \,]$$

α. Συνθήκες Πλήρους Βεβαιότητας

Σε συνθήκες πλήρους βεβαιότητας, η πιθανότητα p θα ισούται με *0* ή *1*. Επομένως, στη περίπτωση αυτή θα έχουμε:

$$H(0,1) = 0 - 0 = 0$$

β. Συνθήκες Πλήρους Αβεβαιότητας

Σε συνθήκες πλήρους αβεβαιότητας, η πιθανότητα p θα ισούται με *0,5*. Επομένως, στη περίπτωση αυτή θα έχουμε:

$$H(0.5, 0.5) = 2\ (0{,}5 \ log_2 2) = 1$$

Από τα παραπάνω γίνεται άμεσα προφανές πως ο στόχος της κατηγοριοποίησης, πρέπει να είναι η ελαχιστοποίηση της αβεβαιότητας, δηλαδή η ελαχιστοποίηση της εντροπίας, κατόπιν της εφαρμογής της διαμέρισης στα πειραματικά δεδομένα της ανάλυσης.

3.3 Αλγόριθμοι βασισμένοι στην Εντροπία της πληροφορίας

Προκειμένου να κατασκευαστεί επαγωγικά ένα δέντρο απόφασης, ο σχετικός αλγόριθμος ο οποίος κατά περίπτωση χρησιμοποιείται, επιχειρεί τη μεγιστοποίηση μίας εκ των προαναφερθέντων συναρτήσεων καταλληλότητας, κατά τρόπον ώστε να επιτευχθούν οι κάτωθι στόχοι:

- ✓ Κατασκευή του μικρότερου δέντρου.
- ✓ Κατασκευή του πλέον ισοζυγισμένου δέντρου.
- ✓ Χρήση κατάλληλου ευριστικού κανόνα, ώστε να μεγιστοποιηθεί η τιμή της συνάρτησης καταλληλότητας.

Ένας ενδεικτικός αλγοριθμικός κανόνας, ο οποίος περιγράφει την υλοποίηση της παραπάνω απαίτησης, σε μορφή '*Ψευδοκώδικα*' (*Pseudo code*), είναι ο παρακάτω:

Input:

D //Training data

Output:

T //Decision tree (DT)

DT Build algorithm: //Simplistic algorithm to illustrate naïve building of a DT

Begin

T = ∅;

Determine splitting criterion;

T = Create root node and label with splitting attribute;

T = Add arc to root node for each split predicate and label;

For each arc **do**

Begin

D = Database created by applying splitting predicate to D;

if stopping point reached for this path, **then**

T' = Create leaf node and label with appropriate class;

Else

T' = DT Build (D); T = Add T' to arc;

End

End

Οι κυριότεροι αλγόριθμοι δημιουργίας δέντρων απόφασης και κατηγοριοποίησης, διακρίνονται αναλόγως της χρησιμοποιούμενης συναρτήσεως καταλληλότητας και είναι οι εξής:

α. Αλγόριθμος ID3

Ο αλγόριθμος *ID3* (*Iterative Dichotomiser 3*), αρχικώς αναπτύχθηκε από τον Quinlan (*1986*) στο Πανεπιστήμιο του Σύδνεϋ, βασιζόμενος στη εργασία του Hunt (1962) και αξιοποιώντας τη γενικότερη θεώρηση του Αλγορίθμου *CLS* (*Concept Learning System*). Ο αλγόριθμος επιλέγει εκείνο το χαρακτηριστικό της ανάλυσης, ως το καταλληλότερο για τη διαμέριση του χώρου των πειραματικών δεδομένων *D*, το οποίο ελαχιστοποιεί την εντροπία του *Shannon*, μετά της εφαρμογής της αντιστοίχου διαμερίσεως *C*. Επιπροσθέτως, μεγιστοποιείται η χρησιμοποιούμενη συνάρτηση καταλληλότητας, η οποία στην περίπτωση του αλγόριθμου *ID3*, ονομάζεται 'Κέρδος Πληροφορίας' (*IG: Information Gain*) και η οποία ορίζεται ως ακολούθως:

$$IG(D, C) = H(D) - \sum_{i=1}^{m} P(D_i)\, H(D_i)$$

όπου

- ✓ $H(D)$: Η συνολική εντροπία του χώρου των δεδομένων, πριν την εφαρμογή της διαμέρισης *C*.

- ✓ $H(D_i)$: Η εντροπία των επιμέρους κλάσεων, μετά την εφαρμογή της διαμέρισης *C*.

Το βασικό μειονέκτημα του αλγόριθμου *ID3* είναι πως μεροληπτεί, υπέρ εκείνων των χαρακτηριστικών, τα οποία εμφανίζουν μεγάλο αριθμό επιπέδων, επομένως εμφανίζει ευαισθησία στο πρόβλημα της υπερπροσαρμογής στα πειραματικά δεδομένα. Απαιτεί δε τη χρήση κατηγορικών μεταβλητών, επομένως, πιθανή χρήση συνεχούς μεταβλητής καθίσταται εφικτή, μόνον εφόσον προηγηθεί κατηγοριοποίηση της αντιστοίχου συνεχούς κλίμακας της μεταβλητής.

Επιπρόσθετα, δεν καθίσταται δυνατή η χρήση εγγραφών / στιγμιότυπων (*records / instances*), οι οποίες ενέχουν μία ή περισσότερες ελλείπουσες τιμές (*missing values*), σε ένα ή περισσότερα χαρακτηριστικά της ανάλυσης. Στην τελευταία περίπτωση, θα πρέπει να χρησιμοποιηθούν οι πλήρεις εγγραφές του συνόλου των δεδομένων εκπαίδευσης και εν συνεχεία να χρησιμοποιηθεί το Δέντρο Απόφασης που προέκυψε, ώστε να εκτιμηθούν τυχόν ελλείπουσες τιμές, με τη πλέον πιθανή τιμή ανά περίπτωση χαρακτηριστικού. Εν συνεχεία, το πλήρες σύνολο των δεδομένων εκπαίδευσης, χρησιμοποιείται ώστε να επαναπροσδιοριστεί το Δέντρο Απόφασης, με απώτερο στόχο τη μείωση του Φαινομενικού Σφάλματος Ταξινόμησης (*APER: Apparent Error Rate*).

β. Αλγόριθμος C 4.5

Ο αλγόριθμος *C_4.5*, λειτουργεί ακριβώς με την ίδια λογική που αναπτύξαμε προηγούμενα στον αλγόριθμο *ID3*, διορθώνοντας όμως σε μεγάλο βαθμό τη προαναφερθείσα ευαισθησία του, σε περιπτώσεις υπερπροσαρμογής (Quinlan, 1993).

Η συνάρτηση καταλληλότητας του αλγόριθμου *C_4.5*, είναι πλέον ο 'Λόγος Κέρδους Πληροφορίας' (*IGR: Information Gain Ratio*), η οποία ορίζεται ως εξής:

$$IGR\ (D, C) = IG(D, C)/H(\frac{|D_{1|}}{|D|}, \frac{|D_2|}{|D|}, \dots, \frac{|D_m|}{|D|})$$

όπου

$$p_i = \frac{|D_i|}{|D|}, i = 1, \dots, m$$

$$H(p_1, p_2, \dots, p_m) = -\sum_{i=1}^{m} [\, p_i\ log_2(p_i) + (1 - p_i)\, log_2(1 - p_i)\,]$$

Ο αλγόριθμος *C_4.5* σε σχέση με τον αλγόριθμο *ID3*, παρέχει την εξαιρετικά χρήσιμη δυνατότητα, μετέπειτα 'κλαδέματος' του προκύπτοντος διαχωριστικού δέντρου απόφασης (Xindong Wu, et. al. , 2008), εξασφαλίζοντας τα ακόλουθα βασικά πλεονεκτήματα:

- ✓ Περιορισμός της πολυπλοκότητας της δομής του δέντρου απόφασης, μέσω κατάλληλης αποκοπής τμημάτων του δέντρου (*sub – trees*), τα οποία δεν εισφέρουν σημαντικά στη προβλεπτική / διαχωριστική ικανότητα του μοντέλου.
- ✓ Αποφεύγεται η υπερπροσαρμογή του δέντρου απόφασης στα πειραματικά δεδομένα και περιορίζεται η ευαισθησία του στο θόρυβο των πειραματικών δεδομένων (*sample noise*), μέσω της αποκοπής τμημάτων του δέντρου, τα οποία πιθανότατα προέκυψαν από ακραία ή εσφαλμένα δεδομένα.

γ. Αλγόριθμος CART

Ο συγκεκριμένος αλγόριθμος παράγει αποκλειστικά δυαδικά / διχοτομικά δέντρα, βασιζόμενος στη Θεωρία ελαχιστοποίησης της εντροπίας της πληροφορίας του συνόλου των πειραματικών δεδομένων και συνάμα στη μεγιστοποίηση της αντιστοίχου συνάρτησης καταλληλότητας (Breiman et al. , 1984). Στην εν λόγω περίπτωση, η χρησιμοποιούμενη συνάρτηση καταλληλότητας, ορίζεται ως ακολούθως:

$$\Phi\ (s|t) = 2\ P(Left)\ P(Right) \sum_{i=1}^{m} \{\, |P(C_i|t_{Left}) - P(C_i|t_{Right})| \,\}$$

Πιο συγκεκριμένα, οι όροι *P(Left)* και *P(Right)* εκφράζουν την πιθανότητα, τυχούσα εγγραφή s της βάσης δεδομένων, να βρεθεί αντιστοίχως στο αριστερό ή στο δεξί μονοπάτι του δέντρου, που ξεκινά από τον κόμβο *t*. Κατά ανάλογο τρόπο, οι πιθανότητες $P(C_i|t_{Left})$ και $P(C_i|t_{Right})$ εκφράζουν τη δεσμευμένη πιθανότητα, τυχούσα εγγραφή s της βάσης δεδομένων, να ανήκει στη κλάση C_i , $i = 1, \dots, m$, δεδομένου του γεγονότος πως η εν λόγω εγγραφή, βρίσκεται είτε στο αριστερό, είτε στο δεξί μονοπάτι αντιστοίχως, που ξεκινά από το κόμβο *t*.

3.4 Αλγόριθμοι βασισμένοι στην απόσταση

Η εν λόγω μεθοδολογία αναπτύσσει αλγόριθμους διαχωριστικής ανάλυσης, οι οποίοι κατατάσσουν τυχούσα πειραματική μονάδα προερχόμενη από το σύνολο των δεδομένων εκπαίδευσης, σε μία εκ των προκαθορισμένων κλάσεων του προβλήματος κατηγοριοποίησης, με κριτήριο την απόσταση μεταξύ πειραματικών μονάδων, ή πειραματικών μονάδων και κέντρου βάρους (*centroid*) υφιστάμενης κλάσεως του προβλήματος.

Ο πλέον αντιπροσωπευτικός αλγόριθμος αυτής της κατηγορίας, είναι ο αλγόριθμος των Κ Πλησιέστερων Γειτόνων (*KNN: K Nearest Neighbors*). Ο τρόπος ορισμού της απόστασης κατά περίπτωση διαφέρει. Στο πλείστο των περιπτώσεων χρησιμοποιείται η Ευκλείδεια απόσταση, είτε η απόσταση κατά *Pearson*, είτε η απόσταση κατά *Ward*. Ο αλγόριθμος *KNN*, κατατάσσει την εξεταζόμενη πειραματική μονάδα, ακολουθώντας επαναληπτικά τα εξής βήματα:

- ✓ Αρχικά προσδιορίζονται οι *K* το πλήθος πλησιέστερες πειραματικές μονάδες, από την υπό κατάταξη πειραματική μονάδα.

- ✓ Η υπό κατάταξη πειραματική μονάδα, κατατάσσεται σε εκείνη την κλάση, η οποία περιέχει το μεγαλύτερο πλήθος ‘Πλησιέστερων Γειτόνων’.

Η προαναφερθείσα αλγοριθμική διαδικασία, εκφρασμένη σε όρους ‘Ψευδοκώδικα’ έχει ως ακολούθως

Input:
T //training data
K //Number of neighbors
t //Input tuple to classify

Output:
c //Class to which t is assigned

KNN algorithm: //Algorithm to classify tuple using KNN
begin
N = ∅;
//Find set of neighbors, N, for t
for each d ∈ T **do**

if |N| ≤ K
then N = N ∪ {d};
else if ∃ u ∈ N such that

dis (t, u) ≥ dis (t, d) AND dis (t, u) ≥ dis (t, u') ∀ u' ∈ N

then N = N – {u}; N = N ∪ {d};

//Find class for classification
c = class to which the most u ∈ N are classified

end

4 Αξιολόγηση επάρκειας και καλής προσαρμογής Μοντέλων Διαχωριστικής Ανάλυσης

Η έννοια της επάρκειας και καλής προσαρμογής ενός Στατιστικού Μοντέλου στα διαθέσιμα πειραματικά δεδομένα, έχει να κάνει με το πόσο κοντά βρίσκονται οι εκτιμήσεις του μοντέλου, αναφορικά με τις τιμές της μεταβλητής Απόκρισης, σε σχέση με τις αντίστοιχες παρατηρηθείσες δειγματικές τιμές. Με διαφορετικά αλλά ισοδύναμα λόγια, πόσο αξιόπιστα περιγράφεται η μεταβλητή απόκρισης από το μοντέλο, δεδομένων των πειραματικών τιμών των ερμηνευτικών μεταβλητών. Ο συνήθης έλεγχος υποθέσεων ο οποίος πραγματοποιείται στη περίπτωση αυτή, είναι ο ακόλουθος:

H_0: *Το μοντέλο είναι επαρκές (δε διαφέρει στατιστικώς σημαντικά από το κορεσμένο)*

H_1: *Το μοντέλο δεν είναι επαρκές*

Η αδυναμία απόρριψης της παραπάνω μηδενικής υπόθεσης, μας οδηγεί στην αναγκαστική κατάσταση αποδοχής της, γεγονός το οποίο σε καμία περίπτωση δε πρέπει να εκληφθεί ως τεκμήριο απόλυτης και άρτιας προσαρμογής του μοντέλου, σε επίπεδο πληθυσμού. Γνωρίζουμε εκ των προτέρων πως κανένα μοντέλο δεν είναι απολύτως αξιόπιστο. Απλά ένα στατιστικώς επαρκές μοντέλο, περιγράφει σχετικά αξιόπιστα τη μεταβλητή απόκρισης (σε δεδομένο επίπεδο σημαντικότητας του ελέγχου), υπό της έννοιας πως, οι προσαρμοσμένες από το μοντέλο τιμές της αποκρίσεως, προσεγγίζουν σε ικανοποιητικό βαθμό τις αντίστοιχες δειγματικά παρατηρηθείσες τιμές της μεταβλητής απόκρισης (Maxwell , 2009). Επιπλέον των παραπάνω, ενδέχεται να υφίσταται η περίπτωση κατά την οποία, περισσότερα του ενός μοντέλα, μετά της προσαρμογής αυτών στα πειραματικά δεδομένα, να κρίνονται ως στατιστικώς επαρκή με τη παραπάνω λογική. Στη περίπτωση αυτή, θα πρέπει να δρομολογηθεί ειδικός έλεγχος, ο οποίος να επιλέγει το βέλτιστο μοντέλο μεταξύ του συνόλου των επιλεχθέντων. Σε κάθε περίπτωση, η διαδικασία ελέγχου επάρκειας εκάστου μοντέλου, πρέπει να πραγματοποιηθεί μετά την ολοκλήρωση της διαδικασίας προσαρμογής του στα διαθέσιμα πειραματικά δεδομένα (Hosmer & Lemeshow, 2000).

4.1 Στατιστικοί έλεγχοι συνολικής Επάρκειας Μοντέλου

Τα βασικότερα εργαλεία ελέγχου της συνολικής επάρκειας ενός μοντέλου, είναι τα ακόλουθα:

- ✓ Έλεγχος καλής προσαρμογής χ^2 του *Pearson*
  (*χ^2 Goodness of fit test* ή εναλλακτικά το *G^2 Homogeneous test*)
- ✓ Έλεγχος Απόκλισης (*Deviance*)
- ✓ Έλεγχος των *Hosmer & Lemeshow* για δίτιμα δεδομένα
- ✓ Πίνακες κατάταξης ή συνάφειας *(Misclassification / Confusion / Contingency Tables)*
- ✓ Λειτουργικές Καμπύλες *ROC* (*Received Operating Characteristic Curves*)
- ✓ Συντελεστής Προσδιορισμού R^2

Αναλυτικότερα έχουμε,

α. Έλεγχος καλής προσαρμογής μέσω των στατιστικών συναρτήσεων χ2 και G^2

Η στατιστική συνάρτηση χ^2 του *Pearson*, ορίζεται ως ακολούθως

$$\chi^2 = \sum_{i=1}^{m} \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i}$$

y_i: οι παρατηρούμενες τιμές της i ομάδας των δίτιμων δεδομένων.

$\hat{y}_i$: οι εκτιμώμενες τιμές της i ομάδας των δίτιμων δεδομένων που προκύπτουν από τον τύπο

$$\hat{y}_i = n_i \hat{\pi}_i \, , i = 1, \dots, m$$

Προκειμένου για ομαδοποιημένα δίτιμα δεδομένα και κάτω από την ισχύ της μηδενικής υπόθεσης περί επάρκειας του μοντέλου, η στατιστική συνάρτηση χ^2 του *Pearson* ακολουθεί X^2 κατανομή πιθανότητας, με *m-p* βαθμούς ελευθερίας, όπου *m* το πλήθος των ομάδων και *p* το πλήθος των εκτιμώμενων παραμέτρων του μοντέλου, της σταθεράς συμπεριλαμβανομένης.

Στα ίδιο πλαίσιο στατιστικού ελέγχου της επαρκούς προσαρμογής του μοντέλου στα πειραματικά δεδομένα, μπορεί κανείς να πράξει τον ακόλουθο έλεγχο Ομοιογένειας (*Homogeneity*), μέσω χρήσης της στατιστικής συνάρτησης

$$G^2 = 2 \sum_{i=1}^{m} \left[y_i \ \frac{\log(y_i)}{\hat{y}_i} \right]$$

Η στατιστική συνάρτηση G^2, κάτω από την ισχύ της μηδενικής υπόθεσης περί Ομοιογένειας, ακολουθεί χ^2 κατανομή πιθανότητας, με *m-p* βαθμούς ελευθερίας. Ουσιαστικά τα στατιστικά χ^2 και G^2 είναι μεταξύ των ασυμπτωτικά ισοδύναμα, με μόνη διαφορά πως το χ^2 στατιστικό του *Pearson*, συγκλίνει στη X^2_{m-p} κατανομή, γρηγορότερα από το στατιστικό G^2. Απαραίτητη προϋπόθεση για να ισχύουν τα παραπάνω, είναι στο σχετικό πίνακα συνάφειας των κατηγορικών δεδομένων, ένα ποσοστό το πολύ 20% των αναμενόμενων συχνοτήτων, να είναι μικρότερο του 5.

β. Απόκλιση Μοντέλου

Μετά την προσαρμογή του μοντέλου στα πειραματικά δεδομένα, ο σημαντικότερος δείκτης σύγκρισης των εκτιμώμενων τιμών της εξαρτημένης μεταβλητής, σε σχέση με τις αντίστοιχες παρατηρηθείσες τιμές, είναι το στατιστικό της απόκλισης. Αξίζει δε να σημειωθεί πως, ασυμπτωτικά η στατιστική συνάρτηση χ^2 του *Pearson*, συγκλίνει στην απόκλιση του μοντέλου. Το στατιστικό της απόκλισης του προσαρμοσμένου μοντέλου, ορίζεται ως ακολούθως

Deviance (model) = - 2 log [Likelihood (model)] =

$$= -2 \sum_{i=1}^{m} \left[y_i \ log\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i) \ log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right) \right]$$

Ένα μοντέλο το οποίο κρίνεται ως συνολικά επαρκές, με καλή προσαρμογή στα πειραματικά δεδομένα, θα πρέπει να εμφανίζει τιμή πιθανοφάνειας ίση με ένα και άρα λογάριθμο πιθανοφάνειας, δηλαδή τιμή αποκλίσεως, ίση με το μηδέν. Τηρουμένων των αναλογιών, η ποσότητα εντός της αγκύλης του παραπάνω αθροίσματος, είναι το ανάλογο των τετραγωνικών σφαλμάτων στην πολλαπλή γραμμική παλινδρόμηση και η απόκλιση το ανάλογο του αθροίσματος των τετραγωνικών σφαλμάτων (*SSE: Sum of Squared Errors*).

Επομένως, ένα επαρκές μοντέλο θα πρέπει να εμφανίζει μικρή τιμή απόκλισης και καθώς το μέγεθος του δείγματος αυξάνει, κάτω από την ισχύ της μηδενικής υπόθεσης, η απόκλιση θα ακολουθεί ασυμπτωτικά κατανομή χ^2 με $m - p$ βαθμούς ελευθερίας. Η κρίσιμη περιοχή απόρριψης της μηδενικής υποθέσεως, σε δεδομένο επίπεδο σημαντικότητας α, είναι η ακόλουθη

$$D > \chi^2_{(m-p)}(\alpha)$$

όπου

$\chi^2_{m-p}(\alpha)$ το α ποσοστιαίο σημείο της κατανομής χ^2

Να σημειώσουμε πως τα παραπάνω ισχύουν μόνο στην περίπτωση κατά την οποία, τα δίτιμα δεδομένα, έχουν ομαδοποιηθεί και άρα έχουν προκύψει δυωνυμικά δεδομένα. Επιπλέον, απαιτείται μεγάλο μέγεθος δείγματος για κάθε ομάδα $i,\ n_i \to \infty,\ i = 1, \dots, m.$

Στην ειδική περίπτωση των μη ομαδοποιημένων δίτιμων δεδομένων, η σύγκλιση της απόκλισης δεν είναι εφικτή ούτε ασυμπτωτικά και θα πρέπει να υιοθετήσουμε μία από τις ακόλουθες μεθόδους ελέγχου επάρκειας και καλής προσαρμογής του μοντέλου.

