

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ

Κατεύθυνση : «ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΔΙΟΙΚΗΣΗΣ
ΕΠΙΧΕΙΡΗΣΕΩΝ»

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

**Η ΧΡΗΣΗ ΤΩΝ ΤΕΧΝΙΚΩΝ
ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ ΣΤΗ
ΦΟΡΟΛΟΓΙΚΗ ΔΙΟΙΚΗΣΗ**

A.M. 12602

ΑΘΗΝΑ 2016

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ

Κατεύθυνση : «ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ»

Η ΧΡΗΣΗ ΤΩΝ ΤΕΧΝΙΚΩΝ ΕΞΟΡΥΞΗΣ ΣΤΗΝ ΦΟΡΟΛΟΓΙΚΗ ΔΙΟΙΚΗΣΗ

Γλυτσός Αντώνης Α.Μ. 12602

Επιβλέπων:

Ηρακλής Βαρλάμης, Επίκουρος καθηγητής

Μέλη της εξεταστικής επιτροπής:

Δημοσθένης Αναγνωστόπουλος, Καθηγητής

Μαλβίνα Βαμβακάρη, Αναπληρώτρια Καθηγήτρια

ΑΘΗΝΑ 2016

Περίληψη

Η απατηλή συμπεριφορά των φορολογούμενων δημιουργεί μια σειρά από αλλοιώσεις στο οικονομικό περιβάλλον, οι οποίες με τη σειρά τους επηρεάζουν το σύνολο της οικονομίας και της κοινωνίας. Η οικονομική παραβατικότητα μπορεί να πάρει διάφορες μορφές, φοροδιαφυγή, φοροαπαλλαγή, απόκρυψη εσόδων κλπ.

Μια από τις πιο διαδεδομένες μορφές οικονομικού εγκλήματος είναι η απάτη στον τομέα του ΦΠΑ.

Δεδομένου του μεγάλου όγκου πληροφοριών που πρέπει να αναλυθούν για να εντοπισθούν πιθανές υποθέσεις απάτης, αλλά και του μεγάλου κόστους που επιφέρουν οι διαδικασίες ελέγχων, ο έλεγχος γίνεται συνήθως εστιάζοντας σε συγκεκριμένες περιπτώσεις και όχι αυτοματοποιημένα ή μαζικά. Πρόσθετες δυσκολίες για τους ελέγχους, πηγάζουν από το γεγονός ότι συνήθως τα μέσα που έχει στη διάθεση της η διοίκηση είναι περιορισμένα.

Στην παρούσα εργασία θα ελέγξουμε κατά πόσο οι τεχνικές εξόρυξης γνώσης μπορούν να χρησιμοποιηθούν ώστε να βελτιωθούν οι διαδικασίες φορολογικών ελέγχων, αν υπάρχουν εργαλεία ανοικτού κώδικα/ελεύθερης διανομής μπορούν να χρησιμοποιηθούν για την ταξινόμηση των υποθέσεων και ποια γνωρίσματα είναι αυτά που έχουν τη μεγαλύτερη συμμετοχή στη διαμόρφωση της αποτελεσμάτων.

Για το σκοπό της εργασίας το πεδίο ελέγχου περιορίστηκε στον τομέα του ενδοκοινοτικού εμπορίου και των προβλημάτων που εμφανίζονται στο συγκεκριμένο τομέα και ειδικά στην απάτη τύπου «καρουσέλ».

Λέξεις κλειδιά : εξόρυξη δεδομένων, πρόβλεψη, οικονομικό έγκλημα, φορολογική διοίκηση, φορολογία, WEKA, oracle data miner, καρουσέλ, εξαφανισμένος έμπορος, ενδοκοινοτικό εμπόριο

Abstract

Fraudulent behavior of the taxpayers creates a series of alterations to the economic environment, which in turn affect the economy and the society as a whole. Economic delinquency can take various forms like, tax evasion, tax exemption, income concealment etc.

One of the most prevalent forms of financial crime is VAT fraud. The selection of cases to be checked and the detection of the fraud are both painful procedures. The large number of financial entities and respectively the large volume of transactions, as well as the limited means of the tax administration contribute to this.

In this study we will check if the data mining techniques can be used to improve fiscal control procedures. We will also check if existing free/open source software can be used to classify cases and which features are best fitted to describe the problem in hand, in order to make successful predictions.

For the purpose of this study we have restrict the universe of discourse to European intra-Community trade and a special kind of fraud encounter in this field, namely “Carousel” fraud.

Keywords : Data mining, predictive analytics, tax administration, tax administration, taxation, WEKA, oracle data miner, carousel, missing trader, intra-community trade

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΛΗΨΗ.....	3
ABSTRACT	4
ΠΕΡΙΕΧΟΜΕΝΑ.....	5
ΠΙΝΑΚΕΣ – ΣΧΗΜΑΤΑ	7
ΕΙΣΑΓΩΓΗ	9
ΚΕΦΑΛΑΙΟ 1 - ΕΙΣΑΓΩΓΗ ΣΤΟ ΠΡΟΒΛΗΜΑ	11
ΚΕΦΑΛΑΙΟ 2 - ΜΕΛΕΤΕΣ ΚΑΙ ΕΦΑΡΜΟΓΕΣ	14
1. ΥΠΑΡΧΟΥΣΕΣ ΜΕΛΕΤΕΣ	14
2. ΠΡΑΚΤΙΚΕΣ ΕΦΑΡΜΟΓΕΣ	16
ΚΕΦΑΛΑΙΟ 3 - ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΦΟΡΟΛΟΓΙΚΗ ΔΙΟΙΚΗΣΗ	20
1. ΔΙΟΙΚΗΣΗ ΚΑΙ ΦΟΡΟΛΟΓΙΚΟΙ ΕΛΕΓΧΟΙ	20
2. Η ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΣΤΗ ΦΟΡΟΛΟΓΙΚΗ ΔΙΟΙΚΗΣΗ	21
3. ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ	23
4. ΔΗΜΟΦΙΛΕΙΣ ΔΙΑΔΙΚΑΣΙΕΣ ΕΞΟΡΥΞΗΣ	25
5. ΔΗΜΟΦΙΛΕΙΣ ΑΛΓΟΡΙΘΜΟΙ ΕΞΟΡΥΞΗΣ	27
<i>Δέντρα Απόφασης – Decision Trees</i>	27
<i>Κανόνες – Rules</i>	28
<i>Bagging - Boosting</i>	29
<i>Δίκτυα Bayes</i>	29
<i>Self-Organizing Map</i>	30
<i>Support Vector Machine (SVM)</i>	30
<i>Μετρικές απόδοσης μοντέλων ταξινόμησης</i>	30
ΚΕΦΑΛΑΙΟ 4 – ΤΟ ΠΕΔΙΟ ΜΕΛΕΤΗΣ	33
1. ΤΟ ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΛΑΙΣΙΟ	33
2. ΠΩΣ ΛΕΙΤΟΥΡΓΕΙ Ο ΦΠΑ	34
3. ΤΟ ΠΡΟΒΛΗΜΑ ΠΟΥ ΠΡΟΚΥΠΤΕΙ	35
ΚΕΦΑΛΑΙΟ 5 - ΕΡΕΥΝΗΤΙΚΑ ΕΡΩΤΗΜΑΤΑ ΚΑΙ ΣΥΛΛΟΓΗ ΣΤΟΙΧΕΙΩΝ	38
1. ΣΤΟΧΟΣ/ ΕΡΕΥΝΗΤΙΚΑ ΕΡΩΤΗΜΑΤΑ	38
2. ΣΥΛΛΟΓΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ	38
3. ΑΝΩΝΥΜΙΑ ΔΕΔΟΜΕΝΩΝ	40
4. ΤΑ ΔΕΔΟΜΕΝΑ (ΓΝΩΡΙΣΜΑΤΑ) ΠΟΥ ΘΑ ΧΡΗΣΙΜΟΠΟΙΗΘΟΥΝ	41
ΚΕΦΑΛΑΙΟ 6 - ΤΟ ΤΕΧΝΙΚΟ ΥΠΟΒΑΘΡΟ	44
1. ΕΡΓΑΛΕΙΑ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΗΘΗΚΑΝ	44
2. SQL DEVELOPER – DATA MINER	44
3. WEKA	46
4. KNIME	47
5. CYTOSCAPE	48
ΚΕΦΑΛΑΙΟ 7- ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ	50
1. ΔΙΑΜΟΡΦΩΣΗ ΣΤΟΙΧΕΙΩΝ	50
2. ΕΠΙΣΚΟΠΗΣΗ ΣΤΟΙΧΕΙΩΝ	50
3. ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΜΕ ΤΟ ΕΡΓΑΛΕΙΟ ORACLE DATA MINER	57
4. ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΜΕ ΤΟ ΕΡΓΑΛΕΙΟ WEKA	65

5. ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΜΕ ΤΑ ΕΡΓΑΛΕΙΑ KNIME - CYTOSCAPE.....	73
ΚΕΦΑΛΑΙΟ 8 – ΣΥΜΠΕΡΑΣΜΑΤΑ	81
1. ΔΙΑΘΕΣΙΜΑ ΕΡΓΑΛΕΙΑ	81
2. ΣΥΜΠΕΡΑΣΜΑΤΑ ΑΠΟ ΤΗ ΧΡΗΣΗ ΤΩΝ ΕΡΓΑΛΕΙΩΝ	82
<i>Ευκολία χρήσης</i>	<i>82</i>
<i>Δυνατότητες εργαλείων</i>	<i>82</i>
<i>Χρήση των εργαλείων</i>	<i>83</i>
3. ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΝΑΛΥΣΗΣ ΚΑΙ ΕΞΟΡΥΞΗΣ.....	84
4. ΌΡΙΑ ΚΑΙ ΠΕΡΙΟΡΙΣΜΟΙ.....	86
5. ΣΥΝΟΨΗ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΕΝΕΡΓΕΙΕΣ	87
6. ΓΕΝΙΚΑ ΣΥΜΠΕΡΑΣΜΑΤΑ	88
ΒΙΒΛΙΟΓΡΑΦΙΑ	90
ΠΑΡΑΡΤΗΜΑ 1.....	94
ΠΑΡΑΡΤΗΜΑ 1.....	94
ΠΑΡΑΡΤΗΜΑ 2.....	96
1. ΔΙΑΧΕΙΡΙΣΗ ΜΝΗΜΗΣ.....	96
ΠΑΡΑΡΤΗΜΑ 3.....	97
1. ΑΝΩΝΥΜΟΠΟΙΗΣΗ ΤΟΥ ΑΦΜ	97
2. ΜΕΤΑΣΧΗΜΑΤΙΣΜΟΣ ΑΠΟΛΥΤΩΝ ΠΟΣΩΝ ΣΕ ΣΧΕΤΙΚΟ ΠΟΣΟΣΤΟ	98
ΠΑΡΑΡΤΗΜΑ 4.....	99
1. ΔΗΜΙΟΥΡΓΙΑ ΣΕΤ ΔΕΔΟΜΕΝΩΝ ΓΙΑ ΤΟ ΕΡΓΑΛΕΙΟ WEKA	99
ΠΑΡΑΡΤΗΜΑ 5.....	102

ΠΙΝΑΚΕΣ – ΣΧΗΜΑΤΑ

ΕΙΚΟΝΑ 1- ΤΡΙΓΩΝΟ ΤΗΣ ΑΠΑΤΗΣ (ΠΡΟΣΑΡΜΟΓΗ ΑΠΟ Β.ΒAESENS ET AL - 2015)	12
ΕΙΚΟΝΑ 2- ΔΙΑΔΙΚΑΣΙΕΣ ΕΞΟΥΡΕΣΗΣ ΓΝΩΣΗΣ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΟΥΝΤΑΙ ΑΠΟ ΤΙΣ ΦΟΡΟΛΟΓΙΚΕΣ ΑΡΧΕΣ (P.CASTELLON ET AL- 2012)	16
ΕΙΚΟΝΑ 3- ΔΙΑΓΡΑΜΜΑ ΡΟΗΣ ΕΡΓΑΣΙΩΝ ΚΑΙ ΛΕΙΤΟΥΡΓΙΩΝ (ΑΤΟ, 2008)	17
ΕΙΚΟΝΑ 4- ΧΡΗΣΗ ΤΕΧΝΙΚΩΝ ΕΞΟΥΡΕΣΗΣ ΣΤΗΝ ΙΝΔΙΑ (FICCI, 2015)	18
ΕΙΚΟΝΑ 5- ΕΠΙΣΚΟΠΗΣΗ ΤΩΝ ΒΗΜΑΤΩΝ ΠΟΥ ΑΠΟΤΕΛΟΥΝ ΜΕΡΟΣ ΜΙΑΣ ΔΙΑΔΙΚΑΣΙΑΣ ΕΞΟΥΡΕΣΗΣ ΔΕΔΟΜΕΝΩΝ(FAYYAD ET AL, 1996)	23
ΕΙΚΟΝΑ 6- Η ΕΞΕΛΙΞΗ ΤΗΣ ΕΞΟΥΡΕΣΗΣ ΓΝΩΣΗΣ (ΠΡΟΣΑΡΜΟΓΗ ΑΠΟ THEARLING,2012)	ΣΦΑΛΜΑ! ΔΕΝ ΕΧΕΙ ΟΡΙΣΤΕΙ ΣΕΛΙΔΟΔΕΙΚΤΗΣ.
ΕΙΚΟΝΑ 7- Η ΕΞΕΛΙΞΗ ΤΗΣ ΕΞΟΥΡΕΣΗΣ ΓΝΩΣΗΣ (ΠΡΟΣΑΡΜΟΓΗ ΑΠΟ THEARLING,2012)	25
ΕΙΚΟΝΑ 8- ΤΑ ΒΗΜΑΤΑ ΤΗΣ ΔΙΑΔΙΚΑΣΙΑΣ CRISP (WIKIPEDIA,2015)	26
ΕΙΚΟΝΑ 9- ΑΝΤΙΣΤΟΙΧΙΣΕΙΣ ΜΕΤΑΞΥ ΤΩΝ ΔΙΑΔΙΚΑΣΙΩΝ ΕΞΟΥΡΕΣΗΣ (AZEVEDO & SANTOS, 2008)	27
ΕΙΚΟΝΑ 10- DECISION TREE (WEKA-IRIS DATA SET).....	27
ΕΙΚΟΝΑ 11- RULES (WEKA- WEATHER DATA SET)	28
ΕΙΚΟΝΑ 12- BAGGING (L. ROKACH, O. MAIMON, 2015, ,ΚΕΦ.9.4.1).....	29
ΕΙΚΟΝΑ 13- BAYESNET (WEKA WEATHER DATA SET)	29
ΕΙΚΟΝΑ 14- SVM (WITTEN FIG. 6.9, 2011)	30
ΕΙΚΟΝΑ 15 – CONFUSION MATRIX (STEFANO PISANI, PAOLA DE SISTI, 2007).....	31
ΕΙΚΟΝΑ 16- ΜΕΤΡΙΚΕΣ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΟΥΝΤΑΙ ΓΙΑ ΤΟ ΣΧΗΜΑΤΙΣΜΟ ΤΗΣ ΚΑΜΠΥΛΗΣ ROC	32
ΕΙΚΟΝΑ 17- ΚΑΜΠΥΛΗ ROC	32
ΕΙΚΟΝΑ 18- ΛΕΙΤΟΥΡΓΙΑ ΚΥΚΛΩΜΑΤΟΣ CAROUSEL(1)	35
ΕΙΚΟΝΑ 19 - ΛΕΙΤΟΥΡΓΙΑ ΚΥΚΛΩΜΑΤΟΣ CAROUSEL (2)	36
ΕΙΚΟΝΑ 20 - ΓΝΩΡΙΣΜΑΤΑ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΗΘΗΚΑΝ ΣΤΗΝ ΕΡΓΑΣΙΑ.....	43
ΕΙΚΟΝΑ 21- ORACLE DATA MINER ΓΡΑΦΙΚΟ ΠΕΡΙΒΑΛΛΟΝ(ORACLE.COM, 2015).....	45
ΕΙΚΟΝΑ 22- ORACLE (2012), ORACLE DATA MINING 11g RELEASE 2 - COMPETING ON IN-DATABASE ANALYTICS	45
ΕΙΚΟΝΑ 23 WEKA INTERFACE(I.WITTEN, E.FRANK, M.HALL (2011). "DATA MINING: PRACTICAL MACHINE LEARNING TOOLS AND TECHNIQUES, 3RD EDITION")	46
ΕΙΚΟΝΑ 24- KNIME PRODUCT OVERVIEW (KNIME.COM,2016)	47
ΕΙΚΟΝΑ 25- KNIME ΒΑΣΙΚΟ INTERFACE(KNIME.COM,2016)	47
ΕΙΚΟΝΑ 26- CYTOSCAPE ΑΠΕΙΚΟΝΙΣΗ ΔΙΚΤΥΟΥ (CYTOSCAPE.COM, 2015)	48
ΕΙΚΟΝΑ 27- CYTOSCAPE ΚΕΝΤΡΙΚΟ INTERFACE (CYTOSCAPE.COM,2015).....	49
ΕΙΚΟΝΑ 28- ΠΡΟ-ΕΠΙΣΚΟΠΗΣΗ ΣΤΟΙΧΕΙΩΝ ΜΕ ΤΟ ORACLE DATA MINER	50
ΕΙΚΟΝΑ 29- ΠΡΟ-ΕΠΙΣΚΟΠΗΣΗ ΓΝΩΡΙΣΜΑΤΩΝ ΜΕ ΤΟ WEKA	51
ΕΙΚΟΝΑ 30 - ΠΡΟ-ΕΠΙΣΚΟΠΗΣΗ ΓΝΩΡΙΣΜΑΤΩΝ ΜΕ ΤΟ KNIME (I)	51
ΕΙΚΟΝΑ 31- ΡΟΗ ΓΙΑ ΠΡΟΕΠΙΣΚΟΠΗΣΗ ΣΤΟΙΧΕΙΩΝ ΣΤΟΝ DATA MINER.....	52
ΕΙΚΟΝΑ 32- ΔΙΕΡΕΥΝΗΣΗ ΓΝΩΡΙΣΜΑΤΩΝ	52
ΕΙΚΟΝΑ 33- ΈΛΕΓΧΟΣ ΓΝΩΡΙΣΜΑΤΩΝ	53
ΕΙΚΟΝΑ 34 - ΣΥΓΚΡΙΣΗ ΓΝΩΡΙΣΜΑΤΩΝ ΠΡΙΝ ΚΑΙ ΜΕΤΑ ΤΗΝ ΕΠΕΞΕΡΓΑΣΙΑ.....	53
ΕΙΚΟΝΑ 35- ΑΝΙΣΟΡΡΟΠΙΑ ΚΛΑΣΕΩΝ	54
ΕΙΚΟΝΑ 36- ΠΛΗΘΟΣ ΣΤΟΙΧΕΙΩΝ ΔΕΙΓΜΑΤΟΣ	56
ΕΙΚΟΝΑ 37- ΠΡΟ-ΕΠΙΣΚΟΠΗΣΗ ΣΤΟΙΧΕΙΩΝ ΠΟΥ ΣΥΜΠΕΡΙΛΑΜΒΑΝΟΥΝ ΕΓΓΡΑΦΕΣ ΤΙΣ ΚΛΑΣΗΣ ΜΕΙΟΨΗΦΙΑΣ (<2010).....	57
ΕΙΚΟΝΑ 38- ΕΚΤΕΛΕΣΗ ΤΑΞΙΝΟΜΗΣΗΣ(CLASSIFICATION) ΜΕ ΤΟΝ DATA MINER	59
ΕΙΚΟΝΑ 39- ΑΠΟΔΟΣΗ ΜΟΝΤΕΛΩΝ ΑΝΑΛΟΓΑ ΜΕ ΤΑ ΔΕΔΟΜΕΝΑ ΕΚΠΑΙΔΕΥΣΗΣ	59
ΕΙΚΟΝΑ 40- ΚΑΜΠΥΛΕΣ ROC ΓΙΑ ΔΕΔΟΜΕΝΑ ΕΚΠΑΙΔΕΥΣΗΣ 2010-14+ ΣΤΟΙΧΕΙΑ ΚΛΑΣΗΣ ΜΕΙΟΨΗΦΙΑΣ	60
ΕΙΚΟΝΑ 41- ΚΑΜΠΥΛΕΣ ROC ΓΙΑ ΔΕΔΟΜΕΝΑ ΕΚΠΑΙΔΕΥΣΗΣ 2010-14	60
ΕΙΚΟΝΑ 42- ΚΑΜΠΥΛΕΣ ROC, ΜΕΤΡΙΚΕΣ	61
ΕΙΚΟΝΑ 43- ΒΑΡΥΤΗΤΑ ΤΩΝ ΚΛΑΣΕΩΝ ΣΤΟΝ SVM (DATA MINER)	61
ΕΙΚΟΝΑ 44- ΓΝΩΡΙΣΜΑΤΑ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΕΙ Ο SVM (DATA MINER)	62
ΕΙΚΟΝΑ 45- DECISION TREE (DATA MINER)	62
ΕΙΚΟΝΑ 46- ΟΜΑΔΟΠΟΙΗΣΗ (CLUSTERING - DATA MINER).....	62
ΕΙΚΟΝΑ 47- ΟΜΑΔΟΠΟΙΗΣΗ ΜΕ ΤΟΝ O-CLUSTER (DATA MINER).....	63
ΕΙΚΟΝΑ 48- ΟΜΑΔΟΠΟΙΗΣΗ ΜΕ ΤΟΝ ΑΛΓΟΡΙΘΜΟ K-MEANS (DATA MINER)	64
ΕΙΚΟΝΑ 49- ΑΠΟΔΟΣΗ ΑΛΓΟΡΙΘΜΟΥ ONE1 (WEKA)	65
ΕΙΚΟΝΑ 50- ΑΠΟΔΟΣΗ ΑΛΓΟΡΙΘΜΟΥ J-48 (WEKA).....	66
ΕΙΚΟΝΑ 51- ΑΠΟΔΟΣΗ ΑΛΓΟΡΙΘΜΟΥ RANDOM TREE (WEKA)	66
ΕΙΚΟΝΑ 52- ΑΠΟΔΟΣΗ ΑΛΓΟΡΙΘΜΟΥ RANDOM FOREST (WEKA).....	66
ΕΙΚΟΝΑ 53- ΑΠΟΔΟΣΗ ΑΛΓΟΡΙΘΜΟΥ NAIVE BAYES (WEKA).....	66

ΕΙΚΟΝΑ 54- ΑΠΟΔΟΣΗ ΑΛΓΟΡΙΘΜΟΥ KSTAR (WEKA).....	66
ΕΙΚΟΝΑ 55- ΑΠΟΔΟΣΗ ΑΛΓΟΡΙΘΜΟΥ SMO (WEKA).....	66
ΕΙΚΟΝΑ 56- ΧΡΗΣΗ ΒΑΡΩΝ (WEKA)	67
ΕΙΚΟΝΑ 57- ΑΛΓΟΡΙΘΜΟΣ SMO ΜΕ ΒΑΡΗ (WEKA)	67
ΕΙΚΟΝΑ 58- ΑΛΓΟΡΙΘΜΟΣ ONER ΜΕ ΒΑΡΗ (WEKA)	67
ΕΙΚΟΝΑ 59- ΑΛΓΟΡΙΘΜΟΣ IBK ΜΕ ΒΑΡΗ (K=3) (WEKA)	67
ΕΙΚΟΝΑ 60- ΑΛΓΟΡΙΘΜΟΣ REP TREE ΜΕ ΒΑΡΗ (WEKA).....	68
ΕΙΚΟΝΑ 61- ΑΛΓΟΡΙΘΜΟΣ ONER (BAGGING - WEKA)	68
ΕΙΚΟΝΑ 62- ΑΛΓΟΡΙΘΜΟΣ REP TREE (BAGGING - WEKA)	68
ΕΙΚΟΝΑ 63- ΙΕΡΑΡΧΗΣΗ ΓΝΩΡΙΣΜΑΤΩΝ (INFOGAINATTRIBUTEVAL, WEKA)	69
ΕΙΚΟΝΑ 64- ΓΝΩΡΙΣΜΑΤΑ ΠΟΥ ΣΥΝΔΕΟΝΤΑΙ ΜΕ ΤΗΝ ΚΛΑΣΗ (INFOGAINATTRIBUTEVAL, WEKA)	69
ΕΙΚΟΝΑ 65- ΑΛΓΟΡΙΘΜΟΣ JRIP (WEKA)	70
ΕΙΚΟΝΑ 66- ΑΛΓΟΡΙΘΜΟΣ KSTAR (WEKA)	70
ΕΙΚΟΝΑ 67- ΑΛΓΟΡΙΘΜΟΣ NAIVE BAYES (WEKA).....	70
ΕΙΚΟΝΑ 68- ΑΛΓΟΡΙΘΜΟΣ NAIVE BAYES (COSTBASED, WEKA)	70
ΕΙΚΟΝΑ 69- STACKING JRIP, NAIVE BAYES, J-48 ΜΕ ΑΛΓΟΡΙΘΜΟ ΕΠΙΛΟΓΗΣ SIMPLELOGISTIC (WEKA)	70
ΕΙΚΟΝΑ 70- ΑΛΓΟΡΙΘΜΟΣ IBK (COSTBASED, WEKA)	71
ΕΙΚΟΝΑ 71- ΑΛΓΟΡΙΘΜΟΣ JRIP (WEKA)	71
ΕΙΚΟΝΑ 72- ΑΛΓΟΡΙΘΜΟΣ NAIVE BAYES (WEKA).....	72
ΕΙΚΟΝΑ 73- ΣΧΗΜΑΤΙΣΜΟΣ ΟΜΑΔΩΝ (WEKA)	72
ΕΙΚΟΝΑ 74- ΣΥΣΧΕΤΙΣΗ ΟΜΑΔΩΝ ΜΕ ΤΗΝ ΚΛΑΣΗ (WEKA)	72
ΕΙΚΟΝΑ 75- ΔΙΑΘΕΣΙΜΑ ΕΡΓΑΛΕΙΑ (KNIME).....	73
ΕΙΚΟΝΑ 76- ΣΧΗΜΑΤΙΣΜΟΣ ΔΙΚΤΥΟΥ ΑΠΟ ΤΟ ΠΡΟΓΡΑΜΜΑ KNIME.....	75
ΕΙΚΟΝΑ 77- ΕΠΙΚΟΙΝΩΝΙΑ KNIME - CYTOSCAPE	75
ΕΙΚΟΝΑ 78- ΑΡΧΙΚΗ ΑΠΕΙΚΟΝΙΣΗ ΔΙΚΤΥΟΥ ΣΥΝΑΛΛΑΓΩΝ (CYTOSCAPE).....	76
ΕΙΚΟΝΑ 79- ΧΡΩΜΑΤΙΣΜΟΣ ΚΟΜΒΩΝ (CYTOSCAPE).....	76
ΕΙΚΟΝΑ 80- ΚΑΘΟΡΙΣΜΟΣ ΕΤΙΚΕΤΑΣ ΚΟΜΒΟΥ (CYTOSCAPE).....	77
ΕΙΚΟΝΑ 81- ΚΑΘΟΡΙΣΜΟΣ ΣΧΗΜΑΤΟΣ ΚΟΜΒΟΥ (CYTOSCAPE).....	77
ΕΙΚΟΝΑ 82- ΔΙΚΤΥΟ ΣΧΕΣΕΩΝ ΕΠΙΧΕΙΡΗΣΕΩΝ (CYTOSCAPE).....	77
ΕΙΚΟΝΑ 83- ΔΙΚΤΥΟ ΜΕ "ΦΤΩΧΗ" ΠΛΗΡΟΦΟΡΙΑ.....	78
ΕΙΚΟΝΑ 84- ΔΙΚΤΥΑ ΣΥΝΑΛΛΑΓΩΝ (CYTOSCAPE).....	79
ΕΙΚΟΝΑ 85- ΠΟΛΥΠΛΟΚΟ ΔΙΚΤΥΟ ΣΥΝΑΛΛΑΓΩΝ (CYTOSCAPE).....	80
ΕΙΚΟΝΑ 86- ΑΝΑΛΥΣΗ ΑΓΟΡΑΣ ΑΠΟ ΤΗΝ GARTNER(2014).....	81
ΕΙΚΟΝΑ 87- ΒΑΘΜΟΛΟΓΗΣΗ ΤΩΝ ΕΡΓΑΛΕΙΩΝ (DM, WEKA, KNIME).....	83
ΕΙΚΟΝΑ 88- ΑΠΟΔΟΣΗ ΑΛΓΟΡΙΘΜΟΥ J-48.....	84
ΕΙΚΟΝΑ 89- ΑΠΟΔΟΣΗ ΑΛΓΟΡΙΘΜΟΥ NAIVE BAYES.....	85
ΕΙΚΟΝΑ 90- ΑΠΟΔΟΣΗ ΑΛΓΟΡΙΘΜΟΥ KSTAR.....	85
ΕΙΚΟΝΑ 91- Η ΣΗΜΑΣΙΑ ΕΝΟΣ ΟΛΟΚΛΗΡΩΜΕΝΟΥ ΣΧΕΔΙΑΣΜΟΥ (IBM,PARRY 2010-2012).....	89
ΕΙΚΟΝΑ 92 - ΑΠΟΔΟΣΗ ΦΠΑ ΣΕ ΕΘΝΙΚΕΣ ΣΥΝΑΛΛΑΓΕΣ	94
ΕΙΚΟΝΑ 93 - ΑΠΟΔΟΣΗ ΦΠΑ ΣΕ ΕΝΔΟΚΟΙΝΟΤΙΚΕΣ ΣΥΝΑΛΛΑΓΕΣ	95
ΕΙΚΟΝΑ 94- ΑΠΑΙΤΗΣΕΙΣ ΣΕ ΜΝΗΜΗ (WEKA).....	96
ΕΙΚΟΝΑ 95- ORACLE, ΔΗΜΙΟΥΡΓΙΑ ΚΡΥΠΤΟΓΡΑΦΗΜΕΝΟΥ HASH	97
ΕΙΚΟΝΑ 96- ΔΗΜΙΟΥΡΓΙΑ ΤΥΧΑΙΟΥ ΚΛΕΙΔΙΟΥ	98
ΕΙΚΟΝΑ 97- ΕΥΡΕΣΗ % ΜΕΤΑΒΟΛΗΣ ΜΕΤΑΞΥ ΣΥΝΑΛΛΑΓΩΝ ΔΥΟ ΤΡΙΜΗΝΩΝ	98
ΕΙΚΟΝΑ 98- ΠΡΟΣΒΑΣΗ ΣΕ ΒΑΣΗ ΔΕΔΟΜΕΝΩΝ (WEKA).....	99
ΕΙΚΟΝΑ 99- ΆΝΤΛΗΣΗ ΔΕΔΟΜΕΝΩΝ ΑΠΟ ΒΑΣΗ (WEKA).....	100
ΕΙΚΟΝΑ 100- ΔΙΑΓΡΑΦΗ ΓΝΩΡΙΣΜΑΤΩΝ (WEKA)	100
ΕΙΚΟΝΑ 101- ΑΠΟΘΗΚΕΥΣΗ ΤΡΕΧΟΝΤΟΣ ΣΕΤ ΔΕΔΟΜΕΝΩΝ ΣΕ ΑΡΧΕΙΟ ARFF (WEKA).....	101
ΕΙΚΟΝΑ 102- ΠΕΡΙΓΡΑΦΗ ΓΝΩΡΙΣΜΑΤΩΝ (WEKA).....	101
ΕΙΚΟΝΑ 103- ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΕΛΕΓΧΟΣ ΜΟΝΤΕΛΩΝ ΜΕ ΠΛΗΡΕΣ ΣΕΤ ΔΕΔΟΜΕΝΩΝ (WEKA)	102
ΕΙΚΟΝΑ 104- ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΕΛΕΓΧΟΣ ΜΟΝΤΕΛΩΝ ΜΕ ΠΛΗΡΕΣ ΣΕΤ ΔΕΔΟΜΕΝΩΝ (WEKA)	103
ΕΙΚΟΝΑ 105- ΕΚΠΑΙΔΕΥΣΗ ΜΟΝΤΕΛΩΝ ΜΕ ΤΑ ΔΕΔΟΜΕΝΑ ΕΛΕΓΧΟΥ (WEKA)	104
ΕΙΚΟΝΑ 106- ΕΚΠΑΙΔΕΥΣΗ ΜΟΝΤΕΛΩΝ ΧΡΗΣΙΜΟΠΟΙΩΝΤΑΣ ΜΟΝΟ ΤΑ ΣΗΜΑΝΤΙΚΟΤΕΡΑ ΓΝΩΡΙΣΜΑΤΑ (WEKA)	104

Εισαγωγή

Η απατηλή συμπεριφορά των φορολογούμενων, ειδικά των επιχειρήσεων 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 φορολογούμενων.

Στόχος της φορολογικής διοίκησης είναι να περιορίσει αυτή την κακή συμπεριφορά των φορολογούμενων ώστε να σταματήσει η οικονομική αιμορραγία, αλλά και να αποκατασταθεί η ομαλή κοινωνική και οικονομική λειτουργία.

Για να συμβεί αυτό θα πρέπει η φορολογική διοίκηση να έχει μια σαφή εικόνα των τρόπων που λειτουργούν οι παραβάτες καθώς και των αιτιών που προκαλούν αυτά τα φαινόμενα. Με τη γνώση των παραπάνω η φορολογική διοίκηση μπορεί να λάβει τα απαραίτητα μέτρα ώστε να αντιμετωπίσει τα αποτελέσματα αλλά κυρίως τις αιτίες αυτών των συμπεριφορών.

Ένας από τους σημαντικότερους φόρους που συλλέγει το κράτος είναι ο Φόρος Προστιθέμενης Αξίας (Φ.Π.Α.). Η κακή φορολογική συμπεριφορά των επιτηδευματιών και των επιχειρήσεων στο τομέα του ΦΠΑ πέρα από τις άμεσες συνέπειες που επιφέρει, δηλαδή την μείωση των φορολογικών εσόδων, επιφέρει και μια σειρά παράπλευρων επιπτώσεων στην υπόλοιπη οικονομία και κοινωνία.

Οι επιπτώσεις αφορούν κυρίως τρεις τομείς, (1) στέρηση εσόδων από τον προϋπολογισμό, (2) επιβάρυνση του κράτους με το κόστος της καταπολέμησης του φαινομένου, (3) στρέβλωση της αγοράς λόγω αθέμιτου ανταγωνισμού. Εξαιτίας των σοβαρών προβλημάτων που προκαλεί το φαινόμενο είναι απαραίτητη η διαρκής προσπάθεια καταπολέμησης αυτού. Ωστόσο οι διαδικασίες ελέγχου είναι επίπονες, χρονοβόρες και υψηλής εξειδίκευσης, γεγονός που τις κατατάσσει στις κοστοβόρες διαδικασίες.

Μια μέθοδος που θα μπορούσε να βοηθήσει στη μείωση του κόστους του εντοπισμού και της διερεύνησης των υποθέσεων είναι η τεχνική εξόρυξης γνώσης από τα υπάρχοντα φορολογικά δεδομένα.

Ουσιαστικά με την εφαρμογή των τεχνικών εξόρυξης προσπαθούμε να δημιουργήσουμε μια επαναλαμβανόμενη διαδικασία κατά την οποία επιτυγχάνεται πρόοδος μέσω αυτόματων ή ημιαυτόματων μεθόδων. Η διαδικασία είναι χρήσιμη σε περιπτώσεις που δεν γνωρίζουμε εκ των προτέρων ποιες περιπτώσεις μπορούν να χαρακτηρισθούν ως ενδιαφέρουσες καθώς και ποια είναι τα χαρακτηριστικά αυτών (Kirkos et all, 2006).

Στην παρούσα εργασία θα προσπαθήσουμε να εντοπίσουμε ορισμένα στοιχεία, διαδικασίες, εργαλεία, που θα μπορούσαν να βοηθήσουν στην εφαρμογή των διαδικασιών εξόρυξης γνώσης σε πρακτικό επίπεδο.

Η εργασία που ακολουθεί έχει την παρακάτω δομή, στο κεφάλαιο 1 θα αναφερθούμε στην έννοια της φορολογικής απάτης, τις μορφές που αυτή εμφανίζεται και πώς αντιμετωπίζεται. Στο κεφάλαιο 2, θα

συνοψίσουμε άλλες παρόμοιες μελέτες και τα αποτελέσματα αυτών. Επίσης θα αναφερθούμε σε πρακτικές εφαρμογές της εξόρυξης από άλλες χώρες. Στο κεφάλαιο 3 θα κάνουμε την σύνδεση των τεχνικών εξόρυξης και των αναγκών της διοίκησης. Θα αναφερθούμε εν συντομία στα χαρακτηριστικά των αλγόριθμων εξόρυξης, στις διαδικασίες εξόρυξης, καθώς και στο τρόπο μέτρησης της απόδοσης των αλγόριθμων. Στο κεφάλαιο 4 περιγράφεται το επιχειρηματικό πεδίο στο οποίο θα ασχοληθούμε. Στο κεφάλαιο 5 τίθενται τα ερευνητικά ερωτήματα της παρούσας εργασίας και εξετάζουμε τα διαθέσιμα γνωρίσματα πάνω στα οποία θα εφαρμοσθούν οι διαδικασίες εξόρυξης γνώσης. Στο κεφάλαιο 6 εξετάζουμε τα διαθέσιμα εργαλεία που θα χρησιμοποιηθούν. Στο κεφάλαιο 7 παρουσιάζουμε μια σειρά αποτελεσμάτων τα οποία προέκυψαν κατά την εφαρμογή των αλγόριθμων εξόρυξης γνώσης, μέσω των διάφορων εργαλείων. Τέλος στο κεφάλαιο 8 παραθέτουμε τα συμπεράσματα από τα εργαλεία, τους αλγόριθμους που χρησιμοποιήθηκαν καθώς και από την διαδικασία συνολικά. Τα παραρτήματα που ακολουθούν περιέχουν πρόσθετες πληροφορίες σχετικά με θέματα που προέκυψαν κατά την διάρκεια της εργασίας. Στο παράρτημα 1 παρατίθεται παραδείγματα απόδοσης του ΦΠΑ για να γίνουν κατανοητά τα διάφορα βήματα που λαμβάνουν χώρα κατά τη διακίνηση των αγαθών από χώρα σε χώρα. Στο παράρτημα 2 γίνεται αναφορά σε θέματα ρυθμίσεων τα οποία προέκυψαν κατά τη διάρκεια της εργασίας. Στο παράρτημα 3 γίνεται αναφορά στην διαδικασία ανωνυμοποίησης των δεδομένων. Στο παράρτημα 4 περιγράφεται η δημιουργία αρχείων για χρήση από το εργαλείο WEKA. Στο παράρτημα 5 παρουσιάζονται όλα τα αποτελέσματα που προέκυψαν από την ανάλυση των δεδομένων από το εργαλείο WEKA.

ΚΕΦΑΛΑΙΟ 1 - Εισαγωγή στο πρόβλημα

Σύμφωνα με τον ορισμό της εγκυκλοπαίδειας Britanica (2015), η διαδικασία εξόρυξης γνώσης έχει σαν σκοπό, από δεδομένα τα οποία είναι διαθέσιμα, να ανακαλύψει ενδιαφέροντα και χρήσιμα επαναλαμβανόμενα μοτίβα. Από τον ορισμό της εξόρυξης γνώσης γίνεται κατανοητό ότι αναφερόμαστε σε μια διαδικασία η οποία θα λάβει ως είσοδο δεδομένα και θα δώσει ως έξοδο γνώση σχετικά με τον τρόπο με τον οποίο «οργανώνονται» τα δεδομένα.

Η μετατροπή των δεδομένων σε χρήσιμη γνώση επιτυγχάνεται με τη χρήση μαθηματικών μεθόδων και στατιστικών τεχνικών. Η ερμηνεία και η βαρύτητα των αποτελεσμάτων που δίνουν οι μέθοδοι εξαρτάται από το πεδίο (context) στο οποίο εφαρμόζονται, τη γνώση την οποία αναζητούμε, αλλά και τις συνθήκες πχ η ανακάλυψη, ότι τα αποτελέσματα ενός ποδοσφαιρικού αγώνα δύο συγκεκριμένων ομάδων λήγουν πάντα υπέρ της μιας ομάδας αν ο αγώνας γίνεται σε συνθήκες βροχής, μπορεί να είναι ισχυρός κανόνας ή αδιάφορος, ανάλογα με το πεδίο στο οποίο αναζητάμε τη γνώση και τις συνθήκες που ισχύουν. Στο συγκεκριμένο παράδειγμα γνωρίσματα στα οποία μπορούμε να βασιστούμε είναι οι συνθέσεις των ομάδων, το ζητούμενο αποτέλεσμα, την πιθανότητα να υπάρχει βροχερός καιρός την ημέρα διεξαγωγής του αγώνα κλπ.

Οι επιλογή των γνωρισμάτων προϋποθέτει ότι υπάρχει η προηγούμενη γνώση του πεδίου όπως, τι είναι ένας ποδοσφαιρικός αγώνας, πώς διεξάγεται, ποιοι παράγοντες επηρεάζουν τη διεξαγωγή του κλπ. Η προϋπάρχουσα γνώση είναι αυτή που θα μας καθοδηγήσει (α) στη συλλογή των καταλληλότερων δεδομένων, (β) στην ερμηνεία και αποδοχή των αποτελεσμάτων της διαδικασίας εξόρυξης γνώσης.

Στη φορολογική διοίκηση οι διαδικασίες εξόρυξης γνώσης μπορούν να χρησιμοποιηθούν για έναν από τους παρακάτω σκοπούς :

- Τον εντοπισμό της απάτης
- Την ανακάλυψη και ομαδοποίηση των φορολογούμενων

Καθώς στη παρούσα εργασία θα επικεντρωθούμε στην πρώτη περίπτωση θα πρέπει να καθορίσουμε αρχικά τι είναι «απάτη». Ο ορισμός της έννοιας «απάτη», θα μας βοηθήσει ώστε να επιλέξουμε τα γνωρίσματα και τα δεδομένα που θα χρησιμοποιήσουμε αλλά και στην ερμηνεία των αποτελεσμάτων.

Σύμφωνα με το λεξικό του Cambridge(2015), ως απάτη ορίζεται «*Το έγκλημα της απόσπασης χρημάτων μέσω της εξαπάτησης ανθρώπων*».

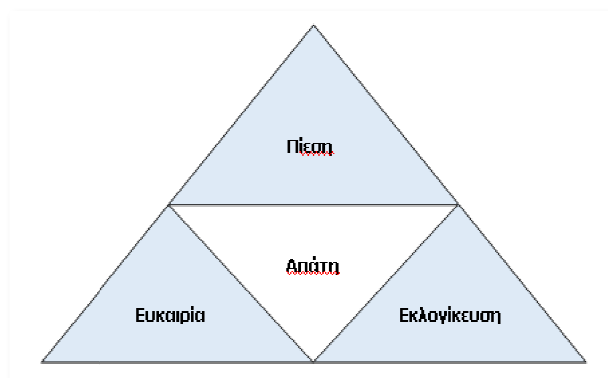
Όμοια στο λεξικό του Oxford(2015), ως απάτη ορίζεται «*Η Παράνομη ή εγκληματική παραπλάνηση που προορίζεται να οδηγήσει σε οικονομικό ή προσωπικό κέρδος*».

