



# ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΣΧΟΛΗ ΨΗΦΙΑΚΗΣ ΤΕΧΝΟΛΟΓΙΑΣ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΜΑΤΙΚΗΣ

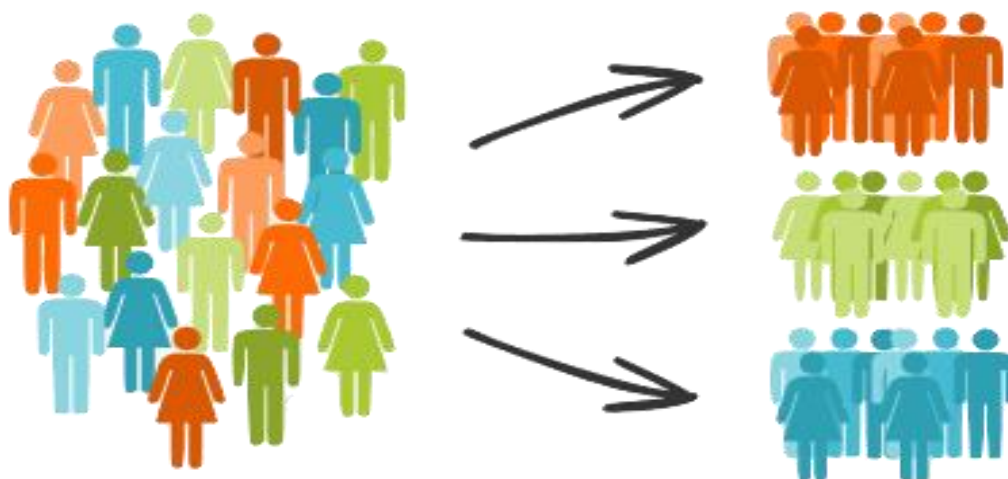
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ: ΠΛΗΡΟΦΟΡΙΚΗ ΚΑΙ  
ΤΗΛΕΜΑΤΙΚΗ

ΚΑΤΕΥΘΥΝΣΗ: ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΣΤΗ ΔΙΟΙΚΗΣΗ  
ΕΠΙΧΕΙΡΗΣΕΩΝ

**Αλγόριθμοι και Πληροφοριακά Συστήματα για την αυτοματοποιημένη  
αξιολόγηση και βαθμολόγηση στο τομέα της φορολογίας**

Διπλωματική Εργασία

Παπαστεργίου Αθανασία



Αθήνα, 2019



# **ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ**

**ΣΧΟΛΗ ΨΗΦΙΑΚΗΣ ΤΕΧΝΟΛΟΓΙΑΣ**

**ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΜΑΤΙΚΗΣ**

**ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ: ΠΛΗΡΟΦΟΡΙΚΗ ΚΑΙ  
ΤΗΛΕΜΑΤΙΚΗ**

**ΚΑΤΕΥΘΥΝΣΗ: ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΣΤΗ ΔΙΟΙΚΗΣΗ  
ΕΠΙΧΕΙΡΗΣΕΩΝ**

## **Τριμελής Εξεταστική Επιτροπή**

**Αναγνωστόπουλος Δημοσθένης (Επιβλέπων)**

**Καθηγητής, Τμήμα Πληροφορικής και Τηλεματικής,  
Χαροκόπειο Πανεπιστήμιο**

**Νικολαΐδη Μαρία**

**Καθηγήτρια, Τμήμα Πληροφορικής και Τηλεματικής,  
Χαροκόπειο Πανεπιστήμιο**

**Μπαρδάκη Κλεοπάτρα**

**Εξωτερική Διδάσκουσα, Χαροκόπειο Πανεπιστήμιο**

Η Αθανασία Παπαστεργίου

δηλώνω υπεύθυνα ότι:

- 1) Είμαι ο κάτοχος των πνευματικών δικαιωμάτων της πρωτότυπης αυτής εργασίας και από όσο γνωρίζω η εργασία μου δε συκοφαντεί πρόσωπα, ούτε προσβάλει τα πνευματικά δικαιώματα τρίτων.
- 2) Αποδέχομαι ότι η ΒΚΠ μπορεί, χωρίς να αλλάξει το περιεχόμενο της εργασίας μου, να τη διαθέσει σε ηλεκτρονική μορφή μέσα από τη ψηφιακή Βιβλιοθήκη της, να την αντιγράψει σε οποιοδήποτε μέσο ή/και σε οποιοδήποτε μορφότυπο καθώς και να κρατά περισσότερα από ένα αντίγραφα για λόγους συντήρησης και ασφάλειας.

## ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

|                                                                   |           |
|-------------------------------------------------------------------|-----------|
| Περίληψη στα Ελληνικά .....                                       | 7         |
| Abstract.....                                                     | 8         |
| Κατάλογος Εικόνων .....                                           | 9         |
| Κατάλογος Πινάκων .....                                           | 12        |
| Συντομογραφίες.....                                               | 13        |
| <b>ΚΕΦΑΛΑΙΟ 1. ΕΙΣΑΓΩΓΗ .....</b>                                 | <b>14</b> |
| 1.1 Περιγραφή του προβλήματος .....                               | 14        |
| 1.2 Αντικείμενο της εργασίας.....                                 | 15        |
| 1.3 Δομή της διπλωματικής.....                                    | 16        |
| <b>ΚΕΦΑΛΑΙΟ 2. ΜΕΛΕΤΕΣ ΚΑΙ ΕΦΑΡΜΟΓΕΣ.....</b>                     | <b>18</b> |
| 2.1 Υπάρχουσες Μελέτες .....                                      | 18        |
| 2.2 Εφαρμογές.....                                                | 19        |
| 2.2.1 Σύστημα Κοινωνικής και Φορολογικής Πίστωσης της Κίνας ..... | 19        |
| 2.2.2 Σύστημα αξιολόγησης φορολογουμένων Λετονίας.....            | 23        |
| 2.2.3 Σύστημα κατηγοριοποίησης φορολογουμένων Αυστραλίας .....    | 26        |
| 2.2.4 Σύστημα αξιολόγησης φορολογικής συμμόρφωσης Ινδίας.....     | 27        |
| 2.2.5 Σύστημα βαθμολόγησης Φορολογουμένων Ουγγαρίας .....         | 29        |
| 2.2.6 Σύστημα αξιολόγησης φορολογουμένων στη Καμπότζη.....        | 29        |
| <b>ΚΕΦΑΛΑΙΟ 3. ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ.....</b>                        | <b>32</b> |
| 3.1 Ορισμοί.....                                                  | 32        |
| 3.2 Στάδια της Διαδικασίας Ανακάλυψης Γνώσης.....                 | 34        |
| 3.3 Μεθοδολογίες Εξόρυξης δεδομένων .....                         | 36        |
| 3.4 Βασικά στάδια της εξόρυξης δεδομένων .....                    | 40        |
| 3.5 Κατηγοριοποίηση .....                                         | 41        |
| 3.6 Μέθοδοι κατηγοριοποίησης.....                                 | 47        |
| 3.7 Δέντρα Αποφάσεων.....                                         | 48        |
| 3.7.1 Δημιουργία κανόνων από Δέντρα Αποφάσεων.....                | 49        |
| 3.7.2 Πλεονεκτήματα και Μειονεκτήματα των Δέντρων Αποφάσεων ..... | 49        |
| 3.7.3 Αλγόριθμοι κατασκευής Δέντρων Αποφάσεων .....               | 50        |

|                                                                                 |           |
|---------------------------------------------------------------------------------|-----------|
| <b>ΚΕΦΑΛΑΙΟ 4. ΠΕΡΙΦΡΑΦΗ ΠΡΟΤΕΙΝΟΜΕΝΟΥ ΣΥΣΤΗΜΑΤΟΣ .....</b>                     | <b>56</b> |
| 4.1 Στόχος του Συστήματος .....                                                 | 56        |
| 4.2 Σε ποιους εφαρμόζεται .....                                                 | 58        |
| 4.3 Κατηγορίες Δεικτών.....                                                     | 58        |
| 4.4 Κλίμακα Βαθμολόγησης .....                                                  | 60        |
| 4.5 Δεδομένα Αξιολόγησης .....                                                  | 62        |
| <b>ΚΕΦΑΛΑΙΟ 5. ΔΕΙΚΤΕΣ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ.....</b>                                  | <b>64</b> |
| 5.1 Αξιολόγηση με βάση τα στοιχεία του Μητρώου .....                            | 64        |
| 5.1.1 Δείκτης 1: Καταχωρημένων Στοιχείων Μητρώου .....                          | 64        |
| 5.1.2 Δείκτης 2: Ενημέρωσης Μητρώου .....                                       | 66        |
| 5.2 Αξιολόγηση με βάση την υποβολή των Δηλώσεων .....                           | 68        |
| 5.2.1 Δείκτης 3: Μη υποβολής δηλώσεων.....                                      | 68        |
| 5.2.2 Δείκτης 4: Ανακριβούς υποβολής δηλώσεων .....                             | 72        |
| 5.2.3 Δείκτης 5: Υποβολής εκπροθέσμων δηλώσεων .....                            | 73        |
| 5.2.4 Δείκτης 6: Υποβολής τροποποιητικών δηλώσεων .....                         | 75        |
| 5.3 Αξιολόγηση με βάση τις Βεβαιώσεις, Πληρωμές και Ληξιπρόθεσμες Οφειλές ..... | 77        |
| 5.3.1 Δείκτης 7: Συνολικού ληξιπρόθεσμου χρέους .....                           | 78        |
| 5.3.2 Δείκτης 8: Μεταβολής Συνολικού ληξιπρόθεσμου χρέους.....                  | 79        |
| 5.3.3 Δείκτης 9: Ρυθμισμένων Οφειλών .....                                      | 80        |
| 5.3.4 Δείκτης 10: Συνέπειας Διατήρησης Ρύθμισης .....                           | 81        |
| 5.3.5 Δείκτης 11: Δυνατότητας Εξόφλησης του Χρέους.....                         | 82        |
| 5.4 Αξιολόγηση με βάση τις Παραβάσεις.....                                      | 83        |
| 5.4.1 Δείκτης 12 : Παραβάσεων.....                                              | 83        |
| <b>ΚΕΦΑΛΑΙΟ 6. ΔΕΔΟΜΕΝΑ .....</b>                                               | <b>85</b> |
| 6.1 Συλλογή δεδομένων .....                                                     | 85        |
| 6.1.1 TAXIS - Υποσύστημα Μητρώου.....                                           | 85        |
| 6.1.2 TAXIS - Υποσύστημα Φ.Π.Α και Υποσύστημα Φ.Ε. ....                         | 86        |
| 6.1.3 TAXIS - Υποσύστημα Εσόδων.....                                            | 87        |
| 6.1.4 ELENXIS και TAXIS.....                                                    | 88        |
| 6.2 Γνωρίσματα που θα χρησιμοποιηθούν .....                                     | 89        |
| <b>ΚΕΦΑΛΑΙΟ 7. ΕΞΟΥΣΗ ΔΕΔΟΜΕΝΩΝ .....</b>                                       | <b>94</b> |
| 7.1 Εργαλεία που θα χρησιμοποιηθούν.....                                        | 94        |
| 7.2 Επισκόπηση Δεδομένων.....                                                   | 95        |

|                                                        |            |
|--------------------------------------------------------|------------|
| 7.3 Δεδομένα εφαρμογής για εκπαίδευση και δοκιμή ..... | 105        |
| 7.4 Εξόρυξη Δεδομένων .....                            | 107        |
| 7.4.1 Εφαρμογή Αλγορίθμου C5.1.....                    | 110        |
| 7.4.2 Εφαρμογή Αλγορίθμου C&R Tree.....                | 116        |
| 7.4.2 Εφαρμογή Αλγορίθμου CHAID.....                   | 122        |
| <b>ΚΕΦΑΛΑΙΟ 8. ΣΥΜΠΕΡΑΣΜΑΤΑ.....</b>                   | <b>129</b> |
| 8.1 Συμπεράσματα .....                                 | 129        |
| 8.2 Βελτιώσεις .....                                   | 130        |
| <b>Βιβλιογραφία .....</b>                              | <b>132</b> |
| <b>Παραρτήματα .....</b>                               | <b>137</b> |
| Παράρτημα Α .....                                      | 137        |
| Παράρτημα Β.....                                       | 149        |
| Παράρτημα Γ .....                                      | 153        |

## Περίληψη στα Ελληνικά

Η φοροδιαφυγή και η φοροαποφυγή αποτελούν σημαντική τροχοπέδη για την εύρυθμη λειτουργία του κράτους. Πέρα από τη διασφάλιση της ομαλής είσπραξης των προβλεπόμενων δημοσιονομικών εσόδων, ο περιορισμός αυτών των φαινομένων αποτελεί προϋπόθεση για μια ευνομούμενη και δημοκρατική κοινωνία, για την ύπαρξη κράτους δικαίου και την καθιέρωση αντιλήψεων δικαιοσύνης.

Οι Ελληνικές φορολογικές αρχές, στην προσπάθειά τους να περιορίσουν τα φαινόμενα αυτά, δίνουν αυξανόμενη έμφαση στη διενέργεια φορολογικών ελέγχων, ενώ η σχέση μεταξύ πλήθους ελέγχων και ποσοστού είσπραξης των φορολογικών οφειλών που προκύπτουν από αυτούς, είναι αρνητική. Η αποκλειστική και εντεινόμενη έμφαση σε κλασικές μεθόδους φορολογικής επιβολής εκτός από το ότι αποδεικνύεται αναποτελεσματική, μπορεί να επιφέρει πολλαπλά αρνητικά αποτελέσματα σε δημοσιονομικό και κοινωνικό επίπεδο, αμβλύνοντας περαιτέρω την φορολογική συνείδηση των πολιτών. Η καλλιέργεια φορολογικής συνείδησης, μέσω της ένταξης σύγχρονων δράσεων στον επιχειρησιακό σχεδιασμό της φορολογικής διοίκησης κρίνεται απαραίτητη για τη διευκόλυνση της επίτευξης των δημοσιονομικών στόχων και την καλλιέργεια κλίματος εμπιστοσύνης μεταξύ φορολογικής διοίκησης και πολιτών.

Η σύγχρονη πραγματικότητα απαιτεί την αναγνώριση της σημασίας της εκούσιας συμμόρφωσης, ως απόρροιας της καθιέρωσης και διατήρησης φορολογικής συνείδησης στο κοινωνικό σύνολο.

Σκοπός της εργασίας αυτής είναι η ολοκληρωμένη πρόταση αλγορίθμου αυτοματοποιημένης αξιολόγησης και βαθμολόγησης των φορολογουμένων στην Ελλάδα βάσει της συμπεριφοράς τους με στόχο την επίτευξη εκούσιας φορολογικής συμμόρφωσης. Περιγράφονται κριτήρια αξιολόγησης, εξετάζονται αλγόριθμοι κατηγοριοποίησης και μετά από σύγκριση επιλέγεται ο καλύτερος.

**Λέξεις κλειδιά:** Βαθμολόγηση φορολογουμένων, Ταξινόμηση, Αλγόριθμοι, Εξόρυξη Δεδομένων, Φορολογική Συμμόρφωση.

## Abstract

The phenomena of tax evasion and avoidance are a major problem for the proper functioning of the state. In addition to ensuring that the planned budget revenue is properly recovered, finding a solution to these issues is a prerequisite for well-founded and democratic societies, the existence of the rule of law and the establishment of the concepts of justice.

The normalization of tax evasion and tax avoidance is imperative. Greek tax authorities are increasingly focusing on tax audits, and the relationship between the number of audits and the rate of recovery of tax debts resulting from these controls is negative. Exclusive and growing emphasis on classic tax enforcement methods can have multiple negative effects at the financial and social level, further limiting the citizens' tax conscience. The fostering of tax consciousness through the integration of modern actions into the operational planning of the tax administration is deemed necessary to facilitate the achievement of fiscal targets and to foster a climate of trust between tax administration and citizens.

Modern reality requires recognition of the importance of voluntary compliance as a consequence of establishing and maintaining fiscal consciousness in society.

The aim of this work is to propose an algorithm for automated assessment and rating of taxpayers in Greece based on their behavior with the aim of voluntary tax compliance. The evaluation criteria – features will be described, classification algorithms will be examined and the best one is selected after comparison.

**Keywords:** Taxpayer Rating System, Classification, Algorithms, Data mining, Tax Compliance.



## Κατάλογος Εικόνων

|                                                                                                                  |       |
|------------------------------------------------------------------------------------------------------------------|-------|
| Εικ.3.1 Η Εξόρυξη Δεδομένων ως αποτέλεσμα συμβολής άλλων κλάδων .....                                            | σ.34  |
| Εικ.3.2 Επισκόπηση των βημάτων που αποτελούν μέρος μιας διαδικασίας εξόρυξης δεδομένων (Fayyad et al, 1996)..... | σ.35  |
| Εικ.3.3: Μια ταξινόμηση των κυριότερων εργασιών ΕΔ.....                                                          | σ.38  |
| Εικ.3.4: Decision Trees, Nearest Networks, Neural Networks, Rule Induction .....                                 | σ.39  |
| Εικ.3.5: Ένα απλό γραμμικό όριο κατηγοριοποίησης με τρεις κατηγορίες.....                                        | σ.43  |
| Εικ.3.6: Λειτουργία μοντέλου ταξινόμησης.....                                                                    | σ.44  |
| Εικ.3.7: Γενικό σχήμα ταξινόμησης-κατηγοριοποίησης.....                                                          | σ.45  |
| Εικ.3.8: Confusion Matrix, σύγκριση της επίδοσης της κατηγοριοποίησης με την ανάκτηση της πληροφορίας.....       | σ.46  |
| Εικ.4.1 Χρωματική κλίμακα συμμόρφωσης.....                                                                       | σ.59  |
| Εικόνα 7.1 Ροή για προεπισκόπηση στοιχείων .....                                                                 | σ.95  |
| Εικόνα 7.2 Προεπισκόπηση Δεδομένων (Audit) .....                                                                 | σ.96  |
| Εικόνα 7.3 Προεπισκόπηση Δεδομένων (Quality) .....                                                               | σ.97  |
| Εικόνα 7.4 Ιστογράμματα Δεδομένων .....                                                                          | σ.98  |
| Εικόνα 7.5 Ιστόγραμμα Πλήθους Τροποποιητικών Δηλώσεων .....                                                      | σ.98  |
| Εικόνα 7.6 Ιστόγραμμα Πλήθους Εκπροθέσμων Δηλώσεων .....                                                         | σ.99  |
| Εικόνα 7.7 Ιστόγραμμα Ληξιπροθέσμων χρεών χωρίς ρύθμιση .....                                                    | σ.99  |
| Εικόνα 7.8 Παραβάσεις ανά κλάση.....                                                                             | σ.100 |
| Εικόνα 7.9 Διάγραμμα διασποράς Πλήθους τροποποιητικών και εκπροθέσμων δηλώσεων .....                             | σ.100 |

|             |                                                                                                             |       |
|-------------|-------------------------------------------------------------------------------------------------------------|-------|
| Εικόνα 7.10 | Νέο Μοντέλο .....                                                                                           | σ.101 |
| Εικόνα 7.11 | Γραφική απεικόνιση τιμών Πλήθους Διακοπών Εργασιών.....                                                     | σ.102 |
| Εικόνα 7.12 | Γραφική απεικόνιση εξάρτησης της κλάσης από το πλήθος διακοπών εργασιών.....                                | σ.102 |
| Εικόνα 7.13 | Ανισορροπία κλάσεων .....                                                                                   | σ.104 |
| Εικόνα 7.14 | Διαχωρισμός Δεδομένων .....                                                                                 | σ.106 |
| Εικόνα 7.15 | Κατανομή των δεδομένων εκπαίδευσης και των δεδομένων επαλήθευσης.....                                       | σ.107 |
| Εικόνα 7.16 | Ροή για υπόδειξη των τριών καλύτερων αλγορίθμων .....                                                       | σ.108 |
| Εικόνα 7.17 | Καλύτεροι Αλγόριθμοι .....                                                                                  | σ.109 |
| Εικόνα 7.18 | Τελικό Μοντέλο .....                                                                                        | σ.110 |
| Εικόνα 7.19 | Παράμετροι αλγορίθμου C5.1. ....                                                                            | σ.111 |
| Εικόνα 7.20 | Αποτελέσματα πρόβλεψης αλγορίθμου C5.1 .....                                                                | σ.111 |
| Εικόνα 7.21 | Δέντρο απόφασης αλγορίθμου C5.1 συνοπτικά και διάγραμμα σπουδαιότητας χαρακτηριστικών στη πρόβλεψη.....     | σ.112 |
| Εικόνα 7.22 | Αποτελέσματα αλγορίθμου C5.1. ....                                                                          | σ.113 |
| Εικόνα 7.23 | Καμπύλη κέρδους αλγορίθμου C5.1. ....                                                                       | σ.116 |
| Εικόνα 7.24 | Παράμετροι αλγορίθμου C&R Tree.....                                                                         | σ.117 |
| Εικόνα 7.25 | Αποτελέσματα πρόβλεψης αλγορίθμου C&R Tree.....                                                             | σ.117 |
| Εικόνα 7.26 | Δέντρο απόφασης αλγορίθμου C&R Tree συνοπτικά και διάγραμμα σπουδαιότητας χαρακτηριστικών στη πρόβλεψη..... | σ.118 |
| Εικόνα 7.27 | Αποτελέσματα αλγορίθμου C&R Tree .....                                                                      | σ.120 |
| Εικόνα 7.28 | Καμπύλη κέρδους αλγορίθμου C&R Tree.....                                                                    | σ.122 |

|                                                                                                                         |       |
|-------------------------------------------------------------------------------------------------------------------------|-------|
| Εικόνα 7.29 Παράμετροι αλγορίθμου CHAID.....                                                                            | σ.123 |
| Εικόνα 7.30 Αποτελέσματα πρόβλεψης αλγορίθμου CHAID .....                                                               | σ.123 |
| Εικόνα 7.32 Δέντρο απόφασης αλγορίθμου CHAID συνοπτικά και διάγραμμα<br>σπουδαιότητας χαρακτηριστικών στη πρόβλεψη..... | σ.124 |
| Εικόνα 7.33 Αποτελέσματα αλγορίθμου CHAID .....                                                                         | σ.126 |
| Εικόνα 7.34 Καμπύλη κέρδους αλγορίθμου CHAID .....                                                                      | σ.128 |

## Κατάλογος Πινάκων

|                                                           |       |
|-----------------------------------------------------------|-------|
| Πίνακας 1. Δείκτες του συστήματος διαβάθμισης .....       | σ.59  |
| Πίνακας 2. Βαθμίδες αξιολόγησης .....                     | σ.61  |
| Πίνακας 3. Περιγραφή πεδίων .....                         | σ.93  |
| Πίνακας 4. Βαρύτητα Γνωρισμάτων αλγορίθμου C5.1 .....     | σ.113 |
| Πίνακας 5. Confusion matrix αλγορίθμου C5.1 .....         | σ.114 |
| Πίνακας 6. Ανάκληση και Ακρίβεια αλγορίθμου C5.1 .....    | σ.115 |
| Πίνακας 7. Βαρύτητα Γνωρισμάτων αλγορίθμου C&R Tree ..... | σ.119 |
| Πίνακας 8. Confusion matrix αλγορίθμου C&R Tree .....     | σ.120 |
| Πίνακας 9. Ανάκληση και Ακρίβεια αλγορίθμου C&R Tree..... | σ.121 |
| Πίνακας 10. Βαρύτητα Γνωρισμάτων αλγορίθμου CHAID .....   | σ.125 |
| Πίνακας 11. Confusion matrix αλγορίθμου CHAID .....       | σ.126 |
| Πίνακας 12. Ανάκληση και Ακρίβεια αλγορίθμου CHAID .....  | σ.127 |

## Συντομογραφίες

|          |                                                                      |
|----------|----------------------------------------------------------------------|
| Α.Π.Δ.   | Αναλυτική Περιοδική Δήλωση                                           |
| ΕΔ       | Εξόρυξη Δεδομένων                                                    |
| Ι.Κ.Α.   | Ίδρυμα Κοινωνικών Ασφαλίσεων                                         |
| Κ.Β.Σ.   | Κώδικας Βιβλίων και Στοιχείων                                        |
| ΚΠ       | Κέρδος Πληροφορίας                                                   |
| Φ.Ε.     | Φορολογία Εισοδήματος                                                |
| Φ.Μ.Υ.   | Φόρος Μισθωτών Υπηρεσιών                                             |
| Φ.Π.Α.   | Φόρος Προστιθέμενης Αξίας                                            |
|          |                                                                      |
| AI       | Artificial Intelligence                                              |
| ATO      | Australian Taxation Office                                           |
| CART     | Classification And Regression Tree                                   |
| C&R Tree | Classification And Regression Tree                                   |
| CHAID    | Chi-Squared Automatic Interaction Detection                          |
| DM       | Data Mining                                                          |
| GDT      | General Department of Taxation                                       |
| GST      | Good and Service Tax                                                 |
| KDD      | Knowledge Discovery in Database                                      |
| NAV      | National Tax and Customs Administration (Nemzeti Adó- és Vámhivatal) |
| SAT      | State Taxation Administration                                        |
| SRS      | State Revenue Service                                                |
| SVM      | Support Vector Machine                                               |
| VIES     | VAT Information Exchange System (Ανακεφαλαιωτικοί πίνακες)           |

### 1.1 Περιγραφή του προβλήματος

### 1.2 Αντικείμενο της εργασίας

### 1.3 Δομή της εργασίας

---

#### 1.1 Περιγραφή του προβλήματος

Οι χώρες, μεταξύ αυτών και η Ελλάδα μαστίζονται από φοροδιαφυγή και φοροαποφυγή. Η απατηλή συμπεριφορά των φορολογουμένων, εκτός από προβλήματα στην άσκηση της οικονομικής πολιτικής ενός κράτους, προκαλεί και στρεβλώσεις στη λειτουργία της αγοράς καθώς οι φοροδιαφεύγοντες αποκτούν ανταγωνιστικό πλεονέκτημα έναντι των έντιμων φορολογουμένων. Η βελτίωση και η διατήρηση της φορολογικής συμμόρφωσης αποτελεί αυτονόητη πρόκληση για τη φορολογική διοίκηση σε όλο τον κόσμο.

Οι φορολογικές υπηρεσίες χρησιμοποιούν κυρίως ελέγχους για να διασφαλίσουν τη συμμόρφωση με τους φορολογικούς νόμους και να διατηρήσουν τις σχετικές ροές εσόδων. Οι έλεγχοι οδηγούν έμμεσα στην εθελοντική συμμόρφωση και παράγουν άμεσα πρόσθετες εισπράξεις φόρων, οι οποίες βοηθούν τις φορολογικές υπηρεσίες να μειώσουν το «φορολογικό χάσμα» μεταξύ του οφειλόμενου φόρου και του ποσού που έχει εισπραχθεί. Επιπρόσθετα εξασφαλίζουν ίσους όρους ανταγωνισμού για έναν έντιμο φορολογούμενο.

Ως εκ τούτου, οι έλεγχοι είναι κρίσιμης σημασίας για την επιβολή της φορολογικής νομοθεσίας και την παροχή βοήθειας στις φορολογικές υπηρεσίες προκειμένου να επιτύχουν τους στόχους των εσόδων, εξασφαλίζοντας έτσι τη δημοσιονομική υγεία της χώρας και τον υγιή ανταγωνισμό.

Από την άλλη πλευρά, ο έλεγχος δεν είναι μια ευπρόσδεκτη διαδικασία τόσο για τους φορολογούμενους όσο και για την οικονομία. Η διεξαγωγή των ελέγχων συνεπάγεται δαπάνες του φορολογικού τμήματος για τον φορολογούμενο. Είναι

αναγκαίο όμως οι οργανισμοί φορολογικής διοίκησης, να χρησιμοποιούν τους περιορισμένους πόρους τους με πολύ συνετό τρόπο και να επιτύχουν τη μέγιστη δυνατή συμμόρφωση, με την ελάχιστη εισβολή και τα ελάχιστα κόστη.

Κύριο μειονέκτημα του ελέγχου είναι η τυχαία επιλογή, η ίση μεταχείριση προς τους έντιμους και ανέντιμους φορολογούμενους, καθώς η πιθανότητα επιλογής είναι ίδια και για τους δύο. Η αποτελεσματικότητα των φορολογικών ελέγχων αυξάνεται όταν αυτοί γίνονται στοχευμένα. Η αξιολόγηση των φορολογουμένων και η επιλογή προς έλεγχο των περισσότερο ύποπτων για παραβατική συμπεριφορά βελτιώνει την απόδοσή τους.

Η διαχείριση της φορολογικής συμμόρφωσης μιας δημοκρατικής πολιτείας, θα πρέπει να χαρακτηρίζεται από διάκριση και να αντιμετωπίζει τον κάθε φορολογούμενο αντίστοιχα προς την φορολογική του συμπεριφορά. Εκτός από την παρακολούθηση, τον έλεγχο και την τιμωρία, τακτικές, που απευθύνονται στους ασυνεπείς φορολογούμενους, μέσο δράσης της φορολογικής διοίκησης θα πρέπει να είναι και η επιβράβευση για τους συνεπείς και έντιμους φορολογούμενους ώστε αυτοί να ενθαρρύνονται στην συμμόρφωση και να τους δημιουργείται αίσθηση δικαίου και αναγνώρισης της έντιμης φορολογικής συμπεριφοράς τους. Απαιτείται λοιπόν, κατηγοριοποίηση των φορολογουμένων σε κλάσεις, με κριτήριο την φορολογική τους συμπεριφορά. Οι φορολογούμενοι γνωρίζοντας ότι, αν καταταγούν στην κλίμακα των πλήρως συμμορφούμενων θα απολαμβάνουν προνόμια που θα επιβραβεύουν τη συνεπή συμπεριφορά τους, θα έχουν ένα σημαντικό κίνητρο για την εθελοντική φορολογική τους συμμόρφωση. Επιβεβαιώνεται λοιπόν και πάλι η ανάγκη αξιολόγησης και βαθμολόγησης των φορολογουμένων με στόχο αυτή τη φορά τη διάκριση των συνεπών.

## 1.2 Αντικείμενο της εργασίας

Στις μέρες μας η απόκτηση πληροφορίας είναι ευκολότερη από ότι στο παρελθόν. Τεχνολογικά προηγμένες μέθοδοι δίνουν τη δυνατότητα καταγραφής δεδομένων με ολοένα αυξανόμενο ρυθμό. Οι φορολογικές υπηρεσίες έχουν πια τεράστιους

όγκους δεδομένων στην διάθεση τους. Για να επιτύχουν τη μέγιστη συμμόρφωση φορολογουμένων πρέπει να αξιοποιήσουν τις συσσωρευμένες πληροφορίες. Η εξόρυξη δεδομένων βοηθά τους φορολογικούς οργανισμούς να επιτύχουν στόχους συμμόρφωσης και να βελτιώσουν την αποτελεσματικότητα λειτουργίας χρησιμοποιώντας τους υπάρχοντες πόρους τους.

Η διπλωματική εργασία αποσκοπεί στην περιγραφή αλγορίθμου για την αυτοματοποιημένη αξιολόγηση και βαθμολόγηση φορολογουμένων με βάση τη συμπεριφορά τους κάνοντας χρήση τεχνικών εξόρυξης δεδομένων.

Αρχικά θα περιγραφούν τα κριτήρια αξιολόγησης. Από τα τεράστια φορολογικά δεδομένα που αναπτύσσονται συνεχώς θα δημιουργηθεί ένα υποσύνολο δεδομένων, που θα αποτελείται από μεταβλητές που επηρεάζουν το επίπεδο συμμόρφωσης των φορολογουμένων. Στη συνέχεια θα εξεταστεί η χρήση αλγορίθμων κατηγοριοποίησης για τον διαχωρισμό των φορολογουμένων σε κλάσεις. Συγκεκριμένα θα εξεταστούν οι αλγόριθμοι C5.1, CART και CHAID και τα αποτελέσματα ταξινόμησης του κάθε αλγορίθμου θα συγκριθούν με βάση τα κριτήρια Accuracy και Time.

Η κατηγοριοποίηση αυτή θα δώσει τη δυνατότητα εφαρμογής διαφορετικών εργαλείων σε κάθε κατηγορία φορολογουμένων, με σκοπό την επίτευξη μεγαλύτερων εσόδων και υψηλότερων ποσοστών συμμόρφωσης.

### 1.3 Δομή της εργασίας

Μετά την εισαγωγική πρώτη ενότητα, στο δεύτερο κεφάλαιο η διπλωματική αναφέρεται σε παρόμοιες μελέτες και πρακτικές εφαρμογές σε άλλες χώρες. Το τρίτο κεφάλαιο επιδιώκει τη διενέργεια θεωρητικής επισκόπησης της έννοιας της Εξόρυξης Δεδομένων, περιγράφεται το πώς εξελίσσεται σαν διαδικασία και σε ποιο μέρος της λειτουργίας της δημιουργίας γνώσης αντιστοιχεί. Επίσης παρουσιάζεται η εξέλιξη της Εξόρυξης Γνώσης με τη χρήση Δέντρων Απόφασης και βασικό στόχο την κατηγοριοποίηση όπως και τα βασικά χαρακτηριστικά των κυριότερων σχετικών



αλγορίθμων. Στο τέταρτο κεφάλαιο επιδιώκεται μία προσέγγιση των βασικών χαρακτηριστικών του προτεινόμενου συστήματος. Στο πέμπτο κεφάλαιο περιγράφονται οι δείκτες αξιολόγησης και βαθμολόγησης φορολογουμένων βάσει των οποίων θα γίνει η κατηγοριοποίηση. Στο έκτο κεφάλαιο γίνεται αναφορά στη συλλογή και στο μετασχηματισμό των δεδομένων. Επίσης εξετάζονται τα διαθέσιμα γνωρίσματα πάνω στα οποία θα εφαρμοστούν οι διαδικασίες εξόρυξης. Στο έβδομο κεφάλαιο εφαρμόζονται οι τεχνικές εξόρυξης και γίνεται επιλογή αλγορίθμων. Στο όγδοο κεφάλαιο, η διπλωματική προβαίνει στην ανάλυση των ευρημάτων και στην εξαγωγή συμπερασμάτων. Επίσης προτείνονται μελλοντικές βελτιώσεις. Η εργασία ολοκληρώνεται με την παράθεση των βιβλιογραφικών αναφορών και των παραρτημάτων.

## ΚΕΦΑΛΑΙΟ 2. ΜΕΛΕΤΕΣ ΚΑΙ ΕΦΑΡΜΟΓΕΣ

---

### 2.1 Υπάρχουσες Μελέτες

### 2.2 Εφαρμογές

---

#### 2.1 Υπάρχουσες Μελέτες

Η ανίχνευση της φοροδιαφυγής είναι μια ενδιαφέρουσα έρευνα επειδή ο φόρος έχει τη φύση του εξαναγκασμού και τα φορολογικά οφέλη για τους φορολογούμενους δεν μπορούν να ληφθούν άμεσα. Ως εκ τούτου, η τάση των φορολογουμένων είναι να μην αναφέρουν τα κέρδη τους σωστά.

Η φορολογική διοίκηση καθημερινά μαζεύει μεγάλο όγκο δεδομένων που την αφορούν. Το περιβάλλον που δημιουργείται είναι δύσκολο να ελεγχθεί με συμβατικά μέσα, αντίθετα διαδικασίες στατιστικής ανάλυσης και εξόρυξης δεδομένων απαιτούνται όλο και περισσότερο για την ομαλή οργάνωση και εύκολη κατανόηση των πληροφοριών. Αρκετές μελέτες έχουν γίνει μέχρι και σήμερα επισημαίνοντας διάφορα κριτήρια αξιολόγησης της συμπεριφοράς των φορολογουμένων και της ανίχνευσης πιθανής παραβατικής συμπεριφοράς.

Οι Gupta και Gill (2012) είχαν χαρακτηρίσει οργανισμούς ως δόλιους ή μη δόλιους χρησιμοποιώντας χρηματοοικονομικούς δείκτες που σχετίζονται με την αποδοτικότητα, τη ρευστότητα και την επιχειρησιακή αποδοτικότητα.

Οι DeBarr D. et al (2005) στην έρευνά τους προσπάθησαν να εντοπίσουν παραμέτρους που να καθορίζουν αν κάποιος έχει κύκλωμα φοροδιαφυγής χρησιμοποιώντας μεγάλες εικονικές ζημιές για να κρύψουν μεγάλα κέρδη. Με τη χρήση SVM αλγορίθμου δημιούργησαν μοντέλο και έγινε ταξινόμηση των περιπτώσεων.

Οι M.Kallio και B.Black (2011) δημιούργησαν μοντέλα για τον εντοπισμό περιπτώσεων που θα έπρεπε να δοθούν για έλεγχο χρησιμοποιώντας μεθοδολογία νευρωνικών δικτύων.

Ο Wilson (2009) ισχυρίστηκε ότι η φοροαποφυγή σχετίζεται θετικά με το μέγεθος της επιχείρησης, τις μεγάλες διαφορές μεταξύ λογιστικής και φορολογίας, όπως και με επιθετικές πρακτικές που χρησιμοποιούνται στη χρηματοοικονομική πληροφόρηση.

Οι Chow και Rice (1992) υποστηρίζουν ότι τα υψηλά επίπεδα χρέους αυξάνουν την πιθανότητα δόλιων οικονομικών αναφορών.

Οι P.Castellon et al (2012), κατέληξαν ότι τα δέντρα αποφάσεων μπορούν να χρησιμοποιηθούν για να εντοπιστούν τα γνωρίσματα που θα χαρακτηρίζουν τις κλάσεις των επιχειρήσεων με ψευδή τιμολόγια. Χρησιμοποίησαν αλγορίθμους νευρωνικών δικτύων για την κατασκευή των μοντέλων.

Οι Wu et al (2012), χρησιμοποίησαν κανόνες συσχέτισης για την επιλογή υποθέσεων για έλεγχο παίρνοντας στοιχεία από το μητρώο εταιρειών και τις δηλώσεις Φ.Π.Α.

Οι Spathis (2002) και Spathis et al (2002) δημιούργησαν μοντέλο ταξινόμησης επιχειρήσεων με 10 μεταβλητές στηριζόμενοι στην υπόθεση ότι οι επιχειρήσεις υποβάλουν στρεβλές οικονομικές καταστάσεις όταν τα περιουσιακά στοιχεία ή τα κέρδη έχουν διογκωθεί ή όταν τα έξοδα ή οι ζημιές έχουν υποτιμηθεί.

## 2.2 Εφαρμογές

Υπάρχουν αρκετά παραδείγματα αξιολόγησης, βαθμολόγησης και κατηγοριοποίησης φορολογουμένων παγκόσμια. Παρακάτω γίνεται αναφορά σε κάποια από αυτά.

### 2.2.1 Σύστημα Κοινωνικής και Φορολογικής Πίστωσης της Κίνας

Η κινεζική κυβέρνηση προωθεί το σχέδιο Κοινωνικής Πίστωσης (**China's Social Credit System and Taxes**), το οποίο αξιολογεί την αξιοπιστία κάθε πολίτη και

επιχείρησης χρησιμοποιώντας ένα σύστημα σημείων (point system). Έχει στόχο την τυποποίηση της αξιολόγησης της οικονομικής και κοινωνικής φήμης των πολιτών και των επιχειρήσεων αποσκοπώντας στη δημιουργία ενός ειλικρινούς και αξιόπιστου οικονομικού και κοινωνικού περιβάλλοντος.

Ήδη το 2014, η κινεζική παράκτια πόλη Rongcheng στην επαρχία Shandong ξεκίνησε μια πρώτη δοκιμασία του Συστήματος Κοινωνικής Πίστωσης και μέχρι το 2020 όλες οι κυβερνητικές βάσεις δεδομένων μαζί με ορισμένες ιδιωτικές βάσεις δεδομένων πρέπει να συγχωνευθούν. Το σύστημα θεωρείται μια μορφή μαζικής επιτήρησης που χρησιμοποιεί μεγάλη τεχνολογία ανάλυσης δεδομένων.

Τα δεδομένα προέρχονται τόσο από τους λογαριασμούς των ίδιων των πολιτών, όσο και από τις δραστηριότητες του δικτύου τους. Οι φορείς εκμετάλλευσης ιστοτόπων μπορούν να εντοπίζουν τα ίχνη των δεδομένων που ανταλλάσσουν χρήστες με ιστότοπους και να αντλούν ένα πλήρες κοινωνικό προφίλ, συμπεριλαμβανομένης της τοποθεσίας, των φίλων, των ιατρικών αρχείων, της ασφάλισης, των ιδιωτικών μηνυμάτων, της οικονομικής κατάστασης, των αγορών, των λογαριασμών και φόρων που πληρώνουν, των έξυπνων στατιστικών για το σπίτι, των προτιμώμενων εφημερίδων, του χρόνου που αφιερώνουν σε παιχνίδια διαδικτύου και της συμπεριφοράς σε ιστότοπους γνωριμιών.

Το 2014 επετράπη σε οκτώ ιδιωτικούς παρόχους να αναπτύξουν τα δικά τους ψηφιακά συστήματα βαθμολόγησης. Μεταξύ αυτών των οκτώ επιχειρήσεων είναι η Baidu, η Sesame Credit η οποία ανήκει στον όμιλο Alibaba, η Tencent, η Didi Chuxing καθώς και την Baihe.com.

Όλα τα συστήματα χρησιμοποιούν αλγόριθμους και τεχνητή νοημοσύνη (AI) για τον υπολογισμό σημείων ή βαθμολογιών σε μια αυτοματοποιημένη διαδικασία που δεν είναι διαφανής για τους χρήστες. Εκτός από την καταπολέμηση της ανεπιθύμητης κοινωνικής συμπεριφοράς και την εκπαίδευση των πολιτών ώστε να είναι πιο ειλικρινείς και αφοσιωμένοι, οι κοινωνικές δραστηριότητες, η φοροδιαφυγή ή τα εγκλήματα του λευκού κολάρου θα παρακολουθούνται περισσότερο.

Το Κρατικό Συμβούλιο της Κίνας έχει θέσει ένα νέο πλαίσιο για τους φορολογούμενους, το Σύστημα Φορολογικής Πίστωσης, το οποίο αποτελεί ουσιαστικό μέρος του σχεδιασμού της κατασκευής του εθνικού συστήματος κοινωνικών πιστώσεων στην Κίνα. Τον τελευταίο καιρό, η Κρατική Διοίκηση Φορολογίας (SAT) εξέδωσε Διοικητικά Μέτρα για την πίστωση των φορολογουμένων και Μέτρα Δημοσίευσης Πληροφοριών για σοβαρές φορολογικές υποθέσεις σε μια προσπάθεια δημιουργίας μηχανισμού παροχής κινήτρων για τους έντιμους φορολογούμενους και μηχανισμού κυρώσεων για ανέντιμους φορολογούμενους για να αναπτύξει την αναβαθμισμένη έκδοση του συστήματος Φορολογικής Πίστωσης.

Το φορολογικό σύστημα θα γίνει πιο επιστημονικό και θα διαδραματίσει σημαντικότερο ρόλο στην οικοδόμηση του συστήματος κοινωνικών πιστώσεων, βελτιώνοντας τον δείκτη αξιολόγησης, περιορίζοντας την περίοδο αξιολόγησης, αξιοποιώντας τα αποτελέσματα της αξιολόγησης, δημοσιοποιώντας έγκαιρα τις φορολογικές υποθέσεις και εφαρμόζοντας ταυτόχρονα την ανταμοιβή και την τιμωρία.

Με βάση ένα σύστημα με ποσοστιαίες μονάδες, ο φορολογούμενος βαθμολογείται από τον βαθμό Α έως τον βαθμό Δ, ο οποίος αντιστοιχεί στις ακόλουθες εκατοστιαίες μονάδες.

- ✓ Βαθμός Α: Πάνω από 90 βαθμούς
- ✓ Βαθμός Β: Μεταξύ 70-90 βαθμών
- ✓ Βαθμός Γ: Μεταξύ 40-70 βαθμών
- ✓ Βαθμός Δ: Κάτω από 40 βαθμούς

Τον Απρίλιο κάθε έτους, οι φορολογικές αρχές θα αξιολογήσουν τις επιδόσεις του προηγούμενου έτους. Εάν ο φορολογούμενος δεν είναι ικανοποιημένος με την αξιολόγησή του, μπορεί να υποβάλει αίτηση για νέα αξιολόγηση.

Το 2015 εισήχθησαν 8.5 εκατομμύρια φορολογούμενοι σε αυτό το νέο σύστημα διαβάθμισης. Από αυτούς, οι 653 χιλιάδες φορολογούμενοι δηλαδή το 7,6% του

συνόλου ανήκουν στο επίπεδο Α, ενώ οι 649 χιλιάδες φορολογούμενοι ή το 7,6% του συνόλου ανήκουν στο επίπεδο Δ.

Έχει δοθεί μεγάλη σημασία στην εφαρμογή και τη χρήση των αξιολογήσεων των φορολογουμένων. Στο εσωτερικό, οι φορολογούμενοι σε επίπεδο Α αντιμετωπίζονται κατά προτεραιότητα σε φορολογικές υποθέσεις όπως επιστροφές, επεξεργασία τιμολογίων και άλλα θέματα ημερήσιας φορολογίας. Αντίθετα οι φορολογούμενοι σε επίπεδο Δ υπόκεινται σε αυστηρότερο έλεγχο από τις φορολογικές διοικήσεις και λαμβάνουν αυστηρότερα πρόστιμα όταν υποπέσουν σε παραπτώματα. Επιπλέον και άλλες αρχές ενημερώνονται με τη βαθμολογία. Ως εκ τούτου, οι φορολογούμενοι επιπέδου Δ θα πρέπει να περιμένουν περιορισμούς σε έργα που απαιτούν κρατική έγκριση όπως επενδύσεις και χρηματοδότηση, πρόσβαση σε κρατικές ιδιοκτησίες, εισαγωγές και εξαγωγές. Δημιουργείται μια γενική συνείδηση ότι οποιαδήποτε συμπεριφορά έχει τις συνέπειές της και ότι οι φορολογούμενοι έχουν κίνητρο να συμμορφωθούν με τους φορολογικούς κανόνες.

Παρόλο που οι περιορισμοί για τους φορολογούμενους Δ ακούγονται σκληροί, η μαύρη λίστα είναι ακόμη χειρότερη. Τα ακριβή κριτήρια της μαύρης λίστας ολοκληρώθηκαν στο τέλος του 2013 και αντιμετωπίζουν σοβαρά φορολογικά αδικήματα.

Εάν υπάρχει σοβαρό φορολογικό αδίκημα, το όνομα του δράστη θα δημοσιευθεί αμέσως. Η δημοσίευση θα περιλαμβάνει το όνομα της εταιρείας, τον αριθμό καταχώρισης του φορολογούμενου, τη διεύθυνση, τη φύση των αδικημάτων και το ύψος των διοικητικών κυρώσεων. Επιπλέον, η μαύρη λίστα δημοσιεύει επίσης το όνομα, το φύλο και τον αριθμό ταυτότητας του νόμιμου εκπροσώπου και του οικονομικού προσωπικού της εταιρείας καθώς και τις πληροφορίες των οργανισμών και των υπαλλήλων τους που είναι υπεύθυνοι για φορολογικά θέματα.

