ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ (Τ.Ε.Ι.) Α.Μ.Θ
ΤΜΗΜΑ ΛΟΓΙΣΤΙΚΗΣ ΚΑΙ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗΣ
ΠΜΣ 'ΛΟΓΙΣΤΙΚΗ, ΕΛΕΓΚΤΙΚΗ ΚΑΙ ΔΙΕΘΝΕΙΣ ΣΥΝΑΛΛΑΓΕΣ'

ΔΙΠΛΩΜΑΤΙΚΗ ΔΙΑΤΡΙΒΗ

ΕΝΑ ΠΛΑΙΣΙΟ ΓΙΑ ΤΗΝ ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΕΛΕΓΚΤΙΚΑ ΔΕΔΟΜΕΝΑ

της

Αικατερίνης Ιωάννου

Υπεύθυνος καθηγητής: Σταύρος Βαλσαμίδης

Καβάλα, Οκτώβριος, 2020

Εκπονηθείσα Διπλωματική Εργασία απαραίτητη
για τη λήψη του Μεταπτυχιακού Διπλώματος

ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ (Τ.Ε.Ι.) Α.Μ.Θ

ΤΜΗΜΑ ΛΟΓΙΣΤΙΚΗΣ ΚΑΙ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗΣ

ΠΜΣ 'ΛΟΓΙΣΤΙΚΗ, ΕΛΕΓΚΤΙΚΗ ΚΑΙ ΔΙΕΘΝΕΙΣ ΣΥΝΑΛΛΑΓΕΣ'

ΔΙΠΛΩΜΑΤΙΚΗ ΔΙΑΤΡΙΒΗ

ΕΝΑ ΠΛΑΙΣΙΟ ΓΙΑ ΤΗΝ ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΕΛΕΓΚΤΙΚΑ ΔΕΔΟΜΕΝΑ

της

Αικατερίνης Ιωάννου

Υπεύθυνος καθηγητής: Σταύρος Βαλσαμίδης

Καβάλα, Οκτώβριος, 2020

Εκπονηθείσα Διπλωματική Εργασία απαραίτητη
για τη λήψη του Μεταπτυχιακού Διπλώματος

Η παρούσα διπλωματική εργασία
εγκρίνεται για παρουσίαση.

Σταύρος Βαλσαμίδης,

Υπογραφή: ……..

Ημερομηνία: ……..

Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο των απαιτήσεων του Προγράμματος Μεταπτυχιακών Σπουδών ‘Λογιστική, Ελεγκτική και Διεθνείς Συναλλαγές’ του Τμήματος Λογιστικής και Χρηματοοικονομικής του ΤΕΙ Α.Μ.Θ.. Η έγκριση της δεν υποδηλώνει απαραιτήτως και την αποδοχή των απόψεων του συγγραφέα εκ μέρους του ΤΕΙ Α.Μ.Θ..

Βεβαιώνω ότι είμαι αποκλειστικός συγγραφέας της παρούσας μεταπτυχιακής διπλωματικής εργασίας και ότι κάθε βοήθεια που είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία.

Βεβαιώνω, επίσης, ότι έχω σαφώς αναφέρει όλες τις δευτερογενείς πηγές συλλογής δεδομένων τις οποίες χρησιμοποίησα για την συγγραφή της παρούσας εργασίας. Το κείμενο της εργασίας είναι γραμμένο με τα δικά μου λόγια και δεν αποτελεί προϊόν λογοκλοπής από τρίτες πηγές. Σε περίπτωση αυτούσιας αντιγραφής προτάσεων από τρίτες πηγές έχω χρησιμοποιήσει εισαγωγικά.

Αικατερίνη Ιωάννου,

Υπογραφή: ……..

Ημερομηνία: ……..

ΑΦΙΕΡΩΣΗ

Η παρούσα μεταπτυχιακή διατριβή αφιερώνεται στην οικογένειά μου από την οποία αντλώ καθημερινά δύναμη και ενέργεια.

ΠΡΟΛΟΓΟΣ - ΕΥΧΑΡΙΣΤΙΕΣ

Με το πέρας της παρούσας μεταπτυχιακής εργασίας θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή κ. Σταύρο Βαλσαμίδη για τη συνεργασία μας και τη μεγάλη βοήθεια που μου παρείχε σε ένα αντικείμενο αρχικά άγνωστο για μένα, αυτό της εξόρυξης δεδομένων. Αντί προλόγου θα ήθελα να μεταφέρω αυτούσια μια φράση που συνάντησα κατά τη διάρκεια της μελέτης της σχετικής βιβλιογραφίας: Έχουμε κατακλυστεί από δεδομένα, όμως μας λείπει η πληροφορία.