Οι παραπάνω ορισμοί δίνουν μια γενική προσέγγιση στον ορισμό. Αν θελήσουμε να εξειδικεύσουμε τον ορισμό για το φορολογικό πεδίο, βρίσκουμε περισσότερο στοχευμένη την περιγραφή που δίνεται από τους Β.

Baesens et al (2015):

Η απάτη είναι ένα ασυνήθιστο, καλά μελετημένο, έγκλημα το οποίο εξελίσσεται στο πέρασμα του χρόνου και είναι συνήθως προσεκτικά οργανωμένο και ανεπαίσθητα ορατό.

Όπως αναφέρεται στους B. Baesens et al (2015), οι συνθήκες κάτω από τις οποίες εμφανίζεται η απάτη μπορούν να απεικονισθούν με το «τρίγωνο της απάτης».



Εικόνα 1- Τρίγωνο της απάτης (Προσαρμογή από B.Baesens et al - 2015)

Το μοντέλο προσπαθεί να εξηγήσει τους παράγοντες οι οποίοι μπορούν να οδηγήσουν κάποιον να διαπράξει μια «επαγγελματική» απάτη, αλλά ταυτόχρονα παρέχει και μια ευρύτερη ματιά στο φαινόμενο της απάτης. Σύμφωνα με το μοντέλο υπάρχουν τρεις βασικοί παράγοντες οι οποίοι συνδυαζόμενοι οδηγούν σε απατηλή συμπεριφορά:

- Πίεση

Αποτελεί ίσως το σημαντικότερο κίνητρο για τη διάπραξη της απάτης. Κάποιος θα διαπράξει απάτη όταν αντιμετωπίζει ένα πρόβλημα οικονομικό, κοινωνικό ή οποιασδήποτε άλλης μορφής και δεν μπορεί να το λύσει με νόμιμο τρόπο.

- Ευκαιρία

Η ευκαιρία είναι προϋπόθεση για τη διάπραξη απάτης. Κάποιος θα διαπράξει απάτη μόνο αν υπάρχει η ευκαιρία να την υποκρύψει.

- Εκλογίκευση

Είναι η ενεργοποίηση ενός ψυχολογικού μηχανισμού που προσπαθεί να δικαιολογήσει γιατί διαπράττεται η απάτη.

Σύμφωνα με το ορισμό η διάπραξη μιας απάτης είναι ένα σύνθετο φαινόμενο στο οποίο υπεισέρχονται πολλαπλοί παράγοντες. Θα μπορούσε δε να χαρακτηριστεί ως κοινωνικό φαινόμενο, με την έννοια ότι

εμπλεκόμενοι είναι άτομα, ομάδες, επιχειρήσεις και τα κράτη μέσω των φορολογικών μηχανισμών τους.

Η απάτη στο φορολογικό τομέα μπορεί να εντοπισθεί σε διάφορα επίπεδα. Είτε σε πολύ υψηλό, με την εμφάνιση δομών υψηλής ικανότητας, όπως στις περιπτώσεις απάτης με τις εκπομπές ρύπων ή το σύστημα Carousel είτε σε χαμηλότερο επίπεδο με την έκδοση εικονικών ή πλαστών παραστατικών.

Κάθε περίπτωση έχει τα δικά της ιδιαίτερα χαρακτηριστικά τόσο από πλευράς σχεδιασμού, όσο και από πλευράς μεθόδων, τρόπου και χρόνου διάπραξης.

Κάθε μορφή απάτης είναι σημαντική καθώς η οικονομική ζημιά μπορεί να είναι σε κάθε περίπτωση μεγάλη. Χαρακτηριστικό παράδειγμα απάτης υψηλού σχεδιασμού είναι η απάτη στο σύστημα εμπορίου ρύπων. Το 2009 η Europol, ειδοποίησε τα Ευρωπαϊκά κράτη για ζημιές περίπου 5 δις ευρώ από απώλειες σε ΦΠΑ, που προκάλεσε η απάτη στο πλαίσιο του συστήματος εμπορίας εκπομπών της ΕΕ (ETS). Η απάτη αφορούσε την απόκρυψη ΦΠΑ κατά την αγοροπωλησία τίτλων ρύπων και την εκμετάλλευση νομικών κενών στα φορολογικά συστήματα των κρατών.

Ζημιές επίσης μπορούν να προέλθουν από πολλές περιπτώσεις μικρότερης απάτης. Οι Gebauer et al. (2007), όπως αναφέρεται από τους Wu et al (2012), υποστηρίζουν ότι, πολλές μικρές περιπτώσεις απάτης μπορούν να προκαλέσουν ένα τεράστιο ποσό των φορολογικών ζημιών, οι οποίες να δημιουργήσουν μια πιο σοβαρή ζημιά από την ύπαρξη λίγων μεγάλων περιπτώσεων. Επιπλέον, η διοικητική επιβάρυνση, το σχετικό προσωπικό, και το κόστος του εξοπλισμού που αυτή η κατάσταση προκαλεί είναι υποτιμημένα.

Σε κάθε περίπτωση οι φορολογικές αρχές θα πρέπει να ερευνήσουν τις διαθέσιμες επιλογές για να κάνουν το σημερινό σύστημα ΦΠΑ πιο αποτελεσματικό στην πρόληψη και τον εντοπισμό της φοροδιαφυγής, ανεξάρτητα μεγέθους και σχηματισμού της απάτης.

ΚΕΦΑΛΑΙΟ 2 - Μελέτες και εφαρμογές

1. Υπάρχουσες μελέτες

Η φορολογική διοίκηση καθημερινά μαζεύει ένα μεγάλο εύρος πληροφοριών που την αφορούν ενώ παράλληλα οι διαδικασίες που είναι υποχρεωμένη να εφαρμόζει αποτελούνται από ένα σύνθετο πλαίσιο κανόνων, οι οποίοι πολλές φορές μεταβάλλονται στο πέρασμα του χρόνου. Το περιβάλλον που δημιουργείται είναι δύσκολο να ελεγχθεί με συμβατικά μέσα, αντίθετα αποτελεί ένα ιδανικό πεδίο έρευνας για την εφαρμογή μεθόδων εξόρυξης γνώσης. Αρκετές μελέτες έχουν γίνει για την εφαρμογή των τεχνικών εξόρυξης σε οικονομικές-φορολογικές περιπτώσεις.

Ο Spathis (2002) και Spathis et al (2002) χρησιμοποίησε τεχνικές εξόρυξης δεδομένων για να βρει παράγοντες που επηρεάζουν την πιθανότητα μια επιχείρηση να έχει υποβάλει στρεβλές οικονομικές καταστάσεις. Για το μοντέλο τους χρησιμοποίησαν ποσοστιαίες αναλογίες. Σκοπός της μελέτης ήταν να εξετάσει την υπόθεση ότι όταν τα περιουσιακά στοιχεία ή τα κέρδη είχαν διογκωθεί ή όταν οι υποχρεώσεις, τα έξοδα ή οι ζημιές είχαν υποτιμηθεί τότε η πιθανότητα να έχουν στρεβλωθεί οι οικονομικές καταστάσεις ήταν μεγάλη. Με βάση αυτή την παρατήρηση και αφού εξέτασαν πλήθος μεταβλητών, κατέληξαν σε 10 μεταβλητές οι οποίες χρησιμοποιήθηκαν για να δημιουργήσουν το κατάλληλο μοντέλο ταξινόμησης των επιχειρήσεων.

Οι DeBarr D. et al (2005) χρησιμοποίησαν στοιχεία της Αμερικανικής υπηρεσίας IRS. Σκοπός της έρευνας ήταν να εντοπισθούν οι παράμετροι εκείνοι που καθορίζουν αν κάποιος έχει στήσει ένα κύκλωμα φοροδιαφυγής, χρησιμοποιώντας εικονικές μεγάλες ζημιές για να αποκρύψουν μεγάλα κέρδη. Για την έρευνα των παραμέτρων χρησιμοποιήθηκε ο αλγόριθμος SVM. Με τη χρήση του αλγόριθμου η μελέτη μπόρεσε να δημιουργήσει ένα μοντέλο για την ταξινόμηση των περιπτώσεων.

Οι M. Kallio και B. Back (2011) βασίστηκαν στον αλγόριθμο SOM (Self Organized Map). Ο αλγόριθμος ανήκει στην οικογένεια των νευρωνικών δικτύων. Σκοπός της έρευνας ήταν να μελετηθεί η δυνατότητα να χρησιμοποιηθεί ο συγκεκριμένος αλγόριθμος για τον εντοπισμό περιπτώσεων οι οποίες θα πρέπει να δοθούν για περαιτέρω έλεγχο. Τα αποτελέσματα της μελέτης ήταν μικτά. Δημιουργήθηκαν διάφορα μοντέλα τα οποία παρουσιάζουν προβλήματα στη γενίκευση των αποτελεσμάτων. Πιθανές αιτίες σύμφωνα με τους μελετητές είναι η υπερεκπαίδευση των μοντέλων, η επιλογή των κριτηρίων και η επιλογή των αρχικών δεδομένων.

Οι Kotsiantis et al(2007), στη μελέτη τους χρησιμοποίησαν πολλαπλούς αλγόριθμους εξόρυξης γνώσης με τη μορφή «στοίβας» (stacking). Τα αποτελέσματα της μελέτης συμφωνούσαν με αυτά προηγούμενων εργασιών. Οι χρήση της τεχνικής «στοίβας» εμφανίζει καλύτερα αποτελέσματα από τους μεμονωμένους αλγόριθμους,

και συνολικά η διαδικασία εξόρυξης γνώσης μπορεί να δώσει αποτελέσματα με μεγάλο αριθμό ακρίβειας. Ο εμπλουτισμός των παραμέτρων με επιπλέον παραμέτρους από διαφορετικές περιοχές είναι επίσης πιθανό να αυξάνει την ακρίβεια των μοντέλων.

Οι P. Castellon et al (2012), μελέτησαν τη χρήση διάφορων τεχνικών εξόρυξης γνώσης στην προσπάθεια τους να κατανοήσουν τις παραμέτρους που μπορούν να χρησιμοποιηθούν για τον εντοπισμό επιχειρήσεων με ψευδή τιμολόγια. Σύμφωνα με τα αποτελέσματα της μελέτης τους τα δέντρα αποφάσεων (decision trees), μπορούν να χρησιμοποιηθούν ώστε να εντοπισθούν τα γνωρίσματα που θα χαρακτηρίζουν την κλάση. Έχοντας εντοπίσει τα κυριότερα γνωρίσματα στη συνέχεια χρησιμοποίησαν αλγόριθμους νευρωνικών δικτύων για να κατασκευάσουν τα μοντέλα.

Οι Wu et al (2012), συγκέντρωσαν στοιχεία από αποτελέσματα ελέγχων, το μητρώο των εταιρειών και δηλώσεις ΦΠΑ των εταιρειών. Με βάση τα στοιχεία αυτά και χρησιμοποιώντας τεχνικές εξόρυξης προσπάθησαν να δημιουργήσουν ένα πλαίσιο ελέγχου, ώστε να διαχωρίσουν τις πιθανότερες για έλεγχο εταιρείες από τις υπόλοιπες. Για την μελέτη τους επέλεξαν να χρησιμοποιήσουν κανόνες συσχέτισης (fast frequent-pattern tree algorithm). Σύμφωνα με τα αποτελέσματα της μελέτης τους η χρήση τεχνικών εξόρυξης θα μπορούσε να χρησιμοποιηθεί ώστε να επιλέγονται υποθέσεις για έλεγχο, καθώς περιορίζει τους απαιτούμενους πόρους για την επιλογή των υποθέσεων, ενώ παράλληλα στατιστικά παρουσιάζει καλύτερα ποσοστά επιτυχών προβλέψεων.

Οι Phua et al(2006), δημιούργησαν ένα μοντέλο πρόβλεψης βασισμένοι σε μια υβριδική προσέγγιση. Αρχικά χρησιμοποιούν τεχνικές «στοίβας» για να επιλέξουν τους αλγόριθμους με τις καλύτερες επιδόσεις στην εύρεση αποτελεσμάτων για τη μειωτική κλάση και στη συνέχεια συνδυάζουν τους αλγόριθμους με τη τεχνική bagging. Σύμφωνα με τα αποτελέσματα της εργασίας τους ο συνδυασμός αυτός παρουσιάζει ελαφρά καλύτερα αποτελέσματα από την χρήση πολλαπλών τεμαχισμών (folding) του δείγματος σε μέρη τα οποία περιέχουν διαφορετικά ποσοστά συμμετοχής των κλάσεων.

Οι M. Gupta και V. Nagadevara, (2010), χρησιμοποίησαν διάφορους αλγόριθμους για να μετρήσουν την απόδοση τους στο πεδίο του ΦΠΑ. Ο ΦΠΑ στην Ινδία, με την έννοια που γνωρίζουμε στην Ευρώπη το 2010 ήταν ένας καινούργιος φόρος. Σκοπός της μελέτης ήταν να εντοπίσει, εάν η χρήση τεχνικών εξόρυξης μπορεί να βοηθήσει στον εντοπισμό περιπτώσεων για έλεγχο από τις φορολογικές αρχές. Σύμφωνα με τα αποτελέσματα της μελέτης η χρήση των αλγόριθμων εξόρυξης μπορεί να μην εμφάνισε μεγάλα ποσοστά ακρίβειας, ωστόσο σε κάθε περίπτωση κυμάνθηκε από 2 – 3.5 φορές σε υψηλότερα επίπεδα από την τυχαία επιλογή περιπτώσεων, ανάλογα με την χρησιμοποιούμενη μέθοδο.

Εκτός από τις μελέτες που σχετίζονται με το θέμα αρκετά κράτη μέσω των ελεγκτικών μηχανισμών που διαθέτουν έχουν αναπτύξει και χρησιμοποιούν διάφορες μεθόδους εξόρυξης γνώσης για να βοηθηθούν στο

έργο τους.

Στη μελέτη των P. Castellon et al (2012) παρουσιάζεται ο παρακάτω μη εξαντλητικός πίνακας των μεθόδων εξόρυξης που χρησιμοποιούν ορισμένες χώρες:

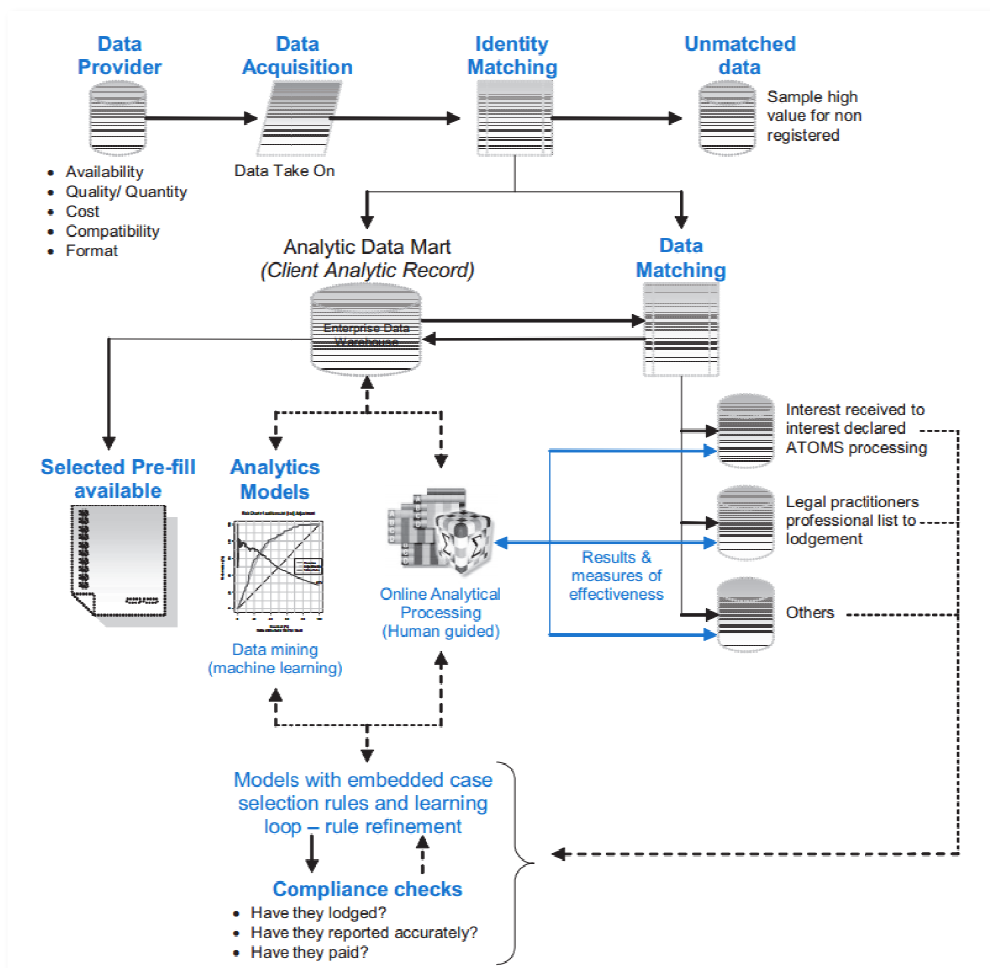
Technique Applied	USA	Canada	Australia	UK	Bulgaria	Brazil	Peru	Chile
Neural Networks	✓	✓		✓	✓		✓	✓
Decision Tree	✓	✓	✓				✓	✓
Logistic Regression	✓		✓	✓	✓			
SOM			✓					✓
K-means			✓					✓
Support Vector Machines	✓		✓					
Visualization Techniques	✓					✓		
Bayesian Networks			✓					
K-Nearest Neighbour			✓					
Association Rules							✓	
Fuzzy Rules							✓	
Markov Chains						✓		
Time Series		✓						
Regression				✓				
Simulations	✓							

Εικόνα 2- Διαδικασίες εξόρυξης γνώσης που χρησιμοποιούνται από τις φορολογικές αρχές (P.Castellon et al- 2012)

2. Πρακτικές εφαρμογές

Όπως αναφέρεται και στη μελέτη των P. Castellon et al (2012), αρκετές φορολογικές αρχές έχουν ενσωματώσει διαδικασίες εξόρυξης στις φορολογικές τους αρχές.

Σύμφωνα με την ετήσια έκθεση του, το Φορολογικό Γραφείο της Αυστραλίας (Australian Taxation Office, 2008) έχει ενσωματώσει τέτοιες διαδικασίες στο καθημερινό έργο του, από το 2007.



Εικόνα 3- Διάγραμμα ροής εργασιών και λειτουργιών (ATO, 2008)

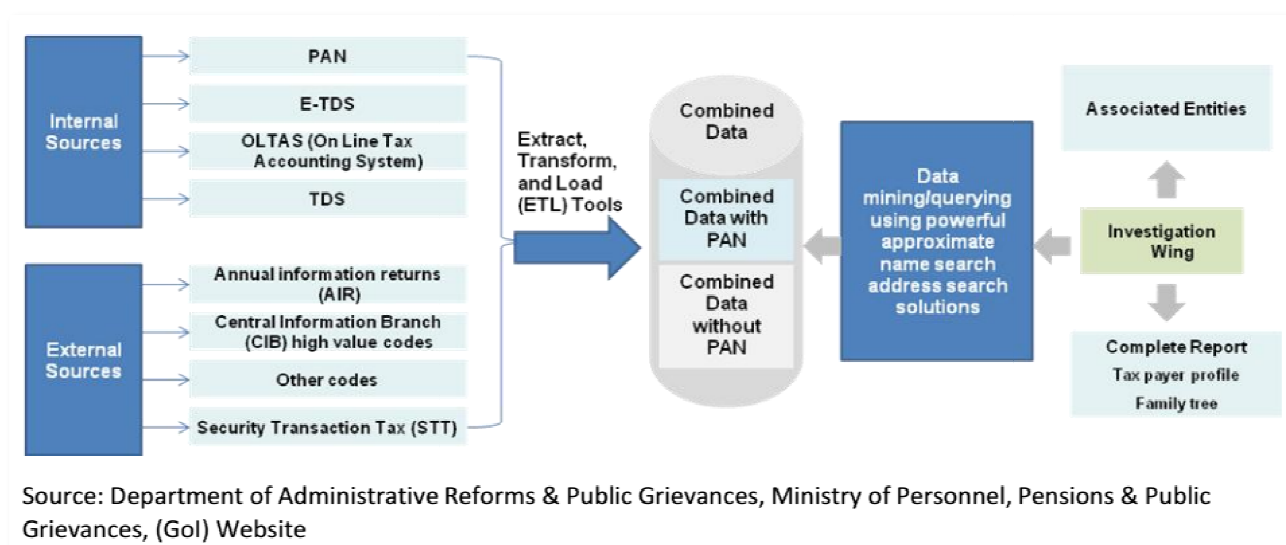
Σύμφωνα με πρόσφατα άρθρα στον τύπο (Terry Hayes, 11/04/2013) το Γραφείο Φορολογίας εφαρμόζει ένα ευρύ πρόγραμμα διασταυρώσεων και εξόρυξης το οποίο εκτείνεται στις πωλήσεις ακινήτων, στις επενδύσεις, στις πωλήσεις αυτοκινήτων, στις πωλήσεις καφέ, στους offshore λογαριασμούς, στα επιδόματα κλπ.

Όπως φαίνεται και στο διάγραμμα της εικόνας 3 το ίδιο το γραφείο φορολογίας έχει δώσει στη δημοσιότητα, μέσω της ετήσιας έκθεσης του, τις διαδικασίες άντλησης πληροφοριών. Σύμφωνα με το διάγραμμα το γραφείο αντλεί στοιχεία από διάφορες πηγές. Τα στοιχεία ελέγχονται ως προς την ποιότητα τους, το κόστος άντλησης τους, τη γραμμογράφηση τους κλπ. Εφόσον αυτό το βήμα ολοκληρωθεί, τα στοιχεία προωθούνται σε ένα προσωρινό αποθετήριο στο οποίο παραμένουν έως ότου ταυτοποιηθούν με τους υπάρχοντες φορολογούμενους. Η διαδικασία έστω και μέχρι αυτό το σημείο είναι σημαντική καθώς επιτρέπει στον οργανισμό να εντοπίζει άγνωστους φορολογούμενους για τους οποίους οι αρχές δεν έχουν προηγούμενη γνώση, ενώ ταυτόχρονα εμπλουτίζει τις πληροφορίες των γνωστών φορολογούμενων.

Με το σύνολο των εμπλουτισμένων στοιχείων η αρχή είναι σε θέση να προσυμπληρώσει φόρμες (πχ. εισόδημα φορολογούμενου), να τα χρησιμοποιήσει για την δημιουργία δομών data warehouse, και να

εκτελέσει διαδικασίες ανάλυσης γνώσης. Τα αποτελέσματα της ανάλυσης γνώσης εμπλουτίζονται και ελέγχονται από αποτελέσματα ανάλυσης που εκτελούνται από το ανθρώπινο δυναμικό της αρχής. Τα αποτελέσματα όλης της διαδικασίας αντιπαραβάλλονται επίσης τα στοιχεία που οι ίδιοι οι φορολογούμενοι έχουν δηλώσει. Στην πράξη το σύνολο των διαδικασιών δημιουργεί ένα κλειστό βρόγχο αποτελεσμάτων και επαληθεύσεων ώστε το ίδιο το σύστημα να προσαρμόζεται στις αλλαγές που δημιουργούνται στη πάροδο του χρόνου.

Σε έκθεση της η ομοσπονδία Ινδικών επιμελητηρίων και Βιομηχανίας (Federation of Indian Chambers of Commerce and Industry, 2015), αναφέρεται στις προσπάθειες που καταβάλει το κράτος για τον εντοπισμό του μαύρου χρήματος. Σύμφωνα με την έκθεση οι Ι.Τ. εργασίες σταδιακά μετακινούνται από κλασικές μεθόδους έρευνας και μέτρησης σε μη παρεμβατικές μεθόδους της εξόρυξης δεδομένων από μια πληθώρα ηλεκτρονικών πηγών. Ειδικά για τον εντοπισμό του μαύρου χρήματος έχει αναπτύξει το εργαλείο ITDMS (Integrated Taxpayer Data Management System), το οποίο χρησιμοποιεί ώστε να δημιουργήσει ένα προφίλ των υψηλότερα αξιολογημένων υποθέσεων, συγκεντρώνοντας στοιχεία από εσωτερικές και εξωτερικές πηγές. Μέρος της λειτουργικότητας του εργαλείου είναι η αυτόματη εύρεση των σχέσεων μιας φορολογικής οντότητας με άλλες, χρησιμοποιώντας στοιχεία όπως πχ την ιδιότητα της οντότητας (πρόεδρος, μέλος διοικητικού συμβουλίου κλπ.), το όνομα, την ημερομηνία γέννησης κλπ.



Εικόνα 4- Χρήση τεχνικών εξόρυξης στην Ινδία (FICCI, 2015)

Η Ινδικές αρχές ακολουθούν μια παρόμοια διαδικασία με αυτή των Αυστραλιανών αρχών. Δεδομένα από διάφορες πηγές, εσωτερικές ή εξωτερικές διαμορφώνονται συγκεντρώνονται σε μια κεντρική αποθήκη δεδομένων. Ιδιαιτερότητα της συγκεκριμένης περίπτωσης είναι ότι τα στοιχεία συμπληρώνονται και με τα αποτελέσματα διαδικασιών ανάλυσης/εξόρυξης για να εμπλουτισθούν με τις σχέσεις που έχουν οι

φορολογούμενοι μεταξύ τους πχ συγγενείς, μέλη διοικητικού συμβουλίου κλπ. Ουσιαστικά δημιουργείται ένα κοινωνικό δίκτυο των φορολογούμενων, το οποίο μπορεί να χρησιμοποιηθεί σε μελλοντικές έρευνες.

Σύμφωνα με άρθρο στην ηλεκτρονική έκδοση του περιοδικού Forbes (Tariq Malik, 8/10/2015) στο Πακιστάν μόλις 925.000 από τα 180 εκατομμύρια ανθρώπους καταβάλουν άμεσους φόρους. Η ενσωμάτωση δεδομένων από διάφορες κρατικές βάσεις δεδομένων, και η βοήθεια της εθνικής αρχής εγγραφών και βάσεων δεδομένων (Nadra) που εφάρμοσε αρχές εξόρυξης σε μεγάλα δεδομένα (big data), βοήθησε στον εντοπισμό 3,5 εκ. φοροφυγάδων. Εκτιμάται ότι αν εφαρμοστεί ένας ελάχιστος βασικός φορολογικός συντελεστής, το Πακιστάν θα έχει όφελος 3,5 δισεκατομμύρια δολάρια, ποσό που ξεπερνά την ανάγκη του Πακιστάν για τα δάνεια από το Διεθνές Νομισματικό Ταμείο σε ετήσια βάση.

ΚΕΦΑΛΑΙΟ 3 - Εξόρυξη δεδομένων και φορολογική διοίκηση

1. Διοίκηση και φορολογικοί έλεγχοι

Μια από τις σημαντικότερες λειτουργίες της φορολογικής διοίκησης είναι ο έλεγχος των φορολογούμενων. Ο έλεγχος είναι σημαντικός γιατί μέσω αυτού η φορολογική διοίκηση έχει να πετύχει πολλαπλούς στόχους:

- (1) να μπορέσει η φορολογική διοίκηση να εντοπίσει φορολογούμενους οι οποίοι δεν υποβάλουν ειλικρινά στοιχεία
- (2) να μπορέσει να εισπράξει, ποσά που δεν έχουν εισπραχθεί
- (3) να επαναφέρει την ισορροπία στην αγορά μεταξύ των τίμιων επιχειρήσεων και των ανειλικρινών επιχειρήσεων, ώστε οι δεύτερες να μην αποκτούν οικονομικό πλεονέκτημα απέναντι στις πρώτες.
- (4) να επικοινωνεί τις συνθήκες της “αγοράς” προς τη διοίκηση, ώστε να λαμβάνονται τα κατάλληλα μέτρα.
- (5) να διαχύσει στην κοινωνία την αίσθηση της ισονομίας και της ισότητας
- (6) να αφουγκρασθεί τις αλλαγές που συμβαίνουν στην αγορά και να προσαρμόσει κατάλληλα τις μελλοντικές φορολογικές πολιτικές

Το πρόβλημα με τον έλεγχο των φορολογούμενων αφορά τα διαθέσιμα μέσα που έχει στην διάθεση της η φορολογική αρχή. Ιδανικά θα έπρεπε να ελέγχεται το σύνολο των φορολογούμενων, ωστόσο ο περιορισμός των διαθέσιμων μέσων, κάνει το εγχείρημα αδύνατο.

Σήμερα, οι φορολογικές αρχές προσπαθούν να αντιμετωπίσουν την πρόκληση της καταπολέμησης του φαινομένου της φοροδιαφυγής, με τον εντοπισμό των επιχειρήσεων που φοροδιαφεύγουν. Οι προσπάθειες επικεντρώνονται στη συλλογή πληροφοριών για τις επιχειρήσεις, με σκοπό την κατανόηση της καλής ή όχι λειτουργίας τους. Καθώς όμως οι φορολογικές αρχές είναι εξοπλισμένες με περιορισμένους πόρους και οι διαδικασίες ελέγχου είναι χρονοβόρες και κοπιώδεις, το αποτέλεσμα είναι η απώλεια σημαντικού ποσού των φορολογικών εσόδων (Wu et al, 2012).

Πρόσθετες δυσκολίες και καθυστερήσεις προκαλούν και οι διαφοροποιήσεις που προέρχονται από το τομέα στον οποίο δραστηριοποιούνται οι επιχειρήσεις (τραπεζικός, καυσίμων, λιανικό εμπόριο κλπ). Κάθε τομέας απαιτεί διαφορετική προσέγγιση και γνώσεις από τον ελεγκτικό μηχανισμό (Gonzalez, D.Velasquez,2012).

Οι διαδικασίες επιλογής των υποθέσεων ποικίλουν, συνήθως ανάλογα με το τύπο του φόρου (ΦΠΑ, εισόδημα κλπ.) ή με τον τύπο των επιχειρήσεων (εμπορικές, παροχής υπηρεσιών κλπ.). Άλλα κριτήρια που λαμβάνονται υπόψη στην επιλογή των υποθέσεων μπορεί να είναι η τυχαία επιλογή ενός δείγματος, η επιλογή βάση του παρελθόντος μιας επιχείρησης κλπ. Συχνά το δείγμα που προκύπτει από την πρώτη διαλογή είναι μεγάλο και υπάρχει επιπλέον επιλογή με πρόσθετα κριτήρια, όπως ο τύπος της δραστηριότητας τους, η περιοχή στην

οποία δραστηριοποιούνται, ο κύκλος πωλήσεων κλπ.

Στην πλειοψηφία των περιπτώσεων οι κανόνες επιλογής έχουν διαμορφωθεί από την εμπειρία των υπαλλήλων, η οποία μπορεί να βασίζεται σε τυχαίες παρατηρήσεις ή και παγιωμένες αντιλήψεις. Η αξιοπιστία και η χρησιμότητα τέτοιων συστημάτων επιλογής μπορεί να τεθεί σε αμφισβήτηση από μια άποψη, καθώς ο άνθρωπος έχει την τάση να ανακαλεί προϋπάρχουσες εμπειρίες τις οποίες χρησιμοποιεί για να επιλύσει νέα προβλήματα, ή περιορίζεται από βιολογικά όρια, όταν ψάχνει να ανακαλύψει νέες συσχετίσεις μεταξύ πολλών μεταβλητών σε πολλές υποθέσεις (Jani Martikainen , 2012).

Δυσκολίες, στην επιλογή των φορολογούμενων, επίσης προκύπτουν από τον τρόπο που διαμορφώνεται η οικονομική δραστηριότητα σε ένα κράτος. Στην Ιταλία η φοροδιαφυγή είναι δύσκολο να εντοπισθεί καθώς το οικονομικό σύστημα χαρακτηρίζεται από μεγάλο αριθμό μικρών και μεσαίων επιχειρήσεων.

Ο μεγάλος αριθμός των μικρομεσαίων επιχειρήσεων στην Ιταλία καθιστούν το φαινόμενο της φοροδιαφυγής, μαζικό φαινόμενο. Ο συνδυασμός της μεγάλης έκτασης του φαινομένου, με το μικρό μέγεθος των επιχειρήσεων που δραστηριοποιούνται στην αγορά, καθιστούν ιδιαίτερα δαπανηρή την καταπολέμηση του (Pisani and Sisti, 2007).

Εκτός από το μέγεθος της φοροδιαφυγής σημαντικό ρόλο στην επιτυχή και ταχεία ολοκλήρωση των ελέγχων, έχουν και οι διαδικασίες ελέγχου που χρησιμοποιούνται. Οι διαδικασίες αυτές καθορίζονται από ένα αυστηρό νομικό πλαίσιο το οποίο πρέπει να εφαρμοσθεί για την ολοκλήρωση ενός ελέγχου. Το νομοθετικό πλαίσιο, χαρακτηρίζεται ως αυστηρό γιατί η ερμηνεία των νόμων, διατάξεων κλπ. δεν επιδέχεται «ευρείας» ερμηνείας, δηλαδή δεν μπορεί να ερμηνευθεί ως προς την πρόθεση του νομοθέτη, , αλλά τυγχάνει εφαρμογής αυστηρά και μόνο του κειμένου. Το νομικό αυτό πλαίσιο, καθώς και ο τρόπος εφαρμογής του, συνήθως αποτελεί τροχοπέδη στην ταχεία ολοκλήρωση των ελέγχων (Pisani and Sisti, 2007) (Jani Martikainen, 2012).

Άλλα πιθανά προβλήματα που επίσης πηγάζουν από το νομικό πλαίσιο είναι κατά περίπτωση η πολυνομοθεσία και η διάχυση αρμοδιοτήτων.

2. Η εξόρυξη δεδομένων στη φορολογική διοίκηση

Οι δράστες του οικονομικού εγκλήματος χρησιμοποιούν πολλές και διαφορετικές τεχνικές, καθώς και διαφορετικά συστήματα για να ολοκληρώσουν τους στόχους τους. Ωστόσο, μια μεγάλη κατηγορία των οικονομικών εγκλημάτων δεν μπορούν να ξεφύγουν από ορισμένες απλές αρχές. π.χ. στην ενδοεπιχειρησιακή απάτη το μοτίβο που υπερισχύει αφορά την ανάγκη τα παράνομα έσοδα να πρέπει να επιστρέψουν ή να είναι υπό τον έλεγχο των διαχειριστών, οι οποίοι τα χρησιμοποιούν για την επίτευξη προσωπικών τους σκοπών,

χαρακτηρίζεται δηλαδή από την ανάγκη επίτευξης προσωπικού κέρδους. Η ανάγκη αυτή δημιουργεί ένα μοτίβο που χαρακτηρίζεται από μια κυκλική ροή του χρήματος.

Αυτή η κυκλική ροή της συναλλαγής παρουσιάζει χαρακτηριστικά όμοια με ένα άλλα είδη απάτης, πχ όπως το σχήμα καρουσέλ στο ΦΠΑ ή με ένα Πολωνικό σύστημα απάτης στα καύσιμα (Jedrzejek et al, 2009).

Οι Wu et al (2012), χρησιμοποιώντας τεχνικές εξόρυξης δεδομένων, έκαναν δυνατή τη δημιουργία ενός πλαισίου ελέγχου με τη χρήση του οποίου είναι δυνατός ο διαχωρισμός των δηλώσεων ΦΠΑ οι οποίες μπορεί να αποτελέσουν αντικείμενο περαιτέρω έρευνας. Τα αποτελέσματα της εφαρμογής έδειξαν ότι η προτεινόμενη τεχνική εξόρυξης δεδομένων ενίσχυσε πραγματικά τη δυνατότητα ανίχνευσης της φοροδιαφυγής, και ως εκ τούτου μπορεί να χρησιμοποιηθεί για την αποτελεσματική μείωση ή ελαχιστοποίηση των απωλειών.

Οι χρήσεις των τεχνικών εξόρυξης μπορούν να χωρισθούν σε δύο βασικές κατηγορίες, ανάλογα με τα επιθυμητά αποτελέσματα :

1. Αυτοματοποίηση ή απλούστευση διαδικασιών.

Η φορολογική διοίκηση μπορεί να δημιουργήσει διαδικασίες επεξεργασίας των πληροφοριών που λαμβάνει πχ κατά την διαδικασία της εγγραφής ή της μεταβολής στοιχείων, της υποβολής δηλώσεων, των πληρωμών κλπ., και να τις ιεραρχήσει σύμφωνα με τον κίνδυνο που αυτές κρύβουν, ώστε να τύχουν της καταλληλότερης αντιμετώπισης.

Η επικινδυνότητα των υποθέσεων μπορεί να υπολογισθεί με βάση τα μοντέλα που έχουν δημιουργηθεί με τη χρήση τεχνικών εξόρυξης. Με αυτό το τρόπο είναι δυνατή η ιεράρχηση των υποθέσεων, βάση της οποίας μπορεί να γίνει μια αυτοματοποίηση των διαδικασιών. Οι διαδικασίες με χαμηλό βαθμό επικινδυνότητας μπορούν να ολοκληρωθούν με τυποποιημένες διαδικασίες ενώ αυτές που εμπεριέχουν υψηλό ρίσκο μπορούν να χαρακτηρισθούν κατάλληλα ώστε να τύχουν περαιτέρω ελέγχου από το ανθρώπινο δυναμικό.

2. Διερεύνηση υποθέσεων.

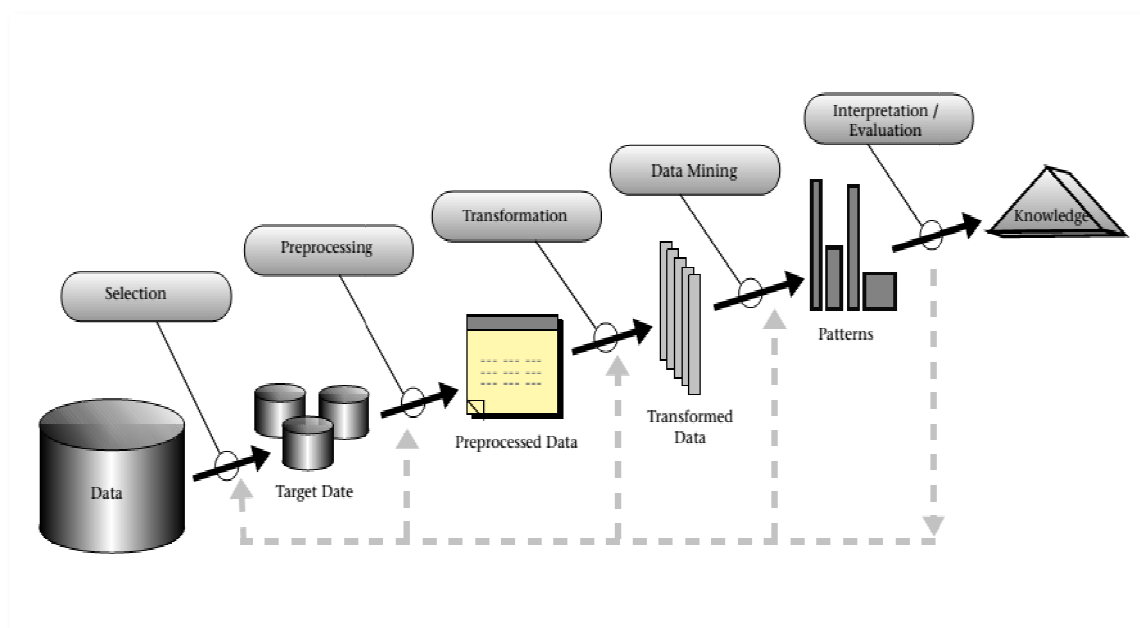
Η φορολογική διοίκηση μπορεί να ομαδοποιήσει τους φορολογούμενους σε ομάδες με συγκεκριμένα χαρακτηριστικά ως προς την συμπεριφορά τους. Αυτή η ομαδοποίηση μπορεί να βοηθήσει τις αρχές να σχεδιάσουν καλύτερα τις ενέργειες τους ώστε να παράσχουν προσαρμοσμένες υπηρεσίες στις ανάγκες τις κάθε ομάδας. Η ταξινόμηση των φορολογούμενων ενδεχόμενα να βοηθούσε και στο σχεδιασμό των κατάλληλων ελεγκτικών ενεργειών (Martikainen, 2012).

3. Εξόρυξη δεδομένων

Ο Fayyad (1996), όπως αναφέρεται από τους Wu et al (2012) διαχωρίζει μεταξύ της διαδικασίας ανακάλυψης γνώσης (knowledge discovery) και των τεχνικών εξόρυξης γνώσης (data mining).

Η ανακάλυψη γνώσης μέσα από τα δεδομένα είναι στην πραγματικότητα μια σειρά διεργασιών, οι οποίες τελικά θα οδηγήσουν στην γνώση. Οι διεργασίες αυτές περιλαμβάνουν ενέργειες όπως η προεπεξεργασία των δεδομένων, ο μετασχηματισμός τους, η εξόρυξη γνώσης και τέλος η εκτίμηση των “αξίας” της γνώσης που ανακαλύφθηκε. Στα διαδοχικά αυτά βήματα γίνεται κατανοητό ότι έννοια «εξόρυξη γνώσης», αποτελεί ένα βήμα από το σύνολο.

Το παρακάτω σχήμα αποδίδει συνοπτικά αυτή τη διαδικασία :



Εικόνα 5- Επισκόπηση των βημάτων που αποτελούν μέρος μιας διαδικασίας εξόρυξης δεδομένων (Fayyad et al, 1996)

Για την παρούσα εργασία δεν θα γίνει αυστηρός διαχωρισμός των όρων «ανακάλυψη γνώσης» και «εξόρυξη γνώσης». Οι όροι θα χρησιμοποιηθούν εναλλάξιμα μεταξύ τους. Θεωρούμε ότι η «εξόρυξη γνώσης» δεν μπορεί να υλοποιηθεί από μόνη της αλλά προϋποθέτει την υλοποίηση όλων των παραπάνω βημάτων, δηλαδή επεξεργασία δεδομένων, καθάρισμα κλπ., αλλιώς η εφαρμογή της δεν μπορεί να παράγει αξιοποιήσιμα αποτελέσματα.

Αν και ο όρος “εξόρυξη δεδομένων” (data mining) είναι σχετικά καινούργιος τόσο η μαθηματική θεωρία, στην οποία βασίζεται όσο και οι εφαρμογές της είναι γνωστές από παλαιότερα. Αυτό που έχει δώσει ιδιαίτερη ώθηση στην εξόρυξη δεδομένων είναι η εξέλιξη των υπολογιστών. Ειδικότερα οι εξελίξεις τις τελευταίες

δεκαετίες, με την ενίσχυση της υπολογιστικής δύναμης των συστημάτων αλλά και την δυνατότητα διαχείρισης μεγάλου όγκου δεδομένων βοήθησε στην εξάπλωση και την πρακτική εφαρμογή της.