Οι κατάλογοι δημοσιεύονται από τις τοπικές και δημοτικές φορολογικές αρχές μέσω διαδικτύου, εφημερίδων, ραδιοφώνου, τηλεόρασης, κοινωνικών μέσων ενημέρωσης και συνεντεύξεων Τύπου. Κάθε εταιρεία θα παραμείνει στη μαύρη λίστα για τουλάχιστον δύο χρόνια μετά τη δημοσίευσή της.

Σημαντική προϋπόθεση για την εφαρμογή του συστήματος είναι ο μηχανισμός συνεργασίας που έχει δημιουργηθεί μεταξύ των διαφόρων φορολογικών αρχών και άλλων αρχών όπως οι αρχές από τη βιομηχανία, το εμπόριο ή τα τελωνεία. Στόχος είναι να δημιουργηθεί ένα πιο αποτελεσματικό σύστημα ελέγχου μέσω διαφοροποιημένων ευθυνών και ανταλλαγής πληροφοριών. Για παράδειγμα, οι εταιρείες, στις οποίες απαγορεύεται η είσοδος ή έξοδος νομίμων εκπροσώπων λόγω φορολογικών αδικημάτων, αναφέρονται στις αρμόδιες αρχές.

Το νέο Σύστημα Φορολογικής Πίστωσης όχι μόνο θα ενισχύει τους ελέγχους αλλά και θα αυξάνει τις κυρώσεις. Ένας βαθμός Γ, Δ ή ακόμα και μαύρη λίστα, μπορεί εύκολα να προκαλέσει μακροπρόθεσμη βλάβη στην επιχείρηση, καθώς όλες οι πληροφορίες και οι κυρώσεις δημοσιοποιούνται.

### 2.2.2 Σύστημα αξιολόγησης φορολογουμένων Λετονίας

Η Υπηρεσία Κρατικών Εσόδων (SRS - State Revenue Service) της Λετονίας έχει θέσει σε εφαρμογή Σύστημα Αξιολόγησης Φορολογουμένων (**Taxpayer Rating System**) με τρεις κύριους στόχους:

- Να ενημερώνει τον φορολογούμενο για τη συμμόρφωση των δεικτών που χαρακτηρίζουν τις οικονομικές του δραστηριότητες με τις γενικές οικονομικές τάσεις.
- Να παρακινήσει τον φορολογούμενο να τηρήσει και να βελτιώσει την πειθαρχία του στον τομέα της φορολογίας.
- Να παράσχει στον φορολογούμενο ένα εργαλείο το οποίο ο φορολογούμενος μπορεί να χρησιμοποιήσει για να ενημερώσει τους εταίρους για τη φήμη του και την πειθαρχία για την πληρωμή του φόρου καθώς και τη συμμόρφωση των δεικτών που χαρακτηρίζουν τις οικονομικές του δραστηριότητες με τις γενικές οικονομικές τάσεις.

Η βαθμολογία είναι διαθέσιμη στις ακόλουθες κατηγορίες φορολογούμενων που έχουν εγγραφεί στο Μητρώο Φορολογικού Φορέα για τουλάχιστον έξι μήνες:

- ✓ Εταιρείες Περιορισμένης Ευθύνης

- ✓ Μετοχικές εταιρείες
- ✓ Γεωργικές και αλιευτικές επιχειρήσεις, καθώς και ατομικές επιχειρήσεις που υποβάλλουν ετήσιες δηλώσεις

Το σύστημα βαθμολόγησης των φορολογουμένων περιλαμβάνει 5 διαστάσεις ανάλυσης που περιγράφονται από 12 δείκτες. Για να υπάρξει μια αντικειμενική αξιολόγηση για διάφορες πτυχές των δραστηριοτήτων των φορολογουμένων, ορισμένες διαστάσεις περιλαμβάνουν περισσότερους από έναν δείκτες. Οι 5 διαστάσεις και οι 12 δείκτες αφορούν:

- Την αξιολόγηση των δεδομένων καταχώρισης
- Την υποχρέωση υποβολής φορολογικών δηλώσεων
- Το χρέος οφειλών που διαχειρίζεται η υπηρεσία κρατικών εσόδων.

Σε αυτή τη διάσταση αντιστοιχούν 3 δείκτες:

- Το συνολικό χρέος οφειλών που διαχειρίζεται το SRS
- Η αναλογία του συνολικού χρέους οφειλών που διαχειρίζεται το SRS στις δόσεις πληρωμών που διαχειρίζεται το SRS
- Οι αλλαγές στο συνολικό χρέος οφειλών που διαχειρίζεται το SRS

- Τους δείκτες μισθών.

Σε αυτή τη διάσταση αντιστοιχούν 3 δείκτες:

- Η συμμόρφωση του μέσου μηνιαίου μισθού με τον μέσο μηνιαίο μισθό στη χώρα
- Η συμμόρφωση του μέσου μηνιαίου μισθού με τον μέσο μηνιαίο μισθό στον τομέα
- Οι αλλαγές στον μέσο μηνιαίο μισθό

- Τους δείκτες οικονομικών δραστηριοτήτων.

Σε αυτή τη διάσταση αντιστοιχούν 4 δείκτες:

- Η αναλογία κερδοφορίας
- Ο συνολικός δείκτης ρευστότητας
- Η αναλογία απόλυτης ρευστότητας
- Η αναλογία χρηματοοικονομικής ανεξαρτησίας



Χρησιμοποιείται μια κλίμακα πέντε σημείων για την ερμηνεία των δεικτών του συστήματος αξιολόγησης των φορολογουμένων. Συγκεκριμένα:

- Δύο θετικές βαθμίδες - "4" και "5" σημεία.
- Ένα ουδέτερο βαθμό - "3" σημεία, το οποίο δηλώνει προβλήματα περιστασιακής φύσης.
- Δυο αρνητικές βαθμίδες - "2" και "1".

Το σύστημα βαθμολόγησης των φορολογουμένων και ο αλγόριθμος καθορισμού σκορ σχεδιάζονται με βάση τις ακόλουθες θεμελιώδεις αρχές:

- Μεσο-μακρές περίοδοι ανάλυσης - για την πλειονότητα των δεικτών που χρησιμοποιούνται στην αξιολόγηση, η περίοδος ανάλυσης περιλαμβάνει 12 μήνες. Έχει σχεδιαστεί ώστε να περιλαμβάνει τον παράγοντα της εποχικότητας της επιχείρησης στην ανάλυση.
- Σύγκριση παρόμοιων δεικτών - τα στοιχεία του προς εξέταση φορολογούμενου συγκρίνονται με τα στοιχεία παρόμοιων φορολογουμένων. Τα μεμονωμένα δεδομένα ενός εμπόρου συγκρίνονται με τους μέσους δείκτες στη χώρα ή τον τομέα.
- Την πιθανότητα να αρχίσει να βελτιώνεται η βαθμολογία ήδη από τον επόμενο μήνα - τα δεδομένα βαθμολόγησης ενημερώνονται κάθε μήνα, μεταβάλλοντας την περίοδο ανάλυσης για ένα μήνα προς τα εμπρός.
- Χαμηλότερες απαιτήσεις ανάπτυξης για τις μεγάλες επιχειρήσεις - τα κριτήρια των δεικτών ανάπτυξης για τις μεγάλες επιχειρήσεις είναι πιο μετριοπαθή. Εάν ένας έμπορος πληρώσει μισθούς που υπερβαίνουν σημαντικά τον μέσο μισθό που καθορίζεται στη χώρα, είναι αντικειμενικά δυσκολότερο για τον έμπορο να εξασφαλίσει μια γρήγορη αύξηση των μισθών.
- Μικρές "προκαταβολές" για επιχειρηματίες - μια νέα επιχείρηση ενδέχεται να μην έχει δεδομένα σχετικά με την κατάσταση των δεικτών πριν από 12

μήνες. Ως εκ τούτου, κατά την αξιολόγηση της δυναμικής του μισθού, η δημιουργία θέσεων εργασίας θεωρείται θετικό φαινόμενο, το οποίο πρέπει να αξιολογηθεί με 4 βαθμούς.

Ορισμένοι δείκτες που χρησιμοποιούνται στο σύστημα διαβάθμισης περιλαμβάνουν εμπορικό μυστικό γι αυτό η ταξινόμηση των φορολογουμένων είναι διαθέσιμη μόνο στον ίδιο τον φορολογούμενο στο Σύστημα Ηλεκτρονικής Δήλωσης. Η γνωστοποίηση της αξιολόγησης είναι ευθύνη του ίδιου του φορολογούμενου.

### 2.2.3 Σύστημα κατηγοριοποίησης φορολογουμένων Αυστραλίας

Η Αυστραλιανή Κυβέρνηση (Australian Government - Australian Taxation Office) έχει κατηγοριοποιήσει μεγάλες δημόσιες και πολυεθνικές επιχειρήσεις που έχουν σημαντική οικονομική δραστηριότητα. Οι πρώτοι 100 πελάτες προσδιορίζονται αρχικά με βάση το μέγεθος των δραστηριοτήτων τους στην Αυστραλία, τον φόρο εισοδήματος, το GST (Good and Service Tax) ή τον ειδικό φόρο κατανάλωσης και την επιρροή που μπορεί να έχει ο πελάτης στο τμήμα της αγοράς. Οι 100 καλύτεροι πελάτες λαμβάνουν μια ετήσια επιστολή από τον Επίτροπο που τους ενημερώνει για την κατηγοριοποίηση κινδύνου τους, τι σημαίνει αυτό για κάθε ισχύοντα φόρο, και πώς σκοπεύει να ασχοληθεί με αυτούς η Αυστραλιανή Κυβέρνηση τους επόμενους 12 μήνες.

Χρησιμοποιούν τρεις κατηγορίες κινδύνου:

- **Βασικός φορολογούμενος:** Ένας βασικός φορολογούμενος έχει γενικά χαμηλότερο επίπεδο κινδύνου σε σχέση με άλλους φορολογούμενους. Αυτό δεν σημαίνει ότι δεν υπάρχει κανένας κίνδυνος και ότι δεν θα υπάρχουν διαφωνίες ή διαφορές απόψεων.
- **Βασικός φορολογούμενος με σημαντικές ανησυχίες:** Ο βασικός φορολογούμενος με σημαντικές ανησυχίες ενδέχεται να έχει πολλαπλούς προσδιορισμένους κινδύνους ή / και οικονομικά αποτελέσματα που δεν

φαίνεται να αντικατοπτρίζονται στα φορολογικά του αποτελέσματα. Ωστόσο, η σχέση με τον πελάτη είναι θετική. Υπάρχει στενή συνεργασία με τους πελάτες αυτής της κατηγορίας για να βελτιωθεί η μελλοντική τους κατηγοριοποίηση κινδύνου.

- **Υψηλότερος κίνδυνος:** Ο φορολογούμενος με υψηλότερο κίνδυνο μπορεί να έχει πολλούς πολύπλοκους και διαρθρωτικούς κινδύνους σε διάφορα μέρη του φορολογικού νόμου, δεν τηρεί προθεσμίες για αιτήματα πληροφοριών και μερικές φορές καθυστερεί να εκπληρώσει τις φορολογικές του υποχρεώσεις.

Η Αυστραλιανή Κυβέρνηση εξετάζει συνεχώς τους φορολογούμενους υψηλότερου κινδύνου είτε με ολοκληρωμένο έλεγχο είτε με άλλες προσεγγίσεις. Στόχος της είναι να προσδιορίσει και να κατανοήσει τους κινδύνους και να επικοινωνήσει τις ανησυχίες της όσο το δυνατόν νωρίτερα. Αυτό επιτρέπει στον φορολογούμενο να κάνει ενημερωμένες επιλογές σχετικά με την φορολογική του συμμόρφωση.

#### 2.2.4 Σύστημα αξιολόγησης φορολογικής συμμόρφωσης Ινδίας

Ο νόμος περί φόρου επί των αγαθών και υπηρεσιών (GST - Good and Service Tax) εισάγει σαφή αλλαγή στο ινδικό φορολογικό σύστημα. Εκτός από το ότι αναθεωρεί πλήρως τους υφιστάμενους φορολογικούς συντελεστές και διαδικασίες, δημιουργεί μια διαδικτυακή πύλη συμμόρφωσης και φέρνει διαφάνεια σε ολόκληρη τη διαδικασία μέσω των αξιολογήσεων συμμόρφωσης (**GST Compliance Rating System**). Όλοι οι καταχωρημένοι φορολογούμενοι θα αξιολογούνται δημόσια ανάλογα με το πόσο καλά - ή πόσο κακά - συμμορφώνονται με τους κανονισμούς της GST.

Οι φορολογούμενοι θα αποκτήσουν μια «βαθμολογία συμμόρφωσης» με βάση την αξιοπιστία τους όσον αφορά την έγκαιρη κατάθεση των φόρων στο δημόσιο ταμείο και την κατάθεση των αποδόσεων στο πλαίσιο της φορολογικής διάρθρωσης των αγαθών και των υπηρεσιών. Αυτή θα είναι ανεξάρτητη από τη φύση, το μέγεθος ή τον κύκλο εργασιών της επιχείρησης.

Η αξιολόγηση GST θα βασίζεται σε ορισμένα κριτήρια που δεν έχουν ακόμη καθοριστεί. Ωστόσο, οι εμπειρογνώμονες πιστεύουν ότι οι ακόλουθοι παράγοντες θα μπορούσαν να διαδραματίσουν βασικό ρόλο κατά την επιλογή των κριτηρίων συμμόρφωσης:

- Η έγκαιρη καταβολή των φόρων
- Η έγκαιρη κατάθεση των επιστροφών
- Η σύναψη διαφανών και χωρίς λάθη συμφωνιών
- Η συμμόρφωση με διάφορα άλλα χρονικά όρια βάσει του GST
- Η συνεργασία με τις φορολογικές αρχές του GST

Η βαθμολογία συμμόρφωσης θα ενημερώνεται περιοδικά. Θα υποδεικνύεται στον φορολογούμενο και επίσης θα είναι διαθέσιμη δημόσια.

Μερικά από τα οφέλη που μπορούν να απολαμβάνουν οι συμβατοί πωλητές περιλαμβάνουν:

- Απόκτηση επιστροφών αμέσως
- Οι αγοραστές μπορούν να λάβουν πίστωση φόρου εισροών αμέσως
- Προσέλκυση περισσότερων επιχειρήσεων
- Μείωση πιθανοτήτων ελέγχου από τις φορολογικές αρχές
- Καλύτερη φήμη

Οι αγοραστές θα αναζητήσουν πωλητές με υψηλότερη βαθμολογία για να διασφαλίσουν ότι θα μπορούν να επωφεληθούν ταχύτερα από την πίστωση φόρου εισροών, ενώ οι πωλητές με υψηλότερη βαθμολογία θα προσελκύσουν περισσότερους πελάτες. Δηλαδή η αξιολόγηση GST θα φέρει έναν υγιή ανταγωνισμό μεταξύ των επιχειρήσεων. Ένας υποψήφιος πελάτης θα ενημερωθεί σχετικά με τις αξιολογήσεις συμμόρφωσης των προμηθευτών πριν συνάψει μαζί τους συμφωνίες.

Η βαθμολογία GST είναι ένα λαμπρό σύστημα που θα παροτρύνει τις επιχειρήσεις να γίνουν πιο συμμορφωμένες και θα βοηθήσουν στην αποτροπή της φοροδιαφυγής.

### 2.2.5 Σύστημα βαθμολόγησης Φορολογουμένων Ουγγαρίας

Η βαθμολογία των φορολογουμένων ξεκίνησε το 2016 από τη ουγγρική φορολογική και τελωνιακή αρχή (NAV). Η αξιολόγηση γίνεται με βάση διάφορες θεσμικές παραμέτρους γνωστές στους φορολογούμενους. Ανάλογα με την αξιολόγηση, οι φορολογούμενοι χαρακτηρίζονται αξιόπιστοι ή επικίνδυνοι και υπόκεινται σε ευνοϊκότερους ή λιγότερο ευνοϊκούς κανόνες, όσον αφορά τις κυρώσεις που επιβάλλονται μετά από επιθεώρηση της NAV. Οι κανόνες για την αξιολόγηση των φορολογουμένων τροποποιήθηκαν στο νόμο περί των κανόνων φορολογίας από το 2017.

Ο στόχος είναι να δοθεί η δυνατότητα στη φορολογική αρχή να διαφοροποιήσει τις ομάδες φορολογουμένων που παραβλέπουν και παραβιάζουν εκ προθέσεως τους φορολογικούς κανόνες και εκείνους που ενεργούν καλή τη πίστει, προσπαθώντας να συμμορφωθούν πλήρως με το νόμο.

Οι φορολογούμενοι που δεν εμπίπτουν σε καμία κατηγορία εξακολουθούν να εκπληρώνουν τις υποχρεώσεις τους σύμφωνα με τους γενικούς κανόνες. Η NAV αναθεωρεί ανά τρίμηνο τους όρους των αξιολογήσεων και αποστέλλει σχετική ειδοποίηση σε πρώτο στάδιο. Στη συνέχεια, η ειδοποίηση αποστέλλεται μόνο εάν υπάρξει αλλαγή. Οι αξιολογήσεις μπορούν να ανακτηθούν μέσω της κυβερνητικής πύλης και να εμφανιστούν μεταξύ των βασικών δεδομένων.

### 2.2.6 Σύστημα αξιολόγησης φορολογουμένων στη Καμπότζη

Η πιο πρόσφατη φορολογική πρωτοβουλία της κυβέρνησης της Καμπότζης βαθμολογεί τους φορολογούμενους για τη συμμόρφωσή τους με τον φορολογικό κώδικα της Καμπότζης και προσφέρει οφέλη σε εκείνους που κατέχουν την υψηλότερη θέση.

Το νέο σύστημα βαθμολόγησης και ταξινόμησης περιγράφηκε σε μια πρακτική έκδοση της Γενικής Διεύθυνσης Φορολογίας (GDT – General Department of

Taxation), η οποία απαριθμεί 12 διαφορετικά κριτήρια, καθένα από τα οποία αξίζει μία ή δύο μονάδες. Συνολικά ο φορολογούμενος βαθμολογείται με 20 το πολύ μονάδες. Τα κριτήρια περιλαμβάνουν εάν μια εταιρεία έχει ολοκληρώσει τη φορολογική της εγγραφή και τις φορολογικές ενημερώσεις, εάν έχει καταθέσει τις φορολογικές δηλώσεις εντός της προθεσμίας και αν έχει διατηρήσει όλα τα απαιτούμενα λογιστικά αρχεία και νομικά έγγραφα.

Το βαθμολογημένο αποτέλεσμα χρησιμοποιείται για να τοποθετηθεί ο φορολογούμενος σε μία από τις τρεις κατηγορίες συμμόρφωσης:

- ✓ Χρυσοί φορολογούμενοι: Ανήκουν οι φορολογούμενοι που έχουν συγκεντρώσει περισσότερα από 16 σημεία.

Οι χρυσοί φορολογούμενοι θα μπορούν να ζητήσουν επιστροφή ΦΠΑ κάτω από το κατώτατο όριο των 500 εκατομμυρίων KHR (περίπου 125.000 δολαρίων ΗΠΑ) χωρίς να χρειάζεται να υποβληθούν σε έλεγχο ΦΠΑ. Επιπλέον, θα υποβάλλονται μόνο σε ένα Συνολικό Φορολογικό Έλεγχο κάθε δύο χρόνια και δεν θα χρειαστεί να υποβληθούν σε Περιορισμένο ή Επιτόπιο Έλεγχο κατά τη διάρκεια της ίδιας περιόδου.

- ✓ Ασημένιοι φορολογούμενοι: Ανήκουν οι φορολογούμενοι που έχουν συγκεντρώσει από 11 έως 15 σημεία.

Οι φορολογούμενοι της ασημένιας κατηγορίας μπορούν να ζητήσουν επιστροφή ΦΠΑ κάτω από 200 εκατομμύρια KHR (περίπου 50.000 δολάρια ΗΠΑ) χωρίς να υποβληθούν σε έλεγχο ΦΠΑ. Θα υπόκεινται μόνο σε έναν Συνολικό Φορολογικό Έλεγχο κάθε δύο χρόνια, έναν Περιορισμένο Φορολογικό Έλεγχο κάθε χρόνο και κανένα Έλεγχο Γραφείου κατά την ίδια περίοδο.

- ✓ Χάλκινοι φορολογούμενοι: Ανήκουν οι φορολογούμενοι που έχουν συγκεντρώσει από ένα έως 10 βαθμούς.

Πιστοποιητικό συμμόρφωσης απονέμεται σε όσους ανήκουν στην χρυσή κατηγορία και μπορεί να ζητηθεί από τους φορολογούμενους που ανήκουν στην ασημένια και χάλκινη κατηγορία. Τα πιστοποιητικά ισχύουν για δύο χρόνια, αν και το GDT διατηρεί το δικαίωμα να επαναξιολογήσει την πιστοποίησή του σε περίπτωση ύποπτης μη συμμόρφωσης.

## ΚΕΦΑΛΑΙΟ 3. ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

---

### 3.1 Ορισμοί

### 3.2 Στάδια της Διαδικασίας Ανακάλυψης Γνώσης

### 3.3 Μεθοδολογίες Εξόρυξης δεδομένων

### 3.4 Βασικά στάδια της εξόρυξης δεδομένων

### 3.5 Κατηγοριοποίηση

### 3.6 Μέθοδοι κατηγοριοποίησης

### 3.7 Δέντρα Αποφάσεων

---

### 3.1 Ορισμοί

Αξίζει να αναφέρουμε ότι υπάρχουν αντικρουόμενες απόψεις γύρω από το ποιος θα μπορούσε να είναι ένας σαφής και περιεκτικός ορισμός για την Εξόρυξη Δεδομένων.

Σύμφωνα με τους Witten and Frank (2000), η Εξόρυξη Δεδομένων (ΕΔ) (Data Mining (DM)) ορίζεται ως η διαδικασία ανακάλυψης προτύπων μέσα από δεδομένα, δίνοντας έτσι έμφαση στη διάσταση της Μηχανικής Μάθησης.

Σύμφωνα με τους Han and Kamber (2001), η Εξόρυξη Δεδομένων συνίσταται στην ανακάλυψη ή «εξόρυξη» γνώσης από μεγάλους όγκους δεδομένων. Ο ορισμός αυτός τονίζει τη διάσταση του όγκου των δεδομένων.

Άλλοι συγγραφείς, όπως οι Maimon and Rokach (2005), χρησιμοποιούν τον όρο Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων (Knowledge Discovery in Databases – KDD) για τη συνολική διαδικασία ανακάλυψης προτύπων μέσα από μεγάλα και περίπλοκα σύνολα δεδομένων.

Η ΕΔ έχει προκύψει ως μια επιτακτική ανάγκη για την παραγωγή νέας γνώσης από πληροφορίες που έχουν συσσωρευτεί από το παρελθόν. Η παραγωγή και

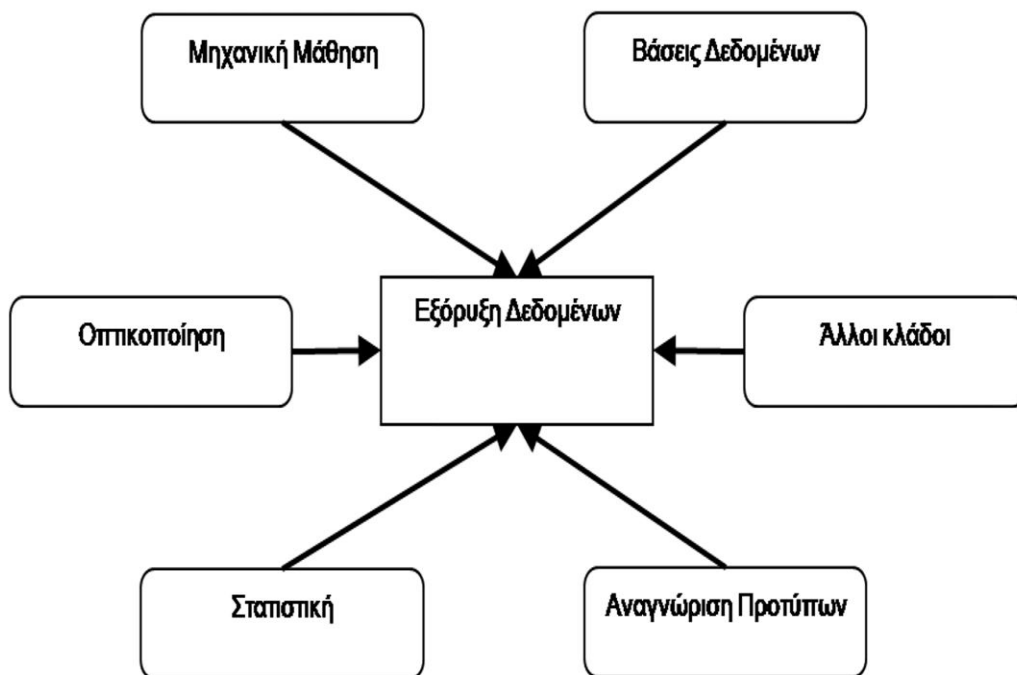


αποθήκευση δεδομένων με καταγιστικό ρυθμό δεν είναι πρόσφατη διαπίστωση. Μόλις το 1996, οι Fayyad, Piatetsky-Shapiro and Smyth (1996) ισχυρίζονται ότι τα δεδομένα συλλέγονται και συσσωρεύονται με δραματικό ρυθμό.

Η ΕΔ ανήκει στη γενικότερη μεθοδολογία της ανακάλυψης της γνώσης από βάσεις δεδομένων (KDD). Η ονομασία αυτή της KDD χρησιμοποιείται από το 1989 (πρώτο συνέδριο KDD) με στόχο να φανεί ότι η γνώση είναι το τελικό προϊόν μιας ανακάλυψης καθοδηγούμενης από τα δεδομένα (Frawley, U., Piatetsky – Shapiro & Matheus, 1991).

Για να ξεκαθαρίσουμε τις δύο έννοιες, αναφέρουμε το σχολιασμό που γίνεται από τους Fayyad et al. (1996), με βάση τους οποίους η KDD είναι η διαδικασία εύρεσης χρήσιμων πληροφοριών και προτύπων στα δεδομένα, ενώ η ΕΔ είναι η χρήση αλγορίθμων με στόχο την εξαγωγή πληροφοριών και προτύπων που παράγονται από τη διαδικασία KDD.

Η ΕΔ αντλώντας μεθοδολογίες από όλους τους επιστημονικούς κλάδους αποτελεί τέκνο της ανάγκης για επεξεργασία των αποθηκευμένων δεδομένων και εξαγωγή χρήσιμης πληροφορίας



**Εικόνα 3.1** Η Εξόρυξη Δεδομένων ως αποτέλεσμα συμβολής άλλων κλάδων.

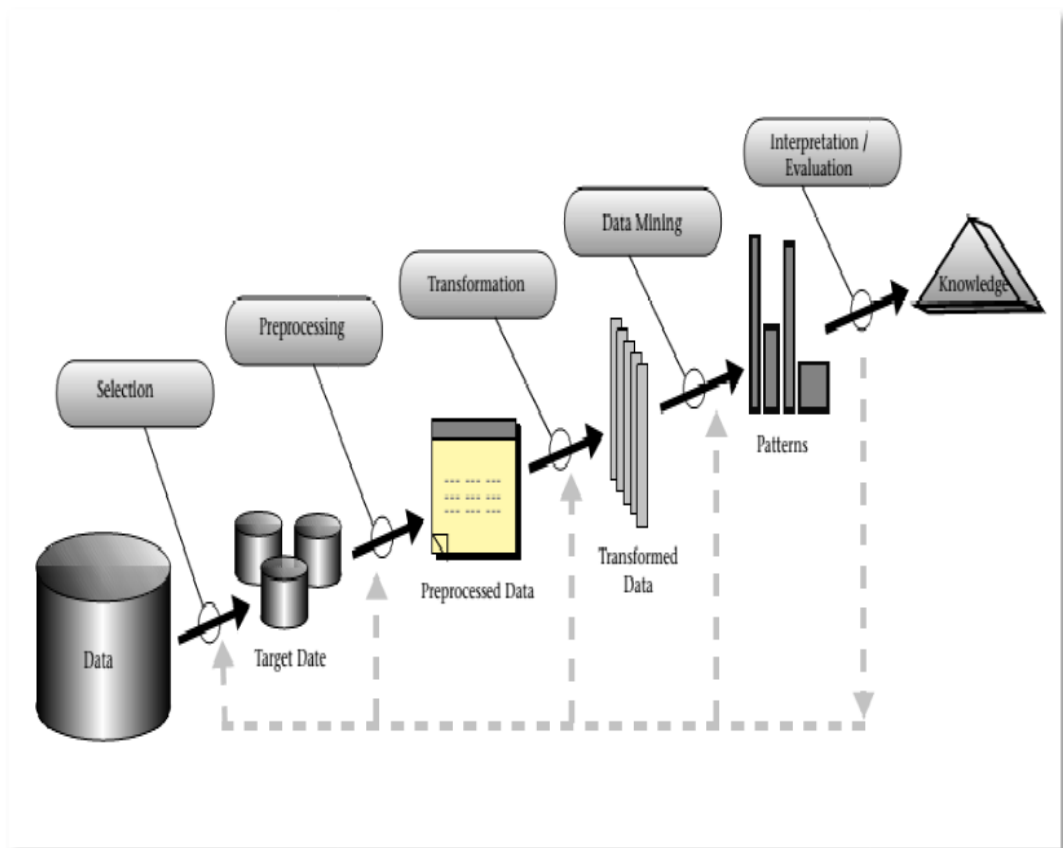
### 3.2 Στάδια της Διαδικασίας Ανακάλυψης Γνώσης

Όπως σημειώνουν οι Elovici and Braha (2003) η Ε.Δ. εστιάζει στην εξαγωγή πληροφοριών και όχι γνώσης. Με αυτό τον τρόπο διαπιστώνουν ότι η ανακάλυψη γνώσης από τις βάσεις δεδομένων αποτελείται από τα ακόλουθα πέντε στάδια:

- Πλήρης κατανόηση του κλάδου στον οποίο χρησιμοποιείται η Ε.Δ. και ακριβής προσδιορισμός των στόχων για τον οποίο αυτή πραγματοποιείται,
- Επιλογή και έλεγχος του συνόλου των δεδομένων που χρησιμοποιούνται,

- Μετασχηματισμός των δεδομένων προκειμένου αυτά να είναι κατάλληλα για επεξεργασία, αλλά και καθαρισμός των δεδομένων από ελλιπή στοιχεία ή λανθασμένη πληροφορία,
- Εξόρυξη των δεδομένων μέσω των προτύπων, δηλαδή μέσω της χρήσης του κατάλληλου αλγορίθμου και μοντέλου,
- Παρουσίαση των αποτελεσμάτων της διαδικασίας ΕΔ.

Το παρακάτω σχήμα αποδίδει συνοπτικά αυτή τη διαδικασία:



**Εικόνα 3.2** Επισκόπηση των βημάτων που αποτελούν μέρος μιας διαδικασίας εξόρυξης δεδομένων (Fayyad et al, 1996)

### 3.3 Μεθοδολογίες Εξόρυξης δεδομένων

Υπάρχουν διάφορες μεθοδολογίες ΕΔ και χωρίζονται σε μεθοδολογίες **επιβλεπόμενης μάθησης** (supervised) και μεθοδολογίες **μη επιβλεπόμενης μάθησης** learning (unsupervised learning). Η επιβλεπόμενη μάθηση έχει στόχο τη μοντελοποίηση των σχέσεων ανάμεσα σε ένα εξαρτημένο γνώρισμα – στόχο και σε άλλα ανεξάρτητα γνωρίσματα. Η ανάλυση συνίσταται στην τυποποίηση των σχέσεων ανάμεσα στην εξαρτημένη και στις ανεξάρτητες μεταβλητές, συνήθως με τη δημιουργία ενός μοντέλου, που επιτρέπει τον υπολογισμό της εξαρτημένης μεταβλητής από τις ανεξάρτητες. Το μοντέλο μπορεί να χρησιμοποιηθεί για τη διατύπωση προβλέψεων. Το όνομα «επιβλεπόμενη μάθηση» σημαίνει ότι το γνώρισμα στόχος και οι τιμές του καθοδηγούν τη διαδικασία μάθησης. Στη μη επιβλεπόμενη μάθηση δεν υπάρχει κάποια στήλη στόχος και οι αλγόριθμοι προσπαθούν να ομαδοποιήσουν τα δεδομένα σε ομάδες που δεν είναι γνωστές εκ των προτέρων.

Οι κυριότερες μεθοδολογίες ΕΔ (Fayyad et al, 1996; Shaw et al, 2001) είναι οι ακόλουθες:

#### ➤ Κατηγοριοποίηση (Classification)

Η ταξινόμηση είναι η εκμάθηση μιας συνάρτησης που χαρτογραφεί (ταξινομεί) ένα στοιχείο δεδομένων σε μία από αρκετές προκαθορισμένες κατηγορίες (Weiss και Kulikowski 1991, Hand 1981).

Παραδείγματα μεθόδων ταξινόμησης που χρησιμοποιούνται ως μέρος των εφαρμογών ανακάλυψης γνώσης περιλαμβάνουν την ταξινόμηση των τάσεων στις χρηματοπιστωτικές αγορές (Apte και Hong 1996) και την αυτοματοποιημένη αναγνώριση των αντικειμένων που ενδιαφέρουν τις βάσεις δεδομένων μεγάλης εικόνας (Fayyad, Djorgovski και Weir 1996).

#### ➤ Παλινδρόμηση (Regression)

Η παλινδρόμηση είναι η εκμάθηση μιας συνάρτησης που χαρτογραφεί ένα στοιχείο δεδομένων σε μια πραγματική μεταβλητή πρόβλεψης. Οι εφαρμογές παλινδρόμησης είναι πολλές, για παράδειγμα, εκτιμώντας την

πιθανότητα επιβίωσης ενός ασθενούς, λαμβάνοντας υπόψη τα αποτελέσματα ενός συνόλου διαγνωστικών εξετάσεων (Fayyad et al, 1996).

➤ **Συσταδοποίηση (Clustering)**

Η συσταδοποίηση (clustering) είναι μια από τις πιο χρήσιμες διεργασίες στη διαδικασία εξόρυξης γνώσης για την ανακάλυψη συστάδων και για τον προσδιορισμό κατανομών ή προτύπων (patterns). Το πρόβλημα της συσταδοποίησης σχετίζεται με την τμηματοποίηση (partitioning) ενός συνόλου δεδομένων σε συστάδες έτσι ώστε τα στοιχεία του συνόλου των δεδομένων που ανήκουν σε μια συστάδα να είναι περισσότερο όμοια μεταξύ τους από ότι είναι με τα στοιχεία των άλλων συστάδων.

Στη διαδικασία της συσταδοποίησης δεν υπάρχουν προκαθορισμένες κατηγορίες ούτε κάποιο παράδειγμα που θα έδειχνε ποιες επιθυμητές σχέσεις θα ήταν έγκυρες μεταξύ των δεδομένων. Για το λόγο αυτό είναι γνωστή ως διαδικασία μη εποπτευόμενης μάθησης.

➤ **Συσχετίσεις (Correlations)**

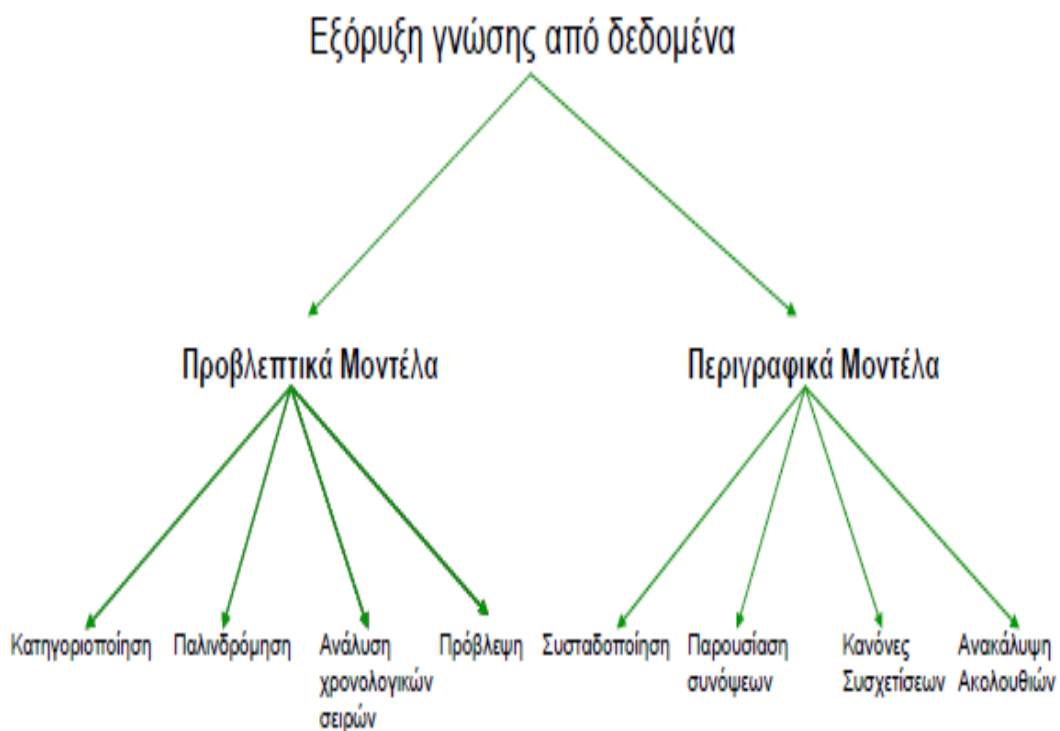
Οι κανόνες συσχέτισης, ή αλλιώς association rules, είναι κανόνες οι οποίοι εκφράζουν συσχετίσεις μεταξύ αντικειμένων. Πιο συχνά χρησιμοποιούνται σε συστήματα σημείων πώλησης προϊόντων, οπότε οι συσχετίσεις που εκφράζουν είναι μεταξύ των διαφόρων προϊόντων που αγοράζουν οι πελάτες.

➤ **Περιληπτική παρουσίαση (Descriptives)**

Η συνοπτική παρουσίαση πληροφορίας περιλαμβάνει τη διαδικασία ανεύρεσης μιας συμπαγούς περιγραφής για ένα σύνολο δεδομένων (Barbara, et.al., 1996). Για παράδειγμα, η περιληπτική παρουσίαση της πληροφορίας θα μπορούσε να χρησιμοποιηθεί για να υπολογίσει τη μέση τιμή και τη σταθερή απόκλιση για όλες τις ιδιότητες ενός συνόλου δεδομένων. Άλλες πιο περίπλοκες εφαρμογές της περιληπτικής παρουσίασης της πληροφορίας είναι η παραγωγή των συνοπτικών κανόνων, τεχνικών απεικόνισης πολλών μεταβλητών. Οι τεχνικές περιληπτικής παρουσίασης της πληροφορίας εφαρμόζονται συχνά στη διαλογική διερευνητική ανάλυση δεδομένων και την αυτοματοποιημένη παραγωγή εκθέσεων (Fayyad et al, 1996).

➤ **Ανίχνευση μεταβολών και αποκλίσεων (Deviation Analysis)**

Η ανίχνευση μεταβολών και αποκλίσεων επικεντρώνεται στην ανακάλυψη των σημαντικότερων αλλαγών στα δεδομένα από προηγούμενες μετρημένες ή κανονιστικές τιμές (Berndt and Clifford 1996; Guyon, Matic, and Vapnik 1996; Kloesgen 1996; Matheus, Piatetsky-Shapiro, and McNeill 1996; Basseville and Nikiforov 1993).

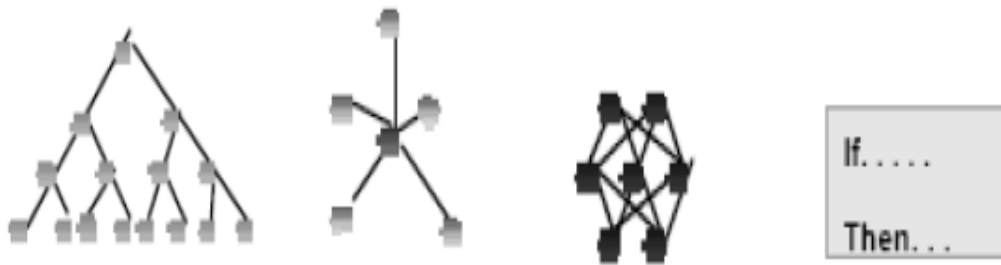


**Εικόνα 3.3:** Μια ταξινόμηση των κυριότερων εργασιών ΕΔ.(Dunham N 2004)

Τα διαφορετικά επίπεδα ανάλυσης που είναι διαθέσιμα:

- Τεχνητά νευρωνικά δίκτυα (artificial Neural Networks)
- Δέντρα αποφάσεων (Decision Trees)
- Γενετικοί αλγόριθμοι (Genetic Algorithms)
- Επαγωγή αποτελεσμάτων μέσω διαφόρων κανόνων απόφασης (Rule Induction)

- Μέθοδος του κοντινότερου γείτονα (Nearest Neighbor Method)
- Απεικόνιση των δεδομένων (Data Visualization)
- Μοντέλα πρόβλεψης (Predictive Models)



**Εικόνα 3.4:** *Decision Trees, Nearest Networks, Neural Networks, Rule Induction*

Όπως θα δούμε στη συνέχεια, όλοι οι αλγόριθμοι της διαδικασίας της ΕΔ αποτελούνται από κάποιες συναρτήσεις αποτελεσμάτων, κάποια μοντέλα και κάποια μέθοδο βελτιστοποίησης. Έτσι αφού προσδιοριστεί ο στόχος, γίνεται η αναπαράσταση του μοντέλου που θα χρησιμοποιηθεί και αποφασίζεται η συνάρτηση μέσω της οποίας γίνεται η αξιολόγηση. Στη συνέχεια, εφαρμόζεται η διαδικασία της βελτιστοποίησης και διευκρινίζεται ο τρόπος με τον οποίο διαχειρίζονται τα δεδομένα, όπως και οι παράμετροι των μοντέλων. Αυτή η διαδικασία μπορεί να γίνει αντιληπτή και από το παρακάτω διάγραμμα το οποίο μπορεί να γενικευτεί σε όλες τις εξεταζόμενες μεθοδολογίες.

Στόχος → Αναπαράσταση → Συνάρτηση αξιολόγησης → Βελτιστοποίηση  
→ Διαχείριση → Παράμετροι μοντέλων

### 3.4 Βασικά στάδια της εξόρυξης δεδομένων

Υπάρχουν τρία στάδια στη διαδικασία της ΕΔ, έτσι όπως περιγράφεται από τους Fayyad et al. (1996).

#### 1.Περιγραφή μοντέλου

Στο πρώτο στάδιο της ΕΔ, επιχειρούμε να δηλώσουμε τη **λειτουργία** του μοντέλου, δηλαδή να δηλώσουμε το στόχο μας, όπως για παράδειγμα την ταξινόμηση (classification), την παλινδρόμηση (regression) ή τη συσταδοποίηση (clustering).

Επίσης, μας ενδιαφέρει η **παραστατική μορφή** του μοντέλου, δηλαδή η απεικόνισή του έτσι ώστε να ταιριάζει με την απεικόνιση των δεδομένων και να είναι δυνατό να ερμηνευθεί. Χαρακτηριστικά παραδείγματα μοντέλων είναι τα δέντρα απόφασης, τα γραφικά μοντέλα, τα νευρωνικά δίκτυα και τα μοντέλα – συστήματα που βασίζονται σε παραδείγματα ή πιθανότητες (δίκτυα Bayes).

#### 2.Αξιολόγηση μοντέλου

Ύστερα από τη δημιουργία του μοντέλου, οφείλουμε να εξετάσουμε κατά πόσο ταιριάζει με τις συνθήκες της KDD. Προχωράμε, δηλαδή, στην αξιολόγηση του μοντέλου, ώστε να κρίνουμε την εγκυρότητα των προτύπων και την ακρίβεια και χρησιμότητα του μοντέλου. Υπάρχουν διάφορα κριτήρια αξιολόγησης, όπως αυτό της μέγιστης πιθανότητας.

#### 3.Αλγόριθμος αναζήτησης

Στόχος του σταδίου αυτού είναι η σύγκριση μοντέλων και παραμέτρων, δοθέντων του συνόλου δεδομένων, της οικογένειας μοντέλων και του κριτηρίου αξιολόγησης. Οι βασικότεροι τύποι αλγορίθμων αναζήτησης είναι αυτοί που αναζητούν παραμέτρους βελτιστοποίησης ενός κριτηρίου αξιολόγησης και αυτοί που αναζητούν μοντέλα αντιπροσώπευσης των δεδομένων.

Το ερώτημα που προκύπτει είναι αν η επιλεχθείσα μεθοδολογία μπορεί να αποδώσει και τα αναμενόμενα αποτελέσματα.



Η κυριότερη αδυναμία σύμφωνα με τους Fayyad et al. (1996) είναι ότι οι βάσεις δεδομένων είναι πολύ μεγάλες με αποτέλεσμα να απαιτείται η χρήση αρκετά δύσκολων αλγορίθμων, προκειμένου να μπορέσει να εξαχθεί γνώση από αυτές. Επίσης, τα δεδομένα μπορεί να είναι αρκετά στενά συνδεδεμένα μεταξύ τους με αποτέλεσμα η υπό εξέταση μεταβλητή να μπορεί να ερμηνευθεί από πολλές επιμέρους μεταβλητές. Ένας ακόμη ανασταλτικός παράγοντας είναι και το γεγονός ότι τα δεδομένα μπορεί να είναι ελλιπή ή και να παρουσιάζουν υψηλά επίπεδα θορύβου. Ο τελευταίος ανασταλτικός παράγοντας που παρουσιάζεται στην ίδια μελέτη αφορά στο ότι τα δεδομένα αρκετές φορές μπορεί να χρειάζεται να συνδυαστούν και με άλλα συστήματα προκειμένου να εξαχθούν ασφαλή συμπεράσματα κι έτσι η διαδικασία να γίνεται ακόμη πιο περίπλοκη. Οι Feelders et al. (2000) προσδιορίζουν ακόμη έναν ανασταλτικό παράγοντα στη διαδικασία αυτή ο οποίος αφορά στο ότι η Ε.Δ. θα πρέπει να γίνεται από άτομα τα οποία χαρακτηρίζονται ως ειδικοί, αφού αυτοί έχουν την ικανότητα να συνδυάζουν με ιδιαίτερα αποτελεσματικό τρόπο τα δεδομένα μεταξύ τους αλλά και να εντοπίσουν πολύ λεπτά και ιδιαίτερα πρότυπα.

### 3.5 Κατηγοριοποίηση

Στην παρούσα εργασία θα εφαρμόσουμε τη μέθοδο της Κατηγοριοποίησης (classification).

Το πρόβλημα της κατηγοριοποίησης έχει μελετηθεί εκτενώς στη στατιστική, στην αναγνώριση προτύπων (patterns) και μηχανικής μάθησης (machine learning) στα πλαίσια του προβλήματος της ανάκτησης ή εξαγωγής γνώσης από σύνολα δεδομένων (Duda, et. al., 1976).

