ΠΑΝΤΕΙΟΝ ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΟΙΝΩΝΙΚΩΝ ΚΑΙ ΠΟΛΙΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

PANTEION UNIVERSITY OF SOCIAL AND POLITICAL SCIENCES

ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΗΣ & ΠΕΡΙΦΕΡΕΙΑΚΗΣ ΑΝΑΠΤΥΞΗΣ

ΠΜΣ ΕΦΗΡΜΟΣΜΕΝΩΝ ΟΙΚΟΝΟΜΙΚΩΝ ΚΑΙ ΔΙΟΙΚΗΣΗΣ

Η ΜΕΘΟΔΟΣ C&RT (CLASSIFICATION AND REGRESSION TREES) ΣΤΗΝ ΤΑΞΙΝΟΜΗΣΗ ΚΑΙ ΑΝΑΛΥΣΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Παναγιώτης ΝΙΚΟΛΑΪΔΗΣ

Αθήνα Μάρτιος 2020

Τριμελή Επιτροπή

Clive Richardson Ομότιμος Καθηγητής Παντείου Πανεπιστημίου (Επιβλέπων)

Σταύρος ΝΤΕΓΙΑΝΑΚΗΣ Αναπληρωτής Καθηγητής Παντείου Πανεπιστημίου
Γρηγόριος ΣΙΟΥΡΟΥΝΗΣ Επίκουρος Καθηγητής Παντείου Πανεπιστημίου

Περίληψη

Με τον όρο Classification and Regression Tree (CaRT), εννοείται μία μέθοδος μηχανικής εκμάθησης, κατά την οποία χρησιμοποιείται ένα ορισμένο σύνολο δεδομένων, προκειμένου να συσταθεί ένα μοντέλο πρόβλεψης. Η διαδικασία που ακολουθείται, βασίζεται στον επαναλαμβανόμενο διαχωρισμό των διαθέσιμων παρατηρήσεων και στην εφαρμογή ενός μη πολύπλοκου προβλεπτικού αλγορίθμου σε κάθε διαχωρισμένο υποσύνολο. Ο αλγόριθμος που δημιουργείται, παρουσιάζεται με τη μορφή ενός δέντρου απόφασης το οποίο αποτελείται από ρίζα, διακλαδώσεις και φύλλα. Στην παρούσα εργασία, πραγματοποιήθηκε αρχικά θεωρητική παρουσίαση της μεθόδου CaRT και εν συνεχεία ανάπτυξη παραδείγματος με τη χρήση του στατιστικού λογισμικού πακέτου R-statistic με τις μεθόδους ταξινόμησης Classification Tree και Logistic Regression. Στόχος της εργασίας είναι αφενός, η παρουσίαση της μεθόδου CaRT, αφετέρου, η εξέταση της προβλεπτικής της ικανότητας σε σχέση με παραδοσιακές μεθόδους ανάλυσης. Προς τον σκοπό αυτό, κατά την ανάπτυξη του παραδείγματος, υλοποιήθηκε σύγκριση αποτελεσμάτων που προέκυψαν από την μέθοδο CaRT, των αποτελεσμάτων από την μέθοδο Logistic Regression καθώς και της και υφιστάμενης ανάλυσης, που διενεργήθηκε με την μέθοδο της Λογιστικής Παλινδρόμησης το 2015 με την εργασία των A. Fotiou, E. Kanavou, M. Stavrou, C. Richardson και A. Kokkevi " Prevalence and correlates of electronic cigarette use among adolescents in Greece: A preliminary cross-sectional analysis of nationwide survey data" . Από τα αποτελέσματα προέκυψαν οι μεταβλητές που καθορίζουν την χρήση του παραδοσιακού και του ηλεκτρονικού τσιγάρου, οι οποίοι είναι το φύλο, η χρήση άλλων ουσιών όπως η κάνναβη και το αλκοόλ, η επίβλεψη της οικογένειας, καθώς και η χρήση τσιγάρου από τα άτομα που συναναστρέφεται ο νέος. Επιπλέον και οι τρεις μέθοδοι οδήγησαν σε παρόμοια αποτελέσματα, κάτι το οποίο ενισχύει την λειτουργικότητα της μεθόδου Classification and Regression Tree.

Summary

By Classification and Regression Tree (CaRT), we mean a machine learning method, in which a certain set of data is used to form a prediction model. The procedure which is followed, is based on the repeated division of the available observations and the application of a non-complex predictive algorithm to each separated subset. The algorithm that is created is presented in the form of a decision tree consisting of roots, branches and leaves. The present work, presents the theory of the CaRT method and then an example, using the R statistical software package with the Classification Tree and Logistic Regression classification methods. The aim of the work is to present the CaRT method and to test its ability to make predictions . The development of the example, includes a comparison of the results from the CaRT method, the results from the Logistic Regression method as well as the existing analysis carried out by the method of Logistic Regression in 2015 with the work of A Fotiou, E. Kanavou, M. Stavrou, C. Richardson and A. Kokkevi "Prevalence and correlates of electronic cigarette use among adolescents in Greece: A preliminary cross-sectional analysis of nationwide survey data". The results revealed the variables that determine the use of traditional and electronic cigarettes, which are gender, use of other substances such as cannabis and alcohol, family supervision, and cigarette use by the adolescent's peers . In addition, all three methods yielded similar results, which enhances the functionality of the Classification and Regression Tree method.

Πίνακας περιεχομένων

Εισαγωγή

Η παρούσα εργασία έχει ως στόχο την ανάλυση των παραγόντων που επηρεάζουν, την χρήση του ηλεκτρονικού τσιγάρου, στους εφήβους ηλικίας 15 ετών, στην Ελλάδα, σύμφωνα με στοιχεία που συλλέχθηκαν από την Health Behaviour in School-aged Children Survey (HBSC) το 2014.

Αρχικά, θα πραγματοποιηθεί θεωρητική παρουσίαση της διαδικασίας Knowledge Discovery in Databases (KDD) και της Μηχανική Εκμάθησης «Machine Learning» ενώ εν συνεχεία θα αναλυθεί η μέθοδος της δημιουργίας δέντρων αποφάσεων . Η ανωτέρω μέθοδος χωρίζεται, ουσιαστικά, σε δύο κατηγορίες, αναφορικά με την φύση της υπό εξέταση μεταβλητής (εξαρτημένη). Η πρώτη κατηγορία είναι το δέντρο ταξινόμησης το οποίο χρησιμοποιείται στη περίπτωση που η εξαρτημένη μεταβλητή είναι κατηγορική, κατά την οποία το σφάλμα του μοντέλου υπολογίζεται βάσει του ποσοστού των λανθασμένων κατηγοριοποιήσεων. Η δεύτερη κατηγορία είναι το δέντρο παλινδρόμησης, στην οποία η εξαρτημένη μεταβλητή είναι ποσοτική (συνεχής ή διατεταγμένη διακριτή) όπου το σφάλμα μέτρησης υπολογίζεται βάσει του τετραγώνου της διαφοράς ανάμεσα στην προβλεπόμενη και στην πραγματική τιμή.

Ακολούθως, θα πραγματοποιηθεί η ανάπτυξη του αλγορίθμου, με την χρήση της συνάρτησης C.5.0 μέσω του λογισμικού πακέτου R- Statistic, καθώς και η εξαγωγή των σχετικών συμπερασμάτων.

Έπειτα, θα πραγματοποιηθεί ανάλυση των ίδιων δεδομένων με μία διαφορετική μέθοδο ταξινόμησης, την Logistic Regression (Λογιστική Παλινδρόμηση).

Ένας από τους βασικούς σκοπούς της παρούσας είναι η σύγκριση των τελικών αποτελεσμάτων που προέκυψαν από την μέθοδο Classification Tree, την μέθοδο Logistic Regression καθώς και των αποτελεσμάτων ανάλυσης που πραγματοποιήθηκε με τη μέθοδο της λογιστικής παλινδρόμησης από τους A. Fotiou, E. Kanavou, M. Stavrou, C. Richardson και A. Kokkevi (2015) “ Prevalence and correlates of electronic cigarette use among adolescents in Greece: A preliminary cross-sectional analysis of nationwide survey data”.

Κεφάλαιο 1. Knowledge Discovery in Databases (KDD) και Machine Learning

1.1 Knowledge Discovery in Databases (KDD)

1.1.1 Ορισμός KDD

Με τον όρο KDD νοείται η διαδικασία για την ανάπτυξη μεθόδων για την εξαγωγή πληροφορίας μέσω των δεδομένων. Σύμφωνα με τους Fayyad, Piatetsky-Shapiro, and Smyth (1997), η εν λόγω μέθοδος αφορά στη μετατροπή των δεδομένων με τρόπο που να καθίστανται περισσότερο συμπαγή, περιληπτικά και χρήσιμα.

Στη σύγχρονη εποχή της πληροφορίας, έχει καταστεί εφικτή η συλλογή τεράστιου όγκου δεδομένων και κρίνεται αναγκαία η επεξεργασία και ανάλυση τους, προκειμένου να εξαχθεί η ζητούμενη, κάθε φορά, πληροφορία. Η πληροφορία είναι απαραίτητο στοιχεία για τη δημιουργία θεωριών και μοντέλων. Το Data Mining είναι ένα μέρος της διαδικασίας του KDD.

1.1.2 Ανάπτυξη της διαδικασίας KDD

Για την ολοκλήρωση της διαδικασίας KDD, ακολουθείται μία σειρά βημάτων ως εξής:

1. Προετοιμασία δεδομένων
2. Επιλογή δεδομένων
3. Καθαρισμός δεδομένων
4. Ενσωμάτωση της υφιστάμενης γνώσης
5. Ορθή ερμηνεία των αποτελεσμάτων

Ειδικότερα, κατά τη διαδικασία του KDD ο ερευνητής διαθέτει μία θεωρία και έχει στη διάθεση του ένα σύνολο δεδομένων και εργαλείων προκειμένου να ελεγχθεί εάν στην πράξη επαληθεύεται η θεωρία.

Καταρχάς, υλοποιείται η παρατήρηση και η κατανόηση των δεδομένων και του γενικού προβλήματος που χρήζει επίλυσης.

Εν συνεχεία επιλέγονται τα δεδομένα τα οποία εμπεριέχουν χρήσιμη πληροφορία, συγκεκριμένα επιλέγονται οι συνιστώσες οι οποίες κρίνεται ότι παρουσιάζουν εξάρτηση με την υπό πρόβλεψη- μελέτη μεταβλητή.

Ακολούθως, πραγματοποιείται ο καθαρισμός των δεδομένων. Ειδικότερα, στο σύνολο τους τα δεδομένα ενδέχεται να παρουσιάζουν απουσία ορισμένων των τιμών των μεταβλητών μερικών παρατηρήσεων η να εμπεριέχουν θόρυβο με παρουσία μεγάλου

αριθμού ακραίων παρατηρήσεων. Σε αυτό το στάδιο απαιτείται αρχικά ο εντοπισμός των προβλημάτων και στη συνέχεια η διαχείριση τους.

Έπειτα, διενεργείται μείωση του όγκου των δεδομένων, δηλαδή γίνεται προσπάθεια προκειμένου να διατηρηθούν στην έρευνα μόνο τα δεδομένα που εμπεριέχουν χρήσιμη πληροφορία και να αφαιρεθούν ή να ομαδοποιηθούν δεδομένα που είτε δεν προσφέρουν πληροφορία είτε η πληροφορία που περιέχουν εμπεριέχεται ήδη σε καλύτερο επίπεδο από άλλη κατηγορία μεταβλητών.

Στο επόμενο στάδιο, επιλέγεται η μέθοδος που θα ακολουθηθεί ανάλογα με την φύση των δεδομένων.

1.1.3 Μέθοδοι KDD

Οι σημαντικότερες μέθοδοι, περιληπτικά, όπως αναφέρονται στο άρθρο των Fayyad, Piatetsky-Shapiro and Smyth, (1996) με τίτλο «Knowledge Discovery and Data Mining» είναι οι κάτωθι:

1. Ταξινόμηση : Σχηματισμός μίας συνάρτησης η οποία ταξινομεί κάθε παρατήρηση των δεδομένων σε μία συγκεκριμένη κλάση.
2. Παλινδρόμηση : Βρίσκει μία εκτίμηση για την πραγματική τιμή μίας μεταβλητής, μέσω συνάρτησης, καθώς και τις σχέσεις μεταξύ των μεταβλητών.
3. Συσταδοποίηση : Καθορίζει συστάδες στις οποίες κατατάσσει τα δεδομένα και βάσει των κοινών στοιχείων των δεδομένων επιχειρείται η περιγραφή των βασικότερων στοιχείων τους.
4. Συνόψιση : βρίσκει ένα κοινό στοιχείο για ένα υποσύνολο των δεδομένων.
5. Μοντέλο Εξάρτησης : Καθορίζεται ένα μοντέλο που ορίζει τις εξαρτώμενες σχέσεις μεταξύ των μεταβλητών.

Στο επόμενο στάδιο, επιλέγεται το μοντέλο που θα χρησιμοποιηθεί προκειμένου να εφαρμοστεί η ανωτέρω μέθοδος καθώς και οι παράμετροι που θα καθοριστούν.

1.1.4 Data Mining

Ακολούθως, πραγματοποιείται το στάδιο του Data Mining. Ο όρος εξόρυξη δεδομένων (data mining) αναφέρεται στη διαδικασία της εξέτασης ενός συγκεκριμένου

συνόλου δεδομένων προκειμένου να εντοπισθούν χρήσιμα πρότυπα και να εξαχθούν συμπεράσματα. Ειδικότερα, μέσω του data mining, είναι εφικτό, αφενός να εντοπιστούν σχέσεις μεταξύ των χαρακτηριστικών των παρατηρήσεων, αφετέρου να δημιουργηθούν αλγόριθμοι με προβλεπτική ικανότητα για παρατηρήσεις που δεν περιέχονται στα δεδομένα και παρουσιάζουν κοινά χαρακτηριστικά με αυτά. Κατά την πρόβλεψη χρησιμοποιούνται τα δεδομένα για την παραγωγή μοντέλων που χρησιμοποιούνται στον υπολογισμό μελλοντικών τιμών ενώ κατά την περιγραφή αναζητούνται σχέσεις μεταξύ των μεταβλητών που απαρτίζουν τα δεδομένα. (σελ. 42 “From Data Mining to Knowledge Discovery in Databases” των Fayyad, Piatetsky-Shapiro, and Smyth (1997)).

Σε αυτό το στάδιο καθορίζεται ο τρόπος παρουσίασης των αποτελεσμάτων όπως πίνακες, διαγράμματα, γραφήματα, δέντρα αποφάσεων . Επιπλέον τονίζεται το στοιχείο των αποτελεσμάτων που κρίνεται σημαντικότερο.

Το τελευταίο στάδιο αφορά στην εξαγωγή πληροφορίας και δημιουργία συμπερασμάτων .

1.2 Machine Learning (Μηχανική Εκμάθηση)

1.2.1 Ορισμός Machine Learning

Το πεδίο της επιστήμης που ασχολείται με την ανάπτυξη αλγορίθμων μέσω της χρήσης υπολογιστικών συστημάτων, για την εξαγωγή γνώσης από ένα σύνολο δεδομένων, ονομάζεται Machine Learning. Οι εφαρμοζόμενες μέθοδοι, οι δυνατότητες των υπολογιστικών συστημάτων καθώς και τα διαθέσιμα δεδομένα συνεχώς εξελίσσονται με ραγδαίο ρυθμό. Η διαφορά μεταξύ του Machine Learning και του Data Mining, βασίζεται στο γεγονός ότι το πρώτο χρησιμοποιείται προκειμένου να εκτελεστεί μία γνωστή διαδικασία ενώ το δεύτερο προκειμένου να ανευρεθούν στοιχεία που παρέχουν πληροφορία. Η βασική εργασία του Machine Learning είναι να εξαχθεί πληροφορία μέσω περίπλοκων δεδομένων. Ουσιαστικά πρόκειται για έναν μηχανισμό ο οποίος λαμβάνει δεδομένα και εξάγει πρότυπα μεταξύ τους.

1.2.2 Διαδικασία εκτέλεση της μηχανικής εκμάθησης (Machine Learning)

Ειδικότερα, η διαδικασία του Machine Learning, μπορεί να χωριστεί σε επιμέρους στάδια ως εξής:

1. Συλλογή δεδομένων. Σε αυτό το στάδιο, συγκεντρώνονται τα δεδομένα και καταχωρούνται στην κατάλληλη κάθε φορά μορφή.
2. Εξερεύνηση και ετοιμασία των δεδομένων. Είναι πολύ σημαντικό να γίνουν κατανοητά τα δεδομένα και τα χαρακτηριστικά αυτών από τον ερευνητή.
3. Εκπαίδευση του μοντέλου στα δεδομένα. Με το πέρας του δεύτερου βήματος ο ερευνητής έχει δημιουργήσει μία εικόνα για το αναμενόμενο αποτέλεσμα και σε αυτό το στάδιο επιλέγεται ο κατάλληλος αλγόριθμός για να ελεγχθεί η προσδοκία.
4. Εκτίμηση της απόδοσης του μοντέλου.
5. Βελτίωση της απόδοσης του μοντέλου.

Έστω για παράδειγμα ότι απαιτείται να αντιμετωπιστεί η αυξημένη εγκληματική δραστηριότητα, σε σχέση με τις κλοπές αυτοκινήτων, σε μία συγκεκριμένη γεωγραφική περιοχή, προκειμένου το αστυνομικό έργο να επικεντρωθεί σε περισσότερο στοχευόμενες δράσεις.

Αρχικά, στο στάδιο της συλλογής δεδομένων, δύναται να αναζητηθούν όλα τα περιστατικά κλοπής οχημάτων που έχουν λάβει χώρα και για κάθε ένα να πραγματοποιηθεί συλλογή των ειδικότερων στοιχείων. Συγκεκριμένα, θα συλλεχθούν στοιχεία αναφορικά με την μάρκα κάθε οχήματος, την παλαιότητα του, την οδό του περιστατικού, την ώρα που έλαβε χώρα, την απόσταση από τον κοντινότερο αυτοκινητόδρομο, την ύπαρξη η μη συναγερμού. Κάθε χαρακτηριστικό θα αποτελεί μία ανεξάρτητη μεταβλητή.

Ακολούθως, θα γίνει διόρθωση των δεδομένων και θα αντιμετωπιστεί η απουσία των μεταβλητών ορισμένων παρατηρήσεων. Εν συνεχεία θα μελετηθούν τα δεδομένα από τον ερευνητή, προκειμένου να υπάρχει η σχετική εξοικείωση που θα οδηγήσει σε μία προσδοκία σχετικά με το αποτέλεσμα της ανάλυσης.

Στη συγκεκριμένη περίπτωση, μία πιθανή λύση είναι να καταγραφούν, με τυχαίο τρόπο, ένας αριθμός οχημάτων, τα οποία δεν έχουν κλαπεί και να καταγραφούν τα στοιχεία για αυτά τα οχήματα. Ακολούθως, να ενοποιηθούν όλα τα δεδομένα σε μία βάση και θα δημιουργηθεί μία δίτιμη μεταβλητή, η οποία θα λαμβάνει τιμή 1 εάν το όχημα έχει κλαπεί και 0 σε διαφορετική περίπτωση. Με τον τρόπο αυτό θα γίνει ανάλυση των παραγόντων που επηρεάζουν την κλοπή ή μη του οχήματος. Στη συνέχεια θα βρεθεί η κατάλληλη μέθοδος που θα χρησιμοποιηθεί. Στη συγκεκριμένη περίπτωση, επειδή η εξαρτημένη

μεταβλητή είναι κατηγορική μπορώ να χρησιμοποιήσω Classification Tree,ή και Logistic Regression.

Ακολούθως, δημιουργείται με την χρήση ενός λογισμικού (SPSS, R, E-VIEWS, κτλ) το κατάλληλο μοντέλο στα δεδομένα. Έπειτα υπολογίζεται η απόδοση του, δηλαδή ο βαθμός προσαρμογής στα δεδομένα του δείγματος. Παράλληλα, τυχόν μεταβλητές οι οποίες δεν είναι στατιστικά σημαντικές αφαιρούνται από το μοντέλο και πραγματοποιούνται μέθοδοι για βελτίωση της απόδοσης.

Κεφάλαιο 2. Δέντρα Αποφάσεων

2.1 Ορισμός δέντρου απόφασης

Ένα δέντρο απόφασης είναι ένα διάγραμμα ροής του οποίου η δομή ομοιάζει, οπτικά, με δέντρο, κάθε κόμβος (διακλάδωση) του δέντρου αποτελεί έναν έλεγχο σε μία

μεταβλητή και κάθε κλαδί αναπαριστά το αποτέλεσμα του ελέγχου. Η κλάση αναπαριστάται από το τελευταίο φύλλο δέντρου. Μία μέθοδος μηχανικής εκμάθησης αποτελεί και αυτή του δέντρου απόφασης.