Σκοπός της διαδικασίας της εξόρυξης είναι να εξερευνήσουμε και να αναδείξουμε τη γνώση που υπάρχει στα δεδομένα. Η εξόρυξη γνώσης γίνεται ερευνώντας για επαναλαμβανόμενα μοτίβα ή/και σχέσεις μεταξύ των μεταβλητών. Για να είναι επιτυχής η διαδικασία θα πρέπει τα αποτελέσματα της διαδικασίας να επαληθεύονται εφαρμόζοντας τα μοντέλα που παράγονται σε νέες ομάδες δεδομένων.

Τυπικές εφαρμογές της εξόρυξης δεδομένων είναι :

- Ο εντοπισμός ανωμαλιών, όπως οι αποκλίσεις από το μέσο όρο ή ακραία δεδομένα.
- Ο εντοπισμός συσχετίσεων μεταξύ των δεδομένων.
- Η δημιουργία ομάδων από ομοειδή δεδομένα και ο χαρακτηρισμός των ομάδων.
- Η ταξινόμηση νέων δεδομένων σε υπάρχουσες ομάδες.
- Η δημιουργία προβλέψεων με τη μέθοδο της παλινδρόμησης.
- Η οπτικοποίηση και συμπίεση των δεδομένων.
- Η ανακάλυψη κανόνων συσχέτισης μεταξύ των δεδομένων.

Τελικός σκοπός όλων των παραπάνω εφαρμογών εξόρυξης δεδομένων είναι η ανακάλυψη γνώσης με σκοπό την πρόβλεψη, η οποία αποτελεί και την συνηθέστερη εφαρμογή της εξόρυξης δεδομένων.

Εξελικτικά βήματα	Επιχειρηματικές Ερωτήσεις	Τεχνολογίες	Προμηθευτές	Χαρακτηριστικά
Συλλογή δεδομένων (δεκαετία 1960)	Ποια τα συνολικά έσοδα τα τελευταία 5 έτη?	Υπολογιστές, ταινίες, δίσκοι	IBM, IDC	Αναδρομικά και κυρίως στατικά δεδομένα
Εύκολη πρόσβαση στα δεδομένα (δεκαετία 1980)	Ποιες οι πωλήσεις ανά περιοχή τον περασμένο Μάρτιο?	Σχεσιακές βάσεις (RDBMS), SQL, ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Αναδρομικά δυναμικά δεδομένα με δυνατότητα ανάκτησης σε επίπεδο εγγραφής
Data Warehouse και συστήματα υποστήριξης αποφάσεων (δεκαετία 1990)	Ποιες ήταν οι συγκεντρωτικές πωλήσεις τον περασμένο Μάρτιο? Με δυνατότητα ανάλυσης ανά είδος/περιοχή κλπ.	On-Line analytic processing (OLAP), πολυδιάστατες βάσεις, σιλό δεδομένων	Pilot, ComShare, Arbor, Cognos, Microstrategy	Αναδρομικά δυναμικά δεδομένα με δυνατότητα ανάκτησης σε πολλαπλά επίπεδα

Εξελικτικά βήματα	Επιχειρηματικές Ερωτήσεις	Τεχνολογίες	Προμηθευτές	Χαρακτηριστικά
Συλλογή δεδομένων (δεκαετία 1960)	Ποια τα συνολικά έσοδα τα τελευταία 5 έτη?	Υπολογιστές, ταινίες, δίσκοι	IBM, IDC	Αναδρομικά και κυρίως στατικά δεδομένα
Data Mining (σήμερα)	Πως μπορεί να εξελιχθούν οι πωλήσεις τον επόμενο μήνα? Γιατί?	Υπολογιστές με πολλαπλούς επεξεργαστές, big data, αλγόριθμοι	Lockheed, IBM, SGI και πολλές startup	Πρόβλεψη, δυναμική εμφάνιση πληροφορίας

Εικόνα 6- Η εξέλιξη της εξόρυξης γνώσης (προσαρμογή από Thearling,2012)

Όπως φαίνεται και από την παράθεση των χαρακτηριστικών στον προηγούμενο πίνακα σκοπός της εξόρυξης γνώσης είναι η δημιουργία προβλέψεων. Για να συμβεί αυτό εφαρμόζονται τα βήματα του πίνακα (5). Το αποτέλεσμα είναι η εξαγωγή συμπερασμάτων που αφορούν τις σχέσεις μεταξύ των δεδομένων εισόδου. Ωστόσο για να προχωρήσουμε στην πρόβλεψη είναι απαραίτητες ορισμένες επιπλέον φάσεις επεξεργασίας.

Μια διαδικασία εξόρυξης δεδομένων αποτελείται από τρεις φάσεις :

1. Την αρχική διερεύνηση. Κατά την αρχική διερεύνηση ελέγχονται τα στοιχεία και αξιολογούνται.
2. Τη δημιουργία ενός μοντέλου και την επαλήθευση αυτού. Στην φάση αυτή τα αποτελέσματα κωδικοποιούνται ώστε να μπορούν να εφαρμοσθούν σε νέα δεδομένα. Η κωδικοποίηση τους δημιουργεί ένα μοντέλο, το οποίο εφαρμόζεται σε νέα δεδομένα και αξιολογείται για τις επιδόσεις του.
3. Τη χρήση του μοντέλου σε παραγωγική λειτουργία, για την δημιουργία προβλέψεων. Τέλος εφόσον το θεωρητικό μοντέλο επαληθευτεί, εφαρμόζεται σε νέα μη πειραματικά δεδομένα για την δημιουργία προβλέψεων.
4. Επανάληψη διαδικασίας. Τα βήματα 1-3 επαναλαμβάνονται καθώς νέα δεδομένα εμπλουτίζουν το πεδίο γνώσης και κατά συνέπεια το μοντέλο.

4. Δημοφιλείς διαδικασίες εξόρυξης

Οι Fayyad et al (1996) διαχώρισαν την διαδικασία ανακάλυψης γνώσης (Knowledge Discovery) σε διακριτά βήματα (επιλογή των στοιχείων, προ-επεξεργασία κλπ.). Το μοντέλο που πρότειναν βοήθησε στην κωδικοποίηση των ενεργειών, με αποτέλεσμα την καλύτερη οργάνωση μιας διαδικασίας εξόρυξης. Όμως το

προτεινόμενο μοντέλο δεν κάλυπτε πλήρως τις πρακτικές ανάγκες που προκύπτουν κατά την εφαρμογή μιας διαδικασίας ανακάλυψης γνώσης.

Ένα σύνολο εταιρειών δημιούργησαν ένα τροποποιημένο μοντέλο εξόρυξης το οποίο και ονόμασαν CRISP-DM. Το ακρωνύμιο προέρχεται από τα αρχικά των λέξεων (CRoss-Industry Standard Process for Data Mining).

Το μοντέλο δημιουργείται από την κυκλική ροή 6 διακριτών φάσεων. Οι φάσεις είναι:

1. Η κατανόηση του επιχειρηματικού πλαισίου – Η αρχική φάση επικεντρώνεται στην κατανόηση των στόχων και απαιτήσεων του έργου από την πλευρά του επιχειρησιακού τομέα και τη μετατροπή της γνώσης σε σχέδιο δράσης.
2. Η κατανόηση των δεδομένων – Σε αυτή τη φάση γίνεται αρχική συλλογή των στοιχείων και έλεγχος αυτών για να βρεθούν προβλήματα ποιότητας, επιλογής, να βρεθούν συσχετίσεις κλπ.
3. Προετοιμασία δεδομένων – Η συλλογή και προετοιμασία των δεδομένων που θα χρησιμοποιηθούν στην εξόρυξη.
4. Μοντελοποίηση – Εφαρμόζονται διάφορες τεχνικές μοντελοποίησης και δοκιμάζονται οι παράμετροι για την βελτιστοποίηση τους.
5. Αξιολόγηση – Το μοντέλο ή τα μοντέλα αξιολογούνται και ελέγχονται για την καταλληλότητά τους.
6. Παραγωγική εκμετάλλευση – οι ενέργειες μετατροπής της γνώσης που αναδεικνύει το μοντέλο σε χρήσιμη γνώση.



Εικόνα 7- Τα βήματα της διαδικασίας CRISP (Wikipedia,2015)

Η εταιρεία SAS ανέπτυξε ένα παρόμοιο μοντέλο το οποίο ονόμασε SEMMA Το ακρωνύμιο προέρχεται από τα αρχικά των λέξεων Sample, Explore, Modify, Model, Assess. Σύμφωνα με το μοντέλο υπάρχουν 5 στάδια:

1. Δειγματοληψία – Σε αυτό το στάδιο γίνεται δειγματοληψία από το σύνολο των δεδομένων για να είναι ταχείς οι έλεγχοι των μοντέλων. Το βήμα αυτό μπορεί να θεωρηθεί προαιρετικό.
2. Εξερεύνηση – Σε αυτό το στάδιο γίνεται εξερεύνηση των στοιχείων για να εντοπισθούν τάσεις και ανωμαλίες.
3. Μετατροπή – Στο στάδιο αυτό υλοποιούνται μετατροπές των δεδομένων για να βοηθηθεί η δημιουργία μοντέλων.
4. Μοντελοποίηση – Η δημιουργία μοντέλων
5. Εκτίμηση – Σε αυτό το στάδιο γίνεται εκτίμηση των στοιχείων εξετάζοντας την χρησιμότητα των μοντέλων που δημιουργήθηκαν.

Οι Azevedo & Santos (2008), συνέκριναν τα δύο μοντέλα σε σχέση με την διαδικασία που προτάθηκε από τους Fayyad et al (1996). Σύμφωνα με τις παρατηρήσεις τους τα δύο προτεινόμενα μοντέλα μπορούν να θεωρηθούν σαν διαφορετικές υλοποιήσεις της διαδικασίας ανάλυσης που προτάθηκε από τον Fayyad.

Ως συμπέρασμα της μελέτης τους παρουσίασαν τον παρακάτω συγκεντρωτικό πίνακα στον οποίο παρουσιάζονται οι σχέσεις μεταξύ των διαφορετικών διαδικασιών.

KDD	SEMMA	CRISP-DM
Pre KDD	-----	Business understanding
Selection	Sample	Data Understanding
Pre processing	Explore	
Transformation	Modify	Data preparation
Data mining	Model	Modeling
Interpretation/Evaluation	Assessment	Evaluation
Post KDD	-----	Deployment

Εικόνα 8- Αντιστοιχίσεις μεταξύ των διαδικασιών εξόρυξης (Azevedo & Santos, 2008)

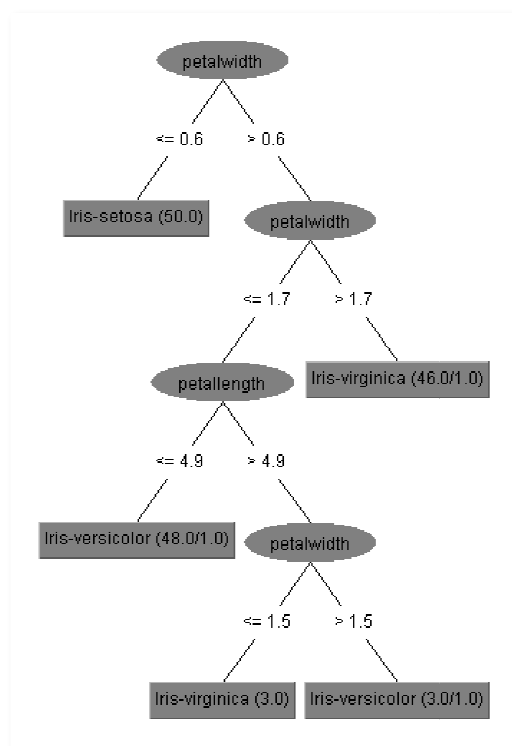
5. Δημοφιλείς αλγόριθμοι εξόρυξης

Αλγόριθμοι εξόρυξης γνώσης υπάρχουν πολλοί, κάποιοι είναι γενικότεροι και κάποιοι πιο εξειδικευμένοι. Στη συνέχεια θα αναφερθούμε σε αυτούς που θα χρησιμοποιήσουμε για την εκτέλεση της εργασίας.

Δέντρα Απόφασης – Decision Trees

Τα δένδρα αποφάσεων (decision trees) είναι μια από τις συνηθέστερα χρησιμοποιούμενες μεθόδους. Η απλότητα της μεθόδου, το γεγονός ότι δεν απαιτεί πρόσθετες παραμέτρους, η ουδετερότητα της ως προς το είδος δεδομένων και το γεγονός ότι τα αποτελέσματα που προκύπτουν είναι άμεσα ερμηνεύσιμα αποτελούν ισχυρά χαρακτηριστικά της μεθόδου.

Υπάρχουν διάφοροι αλγόριθμοι για την κατασκευή των δένδρων, οι οποίοι διαφοροποιούνται από τη στρατηγική που χρησιμοποιούν για την κατασκευή των κλάδων του δένδρου και του βάθους.



Εικόνα 9- Decision tree (WEKA-Iris Data Set)

Οι Gonzalez και Velasquez (2012), στην έρευνα τους χρησιμοποίησαν τον αλγόριθμο CHAID για την κατασκευή των δένδρων. Η επιλογή τους βασίστηκε στο γεγονός ότι η συγκεκριμένη μέθοδος δημιουργεί κλάδους από ένα κόμβο βασιζόμενη τόσο σε συνεχείς όσο και σε κατηγορικές μεταβλητές.

Ο Quinlan (Wikipedia, 1993), πρότεινε τον αλγόριθμο C4.5, εξέλιξη του ID3. Ο αλγόριθμος δημιουργεί ένα δέντρο αποφάσεων βασισμένος στην αλλαγή της εντροπίας της πληροφορίας. Κάθε φορά δημιουργεί ένα κλάδο ελέγχοντας ποιο από τα γνωρίσματα θα αποφέρει το μεγαλύτερο κέρδος και στη συνέχεια επιλέγοντας το ως το γνώρισμα επιλογής,

Ο Ho (1995), προσπάθησε να λύσει το πρόβλημα της υπερ-προσαρμογής των δέντρων στα δεδομένα εκπαίδευσης. Για το σκοπό αυτό πρότεινε την δημιουργία πολλαπλών δέντρων επιλέγοντας τυχαία υποσύνολα από το σύνολο των εκπαιδευτικών δεδομένων και στη συνέχεια τη να συνδυάσει αυτά για την εξαγωγή του συνηθέστερου αποτελέσματος.

Κανόνες – Rules

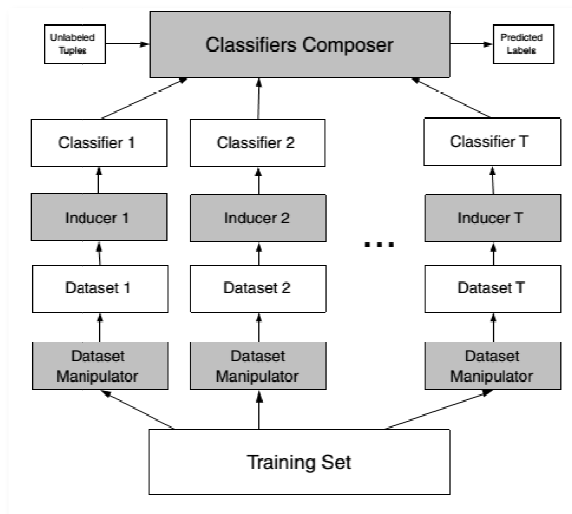
Οι κανόνες (rules) είναι επίσης μια δημοφιλής τεχνική εξόρυξης γνώσης. Σκοπός της είναι από μια σειρά γνωστών δεδομένων να εξαχθεί μια σειρά λογικών κανόνων οι οποίοι με τη σειρά τους οδηγούν σε συμπεράσματα. Όπως και στα δένδρα αποφάσεων υπάρχουν διάφοροι αλγόριθμοι που μπορούν να ανακαλύψουν κανόνες μεταξύ των δεδομένων πχ. Apriori, FP-growth, Opus κλπ.

```
JRIP rules:
=====

(K <= 0) and (Al <= 1.74) and (RI <= 1.51969) => Type=tableware (9.0/1.0)
(Mg <= 2.68) and (Na <= 13.38) and (RI <= 1.52171) => Type=containers (12.0/1.0)
(Si <= 72.77) and (RI <= 1.51776) and (Ca >= 8.32) => Type=vehic wind float (13.0/5.0)
(Ba >= 0.4) => Type=headlamps (28.0/3.0)
(RI >= 1.52247) and (Ca <= 9.76) => Type=headlamps (2.0/0.0)
(Al <= 1.41) and (RI <= 1.51808) and (RI >= 1.5172) => Type=build wind float (37.0/4.0)
(RI >= 1.51869) and (Mg >= 3.33) and (Si <= 72.36) => Type=build wind float (26.0/3.0)
=> Type=build wind non-float (87.0/17.0)
```

Εικόνα 10- Rules (WEKA- weather data set)

Bagging - Boosting



Εικόνα 11- Bagging (L. Rokach, O. Maimon, 2015, ,κεφ.9.4.1)

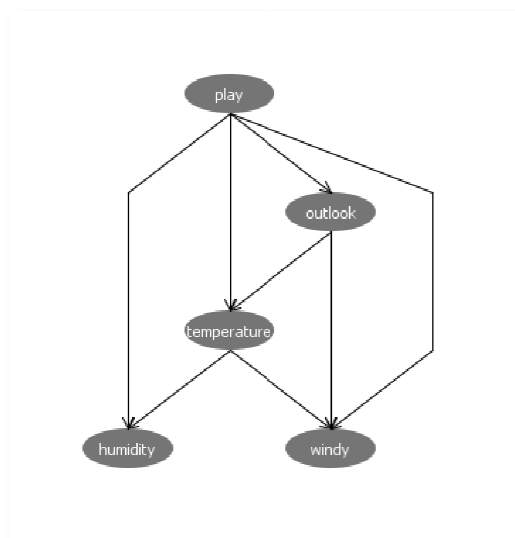
Όπως περιγράφεται από τον I. Witten στο βιβλίο του «Data Mining Practical Machine Learning Tools» :

Ο συνδυασμός αποφάσεων διαφόρων μοντέλων σημαίνει την συγχώνευση των διαφόρων εξόδων σε μια πρόβλεψη. Στις περιπτώσεις της κατηγοριοποίησης αυτό γίνεται κατόπιν ψηφοφορίας, ενώ στην περίπτωση της αριθμητικής πρόβλεψης με τον υπολογισμό του μέσου όρου. Οι διαδικασίες bagging και boosting υλοποιούν τη βασική αυτή αρχή αλλά παράγουν τα μοντέλα με διαφορετικό τρόπο.

Στην περίπτωση του bagging τα χρησιμοποιούμενα μοντέλα έχουν το ίδιο βάρος, ενώ στο boosting χρησιμοποιούμε διαφορετικά βάρη για αναδειχθούν τα περισσότερα πετυχημένα.

Δίκτυα Bayes

Τα δίκτυα Bayes είναι μη κυκλικοί κατευθυνόμενοι γράφοι, που χρησιμοποιούνται για να προβλέψουν τα αποτελέσματα, βασισμένοι σε μια ομάδα γεγονότων. Το δίκτυο αποτελείται από ένα σύνολο κόμβων που αντιπροσωπεύουν τις μεταβλητές του προβλήματος και ένα σύνολο κατευθυνόμενων τόξων που συνδέουν τους κόμβους και δείχνουν τη σχέση εξάρτησης μεταξύ των ιδιοτήτων των παρατηρούμενων δεδομένων. Τα δίκτυα Bayes περιγράφουν την πιθανότητα κατανομής που διέπει ένα σύνολο μεταβλητών, προσδιορίζοντας τις υποθέσεις των εξαρτώμενων μεταβλητών από τις εξαρτώμενες πιθανότητες.



Εικόνα 12- BayesNet (WEKA weather data set)

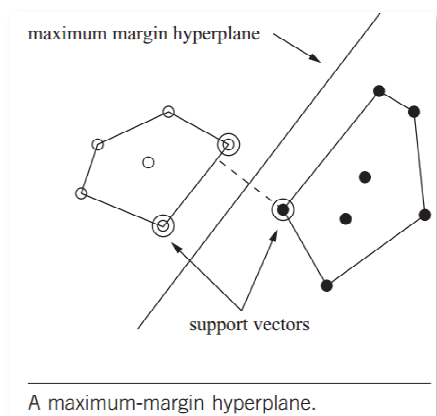
Συνήθως τέτοιου είδους προβλήματα χωρίζονται σε δύο μέρη: τη μάθηση της δομής, η οποία χρησιμοποιείται για να δομηθεί το δίκτυο και τη μάθηση των παραμέτρων με τις οποίες και μέσω του ήδη γνωστού δικτύου λαμβάνουμε τις πιθανότητες για τον κάθε κόμβο.

Το κύριο πλεονέκτημά τους είναι ότι η πιθανότητα εμφάνισης ενός συγκεκριμένου γεγονότος βασίζεται σε μια δέσμη ενεργειών μπορεί να υπολογισθεί, δίνοντας μια σαφή εικόνα των σχέσεων μέσα από ένα γράφημα τύπου ιστού (Gonzalez και Velasquez, 2012).

Self-Organizing Map

Όπως περιγράφουν και οι Gonzalez and Velasquez (2012) μια επίσης δημοφιλή μέθοδος εξόρυξης δεδομένων αποτελούν και τα αυτο-οργανώμενα δίκτυα (Self-Organizing Map). Η τεχνική αυτή έχει τις ρίζες της στα τεχνητά νευρωνικά δίκτυα. Η τεχνική έχει εφαρμογή όταν υπάρχουν πολυδιάστατα δεδομένα και οι σχέσεις που τα συνδέουν δεν είναι γνωστές. Ουσιαστικά δημιουργείται ένα δίκτυο από κόμβους οι οποίοι οδηγούν στα αποτελέσματα. Σκοπός του δικτύου είναι να βρεθούν οι κόμβοι αυτοί, οι οποίοι παράγουν παρόμοια αποτελέσματα και στη συνέχεια οι ομάδες των κόμβων να ομαδοποιηθούν ώστε να προκύψει μια ομάδα κόμβων που θα παράγει τα επιθυμητά αποτελέσματα με την μεγαλύτερη ακρίβεια.

Support Vector Machine (SVM)



Εικόνα 13- SVM (Witten fig. 6.9, 2011)

Αν δοθεί ένα σετ εκπαιδευτικών δεδομένων, τα οποία έχουν χαρακτηριστεί ότι ανήκουν σε μια από δύο κατηγορίες, ο αλγόριθμος SVM μπορεί να δημιουργήσει ένα μοντέλο το οποίο θα κατατάσσει ένα νέο δείγμα σε μια από τις δύο κατηγορίες. Σε ένα μοντέλο SVM, απεικονίζονται τα δείγματα ως σημεία στο χώρο με τέτοιο τρόπο ώστε να υπάρχει ένα ευδιάκριτο κενό μεταξύ των δύο κατηγοριών. Τα νέα δείγματα τοποθετούνται στο χώρο και ανάλογα της περιοχής που τοποθετήθηκαν μπορεί να γίνει και η πρόβλεψη για την κατηγορία που ανήκουν.

Μετρικές απόδοσης μοντέλων ταξινόμησης

Κάθε μια από τις παραπάνω μεθόδους παρουσιάζει πλεονεκτήματα τα οποία εξαρτώνται από το είδος των δεδομένων, το πλήθος, το επιθυμητό αποτέλεσμα και το πεδίο εφαρμογής της.

Ειδικά για τις τεχνικές “πρόβλεψης”, όπως πχ τα δένδρα αποφάσεων, σημαντικές μετρικές που περιγράφουν αν το μοντέλο αποδίδει καλά είναι, το ποσοστό επιτυχίας των προβλέψεων, η επαναληψιμότητα των επιτυχιών, η αποτελεσματικότητα του μοντέλου κλπ.

Η μέτρηση της αποτελεσματικότητας γίνεται συνήθως με την υλοποίηση ενός “Confusion matrix”(Pisani and Sisti, 2007). Για προβλήματα με 2 αποφάσεις (π.χ. ΝΑΙ/ΟΧΙ) είναι ένας 2x2 πίνακας που περιέχει τα αποτελέσματα των επιτυχημένων ή λανθασμένων προβλέψεων σε σχέση με τις πραγματικές τιμές.

TEST SET: The Confusion Matrix				
		<i>Predicted</i>		
		<i>Non Evader</i>	<i>Evader</i>	
<i>Actual Value</i>	<i>Non Evader</i>	8.377 (TN)	1.362 (FP)	9.739
	<i>Evader</i>	3.778 (FN)	2.502 (TP)	6.280
		12.155	3.864	16.019

Εικόνα 14 – Confusion matrix (Stefano Pisani, Paola De Sisti, 2007).

Από τον πίνακα προκύπτουν οι διάφορες μετρικές του μοντέλου. Μερικές από αυτές είναι, η ευαισθησία, η ακρίβεια, την ορθότητα κλπ.

Οι μετρικές που προκύπτουν από τον πίνακα σύγχυσης (confusion matrix) δίνουν μια στατική εικόνα του μοντέλου. Συχνά είναι απαραίτητο να συγκρίνουμε ένα μοντέλο με ένα άλλο ή να ελέγξουμε την απόδοση του μοντέλου με διαφορετικό αριθμό δειγμάτων. Σε αυτές τις περιπτώσεις μπορούμε να χρησιμοποιήσουμε τις καμπύλες απόδοσης.

Καμπύλη ROC (Receiver Operating Curve)

Η καμπύλη δημιουργείται από το ποσοστό των αληθώς θετικών (True Positive Rate) προς το ποσοστό των ψευδώς θετικών (False Positive Rate), όταν αυτά τα μετρήσουμε σε διάφορα σημεία

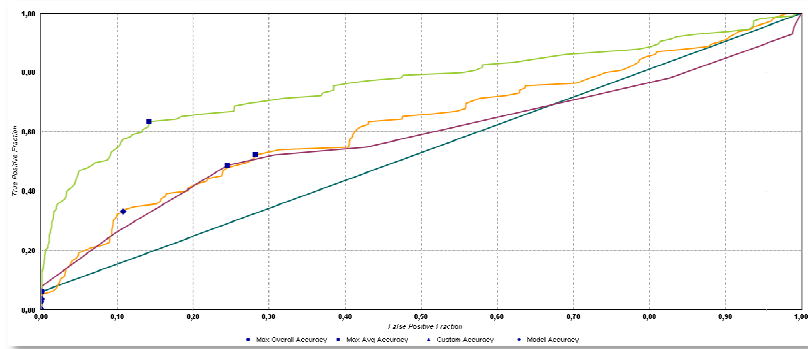
Το ποσοστό αληθώς θετικών είναι γνωστό και ως ευαισθησία, ενώ το ποσοστό των ψευδώς θετικών είναι γνωστό και ως fall-out και υπολογίζεται ως (1 - ακρίβεια).

$$TPR = \frac{TP}{TP + FN}$$

$$\text{Ακρίβεια/specificity} = \frac{TN}{FP + TN}$$

$$FPR = \frac{PF}{TN + FP}$$

Εικόνα 15- Μετρικές που χρησιμοποιούνται για το σχηματισμό της καμπύλης ROC



Εικόνα 16- Καμπύλη ROC

ΚΕΦΑΛΑΙΟ 4 – Το πεδίο μελέτης

1. Το επιχειρησιακό πλαίσιο.

Στην Ε.Ε. μπορούμε να διακρίνουμε τρεις γενικές κατηγορίες εμπορίου, ανάλογα με τον τρόπο του αγοραστή και του πωλητή :

- Τις εισαγωγές/εξαγωγές. Αναφερόμαστε σε λήψεις ή αποστολές αγαθών σε χώρες εκτός Ε.Ε..
- Το εθνικό εμπόριο. Αναφερόμαστε σε όλες τις πράξεις που λαμβάνουν μέρος εντός των συνόρων μιας χώρας.
- Το ενδοκοινοτικό εμπόριο. Αναφερόμαστε σε λήψεις ή αποστολές αγαθών εντός της Ε.Ε. αλλά από μια χώρα της Ε.Ε. σε μια άλλη.

Η κατηγοριοποίηση γίνεται κυρίως για λόγους διοικητικής διευκόλυνσης και εφαρμογής διαδικασιών. Ουσιαστικά ο διαχωρισμός γίνεται με βάση την ανάγκη να διαχωρισθούν οι περιπτώσεις στις οποίες κάποιο αγαθό διασχίζει τα εθνικά σύνορα μιας χώρας. Στις περιπτώσεις στις οποίες τα αγαθά διασχίζουν τα εθνικά σύνορα προερχόμενο από κάποια μη κοινοτική χώρα, οι τελωνειακές υπηρεσίες αναλαμβάνουν να εφαρμόσουν τις σχετικές διατάξεις που περιγράφουν το τρόπο με τον οποίο θα γίνει ο χειρισμός της εισαγωγής/εξαγωγής.

Από τη στιγμή που ένα αγαθό βρίσκεται εντός του εθνικού χώρου τότε υπεύθυνες για θέματα φορολογίας είναι οι εθνικές φορολογικές υπηρεσίες της χώρας. Στην Ε.Ε. αυτό συνήθως είναι το τμήμα ΦΠΑ της φορολογικής διοίκησης, κύρια εργασία του οποίου είναι η είσπραξη του φόρου προστιθέμενης αξίας.

Το ενδοκοινοτικό εμπόριο αποτελεί μια ιδιαίτερη περίπτωση εμπορίου μεταξύ των κρατών μελών της Ε.Ε. Σύμφωνα με τη Ενιαία Ευρωπαϊκή Πράξη του 1986, αλλά και με μια σειρά κειμένων (κανονισμός EC/765/2008, απόφαση 768/2008, Κανονισμός EC/764/2008), η Ε.Ε. προσπαθεί θεμελιώσει την ελεύθερη διακίνηση αγαθών, προσώπων και υπηρεσιών εντός της Ε.Ε. Στα πλαίσια της ελεύθερης διακίνησης των αγαθών εντός της Ε.Ε. τα αγαθά που διακινούνται μεταξύ των κρατών μελών δεν αντιμετωπίζονται ως αγαθά τα οποία προέρχονται από ξένη χώρα, αλλά ως αγαθά τα οποία διακινούνται εντός του κοινοτικού χώρου. Πρακτικά αυτό σημαίνει ότι δεν εφαρμόζονται οι τελωνειακές διαδικασίες. Η ενοποίηση των διαδικασιών στο ενδοκοινοτικό εμπόριο θα ήταν απλή αν οι διαφορετικές χώρες εφάρμοζαν ενιαίες διαδικασίες και κανόνες εφαρμογής του φόρου. Σήμερα τα φορολογικά συστήματα των κρατών μελών δεν είναι ενοποιημένα. Κάθε χώρα έχει το δικαίωμα να εφαρμόζει τις καταλληλότερες για την εθνική της οικονομία διαδικασίες και νόμους ώστε να επιτυγχάνει την ρύθμιση της οικονομικής δραστηριότητας στις τοπικές ιδιαιτερότητες.

2. Πώς λειτουργεί ο ΦΠΑ

Βασική πηγή εσόδων για τα Ευρωπαϊκά κράτη αποτελεί ο φόρος προστιθέμενης αξίας. Ο φόρος πιο γνωστός και ως ΦΠΑ, καταβάλλεται σε κάθε στάδιο της μεταπώλησης ενός αγαθού και επιβάλλεται μόνο στην προστιθέμενη αξία που προκύπτει μεταξύ της αγοράς και της πώλησης. Όταν το αγαθό πωλείται στον τελικό καταναλωτή τότε αυτός καλείται να καταβάλει το σύνολο του φόρου.

Για να γίνει κατανοητός ο τρόπος λειτουργίας και απόδοσης του ΦΠΑ παρατίθεται το παράρτημα (1) . Η υπόθεση στην οποία έχει βασισθεί το παράδειγμα είναι ότι πριν την τελική πώληση του αγαθού έχουν προηγηθεί τρία στάδια κατασκευής/μεταπώλησης (εικόνα 85).

Πρακτικά το ΦΠΑ τον πληρώνει ο τελικός καταναλωτής ενώ για τις επιχειρήσεις είναι ένας δημοσιονομικά ουδέτερος φόρος. Η απόδοση του γίνεται τμηματικά ώστε να περιορισθούν φαινόμενα απόκρυψης του στο τελικό στάδιο της λιανικής πώλησης.

Ακολουθώντας το παραπάνω παράδειγμα είναι εύκολο να καταλάβει κάποιος ότι αν στις συναλλαγές μεταξύ των επιχειρήσεων υπάρξει μια πώληση αγαθού από μια χώρα της Ε.Ε. σε μια άλλη θα προκύψουν ζητήματα, ως προς την απόδοση του ΦΠΑ. Τα θέματα που προκύπτουν αφορούν περιπτώσεις που μεταξύ των χωρών υπάρχουν διαφορετικοί συντελεστές ΦΠΑ, διαφορετικές περιόδους υπολογισμού του φόρου κλπ.

Διάφορα συστήματα και διαδικασίες μπορούν να προβλεφθούν, τα οποία θα διευθετούσαν τέτοιου είδους ζητήματα. Ωστόσο, η ύπαρξη πρόσθετων διαδικασιών θα επανάφερε στο προσκήνιο εμπόδια στην ελεύθερη διακίνηση των αγαθών, αλλά και πρόσθετο κόστος καθώς οι επιχειρήσεις θα καλούνταν να προσαρμοσθούν στις ξεχωριστές ιδιαιτερότητες κάθε χώρας.

Στα παραπάνω προβλήματα μερική λύση ήρθε να δώσει το σύστημα VIES. Στο σύστημα αυτό εγγράφονται οι επιχειρήσεις οι οποίες επιθυμούν να πουλήσουν ή να αγοράσουν αγαθά από μια άλλη ευρωπαϊκή χώρα. Σε περίπτωση που και οι δύο συναλλασσόμενες επιχειρήσεις είναι εγγεγραμμένες στο σύστημα τότε η αλυσίδα του ΦΠΑ καταργείται για τις μεταξύ τους αγοροπωλησίες και μπορούν να εμπορευθούν αγαθά χωρίς την χρέωση ΦΠΑ (πίνακας 12).

Όπως φαίνεται από το παράδειγμα για τη χώρα της τελικής κατανάλωσης δεν αλλάζει η διαδικασία υπολογισμού ή απόδοσης του ΦΠΑ. Για τη χώρα πώλησης τα αγαθά αντιμετωπίζονται όμοια με τις περιπτώσεις εξαγωγών δηλαδή χωρίς να επιβάλλονται φόροι κατά τη μεταφορά στην επόμενη χώρα.

3. Το πρόβλημα που προκύπτει

Το γεγονός ότι υπάρχει μια ασυνέχεια στην αλυσίδα του ΦΠΑ δημιουργεί κενό το οποίο μπορούν να το εκμεταλλευτούν κυκλώματα φοροδιαφυγής.

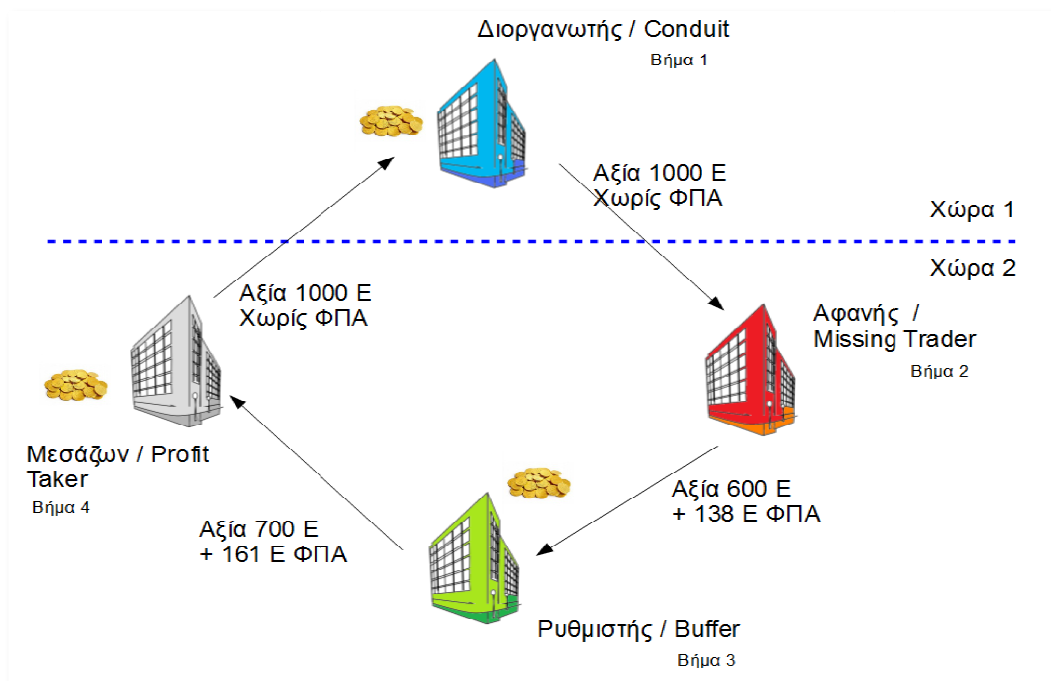
Τα κυκλώματα που αναπτύσσονται εκμεταλλεύονται δύο χαρακτηριστικά του ενδοκοινοτικού εμπορίου :

- Απαιτείται να υπάρξει συνεργασία – πληροφόρηση μεταξύ δύο ή περισσότερων χωρών για να ελεγχθεί μια ενδοκοινοτική συναλλαγή.
- Η συνεργασία αυτή, των αρχών είναι συνήθως αργή.

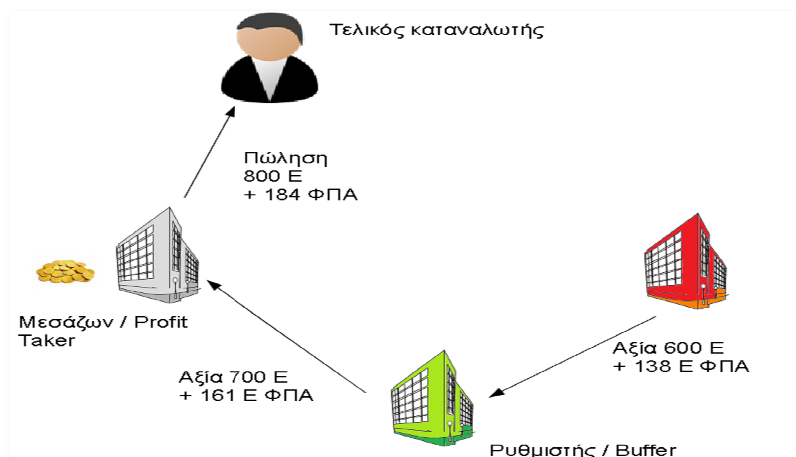
Ο τρόπος δράσης αυτού του είδους των κυκλωμάτων παρουσιάζει διάφορες παραλλαγές, αλλά ο πυρήνας λειτουργίας παραμένει ο ίδιος σε όλες τις περιπτώσεις.

Ουσιαστικά το κύκλωμα δρα εκμεταλλευόμενο την μεγάλη ταχύτητα που αναπτύσσεται, μεταξύ των μελών/επιχειρήσεων που συμμετέχουν σε αυτό, στην ανταλλαγή των αγαθών και την αδυναμία των αρχών να συνεργασθούν αποτελεσματικά.

Στο παρακάτω σχήμα εμφανίζεται η βασική λειτουργία του κυκλώματος :



Εικόνα 17- Λειτουργία κυκλώματος Carousel(1)



Εικόνα 18 - Λειτουργία κυκλώματος Carousel (2)

Απλοποιημένα η λειτουργία του κυκλώματος είναι η εξής :

1. Η αρχική επιχείρηση πουλάει σε μια άλλη χώρα ένα αγαθό, χωρίς να χρεώσει ΦΠΑ.
2. Η επιχείρηση το παραλαμβάνει αλλά αλλοιώνει τη τιμή απόκτησης/αγοράς.
3. Για να διατηρηθεί το κύκλωμα χρησιμοποιεί μια ή παραπάνω επιχειρήσεις, η οποίες αναλαμβάνουν να αναμείξουν την νόμιμη εργασία τους με τα αγαθά που παρέλαβαν στο βήμα 2. Ουσιαστικά “ξεπλένουν” την απάτη.
4. Η τελική εταιρεία πουλάει τα αγαθά προσποριζόμενη την διαφορά που έχει προκύψει στην τελική τιμή ή πουλώντας σε χαμηλότερες τιμές.

Στο παραπάνω κύκλο υπάρχουν δύο κύρια σημεία στα οποία δημιουργούνται ζημιές από την μη καταβολή του ΦΠΑ.

Στο βήμα (1) η χώρα στην οποία δραστηριοποιείται η επιχείρηση χάνει το ΦΠΑ στις περιπτώσεις που η πώληση δεν είναι πραγματική, αλλά διακινούνται μόνο εικονικά έγγραφα. Συνήθως τα αγαθά παραμένουν εντός των συνόρων και την κατάλληλη στιγμή διοχετεύονται στην αγορά, δημιουργώντας στρεβλώσεις. Στο βήμα (4) η χώρα στην οποία κατέληξαν τα αγαθά χάνει από τη διαφορά τις πραγματικής τιμής και της τιμής πώλησης, ενώ παράλληλα αντιμετωπίζει παρόμοιες στρεβλώσεις στην αγορά όπως και η χώρα στο βήμα (1).

Οι ρόλοι που διακρίνονται σε ένα τέτοιο κύκλωμα είναι οι εξής :

- Αφανής επιτηδευματίας ή Εξαφανισμένος έμπορος

Ο εξαφανισμένος έμπορος είναι ο σημαντικότερος ρόλος μέσα στο κύκλωμα. Είναι η επιχείρηση που θα εξαφανισθεί όταν απαιτηθεί για να καλυφθούν τα ίχνη του κυκλώματος.

- Ρυθμιστής ή απομονωτής

Οι απομονωτές χρησιμοποιούνται για να δώσουν μια όψη νομιμοφάνειας στο κύκλωμα. Χρησιμοποιούνται για να περιπλέξουν την ανίχνευση της απάτης.

- Μεσάζων

Ο μεσάζων, είναι μια ειδική μορφή των απομονωτών. Είναι ο τελευταίος κρίκος της αλυσίδας και αυτός που ωφελείται από τη λειτουργία του κυκλώματος

- Διοργανωτής ή αρχική επιχείρηση

Ο διοργανωτής είναι το κεντρικό πρόσωπο μέσα στο κύκλωμα της απάτης ΦΠΑ του τύπου Carousel.

- Κράτος

Τα κράτη είναι ένα από τα θύματα της λειτουργίας των κυκλωμάτων καθώς χάνουν σημαντικό μέρος εσόδων του ΦΠΑ.