γ .Έλεγχος Επάρκειας των Hosmer & Lemeshow

Στην περίπτωση των μη ομαδοποιημένων δίτιμων δεδομένων, ο στατιστικός έλεγχος των *Hosmer – Lemeshow,* ενδείκνυται για την πραγματοποίηση του κάτωθι ελέγχου υποθέσεων

$$\boldsymbol{H_0}: \pi_i = \hat{\pi}_i \ \ vs.\ \boldsymbol{H_1}: \pi_i \neq \hat{\pi}_i\ , i = 1, \dots, n$$

Κατόπιν της προσαρμογής του μοντέλου, οι παρατηρήσεις ταξινομούνται με βάση τις εκτιμώμενες πιθανότητες $\hat{\pi}_i$ και εν συνεχεία δημιουργούμε g ομάδες, ει δυνατόν πιο ισορροπημένες (*balanced*), σε επίπεδο πληθικού αριθμού εκάστης ομάδας.

Κατόπιν διαμορφώνουμε έναν *g x 2* πίνακα συνάφειας, με g γραμμές και δύο στήλες (επιτυχία *vs.* αποτυχία), καταγράφοντας συγκεντρωτικά το συνολικό πλήθος των παρατηρούμενων επιτυχιών ανά ομάδα. Το πλήθος των παρατηρούμενων επιτυχιών *S*, στο σύνολο των ομάδων του πίνακα συνάφειας, κάτω από την ισχύ της μηδενικής υποθέσεως, ακολουθεί κατανομή χ^2 με *g-2* βαθμούς ελευθερίας (Hosmer and Lemeshow, 2013). Η κρίσιμη περιοχή απόρριψης της μηδενικής υπόθεσης περί επάρκειας, σε δεδομένο επίπεδο σημαντικότητας α, είναι η ακόλουθη

$$S > \chi^2_{(g-2)}(\alpha)$$

όπου

$\chi^2_{g-2}(\alpha)$ το α ποσοστιαίο σημείο της κατανομής χ^2

δ. Πίνακες Κατάταξης ή Συνάφειας

Ένας πίνακας συνάφειας αποτυπώνει τις συχνότητες (*frequencies / counts*), για όλους τους δυνατούς συνδυασμούς επιπέδων, μεταξύ δύο ή και περισσότερων κατηγορικών μεταβλητών. Στην ειδική περίπτωση κατά την οποία χρησιμοποιούνται δύο κατηγορικές μεταβλητές, δύο επιπέδων η κάθε μία, στις γραμμές του πίνακα παριστούμε τις αναμενόμενες συχνότητες, ανά επίπεδο κατηγορικής μεταβλητής (*expected counts*) και στις στήλες τις παρατηρούμενες συχνότητες (observed *counts*), ανά επίπεδο κατηγορικής μεταβλητής.

Ως εκ τούτου, η μορφή ενός πίνακα συνάφειας διπλής εισόδου (δύο κατηγορικές μεταβλητές), με δύο επίπεδα ανά μεταβλητή (επιτυχία ή αποτυχία), είναι η ακόλουθη:

Πίνακας Συνάφειας (*Classification Table*)		**Παρατηρούμενες Συχνότητες (*Observed Counts*)**		**Σύνολα (*Totals*)**
		Επιτυχία (*Positive*)	**Αποτυχία (*Negative*)**	
Αναμενόμενες Συχνότητες (*Expected Counts*)	**Επιτυχία (*Positive*)**	*a (TP)*	*b (FP)*	*a + b*
	Αποτυχία (*Negative*)	*c (FN)*	*d (TN)*	*c + d*
	Σύνολα (*Totals*)	*a + c*	*b + d*	*a+ b + c + d*

Με βάση της σημειούμενες συχνότητες του πίνακα συνάφειας, ορίζουμε τα ακόλουθα μεγέθη (Cizek & Fitzgerald, 1999, Fraas & Newman, 2003).

Ευαισθησία (sensitivity ratio) $E = \frac{a}{(a+c)}$

Ειδικότητα (specificity ratio) $E\Delta = \frac{d}{(b+d)}$

TP (True Positive): Πλήθος σωστών θετικών ταξινομήσεων (*# a*)

TN (True Negative): Πλήθος σωστών αρνητικών ταξινομήσεων (*# d*)

FP (False Positive): Πλήθος εσφαλμένων θετικών ταξινομήσεων (*# b*)

FN (False Negative): Πλήθος εσφαλμένων αρνητικών ταξινομήσεων (*# c*)

Σκοπός της μοντελοποίησης θα πρέπει να είναι η μεγιστοποίηση των σωστών ταξινομήσεων (*TP, TN*) και συνάμα η ελαχιστοποίηση των εσφαλμένων ταξινομήσεων (*FP, FN*). Ισοδύναμα, η βασική επιδίωξη κατά τη διαδικασία μοντελοποίησης θα πρέπει να είναι, η μεγιστοποίηση και των δύο προαναφερθέντων στατιστικών, δηλαδή του λόγου ευαισθησίας και του λόγου ειδικότητας, ώστε αμφότερα να συγκλίνουν στη μονάδα. Πρακτικά, η παραπάνω επιδίωξη κρίνεται ως ανέφικτη, διότι η μεγιστοποίηση του ενός στατιστικού, οδηγεί σε ελαχιστοποίηση του άλλου. Μεταξύ των δύο παραπάνω στατιστικών, συνήθως επιλέγεται η μεγιστοποίηση του λόγου ευαισθησίας, έναντι της μεγιστοποίησης του λόγου ειδικότητας, διότι το κόστος μίας εσφαλμένα αρνητικής ταξινόμησης (*FN*), είναι δραματικά μεγαλύτερο και σημαντικότερο, του κόστους μίας εσφαλμένα θετικής ταξινόμησης (*FP*).

Με τρόπο συνήθως εμπειρικό, καθοριζόμενο από την πρακτική ή τις ερευνητικές επιδιώξεις του αναλυτή, καθορίζεται το διαχωριστικό σημείο αποκοπής (*cut off point / threshold*), έστω π_0. Σε περίπτωση κατά την οποία η εκτιμώμενη τιμή $\hat{\pi}_i$, υπερβαίνει το τιθέμενο διαχωριστικό σημείο, τότε η πειραματική μονάδα *i*, κατατάσσεται στις επιτυχίες, άλλως στις αποτυχίες.

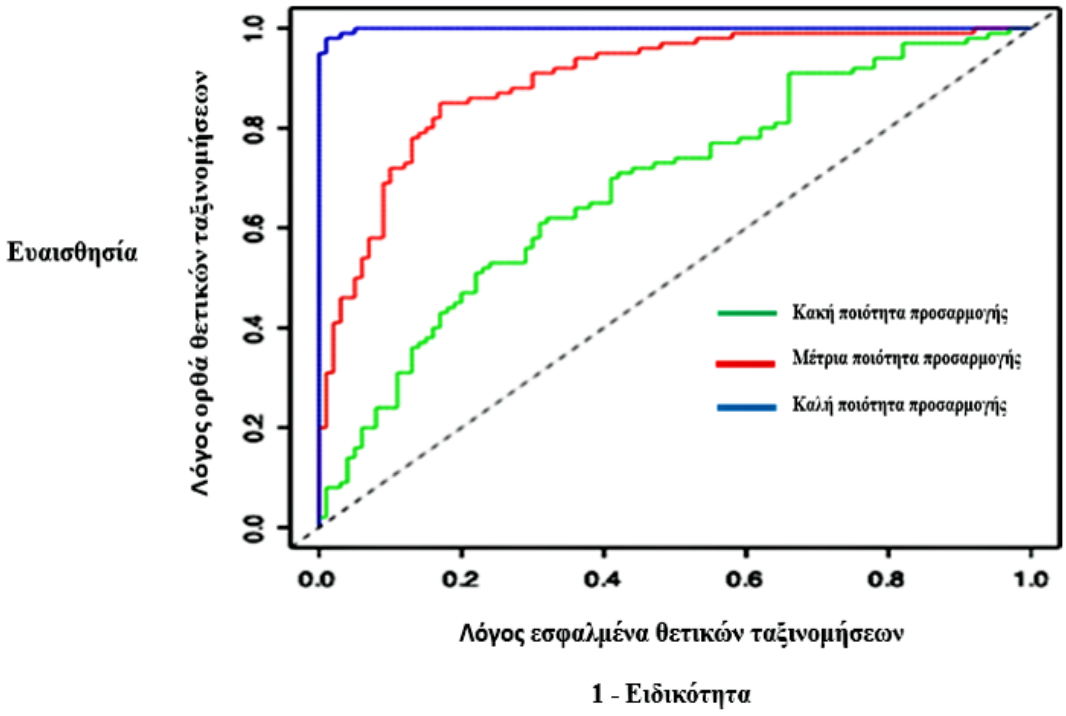
$$\hat{Y}_i = \begin{cases} 1\ (\text{Επιτυχία}), & \hat{\pi}_i > \pi_0 \\ 0\ (\text{Αποτυχία}), & \hat{\pi}_i \leq \pi_0 \end{cases}, \forall\, i = 1, \dots, n$$

Το διαχωριστικό σημείο αποκοπής θα πρέπει να καθορίζεται κατά τέτοιον τρόπο, ώστε να μεγιστοποιείται ο λόγος ευαισθησίας του ελέγχου. Πρακτικά, μία αποδεκτή τιμή για το λόγο ευαισθησίας είναι το *0.7 – 0.8*, με τιμή για το λόγο ειδικότητας του ελέγχου να διαμορφώνεται ιδεατά γύρω στο *0.6 – 0.7*

ε. Λειτουργικές Καμπύλες ROC

Η κατασκευή της γραφικής παράστασης του 'λόγου ευαισθησίας' (*TPR: True Positive Rate* στον άξονα των y), ενάντια στο '1- λόγος ειδικότητας' (*FPR: False Positive Rate* στον άξονα των x), διαμορφώνει τη λειτουργική χαρακτηριστική καμπύλη του προσαρμοσμένου μοντέλου. Επί της κύριας διαγώνιου, η πιθανότητα ισούται πάντοτε με 50%, γεγονός το οποίο σημαίνει πως το μοντέλο δεν εμφανίζει καμία προβλεπτική ικανότητα.

Κατά τη μοντελοποίηση, θα πρέπει να επιδιώκεται η μετακίνηση της κύριας διαγώνιου προς την άνω και αριστερή γωνία του τετραγώνου, συνολικού εμβαδού ίσου με 1, ώστε να μεγιστοποιείται το εμβαδό της περιοχής κάτω της λειτουργικής χαρακτηριστικής καμπύλης (*AUC: Area under Curve*). Η διαμόρφωση του εμβαδού κάτω της λειτουργικής χαρακτηριστικής καμπύλης, σε επίπεδα ανώτερα του *0.7 – 0.8*, αποτελεί ένδειξη καλής προσαρμογής του μοντέλου στα πειραματικά δεδομένα. Στο ακόλουθο σχήμα, αποτυπώνονται ενδεικτικά τρείς διαφορετικές λειτουργικές καμπύλες, αναλόγως της ποιότητας προσαρμογής εκάστου μοντέλου στα πειραματικά δεδομένα.



στ. Συντελεστής προσδιορισμού R^2

Ο συντελεστής προσδιορισμού R^2 (*Coefficient of Determination*), σε ένα μοντέλο κλασσικής γραμμικής παλινδρόμησης, παριστά το ποσοστό της συνολικής μεταβλητότητας των πειραματικών δεδομένων, που επεξηγείται από το μοντέλο, μετά τη προσαρμογή του στα πειραματικά δεδομένα και δίνεται από τον τύπο

$$R^2 = \frac{SSR}{SST}$$

όπου, *SSR* είναι το 'άθροισμα τετραγώνων της παλινδρόμησης' (*Sum of Squares Regression*) και *SST* το 'συνολικό άθροισμα τετραγώνων' (*Total Sum of Squares*). Όταν ο συγκεκριμένος συντελεστής μεγιστοποιείται, λαμβάνοντας τιμή κοντά στη μονάδα, αυτό αποτελεί ένδειξη καλής προσαρμογής του μοντέλου της κλασσικής παλινδρόμησης στα πειραματικά δεδομένα.

Στην περίπτωση των μοντέλων λογιστικής παλινδρόμησης, η προαναφερθείσα έννοια του συντελεστή προσδιορισμού επεκτείνεται, μέσω των αντίστοιχων συντελεστών προσδιορισμού κατά *Cox & Snell (Uncorrected R^2), Nagelkerke (Corrected R^2)* και *Mc Fadden' s (Pseudo R^2)*.

(I) Συντελεστής προσδιορισμού R^2 κατά Mc Fadden

Κατόπιν της προσαρμογής του μοντέλου στα πειραματικά δεδομένα και μεγιστοποίησης της αντιστοίχου πιθανοφάνειας, ο συντελεστής προσδιορισμού *Mc Fadden's R^2 (pseudo)* (McFadden, 1974), ορίζεται ως ακολούθως

$$R^2_{McF} = 1 - \frac{log(L_M)}{log(L_0)}$$

όπου, $log(L_M)$ είναι ο φυσικός λογάριθμος της μεγιστοποιημένης πιθανοφάνειας, μετά την προσαρμογή του μοντέλου Μ, ενώ $log(L_0)$ είναι ο φυσικός λογάριθμος της πιθανοφάνειας μοντέλου που περιέχει μόνον το σταθερό όρο. Ο λόγος των παραπάνω λογαρίθμων, αποτελεί το ανάλογο του αθροίσματος των τετραγωνικών σφαλμάτων (*SSE)* της κλασσικής παλινδρόμησης, επομένως σε ένα μοντέλο καλώς προσαρμοσμένο στα πειραματικά δεδομένα, θα πρέπει να συγκλίνει στο μηδέν και άρα ο *pseudo* συντελεστής προσδιορισμού, θα πρέπει να συγκλίνει στη μονάδα.

(II) Συντελεστής προσδιορισμού R^2 κατά Cox & Snell

Ο συντελεστής προσδιορισμού R^2 *(Uncorrected)* των Cox & Snell (Cox & Snell, 1989), ορίζεται ως ακολούθως

$$R^2_{C\&S} = 1 - (\frac{L_0}{L_M})^{\frac{2}{n}}$$

όπου, *n* το μέγεθος του δείγματος.

Ο εν λόγω συντελεστής προσδιορισμού, αποτελεί τη γενίκευση του κλασσικού συντελεστή προσδιορισμού R^2, με τη μόνη διαφορά πως παρουσιάζει ανώτατο όριο αρκετά μικρότερο της μονάδας. Πιο συγκεκριμένα, το ανώτατο όριο του γενικευμένου συντελεστή προσδιορισμού των *Cox & Snell* είναι το εξής

$$max(R^2_{C\&S}) = \max_{L_M}[\, 1 - (\frac{L_0}{L_M})^{\frac{2}{n}}] \; = [1 - (L_0)^{\frac{2}{n}}] < 1$$

Στο σημείο αυτό, αξίζει να σημειωθούν οι ακόλουθες παρατηρήσεις

- ✓ Όταν το οριακό ποσοστό πραγματοποίησης του επιτυχούς ενδεχόμενου, ενάντια στο πλήθος των παρατηρήσεων, είναι κοντά στο *0.5*, τότε το άνω όριο του συντελεστή των *Cox & Snell* διαμορφώνεται στο *0.75*
- ✓ Όταν το προαναφερθέν οριακό ποσοστό, διαμορφώνεται κοντά στο *0.1* ή *0.9*, τότε το ανώτατο όριο του συντελεστή διαμορφώνεται στο *0.48*

Προκειμένου να διορθωθεί αυτή η αδυναμία του γενικευμένου συντελεστή προσδιορισμού των *Cox & Snell*, έχει προταθεί η διορθωμένη έκδοση κατά *Nagelkerke* (1991), η οποία προκύπτει κατόπιν της διαιρέσεως του μη διορθωμένου συντελεστή, με το προαναφερθέν άνω όριο του.

Ο συντελεστής προσδιορισμού κατά *Mc Fadden* ικανοποιεί (Menard, 2000) και τα οκτώ κριτήρια (Kvalseth's, 1985), τα οποία πρέπει να πληροί ένας αξιόπιστος συντελεστής προσδιορισμού. Επιπλέον, όταν το οριακό ποσοστό πραγματοποίησης του επιτυχούς ενδεχόμενου, έναντι του πλήθους των παρατηρήσεων, είναι κοντά στο 0.5, τότε ο συντελεστής κατά *Mc Fadden*, είναι οριακά μικρότερος του αντιστοίχου κατά *Cox & Snell*. Επιπλέον, όταν το οριακό ποσοστό πραγματοποίησης του επιτυχούς ενδεχόμενου, είναι κοντά στο 0.1 ή στο 0.9, τότε ο συντελεστής κατά *Mc Fadden* υπερβαίνει οριακά εκείνον των *Cox & Snell*.

4.2 Στατιστικός έλεγχος σημαντικότητας ανεξάρτητων μεταβλητών

Προκειμένου να ελεγχθεί η στατιστική σημαντικότητα των ερμηνευτικών μεταβλητών ενός μοντέλου λογιστικής παλινδρόμησης, ο κυριότερος στατιστικός έλεγχος ο οποίος χρησιμοποιείται στην πράξη, είναι ο έλεγχος του *Wald* (Lea,1997, Garson,2006). Ο συγκεκριμένος στατιστικός έλεγχος, διαμορφώνεται ως ακολούθως (Cizek & Fitzgerald, 1999)

H₀: *Ο συντελεστής της μεταβλητής* X_j *δεν είναι στατιστικά σημαντικός* ($b_j = 0$)

H₁: *Ο συντελεστής της μεταβλητής* X_j *είναι στατιστικά σημαντικός* ($b_j \neq 0$)

Η στατιστική συνάρτηση του ελέγχου, είναι η ακόλουθη:

$$W_j = \frac{\widehat{b_j}}{S.E.(\hat{b}_j)} \ , \ \ j = 1, \dots, p$$

όπου, *p* το πλήθος των ερμηνευτικών μεταβλητών του μοντέλου, συμπεριλαμβανομένου του σταθερού όρου.

Το στατιστικό του *Wald*, κάτω από την ισχύ της μηδενικής υποθέσεως, ακολουθεί Τυπική Κανονική Κατανομή *N(0,1)*, όπου η κρίσιμη περιοχή απόρριψης της μηδενικής υποθέσεως (Scott ,2002), για δεδομένο επίπεδο σημαντικότητας *α*, ορίζεται ως εξής

$$W_j > Z_{\frac{\alpha}{2}}, \quad j = 1, \dots, p$$

όπου, $z_{\alpha/2}$ είναι το άνω *α/2* ποσοστιαίο σημείο της Τυπικής Κανονικής Κατανομής *N(0,1)*.

4.3 Συγκριτικός έλεγχος επαρκών μοντέλων

Σε περίπτωση ύπαρξης περισσότερων από δύο, συνολικά επαρκών μοντέλων λογιστικής παλινδρόμησης, οι μεθοδολογίες που ευρέως χρησιμοποιούνται προκειμένου να επιλεχθεί το καταλληλότερο εξ΄ αυτών, εις όρους επάρκειας και καλής προσαρμογής, είναι οι ακόλουθες

α. Πληροφοριακό κριτήριο AIC

Το πληροφοριακό κριτήριο επιλογής *AIC (Akaike Information Criterion)*, ορίζεται ως ακολούθως

$$AIC(M) = 2p - log[L(M)]$$

όπου, *p* είναι το πλήθος των ερμηνευτικών μεταβλητών του μοντέλου M και L η μεγιστοποιημένη πιθανοφάνεια του, μετά της προσαρμογής στα πειραματικά δεδομένα. Σε κάθε περίπτωση, θα πρέπει να επιλέγεται ως επικρατέστερο εκείνο το μοντέλο, που εμφανίζει τη μικρότερη τιμή στο εν λόγω κριτήριο, ώστε να περιορίζεται κατά το δυνατόν περισσότερο, το φαινόμενο της υπερπροσαρμογής.

Επομένως, μεταξύ δύο επαρκών μοντέλων με τον ίδιο αριθμό ερμηνευτικών μεταβλητών *p*, θα πρέπει να επιλεχθεί ως επικρατέστερο εκείνο το μοντέλο, με τη μεγαλύτερη πιθανοφάνεια. Αντιστοίχως, μεταξύ δύο μοντέλων με την ίδια πιθανοφάνεια, θα πρέπει να επιλεγεί εκείνο με το μικρότερο αριθμό ερμηνευτικών μεταβλητών.

β. Συντελεστής μερικής Απόκλισης

Έστω δύο μοντέλα λογιστικής παλινδρόμησης, το περιορισμένο μοντέλο M1 (*restricted*), με *df1* βαθμούς ελευθερίας και το εκτενές μοντέλο M2 (*extended*), με *df2* βαθμούς ελευθερίας. Το σύνολο των ερμηνευτικών μεταβλητών του μοντέλου M1, είναι γνήσιο υποσύνολο του συνόλου των ερμηνευτικών μεταβλητών του μοντέλου M2, γεγονός το οποίο εξασφαλίζει πως (*df1 > df2*).