Η μέθοδος αυτή ανήκει στην κατηγόρια «Διαίρε και Βασίλευε» καθώς χωρίζει τα δεδομένα σε όλο και μικρότερες κατηγορίες μέχρι να καταλήξει στην τελική. Επιπλέον, πρόκειται για μέθοδο επιβλεπόμενης μάθησης (supervised learning). Η επιβλεπόμενη μάθηση είναι μία μέθοδος μηχανικής μάθησης η οποία έχει ως στόχο να χαρακτηριστεί ένα σύνολο δεδομένων βάση ενός διαφορετικού συνόλου που χρησιμοποιείται για την εκπαίδευση του μοντέλου. Το μοντέλο που προκύπτει δύναται να χρησιμοποιηθεί για να προβλέψει νέα δεδομένα.

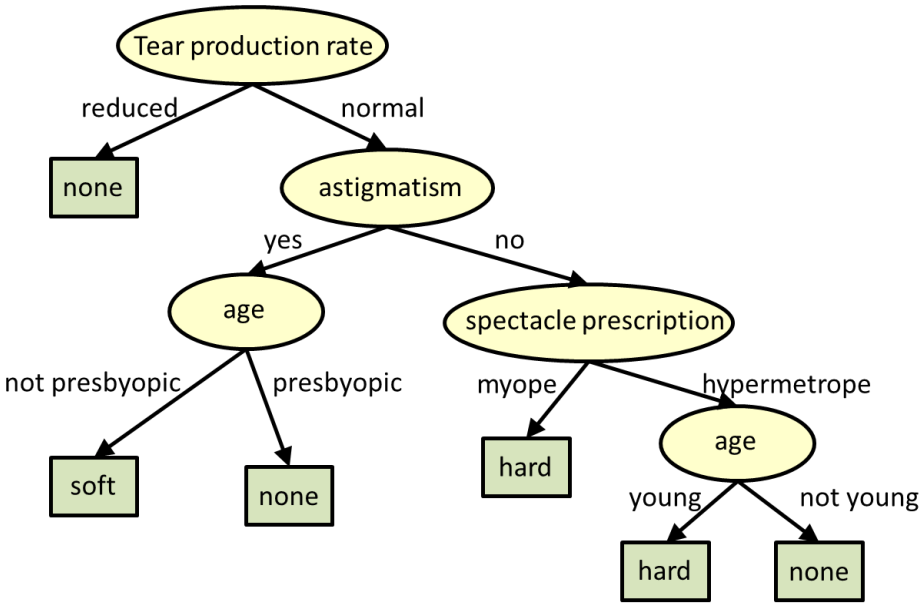
Προκειμένου να καταστεί πιο κατανοητό, ας υποθέσουμε ότι έχουμε ένα σύνολο παρατηρήσεων και ότι η κάθε παρατήρηση αποτελείται από έναν αριθμό ανεξάρτητων μεταβλητών και μία εξαρτημένη μεταβλητή. Κάθε παρατήρηση εισέρχεται στη βάση του δέντρου, ακολούθως σε κάθε διακλάδωση ελέγχεται η τιμή μίας ανεξάρτητης μεταβλητής και ανάλογα με το αποτέλεσμα αυτής ακολουθεί διαφορετική πορεία, η ίδια διαδρομή ακολουθείται προκειμένου να ελεγχθούν όλες οι ανεξάρτητες ή όσες κρίνονται σημαντικές.

Στο τέλος η κάθε παρατήρηση τελειώνει σε μία συγκεκριμένη κλάση ή φύλλο. Η διαδικασία ακολουθείται για όλες τις παρατηρήσεις με αποτέλεσμα να δημιουργείται ένα σύνολο κλάσεων όπου στην κάθε μία εμπεριέχεται ένας αριθμός παρατηρήσεων.

Με την ολοκλήρωση ενός δέντρου δημιουργείται και ένα σύνολο κανόνων οι οποίοι αφορούν την εκτίμηση της εξαρτημένης μεταβλητής για συγκεκριμένες τιμές των ανεξάρτητων. Για παράδειγμα εάν το δέντρο απόφασης έχει ως εξαρτημένη μεταβλητή την χρήση τσιγάρου και ανεξάρτητες μεταβλητές την επικοινωνία με την οικογένεια και την ψυχολογική κατάσταση ο κανόνας ενδέχεται να είναι της μορφής «Ένας που έχει καλές οικογενειακές σχέσεις και καλή ψυχολογική κατάσταση δεν καπνίζει».

Ακολούθως, προκειμένου να καταστεί περισσότερο κατανοητός ο τρόπος λειτουργίας ενός δέντρου αποφάσεων, παρουσιάζεται η γραφική μορφή του στο Διάγραμμα 2.1. Στην συγκεκριμένη περίπτωση, σκοπός είναι να καθοριστεί το είδος των φακών επαφής που θα χρησιμοποιηθεί ανά άτομο. Οι εξαρτημένη μεταβλητή «Είδος φακού» χωρίζεται σε τρεις κατηγορίες, (soft – none – hard). Οι ανεξάρτητες είναι, Tear

Production Rate (reduced – normal), astigmatism (yes-no), age (presbyopic or not, young or not), spectacle prescription (myopia, hypermetropia)



Διάγραμμα 2.1

Παράδειγμα δέντρου απόφασης

(Αναδημοσίευση από την ιστοσελίδα του Carnegie Mellon University, School of Computer Scienc, https://www.cs.cmu.edu/~bhiksha/courses/10-601/decisiontrees/)

Όπως εμφαίνεται στο διάγραμμα, όταν δεν υπάρχει κανονική παραγωγή δακρύων, δεν χρησιμοποιούνται φακοί επαφής, σε διαφορετική περίπτωση όταν υπάρχει αστιγματισμός και το άτομο είναι στην ηλικία που συνηθίζεται να παρουσιάζεται η πρεσβυωπία, ομοίως δεν χρησιμοποιεί φακούς επαφής. Αντιθέτως, όταν υπάρχει ομαλή παραγωγή δακρύων, υπάρχει αστιγματισμός αλλά το άτομο είναι σε ηλικία μικρότερη αυτής που αρχίζει η πρεσβυωπία, τότε χρησιμοποιούνται μαλακοί φακοί επαφής,

διαφορετικά εάν δεν έχει αστιγματισμό αλλά έχει μυωπία σκληροί και αν έχει υπερμετρωπία και είναι μεγάλος δεν χρησιμοποιούνται και αν είναι νέος ομοίως σκληροί.

2.2 Διαδικασία ανάπτυξης Δέντρου Απόφασης

Κατά την εξέλιξη ενός δέντρου απόφασης ο αλγόριθμος επιλέγει ως πρώτο κόμβο, την ανεξάρτητη μεταβλητή που εκτιμάται ότι εμπεριέχει τη σημαντικότερη πληροφορία, δηλαδή που έχει τη μεγαλύτερη προβλεπτική ικανότητα της τιμής της ανεξάρτητης μεταβλητής σε σχέση με τις υπόλοιπες ανεξάρτητες. Ακολούθως και εφόσον ολοκληρωθεί ο διαχωρισμός βάσει της πρώτης ανεξάρτητης, συνεχίζει η διαδικασία με την αμέσως σημαντικότερη ανεξάρτητη και συνεχίζεται ή όλη διαδικασία καθ΄ αυτόν τον τρόπο.

Ο διαχωρισμός τερματίζεται μόλις ολοκληρωθεί ένα κριτήριο που έχει τεθεί ή χρησιμοποιηθεί το σύνολο των ανεξάρτητων μεταβλητών.

2.2.1 Κατηγορίες δέντρων αποφάσεων

Στη μηχανική εκμάθηση τα δέντρα αποφάσεων διαχωρίζονται σε δύο μεγάλες κατηγορίες τα δέντρα ταξινόμησης (Classification Tree) και τα δέντρα παλινδρόμησης (Regression Tree). Κάθε ένα στοχεύει στην πρόβλεψη μίας εξαρτημένης μεταβλητής με δεδομένες τις τιμές ενός συνόλου ανεξαρτήτων μεταβλητών. Για παράδειγμα, έστω Y η εξαρτημένη και X η ανεξάρτητη, το Y αποτελεί διάνυσμα με μέγεθος ίσο με των αριθμό των παρατηρήσεων ενώ το X αποτελεί πίνακα με αριθμό στηλών ίσο με τον αριθμό των ανεξάρτητων μεταβλητών και αριθμό γραμμών ίσο με τον αριθμό των παρατηρήσεων.

Όταν η μεταβλητή είναι συνεχής ή διακριτή που λαμβάνει πραγματικές τιμές τότε αντιμετωπίζεται με τη μέθοδο της παλινδρόμησης (regression), σε διαφορετική περίπτωση όπου η εξαρτημένη μεταβλητή είναι κατηγορική αντιμετωπίζεται με την μέθοδος της ταξινόμησης (classification). Το τελευταίο αποτέλεσμα (φύλλο) είναι η τιμή της μεταβλητής και για να είναι αποτελεσματικό ένα δέντρο απόφασης θα πρέπει να έχει αφενός μικρό ποσοστό σφάλματος αφετέρου να έχει το μικρότερο δυνατό μέγεθος.

2.2.1.1 Regression Tree

Κατά τη μέθοδο αυτή δημιουργείται ένα δέντρο αποφάσεων και ταξινομούνται οι παρατηρήσεις στα φύλλα, όπως φαίνεται στον ανωτέρω διάγραμμα 2.1, όπου τα φύλλα συμβολίζονται με πράσινο χρώμα, ακολούθως υπολογίζεται για κάθε φύλλο η εκτιμώμενη

τιμή της εξαρτημένης μεταβλητής ως το μέσο όρο των τιμών της εξαρτημένης των παρατηρήσεων που αποτελούν το φύλλο. Δηλαδή εάν σε ένα φύλλο έχω c=10 παρατηρήσεις και η κάθε μία έχει τιμή εξαρτημένης μεταβλητής yi τότε :

$$\bar{y} = 1/c * \sum_{i=1}^{10} y_i$$

Κατά την διαδικασία της πρόβλεψης, θεωρείται ότι οι τιμές που ταξινομούνται σε ένα φύλλο έχουν τιμή της εξαρτημένης μεταβλητής ίση με $\bar{y}$.

Τα οφέλη της μεθόδου αυτής είναι:

Πραγματοποίηση γρήγορων προβλέψεων.

Εύκολη παρατήρηση των δεδομένων και εξαγωγή συμπερασμάτων για την σπουδαιότητα των ανεξάρτητων μεταβλητών.

Στην περίπτωση όπου παρατηρείται απουσία της τιμής μίας ανεξάρτητης μεταβλητής μπορεί να γίνει η υπόθεση ότι έχει την τιμή του Μ.Ο των αντίστοιχων τιμών της ίδιας μεταβλητής των λοιπών παρατηρήσεων στο ίδιο φύλλο.

Επί της ουσίας, βασίζεται στη λογική της μεθόδου «Εντόπισε τι κάνουν οι όμοιοι και πράξε το ίδιο». Η ευελιξία κάθε δέντρου υπολογίζεται βάσει του αριθμού των φύλλων του δηλαδή των τελικών κλάσεων του. Με την αύξηση των φύλλων επιτυγχάνεται αφενός μείωση του σφάλματος αφετέρου αύξηση της πολυπλοκότητα του. Για τον λόγο αυτό απαιτείται να ανευρεθεί η χρυσή τομή στον αριθμό των φύλλων του δέντρου.

Κατά τη δημιουργία ενός Regression Tree, ο στόχος είναι να αυξηθεί η πληροφορία που περιέχουν οι ανεξάρτητες μεταβλητές, στον υπολογισμό της εξαρτημένης. Δηλαδή κάθε y_i να είναι όσο το δυνατόν πλησιέστερα στο $\bar{y}$ της κλάσης του και όσο τον δυνατόν διαφορετικό από τα $\bar{y}$ των άλλων κλάσεων.

Το πρώτο στάδιο κατά τη δημιουργία του Regression Tree, είναι να βρεθεί η κατάλληλη ερώτηση, του τύπου "είναι η μεταβλητή x_i>5" ή "είναι η μεταβλητή x_i=κόκκινο", βάση της οποίας θα διχοτομηθούν τα δεδομένα, στον πρώτο κόμβο, με τρόπο που να μεγιστοποιείται η πληροφορία σχετικά με τον μέσο όρο της εξαρτημένης μεταβλητής. Βάσει της σημαντικότερης ερώτησης, δημιουργείται ο πρώτος κόμβος, ο οποίος οδηγεί σε δύο διαφορετικούς κόμβους όπου επαναλαμβάνεται η ίδια διαδικασία εξ

αρχής. Η ίδια διαδικασία επαναλαμβάνεται μέχρι την εφαρμογή ενός "Stop Criterion". Οι κόμβοι που δεν διαιρούνται περαιτέρω αποτελούν τα φύλλα.

Το "stop criterion", όπως προκύπτει και από το όνομα του, είναι ένα κριτήριο που τίθεται από τον ερευνητή και όταν επέλθει περατώνεται η διαδικασία. Στη μέθοδο Classification Tree, υπάρχουν αρκετά κριτήρια που χρησιμοποιούνται καθ' αυτόν τον τρόπο, ενδεικτικά αναφέρονται τα ακόλουθα:

Όταν όλες η παρατηρήσεις της εξαρτημένης μεταβλητής, λαμβάνουν την ίδια τιμή.

Να σχηματιστεί ο μέγιστος αριθμός φύλλων που έχουν οριστεί, δηλαδή το μέγιστο μέγεθος του δέντρου.

Να σχηματιστεί ο μέγιστος αριθμός των κόμβων που έχουν οριστεί.

Να χρησιμοποιηθεί ο μέγιστος αριθμός των ανεξάρτητων μεταβλητών που έχουν οριστεί.

y_i

Προκειμένου να μετρηθεί το μέγεθος της πληροφορίας, που αναφέρθηκε προηγουμένως, χρησιμοποιείται το "Mean Squared Error" (MSE).

$$MSE = 1/n \sum_{i=1}^{t} \sum_{j-1}^{ni} (y_j - m_i)^2$$

$$m_i = 1/n \sum_{i}^{n} y_i$$

Το yj είναι η τιμή των n_i παρατηρήσεων στο φύλλο i με μέση τιμή mi.

Το mc είναι ο μέσος όρος της τιμής της εξαρτημένης μεταβλητής σε κάθε φύλλο.

Το c είναι ο αριθμός των παρατηρήσεων κάθε φύλου.

Το n είναι ο συνολικός αριθμός των παρατηρήσεων των δεδομένων.

Το t είναι ο συνολικός αριθμός των φύλλων.

Αρχικά, διαχωρίζεται το σύνολο των δεδομένων σε δύο υποσύνολα, το Training_set υποσύνολο από το 80% των συνολικών παρατηρήσεων και το Test_set από το υπόλοιπο 20% των δεδομένων.

Ακολούθως, ελέγχεται κάθε πιθανός αρχικός διαχωρισμός των δεδομένων βάσει των ανεξάρτητων μεταβλητών και επιλέγεται ο διαχωρισμός που θα επιφέρει τη μεγαλύτερη μείωση στο MSE. Βάσει των δεδομένων και του stop criterion ενδέχεται να δημιουργηθεί ένα αρκετά μεγάλο δέντρο το οποίο είναι αρκετά περίπλοκο. Ακολούθως, εκτελείται ένα σύνολο ενεργειών για να μειωθεί το μέγεθος του δέντρου, η εν λόγω διαδικασία ονομάζεται κλάδεμα του δέντρου. Αφαιρείται, αρχικά, ένα φύλλο και το

αντίστοιχό του, που ανήκουν στον ίδιο κόμβο και υπολογίζεται το σφάλμα του δέντρου με την χρήση των δεδομένων που δεν έχουν χρησιμοποιηθεί κατά την δημιουργία του (test set), πριν και μετά την περικοπή. Συνεχίζεται η διαδικασία μέχρι την ρίζα. Έπειτα, ελέγχεται πιο υπό – δέντρο παρουσιάζει τα καλύτερα αποτελέσματα, δηλαδή βρίσκονται σχέσεις σφάλματος – μεγέθους. Με την μέθοδο αυτή ουσιαστικά ελέγχεται εάν ένας κόμβος παρέχει χρήσιμη πληροφορία και σε αρνητική περίπτωση αφαιρείται.

2.2.1.2 Classification Tree

Ουσιαστικά, λειτουργεί με τον ίδιο τρόπο που λειτουργεί και το Regression Tree με τη διαφορά ότι η εξαρτημένη μεταβλητή είναι κατηγορική και όχι αριθμητική. Οι ανεξάρτητες μεταβλητές μπορεί να είναι είτε κατηγορικές είτε αριθμητικές όπως και στην περίπτωση του Classification Tree.

Οι κύριες διαφορές με το Regression Tree εντοπίζονται στα κάτωθι:

- Στον τρόπο υπολογισμού του μεγέθους της προσφερόμενης πληροφορίας των ανεξάρτητων μεταβλητών.
- Στο είδος της πρόβλεψης.
- Στον υπολογισμός του σφάλματος της πρόβλεψης.

Για την κατασκευή ενός δέντρου ταξινόμησης, χρησιμοποιούνται διάφοροι μέθοδοι που ελέγχουν την καθαρότητα (purity), μία εξ αυτών είναι η εντροπία.

Η εντροπία χρησιμοποιείται στη στατιστική ως ένα μέτρο της ανωμαλίας ή της μη – καθαρότητας και ο μαθηματικός τύπος είναι :

$$\mathbf{E(S) = -\sum_{i}^{f} (p_i * \log_2 p_i).}$$

p_i είναι η πιθανότητα μία παρατήρηση στα δεδομένα να ανήκει στην κλάση i και f ο αριθμός των κλάσεων.

Για να καταστεί περισσότερο κατανοητό, έστω ότι έχω μόνο δύο κλάσεις A και B και η πιθανότητα μία παρατήρηση να ανήκει στο A είναι 30% και στο B 70%, τότε : $E(S) = - 0.3 * \log_2 0.3 - 0.7 * \log_2 0.7 = 0.8812$.

Η εντροπία λαμβάνει τιμές από το 0 έως το 1 και όσο μεγαλύτερη είναι τόσο μεγαλύτερη και η ανωμαλία στα δεδομένα. Εφόσον υπολογισθεί η εντροπία, χρειάζεται ένα μέτρο προκειμένου να υπολογιστεί πως μειώνεται αυτή με την εισαγωγή περαιτέρω πληροφορίας στα δεδομένα, αυτό πραγματοποιείται με το Information Gain, (IG).

$$\mathbf{IG(Y,X) = E(Y) - E(Y|X)}$$

Υπολογίζεται η μείωση στην αρχική εντροπία E(Y), με την χρησιμότητα της πληροφορίας του X στα δεδομένα E(Y|X). Όσο μεγαλύτερη η μείωση της εντροπίας τόσο καλύτερη η πληροφορία του X. Προκειμένου να γίνει κατανοητή η διαδικασία που ακολουθείται, παρατίθεται το κάτωθι παράδειγμα.

Έστω τα δεδομένα:

	Y		
X	Normal	High	TOTAL
Excellent	3	1	4
Good	4	2	6
Poor	0	4	4
TOTAL	7	7	14

E(Y|Xexcel)= -3/4 $\log_2$(3/4) – 1/4 log2(1/4) = 0.811

E(Y|Xgood) = -4/6$\log_2$(4/6) – 2/6 log2(2/6) = 0.918

E(Y|XPoor) = -0$\log_2$(0) – 4/4log2(4/4) = 0

(ιστοσελίδα του Towards data science: https://towardsdatascience.com/entropy-how-decision-trees-make-decisions-2946b9c18c8)

Εφόσον υπολογιστεί η εντροπία κάθε στοιχείου ακολούθως σταθμίζεται: 0,811 * (4/14) + 0,918* (6/14) +0 = 0,63. Η αρχική εντροπία χωρίς την ανεξάρτητη μεταβλητή είναι E(Y) = -7/14*log2(7/14) -7/14*log2(7/14) =1, άρα το Information Gain είναι : 1 – 0,63 = 0,37. Άρα η ανεξάρτητη μεταβλητή παρείχε σημαντική πληροφορία για τον διαχωρισμό της εξαρτημένης.