- Φορολογούμενοι

Ο φορολογούμενος είναι έμμεσο θύμα. Ενώ οι φόροι που πληρώνει θα έπρεπε να καταλήγουν στο κράτος καταλήγουν ολικά ή εν μέρη στο κύκλωμα. Επίσης επειδή τα έσοδα του κράτους μειώνονται από τη λειτουργία του κυκλώματος, καλείται με διάφορους τρόπους να αντισταθμίσει τα μειωμένα έσοδα.

- Τίμιοι έμποροι

Η απάτη στο ΦΠΑ δημιουργεί στρεβλώσεις στην αγορά. Οι τυπικές επιχειρήσεις επιβαρύνονται είτε έμμεσα με αύξηση της φορολογίας, όπως και οι φορολογούμενοι, είτε άμεσα με την ύπαρξη αθέμιτου ανταγωνισμού.

ΚΕΦΑΛΑΙΟ 5 - Ερευνητικά ερωτήματα και συλλογή στοιχείων

1. Στόχος/ ερευνητικά ερωτήματα

Στην παρούσα εργασία θα γίνει εφαρμογή των μεθόδων εξόρυξης σε δεδομένα φορολογικού περιεχομένου στον τομέα του Φ.Π.Α.

Σκοπός της εργασίας είναι να ερευνηθεί :

- Αν οι γνωστότερες από τις τεχνικές εξόρυξης μπορούν να χρησιμοποιηθούν σε φορολογικά δεδομένα και τι απαιτείται ώστε να υλοποιηθούν.
- Αν ορισμένα από τα υπάρχοντα εργαλεία ανοικτού κώδικα μπορούν να χρησιμοποιηθούν ώστε να υλοποιηθούν οι τεχνικές εξόρυξης
- Τα προβλήματα και οι περιορισμοί που αναδεικνύονται από την πρακτική εφαρμογή των τεχνικών εξόρυξης.

Στη βιβλιογραφία υπάρχουν αρκετές μελέτες που αφορούν την εξόρυξη γνώσης από φορολογικά – οικονομικά στοιχεία καθώς η πρόβλεψη της οικονομικής απάτης είναι ένας τομέας που συγκεντρώνει μεγάλο ενδιαφέρον.

Συνήθως οι υπάρχουσες εργασίες επικεντρώνονται στην εξέταση της αποτελεσματικότητας μιας συγκεκριμένης μεθόδου σε σχέση με τα διαθέσιμα στοιχεία.

Στην παρούσα μελέτη θα προσπαθήσουμε να εξετάσουμε συνολικότερα την διαδικασία, από τα αρχικά στάδια έως την παραγωγή των τελικών αποτελεσμάτων. Σκοπός της μελέτης αυτής είναι να εντοπισθούν και να αναδειχθούν οι παράμετροι αυτοί που επηρεάζουν το τελικό αποτέλεσμα μιας διαδικασίας εξόρυξης γνώσης.

2. Συλλογή των δεδομένων.

Το πρώτο βήμα σε κάθε διαδικασία εξόρυξης είναι η συλλογή των κατάλληλων στοιχείων – χαρακτηριστικών που θα χρησιμοποιηθούν. Βασικός οδηγός σε κάθε περίπτωση αποτελεί το επιχειρησιακό πλαίσιο μέσα στο οποίο θα υλοποιηθεί η δράση.

Οι ιδιαιτερότητες, το νομοθετικό πλαίσιο, τα διαθέσιμα στοιχεία, ο χρόνος που αυτά γίνονται διαθέσιμα, το πλήθος των στοιχείων, η απαιτήσις για επανάληψη της διαδικασίας ή όχι, το χρονικό περιθώριο για την εκτέλεση της διαδικασίας (μια φορά, επείγουσα ανάγκη κλπ. είναι ορισμένες από τις απαιτήσεις που θα πρέπει να ληφθούν υπόψη κατά το σχεδιασμό της διαδικασίας.

Παρακάτω ακολουθούν μερικές παράμετροι που λήφθηκαν υπόψη για την υλοποίηση της παρούσας εργασίας :

- Είδος στοιχείων

Καθώς η παρούσα εργασία είναι διερευνητική επιλέχθηκαν στοιχεία που εμπειρικά έχουν παρατηρηθεί ότι συμβάλουν στην διαμόρφωση του προφίλ μιας επιχείρησης που ασκεί ενδοκοινοτικό εμπόριο. Τα στοιχεία αποτελούνται από ένα συνδυασμό στατικών γνωρισμάτων (στοιχεία μητρώου) και δυναμικών (στοιχεία συναλλαγών).

Ειδικά για τα δεύτερα και για να μειωθεί η εξάρτηση από αριθμητικές τιμές και τα προβλήματα που αυτές προκαλούν, όπως ανάγκη κανονικοποίησης, τμηματοποίησης κλπ. επιλέχθηκε τα γνωρίσματα να μετατραπούν εξ αρχής σε κατηγορικές τιμές. Αντί των γνωρισμάτων της μορφής πχ «ύψος συναλλαγών», δημιουργήθηκαν γνωρίσματα, όπως «επιχείρηση με μεγάλη αύξηση συναλλαγών». Στην παράγραφο 5.4 γίνεται πιο λεπτομερής αναφορά στο είδος των γνωρισμάτων.

- Επιχειρησιακό – νομοθετικό πλαίσιο

Οι Ελληνικές επιχειρήσεις έχουν υποχρέωση να υποβάλουν κάθε μήνα δηλώσεις αγορών (αποκτήσεων), καθώς και δηλώσεις πωλήσεων (παραδόσεων). Στις δηλώσεις πρέπει να αναγράφουν το σύνολο της αξίας των αγορών και των πωλήσεων αντίστοιχα ανά ξένο ΑΦΜ με το οποίο είχαν συναλλαγές. Μέσω αυτών των στοιχείων μπορούν να δημιουργηθούν διάφορες μετρικές, πχ ύψος συναλλαγών, κατηγοριοποίηση των επιχειρήσεων σε μικρές/μεσαίες/μεγάλες, χαρακτηρισμό των συναλλαγών ως «ακραίων» κλπ.

- Ιδιαιτερότητες

Σε αντίθεση με το Ελληνικό νομοθετικό πλαίσιο οι υπόλοιπες Ευρωπαϊκές χώρες εφαρμόζουν διαφορετικές περιόδους συλλογής των στοιχείων. Υπάρχουν χώρες που συλλέγουν τα στοιχεία ανά τρίμηνο ή μήνα ή καθόλου ή η περίοδος εξαρτάται από το ύψος των συναλλαγών της επιχείρησης. Για το λόγο αυτό επιλέχθηκε τα χαρακτηριστικά που αφορούν συναλλαγές να αναχθούν σε επίπεδο τριμήνου.

- Στατικά στοιχεία – Στοιχεία μητρώου

Εκτός από τα στοιχεία που αφορούν συναλλαγές και αφορούν δυναμικά στοιχεία, δηλαδή στοιχεία που έχουν την τάση να αλλάζουν στη διάρκεια ενός οικονομικού έτους συλλέχθηκαν στοιχεία που αποτελούν την ταυτότητα της επιχείρησης, όπως οι κωδικοί δραστηριότητας, αν η επιχείρηση είναι νέα, η ηλικία της επιχείρησης, ο ταχυδρομικός κωδικός της έδρας της επιχείρησης κλπ.

- Απαιτήσεις

Για την παρούσα εργασία δεν υπήρχαν συγκεκριμένοι περιορισμοί για την διαμόρφωση της διαδικασίας ανάλυσης, όπως πχ τη χρήση συγκεκριμένου εργαλείου, αυστηρές ημερομηνίες παράδοσης κλπ. Ήταν όμως απαραίτητο να υιοθετηθούν ορισμένα επιπλέον βήματα, ώστε να καλυφθούν οι απαιτήσεις που προκύπτουν για λόγους ανωνυμίας των δεδομένων ή επιβάλλονται λόγω τεχνικών θεμάτων.

3. *Ανωνυμία δεδομένων*

Διάφορες τεχνικές είναι διαθέσιμες για να πετύχουμε την ανωνυμία των δεδομένων (B. Baesens 2015):

- Ομαδοποίηση των δεδομένων

Τα ομαδοποιημένα δεδομένα παρέχουν παρόμοια πληροφόρηση με τα αναλυτικά στοιχεία χωρίς να αποκαλύπτουν πληροφορίες σχετικές με τα άτομα.

- Γενίκευση των δεδομένων

Οι αριθμητικές τιμές μπορούν να ανωνυμοποιηθούν με την μετατροπή τους σε ομάδες τιμών. Στην περίπτωση αυτή αντί να υπάρχει σαφής αναφορά σε αριθμητική τιμή, αυτές χωρίζονται σε διακριτές ομάδες.

- Παραμόρφωση των στοιχείων

Τα αριθμητικά δεδομένα μπορούν επίσης ανωνυμοποιηθούν προσθέτοντας θόρυβο. Συνήθως αυτό γίνεται προσθέτοντας στις τιμές τους τιμές από μια συγκεκριμένη κατανομή.

- Μετατροπή των τιμών σε διακριτές

Εναλλακτικά οι αριθμητικές τιμές θα μπορούσαν να γενικευθούν χρησιμοποιώντας μια πιο γενική αλλά σχετική τιμή, πχ μια διεύθυνση μπορεί να αντικατασταθεί με τον ταχυδρομικό κωδικό ή την πόλη.

Στην παρούσα μελέτη δόθηκε προσοχή ώστε τα στοιχεία που χρησιμοποιήθηκαν να μην περιέχουν αναφορές οι οποίες θα μπορούσαν να χρησιμοποιηθούν για να ταυτοποιηθούν οι επιχειρήσεις.

Η βασική τεχνική που χρησιμοποιήθηκε για την ανωνυμοποίηση των στοιχείων είναι η μετατροπή τους από αριθμητικές τιμές σε κατηγορικές. Με την συγκεκριμένη μετατροπή ουσιαστικά εφαρμόζουμε στα δεδομένα μας την ταυτόχρονα την γενίκευση, την μετατροπή σε διακριτές τιμές και την ομαδοποίηση.

Μετά την μετατροπή, στα αρχικά στοιχεία παραμένει το σημαντικότερο στοιχείο ταυτοποίησης που οδηγεί απευθείας σε αναγνώριση των επιχειρήσεων, το ΑΦΜ.

Για την εκτέλεση των διαδικασιών εξόρυξης η ύπαρξη του αντιμετωπίστηκε με δύο τρόπους :

- Για το εργαλείο WEKA και καθώς η παρουσία του δεν εξυπηρετεί κάποιο συγκεκριμένο σκοπό, το ΑΦΜ διαγράφηκε εντελώς.
- Για την χρήση του σετ δεδομένων με τον Data Miner καθώς και για την δημιουργία των δικτύων (εργαλεία Kpime και Cytoscape) και εφόσον τα εργαλεία χρειάζονται κάποια τιμή ως αναγνωριστικό των περιπτώσεων για να χαρακτηρίσουν τα αποτελέσματα το ΑΦΜ αντικαταστάθηκε με ένα hash κωδικό.

Η μετατροπή των τιμών σε κατηγορικές τιμές εξυπηρετεί επίσης και την κανονικοποίηση των τιμών. Για το θέμα αυτό θα αναφερθούμε περαιτέρω παρακάτω.

4. Τα δεδομένα (γνωρίσματα) που θα χρησιμοποιηθούν.

Τα στοιχεία που χρησιμοποιήθηκαν αφορούν ελληνικές επιχειρήσεις οι οποίες δραστηριοποιούνται στο ενδοκοινοτικό εμπόριο τα έτη 2010-2015. Για να βρεθούν οι επιχειρήσεις αυτές αντλήθηκαν στοιχεία από :

- Στοιχεία που προέρχονται από το μητρώο των επιχειρήσεων. Η επιλογή των στοιχείων έγινε με τρόπο ώστε αυτά να περιγράφουν με τον απλούστερο τρόπο την επιχείρηση και τον βασικό πυρήνα λειτουργίας αυτής.
- Στοιχεία από τις συναλλαγές τις επιχειρήσεις στο ενδοκοινοτικό εμπόριο. Τα στοιχεία αφορούν ποσοτικούς δείκτες και όχι στοιχεία συναλλαγών. Πχ μας ενδιαφέρει εάν υπάρχουν απότομες εναλλαγές στο ύψος των συναλλαγών αλλά δεν μας ενδιαφέρει το ύψος των συναλλαγών.

Θα πρέπει να επαναλάβουμε ότι η παρούσα εργασία δεν αποσκοπεί στον εντοπισμού της απάτης, ούτε στην δημιουργία ενός αποτελεσματικού μοντέλου. Κύριος σκοπός είναι να διερευνήσει κατά πόσο και με πιο τρόπο οι διαδικασίες εξόρυξης γνώσης μπορούν να εφαρμοσθούν στον συγκεκριμένο τομέα. Το πλήθος των χαρακτηριστικών που επιλέχθηκαν είναι το ελάχιστο δυνατό και έχει στόχο να κρατήσει σε διαχειρίσιμο επίπεδο την πολυπλοκότητα τόσο των μοντέλων όσο και των υπολογιστική ισχύ που απαιτείται για την δημιουργία και επαλήθευση των μοντέλων.

Με βάση τα παραπάνω τα χαρακτηριστικά που επιλέχθηκαν είναι τα :

Όνομα	Τύπος Μεταβλητής	Τιμές	Πηγή	Περιγραφή
VAT	Label/ID		Μητρώο	ΑΦΜ
KAD1	nominal	Διακριτές τιμές	Μητρώο	Κωδικός Δραστηριότητας 1
KAD2	nominal	Διακριτές τιμές	Μητρώο	Κωδικός Δραστηριότητας 2
KAD3	nominal	Διακριτές τιμές	Μητρώο	Κωδικός Δραστηριότητας 3
KAD4	nominal	Διακριτές τιμές	Μητρώο	Κωδικός Δραστηριότητας 4
KAD5	nominal	Διακριτές τιμές	Μητρώο	Κωδικός Δραστηριότητας 5
COMPANY_TYPE	nominal	Διακριτές τιμές	Μητρώο	Τύπος Επιχείρησης
AGEOFFIRM		Αριθμός	Μητρώο	Ηλικία επιχείρησης
SMALLLIFETIME	nominal	0 ή 1	Μητρώο	Επιχείρηση ηλικίας < 1.5 έτους
DIFF	nominal	0 ή 1	Συναλλαγές	Παρουσιάζονται διαφορές μεταξύ των αγορών και των πωλήσεων της επιχείρησης μέσα στο ίδιο χρονικό διάστημα μεγαλύτερες από 110000 Ευρώ.
ALT_MARKED	nominal	0 ή 1	Μητρώο	Το ΑΦΜ της επιχείρησης εμφανίζεται σε λίστες ελέγχου που συντάσσουν άλλες χώρες
CONNECTED	nominal	0 ή 1	Συναλλαγές	Παρουσιάζει αγορές από ξένες επιχειρήσεις οι οποίες πιθανά έχουν προβλήματα
TRANSACTIONS	nominal	0 ή 1	Συναλλαγές	Διαφορά μεταξύ αυτών που δηλώνει η επιχείρηση ότι αγόρασε με αυτά που δηλώνουν οι ξένοι ότι της πούλησαν
VAT_FLAG	nominal	Διακριτές τιμές	Μητρώο	Ανήκει σε κατηγορία που εφαρμόζονται η διατάξεις για τον ΦΠΑ
VAT_STATUS	nominal	Διακριτές τιμές	Μητρώο	Η κατηγορία ΦΠΑ που ανήκει η επιχείρηση
VIES_OPTION	nominal	Διακριτές τιμές	Μητρώο	Μπορεί να ασκήσει ενδοκοινοτικό εμπόριο ακόμα και αν δεν ανήκει σε κατηγορία που εφαρμόζονται οι διατάξεις του ΦΠΑ
BOOKS_KIND	nominal	Διακριτές τιμές	Μητρώο	Το είδος των λογιστικών βιβλίων που τηρεί η επιχείρηση
PAR_BUSINESS	nominal	Διακριτές τιμές	Μητρώο	Ταχυδρομικός κωδικός
ABROADMS	nominal	Διακριτές τιμές	Μητρώο	Η χώρα προέλευσης της μητρικής εταιρείας ή του επιχειρηματία
QUARTER_EXTRD IFF	nominal	0 ή 1	Συναλλαγές	Εμφανίζονται μεγάλες διακυμάνσεις μεταξύ των αγορών που δηλώνει η επιχείρηση μεταξύ των τριμήνων
FIRST_TIME	nominal	0 ή 1	Μητρώο	Είναι η πρώτη φορά που δηλώνει έναρξη λειτουργίας
ORIGIN	nominal	Διακριτές τιμές	Μητρώο	Λόγος έναρξης
IMPORTS	nominal	0 ή 1	Συναλλαγές	Έχει μόνο αγορές από το εξωτερικό και

Όνομα	Τύπος Μεταβλητής	Τιμές	Πηγή	Περιγραφή
				καθόλου πωλήσεις
CONNECTED_T	nominal	0 ή 1	Συναλλαγές	Παρουσιάζει πωλήσεις σε ξένες επιχειρήσεις που πιθανά έχουν προβλήματα
ONLY_VOW	nominal	0 ή 1	Συναλλαγές	Το ΑΦΜ της επιχείρησης δεν έχει συναλλαγές αλλά εμφανίζεται να το αναζητούν άλλες επιχειρήσεις
VAT_OPTION	nominal	0 ή 1	Μητρώο	Συμμετέχει στο σύστημα VIES, επειδή ανήκει σε κατηγορία που εφαρμόζονται οι διατάξεις του ΦΠΑ
REQ_OPTION	nominal	0 ή 1	Μητρώο	Συμμετέχει στο σύστημα VIES, με βάση τη δήλωση της επιχείρησης ότι θα ασκήσει ενδοκοινοτικό εμπόριο
CLASS	Class	0 ή 1		Αφανής έμπορος

Εικόνα 19 - Γνωρίσματα που χρησιμοποιήθηκαν στην εργασία

ΚΕΦΑΛΑΙΟ 6 - Το τεχνικό υπόβαθρο

1. *Εργαλεία που χρησιμοποιήθηκαν*

Για την εκτέλεση της εργασίας επιλέχθηκαν τα εξής εργαλεία :

- Oracle Sql Developer με plugin τον Data Miner
- WEKA
- Knime
- Cytoscape

2. *SQL Developer – Data Miner*

Η εταιρεία Oracle, γνωστή για την βάση δεδομένων και τα εργαλεία που σχετίζονται με αυτή, εντάσσει στο εύρος των υπηρεσιών που παρέχει και υπηρεσίες data-mining.

Οι υπηρεσίες data mining παρέχονται με τους εξής τρόπους:

- Τη δημιουργία υποδομής εντός της βάσης.

Η εταιρεία έχει αναπτύξει και ενσωματώσει στη υποδομή της βάσης αλγόριθμους εξόρυξης (decision tress, SVM κλπ).

- Την ενσωμάτωση της γλώσσας R.

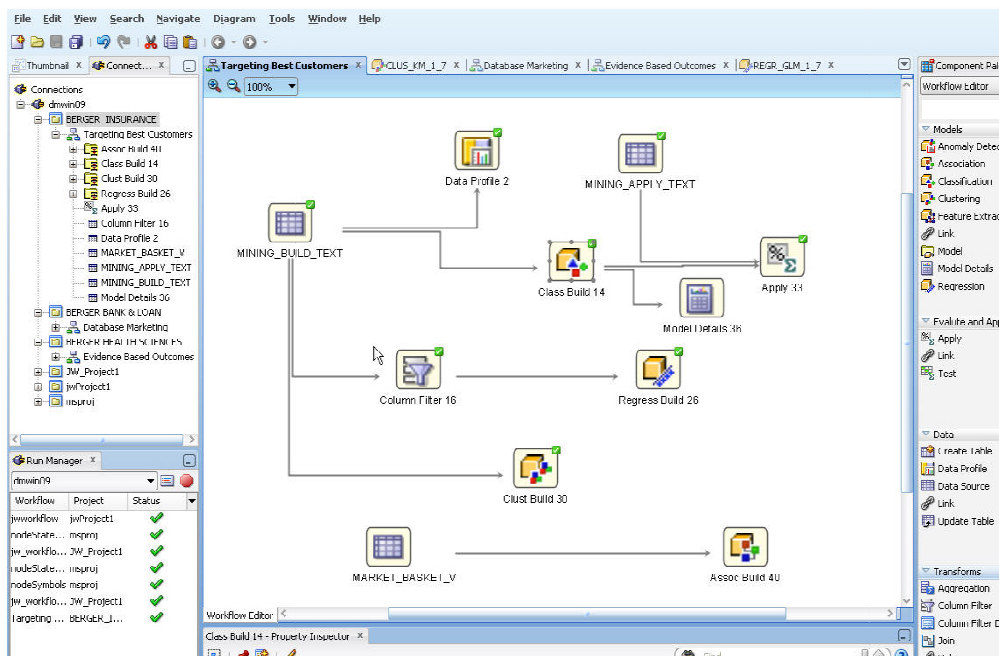
Η γλώσσα R, έχει αναδειχθεί ως η κατεξοχήν γλώσσα που χρησιμοποιείται για την δημιουργία αλγορίθμων εξόρυξης. Με την ενσωμάτωση της η εταιρεία αποκτά πρόσβαση στις βιβλιοθήκες που έχουν αναπτυχθεί σε αυτή τη γλώσσα.

Και οι δύο προσεγγίσεις απαιτούν την γνώση του χρήστη στο προγραμματιστικό περιβάλλον με το οποίο θα ασχοληθεί. Για να βοηθήσει στην εξάπλωση των διαδικασιών εξόρυξης η εταιρεία έχει δημιουργήσει (για την πρώτη περίπτωση) το εργαλείο “Data Miner”. Το εργαλείο παρέχεται ως plugin στο εργαλείο ανάπτυξης “SQL developer”. Τόσο το plugin όσο και το εργαλείο ανάπτυξης είναι διαθέσιμα άνευ χρέωσης από την εταιρεία. Για να χρησιμοποιηθούν απαιτείται η διασύνδεση με μια βάση Oracle έκδοσης ίσης ή μεγαλύτερης από 10g.

Η διαδικασία δημιουργίας ενός μοντέλου μπορεί να γίνει με δύο τρόπους :

- Είτε μέσω της δημιουργίας εντολών, δηλαδή με τρόπο ανάλογο με την δημιουργία ενός query προς την βάση.
- Είτε μέσω του γραφικού περιβάλλοντος που παρέχει το εργαλείο

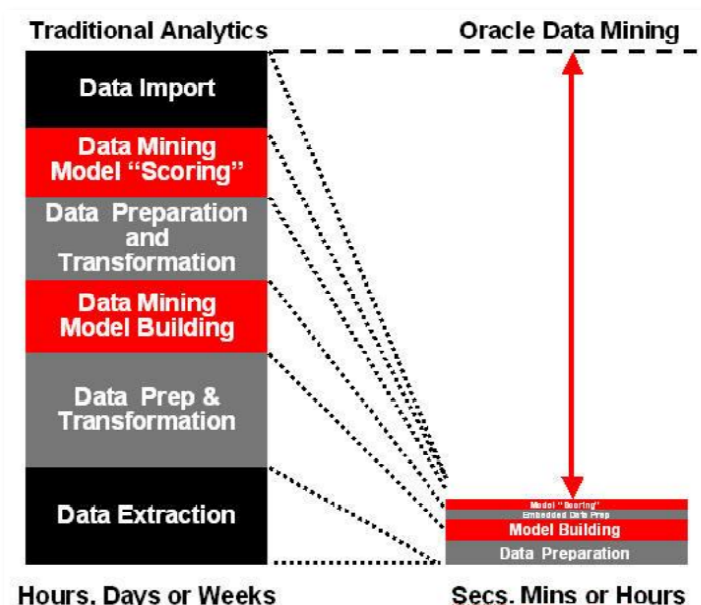
Βασικό χαρακτηριστικό του εργαλείου Data Miner είναι η ευχρηστία του, επιτρέποντας ακόμα και σε σχετικά νέους χρήστες να δημιουργήσουν γρήγορα διαδικασίες εξόρυξης.



Εικόνα 20- Oracle data miner γραφικό περιβάλλον(oracle.com, 2015)

Σύμφωνα με την εταιρεία η ενσωμάτωση στη βάση των αλγόριθμων εξόρυξης αναμένεται να διευκολύνει σημαντικά την εξόρυξη γνώσης, αλλά και να μειώσει το συνολικό κόστος αυτής.

Σύμφωνα πάντα με την εταιρεία η ενσωμάτωση των αλγόριθμων εξόρυξης στην βάση μειώνει τον χρόνο ανάπτυξης ενός έργου εξόρυξης, μειώνοντας το χρόνο που σε άλλες περιπτώσεις διατίθεται για την εξαγωγή και την εκκαθάριση των δεδομένων τα οποία στη συνέχεια θα χρησιμοποιηθούν από εξωτερικά εργαλεία.



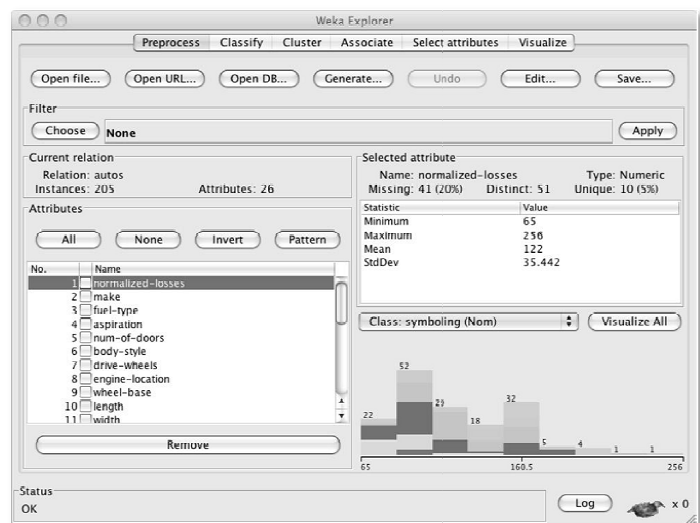
Εικόνα 21- Oracle (2012), Oracle Data Mining 11g Release 2 - Competing on In-Database Analytics

3. WEKA

Το WEKA (Waikato Environment for Knowledge Analysis) είναι ένα open source εργαλείο γραμμένο σε Java, γεγονός που του επιτρέπει να τρέχει σε οποιαδήποτε πλατφόρμα υπάρχει υλοποίηση της JVM.

Όπως φαίνεται και από το όνομα του, δεν είναι ένα μονολιθικό εργαλείο αλλά ένα περιβάλλον για την ανάπτυξη και χρήση αλγορίθμων εξόρυξης. Το περιβάλλον αποτελείται από ένα γραφικό interface και ένα σύνολο αλγορίθμων όλα υλοποιημένα σε Java. Το γραφικό interface παρέχει προς τον χρήστη ένα απλοποιημένο περιβάλλον για την εκτέλεση των εργασιών.

Οι αλγόριθμοι που περιλαμβάνονται στο εργαλείο είναι Java προγράμματα (κλάσεις), τα οποία μπορούν να εκτελεστούν αυτόνομα από την γραμμή εντολών (command line interface) ή μέσω του γραφικού interface ή να ενσωματωθούν σε άλλα προγράμματα.



Εικόνα 22 Weka Interface (I. Witten, E. Frank, M. Hall (2011). "Data Mining: Practical machine learning tools and techniques, 3rd Edition")

Κύρια πηγή εισόδου των δεδομένων είναι flat ascii αρχεία τα οποία έχουν συγκεκριμένη δομή.

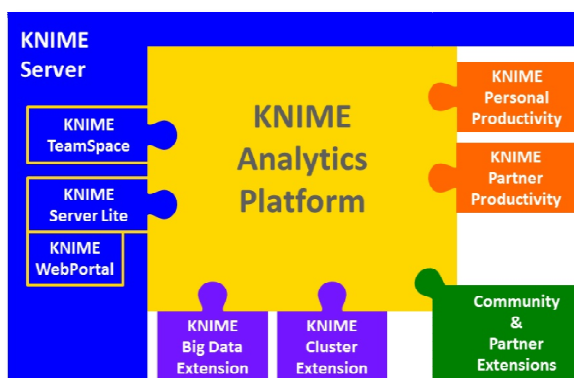
Σαν πηγή εισόδου μπορούν επίσης να χρησιμοποιηθούν βάσεις δεδομένων μέσω jdbc σύνδεσης ή δομές Big Data, όπως το σύστημα Cassandra μέσω πρόσθετων plugins.

Οι βασικές λειτουργίες του εργαλείου μπορούν να χωρισθούν σε τρεις τομείς:

- Εφαρμογή των αλγορίθμων εξόρυξης στα στοιχεία, για την διερεύνηση των στοιχείων και την υλοποίηση μοντέλων
- Εφαρμογή των μοντέλων που δημιουργήθηκαν από το προηγούμενο βήμα, για πρόβλεψη
- Σύγκριση της απόδοσης των διάφορων μεθόδων/αλγορίθμων εξόρυξης πάνω σε νέα στοιχεία

Ο τρόπος υλοποίησης του εργαλείου καθώς και το πλήθος των αλγορίθμων που περιλαμβάνει το καθιστούν ένα ευέλικτο εργαλείο, το οποίο για να αποδώσει θα πρέπει ο χρήστης να αφιερώσει λίγο χρόνο, ώστε να εξοικειωθεί αρχικά με το interface λειτουργίας του, και σε επόμενη φάση με τους διαθέσιμους αλγόριθμους και τις παραμέτρους αυτών.

4. *Knime*



Εικόνα 23- *Knime product overview*
(*Knime.com,2016*)

απεικόνιση των δεδομένων και των αναλύσεων, η χρήση κεντρικού repository για τα μοντέλα, connectors για δομές big data κλπ.

Το εργαλείο Knime αποτελεί ουσιαστικά μια ολοκληρωμένη πλατφόρμα εκτέλεσης εργασιών εξόρυξης. Βασικός κορμός αλλά και το εργαλείο που χρησιμοποιήθηκε στην παρούσα εργασία αποτελεί το module «Knime Analytics Platform».

Η βασική πλατφόρμα αποτελεί το συνδετικό κρίκο για την προσθήκη επιπλέον modules τα οποία μπορούν να επανξήσουν την λειτουργικότητα του εργαλείου. Ενδεικτικά αναφέρονται η δημιουργία web portal για την

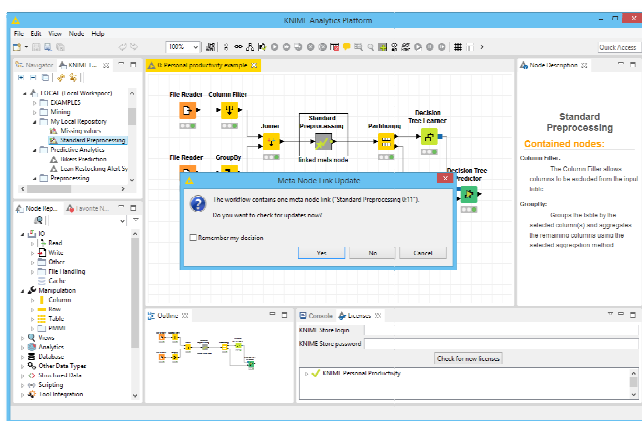
Η βασική πλατφόρμα χρησιμοποιεί το περιβάλλον eclipse για την υλοποίηση ενός γραφικού interface προς τον χρήστη.

Ο χρήστης έχει στη διάθεση του εικονίδια τα οποία εκτελούν συγκεκριμένες εργασίες (πχ decision tree, association, database query, filter κλπ.).

Με τα εικονίδια μπορεί να συνθέσει μια ροή εργασιών τοποθετώντας τα στον καμβά, ενώνοντας τα και τροποποιώντας τις παραμέτρους τους.

Σύμφωνα με την ομάδα ανάπτυξης το βασικά πλεονεκτήματα του εργαλείου εντοπίζονται :

- Στη δυνατότητα γραφικής αναπαράστασης της διαδικασίας εξόρυξης



Εικόνα 24- *Knime βασικό Interface*(*Knime.com,2016*)

- Στη δυνατότητα δημιουργίας νέων «εικονιδίων» από την ομαδοποίηση ροών που δημιουργήθηκαν από το χρήστη
- Στη δυνατότητα διασύνδεσης του εργαλείου με διαδικασίες που έχουν γραφεί για άλλα εργαλεία/περιβάλλοντα, όπως πχ χρήση των αλγόριθμων που έχουν αναπτυχθεί για το WEKA ή η χρήση αλγόριθμων σε γλώσσα R

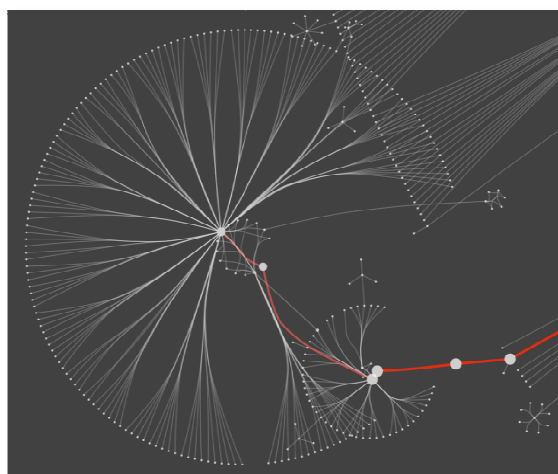
5. *Cytoscape*

Το συγκεκριμένο εργαλείο ξεφεύγει από το μοτίβο των προηγούμενων, καθώς δεν είναι ένα γενικό εργαλείο εξόρυξης, αλλά επικεντρώνεται στην απεικόνιση και διερεύνηση δικτύων.

Το εργαλείο υποστηρίζεται και αναπτύσσεται από ένα συνεργατικό σχήμα στο οποίο συμμετέχουν αρκετά πανεπιστήμια της Αμερικής και του Καναδά. Βασικός στόχος του εργαλείου είναι η διερεύνηση βιολογικών δικτύων, όπως πχ την ανακάλυψη των σχέσεων μεταξύ σειρών γονιδιωμάτων, την απεικόνιση των κυτταρικών δικτύων επικοινωνίας κλπ.

Το ίδιο το εργαλείο είναι δομημένο ώστε να είναι ουδέτερο ως προς τα στοιχεία που θα απεικονίσει. Παρέχει γενικά εργαλεία εξερεύνησης και απεικόνισης οποιουδήποτε τύπου δικτύου.

Βασικό χαρακτηριστικό του εργαλείου είναι η δυνατότητα του να επεκτείνεται εύκολα με την χρήση plugins.



Εικόνα 25 -Cytoscape απεικόνιση δικτύου (cytoscape.com, 2015)

Το εργαλείο παρέχει δυνατότητα ανάγνωσης από πολλαπλά πρότυπα αρχείων (SBML, GraphML, GML κλπ).

Για την διερεύνηση των δικτύων παρέχει interface ώστε ο χρήστης να μπορεί να μεταβάλλει χαρακτηριστικά του δικτύου (χρώμα κόμβων, μέγεθος συνδέσμων κλπ.).

ΚΕΦΑΛΑΙΟ 7- Εξόρυξη δεδομένων

1. Διαμόρφωση στοιχείων

Έχοντας εκτελέσει τις πρώτες φάσεις της διαδικασίας εξόρυξης, δηλαδή τα βήματα:

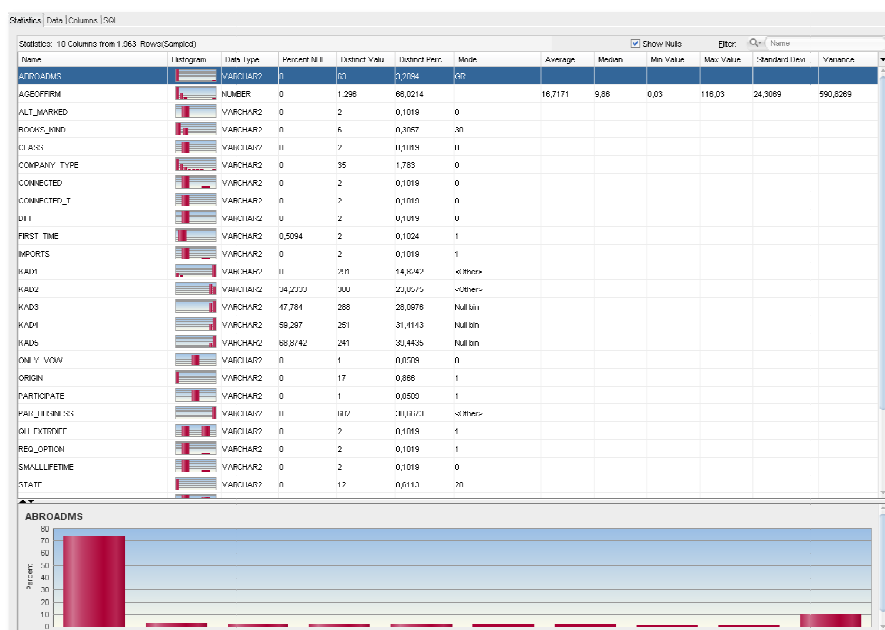
- Διερεύνησης του επιχειρηματικού πεδίου
- Συλλογής των στοιχείων
- Του μετασχηματισμού των στοιχείων
- Της εκκαθάρισης των στοιχείων

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

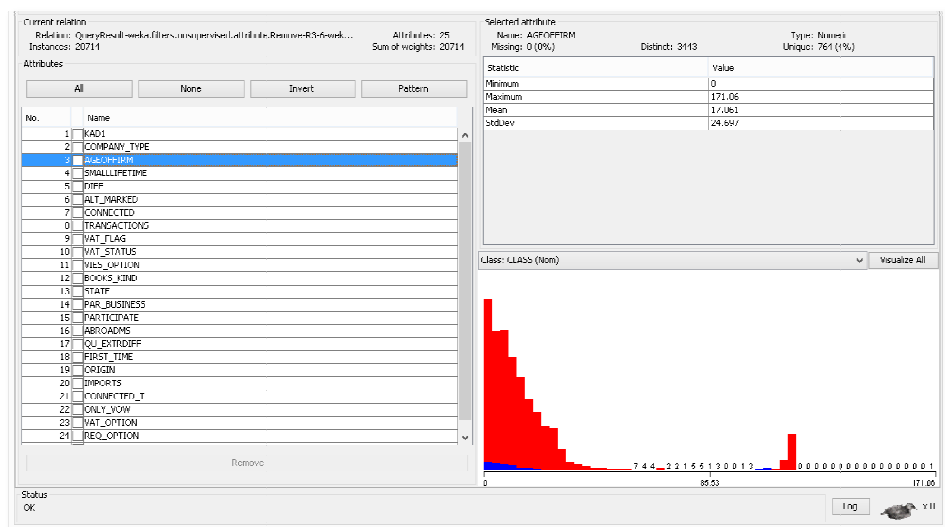
2. Επισκόπηση στοιχείων

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.

Όλα τα εργαλεία που χρησιμοποιήθηκαν σε αυτή την εργασία παρέχουν είτε σε μεγαλύτερο ή σε μικρότερο βαθμό εργαλεία επισκόπησης των δεδομένων.

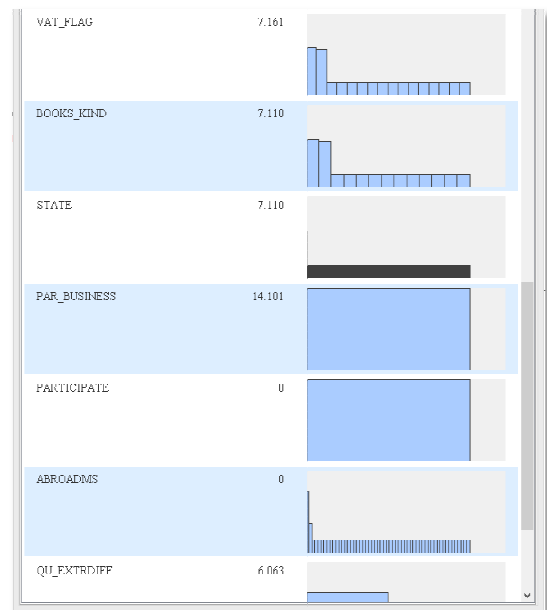


Εικόνα 27- Προ-επισκόπηση στοιχείων με το Oracle Data Miner



Εικόνα 28- Προ-επισκόπηση γνωρισμάτων με το WEKA

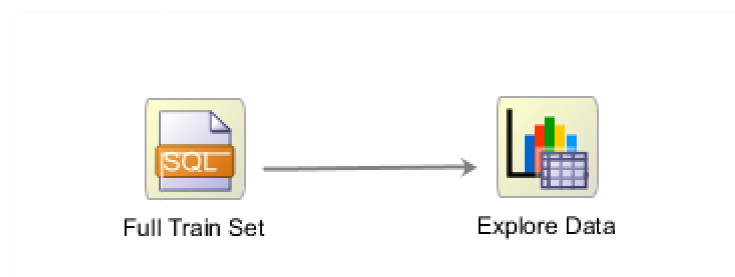
KADI	COMP_TYPE	SMALLLIFETIME	DIFF	EUROFISC	CONNECTED	TRANSACTIONS	COMPTYPE
No. missing: 0	No. missing: 0	No. missing: 0	No. missing: 0	No. missing: 0	No. missing: 0	No. missing: 0	No. missing: 0
Top 20:	Top 20:	Top 20:	Top 20:	Top 20:	Top 20:	Top 20:	Top 20:
9999999 : 2541	0 : 4184	0 : 12220	0 : 8628	0 : 14074	0 : 14101	0 : 14101	0 : 14101
4669 : 1172	7 : 3080	1 : 1881	1 : 5473	1 : 27			
4752 : 390	1 : 2342						
4673 : 385	4 : 1183						
7022 : 330	2 : 965						
4772 : 326	5 : 730						
4642 : 258	89 : 166						
4771 : 218	57 : 92						
4671 : 193	96 : 77						
4759 : 183	97 : 62						
4764 : 179	38 : 61						
1634 : 174	75 : 55						
3312 : 157	51 : 45						
5510 : 141	81 : 37						
4646 : 140	61 : 33						
4631 : 136	60 : 21						
4511 : 124	35 : 20						
4540 : 120	78 : 20						
1610 : 121	9 : 18						
5220 : 121	82 : 15						
Bottom 20:	Bottom 20:	Bottom 20:	Bottom 20:	Bottom 20:	Bottom 20:	Bottom 20:	Bottom 20:
2630 : 1	76 : 2						
3210 : 1	3 : 2						
3012 : 1	66 : 1						
11110 : 1	79 : 1						
4291 : 1	9 : 1						
1231 : 1	20 : 1						
1461 : 1	30 : 1						
8422 : 1	37 : 1						
9999 : 1	32 : 1						
2811 : 1	28 : 1						
4942 : 1	36 : 1						
4763 : 1	19 : 1						
2223 : 1	45 : 1						
7237 : 1	34 : 1						
9315 : 1	27 : 1						
7340 : 1	55 : 1						
4331 : 1	54 : 1						
4331 : 1	54 : 1						



Εικόνα 29 - Προ-επισκόπηση γνωρισμάτων με το Knime (I)

Για το σκοπό αυτό χρησιμοποιήθηκε το εργαλείο Oracle Data Miner.