Προκειμένου να ελεγχθεί η στατιστική σημαντικότητα εκείνων των ερμηνευτικών μεταβλητών του μοντέλου M2, οι οποίες απουσιάζουν από το μοντέλο M1, εις όρους συνεισφοράς στην επάρκεια και την καλή προσαρμογή στα πειραματικά δεδομένα, πραγματοποιείται ο κάτωθι έλεγχος υποθέσεων

H_0: *Τα μοντέλα M1 & M2 δε διαφέρουν στατιστικώς σημαντικά ως προς την επάρκεια*
H_1: *Το εκτενές μοντέλο M2 είναι επαρκέστερο του περιορισμένου μοντέλου M1*

Η στατιστική συνάρτηση του ελέγχου είναι η ακόλουθη

$$PD = -2\ log\left[\frac{L(M1)}{L(M2)}\right] = -2\left[log\big(L(M1)\big) - log\big(L(M2)\big)\right]$$

Το παραπάνω στατιστικό *PD (Partial Deviance),* κάτω από την ισχύ της μηδενικής υποθέσεως, ακολουθεί χ^2 κατανομή με $(df_1 - df_2)$ βαθμούς ελευθερίας, η οποία διαφορά βαθμών ελευθερίας, στην ουσία είναι ίση με το πλεονάζον πλήθος μεταβλητών του εκτεταμένου μοντέλου M2 (Babtain & Taha , 2009). Για δοθέν επίπεδο σημαντικότητας a, η κρίσιμη περιοχή απόρριψης της μηδενικής υποθέσεως είναι

$$PD > \chi^2_{(df1-df2)}(1-a)$$

4.4 Μεθοδολογίες επικύρωσης Μοντέλων Διαχωριστικής Ανάλυσης

Προκειμένου να αξιολογηθεί αμερόληπτα η ακρίβεια ταξινόμησης ενός διαχωριστικού μοντέλου, απαιτείται να προσδιοριστεί το ποσοστό των εσφαλμένων ταξινομήσεων, δηλαδή το 'Φαινομενικό σφάλμα ταξινόμησης' (*APER: Apparent Error Rate*), το οποίο προκύπτει κατά την εφαρμογή του διαχωριστικού μοντέλου, σε ένα σύνολο πειραματικών δεδομένων, τα οποία ο αλγόριθμος δεν είχε διαχειριστεί κατά τη φάση εκπαίδευσης του.

Στην περίπτωση κατά την οποία, ο προσδιορισμός της ακρίβειας ταξινόμησης του διαχωριστικού μοντέλου, πραγματοποιηθεί μέσω αξιοποίησης των δεδομένων εκπαίδευσης, δηλαδή των δεδομένων επί των οποίων εκπαιδεύτηκε ο αλγόριθμος και εκτιμήθηκαν οι παράμετροι του διαχωριστικού μοντέλου, τότε υπερεκτιμούμε τη δυνατότητα ορθής ταξινόμησης και τελικώς προκύπτει σφάλμα ταξινόμησης μικρότερο του πραγματικού.

α. Cross Validation Method

Κατά την εφαρμογή της μεθόδου *Cross Validation*, το αρχικό σύνολο των δειγματικών παρατηρήσεων διακρίνεται σε δύο τμήματα. Το πρώτο τμήμα ονομάζεται 'Σύνολο δεδομένων εκπαίδευσης' (*Training Data Set*), μέσω του οποίου εκτιμάται η διαχωριστική συνάρτηση του μοντέλου και το δεύτερο τμήμα ονομάζεται 'Σύνολο δεδομένων επικύρωσης' (*Test Data Set*), το οποίο χρησιμοποιείται ώστε να προσδιοριστεί το ποσοστό των εσφαλμένων ταξινομήσεων. Αναφορικά με τα μεγέθη των δύο προαναφερθέντων συνόλων δεδομένων, ο πρακτικός κανόνας ο οποίος συνήθως χρησιμοποιείται στην πράξη υποδεικνύει, διάκριση του αρχικού συνόλου των πειραματικών δεδομένων σε δύο επιμέρους τμήματα, με αναλογίες *80:20* ή *70:30*, πάντοτε υπέρ του συνόλου των δεδομένων εκπαίδευσης.

Επιπλέον των παραπάνω, αξίζει να σημειωθεί πως το φαινομενικό σφάλμα ταξινόμησης που προκύπτει κατά την εφαρμογή του μοντέλου επί του συνόλου των δεδομένων εκπαίδευσης, σε κάθε περίπτωση αναμένεται μικρότερο ή ίσο, του αντίστοιχου σφάλματος που προκύπτει, κατά την εφαρμογή του μοντέλου επί του συνόλου των δεδομένων επικύρωσης.

β. U method (Jack Knife method)

Στη συγκεκριμένη μέθοδο, το αρχικό σύνολο των πειραματικών δεδομένων, διακρίνεται με τυχαίο τρόπο σε *n* διακεκριμένα υποσύνολα, τα οποία πρέπει να είναι συγκρίσιμου μεγέθους και ισοζυγισμένα επί των επιπέδων της κατηγορικής μεταβλητής απόκρισης. Το εν λόγω σημαίνει πως, κάθε επίπεδο της εξαρτημένης κατηγορικής μεταβλητής, θα πρέπει να εμφανίζει συχνότητα εμφάνισης εντός εκάστου υποσυνόλου δεδομένων, συγκρίσιμη με τη συχνότητα εμφάνισης του αντίστοιχου επιπέδου, επί του αρχικού συνόλου δεδομένων.

Εν συνεχεία, με τυχαίο τρόπο επιλέγεται μία δειγματική παρατήρηση ανά υποσύνολο δεδομένων, η οποία εξαιρείται του αντιστοίχου υποσυνόλου και κατόπιν εκτιμάται η διαχωριστική συνάρτηση του μοντέλου ανά υποσύνολο, χρησιμοποιώντας τις εναπομείνασες παρατηρήσεις του υποσυνόλου αυτού. Καταληκτικά, για κάθε υποσύνολο δεδομένων εκπαίδευσης, ελέγχεται η ακρίβεια ορθής ταξινόμησης της εξαιρεθείσας δειγματικής παρατήρησης. Το συνολικό φαινομενικό σφάλμα ταξινόμησης του διαχωριστικού μοντέλου, προκύπτει ως ο αριθμητικός μέσος όρος των *n* επιμέρους εσφαλμένων ταξινομήσεων. (Johnson & Wichern , 2007).

5 Ενδεικτικά παραδείγματα εφαρμογής μεθόδων Διαχωριστικής Ανάλυσης

Ακολούθως δίδονται ενδεικτικά παραδείγματα εφαρμογής των διαχωριστικών αλγορίθμων *ID3* και *C_4.5*, καθώς και ενδεικτικό παράδειγμα εφαρμογής της μεθόδου Πολυωνυμικής Λογιστικής Παλινδρόμησης.

5.1 Παράδειγμα συγκριτικής εφαρμογής Αλγορίθμων *ID3* και *C4_5*

Κατά τη χρονική διάρκεια δύο ημερολογιακών εβδομάδων, συλλέχθηκαν τα ακόλουθα πρωτογενή πειραματικά δεδομένα, τα οποία θα χρησιμοποιηθούν ως τα απαραίτητα δεδομένα εκπαίδευσης (*training data set*), προκειμένου να κατασκευαστούν τα αντίστοιχα δέντρα απόφασης, μέσω εφαρμογής των αλγορίθμων που παρουσιάστηκαν στις προηγούμενες ενότητες.

Ο πίνακας των δεδομένων εκπαίδευσης, είναι ο ακόλουθος:

Day	**Outlook**	**Temperature**	**Humidity**	**Wind**	**Play**
D1	Sun	Hot	High	Low	No
D2	Sun	Hot	High	High	No
D3	Overcast	Hot	High	Low	Yes
D4	Rain	Sweet	High	Low	Yes
D5	Rain	Cold	Normal	Low	Yes
D6	Rain	Cold	Normal	High	No
D7	Overcast	Cold	Normal	High	Yes
D8	Sun	Sweet	High	Low	No
D9	Sun	Cold	Normal	Low	Yes
D10	Rain	Sweet	Normal	Low	Yes
D11	Sun	Sweet	Normal	High	Yes
D12	Overcast	Sweet	High	High	Yes
D13	Overcast	Hot	Normal	Low	Yes
D14	Rain	Sweet	High	High	No

Η μεταβλητή απόκρισης του προβλήματος είναι η μεταβλητή *Play*, η οποία είναι διχοτομική μεταβλητή, με δύο εναλλακτικές κατηγορίες, *Yes* και *No*, η οποία εκτιμά την πιθανότητα πραγματοποίησης ποδοσφαιρικού αγώνα, δεδομένων των κλιματολογικών συνθηκών που επικρατούν.

Οι ερμηνευτικές μεταβλητές είναι επίσης κατηγορικές, τα επίπεδα των οποίων είναι τα εξής

Outlook: Rain, Overcast, Sun

Temperature: Cold, Sweet, Hot

Humidity: Normal, High

Wind: Low, High

Η αλγοριθμική διαδικασία ξεκινά με τον προσδιορισμό, εκείνης της σημαντικότερης ερμηνευτικής μεταβλητής, η οποία θα αποτελέσει τη ρίζα (*root*) του δέντρου απόφασης. Σύμφωνα με την προαναφερθείσα μεθοδολογία, θα πρέπει να επιλεγεί εκείνο το χαρακτηριστικό της ανάλυσης, το οποίο μεγιστοποιεί τη συνάρτηση καταλληλότητας, του χρησιμοποιούμενου αλγόριθμου. Στην περίπτωση του αλγορίθμου *ID3*, η χρησιμοποιούμενη συνάρτηση καταλληλότητας είναι το 'Κέρδος Πληροφορίας' (*IG:Information Gain*).

Η συνολική εντροπία των δεδομένων εκπαίδευσης *D* είναι ίση με

$$Entropy\ (D) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14}) = 0.94$$

Η εντροπία των δεδομένων υπό του διαχωρισμού της μεταβλητής '*Outlook'* είναι ίση με

$$Entropy\ (Outlook) = \frac{5}{14}\ Entropy\ (Rain) + \frac{4}{14}\ Entropy\ (Overcast) + \frac{5}{14}\ Entropy\ (Sun)$$

$$Entropy\ (Rain) = -\frac{3}{5}\ log_2(\frac{3}{5}) - \frac{2}{5}\ log_2(\frac{2}{5}) = 0.971$$

$$Entropy\ (Overcast) = -\frac{4}{4}\ log_2\left(\frac{4}{4}\right) = 0$$

$$Entropy\ (Sun) = -\frac{2}{5}\ log_2\left(\frac{2}{5}\right) - \frac{3}{5}\ log_2(\frac{3}{5}) = 0.971$$

Επομένως έχουμε

$$Entropy\ (Outlook) = \frac{5}{14}\ 0.971 + \frac{4}{14}\ 0 + \frac{5}{14}\ 0.971 = 0.694$$

Καταληκτικά, το κέρδος πληροφορίας (*IG*), εάν ως ρίζα του δέντρου επιλεγεί το χαρακτηριστικό '*Outlook*', διαμορφώνεται ως ακολούθως

$$IG\ (Outllok) = Entropy\ (D) - Entropy\ (Outlook) = 0.94 - 0.694 = 0.246$$

Η εντροπία των δεδομένων υπό του διαχωρισμού της μεταβλητής '*Temperature'* είναι ίση με

$$Entropy\ (Temperature) = \frac{4}{14}\ Entropy\ (Cold) + \frac{6}{14}\ Entropy\ (Sweat) + \frac{4}{14}\ Entropy\ (Hot)$$

$$Entropy\ (Cold) = -\frac{3}{4}\ log_2(\frac{3}{4}) - \frac{1}{4}\ log_2(\frac{1}{4}) = 0.811$$

$$Entropy\ (Sweat) = -\frac{4}{6}\ log_2\left(\frac{4}{6}\right) - \frac{2}{6}\ log_2(\frac{2}{6}) = 0.918$$

$$Entropy\ (Hot) = -\frac{2}{4}\ log_2\left(\frac{2}{4}\right) - \frac{2}{4}\ log_2(\frac{2}{4}) = 1$$

Επομένως έχουμε

$$Entropy\ (Temperature) = \frac{4}{14}\ 0.811 + \frac{6}{14}\ 0.918 + \frac{4}{14}\ 1 = 0.9111$$

Καταληκτικά, το κέρδος πληροφορίας (*IG*), εάν ως ρίζα του δέντρου επιλεγεί το χαρακτηριστικό '*Temperature*', διαμορφώνεται ως ακολούθως

$$IG\ (Temperature) = Entropy\ (D) - Entropy\ (Temperature) =$$

$$= 0.94 - 0.9111 = 0.0289$$

Η εντροπία των δεδομένων υπό του διαχωρισμού της μεταβλητής '*Humidity'* είναι ίση με

$$Entropy\ (Humidity) = \frac{7}{14}\ Entropy\ (Normal) + \frac{7}{14}\ Entropy\ (High)$$

$$Entropy\ (Normal) = -\frac{6}{7}\ log_2(\frac{6}{7}) - \frac{1}{7}\ log_2(\frac{1}{7}) = 0.59167$$

$$Entropy\ (High) = -\frac{3}{7}\ log_2\left(\frac{3}{7}\right) - \frac{4}{7}\ log_2(\frac{4}{7}) = 0.98523$$

Επομένως έχουμε

$$Entropy\ (Humidity) = \frac{7}{14}\ 0.59167 + \frac{7}{14}\ 0.98523 = 0.7885$$

Καταληκτικά, το κέρδος πληροφορίας (*IG*), εάν ως ρίζα του δέντρου απόφασης επιλεγεί το Χαρακτηριστικό '*Humidity*', διαμορφώνεται ως ακολούθως

$$IG\ (Humidity) = Entropy\ (D) - Entropy\ (Humidity) =$$

$$0.94 - 0.7885 = 0.1515$$

Η εντροπία των δεδομένων υπό του διαχωρισμού της μεταβλητής '*Wind'* είναι ίση με

$$Entropy\ (Wind) = \frac{8}{14}\ Entropy\ (Low) + \frac{6}{14}\ Entropy\ (High)$$

$$Entropy\ (Low) = -\frac{6}{8}\ log_2(\frac{6}{8}) - \frac{2}{8}\ log_2(\frac{2}{8}) = 0.5714$$

$$Entropy\ (High) = -\frac{3}{6}\ log_2\left(\frac{3}{6}\right) - \frac{3}{6}\ log_2(\frac{3}{6}) = 0.4286$$

Επομένως έχουμε

$$Entropy\ (Wind) = \frac{8}{14}\ 0.5714 + \frac{6}{14}\ 0.4286 = 0.8922$$

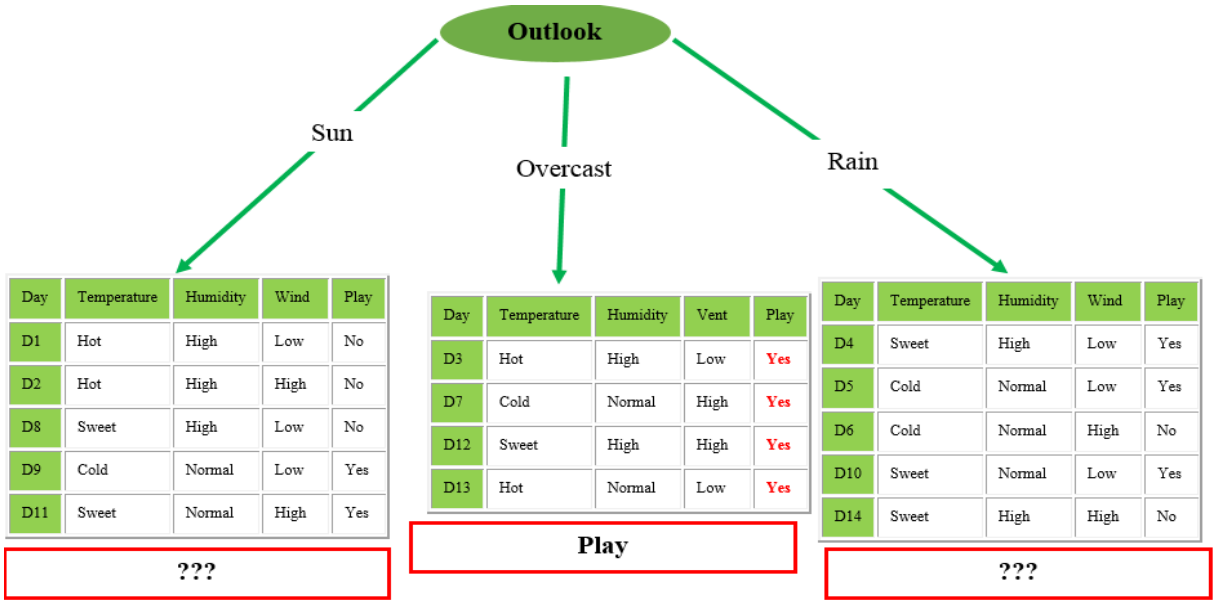
Καταληκτικά, το κέρδος πληροφορίας (*IG*), εάν ως ρίζα του δέντρου απόφασης επιλεγεί το Χαρακτηριστικό '*Wind*', διαμορφώνεται ως ακολούθως

$$IG(Wind) = Entropy\ (D) - Entropy\ (Wind) = 0.94 - 0.8922 = 0.0478$$

Συγκεντρωτικά τα αποτελέσματα έχουν ως εξής

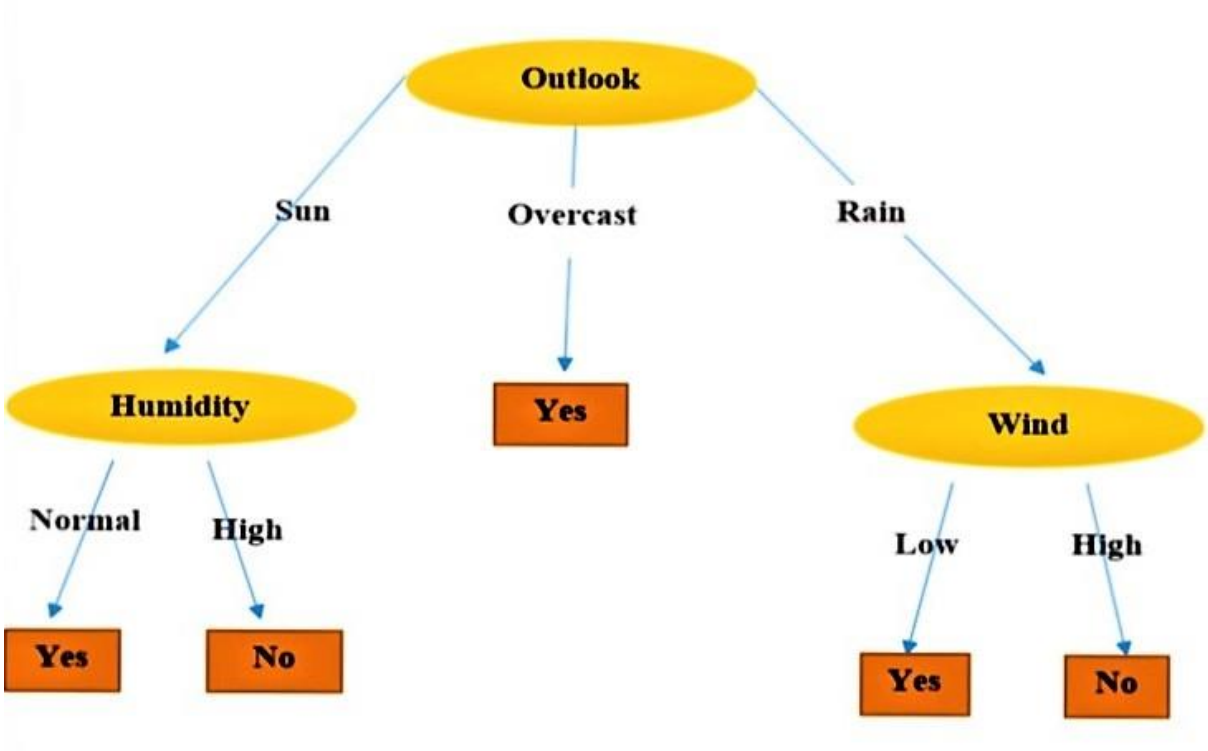
Χαρακτηριστικό Διάσπασης	Κέρδος Πληροφορίας Δεδομένων
Outlook	0.246
Humidity	0.1515
Wind	0.0478
Temperature	0.0289

Από τα παραπάνω αποτελέσματα καταλήγουμε πως, ως ρίζα (*root*) του δέντρου, θα πρέπει να επιλεγεί το χαρακτηριστικό '*Outlook*', αφού για αυτό έχει προκύψει η μεγαλύτερη τιμή του κέρδους πληροφορίας.



Η εφαρμογή της διαμέρισης του αρχικού συνόλου των δεδομένων εκπαίδευσης, με βάση τα επίπεδα της μεταβλητής '*Outlook'*, δημιουργεί τρία επιμέρους υποσύνολα δεδομένων εκπαίδευσης. Παρατηρούμε πως, στο υποσύνολο που αντιστοιχεί στο επίπεδο '*Overcast'*, η μεταβλητή απόκρισης '*Play*', λαμβάνει πάντοτε τη τιμή '*Yes',* με αποτέλεσμα να μην απαιτείται περαιτέρω διαμέριση. Αναφορικά όμως με τα υποσύνολα δεδομένων, των υπολοίπων δύο επιπέδων '*Sun*' και '*Rain*', παρατηρούμε πως η μεταβλητή απόκρισης '*Play*', λαμβάνει και τις δύο δυνατές τιμές '*Yes*' ή '*No*', σε συγκρίσιμα μεταξύ των ποσοστά. Κατά συνέπεια κρίνεται απαραίτητη η συνέχιση της διαδικασίας διαμερίσεως, αυτή τη φορά με αξιοποίηση των επιπέδων των χαρακτηριστικών '*Temperature*', '*Humidity*' και '*Wind*', έως ότου προκύψουν υποσύνολα δεδομένων εκπαίδευσης, με όσο το δυνατόν πιο ομοιογενή κατανομή τιμών για τη μεταβλητή απόκρισης '*Play'*.