Κατά το σχηματισμό ενός δέντρου αποφάσεων, σε περιπτώσεις με πλήθος ανεξάρτητων μεταβλητών, ακολουθείται η ίδια διαδικασία προκειμένου να βρεθεί ποια ανεξάρτητη θα αποτελέσει τον πρώτο κόμβο. Ακολούθως, σε κάθε στάδιο υπολογίζονται εκ νέου το IG των ανεξάρτητων μεταβλητών, καθώς αλλάζει διότι πλέον υπολογίζεται βάσει των προηγούμενων ανεξάρτητων. Η ίδια διαδικασία ακολουθείται μέχρι να ολοκληρωθεί η διαδικασία.

Για τη δημιουργία προβλέψεων, τα δέντρα ταξινόμησης μπορούν να χρησιμοποιηθούν με δύο διαφορετικούς τρόπους, ο πρώτος είναι για να προβλέψουν την

τιμή μίας κατηγορικής μεταβλητής και ο δεύτερος για να υπολογίσουν την κατανομή της πιθανότητας μίας εξαρτημένης μεταβλητής

2.2.2 Ο Αλγόριθμος c 5.0

Για τη δημιουργία δέντρων αποφάσεων έχουν δημιουργηθεί αρκετοί αλγόριθμοι, στην παρούσα εργασία θα γίνει ανάλυση με την χρήση του αλγορίθμου c5.0 ο οποίος βασίζεται στον αλγόριθμο c.4.5.

Αρχικός δημιουργός του C4.5 είναι ο Quinlan, (1993) και του C5.0 οι Quinlan και Kohavi (1999).

Όπως προαναφέρθηκε προηγουμένως, ένας αλγόριθμος ταξινόμησης, η πρώτη δυσκολία που καλείται να αντιμετωπίσει είναι βάσει ποιας ανεξάρτητης μεταβλητής θα γίνει ο αρχικός διαχωρισμός των δεδομένων δηλαδή ποια ερώτηση θα περιέχεται στον πρώτο κόμβο.

Για τη μέτρηση της καθαρότητας και την υλοποίηση του διαχωρισμού των δεδομένων, ο c5.0 χρησιμοποιεί αρχικά την εντροπία και εν συνεχεία το information gain, με τον τρόπο που αναλύθηκαν στην προηγούμενη ενότητα.

Ο αλγόριθμος που δημιουργείται συνεχίζει μέχρι το σημείο όπου ή θα χρησιμοποιηθούν για διαχωρισμός όλες οι ανεξάρτητες μεταβλητές ή θα εφαρμοστεί ένα stop criterion.

2.2.3 Βελτίωση της Απόδοσης των Δέντρων Αποφάσεων

Υπάρχουν αρκετές μέθοδοι, οι οποίοι χρησιμοποιούνται προκειμένου αφενός να βελτιωθεί η απόδοσης ενός δέντρου απόφασης, ακολούθως παρατίθενται οι σημαντικότεροι.

Προκειμένου να καθίσταται δυνατή η μέτρηση της προβλεπτικής ικανότητας του μοντέλου, αρχικά διαχωρίζονται τα μοντέλα σε δύο υποσύνολα, το ένα υποσύνολο αποτελεί το Training Set και είναι αυτό το οποίο χρησιμοποιείται για την δημιουργία του μοντέλου. Το δεύτερο είναι το Test Set και είναι αυτό το οποίο χρησιμοποιείται για τον έλεγχο του μοντέλου. Δηλαδή εφόσον έχει δημιουργηθεί το μοντέλο, εφαρμόζεται για κάθε μία εκ των παρατηρήσεων του Test Set και προκύπτει μία τιμή, για κάθε παρατήρηση, για την εξαρτημένη μεταβλητή. Εν συνεχεία συγκρίνονται οι τιμές της εξαρτημένης

μεταβλητής που έχουν προκύψει από την πρόβλεψη με τις πραγματικές τιμές της όπως είναι στο Test Set και υπολογίζεται το ποσοστό των ορθών και εσφαλμένων προβλέψεων.

2.2.3.1 Bootstrap

Με τον όρο αυτό, νοείται μία στατιστική μέθοδος η οποία χρησιμοποιείται για τον περιορισμό της αβεβαιότητας. Κατά τη διαδικασία αυτή δημιουργώ υποσύνολα, με τυχαίες παρατηρήσεις, από το Training Set. Επειδή τα υποσύνολα δημιουργούνται με τυχαίες παρατηρήσεις, κάθε παρατήρηση παρουσιάζεται σε περισσότερα από ένα υποσύνολα. Εν συνεχεία, χρησιμοποιώ κάθε ένα υποσύνολο για τον υπολογισμό στατιστικών μεγεθών και υπολογίζω τον μέσο όρο των αποτελεσμάτων από κάθε υποσύνολο. Ο εν λόγω μέσος όρος λαμβάνεται ως τελική εκτίμηση.

2.2.3.2 Bagging

Κατά την δημιουργία δέντρων αποφάσεων, εάν χρησιμοποιηθούν διαφορετικά υποσύνολα δεδομένων δημιουργούνται διαφορετικά δέντρα.

Με τη διαδικασία του Bagging, χρησιμοποιούνται τα υποσύνολα των δεδομένων που προέκυψαν μέσω του Bootstrap και με την χρήση του κατάλληλου αλγόριθμου, δημιουργούνται, ισάριθμα με τα υποσύνολα, δέντρα αποφάσεων.

Στην περίπτωση του Regression Tree, εφαρμόζεται το σύνολο των παρατηρήσεων σε κάθε ένα από τα δέντρα και για κάθε παρατήρηση προκύπτει ένας αριθμός προβλέψεων, ίσος με τον αριθμό των δέντρων, για την ανεξάρτητη μεταβλητή. Ακολούθως, λογίζεται ως τελική πρόβλεψη της εξαρτημένης μεταβλητής, για κάθε παρατήρηση, ο μέσος όρος των προβλέψεων που προέκυψαν.

$$\bar{y}_i = 1/b * \sum_{k=1}^{b} y_{ki}$$

$\bar{y}_i$ = Η πρόβλεψη της τιμή της εξαρτημένης μεταβλητής της παρατήρησης i.

b = Ο αριθμός των υποσυνόλων που δημιουργήθηκαν με το bootstrap, ο οποίος είναι ίδιος με τον αριθμό των δέντρων που δημιουργήθηκαν.

yk_i = η πρόβλεψη της τιμής της εξαρτημένης μεταβλητής της παρατήρησης i που προκύπτει από το δέντρο k.

Στην περίπτωση του Classification Tree, δημιουργείται ομοίως ένας αριθμός δέντρων ίσος με τον αριθμό των υποσυνόλων που έχουν δημιουργηθεί με το Bootstrap. Ακολούθως, δοκιμάζεται το σύνολο των δεδομένων σε όλα τα δέντρα και προκύπτει ένα σύνολο προβλέψεων για την τιμή της εξαρτημένης μεταβλητής κάθε παρατήρησης. Η τιμή, της εξαρτημένης μεταβλητής κάθε παρατήρησης, η οποία προκύπτει ως αποτέλεσμα από τα περισσότερα δέντρα, λογίζεται ως η πρόβλεψη για την τιμή αυτή. Επί της ουσίας δηλαδή τα δέντρα ψηφίζουν για την πρόβλεψη.

2.2.3.3 Random Forest

Η μέθοδος αυτή, πρώτη φορά χρησιμοποιήθηκε από τους Breiman και Adele Cutler (n.d). Κατά τη διαδικασία αυτή, δημιουργείται αρχικά μέσω της διαδικασίας του Bootstrap, ένας αριθμός υποσυνόλων παρατηρήσεων και ακολούθως, με την χρησιμοποίηση κάθε υποσυνόλου δημιουργείται ένα δέντρο αποφάσεων. Η ουσιαστική διαφορά με το Bagging είναι ότι σε αυτή τη διαδικασία για κάθε δέντρο δεν χρησιμοποιείται το σύνολο των ανεξάρτητων μεταβλητών αλλά σε κάθε δέντρο επιλέγεται, τυχαία, ένας συγκεκριμένος αριθμός ανεξάρτητων μεταβλητών. Συνηθίζεται, ο αριθμός των ανεξάρτητων μεταβλητών που χρησιμοποιείται να είναι η ρίζα του συνόλου αυτών, δηλαδή $m=\sqrt{p}$.

m = ο αριθμός των ανεξάρτητων που χρησιμοποιούνται για την κατασκευή του δέντρου.
p = ο συνολικός αριθμός των ανεξάρτητων μεταβλητών.

Το πλεονέκτημα της συγκεκριμένης μεθόδου βασίζεται στο γεγονός, ότι σε περιπτώσεις ύπαρξης μίας μεταβλητής με ισχυρή εξάρτηση υπερκαλύπτεται η πληροφορία των υπολοίπων καθώς επιλέγεται συνεχώς αυτή ως πρώτη μεταβλητή διαχωρισμού.

2.2.3.4 Boosting

Η ιδέα του Boosting, στηρίζεται στην ιδέα ότι μπορούν να συνδυαστούν μεταβλητές που παρέχουν μικρό μέγεθος πληροφορίας προκειμένου να προσφέρουν από κοινού μεγαλύτερη πληροφορία. Στη διαδικασία αυτή, ακολουθείται μία διαδικασία, παρόμοια με το Bagging, με τη διαφορά ότι, σε αυτή τη μέθοδο τα δέντρα δημιουργούνται διαδοχικά και κάθε ένα κατά το σχηματισμό του λαμβάνει υπόψη τα αποτελέσματα των προηγούμενων. Ειδικότερα, κάθε πρόβλεψη που έχει γίνει από τα δημιουργούμενα δέντρα

σταθμίζεται βάσει της σημαντικότητας του, δηλαδή επί της ουσίας, ομοίως τα δέντρα ψηφίζουν αλλά κάθε ψήφος παρουσιάζει διαφορετική βαρύτητα ανάλογα με την προβλεπτική ικανότητα του κάθε μοντέλου. Ο αλγόριθμος c.5.0 επιτρέπει την διαδικασία αυτή με αρκετά εύκολο τρόπο.

Κεφάλαιο 3. Ανάπτυξη Παραδείγματος

3.1 Παρουσίαση Προβλήματος

Ο σκοπός της ανάλυσης είναι η εξέταση των παραγόντων, οι οποίοι αποτελούν 17 διαφορετικές μεταβλητές και 4 διαφορετικές κατηγορίες, όπως αναλύεται στον Πίνακα 3.2Α, σχετικά με την χρήση του παραδοσιακού και του ηλεκτρονικού τσιγάρου. Ειδικότερα, ο στόχος είναι να εντοπιστούν τα στοιχεία της συμπεριφοράς, των νέων ηλικίας 15 ετών, τα οποία παρουσιάζουν συσχέτιση με την χρήση του τσιγάρου.

Αρχικά θα αναλυθούν οι παράγοντες που επιδρούν στην χρήση του παραδοσιακού τσιγάρου και ακολούθως αυτοί που επιδρούν στην χρήση του ηλεκτρονικού.

Η ανάλυση θα γίνει με δύο διαφορετικές μεθόδους, Classification Tree και Logistic Regression.

Εν συνεχεία, θα εξαχθούν συμπεράσματα, σχετικά με τον τρόπο συσχέτισης των μεταβλητών, οι οποίες αναπαριστούν τη συμπεριφορά των νέων και έπειτα θα συγκριθούν τα συμπεράσματα που προέκυψαν από κάθε μέθοδο καθώς και τα αποτέλεσμα της υφιστάμενης έρευνας που έχει πραγματοποιηθεί με τη μέθοδο του Logistic Regression.

3.2 Παρουσίαση δεδομένων

Η συλλογή των δεδομένων που θα χρησιμοποιηθούν, πραγματοποιήθηκε στην Ελλάδα, το 2014, από το (HBSC) «Health Behaviour in School-aged Children» και αφορούν απαντήσεις νέων εφήβων ηλικίας 15 ετών.

Τα δεδομένα αποτελούν ένα δείγμα 1127 παρατηρήσεων και 19 διαφορετικών μεταβλητών. Οι ανεξάρτητες μεταβλητές είναι 17 και χωρίζονται σε τέσσερις μεγάλες κατηγορίες, είναι κατηγορικές μεταβλητές, τύπου "Binary", ενώ αναλύονται στον Πίνακα 3.2Α ως εξής:

Πίνακας 3.2Α

«Ανεξάρτητες Μεταβλητές»

Α. Κατηγορία: Κοινωνικοδημογραφικές (Social)		
1.Φύλο (Gender_Social),	1= Αγόρι	2 = Κορίτσι
2.Καταγωγή γονέων (Parents_Social)	1 = Τουλάχιστον ένας όχι Έλληνας	2= Δύο γονείς Έλληνες
3.Οικονομική κατάσταση γονέων (Affluence_Social)	1 = χαμηλή	2= μέσου όρου και άνω
Β.Κατηγορία : Οικογενειακές Σχέσεις (Family)		

1.Ευκολία συζήτησης με γονείς (Talk_Family)	1= όχι εύκολη	2= εύκολη τουλάχιστον με έναν
2.Επικοινωνία με την οικογένεια (Communic_Family)	1= όχι καλή	2= καλή
3. Υποστήριξη Οικογένειας (Support_Family)	1= δεν υπάρχει	2 = υπάρχει
4. Γνώση του γονέα σχετικά με τις δραστηριότητες του τέκνου (Know_Family)	1= χαμηλή επίβλεψη	2= υψηλή επίβλεψη
5. Οικογενειακές Σχέσεις (Relations_Family)	1= όχι καλές	2 = καλές
Γ. Κατηγορία : Ψυχοσωματική Υγεία (Health)		
1. Σωματική Δραστηριότητα (Activity_Health)	1= χαμηλή δραστηριότητα (έως 2 φορές την εβδομάδα)	2= υψηλή (3 και άνω)
2. Αυτοεκτίμηση σχετικά με την κατάσταση της υγείας (SelRe_Health	1= κακή υγεία	2 = καλή υγεία
3. Ικανοποίηση από την ζωή (Satisfaction_Health)	1= Όχι και τόσο καλή	2 =Πολύ καλή
Δ. Κατηγορία : Χρήση Ουσιών (Substance)		
1.Τωρινός καπνιστής παραδοσιακού τσιγάρου (Smoke_Substance)	1= καπνιστής	0 = μη καπνιστής
2. Heavy Smoker (Heavy_Substance)	1= ναι	0 = όχι
3. Κατανάλωση Αλκοόλ (Alcohol_Substance)	1= χρήση αλκοόλ 0 έως 5 μέρες	2= 6 μέρες και άνω
4. Χρήση κάνναβης σε όλη τη ζωή του (Cannabis_ Substance)	1= τουλάχιστον μία φορά	2= όχι

5. Χρήση άλλων ουσιών σε όλη τη ζωή του (Other_Substance)	1= τουλάχιστον μία φορά	2= όχι
6. Χρήση τσιγάρου από οικείους (Peer_Substance)	1= σχεδόν όλοι	2= κανένας

Πίνακας 3.2Β'

«Εξαρτημένες μεταβλητές»

Οι εξαρτημένες μεταβλητές είναι δύο και αναλύονται ως εξής:

1. Χρήση Παραδοσιακού τσιγάρου (smokeLT)	1 = τουλάχιστον μία φορά	0= όχι
2. Χρήση ηλεκτρονικού τσιγάρου (esmokeLT)	1 = τουλάχιστον μία φορά	0= όχι

3.3 Περιγραφική Ανάλυση

Στον πίνακα 3.3 που ακολουθεί, εμφαίνονται τα βασικά στοιχεία σχετικά με την χρήση του παραδοσιακού και του ηλεκτρονικού τσιγάρου από τα αγόρια και τα κορίτσια του δείγματός μας ενώ παρουσιάζονται και τα ακόλουθα:

Το σύνολο των δεδομένων αποτελείται από 1127 παρατηρήσεις το οποίο αντιστοιχεί στις απαντήσεις 1127 νέων ηλικίας 15 ετών, εκ των οποίων οι 602 (53,4%) είναι κορίτσια και οι 525 (46,6%) είναι αγόρια.

Από το σύνολο των νέων του δείγματος, έχουν κάνει χρήση παραδοσιακού τσιγάρου, έστω και μία φορά στη ζωή τους, οι 406 (36%), ενώ έχουν κάνει χρήση ηλεκτρονικού τσιγάρου, έστω και μία φορά στη ζωή τους, οι 177 (15,7%).

Από το σύνολο των εφήβων που έχουν δοκιμάσει ηλεκτρονικό τσιγάρο οι 151 (85,3%) έχουν δοκιμάσει και παραδοσιακό τσιγάρο. Αδιαμφισβήτητα υπάρχει ιδιαίτερα υψηλή συσχέτιση μεταξύ της χρήσης παραδοσιακού και ηλεκτρονικού τσιγάρου.

Από το σύνολο των αγοριών, έχουν κάνει χρήση παραδοσιακού τσιγάρου, 181 έφηβοι (34,5%), ενώ ηλεκτρονικού τσιγάρου 114 έφηβοι (21,7%).

Από το σύνολο των κοριτσιών, έχουν κάνει χρήση παραδοσιακού τσιγάρου 225 έφηβες (37,4%), ενώ ηλεκτρονικού 63 έφηβες (10%).

Παρατηρείται, ότι τα αγόρια ηλικίας 15 ετών έχουν ελάχιστα μικρότερη τάση προς τη χρήση τσιγάρου σε σχέση με τα κορίτσια της ίδιας ηλικίας, ενώ αντιθέτως έχουν αρκετά μεγαλύτερη (περίπου 1 προς 2) τάση για τη χρήση ηλεκτρονικού τσιγάρου.

Από το σύνολο των 406, εφήβων που έχουν δοκιμάσει παραδοσιακό τσιγάρο, οι 188 είναι τωρινοί καπνιστές (46,3%) και οι 77 (19%) δεν είναι περιστασιακοί αλλά τακτικοί καπνιστές. Το οποίο υποδηλώνει ότι, ένας έφηβος, που έχει δοκιμάσει παραδοσιακό τσιγάρο έχει πιθανότητες περίπου 50% να συνεχίσει το κάπνισμα έστω και ως περιστασιακός καπνιστής και 20% να συνεχίσει ως τακτικός καπνιστής, μέχρι την ηλικία των 15 ετών.

Από τους νέους που είναι τωρινοί καπνιστές παραδοσιακού τσιγάρο, οι 100 (53,2%) έχουν κάνει χρήση ηλεκτρονικού τσιγάρου.

Από τον συνολικό αριθμό των νέων που έχουν δοκιμάσει ηλεκτρονικό τσιγάρο (177) το 85,3% έχει δοκιμάσει και παραδοσιακό, ενώ από τους νέους πους δεν έχουν δοκιμάσει ηλεκτρονικό τσιγάρο μόλις το 26.8% έχει δοκιμάσει παραδοσιακό.

Πίνακας 3.3

«Περιγραφικά Στοιχεία»

	Σύνολο n= 1127	Αγόρια n= 525	Κορίτσια n= 602
Έχουν δοκιμάσει παραδοσιακό τσιγάρο έστω και μία φορά	36,0%	34,5%	37,4%
Αυτοί που έχουν δοκιμάσει έστω και μία φορά παραδοσιακό και ηλεκτρονικό τσιγάρο	13,4%	17,3%	10,0%
Αυτοί που καπνίζουν παραδοσιακό τσιγάρο την παρούσα χρονική στιγμή	16,7%	16,8%	16,6%
Αυτοί που καπνίζουν παραδοσιακό τσιγάρο την παρούσα χρονική στιγμή και έχουν δοκιμάσει ηλεκτρονικό τσιγάρο	8,9%	11,0%	7,0%
Αυτοί που κάνουν μεγάλη κατανάλωση παραδοσιακών τσιγάρων (Heavy Smokers)	6,9%	7,8%	6,1%
Αυτοί που κάνουν μεγάλη κατανάλωση παραδοσιακών τσιγάρων και έχουν δοκιμάσει ηλετρονικό τσιγάρο	4,5%	5,5%	3,7%
Αυτοί που έχουν δοκιμάσει έστω και μία φορά ηλεκτρονικό τσιγάρο	15,7%	21,7%	10,5%
Αυτοί που έχουν δοκιμάσει έστω και μία φορά ηλεκτρονικό τσιγάρο και δεν έχουν δοκιμάσει παραδοσιακό	2,3%	4,4%	0,5%

Κεφάλαιο 4 Εφαρμογή Classification Tree

4.1 Ανάπτυξη του Classification Tree

Η δημιουργία του Classification Tree θα πραγματοποιηθεί με την χρήση του αλγορίθμου c.5.0, όπως αναλύθηκε ανωτέρω και με την χρήση του λογισμικού πακέτου R.