Στο εργαλείο δημιουργήθηκε το παρακάτω μοντέλο :



Εικόνα 30- Ροή για προεπισκόπηση στοιχείων στον Data Miner

Το μοντέλο παρέχει μια βασική επισκόπηση των στοιχείων. Χρησιμοποιώντας τα αποτελέσματα από την εμφάνιση των ποσοτικών χαρακτηριστικών των στοιχείων θα τροποποιήσουμε το σετ των στοιχείων ή/και των γνωρισμάτων.

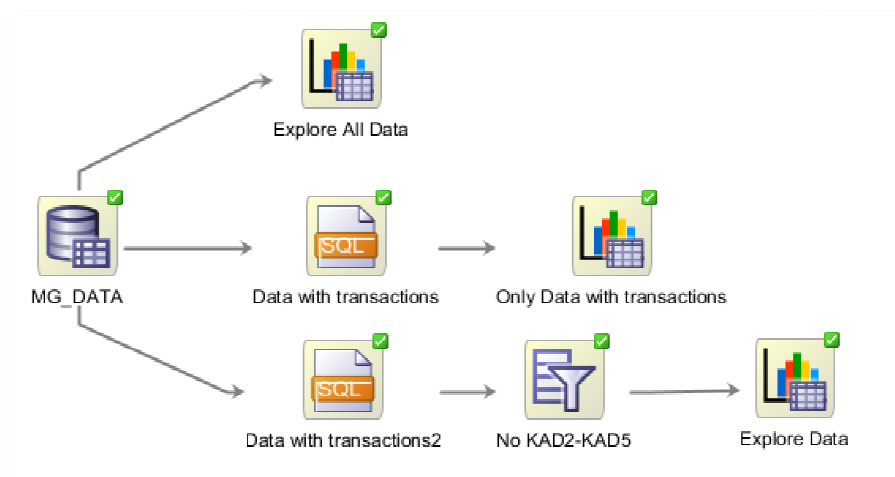
Από την διερεύνηση των στοιχείων προκύπτει ότι η παρότι στα δεδομένα έχουν συμπεριληφθεί πέντε κωδικοί δραστηριότητας η μεγάλη πλειοψηφία των πρόσθετων κωδικών δραστηριοτήτων είναι κενές.

	ATTR	DATA_TYPE	NULL_PERCENT	DISTINCT_CNT	DISTINCT_PERCENT	MODE_VALUE
1	KAD5	VARCHAR2	75,7576	199	40,7787	Null bin
2	KAD4	VARCHAR2	65,4247	222	31,8966	Null bin
3	KAD3	VARCHAR2	51,4158	280	28,6299	Null bin
4	KAD2	VARCHAR2	34,3269	308	23,298	<Other>
5	VAT_...	VARCHAR2	1,9374	13	0,6586	4
6	STATE	VARCHAR2	0,5961	9	0,4498	20
7	PAR_...	VARCHAR2	0,5961	658	32,8836	<Other>
8	VAT_...	VARCHAR2	0,5961	2	0,1	1
9	VIES_...	VARCHAR2	0,5961	17	0,8496	10
10	BOOK...	VARCHAR2	0,5961	7	0,3498	30
11	FIRST...	VARCHAR2	0,2981	3	0,1495	1
12	COMP...	VARCHAR2	0	32	1,5897	0

Εικόνα 31- Διερεύνηση γνωρισμάτων

Στις γραμμές KAD2/KAD3/KAD4/KAD5 τα ποσοστά μη κενών γνωρισμάτων είναι από 25% - 65%. Δηλαδή οι περισσότερες από τις εταιρείες έχουν μια με δύο δραστηριότητες, ενώ μέχρι τρεις δραστηριότητες έχουν περίπου οι μισές επιχειρήσεις. Με αυτό το εύρημα και για την απλοποίηση των μοντέλων θα εξετασθεί αν οι επιπλέον δραστηριότητες (KAD2-KAD5) μπορούν να αγνοηθούν κατά τη δημιουργία μοντέλων.

Για τον έλεγχο της παραπάνω υπόθεσης δημιουργήθηκε η επόμενο ροή. Η ροή έχει δύο παράλληλους κλάδους. Στόχος της ροής είναι να ελεγχθεί αν τα στατιστικά των δεδομένων αλλάζουν όταν από τα στοιχεία αφαιρεθούν οι επιπλέον κωδικοί δραστηριοτήτων.



Εικόνα 32- Έλεγχος γνωρισμάτων

Από την εξέταση των αποτελεσμάτων, προκύπτει ότι η διαγραφή των κωδικών δραστηριότητας από τα δεδομένα δεν επηρεάζει σημαντικά την κατανομή των υπολοίπων γνωρισμάτων.

Statistics: 10 Columns from 2.013 Rows(Sampled)						
Name	Histogram	Data Type	Percent NUL...	Distinct Valu...	Distinct Perc...	Mode
ABROADMS		VARCHAR2	0	24	1,1923	GR
AGEOFFIRM		NUMBER	0	1,347	66,9151	
ALT_MARKED		VARCHAR2	0	1	0,0497	0
BOOKS_KIND		VARCHAR2	0,5981	7	0,3498	30
CLASS		VARCHAR2	0	2	0,0894	0
COMPANY_TYPE		VARCHAR2	0	32	1,5897	0
CONNECTED		VARCHAR2	0	2	0,0894	0
CONNECTED_T		VARCHAR2	0	2	0,0894	0
DIFF		VARCHAR2	0	2	0,0894	0
FIRST_TIME		VARCHAR2	0,2981	3	0,1495	1
IMPORTS		VARCHAR2	0	2	0,0894	1
KAD1		VARCHAR2	0	307	15,2509	<Other>
KAD2		VARCHAR2	34,3269	308	23,298	<Other>
KAD3		VARCHAR2	51,4158	280	28,6239	Null bin
KAD4		VARCHAR2	65,4247	222	31,8866	Null bin
KAD5		VARCHAR2	75,7576	199	40,7787	Null bin
ONLY_VOW		VARCHAR2	0	2	0,0894	0
ORIGIN		VARCHAR2	0	17	0,8445	1
PARTICIPATE		VARCHAR2	0	4	0,1987	0
PAR_BUSINESS		VARCHAR2	0,5961	658	32,8636	<Other>
QU_EXTRDIFF		VARCHAR2	0	2	0,0894	0
REQ_OPTION		VARCHAR2	0	2	0,0894	1
SMALLLIFETIME		VARCHAR2	0	2	0,0894	0
STATE		VARCHAR2	0,5961	9	0,4498	20
TRANSACTIONS		VARCHAR2	0	2	0,0894	0

Statistics: 10 Columns from 2.045 Rows(Sampled)						
Name	Histogram	Data Type	Percent NUL...	Distinct Valu...	Distinct Perc...	Mode
ABROADMS		VARCHAR2	0	21	1,0269	GR
AGEOFFIRM		NUMBER	0	1,348	66,9169	
ALT_MARKED		VARCHAR2	0	1	0,0489	0
BOOKS_KIND		VARCHAR2	0,0978	6	0,2837	30
CLASS		VARCHAR2	0	2	0,0878	0
COMPANY_TYPE		VARCHAR2	0	35	1,7115	0
CONNECTED		VARCHAR2	0	2	0,0878	0
CONNECTED_T		VARCHAR2	0	2	0,0878	0
DIFF		VARCHAR2	0	2	0,0878	0
FIRST_TIME		VARCHAR2	0,5868	3	0,1476	1
IMPORTS		VARCHAR2	0	2	0,0878	1
KAD1		VARCHAR2	0	319	15,599	<Other>
ONLY_VOW		VARCHAR2	0	1	0,0489	0
ORIGIN		VARCHAR2	0	20	0,978	1
PARTICIPATE		VARCHAR2	0	4	0,1855	0
PAR_BUSINESS		VARCHAR2	0,0978	650	31,815	<Other>
QU_EXTRDIFF		VARCHAR2	0	2	0,0878	0
REQ_OPTION		VARCHAR2	0	2	0,0878	1
SMALLLIFETIME		VARCHAR2	0	2	0,0878	0
STATE		VARCHAR2	0,0978	11	0,5384	20
TRANSACTIONS		VARCHAR2	0	2	0,0878	0
VAT		VARCHAR2	0	2.045	100	<Other>
VAT_FLAG		VARCHAR2	0,0978	2	0,0878	1
VAT_OPTION		VARCHAR2	0	2	0,0878	1
VAT_STATUS		VARCHAR2	1,8093	11	0,5478	4
VIES_OPTION		VARCHAR2	0,0978	18	0,8811	22

Εικόνα 33 - Σύγκριση γνωρισμάτων πριν και μετά την επεξεργασία

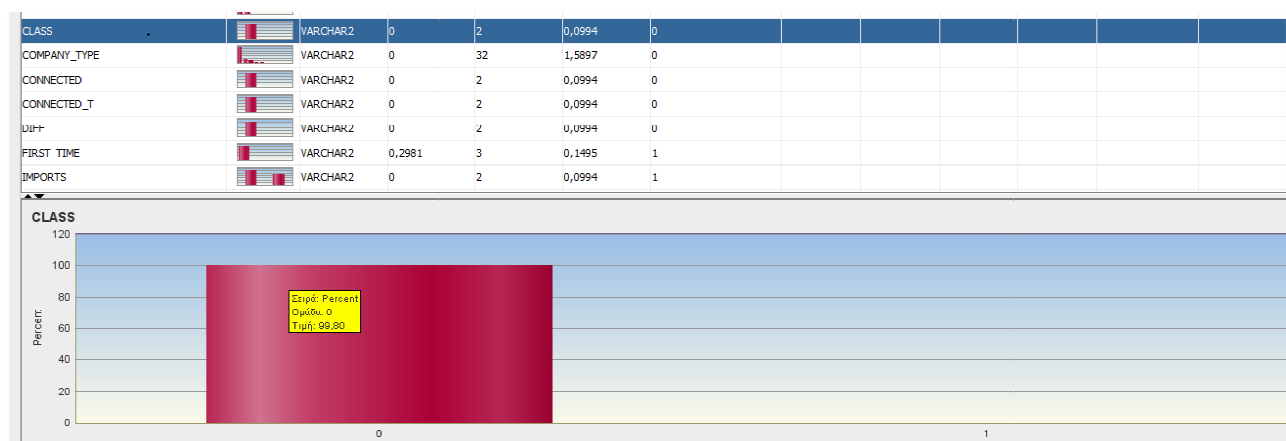
Σύμφωνα με την προεπισκόπηση των στοιχείων η διαγραφή από τα γνωρίσματα των γνωρισμάτων KAD2/KAD3/KAD4/KAD5 δεν επιφέρει παρατηρούμενες μεταβολές, Άρα σε πρώτη ανάγνωση τα γνωρίσματα θα μπορούσαν να αγνοηθούν. Ωστόσο θα διατηρηθούν έως την υλοποίηση των μοντέλων ώστε να υπάρχει μια πιο ολοκληρωμένη εικόνα.

Ο επόμενος έλεγχος αφορούσε το πλήθος των διαθέσιμων στοιχείων. Τα στοιχεία που συλλέχθηκαν στην αρχική φάση αφορούσαν το σύνολο των επιχειρήσεων οι οποίες συμμετείχαν με οποιοδήποτε τρόπο σε ενδοκοινοτικές συναλλαγές.

Η προσέγγιση αυτή δημιούργησε ένα σύνολο περίπου 260000 περιπτώσεων. Προφανώς το σύνολο είναι ιδιαίτερα μεγάλο για επεξεργασθεί εντός των πλαισίων αυτής της εργασίας. Παράλληλα η ύπαρξη ενός τόσο εκτεταμένου συνόλου δημιουργεί πρόβλημα ανισοβαρούς κατανομής των κλάσεων. Οι χαρακτηρισμένες εγγραφές (κλάση = '1') αφορούν ένα πλήθος περίπου 900 περιπτώσεων, δίνοντας ένα λόγο ~0.35% μεταξύ των δύο κλάσεων.

Ένα σημαντικό 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 την συνολική ακρίβεια τους αλλά την υπερ-εκπαίδευση τους στην αναγνώριση της κλάσης πλειοψηφίας, πχ αν στην συγκεκριμένη περίπτωση και χωρίς οποιαδήποτε αλλαγή επιλέξουμε σαν αλγόριθμο τον zeroR¹ θα πετύχουμε ακρίβεια 99.9965%.

Η ανισορροπία εμφανίζεται έντονα στο παρακάτω διάγραμμα:



Εικόνα 34- Ανισορροπία κλάσεων

¹ Ο zeroR «προβλέπει» πάντα την κλάση πλειοψηφίας.

Η μεγάλη αυτή έλλειψη ισορροπίας μεταξύ των δύο κλάσεων δημιουργεί προβλήματα στους περισσότερους αλγόριθμους ταξινόμησης καθώς αυτοί προσπαθούν να περιορίσουν το συνολικό ποσοστό σφάλματος, παρά να δώσουν ειδική προσοχή στη θετική κλάση (Chao Chen et al, 2010).

Στη βιβλιογραφία (Barandela et al, 2004) (Baesens et al, 2015), (Wang-Japkowic, 2009) αναφέρονται διάφορες τεχνικές που προσπαθούν να αντιμετωπίσουν το πρόβλημα. Ορισμένες εφαρμόζονται απλούστερα, ενώ άλλες απαιτούν σύνθετες ενέργειες για την εφαρμογή τους.

Σύμφωνα με την παραπάνω βιβλιογραφία, οι εξής τεχνικές έχουν χρησιμοποιηθεί ή αναφερθεί ως πιθανές λύσεις στο πρόβλημα της άνισης κατανομής των κλάσεων:

- Ο υπερδειγματισμός της κλάσης μειοψηφίας. Πρακτικά τα δείγματα της μειοψηφίας λαμβάνονται υπόψη πολλές φορές, με αποτέλεσμα να αυξάνεται ο απόλυτος αριθμός τους.
- Η υπο-δειγματοληψία της κλάσης πλειοψηφίας. Η χρήση λιγότερων δειγμάτων από την κλάση πλειοψηφίας, είτε με τυχαία επιλογή είτε όχι.
- Η χρήση τεχνικών σύνθεσης δειγμάτων για την κλάση μειοψηφίας (πχ. SMOTE- Synthetic Minority over-sampling Technique).
- Η χρήση βαρών στις κλάσεις.
- Η χρήση αλγορίθμων οι οποίοι μπορούν να παραμετροποιηθούν ώστε να δείξουν εύνοια προς κάποια γνωρίσματα (πχ RIPPER, K-NN κλπ.)
- Η χρήση μεγαλύτερου, συνολικά, δείγματος.

Ως αρχική προσέγγιση στο πρόβλημα της ανισοκατανομής του δείγματος, προκρίθηκε ο υπο-δειγματισμός της κλάσης πλειοψηφίας. Τα αρχικά δεδομένα σχηματίστηκαν από το σύνολο των στοιχείων που αφορούν το επιχειρηματικό πεδίο των ενδοκοινοτικών συναλλαγών. Σε αυτά τα δεδομένα πολλά στοιχεία περικλείονται χωρίς να έχουν συναλλαγές.

Στο πεδίο που εξετάζουμε ο όρος του «εξαφανισμένου εμπόρου» έχει νόημα μόνο αν περιληφθούν φορολογικές οντότητες οι οποίες έχουν αποδεδειγμένα την ιδιότητα του «εμπόρου». Βάση αυτής της παραδοχής αποκλείστηκαν όλα τα στοιχεία που αφορούσαν φορολογικές οντότητες που δεν εμφανίζουν συναλλαγές.

	Αρχικό Δείγμα	Δείγμα μετά την διαγραφή των επιχειρήσεων χωρίς συναλλαγές	Δείγμα για τα έτη 2010-15	Δείγμα για τα έτη 2010-14 (Δεδομένα εκπαίδευσης)	Δείγμα για το έτος 2015 (Δεδομένα ελέγχου)
	266622	209073	45149	19975	25174
Κλάση 1	889	888	208	93	115
Κλάση 0	265733	208185	44941	19882	25059
Ποσοστό κλάσης 1 στο σύνολο των δειγμάτων (%)	0.33	0.43	0.46	0.47	0.46

Εικόνα 35- Πλήθος στοιχείων δείγματος

Ουσιαστικά με τον παραπάνω αποκλεισμό επιτεύχθηκε μείωση του δείγματος (υπο-δειγματισμός) σε περίπου 209000 γραμμές.

Για να γίνει πιο αντιπροσωπευτικό το δείγμα αποφασίστηκε να μειωθεί η χρονική περίοδος για την οποία θα συλλεχθούν στοιχεία, ώστε να περιλαμβάνει αυστηρά τα έτη 2010-2015. Η διερεύνηση των στοιχείων έδειξε ότι για προγενέστερα του 2010 υπήρχαν ελλιπή δεδομένα, οπότε και θα έπρεπε να γίνουν υποθέσεις για τις κενές τιμές ή να γίνουν χειροκίνητες διορθώσεις, ενέργειες οι οποίες θα περιέπλεκαν τα επόμενα βήματα.

Χρησιμοποιώντας τα παραπάνω κριτήρια τροποποιήθηκε το αρχικό δείγμα καθώς το συνολικό πλήθος των χαρακτηριστικών μειώθηκε από 266000 σε περίπου 45000.

Παράλληλα μειώθηκε και το δείγμα της κλάσης μειοψηφίας από 889 εγγραφές σε 208. Μετά την μείωση το ποσοστό της κλάσης μειοψηφίας διαμορφώθηκε σε $208 / 45000 \rightarrow \sim 0,462\%$, ελαφρώς βελτιωμένο σε σχέση με το αρχικό $\sim 0.35\%$.

Επιβεβαιώθηκε ότι η διαφορά των $889 - 208 \rightarrow 681$ δειγμάτων από την κλάση μειοψηφίας ανήκουν σε δεδομένα που αφορούν στοιχεία πριν το 2010.

Το ποσοστό της κλάσης μειοψηφίας θα μπορούσε να αυξηθεί εφόσον στο τελικό δείγμα συμπεριληφθούν τα επιπλέον 681 δείγματα τα οποία έχουν αποκλειστεί, χωρίς ταυτόχρονα να συμπεριληφθούν άλλα στοιχεία πριν το 2010. Μια τέτοια ενέργεια θα βελτίωνε το ποσοστό της κλάσης μειοψηφίας σε $889 / 45000 \rightarrow \sim 1,97\%$.

Στο κεφάλαιο «Εξόρυξη γνώσης με το εργαλείο Oracle Data Miner» θα εξετάσουμε αν τα αποκλεισμένα δείγματα μπορούν να συνεισφέρουν στην βελτίωση της απόδοσης.

Statistics Data Columns SQL						
Statistics: 10 Columns from 1.944 Rows(Sampled)						
Name	Histogram	Data Type	Percent NUL...	Distinct Valu...	Distinct Perc...	Mode
ABROADMS		VARCHAR2	0	61	3,1379	GR
AGEOFFIRM		NUMBER	0	1.287	66,2037	
ALT_MARKED		VARCHAR2	0	2	0,1029	0
BOOKS_KIND		VARCHAR2	0,0514	7	0,3603	30
CLASS		VARCHAR2	0	2	0,1029	0
COMPANY_TYPE		VARCHAR2	0	33	1,6975	0
CONNECTED		VARCHAR2	0	2	0,1029	0
CONNECTED_T		VARCHAR2	0	2	0,1029	0
DIFF		VARCHAR2	0	2	0,1029	0
FIRST_TIME		VARCHAR2	0,3601	3	0,1549	1
IMPORTS		VARCHAR2	0	2	0,1029	1
KAD1		VARCHAR2	0	288	14,8148	<Other>
ORIGIN		VARCHAR2	0	19	0,9774	1
PARTICIPATE		VARCHAR2	0	2	0,1029	1
PAR_BUSINESS		VARCHAR2	0,0514	605	31,1374	<Other>
QU_EXTRDIFF		VARCHAR2	0	2	0,1029	0
REQ_OPTION		VARCHAR2	0	2	0,1029	1
SMALLLIFETIME		VARCHAR2	0	2	0,1029	0
STATE		VARCHAR2	0,0514	11	0,5661	20
TRANSACTIONS		VARCHAR2	0	2	0,1029	1
VAT		VARCHAR2	0	1.944	100	<Other>
VAT_FLAG		VARCHAR2	0,0514	2	0,1029	1
VAT_OPTION		VARCHAR2	0	2	0,1029	1
VAT_STATUS		VARCHAR2	1,3889	11	0,5738	4
VIES_OPTION		VARCHAR2	0,0514	20	1,0293	22

Εικόνα 36- Προ-επισκόπηση στοιχείων που συμπεριλαμβάνουν εγγραφές τις κλάσης μειοψηφίας (<2010)

Συγκρίνοντας τα στατιστικά στοιχεία μεταξύ των τριών περιπτώσεων βρέθηκε ότι η διαγραφή των δραστηριοτήτων δεν επηρέασε τα στατιστικά των δεδομένων (εικ 37). Αντίθετα η διαγραφή των περιπτώσεων που δεν είχαν συναλλαγές πριν το 2010 επηρέασε ελαφρά τα στατιστικά των δεδομένων, μειώνοντας κυρίως το πλήθος των κενών (null) και το πλήθος των διακριτών τιμών. Οι παρατηρούμενες αλλαγές είναι αποδεκτές και εντός των αναμενόμενων αποτελεσμάτων.

3. Εξόρυξη γνώσης με το εργαλείο Oracle Data Miner

Δημιουργήθηκαν οι παρακάτω διαδικασίες εξόρυξης :

- Δημιουργία ταξινόμησης - classification

- Εντοπισμός ανωμαλίας - Anomaly detection
- Δημιουργία ομαδοποίησης – clustering

Καθώς το εργαλείο, στην έκδοση που ήταν διαθέσιμη δεν υποστήριζε αλγόριθμους όπως οι προαναφερόμενοι (Ripper, KNN), θα δοκιμασθούν τα μοντέλα που παρέχει με δύο σενάρια :

- Με το σύνολο των στοιχείων
- Με περιορισμό των στοιχείων (υποδειγματοληψία-undersampling).

Οι διαδικασίες που υποστηρίζει το εργαλείο είναι :

- Classification
- Anomaly Detection
- Association
- Clustering
- Regression

Από τα παραπάνω δεν θα εξετασθούν καθόλου τα :

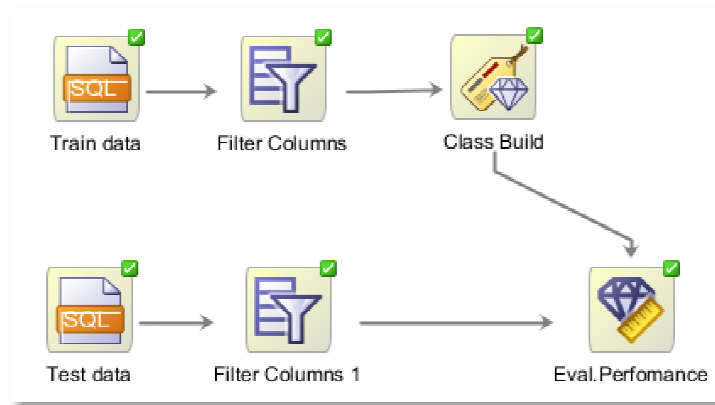
- Regression
- Association

Το πρώτο γιατί στην συγκεκριμένη υλοποίηση του δεν υποστηρίζει κατηγορικές τιμές (0,1), οι οποίες ήταν η συντριπτική πλειοψηφία για των γνωρισμάτων.

Το δεύτερο γιατί κατά την διαδικασία εκτέλεσης των υπολογισμών το μοντέλο δημιουργούσε συστημικά λάθη ώστε να είναι αδύνατη η ολοκλήρωση της διαδικασίας. Εκτιμάται ότι η συμπεριφορά αυτή του συστήματος οφείλεται στη χρήση των κατηγορικών γνωρισμάτων που χρησιμοποιούνται καθώς και στις διαθέσιμη υπολογιστική πλατφόρμα (μνήμη, heap size κλπ.).

Ταξινόμηση στοιχείων (classify)

Για την εκτέλεση της ταξινόμησης δημιουργήθηκε η παρακάτω ροή:



Εικόνα 37- Εκτέλεση ταξινόμησης(classification) με τον Data Miner

Το μοντέλο έτρεξε για δύο διαφορετικά δεδομένα εκπαίδευσης :

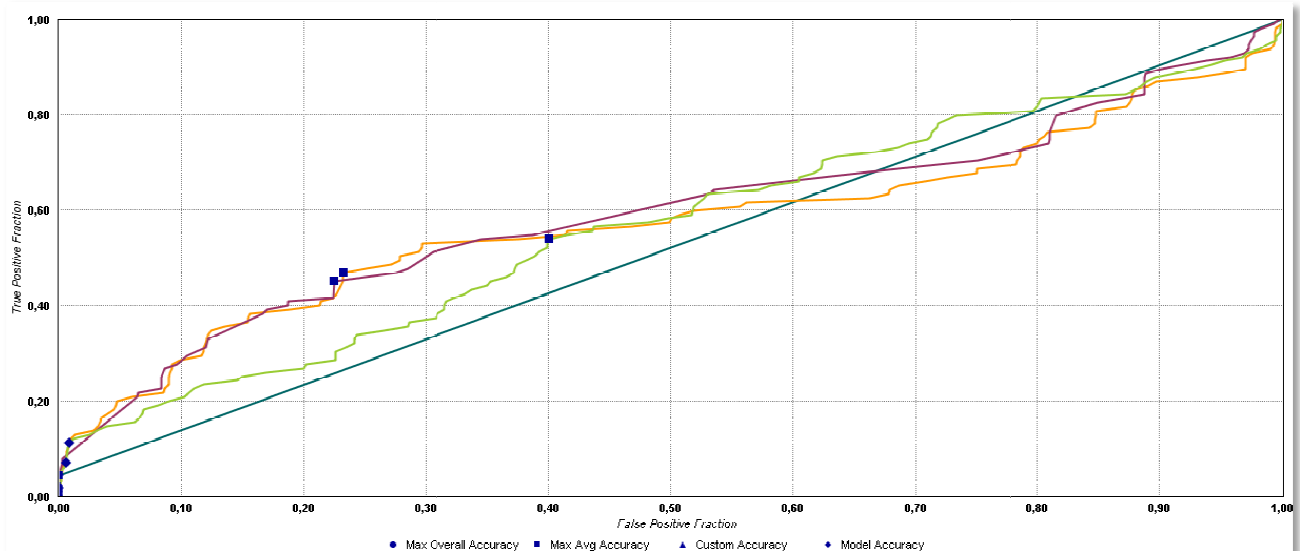
- Δεδομένα εκπαίδευσης από τα έτη 2010-2014
- Δεδομένα εκπαίδευσης από τα έτη 2010-2014 με επιπλέον εμπλουτισμό με τα στοιχεία της κλάσης μειοψηφίας πριν το 2010.

Τα αποτελέσματα των μοντέλων είναι :

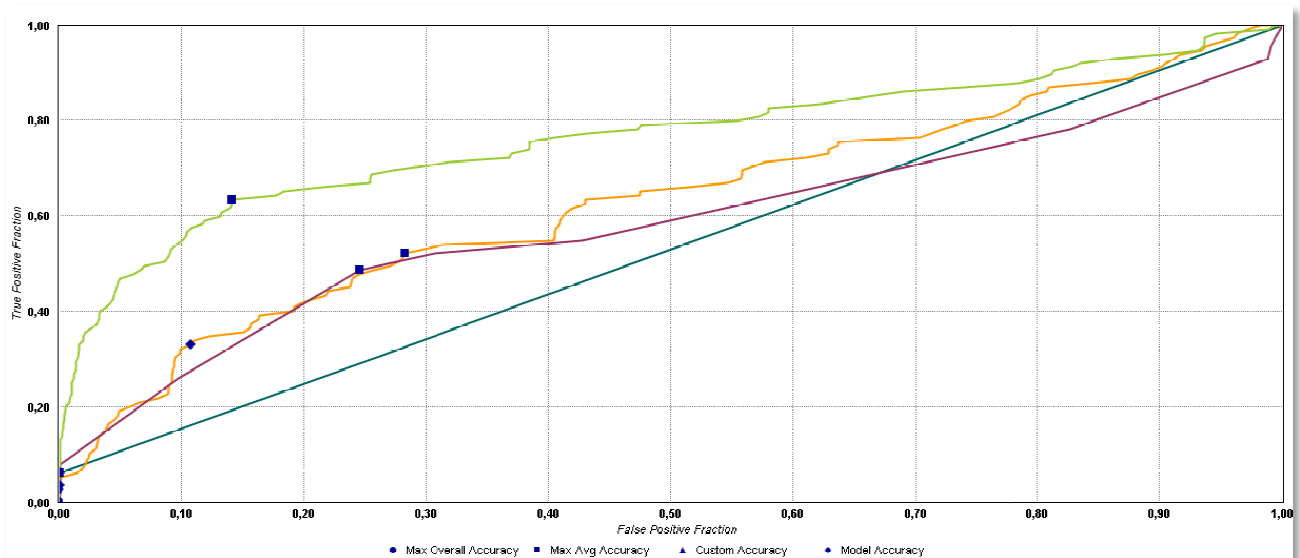
	Predictive Confidence		Overall Accuracy		Average Accuracy	
	2010-14	2010-14+	2010-14	2010-14+	2010-14	2010-14+
Naïve Bayes	5,9672	1,6912	99,4518	99,5511	52,9836	61,4071
SVM	2,6087	6,334	99,5551	99,5472	51,3043	56,9378
Generalized Linear Model	22,2773	10,4025	89,8387	99,5511	61,8396	61,8896
Decision Tree	6,0151	4,3079	99,4995	99,5432	53,0076	52,154

Εικόνα 38- Απόδοση μοντέλων ανάλογα με τα δεδομένα εκπαίδευσης

Τα αποτελέσματα τείνουν να είναι μια στατική απεικόνιση του μοντέλου. Για την καλύτερη εκτίμηση των αποτελεσμάτων θα χρησιμοποιηθούν οι καμπύλες ROC και τα χαρακτηριστικά που απορρέουν από αυτές.



Εικόνα 39- Καμπύλες ROC για δεδομένα εκπαίδευσης 2010-14+ στοιχεία κλάσης μειοψηφίας



Εικόνα 40- Καμπύλες ROC για δεδομένα εκπαίδευσης 2010-14

Η ερμηνεία των χρωμάτων είναι :

Ανοικτό πράσινο – SVM

Πορτοκαλί – GLM

Μωβ – Naïve Bayes

Σκούρο Πράσινο – Decision Trees

	Area Under Curve		Max Overall Accuracy %		Max average Accuracy %		Model Accuracy%	
	2010-14	2010-14+	2010-14	2010-14+	2010-14	2010-14+	2010-14	2010-14+
Naïve Bayes	0,5828	0,5899	99,5432	99,5511	62,0947	61,4071	99,4518	99,5511
SVM	0,764	0,5638	99,5551	99,5472	74,6898	56,9378	99,5551	98,9632
Generalized Linear Model	0,6272	0,5735	99,5432	99,5511	61,9643	61,8896	88,9727	98,6971
Decision Tree	0,5301	0,5215	99,5432	99,5432	53,0076	52,154	99,4876	99,5233

Εικόνα 41- Καμπύλες ROC, μετρικές

Σε αντίθεση με τα γενικά αποτελέσματα των μοντέλων που διαμορφώθηκε μια μικτή κατάσταση, δηλαδή η απόδοση κάποιων αλγόριθμων βελτιώθηκε και κάποιων χειροτέρευσε, με την χρήση της καμπύλης ROC, έχουμε μια πιο ολοκληρωμένη εικόνα της απόδοσης των αλγόριθμων.

Σύμφωνα με τα αποτελέσματα που προκύπτουν από τις καμπύλες ROC, σε όλες τις περιπτώσεις τα μοντέλα που δημιουργήθηκαν από το δείγμα εκπαίδευσης που περιείχε στοιχεία του 2010-14 απέδωσαν καλύτερα από τα μοντέλα που δημιουργήθηκαν στις περιπτώσεις που το δείγμα εκπαίδευσης περιείχε επιπλέον στοιχεία από την κλάση μειοψηφίας πριν από το 2010.

Στη μοναδική περίπτωση που το δεύτερο μοντέλο απέδωσε καλύτερα ήταν για τον αλγόριθμο “Decision Tree”.

Επίσης σημαντικό είναι το γεγονός ότι ο SVM σε κάθε περίπτωση παρουσιάζει την καλύτερη επίδοση μεταξύ των υπολοίπων. Η επίδοση του φαίνεται να προέρχεται από την απόδοση βαρών στα στοιχεία όπως φαίνεται από τον παρακάτω πίνακα :

Class	Weight
1	0,960287
0	0,039713

Εικόνα 42- Βαρύτητα των κλάσεων στον SVM (Data Miner)

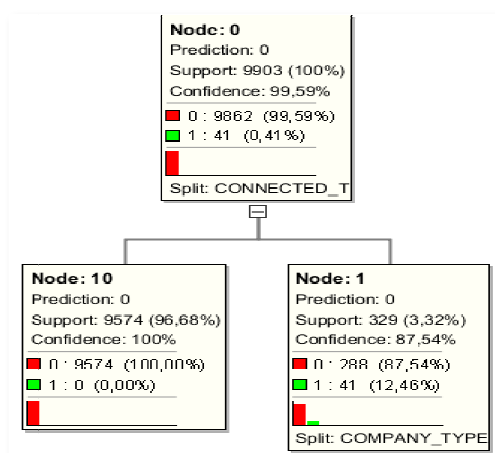
Από την εξέταση των γνωρισμάτων που χρησιμοποιεί ο αλγόριθμος παρουσιάζει ενδιαφέρον το γεγονός ότι δίνει μεγάλη βαρύτητα στο χαρακτηριστικό «Connected_T», δηλαδή αν υπάρχουν συναλλαγές με άλλη ύποπτη επιχείρηση.

Propensities: 20 out of 71		Attribute	
Attribute	Value	Propensity for 0	Propensity for 1
<Intercept>		4,00638224	-4,00638224
CONNECTED_T	0	1,30283751	1,30283751
CONNECTED_T	1	-1,30283751	1,30283751
STATE	40	-1,23899981	1,23899981
ABROADMS	GR	-1,06284604	1,06284604

Εικόνα 43- Γνωρίσματα που χρησιμοποιεί ο SVM (Data Miner)

Η εξέταση των γνωρισμάτων που χρησιμοποιούν οι υπόλοιποι αλγόριθμοι δείχνει ότι το ίδιο χαρακτηριστικό εντοπίζει και ο αλγόριθμος «Decision Tree».

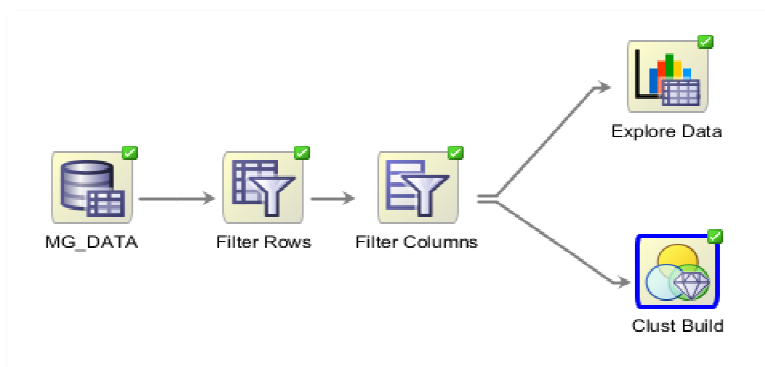
Καταλήγοντας όμως σε διαφορετικά αποτελέσματα και απόδοση.



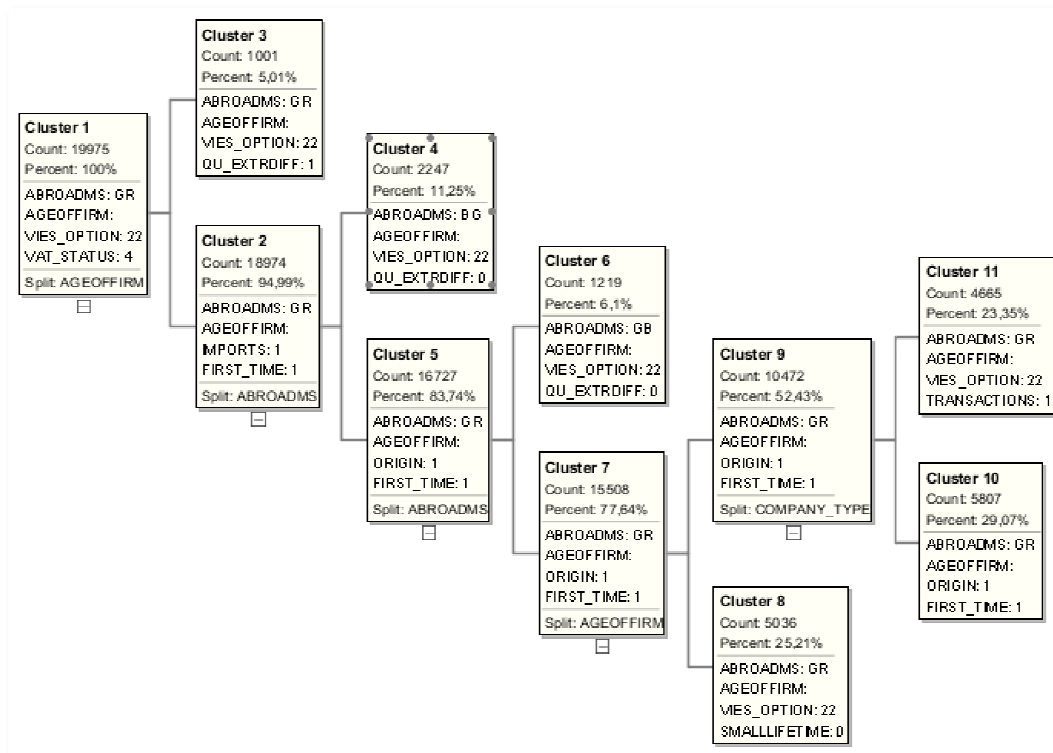
Εικόνα 44- Decision Tree (Data Miner)

Σχηματισμός ομάδων (clustering)

Για την εκτέλεση της ομαδοποίησης δημιουργήθηκε η παρακάτω ροή:



Εικόνα 45- Ομαδοποίηση (clustering - Data Miner)



Εικόνα 46- Ομαδοποίηση με τον O-Cluster (Data Miner)

Αρχικά δοκιμάζουμε τον αλγόριθμο O-Cluster. Σύμφωνα με τα αποτελέσματα του αλγόριθμου τα γνωρίσματα που επηρεάζουν το σχηματισμό ομάδων είναι : η χώρα καταγωγής του επιτηδευματία (AbroadMS), η ύπαρξη απότομων μεταβολών (Qu_ExtrDiff) μεταξύ των τριμήνων, οι διαφορές που προκύπτουν μεταξύ των στοιχείων που δηλώνει η επιχείρηση και των στοιχείων που δηλώνουν οι ξένες επιχειρήσεις (Transactions), η επιχείρηση έχει κάνει πρώτη φορά έναρξη εργασιών (first_time).

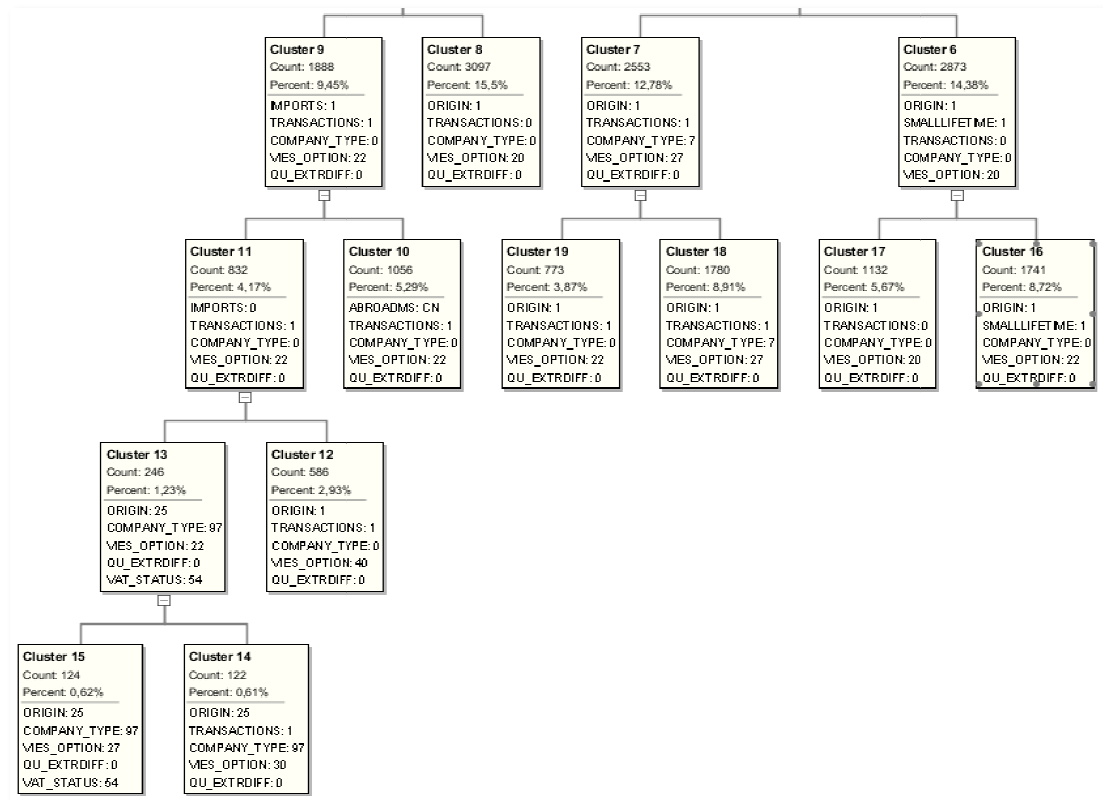
Αν θελήσουμε να δώσουμε τίτλους στις ομάδες που δημιουργούνται, για τις χαρακτηριστικότερες από αυτές θα έχουμε :

«Οι γείτονες» (11,25%): Επιχειρήσεις από ξένους (Βουλγαρία), που δραστηριοποιούνται αποκλειστικά σε εισαγωγές.

«Οικονομική διείσδυση» (6,1%) : Επιχειρήσεις από ξένους (Αγγλία), που δραστηριοποιούνται αποκλειστικά σε εισαγωγές.

«Ασταθείς» (23,35%): Επιχειρήσεις με τζίρους που παρουσιάζουν μεγάλη διακύμανση σε ορισμένα τρίμηνα.