Ένα πλαίσιο για την εξόρυξη γνώσης από ελεγκτικά δεδομένα

Αικατερίνη Ιωάννου, gikomo92@gmail.com

ΤΕΙ Ανατολική Μακεδονίας και Θράκης, Τμήμα Λογιστικής και Χρηματοοικονομικής,
Π.Μ.Σ. 'Λογιστική, Ελεγκτική και Διεθνείς Συναλλαγές', 2020

Επόπτης καθηγητής: Σταύρος Βαλσαμίδης

Περίληψη:

Οι εταιρείες βρίσκονται σε συνεχή αναζήτηση νέων τεχνολογιών με σκοπό να βελτιώσουν τις επιχειρηματικές τους διαδικασίες. Καθώς όμως τα συστήματα πληροφοριών γίνονται ολοένα και πιο περίπλοκα, οι παραδοσιακές τεχνικές ελέγχου μειώνονται ή και εξαλείφονται. Η σημασία της αυτοματοποίησης των ελέγχων και η χρήση της πληροφορικής στους σύγχρονους ελέγχους έχει αυξηθεί σημαντικά τα τελευταία χρόνια λόγω τόσο των τεχνολογικών εξελίξεων όσο και του μεταβαλλόμενου ρυθμιστικού περιβάλλοντος. Η αυτοματοποίηση των επιχειρηματικών διαδικασιών οδήγησε αναπόφευκτα σε αλλαγές στις διαδικασίες και στα πρότυπα ελέγχου. Πρόσθετοι παράμετροι για την υιοθέτηση ενός αυτοματοποιημένου ελέγχου περιλαμβάνουν τη συνεχώς αυξανόμενη πολυπλοκότητα των επιχειρηματικών συναλλαγών καθώς και την αυξανόμενη έκθεση σε κίνδυνο των σύγχρονων επιχειρήσεων. Επομένως, ο σκοπός του ελέγχου, δηλαδή της εξέτασης των οικονομικών καταστάσεων μιας εταιρείας από μια αληθινή και δίκαιη σκοπιά, αυξάνεται σε μεγάλο βαθμό ως προς την πολυπλοκότητα. Για την κάλυψη των απαιτήσεων ενός Πληροφοριακού Συστήματος Ελέγχου, μελετάται η ανάπτυξη ενός πλαισίου για την εξόρυξη πληροφοριών από ήδη γνωστά δεδομένα ελέγχου.

Αναπτύσσεται μια προσέγγιση για την ικανοποίηση αυτών των απαιτήσεων χρησιμοποιώντας τεχνικές εξόρυξης δεδομένων. Αναλύονται εδραιωμένα δεδομένα ελέγχου από ένα γνωστό αποθετήριο τα οποία οδηγούν σε μια δοκιμαστική πρόταση ενός εννοιολογικού μηχανισμού για μια ολοκληρωμένη προσέγγιση ελέγχου. Με τον αυξανόμενο αριθμό περιπτώσεων οικονομικής απάτης, η εφαρμογή τεχνικών εξόρυξης δεδομένων θα μπορούσε να διαδραματίσει σημαντικό ρόλο στη βελτίωση της ποιότητας της διενέργειας ελέγχου στο μέλλον.

Λέξεις - Κλειδιά: Ελεγκτικά δεδομένα, τεχνικές εξόρυξης δεδομένων, ανίχνευση δόλιων επιχειρήσεων

Dissertation Title

Aikaterini Ioannou, gikomo92@gmail.com

Eastern Macedonia Institute of Technology, Department of Accounting and Finance, Postgraduate Program ‘Accounting, Audit and International Transactions’, year

Supervisor: Stavros Valsamidis

Abstract:

Companies are constantly looking for new technologies in order to improve their business processes. But as information systems become more sophisticated, traditional control techniques are being reduced or eliminated. The importance of audit automation and the use of information technology in modern audits has increased significantly in recent years due to both technological developments and the changing regulatory environment. Business process automation has inevitably led to changes in control processes and standards. Additional parameters for adopting an automated control include the ever-increasing complexity of business transactions as well as the increasing risk exposure of modern businesses. Therefore, the purpose of control, that is, to examine the financial statements of a company from a true and fair point of view, is greatly increased in terms of complexity. To meet the requirements of an Audit Information System, the development of a framework for extracting information from already known audit data is being considered. An approach is developed to meet these requirements using data mining techniques. Established control data from a known repository are analyzed which leads to a test proposal of a conceptual mechanism for an integrated control approach. With the growing number of cases of financial fraud, the application of data mining techniques could play an important role in improving the quality of auditing in the future.