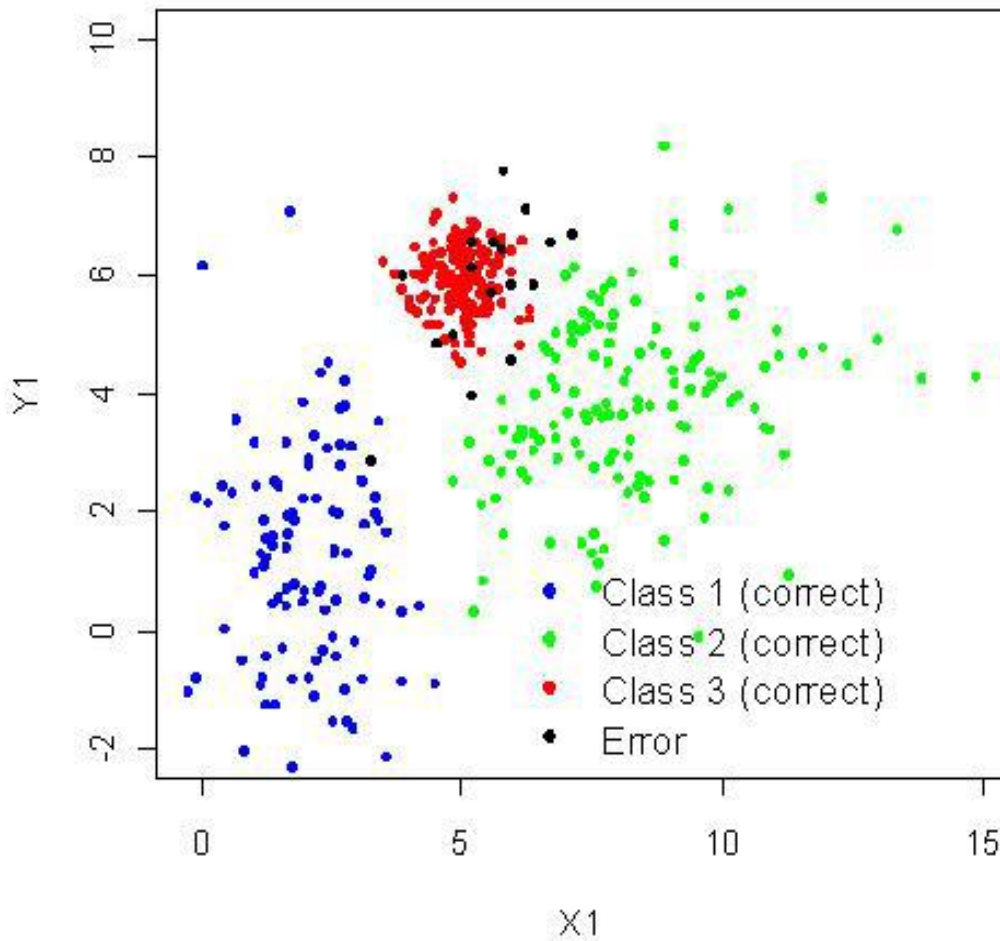
Χαρακτηρίζεται ως μία από τις βασικές εργασίες στη διαδικασία εξόρυξης γνώσης, η οποία αποσκοπεί στην ανάθεση ενός στοιχείου σε ένα προκαθορισμένο σύνολο κατηγοριών (classes). Η κατηγοριοποίηση (classification) μπορεί να περιγραφεί ως

μία λειτουργία που αντιστοιχίζει (κατηγοριοποιεί) ένα στοιχείο σε μία από τις διαφορετικές κατηγορίες που έχουν προκαθοριστεί (Fayyad, et.al., 1996).

Η κατηγοριοποίηση χαρακτηρίζεται από ένα καλά καθορισμένο σύνολο κατηγοριών καθώς και ένα σύνολο από προκατηγοριοποιημένα (pre-classified) παραδείγματα (αντίθετα, η διαδικασία συσταδοποίησης δεν στηρίζεται σε προκαθορισμένες κατηγορίες ή παραδείγματα). Γενικά, ο στόχος της διαδικασίας κατηγοριοποίησης είναι η δημιουργία ενός μοντέλου που θα μπορεί να χρησιμοποιηθεί για την κατηγοριοποίηση μελλοντικών δεδομένων των οποίων η κατηγοριοποίηση είναι άγνωστη (Berry, et.al., 1996).

Στις περισσότερες περιπτώσεις, υπάρχει ένας περιορισμένος αριθμός κατηγοριών και θα πρέπει κάθε εγγραφή να ανατεθεί στην κατάλληλη κατηγορία. Για το σκοπό αυτό χρησιμοποιούνται κάποιες τεχνικές, οι οποίες κατατάσσονται σε δύο κατηγορίες. Η πρώτη χρησιμοποιεί Δέντρα Αποφάσεων (Decision Trees) (Han, et. al., 2001) και η δεύτερη Νευρωνικά Δίκτυα (Neural Networks). Και οι δύο στηρίζονται στην ιδέα της «εκπαίδευσης» (training) με τη βοήθεια ενός υποσυνόλου δεδομένων που ονομάζεται σύνολο εκπαίδευσης (training set). Το υποσύνολο αυτό επιλέγεται σαν αντιπροσωπευτικό δείγμα του συνολικού όγκου δεδομένων. Με την εφαρμογή της διαδικασίας εκπαίδευσης καθορίζονται κάποια πρότυπα για τις κατηγορίες δεδομένων.

Έτσι, όταν προκύψει ένα νέο στοιχείο τότε μπορεί εύκολα να κατηγοριοποιηθεί. Για τη διαδικασία αυτή χρησιμοποιούνται είτε τεχνικές βασισμένες στα νευρωνικά δίκτυα είτε συμβολικές τεχνικές. Στις πρώτες υπάρχει το φαινόμενο της αμφίδρομης αναμετάδοσης και επεξεργασίας δεδομένων ενώ στη δεύτερη υπάρχουν μοντέλα δέντρων αποφάσεων ή μοντέλα για IF...THEN...ELSE ανάλυση.

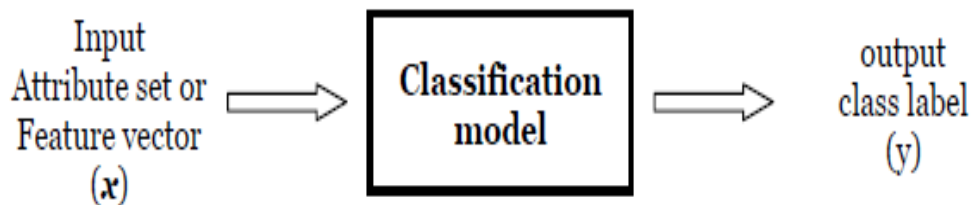


**Εικόνα 3.5:** Ένα απλό γραμμικό όριο κατηγοριοποίησης με τρεις κατηγορίες.

Πιο συγκεκριμένα, η κατηγοριοποίηση δεδομένων μπορεί να περιγραφεί ως μία διαδικασία δύο βημάτων:

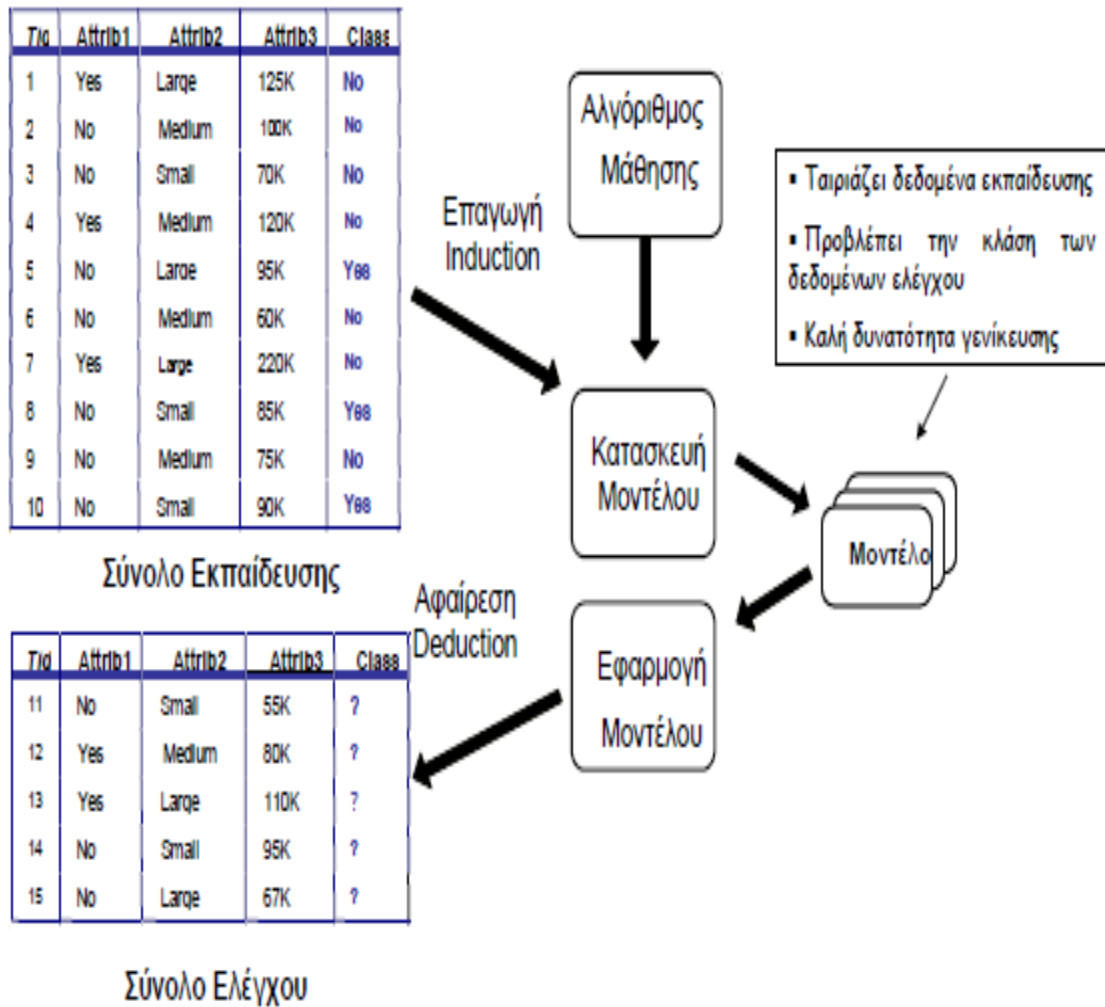
**Βήμα 1ο:** Εκμάθηση (Learning). Σε αυτό το βήμα χτίζεται ένα μοντέλο (model), περιγράφοντας ένα προκαθορισμένο σύνολο από κατηγορίες δεδομένων. Τα δεδομένα εκπαίδευσης (training data) αναλύονται από έναν αλγόριθμο κατηγοριοποίησης για να κατασκευάσουν στη συνέχεια το μοντέλο. Τα στοιχεία που αποτελούν το σύνολο κατάρτισης επιλέγονται τυχαία από έναν πληθυσμό δεδομένων και ανήκουν σε μία από τις προκαθορισμένες κατηγορίες. Δεδομένου ότι η κατηγορία των δειγμάτων εκπαίδευσης είναι γνωστή, αυτό το βήμα είναι επίσης γνωστό σαν «εποπτευμένη μάθηση» (supervised learning). Το μοντέλο που

ορίζεται, γνωστό και ως κατηγοριοποιητής (classifier), αναπαριστάται με τη μορφή κανόνων κατηγοριοποίησης (classification rules), δέντρων αποφάσεων (decision trees) ή μαθηματικών τύπων (mathematical formulas) (Han, et.al., 2001).



**Εικόνα 3.6:** Λειτουργία μοντέλου ταξινόμησης

**Βήμα 2ο:** Κατηγοριοποίηση (Classification). Σε αυτό το βήμα χρησιμοποιούνται τα δοκιμαστικά δεδομένα (test data) για να υπολογίσουν την ακρίβεια (accuracy) του μοντέλου. Υπάρχουν διάφορες μέθοδοι για να εκτιμηθεί η ακρίβεια του κατηγοριοποιητή (classifier). Τα δεδομένα εκπαίδευσης επιλέγονται τυχαία και είναι ανεξάρτητα. Το μοντέλο κατηγοριοποιεί κάθε ένα από τα δοκιμαστικά παραδείγματα (training samples). Στη συνέχεια η κατηγορία που ανήκουν τα δεδομένα με βάση το σύνολο δοκιμαστικών δεδομένων συγκρίνεται με την πρόβλεψη που έκανε το μοντέλο για την κατηγορία. Η ακρίβεια του μοντέλου σε ένα καθορισμένο σύνολο δεδομένων δοκιμής είναι το ποσοστό των δειγμάτων δοκιμής που κατηγοριοποιήθηκαν σωστά από το υπό εκπαίδευση μοντέλο.



**Εικόνα 3.7:** Γενικό σχήμα ταξινόμησης-κατηγοριοποίησης

Εάν η ακρίβεια του μοντέλου θεωρείται ως αποδεκτή, το μοντέλο μπορεί πλέον να χρησιμοποιηθεί για να κατηγοριοποιήσει τα μελλοντικά δείγματα δεδομένων (αντικείμενα), των οποίων η κατηγοριοποίηση είναι άγνωστη.

Όταν έχουμε ένα σύνολο δεδομένων που κατηγοριοποιούνται σε δυο μόνο κατηγορίες, τότε προκύπτουν τέσσερα πιθανά ενδεχόμενα για την κατηγοριοποίηση. Μία πλειάδα της βάσης δεδομένων θα εκχωρηθεί σε μια από τις δυο κατηγορίες, ενώ στην πραγματικότητα μπορεί να ανήκει ή να μην ανήκει σε αυτή τη κατηγορία.

|            |       | Actual         |                |
|------------|-------|----------------|----------------|
|            |       | True           | False          |
| Classifier | True  | True Positive  | False Positive |
|            | False | False Negative | True Negative  |

**Εικόνα 3.8 :** *ConfusionMatrix*, σύγκριση της επίδοσης της κατηγοριοποίησης με την ανάκτηση της πληροφορίας

Έτσι οδηγούμαστε στα τέσσερα τεταρτημόρια τα οποία είναι τα εξής (Dunham, M., 2004).:

- Αληθώς θετικό (Truepositive – TP): η πλειάδα  $x_1$  εκτιμάται ότι ανήκει στην κατηγορία C και όντως στην πραγματικότητα ανήκει σε αυτήν.
- Ψευδώς θετικό (FalsePositive – FP): η πλειάδα  $x_1$  εκτιμάται ότι ανήκει στην κατηγορία C ενώ στην πραγματικότητα δεν ανήκει σε αυτήν.
- Αληθώς αρνητικό (TrueNegative – TN): η πλειάδα  $x_1$  εκτιμάται ότι δεν ανήκει στην κατηγορία C και όντως στην πραγματικότητα δεν ανήκει σε αυτήν.
- Ψευδώς αρνητικό (FalseNegative – FN): η πλειάδα  $x_1$  εκτιμάται ότι δεν ανήκει στην κατηγορία C ενώ στην πραγματικότητα ανήκει σε αυτήν.

Κατά την εκπαίδευση πρέπει να αποφευχθεί η υπερπροσαρμογή του μοντέλου, η απομνημόνευση δηλαδή του συγκεκριμένου συνόλου εκπαίδευσης. Αποτέλεσμα της υπερπροσαρμογής είναι η πτώση της επίδοσης έναντι άγνωστων παρατηρήσεων.

Κριτήρια για την αξιολόγηση των μεθόδων κατηγοριοποίησης είναι η ακρίβεια πρόβλεψης, η ταχύτητα, η ερμηνευσιμότητα, η επεκτασιμότητα και η ανθεκτικότητα.

### 3.6 Μέθοδοι κατηγοριοποίησης

Μέθοδοι κατηγοριοποίησης υπάρχουν πολλές, όπως πχ τα Δέντρα Αποφάσεων, τα Μπαϋεσιανά Δίκτυα, τα Νευρωνικά Δίκτυα τύπου Multilayer Perceptron κλπ. Τα μοντέλα τα οποία κατασκευάζει η κάθε μέθοδος είναι τελείως διαφορετικά. Για παράδειγμα, ένα μοντέλο Δέντρου Αποφάσεων είναι μια δενδρική δομή, όπου κάθε κόμβος είναι ένας έλεγχος σε κάποιο γνώρισμα, κάθε κλάδος είναι ένα αποτέλεσμα του ελέγχου και κάθε φύλο είναι μια απόφαση κατηγοριοποίησης. Ένα μοντέλο Μπαϋεσιανού Δικτύου είναι ένας κατευθυνόμενος ακυκλικός γράφος και μια κατανομή πιθανοτήτων συσχέτισης μεταξύ των μεταβλητών. Στον γράφο κάθε κόμβος αντιστοιχεί σε μια μεταβλητή.

Όπως αναφέρθηκε και προηγουμένως, ένας επαγωγικός αλγόριθμος επεξεργάζεται ένα σύνολο δεδομένων και παράγει ένα μοντέλο. Αν συμβολίσουμε με το  $I$  ένα επαγωγικό αλγόριθμο και με  $D$  ένα σύνολο δεδομένων, τότε το  $I(D)$  συμβολίζει το μοντέλο που θα παραχθεί από την επεξεργασία του  $D$  από τον  $I$ . Αν ο ίδιος αλγόριθμος επεξεργαστεί ένα διαφορετικό σύνολο δεδομένων  $D'$ , τότε θα παραχθεί ένα διαφορετικό μοντέλο  $I(D')$ . Εάν το  $D'$  περιέχει διαφορετικά πεδία από το  $D$ , είναι προφανές ότι τα δύο μοντέλα θα είναι διαφορετικά. Αν τα δύο σύνολα δεδομένων περιέχουν τα ίδια πεδία, αλλά διαφορετικές παρατηρήσεις, τότε και πάλι θα προκύψουν δύο διαφορετικά μοντέλα. Υπάρχουν μάλιστα μέθοδοι, όπως τα Δέντρα Αποφάσεων, όπου μικρές διαφορές στα σύνολα δεδομένων έχουν σαν αποτέλεσμα τη δημιουργία σημαντικά διαφορετικών μοντέλων.

Μια μέθοδος κατηγοριοποίησης μπορεί να είναι **αιτιοκρατική (deterministic)** ή **στοχαστική (stochastic)**.

Οι στοχαστικές μέθοδοι καλούνται και πιθανολογικές (probabilistic). Μια αιτιοκρατική μέθοδος δημιουργεί μοντέλα, τα οποία εκχωρούν μια παρατήρηση σε μια κλάση. Τα στοχαστικά μοντέλα υπολογίζουν την πιθανότητα να ανήκει η παρατήρηση σε κάθε μια από τις δυνατές κλάσεις. Παράδειγμα αιτιοκρατικής

μεθόδου είναι τα Δέντρα αποφάσεων τύπου C4.5, ενώ παράδειγμα στοχαστικής μεθόδου είναι τα Μπαϋεσιανά Δίκτυα.

Θα αναλύσουμε την τεχνική των δέντρων αποφάσεων και τους κυριότερους αλγόριθμους που είναι βασισμένοι στην μέθοδο αυτή.

### 3.7 Δέντρα Αποφάσεων

Τα Δέντρα Αποφάσεων είναι μια από τις βασικότερες και πιο δημοφιλείς μεθόδους κατηγοριοποίησης. Βασική λογική της κατασκευής τους είναι η διαδοχική διάσπαση του συνόλου των παρατηρήσεων σε υποσύνολα.

Κριτήριο για τη διάσπαση είναι οι τιμές των μεταβλητών. Η διαδικασία των διαδοχικών διασπάσεων αναπαρίσταται με μια ανεστραμμένη δενδρική δομή. Στην κορυφή βρίσκεται ο κόμβος-ρίζα του δέντρου. Σε κατώτερα επίπεδα βρίσκονται επιπλέον κόμβοι, οι οποίοι συνδέονται με ακμές με άλλα στοιχεία του δέντρου. Στο κατώτερο επίπεδο κάθε κλάδου βρίσκονται τα φύλλα του δέντρου. Ο κόμβος - ρίζα έχει μόνο εξερχόμενες ακμές που τον συνδέουν με στοιχεία του κατώτερου επιπέδου. Οι υπόλοιποι κόμβοι έχουν εισερχόμενες ακμές που τους συνδέουν με τους κόμβους του ανώτερου επιπέδου και εξερχόμενες ακμές που τους συνδέουν με στοιχεία του κατώτερου επιπέδου. Τέλος, τα φύλλα έχουν μόνο εισερχόμενες ακμές, οι οποίες τα συνδέουν με τους κόμβους του ανώτερου επιπέδου. Κάθε κόμβος αντιπροσωπεύει έναν έλεγχο στα δεδομένα και αντίστοιχη διάσπαση τους σε δύο ή περισσότερα υποσύνολα, ανάλογα με το αποτέλεσμα του ελέγχου. Η συνηθέστερη εκδοχή είναι ο έλεγχος να περιλαμβάνει μία μόνο μεταβλητή, έχουν προταθεί ωστόσο αλγόριθμοι όπου σε έναν κόμβο ελέγχονται περισσότερες μεταβλητές. Κάθε ακμή αντιπροσωπεύει ένα αποτέλεσμα του ελέγχου και το αντίστοιχο υποσύνολο των δεδομένων. Τέλος, κάθε φύλλο αντιπροσωπεύει μια απόφαση κατηγοριοποίησης.



### 3.7.1 Δημιουργία κανόνων από Δέντρα Αποφάσεων

Στα Δέντρα Αποφάσεων η ανακαλυφθείσα γνώση αναπαρίσταται με τέτοιον τρόπο, ώστε είναι εύκολη η εξαγωγή επαγωγικών κανόνων της μορφής IF-THEN. Για κάθε φύλο δημιουργείται ένας κανόνας, ο οποίος περιλαμβάνει τις λογικές συνθήκες όλων των ελέγχων από τη ρίζα έως το φύλο. Οι λογικές αυτές συνθήκες συνδέονται με τον λογικό τελεστή AND. Ωστόσο, μια τέτοιου τύπου εξαγωγή κανόνων, οδηγεί συχνά σε κανόνες περιττά περίπλοκους. Για την απλοποίηση των κανόνων μπορεί να υιοθετηθεί μια προσέγγιση κλαδέματος, η οποία συνίσταται στην απομάκρυνση των μη σημαντικών συνθηκών. Ο καθορισμός των μη σημαντικών συνθηκών επιτυγχάνεται με τη σύγκριση του ρυθμού σφάλματος του απλουστευμένου κανόνα με τον ρυθμό σφάλματος του πλήρους κανόνα.

### 3.7.2 Πλεονεκτήματα και Μειονεκτήματα των Δέντρων Αποφάσεων

Τα πλεονεκτήματα της χρήσης δέντρων αποφάσεων για την μέθοδο της κατηγοριοποίησης είναι ότι (Dunham, M., 2004).:

- ✓ είναι εύχρηστα και αποτελεσματικά
- ✓ έχουν την δυνατότητα να εξάγουν κανόνες εύκολους στην κατανόηση και στην ερμηνεία
- ✓ τα δέντρα αποφάσεων αποδίδουν καλά σε μεγάλες βάσεις δεδομένων επειδή το μέγεθος του δέντρου είναι ανεξάρτητο από το μέγεθος της βάσης δεδομένων
- ✓ τέλος τα δέντρα είναι ικανά να χειριστούν δεδομένα με πολλά γνωρίσματα (κόμβους).

Τα μειονεκτήματα των δέντρων αποφάσεων είναι ότι (Dunham, M., 2004).:

- ✓ δεν μπορούν να χειριστούν εύκολα συνεχή δεδομένα. Τα γνωρίσματα που περιέχουν συνεχή δεδομένα θα πρέπει να χωριστούν σε κατηγορίες για να μπορέσουν να χρησιμοποιηθούν
- ✓ τα ελλιπή δεδομένα γίνονται δύσκολα στη χρήση τους γιατί δεν μπορούν να βρεθούν τα σωστά κλαδιά του δέντρου για να ακολουθηθούν

- ✓ τα δέντρα απόφασης λόγω ότι δημιουργούνται από τα δεδομένα εκπαίδευσης κάποιες φορές μπορεί να εμφανίσουν υπέρ-προσαρμογή
- ✓ τέλος παραλείπονται από τη διαδικασία των δέντρων οι συσχετίσεις μεταξύ των γνωρισμάτων της βάσης δεδομένων.

### 3.7.3 Αλγόριθμοι κατασκευής Δέντρων Αποφάσεων

Κατά τη διάρκεια των τελευταίων ετών έχουν αναπτυχθεί διάφοροι αλγόριθμοι κατασκευής των δέντρων αποφάσεων. Μερικοί από τους πιο γνωστούς αλγορίθμους που προτείνονται στη βιβλιογραφία είναι οι: ID3 (Mitchell, 1997), C4.5 (Quinlan, 1993), SPRINT (Shafer, et.al., 1996), SLIQ (Melta, 1996), CHAID, CART (Breiman, et.al., 1984), , κ.λπ. Γενικά, οι περισσότεροι από τους αλγόριθμους έχουν δύο διακριτές φάσεις: τη φάση οικοδόμησης (building phase,) και τη φάση περικοπής (pruning phase) (Mitchell, 1997). Στη φάση οικοδόμησης, το σύνολο των δεδομένων εκπαίδευσης χωρίζεται κατ' επανάληψη μέχρις ότου όλα τα δείγματα σε ένα τμήμα (partition) να ανήκουν στην ίδια κατηγορία. Το αποτέλεσμα είναι ένα δέντρο που κατηγοριοποιεί κάθε στοιχείο του συνόλου εκπαίδευσης. Ωστόσο, το δέντρο που κατασκευάζεται μπορεί να είναι ευαίσθητο στις στατιστικές παρατυπίες (irregularities) του συνόλου κατάρτισης. Κατά συνέπεια, οι περισσότεροι από τους αλγορίθμους εκτελούν μια φάση περικοπής μετά από τη φάση κατασκευής του δέντρου, στην οποία οι κόμβοι περικόπτονται για να αποτραπούν οι επικαλύψεις και για να δημιουργηθεί ένα δέντρο με υψηλότερη ακρίβεια (Χαλκίδης, Βαζιργιάννης, 2005).

Οι διάφοροι αλγόριθμοι κατασκευής δέντρων αποφάσεων χρησιμοποιούν διαφορετικούς αλγορίθμους για την επιλογή του κριτηρίου ελέγχου για την κατηγοριοποίηση ενός συνόλου δεδομένων.

Οι αλγόριθμοι ID3 (Mitchell, 1997) και C4.5 (Quinlan, 1993), βασίζονται σε μια στατιστική ιδιότητα, καλούμενη κέδρος πληροφορίας (Information Gain), προκειμένου να επιλέξουμε το γνώρισμα που θα ελέγξουμε σε κάθε κόμβο του δέντρου. Ο ορισμός του μέτρου βασίζεται στην εντροπία, η οποία χαρακτηρίζει την καθαρότητα μιας αφηρημένης επιλογής των δειγμάτων.

Εναλλακτικά οι αλγόριθμοι όπως ο SLIQ (Melta, 1996), SPRINT (Shafer,et.al., 1996), επιλέγουν το γνώρισμα που θα ελεγχθεί με βάση το δείκτη GINI και όχι το μέτρο εντροπίας. Το καλύτερο γνώρισμα για τον έλεγχο (δηλαδή το γνώρισμα που δίνει την καλύτερη τμηματοποίηση) δίνει τη χαμηλότερη τιμή για το δείκτη GINI.

Στη συνέχεια θα αναλυθούν οι κυριότεροι αλγόριθμοι των δέντρων αποφάσεων.

### **Αλγόριθμος ID3**

Έχουν προταθεί διάφοροι αλγόριθμοι για τη δημιουργία Δέντρων Αποφάσεων. Ένας από τους πιο διαδεδομένους είναι ο ID3, καθώς και οι μετεξελίξεις του, ο C4.5 και η εμπορική του εκδοχή C5.0. Ο ID3 προτάθηκε από τον Quinlan (1986) και υλοποιεί μια καθοδική (top-down) στρατηγική διαίρεσης. Ένα χαρακτηριστικό του είναι ότι απαιτεί την ύπαρξη μόνο ονομαστικών πεδίων. Η βασική ακολουθία βημάτων του ID3 είναι η εξής:

- ❖ Δημιουργείται ένας αρχικός κόμβος που αντιπροσωπεύει ολόκληρο το δείγμα.
- ❖ Εάν όλες οι παρατηρήσεις του δείγματος ανήκουν στην ίδια κλάση, τότε ο κόμβος μετατρέπεται σε φύλο.
- ❖ Διαφορετικά επιλέγεται το γνώρισμα που βέλτιστα διαχωρίζει τις παρατηρήσεις του δείγματος ανάλογα με την κλάση στην οποία ανήκουν.
- ❖ Δημιουργούνται κλάδοι που διαχωρίζουν τις παρατηρήσεις του δείγματος. Για κάθε δυνατή τιμή του γνωρίσματος δημιουργείται ένας κλάδος.
- ❖ Η διαδικασία επαναλαμβάνεται για κάθε ένα από τα υποσύνολα του δείγματος που δημιουργήθηκαν από τους κλάδους του προηγούμενου βήματος. Η επανάληψη τερματίζεται όταν ικανοποιηθεί τουλάχιστον μία από τις επόμενες συνθήκες εξόδου:
  - ✓ Όλες οι παρατηρήσεις ενός κόμβου ανήκουν στην ίδια κλάση.
  - ✓ Δεν υπάρχουν άλλα γνωρίσματα για τον διαχωρισμό του δείγματος. Σε αυτήν την περίπτωση ο κόμβος μετατρέπεται σε φύλο. Η απόφαση κατηγοριοποίησης στο φύλο είναι η κλάση που πλειοψηφεί στο συγκεκριμένο υποσύνολο παρατηρήσεων.

- ✓ Δεν υπάρχουν παρατηρήσεις που να ανήκουν στο υποσύνολο του δείγματος που ορίζει ο κλάδος.

Το βασικότερο πρόβλημα στα Δέντρα Αποφάσεων είναι ο καθορισμός του κριτηρίου, βάση του οποίου θα γίνει ο διαχωρισμός των παρατηρήσεων. Έχουν προταθεί αλγόριθμοι οι οποίοι πραγματοποιούν ελέγχους σε συνδυασμούς μεταβλητών (multivariate). Όμως στους περισσότερους αλγορίθμους, συμπεριλαμβανομένου του ID3, ο έλεγχος πραγματοποιείται σε μία μόνο μεταβλητή (univariate). Διάφοροι ερευνητές έχουν επινοήσει και προτείνει πλήθος μονομεταβλητών κριτηρίων. Σε ένα δεδομένο σημείο του δέντρου, το ερώτημα που τίθεται είναι ποιο γνώρισμα πρέπει να χρησιμοποιηθεί για τον διαχωρισμό των παρατηρήσεων. Στον ID3 το κριτήριο που χρησιμοποιείται ονομάζεται Κέρδος Πληροφορίας (ΚΠ). Για κάθε διαθέσιμο γνώρισμα υπολογίζεται το Κέρδος Πληροφορίας και επιλέγεται το γνώρισμα με τη μεγαλύτερη τιμή ΚΠ.

Το Κέρδος Πληροφορίας ( $S, A$ ) ( $\text{Information Gain}(S, A)$ ) εκφράζει τη μείωση της εντροπίας που θα προκύψει, εάν ένα σύνολο παρατηρήσεων  $S$  διαχωριστεί σε υποσύνολα με βάση τις τιμές του γνωρίσματος  $A$ . Η εντροπία μετρά την ανομοιογένεια του συνόλου  $S$ , ανάλογα με τη διασπορά των παρατηρήσεων ως προς την κλάση στην οποία ανήκουν.

#### **Αλγόριθμος C4.5**

Ο αλγόριθμος C4.5 αποτελεί επέκταση του ID3 και προτάθηκε από τον ίδιο ερευνητή (Quinlan, 1993). Μια από τις βασικές βελτιώσεις αφορά το κριτήριο διαχωρισμού. Σύμφωνα με τον Quinlan το Κέρδος Πληροφορίας τείνει να ευνοεί γνωρίσματα με μεγάλο πλήθος τιμών. Τα γνωρίσματα αυτά οδηγούν σε μεγάλο αριθμό μικρών και πολύ ομοιογενών υποσυνόλων. Σε πολλές περιπτώσεις όμως, τα γνωρίσματα αυτά δεν περιέχουν ουσιαστική πληροφορία. Αν για παράδειγμα τα δεδομένα περιέχουν πεδίο για κάποιον κωδικό, όπως ο αριθμός ταυτότητας, τότε το πεδίο αυτό θα έχει μεγάλο κέρδος πληροφορίας και θα επιλεγεί. Ωστόσο, δεν περιέχει πληροφορία χρήσιμη για την κατηγοριοποίηση. Για την αντιμετώπιση αυτού του προβλήματος, στον C4.5 χρησιμοποιείται το κριτήριο Λόγος Κέρδους (Gain Ratio), το οποίο ορίζεται με την εξίσωση:

$$\text{Gain Ratio}(S, A) = \text{Information Gain}(S, A) / \text{Entropy}(S, A)$$

Ο Λόγος Κέρδους κανονικοποιεί το κέρδος πληροφορίας ως προς την εντροπία. Μελέτες έχουν δείξει ότι ο Λόγος Κέρδους βελτιώνει την ακρίβεια και μειώνει την πολυπλοκότητα των δέντρων.

Μια άλλη σημαντική βελτίωση στον C4.5 είναι ότι, σε αντίθεση με τον ID3, μπορεί και χειρίζεται πεδία αριθμητικών τιμών. Για κάθε αριθμητικό πεδίο, ο αλγόριθμος ταξινομεί τις τιμές του, το πλήθος των οποίων είναι πεπερασμένο, σε αύξουσα σειρά, και ορίζει μια τιμή κατωφλίου. Με τον τρόπο αυτόν οι παρατηρήσεις χωρίζονται σε εκείνες των οποίων η τιμή στο συγκεκριμένο πεδίο είναι μικρότερη ή ίση με την τιμή κατωφλίου και σε εκείνες που η τιμή τους είναι μεγαλύτερη. Ακολούθως, το γνώρισμα αντιμετωπίζεται σαν να έχει διακριτές τιμές, όπου οι δύο διακριτές τιμές είναι οι δύο καθορισμένες περιοχές συνεχών τιμών. Επίσης ο C4.5 μπορεί και χειρίζεται δεδομένα με χαμένες τιμές.

#### **Αλγόριθμος C5.0**

Ο αλγόριθμος C5.0 δημιουργήθηκε επίσης από τον Quinlan (C5.0 for Unix/Linux, See5 for Windows) και αποτελεί μια εμπορική έκδοση του C4.5 και χρησιμοποιείται σε μεγάλα σύνολα δεδομένων. Διαφέρει με τον C4.5 στη δημιουργία κανόνων όπου περιλαμβάνει βελτιώσεις ενώ οι αλγόριθμοι που χρησιμοποιούνται δεν έχουν γίνει ακόμα γνωστοί. Ο C5.0 ανεβάζει το επίπεδο της χρήσης της μνήμης κατά 90% και είναι πιο γρήγορος από τον C4.5 κατά 5,7 έως 240 φορές και παράγει πιο ακριβείς κανόνες (Dunham, M., 2004).

#### **Αλγόριθμος CART**

Τα δέντρα τύπου CART (Classification And Regression Trees) προτάθηκαν από τους Breiman, Friedman, Olshen and Stone (1984). Ο πρωτότυπος αλγόριθμος τους είναι ενσωματωμένος σε λογισμικά της Salford Systems. Τα δέντρα CART παρουσιάζουν αρκετά ενδιαφέροντα χαρακτηριστικά. Ένα από τα σημαντικότερα χαρακτηριστικά τους είναι ότι μπορούν να χρησιμοποιηθούν και για κατηγοριοποίηση και για παλινδρόμηση, μπορούν δηλαδή να προβλέψουν και ονομαστικές τιμές κλάσης και τιμές αριθμητικών πεδίων. Τα δέντρα CART είναι δυαδικά και κάθε κόμβος μπορεί

να έχει μόνο δύο κλάδους. Για τον διαχωρισμό των παρατηρήσεων χρησιμοποιείται το κριτήριο Twoing. Στα πλεονεκτήματα τους περιλαμβάνονται οι υψηλές επιδόσεις σε ταχύτητα και ακρίβεια, καθώς και η ικανότητα τους να χειρίζονται δεδομένα με χαμένες τιμές. Επιπλέον, τα δέντρα CART είναι ικανά να εκτελέσουν κατηγοριοποίηση που λαμβάνει υπόψη το διαφορετικό κόστος σφάλματος (cost sensitive classification). Σε αυτήν την περίπτωση ο αλγόριθμος επιδιώκει να μειώσει τις εσφαλμένες προβλέψεις της πιο ακριβής κλάσης.

### **Αλγόριθμος CHAID**

Chi-Squared Automatic Interaction Detection (CHAID) είναι μια τεχνική του δέντρου απόφασης ή δέντρου παλινδρόμησης που προτάθηκε από τον Kass (1980). Είναι το καλύτερο εργαλείο που χρησιμοποιείται για να ανακαλύψει τη σχέση μεταξύ των μεταβλητών. Η ανάλυση CHAID καθορίζει πως οι μεταβλητές συνδυάζονται καλύτερα για να εξηγήσουν το αποτέλεσμα στις δεδομένες εξαρτημένες μεταβλητές. Χρησιμοποιεί κατηγορικά ή τακτικά δεδομένα. Η τεχνική CHAID μετατρέπει συνεχή δεδομένα σε τακτικά δεδομένα κατά τη διάρκεια της ανάλυσης. Η καλύτερη χρήση της ανάλυσης CHAID στους πίνακες έκτακτης ανάγκης είναι να αποφασιστεί ποια μεταβλητή έχει τη μέγιστη ανικανότητα στην κατάταξη. Έχει επίσης τη δυνατότητα να χτίσει μη-δυαδικά δέντρα ταξινόμησης. Αυτό είναι που πάνω από δύο κλάδοι μπορούν να περάσουν από τον κόμβο. Στην τεχνική CHAID, μπορούμε να δούμε οπτικά τη σχέση μεταξύ της μεταβλητής και του συνδεδεμένου σχετικού παράγοντα με ένα δέντρο (Nisbet, et.al., 2009).

Η τεχνική CHAID είναι καλύτερη από την τεχνική ανάλυσης παλινδρόμησης, διότι στην ανάλυση CHAID δεν απαιτείται κανονική κατανομή. Όπως στην ανάλυση συστάδων, στην ανάλυση CHAID, οι κατηγορίες των μεταβλητών συγχωνεύονται.

Οι συνιστώσες των δέντρων αποφάσεων στην ανάλυση CHAID περιγράφονται παρακάτω:

1. Ριζικός Κόμβος. Στην ανάλυση CHAID ριζικός κόμβος είναι η εξαρτημένη μεταβλητή ή η μεταβλητή στόχου. Για παράδειγμα, CHAID μπορεί να χρησιμοποιηθεί εάν η τράπεζα θέλει να προβλέψει τον κίνδυνο της πιστωτικής

κάρτας βασιζόμενη σε πληροφορίες όπως η ηλικία, το εισόδημα, ο αριθμός πιστωτικών καρτών κλπ., σε αυτό το παράδειγμα η εξαρτημένη μεταβλητή κινδύνου των πιστωτικών καρτών θα είναι ο ριζικός κόμβος.

2. Γονεακός Κόμβος. Ο αλγόριθμος Ehen χωρίζει τη μεταβλητή στόχο σε δύο ή περισσότερες κατηγορίες. Οι κατηγορίες αυτές ονομάζονται κόμβος γονέα ή αρχικός κόμβος. Για το παράδειγμα της τράπεζας, οι κατηγορίες υψηλή, μέση και χαμηλή είναι οι γονεακοί κόμβοι.

3. Κόμβος Παιδιού. Ανεξάρτητες μεταβλητές κατηγορίες που έρχονται κάτω από τις κατηγορίες του γονέα στο δέντρο ανάλυσης CHAID.

4. Τερματικός Κόμβος. Οι τελευταίες κατηγορίες του δέντρου ανάλυσης CHAID ονομάζονται τερματικοί κόμβοι. Στο δέντρο ανάλυσης CHAID, η κατηγορία που έχει σημαντική επίδραση στην εξαρτημένη μεταβλητή έρχεται πρώτη και η λιγότερο σημαντική κατηγορία έρχεται τελευταία. Ως εκ τούτου, ονομάζεται τερματικός κόμβος (Nisbet, et.al., 2009).

Πρόσθετοι αλγόριθμοι δημιουργίας δέντρων αποφάσεων έχουν προταθεί από διάφορους ερευνητές. Ορισμένα δέντρα αποφάσεων είναι τα AID (Sonquist, Baker & Morgan, 1971), και QUEST (Loh & Shih, 1997).

## ΚΕΦΑΛΑΙΟ 4. ΠΕΡΙΦΡΑΦΗ ΠΡΟΤΕΙΝΟΜΕΝΟΥ ΣΥΣΤΗΜΑΤΟΣ

---

### 4.1 Στόχος του Συστήματος

### 4.3 Σε ποιους εφαρμόζεται

### 4.4 Κατηγορίες Δεικτών

### 4.5 Κλίμακα Βαθμολόγησης

### 4.6 Δεδομένα Αξιολόγησης

---

#### 4.1 Στόχος του Συστήματος

Στο σύνθετο περιβάλλον της κρίσης που διανύουμε και το οποίο χαρακτηρίζεται από ραγδαίες εξελίξεις και υψηλή αβεβαιότητα, τόσο οι φορολογούμενοι όσο και οι επιχειρήσεις δέχονται έντονες πιέσεις με αποτέλεσμα η συμμόρφωσή τους ως προς τις φορολογικές τους υποχρεώσεις να παρουσιάζει έντονες διακυμάνσεις από έτος σε έτος.

Αξίζει να σημειωθεί ότι η ασυνέπεια των φορολογουμένων στις φορολογικές τους υποχρεώσεις, δηλαδή η μη «**συμμόρφωση**», έχει ως άμεση επίπτωση την ανατροπή των φορολογικών στόχων της Φορολογικής Αρχής. Αυτό με την σειρά του συνεπάγεται απώλεια εσόδων, μη επίτευξη των δημοσιονομικών στόχων, περιορισμό του κοινωνικού κράτους (υγεία, παιδεία, υποδομές), περαιτέρω συμπίεση των ασθενέστερων κοινωνικών ομάδων με την επιβολή σκληρότερων οριζόντιων μέτρων φορολόγησης προς αναπλήρωση των απωλειών και άνιση κατανομή των φορολογικών βαρών εις βάρος των συνεπών φορολογουμένων. Η διαμόρφωση μιας τέτοιας αλυσίδας, οδηγεί στην αδυναμία πλέον ανταπόκρισης και των συνεπών κατά πεποίθηση φορολογουμένων με αποτέλεσμα να ανατροφοδοτείται το πρόβλημα της μη συμμόρφωσης δημιουργώντας έτσι έναν φαύλο κύκλο ο οποίος απειλεί την βιωσιμότητα της ελληνικής οικονομίας.



Η Φορολογική Διοίκηση, όπως είναι φυσικό, καλείται να αντιδράσει σε αυτήν την απειλή αυξάνοντας την φορολογική συμμόρφωση. Για να το πετύχει αυτό χρειάζεται να εφαρμόσει μια πολιτική που θα αντιμετωπίζει τον κάθε φορολογούμενο με τον ενδεδειγμένο, σύμφωνα με την φορολογική του συμπεριφορά, τρόπο.

Κατά συνέπεια, χρειάζεται ο σχεδιασμός και η υιοθέτηση μιας ροής παρακολούθησης της συμμόρφωσης και ένα εργαλείο κατηγοριοποίησης των φορολογουμένων, ώστε να εφαρμοστεί ανά κατηγορία η ενδεδειγμένη πολιτική συμμόρφωσης, που εκτός από την επιβολή, θα μπορούσε να είναι η επιβράβευση ή και η έμπνευση για φορολογική συνέπεια.

Σε κάθε περίπτωση, χρειάζονται τεχνικές που θα βοηθήσουν στην άσκηση πολιτικής που θα αυξήσει την συμμόρφωση και την εμπιστοσύνη του πολίτη στο φορολογικό σύστημα.

Στα επόμενα κεφάλαια θα περιγραφεί το προτεινόμενο σύστημα αξιολόγησης, βαθμολόγησης και κατηγοριοποίησης φορολογουμένων βάσει της συμπεριφοράς τους με χρήση των μεθόδων εξόρυξης σε δεδομένα φορολογικού περιεχομένου έχοντας σαν πρότυπο το σύστημα βαθμολόγησης φορολογουμένων της Λετονίας. Συγκεκριμένα θα προσδιορισθούν:

- η ομάδα-στόχος των φορολογουμένων όπου δύναται να εφαρμοστεί το σύστημα,
- η προέλευση και η φύση των δεδομένων στα οποία θα στηριχθεί η αξιολόγηση,
- η κλίμακα βαθμολόγησης,
- τα κριτήρια αξιολόγησης,
- η μεθοδολογία που θα εφαρμοστεί,
- ο αλγόριθμος που θα εφαρμοστεί.
- Θα εφαρμοστούν τα παραπάνω και θα εξαχθούν συμπεράσματα.
- Στο τέλος θα προταθούν μελλοντικές βελτιώσεις.

## 4.2 Σε ποιους εφαρμόζεται

Αρχικά το σύστημα αξιολόγησης και βαθμολόγησης των φορολογουμένων θα δύναται να εφαρμοστεί σε:

- Φυσικά Πρόσωπα- επιτηδευματίες,
- Νομικά Πρόσωπα

και σε όσους έχουν κάνει έναρξη τουλάχιστον ένα έτος πριν, ώστε να υπάρχουν στοιχεία στα συστήματα «TAXIS» και «ELENXIS».

Επειδή σημαντικό μέρος των πληροφοριών προέρχεται από το υποσύστημα του Φ.Π.Α. οι προς αξιολόγηση φορολογούμενοι θα πρέπει να είναι υπόχρεοι Φ.Π.Α.

Στη συνέχεια μπορεί να επεκταθεί σε όλους τους φορολογούμενους.

## 4.3 Κατηγορίες Δεικτών

Το σύστημα βαθμολόγησης των φορολογουμένων περιλαμβάνει 4 διαστάσεις ανάλυσης που περιγράφονται από 12 δείκτες.

Οι τέσσερις αυτές διαστάσεις εστιάζουν:

- στην τήρηση ενός έγκυρου και αξιόπιστου Μητρώου στη βάση δεδομένων του TAXIS,
- στην ορθή και έγκαιρη υποβολή φορολογικών δηλώσεων,
- στην καταβολή φόρων που προκύπτουν από τις φορολογικές δηλώσεις ώστε να μην συσσωρεύονται ληξιπρόθεσμες οφειλές,
- στην παραβατικότητα που διαπιστώνεται κατά τη διενέργεια μερικών επιτόπιων ελέγχων.