«Προσωπικές» (25,21%) : Επιχειρήσεις του ενός προσώπου – ελεύθεροι επαγγελματίες



Εικόνα 47- Ομαδοποίηση με τον αλγόριθμο K-means (Data Miner)

Όμοια με προηγουμένως δοκιμάζουμε τον αλγόριθμο k-means. Γνωρίσματα που αναδεικνύει ο αλγόριθμος και λαμβάνει υπόψη για να δημιουργήσει τις ομάδες είναι: : η χώρα καταγωγής του επιτηδευματία (AbroadMS), η ύπαρξη απότομων μεταβολών (Qu_ExtraDiff), η διαφορές που προκύπτουν μεταξύ των στοιχείων που δηλώνει η επιχείρηση και των στοιχείων που δηλώνουν οι ξένες επιχειρήσεις (Transactions), η επιχείρηση έχει μικρή διάρκεια ζωής (smalllifetime).

Ορισμένες ομάδες που δημιουργούνται είναι :

«Ασταθείς» (47,88%) : Επιχειρήσεις με τζίρους που παρουσιάζουν μεγάλη διακύμανση σε ορισμένα τρίμηνα.

«Οι νεοεισερχόμενοι» (8,72%) : Νέες προσωπικές επιχειρήσεις, με διάρκεια ζωής μικρότερο του 1,5 έτους.

«Τσάινα-τάουν» (5,29%) : Επιχειρήσεις από ξένους (Κινέζοι).

4. Εξόρυξη γνώσης με το εργαλείο WEKA

Το εργαλείο WEKA αποτελείται από ένα σύνολο αλγόριθμων εξόρυξης οι οποίοι μπορούν να εκτελεσθούν είτε από την γραμμή εντολών, είτε μέσω του γραφικού περιβάλλοντος που παρέχεται. Η ύπαρξη αρκετών αλγόριθμων και το ενιαίο περιβάλλον που παρέχεται το καθιστούν ιδανικό εργαλείο για πειραματισμό και αξιολόγηση της απόδοσης των αλγορίθμων στο πεδίο της έρευνας που μας απασχολεί.

Θα πρέπει να αναφερθεί ότι πέρα από τους αλγόριθμους που παρέχονται μαζί με το εργαλείο υπάρχει ένα repository από επιπλέον αλγόριθμους οι οποίοι μπορούν να εγκατασταθούν όταν αυτό απαιτηθεί.

Για την εργασία δοκιμάστηκαν διάφοροι αλγόριθμοι και συνδυασμοί αυτών ώστε να αποκτήσουμε μια σφαιρική εικόνα της καταλληλότητας τους και της απόδοσης τους στον συγκεκριμένο τομέα.

Πριν την δημιουργία των μοντέλων εξόρυξης και της αξιολόγησης, χρειάστηκε να δημιουργηθούν τα αρχεία με τα στοιχεία που θα τροφοδοτούσαν το εργαλείο με τα δεδομένα εκπαίδευσης και ελέγχου. Αν και το εργαλείο παρέχει τη δυνατότητα σύνδεσης απευθείας σε βάσεις δεδομένων, μέσω του jdbc driver, η κατηγοριοποίηση των γνωρισμάτων ως αριθμητικά, κατηγορικά κλπ., γίνεται απευθείας από το εργαλείο χωρίς δυνατότητα επέμβασης από το χρήστη. Αυτή η συμπεριφορά στις συγκεκριμένες δοκιμές αλλοίωσε την έννοια των γνωρισμάτων χαρακτηρίζοντας κατηγορικές τιμές ως αριθμητικές (πχ κωδικοί δραστηριότητας). Επίσης δυσκολία προκαλούσε και το γεγονός ότι το γραφικό περιβάλλον δεν παρείχε κάποιο είδος ρυθμίσεων ώστε ο χρήστης να μπορεί να επαναχρησιμοποιήσει ορισμένες παραμέτρους με προκαθορισμένες τιμές (πχ τη σύνδεση με τη βάση).

Εξαιτίας των παραπάνω δυσκολιών τα δεδομένα έγινα export από τη βάση και δημιουργήθηκαν τα δύο σετ δεδομένων. Λεπτομερή αναφορά στους λόγους και στην διαδικασία γίνεται στο παράρτημα 4.

Ταξινόμηση στοιχείων (classify)

Τα σετ που δημιουργήθηκαν αφορούσαν δύο διαφορετικές περιόδους, όπως και στην προηγούμενη περίπτωση. Το σετ εκπαίδευσης περιέχει στοιχεία από τα έτη 2010-14 και το σετ ελέγχου από τα έτη 2014-15. Το σύνολο των αποτελεσμάτων παρατίθεται στον πίνακα του παραρτήματος 5. Εδώ αναφέρονται ενδεικτικά ορισμένα από αυτά:

	Correctly	Incorrectly	TN	FN	FP	TP	1-ROC	Precision	Recall
Train	99.5745	0.4255	19873	9	76	17	0.591	0.654	0.183
Test	99.5114	0.4886	25050	9	114	1	0.504	0.1	0.009

Εικόνα 48- Απόδοση αλγόριθμου oneR (WEKA)

	Correctly	Incorrectly	TN	FN	FP	TP	1-ROC	Precision	Recall
Train	99.9549	0.0451	19882	0	9	84	0.998	1	0.903
Test	99.5432	0.4568	25059	0	115	0	0.539	0	0

Εικόνα 49- Απόδοση αλγόριθμου J-48 (WEKA)

	Correctly	Incorrectly	TN	FN	FP	TP	1-ROC	Precision	Recall
Train	99.6546	0.3454	19877	5	64	29	0.84	0.853	0.312
Test	99.5432	0.4568	25058	1	114	1	0.513	0.5	0.009

Εικόνα 50- Απόδοση αλγόριθμου Random Tree (WEKA)

	Correctly	Incorrectly	TN	FN	FP	TP	1-ROC	Precision	Recall
Train	99,7146	0,2854	19987	3	54	39	0,988	0,929	0,419
Test	99,5273	0,4727	25055	4	115	0	0,621	0	0

Εικόνα 51- Απόδοση αλγόριθμου Random Forest (WEKA)

	Correctly	Incorrectly	TN	FN	FP	TP	1-ROC	Precision	Recall
Train	99.6496	0.3504	19841	41	29	64	0.966	0.61	0.688
Test	99.3883	0.6117	25013	46	108	7	0.574	0.132	0.061

Εικόνα 52- Απόδοση αλγόριθμου Naive Bayes (WEKA)

	Correctly	Incorrectly	TN	FN	FP	TP	1-ROC	Precision	Recall
Train	99.7096	0.2904	19867	15	43	50	0.876	0.769	0.538
Test	99.428	0.572	25025	34	110	5	0.713	0.128	0.043

Εικόνα 53- Απόδοση αλγόριθμου KStar (WEKA)

	Correctly	Incorrectly	TN	FN	FP	TP	1-ROC	Precision	Recall
Train	99.9499	0.0501	19881	1	9	81	0.952	0.998	0.903
Test	99.5432	0.4568	25059	0	115	0	0.5	0	0

Εικόνα 54- Απόδοση αλγόριθμου SMO (WEKA)

Οι μετρικές Correctly (Classified Instances), Incorrectly (Classified Instances), TN, FN, FP, TP αναφέρονται στην συνολική απόδοση του μοντέλου. Οι μετρικές 1-ROC, Precision, Recall αναφέρονται στην απόδοση του μοντέλου για την κλάση μειοψηφίας (κλάση=1).

Όπως παρατηρούμε τα παραπάνω μοντέλα ταξινόμησης αντιμετωπίζουν όλα τα ίδια προβλήματα, χαμηλό ποσοστό επιτυχίας και χαμηλό ή μηδενικό ποσοστό επαναληψιμότητας (recall), για την κλάση που μας ενδιαφέρει (κλάση=1), ενώ το ποσοστό επιτυχίας είναι ιδιαίτερα υψηλό (> 99%).

Η συμπεριφορά αυτή είναι τυπική ενός μοντέλου στο οποίο οι κλάσεις είναι σοβαρά ανισοβαρείς.

Για να αναιρεθεί κατά το δυνατόν η επήρεια της μηδενικής κλάσης στο τελικό αποτέλεσμα εφαρμόστηκαν βάρη στα αποτελέσματα ορισμένων αλγόριθμων. Με την απόδοση βαρών θα επιβαρυνθούν τα λάθος αποτελέσματα (τα FP) της κλάσης μειοψηφίας (κλάση = 1).

	0	1
0	0.0	1.0
1	10.0	0

Εικόνα 55- Χρήση βαρών (WEKA)

Μετά την χρήση βαρών τα αποτελέσματα είναι :

	Correctly	Incorrectly	TN	FN	FP	TP	1-ROC	Precision	Recall
Train	99,9299	0,0701	19879	3	11	82	0,941	0,965	0,882
Test	99,5432	0,4568	25059	0	115	0	0,5	0	0

Εικόνα 56- Αλγόριθμος SMO με βάρη (WEKA)

	Correctly	Incorrectly	TN	FN	FP	TP	1-ROC	Precision	Recall
Train	97,0713	2,9287	19297	585	0	93	0,985	0,137	1
Test	99,4598	0,5402	25029	30	106	9	0,539	0,231	0,078

Εικόνα 57- Αλγόριθμος oneR με βάρη (WEKA)

	Correctly	Incorrectly	TN	FN	FP	TP	1-ROC	Precision	Recall
Train	99,8298	0,1702	19863	19	15	78	0,912	0,804	0,839
Test	99,4478	0,5522	25031	28	111	4	0,517	0,125	0,035

Εικόνα 58- Αλγόριθμος Ibk με βάρη (k=3) (WEKA)

	Correctly	Incorrectly	TN	FN	FP	TP	1-ROC	Precision	Recall
Train	97,0713	2,29287	19297	585	0	93	0,984	0,137	1
Test	99,4598	0,5402	25029	30	106	9	0,539	0,231	0,078

Εικόνα 59- Αλγόριθμος Rep Tree με βάρη (WEKA)

Μετά την χρήση των βαρών οι αλγόριθμοι εμφάνισαν ελαφρά καλύτερα αποτελέσματα στον έλεγχο των αποτελεσμάτων με τα δεδομένα ελέγχου, ωστόσο και πάλι τα ποσοστά επιτυχίας ήταν ιδιαίτερα χαμηλά.

Σε επόμενο στάδιο δοκιμάστηκαν οι αλγόριθμοι oneR και REPTree με την μέθοδο bagging

	Correctly	Incorrectly	TN	FN	FP	TP	1-ROC	Precision	Recall
Train	99,5845	0,4155	19875	7	76	17	0,613	0,708	0,183
Test	99,5273	0,4727	25055	4	115	0	0,508	0	0

Εικόνα 60- Αλγόριθμος oneR (Bagging - WEKA)

	Correctly	Incorrectly	TN	FN	FP	TP	1-ROC	Precision	Recall
Train	99,5645	0,4355	19871	11	76	17	0,806	0,607	0,183
Test	99,5352	0,4646	25057	2	115	0	0,738	0	0

Εικόνα 61- Αλγόριθμος Rep Tree (Bagging - WEKA)

Γενικά παρατηρούμε ότι όλα τα μοντέλα τα παρουσιάζουν καλύτερα ποσοστά επιτυχίας κατά την εκπαίδευση παρά κατά τον έλεγχο. Το γεγονός αυτό είναι αναμενόμενο, ωστόσο σε όλες τις περιπτώσεις (bagging oneR, CostSensitive Ibk κλπ) η διαφορά που προκύπτει είναι ιδιαίτερα μεγάλη.

Αιτίες για αυτή τη συμπεριφορά των μοντέλων μπορεί να είναι ορισμένα από τα παρακάτω ή συνδυασμός αυτών :

- Χρήση λάθος γνωρισμάτων
- Ανισοβαρείς κλάσεις
- Λάθος κατανομή εκπαιδευτικών δεδομένων και δεδομένων ελέγχου

Θα προσπαθήσουμε να ελέγξουμε κατά πόσο οι παραπάνω υποθέσεις, για την ποιότητα των δεδομένων, επηρεάζουν το αποτέλεσμα μας.

Χρήση λάθος γνωρισμάτων

Θα ελέγξουμε κατά πόσο τα γνωρίσματα που χρησιμοποιήθηκαν σχετίζονται με τα την κλάση που ελέγχουμε. Για τον έλεγχο θα χρησιμοποιήσουμε την δυνατότητα του WEKA, να ταξινομεί τα γνωρίσματα (attributes) σε σχέση με την κλάση.

Με τη χρήση του αλγόριθμου 'InfoGainAttributeEvaluator' προκύπτει η παρακάτω λίστα ταξινόμησης των γνωρισμάτων.

average merit	average rank	attribute
0.026 +- 0	1 +- 0	1 KAD1
0.023 +- 0	2 +- 0	23 CONNECTED_T
0.016 +- 0.001	3.2 +- 0.4	2 KAD2
0.015 +- 0	3.8 +- 0.4	17 PAR_BUSINESS
0.011 +- 0	5.2 +- 0.4	18 ABROADMS
0.01 +- 0	5.8 +- 0.4	22 IMPORTS
0.008 +- 0	7 +- 0	3 KAD3
0.005 +- 0	8.1 +- 0.3	4 KAD4
0.004 +- 0	9.1 +- 0.54	19 QU_EXTRDIFF
0.004 +- 0	10.2 +- 0.75	7 AGEOFFIRM
0.004 +- 0	10.6 +- 0.49	6 COMPANY_TYPE
0.003 +- 0	12 +- 0	5 KAD5
0.003 +- 0	13 +- 0	12 TRANSACTIONS
0.002 +- 0	14 +- 0	15 VIES_OPTION
0.001 +- 0	15 +- 0	25 REQ_OPTION
0.001 +- 0	16.1 +- 0.3	21 ORIGIN
0 +- 0	16.9 +- 0.3	16 BOOKS_KIND
0 +- 0	18.4 +- 0.66	9 DIFF
0 +- 0	19.9 +- 1.37	11 CONNECTED

Εικόνα 62- Ιεράρχηση γνωρισμάτων (InfoGainAttributeEval, WEKA)

Σύμφωνα με τα αποτελέσματα δεν έχουν όλα τα γνωρίσματα συμμετοχή στην αξιολόγηση των αποτελεσμάτων.

Με βάση την παρατήρηση αυτή δημιουργήθηκαν δύο νέα σετ δεδομένων τα οποία περιέχουν μόνο τα γνωρίσματα τα οποία έχουν συμμετοχή (average merit) $\geq 0,01$, δηλαδή τα :

average merit	average rank	attribute
0.026 +- 0	1 +- 0	1 KAD1
0.023 +- 0	2 +- 0	23 CONNECTED_T
0.016 +- 0.001	3.2 +- 0.4	2 KAD2
0.015 +- 0	3.8 +- 0.4	17 PAR_BUSINESS
0.011 +- 0	5.2 +- 0.4	18 ABROADMS
0.01 +- 0	5.8 +- 0.4	22 IMPORTS

Εικόνα 63- Γνωρίσματα που συνδέονται με την κλάση (InfoGainAttributeEval, WEKA)

Χρησιμοποιώντας τα δύο νέα σετ δεδομένων δημιουργήθηκαν και αξιολογήθηκαν τα παρακάτω μοντέλα :

	Correctly	Incorrectly	TN	FN	FP	TP	1-ROC	Precision	Recall
Train	99,7447	0,2553	19845	37	14	79	0,89	0,681	0,846
Test	99,4995	0,5005	25046	13	113	2	0,508	0,133	0,017

Εικόνα 64- Αλγόριθμος JRip (WEKA)

	Correctly	Incorrectly	TN	FN	FP	TP	1-ROC	Precision	Recall
Train	99,7397	0,2603	19878	4	48	45	0,97	0,918	0,484
Test	99,5432	0,4568	25058	1	114	1	0,73	0,5	0,009

Εικόνα 65- Αλγόριθμος KStar (WEKA)

	Correctly	Incorrectly	TN	FN	FP	TP	1-ROC	Precision	Recall
Train	99,7347	0,2653	19865	17	36	57	0,988	0,77	0,613
Test	99,5472	0,4528	25059	0	114	1	0,57	1	0,009

Εικόνα 66- Αλγόριθμος Naive Bayes (WEKA)

	Correctly	Incorrectly	TN	FN	FP	TP	1-ROC	Precision	Recall
Train	99,6095	0,3905	19840	42	36	57	0,962	0,576	0,613
Test	99,4915	0,5085	25042	17	111	4	0,561	0,19	0,035

Εικόνα 67- Αλγόριθμος Naive Bayes (CostBased, WEKA)

	Correctly	Incorrectly	TN	FN	FP	TP	1-ROC	Precision	Recall
Train	99,8048	0,1952	19871	11	28	65	0,919	0,855	0,699
Test	99,5432	0,4568	25058	1	114	1	0,57	0,5	0,009

Εικόνα 68- Stacking JRip, Naive Bayes, J-48 με αλγόριθμο επιλογής SimpleLogistic (WEKA)

Τα μοντέλα συνεχίζουν να παρουσιάζουν την ίδια συμπεριφορά, καλά αποτελέσματα εκπαίδευσης, αλλά ελλιπή αποτελεσματικότητα στα δεδομένα ελέγχου.

Ανισοβαρείς κλάσεις

Σύμφωνα με την βιβλιογραφία που εξετάσαμε νωρίτερα (Chao Chen et al, 2010), (Barandela et al, 2004) (Baesens et al, 2015), (Wang-Japkowic, 2009), οι ανισοβαρείς κλάσεις είναι ένα τυπικό πρόβλημα. Εμφανίζεται σε εκείνες τις περιπτώσεις στις οποίες η επιβεβαίωση του αποτελέσματος είναι δύσκολο να γίνει

(πχ περιπτώσεις απάτης, ανάλυση εικόνας για εντοπισμό αντικειμένου κλπ.).

Στη συγκεκριμένη εργασία τα στοιχεία που χρησιμοποιήθηκαν περιέχουν ένα ποσοστό περίπου 0,5% στοιχεία της κλάσης που μας ενδιαφέρει.

Πιθανή βελτίωση των στοιχείων θα υπήρχε αν οι κλάσεις είχαν μεγαλύτερη σαφήνεια. Η πληροφορία που περιέχουν οι σημερινές κλάσεις είναι :

Κλάση '1' – Χαρακτηρισμένες εγγραφές

Κλάση '0' – Εγγραφές για τις οποίες δεν έχουμε πληροφορίες ή μη χαρακτηρισμένες εγγραφές

Αν ήταν δυνατός ο διαχωρισμός μεταξύ των εγγραφών για τις οποίες δεν έχουμε πληροφορίες και των μη χαρακτηρισμένων θα υπήρχαν ποιοτικότερα δεδομένα προς ανάλυση και οι κλάσεις θα ήταν σαφώς καλύτερα καταναεμημένες στο δείγμα.

Λάθος κατανομή εκπαιδευτικών δεδομένων και δεδομένων ελέγχου

Για την έλεγχο των στοιχείων αντιστράφηκε ο ρόλος των δεδομένων εκπαίδευσης και ελέγχου. Τα αρχικά δεδομένα ελέγχου χρησιμοποιήθηκαν ως δεδομένα εκπαίδευσης και τα δεδομένα εκπαίδευσης ως δεδομένα ελέγχου. Η υπόθεση που θέλουμε να ελέγξουμε είναι αν τα νεότερα δεδομένα είναι ποιοτικά καλύτερα από τα δεδομένα των παλαιότερων ετών, άρα τα μοντέλα που δημιουργούν θα εμφανίσουν καλύτερη απόδοση από τα προηγούμενα.

Μετά την αντιστροφή τα αποτελέσματα ήταν :

	Correctly	Incorrectly	TN	FN	FP	TP	1-ROC	Precision	Recall
Train	99,2373	0,7627	24961	98	94	21	0,62	0,176	0,183
Test	98,0175	1,9825	19555	327	69	24	0,629	0,068	0,258

Εικόνα 69- Αλγόριθμος Ibk (CostBased, WEKA)

	Correctly	Incorrectly	TN	FN	FP	TP	1-ROC	Precision	Recall
Train	99,5948	0,4052	25048	11	91	24	0,616	0,686	0,209
Test	97,3967	2,6033	19430	452	68	25	0,624	0,052	0,269

Εικόνα 70- Αλγόριθμος JRip (WEKA)

	Correctly	Incorrectly	TN	FN	FP	TP	1-ROC	Precision	Recall
Train	99,3485	0,6515	24992	67	97	18	0,802	0,212	0,157
Test	97,2816	2,7148	19357	525	18	75	0,962	0,125	0,806

Εικόνα 71- Αλγόριθμος Naive Bayes (WEKA)

Τα αποτελέσματα που προκύπτουν από τα δεδομένα ελέγχου όταν εφαρμοσθούν στα δεδομένα εκπαίδευσης, παρουσιάζουν μικρότερες αποκλίσεις ή σε κάποιες περιπτώσεις και μικρή βελτίωση σε σχέση με τα μοντέλα που εξετάσαμε έως τώρα.

Σχηματισμός ομάδων (clustering)

Εφαρμόζοντας τους αλγόριθμους αναζήτησης ομάδων, στα δεδομένα σχηματίζονται ανάλογα με τον αλγόριθμο οι παρακάτω ομάδες:

Attribute	Full Data (19975)	0 (8073)	1 (2905)	2 (6041)	3 (463)	4 (374)	5 (1956)	6 (163)
KAD1	9999999	4669	4752	9999999	4631	4669	9999999	6810
KAD2	4771	4669	4673	4771	4631	4772	4771	4774
KAD3	4771	4642	4771	4771	4771	4771	4771	3511
PAR_BUSINESS	54627	85100	15125	54627	19300	85100	54625	10681
ABROADMS	GR	GR	GR	GR	GR	GR	GR	GR
IMPORTS	1	1	1	1	1	1	0	1
CONNECTED_T	0	0	0	0	1	0	0	0

Εικόνα 72- Σχηματισμός ομάδων (WEKA)

Οι παραπάνω ομάδες δημιουργήθηκαν με τον αλγόριθμο K-means και περιορισμό τον σχηματισμό 7 ομάδων. Ως δεδομένα εκπαίδευσης χρησιμοποιήθηκε το αρχικό σετ με, μειωμένο στα γνωρίσματα που αναδείχθηκαν μέσω της επιλογής γνωρισμάτων (select attribute) του εργαλείου.

Πρόσθετη απαίτηση είναι να γίνει αποτίμηση των ομάδων σε σχέση με τις κλάσεις των δεδομένων. Σύμφωνα με την αποτίμηση του εργαλείου οι κλάσεις κατανέμονται στις ομάδες ως εξής :

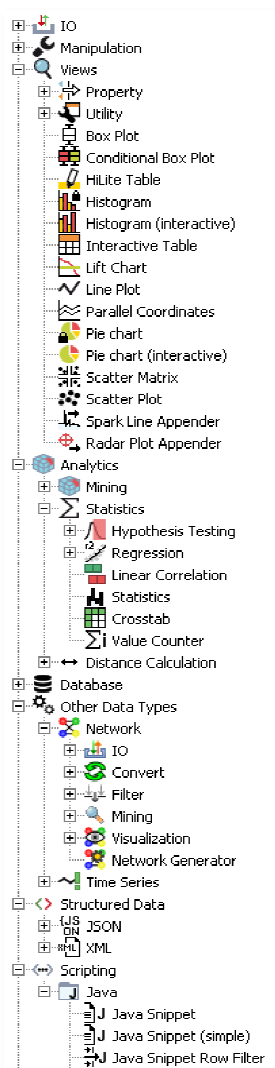
0	1	2	3	4	5	6	<-- assigned to cluster
8067	2904	6036	406	374	1932	163	0
6	1	5	57	0	24	0	1

Εικόνα 73- Συσχέτιση ομάδων με την κλάση (WEKA)

Όπως φαίνεται από τις ομάδες που σχηματίζονται η πλειοψηφία των χαρακτηρισμένων εγγραφών ανήκουν στις ομάδες 3 και 5. Ενδιαφέρον παρουσιάζουν τα γνωρίσματα που συμμετέχουν στις ομάδες 3 και 5. Στην ομάδα 3 ξεχωρίζουν τα χαρακτηριστικά 'imports' και 'Connected_T', δηλαδή ότι αφορούν επιχειρήσεις οι οποίες έχουν μόνο εισαγωγές και έχουν σχέσεις με ύποπτες επιχειρήσεις. Όμοια στην ομάδα 5 ξεχωρίζει ο κωδικός δραστηριότητας '99999999', που αντιστοιχεί σε επιχειρήσεις οι οποίες δεν έχουν προβεί σε ενημέρωση του κωδικού, δηλαδή επιχειρήσεις που δεν έχουν ενημερώσει το μητρώο τους τα τελευταία χρόνια.

5. Εξόρυξη γνώσης με τα εργαλεία Knime - Cytoscape

Το εργαλείο Knime είναι μια ολοκληρωμένη πλατφόρμα ανάλυσης και εξόρυξης. Παρέχει όλα τα εργαλεία



που είναι απαραίτητα ώστε να εκτελεστούν όλα τα βήματα μιας διαδικασίας εξόρυξης. Περιλαμβάνει εργαλεία σύνδεσης σε πηγές δεδομένων, απεικόνισης γραφημάτων-πινάκων, εργαλεία ανάλυσης, στατιστικών, δημιουργίας δικτύων κλπ.

Ανάμεσα στις δυνατότητες που υποστηρίζει η πλατφόρμα είναι η ανάπτυξη ροών, η δημιουργία script με java, η ενσωμάτωση (Plugin) της γλώσσας R, η ενσωμάτωση των αλγόριθμων που έχουν αναπτυχθεί για το WEKA.

Γενικά είναι ένα πολύ-εργαλείο το οποίο καλύπτει το σύνολο των ενεργειών που απαιτείται για την εξόρυξη.

Από τις δυνατότητες του εργαλείου στην παρούσα εργασία θα χρησιμοποιήσουμε μόνο την δυνατότητα του να δημιουργεί δίκτυα αν τροφοδοτηθεί με τις πληροφορίες των ακμών και κόμβων. Στο παρόν στάδιο δεν θα ασχοληθούμε με τις υπόλοιπες δυνατότητες ανάλυσης καθώς αυτές έχουν ήδη με τα προηγούμενα εργαλεία.

Το δίκτυο που θα δημιουργηθεί θα κατευθυνθεί προς το εργαλείο Cytoscape, με το οποίο θα χειρισθούμε το γράφο.

Για την δημιουργία του δικτύου χρησιμοποιήθηκε ένα διαφορετικό σετ δεδομένων από τα προηγούμενα, χρησιμοποιώντας όμως τα στοιχεία που αυτά περιείχαν ως σημείο αφετηρίας.

Εικόνα 74- Διαθέσιμα εργαλεία (Knime)

Για το σχηματισμό του δικτύου ερευνήθηκαν δύο πιθανές διαδικασίες :

- Χρήση όλων των διαθέσιμων στοιχείων συναλλαγών για την περίοδο 2010-15.
- Χρήση περιορισμένου εύρους στοιχείων συναλλαγών

Στην πρώτη περίπτωση χρησιμοποιούνται όλα τα διαθέσιμα στοιχεία συναλλαγών, ώστε να δημιουργηθεί το πλήρες δίκτυο των συναλλαγών, άρα και των σχέσεων μεταξύ των επιχειρήσεων. Στη συνέχεια

βαθμονομώντας τις σχέσεις (υπολογίζοντας το βάρος των ακμών), να αναδειχθούν εκείνα τα υπο-δίκτυα που παρουσιάζουν μεγαλύτερο ενδιαφέρον.

Οι βαθμολογήσεις μπορεί να περιλαμβάνει στοιχεία/γνωρίσματα που εξετάσαμε στις προηγούμενες παραγράφους, όπως ο κωδικός δραστηριότητας, το ύψος της συναλλαγής, η συχνότητα συναλλαγής, η απότομη μεταβολή του ποσού συναλλαγής κλπ.

Στην περίπτωση αυτή ουσιαστικά έχουμε μια ανάδυση των σημαντικότερων υπο-δικτύων μέσα από την αρχική «χαώδη» κατάσταση.

Στην δεύτερη περίπτωση για την υλοποίηση του δικτύου χρησιμοποιούνται τα ελάχιστα δυνατά στοιχεία, ενώ το δίκτυο σχηματίζεται και διευρύνεται με την προσθήκη επιπλέον στοιχείων στην πάροδο του χρόνου.

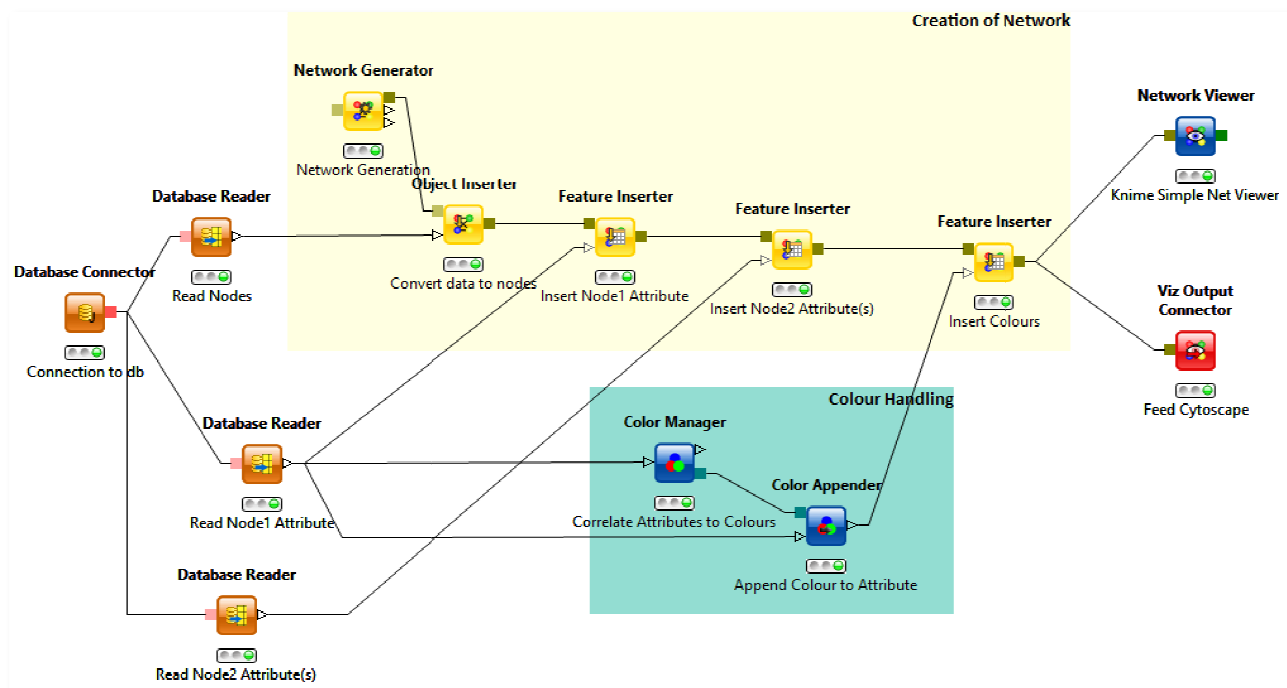
Για την εργασία προτιμήθηκε η δεύτερη επιλογή καθώς η πρώτη απαιτούσε μεγάλο χρόνο επεξεργασίας και υπολογιστική ισχύ, ενώ παράλληλα τα κριτήρια με βάση τα οποία θα αναδειχθούν τα υποδίκτυα χρειάζονται μεγαλύτερη διερεύνηση.

Για την άντληση των αρχικών δεδομένων και τη δημιουργία του δικτύου χρησιμοποιήθηκαν, από τα υπάρχοντα στοιχεία, οι εγγραφές για τις οποίες υπήρχε χαρακτηρισμός (κλάση = 1). Για τις εγγραφές αυτές εντοπίστηκαν όλες οι συναλλαγές που είχαν στο διάστημα 2010-15.

Δημιουργήθηκε έτσι ένα σύνολο από κόμβους οι οποίοι συνδέονται μεταξύ τους με κριτήριο την ύπαρξη εμπορικής συναλλαγής.

Ο αρχικός εντοπισμός έδωσε περίπου 1290 επιχειρήσεις/κόμβους με 1445 συναλλαγές/ακμές.

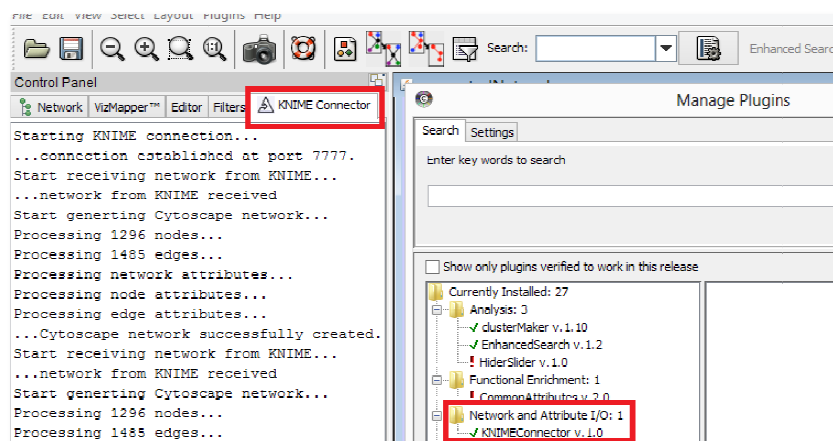
Για να απεικονίσουμε το δίκτυο χρησιμοποιήθηκαν τα προγράμματα Knime και Cytoscape σε συνεργασία. Αρχικά δημιουργήθηκε το παρακάτω διάγραμμα ροής στο Knime :



Εικόνα 75- Σχηματισμός δικτύου από το πρόγραμμα Knime

Το διάγραμμα είναι ιδιαίτερα απλό, μοναδική λειτουργία του είναι η μετατροπή των συσχετίσεων των επιχειρήσεων σε μορφή αναγνώσιμη από το εργαλείο Cytoscape (κόκκινος κόμβος στο διάγραμμα).

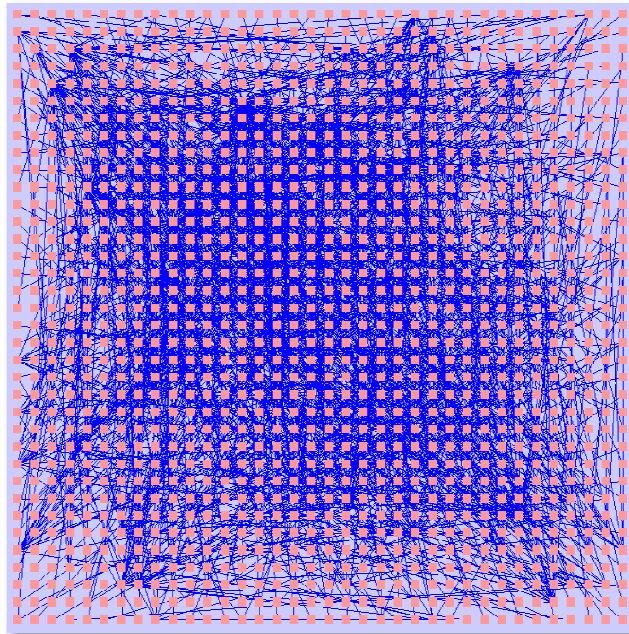
Η ανταλλαγή των στοιχείων γίνεται σε πραγματικό χρόνο απευθείας μέσω ενός plugin που φορτώνεται στο cytoscape.



Εικόνα 76- Επικοινωνία Knime - Cytoscape

Με την δημιουργία του δικτύου στο Knime μεταφέρονται σε πραγματικό χρόνο τα στοιχεία και στο πρόγραμμα Cytoscape.

Το δίκτυο που σχηματίζεται μοιάζει αρχικά με το παρακάτω :



Εικόνα 77- Αρχική απεικόνιση δικτύου συναλλαγών (Cytoscape)

Εκμεταλλευόμενοι τις δυνατότητες του εργαλείου απεικόνισης δικτύων και εφαρμόζοντας τους παρακάτω μετασχηματισμούς του δικτύου :

- Layout → Cytoscape Layout → Edge Weighted force Directed(BioLayout) → All Nodes (για να σχηματισθούν τα ξεχωριστά δίκτυα)
- VizMapper → Node Color → Mapping Type → Discrete Mapping με τις παρακάτω τιμές :

Node Color	MissingMS
Mapping Type	Discrete Mapping
102	
105	
107	
108	
109	
240	
350	
370	
801	

Εικόνα 78- Χρωματισμός κόμβων (Cytoscape)

- VizMapper → Node Label :

☒ Node Label	1
Mapping Type	Discrete Mapping
1	MISSING

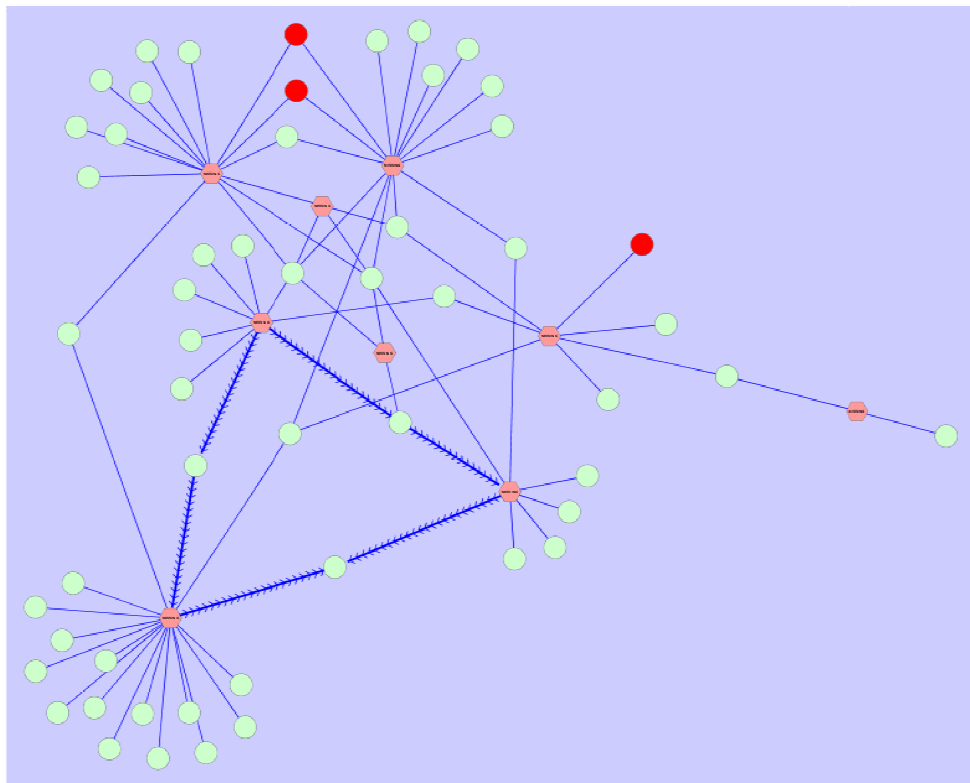
Εικόνα 79- Καθορισμός ετικέτας κόμβου (Cytoscape)

- VizMapper → Node Shape :

☒ Node Shape	1
Mapping Type	Discrete Mapping
1	 HEXAGON

Εικόνα 80- Καθορισμός σχήματος κόμβου (Cytoscape)

Πετυχαίνουμε να μετασχηματίσουμε το αρχικό άμορφο δίκτυο σε επιμέρους δίκτυα με πιο κατανοητή μορφή.



Εικόνα 81- Δίκτυο σχέσεων επιχειρήσεων (Cytoscape)

Για την επεξήγηση του δικτύου θα ανακαλέσουμε το σχήμα στην εικόνα 18.

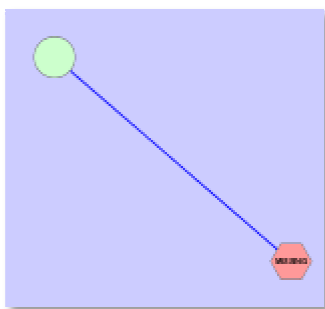
Με ροζ χρώμα εμφανίζονται οι επιχειρήσεις με κλάση '1', οι οποίες είναι οι Ελληνικές επιχειρήσεις με χαρακτηρισμό αφανής έμπορος (Missing trader).

Με κόκκινο χρώμα εμφανίζονται οι ξένες επιχειρήσεις με χαρακτηρισμό αφανής έμπορος.

Με πράσινο χρώμα εμφανίζονται οι επιχειρήσεις με χαρακτηρισμό μεσάζων (broker).

Σύμφωνα με την διαμόρφωση του δικτύου οι υφιστάμενες χαρακτηρισμένες επιχειρήσεις έχουν σημαντικούς δεσμούς που τις συνδέουν μεταξύ τους, αλλά και τις ίδιες με άλλες ξένες επιχειρήσεις η οποίες έχουν χαρακτηριστεί ότι έχουν συγκεκριμένες ιδιότητες (αφανής έμπορος, μεσάζων, κλπ).