Αρχικά θα πραγματοποιηθεί η δημιουργία ενός Classification Tree, με εξαρτημένη μεταβλητή την smokeLT, προκειμένου να εξεταστεί ποιοι από τους ανωτέρω παράγοντες σχετίζονται με την χρήση παραδοσιακού τσιγάρου στους εφήβους ηλικίας 15 ετών.

Ακολούθως, θα πραγματοποιηθεί η δημιουργία ενός Classification Tree, με εξαρτημένη μεταβλητή την esmokeLT, προκειμένου να εξεταστεί ποιοι από τους ανωτέρω παράγοντες σχετίζονται με την χρήση ηλεκτρονικού τσιγάρου στους εφήβους ηλικίας 15ετών, ανάμεσα στους καπνιστές παραδοσιακού τσιγάρου.

4.1.1 Classification Tree – Παραδοσιακό Τσιγάρο

4.1.1.1 Ανάπτυξη Αλγορίθμου

Για την ανάλυση χρησιμοποιήθηκαν 15 ανεξάρτητες μεταβλητές, καθώς αφαιρέθηκαν οι μεταβλητές «τωρινός καπνιστής παραδοσιακού τσιγάρου (Smoke_Substance)» και τακτικός καπνιστής παραδοσιακού τσιγάρου (Heavy_Substance)», καθώς οι εν λόγω προϋποθέτουν θετική τιμή στην εξαρτημένη μεταβλητή και δεν παρέχουν οποιαδήποτε πληροφορία.

Ο αλγόριθμός που χρησιμοποιήθηκε και το αποτέλεσμα αυτού εμφαίνεται στο Παράρτημα 1 και αναλύεται ως ακολούθως.

Το σύνολο των δεδομένων διαχωρίστηκε σε δύο μερικά σύνολα, το πρώτο είναι το Training_set, υποσύνολο, το οποίο αποτελείται από το 80% των συνολικών παρατηρήσεων και το οποίο θα χρησιμοποιηθεί προκειμένου να δημιουργηθεί το μοντέλο πρόβλεψης. Το δεύτερο είναι το Test_set και το οποίο αποτελείται από το υπόλοιπο 20% των δεδομένων και το οποίο θα χρησιμοποιηθεί προκειμένου να υπολογιστεί η προβλεπτική ικανότητα του μοντέλου που θα δημιουργηθεί.

Τα δύο υποσύνολα δημιουργούνται με τυχαία επιλογή παρατηρήσεων από το αρχικό σύνολο προκειμένου να είναι αντικειμενικά κατανεμημένες οι παρατηρήσεις.

Ακολούθως δημιουργήθηκε το μοντέλο με τη χρήση του αλγορίθμου c5.0 και το υποσύνολο των δεδομένων Training_Set.

Δημιουργήθηκε ένα δέντρο, μεγέθους τριών φύλλων

Classification Tree

Number of samples: 902

Number of predictors: 15

Tree size: 3

Cannabis_ Substance <= 1: 1 (78/8)

Cannabis_ Substance > 1:

:...Peer_Substance <= 1: 1 (108/36)

Peer_Substance > 1: 0 (716/183)

Δημιουργήθηκαν τρεις κανόνες, καθώς τρεις είναι οι «διαδρομές» που οδηγούν στα τρία φύλλα, για τους έφηβους ηλικίας 15 ετών, οι οποίοι είναι οι εξής;

Εάν έχει δοκιμάσει κάνναβη τότε έχει δοκιμάσει και παραδοσιακό τσιγάρο με πιθανότητα (90,6%, 78/86).

Εάν δεν έχει δοκιμάσει κάνναβη και τα άτομα συναναστροφής τους είναι καπνιστές τότε έχει δοκιμάσει παραδοσιακό τσιγάρο με πιθανότητα (75%, 108/144).

Εάν δεν έχει δοκιμάσει κάνναβη και τα άτομα συναναστροφής του δεν είναι καπνιστές τότε δεν έχει δοκιμάσει παραδοσιακό τσιγάρο με πιθανότητα (79,6%, 716/899)

Προκύπτει ότι, οι σημαντικότεροι παράγοντες που καθορίζουν εάν κάποιος έχει δοκιμάσει παραδοσιακό τσιγάρο είναι η χρήση κάνναβης καθώς και η χρήση τσιγάρου από τα άτομα συναναστροφής του.

Ακολούθως, πραγματοποιήθηκε έλεγχος την προβλεπτικής ικανότητας και προέκυψαν τα κάτωθι :

		Predicted		
		0	1	TOTAL
Real	0	533	44	577
	1	183	142	325
	TOTAL	716	186	902

Το μοντέλο που δημιουργήθηκε παρουσιάζει στο υποσύνολο, των δεδομένων που χρησιμοποιήθηκαν για τον σχηματισμό του, ποσοστό σφάλματος 25%. Ειδικότερα, προβλέπει με επιτυχία 92,4% την μη χρήση παραδοσιακού τσιγάρου, ενώ δεν έχει

προβλεπτική ικανότητα στην περίπτωση που κάποιος είναι χρήστης παραδοσιακού τσιγάρου καθώς το σφάλμα του είναι 56,3%.

Ακολούθως, χρησιμοποιήθηκε το Test Set προκειμένου να υπολογιστεί η προβλεπτική ικανότητα του μοντέλου.

		Predicted		
		0	1	TOTAL
Real	0	140	4	144
	1	39	42	81
	TOTAL	179	46	225

Παρατηρείται ότι το συνολικό ποσοστό σφάλματος είναι, 19,1% (43/225) εκ των οποίων ποσοστό 3% είναι στην περίπτωση που ο νέος δεν έχει δοκιμάσει παραδοσιακό τσιγάρο και 48,1% στην περίπτωση που έχει δοκιμάσει παραδοσιακό τσιγάρο.

Προκύπτει ότι, στο Training Set προβλέπονται ορθά το 75% των παρατηρήσεων ενώ στο Test Set το 80% των παρατηρήσεων.

Στη συνέχεια, προκειμένου να βελτιωθεί η απόδοση του δέντρου, εφαρμόστηκε Boosting με αριθμό επαναλήψεων δέκα, με την χρήση του αλγορίθμου c.5.0. Με την διαδικασία αυτή, δημιουργήθηκαν δέκα διαφορετικά δέντρα, από δέκα διαφορετικά υποσύνολα, που δημιουργήθηκαν από τυχαίες παρατηρήσεις από το Training Set, όπως αναλυτικά εμφαίνονται στο Παράρτημα 1. Οι μεταβλητές ταξινομήθηκαν, βάσει της προβλεπτικής ικανότητας της κάθε μίας.

Από τον έλεγχο της προβλεπτικής ικανότητας προέκυψαν τα κάτωθι:

		Predicted		
		0	1	TOTAL
Real	0	529	48	577
	1	180	145	325
	TOTAL	709	193	902

Η απόδοση της προβλεπτική ικανότητας του μοντέλου είναι συνολικά 75% (25% σφάλμα).

Ενώ από τον έλεγχο της προβλεπτικής ικανότητας με το Test Set:

		Predicted		
		0	1	TOTAL
Real	0	137	7	144
	1	38	43	81
	TOTAL	175	44	225

Η απόδοση της προβλεπτική ικανότητας του μοντέλου είναι συνολικά 80% (20% σφάλμα).

Τα αποτελέσματα που προέκυψαν είχαν ως εξής :

Attribute usage:

100.00% Cannabis_ Substance
91.35% Support_Family
91.35% Know_Family
91.35% Alcohol_Substance
91.35% Other_Substance
91.35% Peer_Substance
82.93% Parents_Social
82.93% Relations_Family
69.62% Affluence_Social

Εκφράζεται το ποσοστό των παρατηρήσεων κάθε μεταβλητής που χρησιμοποιήθηκαν για την δημιουργία του δέντρου.

Από την ανωτέρω ανάλυση προέκυψε ότι οι 6 σημαντικότερες μεταβλητές στο καθορισμό της υπό εξέταση μεταβλητής, ήταν οι :

Cannabis_ Substance
Support_Family
Know_Family
Alcohol_Substance
Other_Substance
Peer_Substance

4.1.2 Classification Tree –Ηλεκτρονικό Τσιγάρο

4.1.2.1 Ανάπτυξη Αλγορίθμου

Στην ανάλυση χρησιμοποιήθηκαν μόνο οι μεταβλητές οι οποίες έχουν τιμή 1 στη μεταβλητή smokeLT, δηλαδή αφορά μόνο τους νέους οι οποίοι έχουν δοκιμάσει παραδοσιακό τσιγάρο.

Τα δεδομένα που χρησιμοποιήθηκαν αποτελούνται από 406 παρατηρήσεις, 17 ανεξάρτητες μεταβλητές και μία εξαρτημένη της esmokeLT η οποία αντιστοιχεί στη χρήση ηλεκτρονικού τσιγάρου έστω και μία φορά.

Ο αλγόριθμός που χρησιμοποιήθηκε και το αποτέλεσμα αυτού εμφαίνεται στο Παράρτημα 2 και αναλύεται ως ακολούθως.

Το σύνολο των παρατηρήσεων είναι 406 έφηβοι, οι οποίοι στο σύνολό τους έχουν δοκιμάσει παραδοσιακό τσιγάρο και εκ των οποίων οι 255 (63%) έχουν δοκιμάσει ηλεκτρονικό τσιγάρο ενώ οι 151 (37%) δεν έχουν δοκιμάσει. Ακολούθως, πραγματοποιήθηκε, διαχωρισμός των δεδομένων σε Training Data, για την κατασκευή του Classification Tree (80%, 325) και σε Test Data, για τον έλεγχο του δέντρου (20%,81).

Ακολούθως δημιουργήθηκε το μοντέλο με τη χρήση του αλγορίθμου c5.0 και το υποσύνολο των δεδομένων Training_Set.

Δημιουργήθηκε ένα δέντρο, μεγέθους 8 φύλλων, το οποίο σημαίνει ότι δημιουργήθηκαν 8 διαφορετικές «διαδρομές» που οδηγούν στα φύλλα, ως εξής:

```
Decision tree:
Smoke_Substance <= 0:
:...Can_Substance > 1: 0 (160/35)
:   Can_Substance <= 1:
:   :...Peer_Substance <= 1: 1 (2)
:       Peer_Substance > 1:
:       :...Communic_Family <= 1: 1 (6/1)
:           Communic_Family > 1: 0 (6)
Smoke_Substance > 0:
```

:...Talk_Family <= 1: 1 (33/8)

Talk_Family > 1:

:...Gender_Social > 1: 0 (59/18)

Gender_Social <= 1:

:...Relations_Family <= 1: 0 (14/5)

Relations_Family > 1: 1 (45/14)

Οι κανόνες είναι οι εξής:

1. Εάν δεν είναι τωρινός καπνιστής παραδοσιακού τσιγάρου και δεν έχουν δοκιμάσει κάνναβη τότε δεν έχει δοκιμάσει ηλεκτρονικό τσιγάρο με πιθανότητα σφάλματος (18%, 35/195)
2. Εάν δεν είναι τωρινός καπνιστής παραδοσιακού τσιγάρου και έχουν δοκιμάσει κάνναβη και οι οικείοι τους καπνίζουν τότε έχει δοκιμάσει ηλεκτρονικό τσιγάρο.

Πιθανότητα σφάλματος (0%)

3. Εάν δεν είναι τωρινός καπνιστές παραδοσιακού τσιγάρου και έχουν δοκιμάσει κάνναβη και οι οικείοι τους δεν καπνίζουν και η επικοινωνία με την οικογένεια δεν είναι καλή τότε έχει δοκιμάσει ηλεκτρονικό τσιγάρο.

Πιθανότητα σφάλματος (14%, 1/7)

4. Εάν δεν είναι τωρινός καπνιστές παραδοσιακού τσιγάρου και έχουν δοκιμάσει κάνναβη και οι οικείοι τους δεν καπνίζουν και η επικοινωνία με την οικογένεια είναι καλή τότε δεν έχει δοκιμάσει ηλεκτρονικό τσιγάρο.

Πιθανότητα σφάλματος (0%)

5. Εάν είναι τωρινός καπνιστής παραδοσιακού τσιγάρου και η ευκολία συζήτησης με τους γονείς δεν είναι εύκολη τότε έχει δοκιμάσει ηλεκτρονικό τσιγάρο.

Πιθανότητα σφάλματος (19%, 8/41)

6. Εάν είναι τωρινός καπνιστής παραδοσιακού τσιγάρου και η ευκολία συζήτησης με τους γονείς είναι εύκολη και είναι κορίτσι τότε δεν έχει δοκιμάσει ηλεκτρονικό τσιγάρο.

Πιθανότητα σφάλματος (23%,18/77)

7. Εάν είναι τωρινός καπνιστής παραδοσιακού τσιγάρου και η ευκολία συζήτησης με τους γονείς είναι εύκολη και είναι αγόρι και δεν υπάρχουν καλές οικογενειακές σχέσεις τότε δεν έχει δοκιμάσει ηλεκτρονικό τσιγάρο.

Πιθανότητα σφάλματος (26%,5/19)

8. Εάν είναι τωρινός καπνιστής παραδοσιακού τσιγάρου και η ευκολία συζήτησης με τους γονείς είναι εύκολη και είναι αγόρι και υπάρχουν καλές οικογενειακές σχέσεις τότε έχει δοκιμάσει ηλεκτρονικό τσιγάρο.

Πιθανότητα σφάλματος (24%, 14/59)

Ακολούθως, πραγματοποιήθηκε έλεγχος της προβλεπτικής ικανότητας του μοντέλου με την χρήση του training set.

		Predicted		
		0	1	TOTAL
Real	0	181	23	204
	1	58	63	121
	TOTAL	239	86	325

Το ποσοστό λάθους ήταν (81/325, 25%) συνολικά και συγκεκριμένα, στη περίπτωση 0, 11% και στην περίπτωση 1 , 48%.

Ακολούθως, πραγματοποιήθηκε ό έλεγχος της προβλεπτικής ικανότητας του μοντέλου με την χρήση του test set, δηλαδή των παρατηρήσεων που δεν χρησιμοποιήθηκαν για την δημιουργία του μοντέλου.

		Predicted		
		0	1	TOTAL
Real	0	44	7	51
	1	18	12	30
	TOTAL	62	19	81

Το ποσοστό λάθους ήταν (25/81, 31%) συνολικά και συγκεκριμένα, στη περίπτ ωση 0, 14% και στην περίπτωση 1 , 60%.

Στη συνέχεια, προκειμένου να βελτιωθεί η απόδοση του δέντρου, εφαρμόστηκε Boosting με αριθμό επαναλήψεων δέκα, με την χρήση του λγορίθμου c.5.0. Με την διαδικασία αυτή, δημιουργήθηκαν δέκα διαφορετικά δέντρα, από δέκα διαφορετικά υποσύνολα, που δημιουργήθηκαν από τυχαίες παρατηρήσεις από το Training Set, όπως αναλυτικά εμφαίνονται στο Παράρτημα 2.

Η απόδοση της προβλεπτική ικανότητας του μοντέλου :

		Predicted		
		0	1	TOTAL
Real	0	45	6	51
	1	15	15	30
	TOTAL	60	21	81

Το συνολικό σφάλμα είναι : (21/81, 26%) και ειδικότερα, στη περίπτωση 0, 12% και στη περίπτωση 1 50%.

Από την ανωτέρω ανάλυση προέκυψε ότι οι 5 σημαντικότερες, με σειρά προτερ αιότητας, στον καθορισμό της υπό εξέταση μεταβλητής, ήταν οι :

100.00% Gender_Social

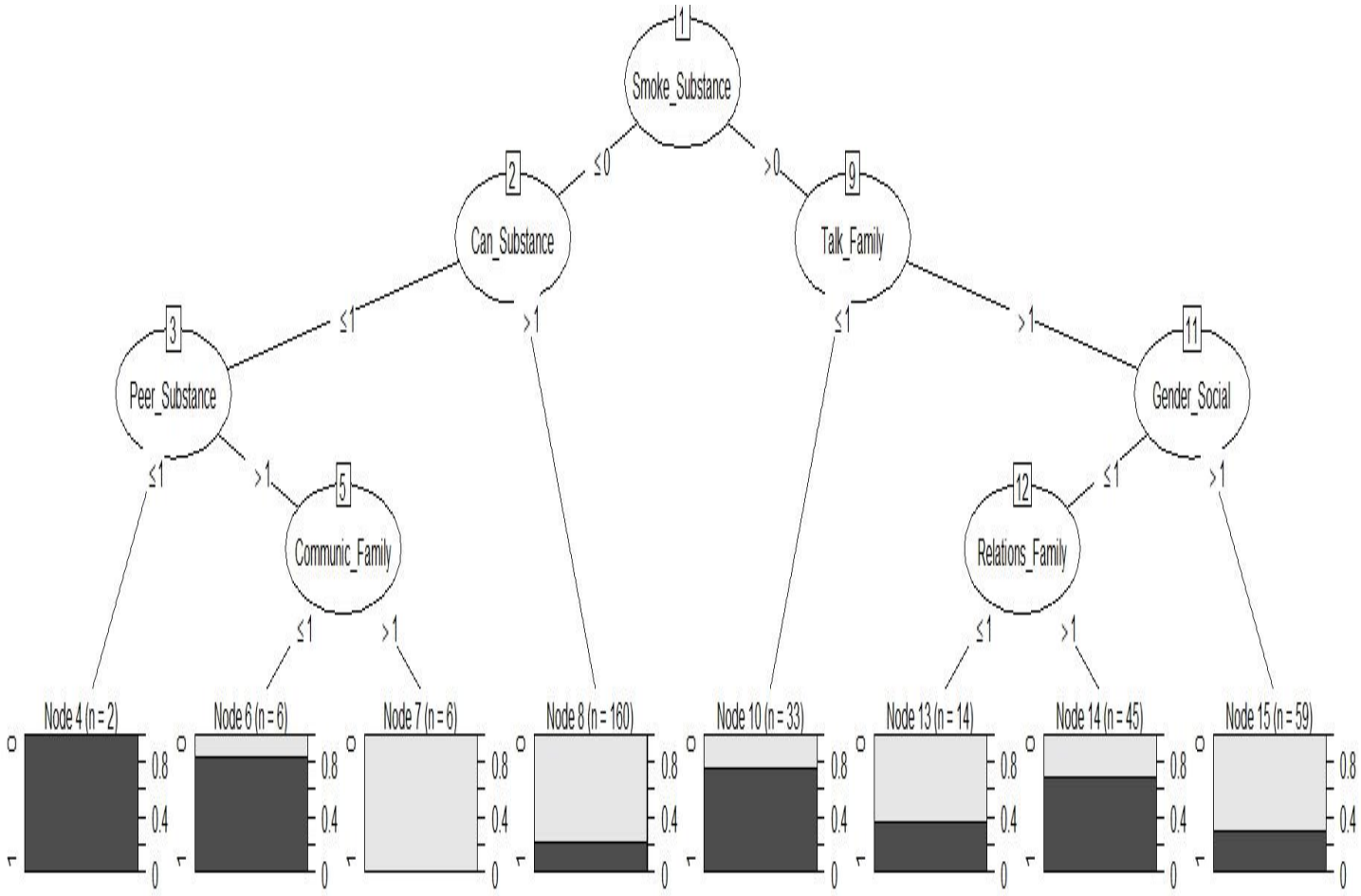
100.00% Smoke_Substance

100.00% Peer_Substance

95.08% Satisfaction_Health

90.77% Can_Substance

Ακολούθως, εμφανίζεται το διάγραμμα που αναπαριστά το δέντρο που δημιουργήθηκε :



Διάγραμμα 4.1.2.1

Δέντρο απόφασης «Classification Tree – EsmokeLT)

Κεφάλαιο 5ο Logistic Regression.

5.1 Μέθοδος Ταξινόμησης – Λογιστική Παλινδρόμηση (Logistic Regression)

Αποτελεί μία μέθοδο ταξινόμησης, με την οποία υπολογίζεται η πιθανότητα της εξαρτημένης μεταβλητής, η οποία είναι δίτιμη κατηγορική μεταβλητή, να λαμβάνει μία συγκεκριμένη τιμή.

Η συνάρτηση που χρησιμοποιείται είναι της μορφής :

$$p(j) = \frac{e^{bx}}{1+e^{bx}} \qquad 5.1$$

b = διάνυσμα με αριθμό γραμμών ίσο με τον αριθμό των ανεξάρτητων μεταβλητών +1 το οποίο αντιστοιχεί στον σταθερό όρο

x = διάνυσμα με p+1 στοιχεία.