Keywords: Audit data, data mining techniques, detection of fraudulent companies

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΚΑΤΑΛΟΓΟΣ ΣΧΕΔΙΑΓΡΑΜΜΑΤΩΝ

ΕΙΣΑΓΩΓΗ ΣΤΟ ΕΡΕΥΝΗΤΙΚΟ ΑΝΤΙΚΕΙΜΕΝΟ: ΣΤΟΧΟΙ & ΕΡΩΤΗΜΑΤΑ

Η απάτη είναι ένα κρίσιμο και διαχρονικό θέμα παγκοσμίως. Οι επιχειρήσεις που καταφεύγουν σε αθέμιτες πρακτικές χωρίς τον φόβο των νομικών επιπτώσεων δημιουργούν σοβαρές συνέπειες τόσο για την οικονομία μίας χώρας όσο και για την ίδια την κοινωνία (Hooda, 2018). Οι ελεγκτικές πρακτικές είναι υπεύθυνες για την ανίχνευση απάτης. Ο έλεγχος ορίζεται ως η διαδικασία εξέτασης των οικονομικών αρχείων κάθε επιχείρησης για να επιβεβαιωθεί ότι οι οικονομικές τους καταστάσεις βρίσκονται σε συμμόρφωση με τους τυποποιημένους λογιστικούς νόμους και τις αρχές (Cosserat, 2009). Η ανίχνευση επιχειρήσεων αναμειγμένων σε απάτες, η ανίχνευση σφαλμάτων και η αποκάλυψη εργαζόμενων ένοχων για τη συμμετοχή τους σε παράνομες συναλλαγές είναι διεργασίες οι οποίες απαιτούν μεγάλη προσοχή και ακρίβεια. Τα εργαλεία ανάλυσης δεδομένων για μια αποτελεσματική διαχείριση απάτης χρειάζονται περισσότερο από ποτέ την ώρα του ελέγχου. Οι δυνατότητες για το πως η ανάλυση δεδομένων μπορεί να βελτιώσει την ποιότητα της όλης διαδικασίας του ελέγχου έχει δημοσιευτεί στο Emerging Assurance Technologies Task Force of the AICPA Assurance Services Executive Committee (ASEC) (Staff, 2014). Γενικά, οι έλεγχοι ταξινομούνται σε δύο κατηγορίες ως εσωτερικός και ως εξωτερικός έλεγχος (Cosserat, 2009). Πιο αναλυτικά, ο εσωτερικός έλεγχος, αν και αποτελεί ανεξάρτητο τμήμα ενός οργανισμού, εδρεύει στον οργανισμό. Ουσιαστικά εκτελείται από υπαλλήλους της εκάστοτε επιχείρησης οι οποίοι είναι υπόλογοι για την εκτέλεση ελέγχων οικονομικών και μη χρηματοοικονομικών καταστάσεων σύμφωνα πάντοτε με τον ετήσιο σχεδιασμό του ελέγχου. Από την άλλη, ο εξωτερικός έλεγχος είναι μια δίκαιη και ανεξάρτητη αρχή τακτικού ελέγχου, η οποία είναι υπεύθυνη για τον ετήσιο νόμιμο έλεγχο των οικονομικών αρχείων. Για παράδειγμα, το έργο τους είναι ο έλεγχος των εσόδων, των δαπανών, λογαριασμών που σχετίζονται με το εμπόριο, τα κέρδη, τα ταμεία έκτακτης ανάγκης, δημόσιους λογαριασμούς κ.λπ. που διατηρούνται σε οποιοδήποτε κυβερνητικό γραφείο. Είναι καθήκον τους να διασφαλίζουν ότι τα κονδύλια που έχουν διατεθεί σε οποιαδήποτε κυβερνητική υπηρεσία έχουν χρησιμοποιηθεί σύμφωνα με το νόμο. Με την επιτυχή ολοκλήρωση μιας διαδικασίας ελέγχου, οι ελεγκτές παραδίδουν στην εταιρεία μια

συνοπτική έκθεση ελέγχου και επιθεώρησης η οποία ονομάζεται παράταγμα ελέγχου (audit paras) και περιλαμβάνει τις λεπτομέρειες όλων των ευρημάτων από τον έλεγχο. Η έκθεση αυτή μπορεί να περιλαμβάνει οικονομικές αποκλίσεις, μη συμμόρφωση των λογιστικών κανόνων, διαρροή εσόδων, ανακριβείς υπολογισμούς κλπ.