Για να υπάρξει ακριβέστερη αξιολόγηση για τις διάφορες επιμέρους δραστηριότητες των φορολογουμένων, ορισμένες διαστάσεις περιλαμβάνουν περισσότερους από έναν δείκτες

| Διαστάσεις ανάλυσης                                                                                   | Δείκτες του συστήματος διαβάθμισης                                                                                                                                                                                                               |
|-------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Αξιολόγηση με βάση τα στοιχεία του Μητρώου                                                            | <ul style="list-style-type: none"> <li>▪ Αξιολόγηση των δεδομένων Μητρώου</li> <li>▪ Ενημέρωση Μητρώου</li> </ul>                                                                                                                                |
| Αξιολόγηση με βάση την υποβολή φορολογικών δηλώσεων                                                   | <ul style="list-style-type: none"> <li>▪ Μη υποβολή δηλώσεων</li> <li>▪ Υποβολή ανακριβών δηλώσεων</li> <li>▪ Υποβολή εκπροθέσμων δηλώσεων</li> <li>▪ Υποβολή τροποποιητικών δηλώσεων</li> </ul>                                                 |
| Αξιολόγηση με βάση τις βεβαιωμένες οφειλές, το ληξιπρόθεσμο χρέος και τις πληρωμές του φορολογούμενου | <ul style="list-style-type: none"> <li>▪ Συνολικό ληξιπρόθεσμο χρέος</li> <li>▪ Μεταβολή συνολικού ληξιπρόθεσμου χρέους</li> <li>▪ Ρυθμισμένες οφειλές</li> <li>▪ Συνέπεια διατήρησης ρύθμισης</li> <li>▪ Δυνατότητα εξόφλησης χρέους</li> </ul> |
| Αξιολόγηση με βάση την παραβατικότητα                                                                 | <ul style="list-style-type: none"> <li>▪ Αξιολόγηση των δεδομένων παραβάσεων</li> </ul>                                                                                                                                                          |

Πίνακας 1. Δείκτες του συστήματος διαβάθμισης

Ο κάθε φορολογούμενος θα κατατάσσεται σε μια βαθμίδα για κάθε διάσταση λαμβάνοντας υπόψη όλους τους δείκτες της διάστασης.

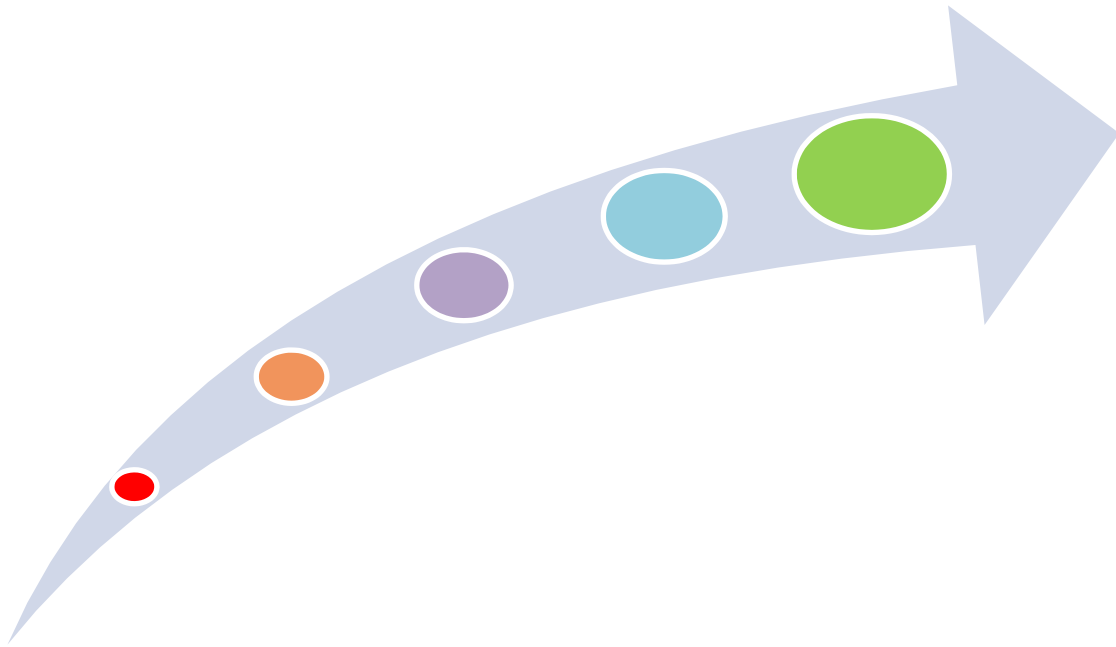
Τόσο οι διαστάσεις όσο και οι δείκτες του συστήματος διαβάθμισης μπορούν να επεκταθούν στο μέλλον.

#### 4.4 Κλίμακα Βαθμολόγησης

Η αυτοματοποιημένη αξιολόγηση και βαθμολόγηση των φορολογούμενων γίνεται βάσει κριτηρίων που θα αναπτυχθούν παρακάτω. Η μορφή που θα έχει το σύστημα βαθμολόγησης είναι η εξής:

Θα υπάρχουν πέντε βαθμίδες για την ερμηνεία των δεικτών του συστήματος αξιολόγησης των φορολογουμένων. Θεωρείται ιδανικό το πλήθος των βαθμίδων, διότι θα ήταν δύσκολο να περιγραφούν με σαφήνεια οι διαβαθμίσεις για μια κλίμακα με μεγαλύτερο αριθμό σημείων. Επίσης, μια μικρότερη κλίμακα δεν θα εξασφάλιζε τη δυνατότητα επαρκούς διαφοροποίησης των μοντέλων συμπεριφοράς των φορολογουμένων.

Οι πέντε βαθμίδες θα χαρακτηρίζουν με χρώματα τον βαθμό συμμόρφωσης των φορολογούμενων: στη βαθμίδα με χρώμα πράσινο θα κατατάσσονται οι φορολογούμενοι που έχουν συμμορφωθεί πλήρως ως προς τις υποχρεώσεις τους και έχουν την υψηλότερη βαθμολογία, ενώ στη βαθμίδα με χρώμα κόκκινο θα κατατάσσονται όσοι είναι μη συμμορφούμενοι και έχουν την χαμηλότερη βαθμολογία. Στις ενδιάμεσες κλίμακες χρωμάτων θα κατατάσσονται οι λοιπές κατηγορίες φορολογούμενων.



Εικόνα 2. Χρωματική κλίμακα συμμόρφωσης

Αυτές οι βαθμίδες θα χαρακτηρίζονται με αριθμούς από το 5 (πράσινη βαθμίδα) έως το 1 (κόκκινη βαθμίδα ) αντίστοιχα.

Επιγραμματικά το σύστημα βαθμολόγησης των φορολογουμένων προβλέπει τα εξής:

- ❖ «5» και «4» βαθμίδες: θετικές κατηγορίες με συνεπείς φορολογούμενους.
- ❖ «3» βαθμίδα: ουδέτερη κατηγορία που δηλώνει προβλήματα περιστασιακής φύσης.
- ❖ «2» και «1» βαθμίδες: αρνητικές κατηγορίες με ασυνεπείς και ύποπτους για παραβατική συμπεριφορά φορολογούμενους.

| Βαθμοί | Περιγραφή                                                                                                                            |
|--------|--------------------------------------------------------------------------------------------------------------------------------------|
| 5      | Η τιμή του δείκτη υποδηλώνει την πλήρωση όλων των φορολογικών υποχρεώσεων και την πλήρη συμμόρφωση του φορολογούμενου.               |
| 4      | Η τιμή του δείκτη υποδηλώνει την καλή πειθαρχία στην εκπλήρωση των φορολογικών υποχρεώσεων.                                          |
| 3      | Η τιμή του δείκτη υποδηλώνει περιοδικά προβλήματα με την πειθαρχία στην εκπλήρωση φορολογικών υποχρεώσεων.                           |
| 2      | Η τιμή του δείκτη, οποία υποδηλώνει κακή πειθαρχία στην εκπλήρωση των φορολογικών υποχρεώσεων.                                       |
| 1      | Η τιμή του δείκτη υποδηλώνει κακή πειθαρχία στην εκπλήρωση των φορολογικών υποχρεώσεων και υψηλό κίνδυνο οικονομικών δραστηριοτήτων. |

Πίνακας 4.2 Βαθμίδες αξιολόγησης

#### 4.5 Δεδομένα Αξιολόγησης

Για την αυτοματοποιημένη αξιολόγηση και βαθμολόγηση φορολογουμένων θα πρέπει να τεθούν κριτήρια αξιολόγησης σύμφωνα με τα οποία θα βαθμολογούνται τα Φυσικά και Νομικά πρόσωπα ανάλογα με τη συμπεριφορά τους και το βαθμό συμμόρφωσής τους στις φορολογικές τους υποχρεώσεις. Οι φορολογικές υποχρεώσεις που καλούνται να εκπληρώσουν οι φορολογούμενοι και των οποίων η μη εκπλήρωση συνιστά την έννοια της «μη συμμόρφωσης», είναι οι ακόλουθες:

- η ορθή εγγραφή και η έγκαιρη ενημέρωση των απαιτούμενων προσωπικών και επαγγελματικών δεδομένων στο Φορολογικό Μητρώο, ώστε να είναι δυνατή η ορθή επιβολή των φόρων που ορίζει η νομοθεσία,
- η έγκαιρη υποβολή των φορολογικών δηλώσεων,
- η έγκαιρη πληρωμή των φόρων που έχουν βεβαιωθεί,
- η ορθή και ακριβής υποβολή δηλώσεων.

Για την υλοποίηση της αξιολόγησης ως προς τις παραπάνω υποχρεώσεις, θα δημιουργηθούν κριτήρια με βάση πληροφορίες που αφορούν την εικόνα του φορολογούμενου ως προς:

- τα στοιχεία Μητρώου,
- την υποβολή δηλώσεων,
- τις βεβαιώσεις φόρου, τις πληρωμές και τις ληξιπρόθεσμες οφειλές,
- τις καταλογισμένες παραβάσεις.

Στην συνέχεια θα αναπτυχθούν τα κριτήρια για την αξιολόγηση και βαθμολόγηση των Φυσικών και Νομικών Προσώπων.

## ΚΕΦΑΛΑΙΟ 5. ΔΕΙΚΤΕΣ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ

---

### 5.1 Αξιολόγηση με βάση τα στοιχεία του Μητρώου

### 5.2 Αξιολόγηση με βάση την υποβολή των Δηλώσεων

### 5.3 Αξιολόγηση με βάση τις Βεβαιώσεις, Πληρωμές και Ληξιπρόθεσμες Οφειλές

### 5.4 Αξιολόγηση με βάση τις παραβάσεις

---

## 5.1 Αξιολόγηση με βάση τα στοιχεία του Μητρώου

### 5.1.1 Δείκτης 1: Καταχωρημένων Στοιχείων Μητρώου

Οι πληροφορίες που προκύπτουν από το Μητρώο TAXIS είναι το πρωταρχικό κριτήριο που χρησιμοποιείται για την βαθμολόγηση των φορολογούμενων.

Αρχικά θα εξαιρούνται από τη βαθμολόγηση όσοι φορολογούμενοι δεν έχουν συμπληρώσει το 18ο έτος της ηλικίας τους ή έχουν αποβιώσει.

Στη συνέχεια, τόσο για τα Φυσικά όσο και για τα Νομικά Πρόσωπα θα χρησιμοποιείται από το Μητρώο TAXIS η ένδειξη της «Πτώχευσης» και όταν αυτή υπάρχει θα αποδίδεται η χαμηλότερη βαθμολογία στο φορολογούμενο.

Επιπλέον για τα Νομικά και τα Φυσικά Πρόσωπα επιτηδευματίες θα λαμβάνονται υπόψη οι παρακάτω ενδείξεις:

- **«Εξαφανισμένος έμπορος»** : Η ένδειξη αυτή, δεδομένου ότι προκύπτει κατόπιν ελέγχου VIES, θα πρέπει να χαρακτηρίζει απευθείας τον φορολογούμενο ως μη συμμορφούμενο. Υποδηλώνει φορολογούμενο που εμπλέκεται σε φορολογικές απάτες (κυρίως μέσω τριγωνικών συναλλαγών) και για τον λόγο αυτό θα λαμβάνει την ελάχιστη βαθμολογία «1».
- **«Αναστολή χρήσης ΑΦΜ»** : Η ένδειξη αυτή θα χαρακτηρίζει απευθείας τον φορολογούμενο ως μη συμμορφούμενο. Η αναστολή χρήσης Α.Φ.Μ. πραγματοποιείται όταν υφίστανται στοιχεία για παραβατική συμπεριφορά



και φοροδιαφυγή και για τον λόγο αυτό θα λαμβάνει την ελάχιστη βαθμολογία «1».

- **«Απενεργοποιημένο ΑΦΜ» :** Η ένδειξη αυτή θα χαρακτηρίζει απευθείας τον φορολογούμενο ως μη συμμορφούμενο. Η απενεργοποίηση Α.Φ.Μ. πραγματοποιείται όταν υπάρχουν διπλά Α.Φ.Μ. ή όταν βρεθούν πλαστά στοιχεία κατόπιν ελέγχου. Πρόκειται για παραβατική συμπεριφορά και φοροδιαφυγή και για τον λόγο αυτό θα λαμβάνει την ελάχιστη βαθμολογία «1».

Στις περιπτώσεις αυτές η βαθμολογία αυτή θα είναι και η τελική. Ο αλγόριθμος δεν θα προχωρά από εκεί και κάτω καθώς οι λοιπές πληροφορίες που θα εξεταστούν είναι δευτερευούσης σημασίας σε σχέση με τους παραπάνω χαρακτηρισμούς.

Επίσης αξιολογείται το **Πλήθος διακοπών**. Η ένδειξη αυτή θα αφορά μόνο τους επιτηδευματίες εκείνους που δεν ασκούν εποχιακό επάγγελμα. Συνήθως μεγάλο πλήθος διακοπών και επανενάρξεων με ομοειδή δραστηριότητα υποδηλώνει έντονη τάση για παραβατικότητα και προσπάθεια απόκρυψης εισοδημάτων ή εκμετάλλευση των φορολογικών ωφελειών που κατά καιρούς προσφέρονται στις νέες επιχειρήσεις.

**Στόχος του δείκτη** - να απαγορεύσει τις οικονομικές δραστηριότητες φορολογουμένων με παραβατική εικόνα στο Μητρώο Φορολογουμένων. Επίσης να εμποδίσει τις επαναλαμβανόμενες διακοπές εργασιών και ενάρξεις που κρύβουν ύποπτη συμπεριφορά.

**Πηγή πληροφοριών για τη μέτρηση του δείκτη** – Μητρώο Φορολογούμενων TAXIS.

**Προσδιορισμός του δείκτη** - ο δείκτης προσδιορίζεται από τις τιμές των παραπάνω ενδείξεων καθώς και το πλήθος διακοπών εργασιών που έχουν γίνει τα τελευταία 3 χρόνια.

**Χορήγηση βαθμών** – βαθμοί χορηγούνται με βάση τις ενδείξεις καθώς και το πλήθος διακοπών εργασιών.

"1" βαθμός χορηγείται στο φορολογούμενο με τουλάχιστον μία από τις παραπάνω ενδείξεις ενεργή.

"5" βαθμοί χορηγούνται στον φορολογούμενο που δεν έχει καμιά από τις παραπάνω ενδείξεις ενεργή και δεν έχει κάνει καμιά διακοπή εργασιών.

Οι βαθμοί από 2 έως 4 αποδίδονται αναλογικά σε όσους δεν έχουν καμιά από τις παραπάνω ενδείξεις ενεργή αλλά έχουν κάνει διακοπές εργασιών και ενάρξεις με την ίδια δραστηριότητα.

### 5.1.2 Δείκτης 2: Ενημέρωσης Μητρώου

Τα προσωπικά και επαγγελματικά δεδομένα του Μητρώου προσδιορίζουν τις υποχρεώσεις του φορολογούμενου ως προς την επιβολή φόρων. Είναι σημαντικό τα δεδομένα αυτά να είναι σωστά και επικαιροποιημένα. Γι αυτό πρέπει η οποιαδήποτε μεταβολή να δηλώνεται από τον φορολογούμενο χωρίς καθυστέρηση.

Με τον ορισμό του επόμενου δείκτη ελέγχεται αν ο φορολογούμενος έχει ενημερώσει εμπρόθεσμα την Φορολογική Αρχή για μεταβολές που τον αφορούν όπως:

- Μεταβολή Ατομικών Στοιχείων  
(π.χ. Αλλαγή Α.Δ.Τ., Αλλαγή διεύθυνσης, Αλλαγή υπηκοότητας, Αλλαγή στοιχείων επικοινωνίας, δήλωση γάμου κλπ.)
- Έναρξη Εργασιών
- Μεταβολή Εργασιών  
(π.χ. για Φυσικά Πρόσωπα : Αλλαγή Έδρας, Αλλαγή Δραστηριότητας, Αλλαγή Κατηγορίας Βιβλίων, Δήλωση Έναρξης Ενδοκοινοτικών Συναλλαγών κλπ),
- Διακοπή Εργασιών
- Εγκαταστάσεις Υποκαταστήματος Εσωτερικού
- Εγκαταστάσεις Υποκαταστήματος Εξωτερικού

- Έναρξη πωλήσεων από απόσταση κλπ.

Οι παραπάνω μεταβολές δεν έχουν όλες την ίδια σπουδαιότητα. Κάποιες από αυτές είναι απλά ενημερωτικές ενώ άλλες καθορίζουν τις υποχρεώσεις του φορολογούμενου. Γι αυτό ορίζεται συντελεστής βαρύτητας για κάθε αιτία μεταβολής.

**Στόχος του δείκτη** - να παρακινήσει τον φορολογούμενο να ενημερώνει την εικόνα του στο Μητρώο Φορολογουμένων άμεσα.

**Προσδιορισμός του δείκτη** - ο δείκτης προσδιορίζεται με την αξιολόγηση:

- ✓ της αιτίας μεταβολής που καθυστερημένα δήλωσε ο φορολογούμενος στο Μητρώο,
- ✓ του χρόνου καθυστέρησης ενημέρωσης του Μητρώου

για το χρονικό διάστημα των τελευταίων 36 μηνών.

Συγκεκριμένα ορίζεται ως το άθροισμα των επιμέρους γινομένων του συντελεστή βαρύτητας κάθε αιτίας μεταβολής επί του χρόνου καθυστέρησης ενημέρωσης του Μητρώου.

**Πηγή πληροφοριών για τη μέτρηση του δείκτη** – Μητρώο Φορολογούμενων TAXIS.

**Χορήγηση βαθμών** - βαθμοί χορηγούνται με βάση τον παραπάνω δείκτη.

"5" βαθμοί χορηγούνται στον φορολογούμενο που ενημερώνει εγκαίρως το Μητρώο με τις μεταβολές που τον αφορούν.

"4" βαθμοί χορηγούνται στον φορολογούμενο που έχει ενημερώσει το Μητρώο με μικρή καθυστέρηση.

"1" βαθμός χορηγείται στον φορολογούμενο που καθυστερεί σημαντικά ειδικά για μεταβολές μεγάλης βαρύτητας.

## 5.2 Αξιολόγηση με βάση την υποβολή των Δηλώσεων

Θα οριστούν τέσσερις δείκτες.

- ✓ Μη Υποβολής Δηλώσεων
- ✓ Ανακριβούς Υποβολής Δηλώσεων
- ✓ Εκπρόθεσμης Υποβολής Δηλώσεων
- ✓ Υποβολής Τροποποιητικών Δηλώσεων

Η συνολική βαθμολογία του φορολογούμενου για την διάσταση θα είναι η ελάχιστη βαθμολογία που θα πάρει σε τουλάχιστον ένα δείκτη.

### 5.2.1 Δείκτης 3: Μη υποβολής δηλώσεων

Ελέγχεται αν οι φορολογούμενοι έχουν υποβάλει όλες τις φορολογικές δηλώσεις για τις οποίες είναι υπόχρεοι καθώς η μη υποβολή δήλωσης έχει ως συνέπεια την μη είσπραξη φόρων.

Οι φορολογικές περιοχές που θα εξεταστούν για μη υποβολή δήλωσης στο χρονικό διάστημα των τελευταίων 36 μηνών είναι:

1. Φόρος Εισοδήματος (Φ.Ε.)
2. Φόρος Προστιθέμενης Αξίας (Φ.Π.Α)
3. Ανακεφαλαιωτικοί πίνακες VIES
4. Δήλωση Ε9
5. Φόρος Μισθωτών Υπηρεσιών (Φ.Μ.Υ.)
6. Καταστάσεις Πελατών-Προμηθευτών

#### 1. Φόρος Εισοδήματος

Η μη υποβολή δηλώσεων εισοδήματος θα αποτελεί βασικό κριτήριο για την αξιολόγηση και βαθμολόγηση των φορολογουμένων. Ειδικότερα, θα εξετάζεται το πλήθος των δηλώσεων που δεν έχουν υποβληθεί τα τελευταία τρία χρόνια.

Για τον εντοπισμό των υπόχρεων σε υποβολή δήλωσης Φυσικών Προσώπων και την αξιολόγηση της σημαντικότητας της μη υποβολής θα δίνεται ιδιαίτερη βαρύτητα στις παρακάτω πηγές εισοδήματος (πραγματικές και τεκμαρτές):

Μισθοί: Η πληροφορία αυτή θεωρείται «βέβαιη», καθώς προέρχεται από στοιχεία που προέκυψαν από την επεξεργασία των ηλεκτρονικά υποβληθεισών βεβαιώσεων αποδοχών ή συντάξεων από εργοδότες και ασφαλιστικούς φορείς. Στην ίδια κατηγορία, εκτός από τους μισθούς, τα ημερομίσθια και τις κύριες συντάξεις, εντάσσεται και το άθροισμα των καθαρών ποσών από επικουρικές συντάξεις, οι αμοιβές μελών ΔΣ καθώς και οι καθαρές αποδοχές των αξιωματικών και του κατώτερου πληρώματος του εμπορικού ναυτικού.

Μερίσματα και τόκοι ημεδαπής και αλλοδαπής προέλευσης: Η ύπαρξη υψηλών ποσών μερισμάτων και τόκων υποδεικνύει την ύπαρξη καταθέσεων και γενικότερα ρευστών περιουσιακών στοιχείων. Οι πληροφορίες αυτές θα αξιολογούνται με αυξανόμενη κλίμακα βάση της οποίας θα γίνει η περαιτέρω κατάταξη των μη συμμορφούμενων.

Εισόδημα από ακίνητα: Επιπλέον βαρύτητα θα δίνεται στην πληροφορία των φορολογουμένων που δηλώνονται ως εκμισθωτές στις υποβληθείσες δηλώσεις εισοδήματος τρίτων φυσικών ή νομικών προσώπων. Στην κατηγορία αυτή εντάσσεται το εισόδημα από ενοίκια όπως προκύπτει από τις υποβληθείσες δηλώσεις τρίτων.

Η πληροφορία που παρέχεται από αυτή την πηγή υποδηλώνει ότι οι εν λόγω φορολογούμενοι απέκτησαν εντός του εξεταζόμενου έτους εισόδημα από ακίνητα και συνεπώς είναι υπόχρεοι σε υποβολή δήλωσης Φορολογίας Εισοδήματος. Το ύψος των εισοδημάτων αυτών θα επηρεάζει την κατάταξη και την αξιολόγηση των μη συμμορφούμενων.

Δαπάνες: Θα αξιολογείται η δαπάνη που καταβλήθηκε για την τοκοχρεωλυτική απόσβεση δανείων οποιασδήποτε μορφής. Η πληροφορία αυτή αντλείται από στοιχεία τρίτων πηγών που έχει στη διάθεσή της η Φορολογική Αρχή.

Κατοχή οχημάτων: Η πληροφορία αυτή θα αντλείται από το Πληροφοριακό Σύστημα Οχημάτων και πληρωμής τελών κυκλοφορίας

Τα κριτήρια που αναλύονται στη συνέχεια αφορούν τόσο τα Φυσικά όσο και τα Νομικά Πρόσωπα προκειμένου να χρησιμοποιηθούν για την αξιολόγησή τους ως συμμορφούμενα ή μη.

Παρακράτηση φόρων: Για τον εντοπισμό των υπόχρεων σε υποβολή δήλωσης Φορολογίας Εισοδήματος, άρα και τη συμμόρφωσή τους, θα πρέπει να αξιολογούνται τα ποσά των προ-συμπληρωμένων κωδικών που αφορούν φόρους που παρακρατήθηκαν στα εισοδήματα από επιχειρηματική δραστηριότητα.

Έναρξη εργασιών: Όσοι έχουν προβεί σε έναρξη εργασιών, σύμφωνα με τον ν. 4172/2013, είναι υπόχρεοι σε υποβολή δήλωσης Φορολογίας Εισοδήματος και για το λόγο αυτό η έναρξη επιτηδεύματος θα επηρεάζει την κατάταξη και την αξιολόγηση των μη συμμορφούμενων.

## 2. Φόρος Προστιθέμενης Αξίας

Στην αξιολόγηση των Νομικών και των Φυσικών Προσώπων επιτηδευματιών για την μη υποβολή δήλωσης ΦΠΑ θα πρέπει να λαμβάνεται υπόψη τόσο το πλήθος των δηλώσεων που δεν έχουν υποβληθεί ενώ υπάρχει σχετική υποχρέωση, όσο και το σύνολο των εκρών που θα περιλαμβάνονταν σε αυτές. Το σύνολο των εκρών θα υπολογίζεται προσεγγιστικά βάσει των εκρών που έχουν περιληφθεί στις δηλώσεις ΦΠΑ ή τις καταστάσεις πελατών – προμηθευτών που έχουν υποβληθεί. Αν δεν έχει υποβληθεί καμία δήλωση ΦΠΑ κατά το τρέχον έτος, το σύνολο των εκρών θα λαμβάνεται από τις καταστάσεις πελατών – προμηθευτών ή αλλιώς προσεγγιστικά από τις τελευταίες υποβληθείσες δηλώσεις. Σε περίπτωση που δεν έχει υποβληθεί καμία δήλωση και κατά τα προηγούμενα έτη, το σύνολο των εκρών θα λαμβάνεται από τις καταστάσεις πελατών – προμηθευτών ή αλλιώς η αξιολόγηση θα γίνεται αποκλειστικά από το πλήθος των μη υποβολών.

## 3. Ανακεφαλαιωτικοί πίνακες VIES

Κριτήριο αξιολόγησης των Νομικών και των Φυσικών Προσώπων επιτηδευματιών θα αποτελέσει το πλήθος των μη υποβαλλόμενων δηλώσεων και το ύψος του ποσού των ενδοκοινοτικών αποκτήσεων. Το ύψος αυτών θα υπολογίζεται από τις διαφορές που προκύπτουν με τα στοιχεία που δηλώνουν οι άλλες Ευρωπαϊκές χώρες ως παραδόσεις στο σύστημα Vies.

#### 4. Δήλωση Ε9

Θα αξιολογείται το πλήθος των δηλώσεων Ε9, των Νομικών και των Φυσικών Προσώπων, που δεν έχουν υποβληθεί εφόσον υπάρχει σχετική υποχρέωση, δηλαδή όταν υπάρχει μεταβολή της ακίνητης περιουσίας. Μεγάλη βαρύτητα στην περαιτέρω αξιολόγηση των μη συμμορφούμενων θα έχει το ύψος της αξίας των φορολογητέων ακινήτων.

#### 5. Δήλωση Φ.Μ.Υ

Θα αξιολογείται το πλήθος των μη υποβαλλόμενων δηλώσεων Φ.Μ.Υ., των Νομικών και των Φυσικών Προσώπων επιτηδευματιών και με αυτόν τον τρόπο θα γίνεται η τελική κατάταξη των φορολογούμενων σε βαθμίδες μη συμμορφούμενων. Η μη υποβολή δήλωσης έχει ως συνέπεια την μη είσπραξη φόρων.

Δεδομένου ότι δεν υποχρεούνται όλοι οι επιτηδευματίες σε μηνιαία υποβολή δήλωσης Φ.Μ.Υ. και στα μηχανογραφικά συστήματα της Φορολογικής Διοίκησης δεν υπάρχει η πληροφορία της απασχόλησης ή μη προσωπικού, για τον εντοπισμό των υπόχρεων στην υποβολή της εν λόγω δήλωσης κρίνεται απαραίτητο να πραγματοποιηθούν οι διασταυρώσεις σε όσους:

- υπέβαλαν Αναλυτική Περιοδική Δήλωση (Α.Π.Δ.) των εργαζομένων στο Ίδρυμα Κοινωνικών Ασφαλίσεων (Ι.Κ.Α.) και δεν υπέβαλαν ως όφειλαν δήλωση Φ.Μ.Υ..
- υπέβαλαν ηλεκτρονικά στο Σύστημα «ΕΡΓΑΝΗ» αναλυτικά στοιχεία που αφορούν το προσωπικό τους, όπως αναγγελία πρόσληψης προσωπικού, αναγγελία έναρξης/μεταβολών απασχόλησης, πίνακα προσωπικού,

αναγγελία οικειοθελούς αποχώρησης μισθωτού, καταγγελία σύμβασης εργασίας αορίστου χρόνου και δεν υπέβαλαν ως όφειλαν δήλωση Φ.Μ.Υ..

- έχουν δηλωθεί ως εργοδότες από μισθωτούς, ενώ οι ίδιοι δεν απέστειλαν ως όφειλαν ηλεκτρονικά το αρχείο των ετήσιων αποδοχών των εργαζομένων ή δεν υπέβαλλαν περιοδική δήλωση Φ.Μ.Υ..

#### 6. Καταστάσεις Πελατών-Προμηθευτών

Στην περίπτωση των Νομικών και των Φυσικών Προσώπων επιτηδευματιών θα αποτελέσουν κριτήριο αξιολόγησης οι καταστάσεις Πελατών – Προμηθευτών. Συγκεκριμένα για την αξιολόγηση θα ληφθεί υπόψη τόσο η μη υποβολή των καταστάσεων ενώ υπάρχει σχετική υποχρέωση, όσο και οι διαφορές που πιθανόν θα παρουσιάζονται στις καταστάσεις των υπόχρεων με τους αντισυμβαλλόμενους.

**Στόχος του δείκτη** - να παρακινήσει τον φορολογούμενο να είναι συνεπής στην υποχρέωση υποβολής δήλωσης.

**Προσδιορισμός του δείκτη** - ο δείκτης προσδιορίζεται με την αξιολόγηση του αριθμού των φορολογικών δηλώσεων που δεν υποβλήθηκαν τα τελευταία 3 χρόνια.

**Πηγή πληροφοριών για τη μέτρηση του δείκτη** – Υποσυστήματα Φ.Ε., Φ.Π.Α και Κ.Β.Σ του TAXIS.

**Χορήγηση βαθμών** - βαθμοί χορηγούνται με βάση τον αριθμό των δηλώσεων που δεν υποβλήθηκαν.

"5" βαθμοί χορηγούνται στον φορολογούμενο που έχει υποβάλει όλες τις φορολογικές δηλώσεις.

#### **5.2.2 Δείκτης 4: Ανακριβούς υποβολής δηλώσεων**

Η ανακριβής υποβολή δήλωσης των Νομικών και των Φυσικών Προσώπων προκύπτει κατόπιν διενέργειας ελέγχου και επιβολής προστίμου ανακρίβειας. Ο



εντοπισμός της ανακριβούς δήλωσης υποδηλώνει παραβατική συμπεριφορά και συνεπάγεται απώλεια φορολογικών εσόδων. Για την κατηγοριοποίηση των μη συμμορφούμενων θα πρέπει να λαμβάνεται υπόψη τόσο το πλήθος των προστίμων ανακρίβειας που έχουν επιβληθεί κατόπιν ελέγχου όσο και το συνολικό ύψος των φόρων που έχουν επιβληθεί λόγω ανακρίβειας.

**Στόχος του δείκτη** - να παρακινήσει τον φορολογούμενο να είναι ακριβής στην συμπλήρωση της δήλωσης και κατά συνέπεια στα ποσά φόρου που οφείλει.

**Προσδιορισμός του δείκτη** - ο δείκτης προσδιορίζεται με την αξιολόγηση:

- ✓ του πλήθους των προστίμων ανακρίβειας που έχουν επιβληθεί κατόπιν ελέγχου
- ✓ και του συνολικού ύψους των φόρων που έχουν επιβληθεί λόγω ανακρίβειας

των τελευταίων τριών ετών.

**Πηγή πληροφοριών για τη μέτρηση του δείκτη** – Υποσυστήματα Φ.Ε., Φ.Π.Α και Κ.Β.Σ του TAXIS.

**Χορήγηση βαθμών** - βαθμοί χορηγούνται με βάση τον παραπάνω δείκτη.

Όσα περισσότερα και μεγαλύτερου ύψους φόροι και πρόστιμα ανακρίβειας έχουν βεβαιωθεί, τόσο λιγότερο συμμορφούμενος θα θεωρείται ο φορολογούμενος.

"5" βαθμοί χορηγούνται στον φορολογούμενο που έχει υποβάλει όλες τις φορολογικές δηλώσεις με ακρίβεια.

### 5.2.3 Δείκτης 5: Υποβολής εκπροθέσμων δηλώσεων

Ελέγχεται αν οι φορολογούμενοι έχουν υποβάλει εμπρόθεσμα τις φορολογικές τους δηλώσεις οι οποίες έχουν καταχωρηθεί στο μηχανογραφικό σύστημα «TAXIS». Η εκπρόθεσμη υποβολή δήλωσης μπορεί να εντοπιστεί:

- από την ίδια τη δήλωση όπου αναγράφονται τόσο η ημερομηνία υποβολής όσο και η ημερομηνία λήξης φορολογικής περιόδου,
- από το πρόστιμο που επιβάλλεται κατά την υποβολή των δηλώσεων στις οποίες ο φορολογούμενος είναι υπόχρεος.

Για εκείνους που κατ' εξακολούθηση υποβάλουν εκπρόθεσμα δηλώσεις υπάρχει υπόνοια παραβατικής συμπεριφοράς.

Θα πρέπει να εξετάζεται το συνολικό πλήθος των προστίμων εκπρόθεσμης υποβολής από όλες τις φορολογικές περιοχές καθώς και το αν προκύπτει φόρος προς απόδοση.

Συγκεκριμένα θα αξιολογηθούν οι εξής περιπτώσεις:

#### 1. Εκπρόθεσμη υποβολή δήλωσης με χρεωστικό ποσό.

Διακρίνουμε δύο περιπτώσεις:

- ✓ η εκπρόθεσμη δήλωση είναι χρεωστική και έχουν εξοφληθεί εμπρόθεσμα οι δόσεις του φόρου. Σε αυτήν την περίπτωση θα αξιολογηθεί μόνο η εκπρόθεσμη υποβολή.
- ✓ η εκπρόθεσμη δήλωση είναι χρεωστική και δεν έχουν εξοφληθεί εμπρόθεσμα οι βεβαιωθείσες δόσεις του φόρου. Σε αυτήν την περίπτωση απαιτείται περεταίρω αξιολόγηση αφενός λόγω της εκπρόθεσμης υποβολής της δήλωσης και αφετέρου λόγω των ληξιπρόθεσμων οφειλών από τον βεβαιωμένο φόρο.

#### 2. Εκπρόθεσμη υποβολή δήλωσης με πιστωτικό ποσό.

Στην περίπτωση εκπρόθεσμης υποβολής δήλωσης που το αποτέλεσμα είναι πιστωτικό δεν θα χρήζει περαιτέρω ανάλυσης, ωστόσο θα αξιολογηθεί ως προς την εκπρόθεσμη υποβολή.

#### 3. Εκπρόθεσμη υποβολή δήλωσης με μηδενικό φόρο.

Στην περίπτωση εκπρόθεσμα υποβληθείσας μηδενικής δήλωσης, δεν θα χρήζει περεταίρω ανάλυσης, ωστόσο θα αξιολογηθεί ως προς την εκπρόθεσμη υποβολή.

Σε κάθε περίπτωση θα πρέπει να αξιολογείται η εκπρόθεσμη υποβολή δήλωσης με μεγαλύτερη βαρύτητα όταν προκύπτει χρεωστικό υπόλοιπο καθώς και όταν κρίνεται μεγάλος ο αριθμός των εκπροθέσμων δηλώσεων.

**Στόχος του δείκτη** - να παρακινήσει τον φορολογούμενο να υποβάλλει εγκαίρως τις φορολογικές του δηλώσεις.

**Προσδιορισμός του δείκτη** - ο δείκτης προσδιορίζεται με την αξιολόγηση:

- ✓ του πλήθους εκπροθέσμων φορολογικών δηλώσεων,
- ✓ του συνολικού χρόνου καθυστέρησης υποβολής των εκπρόθεσμων δηλώσεων
- ✓ καθώς και του ύψους ποσού επιπρόσθετου φόρου που προκύπτει από τις εκπρόθεσμες δηλώσεις

που υποβλήθηκαν τα τελευταία 3 χρόνια.

**Πηγή πληροφοριών για τη μέτρηση του δείκτη** – Υποσυστήματα Φ.Π.Α και Φ.Ε του TAXIS.

**Χορήγηση βαθμών** - βαθμοί χορηγούνται με βάση τη τιμή του παραπάνω δείκτη.

"5" βαθμοί χορηγούνται στον φορολογούμενο που έχει υποβάλει εγκαίρως όλες τις φορολογικές δηλώσεις.

#### 5.2.4 Δείκτης 6: Υποβολής τροποποιητικών δηλώσεων

Η υποβολή τροποποιητικής δήλωσης αποτελεί υπόνοια για παραβατική συμπεριφορά ειδικά όταν υποβάλλεται με μεγάλη καθυστέρηση. Για παράδειγμα, υποβολή αρχικής εμπρόθεσμης δήλωσης με λανθασμένο ποσό (μικρό χρεωστικό ή μεγάλο πιστωτικό) και στη συνέχεια διόρθωση του ποσού με υποβολή

τροποποιητικής δήλωσης (με μεγαλύτερο χρεωστικό ή μικρότερο πιστωτικό ποσό σε σχέση με το ποσό της αρχικής) υποκρύπτει δόλο με σκοπό τον “έμμεσο δανεισμό”.

Στην συγκεκριμένη περίπτωση θα λαμβάνεται υπόψη το πλήθος των τροποποιητικών δηλώσεων σε συνδυασμό με τους λόγους υποβολής αυτών.

Για παράδειγμα:

- Στην Δήλωση Φορολογίας Εισοδήματος η υποβολή τροποποιητικών δηλώσεων σε περίπτωση αναδρομικών συντάξεων δεν οφείλεται σε υπαιτιότητα του φορολογούμενου. Συνεπώς στην περίπτωση αυτή δεν θα πρέπει να λαμβάνεται υπόψη ως κριτήριο μη συμμόρφωσης. Αντίθετα σε περίπτωση που γίνει τροποποίηση για ακίνητο, που ενώ ήταν στην κατοχή του υπόχρεου, δεν περιλαμβανόταν στην αρχική δήλωση τότε θεωρείται ότι υπάρχει παραβατική συμπεριφορά.
- Στην δήλωση Φ.Π.Α. η υποβολή μιας τροποποιητικής δήλωσης ανά υποβαλλόμενη περίοδο που αφορά το πιστωτικό υπόλοιπο λόγω επιστροφής Φ.Π.Α. δεν οφείλεται σε υπαιτιότητα του φορολογούμενου και συνεπώς δεν θα πρέπει να λαμβάνεται υπόψη ως κριτήριο μη συμμόρφωσης. Αντίθετα στην περίπτωση υποβολής τροποποιητικών δηλώσεων που οδηγούν σε μείωση του φόρου θεωρείται ότι υπάρχει παραβατική συμπεριφορά. Συνδυασμός μεγάλου πλήθους τροποποιητικών δηλώσεων και μείωσης φόρου συνεπάγεται μεγαλύτερη πιθανότητα μη συμμόρφωσης.

Έτσι ορίζεται ο παρακάτω δείκτης.

**Στόχος του δείκτη** - να παρακινήσει τον φορολογούμενο να μην υποβάλλει τροποποιητικές δηλώσεις.

**Προσδιορισμός του δείκτη** - ο δείκτης προσδιορίζεται βάση:

- ✓ του πλήθους τροποποιητικών φορολογικών δηλώσεων που οφείλονται σε υπαιτιότητα του φορολογούμενου
- ✓ καθώς και του ύψους ποσού επιπρόσθετου φόρου που προκύπτει από αυτές τις τροποποιητικές δηλώσεις

που υποβλήθηκαν τα τελευταία 3 χρόνια.

**Πηγή πληροφοριών για τη μέτρηση του δείκτη** – Υποσυστήματα Φ.Π.Α και Φ.Ε. TAXIS.

**Χορήγηση βαθμών** - βαθμοί χορηγούνται με βάση τον παραπάνω δείκτη.

"5" σημεία χορηγούνται στον φορολογούμενο που δεν έχει υποβάλει τροποποιητικές φορολογικές δηλώσεις.

"1" βαθμός χορηγείται στον φορολογούμενο που έχει υποβάλει τροποποιητικές φορολογικές δηλώσεις με υπαιτιότητά του και σημαντική διαφορά φόρου.

### 5.3 Αξιολόγηση με βάση τις Βεβαιώσεις, Πληρωμές και Ληξιπρόθεσμες Οφειλές

Οι βεβαιώσεις οφειλών, οι πληρωμές και ειδικά οι ληξιπρόθεσμες οφειλές θα μπορούσαν να αποτελέσουν το κύριο κριτήριο βαθμολόγησης των φορολογούμενων.

Θα οριστούν πέντε δείκτες.

- ✓ Συνολικού ληξιπρόθεσμου χρέους
- ✓ Μεταβολής συνολικού ληξιπρόθεσμου χρέους
- ✓ Ρυθμισμένων ή μη οφειλών
- ✓ Συνέπειας διατήρησης ρύθμισης
- ✓ Δυνατότητας εξόφλησης χρέους

Και εδώ η συνολική βαθμολογία θα αποδοθεί στο φορολογούμενο για τη συγκεκριμένη διάσταση θα είναι η ελάχιστη των επιμέρους βαθμολογιών των δεικτών της διάστασης.

#### **5.3.1 Δείκτης 7: Συνολικού ληξιπρόθεσμου χρέους**

Οι φορολογούμενοι θα αξιολογούνται αναλογικά με το ύψος των ληξιπρόθεσμων οφειλών τους, δηλαδή όσο μεγαλύτερες είναι οι οφειλές που έχει κάποιος τόσο χαμηλότερα θα κατατάσσεται στη κλίμακα αξιολόγησης.

Ελέγχεται το ύψος του συνολικού ληξιπρόθεσμου φορολογικού χρέους του φορολογούμενου που είναι εκτός ρύθμισης.

**Στόχος του δείκτη** - να παρακινήσει τον φορολογούμενο να σβήσει το συσσωρευμένο ληξιπρόθεσμο φορολογικό χρέος.

**Προσδιορισμός του δείκτη** - η μέτρηση του συνολικού ποσού των ληξιπρόθεσμων οφειλών που διαχειρίζεται η Φορολογική Αρχή πραγματοποιείται με βάση την κατάσταση των χρεών στο τέλος του προηγούμενου μήνα.

*Παρατήρηση: Στον υπολογισμό του συνολικού ποσού των οφειλών θα ληφθούν υπόψη τυχόν συμψηφισμοί που δεν έχουν εκτελεστεί.*

**Πηγή πληροφοριών για τη μέτρηση του δείκτη** – Υποσύστημα ΕΣΟΔΩΝ -TAXIS.

**Χορήγηση βαθμών** - βαθμοί χορηγούνται βάσει του ποσού του ληξιπρόθεσμου χρέους του φορολογούμενου.

Τα "5" σημεία χορηγούνται στον φορολογούμενο χωρίς χρέη ή με ληξιπρόθεσμο χρέη έως 30 ευρώ.

Τα "4" σημεία χορηγούνται στους φορολογούμενους των οποίων το φορολογικό χρέος είναι ασήμαντο.

### 5.3.2 Δείκτης 8: Μεταβολής Συνολικού ληξιπρόθεσμου χρέους

Εκτός από το ύψος του συνολικού ληξιπρόθεσμου φορολογικού χρέους του φορολογούμενου, είναι σημαντικό να γνωρίζουμε αν αυτό αυξάνεται ή μειώνεται στο χρόνο. Για αυτό ορίζεται ο παρακάτω δείκτης.

**Στόχος του δείκτη** - να παρακινήσει τον φορολογούμενο να μειώσει το συσσωρευμένο ληξιπρόθεσμο φορολογικό χρέος εκτός ρύθμισης.

**Προσδιορισμός του δείκτη** - οι μεταβολές στο συνολικό ληξιπρόθεσμο χρέος που διαχειρίζεται η Φορολογική Αρχή υπολογίζονται με τη σύγκριση της κατάστασης των ληξιπρόθεσμων χρεών στο τέλος του προηγούμενου μήνα και της κατάστασης των ληξιπρόθεσμων χρεών τη σχετική ημερομηνία του προηγούμενου έτους.

**Πηγή πληροφοριών για τη μέτρηση του δείκτη** – Υποσύστημα Εσόδων -TAXIS.

**Η χορήγηση βαθμών** γίνεται με δύο δείκτες:

- ✓ την απόλυτη αξία της αύξησης του ποσού των ληξιπρόθεσμων οφειλών.
- ✓ τη σχετική αύξηση του ληξιπρόθεσμου χρέους, η οποία εκφράζεται σε ποσοστιαίες μονάδες και υπολογίζεται ως ο λόγος της απόλυτης αξίας της αύξησης του συνολικού ποσού του ληξιπρόθεσμου χρέους την αντίστοιχη ημερομηνία του προηγούμενου έτους.

Για παράδειγμα, αν το συνολικό ληξιπρόθεσμο χρέος του φορολογούμενου ανερχόταν σε 170 ευρώ την 31η Αυγούστου 2018 και σε 140 ευρώ την 31η Αυγούστου 2017, η σχετική αύξηση του χρέους είναι  $(170-140): 140 \times 100\% = 21\%$ .

Σημεία "5" χορηγούνται στον φορολογούμενο που δεν είχε ληξιπρόθεσμο χρέη στο τέλος του προηγούμενου μήνα, ανεξάρτητα από την παρουσία και το ύψος του χρέους την αντίστοιχη ημερομηνία του προηγούμενου έτους.

Τα "4" σημεία χορηγούνται στους φορολογούμενους που έχουν ληξιπρόθεσμες φορολογικές οφειλές, αλλά σε σύγκριση με την αντίστοιχη περίοδο του προηγούμενου έτους το χρέος μειώθηκε τουλάχιστον κατά 0,01 ευρώ.

Βαθμοί "3" χορηγούνται στους φορολογούμενους για τους οποίους η αύξηση του ληξιπρόθεσμου χρέους είναι ασήμαντη.

Το σημείο "1" χορηγείται στους φορολογούμενους για τους οποίους η αύξηση του ληξιπρόθεσμου χρέους είναι σημαντική.

### 5.3.3 Δείκτης 9: Ρυθμισμένων Οφειλών

Μία ρυθμισμένη οφειλή δείχνει προθυμία για καταβολή των οφειλόμενων φόρων και τάση για συμμόρφωση.

**Στόχος του δείκτη** - να παρακινήσει τον φορολογούμενο να ρυθμίσει το ληξιπρόθεσμο φορολογικό χρέος.

**Προσδιορισμός του δείκτη** - ο δείκτης προσδιορίζεται βάσει του χρόνου που έχει περάσει από τη δημιουργία του τελευταίου ληξιπρόθεσμου χρέους του φορολογούμενου.

**Πηγή πληροφοριών για τη μέτρηση του δείκτη** – Υποσύστημα Εσόδων -TAXIS.

**Χορήγηση βαθμών** - βαθμοί χορηγούνται βάσει του χρόνου που το χρέος παραμένει ληξιπρόθεσμο χωρίς ρύθμιση.

Τα "5" σημεία χορηγούνται στον φορολογούμενο χωρίς ληξιπρόθεσμο χρέη ή με ληξιπρόθεσμο χρέη έως 30 ευρώ.