Με την εν λόγω συνάρτηση υπολογίζεται η πιθανότητα η εξαρτημένη μεταβλητή να λαμβάνει τιμή ίση με j. Για την εκτίμηση των παραμέτρων, χρησιμοποιείται η μέθοδος της μέγιστης πιθανοφάνειας. Προκειμένου να καθίσταται απλούστερη η παρουσίαση αποτελεσμάτων χρησιμοποιούνται τα odds, για να εκφραστούν οι πιθανότητες

Τα odds, στην ουσία είναι ένας τρόπος παρουσίασης των πιθανοτήτων και παρουσιάζονται με τη μορφή i/k, το οποίο σημαίνει ότι εάν έχω συνολικά n παρατηρήσεις (n=i+k), οι i εξ αυτών έχουν θετικό αποτέλεσμα στην επίτευξη ενός συμβάντος, που έχω ορίσει, και οι υπόλοιπες k αρνητικό. Για παράδειγμα η πιθανότητα να έρθει ένα συγκεκριμένο νούμερο στη ρίψη ενός ζαριού είναι odds 1/5, καθώς κάθε 6 προσπάθειες στη 1 θα έχω επιτυχία και στις υπόλοιπες 5 αποτυχία.

Ο μαθηματικό τύπος odds :

$$\frac{p(x)}{1-p(x)} \qquad 5.2$$

Από την ανωτέρω εξίσωση 5.1 προκύπτει με μετασχηματισμό

$p(x) * (1 + e^{bx}) = e^{bx}$

$p(x) + p(x) * (e^{bx}) = e^{bx}$

$p(x) = e^{bx} (1-p(x))$

$\frac{p(x)}{1-p(x)} = e^{bx}$

$Log(\frac{p(x)}{1-p(x)}) = bx$

Με τον τρόπο αυτό, επιτυγχάνεται η δημιουργία μίας εξίσωσης γραμμικής μορφής. Κάθε bi, δείχνει ότι, με την μεταβολή της xi ανεξάρτητης μεταβλητής κατά μία μονάδα, μεταβάλλεται το log(odds) κατά b_i, όταν οι υπόλοιπες τιμές παραμένουν σταθερές.

Όπως προαναφέρθηκε, στη λογιστική παλινδρόμηση οι εκτιμητές bi υπολογίζονται με την μέθοδο της μέγιστης πιθανοφάνειας, ουσιαστικά, στη περίπτωη που έχω εξαρτημένη μεταβλητή κατηγορική τύπου Binary, απαιτείται να ανευρεθούν τα κατάλληλα bi που μεγιστοποιούν την πιθανοφάνειας.

$$L(bi) = \prod_{i:yi=1} p_{xi} \prod_{i\prime:yi\prime=0} (1 - p_{xi}') \quad 5.3$$

5.2 Ανάπτυξης Αλγορίθμου με τη μέθοδο Logistic Regression.

Θα γίνει ανάλυση των δεδομένων, που χρησιμοποιήθηκαν, για την δημιουργία του Classification Tree, όπως παρουσιάζονται στο Κεφάλαιο 4, με την χρήση της μεθόδου της Λογιστικής Παλινδρόμησης (Logistic Regression).

Αρχικά θα πραγματοποιηθεί ανάλυση για την μεταβλητή smokeLT, προκειμένου να εξεταστεί ποιοι από τους ανωτέρω παράγοντες σχετίζονται με την χρήση παραδοσιακού τσιγάρου στους εφήβους ηλικίας 15 ετών.

Ακολούθως, θα πραγματοποιηθεί ανάλυση, με εξαρτημένη μεταβλητή την esmokeLT, προκειμένου να εξεταστεί ποιοι από τους ανωτέρω παράγοντες σχετίζονται με την χρήση ηλεκτρονικού τσιγάρου στους εφήβους ηλικίας 15ετών, οι οποίοι έχουν δοκιμάσει παραδοσιακό τσιγάρο.

5.2.1 Παραδοσιακό Τσιγάρο

5.2.1.1 Ανάπτυξη Αλγορίθμου.

Για την ανάλυση χρησιμοποιήθηκαν 15 ανεξάρτητες μεταβλητές, καθώς αφαιρέθηκαν οι μεταβλητές «τωρινός καπνιστής παραδοσιακού τσιγάρου (Smoke_Substance)» και τακτικός καπνιστής παραδοσιακού τσιγάρου (Heavy_Substance)», καθώς οι εν λόγω προϋποθέτουν θετική τιμή στην εξαρτημένη μεταβλητή και δεν παρέχουν οποιαδήποτε πληροφορία.

Ο αλγόριθμός που χρησιμοποιήθηκε και το αποτέλεσμα αυτού εμφανίζεται στο Παράρτημα 3.

Ομοίως με την μέθοδο Classification Tree, το σύνολο των δεδομένων διαχωρίστηκε σε δύο σύνολα, το Training_set υποσύνολο από το 80% των συνολικών παρατηρήσεων και το οποίο χρησιμοποιήθηκε προκειμένου να δημιουργηθεί το μοντέλο πρόβλεψης και το Test_set από το υπόλοιπο 20% των δεδομένων για να υπολογιστεί η προβλεπτική ικανότητα του μοντέλου.

Με την εφαρμογή του αλγορίθμου (glm), με εξαρτημένη μεταβλητή την smoke_LT, προέκυψαν τα ακόλουθα:

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	11.543.495	1.370.391	8.424	< 2e-16	***
Gender_Social	0.301670	0.171390	1.760	0.078385	.
Parents_Social	-0.044324	0.212313	-0.209	0.834631	
Affluence_Social	0.268409	0.229374	1.170	0.241928	
Talk_Family	0.059070	0.248333	0.238	0.811986	
Communic_Family	-0.343571	0.212583	-1.616	0.106056	
Support_Family	-0.230221	0.262217	-0.878	0.379954	
Know_Family	-0.506092	0.180274	-2.807	0.004995	**
Relations_Family	-0.293974	0.231203	-1.272	0.203551	
Activity_Health	0.097205	0.170729	0.569	0.569116	
SelRe_Health	-0.149404	0.295104	-0.506	0.612663	
Satisfaction_Health	0.003865	0.175817	0.022	0.982461	
Alcohol_Substance	-0.990798	0.272800	-3.632	0.000281	***
`Cannabis_ Substance`	-2.424.961	0.412086	-5.885	3.99e-09	***
Other_Substance	-0.369851	0.263518	-1.404	0.160464	
Peer_Substance	-1.837.560	0.240365	-7.645	2.09e-14	***

Residual deviance: 930.86

AIC: 962.86

Γίνεται ιδιαίτερη μνεία ότι το AIC, αντιπροσωπεύει το Akaike Criterion (AIC= 2T-1(k-L), όπου το k είναι ο αριθμός των ανεξάρτητων μεταβλητών και L η μέγιστη τιμή της λογαριθμοποιημένης συνάρτησης πιθανοφάνειας όπως εμφαίνεται στη σχέση 5.3. Το Akaike αποτελεί κριτήριο για την σύγκριση μοντέλων μεταξύ τους το μοντέλο με την μικρότερη τιμή θεωρείται το καλύτερο.

Όπως εμφαίνεται στον ανωτέρω πίνακα, οι μεταβλητές Peer_Substance, Cannabis_ Substance και Alcohol_Substance είναι στατιστικά σημαντικές για κάθε επίπεδο εμπιστοσύνης, η μεταβλητή Know_Family είναι στατιστικά σημαντική για επίπεδο σημαντικότητας 0,01 και η μεταβλητή Gender_Social για επίπεδο 0,1.

Ακολούθως, εφαρμόστηκε ο ίδιος αλγόριθμος δέκα επιπλέον φορές με την αφαίρεση μίας μη στατιστικά σημαντικής μεταβλητής κάθε φορά και υπολογίστηκε το AIC για κάθε ένα μοντέλο.

ALL VARIABLES	962.86
- VARIABLE	AIC:
Satisfaction_Health	960.86
Parents_Social	960.91
Talk_Family	960.92
SelRe_Health	961.12
Activity_Health	961.19
Support_Family	961.63
Affluence_Social	962.26
Relations_Family	962.47
Other_Substance	962.86
Communic_Family	963.45

Προέκυψε ότι, με την αφαίρεση των μεταβλητών Satisfaction_Health, Parents_Social, Talk_Family, SelRe_Health, Activity_Health, Support_Family, Affluence_Social και Relations_Family, προκλήθηκε μείωση του συνολικού AIC.

Εν συνεχεία αφαιρέθηκε το σύνολο των εν λόγω μεταβλητών και εφαρμόστηκε το μοντέλο με τις υπόλοιπες εξαρτημένες.

Το νέο μοντέλο έχει τα κάτωθι αποτελέσματα :

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	114.932	11.517	9.979	< 2e-16	***
Gender_Social	0.3181	0.1655	1.921	0.054686	.
Communic_Family	-0.5413	0.1841	-2.940	0.003281	**
Know_Family	-0.5686	0.1734	-3.279	0.001042	**
Alcohol_Substance	-0.9839	0.2702	-3.642	0.000271	***
`Cannabis_ Substance`	-24.500	0.4096	-5.981	2.22e-09	***
Other_Substance	-0.4033	0.2600	-1.551	0.120808	
Peer_Substance	-18.146	0.2375	-7.640	2.17e-14	***

Residual deviance: 936.77

AIC: 952.77

Όπως παρατηρείται, οι μεταβλητές Peer_Substance, Cannabis_ Substance και Alcohol_Substance είναι στατιστικά σημαντικές για κάθε επίπεδο σημαντικότητας, οι μεταβλητές Communic_Family και Know_Family είναι σημαντικές για επίπεδο σημαντικότητας 0,01, ενώ η μεταβλητή Gender_Social είναι για επίπεδο 0,1.

Επιπλέον, το συνολικό AIC έχει μειωθεί από 962.86 σε 952.77, ένδειξη ότι το νέο μοντέλο προσαρμόζεται καλύτερα στα δεδομένα από το προηγούμενο.

Εν συνεχεία εφαρμόστηκε έλεγχος της προβλεπτικής ικανότητας του μοντέλου, με την χρήση του test set και προέκυψες ότι, από τις 144 περιπτώσεις όπου δεν είχε γίνει

χρήση παραδοσιακού τσιγάρου οι 131 (91%) προβλέφθηκαν ορθά και οι 13 (9%) εσφαλμένα, ενώ από τις 81 περιπτώσεις όπου είχε γίνει χρήση τσιγάρου οι μισές μόνο προβλέφθηκαν σωστά ως εκ τούτου, το μοντέλο παρουσιάζει αδυναμία στην προβλεπτική ικανότητα σχετικά με την κατανάλωση παραδοσιακού τσιγάρου. Η συνολική προβλεπτική ικανότητα είναι 172/225 76,4%.

		Predicted		
		0	1	TOTAL
Real	0	131	13	144
	1	40	41	81
	TOTAL	171	54	225

5.2.1.2 Αποτελέσματα Ανάλυσης

Από την ανάλυση που προηγήθηκε εξάγονται, υπό μορφή κανόνων τα ακόλουθα συμπεράσματα.

1. Η χρήση παραδοσιακού τσιγάρου από του οικείους επηρεάζει την χρήση τσιγάρου, το αρνητικό πρόσημο δηλώνει ότι όταν αυξάνεται η τιμή της μεταβλητής από το 1 στο 2, τότε τείνει να μειωθεί η τιμή της εξαρτημένη από το 1 στο μηδέν. Δηλαδή η μη χρήση παραδοσιακού τσιγάρου από τους οικείους (2) σχετίζεται με στη μη χρήση παραδοσιακού τσιγάρου από τους νέους (0).

2. Η χρήση κάνναβης, είναι καθοριστικός παράγοντας καθώς σχετίζεται με την χρήση παραδοσιακού τσιγάρου, ομοίως το αρνητικό πρόσημο δηλώνει ότι η μη χρήση κάνναβης σχετίζεται με την μη χρήση παραδοσιακού τσιγάρου και αντίστροφα.

3. Η χρήση αλκοόλ σχετίζεται με την χρήση παραδοσιακού τσιγάρου, ομοίως το αρνητικό πρόσημο σημαίνει ότι η μη χρήση αλκοόλ σχετίζεται με την μη χρήση παραδοσιακού τσιγάρου.

4. Η υψηλή επίβλεψη του γονέα σχετίζεται με την μη χρήση παραδοσιακού τσιγάρου.

5. Η καλή επικοινωνία του γονέα με τον νέο σχετίζεται με την μη χρήση παραδοσιακού τσιγάρου.

6. Το φύλο επηρεάζει και ειδικότερα, επειδή είναι θετικό το πρόσημο σημαίνει ότι τα κορίτσια (2) έχουν μεγαλύτερη τάση για κατανάλωση παραδοσιακού τσιγάρου.

5.2.2 Ηλεκτρονικό Τσιγάρο

5.2.2.1 Ανάπτυξη Αλγορίθμου

Στην ανάλυση χρησιμοποιήθηκαν μόνο οι παρατηρήσεις οι οποίες έχουν τιμή 1 στη μεταβλητή smokeLT, δηλαδή αφορά μόνο τους νέους οι οποίοι έχουν δοκιμάσει παραδοσιακό τσιγάρο.

Τα δεδομένα που χρησιμοποιήθηκαν αποτελούνται από 406 παρατηρήσεις, 17 ανεξάρτητες μεταβλητές και μία εξαρτημένη της esmokeLT η οποία αντιστοιχεί στη χρήση ηλεκτρονικού τσιγάρου έστω και μία φορά.

Ο αλγόριθμός που χρησιμοποιήθηκε και το αποτέλεσμα αυτού εμφαίνεται στο Παράρτημα 4.

Για το σύνολο των μεταβλητών εξήχθησαν τα ακόλουθα:

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	291.741	176.777	1.650	0.098874	.
Gender_Social	-104.088	0.28128	-3.700	0.000215	***
Parents_Social	-0.35671	0.31857	-1.120	0.262835	
Affluence_Social	0.39253	0.36523	1.075	0.282481	
Talk_Family	-0.82617	0.36764	-2.247	0.024624	*
Communic_Family	0.04231	0.33848	0.125	0.900517	
Support_Family	-0.25877	0.37287	-0.694	0.487685	
Know_Family	0.19086	0.29282	0.652	0.514528	
Relations_Family	0.20971	0.34613	0.606	0.544607	
Activity_Health	0.42263	0.28351	1.491	0.136034	
SelRe_Health	-0.18903	0.39857	-0.474	0.635307	
Satisfaction_Health	-0.19955	0.28579	-0.698	0.485029	
Smoke_Substance	0.92630	0.31913	2.903	0.003701	**
Heavy_Substance	0.32228	0.39827	0.809	0.418392	
Alcohol_Substance	0.19873	0.34292	0.580	0.562226	
Can_Substance	-0.53184	0.35619	-1.493	0.135408	
Other_Substance	0.15779	0.37105	0.425	0.670647	
Peer_Substance	-0.61502	0.30329	-2.028	0.042577	*

Residual deviance: 363.05

AIC: 399.05

Όπως εμφαίνεται, η μεταβλητές που είναι στατιστικά σημαντικές είναι η Gender_Social για οποιοδήποτε επίπεδο εμπιστοσύνης, η Smoke_Substance για επίπεδο εμπιστοσύνης 0,01, οι μεταβλητές Talk_Family και Peer_Substance για επίπεδο 0,05.

Ακολούθως, εφαρμόστηκε ο ίδιος αλγόριθμος δεκατρείς επιπλέον φορές με την αφαίρεση μίας μη στατιστικά σημαντικής μεταβλητής κάθε φορά και υπολογίστηκε το AIC για κάθε περίπτωση, τα αποτελέσματα που προέκυψαν είχαν ως εξής :

All Variables 399.05

	AIC:
Communic_Family	397.06
Other_Substance	397.23
SelRe_Health	397.27
Alcohol_Substance	397.39
Relations_Family	397.42
Know_Family	397.48
Support_Family	397.53
Satisfaction_Health	397.54
Heavy_Substance	397.71
Affluence_Social	398.23
Parents_Social	398.29
Can_Substance	399.28
Activity_Health	399.3

Προέκυψε ότι, με την αφαίρεση των μεταβλητών Communic_Family, Other_Substance, SelRe_Health, Alcohol_Substance, Relations_Family, Know_Family Support_Family, Satisfaction_Health , Heavy_Substance Affluence_Social και Parents_Social, προκλήθηκε μείωση του συνολικού AIC.

Εν συνεχεία αφαιρέθηκε το σύνολο των εν λόγω μεταβλητών και εφαρμόστηκε το μοντέλο με τις υπόλοιπες εξαρτημένες.

Το νέο μοντέλο έχει τα κάτωθι αποτελέσματα :

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	35.713	11.876	3.007	0.002637	**
Gender_Social	-10.057	0.2673	-3.763	0.000168	***
Talk_Family	-0.8594	0.3103	-2.770	0.005606	**
Activity_Health	0.3619	0.2720	1.330	0.183386	
Smoke_Substance	0.9401	0.2899	3.243	0.001184	**
Can_Substance	-0.6038	0.3144	-1.921	0.054770	.
Peer_Substance	-0.6181	0.2876	-2.149	0.031636	*

Residual deviance: 368.30

AIC: 382.3

Όπως εμφαίνεται, οι μεταβλητές που είναι στατιστικά σημαντικές είναι η Gender_Social για οποιοδήποτε επίπεδο εμπιστοσύνης, η Smoke_Substance και η Talk_Family για επίπεδο εμπιστοσύνης 0,01, η μεταβλητή Peer_Substance για επίπεδο 0,05 και η Can_ Substance για 0,1.

Επιπλέον, το συνολικό AIC έχει μειωθεί από 399,05 σε 382,3, ένδειξη ότι το νέο μοντέλο προσαρμόζεται καλύτερα στα δεδομένα από το προηγούμενο.

Εν συνεχεία εφαρμόστηκε έλεγχος της προβλεπτικής ικανότητας του μοντέλου με την χρήση του Test Set και προέκυψες ότι, από τις 51 περιπτώσεις όπου δεν είχε γίνει χρήση παραδοσιακού τσιγάρου οι 44 (86%) προβλέφθηκαν ορθά και οι 7 (13,7%) εσφαλμένα, ενώ από τις 30 περιπτώσεις όπου είχε γίνει χρήση τσιγάρου οι 17 (57%) προβλέφθηκαν σωστά ως εκ τούτου, το μοντέλο παρουσιάζει αδυναμία στην προβλεπτική ικανότητα σχετικά με την κατανάλωση παραδοσιακού τσιγάρου. Η συνολική προβλεπτική ικανότητα ήταν (61/81 , 75%).

		Predicted		
		0	1	TOTAL
Real	0	44	7	51
	1	13	17	30
	TOTAL	57	24	81

5.2.2.2 Αποτελέσματα Ανάλυσης

Τα αποτελέσματα είναι τα ακόλουθα.

1. Η χρήση του ηλεκτρονικού τσιγάρου επηρεάζεται από το φύλλου του νέου και ειδικότερα λόγω του αρνητικού προσήμου, τα αγόρια έχουν μεγαλύτερη τάση προς το τσιγάρο από ότι τα κορίτσια.
2. Η ευκολία συζήτησης με τους γονείς, επηρεάζει την χρήση ηλεκτρονικού τσιγάρου και συγκεκριμένα εάν υπάρχει ευκολία συζήτησης τότε επιδρά στην μη χρήση ηλεκτρονικού τσιγάρου.
3. Η χρήση παραδοσιακού τσιγάρου την παρούσα στιγμή σχετίζεται με την χρήση ηλεκτρονικού τσιγάρου, λόγω του θετικού προσήμο υπάρχει θετική σχέση μεταξύ της χρήση παραδοσιακού και ηλεκτρονικού τσιγάρου.
4. Η χρήση παραδοσιακού τσιγάρου από του οικείους επηρεάζει την χρήση ηλεκτρονικού τσιγάρου, το αρνητικό πρόσημο δηλώνει ότι όταν αυξάνεται η τιμή της μεταβλητής από το 1 στο 2, τότε τείνει να μειωθεί η τιμή της εξαρτημένη από το 1 στο μηδέν. Δηλαδή η μη χρήση παραδοσιακού τσιγάρου από τους οικείους (2) οδηγεί στη μη χρήση ηλεκτρονικού τσιγάρου από τους νέους (0).
5. Η χρήση κάνναβης, είναι καθοριστικός παράγοντας καθώς σχετίζεται με την χρήση ηλεκτρονικού τσιγάρου, ομοίως το αρνητικό πρόσημο δηλώνει ότι η μη χρήση κάνναβης σχετίζεται με την μη χρήση παραδοσιακού τσιγάρου και αντίστροφα.