Η συνέχιση εφαρμογής του αλγορίθμου *ID3*, στο υποσύνολο δεδομένων εκπαίδευσης που αντιστοιχεί στο επίπεδο '*Sun*', μεγιστοποιεί το κέρδος πληροφορίας, όταν η διαμέριση πραγματοποιηθεί επί των επιπέδων του χαρακτηριστικού '*Humidity*'. Προκειμένου για το υποσύνολο δεδομένων εκπαίδευσης το οποίο αντιστοιχεί στο επίπεδο '*Rain*', η μεγιστοποίηση του κέρδους πληροφορίας πραγματοποιείται με διάσπαση των δεδομένων, επί των επιπέδων του χαρακτηριστικού '*Wind*'. Η τελική δομή του προκύπτοντος δέντρου, είναι η ακόλουθη



Προκειμένου να αναπτύξουμε ενδεικτικό παράδειγμα εφαρμογής του αλγορίθμου *C_4.5*, θα χρησιμοποιήσουμε το ίδιο βασικό σύνολο δεδομένων εκπαίδευσης, με τη μόνη διαφορά πως θα θεωρήσουμε το χαρακτηριστικό '*Humidity*'ως συνεχή μεταβλητή, η οποία λαμβάνει πλέον συγκεκριμένες αριθμητικές τιμές. Ακολούθως έχουμε

Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sun	Hot	**85**	Low	No
D2	Sun	Hot	**90**	High	No
D3	Overcast	Hot	**78**	Low	Yes
D4	Rain	Sweet	**96**	Low	Yes
D5	Rain	Cold	**80**	Low	Yes
D6	Rain	Cold	**70**	High	No
D7	Overcast	Cold	**65**	High	Yes
D8	Sun	Sweet	**95**	Low	No
D9	Sun	Cold	**70**	Low	Yes
D10	Rain	Sweet	**80**	Low	Yes
D11	Sun	Sweet	**70**	High	Yes
D12	Overcast	Sweet	**90**	High	Yes
D13	Overcast	Hot	**75**	Low	Yes
D14	Rain	Sweet	**80**	High	No

Οι μοναδικές τιμές του χαρακτηριστικού '*Humidity*', σε αύξουσα τάξη μεγέθους, είναι οι εξής

{ 65, 70, 75, 78, 80, 85, 90, 95, 96 }

Με βάση τις παραπάνω τιμές, εφαρμόζουμε διαμέριση επί του αρχικού συνόλου δεδομένων εκπαίδευσης, λαμβάνοντας συγκεντρωτικά τα ακόλουθα αποτελέσματα

	65		70		75		78		80		85		90		95		96	
interval	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>
Yes	1	8	3	6	4	5	5	4	7	2	7	2	8	1	8	1	9	0
No	0	5	1	4	1	4	1	4	2	3	3	2	4	1	5	0	5	0
Entropy	0	0.961	0.811	0.971	0.721	0.991	0.65	1	0.764	0.971	0.881	1	0.918	1	0.961	0	0.94	0
Info	0.892		0.925		0.8950		0.85		0.838		0.915		0.929		0.892		0.94	
Gain	0.048		0.015		0.045		0.09		0.102		0.025		0.011		0.048		0	

Η διαμέριση με βάση τις τιμές του συνεχούς χαρακτηριστικού '*Humidity*', θα πρέπει να πραγματοποιηθεί στο επίπεδο '*80'*, διαμορφώνοντας τα ακόλουθα διαστήματα τιμών

$$(-\infty\,, \mathbf{80}], (\mathbf{80}\,, +\infty)$$

Πράγματι, η εντροπία των δεδομένων υπό του παραπάνω διαχωρισμού της μεταβλητής '*Humidity'*, είναι ίση με

$$Entropy\,(Humidity) = \frac{9}{14}\,Entropy\,(\leq 80) + \frac{5}{14}\,Entropy\,(> 80)$$

$$Entropy\,(\leq 80) = -\frac{7}{9}\,log_2(\frac{7}{9}) - \frac{2}{9}\,log_2(\frac{2}{9}) = 0.7642$$

$$Entropy\,(> 80) = -\frac{2}{5}\,log_2\left(\frac{2}{5}\right) - \frac{3}{5}\,log_2(\frac{3}{5}) = 0.971$$

Επομένως έχουμε

$$Entropy\,(Humidity) = = \frac{9}{14}\,0.7642 + \frac{5}{14}\,0.971 = 0.838$$

Καταληκτικά, το κέρδος πληροφορίας (*IG*), εάν ως ρίζα του δέντρου απόφασης επιλεγεί το Χαρακτηριστικό '*Humidity*', διαμορφώνεται ως ακολούθως

$$IG\,(Humidity) = Entropy\,(D) - Entropy\,(Humidity) =$$

$$= 0.94 - 0.838 = 0.1022$$

Η τιμή του λόγου κέρδους πληροφορίας (*IGR*), δηλαδή της συνάρτησης καταλληλότητας την οποία χρησιμοποιεί ο αλγόριθμος *C_4.5*, προκύπτει εάν το παραπάνω κέρδος πληροφορίας, διαιρεθεί με την ποσότητα

$$Entropy\left(\frac{|\,(-\infty, \mathbf{80}]\,|}{|D|}, \frac{|\,[\mathbf{80}, +\infty\,|}{|D|}\right) = Entropy\left(\frac{9}{14}, \frac{5}{14}\right) =$$

$$= -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.94$$

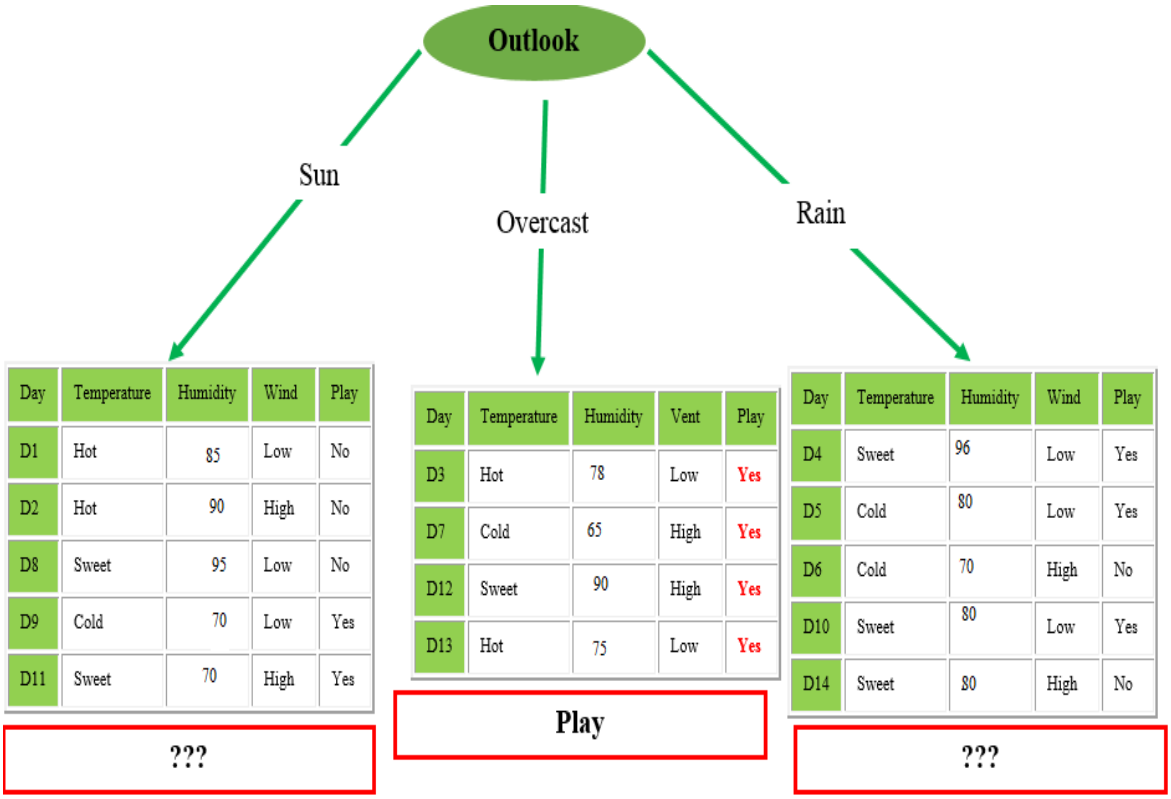
Τελικά λαμβάνουμε,

$$IGR\,(Humidity) = \frac{0.1022}{0.94} = 0.1087$$

Με χρήση της συνάρτησης καταλληλότητας 'Λόγος Κέρδους Πληροφορίας' (*IGR*), που χρησιμοποιεί ο Αλγόριθμος *C_4.5* και προκειμένου για τα υπόλοιπα κατηγορικά χαρακτηριστικά της ανάλυσης, λαμβάνουμε συγκεντρωτικά τα εξής αποτελέσματα

Χαρακτηριστικό Διάσπασης	**Κέρδος Πληροφορίας**	**Λόγος Κέρδους Πληροφορίας**
Outlook	0.2467	0.1564
Humidity (continuous)	0.1022	0.1087
Wind	0.0481	0.0488
Temperature	0.0292	0.0188

Προκειμένου για τον αλγόριθμο *C_4.5*, διαμορφώνονται τα ακόλουθα υποσύνολα δεδομένων εκπαίδευσης



Σχετικά με το υποσύνολο δεδομένων εκπαίδευσης που αντιστοιχεί στο επίπεδο '*Overcast'*, δεν απαιτείται περαιτέρω διάσπαση, λόγω της ομοιογένειας τιμών της μεταβλητής απόκρισης '*Play*'.

Αναφορικά με τα άλλα δύο υποσύνολα δεδομένων εκπαίδευσης, που αντιστοιχούν στα επίπεδα '*Sun*' και '*Rain*', η περαιτέρω διάσπαση κρίνεται αναγκαία, ώστε να προκύψει ει δυνατόν πιο ομοιογενής κατανομή τιμών της μεταβλητής απόκρισης.

Με εφαρμογή των προαναφερθέντων κανόνων υπολογισμού της συνάρτησης καταλληλότητας 'Λόγος Κέρδους Πληροφορίας' (*IGR)*, προκύπτει πως για το υποσύνολο δεδομένων εκπαίδευσης που αντιστοιχεί στο επίπεδο '*Rain*', η διάσπαση θα συνεχιστεί επί των επιπέδων της μεταβλητής '*Wind*'.

Για το υποσύνολο δεδομένων εκπαίδευσης που αντιστοιχεί στο επίπεδο '*Sun*', η διάσπαση θα συνεχιστεί με βάση το συνεχές χαρακτηριστικό '*Humidity*', στη θέση διάσπασης '*70*'. Τα παραπάνω αποτυπώνονται συγκεντρωτικά, στην ακόλουθη συλλογή εντολών, υπό τη μορφή 'ψευδοκώδικα' (*Pseudo code*)

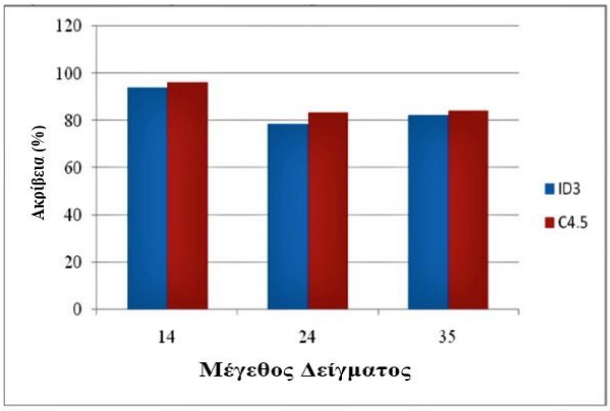
```
If Outlook= Sun
then
    If Humidity <= 70
    then
    Classification = Yes
    else
    Classification = No
else
    if Outlook = Overcast
    then
    Classification = Yes
    Else
     if Outlook= Rain then
            If Wind =High
            Classification = No
            Else
            Classification = Yes
```

Ο αλγόριθμος *C_4.5*, μέσω χρήσης της κανονικοποιημένης συνάρτησης καταλληλότητας 'Λόγου κέρδους πληροφορίας', δε μεροληπτεί υπέρ εκείνων των χαρακτηριστικών με μεγάλο πλήθος τιμών ή επιπέδων, διορθώνοντας τη σχετική αδυναμία του αλγορίθμου *ID3*. Επιπλέον, ο αλγόριθμος *C_4.5*, υπερτερεί του αλγορίθμου *ID3* στα ακόλουθα σημεία

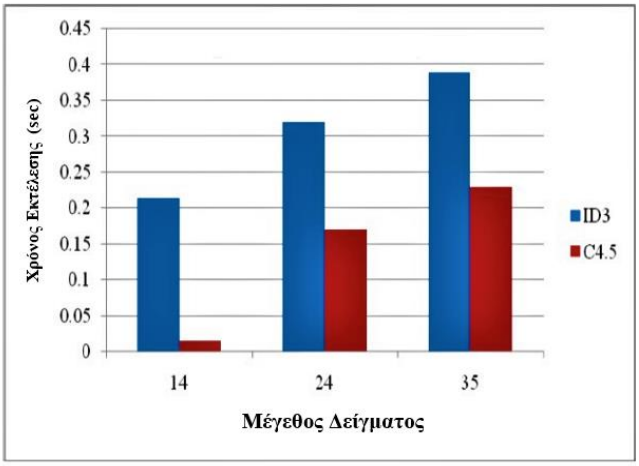
- ✓ Δυνατότητα χρησιμοποίησης χαρακτηριστικών συνεχούς κλίμακας
- ✓ Αξιοποίηση χαρακτηριστικών με αυξημένο ποσοστό ελλειπουσών τιμών
- ✓ Ταυτόχρονη αξιοποίηση χαρακτηριστικών διαφορετικής σημαντικότητας
- ✓ Δυνατότητα μετέπειτα 'κλαδέματος' του προκύπτοντος δέντρου κατάταξης

Η παρούσα ενότητα ολοκληρώνεται, με την παράθεση συγκριτικών πινάκων ακρίβειας αποτελεσμάτων (*Accuracy*), καθώς και πινάκων χρόνου εκτέλεσης (*Execution Time*), των αλγορίθμων *ID3* και *C_4.5*, αναλόγως του μεγέθους του χρησιμοποιούμενου δείγματος

Μέγεθος Δείγματος		
	ID3 (%)	C4.5 (%)
14	94.15	96.2
24	78.47	83.52
35	82.2	84.12



Μέγεθος Δείγματος		
	ID3 (%)	C4.5 (%)
14	0.215	0.0015
24	0.32	0.17
35	0.39	0.23



(Badr HSSINA et.al., TIAD laboratory, Computer Sciences Department, Faculty of sciences and techniques)

5.2 Παράδειγμα εφαρμογής Πολυωνυμικής Λογιστικής Παλινδρόμησης

Στη συγκεκριμένη μελέτη περίπτωσης (Chao-Ying Joanne Peng, Rebecca Naegle Nichols, 2003), παρουσιάζεται η δυνατότητα αξιοποίησης της μεθόδου Πολυωνυμικής Λογιστικής Παλινδρόμησης, ως εναλλακτικής και συνάμα αποτελεσματικής μεθόδου εποπτεύομενης ταξινόμησης (*supervised classification*), προκειμένου να αναγνωριστούν και κατηγοριοποιηθούν εκείνοι οι έφηβοι, οι οποίοι διατρέχουν το μεγαλύτερο κίνδυνο εμπλοκής σε συνήθειες επιβλαβείς για την υγεία τους, με βάση τα προσωπικά τους ιδιωματικά χαρακτηριστικά, καθώς και τα χαρακτηριστικά των οικογενειών στις οποίες ανατράφηκαν.

α. Μεθοδολογία και ταυτότητα μελέτης περιπτώσεως

Το Φθινόπωρο του 1988, από δύο Γυμνάσια της Αμερικής συγκεντρώθηκαν οι απαντήσεις που δόθηκαν σε συγκεκριμένο ερωτηματολόγιο, το οποίο έλαβαν 517 μαθητές. Οι μαθητές που επιλέχθηκαν να συμμετάσχουν στην έρευνα, είχαν επιτύχει μέση βαθμολογία στα μαθήματα τους, η οποία κυμαίνονταν από 7 έως 9, με μέση ηλικία τα 13.9 έτη. Εξ αυτών, οι 85 μαθητές δε συμπλήρωσαν πλήρως το ερωτηματολόγιο που τους δόθηκε, επομένως το πλήθος των πλήρως απαντημένων ερωτηματολογίων, ανήλθε τελικά στα 432 ερωτηματολόγια. Το πλήθος των κοριτσιών που συμμετείχαν στην έρευνα και συμπλήρωσαν πλήρως το ερωτηματολόγιο τους, ανήλθε στα 208 άτομα, ενώ το αντίστοιχο πλήθος των αγοριών, ανήλθε στα 224 άτομα. Το ερωτηματολόγιο είναι το *Health Behavioral Questionnaire* (*HBQ*; Ingersoll & Orr, 1989; Resnick, Harris & Blum, 1993), σε συνδυασμό με το *Rosenberg's self-esteem inventory* (Rosenberg, 1965) και μοιράστηκε στους μαθητές την ίδια ακριβώς χρονική στιγμή, κατά τη διάρκεια παραδόσεως του υποχρεωτικού μαθήματος των Μαθηματικών.

Η εξαρτημένη μεταβλητή (Απόκριση) του μοντέλου, ορίστηκε να είναι η *Y: 'Συμπεριφορικός Κίνδυνος' (Behavioral Risk)*, η οποία αποτυπώνει τη βαθμολογία (*scoring*) εκάστου μαθητή στο *HBQ* ερωτηματολόγιο, με μέση τιμή 47.69, τυπική απόκλιση 10.89 και εύρος τιμών από 40.44 έως 66.81. Προκειμένου να κατηγοριοποιηθεί η μεταβλητή απόκρισης, εφαρμόστηκε ο ακόλουθος κανόνας

- ✓ Μαθητές με τιμή απόκρισης *Y* (βαθμολογία στο *HBQ* ερωτηματολόγιο), η οποία απέχει απόσταση μεγαλύτερη της μίας τυπικής αποκλίσεως, από τη μέση τιμή της μεταβλητής ($Y \geq 60$), κατατάχθηκαν στην ανώτερη κατηγορία συμπεριφορικού κινδύνου (*Y: High Behavioral Risk Level = 3)*

- ✓ Μαθητές με τιμή απόκρισης *Y*, η οποία κυμάνθηκε μεταξύ 45 και 59 ($45 \leq Y \leq 59$), κατατάχθηκαν στη μέση κατηγορία συμπεριφορικού κινδύνου (*Y: Medium Behavioral Risk Level = 2)*

- ✓ Μαθητές με τιμή απόκρισης *Y*, η οποία κυμάνθηκε μεταξύ 40 και 44 ($40 \leq Y \leq 44$), κατατάχθηκαν στη χαμηλή κατηγορία συμπεριφορικού κινδύνου (*Y:Low Behavioral Risk Level = 1)*

Το σημείο αποκοπής (*cut off point*), του μέσου και κατώτερου συμπεριφορικού κινδύνου, ορίστηκε στη τιμή *Y = 45* και προσδιορίστηκε ως εξής: Η διάμεσος των τιμών της απόκρισης *Y*, του δείγματος των μαθητών οι οποίοι δεν κατατάχθηκαν στο επίπεδο του υψηλού συμπεριφορικού κινδύνου, εντοπίστηκε μεταξύ των τιμών *44* και *45*. Η τιμή *45* επιλέχθηκε λόγω της θετικής λοξότητας της αντιστοίχου κατανομής.

Οι ερμηνευτικές μεταβλητές του μοντέλου, είναι οι ακόλουθες

- ✓ *X1: 'Gender'* (*Nominal, Levels: Girls=0, Boys=1)*
- ✓ *X2: 'Drop out of School'* (*Nominal, Levels: No=0, Yes=1)*
- ✓ *X3: 'Family Structure'*
  (*Nominal, Levels: Intact=1, Step parent=2, Single parent=3)*
- ✓ *X4: Emotional Risk score (continuous)*
- ✓ *X5: Self Esteem score (continuous)*

Η υπόθεση η οποία ελέγχθηκε στα πλαίσια της μελέτης διατυπώθηκε ως εξής: *Η πιθανοφάνεια ένας έφηβος να καταταχθεί σε ένα εκ των επιπέδων (nominal levels), της μεταβλητής απόκρισης Y:'Behavioral Risk', σχετίζεται με το Φύλο του, την πρόθεση του να εγκαταλείψει το Σχολείο, την οικογενειακή του κατάσταση, καθώς και με τις βαθμολογίες που έλαβε στα σχετικά ερωτηματολόγια που καθορίζουν, το επίπεδο της Αυτοεκτίμησης του, καθώς και το επίπεδο του Συναισθηματικού του κινδύνου.*

Οι παρατηρούμενες συχνότητες εμφάνισης, οι οποίες καθορίζουν το βαθμό συσχέτισης μεταξύ της μεταβλητής απόκρισης και των κατηγορικών μεταβλητών της μελέτης, παρουσιάζονται στους ακόλουθους πίνακες συνάφειας

Συμπεριφορικός Κίνδυνος (Behavioral Risk)	Φύλο (Gender) Κορίτσι = 0 Αγόρι = 1		Σύνολα
Υψηλός Κίνδυνος (High Risk)	5	24	29
Μέσος Κίνδυνος (Medium Risk)	66	104	170
Χαμηλός Κίνδυνος (Low Risk)	137	96	233
Σύνολα	208	224	432

Τα αγόρια της μελέτης κατατάσσονται κατά κύριο λόγο, στο μέσο και ανώτερο επίπεδο συμπεριφορικού κινδύνου, σε αντιδιαστολή με τα κορίτσια της μελέτης, τα οποία κατατάσσονται στο μέσο και κατώτερο επίπεδο κινδύνου.