Τα "4" σημεία χορηγούνται στους φορολογούμενους με ληξιπρόθεσμο χρέη συνολικού ποσού μεγαλύτερου των 30 ευρώ που το παλαιότερο ληξιπρόθεσμο χρέος δημιουργήθηκε σχετικά πρόσφατα.

"1" βαθμός χορηγείται στο φορολογούμενο που δείχνει απρόθυμος να ρυθμίσει τα χρέη του.



#### 5.3.4 Δείκτης 10: Συνέπειας Διατήρησης Ρύθμισης

Παρατηρείται συχνά οι φορολογούμενοι να προβαίνουν σε ρύθμιση ληξιπρόθεσμων οφειλών αλλά να μην παραμένουν συνεπείς σε αυτή. Αυτό συνήθως οφείλεται σε αδυναμία πληρωμής των δόσεων της ρύθμισης. Είναι όμως πολύ πιθανό η κίνηση της ρύθμισης να έχει γίνει με δόλο για να βελτιωθεί προσωρινά η εικόνα των χρεών του φορολογούμενου προκειμένου να επιτραπεί κάποια ενέργεια όπως π.χ. η έκδοση φορολογικής ενημερότητας. Σε αυτή τη περίπτωση έχει παρατηρηθεί πως η ρύθμιση διακόπτεται άμεσα, συνήθως μέσα στους επόμενους 2 μήνες από την έναρξή της.

**Στόχος του δείκτη** - να παρακινήσει τον φορολογούμενο να παραμένει συνεπής στη ρύθμιση.

**Προσδιορισμός του δείκτη** - η μέτρηση του συνολικού αριθμού των ρυθμίσεων που δεν τηρήθηκαν με υπαιτιότητα του φορολογούμενου τα τελευταία 3 χρόνια. Επιπρόσθετα επιβαρύνεται περαιτέρω ο φορολογούμενος αν κάποια ή κάποιες από τις παραπάνω ρυθμίσεις διακόπηκαν σε λιγότερο από 2 μήνες μετά την έναρξή τους πάντα με υπαιτιότητα του φορολογούμενου.

**Πηγή πληροφοριών για τη μέτρηση του δείκτη** – Υποσύστημα Εσόδων -TAXIS.

**Χορήγηση βαθμών** - βαθμοί χορηγούνται βάσει:

- ✓ του συνολικού αριθμού των ρυθμίσεων του φορολογούμενου που δεν τηρήθηκαν με υπαιτιότητά του
- ✓ του συνολικού αριθμού των ρυθμίσεων του φορολογούμενου που διακόπηκαν σε λιγότερο από 2 μήνες μετά την έναρξή τους με υπαιτιότητα του φορολογούμενου

Τα "5" σημεία χορηγούνται στον φορολογούμενο χωρίς ρυθμίσεις ή με ρυθμίσεις που τηρεί.

### 5.3.5 Δείκτης 11: Δυνατότητας Εξόφλησης του Χρέους

Το ποσό της φορολογικής οφειλής εξαρτάται από το ποσό των φορολογικών υποχρεώσεων. Οι φορολογούμενοι με μεγάλο κύκλο εργασιών και μεγάλο αριθμό εργαζομένων έχουν μεγαλύτερη πιθανότητα να σχηματίσουν μεγάλα ποσά φορολογικών οφειλών. Με τη σειρά τους, για τους μικρούς φορολογούμενους, τα απόλυτα ποσά των φορολογικών οφειλών είναι ως επί το πλείστον μικρά. Λαμβάνοντας υπόψη τα προαναφερθέντα, είναι σημαντικό να μετρήσουμε τη δυνατότητα εξόφλησης του χρέους, διευκρινίζοντας αν το συσσωρευμένο ποσό του χρέους είναι ανάλογο με την ταμειακή ροή των πληρωμών που εξασφαλίζει ο φορολογούμενος.

**Στόχος του δείκτη** - να εξισορροπηθεί η διάσταση του συνολικού χρέους και ο παράγοντας μεγέθους επιχειρήσεων.

**Προσδιορισμός του δείκτη** - ο δείκτης προσδιορίζεται ως ο λόγος του συνολικού ποσού των βεβαιωμένων οφειλών του φορολογούμενου προς τις δόσεις πληρωμών του. Η μέτρηση του συνολικού ποσού των οφειλών πραγματοποιείται με βάση την κατάσταση των οφειλών στο τέλος του προηγούμενου μήνα. Τα ποσά των δόσεων πληρωμών υπολογίζονται με βάση τις δόσεις των πληρωμών που πραγματοποίησαν οι φορολογούμενοι τους τελευταίους 12 μήνες για αποπληρωμή κεφαλαίου.

**Πηγή πληροφοριών για τη μέτρηση του δείκτη** – Υποσύστημα Εσόδων -TAXIS.

**Χορήγηση βαθμών** - βαθμοί χορηγούνται με τη μέτρηση του λόγου του συνολικού χρέους που διαχειρίζεται η υπηρεσία προς τις δόσεις πληρωμών.

Τα "5" σημεία χορηγούνται στον φορολογούμενο χωρίς χρέη.

Τα "4" σημεία χορηγούνται στους φορολογούμενους, για τους οποίους το συνολικό χρέος τους είναι μικρότερο των δόσεων πληρωμών που δύνανται να πληρώνουν ετησίως.

Το σημείο "1" χορηγείται εάν ο φορολογούμενος με φορολογική οφειλή δεν έχει προβεί σε πληρωμές κατά τη διάρκεια των τελευταίων 12 μηνών ή το οφειλόμενο χρέος είναι πολύ μεγαλύτερο των χρηματικών ποσών που πληρώνει ετησίως.

## 5.4 Αξιολόγηση με βάση τις Παραβάσεις

### 5.4.1 Δείκτης 12 : Παραβάσεων

Ελέγχεται αν οι φορολογούμενοι έχουν υποπέσει σε φορολογικές παραβάσεις οι οποίες έχουν καταχωρηθεί στα μηχανογραφικά συστήματα «TAXIS» και «ELENXIS».

Για τα Φυσικά Πρόσωπα θα πρέπει να αξιολογούνται καταλογισμένες παραβάσεις για περιπτώσεις όπως η επαύξηση περιουσίας, η μη έναρξη δραστηριότητας, η λήψη δεύτερου Α.Φ.Μ., κλπ.

Για τα Νομικά και τα Φυσικά Πρόσωπα επιτηδευματίες βασικό κριτήριο θα αποτελέσει το είδος των φορολογικών παραβάσεων – πρόστιμων που έχουν επιβληθεί, το πλήθος αυτών καθώς και τα επιβαλλόμενα ποσά των ανωτέρω προστίμων.

Κάποιες από τις παραβάσεις που πρέπει να αξιολογηθούν με υψηλό συντελεστή βαρύτητας είναι οι εξής:

- Έκδοση και λήψη πλαστών – εικονικών φορολογικών στοιχείων.
- Μη έκδοση ή ανακριβής έκδοση φορολογικών στοιχείων.
- Παρεμπόδιση ελέγχου.
- Χρήση παραποιημένου λογισμικού για τη διαγραφή αποδείξεων που είχαν εκδοθεί.
- Μη δήλωση φορολογικού μηχανισμού, παραβίαση φορολογικού μηχανισμού, μη χρήση εγκεκριμένου φορολογικού μηχανισμού ή χρήση μη εγκεκριμένου φορολογικού μηχανισμού κλπ.

**Στόχος του δείκτη** - να παρακινήσει τους φορολογούμενους να δώσουν μεγαλύτερη προσοχή στην αποφυγή παραβάσεων του φορολογικού συστήματος, που μπορεί να είναι ο λόγος για τον περιορισμό των οικονομικών τους δραστηριοτήτων.

**Προσδιορισμός του δείκτη** - ο δείκτης προσδιορίζεται λαμβάνοντας υπόψη:

- ✓ τον συντελεστής βαρύτητας παράβασης,
- ✓ τον αριθμό των πλήρων μηνών που έχουν περάσει από την ημερομηνία παράβασης,
- ✓ επιβαλλόμενο ποσό προστίμου λόγω της συγκεκριμένης παράβασης.

**Πηγή πληροφοριών για τη μέτρηση του δείκτη** – Υποσύστημα Φ.Π.Α, Υποσύστημα Κ.Β.Σ, Υποσύστημα Φ.Ε. , Υποσύστημα Εσόδων - TAXIS και ELENXIS.

**Χορήγηση βαθμών** - βαθμοί χορηγούνται βάσει του παραπάνω υπολογισμένου δείκτη. Οι χαμηλότεροι βαθμοί αξιολόγησης χορηγούνται σε φορολογούμενους με σχετικά πρόσφατα περιστατικά παράβασης.

Η επανάληψη της ίδιας παράβασης θα έχει σαν αποτέλεσμα την περαιτέρω μείωση της υπολογισθείσας κατηγορίας.

Τα "5" σημεία χορηγούνται στον φορολογούμενο χωρίς παραβάσεις.

### 6.1 Συλλογή δεδομένων

### 6.2 Γνωρίσματα που θα χρησιμοποιηθούν

#### 6.1 Συλλογή δεδομένων

Το πρώτο βήμα σε κάθε διαδικασία εξόρυξης είναι η συλλογή των κατάλληλων στοιχείων – χαρακτηριστικών που θα χρησιμοποιηθούν. Η εξόρυξη δεδομένων έχει σχεδιαστεί με σκοπό να εντοπίσει τη σωστή μεταβλητή έτσι ώστε να καθορίσει την απόφαση που πρόκειται να ληφθεί (Witten I.H. and Frank E., 2000). Η επιλογή της κατάλληλης μεταβλητής καθορίζει σε μεγάλο βαθμό την ακρίβεια της εξόρυξης δεδομένων που πρόκειται να εκτελεστεί. Βασικός οδηγός σε κάθε περίπτωση αποτελεί το επιχειρησιακό πλαίσιο μέσα στο οποίο θα υλοποιηθεί η δράση.

Για την υλοποίηση όσων αναφέρθηκαν στα προηγούμενα κεφάλαια τα στοιχεία που συλλέγουμε αφορούν νομικά πρόσωπα και αναφέρονται στις κινήσεις τους για τα φορολογικά έτη 2015-2017. Για τους σκοπούς της ΕΔ οι φορολογούμενοι χωρίζονται σε πέντε ομάδες βάσει των διαθέσιμων πληροφοριών.

Οι πληροφορίες προκύπτουν από τα πληροφοριακά συστήματα «TAXIS» και «ELENXIS». Συγκεκριμένα:

- TAXIS – Υποσύστημα Μητρώου :
- TAXIS – Υποσύστημα Εσόδων
- TAXIS – Υποσύστημα Φ.Π.Α
- TAXIS – Υποσύστημα Φορολογίας Εισοδήματος
- ELENXIS

##### 6.1.1 TAXIS - Υποσύστημα Μητρώου

Από το Υποσύστημα Μητρώου θα συλλέξουμε τα στοιχεία που ακολουθούν.

Για τον υπολογισμό του Δείκτη 1.

- ✓ Α.Φ.Μ.
- ✓ Ένδειξη Μητρώου  
(Ένδειξη = 1 αν έστω μία από τις ενδείξεις Πτώχευσης, Απενεργοποιημένου Α.Φ.Μ., Εξαφανισμένου Εμπόρου, Α.Φ.Μ σε Αναστολή είναι ενεργή, διαφορετικά Ένδειξη = 0).
- ✓ Πλήθος διακοπών εργασιών

Για τον υπολογισμό του Δείκτη 2.

Από Εκπρόθεσμες Δηλώσεις Μεταβολών Μητρώου τα πεδία:

- ✓ Συντελεστής βαρύτητας αιτίας μεταβολής
- ✓ Χρόνος καθυστέρησης υποβολής Εκπρόθεσμης Δήλωσης  
Υπολογίζεται από τις τιμές των πεδίων :
  - Ημερομηνία υποβολής Εκπρόθεσμης Δήλωσης
  - Ημερομηνία Έναρξης Μεταβολής.

Από τις τιμές των δύο παραπάνω και βάση του τύπου υπολογίζουμε τον Δείκτη 2.

**6.1.2 TAXIS - Υποσύστημα Φ.Π.Α και Υποσύστημα Φ.Ε.**

Από τα Υποσυστήματα Φ.Π.Α και Φ.Ε. θα συλλέξουμε τα στοιχεία που ακολουθούν.

Για τον υπολογισμό του Δείκτη 3.

- ✓ Πλήθος Δηλώσεων που δεν υποβλήθηκαν τα τελευταία 3 έτη : προκύπτει από την αφαίρεση του πλήθους των δηλώσεων που αναμέναμε (βάσει κατηγορίας βιβλίων της επιχείρησης) μείον το πλήθος των δηλώσεων που υποβλήθηκαν.

Για τον υπολογισμό του Δείκτη 4.

Από τα πρόστιμα των Ανακριβών Δηλώσεων:

- ✓ Το συνολικό πλήθος προστίμων ανακρίβειας που έχουν επιβληθεί κατόπιν ελέγχου
- ✓ Το συνολικό ύψος των φόρων που έχουν επιβληθεί λόγω ανακρίβειας

Για τον υπολογισμό του Δείκτη 5.

Από το αρχείο των Δηλώσεων:

- ✓ Το πλήθος εκπροθέσμων φορολογικών δηλώσεων.
- ✓ Το συνολικό χρόνο καθυστέρησης υποβολής όλων των εκπρόθεσμων δηλώσεων.
- ✓ Το συνολικό ύψος ποσού επιπρόσθετου φόρου που προκύπτει από τις εκπρόθεσμες δηλώσεις.

Για τον υπολογισμό του Δείκτη 6.

Από το αρχείο των Δηλώσεων:

- ✓ Το πλήθος τροποποιητικών φορολογικών δηλώσεων με υπαιτιότητα του φορολογούμενου.
- ✓ Το συνολικό ύψος ποσού επιπρόσθετου φόρου που προκύπτει από τις παραπάνω τροποποιητικές δηλώσεις.

### **6.1.3 TAXIS - Υποσύστημα Εσόδων**

Από το Υποσύστημα Εσόδων θα συλλέξουμε τα στοιχεία που ακολουθούν.

Για τον υπολογισμό του Δείκτη 7.

- ✓ Το συνολικό ποσό των ληξιπρόθεσμων οφειλών στο τέλος του προηγούμενου μήνα.

Για τον υπολογισμό του Δείκτη 8.

- ✓ Το συνολικό ποσό των ληξιπρόθεσμων οφειλών στο τέλος του προηγούμενου μήνα του τρέχοντος έτους.
- ✓ Το συνολικό ποσό των ληξιπρόθεσμων οφειλών στο τέλος του προηγούμενου μήνα του προηγούμενου έτους.

Και υπολογίζουμε τη Μεταβολή και τη Σχετική Μεταβολή του Ληξιπρόθεσμου Χρέους.

Για τον υπολογισμό του Δείκτη 9.

- ✓ Το χρόνο που έχει περάσει από τη δημιουργία του τελευταίου ληξιπρόθεσμου χρέους του φορολογούμενου (χωρίς να έχει γίνει ρύθμιση).

Για τον υπολογισμό του Δείκτη 10.

- ✓ Πλήθος ρυθμίσεων που δεν τηρήθηκαν τα τελευταία 3 χρόνια.
- ✓ Πλήθος ρυθμίσεων που ίσχυσαν λιγότερο από 2 μήνες

Για τον υπολογισμό του Δείκτη 11.

- ✓ Το συνολικό ποσό των βεβαιωμένων οφειλών του φορολογούμενου στο τέλος του προηγούμενου μήνα.
- ✓ Το συνολικό ποσό των πληρωμών του φορολογούμενου που έγιναν κατά τη διάρκεια του τελευταίου χρόνου.

#### **6.1.4 ELENXIS και TAXIS**

Για τον υπολογισμό του Δείκτη 12.

- ✓ Το συντελεστής βαρύτητας παράβασης.
- ✓ Τον αριθμό των πλήρων μηνών που έχουν περάσει από την ημερομηνία παράβασης.
- ✓ Το επιβαλλόμενο ποσό προστίμου λόγω της συγκεκριμένης παράβασης.
- ✓ Ένδειξη Επανάληψης Παράβασης



(η παραπάνω Ένδειξη παίρνει τη τιμή 1 αν υπάρχει έστω ένας κωδικός παράβασης που επαναλαμβάνεται στη διάρκεια των τελευταίων 36 μηνών, διαφορετικά η Ένδειξη παίρνει τη τιμή 0).

Ο κατάλογος των μεταβλητών που αναφέρθηκε παραπάνω είναι ένας εφικτός κατάλογος των μεταβλητών εισόδου. Θα μπορούσε να εμπλουτιστεί πολύ περισσότερο αλλά τότε οι νέες μεταβλητές, αν και τεχνικά εφικτές, δεν θα μπορούσαν να υπολογιστούν λόγω μη διαθεσιμότητας στη βάση δεδομένων, ασυνέπειας δεδομένων, υψηλού προγραμματισμού & κ.λπ.

Στα παραπάνω δεδομένα έχει ήδη γίνει **συσσωμάτωση (Data Aggregation)**. Η συσσωμάτωση είναι μια προσπάθεια να γίνει μια περίληψη στα δεδομένα (Fisher, Vickie L. (1985)). Τα στοιχεία των φορολογικών δηλώσεων, των χρεών, των πληρωμών φόρου και των παραβάσεων έχουν συγκεντρωθεί ανά έτος ή τριετία για κάθε φορολογούμενο και για όλα τα είδη φόρου. Το αποτέλεσμα της διαδικασίας συσσωμάτωσης είναι η μεταβλητή που θα καθορίσει εάν οι φορολογούμενοι συμμορφώνονται τυπικά.

## 6.2 Γνωρίσματα που θα χρησιμοποιηθούν

Ακολουθεί πίνακας με τα γνωρίσματα που θα χρησιμοποιηθούν.

| A/A | ΟΝΟΜΑ           | ΤΥΠΟΣ<br>ΜΕΤΑΒΛΗΤ<br>ΗΣ | ΤΙΜΕΣ                  | ΠΗΓΗ               | ΠΕΡΙΓΡΑΦΗ                                                    |
|-----|-----------------|-------------------------|------------------------|--------------------|--------------------------------------------------------------|
| 1   | AFM             | ID                      |                        | Μητρώο             | A.Φ.Μ.                                                       |
| 2   | ENDEIXH_MHTRWOY | Categorical<br>binary   | 0,1                    | Μητρώο             | Ένδειξη Μητρώου                                              |
| 3   | DIAK_ERG_PL     | ratio                   | Διακρι<br>τές<br>τιμές | Μητρώο             | Πλήθος Διακοπών<br>Εργασιών                                  |
| 4   | AITIA_METABOLHS | ratio                   | Διακρι<br>τές<br>τιμές | Μητρώο             | Συντελεστής<br>βαρύτητας αιτίας<br>μεταβολής                 |
| 5   | MHTRWO_XR       | ratio                   |                        | Μητρώο             | Χρόνος<br>καθυστέρησης<br>υποβολής<br>Εκπρόθεσμης<br>Δήλωσης |
| 6   | MH_YPOB_PL      | ratio                   | Διακρι<br>τές<br>τιμές | Φ.Ε. και<br>Φ.Π.Α. | Πλήθος<br>Δηλώσεων που<br>δεν υποβλήθηκαν                    |
| 7   | EKPR_PL         | ratio                   | Διακρι<br>τές<br>τιμές | Φ.Ε. και<br>Φ.Π.Α. | Πλήθος<br>Εκπροθέσμων<br>Δηλώσεων                            |

|    |            |       |                        |                    |                                             |
|----|------------|-------|------------------------|--------------------|---------------------------------------------|
|    |            |       |                        |                    |                                             |
| 8  | EKPR_XR    | ratio |                        | Φ.Ε. και<br>Φ.Π.Α. | Συνολικός χρόνος<br>Εκπροθέσμων<br>Δηλώσεων |
| 9  | EKPR_PS    | ratio |                        | Φ.Ε. και<br>Φ.Π.Α. | Συνολικό ποσό<br>Εκπροθέσμων<br>Δηλώσεων    |
| 10 | TROP_PL    | ratio | Διακρι<br>τές<br>τιμές | Φ.Ε. και<br>Φ.Π.Α. | Πλήθος<br>Τροποποιητικών<br>Δηλώσεων        |
| 11 | TROP_PS    | ratio |                        | Φ.Ε. και<br>Φ.Π.Α. | Συνολικό ποσό<br>Τροποποιητικών<br>Δηλώσεων |
| 12 | LXPR_PS    | ratio |                        | ΕΣΟΔΑ              | Συνολικό<br>ληξιπρόθεσμο χρέος              |
| 13 | LXPR_MET_1 | ratio |                        | ΕΣΟΔΑ              | Απόλυτη μεταβολή<br>ληξιπρόθεσμου<br>χρέους |
| 14 | LXPR_MET_2 | ratio |                        | ΕΣΟΔΑ              | Σχετική μεταβολή<br>ληξιπρόθεσμου<br>χρέους |

|    |                    |       |                 |                                     |                                                    |
|----|--------------------|-------|-----------------|-------------------------------------|----------------------------------------------------|
| 15 | LXPR_XWRIS_RTHM_XR | ratio |                 | ΕΣΟΔΑ                               | Χρόνος ληξιπρόθεσμων οφειλών χωρίς ρύθμιση         |
| 16 | RTHM_DIAK_PL       | ratio | Διακριτές τιμές | ΕΣΟΔΑ                               | Πλήθος ρυθμίσεων που δεν τηρήθηκαν                 |
| 17 | RTHM_2M_PL         | ratio | Διακριτές τιμές | ΕΣΟΔΑ                               | Πλήθος ρυθμίσεων που ίσχυσαν λιγότερο από 2 μήνες  |
| 18 | BEB_ANA_PLRM       | ratio |                 | ΕΣΟΔΑ                               | Συνολικό Χρέος / Σύνολο Πληρωμών τελευταίου χρόνου |
| 19 | ANAKR_DHL_PROST    | ratio |                 | Φ.Π.Α.<br>Φ.Ε.                      | Συνολικό ποσό προστίμων ανακριβών δηλώσεων         |
| 20 | PARAVASI_SYNT      | ratio |                 | Φ.Π.Α.<br>Φ.Ε.<br>Κ.Β.Σ.<br>ELENXIS | Συντελεστής Παραβάσεων                             |
| 21 | PARAVASI_PS        | ratio |                 | Φ.Π.Α.<br>Φ.Ε.<br>Κ.Β.Σ.<br>ELENXIS | Ποσό Προστίμου Παράβασης                           |
| 22 | PARAVASI_XR        | ratio |                 | Φ.Π.Α.                              | Χρόνος που έχει                                    |

|    |               |                       |       |                                      |                                            |
|----|---------------|-----------------------|-------|--------------------------------------|--------------------------------------------|
|    |               |                       |       | Φ.Ε.<br>Κ.Β.Σ.<br>ELENXIS            | περάσει από την<br>ημ/νία της<br>παράβασης |
| 23 | PARAVASI_EPAN | Categorical<br>binary | 0, 1  | Φ.Π.Α.<br>Φ.Ε.<br>Κ.Β.Σ..<br>ELENXIS | Ένδειξη<br>Επανάληψης<br>Παράβασης         |
| 24 | V_CLASS       | ordinal               | 1 – 5 |                                      | Βαθμίδα                                    |

*Πίνακας 3. Περιγραφή πεδίων*

## ΚΕΦΑΛΑΙΟ 7. ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

- 
- 7.1 Εργαλεία που θα χρησιμοποιηθούν
  - 7.2 Επισκόπηση Δεδομένων
  - 7.3 Δεδομένα εφαρμογής για εκπαίδευση και δοκιμή
  - 7.4 Εξόρυξη Δεδομένων
- 

### 7.1 Εργαλεία που θα χρησιμοποιηθούν

Το εργαλείο που θα χρησιμοποιηθεί για την εξόρυξη δεδομένων είναι το SPSS Modeler της IBM.

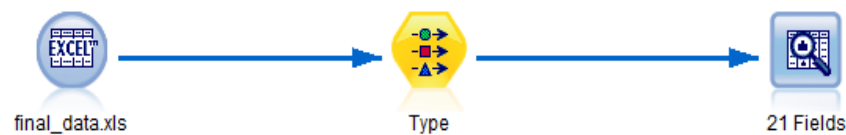
Πρόκειται για μια εφαρμογή λογισμικού εξόρυξης δεδομένων και ανάλυσης κειμένου. Χρησιμοποιείται για τη δημιουργία προγνωστικών μοντέλων και τη διεξαγωγή άλλων αναλυτικών εργασιών. Έχει μια οπτική διεπαφή που επιτρέπει στους χρήστες να χρησιμοποιούν αλγόριθμους στατιστικής και εξόρυξης δεδομένων χωρίς προγραμματισμό. Ένας από τους κύριους στόχους του εργαλείου είναι η απαλλαγή από την περιττή πολυπλοκότητα των μετασχηματισμών δεδομένων και η μετατροπή πολύπλοκων μοντέλων πρόβλεψης σε πολύ εύχρηστα. Η πρώτη έκδοση ενσωμάτωσε δέντρα απόφασης (ID3) και νευρωνικά δίκτυα, τα οποία θα μπορούσαν να εκπαιδευτούν και τα δύο χωρίς να έχουν υποκείμενη γνώση του τρόπου με τον οποίο οι τεχνικές αυτές δουλεύουν.

Το IBM SPSS Modeler ονομάστηκε αρχικά Clementine από τους δημιουργούς του, Integral Solutions Limited. Το όνομα αυτό συνεχίστηκε για λίγο μετά την απόκτηση του προϊόντος από την SPSS. Το SPSS αργότερα άλλαξε το όνομα σε SPSS Clementine και αργότερα σε PASW Modeler. Μετά την εξαγορά της SPSS από την IBM το 2009, το προϊόν μετονομάστηκε σε IBM SPSS Modeler, το σημερινό του όνομα.

## 7.2 Επισκόπηση Δεδομένων

Βασικό τμήμα κάθε διαδικασίας εξόρυξης είναι η επισκόπηση και κατανόηση των στοιχείων (γνωρισμάτων) που θα χρησιμοποιηθούν. Είναι κρίσιμο στάδιο καθώς τα συμπεράσματα από την επισκόπηση των στοιχείων σε ένα σημαντικό βαθμό καθορίζουν και την ποιότητα των αποτελεσμάτων.

Στο εργαλείο δημιουργούμε το παρακάτω μοντέλο:



Εικόνα 7.1 Ροή για προεπισκόπηση στοιχείων

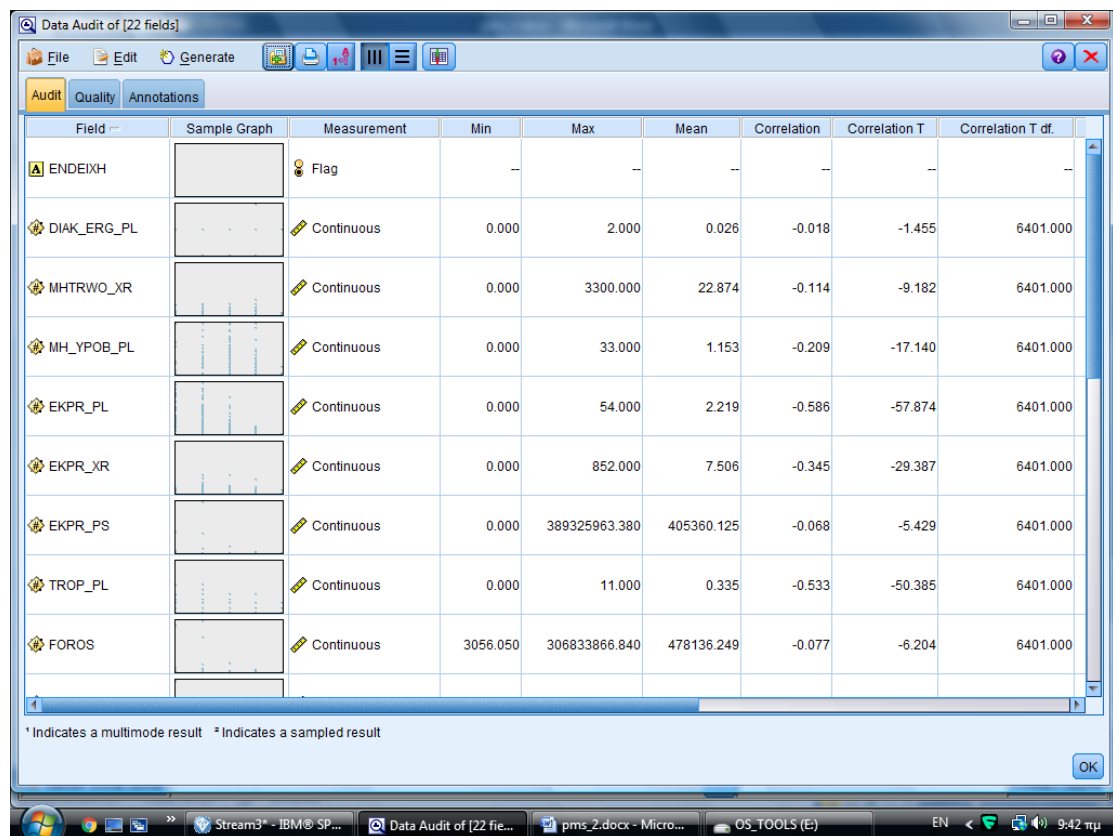
Αμέσως μετά κάνουμε προεπισκόπηση των στοιχείων. Χρησιμοποιώντας τα αποτελέσματα από την εμφάνιση των χαρακτηριστικών των στοιχείων, αν χρειαστεί, θα τροποποιήσουμε το σετ των στοιχείων ή και των γνωρισμάτων.

Μπορούμε να ελέγξουμε τους εξής περιγραφικούς δείκτες (Εικόνα 7.2):

- ✓ τύπος μεταβλητής,
- ✓ ελάχιστη και μέγιστη τιμή,
- ✓ μέσος όρος,
- ✓ τυπική απόκλιση,
- ✓ δείκτης ασυμμετρίας,

- ✓ μοναδικές τιμές για κατηγορικές μεταβλητές,
- ✓ έγκυρες τιμές

και να εξάγουμε συμπεράσματα για την ορθότητα των δεδομένων μας. Π.χ. στα binary πεδία θα πρέπει η ελάχιστη τιμή να είναι 0 και η μέγιστη 1. Διαφορετικά υπάρχει πρόβλημα. Επίσης αν διαπιστώσουμε ότι η ελάχιστη τιμή των πεδίων που μετράνε πλήθος είναι αρνητικός αριθμός πάλι υπάρχει πρόβλημα. Τέλος, η μεγάλη απόκλιση μεταξύ της μέσης τιμής και της εξεταζόμενης τιμής μπορεί να είναι ένδειξη θορύβου. Εκείνες οι τιμές που φαίνονται ύποπτες αφαιρούνται.



Εικόνα 7.2 Προεπισκόπηση Δεδομένων (Audit)

Για την ποιότητα των δεδομένων δίνονται οι εξής δείκτες (Εικόνα 7.3):

- ✓ τύπος μεταβλητής,
- ✓ outliers & extreme τιμές,
- ✓ ποσοστό και σύνολο έγκυρων τιμών,



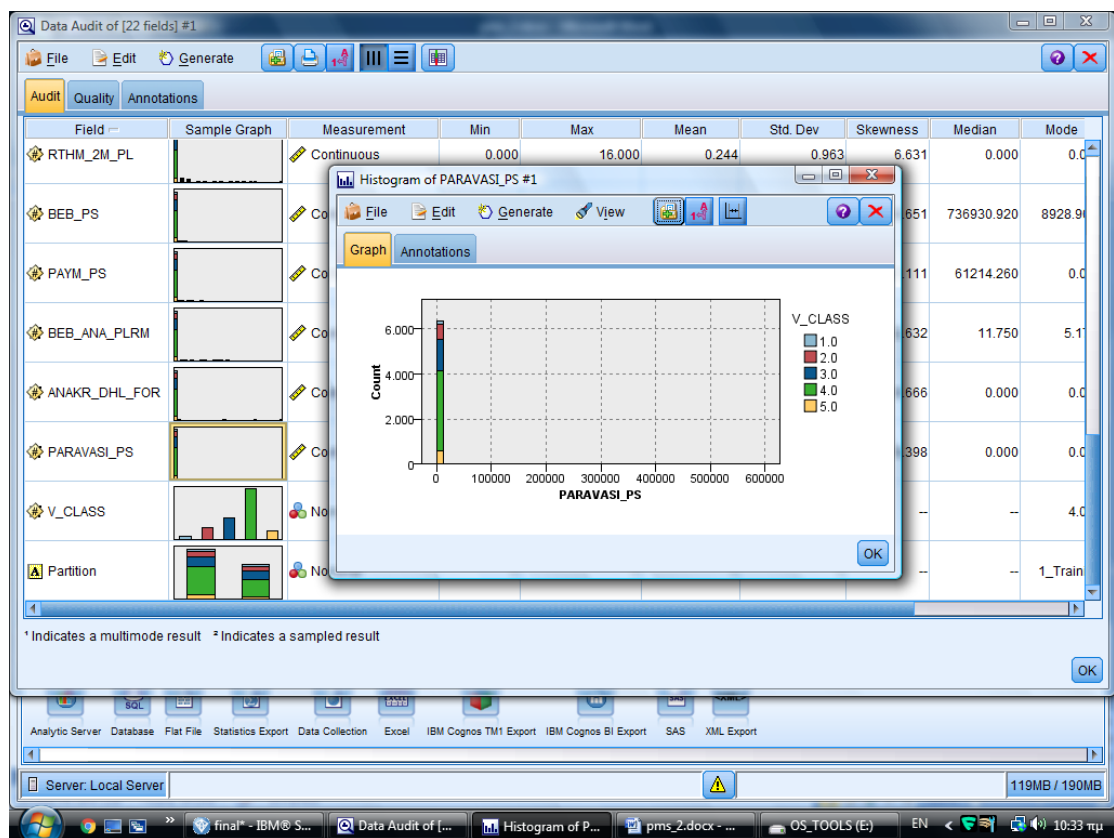
- ✓ Null τιμές,
- ✓ κενές τιμές κ.α.

| Field        | Measurement | Outliers | Extremes | Action | Impute Missing | Method | % Complete | Valid Rec |
|--------------|-------------|----------|----------|--------|----------------|--------|------------|-----------|
| DIAG_ERG_PL  | Continuous  | 0        | 164 None | Never  | Never          | Fixed  | 100        |           |
| MHTRWO_XR    | Continuous  | 34       | 28 None  | Never  | Never          | Fixed  | 100        |           |
| MH_YPOB_PL   | Continuous  | 95       | 72 None  | Never  | Never          | Fixed  | 100        |           |
| EKPR_PL      | Continuous  | 95       | 74 None  | Never  | Never          | Fixed  | 100        |           |
| EKPR_XR      | Continuous  | 62       | 36 None  | Never  | Never          | Fixed  | 100        |           |
| EKPR_PS      | Continuous  | 4        | 11 None  | Never  | Never          | Fixed  | 100        |           |
| TROP_PL      | Continuous  | 81       | 51 None  | Never  | Never          | Fixed  | 100        |           |
| FOROS        | Continuous  | 3        | 11 None  | Never  | Never          | Fixed  | 100        |           |
| AKATH        | Continuous  | 2        | 8 None   | Never  | Never          | Fixed  | 100        |           |
| LXPR_PS      | Continuous  | 6        | 9 None   | Never  | Never          | Fixed  | 100        |           |
| LXPR_LY_PS   | Continuous  | 37       | 42 None  | Never  | Never          | Fixed  | 100        |           |
| LXPR_MET1    | Continuous  | 1        | 8 None   | Never  | Never          | Fixed  | 100        |           |
| LXPR_XWRI... | Continuous  | 2        | 1 None   | Never  | Never          | Fixed  | 100        |           |
| RTHM_DIAK... | Continuous  | 38       | 61 None  | Never  | Never          | Fixed  | 100        |           |
| RTHM_2M_PL   | Continuous  | 69       | 49 None  | Never  | Never          | Fixed  | 100        |           |
| BEB_PS       | Continuous  | 3        | 6 None   | Never  | Never          | Fixed  | 100        |           |
| PAYM_PS      | Continuous  | 5        | 12 None  | Never  | Never          | Fixed  | 100        |           |
| BEB_ANA_P... | Continuous  | 7        | 17 None  | Never  | Never          | Fixed  | 100        |           |
| ANAKR_DHL... | Continuous  | 2        | 6 None   | Never  | Never          | Fixed  | 100        |           |
| PARAVASI_PS  | Continuous  | 9        | 37 None  | Never  | Never          | Fixed  | 100        |           |
| V_CLASS      | Nominal     | --       | --       | Never  | Never          | Fixed  | 100        |           |
| Partition    | Nominal     | --       | --       | Never  | Never          | Fixed  | 100        |           |

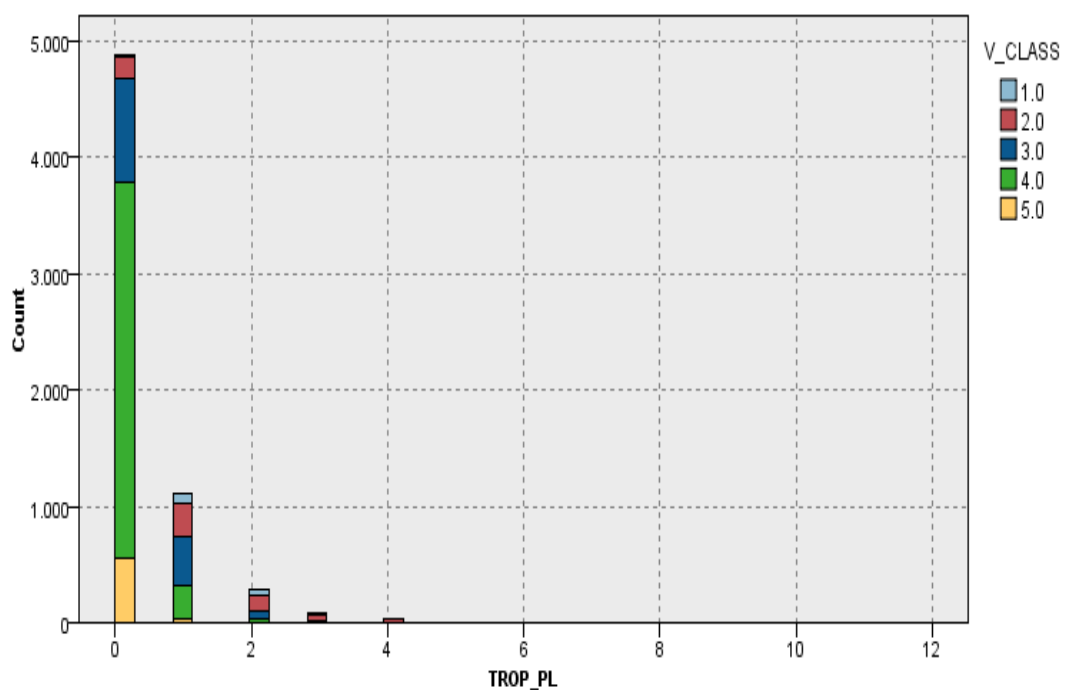
Εικόνα 7.3 Προεπισκόπηση Δεδομένων (Quality)

Τέλος μπορούμε να έχουμε οπτικοποίηση των τιμών των μεταβλητών με χρήση γραφημάτων.

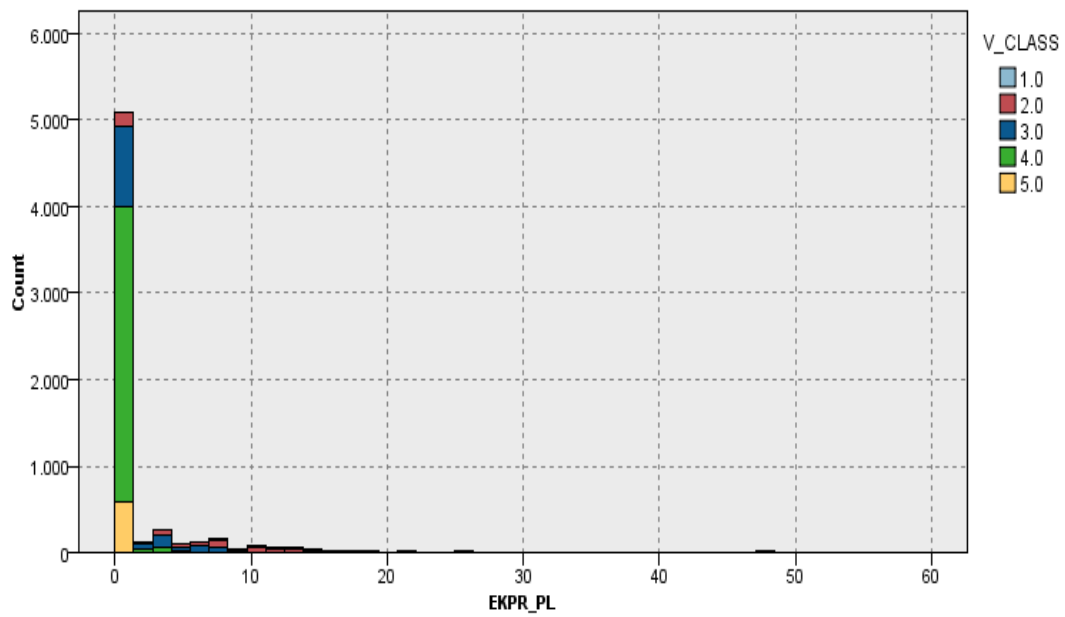
Ακολουθούν ιστογράμματα που δείχνουν μια κατανομή τιμών για ένα απλό χαρακτηριστικό.



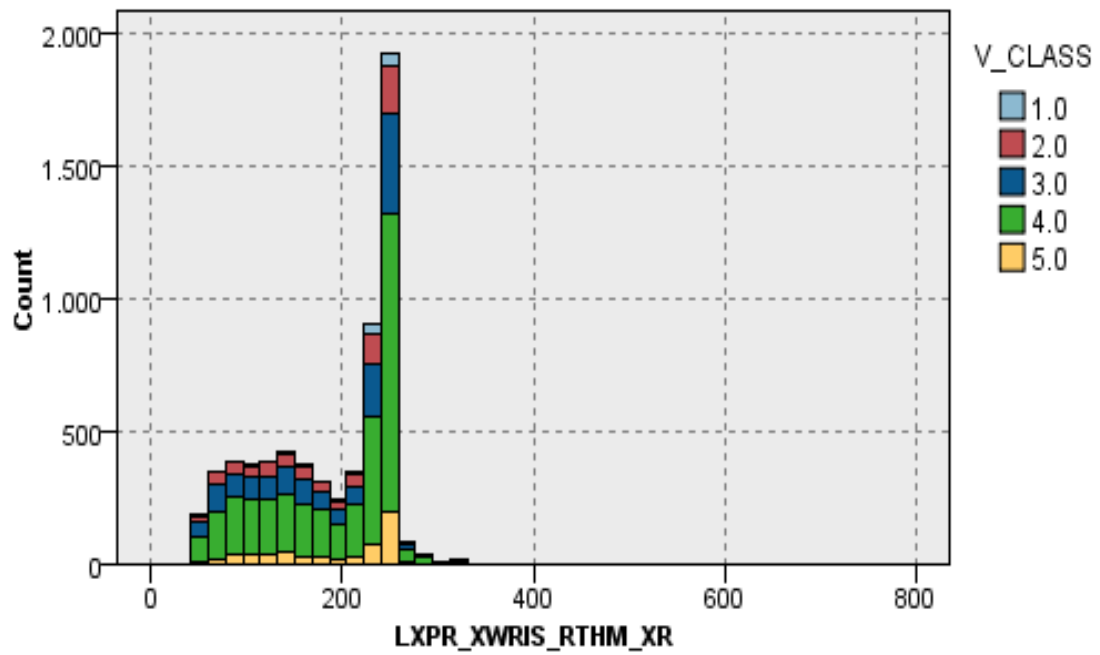
Εικόνα 7.4 Ιστογράμματα Δεδομένων



Εικόνα 7.5 Ιστόγραμμα Πλήθους Τροποποιητικών Δηλώσεων

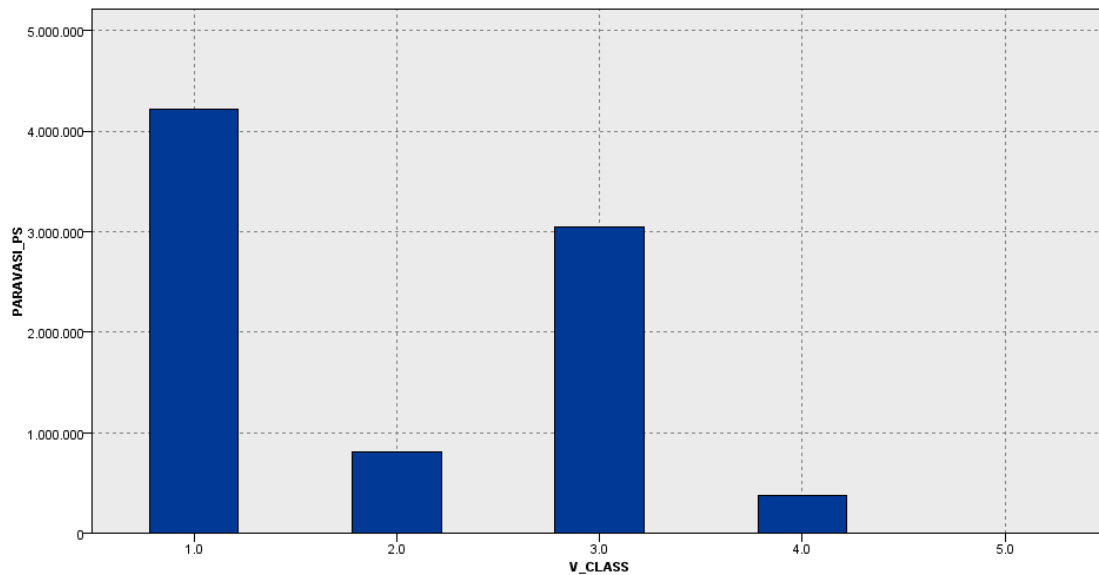


Εικόνα 7.6 Ιστόγραμμα Πλήθους Εκπροθέσμων Δηλώσεων



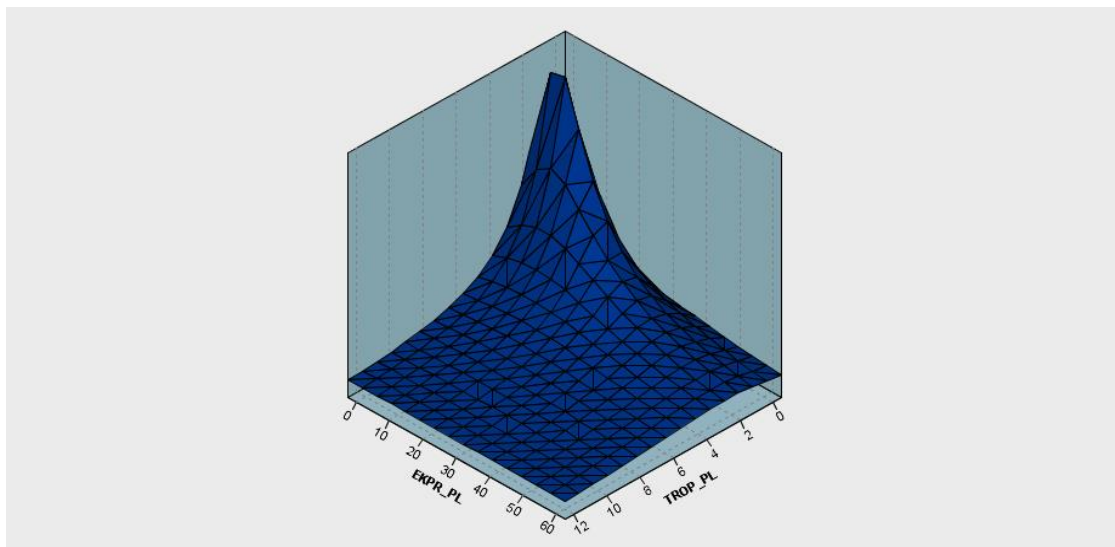
Εικόνα 7.7 Ιστόγραμμα Ληξιπροθέσμων χρεών χωρίς ρύθμιση

Υπάρχει η δυνατότητα εμφάνισης γραφημάτων με τις εξαρτήσεις δύο ή περισσότερων μεταβλητών π.χ. η κλάση του φορολογούμενου σε σχέση με τις παραβάσεις στις οποίες έχει υποπέσει.