Κεφάλαιο 6 Σύγκριση αποτελεσμάτων

6.1 Σύγκριση Αποτελεσμάτων σχετικά με τη χρήση του Παραδοσιακού Τσιγάρου

Στον παρακάτω πίνακα παρουσιάζονται οι ανεξάρτητες μεταβλητές, οι οποίες προέκυψαν ως σημαντικές παράμετροι στον καθορισμό της εξαρτημένης μεταβλητής SmokeLT.

	Logistic Regression άρθρο των A. Fotiou, E. Kanavou, M. Stavrou, C. Richardson και A. Kokkevi (2015)	Logistic Regression Παρούσα Εργασία	Classification Tree Παρούσα εργασία
1	Know_Family (Parental Monitoring)	Know_Family	Know_Family
2	Cannabis_ Substance (Any lifetime cannabis use)	Cannabis_ Substance	Cannabis_ Substance
3	Peer_Substance (Peers who smoke tobacco)	Peer_Substance	Peer_Substance
4	Alcohol_Substance (Frequent alcohol use)	Alcohol_Substance	Alcohol_Substance
5	Gender_Social (Boy)	Gender_Social	
6		Communic_Family	
7			Support_Family
			Other_Substance

Όπως προκύπτει, οι 4 μεταβλητές, Know_ Family, Cannabis_ Substance, Peer_Substance και Alcohol_Substance ήταν σημαντικές και στις 3 αναλύσεις.

Μία διαφορά οφείλεται στο γεγονός ότι στο Classification Tree, αποδείχθηκε σημαντική η μεταβλητή Support _Family (Υποστήριξη Οικογένειας) η οποία μεταβλητή, δεν ήταν σημαντική με την μέθοδο Logistic Regression σε κάμία από τις δύο αναλύσεις.

Μία πιθανή εξήγηση είναι ότι η μεταβλητή Support_Family αντιπροσωπεύει την υποστήριξης της οικογένειας και η μεταβλητή Know_Family την επίβλεψη της οικογένειας. Οι δύο μεταβλητές έχουν τιμή 1 σε περίπτωση χαμηλής υποστήριξης και χαμηλής επίβλεψης και τιμή 2 σε περίπτωση υψηλής επίβλεψης και υψηλής υποστήριξης, αντίστοιχα . Στο 70% των περιπτώσεων, οι εν λόγω μεταβλητές είχαν ίδια τιμή, το οποίο σημαίνει ότι υπάρχει υψηλή συσχέτιση μεταξύ της επίβλεψης και της υποστήριξης της οικογένειας, με αποτέλεσμα μέγεθος της πληροφορίας της Support_Family να εμπεριέχεται στη Know_Family.

Μία ακόμα διαφορά ήταν ότι, κατά τη Λογιστική Παλινδρόμηση η μεταβλητή που καθορίζει το φύλο αναδείχθηκε ως σημαντική ενώ κατά το Classification Tree όχι.

Όπως προαναφέρθηκε κατά την περιγραφική ανάλυση, από το σύνολο των αγοριών, έχουν κάνει χρήση παραδοσιακού τσιγάρου, 181 έφηβοι (34,5%) και Από το σύνολο των κοριτσιών, έχουν κάνει χρήση παραδοσιακού τσιγάρου 225 έφηβες (37,4%). Τα ποσοστά είναι σχεδόν όμοια και ως εκ τούτου το αναμενόμενο αποτέλεσμα ήταν να μην αποτελεί σημαντική ερμηνευτική μεταβλητή.

Επιπλέον, η χρήση άλλων ουσιών (Other_Substance) αποδείχθηκε σημαντική μεταβλητή στην ανάλυση με Classification Tree και όχι στην ανάλυση με Logistic Regression. Η εν λόγω μεταβλητή έχει ιδιαίτερα υψηλή εξάρτηση με την μεταβλητή που καθορίζει την χρήση κάνναβης, καθώς στο 87% των παρατηρήσεων, οι νέοι που είχαν κάνει χρήση κάνναβης είχαν κάνει και χρήση άλλων ουσιών και οι χρήστες που δεν είχαν κάνει χρήση κάνναβης δεν είχαν κάνει χρήση άλλων ουσιών. Μόλις στο 13% των παρατηρήσεων παρουσιαζόταν αναντιστοιχία, δηλαδή οι χρήστες που είναι κάνει χρήση κάνναβης δεν είχαν κάνει χρήση άλλων ουσιών και, αυτοί που είχαν κάνει χρήση άλλων ουσιών δεν είχαν κάνει χρήση κάνναβης.

Μεταξύ των 2 αναλύσεων με την μέθοδο του Logistic Regression, προέκυψε ότι η μεταβλητή Communic_Family, αποδείχθηκε σημαντική στη παρούσα εργασία και όχι στην έρευνα με το προαναφερθέν άρθρο.

6.2 Σύγκριση Αποτελεσμάτων σχετικά με τη χρήση του Ηλεκτρονικού Τσιγάρου

Στον παρακάτω πίνακα παρουσιάζονται οι ανεξάρτητες μεταβλητές, οι οποίες προέκυψαν ως σημαντικές παράμετροι στον καθορισμό της εξαρτημένης μεταβλητής EsmokeLT.

	Logistic Regression άρθρο των A. Fotiou, E. Kanavou, M. Stavrou, C. Richardson και A. Kokkevi (2015)	Logistic Regression Παρούσα Εργασία	Classification Tree Παρούσα εργασία
1	Can_Substance (Any lifetime cannabis use)	Can_Substance	Can_Substance
2	Smoke_Substance	Smoke_Substance	Smoke_Substance
3	Peer_Substance	Peer_Substance	Peer_Substance
4	Gender_Social (Boy)	Gender_Social	Gender_Social
5	Satisfaction_Health (Average or low life satisfaction)		Satisfaction_Health
6		Talk_Family	

Όπως προκύπτει 4 μεταβλητές, οι Can_Substance, Smoke_Substance, Peer_Substance και Gender_Social αποδείχθηκαν ως σημαντικές και από τις τρεις αναλύσεις.

Η μεταβλητή Talk_ Family, αποδείχθηκε ως σημαντική με την μέθοδο του Logistic Regression στην παρούσα εργασία αλλά όχι με την μέθοδο Classification Tree και με την μέθοδο Logistic Regression στο ανωτέρω άρθρο.

6.3 Εξαγωγή Συμπερασμάτων

Από την ανάλυση που προηγήθηκε, προκύπτει, αδιαμφισβήτητα, ότι η χρήση τσιγάρου, είναι μία διαδεδομένη συνήθεια μεταξύ των νέων στην Ελλάδα, καθώς περίπου ένας στους τρεις έχει έρθει σε επαφή με το παραδοσιακό τσιγάρο και ένας στους 6 με το ηλεκτρονικό, μέχρι την ηλικία των 15 ετών.

Επιπλέον, σημαντικός προσδιοριστικός παράγοντας, για τη χρήση του παραδοσιακού τσιγάρου, αποδείχθηκε ότι αποτελεί το οικογενειακό περιβάλλον.

Ειδικότερα, η επίβλεψη που ασκείται στο παιδί, αποδείχθηκε σημαντικός παράγοντας και από τις τρεις αναλύσεις, στην επιλογή του παιδιού να κάνει χρήση τσιγάρου.

Επιπροσθέτως, η χρήση άλλων ουσιών εκτός του τσιγάρου όπως η κάνναβη και το αλκοόλ, αποδείχθηκε ότι επιδρούν καθοριστικά, στην χρήση του τσιγάρου από τους νέους.

Περαιτέρω, επιβεβαιώθηκε ένα είδος μιμητικής συμπεριφοράς, καθώς η χρήση του τσιγάρου από τα άτομα της συναναστροφής του νέου, αποδείχθηκε ότι επηρεάζει και τη συμπεριφορά του ίδιου στην χρήση αυτού.

Αναφορικά με την χρήση του ηλεκτρονικού τσιγάρου, προέκυψε ιδιαίτερα υψηλή συσχέτιση με τη χρήση του παραδοσιακού.

Επιπλέον, και από τις τρεις μεθόδους προέκυψε ότι καθοριστική μεταβλητή είναι το φύλο, καθώς από τα κορίτσι μόλις το 10,4% έχει δοκιμάσει ηλεκτρονικό τσιγάρο ενώ από τα αγόρια το 21,7%.

Το ενδεχόμενο ένας νέος να είναι τωρινός καπνιστής παραδοσιακού τσιγάρου, ή χρήστης κάνναβης καθώς επίσης και η χρήση παραδοσιακού τσιγάρου από τα άτομα συναναστροφής, αποδείχθηκαν ότι είναι ιδιαίτερα σημαντικοί παράγοντες, στη χρησιμοποίηση ηλεκτρονικού τσιγάρου.

Επιπλέον, ο βαθμός ικανοποίησης αναφορικά με τη ζωή, όπως την αντιλαμβάνεται ο νέος, προέκυψε ως ιδιαίτερα σημαντικός παράγοντας, που υποδηλώνει ότι η καλή ψυχολογική κατάσταση αποτρέπει από την χρήση ηλεκτρονικού τσιγάρου.

Κατά την ανάλυση των προσδιοριστικών παραγόντων της χρήσης του ηλεκτρονικού τσιγάρου, οι δύο μέθοδοι είχαν παρόμοια αποτελέσματα, καταδεικνύοντας τις ίδιες σχεδόν μεταβλητές ως τις σημαντικότερες.

Βάσει της ανάλυση προκύπτει ότι, η μέθοδος του Classification Tree, είναι μία ασφαλή μέθοδο για εξαγωγή συμπερασμάτων με γρήγορο εύκολο και κατανοητό από τον χρήστη τρόπο ενώ τα αποτελέσματα που εξήχθησαν και αναλύθηκαν ήταν παρόμοια με την μέθοδο του Logistic Regression.

Παράρτημα 1 SmokeLt – Classification Tree

Αλγόριθμος :

```
#library
library(caTools)
library (ISLR)
library (tree)
library(rpart)
library(C50)
library(gmodels)
library(RWeka)
library(caret)

#data
smoke = as.data.frame(smoke)
set.seed(12345)
smoke <- smoke[order(runif(1127)),]
smoke$smokeLT=factor(smoke$smokeLT,levels=c("0","1"))

#view
str(smoke)
table(smoke$smokeLT)
prop.table((table(smoke$smokeLT)))

#split data
split= sample.split(smoke$smokeLT, SplitRatio = 0.80)
training_set <- subset(smoke, split==TRUE)
test_set<- subset(smoke, split==FALSE)
table(training_set$smokeLT)
table(test_set$smokeLT)
prop.table(table(training_set$smokeLT))
```

```
prop.table(table(test_set$smokeLT))

#model
model1<- C5.0(training_set[ ,-1], training_set$smokeLT)
model1
summary(model1)
plot(model1)

#evaluate performance
smoke_pred <- predict (model1, test_set)
smoke_pred
CrossTable(test_set$smokeLT, smoke_pred, prop.chisq = FALSE, prop.c = FALSE, prop.r
= FALSE,dnn = c('actual smokeLT', 'predicted smokeLT'))

#boost
model_boost10 <- C5.0(training_set[ ,-1],  training_set$smokeLT, trials = 10)
model_boost10
summary(model_boost10)

#evaluate performance after boost
model_boost10_pred <- predict (model_boost10, test_set)
model_boost10_pred
CrossTable(test_set$smokeLT, model_boost10_pred, prop.chisq = FALSE, prop.c =
FALSE, prop.r = FALSE, dnn = c('actual smokeLT', 'predicted smokeLT'))

#view
confusionMatrix (model_boost10_pred, test_set$smokeLT)
```

Αποτελέσματα

```
> #data
> smoke = as.data.frame(smoke)
```

```
> set.seed(12345)
> smoke <- smoke[order(runif(1127)),]
> smoke$smokeLT=factor(smoke$smokeLT,levels=c("0","1"))
>
> #view
> str(smoke)
'data.frame':    1127 obs. of  16 variables:
 $ smokeLT           : Factor w/ 2 levels "0","1": 1 2 1 2 2 2 1 1 2 1 ...
 $ Gender_Social     : num  2 2 1 2 1 2 1 1 2 1 ...
 $ Parents_Social    : num  2 2 2 1 2 1 1 2 2 2 ...
 $ Affluence_Social  : num  2 2 2 2 2 2 1 2 2 2 ...
 $ Talk_Family       : num  2 1 2 2 2 1 2 2 2 2 ...
 $ Communic_Family   : num  2 2 2 2 2 2 2 2 2 2 ...
 $ Support_Family    : num  2 1 2 2 2 2 2 2 2 2 ...
 $ Know_Family       : num  2 1 1 2 2 1 1 1 2 1 ...
 $ Relations_Family  : num  2 1 1 2 2 1 2 2 2 2 ...
 $ Activity_Health   : num  1 1 1 1 2 2 2 2 1 1 ...
 $ SelRe_Health      : num  2 2 2 2 2 2 2 2 2 2 ...
 $ Satisfaction_Health: num  2 1 2 2 2 1 2 1 2 1 ...
 $ Alcohol_Substance : num  2 2 2 2 1 2 2 2 2 2 ...
 $ Cannabis_ Substance: num  2 2 2 2 2 1 2 2 2 2 ...
 $ Other_Substance   : num  2 2 2 2 2 2 2 2 2 2 ...
 $ Peer_Substance    : num  2 2 2 2 1 1 2 2 1 2 ...
> table(smoke$smokeLT)

  0   1
721 406
> prop.table((table(smoke$smokeLT)))

        0         1
0.6397516 0.3602484
```

```
>
> #split data
> split= sample.split(smoke$smokeLT, SplitRatio = 0.80)
> training_set <- subset(smoke, split==TRUE)
> test_set<- subset(smoke, split==FALSE)
> table(training_set$smokeLT)

  0   1
577 325
> table(test_set$smokeLT)

  0   1
144  81
> prop.table(table(training_set$smokeLT))

       0         1
0.6396896 0.3603104
> prop.table(table(test_set$smokeLT))

  0    1
0.64 0.36
>
> #model
> model1<- C5.0(training_set[ ,-1], training_set$smokeLT)
> model1

Call:
C5.0.default(x = training_set[, -1], y = training_set$smokeLT)

Classification Tree
Number of samples: 902
```

Number of predictors: 15

Tree size: 3

Non-standard options: attempt to group attributes

> summary(model1)

Call:
C5.0.default(x = training_set[, -1], y = training_set$smokeLT)

C5.0 [Release 2.07 GPL Edition] Mon Jan 06 17:53:15 2020

Class specified by attribute `outcome'

Read 902 cases (16 attributes) from undefined.data

Decision tree:

Cannabis_ Substance <= 1: 1 (78/8)
Cannabis_ Substance > 1:
:...Peer_Substance <= 1: 1 (108/36)
    Peer_Substance > 1: 0 (716/183)

Evaluation on training data (902 cases):

Decision Tree

Size Errors

3 227(25.2%) <<

(a) (b) <-classified as

---- ----

533 44 (a): class 0

183 142 (b): class 1

Attribute usage:

100.00% Cannabis_ Substance

91.35% Peer_Substance

Time: 0.0 secs

> plot(model1)

Error in parse(text = x, keep.source = FALSE) :

<text>:1:224: unexpected symbol

1: y ~ Gender_Social + Parents_Social + Affluence_Social + Talk_Family + Communic_Family + Support_Family + Know_Family + Relations_Family + Activity_Health + SelRe_Health + Satisfaction_Health +

^

>

> #evaluate performance

> smoke_pred <- predict (model1, test_set)

> smoke_pred

```
 [1] 0 1 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 1
0 1 0 0 0 0 1 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 1 0 1 1 0
 [75] 0 0 0 0 0 0 1 0 1 0 1 1 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 1
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
[149] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0 1 0 1 0 1 0 0 0 0 0 1 1 1 0 1
0 0 0 0 0 1 0 0 0 0 1 1 0 1 0 1 0 0 1 0 0 0 0 0 0 1 0 1 1
[223] 0 0 1
Levels: 0 1
> CrossTable(test_set$smokeLT, smoke_pred, prop.chisq = FALSE, prop.c = FALSE,
prop.r = FALSE,dnn = c('actual smokeLT', 'predicted smokeLT'))


   Cell Contents
|-------------------------|
|                       N |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  225


               | predicted smokeLT
actual smokeLT |         0 |         1 | Row Total |
---------------|-----------|-----------|-----------|
             0 |       140 |         4 |       144 |
               |     0.622 |     0.018 |           |
---------------|-----------|-----------|-----------|
             1 |        39 |        42 |        81 |
               |     0.173 |     0.187 |           |
---------------|-----------|-----------|-----------|
```

```
 Column Total |       179 |        46 |       225 |
--------------|-----------|-----------|-----------|


>
> #boost
> model_boost10 <- C5.0(training_set[ ,-1],  training_set$smokeLT, trials = 10)
> model_boost10

Call:
C5.0.default(x = training_set[, -1], y = training_set$smokeLT, trials = 10)

Classification Tree
Number of samples: 902
Number of predictors: 15

Number of boosting iterations: 10
Average tree size: 3.7

Non-standard options: attempt to group attributes

> summary(model_boost10)

Call:
C5.0.default(x = training_set[, -1], y = training_set$smokeLT, trials = 10)


C5.0 [Release 2.07 GPL Edition]       Mon Jan 06 17:53:16 2020
-------------------------------

Class specified by attribute `outcome'
```

Read 902 cases (16 attributes) from undefined.data

----- Trial 0: -----

Decision tree:

Cannabis_ Substance <= 1: 1 (78/8)
Cannabis_ Substance > 1:
:...Peer_Substance <= 1: 1 (108/36)
    Peer_Substance > 1: 0 (716/183)

----- Trial 1: -----

Decision tree:

Cannabis_ Substance <= 1: 1 (70.3/11.9)
Cannabis_ Substance > 1:
:...Alcohol_Substance <= 1: 1 (76.8/26.7)
    Alcohol_Substance > 1: 0 (754.8/283.1)

----- Trial 2: -----

Decision tree:

Cannabis_ Substance <= 1: 1 (65.9/14)
Cannabis_ Substance > 1:
:...Know_Family <= 1: 1 (286.6/122.4)
    Know_Family > 1:
    :...Peer_Substance <= 1: 1 (74.7/27.8)
        Peer_Substance > 1: 0 (474.8/171.9)

----- Trial 3: -----

Decision tree:

Cannabis_ Substance <= 1: 1 (62.3/15.7)
Cannabis_ Substance > 1:
:...Support_Family <= 1: 1 (167.9/73.8)
    Support_Family > 1:
    :...Alcohol_Substance <= 1: 1 (69.3/30.7)
        Alcohol_Substance > 1: 0 (602.5/250)

----- Trial 4: -----

Decision tree:

Cannabis_ Substance <= 1: 1 (60/16.9)
Cannabis_ Substance > 1:
:...Peer_Substance <= 1: 1 (125.1/56.1)
    Peer_Substance > 1: 0 (717/326.8)

----- Trial 5: -----

Decision tree:

Cannabis_ Substance <= 1: 1 (58.7/17.8)
Cannabis_ Substance > 1:
:...Other_Substance <= 1: 1 (84.5/36.1)
    Other_Substance > 1:
    :...Parents_Social <= 1: 0 (121.1/49.9)
        Parents_Social > 1:

:...Affluence_Social <= 1: 0 (87.8/34.6)
Affluence_Social > 1: 1 (549.9/269.9)

----- Trial 6: -----

Decision tree:

Cannabis_ Substance <= 1: 1 (39.2)
Cannabis_ Substance > 1:
:...Other_Substance <= 1: 1 (85.3/39)
Other_Substance > 1:
:...Relations_Family <= 1: 1 (197.9/96.2)
Relations_Family > 1: 0 (571.7/255.1)

----- Trial 7: -----

Decision tree:

Cannabis_ Substance <= 1: 1 (38.2)
Cannabis_ Substance > 1:
:...Peer_Substance <= 1: 1 (135.3/56.5)
Peer_Substance > 1: 0 (649.5/213.5)

----- Trial 8: -----

Decision tree:

Cannabis_ Substance <= 1: 1 (34.1)
Cannabis_ Substance > 1:
:...Alcohol_Substance <= 1: 1 (107.1/27.6)
Alcohol_Substance > 1:

:...Support_Family <= 1: 1 (202.7/82.1)
Support_Family > 1: 0 (423.1/80.7)

----- Trial 9: -----

Decision tree:

Cannabis_ Substance <= 1: 1 (29.6)
Cannabis_ Substance > 1:
:...Peer_Substance <= 1: 1 (156.7/54.4)
Peer_Substance > 1:
:...Alcohol_Substance <= 1: 1 (96.9/39.6)
Alcohol_Substance > 1: 0 (461.8/74)