Συμπεριφορικός Κίνδυνος (Behavioral Risk)	Εγκατάλειψη Σχολείου (Drop Out) Όχι = 0 Ναι = 1		Σύνολα
Υψηλός Κίνδυνος (High Risk)	15	14	29
Μέσος Κίνδυνος (Medium Risk)	137	33	170
Χαμηλός Κίνδυνος (Low Risk)	227	6	233
Σύνολα	379	53	432

Οι παραπάνω παρατηρούμενες συχνότητες καταδεικνύουν πως, εκείνοι οι έφηβοι που εμφανίζουν πρόθεση εγκατάλειψης του σχολείου, κατατάσσονται στο μέσο και υψηλό επίπεδο κινδύνου, σε αντίθεση με εκείνους οι οποίοι συνεχίζουν το σχολείο, που κατατάσσονται στο μέσο και χαμηλό επίπεδο κινδύνου

Συμπεριφορικός Κίνδυνος (Behavioral Risk)	Οικογενειακή Κατάσταση (Family Structure) Ανάδοχος =1 Υιοθεσία = 2 Μονογονεϊκή = 3 (Intake) (Step) (Single)			Σύνολα
Υψηλός Κίνδυνος (High Risk)	8	7	14	29
Μέσος Κίνδυνος (Medium Risk)	62	38	70	170
Χαμηλός Κίνδυνος (Low Risk)	123	53	57	233
Σύνολα	193	98	141	432

Οι έφηβοι που προέρχονται από οικογένειες με έναν φυσικό γονέα, κατατάσσονται κατά κύριο λόγο στο μέσο και υψηλό επίπεδο κινδύνου, σε αντίθεση με τις οικογένειες στις οποίες υφίστανται ανάδοχοι γονείς, ή ένας θετός γονέας, όπου η κατάταξη κατά κύριο λόγο πραγματοποιείται στο μέσο και χαμηλό επίπεδο κινδύνου.

Προκειμένου να κατασκευάσουμε και εκτιμήσουμε τις εξισώσεις του Πολυωνυμικού Λογιστικού Μοντέλου (*MLR: Multinomial Logistic Regression Model*), θα πρέπει πρότερα να ορίσουμε τις ακόλουθες πιθανότητες

- ✓ p_1: Η πιθανότητα ένας έφηβος, να ανήκει στην ομάδα υψηλού συμπεριφορικού κινδύνου (*High Behavioral Risk: Y = 3*)
- ✓ p_2: Η πιθανότητα ένας έφηβος, να ανήκει στην ομάδα μεσαίου συμπεριφορικού κινδύνου (*Medium Behavioral Risk: Y = 2*)
- ✓ p_3: Η πιθανότητα ένας έφηβος, να ανήκει στην ομάδα χαμηλού συμπεριφορικού κινδύνου (*Low Behavioral Risk: Y = 1*)

Προφανώς θα πρέπει να ισχύει ότι

$$p_1 + p_2 + p_3 = 1$$

Με δεδομένο πως τα επίπεδα της μεταβλητής απόκρισης *Y* είναι τρία, οι εξισώσεις μοντελοποίησης των αθροιστικών σχετικών πιθανοτήτων (*cumulative odds ratios*), στα πλαίσια του Πολυωνυμικού Λογιστικού Μοντέλου, είναι δύο στο πλήθος και διαμορφώνονται ως ακολούθως

$$log(\frac{p_1}{1-p_1}) = \alpha_1 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

$$log(\frac{p_1+p_2}{1-(p_1+p_2)}) = \alpha_2 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

Παρατηρούμε πως, στη παραπάνω μοντελοποίηση διατηρούνται οι συντελεστές των ερμηνευτικών μεταβλητών και μεταβάλλεται μόνον η αντίστοιχη σταθερά. Αντιλογαριθμίζοντας τις παραπάνω εξισώσεις, λαμβάνουμε τους ακόλουθους τύπους υπολογισμού των αθροιστικών πιθανοτήτων

$$p_1 = \frac{exp(\alpha_1 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5)}{1 + exp(\alpha_1 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5)}$$

$$p_1 + p_2 = \frac{exp(\alpha_2 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5)}{1 + exp(\alpha_2 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5)}$$

Η εφαρμογή του χ^2 *test*, κατέδειξε $p - value = 0.6548 > \alpha = 0.05$, με αποτέλεσμα να μην απορρίπτεται η μηδενική υπόθεση περί επάρκειας και καλής προσαρμογής του μοντέλου στα δεδομένα. Επομένως, δεν κρίθηκε αναγκαία η προσαρμογή εναλλακτικού μοντέλου, με διακεκριμένους συντελεστές ερμηνευτικών μεταβλητών (Peterson & Harrell, 1990).

β. Προσαρμογή και ερμηνεία μοντέλου Πολυωνυμικής Λογιστικής Παλινδρόμησης

Predictor	β	*SE* β	Wald's χ^2 (*df*=1)	*p*	e^{β} *(odds ratio)*
CONSTANT 1 (Y_1)	–0.6211	1.0627	0.3416	0.5589	Not necessary
CONSTANT 2 (Y_1+Y_2)	2.5220	1.0723	5.5317	0.0187	Not necessary
GENDER (boys=1,girls=0)	1.1070	0.2111	27.5060	<0.0001	3.0253
DROPOUT (yes=1, no=0)	2.1818	0.3287	44.0618	<0.0001	8.8622
FAMILY	0.4135	0.1179	12.2979	<0.001	1.5121
EMOTION	0.0074	0.0115	0.4118	0.5211	1.0074
ESTEEM	–0.0488	0.0118	16.9867	<0.0001	0.9524

Overall Model Evaluation

Tests	χ^2	*df*	*p*
Likelihood Ratio Test	122.38	5	<0.0001
Score test	110.47	5	<0.0001
Wald test	97.87	5	<0.0001

Notes. Cox and Snell *R* squared=0.2467. Nagelkerke *R* squared (Max rescaled *R* squared)=0.2978. Kendall's Tau-*a* = 0.297. Goodman-Kruskal's Gamma= 0.548. Somers' D_{xy}= 0.539. *c*-statistic = 0.769.

SAS® Programming Codes

```
PROC LOGISTIC DATA=risk432
        MODEL risk= gender dropout family emotion esteem;
        OUTPUT out=probs  predicted=prob xbeta=logit;
RUN;
```

Εκ του συνόλου των χρησιμοποιούμενων ερμηνευτικών μεταβλητών της παρούσας μελέτης, όλες κρίνονται ως στατιστικώς σημαντικές, σε επίπεδο σημαντικότητας *α = 0.05 (5%)*, με εξαίρεση την ερμηνευτική μεταβλητή '*Emotional Risk*', για την οποία δεν συνάδεται από τα προκύπτοντα αποτελέσματα της προσαρμογής, στατιστική σημαντικότητα (*p - value = 0.5211 > 0.05*).

Η ποιοτική ερμηνεία του μοντέλου, με ταυτόχρονη αξιοποίηση και των τεσσάρων ερμηνευτικών μεταβλητών που προέκυψαν ως στατιστικώς σημαντικές, μετά της προσαρμογής στα πειραματικά δεδομένα, μας αποκαλύπτει το ακόλουθο προφίλ εφήβου, με υψηλή πιθανότητα εμφάνισης, αυξημένου επιπέδου συμπεριφορικού κινδύνου

Άνδρας έφηβος, με πρόθεση εγκατάλειψης του σχολείου, προερχόμενος από μονογονεϊκή οικογένεια, με χαμηλό επίπεδο αυτοεκτίμησης και πιθανότατα με υψηλή βαθμολογία στην κλίμακα συναισθηματικού κινδύνου, εμφανίζει υψηλή πιθανότητα εκδήλωσης αυξημένου συμπεριφορικού κινδύνου και υιοθέτησης πρακτικών ή συνηθειών, επιβλαβών για την προσωπική του υγεία και ασφάλεια.

Αναλυτικότερα, ανά ερμηνευτική μεταβλητή (σταθεροποιώντας τις υπόλοιπες), έχουμε την ακόλουθη ερμηνεία

- ✓ *Ερμηνευτική Μεταβλητή : 'Gender' (p – value = 0.0001 < 0.05)*

  Η εκτιμωμένη σχετική πιθανότητα ένας άνδρας έφηβος, να εμφανίσει υψηλό συμπεριφορικό κίνδυνο, έναντι του ενδεχομένου μη εμφάνισης υψηλού συμπεριφορικού κινδύνου, είναι *exp(1.1070) = 3.0253 φορές (ή 200 %) μεγαλύτερη,* της εκτιμώμενης σχετικής πιθανότητας, μία γυναίκα έφηβος να εμφανίσει υψηλό συμπεριφορικό κίνδυνο, έναντι του ενδεχομένου μη εμφάνισης υψηλού συμπεριφορικού κινδύνου.

- ✓ *Ερμηνευτική Μεταβλητή : 'Drop out' (p – value = 0.0001 < 0.05)*

  Η εκτιμωμένη σχετική πιθανότητα ένας έφηβος με πρόθεση εγκατάλειψης του σχολείου, να εμφανίσει υψηλό συμπεριφορικό κίνδυνο, έναντι του ενδεχομένου μη εμφάνισης υψηλού συμπεριφορικού κινδύνου, είναι *exp(2,1818) = 8,8622 φορές (ή 800 %)* μεγαλύτερη, της εκτιμώμενης σχετικής πιθανότητας, ένας έφηβος με πρόθεση παραμονής στο σχολείο, να εμφανίσει υψηλό συμπεριφορικό κίνδυνο, έναντι του ενδεχομένου μη εμφάνισης υψηλού συμπεριφορικού κινδύνου.

- ✓ *Ερμηνευτική Μεταβλητή : 'Family status'(p – value = 0.001 < 0.05)*

  Η εκτιμωμένη σχετική πιθανότητα ένας έφηβος προερχόμενος από οικογένεια με έναν θετό γονέα, να εμφανίσει υψηλό συμπεριφορικό κίνδυνο, έναντι του ενδεχομένου μη εμφάνισης υψηλού συμπεριφορικού κινδύνου, είναι *exp(0,4135) = 1,5121 φορές (ή 51 %)* μεγαλύτερη, της εκτιμώμενης σχετικής πιθανότητας, ένας έφηβος προερχόμενος από ανάδοχη οικογένεια, να εμφανίσει υψηλό συμπεριφορικό κίνδυνο, έναντι του ενδεχομένου μη εμφάνισης υψηλού συμπεριφορικού κινδύνου.

Η εκτιμωμένη σχετική πιθανότητα ένας έφηβος προερχόμενος από οικογένεια με έναν φυσικό γονέα, να εμφανίσει υψηλό συμπεριφορικό κίνδυνο, έναντι του ενδεχομένου μη εμφάνισης υψηλού συμπεριφορικού κινδύνου, είναι *exp(0,4135) = 1,5121 φορές (ή 51 %)* μεγαλύτερη, της εκτιμώμενης σχετικής πιθανότητας, ένας έφηβος προερχόμενος από οικογένεια με έναν θετό γονέα, να εμφανίσει υψηλό συμπεριφορικό κίνδυνο, έναντι του ενδεχομένου μη εμφάνισης υψηλού συμπεριφορικού κινδύνου.

- ✓ *Ερμηνευτική Μεταβλητή : 'Self Esteem'(p – value = 0.0001 < 0.05)*

Κάθε αύξηση κατά μία μονάδα στην κλίμακα αυτοεκτίμησης ενός εφήβου, επιφέρει exp(- 0,0488) = 0,9524 φορές (ή 5 %) μείωση, της εκτιμώμενης σχετικής πιθανότητας, ο εν λόγω έφηβος να εμφανίσει υψηλό συμπεριφορικό κίνδυνο, έναντι του ενδεχομένου μη εμφάνισης υψηλού συμπεριφορικού κινδύνου.

Καταληκτικά, οι εκτιμώμενες πιθανότητες για κάθε δυνατό συνδυασμό επιπέδων, των ερμηνευτικών μεταβλητών, δίδονται στον ακόλουθο πίνακα

Case No.	SEX β= 1.107 1=boy 0=girl	DROPOUT β=2.1818 1=yes 0=no	FAMILY β=0.4135 1=intact, 2=step, 3=single	EMOTION β=0.0074	ESTEEM β= −0.0488	Intercept 1 α_1= −0.6211	Intercept 2 α_2=2.522	Predicted probability of participating in self-injurious behavior p_1 (high)	p_2 (medium)	p_3 (low)	Actual Behavior risk, 1=high, 2=med, 3=low (score on HBO, M=47.69, SD=10.89)
1	1	0	1	62.39	32.68	−0.6211	2.5220	.0818	.5921	.3261	1 (60.40)
2	1	0	1	80.74	32.68	−0.6211	2.5220	.0926	.6102	.2972	2 (52.77)
3	1	0	1	32.07	71.58	−0.6211	2.5220	.0106	.1878	.8016	3 (42.65)
4	1	1	1	72.72	46.41	−0.6211	2.5220	.3038	.6062	.0900	1 (95.21)
5	1	1	1	63.07	37.25	−0.6211	2.5220	.3885	.5479	.0636	2 (50.00)
6	1	1	1	----	----	−0.6211	2.5220	---	---	---	3 (------)
7	0	0	1	47.29	41.83	−0.6211	2.5220	.0166	.2645	.7189	1 (61.53)
8	0	0	1	45.78	44.12	−0.6211	2.5220	.0147	.2422	.7431	2 (47.07)
9	0	0	1	42.05	21.24	−0.6211	2.5220	.0425	.4643	.4932	3 (42.70)
10	0	1	1	51.37	34.97	−0.6211	2.5220	.1772	.6559	.1669	1 (70.23)
11	0	1	1	56.77	37.25	−0.6211	2.5220	.1670	.6559	.1771	2 (53.27)
12	0	1	1	----	----	−0.6211	2.5220	---	---	---	3 (-------)
13	1	0	2	41.36	50.98	−0.6211	2.5220	.0451	.4776	.4773	1 (72.83)
14	1	0	2	46.14	50.98	−0.6211	2.5220	.0467	.4848	.4685	2 (45.84)
15	1	0	2	36.11	41.83	−0.6211	2.5220	.0663	.5559	.3778	3 (40.44)
16	1	1	2	38.59	57.85	−0.6211	2.5220	.2269	.6449	.1282	1 (92.50)
17	1	1	2	54.87	46.41	−0.6211	2.5220	.3665	.5641	.0694	2 (46.99)
18	1	1	2	70.35	34.97	−0.6211	2.5220	.5312	.4321	.0367	3 (43.52)
19	0	0	2	---	---	−0.6211	2.5220	---	---	---	1 (-------)
20	0	0	2	34.21	44.12	−0.6211	2.5220	.0203	.3041	.6756	2 (45.78)
21	0	0	2	50.18	53.27	−0.6211	2.5220	.0147	.2421	.7432	3 (40.44)
22	0	1	2	----	---	−0.6211	2.5220	---	---	----	1 (------)
23	0	1	2	54.84	50.98	−0.6211	2.5220	.1326	.6473	.2201	2 (48.64)
24	0	1	2	50.18	46.41	−0.6211	2.5220	.1559	.6547	.1894	3 (43.08)
25	1	0	3	63.52	23.52	−0.6211	2.5220	.2432	.6384	.1184	1 (67.90)
26	1	0	3	32.07	67.00	−0.6211	2.5220	.0296	.3848	.5856	2 (56.69)
27	1	0	3	50.18	48.70	−0.6211	2.5220	.0786	.5854	.3360	3 (40.44)
28	1	1	3	43.54	48.70	−0.6211	2.5220	.4184	.5250	.0566	1 (85.49)
29	1	1	3	56.74	44.12	−0.6211	2.5220	.4979	.4604	.0417	2 (54.31)
30	1	1	3	---	---	−0.6211	2.5220	---	---	----	3 (-------)
31	0	0	3	---	---	−0.6211	2.5220	---	---	---	1 (-------)
32	0	0	3	64.12	28.10	−0.6211	2.5220	.0786	.5856	.3358	2 (48.41)
33	0	0	3	60.08	39.54	−0.6211	2.5220	.0453	.4781	.4766	3 (44.41)
34	0	1	3	---	---	−0.6211	2.5220	---	---	---	1 (-------)
35	0	1	3	43.63	48.70	−0.6211	2.5220	.1922	.6543	.1535	2 (46.34)
36	0	1	3	---	---	−0.6211	2.5220	---	---	----	3 (-------)

γ. Αξιολόγηση μοντέλου Πολυωνυμικής Λογιστικής Παλινδρόμησης

Κατά τη προσαρμογή του μοντέλου, προκύπτουν οι ακόλουθοι δείκτες

- ✓ R^2 *(Cox & Snell, 1989)* = 0.2467
- ✓ R^2 *(Nagelkerke, 1991)* = 0.2978

Οι εν λόγω δείκτες αποτελούν φυσική επέκταση του δείκτη R^2 της κλασσικής παλινδρόμησης, χωρίς όμως να φέρουν τη φυσική ερμηνεία του εν λόγω δείκτη, ως το ποσοστό της συνολικής μεταβλητότητας των δεδομένων, το οποίο ερμηνεύει το μοντέλο. Ως εκ τούτου, η χρήση των δεικτών αυτών είναι προαιρετική και συμπληρωματική των ελέγχων συνολικής επάρκειας και καλής προσαρμογής του μοντέλου.

Ένας έλεγχος επάρκειας μοντέλου Πολυωνυμικής Λογιστικής Παλινδρόμησης, έχει προταθεί από τους *Begg & Gray (Begg & Gray 1984, cited in Hosmer & Lemeshow, 2001)*. Πιο συγκεκριμένα, έχει προταθεί η εφαρμογή του ελέγχου επάρκειας και καλής προσαρμογής των *Hosmer & Lemeshow*, για όλους τους ανά δύο δυνατούς συνδυασμούς επιπέδων της μεταβλητής απόκρισης. Εν συνέχεια, πραγματοποιείται η ενοποίηση και συνδυαστική ερμηνεία των επιμέρους αποτελεσμάτων. Στη συγκεκριμένη μελέτη περιπτώσεως, προέκυψαν τα ακόλουθα συμπεράσματα

- ✓ Για το μοντέλο Απλής Λογιστικής Παλινδρόμησης, που συνέκρινε το μέσο με το χαμηλό επίπεδο συμπεριφορικού κινδύνου, προέκυψε τιμή για το χ^2 στατιστικό των *H – L*, ίση με 5.8011 στους 8 βαθμούς ελευθερίας και σχετικό $p - value = 0,67 > 0,05$.
- ✓ Για το μοντέλο Απλής Λογιστικής Παλινδρόμησης, που συνέκρινε το υψηλό με το χαμηλό επίπεδο συμπεριφορικού κινδύνου, προέκυψε τιμή για το χ^2 στατιστικό των *H – L*, ίση με 8.2925 στους 8 βαθμούς ελευθερίας και σχετικό $p - value = 0,40 > 0,05$.

Συνδυαστικά, σε κάθε περίπτωση το μοντέλο Πολυωνυμικής Λογιστικής Παλινδρόμησης, κρίνεται ως επαρκές, έχοντας καλή προσαρμογή στα πειραματικά δεδομένα με $p - value \geq 0.40$. Επομένως, η μηδενική υπόθεση περί επαρκούς μοντέλου με καλή προσαρμογή στα δεδομένα, δε μπορεί να απορριφθεί.

Εναλλακτικά, για τον έλεγχο της συνολικής επάρκειας του προκύπτοντος Πολυωνυμικού μοντέλου Λογιστικής Παλινδρόμησης, δίδονται οι ακόλουθοι δείκτες:

- ✓ *Kendall's Tau-a*

  Πρόκειται για έναν δείκτη συσχέτισης των επιπέδων της μεταβλητής απόκρισης (*rank - order correlation coefficient*), μη προσαρμοσμένος στην ταυτόχρονη ύπαρξη *'δεσμών'(ties),* στις κατηγορίες της μεταβλητής απόκρισης, καθώς και στις εκτιμώμενες πιθανότητες. Στη παρούσα μελέτη έλαβε τιμή *0.297*. Τα συγκεκριμένα πειραματικά δεδομένα εμφανίζουν *923* περιπτώσεις δεσμών (ή ποσοστό *1.8%* του συνόλου των δυνατών συνδυασμών), γεγονός το οποίο στη παρούσα μελέτη περίπτωσης, καθιστά προβληματική και μη προτεινόμενη, τη χρήση και σχετική αξιοποίησή του.

- ✓ *Goodman - Kruskal's Gamma*

  Πρόκειται για έναν διορθωμένο δείκτη συσχέτισης των κατηγοριών της μεταβλητής απόκρισης και των εκτιμωμένων από το μοντέλο πιθανοτήτων, προσαρμοσμένος στους δεσμούς των δεδομένων. Έλαβε τιμή ίση με *0.548* και ερμηνεύεται ως ακολούθως

  Πραγματοποιήθηκαν 54.8% λιγότερα σφάλματα ταξινόμησης, κατά τη διαδικασία κατάταξης των εφήβων της μελέτης, στις τρείς κατηγορίες της μεταβλητής απόκρισης 'Behavioral Risk' (High, Medium, Low), μέσω αξιοποίησης των εκτιμώμενων πιθανοτήτων του μοντέλου, κατά τη διαδικασία κατάταξης, έναντι της τυχαίας αναθέσεως.

  Ο συγκεκριμένος δείκτης πρέπει να χρησιμοποιείται με προσοχή, λαμβάνοντάς υπόψιν τα ακόλουθα σημεία κριτικής του. Ο δείκτης έχει τη τάση να υπερεκτιμά την ισχύ της συσχέτισης, μεταξύ των κατηγοριών της απόκρισης και των αντίστοιχα εκτιμωμένων πιθανοτήτων (*Demaris, 1992*). Επιπλέον, σε περιπτώσεις κατά τις οποίες η δομή του πίνακα συνάφειας ξεπερνά τη διάσταση *2 x 2*, τότε πιθανή τιμή του δείκτη η οποία προσεγγίζει το μηδέν, δεν πρέπει να ερμηνεύεται απαραιτήτως ως ανεξαρτησία, μεταξύ των επιπέδων της μεταβλητής απόκρισης (*Siegel & Castellan, 1988*).