Εικόνα 7.8 Παραβάσεις ανά κλάση

Ακολουθεί γραφική παράσταση πλήθους τροποποιητικών δηλώσεων και πλήθους εκπροθέσμων.

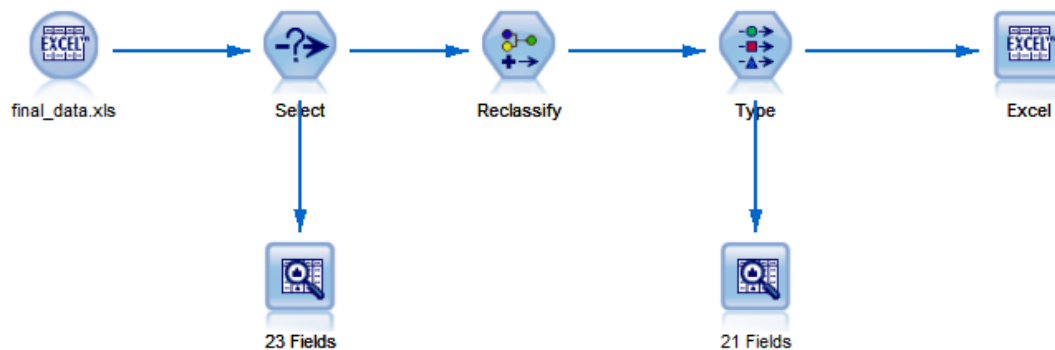


Εικόνα 7.9 Διάγραμμα διασποράς Πλήθους τροποποιητικών και εκπροθέσμων δηλώσεων

Παρατηρώντας τις τιμές διαπιστώνουμε ότι υπάρχουν εγγραφές που θα πρέπει να εξαιρέσουμε γιατί θα αλλοιώσουν το αποτέλεσμα. Πρόκειται για εκείνες τις εγγραφές που έχουν την τιμή 1 στο πεδίο ENDEIXH\_MHTRWOY γιατί όπως έχει ήδη ειπωθεί αυτοί οι φορολογούμενοι θα ανήκουν στη βαθμίδα 1 ανεξάρτητα από τις άλλες παραμέτρους. Αυτό γίνεται με χρήση <Select node>.

Επίσης υπάρχουν δείκτες- γνωρίσματα οι οποίοι λειτουργούν συμπληρωματικά για την αξιολόγηση του φορολογούμενου όπως για παράδειγμα ο Δείκτης 8. Θα υπολογίσουμε τις τελικές τιμές αυτών των δεικτών κάνοντας χρήση του <Reclassify node>.

Το μοντέλο ροής γίνεται ως εξής:

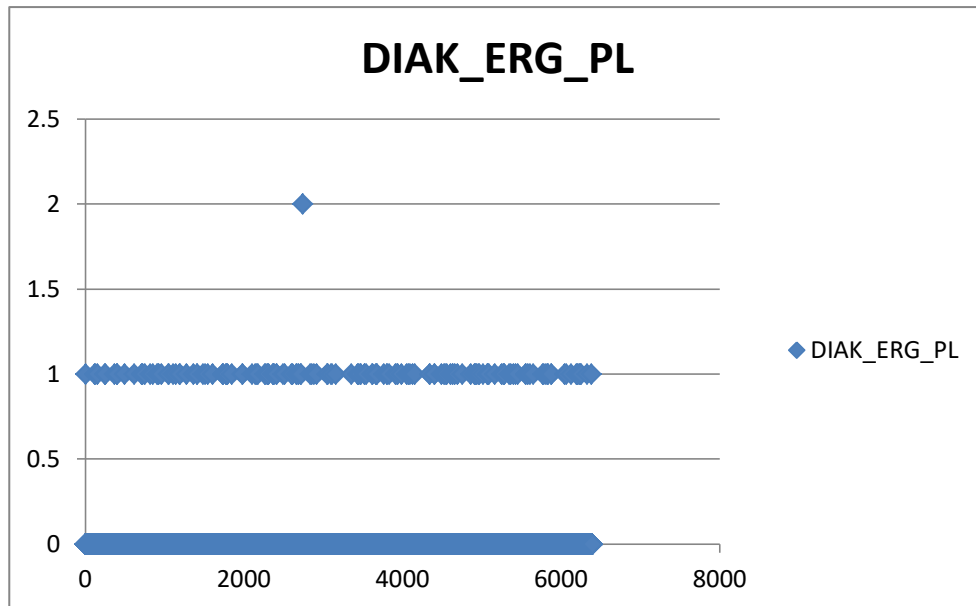


Εικόνα 7.10 Νέο Μοντέλο

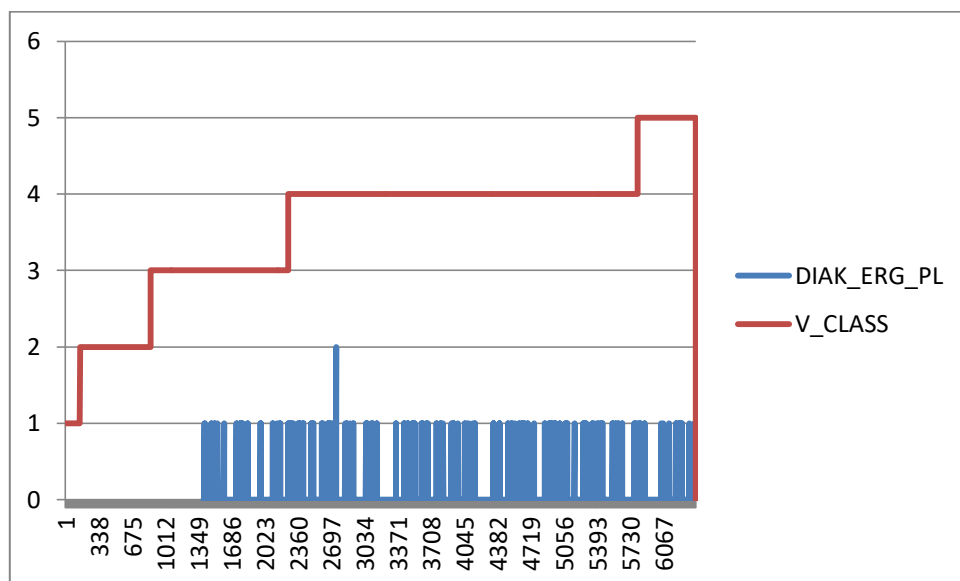
Από τη διερεύνηση των στοιχείων προκύπτει ότι παρότι στα δεδομένα έχει συμπεριληφθεί το γνώρισμα «Πλήθος Διακοπών Εργασιών», στη μεγάλη πλειοψηφία των εγγραφών η τιμή του γνωρίσματος είναι μηδέν (97,45%), ελάχιστες εγγραφές έχουν τιμή 1 (2,55%) και μόνο μία εγγραφή με τιμή 2. Συγκεκριμένα παρατηρούμε ότι:

- Ελάχιστη Τιμή = 0
- Μέγιστη Τιμή = 2
- Μέσος Όρος = 0,026

όπως φαίνεται άλλωστε και στα διαγράμματα που ακολουθούν.



Εικόνα 7.11 Γραφική απεικόνιση τιμών Πλήθους Διακοπών Εργασιών



Εικόνα 7.12 Γραφική απεικόνιση εξάρτησης της κλάσης από το πλήθος διακοπών εργασιών

Επιπρόσθετα οι τιμές 1 και 2 δεν χαρακτηρίζουν ύποπτη συμπεριφορά. Δηλαδή πρόκειται για ένα γνώρισμα που δεν βοηθά στην ταξινόμηση συνεπώς μπορούμε να το αγνοήσουμε.

Αυτό συμβαίνει γιατί το δείγμα μας αποτελείται μόνο από νομικά πρόσωπα. Συνήθως πολλές διακοπές εργασιών και επανενάρξεις εμφανίζονται στα φυσικά πρόσωπα-επιτηδευματίες.

Με τη βοήθεια των διαγραμμάτων ελέγχουμε αν υπάρχουν outliers προκειμένου να τα απομακρύνουμε.

Τέλος, το δείγμα περιορίστηκε λαμβάνοντας υπόψη διάφορα κριτήρια όπως ελλιπείς εγγραφές, null τιμές κλπ.

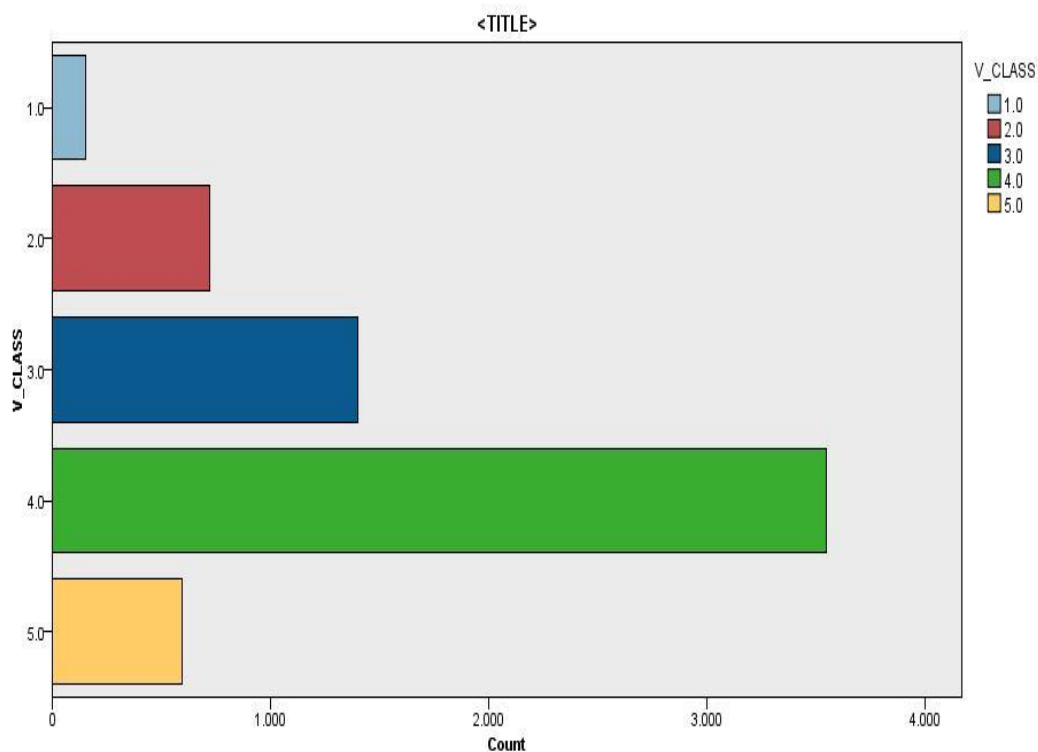
#### Ανισοβαρείς Κλάσεις

Σύμφωνα με τη βιβλιογραφία (Barandela et al, 2004), (Wang- Japkowic, 2009), (Baesens et al, 2015) οι ανισοβαρείς κλάσεις είναι ένα τυπικό πρόβλημα.

Εξετάζοντας προσεκτικότερα το πεδίο V\_CLASS διαπιστώνουμε ότι στο δείγμα φαίνεται να υπάρχει ανισοβαρής κατανομή των κλάσεων. Συγκεκριμένα το δείγμα αποτελείται από:

- 9,25 % φορολογούμενοι που ανήκουν στην βαθμίδα «5».
- 55,36 % φορολογούμενοι που ανήκουν στην βαθμίδα «4».
- 21,82 % φορολογούμενοι που ανήκουν στην βαθμίδα «3».
- 11,2 % φορολογούμενοι που ανήκουν στην βαθμίδα «2».
- 2,37 % φορολογούμενοι που ανήκουν στην βαθμίδα «1».

Αυτό απεικονίζεται στο διάγραμμα που ακολουθεί.



*Εικόνα 7.13 Ανισορροπία κλάσεων*

Το πρόβλημα που δημιουργεί η άνιση κατανομή μεταξύ των κλάσεων είναι η ακρίβεια των μοντέλων που παράγονται. Σε αυτές τις περιπτώσεις δημιουργούνται μοντέλα τα οποία παρουσιάζουν μεγάλη ακρίβεια, η οποία όμως δεν χαρακτηρίζει τη συνολική ακρίβειά τους αλλά την υπερ-εκπαίδευσή τους στην αναγνώριση της κλάσης πλειοψηφίας.

Υπάρχουν αρκετές τεχνικές που προσπαθούν να αντιμετωπίσουν το πρόβλημα, άλλες απλούστερες και άλλες αρκετά πιο σύνθετες όπως:

- Ο υπερ-δειγματισμός της κλάσης μειοψηφίας
- Η υπο-δειγματοληψία της κλάσης πλειοψηφίας
- Η χρήση βαρών στις κλάσεις
- Η χρήση μεγαλύτερου δείγματος κλπ



Στη συγκεκριμένη περίπτωση αυτή η ανισορροπία μέχρι ένα βαθμό είναι αποδεκτή καθώς δεν αναμένουμε οι πέντε κατηγορίες να έχουν ίσο πλήθος φορολογουμένων.

### 7.3 Δεδομένα εφαρμογής για εκπαίδευση και δοκιμή

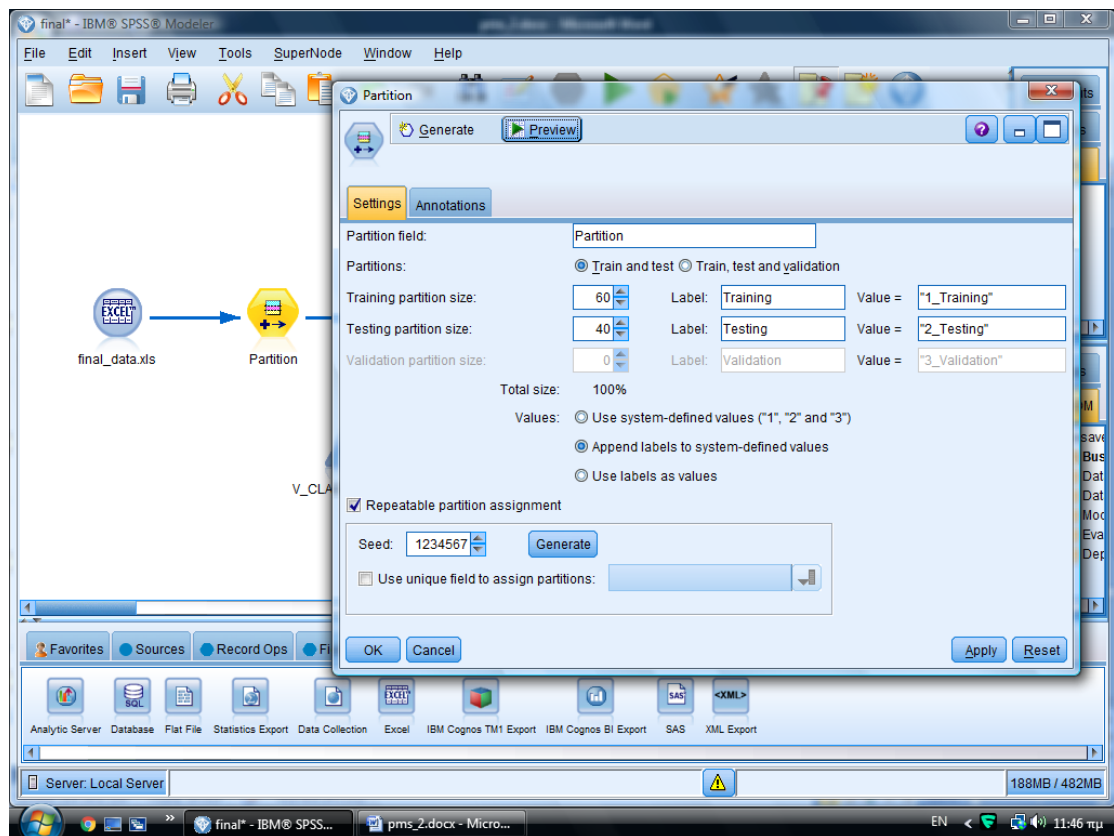
Έχοντας εκτελέσει τις πρώτες φάσεις της διαδικασίας Ε.Δ. δηλαδή τα βήματα:

- Διερεύνηση του επιχειρηματικού πεδίου
- Συλλογής των στοιχείων
- Μετασχηματισμού των στοιχείων
- Εκκαθάρισης των Στοιχείων

Τα στοιχεία είναι έτοιμα για τις επόμενες φάσεις της Ε.Δ. δηλαδή την εφαρμογή των αλγορίθμων και της ανάλυσης των αποτελεσμάτων.

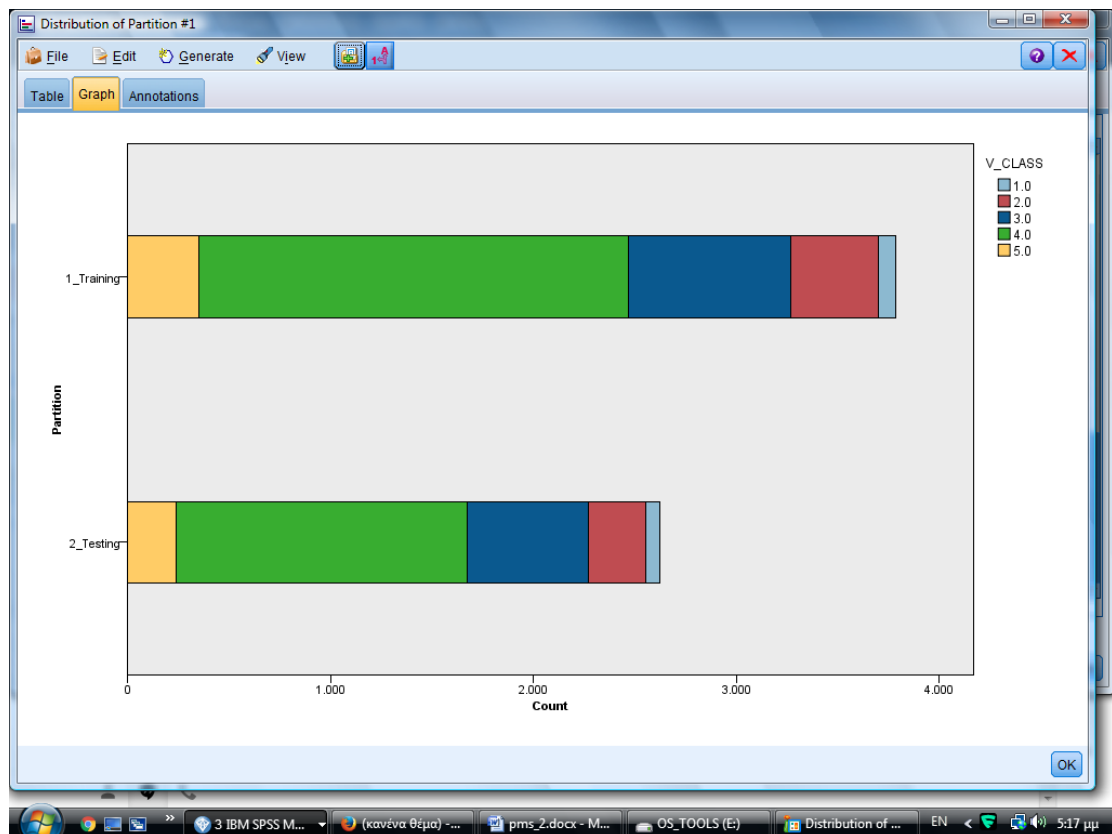
Προηγουμένως όμως πρέπει το σύνολο των δεδομένων να χωριστεί σε δεδομένα εκπαίδευσης (training data set) και δεδομένα επαλήθευσης (validation data set).

Αυτό γίνεται κάνοντας χρήση του <Partition Node> όπου και αναφέρουμε το ποσοστό των δεδομένων που θα χρησιμοποιηθούν για εκπαίδευση του αλγορίθμου (60%) και για επαλήθευση (40%).



Εικόνα 7.14 Διαχωρισμός Δεδομένων

Ακολουθεί διάγραμμα που δείχνει την κατανομή των δεδομένων εκπαίδευσης και των δεδομένων επαλήθευσης. Διαπιστώνουμε ότι ο διαχωρισμός έγινε με ποσοστά 50.9% και 41.1%.



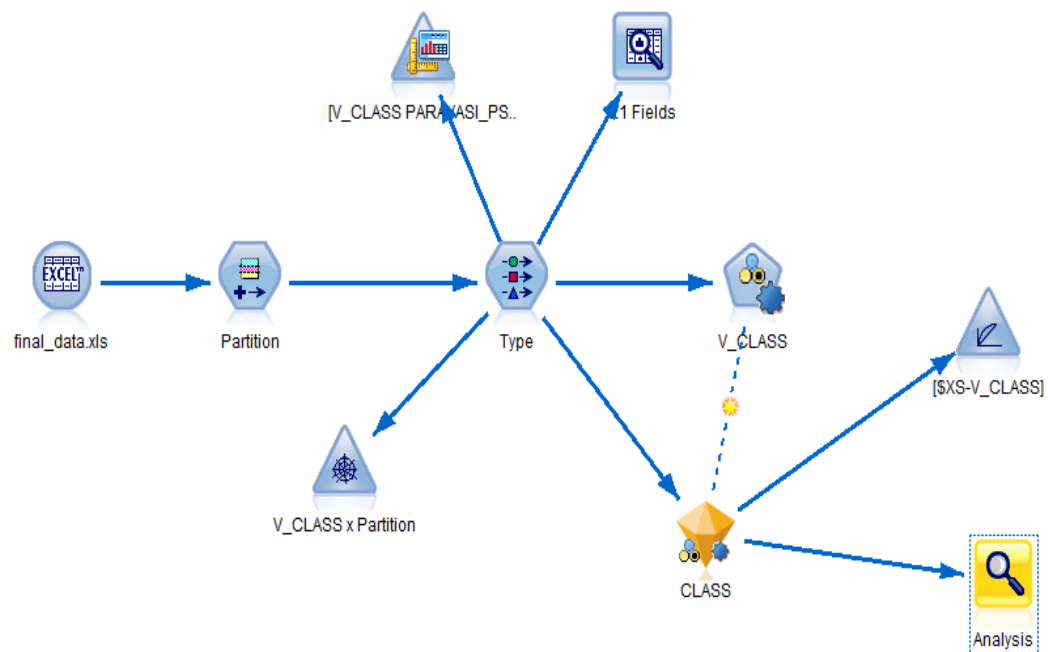
Εικόνα 7.15 Κατανομή των δεδομένων εκπαίδευσης και των δεδομένων επαλήθευσης

## 7.4 Εξόρυξη Δεδομένων

### Ταξινόμηση στοιχείων

Το εργαλείο SPSS Modeler αποτελείται από ένα σύνολο αλγορίθμων εξόρυξης οι οποίοι μπορούν να εκτελεστούν μέσω του γραφικού περιβάλλοντος που διαθέτει. Επίσης έχει τη δυνατότητα να μας υποδείξει τους τρεις καλύτερους αλγορίθμους λαμβάνοντας υπόψη τα δεδομένα της εφαρμογής και συγκρίνοντας τα αποτελέσματα.

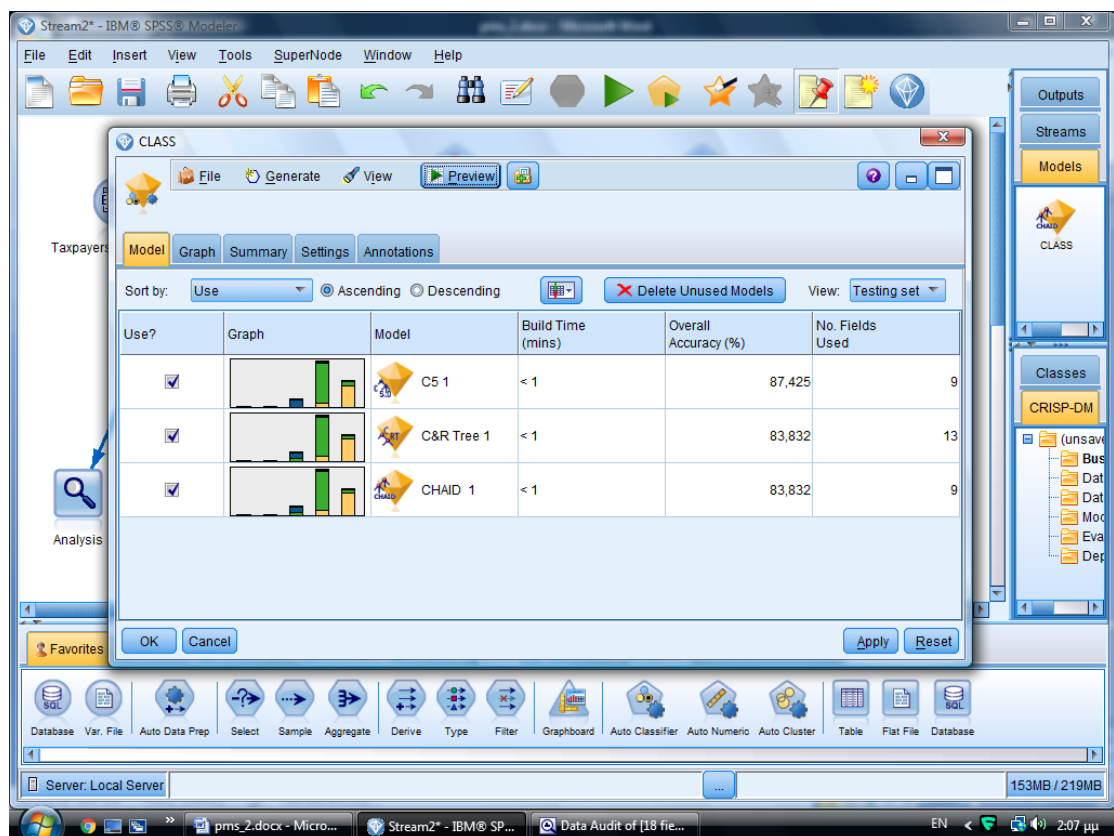
Δημιουργούμε την παρακάτω ροή για δούμε το αποτέλεσμα.



Εικόνα 7.16 Ροή για υπόδειξη των τριών καλύτερων αλγορίθμων

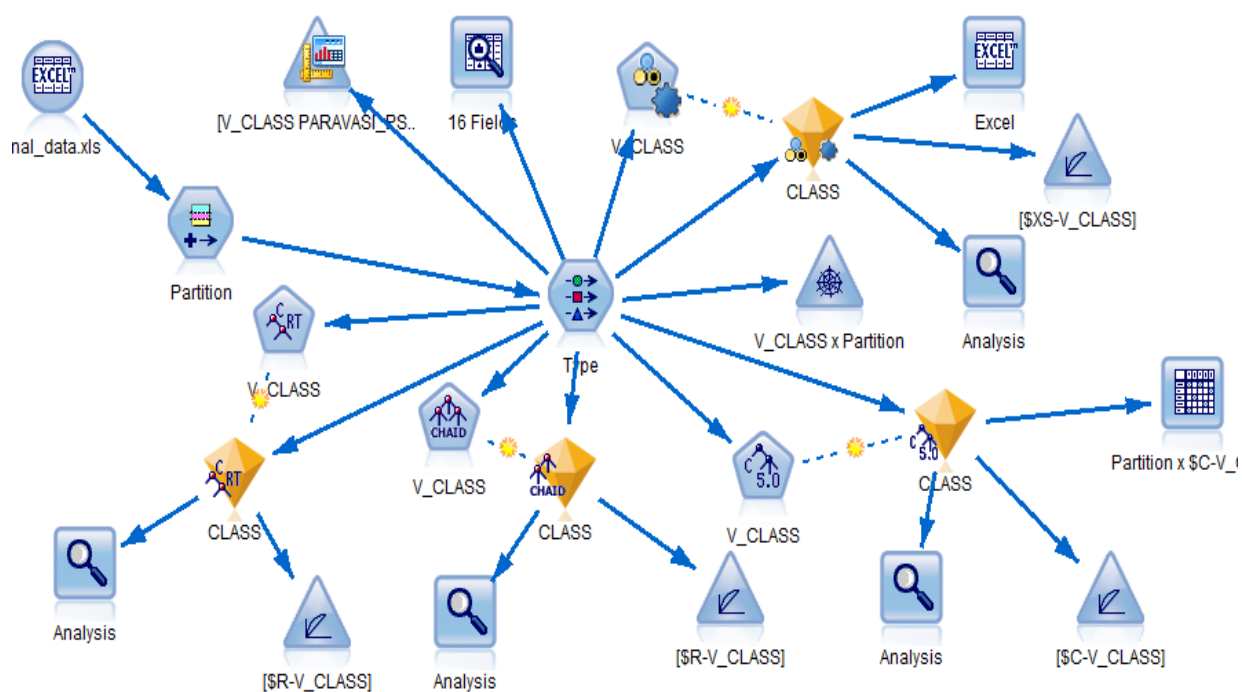
Όπως διαπιστώνουμε και στην εικόνα που ακολουθεί μετά από έρευνα κατέληξε στους παρακάτω με την εξής σειρά:

- C5.1
- C&R Tree
- CHAID



Εικόνα 7.17 Καλύτεροι Αλγόριθμοι

Ακολουθεί η ροή που υλοποιεί τα παραπάνω και με τους τρεις αλγορίθμους.

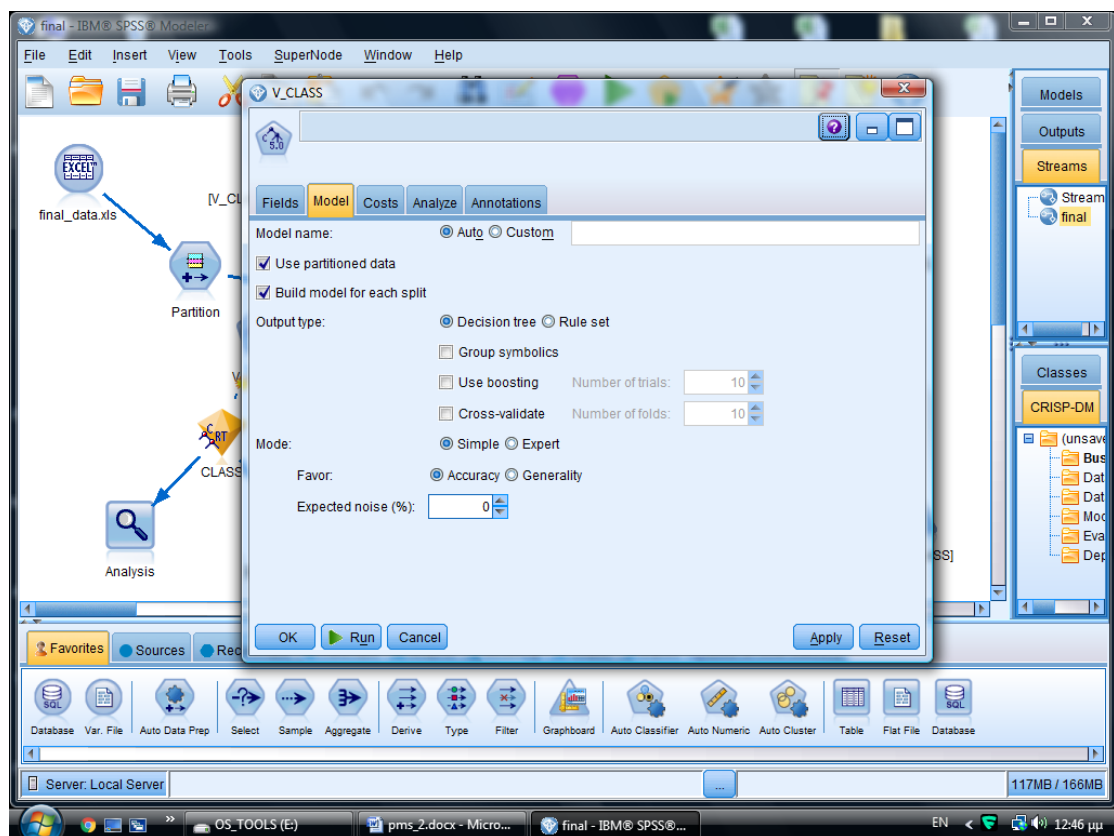


Εικόνα 7.18 Τελικό Μοντέλο

Ακολουθεί αναλυτική μελέτη των τριών προτεινόμενων από το εργαλείο αλγορίθμων.

#### 7.4.1 Εφαρμογή Αλγορίθμου C5.1

Πριν τρέξουμε τον αλγόριθμο επιλέγουμε βασικές παραμέτρους όπως φαίνεται στην εικόνα που ακολουθεί:



Εικόνα 7.19 Παράμετροι αλγορίθμου C5.1.

Ακολουθεί δείγμα από τα αποτελέσματα:

Preview from CLASS Node (25 fields, 10 records) #1

|    | PL    | RTM_2M_PL | BEB_PS       | PAYM_PS      | BEB_ANA_PLRM | ANAKR_DHL_FOR | PARAVASI_PS | V_CLASS | Partition  | \$C-V_CLASS | \$CC-V_CLASS |
|----|-------|-----------|--------------|--------------|--------------|---------------|-------------|---------|------------|-------------|--------------|
| 1  | 0.000 | 1.000     | 876575.960   | 64836.430    | 13.520       | 0.000         | 1253.000    | 1.000   | 1_Training | 1.000       | 0.931        |
| 2  | 0.000 | 0.000     | 125746.010   | 23792.500    | 5.290        | 0.000         | 114525.000  | 1.000   | 1_Training | 1.000       | 0.931        |
| 3  | 0.000 | 0.000     | 22354653.320 | 2450751.8... | 9.120        | 0.000         | 114526.000  | 1.000   | 1_Training | 1.000       | 0.931        |
| 4  | 0.000 | 0.000     | 287016.370   | 19071.160    | 15.050       | 0.000         | 1258.000    | 1.000   | 2_Testing  | 2.000       | 0.774        |
| 5  | 0.000 | 0.000     | 318988.390   | 27341.310    | 11.670       | 0.000         | 336956.000  | 1.000   | 1_Training | 1.000       | 0.931        |
| 6  | 0.000 | 0.000     | 2718286.480  | 1428251.9... | 1.900        | 200.000       | 11123.000   | 1.000   | 1_Training | 1.000       | 0.667        |
| 7  | 0.000 | 1.000     | 1282845.240  | 511248.690   | 2.510        | 0.000         | 1900.000    | 1.000   | 2_Testing  | 1.000       | 0.750        |
| 8  | 0.000 | 0.000     | 20907467.420 | 1770097.4... | 11.810       | 0.000         | 2536.000    | 1.000   | 1_Training | 1.000       | 0.931        |
| 9  | 0.000 | 0.000     | 2069808.560  | 179672.960   | 11.520       | 0.000         | 900.000     | 1.000   | 1_Training | 1.000       | 0.931        |
| 10 | 0.000 | 0.000     | 5101917.830  | 602271.330   | 8.470        | 0.000         | 22369.000   | 1.000   | 1_Training | 1.000       | 0.931        |

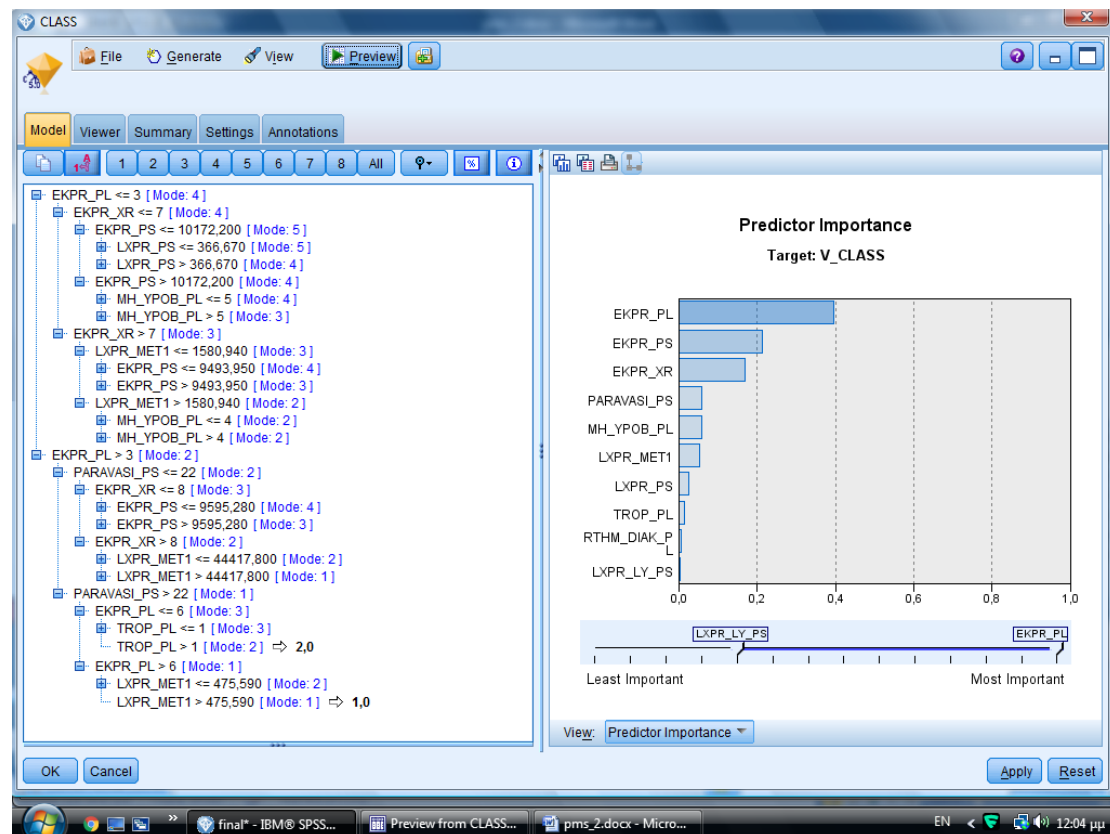
Εικόνα 7.20 Αποτελέσματα πρόβλεψης αλγορίθμου C5.1

Όπου φαίνεται η πρόβλεψη (\$C\_V\_CLASS) και το διάστημα επιστοσύνης (\$CC\_V\_CLASS).

Το δέντρο απόφασης που δημιουργείται έχει βάθος 14 και χρησιμοποιούνται 14 χαρακτηριστικά σαν μεταβλητές εισόδου όπως φαίνεται στο Παράρτημα Α (1).

Στην επόμενη εικόνα φαίνεται συνοπτικά το δέντρο απόφασης που δημιουργείται. Αναλυτικά υπάρχει στο Παράρτημα Α (2 και 3).

Επίσης εμφανίζεται διάγραμμα των χαρακτηριστικών που δείχνει τη σημασία τους στη πρόβλεψη.



Εικόνα 7.21 Δέντρο απόφασης αλγορίθμου C5.1 συνοπτικά και διάγραμμα σπουδαιότητας χαρακτηριστικών στη πρόβλεψη

Είναι φανερό πως το «Πλήθος εκπροθέσμων δηλώσεων» θεωρείται το πιο σπουδαίο χαρακτηριστικό για τον C5.1 αλγόριθμο με αμέσως επόμενο τον «Επιπρόσθετο φόρο εκπροθέσμων δηλώσεων». Ακολουθεί ο «Συνολικός χρόνος εκπροθέσμων δηλώσεων» με επόμενο τις Παραβάσεις.

Ο επόμενος πίνακας δείχνει τη βαρύτητα των γνωρισμάτων στη πρόβλεψη της εξαρτημένης μεταβλητής V\_CLASS.

| Nodes        | Importance | Importance | V4           | V5     |
|--------------|------------|------------|--------------|--------|
| LXPR_LY_PS   | 0,0031     | 0,0031     | LXPR_LY_PS   | 0,0031 |
| RTHM_DIAK_PL | 0,0059     | 0,0059     | RTHM_DIAK_PL | 0,0059 |



|             |        |        |             |        |
|-------------|--------|--------|-------------|--------|
| TROP_PL     | 0,0149 | 0,0149 | TROP_PL     | 0,0149 |
| LXPR_PS     | 0,0259 | 0,0259 | LXPR_PS     | 0,0259 |
| LXPR_MET1   | 0,0538 | 0,0538 | LXPR_MET1   | 0,0538 |
| MH_YPOB_PL  | 0,0575 | 0,0575 | MH_YPOB_PL  | 0,0575 |
| PARAVASI_PS | 0,0584 | 0,0584 | PARAVASI_PS | 0,0584 |
| EKPR_XR     | 0,1686 | 0,1686 | EKPR_XR     | 0,1686 |
| EKPR_PS     | 0,2123 | 0,2123 | EKPR_PS     | 0,2123 |
| EKPR_PL     | 0,3947 | 0,3947 | EKPR_PL     | 0,3947 |

Πίνακας 4. Βαρύτητα Γνωρισμάτων αλγορίθμου C5.1

Σύμφωνα με την επόμενη εικόνα διαπιστώνουμε ότι:

- ✓ Σωστές προβλέψεις στα δεδομένα εκπαίδευσης: 90.88%
- ✓ Λάθος προβλέψεις στα δεδομένα εκπαίδευσης: 9.12%
- ✓ Σωστές προβλέψεις στα δεδομένα δοκιμής: 84.31%
- ✓ Λάθος προβλέψεις στα δεδομένα δοκιμής: 15.69%

Results for output field V\_CLASS

Comparing \$C-V\_CLASS with V\_CLASS

| 'Partition' | 1_Training |        | 2_Testing |        |
|-------------|------------|--------|-----------|--------|
| Correct     | 3.439      | 90,88% | 2.208     | 84,31% |
| Wrong       | 345        | 9,12%  | 411       | 15,69% |
| Total       | 3.784      |        | 2.619     |        |

Εικόνα 7.22 Αποτελέσματα αλγορίθμου C5.1.

Ακολουθεί ο Πίνακας Σύγχυσης (Confusion matrix): Αποτελεί ένα εργαλείο για την μέτρηση της ποιότητας των αλγορίθμων εξόρυξης γνώσης. Κάθε γραμμή και στήλη του πίνακα αντιστοιχεί σε μία κλάση ταξινόμησης. Οι γραμμές του πίνακα αντιστοιχούν στις τιμές που θα έπρεπε να προκύψουν από την εφαρμογή του αλγορίθμου εξόρυξης γνώσης ενώ οι στήλες αντιστοιχούν στις πραγματικές τιμές.

| V_CLASS | 1.0 | 2.0 | 3.0  | 4.0  | 5.0 |
|---------|-----|-----|------|------|-----|
| 1.0     | 129 | 23  | 0    | 0    | 0   |
| 2.0     | 3   | 602 | 107  | 5    | 0   |
| 3.0     | 0   | 85  | 1223 | 89   | 0   |
| 4.0     | 0   | 11  | 153  | 3357 | 24  |
| 5.0     | 0   | 0   | 0    | 256  | 336 |

Πίνακας 5. Confusion matrix αλγορίθμου C5.1

Στη διαγώνιο του παραπάνω πίνακα εμφανίζονται οι σωστές προβλέψεις.

Δηλαδή:

- Πρόκειται για έναν MxM πίνακα, όπου το (i,j) στοιχείο του ισούται με το πλήθος των σημείων που, ενώ προέρχονται από την κλάση i, καταχωρούνται στην κλάση j.
- Δίνει πληροφορίες σχετικά με το αν κάποιες κλάσεις έχουν τη τάση να συγχέονται με άλλες κλάσεις.

Με τη βοήθεια του Confusion matrix θα υπολογίσουμε τα παρακάτω:

- Συνολική Ακρίβεια (Overall Accuracy): Είναι το άθροισμα των σωστών προβλέψεων (διαγώνιων στοιχείων) προς το πλήθος των σημείων του συνόλου δεδομένων:

$$(129+602+1223+3357+336) / 6403 = 0,8819 \text{ άρα } 88,19\%.$$

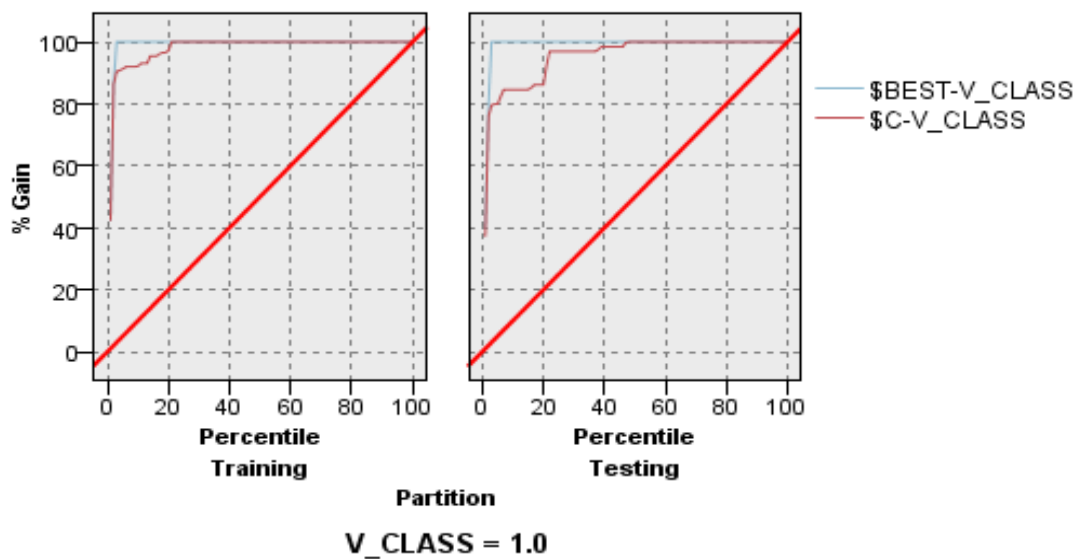
- Ανάκληση (Recall): Το ποσοστό των δεδομένων που προέρχονται από την κλάση i και ταξινομούνται σωστά στη κλάση αυτή.
- Ακρίβεια (Precision): Το ποσοστό των δεδομένων που ταξινομούνται στη κλάση i και πράγματι ανήκουν στη κλάση αυτή.