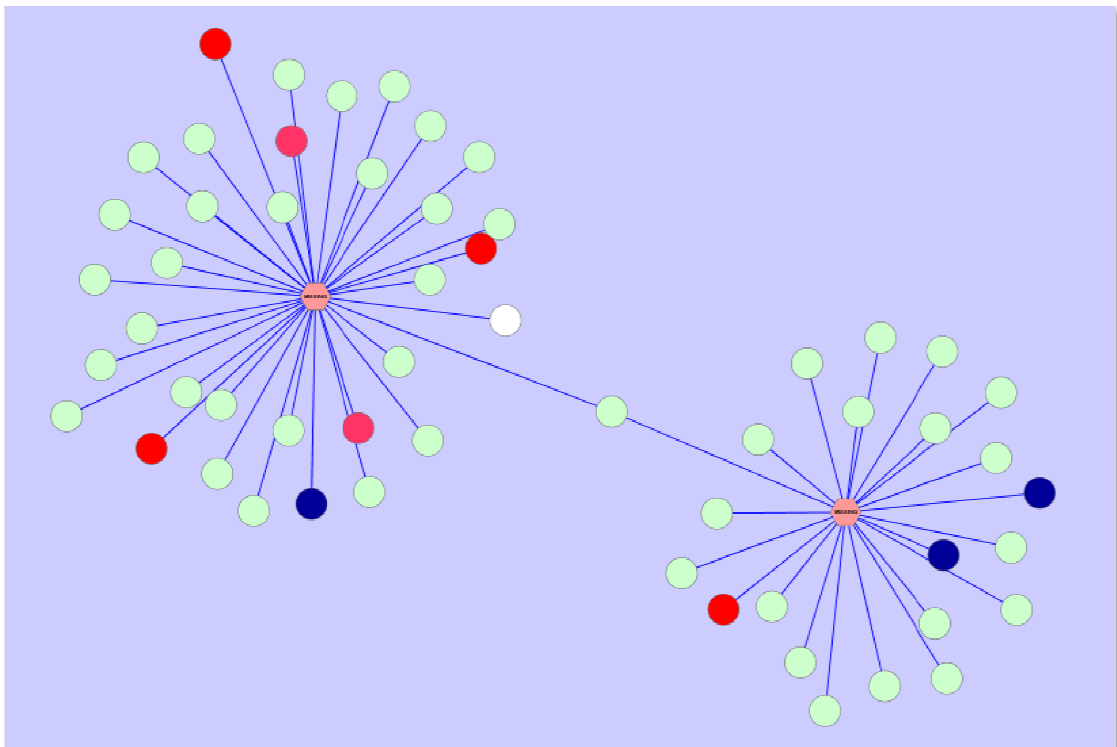
Το γεγονός αυτό χρήζει προσοχής. Θα περιμέναμε ότι η αναζήτηση των συναλλαγών των 208 χαρακτηρισμένων (κλάση=1) επιχειρήσεων, μέσα σε ένα σύνολο περισσότερων από 10 εκ. συναλλαγών θα επιστρέψει «φτωχά» αποτελέσματα. Όπου «φτωχά» αποτελέσματα χαρακτηρίζουμε δίκτυα της μορφής :



Εικόνα 82- Δίκτυο με "φτωχή" πληροφορία

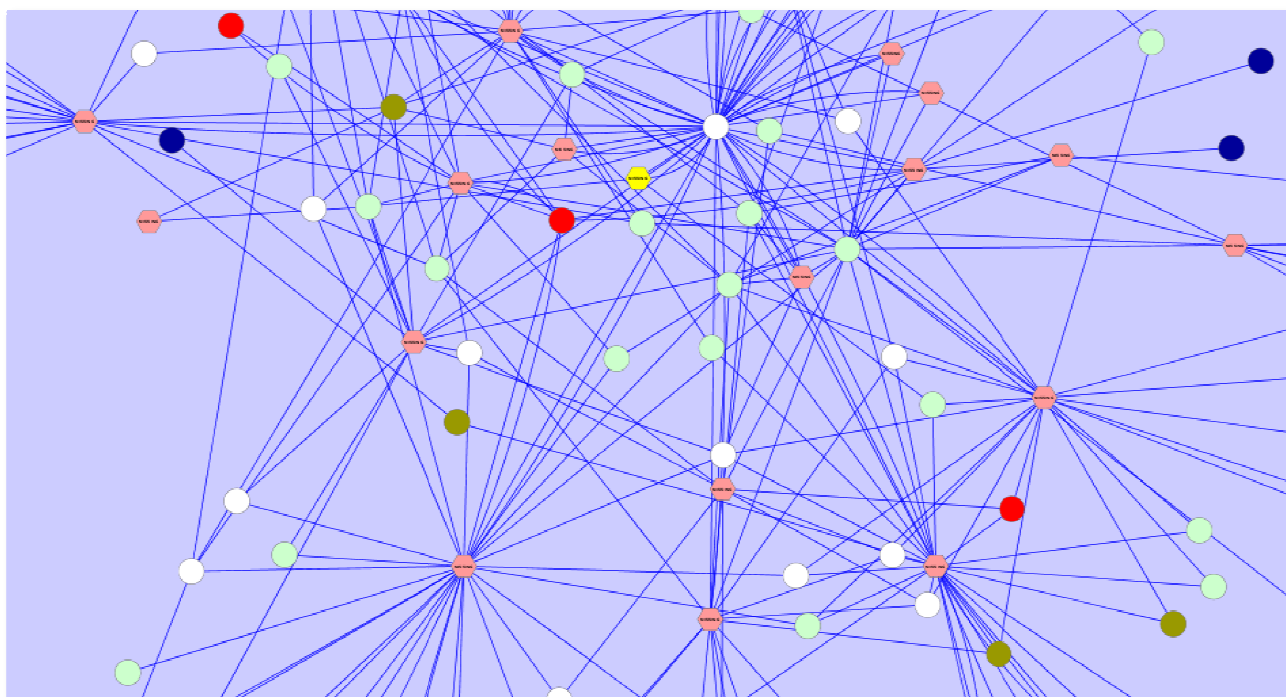
Δηλαδή δίκτυα τα οποία έχουν ελάχιστες συνδέσεις ή συνδέσεις με μικρή ισχύ. Αντί αυτού εμφανίζονται ανεπτυγμένα δίκτυα με ισχυρές/πολλές συσχετίσεις μεταξύ των επιχειρήσεων/κόμβων.

Στο παραπάνω σχήμα (εικόνα 81), ορισμένες ακμές έχουν χαρακτηριστεί με διαφορετικό είδος γραμμής. Οι ακμές αυτές τροποποιήθηκαν ώστε να γίνει εμφανής η ροή που σχηματίζεται και μοιάζει με τη ροή της εικόνας 17. Ουσιαστικά οριοθετούν ένα κύκλωμα το οποίο σε μεγάλο βαθμό μοιάζει με το θεωρητικό μοντέλο της περίπτωσης που εξετάζεται.



Εικόνα 83- Δίκτυα συναλλαγών (Cytoscape)

Παρόμοιες παρατηρήσεις, χαρακτηρίζουν και αρκετά από τα υπόλοιπα δίκτυα που αναδεικνύονται. Σε μια άλλη περίπτωση (εικόνα 83), ενός σαφώς απλούστερου δικτύου δυο αφανείς επιχειρήσεις (ροζ εξάγωνα) συνδέονται μεταξύ τους με έναν μεσάζοντα (πράσινος κόμβος). Παράλληλα πλαισιώνονται από μια σειρά μεσαζόντων, αφανών εμπόρων, πλαστών ΑΦΜ (σκούρος μπλε κόμβος) και ύποπτων επιχειρήσεων.



Εικόνα 84- Πολύπλοκο δίκτυο συναλλαγών (Cytoscape)

Όμοια κατάσταση σε ένα μεγαλύτερο δίκτυο.

ΚΕΦΑΛΑΙΟ 8 – Συμπεράσματα

1. Διαθέσιμα εργαλεία

Τα διαθέσιμα εργαλεία εκτέλεσαν με ευκολία τις εργασίες για τις οποίες χρησιμοποιήθηκαν. Δεν εμφάνισαν ή αντιμετώπισαν σημαντικά προβλήματα ή αδυναμίες. Το γεγονός αυτό φανερώνει ότι από πλευράς εργαλείων η περιοχή της εξόρυξης γνώσης έχει ολοκληρώσει την αρχική περίοδο ανάπτυξης και έχει εισέρθει σε περίοδο ωρίμανσης.

Αν κάποιος θελήσει να κάνει μια έρευνα αγοράς θα παρατηρήσει ότι υπάρχει πλήθος εργαλείων τα οποία είναι διαθέσιμα για την εξόρυξη και ανάλυση δεδομένων. Στην παρούσα εργασία περιορίστηκα σε εργαλεία τα οποία είναι ελεύθερα διαθέσιμα. Δεν θα πρέπει να μας διαφεύγει ότι εμπορικά εκμεταλλεύσιμα εργαλεία ανάλυσης παρέχουν σχεδόν όλες οι εταιρείες, είτε εξειδικευμένες στο τομέα, είτε όχι.

Στον συγκεκριμένο τομέα δραστηριότητα εμφανίζουν εκτός από εταιρείες που σχετίζονται με ανάλυση δεδομένων και εταιρείες που το κύριο αντικείμενο τους δεν έχει άμεση σχέση με τον τομέα, όπως πχ η Amazon, η Microsoft κλπ.



Εικόνα 85- Ανάλυση αγοράς από την Gartner(2014)

Άλλες φορές η πρόταση τους αφορά την επέκταση των υπαρχόντων προϊόντων τους (Oracle, SAP), ενώ άλλες φορές η πρόταση τους αφορά την παροχή υλικοτεχνικής υποστήριξης με την μορφή πρωτογενών πηγών (πχ Amazon, Microsoft).

2. Συμπεράσματα από τη χρήση των εργαλείων

Ευκολία χρήσης

Ο Data Miner της Oracle και η πλατφόρμα Knime παρέχουν γραφικό interface προς τον χρήστη. Ωστόσο απευθύνονται σε διαφορετικές ομάδες χρηστών. Ο Data Miner ξεχωρίζει για την ευκολία χρήσης του. Σημαντικό ρόλο σε αυτό έχει η απλότητα των ρυθμίσεων που απαιτούνται, οι οποίες είναι σχεδόν μηδαμινές. Ο χρήστης θα βρεθεί σε σύντομο χρονικό διάστημα να σχεδιάζει μοντέλα ανάλυσης.

Η πλατφόρμα Knime ενώ παρέχει και αυτή γραφικό περιβάλλον, για την σχεδίαση της ανάλυσης, απαιτεί όμως από τον χρήστη μια περίοδο εξοικείωσης, πριν αυτός να είναι σε θέση να εκμεταλλευτεί τις δυνατότητες που του παρέχονται. Η δυσκολία προέρχεται κυρίως από το γεγονός ότι η πλατφόρμα παρέχει περισσότερες ρυθμίσεις στον χρήστη ο οποίος θα πρέπει να είναι εξοικειωμένος με τους αλγόριθμους εξόρυξης και τον τρόπο παραμετροποίησης αυτών.

Η εφαρμογή WEKA σίγουρα είναι μια ιδιαίτερη εφαρμογή. Προς το παρόν διαθέτει και command line interface και ένα γραφικό interface, ενώ βρίσκεται σε ανάπτυξη ένα νέο γραφικό interface, παρόμοιας λειτουργικότητας με τις προηγούμενες δύο εφαρμογές. Η χρήση του δεν μπορεί να χαρακτηριστεί απλή, αν και όταν ο χρήστης συνηθίσει το περιβάλλον λειτουργίας, θα είναι σε θέση να εκτελέσει εύκολα τις σχετικές εργασίες.

Δυνατότητες εργαλείων

Η πλατφόρμα Knime παρέχει ένα εύρος εργαλείων για την ανάλυση των στοιχείων και την εκτέλεση της εξόρυξης δεδομένων. Εκτός από τους αλγόριθμους εξόρυξης υιοθετεί μια σειρά ευκολιών ανάκτησης δεδομένων από εξωτερικές πηγές (βάσεις δεδομένων, internet, text αρχεία κλπ.), αλλά και μετασχηματισμών αυτών (φίλτρα, μετατροπές, κλπ.). Στόχος της πλατφόρμας είναι να αποτελέσει ένα ολοκληρωμένο περιβάλλον εργασίας μέσω του οποίου θα εκτελούνται όλα τα βήματα της ανάλυσης δεδομένων (Συλλογή, προ-επεξεργασία, μετασχηματισμός, εξόρυξη, ανάλυση δεδομένων).

Ισχυρό πλεονέκτημα της πλατφόρμας είναι η δυνατότητα της να μπορεί να χρησιμοποιεί αλγόριθμους εξόρυξης οι οποίοι έχουν αναπτυχθεί για άλλα εργαλεία (πχ WEKA) ή σε άλλες γλώσσες (πχ R). Ουσιαστικά με την δυνατότητα αυτή ενισχύει την ιδιαιτερότητα του ως πλατφόρμα ανάλυσης και ανακάλυψης της γνώσης.

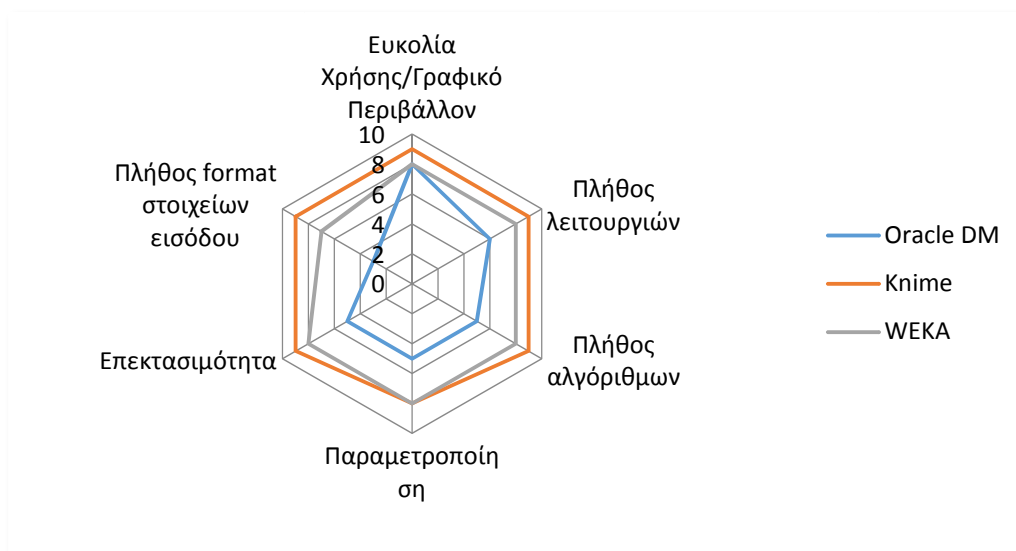
Ο Data Miner είναι μια προσθήκη στη σειρά των εργαλείων που παρέχει η Oracle και έχουν αναπτυχθεί γύρω από τη γνωστή βάση δεδομένων. Όπως και τα περισσότερα εργαλεία της εταιρείας προϋποθέτει την ύπαρξη

και χρήση της συγκεκριμένης βάσης δεδομένων για να μπορέσει λειτουργήσει. Στόχος είναι να αποτελέσει ένα περιβάλλον το οποίο θα διευκολύνει τον χρήστη στην εκτέλεση της εξόρυξης γνώσης και την ερμηνεία των αποτελεσμάτων. Το εργαλείο υλοποιεί ένα γραφικό περιβάλλον το οποίο ο χρήστης μπορεί να χρησιμοποιήσει για να δημιουργήσει το πρόγραμμα αντί να γράψει ο ίδιος τις εντολές. Το πλήθος και η ποικιλία των αλγόριθμων που υποστηρίζει αυτή τη στιγμή είναι μικρό και περιορίζεται σε ότι έχει υλοποιηθεί από την εταιρεία. Περαιτέρω εργασίες που αφορούν την ανάλυση των δεδομένων, όπως προεπεξεργασία, μετασχηματισμούς κλπ δεν υλοποιούνται ή υλοποιούνται σε πολύ βασικό επίπεδο από το εργαλείο, καθώς γραμμή της εταιρείας είναι ότι αυτές οι εργασίες θα πρέπει να εκτελούνται από την υποδομή της βάσης δεδομένων και όχι από το εργαλείο.

Το WEKA είναι ένα πρόγραμμα ανακάλυψης γνώσης το οποίο χαρακτηρίζεται από την εστίαση του, στους αλγόριθμους εξόρυξης γνώσης. Παρέχει μεθόδους διαχείρισης των δεδομένων (κυρίως φίλτρα), αλλά το δυνατό σημείο του είναι η μεγάλη ποικιλία από αλγόριθμους εξόρυξης. Επίσης για κάθε αλγόριθμο που υλοποιείται δίνονται μεγάλες δυνατότητες παραμετροποίησης του. Αποτελεί ένα εργαλείο το οποίο έχει σαφή προσανατολισμό στην διερεύνηση των αλγόριθμων και στην απόδοση αυτών πάνω στα τροφοδοτούμενα δεδομένα.

Χρήση των εργαλείων

Στα πλαίσια της παρούσας εργασίας έγινε φανερό ότι οι τομείς που καλύπτουν τα τρία εργαλεία είναι διαφορετικοί αλλά ταυτόχρονα αλληλοκαλύπτονται.



Εικόνα 86- Βαθμολόγηση των εργαλείων (DM, WEKA, Knime)

Το WEKA απαιτεί μια περίοδο εκμάθησης του user interface, αλλά είναι ιδιαίτερα ικανό στην ανάλυση των στοιχείων. Στις περιπτώσεις που διατίθενται προεπεξεργασμένα δεδομένα ή η προεπεξεργασία μπορεί να γίνει με άλλες μεθόδους, το εργαλείο αποτελεί μια αξιόπιστη και ευέλικτη λύση για την διερεύνηση των στοιχείων.

Το Knime έχει τις περισσότερες δυνατότητες καθώς τοποθετείται στην αγορά ως πλατφόρμα ανάλυσης δεδομένων και ταυτόχρονα είναι σε θέση να συνεργασθεί με αλγόριθμους γραμμένους σε διάφορες γλώσσες. Ουσιαστικά είναι ένα πολυεργαλείο το οποίο όμως απαιτεί μια περίοδο εξοικείωσης για να μπορέσει ο χρήστης να το αξιοποιήσει. Μπορεί να αποτελέσει τη βάση για μια περισσότερο μακροχρόνια επένδυση ενός οργανισμού καθώς μπορεί να υποστηρίξει την τμηματική ανάπτυξη και εξέλιξη των ροών (προσθήκες πηγών, φίλτρων, αντικατάσταση αλγορίθμων κλπ.)

Ο Data Miner είναι ένα ικανοποιητικό εργαλείο για όσους είναι υποχρεωμένοι να χρησιμοποιούν την βάση της εταιρείας και θέλουν να εξερευνήσουν και το πεδίο της εξόρυξης δεδομένων. Ουσιαστικά με την ύπαρξη του εργαλείου, ένας οργανισμός εκμεταλλεύεται την υπάρχουσα τεχνογνωσία που διαθέτει στην χρήση της βάσης και πάνω σε αυτή προσθέτει την ικανότητα ανάλυσης και εξόρυξης.

3. Αποτελέσματα ανάλυσης και εξόρυξης

Η ανάλυση των δεδομένων έγινε μέσω δυο διαφορετικών προσεγγίσεων. Τόσο με τον Data Miner όσο και με το WEKA η ανάλυση βασίστηκε στην δυνατότητα των εργαλείων να ανακαλύψουν συσχετίσεις και να προβλέψουν την κλάση βασισμένα σε γνωστά γνωρίσματα των υπό εξέταση δεδομένων.

Ενώ τα μοντέλα που δημιουργήθηκαν και με τα δύο εργαλεία εμφάνιζαν ικανοποιητικές επιδόσεις στην ταξινόμηση των στοιχείων, όταν στο πειραματικό μοντέλο εφαρμόζονταν τα στοιχεία ελέγχου, η απόδοση του αλγόριθμου έπεφτε σημαντικά.

Χαρακτηριστικότερο παραδείγματα η ταξινόμηση με τον αλγόριθμο J-48 στο WEKA :

Εκπαίδευση μοντέλου		
	19882 TN	0 FN
	9 FN	84 TP

Έλεγχος μοντέλου		
	25059 TN	0 FN
	115 FN	0 TP

Εικόνα 87- Απόδοση αλγόριθμου J-48

Αν ανατρέξει κάποιος στους πίνακες του κεφαλαίου 7 ή/και στο παράρτημα 5, θα παρατηρήσει παρόμοια συμπεράσματα για όλα τα μοντέλα που δημιουργήθηκαν. Η πτώση της απόδοσης των μοντέλων είναι συνεπής ανεξάρτητα με το είδος του αλγόριθμου που χρησιμοποιήθηκε, ορισμένες φορές μάλιστα φαίνεται να επιβεβαιώνεται ότι τα απλά μοντέλα καταφέρνουν να δώσουν αποτελέσματα που είναι ισάξια ή και καλύτερα των πιο σύνθετων.

Τέτοιες επιδόσεις πετυχαίνουν πχ τα μοντέλα που δημιουργήθηκαν με τη μέθοδο Naive Bayes ή με KStar

Εκπαίδευση μοντέλου		
	19841 TN	41 FN
	29 FP	64 TP

Έλεγχος μοντέλου		
	25013 TN	43 FN
	108 FP	7 TP

Εικόνα 88- Απόδοση αλγόριθμου Naive Bayes

Εκπαίδευση μοντέλου		
	19867 TN	15 FN
	43 FP	50 TP

Έλεγχος μοντέλου		
	25025 TN	34 FN
	110 FP	5 TP

Εικόνα 89- Απόδοση αλγόριθμου KStar

Κρίνοντας τα αποτελέσματα έχουμε να παρατηρήσουμε τα εξής :

- Τα αποτελέσματα που επιτυγχάνονται μετά τον έλεγχο του μοντέλου δεν είναι ικανοποιητικά.
- Τα αποτελέσματα μπορεί να μην είναι ικανοποιητικά, ωστόσο το γεγονός ότι καταφέρνουν να υποδείξουν έστω και περιορισμένο αριθμό θετικών περιπτώσεων είναι καλύτερο από την τυφλή διερεύνηση των υποθέσεων.
- Η ίδια η διαδικασία που παράγει τα μοντέλα είναι σημαντική, καθώς συνεισφέρει στην συσσώρευση γνώσης στο πεδίο.

Η αντιστροφή των σετ εκπαίδευσης και ελέγχου και η σημαντική αλλαγή στην συμπεριφορά των μοντέλων υποδεικνύει ότι η υπάρχει μια διαφοροποίηση στο πέραςμα του χρόνου της σύνθεσης των γνωρισμάτων που χαρακτηρίζουν την κλάση. Η αλλαγή στην συμπεριφορά των αλγορίθμων ίσως οφείλεται και στο γεγονός ότι οι υποθέσεις που εντοπίζονται έχουν αυξηθεί σημαντικά τον τελευταίο χρόνο, οπότε και η ανάλυση των στοιχείων έχει στη διάθεση της περισσότερα και πιο λεπτομερή στοιχεία για να δημιουργήσει τα μοντέλα.

Ουσιαστικές πληροφορίες αντλήθηκαν και από την εντοπισμό των σημαντικότερων γνωρισμάτων (select attribute) από το WEKA. Χρησιμοποιώντας τον αλγόριθμο InfoGainAttributeEvaluator εντοπίστηκαν από το σύνολο των 26 γνωρισμάτων 8 τα οποία είχαν την σημαντικότερη συμβολή στην επιλογή της κλάσης.

Η ομαδοποίηση (clustering) των επιχειρήσεων με βάση τα γνωρίσματα ανέδειξε ομάδες που μοιράζονται κοινά χαρακτηριστικά. Αυτές τις ομάδες η φορολογική διοίκηση μπορεί να τις εξυπηρετήσει καλύτερα μέσω εξατομικευμένων ενεργειών. Όμοια η ομαδοποίηση των επιχειρήσεων με βάση τα γνωρίσματα και η

συσχέτιση των ομάδων με την κλάση που ανήκουν, ανέδειξε γνωρίσματα τα οποία έχουν σημαντική συμβολή στην πρόβλεψη της κλάσης.

Ο εντοπισμός γνωρισμάτων που αφορούν συσχετίσεις μεταξύ των επιχειρήσεων, έδωσε το έναυσμα για μια εναλλακτική προσέγγιση στην ανάλυση των στοιχείων. Η χρήση των δεδομένων για την δημιουργία δικτύων φανερώνει την δυναμική της τεχνικής. Στα πλαίσια της συγκεκριμένης μελέτης, η ανάλυση δικτύων περιορίστηκε ουσιαστικά σε περιγραφή της κατάστασης και όχι σε πρόβλεψη. Ο περιορισμός αυτός ήταν αναγκαίος γιατί η τεχνική υποδομή ήταν αδύναμη για να εκτελεσθούν τα απαραίτητα βήματα ανάλυσης. Μια πλήρη ανάλυση δικτύου για τα τελευταία πέντε έτη θα απαιτούσε τον σχηματισμό ενός δικτύου με περίπου 1.8 εκ. συνδέσμους/ακμές και περίπου 650 χιλ. κόμβους.

Αντίθετα η μέθοδος που επιλέχθηκε περιελάμβανε περίπου 1400 συνδέσμους/ακμές και περίπου 1200 κόμβους, αριθμός ο οποίος είναι τεχνικά διαχειρίσιμος.

Από την απεικόνιση του δικτύου και μια μικρή αναδιάταξη των κόμβων σχηματίστηκαν μερικά ιδιαίτερα ενδιαφέροντα υποδίκτυα. Μέσω της μελέτης των υποδικτύων γίνεται φανερό ότι μεταξύ των επιχειρήσεων που ανήκουν στη κλάση '1' και των επιχειρήσεων που έχουν ένα από τα γνωρίσματα (οργανωτής, αφανής, μεσάζων κλπ), υπάρχει στενή σχέση. Ουσιαστικά επιβεβαιώνεται η υπόθεση ότι η πιθανότητα κάποιος να διαπράττει απάτη εξαρτάται (και) από τις σχέσεις που αυτός έχει. Παρόμοια σχέση είχαν εντοπίσει και οι B. Baesens, V. Vlasselaar, W. Verbeke(2011), ότι οι διαδικασίες εντοπισμού της απάτης μπορούν να ευνοηθούν από ανάλυση τύπου κοινωνικών δικτύων.

4. Όρια και περιορισμοί

Για την εκτέλεση της παρούσας εργασίας χρησιμοποιήθηκαν πρωτογενή δεδομένα από στοιχεία ενδοκοινοτικών συναλλαγών. Τα γνωρίσματα που επιλέχθηκαν αποτελούν ένα περιορισμένο σύνολο από τις δυνατές επιλογές. Πχ δεδομένου ότι η απάτη που εξετάστηκε, για να καλυφθεί, συνδυάζεται με την συνήθη ροή του εμπορίου θα ήταν ενδιαφέρον να εμπλουτισθούν τα στοιχεία με μετρικές από τις εθνικές συναλλαγές (περιοδικές ΦΠΑ).

Τα γνωρίσματα που χρησιμοποιήθηκαν είναι κατά βάση κατηγορικά της μορφής 0/1 (ναι/όχι). Η απουσία αριθμητικών δεδομένων (πχ μηνιαίες συναλλαγές) είναι μια έλλειψη που πιθανά επηρεάζει την ποιότητα της ανάλυσης.

Λόγο τεχνικών περιορισμών, ορισμένες τεχνικές εξόρυξης δεν στάθηκε δυνατόν να δοκιμασθούν.

Επίσης δεν εφαρμόστηκαν τεχνικές κατάτμησης των δεδομένων και εκτέλεσης ανάλυσης σε αυτές.

5. Σύνοψη και μελλοντικές ενέργειες

Από την διερεύνηση των δεδομένων προέκυψαν ορισμένα ενδιαφέροντα συμπεράσματα. Σαφώς το σημαντικότερο αφορά την χρήση γράφων για την ανάλυση και τον εντοπισμό περιπτώσεων. Το αποτέλεσμα δικαιολογείται και από τη ανάλυση του επιχειρηματικού πλαισίου. Με τη μέθοδο επιβεβαιώθηκε το θεωρητικό μοντέλο λειτουργίας της απάτης. Συνεπώς επιβεβαιώνεται η δυνατότητα χρήσης τεχνικών ανάλυσης για τον εντοπισμό περιπτώσεων σε φορολογικά δεδομένα.

Με την επιβεβαίωση αυτή δημιουργείται μια ενδιαφέρουσα δυνατότητα χρήσης της τεχνικής σε οποιοδήποτε τομέα υπάρχουν παρόμοια δεδομένα. Τέτοιοι τομείς θα μπορούσε να είναι η αναζήτηση της απάτης των πλαστών/εικονικών παραστατικών, η ανάλυση της φοροαποφυγής, η διερεύνηση της διακίνησης του μαύρου χρήματος κλπ.

Η συνολική ανάλυση βασίσθηκε και ολοκληρώθηκε με τη χρήση εργαλείων ανοικτού κώδικα/δωρεάν διανομής. Τα εργαλεία παρουσίασαν εξαιρετική ωριμότητα τόσο ως προς την σταθερότητα τους, όσο και προς την διαθεσιμότητα των επιλογών (αλγόριθμοι, χρήση κλπ.). Κάποια εργαλεία προσπαθούν να καλύψουν το σύνολο των διαδικασιών ανακάλυψης γνώσης, ενώ άλλα επικεντρώνονται σε περισσότερο εξειδικευμένους τομείς εξόρυξης γνώσης. Ένα θέμα που δημιουργείται από τη χρήση των συγκεκριμένων λογισμικών, αλλά δεν αφορά μόνο τα συγκεκριμένα, είναι η αδυναμία συνεργασίας μεταξύ των εργαλείων για την επίτευξη μιας ολοκληρωμένης εμπειρίας. Σε ένα παραγωγικό περιβάλλον θα πρέπει να σταθμιστεί η ευκολία και οι δυνατότητες που τυχόν παρέχει ένα εμπορικό λογισμικό, σε σχέση με αυτά που είναι δυνατά από το λογισμικό ανοικτού κώδικα/δωρεάν διανομής.

Πιστεύουμε ότι εκτός από τα εργαλεία και τις μεθόδους εξόρυξης, σημαντικό ρόλο για την επιτυχία ενός εγχειρήματος ανάλυσης επιτελεί η διαδικασία που ακολουθείται. Θα πρέπει να είναι σαφές εκ των προτέρων ο στόχος της ανάλυσης, τα διαθέσιμα δεδομένα, το επιχειρηματικό πλαίσιο. Μόνο αν αυτές οι ελάχιστες συνθήκες ικανοποιούνται μπορεί η ανάλυση δεδομένων να αποτελέσει μια ουσιαστική πρόταση. Αν δεν υπάρχουν ή είναι ασαφή τα παραπάνω τότε η ερμηνεία και η επαλήθευση των αποτελεσμάτων καθίσταται δυσχερής.

Σε επίπεδο ανάλυσης ενδιαφέρον παρουσιάζουν οι παρακάτω προεκτάσεις :

- Εμπλουτισμός του γράφου με χαρακτηριστικά και βαθμονόμηση.
- Δημιουργία ενός πλήρους δικτύου και ανάδειξη των σημαντικών υποδικτύων μέσω της βαθμονόμησης.

- Κατάτμηση των δεδομένων εκπαίδευσης σε πολλαπλές ομάδες, με σκοπό την εξομάλυνση της ανισορροπίας μεταξύ των κλάσεων.
- Διεύρυνση των γνωρισμάτων με αριθμητικά δεδομένα συναλλαγών, καθώς και με δεδομένα από άλλα πεδία φορολογικού περιεχομένου.
- Εξέταση των χαρακτηριστικών του δικτύου με εναλλακτικούς αλγόριθμους.
- Χρήση υβριδικού μοντέλου με σύνθεση των αποτελεσμάτων του γράφου με τεχνικές εξόρυξης γνώσης.

6. Γενικά συμπεράσματα

Η διαδικασίας ανάλυσης και ανακάλυψης γνώσης έχουν εισέλθει σε περίοδο ωριμότητας. Υπάρχουν αρκετά εργαλεία διαθέσιμα είτε με τη μορφή «ανοικτού» λογισμικού, είτε με τη μορφή «κλειστού» λογισμικού.

Στην παρούσα εργασία χρησιμοποιήσαμε κυρίως εργαλεία ανοικτού λογισμικού καθώς αυτά είναι εύκολα διαθέσιμα, ενώ επίσης χρησιμοποιήθηκε το εργαλείο Data Miner γιατί παρέχεται δωρεάν για τους χρήστες της συγκεκριμένης βάσης.

Οι δυνατότητες των εργαλείων διέφεραν μεταξύ τους, ωστόσο τα αποτελέσματα και των τριών εργαλείων εμφανίζουν συνοχή μεταξύ τους. Με την χρήση των εργαλείων ήταν δυνατόν να εντοπισθούν ορισμένα ενδιαφέροντα μοτίβα καθώς και να ελεγχθεί η συμβολή των γνωρισμάτων στη διαμόρφωση των μοντέλων και η επιτυχία αυτών στην πρόβλεψη.

Σύμφωνα με τα αποτελέσματα της εξόρυξης δεδομένων τόσο τα γνωρίσματα όσο και οι τυπικές μέθοδοι (αλγόριθμοι ταξινόμησης και ομαδοποίησης) που χρησιμοποιήθηκαν απαιτούν περαιτέρω διερεύνηση. Τα μοντέλα που δημιουργήθηκαν από το εργαλείο Data Miner, όσο και αυτά που δημιουργήθηκαν από το εργαλείο WEKA, παρουσιάζουν φτωχά αποτελέσματα όταν χρησιμοποιούνται για πρόβλεψη. Αντίθετα πολύ ενθαρρυντικά είναι τα αποτελέσματα που δημιουργούνται όταν η ανάλυση των στοιχείων γίνεται με όρους «κοινωνικών δικτύων». Το γεγονός ότι η συγκεκριμένη μέθοδος απάτης βασίζεται στην συνεργασία επιχειρήσεων, και στη ουσία στο σχηματισμό μικρών ομάδων (δικτύων), είναι εκείνο το χαρακτηριστικό που πριμοδοτεί ως μέθοδο ανάλυσης και πρόβλεψης τα κοινωνικά δίκτυα. Βασισμένοι στην τελευταία παρατήρηση, αλλά και στο γεγονός ότι το συγκεκριμένο σχήμα απάτης είναι δυναμικό, δηλαδή εξελίσσεται στην πάροδο του χρόνου, θεωρούμε ότι σε μια επόμενη εξέταση του πεδίου, θα πρέπει τα γνωρίσματα να εμπλουτισθούν με δεδομένα που να αφορούν μεγέθη τα οποία θα περιέχουν την έννοια της μεταβολής (πχ. ποσά συναλλαγών, τζίρος, στοιχεία από άλλου είδους δηλώσεων κλπ.).

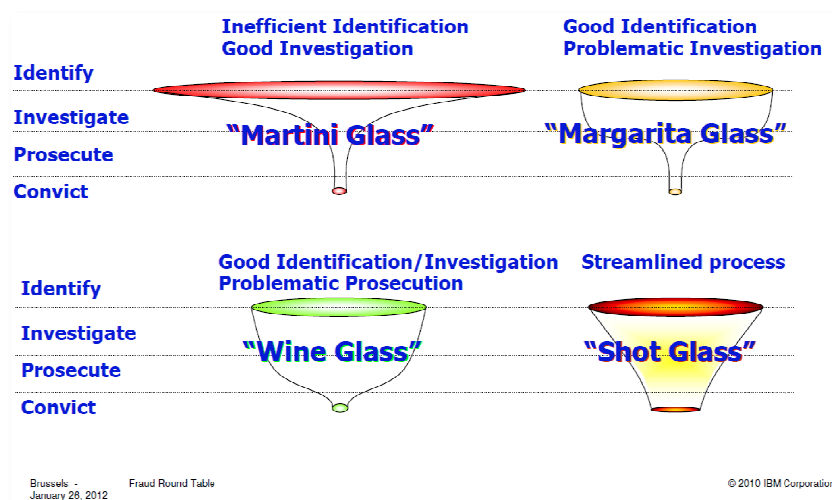
Στη διάρκεια αυτής της εργασίας εκτός από τα εργαλεία και την ανάλυση των δεδομένων παρατέθηκαν και στοιχεία σχετικά με τις διαδικασίες που αφορούν την επεξεργασία των δεδομένων. Θα πρέπει να γίνει

κατανοητό ότι για να υπάρξουν μετρήσιμα αποτελέσματα δεν αρκούν μόνο δεδομένα και εργαλεία αλλά και η ανάπτυξη των κατάλληλων διαδικασιών, ώστε τα αποτελέσματα να είναι αξιοποιήσιμα. Ορισμένοι παράγοντες που θα πρέπει να εξετασθούν για την ενσωμάτωση ενός συστήματος εξόρυξης γνώσης σε ένα επιχειρησιακό περιβάλλον είναι :

- Η γνώσεις των χρηστών του συστήματος
- Η ανάγκες που θα πρέπει να καλύψει (παραγωγική λειτουργία, διερεύνηση κλπ.)
- Συνεργασία με υπάρχοντα εργαλεία / περιβάλλον
- Κόστος συστήματος
- Επιχειρησιακές διαδικασίες που απαιτούνται, είτε πριν είτε μετά την διαδικασία εξόρυξης

Ένα σύστημα εξόρυξης γνώσης δεν είναι ένα αυτοτελές σύστημα αλλά ακόμα ένας κρίκος σε μια μεγαλύτερη αλυσίδα διαδικασιών, στόχος των οποίων είναι αυτό που αναφέρθηκε εξ αρχής, ο περιορισμός της κακής συμπεριφοράς των φορολογούμενων ώστε να σταματήσει η οικονομική απώλεια, αλλά και να υπάρχει μια η ομαλή κοινωνική και οικονομική λειτουργία.

Η παρακάτω εικόνα από την IBM καταφέρει να περιγράψει συνοπτικά το σύνολο των διεργασιών με εύγλωττο τρόπο :



Εικόνα 90- Η σημασία ενός ολοκληρωμένου σχεδιασμού (IBM, Parry 2010-2012)

ΒΙΒΛΙΟΓΡΑΦΙΑ

Artavanis N., Morse A., Tsoutsoura M., 25 September 2015. Measuring Income Tax Evasion using Bank Credit: Evidence from Greece. Διαθέσιμο στο διαδίκτυο: <http://dx.doi.org/10.2139/ssrn.2109500>

Australian Taxation Office, 2008. ANAO Audit Report No.30 2007–08. The Australian Taxation Office's Use of Data Matching and Analytics in Tax Administration.

Azevedo A., Santos M. F., 2008. KDD, SEMMA and CRISP-DM: a parallel overview., in Ajith Abraham, ed., 'IADIS European Conf. Data Mining', IADIS, , pp. 182-185. BibTeX key `conf/iadis/AzevedoS08`

Baesens B., Vlasselaer V.V., Verbeke W., 2015. Fraud Analytics- Using Descriptive, Predictive and Social Network Analysis. ISBN: 978-1-119-13312-4.

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

Brennan P., 2012. A comprehensive survey of methods for overcoming the class imbalance problem in fraud detection. Διαθέσιμο στο διαδίκτυο: <http://www.dataminingmasters.com/uploads/studentProjects/thesis27v12P.pdf>. Πρόσβαση: 25/10/2014

Chawla N. V., 2005. Data Mining and Knowledge discovery handbook. Doi: 10.1007/0-387-25465-X_40

Chen C., Liaw A., Breiman L., 2010. Using Random Forest to Learn Imbalanced Data. Διαθέσιμο στο διαδίκτυο: <http://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf> . Πρόσβαση 20/12/2015

DeBarr D., MITRE, Harwood M., 2005. Relational Mining for Compliance Risk. Διαθέσιμο στο διαδίκτυο: <https://www.irs.gov/pub/irs-soi/04debarr.pdf>. Πρόσβαση 01/06/2015

Efstathios K., Spathis C., Manolopoulos Y., 2007. Data Mining techniques for the detection of fraudulent financial statements. Doi 10.1016/j.eswa.2006.02.016

Europol, 2009. Carbon credit fraud causes more than 5 billion Euros damage for the European taxpayer. Διαθέσιμο στο διαδίκτυο: <https://www.europol.europa.eu/content/press/carbon-credit-fraud-causes-more-5-billion-euros-damage-european-taxpayer-1265>. Πρόσβαση 11/01/2016.

Fayyad U., Piatetsky-Shapiro G., Smyth P., 1996. From Data Mining to Knowledge Discovery in Databases. ISBN 0738-4602-1996.

Federation of Indian Chambers of Commerce and Industry, 2015. A study on widening of tax base and tackling black money. Διαθέσιμο online: <http://ficci.in/spdocument/20548/STUDY-ON-WIDENING-OF-TAX-BASE-AND-TACKLING-BLACK-MONEY.pdf>. Πρόσβαση 01/02/2016.

Financial Action Task Force- FATF, 2007. Laundering the proceeds of VAT carousel fraud. Διαθέσιμο στο διαδίκτυο: <http://www.fatf-gafi.org/media/fatf/documents/reports/Laundering%20the%20Proceeds%20of%20VAT%20Caroussel%20Fraud.pdf>. Πρόσβαση 07/01/2015.

Gonzalez P. C., Velasquez J. D., 2012. Characterization and detection of taxpayers with false invoices using data mining techniques. DOI: 10.1016/j.eswa.2012.08.051

Gupta M., Nagadevara V., 2010. Audit selection strategy for improving tax compliance – application of data mining techniques. Διαθέσιμο στο διαδίκτυο: <http://citeseerx.ist.psu.edu/viewdoc/citations?doi=10.1.1.509.7363>. Πρόσβαση 10/01/2016

Hayes T., 11/04/2013. Online άρθρο: Beware of the ATO's data-matching machine. Διαθέσιμο στο διαδίκτυο: <http://www.smartcompany.com.au/finance/tax/31198-beware-of-the-ato-s-data-matching-machine/>. Πρόσβαση 08/12/2015

Ho T.K. (1995). Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282, Διαθέσιμο στο διαδίκτυο: <http://ect.bell-labs.com/who/tkh/publications/papers/odt.pdf>. Πρόσβαση 02/03/2016

IBM, Parry J.E.(2010-2012), Predictive Analytics and Fraud Prevention. Διαθέσιμο στο διαδίκτυο: [https://www-950.ibm.com/events/wwe/grp/grp004.nsf/vLookupPDFs/James'%20Presentation/\\$file/James'%20Presentation.pdf](https://www-950.ibm.com/events/wwe/grp/grp004.nsf/vLookupPDFs/James'%20Presentation/$file/James'%20Presentation.pdf). Πρόσβαση 01/03/2016

Jedrzejek C., Falkowski M., Bak J., 2009. Graph Mining for Detection of a Large Class of Financial Crimes. Διαθέσιμο στο διαδίκτυο: <http://ceur-ws.org/Vol-483/paper4.pdf>. Πρόσβαση 01/12/2014

Kallio M., Back B., 2011. The Self-Organizing Map in Selecting Companies for Tax Audit. DOI 10.1007/978-3-7908-2739-2_27

- Kotsiantis S., Koumanakos E., Tzelepis D., Tampakas V., 2007. Forecasting Fraudulent Financial Statements using Data Mining. Διαθέσιμο στο διαδίκτυο: <http://waset.org/publications/7180/forecasting-fraudulent-financial-statements-using-data-mining>. Πρόσβαση 22/11/2015
- Malik T., 8/10/2015. Online άρθρο: The Jig Is Up: Combatting Tax Fraud. Διαθέσιμο online <http://www.forbes.com/sites/teradata/2015/10/08/the-jig-is-up-combatting-tax-fraud/#6107f22f28ed>
- Martikainen J., 2012. Data Mining in Tax Administration - Using Analytics to Enhance Tax Compliance. Διαθέσιμο στο διαδίκτυο: http://epub.lib.aalto.fi/en/ethesis/pdf/13054/hse_ethesis_13054.pdf, πρόσβαση 01/12/2014
- Oracle, 2012. Oracle Data Mining 11g Release 2 - Competing on In-Database Analytics. Διαθέσιμο στο διαδίκτυο: <http://www.oracle.com/technetwork/database/options/advanced-analytics/odm/twp-data-mining-11gr2-160025.pdf>. Πρόσβαση 10/01/2016
- Pisani S., Sisti P., 2007. Risk Analysis applied to Tax Evasion using data mining methodology. Διαθέσιμο στο διαδίκτυο: http://www1.agenziaentrate.gov.it/ufficiostudi/documenti/risk_analysis_tax_evasion.pdf, πρόσβαση 01/12/2014
- Rokach L., Maimon O., 2015. Data Mining with decision trees. ISBN 978-9814590082.
- Shannon P., Markiel A., Ozier O., Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks Genome Research 2003 Nov; 13(11):2498-504
- Spathis C. T., 2002 Detecting false financial statements using published data: some evidence from Greece. DOI 10.1108/0268690021042432 1.
- Spathis C., Doumpos M., Zopoudinis C., 2002. Detecting falsified financial statements: a comparative study using multicriteria analysis and multivariate statistical techniques. Διαθέσιμο στο διαδίκτυο: http://users.auth.gr/~hspathis/Spathis-Doumpos-ZopounidisEAR_Site.pdf. Πρόσβαση 10/01/2016
- Thearling K., 2012. An Introduction to Data Mining. Διαθέσιμο στο διαδίκτυο: <http://www.thearling.com/text/dmwhite/dmwhite.htm>. Πρόσβαση 20/10/2015.
- Wang B.X., Japkowic N., 2009. Boosting support vector machines for imbalanced data

Sets. Doi: 10.1007/s10115-009-0198-y

Witten H.I., Frank E., Hall A.M., 2011. "Data Mining: Practical machine learning tools and techniques, 3rd Edition". Morgan Kaufmann, San Francisco

Wikipedia, WEKA (machine learning). [https://en.wikipedia.org/wiki/Weka_\(machine_learning\)](https://en.wikipedia.org/wiki/Weka_(machine_learning)). Πρόσβαση 10/01/2016

Wikipedia, C4.5 algorithm. https://en.wikipedia.org/wiki/C4.5_algorithm#cite_note-1. Πρόσβαση 20/04/2016

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

Εγκυκλοπαίδεια Britannica, 2015, Data Mining. <http://www.britannica.com/technology/data-mining>. Πρόσβαση 10/01/2015

Ευρωπαϊκό Δίκαιο, 1986. Η Ενιαία Ευρωπαϊκή Πράξη. Διαθέσιμο στο διαδίκτυο: <http://eur-lex.europa.eu/legal-content/EL/TXT/?uri=URISERV:xy0027>. Πρόβαση 15/09/2015.