- ✓ *Somers' Delta*

Αποτελεί μία βελτιωμένη επέκταση του δείκτη *Gamma*, λαμβάνοντας τις ακόλουθες δύο διαφορετικές μεταξύ των, μη συμμετρικές παραλλαγές D_{xy} και D_{yx}, *(Siegel & Castellan, 1988),* όπου *X*: *'Εκτιμώμενες πιθανότητες του προσαρμοσμένου μοντέλου'* και *Y: 'Κατηγορίες Μεταβλητής Απόκρισης'*

- *Somers'* D_{xy}
  Στη παρούσα μορφή, υπολογίζεται ο δείκτης συσχέτισης της ανεξάρτητης μεταβλητής *X* και της εξαρτημένης μεταβλητής *Y*. Προκειμένου για τη παρούσα μελέτη περιπτώσεως, έλαβε τιμή *0.539*
- *Somers'* D_{yx}
  Η συγκεκριμένη μορφή του δείκτη, ονομάζεται και *c – statistic*, κυμαίνεται μεταξύ των τιμών *0.5* και *1*, αποτυπώνοντας το πραγματικό επίπεδο συσχέτισης μεταξύ των *X* και *Y* (*Demaris, 1992*). Αποτελεί κατ' ουσία το εμβαδό *AUC (Area Under Curve)* κάτω της καμπύλης *ROC (Received Operating Characteristic Curve)* και έλαβε τιμή *0.769*. Ερμηνεύεται δε ως ακολούθως

  Στο 76.9% όλων των ανά δύο δυνατών συνδυασμών εφήβων, όπου ο ένας έφερε υψηλότερο επίπεδο συμπεριφορικού κινδύνου σε σχέση με τον άλλον, το μοντέλο προέβλεψε υψηλότερη πιθανότητα σε εκείνους τους εφήβους, με υψηλότερη βαθμολογία στο ερωτηματολόγιο HBQ. Επομένως, το προσαρμοσμένο MLR μοντέλο, λειτούργησε αποτελεσματικότερα έναντι της τυχαίας αναθέσεως των εφήβων, σε κάποια εκ των τριών κατηγοριών συμπεριφορικού κινδύνου.

6 Μελέτη περίπτωσης

Στην παρούσα μελέτη περίπτωσης, θα επιχειρήσουμε την πρακτική εφαρμογή των αλγορίθμων διαχωριστικής ανάλυσης, οι οποίοι αναπτύχθηκαν στις προηγούμενες ενότητες. Πιο συγκεκριμένα, θα εφαρμόσουμε το διαχωριστικό αλγόριθμο *C_4.5*, καθώς και τον αλγόριθμο της Πολυωνυμικής Λογιστικής Παλινδρόμησης (*MLR*), στο ίδιο σύνολο πρωτογενών δεδομένων, με σκοπό τη συγκριτική αξιολόγηση της διακριτικής και προβλεπτικής ικανότητας των δέντρων απόφασης τα οποία θα προκύψουν, αναφορικά με την κατάταξη των εξεταζόμενων περιπτώσεων, στην κατάλληλη στάθμη / επίπεδο της κατηγορικής μεταβλητής '*οικογενειακή κατάσταση' (marital status).*

α. Περιγραφή συνόλου δεδομένων

Το πρωτογενές σύνολο δεδομένων (*Primary Data Set*), που χρησιμοποιήθηκε προκειμένου να εφαρμοστούν οι αλγόριθμοι *C 4.5* και *MLR*, ώστε να αξιολογηθεί συγκριτικά η διακριτική τους ικανότητα, προέρχεται από το *UCI ML Repository* και είναι το *Census-Income (KDD) Data Set*. Τα εν λόγω δεδομένα, https://archive.ics.uci.edu/ml/datasets/Census-Income+(KDD) περιλαμβάνουν πληροφορίες σχετικά με τα κοινωνικά, δημογραφικά, καθώς και οικονομικά χαρακτηριστικά του εργατικού δυναμικού του πληθυσμού των Η.Π.Α, της χρονικής περιόδου 1996-1997. Το συνολικό μέγεθος των αρχείου ανέρχεται σε *300.000* εγγραφές, καταγράφοντας συνολικά *41* ερμηνευτικές μεταβλητές.

Οι αρχικές ερμηνευτικές μεταβλητές του συνόλου των δεδομένων, είναι είτε κατηγορικές (*nominal*), είτε συνεχείς (*continuous*). Η μεταβλητή η οποία θα χρησιμοποιηθεί ως εξαρτημένη μεταβλητή απόκρισης, είναι η κατηγορική (*nominal*) μεταβλητή '*marital status*', η οποία αρχικά είχε *7* επίπεδα, τα οποία συγχωνεύτηκαν για τις ανάγκες της ανάλυσης, στα ακόλουθα τρία

1. *Never Married* (Στάθμη Αναφοράς)
2. *Married*
3. *Divorced / Separated*

Στο σημείο αυτό πρέπει να αναφερθεί πως, το μεγαλύτερο πλήθος ερμηνευτικών μεταβλητών του αρχικού συνόλου δεδομένων, χαρακτηρίζονται από μεγάλο ποσοστό ελλειπουσών τιμών, καθώς και από την παρουσία ακραίων παρατηρήσεων. Προκειμένου να αντιμετωπιστούν τα εν λόγω προβλήματα ποιότητας των αρχικών δεδομένων, καθώς και να περιοριστεί το πλήθος των χρησιμοποιούμενων ερμηνευτικών μεταβλητών, επελέγησαν να χρησιμοποιηθούν στην ανάλυση εκείνες οι μεταβλητές, οι οποίες εμφάνιζαν μηδενικό ποσοστό ελλειπουσών τιμών, καθώς και το μεγαλύτερο βαθμό συσχέτισης με τη μεταβλητή απόκρισης. Κατόπιν της εφαρμογής των παραπάνω, το σύνολο των χρησιμοποιούμενων μεταβλητών περιορίστηκε στις ακόλουθες *8* ερμηνευτικές μεταβλητές

Εξαρτημένη Μεταβλητή Απόκρισης			
id	**Όνομα Μεταβλητής**	**Τύπος**	**Πλήθος Επιπέδων**
1	Οικογενειακή Κατάσταση (*marital status*)	*Nominal*	3
Ανεξάρτητες Ερμηνευτικές Μεταβλητές			
id	**Όνομα Μεταβλητής**	**Τύπος**	**Πλήθος Επιπέδων**
1	Μορφωτικό Επίπεδο (*education*)	*Nominal*	6
2	Φύλο (*sex*)	*Nominal*	2
3	Κατάσταση Νοικοκυριού (*household status*)	*Nominal*	6
4	Υπηκοότητα (*citizenship*)	*Nominal*	4
5	Εισοδηματικό Επίπεδο (*income level*)	*Nominal*	2
6	Εργασιακή Κατάσταση (*employment status*)	*Nominal*	3
7	Πλήθος εργαζομένων στην επιχείρηση απασχόλησης (*number of persons worked for employer*)	*Nominal*	6
8	Αριθμός εργάσιμων εβδομάδων ανά έτος *(number of weeks worked in year)*	*Nominal*	10

β. Διαμόρφωση συνόλων δεδομένων Εκπαίδευσης και Επικύρωσης

Προκειμένου να εκπαιδευτούν οι αλγόριθμοι *C_4.5* και *MLR*, θα χρησιμοποιηθούν *6* διαφορετικά υποσύνολα δεδομένων εκπαίδευσης, τα οποία θα προέλθουν από το αρχικό σύνολο δεδομένων, μέσω τυχαίας στρωματοποιημένης δειγματοληψίας (*stratified random sampling*). Συγκεκριμένα, οι αρχικές ποσοστώσεις των τριών επιπέδων της μεταβλητής απόκρισης, οι οποίες ισχύουν στο αρχικό σύνολο πρωτογενών δεδομένων, θα πρέπει να είναι συγκρίσιμες με τις ποσοστώσεις των τριών επιπέδων της μεταβλητής απόκρισης, στα προκύπτοντα υποσύνολα δεδομένων εκπαίδευσης. Το κάθε ένα από τα υποσύνολα δεδομένων εκπαίδευσης, θα έχει μέγεθος περίπου 50000 εγγραφών. Κάθε υποσύνολο δεδομένων εκπαίδευσης το οποίο προκύπτει από την τυχαία στρωματοποιημένη δειγματοληψία, θα συνδέεται με ένα υποσύνολο δεδομένων επικύρωσης συγκρίσιμου μεγέθους, περίπου 15000 εγγραφών (στο 30% του μεγέθους του αντίστοιχου συνόλου εκπαίδευσης), το οποίο θα αντλείται από το αρχικό σύνολο δεδομένων, μετά της αφαίρεσης των εγγραφών του συνδεδεμένου υποσυνόλου εκπαίδευσης, με την ίδια ακριβώς μεθοδολογία τυχαίας στρωματοποιημένης δειγματοληψίας. Συγκεντρωτικά έχουμε

Sample Size	Training Data Set	Test Data Set	~ % of split
1st	43000	18000	70 - 30
2nd	41000	18000	70 - 30
3rd	48000	15000	75 - 25
4th	40000	21000	70 - 30
5th	42000	15000	75 - 25
6th	46000	18000	70 - 30

Αξίζει να σημειωθεί πως, από τα προκύπτοντα υποσύνολα δεδομένων επικύρωσης, θα πρέπει να απομονώνεται η μεταβλητή απόκρισης, ώστε μετά της εφαρμογής του εκπαιδευμένου αλγόριθμου επί των δεδομένων αυτών, να προκύπτει η σχετική πρόβλεψη – κατάταξη εκάστης πειραματικής μονάδας, σε μία εκ των τριών επιπέδων της απόκρισης 'οικογενειακή κατάσταση'. Το πλήθος των εσφαλμένων ταξινομήσεων, θα καθορίζει και το σφάλμα ταξινόμησης του αντιστοίχου υποσυνόλου δεδομένων επικύρωσης.

Με τον τρόπο αυτό, θα προκύψουν 6 ζεύγη δεδομένων εκπαίδευσης – επικύρωσης. Σε κάθε ένα από τα υποσύνολα δεδομένων εκπαίδευσης, θα πραγματοποιείται η εκπαίδευση του αλγόριθμου και θα προσδιορίζεται το σχετικό σφάλμα ταξινόμησης (*training error rate*). Εν συνεχεία, θα εφαρμόζεται ο εκπαιδευμένος αλγόριθμος, επί των αντίστοιχων δεδομένων επικύρωσης, ώστε να προκύψει το σχετικό σφάλμα ταξινόμησης (*test error rate*). Το σφάλμα ταξινόμησης το οποίο προκύπτει από τα δεδομένα εκπαίδευσης, είναι πάντοτε μικρότερο ή ίσο, του σφάλματος ταξινόμησης που προκύπτει από τα δεδομένα επικύρωσης. Αυτό σημαίνει πως, το σφάλμα ταξινόμησης υποεκτιμάται, όταν υπολογίζεται σε εκείνο το σύνολο δεδομένων, επί των οποίων έχει εκπαιδευτεί ο αλγόριθμος.

Το συνολικό σφάλμα ταξινόμησης (εκπαίδευσης και επικύρωσης), θα προκύψει ως ο σταθμισμένος μέσος όρος των προαναφερθέντων έξι σφαλμάτων ταξινόμησης (εκπαίδευσης και επικύρωσης αντίστοιχα).

γ. Αποτελέσματα Διαχωριστικής Ανάλυσης

Εν συνεχεία, θα εκπαιδεύσουμε τους αλγορίθμους *C_4.5* και *MLR*, σε κάθε υποσύνολο δεδομένων εκπαίδευσης, προσδιορίζοντας το σχετικό σφάλμα εκπαίδευσης και ακολούθως θα προσδιορίσουμε το σχετικό σφάλμα επικύρωσης, μέσω της εφαρμογής του εκπαιδευμένου αλγορίθμου, στο αντίστοιχο υποσύνολο δεδομένων επικύρωσης.

1ο ζεύγος συνόλων δεδομένων εκπαίδευσης και επικύρωσης

Αποτελέσματα Δεδομένων Εκπαίδευσης – Πίνακας Συνάφειας

Αλγόριθμος C_4.5

Correct	36.258	84,32%
Wrong	6.742	15,68%
Total	43.000	

	1.000000	**2.000000**	**3.000000**
1.000000	15.646	1.291	1.063
2.000000	596	17.232	1.172
3.000000	1.147	1.473	3.380

Αλγόριθμος MLR

Correct	36.218	84,23%
Wrong	6.782	15,77%
Total	43.000	

	1.000000	**2.000000**	**3.000000**
1.000000	15.399	1.365	1.236
2.000000	496	17.249	1.255
3.000000	1.009	1.421	3.570

Επιπρόσθετα, για το μοντέλο Πολυωνυμικής Λογιστικής Παλινδρόμησης, παραθέτουμε τους ακόλουθους ελέγχους επάρκειας και καλής προσαρμογής.

Σε επίπεδο σημαντικότητας $\alpha = 0.05$, το προσαρμοσμένο μοντέλο *MLR*, κρίνεται ως στατιστικά σημαντικό, διότι $p - value = 0.0001 < 0.05$

Model Fitting Information

Model	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	61587,188	61604,526	61583,188			
Final	10562,107	11116,920	10434,107	51149,081	62	,000

Στο ίδιο συμπέρασμα καταλήγουμε και με τον έλεγχο απόκλισης του μοντέλου

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	11633,351	9150	,000
Deviance	7154,299	9150	1,000

Pseudo R-Square

Cox and Snell	,696
Nagelkerke	,804
McFadden	,595

Στη προκειμένη περίπτωση προκύπτει $p - value = 1 > 0.05$, επομένως δε μπορεί να απορριφθεί η μηδενική υπόθεση περί επάρκειας και καλής προσαρμογής του μοντέλου στα πειραματικά δεδομένα.

Σε επίπεδο σημαντικότητας $\alpha = 0.05$, οι ερμηνευτικές μεταβλητές του μοντέλου, κρίνονται ως στατιστικά σημαντικές, διότι $p - value = 0.0001 < 0.05$

Step Summary

Model		Action	Effect(s)	Model Fitting Criteria			Effect Selection Tests		
				AIC	BIC	-2 Log Likelihood	Chi-Square(a,b)	df	Sig.
Step 0	0	Entered	Intercept	61587,188	61604,526	61583,188	.		
Step 1	1	Entered	household summary status	17193,546	17297,573	17169,546	36407,580	10	,000
Step 2	2	Entered	sex	12495,388	12616,753	12467,388	4568,261	2	,000
Step 3	3	Entered	education	11344,458	11552,513	11296,458	1078,422	10	,000
Step 4	4	Entered	weeks worked in year_transformed	10918,150	11282,247	10834,150	472,973	18	,000
Step 5	5	Entered	citizenship	10733,594	11149,704	10637,594	186,391	6	,000
Step 6	6	Entered	income level	10595,028	11028,475	10495,028	128,101	2	,000
Step 7	7	Entered	num persons worked for employer_transformed	10571,729	11091,866	10451,729	42,414	10	,000
Step 8	8	Entered	employment status_transformed	10562,107	11116,920	10434,107	17,708	4	,001

Stepwise Method: Forward Stepwise

a. The chi-square for entry is based on the score test.

b. The chi-square for removal is based on the likelihood ratio test.

Αποτελέσματα Δεδομένων Επικύρωσης – Πίνακας Συνάφειας

C_4.5 (Test Accuracy)	Predicted			
Actuals	1	2	3	Grand Total
1	5201	463	337	6001
2	193	5448	360	6001
3	1140	1516	3344	6000
Grand Total	6534	7427	4041	18002
C_4.5 Accuracy on Test Data set				77,73%

MLR (Test Accuracy)	Predicted			
Actuals	1	2	3	Grand Total
1	5120	477	404	6001
2	174	5460	367	6001
3	980	1461	3559	6000
Grand Total	6274	7398	4330	18002
MLR Accuracy on Test Data set				78,54%

2ο ζεύγος συνόλων δεδομένων εκπαίδευσης και επικύρωσης

Αποτελέσματα Δεδομένων Εκπαίδευσης – Πίνακας Συνάφειας

Αλγόριθμος C_4.5

Correct	34.862	85,03%
Wrong	6.138	14,97%
Total	41.000	

	1.000000	2.000000	3.000000
1.000000	14.827	1.192	981
2.000000	584	17.239	1.177
3.000000	929	1.275	2.796

Αλγόριθμος MLR

Correct	34.808	84,9%
Wrong	6.192	15,1%
Total	41.000	

	1.000000	2.000000	3.000000
1.000000	14.630	1.242	1.128
2.000000	509	17.258	1.233
3.000000	814	1.266	2.920

Επιπρόσθετα, για το μοντέλο Πολυωνυμικής Λογιστικής Παλινδρόμησης, παραθέτουμε τους ακόλουθους ελέγχους επάρκειας και καλής προσαρμογής

Σε επίπεδο σημαντικότητας $\alpha = 0.05$, το προσαρμοσμένο μοντέλο *MLR*, κρίνεται ως στατιστικά σημαντικό, διότι $p - value = 0.0001 < 0.05$

Model Fitting Information

Model	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	58065,673	58082,915	58061,673			
Final	9771,434	10323,199	9643,434	48418,238	62	,000

Στο ίδιο συμπέρασμα καταλήγουμε και με τον έλεγχο απόκλισης του μοντέλου

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	11855,927	9076	,000
Deviance	6600,502	9076	1,000

Pseudo R-Square

Cox and Snell	,693
Nagelkerke	,807
McFadden	,604

Στη προκειμένη περίπτωση προκύπτει $p - value = 1 > 0.05$, επομένως δε μπορεί να απορριφθεί η μηδενική υπόθεση περί επάρκειας και καλής προσαρμογής του μοντέλου στα πειραματικά δεδομένα.

Σε επίπεδο σημαντικότητας $\alpha = 0.05$, οι ερμηνευτικές μεταβλητές του μοντέλου, κρίνονται ως στατιστικά σημαντικές, διότι $p - value = 0.0001 < 0.05$

Step Summary

Model		Action	Effect(s)	Model Fitting Criteria			Effect Selection Tests		
				AIC	BIC	-2 Log Likelihood	Chi-Square(a,b)	df	Sig.
Step 0	0	Entered	Intercept	58065,673	58082,915	58061,673	.		
Step 1	1	Entered	household summary status	15519,242	15622,698	15495,242	34716,571	10	,000
Step 2	2	Entered	sex	11375,096	11495,794	11347,096	4085,960	2	,000
Step 3	3	Entered	education	10409,835	10616,747	10361,835	903,306	10	,000
Step 4	4	Entered	weeks worked in year_transformed	10058,662	10420,758	9974,662	395,756	18	,000
Step 5	5	Entered	citizenship	9893,565	10307,389	9797,565	167,006	6	,000
Step 6	6	Entered	income level	9817,209	10248,275	9717,209	74,082	2	,000
Step 7	7	Entered	num persons worked for employer_transformed	9786,126	10303,406	9666,126	50,080	10	,000
Step 8	8	Entered	employment status_transformed	9771,434	10323,199	9643,434	23,014	4	,000

Stepwise Method: Forward Stepwise

a. The chi-square for entry is based on the score test.

b. The chi-square for removal is based on the likelihood ratio test.

Αποτελέσματα Δεδομένων Επικύρωσης – Πίνακας Συνάφειας

C_4.5	Predicted			
Actuals	1	2	3	Grand Total
1	5204	425	372	6001
2	218	5433	350	6001
3	1199	1469	3332	6000
Grand Total	6621	7327	4054	18002
C_4.5 Accuracy on Test Data set				77,60%

MLR	Predicted			
Actuals	1	2	3	Grand Total
1	5111	446	444	6001
2	180	5445	376	6001
3	1035	1443	3522	6000
Grand Total	6326	7334	4342	18002
MLR Accuracy on Test Data set				78,20%

3ο ζεύγος συνόλων δεδομένων εκπαίδευσης και επικύρωσης

Αποτελέσματα Δεδομένων Εκπαίδευσης – Πίνακας Συνάφειας

Αλγόριθμος C_4.5

Correct	40.260	83,88%
Wrong	7.740	16,12%
Total	48.000	

	1.000000	2.000000	3.000000
1.000000	18.701	1.564	1.377
2.000000	549	17.693	1.309
3.000000	1.371	1.570	3.866

Αλγόριθμος MLR

Correct	40.363	84,09%
Wrong	7.637	15,91%
Total	48.000	

	1.000000	2.000000	3.000000
1.000000	18.439	1.604	1.599
2.000000	475	17.745	1.331
3.000000	1.050	1.578	4.179

Επιπρόσθετα, για το μοντέλο Πολυωνυμικής Λογιστικής Παλινδρόμησης, παραθέτουμε τους ακόλουθους ελέγχους επάρκειας και καλής προσαρμογής

Σε επίπεδο σημαντικότητας $\alpha = 0.05$, το προσαρμοσμένο μοντέλο *MLR*, κρίνεται ως στατιστικά σημαντικό, διότι $p - value = 0.0001 < 0.05$

Model Fitting Information

Model	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	68382,864	68400,422	68378,864			
Final	11255,460	11782,197	11135,460	57243,404	58	,000

Στο ίδιο συμπέρασμα καταλήγουμε και με τον έλεγχο απόκλισης του μοντέλου

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	13841,960	9788	,000
Deviance	7514,838	9788	1,000

Pseudo R-Square

Cox and Snell	,697
Nagelkerke	,805
McFadden	,595

Στη προκειμένη περίπτωση προκύπτει $p - value = 1 > 0.05$, επομένως δε μπορεί να απορριφθεί η μηδενική υπόθεση περί επάρκειας και καλής προσαρμογής του μοντέλου στα πειραματικά δεδομένα.

Σε επίπεδο σημαντικότητας $\alpha = 0.05$, οι ερμηνευτικές μεταβλητές του μοντέλου, κρίνονται ως στατιστικά σημαντικές, διότι $p - value = 0.0001 < 0.05$

Step Summary

Model		Action	Effect(s)	Model Fitting Criteria			Effect Selection Tests		
				AIC	BIC	-2 Log Likelihood	Chi-Square(a,b)	df	Sig.
Step 0	0	Entered	Intercept	68382,864	68400,422	68378,864	.		
Step 1	1	Entered	household summary status	18452,083	18557,431	18428,083	40981,916	10	,000
Step 2	2	Entered	sex	13197,643	13320,548	13169,643	5060,993	2	,000
Step 3	3	Entered	education	12085,880	12296,575	12037,880	1030,901	10	,000
Step 4	4	Entered	weeks worked in year_transformed	11614,762	11983,478	11530,762	520,704	18	,000
Step 5	5	Entered	citizenship	11410,515	11831,905	11314,515	207,070	6	,000
Step 6	6	Entered	income level	11308,595	11747,543	11208,595	98,548	2	,000
Step 7	7	Entered	num persons worked for employer_transformed	11255,460	11782,197	11135,460	72,187	10	,000

Stepwise Method: Forward Stepwise

a. The chi-square for entry is based on the score test.

b. The chi-square for removal is based on the likelihood ratio test.