Ακολουθεί πίνακας με τους παραπάνω υπολογισμούς:

|              | Predicted Class |                          |                          |                            |                            |                          |        |                            |
|--------------|-----------------|--------------------------|--------------------------|----------------------------|----------------------------|--------------------------|--------|----------------------------|
|              | Κλάσεις         | 1                        | 2                        | 3                          | 4                          | 5                        | Σύνολο | Recall                     |
| Actual Class | 1               | 129                      | 23                       | 0                          | 0                          | 0                        | 152    | $\frac{129}{152} = 0,85$   |
|              | 2               | 3                        | 602                      | 107                        | 5                          | 0                        | 717    | $\frac{602}{717} = 0,84$   |
|              | 3               | 0                        | 85                       | 1223                       | 89                         | 0                        | 1397   | $\frac{1223}{1397} = 0,88$ |
|              | 4               | 0                        | 11                       | 153                        | 3357                       | 24                       | 3545   | $\frac{3357}{3545} = 0,95$ |
|              | 5               | 0                        | 0                        | 0                          | 256                        | 336                      | 592    | $\frac{336}{592} = 0,57$   |
|              | Σύνολο          | 132                      | 721                      | 1483                       | 3707                       | 360                      | 6403   |                            |
|              | Precision       | $\frac{129}{132} = 0,98$ | $\frac{602}{721} = 0,83$ | $\frac{1223}{1483} = 0,82$ | $\frac{3357}{3707} = 0,91$ | $\frac{336}{360} = 0,93$ |        |                            |

Πίνακας 6. Ανάκληση και Ακρίβεια αλγορίθμου C5.1

Η καμπύλη κέρδους έχει την παρακάτω μορφή:



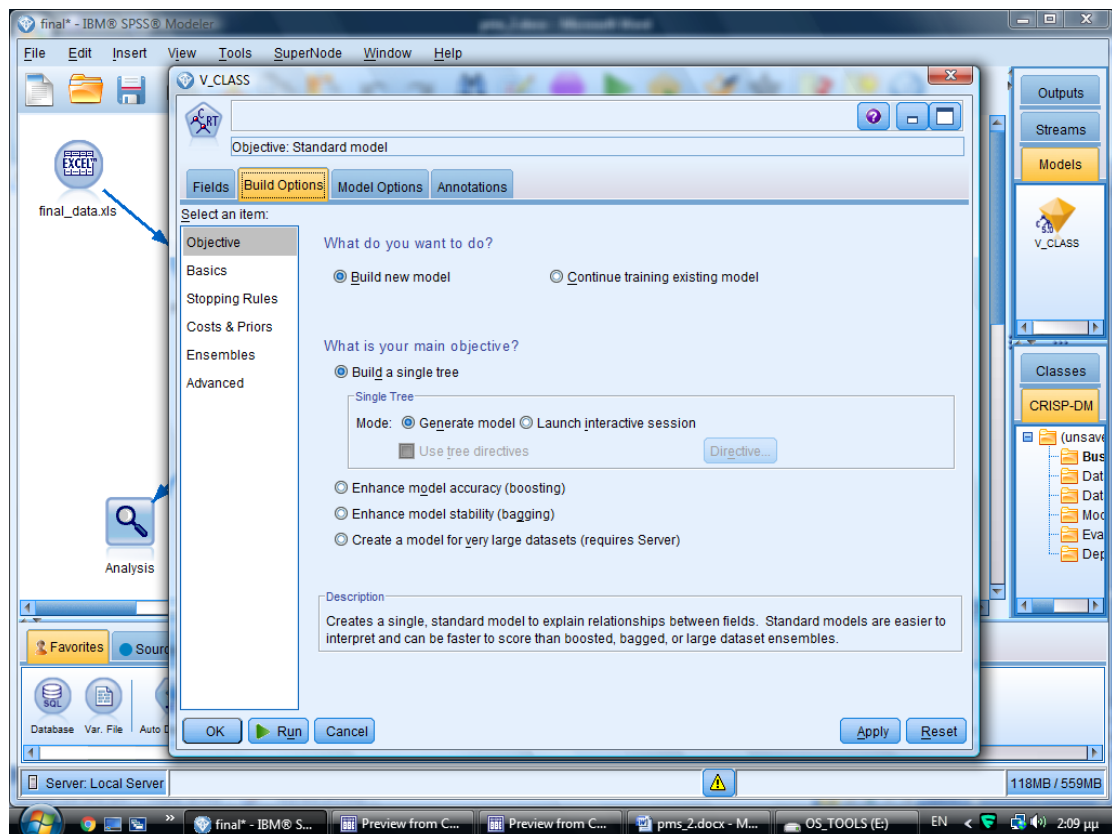
Εικόνα 7.23 Καμπύλη κέρδους αλγορίθμου C5.1.

Η γραφική παράσταση αρχίζει πάντα στο 0% και τελειώνει στο 100% καθώς πηγαίνει από τα αριστερά προς τα δεξιά. Για ένα καλό μοντέλο η καμπύλη του κέρδους θα αυξηθεί απότομα προς το 100%. Ένα μοντέλο που δεν παρέχει πληροφορίες θα ακολουθήσει τη διαγώνιο (κόκκινη γραμμή) από κάτω αριστερά προς τα επάνω δεξιά (baseline).

Επίσης, είναι σημαντικό οι δυο καμπύλες των δεδομένων εκπαίδευσης και δεδομένων δοκιμής να είναι παρόμοιες.

#### 7.4.2 Εφαρμογή Αλγορίθμου C&R Tree

Επιλέγουμε παραμέτρους πριν το χτίσιμο όπως φαίνεται στην εικόνα που ακολουθεί:



Εικόνα 7.24 Παράμετροι αλγορίθμου C&R Tree

Ακολουθεί δείγμα από τα αποτελέσματα:

Preview from CLASS Node (25 fields, 10 records) #4

|    | HM_2M_PL | BEB_PS       | PAYM_PS      | BEB_ANA_PLRM | ANAKR_DHL_FOR | PARAVASI_PS | V_CLASS | Partition  | \$R-V_CLASS | \$RC-V_CLASS |
|----|----------|--------------|--------------|--------------|---------------|-------------|---------|------------|-------------|--------------|
| 1  | 1.000    | 876575.960   | 64836.430    | 13.520       | 0.000         | 1253.000    | 1.000   | 1_Training | 1.000       | 0.703        |
| 2  | 0.000    | 125746.010   | 23792.500    | 5.290        | 0.000         | 114525.000  | 1.000   | 1_Training | 1.000       | 0.703        |
| 3  | 0.000    | 22354653.320 | 2450751.8... | 9.120        | 0.000         | 114526.000  | 1.000   | 1_Training | 1.000       | 0.703        |
| 4  | 0.000    | 287016.370   | 19071.160    | 15.050       | 0.000         | 1258.000    | 1.000   | 2_Testing  | 1.000       | 0.703        |
| 5  | 0.000    | 318988.390   | 27341.310    | 11.670       | 0.000         | 336956.000  | 1.000   | 1_Training | 1.000       | 0.703        |
| 6  | 0.000    | 2718286.480  | 1428251.9... | 1.900        | 200.000       | 11123.000   | 1.000   | 1_Training | 1.000       | 0.703        |
| 7  | 1.000    | 1282845.240  | 511248.690   | 2.510        | 0.000         | 1900.000    | 1.000   | 2_Testing  | 1.000       | 0.703        |
| 8  | 0.000    | 20907467.420 | 1770097.4... | 11.810       | 0.000         | 2536.000    | 1.000   | 1_Training | 1.000       | 0.703        |
| 9  | 0.000    | 2069808.560  | 179672.960   | 11.520       | 0.000         | 900.000     | 1.000   | 1_Training | 1.000       | 0.703        |
| 10 | 0.000    | 5101917.830  | 602271.330   | 8.470        | 0.000         | 22369.000   | 1.000   | 1_Training | 1.000       | 0.703        |

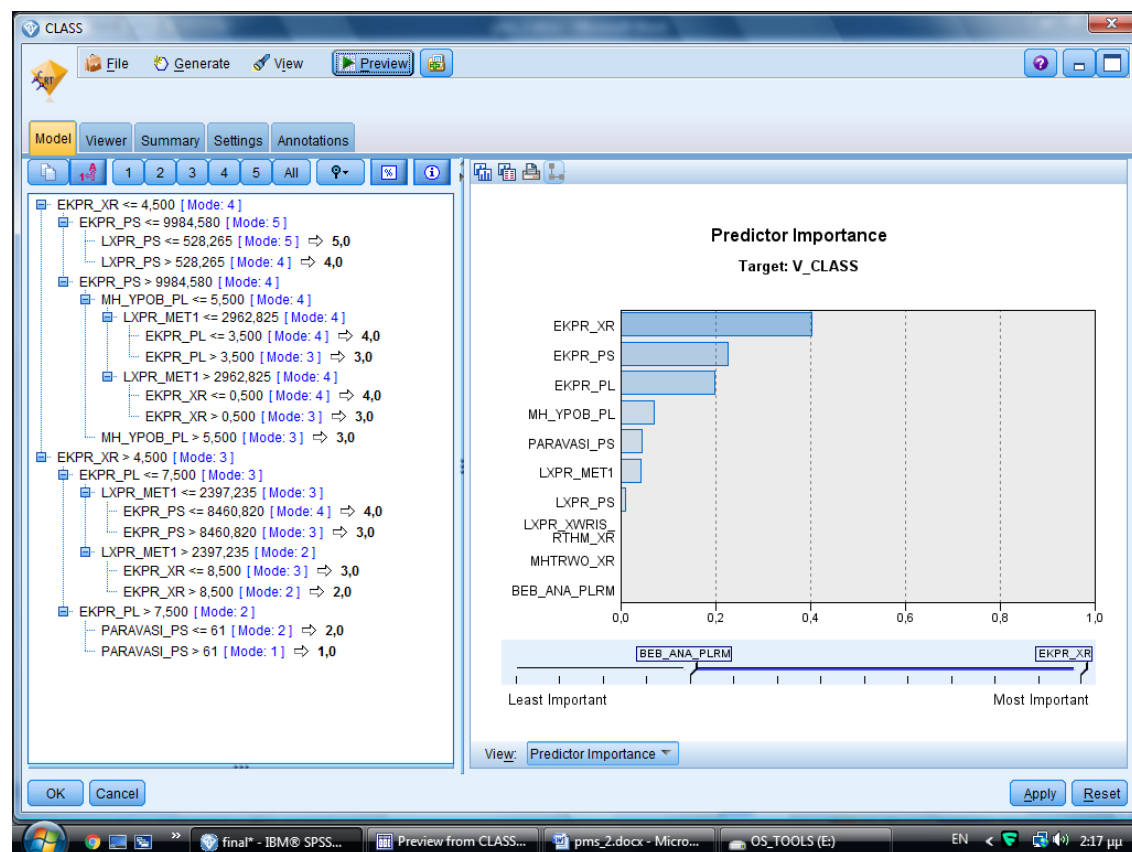
Εικόνα 7.25 Αποτελέσματα πρόβλεψης αλγορίθμου C&R Tree

Όπου φαίνεται η πρόβλεψη (\$RC\_V\_CLASS) και το διάστημα επιστοσύνης της πρόβλεψης (\$R\_V\_CLASS).

Το δέντρο απόφασης που δημιουργείται έχει βάθος 14 και χρησιμοποιούνται 14 χαρακτηριστικά σαν μεταβλητές εισόδου όπως φαίνεται στο Παράρτημα Β (1).

Στην επόμενη εικόνα φαίνεται συνοπτικά το δέντρο απόφασης που δημιουργείται. Αναλυτικά υπάρχει στο Παράρτημα Β (2 και 3). Το δέντρο απόφασης αυτού του αλγορίθμου είναι εμφανώς πολύ πιο μικρό από το δέντρο απόφασης του προηγούμενου.

Επίσης με τη βοήθεια του εργαλείου διαπιστώνουμε τη βαρύτητα των γνωρισμάτων στη πρόβλεψη της εξαρτημένης μεταβλητής CLASS.



Εικόνα 7.26 Δέντρο απόφασης αλγορίθμου C&R Tree συνοπτικά και διάγραμμα σπουδαιότητας χαρακτηριστικών στη πρόβλεψη

Σε αυτό το μοντέλο το πιο σπουδαίο χαρακτηριστικό ο «Συνολικός χρόνος εκπροθέσμων δηλώσεων» με αμέσως επόμενο τον «Επιπρόσθετο φόρο εκπροθέσμων δηλώσεων». Ακολουθεί το «Πλήθος εκπροθέσμων δηλώσεων».

Τέταρτο στη λίστα είναι το «Πλήθος δηλώσεων που δεν υποβλήθηκαν» με επόμενο τις Παραβάσεις. Ο επόμενος πίνακας δείχνει τη σημασία των χαρακτηριστικών στη πρόβλεψη.

| Nodes              | Importance | Importance | V4                 | V5      |
|--------------------|------------|------------|--------------------|---------|
| BEB_ANA_PLRM       | 0,0009     | 8,84E-4    | BEB_ANA_PLRM       | 8,84E-4 |
| MHTRWO_XR          | 0,0009     | 8,84E-4    | MHTRWO_XR          | 8,84E-4 |
| LXPR_XWRIS_RTHM_XR | 0,0009     | 8,84E-4    | LXPR_XWRIS_RTHM_XR | 8,84E-4 |
| LXPR_PS            | 0,0098     | 0,0098     | LXPR_PS            | 0,0098  |
| LXPR_MET1          | 0,0421     | 0,0421     | LXPR_MET1          | 0,0421  |
| PARAVASI_PS        | 0,0436     | 0,0436     | PARAVASI_PS        | 0,0436  |
| MH_YPOB_PL         | 0,0713     | 0,0713     | MH_YPOB_PL         | 0,0713  |
| EKPR_PL            | 0,1978     | 0,1978     | EKPR_PL            | 0,1978  |
| EKPR_PS            | 0,2259     | 0,2259     | EKPR_PS            | 0,2259  |
| EKPR_XR            | 0,4033     | 0,4033     | EKPR_XR            | 0,4033  |

Πίνακας 7. Βαρύτητα Γνωρισμάτων αλγορίθμου C&R Tree

Σύμφωνα με την επόμενηεικόνα διαπιστώνουμε ότι:

- ✓ Σωστές προβλέψεις στα δεδομένα εκπαίδευσης: 79.57%
- ✓ Λάθος προβλέψεις στα δεδομένα εκπαίδευσης: 20.43%
- ✓ Σωστές προβλέψεις στα δεδομένα δοκιμής: 77.43%
- ✓ Λάθος προβλέψεις στα δεδομένα δοκιμής: 22.57%

Διαπιστώνουμε μείωση στο ποσοστό σωστής πρόβλεψης τόσο στα δεδομένα εκπαίδευσης όσο και στα δοκιμής σε σχέση με τον προηγούμενο αλγόριθμο.

Results for output field V\_CLASS

Comparing \$R-V\_CLASS with V\_CLASS

| 'Partition' | 1_Training |        | 2_Testing |        |
|-------------|------------|--------|-----------|--------|
| Correct     | 3.011      | 79,57% | 2.028     | 77,43% |
| Wrong       | 773        | 20,43% | 591       | 22,57% |
| Total       | 3.784      |        | 2.619     |        |

Εικόνα 7.27 Αποτελέσματα αλγορίθμου C&R Tree

Ακολουθεί ο Πίνακας Σύγχυσης (Confusion matrix):

| V_CLASS | 1.0 | 2.0 | 3.0 | 4.0  | 5.0 |
|---------|-----|-----|-----|------|-----|
| 1.0     | 126 | 23  | 3   | 0    | 0   |
| 2.0     | 34  | 443 | 221 | 19   | 0   |
| 3.0     | 6   | 134 | 953 | 297  | 7   |
| 4.0     | 0   | 7   | 275 | 3177 | 86  |
| 5.0     | 0   | 0   | 4   | 248  | 340 |

Πίνακας 8. Confusion matrix αλγορίθμου C&R Tree

Με τη βοήθεια του Confusion matrix θα υπολογίσουμε τα παρακάτω:

- Συνολική Ακρίβεια (Overall Accuracy):

$$(126+443+953+3177+340) / 6403 = 0,7870 \text{ άρα } 78,70\%.$$

Παρατηρούμε ότι η συνολική ακρίβεια μειώθηκε σε σχέση με το προηγούμενο μοντέλο.

- Ανάκληση (Recall) και

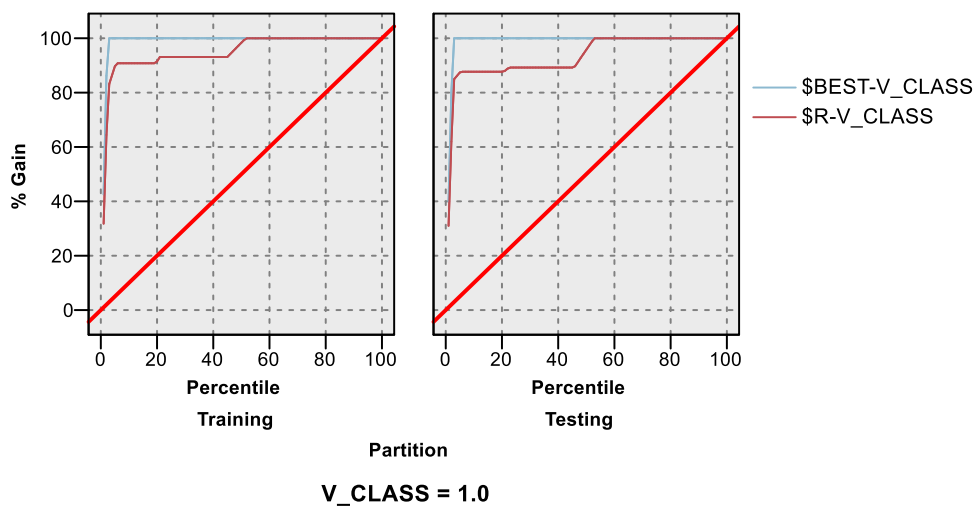


- Ακρίβεια (Precision)

| •<br>•          | Predicted Class |                          |                          |                           |                            |                          |        |                            |
|-----------------|-----------------|--------------------------|--------------------------|---------------------------|----------------------------|--------------------------|--------|----------------------------|
|                 | Κλάσεις         | 1                        | 2                        | 3                         | 4                          | 5                        | Σύνολο | Recall                     |
| Actual<br>Class | 1               | 126                      | 23                       | 3                         | 0                          | 0                        | 152    | $\frac{126}{152} = 0,83$   |
|                 | 2               | 34                       | 443                      | 221                       | 19                         | 0                        | 717    | $\frac{443}{717} = 0,62$   |
|                 | 3               | 6                        | 134                      | 953                       | 297                        | 7                        | 1397   | $\frac{953}{1397} = 0,68$  |
|                 | 4               | 0                        | 7                        | 275                       | 3177                       | 86                       | 3545   | $\frac{3177}{3545} = 0,90$ |
|                 | 5               | 0                        | 0                        | 4                         | 248                        | 340                      | 592    | $\frac{340}{592} = 0,57$   |
|                 | Σύνολο          | 166                      | 607                      | 1456                      | 3741                       | 433                      | 6403   |                            |
|                 | Precision       | $\frac{126}{166} = 0,76$ | $\frac{443}{607} = 0,73$ | $\frac{953}{1456} = 0,65$ | $\frac{3177}{3740} = 0,85$ | $\frac{340}{433} = 0,76$ |        |                            |

Πίνακας 9. Ανάκληση και Ακρίβεια αλγορίθμου C&R Tree

Η καμπύλη κέρδους έχει την παρακάτω μορφή:

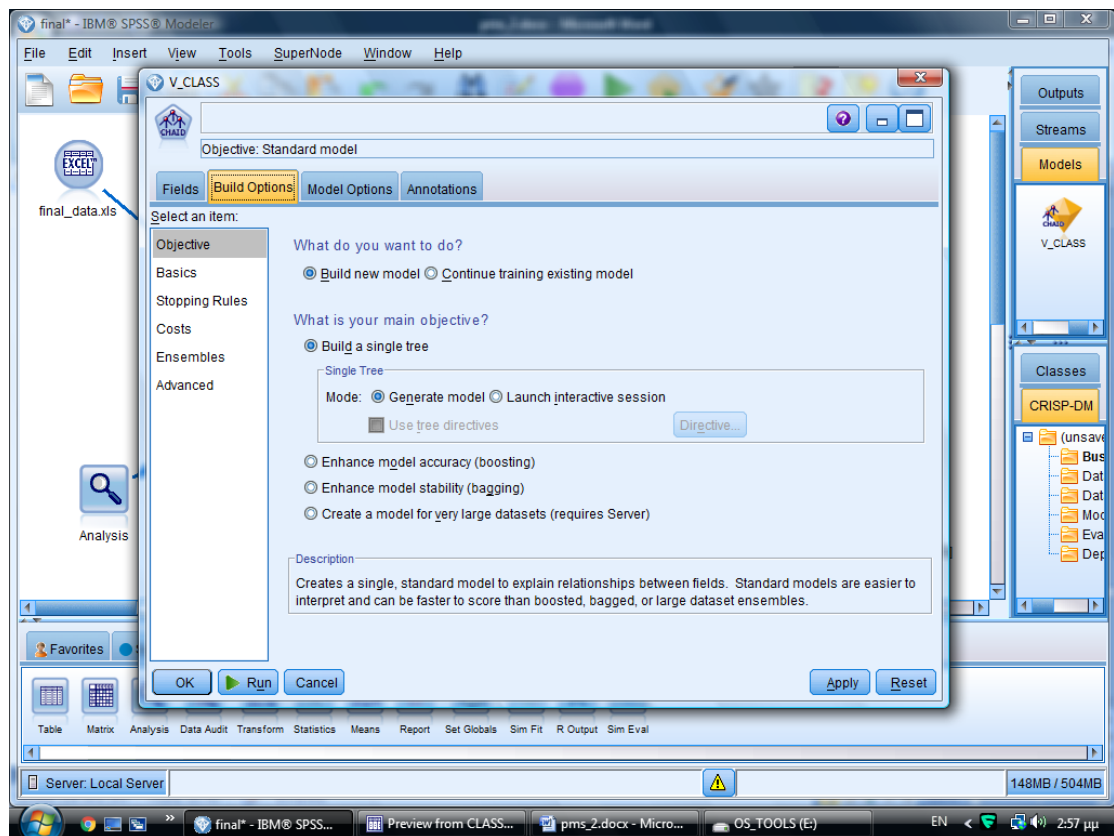


Εικόνα 7.28 Καμπύλη κέρδους αλγορίθμου C&R Tree

Παρατηρούμε ότι οι καμπύλες κέρδους μεταξύ δεδομένων εκπαίδευσης και δοκιμής δεν έχουν μεγάλη απόκλιση.

#### 7.4.2 Εφαρμογή Αλγορίθμου CHAID

Μελετάμε τα ίδια όπως και στους προηγούμενους δυο αλγορίθμους. Επιλέγουμε παραμέτρους πριν το χτίσιμο όπως φαίνεται στην εικόνα που ακολουθεί:



Εικόνα 7.29 Παράμετροι αλγορίθμου CHAID

Ακολουθεί δείγμα από τα αποτελέσματα:

|    | THM_2M_PL | BEB_PS       | PAYM_PS      | BEB_ANA_PLRM | ANAKR_DHL_FOR | PARAVASI_PS | V_CLASS | Partition  | \$R-V_CLASS | \$SRC-V_CLASS |
|----|-----------|--------------|--------------|--------------|---------------|-------------|---------|------------|-------------|---------------|
| 1  | 1.000     | 876575.960   | 64836.430    | 13.520       | 0.000         | 1253.000    | 1.000   | 1_Training | 1.000       | 0.931         |
| 2  | 0.000     | 125746.010   | 23792.500    | 5.290        | 0.000         | 114525.000  | 1.000   | 1_Training | 1.000       | 0.931         |
| 3  | 0.000     | 22354653.320 | 2450751.8... | 9.120        | 0.000         | 114526.000  | 1.000   | 1_Training | 1.000       | 0.931         |
| 4  | 0.000     | 287016.370   | 19071.160    | 15.050       | 0.000         | 1258.000    | 1.000   | 2_Testing  | 2.000       | 0.518         |
| 5  | 0.000     | 318988.390   | 27341.310    | 11.670       | 0.000         | 336956.000  | 1.000   | 1_Training | 1.000       | 0.931         |
| 6  | 0.000     | 2718286.480  | 1428251.9... | 1.900        | 200.000       | 11123.000   | 1.000   | 1_Training | 2.000       | 0.518         |
| 7  | 1.000     | 1282845.240  | 511248.690   | 2.510        | 0.000         | 1900.000    | 1.000   | 2_Testing  | 2.000       | 0.518         |
| 8  | 0.000     | 20907467.420 | 1770097.4... | 11.810       | 0.000         | 2536.000    | 1.000   | 1_Training | 1.000       | 0.931         |
| 9  | 0.000     | 2069808.560  | 179672.960   | 11.520       | 0.000         | 900.000     | 1.000   | 1_Training | 1.000       | 0.931         |
| 10 | 0.000     | 5101917.830  | 602271.330   | 8.470        | 0.000         | 22369.000   | 1.000   | 1_Training | 1.000       | 0.931         |

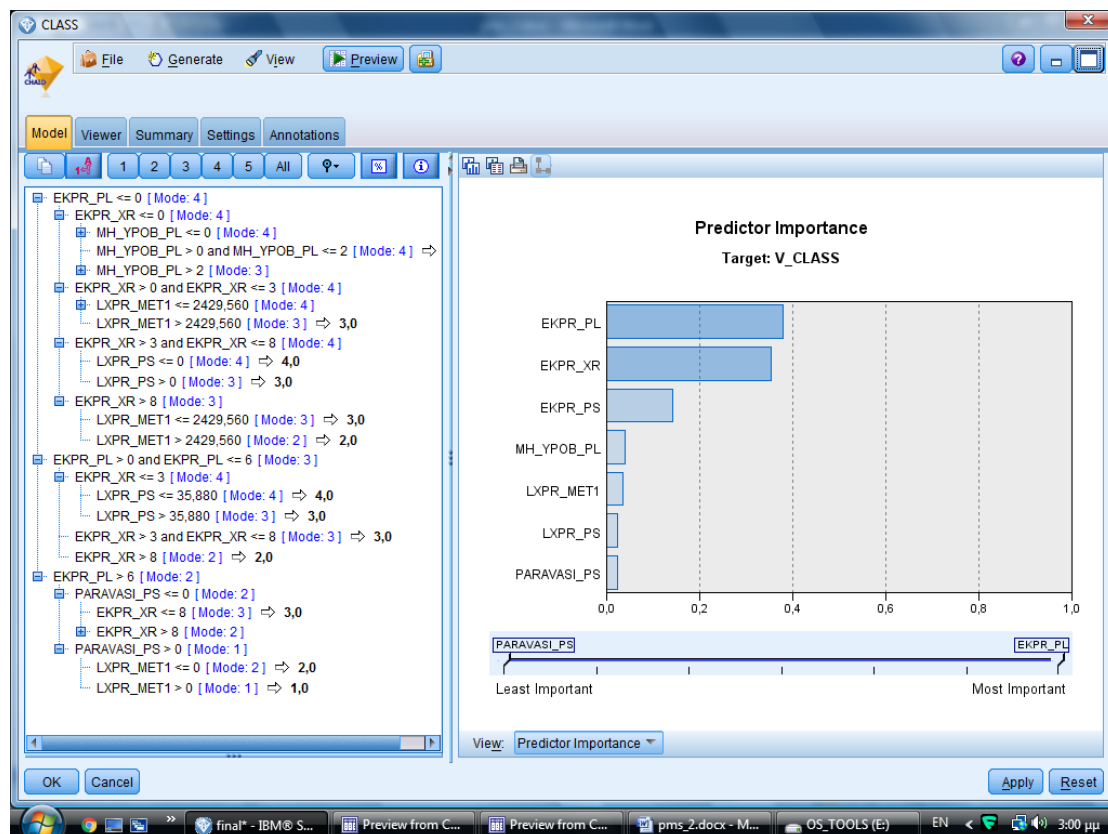
Εικόνα 7.30 Αποτελέσματα πρόβλεψης αλγορίθμου CHAID

Όπου φαίνεται η πρόβλεψη (\$R\_V\_CLASS) και το διάστημα επιστοσύνης της πρόβλεψης (\$SRC\_V\_CLASS).

Το δέντρο απόφασης που δημιουργείται έχει βάθος 5 και χρησιμοποιούνται 7 χαρακτηριστικά σαν μεταβλητές εισόδου όπως φαίνεται στο Παράρτημα Γ (1).

Στην επόμενη εικόνα φαίνεται συνοπτικά το δέντρο απόφασης που δημιουργείται. Αναλυτικά υπάρχει στο Παράρτημα Γ (2 και 3). Το δέντρο απόφασης αυτού του αλγορίθμου είναι εμφανώς πολύ πιο μικρό από το δέντρο απόφασης του προηγούμενου.

Επίσης με τη βοήθεια του εργαλείου διαπιστώνουμε τη βαρύτητα των γνωρισμάτων στη πρόβλεψη της εξαρτημένης μεταβλητής CLASS.



Εικόνα 7.32 Δέντρο απόφασης αλγορίθμου CHAID συνοπτικά και διάγραμμα σπουδαιότητας χαρακτηριστικών στη πρόβλεψη

Σε αυτό το μοντέλο το πιο σπουδαίο χαρακτηριστικό το «Πλήθος εκπροθέσμων δηλώσεων». Ακολουθεί ο «Συνολικός χρόνος εκπροθέσμων δηλώσεων» με αμέσως επόμενο τον «Επιπρόσθετο φόρο εκπροθέσμων δηλώσεων». Τέταρτο στη λίστα είναι το «Πλήθος δηλώσεων που δεν υποβλήθηκαν» με επόμενο τη «Μεταβολή

ληξιπρόθεσμου χρέους». Ο επόμενος πίνακας δείχνει τη βαρύτητα των γνωρισμάτων στη πρόβλεψη.

| Nodes              | Importanc<br>e | Importanc<br>e | V4                 | V5      |
|--------------------|----------------|----------------|--------------------|---------|
| BEB_ANA_PLRM       | 0,0009         | 8,84E-4        | BEB_ANA_PLRM       | 8,84E-4 |
| MHTRWO_XR          | 0,0009         | 8,84E-4        | MHTRWO_XR          | 8,84E-4 |
| LXPR_XWRIS_RTHM_XR | 0,0009         | 8,84E-4        | LXPR_XWRIS_RTHM_XR | 8,84E-4 |
| LXPR_PS            | 0,0098         | 0,0098         | LXPR_PS            | 0,0098  |
| LXPR_MET1          | 0,0421         | 0,0421         | LXPR_MET1          | 0,0421  |
| PARAVASI_PS        | 0,0436         | 0,0436         | PARAVASI_PS        | 0,0436  |
| MH_YPOB_PL         | 0,0713         | 0,0713         | MH_YPOB_PL         | 0,0713  |
| EKPR_PL            | 0,1978         | 0,1978         | EKPR_PL            | 0,1978  |
| EKPR_PS            | 0,2259         | 0,2259         | EKPR_PS            | 0,2259  |
| EKPR_XR            | 0,4033         | 0,4033         | EKPR_XR            | 0,4033  |

Πίνακας 10. Βαρύτητα Γνωρισμάτων αλγορίθμου CHAID

Σύμφωνα με την επόμενηεικόνα διαπιστώνουμε ότι:

- ✓ Σωστές προβλέψεις στα δεδομένα εκπαίδευσης: 78.01%
- ✓ Λάθος προβλέψεις στα δεδομένα εκπαίδευσης: 21.99%
- ✓ Σωστές προβλέψεις στα δεδομένα δοκιμής: 76.10%
- ✓ Λάθος προβλέψεις στα δεδομένα δοκιμής: 23.90%

Διαπιστώνουμε μείωση στο ποσοστό σωστής πρόβλεψης τόσο στα δεδομένα εκπαίδευσης όσο και στα δοκιμής σε σχέση με τον προηγούμενο αλγόριθμο.

Results for output field V\_CLASS

Comparing \$R-V\_CLASS with V\_CLASS

| 'Partition' | 1_Training |        | 2_Testing |       |
|-------------|------------|--------|-----------|-------|
| Correct     | 2.952      | 78,01% | 1.993     | 76,1% |
| Wrong       | 832        | 21,99% | 626       | 23,9% |
| Total       | 3.784      |        | 2.619     |       |

Εικόνα 7.33 Αποτελέσματα αλγορίθμου CHAID

Ακολουθεί ο Πίνακας Σύγχυσης (Confusion matrix):

| V_CLASS | 1.0 | 2.0 | 3.0 | 4.0  | 5.0 |
|---------|-----|-----|-----|------|-----|
| 1.0     | 85  | 64  | 3   | 0    | 0   |
| 2.0     | 3   | 520 | 172 | 22   | 0   |
| 3.0     | 0   | 166 | 820 | 411  | 0   |
| 4.0     | 0   | 13  | 231 | 3295 | 6   |
| 5.0     | 0   | 0   | 5   | 362  | 225 |

Πίνακας 11. Confusion matrix αλγορίθμου CHAID

Με τη βοήθεια του Confusion matrix θα υπολογίσουμε τα παρακάτω:

- Συνολική Ακρίβεια (Overall Accuracy):

$$(85+520+820+3295+225) / 6403 = 0,7723 \text{ άρα } 77,23\%.$$

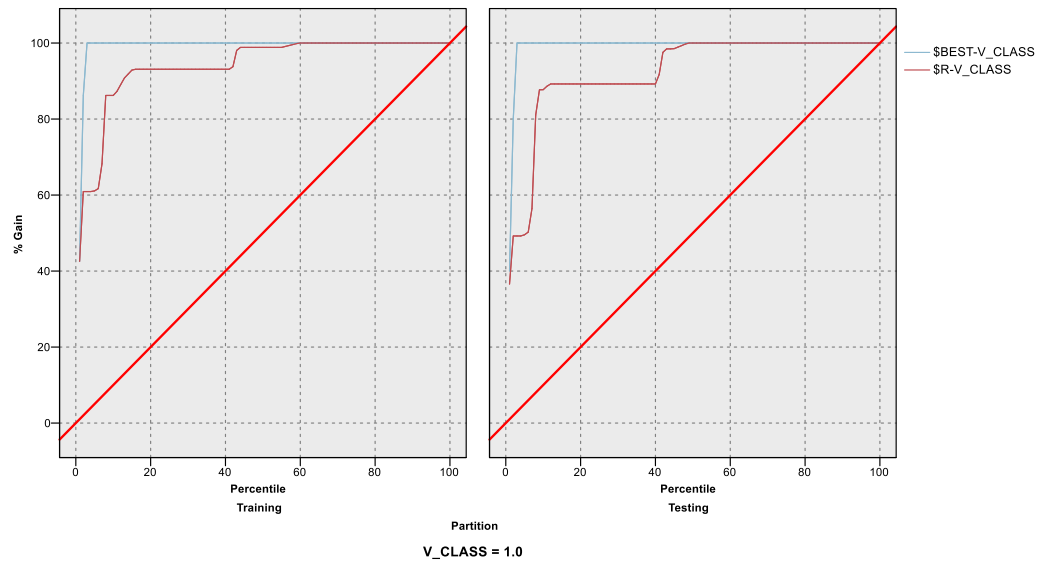
Παρατηρούμε ότι η συνολική ακρίβεια μειώθηκε ελαφρά σε σχέση με το προηγούμενο μοντέλο.

- Ανάκληση (Recall) και
- Ακρίβεια (Precision)

|                     |                        |                         |                          |                           |                            |                          |               |                            |
|---------------------|------------------------|-------------------------|--------------------------|---------------------------|----------------------------|--------------------------|---------------|----------------------------|
| •<br>•              | <b>Predicted Class</b> |                         |                          |                           |                            |                          |               |                            |
| <b>Actual Class</b> | <b>Κλάσεις</b>         | <b>1</b>                | <b>2</b>                 | <b>3</b>                  | <b>4</b>                   | <b>5</b>                 | <b>Σύνολο</b> | <b>Recall</b>              |
|                     | <b>1</b>               | 85                      | 64                       | 3                         | 0                          | 0                        | 152           | <b>85/152<br/>= 0,56</b>   |
|                     | <b>2</b>               | 3                       | 520                      | 172                       | 22                         | 0                        | 717           | <b>520/717<br/>=0,73</b>   |
|                     | <b>3</b>               | 0                       | 166                      | 820                       | 411                        | 0                        | 1397          | <b>820/1397<br/>=0,59</b>  |
|                     | <b>4</b>               | 0                       | 13                       | 231                       | 3295                       | 6                        | 3545          | <b>3295/3545<br/>=0,93</b> |
|                     | <b>5</b>               | 0                       | 0                        | 5                         | 362                        | 225                      | 592           | <b>225/592<br/>=0,38</b>   |
|                     | <b>Σύνολο</b>          | 94                      | 763                      | 1231                      | 4090                       | 231                      | 6403          |                            |
|                     | <b>Precision</b>       | <b>85/94<br/>= 0,90</b> | <b>520/763<br/>=0,68</b> | <b>820/1231<br/>=0,67</b> | <b>3295/4090<br/>=0,81</b> | <b>225/231<br/>=0,97</b> |               |                            |

Πίνακας 12. Ανάκληση και Ακρίβεια αλγορίθμου CHAID

Η καμπύλη κέρδους έχει την παρακάτω μορφή:



Εικόνα 7.34 Καμπύλη κέρδους αλγορίθμου CHAID



### 8.1 Συμπεράσματα

### 8.2 Βελτιώσεις

#### 8.1 Συμπεράσματα

Στη συνέχεια θα συγκρίνουμε τα αποτελέσματα των αλγορίθμων που δοκιμάσαμε.

Τα κριτήρια σύγκρισης θα είναι:

- το ποσοστό σωστών προβλέψεων και
- ο χρόνος δημιουργίας του μοντέλου.

#### Ικανότητα πρόβλεψης

Ο αλγόριθμος C5.1 δείχνει να είναι ο πιο εύστοχος με ποσοστό σωστών προβλέψεων 90,88% στα δεδομένα εκπαίδευσης έναντι των άλλων δύο που τα ποσοστά επιτυχίας είναι αρκετά χαμηλότερα (79,57% στον αλγόριθμο C&R Tree και 78,01% CHAID). Όμως στα δεδομένα δοκιμής το ποσοστό μειώνεται σε 84,31% για τον C5.1 αλγόριθμο, φαίνεται δηλαδή να υπάρχει διαφορά στη σωστή πρόβλεψη μεταξύ δεδομένων εκπαίδευσης και δεδομένων δοκιμής. Αντίθετα οι άλλοι δυο αλγόριθμοι παρουσιάζουν σχετικά σταθερή συμπεριφορά.

Συγκεκριμένα για τον C&R Tree αλγόριθμο ισχύει:

- Επιτυχής πρόβλεψης δεδομένων εκπαίδευσης: 79,57% και
- Επιτυχής πρόβλεψης δεδομένων δοκιμής: 77,43%

Και για τον CHAID αλγόριθμο ισχύει:

- Επιτυχής πρόβλεψης δεδομένων εκπαίδευσης: 78,01% και
- Επιτυχής πρόβλεψης δεδομένων δοκιμής: 76,1%

### Χρόνος δημιουργίας του μοντέλου

Ο αλγόριθμος C5.1 είναι πολύ πιο γρήγορος στο χτίσιμο του μοντέλου σε σχέση με τους άλλους δυο. Χρειάζεται 2 sec σε αντίθεση με τον C&R Tree αλγόριθμο που απαιτεί 14 sec ενώ ο CHAID αλγόριθμος 12 sec.

## 8.2 Βελτιώσεις

Το παραπάνω σύστημα επιδέχεται βελτιώσεις. Αρχικά θα μπορούσε να επεκταθεί στο σύνολο των φορολογουμένων, δηλαδή να συμπεριλάβει και τα Φυσικά Πρόσωπα.

Θα μπορούσε να εμπλουτιστεί ως προς τους δείκτες αξιολόγησης για τον εντοπισμό πιθανής απόκρυψης φορολογητέας ύλης. Υπάρχουν στοιχεία που έχει στη διάθεσή της η φορολογική διοίκηση (Πολ. 1033/2014) και προέρχονται από εξωτερικές πηγές όπως:

- Πιστωτικά και χρηματοδοτικά ιδρύματα
- Ιδρύματα πληρωμών
- Ελληνικά Ταχυδρομεία
- Εταιρείες Παροχής Επενδυτικών Υπηρεσιών
- Επιχειρήσεις ιδιωτικής ασφάλισης
- Ιδιωτικά θεραπευτήρια
- Ιδιωτικά εκπαιδευτήρια
- Εταιρείες σταθερής και κινητής τηλεφωνίας
- Εταιρείες παροχής ηλεκτρικής ενέργειας και ύδρευσης

Τα στοιχεία αυτά μπορούν να διασταυρωθούν με τα δεδομένα των δηλώσεων Φ.Ε. για την ανίχνευση τεκμηρίων.

Επίσης, η σύγκριση των τιμών των δεικτών επιχειρήσεων ενός κλάδου και μιας περιφέρειας με τη μέση τιμή του δείκτη του κλάδου (για το σύνολο της χώρας ή την

περιφέρεια), θα μπορούσε να αποτελέσει ένα ακόμα σημαντικό κριτήριο αξιολόγησης φορολογουμένων.

Τέλος, άμεση εφαρμογή του παραπάνω συστήματος αξιολόγησης και βαθμολόγησης των φορολογουμένων θα μπορούσε να γίνει από το ίδιο το σύστημα, με τον ορισμό ενός ακόμα δείκτη που θα αξιολογούσε τον κάθε φορολογούμενο βάσει του συνολικού βαθμού των προμηθευτών του. Αυτό το κριτήριο θα ενθάρρυνε τους αγοραστές να διατηρήσουν μια εφοδιαστική αλυσίδα από «έντιμους» προμηθευτές. Επίσης θα έδινε ανταγωνιστικό πλεονέκτημα στους συνεπείς προμηθευτές.

## Βιβλιογραφία

1. Apte, C., and Hong, S. J. 1996. Predicting Equity Returns from Securities Data with Minimal Rule Generation. In *Advances in Knowledge Discovery and Data Mining*, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 514–560. Menlo Park, Calif.: AAAI Press
2. Barandela, R., Valdovinos M.R., Sánchez J. S., Francesc J. F., 2004. The Imbalanced Training Sample Problem: Under or over Sampling? Doi 10.1007/978-3-540-27868-9\_88
3. Barbara, D., Faloutsos, C., Hellerstein, J., Ioannidis, Y., Jagadish, H.V., Johnson, T, Ng, R., Poosala, V., Ross, K., and Sevcik, K.V., (1996), The New Jersey Data Reduction Report, Data Engineering Bulletin.
4. Basseville, M., and Nikiforov, I. V. 1993. Detection of Abrupt Changes: Theory and Application. Englewood Cliffs, N.J.: Prentice Hall.
5. Berry, M., Linoff, G. (1996), Data Mining Techniques For marketing, Sales and Customer Support. John Willey & Sons, Inc.
6. Breiman, L., Friedman, J. & Olshen, R. & Stone, C., (1984), Classification and Regression Trees, Wadsworth  
Berndt, D., and Clifford, J. 1996. Finding Patterns in Time Series: A Dynamic Programming Approach. In *Advances in Knowledge Discovery and Data Mining*, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 229–248. Menlo Park, Calif.: AAAI Press.
7. Chow, C.; Rice, S. 1982. Qualified Audit Opinions and Share Prices – an Investigation, *Auditing: A Journal of Practice and Theory*, 1: 35–53.
8. DeBarr D., MITRE, Harwood M., 2005. Relational Mining for Compliance Risk.
9. Duda, R.O., Hart, PE, (1976), Pattern Classification and Scene Analysis. John Wiley and Sons.
10. Dunham, M., «Εισαγωγικά και προηγμένα θέματα εξόρυξης γνώσης από δεδομένα», (2004) (Επιμέλεια ελληνικής έκδοσης: Βερούκιος Β. & Θεοδωρίδης Γ.)

11. Elovici, Y. and Braha, D. (2003), "A decision – theoretic approach to data mining", IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans, 33 (1), pp. 42 – 51.
12. Fayyad, U. M.; Djorgovski, S. G.; and Weir, N. 1996. From Digitized Images to On-Line Catalogs: Data Mining a Sky Survey. AI Magazine 17(2): 51–66.
13. Fayyad, U., Piatetsky – Shapiro, G. and Smyth, P. (1996), "The KDD process for extracting useful knowledge from volumes of data", Communications of the ACM, 39 (11), pp. 27 – 34.
14. Fayyad, U., Piatetsky – Shapiro, G. and Smyth, P. (1996), "From Data Mining to Knowledge Discovery in Databases. Artificial Intelligence Magazine, 17(3), 37-54. doi: 10.1609/aimag.v17i3.1230.
15. Felders, A., Daniels, H. and Holsheimer, M. (2000), "Methodological and practical aspects of data mining", Information & Management, 37, pp. 271–281.
16. Fisher, Vickie L. (1985) Recent Innovations in State Tax Compliance Programs. National Tax Journal. 38, (3), 365-71.
17. Frawley, U., Piatetsky – Shapiro, G. and Matheus, C. J. (1991), *Knowledge discovery in databases*, AAAI/MIT Press.
18. Gupta, R.; Gill, N. 2012. Data Mining Techniques – A Key for Detection of Financial Statement Fraud, International Journal of Computer Science and Information Security, 10(3): 49 – 57.
19. Guyon, O.; Matic, N.; and Vapnik, N. 1996. Discovering Informative Patterns and Data Cleaning. In Advances in Knowledge Discovery and Data Mining, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 181–204. Menlo Park, Calif.: AAAI Press.
20. Han, J., & Kamber, M. (2001). Data Mining Concepts and Techniques. San Francisco, CA: Morgan Kaufmann Publishers.
21. Kallio, M., Back B., 2011. The Self-Organizing Map in Selecting Companies for Tax Audit. DOI 10.1007/978-3-7908-2739-2\_27.
22. Kloesgen, W., and Zytkow, J. 1996. Knowledge Discovery in Databases Terminology. In Advances in Knowledge Discovery and Data Mining, eds. U.

- Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 569–588. Menlo Park, Calif.: AAAI Press.
23. Maimon, O., & Rokach, L. (2005). *The Data Mining and Knowledge Discovery Handbook*. New York, NY: Springer Science + Business Media Inc.
  24. Matheus, C.; Piatetsky-Shapiro, G.; and McNeill, D. 1996. Selecting and Reporting What Is Interesting: The KEfIR Application to Healthcare Data. In *Advances in Knowledge Discovery and Data Mining*, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 495–516. Menlo Park, Calif.: AAAI Press.
  25. Matthieu de Gaudemar , (2017) Taxpayers to be scored on compliance level, The Phnom Penh Post.
  26. Melta, M., Agrawal, R., Rissanen, J., (1996), «SLIQ: A fast scalable classifier for data mining». In EDBT96, Avignon France.
  27. Mitchell, T., (1997), *Machine Learning*. McGraw-Hill.
  28. Nisbet, E. K., Zelenski, J. M., & Murphy, S. A. (2009). The nature relatedness scale: Linking individuals' connection with nature to environmental concern and behavior. *Environment and Behavior*, 41, 715-740.
  29. Quinlan, J.R., (1993), *C4.5: Programs for Machine Learning*. Morgan Kaufman.
  - Rastori, R., Shim, K., (1998), «PUBLIC: A Decision Tree Classifier that Integrates Building and Pruning». *Proceeding of the 24th VLDB Conference*, New York, USA.
  30. Shafer, J., Agrawal, R., Mehta, M., (1996), «SPRINT: A scalable parallel classifier for data mining». In *Proc. of the VLDB Conference*, Bombay, India
  - Vapnik, V., (1995), *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.
  31. Shaw, M. J., Subramaniam, C., Tan, G. W. and Welge, M. E. (2001), "Knowledge management and data mining for marketing", *Decision Support Systems*, 31 (1), pp. 127 -137.
  32. Spathis, C. 2002. Detecting False Financial Statements Using Published Data: Some Evidence from Greece, *Managerial Auditing Journal*, 17(4): 179-191.