Ωστόσο, επειδή τα τελευταία χρόνια επειδή η παραδοσιακή Ελεγκτική στις περισσότερες περιπτώσεις κρίνεται ανεπαρκής, αναπτύχθηκαν τεχνολογικά εργαλεία, μέθοδοι και προγράμματα για την απόσπαση, ανάλυση και έλεγχο των δεδομένων ελέγχου που θα μας βοηθήσουν στον έγκαιρο εντοπισμό σφαλμάτων ή και στην αποτροπή αυτών (Dull, 2006). Επίσης, είναι κοινώς γνωστό ότι μία σύγχρονη επιχείρηση αλληλεπιδρά καθημερινά ένα τεράστιο όγκο δεδομένων (Big Data). Με τη βοήθεια της τεχνολογίας και πιο συγκεκριμένα της επιστήμης της πληροφορικής πλέον η πρόσβαση σε Big data καθίσταται ολοένα και πιο εύκολη. Έτσι, με τον συνδυασμό των επιστημών της πληροφορικής και της ελεγκτικής μπορούμε να εκμεταλλευτούμε Big Data οικονομικού περιεχομένου από επιχειρήσεις για ελεγκτικούς σκοπούς.

Η σύγχρονη εποχή θέτει επιτακτικά νέα και πιο απαιτητικά καθήκοντα στους ελεγκτικούς μηχανισμούς των επιχειρήσεων. Στο παρελθόν δεν ήταν λίγα τα παραδείγματα αποτυχίας. Χαρακτηριστική είναι η περίπτωση του σκανδάλου Enron, το οποίο εκτός των άλλων προκάλεσε και την κατάρρευση της ελεγκτικής εταιρείας Arthur Andersen, ενός από τους πυλώνες του παγκόσμιου ελεγκτικού συστήματος και μέλους της λεγόμενης ομάδας των πέντε μεγάλων ελεγκτών (Big 5). Δυστυχώς οι αποτυχίες δεν περιορίστηκαν εκεί. Η πρόσφατη οικονομική κρίση πυροδοτήθηκε από την κατάρρευση αμερικανικών τραπεζικών κολοσσών, όπως η Fannie Mae και η Freddie Mac, και οι ελεγκτικοί μηχανισμοί απέτυχαν να προβλέψουν και να εμποδίσουν αυτό το φαινόμενο. Οι αρμόδιοι κυβερνητικοί και άλλοι φορείς, στην προσπάθεια τους να αντιμετωπίσουν τέτοια φαινόμενα, θεσπίζουν νέα κανονιστικά πλαίσια, τα οποία διέπουν την εταιρική διακυβέρνηση και ορίζουν ελεγκτικές διαδικασίες. Το έργο των εξωτερικών ελεγκτών είναι ιδιαίτερα δύσκολο, καθώς καλούνται να λάβουν μη δομημένες αποφάσεις σε συνθήκες υψηλού βαθμού αβεβαιότητας. Το έργο τους καθίσταται ακόμα δυσκολότερο σε περιπτώσεις όπου τα διοικητικά στελέχη των επιχειρήσεων εμπλέκονται σε καταχρηστικές πρακτικές. Τα στελέχη αυτά διαθέτουν την κατάλληλη εμπειρία, αλλά και το κίνητρο, ώστε να παρέμβουν και να αποπροσανατολίσουν την ελεγκτική διαδικασία. Οι πολλαπλές δυσκολίες, αλλά και η μεγάλη σημασία του αντικειμένου, καθιστούν τη διαρκή αναβάθμιση των ελεγκτικών πρακτικών αδήριτη