Αποτελέσματα Δεδομένων Επικύρωσης – Πίνακας Συνάφειας

C_4.5	Predicted			
Actuals	1	2	3	Grand Total
1	4338	342	326	5006
2	161	4508	332	5001
3	954	1259	2787	5000
Grand Total	5453	6109	3445	15007
C_4.5 Accuracy on Test Data set				77,52%

MLR	Column Labels			
Row Labels	1	2	3	Grand Total
1	4288	358	360	5006
2	122	4528	351	5001
3	818	1219	2963	5000
Grand Total	5228	6105	3674	15007
MLR Accuracy on Test Data set				78,49%

4ο ζεύγος συνόλων δεδομένων εκπαίδευσης και επικύρωσης

Αποτελέσματα Δεδομένων Εκπαίδευσης – Πίνακας Συνάφειας

Αλγόριθμος C_4.5

Correct	33.975	84,94%
Wrong	6.025	15,06%
Total	40.000	

	1.000000	2.000000	3.000000
1.000000	14.812	1.336	852
2.000000	506	16.626	868
3.000000	895	1.568	2.537

Αλγόριθμος MLR

Correct	33.883	84,71%
Wrong	6.117	15,29%
Total	40.000	

	1.000000	2.000000	3.000000
1.000000	14.607	1.171	1.222
2.000000	446	16.321	1.233
3.000000	820	1.225	2.955

Επιπρόσθετα, για το μοντέλο Πολυωνυμικής Λογιστικής Παλινδρόμησης, παραθέτουμε τους ακόλουθους ελέγχους επάρκειας και καλής προσαρμογής

Σε επίπεδο σημαντικότητας $\alpha = 0.05$, το προσαρμοσμένο μοντέλο *MLR*, κρίνεται ως στατιστικά σημαντικό, διότι $p - value = 0.0001 < 0.05$

Model Fitting Information

Model	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	56961,455	56978,648	56957,455			
Final	9513,936	10064,121	9385,936	47571,519	62	,000

Στο ίδιο συμπέρασμα καταλήγουμε και με τον έλεγχο απόκλισης του μοντέλου

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	10447,705	8890	,000
Deviance	6350,654	8890	1,000

Pseudo R-Square

Cox and Snell	,696
Nagelkerke	,809
McFadden	,605

Στη προκειμένη περίπτωση προκύπτει $p - value = 1 > 0.05$, επομένως δε μπορεί να απορριφθεί η μηδενική υπόθεση περί επάρκειας και καλής προσαρμογής του μοντέλου στα πειραματικά δεδομένα.

Σε επίπεδο σημαντικότητας $\alpha = 0.05$, οι ερμηνευτικές μεταβλητές του μοντέλου, κρίνονται ως στατιστικά σημαντικές, διότι $p - value = 0.0001 < 0.05$

Step Summary

Model		Action	Effect(s)	Model Fitting Criteria			Effect Selection Tests		
				AIC	BIC	-2 Log Likelihood	Chi-Square(a,b)	df	Sig.
Step 0	0	Entered	Intercept	56961,455	56978,648	56957,455	.		
Step 1	1	Entered	household summary status	15101,268	15204,428	15077,268	34066,198	10	,000
Step 2	2	Entered	sex	11065,264	11185,616	11037,264	3954,886	2	,000
Step 3	3	Entered	education	10172,353	10378,672	10124,353	816,980	10	,000
Step 4	4	Entered	weeks worked in year_transformed	9844,426	10205,485	9760,426	373,055	18	,000
Step 5	5	Entered	citizenship	9630,611	10043,250	9534,611	210,824	6	,000
Step 6	6	Entered	income level	9551,663	9981,495	9451,663	75,738	2	,000
Step 7	7	Entered	employment status_transformed	9522,908	9987,126	9414,908	36,897	4	,000
Step 8	8	Entered	num persons worked for employer_transformed	9513,936	10064,121	9385,936	28,691	10	,001

Stepwise Method: Forward Stepwise

a. The chi-square for entry is based on the score test.

b. The chi-square for removal is based on the likelihood ratio test.

Αποτελέσματα Δεδομένων Επικύρωσης – Πίνακας Συνάφειας

C_4.5	Predicted			
Actuals	1	2	3	Grand Total
1	6060	516	425	7001
2	218	6357	425	7000
3	1363	1659	3785	6807
Grand Total	7641	8532	4635	20808
C_4.5 Accuracy on Test Data set				77,86%

MLR	predicted			
Actuals	1	2	3	Grand Total
1	5979	544	478	7001
2	181	6377	442	7000
3	1146	1663	3998	6807
Grand Total	7306	8584	4918	20808
MLR Accuracy on Test Data set				78,59%

5ο ζεύγος συνόλων δεδομένων εκπαίδευσης και επικύρωσης

Αποτελέσματα Δεδομένων Εκπαίδευσης – Πίνακας Συνάφειας

Αλγόριθμος C_4.5

Correct	35.404	84,3%
Wrong	6.596	15,7%
Total	42.000	

	1.000000	2.000000	3.000000
1.000000	15.546	1.189	1.265
2.000000	462	16.275	1.263
3.000000	1.094	1.323	3.583

Αλγόριθμος MLR

Correct	35.372	84,22%
Wrong	6.628	15,78%
Total	42.000	

	1.000000	2.000000	3.000000
1.000000	15.425	1.293	1.282
2.000000	441	16.376	1.183
3.000000	996	1.433	3.571

Επιπρόσθετα, για το μοντέλο Πολυωνυμικής Λογιστικής Παλινδρόμησης, παραθέτουμε τους ακόλουθους ελέγχους επάρκειας και καλής προσαρμογής

Σε επίπεδο σημαντικότητας $\alpha = 0.05$, το προσαρμοσμένο μοντέλο *MLR*, κρίνεται ως στατιστικά σημαντικό, διότι $p - value = 0.0001 < 0.05$

Model Fitting Information

Model	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	60705,342	60722,633	60701,342			
Final	10073,046	10626,353	9945,046	50756,296	62	,000

Στο ίδιο συμπέρασμα καταλήγουμε και με τον έλεγχο απόκλισης του μοντέλου

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	10795,854	9104	,000
Deviance	6700,918	9104	1,000

Pseudo R-Square

Cox and Snell	,701
Nagelkerke	,810
McFadden	,602

Στη προκειμένη περίπτωση προκύπτει $p - value = 1 > 0.05$, επομένως δε μπορεί να απορριφθεί η μηδενική υπόθεση περί επάρκειας και καλής προσαρμογής του μοντέλου στα πειραματικά δεδομένα.

Σε επίπεδο σημαντικότητας $\alpha = 0.05$, οι ερμηνευτικές μεταβλητές του μοντέλου, κρίνονται ως στατιστικά σημαντικές, διότι $p - value = 0.0001 < 0.05$

Step Summary

Model		Action	Effect(s)	Model Fitting Criteria			Effect Selection Tests		
				AIC	BIC	-2 Log Likelihood	Chi-Square(a,b)	df	Sig.
Step 0	0	Entered	Intercept	60705,342	60722,633	60701,342	.		
Step 1	1	Entered	household summary status	16466,205	16569,950	16442,205	36151,124	10	,000
Step 2	2	Entered	sex	11865,561	11986,597	11837,561	4452,346	2	,000
Step 3	3	Entered	education	10791,565	10999,055	10743,565	997,949	10	,000
Step 4	4	Entered	weeks worked in year_transformed	10369,896	10733,004	10285,896	469,374	18	,000
Step 5	5	Entered	citizenship	10186,262	10601,242	10090,262	183,651	6	,000
Step 6	6	Entered	income level	10098,275	10530,546	9998,275	86,037	2	,000
Step 7	7	Entered	num persons worked for employer_transformed	10076,537	10595,263	9956,537	40,930	10	,000
Step 8	8	Entered	employment status_transformed	10073,046	10626,353	9945,046	11,529	4	,021

Stepwise Method: Forward Stepwise

a. The chi-square for entry is based on the score test.

b. The chi-square for removal is based on the likelihood ratio test.

Αποτελέσματα Δεδομένων Επικύρωσης – Πίνακας Συνάφειας

C_4.5	Predicted			
Actuals	1	2	3	Grand Total
1	4330	352	319	5001
2	154	4527	319	5000
3	929	1275	2796	5000
Grand Total	5413	6154	3434	15001
C_4.5 Accuracy on Test Data set				77,68%

MLR	Predicted			
Actuals	1	2	3	Grand Total
1	4256	372	373	5001
2	139	4521	340	5000
3	814	1266	2920	5000
Grand Total	5209	6159	3633	15001
MLR Accuracy on Test Data set				77,97%

6ο ζεύγος συνόλων δεδομένων εκπαίδευσης και επικύρωσης

Αποτελέσματα Δεδομένων Εκπαίδευσης – Πίνακας Συνάφειας

Αλγόριθμος C_4.5

Correct	39.047	84,88%
Wrong	6.953	15,12%
Total	46.000	

	1.000000	2.000000	3.000000
1.000000	18.393	1.431	1.176
2.000000	647	17.221	1.132
3.000000	1.152	1.415	3.433

Αλγόριθμος MLR

Correct	38.894	84,55%
Wrong	7.106	15,45%
Total	46.000	

	1.000000	2.000000	3.000000
1.000000	18.150	1.494	1.356
2.000000	582	17.215	1.203
3.000000	1.038	1.433	3.529

Επιπρόσθετα, για το μοντέλο Πολυωνυμικής Λογιστικής Παλινδρόμησης, παραθέτουμε τους ακόλουθους ελέγχους επάρκειας και καλής προσαρμογής

Σε επίπεδο σημαντικότητας $\alpha = 0.05$, το προσαρμοσμένο μοντέλο *MLR*, κρίνεται ως στατιστικά σημαντικό, διότι $p - value = 0.0001 < 0.05$

Model Fitting Information

Model	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	65234,292	65251,765	65230,292			
Final	10739,938	11264,122	10619,938	54610,354	58	,000

Στο ίδιο συμπέρασμα καταλήγουμε και με τον έλεγχο απόκλισης του μοντέλου

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	67004,886	9532	,000
Deviance	7228,494	9532	1,000

Pseudo R-Square

Cox and Snell	,695
Nagelkerke	,807
McFadden	,600

Στη προκειμένη περίπτωση προκύπτει $p - value = 1 > 0.05$, επομένως δε μπορεί να απορριφθεί η μηδενική υπόθεση περί επάρκειας και καλής προσαρμογής του μοντέλου στα πειραματικά δεδομένα.

Σε επίπεδο σημαντικότητας $\alpha = 0.05$, οι ερμηνευτικές μεταβλητές του μοντέλου, κρίνονται ως στατιστικά σημαντικές, διότι $p - value = 0.0001 < 0.05$

Step Summary

Model		Action	Effect(s)	Model Fitting Criteria			Effect Selection Tests		
				AIC	BIC	-2 Log Likelihood	Chi-Square(a,b)	df	Sig.
Step 0	0	Entered	Intercept	65234,292	65251,765	65230,292	.		
Step 1	1	Entered	household summary status	17247,628	17352,464	17223,628	39079,071	10	,000
Step 2	2	Entered	sex	12581,477	12703,787	12553,477	4528,736	2	,000
Step 3	3	Entered	education	11466,656	11676,330	11418,656	1023,649	10	,000
Step 4	4	Entered	weeks worked in year_transformed	11059,326	11426,255	10975,326	452,265	18	,000
Step 5	5	Entered	citizenship	10865,265	11284,612	10769,265	194,132	6	,000
Step 6	6	Entered	income level	10769,799	11206,619	10669,799	91,896	2	,000
Step 7	7	Entered	num persons worked for employer_transformed	10739,938	11264,122	10619,938	49,901	10	,000

Stepwise Method: Forward Stepwise

a. The chi-square for entry is based on the score test.

b. The chi-square for removal is based on the likelihood ratio test.

Αποτελέσματα Δεδομένων Επικύρωσης – Πίνακας Συνάφειας

C_4.5	Predicted			
Actuals	1	2	3	Grand Total
1	5226	405	370	6001
2	189	5456	356	6001
3	1147	1473	3380	6000
Grand Total	6562	7334	4106	18002
C_4.5 Accuracy on Test Data set				78,11%

MLR	Predicted			
Actuals	1	2	3	Grand Total
1	5151	431	419	6001
2	153	5465	383	6001
3	1009	1421	3570	6000
Grand Total	6313	7317	4372	18002
MLR Accuracy on Test Data set				78,80%

δ. Σύνοψη και αξιολόγηση αποτελεσμάτων

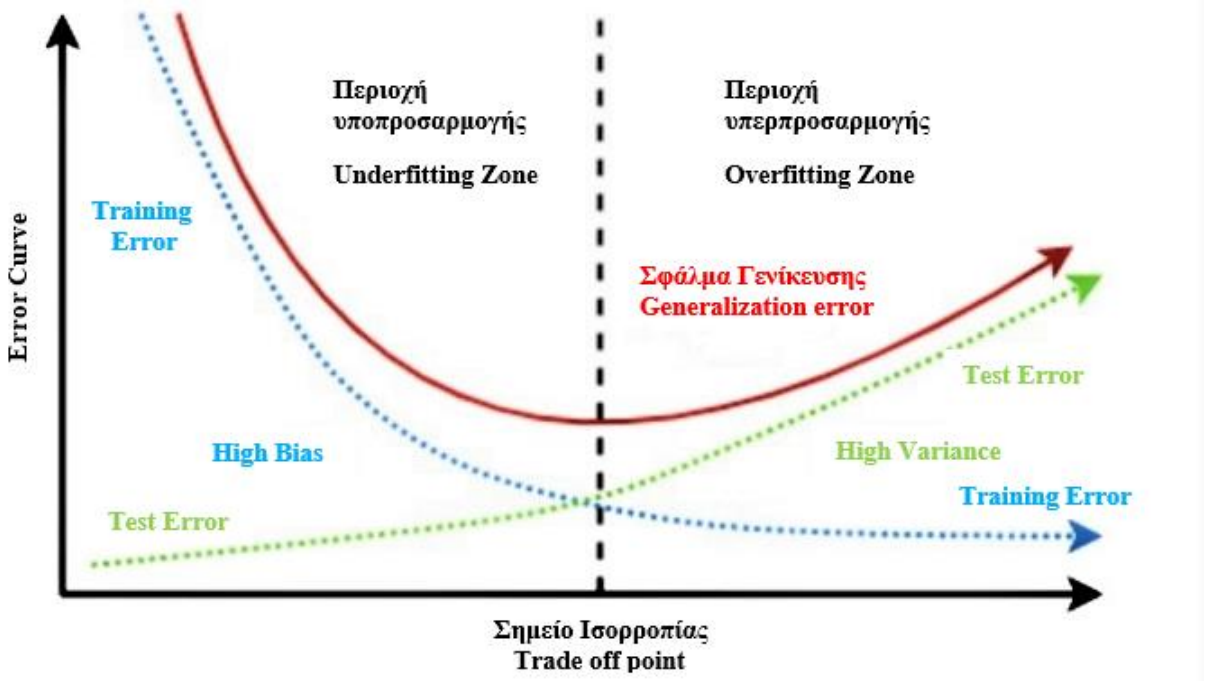
Συγκεντρωτικά, τα αποτελέσματα εφαρμογής των αλγορίθμων, στα έξι διαφορετικά ζεύγη συνόλων εκπαίδευσης και επικύρωσης, δίδονται στον ακόλουθο πίνακα

Model Accuracy	Algorithm C_4.5				Algorithm MLR			
	Training set	Sample size	Test set	Sample size	Training set	Sample size	Test set	Sample size
1st	84,32%	43.000	77,73%	18.000	84,23%	43.000	78,54%	18.000
2nd	85,03%	41.000	77,60%	18.000	84,90%	41.000	78,20%	18.000
3rd	83,88%	48.000	77,52%	15.000	84,09%	48.000	78,49%	15.000
4th	84,94%	40.000	77,86%	20.800	84,71%	40.000	78,59%	20.800
5th	84,30%	42.000	77,68%	15.000	84,22%	42.000	77,97%	15.000
6th	84,88%	46.000	78,11%	18.000	84,55%	46.000	78,80%	18.000
Weighted Averages	84,54%		77,76%		84,44%		78,45%	

Παρατηρούμε πως ο αλγόριθμος *C_4.5*, εμφάνισε ποσοστό ακρίβειας *84.54%* κατά την προσαρμογή του στα δεδομένα εκπαίδευσης, οριακά μεγαλύτερο του ποσοστού *84.44%* που σημείωσε ο αλγόριθμος *MLR*. Αναφορικά με την ακρίβεια των αποτελεσμάτων επί των δεδομένων επικύρωσης, παρατηρούμε πως ο αλγόριθμος *MLR* σημείωσε ελαφρώς καλύτερο ποσοστό, της τάξεως του *78.44%*, έναντι του αλγορίθμου *C_4.5*, ο οποίος σημείωσε ποσοστό *77.75%*. Με διαφορετικά αλλά ισοδύναμα λόγια, αυτό σημαίνει πως ο αλγόριθμος *MLR*, εμφάνισε ελαφρώς καλύτερη ικανότητα γενίκευσης και προσαρμογής επί συνόλων δεδομένων, στα οποία δεν έχει εκπαιδευτεί.

Συνολικά τα αποτελέσματα των δύο αλγορίθμων *C_4.5* και *MLR* κρίνονται ως ισοδύναμα, τα δε προσαρμοσμένα μοντέλα εμφανίζουν ισορροπημένη ικανότητα γενίκευσης σε άγνωστα δεδομένα, χωρίς ωστόσο να υστερούν στην ικανότητα ακριβούς προσαρμογής στα δεδομένα εκπαίδευσης, γεγονός το οποίο αποτελεί προαπαιτούμενο ενός αξιόπιστου μοντέλου. Στο σημείο αυτό αξίζει να σημειωθεί πως, το σημειούμενο σφάλμα επί δεδομένων εκπαίδευσης και το σφάλμα επί δεδομένων επικύρωσης, είναι μεταξύ των ανταγωνιστικά. Δηλαδή πιθανή προσπάθεια μείωσης του ενός, οδηγεί σε αύξηση του άλλου και αντιστρόφως.

Στην παρούσα μελέτη περίπτωσης, παρατηρούμε πως και στους δύο αξιολογούμενους αλγορίθμους, τα σχετικώς προσαρμοσμένα μοντέλα, χαρακτηρίζονται από ισορροπία μεταξύ των δύο ανταγωνιστικών αυτών σφαλμάτων.



Το σφάλμα επί δεδομένων εκπαίδευσης (*training error*), σχετίζεται με την ικανότητα του αλγορίθμου να προσαρμόζεται ικανοποιητικά στα δεδομένα εκπαίδευσης. Υψηλή τιμή στο εν λόγω σφάλμα, ερμηνεύεται ως πρόβλημα υποπροσαρμογής (*underfitting*) του μοντέλου στα δεδομένα (*high biased problem*). Επιπροσθέτως, το σφάλμα επί δεδομένων επικύρωσης (*test error*), συνδέεται με την ικανότητα του αλγορίθμου να προσαρμόζεται ικανοποιητικά επί των δεδομένων επικύρωσης. Υψηλή τιμή στο εν λόγω σφάλμα, ερμηνεύεται ως μειωμένη ικανότητα του αλγορίθμου να γενικεύσει σε άγνωστα δεδομένα στα οποία δεν έχει εκπαιδευτεί και εκλαμβάνεται ως ισχυρή ένδειξη υπερπροσαρμογής (*overfitting*) στα δεδομένα εκπαίδευσης (*high variance problem*).