33. Spathis, C., Doumpos M., Zopoudinis C., 2002. Detecting falsified financial statements: a comparative study using multicriteria analysis and multivariate statistical techniques.
34. Wang B.X., Japkowic N., 2009. Boosting support vector machines for imbalanced data Sets. Doi: 10.1007/s10115-009-0198-y
35. Weiss, S. I., and Kulikowski, C. 1991. Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning, and Expert Systems. San Francisco, Calif.: Morgan Kaufmann.
36. Wilson, R. 2009. An Examination of Corporate Tax Shelter Participants, The Accounting Review 84: 969–999.
37. Witten, I. H., & Frank, E. (2000). Data Mining Practical Machine Learning Tools and Techniques with Java Implementations. San Francisco, CA: Morgan Kaufmann Publishers.
38. Wu R.S., C.S. Ou b, Hui-ying Lin b, She-I Chang b, Yen C.D., 2012. Using data mining technique to enhance tax evasion detection performance. DOI: 10.1016/j.eswa.2012.01.204.
39. Παλιούρας Γεώργιος (2011), Εξόρυξη γνώσης από δεδομένα (Acta Minng).
40. Χαλκίδη Μ. & Βαζιργιάννη Μ., (2005), Εξόρυξη από βάσεις δεδομένων και τον παγκόσμιο ιστό, Δεύτερη έκδοση, Αθήνα: Τυπωθήτω

#### **Ηλεκτρονικά Άρθρα**

41. Australian Government – Australian Taxation Office, Source URL: <https://www.ato.gov.au/Business/Large-business/Top-100-risk-categorisation-approach/>
42. ClearTax, Source URL: <https://cleartax.in/s/demystifying-gst-compliance-rating>
43. Felsmann Tamas (2017), Taxpayer ratings, WTS-Hungary, Source URL: <https://wtsklient.hu/en/2017/02/24/taxpayer-ratings/>
44. General Department of Taxation of Ministry of Economy and Finance, Source URL: <http://www.tax.gov.kh/en/news.php>

45. Hoffmann Richard (4/2/2018), China's Social Credit System and Taxes, ecovis-beijing, Source URL: <https://ecovis-beijing.com/tax-policy/china-social-credit-system-taxes/>
46. RSM, Source URL: <https://www.rsm.hu/en/blog/2017/08/number-of-reliable-taxpayers-growing>
47. State Revenue Service of the Republic of Latvia (28/10/2018), Taxpayer Rating System, Source URL: <https://www.vid.gov.lv/en/taxpayer-rating-system>



## Παραρτήματα

### Παράρτημα Α

#### Αλγόριθμος C5.1

##### 1.Χαρακτηριστικά δημιουργίας δέντρου απόφασης

Analysis

Tree depth: 14

Analysis of final\_data.xls (4 Φεβ 2019 1:17:13 μμ)

Number of records: 3.784

Analysis Accuracy: 90,883%

Fields

Target

V\_CLASS

Inputs

EKPR\_PL

EKPR\_XR

EKPR\_PS

LXPR\_PS

MH\_YPOB\_PL

TROP\_PL

LXPR\_LY\_PS

LXPR\_MET1

PARAVASI\_PS

MHTRWO\_XR

RTHM\_DIAK\_PL

LXPR\_XWRIS\_RTHM\_XR

BEB\_ANA\_PLRM

ANAKR\_DHL\_FOR

Build Settings

Use partitioned data: true  
Partition: Partition  
Calculate predictor importance: true  
Calculate raw propensity scores: false  
Calculate adjusted propensity scores: false  
Use weight: false  
Output type: Decision tree  
Group symbolics: false  
Use boosting: false  
Cross-validate: false  
Mode: Simple  
Favor: Accuracy  
Expected noise (%): 0  
Use misclassification costs: false

#### Training Summary

Algorithm: C5  
Model type: Classification  
Stream: E:\final.str  
User: n.papastergiou  
Date built: 4/2/2019 1:08 μμ  
Application: IBM® SPSS® Modeler 16  
Elapsed time for model build: 0 hours, 0 mins, 2 secs

## 2. Δέντρο απόφασης με κανόνες

Ακολουθεί δέντρο απόφασης με κανόνες όπως δημιουργείται από το μοντέλο C5.1 (με τα διαστήματα εμπιστοσύνης στις παρενθέσεις) :

```
EKPR_PL <= 3 [ Mode: 4 ] (3.160)
  EKPR_XR <= 7 [ Mode: 4 ] (2.762)
    EKPR_PS <= 10172,200 [ Mode: 5 ] (373)
      LXPR_PS <= 366,670 [ Mode: 5 ] (262)
        MH_YPOB_PL <= 2 [ Mode: 5 ] (232)
          EKPR_PL <= 0 [ Mode: 5 ] (226)
            EKPR_XR <= 2 [ Mode: 5 ] => 5,0 (214; 0,944)
            EKPR_XR > 2 [ Mode: 4 ] (12)
              TROP_PL <= 0 [ Mode: 4 ] => 4,0 (10; 0,7)
              TROP_PL > 0 [ Mode: 5 ] => 5,0 (2; 1,0)
            EKPR_PL > 0 [ Mode: 4 ] => 4,0 (6; 1,0)
          MH_YPOB_PL > 2 [ Mode: 4 ] (30)
            TROP_PL <= 0 [ Mode: 4 ] => 4,0 (27; 0,963)
            TROP_PL > 0 [ Mode: 3 ] => 3,0 (3; 0,667)
          LXPR_PS > 366,670 [ Mode: 4 ] (111)
            EKPR_PL <= 0 [ Mode: 4 ] (106)
              MH_YPOB_PL <= 14 [ Mode: 4 ] => 4,0 (101; 0,871)
              MH_YPOB_PL > 14 [ Mode: 3 ] (5)
                LXPR_LY_PS <= 2436,770 [ Mode: 4 ] => 4,0 (2; 1,0)
                LXPR_LY_PS > 2436,770 [ Mode: 3 ] => 3,0 (3; 1,0)
              EKPR_PL > 0 [ Mode: 3 ] (5)
                EKPR_PS <= 4.299 [ Mode: 2 ] => 2,0 (2; 0,5)
                EKPR_PS > 4.299 [ Mode: 3 ] => 3,0 (3; 1,0)
            EKPR_PS > 10172,200 [ Mode: 4 ] (2.389)
              MH_YPOB_PL <= 5 [ Mode: 4 ] (2.212)
                LXPR_MET1 <= 2867,770 [ Mode: 4 ] (1.785)
                  PARAVASI_PS <= 1.380 [ Mode: 4 ] (1.757)
                    EKPR_PL <= 0 [ Mode: 4 ] (1.689)
                      MH_YPOB_PL <= 0 [ Mode: 4 ] => 4,0 (1.553; 0,918)
                      MH_YPOB_PL > 0 [ Mode: 4 ] (136)
                        EKPR_XR <= 6 [ Mode: 4 ] (133)
                          LXPR_MET1 <= 0,060 [
Mode: 4 ] => 4,0 (121; 0,901)
```

|                                           |                                                     |
|-------------------------------------------|-----------------------------------------------------|
| 3 ] (12)                                  | LXPR_MET1 > 0,060 [ Mode:                           |
| Mode: 3 ] (7)                             | EKPR_XR <= 0 [                                      |
| TROP_PL <= 0 [ Mode: 4 ] => 4,0 (4; 0,75) |                                                     |
| TROP_PL > 0 [ Mode: 3 ] => 3,0 (3; 1,0)   |                                                     |
| Mode: 3 ] => 3,0 (5; 1,0)                 | EKPR_XR > 0 [                                       |
| 0,667)                                    | EKPR_XR > 6 [ Mode: 3 ] => 3,0 (3;                  |
|                                           | EKPR_PL > 0 [ Mode: 4 ] (68)                        |
|                                           | MH_YPOB_PL <= 1 [ Mode: 4 ] (66)                    |
|                                           | MHTRWO_XR <= 60 [ Mode: 4 ] (51)                    |
| 4,0 (45; 0,933)                           | EKPR_XR <= 3 [ Mode: 4 ] =>                         |
|                                           | EKPR_XR > 3 [ Mode: 3 ] (6)                         |
|                                           | RTHM_DIAK_PL                                        |
| <= 0 [ Mode: 3 ] => 3,0 (4; 1,0)          | RTHM_DIAK_PL                                        |
|                                           |                                                     |
| > 0 [ Mode: 4 ] => 4,0 (2; 1,0)           | MHTRWO_XR > 60 [ Mode: 3 ] (15)                     |
|                                           | TROP_PL <= 0 [ Mode: 4 ] =>                         |
| 4,0 (4; 1,0)                              |                                                     |
|                                           | TROP_PL > 0 [ Mode: 3 ] =>                          |
| 3,0 (11; 0,909)                           |                                                     |
|                                           | MH_YPOB_PL > 1 [ Mode: 2 ] => 2,0 (2; 0,5)          |
|                                           | PARAVASI_PS > 1.380 [ Mode: 4 ] (28)                |
|                                           | EKPR_XR <= 0 [ Mode: 4 ] (22)                       |
|                                           | PARAVASI_PS <= 3.853 [ Mode: 4 ] => 4,0 (12; 0,917) |
|                                           | PARAVASI_PS > 3.853 [ Mode: 3 ] => 3,0 (10; 0,7)    |
|                                           | EKPR_XR > 0 [ Mode: 3 ] => 3,0 (6; 0,833)           |
|                                           | LXPR_MET1 > 2867,770 [ Mode: 4 ] (427)              |
|                                           | EKPR_XR <= 0 [ Mode: 4 ] (331)                      |
|                                           | PARAVASI_PS <= 22 [ Mode: 4 ] (310)                 |
|                                           | LXPR_MET1 <= 50574,900 [ Mode: 4 ] (261)            |
|                                           | TROP_PL <= 0 [ Mode: 4 ] (248)                      |
|                                           | MH_YPOB_PL <= 0 [ Mode: 4                           |
| ] (227)                                   |                                                     |
|                                           | MHTRWO_XR <=                                        |
| 0 [ Mode: 4 ] => 4,0 (202; 0,995)         |                                                     |
|                                           | MHTRWO_XR > 0                                       |
| [ Mode: 3 ] (25)                          |                                                     |

EKPR\_PS <= 27510,900 [ Mode: 4 ] => 4,0 (6; 0,833)

EKPR\_PS > 27510,900 [ Mode: 3 ] (19)

MHTRWO\_XR <= 150 [ Mode: 3 ] => 3,0 (15; 0,933)

MHTRWO\_XR > 150 [ Mode: 3 ] (4)

EKPR\_PS <= 67579,500 [ Mode: 3 ] => 3,0 (2; 1,0)

EKPR\_PS > 67579,500 [ Mode: 4 ] => 4,0 (2; 1,0)

MH\_YPOB\_PL > 0 [ Mode: 3 ]

(21)

LXPR\_XWRIS\_RTHM\_XR <= 242,350 [ Mode: 3 ] => 3,0 (17; 0,882)

LXPR\_XWRIS\_RTHM\_XR > 242,350 [ Mode: 4 ] => 4,0 (4; 0,75)

TROP\_PL > 0 [ Mode: 3 ] (13)

RTHM\_DIAK\_PL <= 0 [ Mode:

3 ] => 3,0 (10; 1,0)

RTHM\_DIAK\_PL > 0 [ Mode:

4 ] => 4,0 (3; 0,667)

LXPR\_MET1 > 50574,900 [ Mode: 3 ] => 3,0 (49;

0,796)

PARAVASI\_PS > 22 [ Mode: 3 ] (21)

PARAVASI\_PS <= 11.258 [ Mode: 3 ] (18)

LXPR\_MET1 <= 38.769 [ Mode: 3 ] => 3,0

(12; 1,0)

LXPR\_MET1 > 38.769 [ Mode: 3 ] (6)

LXPR\_PS <= 108.315 [ Mode:

4 ] => 4,0 (3; 1,0)

LXPR\_PS > 108.315 [ Mode:

3 ] => 3,0 (3; 1,0)

PARAVASI\_PS > 11.258 [ Mode: 2 ] => 2,0 (3; 1,0)

EKPR\_XR > 0 [ Mode: 3 ] (96)

MH\_YPOB\_PL <= 3 [ Mode: 3 ] (94)

LXPR\_MET1 <= 123.609 [ Mode: 3 ] => 3,0 (89; 0,854)

LXPR\_MET1 > 123.609 [ Mode: 2 ] (5)

MHTRWO\_XR <= 50 [ Mode: 4 ] => 4,0 (3;

0,667)

MHTRWO\_XR > 50 [ Mode: 2 ] => 2,0 (2;

1,0)

MH\_YPOB\_PL > 3 [ Mode: 2 ] => 2,0 (2; 1,0)

MH\_YPOB\_PL > 5 [ Mode: 3 ] (177)

|                |                                                    |
|----------------|----------------------------------------------------|
|                | LXPR_MET1 <= 3050,940 [ Mode: 3 ] (138)            |
|                | EKPR_XR <= 0 [ Mode: 3 ] (111)                     |
|                | PARAVASI_PS <= 665 [ Mode: 3 ] (108)               |
|                | MH_YPOB_PL <= 7 [ Mode: 4 ] (17)                   |
| 4,0 (15; 0,8)  | LXPR_MET1 <= 527,540 [ Mode: 4 ] =>                |
| (2; 1,0)       | LXPR_MET1 > 527,540 [ Mode: 3 ] => 3,0             |
|                | MH_YPOB_PL > 7 [ Mode: 3 ] => 3,0 (91; 0,879)      |
|                | PARAVASI_PS > 665 [ Mode: 3 ] => 3,0 (3; 0,667)    |
|                | EKPR_XR > 0 [ Mode: 3 ] => 3,0 (27; 0,741)         |
|                | LXPR_MET1 > 3050,940 [ Mode: 2 ] (39)              |
|                | MH_YPOB_PL <= 14 [ Mode: 3 ] (28)                  |
|                | RTHM_DIAK_PL <= 0 [ Mode: 3 ] (21)                 |
|                | EKPR_PS <= 25374,400 [ Mode: 2 ] => 2,0 (6; 0,833) |
|                | EKPR_PS > 25374,400 [ Mode: 3 ] (15)               |
| (12; 0,917)    | MHTRWO_XR <= 150 [ Mode: 3 ] => 3,0                |
| 0,667)         | MHTRWO_XR > 150 [ Mode: 2 ] => 2,0 (3;             |
|                | RTHM_DIAK_PL > 0 [ Mode: 2 ] (7)                   |
| 2,0 (4; 1,0)   | LXPR_XWRIS_RTHM_XR <= 206,380 [ Mode: 2 ] =>       |
| 3,0 (3; 0,667) | LXPR_XWRIS_RTHM_XR > 206,380 [ Mode: 3 ] =>        |
|                | MH_YPOB_PL > 14 [ Mode: 2 ] (11)                   |
|                | RTHM_DIAK_PL <= 0 [ Mode: 2 ] (9)                  |
| 3,0 (3; 0,667) | LXPR_XWRIS_RTHM_XR <= 93,250 [ Mode: 3 ] =>        |
| (6; 0,833)     | LXPR_XWRIS_RTHM_XR > 93,250 [ Mode: 2 ] => 2,0     |
|                | RTHM_DIAK_PL > 0 [ Mode: 4 ] => 4,0 (2; 1,0)       |
|                | EKPR_XR > 7 [ Mode: 3 ] (398)                      |
|                | LXPR_MET1 <= 1580,940 [ Mode: 3 ] (298)            |
|                | EKPR_PS <= 9493,950 [ Mode: 4 ] (39)               |
|                | PARAVASI_PS <= 1.600 [ Mode: 4 ] (37)              |
|                | MH_YPOB_PL <= 4 [ Mode: 4 ] => 4,0 (34; 0,912)     |
|                | MH_YPOB_PL > 4 [ Mode: 3 ] => 3,0 (3; 0,667)       |
|                | PARAVASI_PS > 1.600 [ Mode: 3 ] => 3,0 (2; 1,0)    |
|                | EKPR_PS > 9493,950 [ Mode: 3 ] (259)               |
|                | MH_YPOB_PL <= 3 [ Mode: 3 ] (228)                  |

EKPR\_PL <= 0 [ Mode: 3 ] (211)  
 EKPR\_XR <= 8 [ Mode: 4 ] (16)  
 LXPR\_PS <= 139,340 [ Mode: 4 ] (13)  
 MHTRWO\_XR <= 50 [ Mode: 4 ] => 4,0  
 (11; 1,0)  
 MHTRWO\_XR > 50 [ Mode: 3 ] => 3,0 (2;  
 1,0)  
 LXPR\_PS > 139,340 [ Mode: 3 ] => 3,0 (3; 1,0)  
 EKPR\_XR > 8 [ Mode: 3 ] => 3,0 (195; 0,846)  
 EKPR\_PL > 0 [ Mode: 3 ] (17)  
 LXPR\_PS <= 0,220 [ Mode: 3 ] (13)  
 MHTRWO\_XR <= 50 [ Mode: 3 ] => 3,0 (10; 1,0)  
 MHTRWO\_XR > 50 [ Mode: 2 ] => 2,0 (3; 0,667)  
 LXPR\_PS > 0,220 [ Mode: 2 ] => 2,0 (4; 1,0)  
 MH\_YPOB\_PL > 3 [ Mode: 3 ] (31)  
 PARAVASI\_PS <= 122 [ Mode: 3 ] => 3,0 (29; 0,69)  
 PARAVASI\_PS > 122 [ Mode: 2 ] => 2,0 (2; 1,0)  
 LXPR\_MET1 > 1580,940 [ Mode: 2 ] (100)  
 MH\_YPOB\_PL <= 4 [ Mode: 2 ] (86)  
 PARAVASI\_PS <= 4.000 [ Mode: 2 ] (83)  
 EKPR\_PS <= 11740,600 [ Mode: 3 ] => 3,0 (14; 0,929)  
 EKPR\_PS > 11740,600 [ Mode: 2 ] (69)  
 EKPR\_XR <= 8 [ Mode: 3 ] => 3,0 (5; 1,0)  
 EKPR\_XR > 8 [ Mode: 2 ] (64)  
 BEB\_ANA\_PLRM <= 32,040 [ Mode: 2 ] (60)  
 LXPR\_MET1 <= 28188,100 [ Mode: 2 ]  
 (31)  
 LXPR\_PS <= 7774,580 [  
 Mode: 2 ] (10)  
 LXPR\_LY\_PS <=  
 196,540 [ Mode: 2 ] => 2,0 (8; 0,75)  
 LXPR\_LY\_PS >  
 196,540 [ Mode: 3 ] => 3,0 (2; 1,0)  
 LXPR\_PS > 7774,580 [ Mode:  
 2 ] => 2,0 (21; 1,0)  
 LXPR\_MET1 > 28188,100 [ Mode: 2 ] (29)  
 RTHM\_DIAK\_PL <= 0 [ Mode:  
 2 ] (24)  
 LXPR\_LY\_PS <=  
 18406,200 [ Mode: 2 ] => 2,0 (19; 0,789)  
 LXPR\_LY\_PS >  
 18406,200 [ Mode: 3 ] => 3,0 (5; 0,8)

3 ] => 3,0 (5; 1,0)

RTHM\_DIAK\_PL > 0 [ Mode:

BEB\_ANA\_PLRM > 32,040 [ Mode: 3 ] => 3,0 (4; 1,0)

PARAVASI\_PS > 4.000 [ Mode: 2 ] => 2,0 (3; 0,667)

MH\_YPOB\_PL > 4 [ Mode: 2 ] (14)

EKPR\_PL <= 0 [ Mode: 2 ] (12)

PARAVASI\_PS <= 600 [ Mode: 2 ] => 2,0 (10; 1,0)

PARAVASI\_PS > 600 [ Mode: 1 ] => 1,0 (2; 1,0)

EKPR\_PL > 0 [ Mode: 1 ] => 1,0 (2; 1,0)

EKPR\_PL > 3 [ Mode: 2 ] (624)

PARAVASI\_PS <= 22 [ Mode: 2 ] (499)

EKPR\_XR <= 8 [ Mode: 3 ] (211)

EKPR\_PS <= 9595,280 [ Mode: 4 ] (26)

LXPR\_PS <= 3753,360 [ Mode: 4 ] (24)

EKPR\_PL <= 6 [ Mode: 4 ] => 4,0 (14; 1,0)

EKPR\_PL > 6 [ Mode: 3 ] (10)

EKPR\_PS <= 130,750 [ Mode: 4 ] => 4,0 (3; 1,0)

EKPR\_PS > 130,750 [ Mode: 3 ] => 3,0 (7; 0,857)

LXPR\_PS > 3753,360 [ Mode: 2 ] => 2,0 (2; 1,0)

EKPR\_PS > 9595,280 [ Mode: 3 ] (185)

LXPR\_PS <= 0,870 [ Mode: 3 ] (128)

EKPR\_PL <= 9 [ Mode: 3 ] (126)

MH\_YPOB\_PL <= 2 [ Mode: 3 ] => 3,0 (121; 0,851)

MH\_YPOB\_PL > 2 [ Mode: 2 ] (5)

EKPR\_PS <= 14236,700 [ Mode: 3 ] => 3,0 (2; 1,0)

EKPR\_PS > 14236,700 [ Mode: 2 ] => 2,0 (3; 1,0)

EKPR\_PL > 9 [ Mode: 2 ] => 2,0 (2; 1,0)

LXPR\_PS > 0,870 [ Mode: 2 ] (57)

EKPR\_PL <= 6 [ Mode: 3 ] (38)

MH\_YPOB\_PL <= 0 [ Mode: 3 ] (32)

EKPR\_PL <= 4 [ Mode: 3 ] (12)

LXPR\_MET1 <= 3243,960 [ Mode: 3 ] =>

3,0 (4; 1,0)

LXPR\_MET1 > 3243,960 [ Mode: 2 ] (8)

LXPR\_MET1 <= 72781,800 [

Mode: 2 ] => 2,0 (5; 1,0)

LXPR\_MET1 > 72781,800 [

Mode: 3 ] => 3,0 (3; 0,667)



EKPR\_PL > 4 [ Mode: 3 ] (20)  
 TROP\_PL <= 1 [ Mode: 3 ] => 3,0 (16; 1,0)  
 TROP\_PL > 1 [ Mode: 2 ] => 2,0 (4; 0,75)  
 MH\_YPOB\_PL > 0 [ Mode: 2 ] => 2,0 (6; 0,833)  
 EKPR\_PL > 6 [ Mode: 2 ] (19)  
 TROP\_PL <= 0 [ Mode: 2 ] (5)  
 LXPR\_PS <= 23728,300 [ Mode: 3 ] => 3,0 (2; 1,0)  
 LXPR\_PS > 23728,300 [ Mode: 2 ] => 2,0 (3; 1,0)  
 TROP\_PL > 0 [ Mode: 2 ] => 2,0 (14; 0,929)  
 EKPR\_XR > 8 [ Mode: 2 ] (288)  
 LXPR\_MET1 <= 44417,800 [ Mode: 2 ] (281)  
 EKPR\_PS <= 9432,130 [ Mode: 3 ] (37)  
 EKPR\_PS <= 1148,780 [ Mode: 3 ] (19)  
 EKPR\_PL <= 7 [ Mode: 3 ] (8)  
 EKPR\_PL <= 4 [ Mode: 3 ] => 3,0 (5; 0,8)  
 EKPR\_PL > 4 [ Mode: 4 ] => 4,0 (3; 1,0)  
 EKPR\_PL > 7 [ Mode: 3 ] => 3,0 (11; 1,0)  
 EKPR\_PS > 1148,780 [ Mode: 2 ] => 2,0 (18; 0,5)  
 EKPR\_PS > 9432,130 [ Mode: 2 ] (244)  
 MH\_YPOB\_PL <= 2 [ Mode: 2 ] (230)  
 BEB\_ANA\_PLRM <= 25,680 [ Mode: 2 ] (194)  
 LXPR\_XWRIS\_RTHM\_XR <= 224,350 [ Mode: 2 ]  
 (103)  
 EKPR\_PL <= 7 [ Mode: 2 ] (25)  
 EKPR\_PS <= 376.724 [ Mode:  
 2 ] (20)  
 EKPR\_PS <=  
 40655,600 [ Mode: 3 ] (5)  
 TROP\_PL <= 1 [ Mode: 3 ] => 3,0 (3; 1,0)  
 TROP\_PL > 1 [ Mode: 2 ] => 2,0 (2; 1,0)  
 EKPR\_PS >  
 40655,600 [ Mode: 2 ] => 2,0 (15; 1,0)  
 EKPR\_PS > 376.724 [ Mode:  
 3 ] => 3,0 (5; 1,0)  
 EKPR\_PL > 7 [ Mode: 2 ] (78)  
 EKPR\_PL <= 13 [ Mode: 2 ]  
 => 2,0 (37; 1,0)  
 EKPR\_PL > 13 [ Mode: 2 ]  
 (41)

Mode: 2 ] (21) EKPR\_XR <= 28 [

EKPR\_PS <= 115.748 [ Mode: 2 ] => 2,0 (10; 1,0)

EKPR\_PS > 115.748 [ Mode: 3 ] (11)

MHTRWO\_XR <= 50 [ Mode: 3 ] (8)

LXPR\_XWRIS\_RTHM\_XR <= 212,350 [ Mode: 3 ] => 3,0 (6; 1,0)

LXPR\_XWRIS\_RTHM\_XR > 212,350 [ Mode: 2 ] => 2,0 (2; 1,0)

MHTRWO\_XR > 50 [ Mode: 2 ] => 2,0 (3; 1,0)

Mode: 2 ] => 2,0 (20; 1,0) EKPR\_XR > 28 [

LXPR\_XWRIS\_RTHM\_XR > 224,350 [ Mode: 2 ] =>

2,0 (91; 0,989)

BEB\_ANA\_PLRM > 25,680 [ Mode: 2 ] (36)

LXPR\_XWRIS\_RTHM\_XR <= 143,420 [ Mode: 2 ] =>

2,0 (10; 1,0)

LXPR\_XWRIS\_RTHM\_XR > 143,420 [ Mode: 2 ] (26)

MH\_YPOB\_PL <= 0 [ Mode: 2 ] (24)

TROP\_PL <= 0 [ Mode: 3 ] =>

3,0 (3; 1,0)

TROP\_PL > 0 [ Mode: 2 ] =>

2,0 (21; 0,714)

MH\_YPOB\_PL > 0 [ Mode: 3 ] => 3,0 (2;

1,0)

MH\_YPOB\_PL > 2 [ Mode: 2 ] => 2,0 (14; 0,857)

LXPR\_MET1 > 44417,800 [ Mode: 1 ] (7)

EKPR\_PL <= 8 [ Mode: 2 ] => 2,0 (4; 0,75)

EKPR\_PL > 8 [ Mode: 1 ] => 1,0 (3; 1,0)

PARAVASI\_PS > 22 [ Mode: 1 ] (125)

EKPR\_PL <= 6 [ Mode: 3 ] (22)

TROP\_PL <= 1 [ Mode: 3 ] (19)

LXPR\_PS <= 7762,760 [ Mode: 3 ] (16)

MHTRWO\_XR <= 0 [ Mode: 3 ] => 3,0 (13; 0,846)

MHTRWO\_XR > 0 [ Mode: 2 ] => 2,0 (3; 0,667)

LXPR\_PS > 7762,760 [ Mode: 2 ] => 2,0 (3; 1,0)

TROP\_PL > 1 [ Mode: 2 ] => 2,0 (3; 1,0)

EKPR\_PL > 6 [ Mode: 1 ] (103)

LXPR\_MET1 <= 475,590 [ Mode: 2 ] (50)

ANAKR\_DHL\_FOR <= 100 [ Mode: 2 ] (43)

EKPR\_PS <= 9189,330 [ Mode: 3 ] => 3,0 (3; 0,667)

EKPR\_PS > 9189,330 [ Mode: 2 ] (40)

PARAVASI\_PS <= 1.800 [ Mode: 2 ] => 2,0 (26; 0,885)

PARAVASI\_PS > 1.800 [ Mode: 1 ] (14)

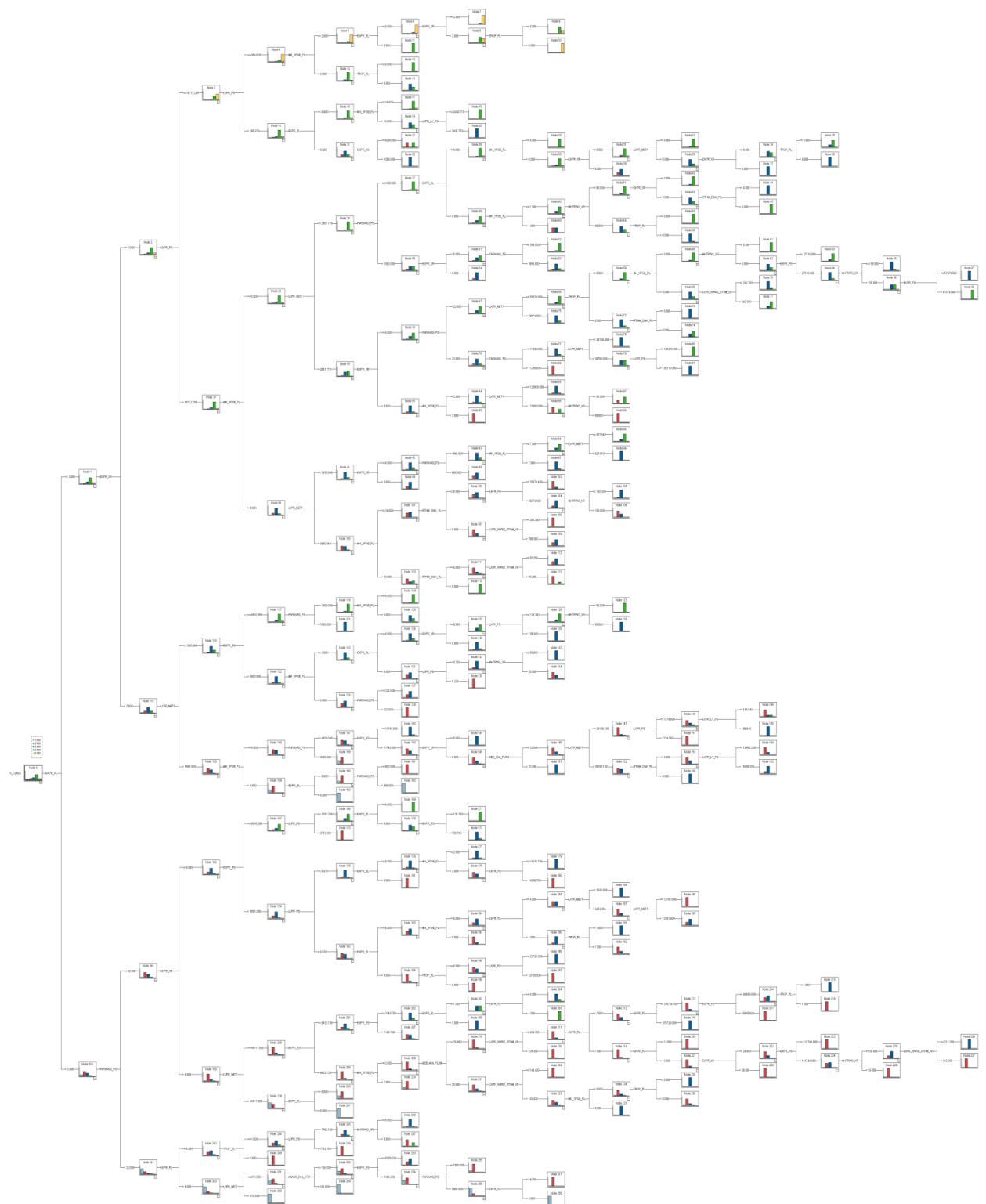
EKPR\_PL <= 9 [ Mode: 2 ] => 2,0 (3; 1,0)

EKPR\_PL > 9 [ Mode: 1 ] => 1,0 (11; 1,0)

ANAKR\_DHL\_FOR > 100 [ Mode: 1 ] => 1,0 (7; 1,0)

LXPR\_MET1 > 475,590 [ Mode: 1 ] => 1,0 (53; 1,0)

### 3.Σχηματική αναπαράσταση δέντρου απόφασης



## Παράρτημα Β

### Αλγόριθμος C&R Tree

#### 1.Χαρακτηριστικά δέντρου απόφασης

##### Analysis

Tree depth: 5

Analysis of final\_data.xls (4 Φεβ 2019 2:37:54 μμ)

Number of records: 3.784

Analysis Accuracy: 79,572%

##### Fields

##### Target

V\_CLASS

##### Inputs

MHTRWO\_XR

MH\_YPOB\_PL

EKPR\_PL

EKPR\_XR

EKPR\_PS

TROP\_PL

LXPR\_PS

LXPR\_LY\_PS

LXPR\_MET1

LXPR\_XWRIS\_RTHM\_XR

RTHM\_DIAK\_PL

BEB\_ANA\_PLRM

ANAKR\_DHL\_FOR

PARAVASI\_PS

##### Build Settings

Use partitioned data: true

Partition: Partition

Calculate predictor importance: true

Calculate raw propensity scores: false  
Calculate adjusted propensity scores: false  
Use frequency: false  
Use weight: false  
Levels below root: 5  
Mode: Expert  
Maximum surrogates: 5  
Minimum change in impurity: 0,0  
Impurity measure for categorical targets: Gini  
Stopping criteria: Use percentage  
Minimum records in parent branch (%): 2  
Minimum records in child branch (%): 1  
Prune tree: true  
Use standard error rule: false  
Prior probabilities: Based on training data  
Adjust priors using misclassification costs: false  
Use misclassification costs: false

#### Training Summary

Algorithm: C&R Tree

Model type: Classification

Stream: E:\final.str

User: n.papastergiou

Date built: 4/2/2019 2:36 μμ

Application: IBM® SPSS® Modeler 16

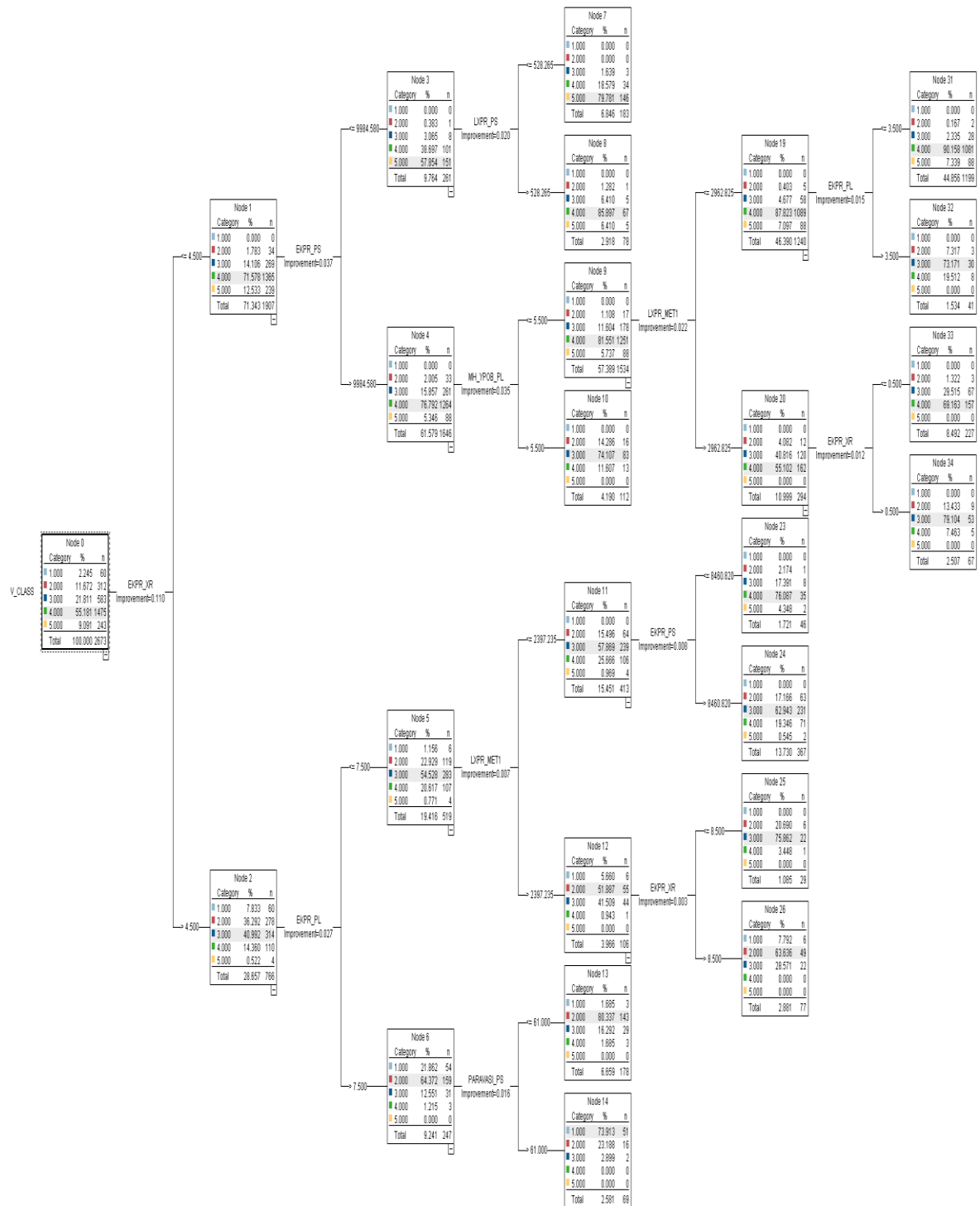
Elapsed time for model build: 0 hours, 0 mins, 14 secs

## 2. Δέντρο απόφασης με κανόνες

Ακολουθεί δέντρο απόφασης με κανόνες όπως δημιουργείται από το μοντέλο C&R Tree (με τα διαστήματα εμπιστοσύνης στις παρενθέσεις) :

```
EKPR_XR <= 4,500 [ Mode: 4 ] (1.907)
  EKPR_PS <= 9984,580 [ Mode: 5 ] (261)
    LXPR_PS <= 528,265 [ Mode: 5 ] => 5,0 (183; 0,798)
    LXPR_PS > 528,265 [ Mode: 4 ] => 4,0 (78; 0,859)
  EKPR_PS > 9984,580 [ Mode: 4 ] (1.646)
    MH_YPOB_PL <= 5,500 [ Mode: 4 ] (1.534)
      LXPR_MET1 <= 2962,825 [ Mode: 4 ] (1.240)
        EKPR_PL <= 3,500 [ Mode: 4 ] => 4,0 (1.199; 0,902)
        EKPR_PL > 3,500 [ Mode: 3 ] => 3,0 (41; 0,732)
      LXPR_MET1 > 2962,825 [ Mode: 4 ] (294)
        EKPR_XR <= 0,500 [ Mode: 4 ] => 4,0 (227; 0,692)
        EKPR_XR > 0,500 [ Mode: 3 ] => 3,0 (67; 0,791)
      MH_YPOB_PL > 5,500 [ Mode: 3 ] => 3,0 (112; 0,741)
    EKPR_XR > 4,500 [ Mode: 3 ] (766)
      EKPR_PL <= 7,500 [ Mode: 3 ] (519)
        LXPR_MET1 <= 2397,235 [ Mode: 3 ] (413)
          EKPR_PS <= 8460,820 [ Mode: 4 ] => 4,0 (46; 0,761)
          EKPR_PS > 8460,820 [ Mode: 3 ] => 3,0 (367; 0,629)
        LXPR_MET1 > 2397,235 [ Mode: 2 ] (106)
          EKPR_XR <= 8,500 [ Mode: 3 ] => 3,0 (29; 0,759)
          EKPR_XR > 8,500 [ Mode: 2 ] => 2,0 (77; 0,636)
      EKPR_PL > 7,500 [ Mode: 2 ] (247)
        PARAVASI_PS <= 61 [ Mode: 2 ] => 2,0 (178; 0,803)
        PARAVASI_PS > 61 [ Mode: 1 ] => 1,0 (69; 0,739)
```

### 3.Σχηματική αναπαράσταση δέντρου απόφασης





## Παράρτημα Γ

### Αλγόριθμος CHAID

#### 1.Χαρακτηριστικά δέντρου απόφασης

##### Analysis

Tree depth: 5

Analysis of final\_data.xls (4 Φεβ 2019 3:05:03 μμ)

Number of records: 3.784

Analysis Accuracy: 78,013%

##### Fields

##### Target

V\_CLASS

##### Inputs

MH\_YPOB\_PL

EKPR\_PL

EKPR\_XR

EKPR\_PS

LXPR\_PS

LXPR\_MET1

PARAVASI\_PS

##### Build Settings

Use partitioned data: true

Partition: Partition

Calculate predictor importance: true

Calculate raw propensity scores: false

Calculate adjusted propensity scores: false

Continue training existing model: false

Use frequency: false

Use weight: false

Levels below root: 5

Alpha for Splitting: 0,05

Alpha for Merging: 0,05  
Epsilon For Convergence: 0,001  
Maximum iterations for convergence: 100  
Use Bonferroni adjustment: true  
Allow splitting of merged categories: false  
Chi-Square method: Pearson  
Stopping criteria: Use percentage  
Minimum records in parent branch (%): 2  
Minimum records in child branch (%): 1  
Use misclassification costs: false

#### Training Summary

Algorithm: CHAID  
Model type: Classification  
Stream: E:\final.str  
User: n.papastergiou  
Date built: 4/2/2019 3:04 μμ  
Application: IBM® SPSS® Modeler 16  
Elapsed time for model build: 0 hours, 0 mins, 12 secs

## 2. Δέντρο απόφασης με κανόνες

Ακολουθεί δέντρο απόφασης με κανόνες όπως δημιουργείται από το μοντέλο CHAID (με τα διαστήματα εμπιστοσύνης στις παρενθέσεις) :

EKPR\_PL ≤ 0 [ Mode: 4 ] (3.027)

EKPR\_XR ≤ 0 [ Mode: 4 ] (2.178)

MH\_YPOB\_PL ≤ 0 [ Mode: 4 ] (1.881)

EKPR\_PS ≤ 7837,370 [ Mode: 5 ] (181)

LXPR\_PS ≤ 0 [ Mode: 5 ] ⇒ 5,0 (129; 0,992)

LXPR\_PS > 0 [ Mode: 4 ] ⇒ 4,0 (52; 0,788)

EKPR\_PS > 7837,370 and EKPR\_PS ≤ 14324,790 [ Mode: 4 ] (208)

LXPR\_PS ≤ 35,880 [ Mode: 4 ] ⇒ 4,0 (147; 0,612)

LXPR\_PS > 35,880 [ Mode: 4 ] ⇒ 4,0 (61; 0,869)

EKPR\_PS > 14324,790 and EKPR\_PS ≤ 29580,040 [ Mode: 4 ] (429)

LXPR\_MET1 ≤ 2429,560 [ Mode: 4 ] ⇒ 4,0 (336; 0,917)

LXPR\_MET1 > 2429,560 [ Mode: 4 ] ⇒ 4,0 (93; 0,86)

EKPR\_PS > 29580,040 [ Mode: 4 ] (1.063)

LXPR\_MET1 ≤ 2429,560 [ Mode: 4 ] ⇒ 4,0 (864; 0,906)

LXPR\_MET1 > 2429,560 [ Mode: 4 ] ⇒ 4,0 (199; 0,658)

MH\_YPOB\_PL > 0 and MH\_YPOB\_PL ≤ 2 [ Mode: 4 ] ⇒ 4,0 (80; 0,75)

MH\_YPOB\_PL > 2 [ Mode: 3 ] (217)

LXPR\_MET1 ≤ 2429,560 [ Mode: 3 ] (168)

LXPR\_PS ≤ 0 [ Mode: 4 ] ⇒ 4,0 (125; 0,52)

LXPR\_PS > 0 [ Mode: 3 ] ⇒ 3,0 (43; 0,744)

LXPR\_MET1 > 2429,560 [ Mode: 3 ] ⇒ 3,0 (49; 0,551)

EKPR\_XR > 0 and EKPR\_XR ≤ 3 [ Mode: 4 ] (312)

LXPR\_MET1 ≤ 2429,560 [ Mode: 4 ] (247)

EKPR\_PS ≤ 56890,870 [ Mode: 4 ] ⇒ 4,0 (139; 0,669)

EKPR\_PS > 56890,870 [ Mode: 4 ] ⇒ 4,0 (108; 0,944)

LXPR\_MET1 > 2429,560 [ Mode: 3 ] ⇒ 3,0 (65; 0,646)

EKPR\_XR > 3 and EKPR\_XR ≤ 8 [ Mode: 4 ] (197)

LXPR\_PS ≤ 0 [ Mode: 4 ] ⇒ 4,0 (134; 0,791)

LXPR\_PS > 0 [ Mode: 3 ] => 3,0 (63; 0,619)  
 EKPR\_XR > 8 [ Mode: 3 ] (340)  
 LXPR\_MET1 <= 2429,560 [ Mode: 3 ] => 3,0 (256; 0,723)  
 LXPR\_MET1 > 2429,560 [ Mode: 2 ] => 2,0 (84; 0,595)  
 EKPR\_PL > 0 and EKPR\_PL <= 6 [ Mode: 3 ] (356)  
 EKPR\_XR <= 3 [ Mode: 4 ] (138)  
 LXPR\_PS <= 35,880 [ Mode: 4 ] => 4,0 (96; 0,635)  
 LXPR\_PS > 35,880 [ Mode: 3 ] => 3,0 (42; 0,571)  
 EKPR\_XR > 3 and EKPR\_XR <= 8 [ Mode: 3 ] => 3,0 (120; 0,758)  
 EKPR\_XR > 8 [ Mode: 2 ] => 2,0 (98; 0,571)  
 EKPR\_PL > 6 [ Mode: 2 ] (401)  
 PARAVASI\_PS <= 0 [ Mode: 2 ] (297)  
 EKPR\_XR <= 8 [ Mode: 3 ] => 3,0 (73; 0,534)  
 EKPR\_XR > 8 [ Mode: 2 ] (224)  
 LXPR\_PS <= 35,880 [ Mode: 2 ] => 2,0 (173; 0,844)  
 LXPR\_PS > 35,880 [ Mode: 2 ] => 2,0 (51; 0,765)  
 PARAVASI\_PS > 0 [ Mode: 1 ] (104)  
 LXPR\_MET1 <= 0 [ Mode: 2 ] => 2,0 (51; 0,549)  
 LXPR\_MET1 > 0 [ Mode: 1 ] => 1,0 (53; 1,0)

### 3.Σχηματική αναπαράσταση δέντρου απόφασης