	Αρχική αξία	Διάφορα κόστη	Κέρδος	Αξία Πώλησης	ΦΠΑ 23%	Απόδοση ΦΠΑ	
Αρχικό προϊόν Α	90		10%	99	22.77	22.77	Επόμενος αγοραστής
Αγορά του Α από εταιρεία για επεξεργασία	99	αξία επεξεργασίας = 21	15%	$99 + 21 + \text{κέρδος} = 126.5$	29.01	$29.01 - 22.77 = 6.24$	Επόμενος αγοραστής
Αγορά του Α από εταιρεία για συσκευασία	126.5	αξία συσκευασίας = 3.5	10%	$126.5 + 3.5 + \text{κέρδος} = 143$	32.89	$32.89 - 29.01 = 3.88$	Επόμενος αγοραστής
Πώληση σε τελικό καταναλωτή	143		10%	$143 + \text{κέρδος} = 157.3$	36.18	$36.18 - 32.89 = 3.29$	Τελικός καταναλωτής
						$22.77 + 6.24 + 3.88 + 3.29 = 36.18$	

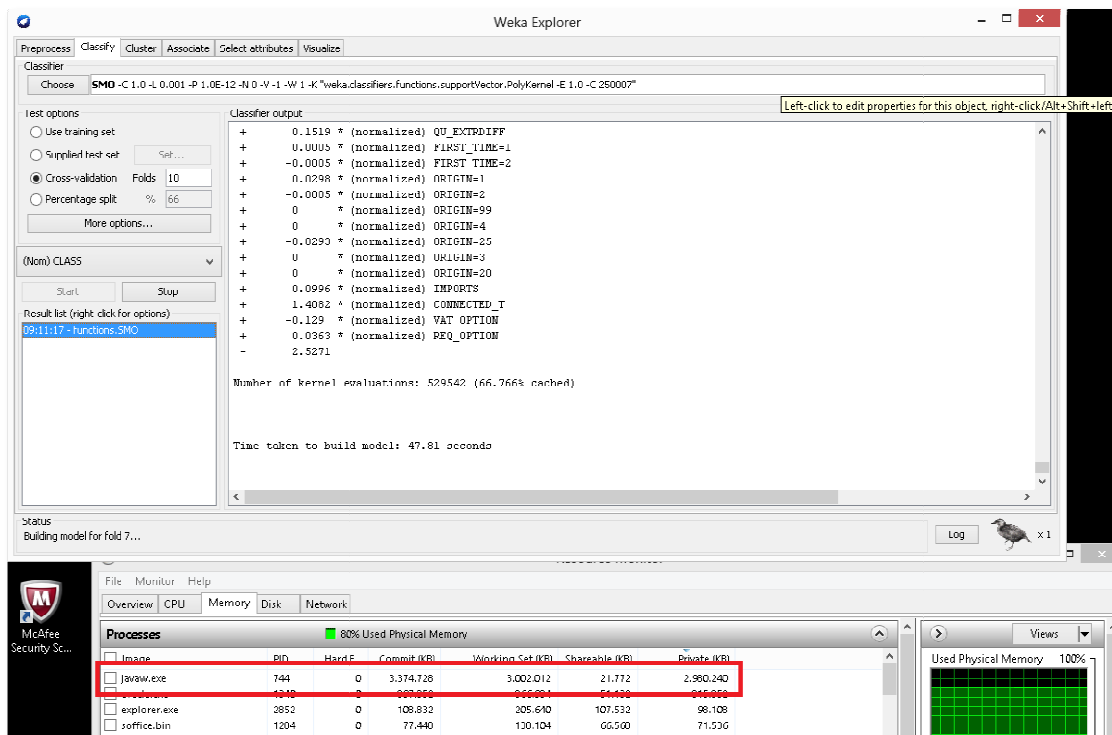
Εικόνα 91 - απόδοση ΦΠΑ σε εθνικές συναλλαγές

	Αρχική αξία	Διάφορα κόστη	Κέρδος	Αξία Πώλησης	ΦΠΑ 20%		
Αρχικό προϊόν Α	90		10%	99	19.8	19.8	Επόμενος αγοραστής
Αγορά του Α από εταιρεία για επεξεργασία	118.8	αξία επεξεργασίας = 21	15%	$99 + 21 + 15\% = 126.5$	0	0	
					ΦΠΑ 23%		
Αγορά του Α από εταιρεία για συσκευασία. Η εταιρεία βρίσκεται σε άλλη χώρα της ΕΕ	126.5	αξία συσκευασίας = 3.5	10%	$126.5 + 3.5 + \text{κέρδος} = 143$	32.89	32.89	Επόμενος αγοραστής
Πώληση σε τελικό καταναλωτή	143		10%	157.3	36.18	$36.18 - 32.89 = 3.29$	Τελικός καταναλωτής

Εικόνα 92 - απόδοση ΦΠΑ σε ενδοκοινοτικές συναλλαγές

ΠΑΡΑΡΤΗΜΑ 2

1. Διαχείριση Μνήμης



Εικόνα 93- Απαιτήσεις σε μνήμη (WEKA)

Η εξόρυξη γνώσης είναι συνάρτηση :

- Στατιστικής / μαθηματικών
- Επιχειρηματικής γνώσης
- Διαχείρισης των υπολογιστικών δομών

Ενώ στην εργασία δίνεται περισσότερη προσοχή στις δύο πρώτες παραμέτρους, δεν θα πρέπει να διαφεύγει της προσοχής μας ότι και ο τρίτος παράγοντας είναι σημαντικός. Στην εικόνα 92 παρατίθεται ένα στιγμιότυπο από την διαδικασία εξόρυξης με το εργαλείο Weka. Όπως φαίνεται η διαδικασία απαιτεί από το σύστημα περίπου 3Gb μνήμης. Η δέσμευση της μνήμης γίνεται με την μορφή heap space της Java.

Ανάλογα με το μέγεθος του δείγματος που θα διερευνηθεί θα πρέπει να έχουμε υπόψη μας και τις τεχνικές απαιτήσεις. Στο παραπάνω παράδειγμα ο υπολογιστής θα πρέπει να έχει αρκετή διαθέσιμη μνήμη. Πιθανά θα πρέπει να απελευθερωθεί μνήμη από άλλες διεργασίες ή να γίνει κατανομή των υπολογιστικών πόρων, ώστε να εκτελεσθεί η διαδικασία. Άλλη παράμετρος, που εξαρτάται από τα εργαλεία που χρησιμοποιούνται είναι ο τύπος του λειτουργικού συστήματος, πχ στην περίπτωση των Windows, ο heap χώρος που μπορεί το σύστημα να αποδώσει στην Java δεν ξεπερνάει τα 2Gb στην περίπτωση του 32bit λειτουργικού, ενώ δεν υφίσταται ο περιορισμός στην έκδοση των 64bit.

ΠΑΡΑΡΤΗΜΑ 3

Στο παράρτημα θα αρχειοθετήσουμε ορισμένες τεχνικές που χρησιμοποιήθηκαν κατά την συλλογή και διαχείριση των στοιχείων μέχρι την εκτέλεση της διαδικασίας εξόρυξης γνώσης.

1. *Ανωνυμοποίηση του ΑΦΜ*

Η απόκρυψη των ΑΦΜ επιλέχθηκε να γίνει με την μετατροπή τους μέσω αλγόριθμου τύπου hash. Η επιλογή έγινε γιατί δεν υπάρχει ανάγκη για επανάκτηση του αρχικού ΑΦΜ από το κωδικοποιημένο κείμενο.

Για την κωδικοποίηση επιλέχθηκε να χρησιμοποιηθούν οι δυνατότητες που παρέχει η συγκεκριμένη βάση δεδομένων, μέσω των πακέτων που εμπεριέχει.

Χρησιμοποιήθηκε το πακέτο DBMS_CRYPTO, το οποίο υποστηρίζει διάφορους αλγόριθμους κρυπτογράφησης και δημιουργίας hash κλειδιών. Για την παρούσα εργασία χρησιμοποιήθηκε η συνάρτηση MAC, η οποία ακολουθεί το πρότυπο IETF RFC 2104²

Σύμφωνα με το εγχειρίδιο της Oracle η συνάρτηση χρησιμοποιείται συνήθως για την δημιουργία hash κωδικών αυθεντικοποίησης ενός μηνύματος (Message Authentication Code). Η συνάρτηση επιλέχθηκε γιατί παρέχει την δυνατότητα να κωδικοποιήσει τα αρχικά δεδομένα με τις γνωστούς αλγόριθμους MD5 και SHA με το επιπλέον χαρακτηριστικό ότι για να επιβεβαιώσει κάποιος το αποτέλεσμα απαιτείται η επιπλέον χρήση ενός κλειδιού.

Παρακάτω παρατίθεται μια τυπική εφαρμογή της συνάρτησης:

```
select sys.dbms_crypto.mac( DATA.vat, 2, SECRET_KEY )HASHED_RAW_DATA
from DATA
```

Εικόνα 94- Oracle, Δημιουργία κρυπτογραφημένου Hash

Οι παράμετροι είναι κατά σειρά :

- Τα αρχικά δεδομένα
- Ο τύπος της κωδικοποίησης (στο παράδειγμα είναι ο αλγόριθμος SHA)
- Το κλειδί

Για την δημιουργία του κλειδιού χρησιμοποιήθηκε η δυνατότητα του ίδιου πακέτου να δημιουργεί τυχαίες σειρές :

²<https://www.ietf.org/rfc/rfc2104.txt>

```
select sys.dbms_crypto.randombytes(2048) SECRET_KEY from dual
```

Εικόνα 95- Δημιουργία τυχαίου κλειδιού

2. Μετασχηματισμός απόλυτων ποσών σε σχετικό ποσοστό

Μια από τις ιδιότητες (attribute) των στοιχείων αφορούσε τον χαρακτηρισμό τους σε περίπτωση που υπήρχε μεταβολή μεγαλύτερη από 12 φορές μεταξύ δύο τριμήνων.

Για την υλοποίηση του χαρακτηρισμού χρησιμοποιήθηκε η δυνατότητα που παρέχει η Oracle για την εκτέλεση analytical functions στα αποτελέσματα ενός query.

```
select AFM, Year, Quarter, Value, Prev_Quarter_Amnt,
       decode( nvl(Prev_Amnt,0), 0, 1, Value/Prev_Amnt ) Percent
from (
  select Afm, Year, Quarter, Value,
         LAG (Value, 1, 0)
         OVER (partition by Afm ORDER BY Afm, Year, Quarter)
         AS Prev_Quarter_Amnt
  from (
    SELECT Afm, Year, Quarter, sum(AMNT) Value
    FROM Agores_Ellinon_Apo_Xenous
    group by Afm, Year, Quarter
    order by Afm, Year, Quarter
  )
)
where Value > 20000
and abs(decode( nvl(Prev_Amnt,0), 0, 1, Value/Prev_Amnt )) > 12;
```

Εικόνα 96- Εύρεση % μεταβολής μεταξύ συναλλαγών δύο τριμήνων

Η παρούσα υλοποίηση βασίζεται στη συνάρτηση LAG, η οποία για κάθε εγγραφή επιστρέφει την αξία της αμέσως προηγούμενης εγγραφής για το συγκεκριμένο ΑΦΜ, ταξινομώντας τις εγγραφές κατά ΑΦΜ, Έτος και τρίμηνο – over (partition by Afm Order By Afm, Year, Quarter).

ΠΑΡΑΡΤΗΜΑ 4

1. Δημιουργία σετ δεδομένων για το εργαλείο WEKA

Για την εκτέλεση της ανάλυσης με το εργαλείο WEKA δημιουργήθηκαν δύο σετ δεδομένων, τα εκπαιδευτικά δεδομένα και τα δεδομένα ελέγχου. Από τα δύο σετ δημιουργήθηκαν δύο αρχεία δεδομένων τύπου .arff. Η δημιουργία των αρχείων επιλέχθηκε γιατί παρουσιάζει ορισμένα πλεονεκτήματα έναντι της ανάγνωσης απευθείας από τη βάση.

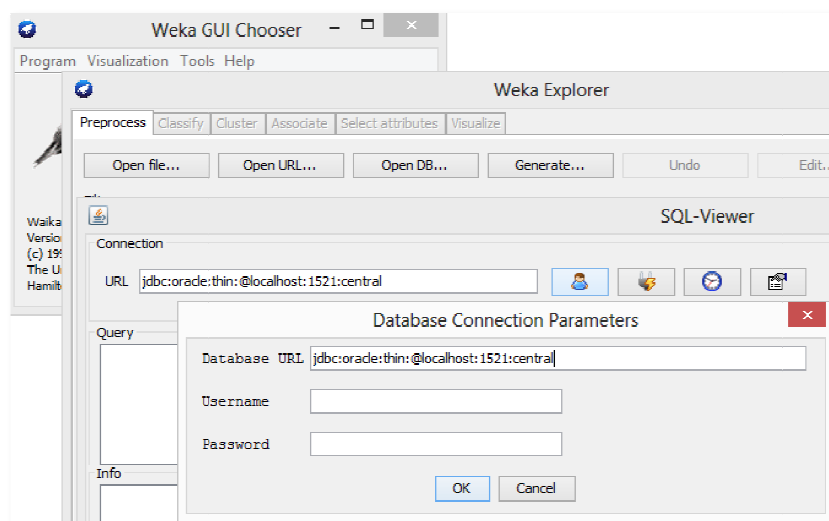
Το εργαλείο για την ανάγνωση από βάσεις δεδομένων προσφέρει ορισμένους drivers προεγκατεστημένους. Σε περίπτωση, όπως η δική μας, που ο driver δεν παρέχεται θα πρέπει για να αποκτήσουμε πρόσβαση στην βάση να εγκατασταθεί ο jdbc driver που αντιστοιχεί στην έκδοση της βάσης που χρησιμοποιούμε.

Η εγκατάσταση γίνεται σε τρία βήματα. Πρώτα αντιγράφουμε τον driver σε ένα directory, κατά προτίμηση κάποιο που είναι ήδη προσβάσιμο από το WEKA. Στη συνέχεια, στο αρχείο RunWeka.ini αλλάζουμε το classpath ώστε να συμπεριλαμβάνει και τον νέο driver, πχ

```
cp=%CLASSPATH%;C:/PROGRAM FILES/WEKA-3-7/ojdbc6.jar;C:/PROGRAM FILES/WEKA-3-7/ojdbc6-12.1.0.2.0.jar
```

Τέλος δημιουργούμε ή αντιγράφουμε ένα που ήδη υπάρχει, αρχείο 'DatabaseUtils.props' στο home directory του WEKA. Το αρχείο περιέχει τις ρυθμίσεις για την ανάγνωση των δεδομένων από τη βάση και την αντιστοίχηση τους σε τύπους δεδομένων που χρησιμοποιεί το WEKA (nominal, numeric κλπ).

Από τη στιγμή που υπάρχει ο driver το WEKA έχει πρόσβαση στη βάση μέσω της επιλογής :



Εικόνα 97- Πρόσβαση σε βάση δεδομένων (WEKA)

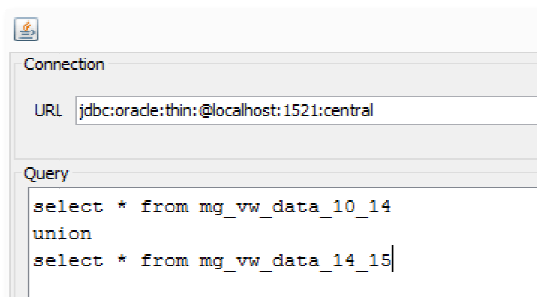
Τα πρόβλημα που προκύπτουν με αυτή την προσέγγιση είναι κυρίως πρακτικά και αφορούν :

- Την αδυναμία να αποθηκευθεί η σύνδεση, ώστε να μην είναι ανάγκη να ξανασυνδεθεί ο χρήστης κάθε φορά που αυτό απαιτείται.

- Ο τρόπος που το WEKA ερμηνεύει τα δεδομένα που διαβάζει από την βάση δεν είναι παραμετροποιήσιμος. Παρατηρείται το φαινόμενο δεδομένα που θέλουμε να ερμηνευθούν ως nominal το εργαλείο να τα ερμηνεύει ως numeric κλπ.

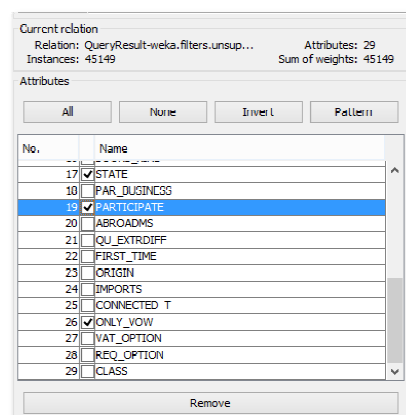
Για να αποφύγουμε τα παραπάνω θέματα δημιουργήθηκαν τα ascii αρχεία με την κατάλληλη τυποποίηση ώστε η ανάλυση των στοιχείων να εκτελεσθεί ταχύτερα. Η δημιουργία των σετ έγινε με συγκεκριμένο τρόπο ώστε αυτά να είναι συμβατά μεταξύ τους. Η ανάγκη για συμβατότητα πηγάζει από την απαίτηση του εργαλείου τα δύο σετ (εκπαίδευσης και ελέγχου) να έχουν ακριβώς τα ίδια χαρακτηριστικά. Δηλαδή ίδιο πλήθος γνωρισμάτων και κάθε γνώρισμα να έχει το ίδιο εύρος τιμών/κατηγοριών. Σε αντίθετη περίπτωση το εργαλείο δεν είναι σε θέση να παρέχει αξιολόγηση του μοντέλου που δημιουργείται σε σχέση με τα δεδομένα ελέγχου.

Τα δύο σετ δεδομένων υπάρχουν ήδη στην βάση ως ξεχωριστοί πίνακες. Αρχικά τα δύο σετ ενώνονται στο interface του εργαλείου ώστε να σχηματίσουν ένα εννιαίο σύνολο στοιχείων.



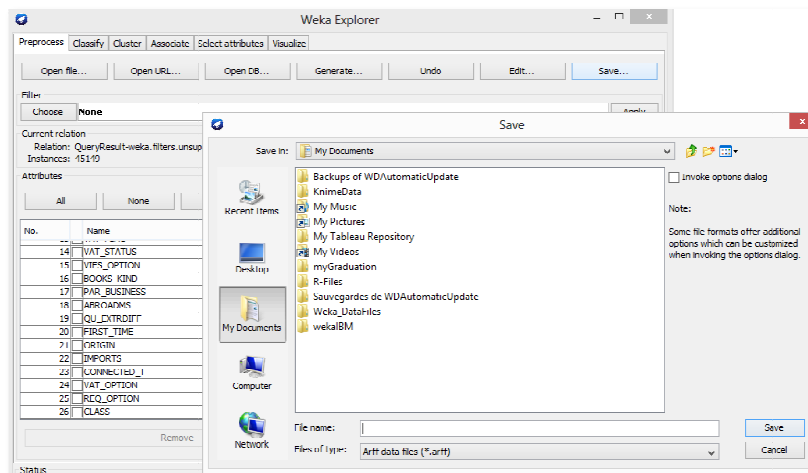
Εικόνα 98- Άντληση δεδομένων από βάση (WEKA)

Από τα γνωρίσματα που προκύπτουν διαγράφονται όλα όσα δεν επιθυμούμε να συμμετέχουν την ανάλυση:



Εικόνα 99- Διαγραφή γνωρισμάτων (WEKA)

Το σετ δεδομένων που απομένει αποθηκεύεται ως αρχείο arff μέσω της δυνατότητας που παρέχει το εργαλείο:



Εικόνα 100- Αποθήκευση τρέχοντος σετ δεδομένων σε αρχείο arff (WEKA)

Το αρχείο που δημιουργείται δεν χρησιμοποιείται για την ανάλυση των δεδομένων αλλά για να εξαχθούν τα δεδομένα του header :

```
@relation 'QueryResult-weka.filters.unsupervised.attribute.Remove-R1,18,26,27'

@attribute KAD1
{9999999,1182,7112,9083,4941,1414,1420,4791,6910,2830,4751,4771,1800,4772,4774,4673,4531,7822,1631,1721,4662,8030,4789,4759,3315,4711,7320,7990,4932,4752,4761,1419,4621,5510,4725,1133,1
411,4631,4669,1084,5810,1812,4661,4778,4329,4540,2825,2712,2599,4765,1051,4754,6010,7556,1623,7311,4719,5010,4775,1520,2222,7756,8219,4671,4532,8121,2512,1020,1421,4642,1261,1231,1401,7
420,1500,8622,5229,5520,4636,4632,8559,3312,4638,1091,4665,4729,4619,9329,2849,4764,4520,6202,7911,3320,4730,9511,4646,2611,4672,6419,4742,6201,4643,3514,1722,7422,4644,4399,3214,1011,4
645,4648,6810,4676,4333,7722,7768,9604,4120,2016,8542,7120,6110,2841,2511,5224,4674,6820,4651,6209,7721,1431,2829,4616,3511,5020,2370,8992,9499,7734,8421,5911,7830,3000,1610,1041,8211,4
633,9529,1052,5630,1470,1439,2030,2451,4624,0011,2630,4222,1071,4931,4762,4312,2229,4641,4332,3513,4675,7430,8690,3299,7790,4639,7500,4690,4777,1512,4781,1413,2592,4652,4677,1611,2932,5
010,8411,7312,7279,2410,4741,5222,1712,4617,2000,4614,4322,1201,4634,1013,1061,2221,9311,1130,3700,1013,1082,2593,7731,1105,8623,4723,6190,4334,4939,4511,5913,5912,9525,4618,7835,4649,7
869,4663,4666,1320,4622,7739,4779,6920,1072,8621,4753,7514,2529,7808,6499,4635,4519,1471,6622,4776,3212,4799,6399,9522,1120,9601,1511,7111,7732,7729,4773,3832,1112,2813,2573,2102,1039,2
042,2893,8230,1729,7300,7711,1139,6130,5320,8551,9319,1412,1191,8299,1137,2211,1392,4722,3102,1113,7600,5621,1472,1396,1111,7789,4331,4391,7410,3103,1211,6311,5811,8130,2651,5629,5210,1
114,7696,6610,3201,2806,3220,1080,4311,2751,2822,1301,8111,4221,6420,4312,2332,6402,1410,3230,2572,3011,1451,1210,4612,2521,7302,7220,4623,2031,2020,1241,2013,1330,2433,7400,7012,1624,4
637,2041,8120,3250,3012,8610,2821,7101,2561,2444,7628,7825,2120,8112,1200,2363,2740,2442,1723,4611,1092,2219,2732,2612,5814,2099,5121,2653,9312,2399,7100,7237,7619,8020,1399,9491,7291,9
609,2361,2420,2720,2530,4212,2733,3319,3030,6491,5914,1400,6512,8413,2891,6411,2562,6020,9200,7219,7449,1310,9512,2341,5223,6391,7803,1132,7469,1391,7811,5110,7294,7210,8552,4721,7624,6
831,7500,7617,3213,4615,7762,7796,7851,5920,7515,7460,2362,3011,9609,7479,7310,4763,7010,7405,5030,9004,7774,6832,8422,1115,7894,7300,9601,7704,1492,4647,4724,1230,6612,8553,7305,4613,1
240,7470,7339,3316,2092,8129,3021,3109,0120,1221,2059,2311,9313,8541,1814,4339,4211,7211,1621,6629,2014,1107,7740,7010,4950,4299,1252,2550,7050,2060,9101,7129,5221,7067,6511,3020,1104,1
251,3222,2640,7070,4942,3217,9521,7396,4782,1101,2365,7845,1711,8531,2369,3900,7537,2894,1110,7091,3831,8291,9602,4726,8425,9002,7550,7549,1394,7221,1073,4743,2052,0012,7476,1701,8010,7
478,7070,3211,9099,1250,1629,7713,7684,1042,5813,7645,1501,3314,7304,7564,8292,2060,7205,7532,2594,2814,5829,8990,7460,7562,7765,3240,8720,2223,2752,7795,2110,2443,5590,1085,3522,7687,7
208,7283,5040,8560,7850,7475,2711,7493,7000,1461,7753,7513,7752,7813,7531,6283,7725,7522,7733,7483,2452,6312,7784,8510,4291,7448,7489,3313,2312,2342,6100,2020,4910,6520,7021,8430,1134,7
821,3812,1242,7669,1622,3221,2651,2020,7464,7540,2319,3210,4313,7284,5530,1192,2815,3112,7529,1200,3101,7755,7340,7735,2364,8412,1200,1161,8931,7589,7574,2823,1020,7650,6630,1130,3116,7
770,3311,1920,7465,7010,7770,1450,7830,1100,7855,7777,7697,7640,3107,1770,7837,9103,8700,7854,9571,2001,7813,1393,7790,5310,7660,2010,1003,9107,7510,1117,7760,1003,7310,1490,7820,7407,7
799,0100,1131,0130,7331,3000,0000,0000,0100,1152,3530,0001,0000,9900,1209,2015,9901,3000,7310,0000,2032,1130,1430,7770,1259,7502,1030,2301,9104,1390,1253,0130,2540,1290,1119,0000,3110,0052,7
004,7102,1031,7145,7792,7665,5021,3092,1150,7013,7302,3091,0010,0001,7707,7265,2320,1151,9412,1110,8790,8520,2011,690,7027,1193,7030,7005,2401,1032,1700,7012,2591,1110,2012,7170,7701,72
67,3114,7640,3111,7834,9411,4110,6611,2434,1300,1062,8099,7304,3512,1011,7325,3040,7015,7571,4664,7379,7892,2200,7207,7709,3117,1000,7766,1001,8122,7615,7512,7839,7710,7471,2011,7620,74
11,7702,7705,7484,9321,1271,2670,1452,7135,7473,7301,3521,7300,7027,8911,7701,2600,8991,7900,2895,2391,1103,7509,2313,4213,7395,6430,7300,7463,7027,1641,7503,1493,7605,2910,7007,7754,30
22,7557,7661,9523,2352,7467,2790,7652,7723,7570,5122,7499,7712,3523,2351,7060,1121,2400,7605,2731,7656,3317,2441,7527,7794,9524,2453,7704,7060,8532,7040,7500,7321,2454,7303,7460,7810,11
36,7227,7593,152,7520,3113,3099,2349,7599,7500,7431,7095,2431,7300,7200,7000,7323,1012,8110,1453,7247,2331,7134,7339,7697,7510,7453,7600,5200,7000,7400,7703,7349,0010,7047,7330,7852,610
1,7461,7290,7643,7790,7494,7303,2017,1910,2100,7575,5201,7613,7700,7370,7496,7282,2103,1640,2445,7603,7543,2432,7736,7822,7710,7890,7000,7005,4920,7857,7653,7177,7030,1141,7570,7700,707
1,7005,7505,1110,7701,9420,6621,7341,2304,7099,7772,2343,7223,7017,7296,5012,7009,7455,7429,7600,7472,0114,7234,7542,7551,7560,7236,7090,7849,7744,7309,7159,7693,7590,7743,8424,7901,700
7,7030,7430,7000}
@attribute KAD2
{9999999,1182,7112,9083,4941,1414,1420,4791,6910,2830,4751,4771,1800,4772,4774,4673,4531,7822,1631,1721,4662,8030,4789,4759,3315,4711,7320,7990,4932,4752,4761,1419,4621,5510,4725,1133,1
411,4631,4669,1084,5810,1812,4661,4778,4329,4540,2825,2712,2599,4765,1051,4754,6010,7556,1623,7311,4719,5010,4775,1520,2222,7756,8219,4671,4532,8121,2512,1020,1421,4642,1261,1231,1401,7
420,1500,8622,5229,5520,4636,4632,8559,3312,4638,1091,4665,4729,4619,9329,2849,4764,4520,6202,7911,3320,4730,9511,4646,2611,4672,6419,4742,6201,4643,3514,1722,7422,4644,4399,3214,1011,4
645,4648,6810,4676,4333,7722,7768,9604,4120,2016,8542,7120,6110,2841,2511,5224,4674,6820,4651,6209,7721,1431,2829,4616,3511,5020,2370,8992,9499,7734,8421,5911,7830,3000,1610,1041,8211,4
633,9529,1052,5630,1470,1439,2030,2451,4624,0011,2630,4222,1071,4931,4762,4312,2229,4641,4332,3513,4675,7430,8690,3299,7790,4639,7500,4690,4777,1512,4781,1413,2592,4652,4677,1611,2932,5
010,8411,7312,7279,2410,4741,5222,1712,4617,2000,4614,4322,1201,4634,1013,1061,2221,9311,1130,3700,1013,1082,2593,7731,1105,8623,4723,6190,4334,4939,4511,5913,5912,9525,4618,7835,4649,7
869,4663,4666,1320,4622,7739,4779,6920,1072,8621,4753,7514,2529,7808,6499,4635,4519,1471,6622,4776,3212,4799,6399,9522,1120,9601,1511,7111,7732,7729,4773,3832,1112,2813,2573,2102,1039,2
042,2893,8230,1729,7300,7711,1139,6130,5320,8551,9319,1412,1191,8299,1137,2211,1392,4722,3102,1113,7600,5621,1472,1396,1111,7789,4331,4391,7410,3103,1211,6311,5811,8130,2651,5629,5210,1
114,7696,6610,3201,2806,3220,1080,4311,2751,2822,1301,8111,4221,6420,4312,2332,6402,1410,3230,2572,3011,1451,1210,4612,2521,7302,7220,4623,2031,2020,1241,2013,1330,2433,7400,7012,1624,4
637,2041,8120,3250,3012,8610,2821,7101,2561,2444,7628,7825,2120,8112,1200,2363,2740,2442,1723,4611,1092,2219,2732,2612,5814,2099,5121,2653,9312,2399,7100,7237,7619,8020,1399,9491,7291,9
609,2361,2420,2720,2530,4212,2733,3319,3030,6491,5914,1400,6512,8413,2891,6411,2562,6020,9200,7219,7449,1310,9512,2341,5223,6391,7803,1132,7469,1391,7811,5110,7294,7210,8552,4721,7624,6
831,7500,7617,3213,4615,7762,7796,7851,5920,7515,7460,2362,3011,9609,7479,7310,4763,7010,7405,5030,9004,7774,6832,8422,1115,7894,7300,9601,7704,1492,4647,4724,1230,6612,8553,7305,4613,1
240,7470,7339,3316,2092,8129,3021,3109,0120,1221,2059,2311,9313,8541,1814,4339,4211,7211,1621,6629,2014,1107,7740,7010,4950,4299,1252,2550,7050,2060,9101,7129,5221,7067,6511,3020,1104,1
251,3222,2640,7070,4942,3217,9521,7396,4782,1101,2365,7845,1711,8531,2369,3900,7537,2894,1110,7091,3831,8291,9602,4726,8425,9002,7550,7549,1394,7221,1073,4743,2052,0012,7476,1701,8010,7
478,7070,3211,9099,1250,1629,7713,7684,1042,5813,7645,1501,3314,7304,7564,8292,2060,7205,7532,2594,2814,5829,8990,7460,7562,7765,3240,8720,2223,2752,7795,2110,2443,5590,1085,3522,7687,7
208,7283,5040,8560,7850,7475,2711,7493,7000,1461,7753,7513,7752,7813,7531,6283,7725,7522,7733,7483,2452,6312,7784,8510,4291,7448,7489,3313,2312,2342,6100,2020,4910,6520,7021,8430,1134,7
821,3812,1242,7669,1622,3221,2651,2020,7464,7540,2319,3210,4313,7284,5530,1192,2815,3112,7529,1200,3101,7755,7340,7735,2364,8412,1200,1161,8931,7589,7574,2823,1020,7650,6630,1130,3116,7
770,3311,1920,7465,7010,7770,1450,7830,1100,7855,7777,7697,7640,3107,1770,7837,9103,8700,7854,9571,2001,7813,1393,7790,5310,7660,2010,1003,9107,7510,1117,7760,1003,7310,1490,7820,7407,7
799,0100,1131,0130,7331,3000,0000,0000,0100,1152,3530,0001,0000,9900,1209,2015,9901,3000,7310,0000,2032,1130,1430,7770,1259,7502,1030,2301,9104,1390,1253,0130,2540,1290,1119,0000,3110,0052,7
004,7102,1031,7145,7792,7665,5021,3092,1150,7013,7302,3091,0010,0001,7707,7265,2320,1151,9412,1110,8790,8520,2011,690,7027,1193,7030,7005,2401,1032,1700,7012,2591,1110,2012,7170,7701,72
67,3114,7640,3111,7834,9411,4110,6611,2434,1300,1062,8099,7304,3512,1011,7325,3040,7015,7571,4664,7379,7892,2200,7207,7709,3117,1000,7766,1001,8122,7615,7512,7839,7710,7471,2011,7620,74
11,7702,7705,7484,9321,1271,2670,1452,7135,7473,7301,3521,7300,7027,8911,7701,2600,8991,7900,2895,2391,1103,7509,2313,4213,7395,6430,7300,7463,7027,1641,7503,1493,7605,2910,7007,7754,30
22,7557,7661,9523,2352,7467,2790,7652,7723,7570,5122,7499,7712,3523,2351,7060,1121,2400,7605,2731,7656,3317,2441,7527,7794,9524,2453,7704,7060,8532,7040,7500,7321,2454,7303,7460,7810,11
36,7227,7593,152,7520,3113,3099,2349,7599,7500,7431,7095,2431,7300,7200,7000,7323,1012,8110,1453,7247,2331,7134,7339,7697,7510,7453,7600,5200,7000,7400,7703,7349,0010,7047,7330,7852,610
1,7461,7290,7643,7790,7494,7303,2017,1910,2100,7575,5201,7613,7700,7370,7496,7282,2103,1640,2445,7603,7543,2432,7736,7822,7710,7890,7000,7005,4920,7857,7653,7177,7030,1141,7570,7700,707
1,7005,7505,1110,7701,9420,6621,7341,2304,7099,7772,2343,7223,7017,7296,5012,7009,7455,7429,7600,7472,0114,7234,7542,7551,7560,7236,7090,7849,7744,7309,7159,7693,7590,7743,8424,7901,700
7,7030,7430,7000}
```

Εικόνα 101- Περιγραφή γνωρισμάτων (WEKA)

Στη συνέχεια δημιουργούνται τα αρχεία εκπαίδευσης και ελέγχου. Στα αρχεία αντικαθίσταται τα header από το header που δημιουργήθηκε στο προηγούμενο στάδιο.

Η διαδικασία αυτή είναι απαραίτητη ώστε να είναι δυνατή η αποτίμηση του μοντέλου που δημιουργείται από τα δεδομένα εκπαίδευσης, με τα δεδομένα ελέγχου.

Αλγόριθμος	Train										Test									
	Correctly	Incorrectly	TN	FN	FP	TP	1-ROC	Precision	Recall	F-Measure	Correctly	Incorrectly	TN	FN	FP	TP	1-ROC	Precision	Recall	F-Measure
OneR	99,5745	0,4255	19873	9	76	17	0,591	0,654	0,183	0,286	99,5114	0,4886	25050	9	114	1	0,504	0,1	0,009	0,016
J48	99,9549	0,0451	19882	0	9	84	0,998	1	0,903	0,949	99,5432	0,4568	25059	0	115	0	0,539	0	0	0
RandomTree	99,6546	0,3454	19877	5	64	29	0,84	0,853	0,312		99,5432	0,4568	25058	1	114	1	0,513	0,5	0,009	
Bagging oneR	99,5845	0,4155	19875	7	76	17	0,613	0,708	0,183	0,291	99,5273	0,4727	25055	4	115	0	0,508	0	0	0
Bagging REPTree	99,5645	0,4355	19871	11	76	17	0,806	0,607	0,183	0,281	99,5352	0,4646	25057	2	115	0	0,738	0	0	0
NaiveBayes	99,6496	0,3504	19841	41	29	64	0,966	0,61	0,688	0,646	99,3883	0,6117	25013	46	108	7	0,574	0,132	0,061	0,083
LWL NaiveBayes	99,7247	0,2753	19857	25	30	63	0,972	0,716	0,677	0,696	99,5035	0,4965	25044	15	110	5	0,59	0,25	0,043	0,74
lazy.Kstar	99,7096	0,2904	19867	15	43	50	0,876	0,769	0,538	0,633	99,428	0,572	25025	34	110	5	0,713	0,128	0,043	0,065
multiBoost j48	99,9499	0,5551	19881	1	10	83	0,998	0,988	0,892	0,938	99,5432	0,4568	25059	0	115	0	0,539	0	0	0
part	99,9349	0,0651	19878	4	9	84	0,998	0,955	0,903	0,928	99,5432	0,4568	25059	0	115	0	0,539	0	0	0
SMO	99,9499	0,0501	19881	1	9	81	0,952	0,998	0,903	0,944	99,5432	0,4568	25059	0	115	0	0,5	0	0	0

Εικόνα 102- Εκπαίδευση και έλεγχος μοντέλων με πλήρες σετ δεδομένων (WEKA)

Αλγόριθμος	Train										Test									
	Correctly	Incorrectly	TN	FN	FP	TP	1-ROC	Precision	Recall	F-Measure	Correctly	Incorrectly	TN	FN	FP	TP	1-ROC	Precision	Recall	F-Measure
meta.CostSensitive SMO	99,9299	0,0701	19879	3	11	82	0,941	0,965	0,882		99,5432	0,4568	25059	0	115	0	0,5	0	0	0
meta.CostSensitive RandomForest	99,7146	0,2854	19987	3	54	39	0,988	0,929	0,419	0,771	99,5273	0,4727	25055	4	115	0	0,621	0	0	0
Meta.CostSensitive RotationForest	99,975	0,025	19882	0	5	88	1	1	0,946	0,972	99,5432	0,4568	25059	0	115	0	0,635	0	0	0
meta.CostSensitive oneR	97,0713	29,287	19297	585	0	93	0,985	0,137	1	0,241	99,4598	0,5402	25029	30	106	9	0,539	0,231	0,078	0,117
meta.CostSensitive IBK	99,8298	0,1702	19863	19	15	78	0,912	0,804	0,839	0,821	99,4478	0,5522	25031	28	111	4	0,517	0,125	0,035	0,054
meta.CostSensitive naive Bayes	99,6646	0,3354	19837	45	22	71	0,976	0,612	0,763	0,679	99,159	0,4897	25046	13	108	7	0,671	0,35	0,061	0,104
meta.CostSensitive LWL	97,0713	2,29287	19297	585	0	93	0,999	0,137	1		99,4598	0,5402	19555	327	69	24	0,629	0,068	0,258	0,108
meta.FilteredClassifier iBK+Resample	99,8598	0,1402	19877	5	23	70	0,876	0,933	0,753	0,833	99,5193	0,4807	25052	7	114	1	0,537	0,125	0,009	0,016

Εικόνα 103- Εκπαίδευση και έλεγχος μοντέλων με πλήρες σετ δεδομένων (WEKA)

Αλγόριθμος	Train										Test									
	Correctly	Incorrectly	TN	FN	FP	TP	1-ROC	Precision	Recall	F-measure	Correctly	Incorrectly	TN	FN	FP	TP	1-ROC	Precision	Recall	F-measure
metaCost lazy.lbk	99,2373	0,7627	24961	98	94	21	0,62	0,176	0,183	0,179	98,0175	1,9825	19555	327	69	24	0,629	0,068	0,258	0,108
jRIP	99,5948	0,4052	25048	11	91	24	0,616	0,686	0,209	0,320	97,3967	2,6033	19430	452	68	25	0,624	0,052	0,269	0,088
lazy.KStar	99,4796	0,5204	25023	36	95	20	0,798	0,357	0,174	0,234	98,8686	1,1314	19736	146	80	13	0,754	0,082	0,14	0,103
naiveBayes	99,3485	0,6515	24992	67	97	18	0,802	0,212	0,157	0,180	97,2816	2,7148	19357	525	18	75	0,962	0,125	0,806	0,216

Εικόνα 104- Εκπαίδευση μοντέλων με τα δεδομένα ελέγχου (WEKA)

Αλγόριθμος	Train										Test									
	Correctly	Incorrectly	TN	FN	FP	TP	1-ROC	Precision	Recall	F-measure	Correctly	Incorrectly	TN	FN	FP	TP	1-ROC	Precision	Recall	F-measure
jRIP	99,7447	0,2553	19845	37	14	79	0,89	0,681	0,846	0,756	994,995	0,5005	25046	13	113	2	0,508	0,133	0,017	0,031
lazy.KStar	99,7397	0,2603	19878	4	48	45	0,97	0,918	0,484	0,634	995,432	0,4568	25058	1	114	1	0,73	0,5	0,009	0,017
naiveBayes	99,7347	0,2653	19865	17	36	57	0,988	0,77	0,613	0,683	995,472	0,4528	25059	0	114	1	0,57	1	0,009	0,017
metaCost naiveBayes	99,6095	0,3905	19840	42	36	57	0,962	0,576	0,613	0,594	994,915	0,5085	25042	17	111	4	0,561	0,19	0,035	0,059
Stacking jRip,naiveBayes, J48/SimpleLogistic	99,9499	0,0501	19881	1	9	84	0,988	0,988	0,903	0,944	99,5432	0,4568	25059	0	115	0	0,539	0	0	0

Εικόνα 105- Εκπαίδευση μοντέλων χρησιμοποιώντας μόνο τα σημαντικότερα γνωρίσματα (WEKA)