ε. Μοντελοποίηση μέσω διαχείρισης ελλειπουσών τιμών

Προκειμένου να βελτιώσουμε τα αποτελέσματα της ανάλυσης, θα εφαρμόσουμε εναλλακτική μοντελοποίηση στην οποία θα πραγματοποιήσουμε διαχείριση των ελλειπουσών τιμών του δείγματος. Στη πρώτη φάση μοντελοποίησης, διατηρήθηκαν στην ανάλυση οκτώ μεταβλητές, οι οποίες εμφάνιζαν πληρότητα εγγραφών στο σύνολο των πειραματικών μονάδων του δείγματος. Στην παρούσα εναλλακτική μοντελοποίηση, πέραν των προαναφερθέντων οκτώ μεταβλητών, θα διατηρηθούν στην ανάλυση εκείνες οι μεταβλητές, οι οποίες εμφανίζουν ποσοστό ελλειπουσών τιμών μικρότερο ή ίσο του *50%*. Οι μεταβλητές αυτές είναι οι ακόλουθες

Εξαρτημένη Μεταβλητή Απόκρισης			
id	**Όνομα Μεταβλητής**	**Τύπος**	**Πλήθος Επιπέδων**
1	Οικογενειακή Κατάσταση (*marital status*)	*Nominal*	3
Ανεξάρτητες Ερμηνευτικές Μεταβλητές (χωρίς ελλείπουσες τιμές)			
id	**Όνομα Μεταβλητής**	**Τύπος**	**Πλήθος Επιπέδων**
1	Μορφωτικό Επίπεδο (*education*)	*Nominal*	6
2	Φύλο (*sex*)	*Nominal*	2
3	Κατάσταση Νοικοκυριού (*household status*)	*Nominal*	6
4	Υπηκοότητα (*citizenship*)	*Nominal*	4
5	Εισοδηματικό Επίπεδο (i*ncome level*)	*Nominal*	2
6	Εργασιακή Κατάσταση (*employment status*)	*Nominal*	3
7	Πλήθος εργαζομένων στην επιχείρηση απασχόλησης (*number of persons worked for employer*)	*Nominal*	6
8	Αριθμός εργάσιμων εβδομάδων ανά έτος *(number of weeks worked in year)*	*Nominal*	10
Ανεξάρτητες Ερμηνευτικές Μεταβλητές (με ελλείπουσες τιμές)			
9	Κατηγορία εργαζομένου *(class of worker)*	*Nominal*	5
10	Κλάδος Οικονομικής Δραστηριότητας *(major industry code)*	*Nominal*	4
11	Φυλή (race)	*Nominal*	4
12	Φορολογική Κατάσταση *(tax filer status)*	*Nominal*	3
13	Χώρα Γέννησης *(country of birth self)*	*Nominal*	4
14	Ηλικία *(age)*	*Continuous*	-

Συνολικά στο επαυξημένο μοντέλο θα συμπεριληφθούν *14* ερμηνευτικές μεταβλητές, εφαρμόζοντας τους ακόλουθους κανόνες διαχείρισης ελλειπουσών τιμών, προκειμένου για τις έξι τελευταίες εξ' αυτών, οι οποίες εμφανίζουν ποσοστό ελλειπουσών τιμών, μικρότερο ή ίσο του *50%*

- ✓ Προκειμένου για τη συνεχή μεταβλητή 'Ηλικία' (*Age*), συμπλήρωση των ελλειπουσών τιμών, με τη μέση τιμή (*mean*) του χαρακτηριστικού
- ✓ Για τα υπόλοιπα κατηγορικά χαρακτηριστικά, συμπλήρωση των ελλειπουσών τιμών, με την αντίστοιχη επικρατούσα τιμή (*mode*), κάθε χαρακτηριστικού

Επιπλέον για τη συνεχή μεταβλητή 'Ηλικία' (*Age*), θα προχωρήσουμε σε κανονικοποίηση (*z- transformation*), δηλαδή αφαιρώντας τη μέση τιμή του χαρακτηριστικού και διαιρώντας με την αντίστοιχη τυπική απόκλιση, θα δημιουργήσουμε ένα νέο συνεχές χαρακτηριστικό, με μέση τιμή μηδέν και τυπική απόκλιση ένα.

Αξιοποιώντας τα παραπάνω έξι ζεύγη συνόλων δεδομένων εκπαίδευσης και επικύρωσης, θα επαναλάβουμε την όλη διαδικασία, καταγράφοντας τα προκύπτοντα αποτελέσματα.

1ο ζεύγος συνόλων δεδομένων εκπαίδευσης και επικύρωσης

Αποτελέσματα Δεδομένων Εκπαίδευσης – Πίνακας Συνάφειας

Αλγόριθμος C_4.5

Correct	40.117	93,3%
Wrong	2.883	6,7%
Total	43.000	

	1.000000	2.000000	3.000000
1.000000	16.889	200	911
2.000000	231	18.320	449
3.000000	736	356	4.908

Αλγόριθμος MLR

Correct	39.685	92,29%
Wrong	3.315	7,71%
Total	43.000	

	1.000000	2.000000	3.000000
1.000000	16.899	227	874
2.000000	330	18.221	449
3.000000	1.055	380	4.565

Αποτελέσματα Δεδομένων Επικύρωσης – Πίνακας Συνάφειας

C_4.5 (Test Accuracy) Predicted				
Actuals	**1**	**2**	**3**	**Grand Total**
1	5602	73	326	6001
2	94	5751	156	6001
3	893	423	4684	6000
Grand Total	6589	6247	5166	18002
C_4.5 Accuracy on Test Data set				**89,08%**

MLR (Test Accuracy) Predicted				
Actuals	**1**	**2**	**3**	**Grand Total**
1	5654	83	264	6001
2	120	5730	151	6001
3	1127	382	4491	6000
Grand Total	6901	6195	4906	18002
MLR Accuracy on Test Data set				**88,18%**

2ο ζεύγος συνόλων δεδομένων εκπαίδευσης και επικύρωσης

Αποτελέσματα Δεδομένων Εκπαίδευσης – Πίνακας Συνάφειας

Αλγόριθμος C_4.5

Correct	38.492	93,88%
Wrong	2.508	6,12%
Total	41.000	

	1.000000	2.000000	3.000000
1.000000	15.991	171	838
2.000000	198	18.375	427
3.000000	585	289	4.126

Αλγόριθμος MLR

Correct	37.856	92,33%
Wrong	3.144	7,67%
Total	41.000	

	1.000000	2.000000	3.000000
1.000000	15.966	241	793
2.000000	335	18.281	384
3.000000	1.034	357	3.609

Αποτελέσματα Δεδομένων Επικύρωσης – Πίνακας Συνάφειας

C_4.5	Predicted			
Actuals	1	2	3	Grand Total
1	5560	54	387	6001
2	98	5761	142	6001
3	923	416	4661	6000
Grand Total	6581	6231	5190	18002
C_4.5 Accuracy on Test Data set				88,78%

MLR	Predicted			
Actuals	1	2	3	Grand Total
1	5631	68	302	6001
2	120	5762	119	6001
3	1209	482	4309	6000
Grand Total	6960	6312	4730	18002
MLR Accuracy on Test Data set				87,22%

3ο ζεύγος συνόλων δεδομένων εκπαίδευσης και επικύρωσης

Αποτελέσματα Δεδομένων Εκπαίδευσης – Πίνακας Συνάφειας

Αλγόριθμος C_4.5

Correct	44.866	93,47%
Wrong	3.134	6,53%
Total	48.000	

	1.000000	2.000000	3.000000
1.000000	20.468	209	965
2.000000	236	18.904	411
3.000000	891	422	5.494

Αλγόριθμος MLR

Correct	44.190	92,06%
Wrong	3.810	7,94%
Total	48.000	

	1.000000	2.000000	3.000000
1.000000	20.358	263	1.021
2.000000	333	18.779	439
3.000000	1.282	472	5.053

Αποτελέσματα Δεδομένων Επικύρωσης – Πίνακας Συνάφειας

C_4.5	Predicted			
Actuals	1	2	3	Grand Total
1	4662	48	296	5006
2	75	4798	128	5001
3	816	352	3832	5000
Grand Total	5553	5198	4256	15007
C_4.5 Accuracy on Test Data set				88,57%

MLR	Column Labels			
Row Labels	1	2	3	Grand Total
1	4690	55	261	5006
2	95	4793	113	5001
3	934	326	3740	5000
Grand Total	5719	5174	4114	15007
MLR Accuracy on Test Data set				88,11%

4ο ζεύγος συνόλων δεδομένων εκπαίδευσης και επικύρωσης

Αποτελέσματα Δεδομένων Εκπαίδευσης – Πίνακας Συνάφειας

Αλγόριθμος C_4.5

Correct	37.629	94,07%
Wrong	2.371	5,93%
Total	40.000	

	1.000000	2.000000	3.000000
1.000000	16.050	174	776
2.000000	213	17.481	306
3.000000	517	385	4.098

Αλγόριθμος MLR

Correct	36.913	92,28%
Wrong	3.087	7,72%
Total	40.000	

	1.000000	2.000000	3.000000
1.000000	15.957	228	815
2.000000	341	17.300	359
3.000000	952	392	3.656

Αποτελέσματα Δεδομένων Επικύρωσης – Πίνακας Συνάφειας

C_4.5	Predicted			
Actuals	1	2	3	Grand Total
1	6498	89	414	7001
2	97	6773	130	7000
3	1070	627	5110	6807
Grand Total	7665	7489	5654	20808
C_4.5 Accuracy on Test Data set				88,34%

MLR	predicted			
Actuals	1	2	3	Grand Total
1	6561	102	338	7001
2	122	6743	135	7000
3	1363	572	4872	6807
Grand Total	8046	7417	5345	20808
MLR Accuracy on Test Data set				87,35%

5ο ζεύγος συνόλων δεδομένων εκπαίδευσης και επικύρωσης

Αποτελέσματα Δεδομένων Εκπαίδευσης – Πίνακας Συνάφειας

Αλγόριθμος C_4.5

Correct	39.106	93,11%
Wrong	2.894	6,89%
Total	42.000	

	1.000000	2.000000	3.000000
1.000000	16.769	186	1.045
2.000000	201	17.352	447
3.000000	681	334	4.985

Αλγόριθμος MLR

Correct	38.716	92,18%
Wrong	3.284	7,82%
Total	42.000	

	1.000000	2.000000	3.000000
1.000000	16.893	202	905
2.000000	265	17.330	405
3.000000	1.101	406	4.493

Αποτελέσματα Δεδομένων Επικύρωσης – Πίνακας Συνάφειας

C_4.5	Predicted			
Actuals	1	2	3	Grand Total
1	4572	75	354	5001
2	56	4817	127	5000
3	675	301	4024	5000
Grand Total	5303	5193	4505	15001
C_4.5 Accuracy on Test Data set				89,41%

MLR	Predicted			
Actuals	1	2	3	Grand Total
1	4641	69	291	5001
2	74	4823	103	5000
3	940	336	3724	5000
Grand Total	5655	5228	4118	15001
MLR Accuracy on Test Data set				87,91%

6ο ζεύγος συνόλων δεδομένων εκπαίδευσης και επικύρωσης

Αποτελέσματα Δεδομένων Εκπαίδευσης – Πίνακας Συνάφειας

Αλγόριθμος C_4.5

Correct	42.816	93,08%
Wrong	3.184	6,92%
Total	46.000	

	1.000000	2.000000	3.000000
1.000000	19.785	205	1.010
2.000000	249	18.248	503
3.000000	882	335	4.783

Αλγόριθμος MLR

Correct	41.339	89,87%
Wrong	4.661	10,13%
Total	46.000	

	1.000000	2.000000	3.000000
1.000000	19.152	272	1.576
2.000000	309	18.147	544
3.000000	1.630	330	4.040

Αποτελέσματα Δεδομένων Επικύρωσης – Πίνακας Συνάφειας

C_4.5	Predicted			
Actuals	1	2	3	Grand Total
1	5626	59	316	6001
2	84	5751	166	6001
3	939	368	4693	6000
Grand Total	6649	6178	5175	18002
C_4.5 Accuracy on Test Data set				89,27%

MLR	Predicted			
Actuals	1	2	3	Grand Total
1	5468	73	460	6001
2	102	5724	175	6001
3	1668	344	3988	6000
Grand Total	7238	6141	4623	18002
MLR Accuracy on Test Data set				84,32%

Συγκριτικά, τα αποτελέσματα των δύο εναλλακτικών προσεγγίσεων, έχουν ως εξής

Model Accuracy	Algorithm C_4.5			
	Training set		Test set	
	with missing values	w/o missing values	with missing values	w/o missing values
1st	93,30%	84,32%	89,08%	77,73%
2nd	93,88%	85,03%	88,78%	77,60%
3rd	93,47%	83,88%	88,57%	77,52%
4th	94,07%	84,94%	88,34%	77,86%
5th	93,11%	84,30%	89,41%	77,68%
6th	93,08%	84,88%	89,27%	78,11%
Weighted Averages	93,47%	84,54%	88,89%	77,76%
Difference	8,93%		11,13%	

Model Accuracy	Algorithm MLR			
	Training set		Test set	
	with missing values	w/o missing values	with missing values	w/o missing values
1st	92,30%	84,23%	88,18%	78,54%
2nd	92,33%	84,90%	87,22%	78,20%
3rd	92,06%	84,09%	88,11%	78,49%
4th	92,28%	84,71%	87,35%	78,59%
5th	92,18%	84,22%	87,91%	77,97%
6th	89,87%	84,55%	84,32%	78,80%
Weighted Averages	91,81%	84,44%	87,14%	78,45%
Difference	7,37%		8,69%	

Ο αλγόριθμος *C_4.5*, μετά της εντάξεως στα δεδομένα ερμηνευτικών μεταβλητών με υψηλό ποσοστό ελλειπουσών τιμών και σχετικής εφαρμογής μεθόδων διαχείρισής τους, βελτίωσε την ακρίβεια του επί των δεδομένων εκπαίδευσης, από *84.54%* σε *93.47%*, ή ποσοστό *8.93%*. Επιπλέον, βελτίωσε την ικανότητα γενίκευσης σε άγνωστα δεδομένα, από *77.76%* σε *88.89%*, ή ποσοστό *11.13%*.

Αντιστοίχως, ο αλγόριθμος *MLR* βελτίωσε την ακρίβεια του επί των δεδομένων εκπαίδευσης, από *84.44%* σε *91.81%*, ή ποσοστό *7.37%*. Επιπλέον, βελτίωσε την ικανότητα γενίκευσης σε άγνωστα δεδομένα, από *78.45%* σε *87.14%*, ή ποσοστό *8.69%*.

Επιπρόσθετα, επαληθεύτηκε η ικανότητα του αλγορίθμου *C_4.5*, να διαχειρίζεται αποτελεσματικότερα δεδομένα με υψηλό ποσοστό ελλειπουσών τιμών, σε σχέση με τον αλγόριθμο *MLR*. Με σχετικά κριτήρια συγκρίσεως, η βελτίωση της ακρίβειας του *C_4.5*, επί των δεδομένων εκπαίδευσης, ήταν *1.56%* μεγαλύτερη της βελτίωσης που σημείωσε ο *MLR*. Επί των δεδομένων επικύρωσης, η σχετική βελτίωση του ποσοστού ακρίβειας ήταν *2.44%* μεγαλύτερη στον αλγόριθμο *C_4.5*, έναντι του αλγόριθμου *MLR*.

7 Βιβλιογραφία

7.1 Ελληνική

Dunham (2003). Επιμέλεια Ελληνικής έκδοσης: Βασίλης Βερύκιος & Γιάννης Θεοδωρίδης (2004-05). *Data Mining – Introductory and Advanced Topics*. Prentice Hall.

Γιάννης Θεοδωρίδης, Νίκος Πελέκης, (2016). '*Στατιστικές Μέθοδοι Εξόρυξης Δεδομένων'*, Πανεπιστημιακές Σημειώσεις, ΠΜΣ 'Εφαρμοσμένη Στατιστική', Πανεπιστήμιο Πειραιώς, Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης.

7.2 Ξένη

Albayrak, A. S. (2009). 'Classification of Domestic and Foreign Commercial Banks in Turkey Based on Financial Efficiency: A Comparison of Decision Tree, Logistic regression and Discriminant Analysis Models'. *Journal of Faculty of Economics and Administrative Sciences, 14(2),* 113–139.

Antonogeorgos, G., Demosthenes, B., Kostas, N., Anastasia, T. (2009). 'Logistic Regression and Linear Discriminant Analyses in Evaluating Factors Associated with Asthma Prevalence: Divergence and Similarity of the Two Statistical Methods'. *International Journal of Pediatrics*.

Anderson, T. W. (2003). '*An Introduction to Multivariate Statistical Analysis'*, Third Edition. J. Wiley & Sons, New York.

Agresti Alan. (2002). *'Categorical Data Analysis'*, 2nd Edition, University of Florida. J. Wiley & Sons, New York.

Begg, C. B., & Gray, R. (1984). 'Calculation of polychotomous logistic regression parameters using individualized regressions'. *Biometrika, 71*.

Brace, N., Kemp, R and Snelgar, R. (2009). '*SPSS for Psychologists'*, 4th Edition. UK: Palgrave MacMillan.

Breiman, Leo, Jerome Friedman, R. Olshen and C. Stone (1984). '*Classification and Regression Trees*'. Belmont, California: Wadsworth.

Cooley, W.W. and Lohnes, P.R. (1962). '*Multivariate Procedures for the Behavioral sciences'*. J. Wiley & Sons, New York.

Chatterjee S, and Hadi A (2006). '*Regression Analysis by Example'*. J. Wiley & Sons, New York.

Chandran, Raj Kumar. (2009). 'The Effectiveness of Stepwise Discriminant Analysis'. *Antimicrob Agents Chemother* 53(7), 2887–2891.

Demaris, A. (1992). '*Logit modeling: Practical applications'*. Newbury Park, CA: Sage.

Dinesh R. Pai. (2009). '*Determining the Efficacy of Mathematical Programming Approaches for Multi-Group Classification*'. Rutgers University Electronic Theses and Dissertations.

Fraas, J. W. and Newman, Is. (2003).'Ordinary Least Squares Regression, Discriminant Analysis, and Logistic Regression', Questions Researchers and Practitioners Should Address When Selecting an Analytic Technique. *Paper Presented at the Annual Meeting of the Eastern Educational Research Association. (Hilton Head Island, GA, February 26-March 1,2003).*

Geoffry J. Mclachlan. (1992). '*Discriminant Analysis and Statistical Pattern Recognition'*. The University of Queensland. J. Wiley & Sons, New York.

Hosmer, D. W. and Lemeshow, S. (2013). *'Applied linear regression'*, 3rd Edition. Johnson Wiley & Sons, Canada.

Hilbe, J. M. (2009). *'Logistic regression models'*. Chapman & Hall/CRC, New York

Ingersoll, G. M. & Orr, D. P. (1989). 'Behavioral and emotional risk in early adolescents'. *Journal of Early Adolescence,* 9.

Ingersoll, G. M., Grizzle, K., Beiter, M. & Orr, D. P. (1993). 'Frequent somatic complaints and psychosocial risk in adolescents' *Journal of Early Adolescence,* 13(1)

Kutner, M., Nachtsheim, C., Neter, J. (2004). '*Applied Linear Statistical Models'* Fifth Edition. New York, Americas.

Klecka, W. R. (1980). *'Discriminant analysis' (Quantitative Applications in the Social Sciences),* Thousand Oaks, CA: Sage Publications.

Khattree, R. and Naik, D.N, (1995). '*Applied Multivariate statistics with SAS Software*'. Cary NC. SAS Institute Inc.

Kachigan, S.K. (1991). *'Multivariate Statistical Analysis'*. New York, Radius Press, 25.

Lachenbruch, P. A. (1975). *'Discriminant Analysis'*. Hafner Press New York.

Maxwell, K. L. H. (2009). Master thesis: *'Logistic Regression Analysis to Determine the Significant Factors Associated with Substance Abuse in School-Aged Children'*. Georgia State University.

Mertler, C. and Vannatta, R. (2002). '*Advanced and multivariate statistical methods'*, 2nd edition. Los Angeles, CA: Pyrczak Publishing.

Onyeagu, S. I. (2003). '*A First Course in Multivariate Statistical Analysis'*. Mega Concept, Awka, Anambra State.

Peterson, B. & Harrell, F. (1990). 'Partial proportional odds models for ordinal response variables'. *Applied Statistics, 39.*

Poulsen J., French A. (2004). '*Discriminant Function Analysis (DA)'*. Retrieved August 18, 2014, from: http://userwww.sfsu.edu/classes/biol710/discrim/discrim.pdf.

Press, S. J. and Wilson, S. (1978). 'Choosing between logistic regression and discriminant analysis'. *Journal of the American Statistical Association*, 73.

Resnick, M. D., Harris, L. J., & Blum R.B. (1993). 'The impact of caring and connectedness on adolescent health and well-being'. *Journal of Pediatrics Child Health*, 29 (1).

Rencher, A. C. and William F. Christensen (2012). '*Methods of Multivariate Analysis'*, 3rd Edition. John Wiley & Sons, New York.

Riemann C., Filzmoser P., Garrett R., Dutter R. (2008). '*Statistical Data Analysis Explained. Applied Environmental Statistics with R'*. John Wiley & Sons, New York.

Rosenberg, M. (1965). '*Society and the adolescent self -image'*. Princeton, NJ: Princeton University Press.

Rodríguez, G. (2007). *'Lecture Notes on Generalized Linear Models'*. from: http://data.princeton.edu/wws509/notes/.

Scott, M. (2002), '*Applied Logistic Regression Analysis'*, 2nd edition. Sage University Paper series on Quantitative Applications in the Social Sciences, No.07-106, Beverly Hills, CA, Sage.

Simar, L., Hardle, W. (2000). '*Applied Multivariate Statistical Analysis'*, 3rd Edition. Springer, Berlin, Germany.

Siegel, S., & Castellan, N. J. (1988). '*Non-parametric statistics for the behavioral science'*, 2nd Edition. McGraw-Hill, New York.

StatSoft, Inc., (1984 – 2000). *'Discriminant Function Analysis'*. Electronic textbook. http://www.uta.edu/faculty/sawasthi/Statistics/stdiscan.html.

Statgun (2008). *Statgun statistics*. from: www.statgun.com/tutorials/logistic-regression.html.

Sonquist, J. A., Morgan, J. N. (1964). '*The Detection of Interaction Effects'*. Survey Research Center, University of Michigan.

Tabachnick, B.G., and Fidell, L. S. (1996). '*Using multivariate statistics'*, 3rd Edition. New York: Harper Collins.

Pang-Ning Tan, Michael Steinbach, Vipin Kumar (2006): '*Introduction to Data Mining'*. Pearson Addison-Wesley.

Timm Neil H. (2002). '*Applied Multivariate Analysis'*. Springer, Berlin, Germany.

Jacob, S. and Cohen, P. (1975). '*Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences'.* University of Michigan, Lawrence Erlbaum Associates, United States.

Richard A Johnson, Wichern, Dean W. (2007). '*Applied Multivariate Statistical Analysis'*, 6th Edition. New Jersey, Prentice Hall.

J. Ross Quinlan. Inc. (1986). '*Induction of Decision Trees - Machine Learning Theory'*. Kluwer Academic Publishers, Boston - Manufactured in The Netherlands

J. Ross Quinlan. Inc. (1993). '*C4.5: Programs for Machine Learning'*. Morgan Kaufmann Publishers, Inc, San Mateo, California.

Wehrens, Ron. (2010). '*Chemometrics with R Multivariate Data Analysis in the Natural Sciences and Life Sciences'*. Springer, Berlin, Germany.

Wang Yingjin. (2008). '*Comparing Linear Discriminant Analysis with Classification Trees Using Forest Landowner Survey Data as a Case Study'*. Master Thesis. The University of Tennessee. Knoxville.

Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg (2007). '*Top 10 algorithms in data mining'*. Springer-Verlag London Limited.