Evaluation on training data (902 cases):

Trial	Decision Tree	
	Size	Errors
0	3	227(25.2%)
1	3	249(27.6%)
2	4	276(30.6%)
3	4	272(30.2%)
4	3	227(25.2%)
5	5	469(52.0%)
6	4	313(34.7%)
7	3	227(25.2%)
8	4	272(30.2%)
9	4	227(25.2%)

boost 228(25.3%) <<

(a) (b) <-classified as
---- ----
529 48 (a): class 0
180 145 (b): class 1

Attribute usage:

100.00% Cannabis_ Substance
91.35% Support_Family
91.35% Know_Family
91.35% Alcohol_Substance
91.35% Other_Substance
91.35% Peer_Substance
82.93% Parents_Social
82.93% Relations_Family
69.62% Affluence_Social

Time: 0.0 secs

>
> #evaluate performance after boost
> model_boost10_pred <- predict (model_boost10, test_set)
> model_boost10_pred
[1] 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 1 1
0 1 0 0 0 0 1 0 0 0 0 1 1 1 0 0 0 0 0 0 1 0 0 1 0 1 1 0

```
 [75] 0 0 0 0 0 0 1 0 1 0 1 1 0 0 0 0 0 0 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 1
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1
[149] 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 1 0 1 0 1 0 1 0 0 0 0 0 1 1 1 0 1
0 0 0 0 0 1 0 0 0 0 1 1 0 1 0 1 0 0 1 0 0 0 0 0 0 1 0 1 1
[223] 0 0 1
Levels: 0 1
> CrossTable(test_set$smokeLT, model_boost10_pred, prop.chisq = FALSE, prop.c =
FALSE, prop.r = FALSE, dnn = c('actual smokeLT', 'predicted smokeLT'))


   Cell Contents
|-------------------------|
|                       N |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  225


               | predicted smokeLT
actual smokeLT |         0 |         1 | Row Total |
---------------|-----------|-----------|-----------|
             0 |       137 |         7 |       144 |
               |     0.609 |     0.031 |           |
---------------|-----------|-----------|-----------|
             1 |        38 |        43 |        81 |
               |     0.169 |     0.191 |           |
---------------|-----------|-----------|-----------|
  Column Total |       175 |        50 |       225 |
---------------|-----------|-----------|-----------|
```

```
>
> #view
> confusionMatrix (model_boost10_pred, test_set$smokeLT)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 137  38
         1   7  43

               Accuracy : 0.8
                 95% CI : (0.7417, 0.8502)
    No Information Rate : 0.64
    P-Value [Acc > NIR] : 1.294e-07

                  Kappa : 0.5263

 Mcnemar's Test P-Value : 7.744e-06

            Sensitivity : 0.9514
            Specificity : 0.5309
         Pos Pred Value : 0.7829
         Neg Pred Value : 0.8600
             Prevalence : 0.6400
         Detection Rate : 0.6089
   Detection Prevalence : 0.7778
      Balanced Accuracy : 0.7411

       'Positive' Class : 0
```

Παράρτημα 2 EsmokeLt – Classification Tree

Αλγόριθμος :

```
#library
library(randomForest)
library(ROCR)
library(caTools)
library (ISLR)
library (tree)
library(rpart)
library(C50)
library(gmodels)
library(RWeka)
library(caret)

#data
smoke = new
smoke <- as.data.frame(smoke)
set.seed(12345)
smoke <- smoke[order(runif(406)),]
smoke$esmokeLT=factor(smoke$esmokeLT,levels=c("0","1"))

#view
str(smoke)
table(smoke$esmokeLT)
prop.table((table(smoke$esmokeLT)))

#split data
```

```
split= sample.split(smoke$esmokeLT, SplitRatio = 0.80)
training_set <- subset(smoke, split==TRUE)
test_set<- subset(smoke, split==FALSE)
table(training_set$esmokeLT)
table(test_set$esmokeLT)
prop.table(table(training_set$esmokeLT))
prop.table(table(test_set$esmokeLT))

#model
model1<- C5.0(training_set[ ,-1], training_set$esmokeLT)
model1
summary(model1)
plot(model1)


#evaluate performance
smoke_pred <- predict (model1, test_set)
smoke_pred
CrossTable(test_set$esmokeLT, smoke_pred, prop.chisq = FALSE, prop.c = FALSE,
prop.r = FALSE,dnn = c('actual esmokeLT', 'predicted esmokeLT'))


#boost
model_boost10 <- C5.0(training_set[ ,-1],  training_set$esmokeLT, trials = 10)
model_boost10
summary(model_boost10)

#evaluate performance after boost
model_boost10_pred <- predict (model_boost10, test_set)
model_boost10_pred
```

```
CrossTable(test_set$esmokeLT, model_boost10_pred, prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE, dnn = c('actual esmokeLT', 'predicted smokeLT'))

#view
confusionMatrix (model_boost10_pred, test_set$esmokeLT)



#random forrest
set.seed((300))
rf <- randomForest(esmokeLT ~ .,data = smoke)
rf
plot(rf)

#evaluate random forrest
ctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 10)
```

Αποτελέσματα

```
> library(randomForest)
> library(ROCR)
> library(caTools)
> library (ISLR)
> library (tree)
> library(rpart)
> library(C50)
> library(gmodels)
> library(RWeka)
> library(caret)
>
> #data
> smoke = new
```

```
> smoke <- as.data.frame(smoke)
> set.seed(12345)
> smoke <- smoke[order(runif(406)),]
> smoke$esmokeLT=factor(smoke$esmokeLT,levels=c("0","1"))
>
> #view
> str(smoke)
'data.frame':    406 obs. of  18 variables:
 $ esmokeLT          : Factor w/ 2 levels "0","1": 2 2 1 2 2 1 1 1 1 2 ...
 $ Gender_Social     : num  1 2 2 1 2 1 2 1 2 2 ...
 $ Parents_Social    : num  2 2 2 2 2 2 2 2 2 1 ...
 $ Affluence_Social  : num  2 2 1 2 2 2 2 2 2 2 ...
 $ Talk_Family       : num  2 2 1 2 1 2 2 2 2 2 ...
 $ Communic_Family   : num  1 1 1 2 2 1 2 1 2 2 ...
 $ Support_Family    : num  1 1 1 2 2 2 2 2 2 2 ...
 $ Know_Family       : num  2 2 1 2 1 1 2 1 2 2 ...
 $ Relations_Family  : num  1 1 1 2 2 1 2 1 2 1 ...
 $ Activity_Health   : num  1 2 1 2 1 1 1 1 2 2 ...
 $ SelRe_Health      : num  2 1 2 2 2 2 2 2 2 2 ...
 $ Satisfaction_Health: num  1 1 1 1 1 1 2 2 1 1 ...
 $ Smoke_Substance   : num  0 1 1 0 1 0 0 0 0 1 ...
 $ Heavy_Substance   : num  0 0 0 0 1 0 0 0 0 0 ...
 $ Alcohol_Substance : num  2 2 2 2 2 2 1 2 2 2 ...
 $ Can_Substance     : num  2 1 1 2 1 2 2 2 2 1 ...
 $ Other_Substance   : num  2 2 1 2 2 2 2 2 2 2 ...
 $ Peer_Substance    : num  2 1 2 2 1 2 2 2 2 2 ...
> table(smoke$esmokeLT)

  0   1
255 151
> prop.table((table(smoke$esmokeLT)))
```

```
       0        1
0.6280788 0.3719212
>
> #split data
> split= sample.split(smoke$esmokeLT, SplitRatio = 0.80)
> training_set <- subset(smoke, split==TRUE)
> test_set<- subset(smoke, split==FALSE)
> table(training_set$esmokeLT)

  0   1
204 121
> table(test_set$esmokeLT)

 0  1
51 30
> prop.table(table(training_set$esmokeLT))

       0        1
0.6276923 0.3723077
> prop.table(table(test_set$esmokeLT))

       0        1
0.6296296 0.3703704
>
> #model
> model1<- C5.0(training_set[ ,-1], training_set$esmokeLT)
> model1

Call:
C5.0.default(x = training_set[, -1], y = training_set$esmokeLT)
```

Classification Tree
Number of samples: 325
Number of predictors: 17

Tree size: 8

Non-standard options: attempt to group attributes

> summary(model1)

Call:
C5.0.default(x = training_set[, -1], y = training_set$esmokeLT)

C5.0 [Release 2.07 GPL Edition] Tue Jan 07 22:35:35 2020

Class specified by attribute `outcome'

Read 325 cases (18 attributes) from undefined.data

Decision tree:

Smoke_Substance <= 0:
:...Can_Substance > 1: 0 (160/35)
: Can_Substance <= 1:
: :...Peer_Substance <= 1: 1 (2)
:   Peer_Substance > 1:
:   :...Communic_Family <= 1: 1 (6/1)
:   Communic_Family > 1: 0 (6)

```
Smoke_Substance > 0:
:...Talk_Family <= 1: 1 (33/8)
    Talk_Family > 1:
    :...Gender_Social > 1: 0 (59/18)
        Gender_Social <= 1:
        :...Relations_Family <= 1: 0 (14/5)
            Relations_Family > 1: 1 (45/14)
```

Evaluation on training data (325 cases):

```
          Decision Tree
        ----------------
        Size      Errors

           8   81(24.9%)   <<


         (a)   (b)    <-classified as
        ----  ----
         181    23    (a): class 0
          58    63    (b): class 1
```

Attribute usage:

```
        100.00%         Smoke_Substance
         53.54%         Can_Substance
         46.46%         Talk_Family
         36.31%         Gender_Social
         18.15%         Relations_Family
```

4.31% Peer_Substance

3.69% Communic_Family

Time: 0.0 secs

```
> plot(model1)
>
>
> #evaluate performance
> smoke_pred <- predict (model1, test_set)
> smoke_pred
 [1] 0 0 1 0 1 1 1 0 0 0 0 0 0 0 1 1 0 0 1 0 0 1 1 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
0 0 0 0 0 0 0 1 1 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
[76] 0 0 0 1 1 0
Levels: 0 1
> CrossTable(test_set$esmokeLT, smoke_pred, prop.chisq = FALSE, prop.c = FALSE,
prop.r = FALSE,dnn = c('actual esmokeLT', 'predicted esmokeLT'))


   Cell Contents
|-------------------------|
|                       N |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  81


                | predicted esmokeLT
```

```
actual esmokeLT |         0 |         1 | Row Total |
----------------|-----------|-----------|-----------|
              0 |        44 |         7 |        51 |
                |     0.543 |     0.086 |           |
----------------|-----------|-----------|-----------|
              1 |        18 |        12 |        30 |
                |     0.222 |     0.148 |           |
----------------|-----------|-----------|-----------|
   Column Total |        62 |        19 |        81 |
----------------|-----------|-----------|-----------|
```

```
>
>
> #boost
> model_boost10 <- C5.0(training_set[ ,-1],  training_set$esmokeLT, trials = 10)
> model_boost10

Call:
C5.0.default(x = training_set[, -1], y = training_set$esmokeLT, trials = 10)

Classification Tree
Number of samples: 325
Number of predictors: 17

Number of boosting iterations: 10
Average tree size: 6.8

Non-standard options: attempt to group attributes

> summary(model_boost10)
```

Call:
C5.0.default(x = training_set[, -1], y = training_set$esmokeLT, trials = 10)

C5.0 [Release 2.07 GPL Edition] Tue Jan 07 22:35:36 2020

Class specified by attribute `outcome'

Read 325 cases (18 attributes) from undefined.data

----- Trial 0: -----

Decision tree:

Smoke_Substance <= 0:
:...Can_Substance > 1: 0 (160/35)
:   Can_Substance <= 1:
:   :...Peer_Substance <= 1: 1 (2)
:   Peer_Substance > 1:
:   :...Communic_Family <= 1: 1 (6/1)
:   Communic_Family > 1: 0 (6)
Smoke_Substance > 0:
:...Talk_Family <= 1: 1 (33/8)
    Talk_Family > 1:
    :...Gender_Social > 1: 0 (59/18)
        Gender_Social <= 1:
        :...Relations_Family <= 1: 0 (14/5)
            Relations_Family > 1: 1 (45/14)

----- Trial 1: -----

Decision tree:

```
Peer_Substance <= 1:
:...Can_Substance <= 1: 1 (35.3/9.2)
:   Can_Substance > 1:
:   :...Communic_Family <= 1: 1 (22.2/7.2)
:       Communic_Family > 1: 0 (55.2/21.9)
Peer_Substance > 1:
:...Gender_Social > 1: 0 (107.4/29)
    Gender_Social <= 1:
    :...Support_Family <= 1: 1 (21.4/8.2)
        Support_Family > 1: 0 (83.6/34.4)
```

----- Trial 2: -----

Decision tree:

```
Smoke_Substance <= 0: 0 (165.1/65.1)
Smoke_Substance > 0:
:...Parents_Social <= 1: 1 (33.3/10.8)
    Parents_Social > 1:
    :...Heavy_Substance > 0: 1 (47.7/19)
        Heavy_Substance <= 0:
        :...Affluence_Social <= 1: 0 (8.8/1.3)
            Affluence_Social > 1:
            :...Other_Substance <= 1: 1 (14.5/4.6)
                Other_Substance > 1:
                :...Know_Family <= 1: 0 (20.5/4)
                    Know_Family > 1: 1 (35/14.4)
```

----- Trial 3: -----

Decision tree:

```
Gender_Social <= 1:
:...Can_Substance <= 1: 1 (30.8/10.3)
:   Can_Substance > 1:
:   :...Activity_Health <= 1: 0 (34.1/14.1)
:       Activity_Health > 1: 1 (90.9/39)
Gender_Social > 1:
:...Satisfaction_Health > 1: 0 (58.7/15.5)
    Satisfaction_Health <= 1:
    :...Activity_Health > 1: 1 (44.9/19.3)
        Activity_Health <= 1:
        :...Heavy_Substance <= 0: 0 (51.8/18.1)
            Heavy_Substance > 0: 1 (13.7/5.5)
```

----- Trial 4: -----

Decision tree:

```
Gender_Social > 1: 0 (166.8/68)
Gender_Social <= 1:
:...Can_Substance <= 1: 1 (30.5/12)
    Can_Substance > 1:
    :...Talk_Family <= 1: 0 (20.3/8)
        Talk_Family > 1:
        :...Support_Family <= 1: 1 (13.9/3.8)
            Support_Family > 1:
            :...Communic_Family <= 1: 1 (18.3/7)
```

Communic_Family > 1:
:...Satisfaction_Health <= 1: 0 (30.1/9.2)
Satisfaction_Health > 1: 1 (45/20.7)

----- Trial 5: -----

Decision tree:

Peer_Substance <= 1:
:...Talk_Family <= 1: 1 (21.5/4.9)
: Talk_Family > 1:
: :...Satisfaction_Health <= 1: 0 (53.9/23.2)
: Satisfaction_Health > 1: 1 (38.2/16)
Peer_Substance > 1:
:...Alcohol_Substance <= 1: 0 (22.9/6.7)
Alcohol_Substance > 1:
:...Heavy_Substance > 0: 1 (14.4/5.7)
Heavy_Substance <= 0:
:...Satisfaction_Health > 1: 0 (67.2/23.4)
Satisfaction_Health <= 1:
:...Know_Family <= 1: 0 (49.4/20.8)
Know_Family > 1: 1 (57.6/25.5)

----- Trial 6: -----

Decision tree:

Peer_Substance <= 1:
:...Talk_Family <= 1: 1 (20.8/5.7)
: Talk_Family > 1:
: :...Gender_Social <= 1: 1 (39.6/16.6)

: Gender_Social > 1: 0 (54.4/24.1)
Peer_Substance > 1:
:...Gender_Social > 1: 0 (94.5/36.4)
   Gender_Social <= 1:
   :...Parents_Social <= 1: 1 (18/6)
      Parents_Social > 1:
      :...Relations_Family <= 1: 0 (27.1/9.8)
         Relations_Family > 1:
         :...Talk_Family <= 1: 1 (10.8/2.5)
            Talk_Family > 1: 0 (58.8/24.9)

----- Trial 7: -----

Decision tree:

Smoke_Substance > 0:
:...Parents_Social <= 1: 1 (34.8/11.4)
: Parents_Social > 1:
: :...Communic_Family <= 1: 0 (53.8/22.9)
: Communic_Family > 1: 1 (74.2/34.5)
Smoke_Substance <= 0:
:...Activity_Health <= 1: 0 (36.3/3.9)
   Activity_Health > 1:
   :...Parents_Social <= 1: 0 (18.7/4.3)
      Parents_Social > 1:
      :...Can_Substance <= 1: 1 (9.8/3.3)
         Can_Substance > 1:
         :...Gender_Social <= 1: 1 (55.7/24.4)
            Gender_Social > 1: 0 (28.7/10)

----- Trial 8: -----

Decision tree:

Smoke_Substance <= 0: 0 (140.5/45.8)
Smoke_Substance > 0:
:...Talk_Family <= 1: 1 (32.2/10.2)
    Talk_Family > 1:
    :...Know_Family <= 1: 0 (44.9/15.6)
        Know_Family > 1:
        :...Alcohol_Substance <= 1: 1 (29.8/8.9)
            Alcohol_Substance > 1: 0 (57.7/25)

----- Trial 9: -----

Decision tree:

Smoke_Substance <= 0:
:...Can_Substance <= 1: 1 (18.2/6.3)
:   Can_Substance > 1: 0 (94.1/15.7)
Smoke_Substance > 0:
:...Gender_Social <= 1: 1 (91.5/31.6)
    Gender_Social > 1: 0 (78.2/32.1)

Evaluation on training data (325 cases):

Trial	Decision Tree	
	Size	Errors
0	8	81(24.9%)

1	6	98(30.2%)
2	7	97(29.8%)
3	7	119(36.6%)
4	7	107(32.9%)
5	8	119(36.6%)
6	8	92(28.3%)
7	8	123(37.8%)
8	5	96(29.5%)
9	4	101(31.1%)
boost		73(22.5%) <<

(a)	(b)	<-classified as
188	16	(a): class 0
57	64	(b): class 1

Attribute usage:

100.00%	Gender_Social
100.00%	Smoke_Substance
100.00%	Peer_Substance
95.08%	Satisfaction_Health
90.77%	Can_Substance
85.23%	Heavy_Substance
84.00%	Parents_Social
83.38%	Activity_Health
78.46%	Alcohol_Substance
73.54%	Talk_Family
64.62%	Know_Family

64.00% Communic_Family

39.69% Support_Family

36.62% Relations_Family

22.15% Affluence_Social

19.69% Other_Substance

Time: 0.0 secs

>

> #evaluate performance after boost

> model_boost10_pred <- predict (model_boost10, test_set)

> model_boost10_pred

[1] 0 0 1 0 1 1 1 0 0 0 0 1 0 0 1 1 0 0 1 0 0 1 1 0 1 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 1

[76] 0 0 0 1 1 0

Levels: 0 1

> CrossTable(test_set$esmokeLT, model_boost10_pred, prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE, dnn = c('actual esmokeLT', 'predicted smokeLT'))

Cell Contents

|-------------------------|

| N |

| N / Table Total |

|-------------------------|

Total Observations in Table: 81

```
             | predicted smokeLT
actual esmokeLT |         0 |         1 | Row Total |
----------------|-----------|-----------|-----------|
              0 |        45 |         6 |        51 |
                |     0.556 |     0.074 |           |
----------------|-----------|-----------|-----------|
              1 |        15 |        15 |        30 |
                |     0.185 |     0.185 |           |
----------------|-----------|-----------|-----------|
   Column Total |        60 |        21 |        81 |
----------------|-----------|-----------|-----------|
```

```
>
> #view
> confusionMatrix (model_boost10_pred, test_set$esmokeLT)
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 45 15
         1  6 15

               Accuracy : 0.7407
                 95% CI : (0.6314, 0.8318)
    No Information Rate : 0.6296
    P-Value [Acc > NIR] : 0.02317

                  Kappa : 0.4075

 Mcnemar's Test P-Value : 0.08086
```

```
            Sensitivity : 0.8824
            Specificity : 0.5000
         Pos Pred Value : 0.7500
         Neg Pred Value : 0.7143
             Prevalence : 0.6296
         Detection Rate : 0.5556
   Detection Prevalence : 0.7407
      Balanced Accuracy : 0.6912

       'Positive' Class : 0

>
>
> #random forrest
> set.seed((300))
> rf <- randomForest(esmokeLT ~ .,data = smoke)
> rf

Call:
 randomForest(formula = esmokeLT ~ ., data = smoke)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 32.76%
Confusion matrix:
    0  1 class.error
0 203 52   0.2039216
1  81 70   0.5364238
> plot(rf)
```

>

> #evaluate random forrest

> ctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 10)

Παράρτημα 3 SmokeLt – Logistic Regression

Αλγόριθμος

```
#library
library(ElemStatLearn)
library(dplyr)
library(caTools)
library (ISLR)
library (tree)
library(rpart)
library(C50)
library(gmodels)
library(RWeka)
library(caret)

#data
smoke=smoke1
smoke <- as.data.frame(smoke)
set.seed(12345)
smoke <- smoke[order(runif(1127)),]
smoke$smokeLT=factor(smoke$smokeLT,levels=c("0","1"))

#view
table(smoke$smokeLT)
prop.table((table(smoke$smokeLT)))

#split data
split= sample.split(smoke$smokeLT, SplitRatio = 0.80)
```

```
training_set <- subset(smoke, split==TRUE)
test_set<- subset(smoke, split==FALSE)
table(training_set$smokeLT)
table(test_set$smokeLT)
prop.table(table(training_set$smokeLT))
prop.table(table(test_set$smokeLT))

#model logistic regression
glm.fits= glm(smokeLT~.,data=training_set,family=binomial)
glm.fits
summary(glm.fits)

#evaluate performance
smoke_pred <- predict (glm.fits, test_set[-1], type= "response")
table_mat<- table(test_set$smokeLT,smoke_pred > 0.5)
table_mat
```

Αποτελέσματα για το σύνολο των μεταβλητών (smoke1) :

```
> #library
> library(ElemStatLearn)
> library(dplyr)
> library(caTools)
> library (ISLR)
> library (tree)
> library(rpart)
> library(C50)
> library(gmodels)
> library(RWeka)
> library(caret)
>
```

```
> #data
> smoke=smoke1
> smoke <- as.data.frame(smoke)
> set.seed(12345)
> smoke <- smoke[order(runif(1127)),]
> smoke$smokeLT=factor(smoke$smokeLT,levels=c("0","1"))
>
> #view
> table(smoke$smokeLT)

  0   1
721 406
> prop.table((table(smoke$smokeLT)))

        0         1
0.6397516 0.3602484
>
> #split data
> split= sample.split(smoke$smokeLT, SplitRatio = 0.80)
> training_set <- subset(smoke, split==TRUE)
> test_set<- subset(smoke, split==FALSE)
> table(training_set$smokeLT)

  0   1
577 325
> table(test_set$smokeLT)

  0   1
144  81
> prop.table(table(training_set$smokeLT))
```

```
       0        1
0.6396896 0.3603104
> prop.table(table(test_set$smokeLT))

  0    1
0.64 0.36
>
> #model logistic regression
> glm.fits= glm(smokeLT~.,data=training_set,family=binomial)
> glm.fits

Call:  glm(formula = smokeLT ~ ., family = binomial, data = training_set)

Coefficients:
        (Intercept)       Gender_Social       Parents_Social     Affluence_Social              T
alk_Family      Communic_Family      Support_Family
          11.543495            0.301670           -0.044324            0.268409            0.059
070           -0.343571           -0.230221
        Know_Family     Relations_Family     Activity_Health        SelRe_Health  Sa
tisfaction_Health    Alcohol_Substance  `Cannabis_ Substance`
          -0.506092           -0.293974            0.097205           -0.149404            0.003
865           -0.990798           -2.424961
    Other_Substance       Peer_Substance
          -0.369851           -1.837560

Degrees of Freedom: 901 Total (i.e. Null);  886 Residual
Null Deviance:          1179
Residual Deviance: 930.9     AIC: 962.9
> summary(glm.fits)

Call:
```

```
glm(formula = smokeLT ~ ., family = binomial, data = training_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9407  -0.7330  -0.5848   0.7576   2.0816

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         11.543495   1.370391   8.424  < 2e-16 ***
Gender_Social        0.301670   0.171390   1.760 0.078385 .
Parents_Social      -0.044324   0.212313  -0.209 0.834631
Affluence_Social     0.268409   0.229374   1.170 0.241928
Talk_Family          0.059070   0.248333   0.238 0.811986
Communic_Family     -0.343571   0.212583  -1.616 0.106056
Support_Family      -0.230221   0.262217  -0.878 0.379954
Know_Family         -0.506092   0.180274  -2.807 0.004995 **
Relations_Family    -0.293974   0.231203  -1.272 0.203551
Activity_Health      0.097205   0.170729   0.569 0.569116
SelRe_Health        -0.149404   0.295104  -0.506 0.612663
Satisfaction_Health  0.003865   0.175817   0.022 0.982461
Alcohol_Substance   -0.990798   0.272800  -3.632 0.000281 ***
`Cannabis_ Substance` -2.424961   0.412086  -5.885 3.99e-09 ***
Other_Substance     -0.369851   0.263518  -1.404 0.160464
Peer_Substance      -1.837560   0.240365  -7.645 2.09e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1179.09  on 901  degrees of freedom
Residual deviance:  930.86  on 886  degrees of freedom
```

AIC: 962.86

Number of Fisher Scoring iterations: 5

```
>
> #evaluate performance
> smoke_pred <- predict (glm.fits, test_set[-1], type= "response")
> table_mat<- table(test_set$smokeLT,smoke_pred > 0.5)
> table_mat

   FALSE TRUE
 0   131   13
 1    40   41
```

Αποτελέσματα με αφαίρεση των μη σημαντικών μεταβλητών (smoke2)

```
> #library
> library(ElemStatLearn)
> library(dplyr)
> library(caTools)
> library (ISLR)
> library (tree)
> library(rpart)
> library(C50)
> library(gmodels)
> library(RWeka)
> library(caret)
>
> #data
> smoke=smoke2
```

```
> smoke <- as.data.frame(smoke)
> set.seed(12345)
> smoke <- smoke[order(runif(1127)),]
> smoke$smokeLT=factor(smoke$smokeLT,levels=c("0","1"))
>
> #view
> table(smoke$smokeLT)

  0   1
721 406
> prop.table((table(smoke$smokeLT)))

       0         1
0.6397516 0.3602484
>
> #split data
> split= sample.split(smoke$smokeLT, SplitRatio = 0.80)
> training_set <- subset(smoke, split==TRUE)
> test_set<- subset(smoke, split==FALSE)
> table(training_set$smokeLT)

  0   1
577 325
> table(test_set$smokeLT)

  0   1
144  81
> prop.table(table(training_set$smokeLT))

       0         1
0.6396896 0.3603104
```

```
> prop.table(table(test_set$smokeLT))

   0    1
0.64 0.36
>
> #model logistic regression
> glm.fits= glm(smokeLT~.,data=training_set,family=binomial)
> glm.fits

Call:  glm(formula = smokeLT ~ ., family = binomial, data = training_set)

Coefficients:
        (Intercept)        Gender_Social       Communic_Family          Know_Family     A
lcohol_Substance  `Cannabis_ Substance`       Other_Substance
              11.4932               0.3181              -0.5413              -0.5686              -0.9839
-2.4500              -0.4033
      Peer_Substance
              -1.8146

Degrees of Freedom: 901 Total (i.e. Null);  894 Residual
Null Deviance:          1179
Residual Deviance: 936.8     AIC: 952.8
> summary(glm.fits)

Call:
glm(formula = smokeLT ~ ., family = binomial, data = training_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8710  -0.7438  -0.5760   0.7666   1.9382
```

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          11.4932     1.1517   9.979  < 2e-16 ***
Gender_Social         0.3181     0.1655   1.921 0.054686 .
Communic_Family      -0.5413     0.1841  -2.940 0.003281 **
Know_Family          -0.5686     0.1734  -3.279 0.001042 **
Alcohol_Substance    -0.9839     0.2702  -3.642 0.000271 ***
`Cannabis_ Substance`  -2.4500     0.4096  -5.981 2.22e-09 ***
Other_Substance      -0.4033     0.2600  -1.551 0.120808
Peer_Substance       -1.8146     0.2375  -7.640 2.17e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1179.09  on 901  degrees of freedom
Residual deviance:  936.77  on 894  degrees of freedom
AIC: 952.77

Number of Fisher Scoring iterations: 5

>
> #evaluate performance
> smoke_pred <- predict (glm.fits, test_set[-1], type= "response")
> table_mat<- table(test_set$smokeLT,smoke_pred > 0.5)
> table_mat

    FALSE TRUE
  0   131   13
  1    40   41
```

Παράρτημα 4 EsmokeLt – Logistic Regression

Αλγόριθμος :

```
#library
library(ElemStatLearn)
library(dplyr)
library(caTools)
library (ISLR)
library (tree)
library(rpart)
library(C50)
library(gmodels)
library(RWeka)
library(caret)

#data
esmoke=esmoke1
esmoke <- as.data.frame(esmoke)
set.seed(12345)
esmoke <- esmoke[order(runif(406)),]
esmoke$esmokeLT=factor(esmoke$esmokeLT,levels=c("0","1"))

#view
glimpse(esmoke)
table(esmoke$esmokeLT)
prop.table((table(esmoke$esmokeLT)))

#split data
split= sample.split(esmoke$esmokeLT, SplitRatio = 0.80)
training_set <- subset(esmoke, split==TRUE)
test_set<- subset(esmoke, split==FALSE)
table(training_set$esmokeLT)
```

```
table(test_set$esmokeLT)
prop.table(table(training_set$esmokeLT))
prop.table(table(test_set$esmokeLT))

#model logistic regression
glm.fits= glm(esmokeLT~.,data=training_set,family=binomial)
glm.fits
summary(glm.fits)

#evaluate performance
esmoke_pred <- predict (glm.fits, test_set, type= "response")
table_mat<- table(test_set$esmokeLT,esmoke_pred > 0.5)
table_mat
```

Αποτελέσματα για smoke 1

```
> #library
> library(ElemStatLearn)
> library(dplyr)
> library(caTools)
> library (ISLR)
> library (tree)
> library(rpart)
> library(C50)
> library(gmodels)
> library(RWeka)
> library(caret)
>
> #data
> esmoke=esmoke1
> esmoke <- as.data.frame(esmoke)
> set.seed(12345)
```

```
> esmoke <- esmoke[order(runif(406)),]
> esmoke$esmokeLT=factor(esmoke$esmokeLT,levels=c("0","1"))
>
> #view
> table(esmoke$esmokeLT)

  0   1
255 151
> prop.table((table(esmoke$esmokeLT)))

       0         1
0.6280788 0.3719212
>
> #split data
> split= sample.split(esmoke$esmokeLT, SplitRatio = 0.80)
> training_set <- subset(esmoke, split==TRUE)
> test_set<- subset(esmoke, split==FALSE)
> table(training_set$esmokeLT)

  0   1
204 121
> table(test_set$esmokeLT)

 0  1
51 30
> prop.table(table(training_set$esmokeLT))

       0         1
0.6276923 0.3723077
> prop.table(table(test_set$esmokeLT))
```

```
       0         1
0.6296296 0.3703704
>
> #model logistic regression
> glm.fits= glm(esmokeLT~.,data=training_set,family=binomial)
> glm.fits

Call:  glm(formula = esmokeLT ~ ., family = binomial, data = training_set)

Coefficients:
       (Intercept)        Gender_Social       Parents_Social     Affluence_Social          Talk_Fa
mily      Communic_Family      Support_Family          Know_Family
           2.91741           -1.04088           -0.35671            0.39253           -0.82617
0.04231           -0.25877            0.19086
  Relations_Family      Activity_Health         SelRe_Health  Satisfaction_Health      Smok
e_Substance     Heavy_Substance    Alcohol_Substance         Can_Substance
           0.20971            0.42263           -0.18903           -0.19955            0.92630
0.32228            0.19873           -0.53184
   Other_Substance       Peer_Substance
           0.15779           -0.61502

Degrees of Freedom: 324 Total (i.e. Null);  307 Residual
Null Deviance:          429.1
Residual Deviance: 363        AIC: 399
> summary(glm.fits)

Call:
glm(formula = esmokeLT ~ ., family = binomial, data = training_set)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
```

-1.9552 -0.8473 -0.5328 0.9994 2.1902

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	2.91741	1.76777	1.650	0.098874	.
Gender_Social	-1.04088	0.28128	-3.700	0.000215	***
Parents_Social	-0.35671	0.31857	-1.120	0.262835	
Affluence_Social	0.39253	0.36523	1.075	0.282481	
Talk_Family	-0.82617	0.36764	-2.247	0.024624	*
Communic_Family	0.04231	0.33848	0.125	0.900517	
Support_Family	-0.25877	0.37287	-0.694	0.487685	
Know_Family	0.19086	0.29282	0.652	0.514528	
Relations_Family	0.20971	0.34613	0.606	0.544607	
Activity_Health	0.42263	0.28351	1.491	0.136034	
SelRe_Health	-0.18903	0.39857	-0.474	0.635307	
Satisfaction_Health	-0.19955	0.28579	-0.698	0.485029	
Smoke_Substance	0.92630	0.31913	2.903	0.003701	**
Heavy_Substance	0.32228	0.39827	0.809	0.418392	
Alcohol_Substance	0.19873	0.34292	0.580	0.562226	
Can_Substance	-0.53184	0.35619	-1.493	0.135408	
Other_Substance	0.15779	0.37105	0.425	0.670647	
Peer_Substance	-0.61502	0.30329	-2.028	0.042577	*

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 429.11 on 324 degrees of freedom

Residual deviance: 363.05 on 307 degrees of freedom

AIC: 399.05

```
Number of Fisher Scoring iterations: 4

>
> #evaluate performance
> esmoke_pred <- predict (glm.fits, test_set, type= "response")
> table_mat<- table(test_set$esmokeLT,esmoke_pred > 0.5)
> table_mat

   FALSE TRUE
0    44    7
1    13   17
```

Αποτελέσματα για smoke 2

```
> #library
> library(ElemStatLearn)
> library(dplyr)
> library(caTools)
> library (ISLR)
> library (tree)
> library(rpart)
> library(C50)
> library(gmodels)
> library(RWeka)
> library(caret)
>
> #data
> esmoke=esmoke2
> esmoke <- as.data.frame(esmoke)
> set.seed(12345)
> esmoke <- esmoke[order(runif(406)),]
> esmoke$esmokeLT=factor(esmoke$esmokeLT,levels=c("0","1"))
```

```
>
> #view
> table(esmoke$esmokeLT)

  0   1
255 151
> prop.table((table(esmoke$esmokeLT)))

        0         1
0.6280788 0.3719212
>
> #split data
> split= sample.split(esmoke$esmokeLT, SplitRatio = 0.80)
> training_set <- subset(esmoke, split==TRUE)
> test_set<- subset(esmoke, split==FALSE)
> table(training_set$esmokeLT)

  0   1
204 121
> table(test_set$esmokeLT)

 0  1
51 30
> prop.table(table(training_set$esmokeLT))

        0         1
0.6276923 0.3723077
> prop.table(table(test_set$esmokeLT))

        0         1
0.6296296 0.3703704
```

```
>
> #model logistic regression
> glm.fits= glm(esmokeLT~.,data=training_set,family=binomial)
> glm.fits

Call:  glm(formula = esmokeLT ~ ., family = binomial, data = training_set)

Coefficients:
   (Intercept)    Gender_Social      Talk_Family  Activity_Health  Smoke_Substance    Ca
n_Substance    Peer_Substance
        3.5713         -1.0057         -0.8594          0.3619          0.9401         -0.6038         -
0.6181

Degrees of Freedom: 324 Total (i.e. Null);  318 Residual
Null Deviance:          429.1
Residual Deviance: 368.3     AIC: 382.3
> summary(glm.fits)

Call:
glm(formula = esmokeLT ~ ., family = binomial, data = training_set)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.7506  -0.8354  -0.5332   1.0436   2.1642

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)       3.5713     1.1876   3.007 0.002637 **
Gender_Social    -1.0057     0.2673  -3.763 0.000168 ***
Talk_Family      -0.8594     0.3103  -2.770 0.005606 **
Activity_Health   0.3619     0.2720   1.330 0.183386
```

```
Smoke_Substance   0.9401     0.2899   3.243 0.001184 **
Can_Substance    -0.6038     0.3144  -1.921 0.054770 .
Peer_Substance   -0.6181     0.2876  -2.149 0.031636 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 429.11  on 324  degrees of freedom
Residual deviance: 368.30  on 318  degrees of freedom
AIC: 382.3

Number of Fisher Scoring iterations: 4

>
> #evaluate performance
> esmoke_pred <- predict (glm.fits, test_set, type= "response")
> table_mat<- table(test_set$esmokeLT,esmoke_pred > 0.5)
> table_mat

   FALSE TRUE
 0    44    7
 1    13   17
```

Πηγές – Βιβλιογραφία

Ξενόγλωσση:

1	R. J. Brachman, T. Anand (1994) ,"The Process of Knowledge Discovery in Databases A first sketch ", Technical Report WS 94-03, (www.aaai.org).
2	L.Breiman, A.Cutler, "Random Forests for Scientific Discovery", UC Berkeley, Utah State University (http://www.math.usu.edu/~adele/RandomForests/ENAR.pdf).
3	G. Casella, S. Fienberg, I. Olkin, (2013) ,"An Introduction to Statistical Learning with Applications in R", Springer.
4	U.Fayyad, G.Piatetsky-Shapiro, and P.Smyth, (1996) ,"Knowledge Discovery and Data Mining Towards a Unifying Framework ", AAAI (www.aaai.org).
5	U.Fayyad, G.Piatetsky-Shapiro, and P.Smyth, (1997) ,"From Data Mining to Knowledge Discovery in Databases", AI Magazine 17: 37-54.
6	A. Fotiou, E. Kanavou, M. Stavrou, C.Richardson, A.Kokkevi, (2015) ,"Prevalence and correlates of electronic cigarette use among adolescents in Greece: A preliminary cross-sectional analysis of nationwide survey data",Addictive Behaviors 51 88–92.
7	B. Lantz, "Machine Learning with R", Packt Publishing .
8	J. Morgan, (2014) ,"Classification and Regression Tree Analysis", Boston University School of Public Health Department of Health Policy & Management (https://www.bu.edu/sph/files/2014/05/MorganCART.pdf).
9	Quinlan J.R, (1986) ,"Induction of Decision Trees ", Kluwer Academic Publishers Boston (https://link.springer.com/article/10.1007/BF00116251).
10	Quinlan J.R, (1993) ,"Programs for Machine Learning ", Morgan Kaufmann Publishers(https://link.springer.com/article/10.1007/BF00993309).
11	Quinlan J.R, Kohavi R., (1999) "Decision Tree Discovery"(http://ai.stanford.edu/~ronnyk/treesHB.pdf).
12	C.Rohilla Shalizi, (2015) ,"Classification and Regression Trees" (https://www.stat.cmu.edu/~cshalizi/mreg/15/lectures/27/lecture-27.pdf),
13	R.E. Schapire, Y. Freund, (2012), “Boosting Foundations and Algorithms”, The MIT Press

	(https://doc.lagout.org/science/0_Computer%20Science/2_Algorithms/Boosting_%20Foundations%20and%20Algorithms%20%5BSchapire%20%26%20Freund%202012-05-18%5D.pdf)
14	H.Sharma, S. Kumar, (2015) ,"A Survey on Decision Tree Algorithms of Classification in Data Mining", International Journal of Science and Research (https://www.researchgate.net/publication/324941161_A_Survey_on_Decision_Tree_Algorithms_of_Classification_in_Data_Mining).
15	Wei-Yin Loh, (2008), "Classification and Regression Tree Methods", Encyclopedia of Statistics in Quality and Reliability. (http://pages.stat.wisc.edu/~loh/treeprogs/guide/eqr.pdf).
16	Wei-Yin Loh(2011) ,"Classification and Regression Trees", University of Wisconsin, Madison, USA (https://www.researchgate.net/publication/227658748_Classification_and_Regression_Trees).
17	Z. Zhang, (2016),"Decision tree modeling using R", Zhejiang University. (http://atm.amegroups.com/article/view/10459/html)

Ελληνική:

1.	Ζαγγανά Ελένη, (2012) "Αναγνώριση και κατάταξη ονομάτων οντοτήτων σε ελληνικά κείμενα με χρήση τυχαίων δασών", Πανεπιστήμιο Πατρών.
2.	Καρανικόλας Ανδρέας, (2010) "Δημιουργία μοντέλου γνώσης από βάση δεδομένων βλαβών ADSL με την χρήση εργαλείων Data Mining", Πανεπιστήμιο Μακεδονίας.
3.	Παιδούση Ελευθερία (2016) "Δέντρα Αποφάσεων", Πανεπιστήμιο Πατρών.