ανάγκη. Σε αυτήν την προσπάθεια, η συμβολή των μεθοδολογιών της Εξόρυξης Δεδομένων μπορεί να αποδειχθεί αποφασιστικής σημασίας. Δύο μεγάλα προβλήματα της Ελεγκτικής, στα οποία βρίσκουν εφαρμογή οι τεχνικές Εξόρυξης Δεδομένων και ειδικότερα οι τεχνικές κατηγοριοποίησης, είναι η πρόβλεψη χρεοκοπίας και ο εντοπισμός παραποιημένων χρηματοοικονομικών καταστάσεων. Η πρόβλεψη χρεοκοπίας είναι ένα από τα σημαντικότερα προβλήματα λήψης αποφάσεων. Εκτός από την Ελεγκτική σχετίζεται και με τραπεζικά ζητήματα, όπως την εκτίμηση πιστοληπτικού κινδύνου. Οι χρεοκοπίες επιχειρήσεων επιφέρουν μεγάλες οικονομικές ζημίες σε επενδυτές και πιστωτές, ενώ σε ακραίες περιπτώσεις μπορούν να επηρεάσουν ολόκληρες κοινωνίες ή και το παγκόσμιο οικονομικό σύστημα. Εξαιτίας της σημασίας των επιπτώσεων, οι εξωτερικοί ελεγκτές είναι υποχρεωμένοι να διατυπώσουν την άποψη τους σχετικά με την ικανότητα να συνεχίσει τις δραστηριότητες της για ένα ουσιαστικό χρονικό διάστημα μετά τη δημοσίευση των οικονομικών της εκθέσεων (σχόλια τύπου going concern). Η υποχρέωση αυτή των ελεγκτών ορίζεται με σαφήνεια στα λεγόμενα Ελεγκτικά Πρότυπα (Statement of Auditing Standards - SAS) και ειδικότερα στα SAS 59, 64, 77 και 96. Η ακαδημαϊκή κοινότητα, θεωρώντας ότι η χρεοκοπία είναι ένα φαινόμενο που εξελίσσεται στη διάρκεια του χρόνου και όχι ένα στιγμιαίο συμβάν, συμβάλλει με τη διατύπωση μοντέλων ικανών να προβλέψουν έγκαιρα τις περιπτώσεις χρεοκοπίας (early warning predictors). Η σχετική έρευνα ξεκίνησε ήδη από τη δεκαετία του ’60 με τις εργασίες των Beaver και Altman. Στη σημερινή εποχή πλήθος ερευνητών έχουν δημοσιεύσει εργασίες στις οποίες χρησιμοποιούν Νευρωνικά Δίκτυα, Μηχανές Διανυσμάτων Υποστήριξης, Δένδρα Αποφάσεων και άλλους κατηγοριοποιητές για την πρόβλεψη της χρεοκοπίας και τα αποτελέσματα είναι πολύ ικανοποιητικά. Οι πρακτικές «μαγειρέματος των βιβλίων» (book cooking practices) είναι ένα μάλλον διαδεδομένο και σε καμία περίπτωση γεωγραφικά περιορισμένο φαινόμενο. Ο Wells (1997) εκτιμά ότι η απάτη κοστίζει στην αμερικανική οικονομία 400 δισεκατομμύρια δολάρια ετησίως, ενώ ο Koskivaara (2004) αποκαλεί το έτος 2002 «φριχτή χρονιά» ως προς την τήρηση των βιβλίων και ισχυρίζεται ότι η χειραγώγηση συνεχίζεται. Ο διαχωρισμός της ιδιοκτησίας από τη διοίκηση στις σύγχρονες μεγάλες επιχειρήσεις δημιουργεί κίνητρα στα διοικητικά στελέχη να δράσουν προς ίδιον όφελος. Η παραποίηση των χρηματοοικονομικών καταστάσεων είναι μια σημαντική πρακτική διοικητικής απάτης και η αντιμετώπιση της εντάσσεται στα καθήκοντα των εξωτερικών ελεγκτών. Ειδικότερα, σύμφωνα με το ελεγκτικό πρότυπο 82 (Statement of Auditing Standards 82 - SAS82) επιβάλλει στους εξωτερικούς ελεγκτές να εκτιμήσουν τον κίνδυνο απάτης κατά τη διάρκεια των ελέγχων. Η

ανάλυση των χρηματοοικονομικών καταστάσεων με χρήση μεθόδων Εξόρυξης Δεδομένων έχει αποδώσει μοντέλα ικανά να εντοπίζουν τις περιπτώσεις απάτης. Ενδεικτικά αναφέρουμε τις εργασίες των Fanning και Cogger (1998) και των Kirkos *et al* (2007).

Λαμβάνοντας υπόψιν όλες τις παραπάνω εισαγωγικές πληροφορίες που συνδυάζουν τα ερευνητικά αντικείμενα της λογιστικής-ελεγκτικής, της πληροφορικής καθώς και της μηχανικής μάθησης, επιλέχθηκε ως στόχος της παρούσας εργασίας η δημιουργία ενός μοντέλου ταξινόμησης που να μπορεί να προβλέψει την δόλια επιχείρηση βάσει των σημερινών αλλά και των ιστορικών παραγόντων κινδύνου. Με άλλα λόγια, το ερευνητικό ερώτημα που τίθεται στην παρούσα μελέτη είναι εάν μπορεί να δημιουργηθεί ένα μοντέλο ελέγχου-ταξινόμησης των επιχειρήσεων με δυνατότητα ανίχνευσης των δόλιων (ύποπτων για εμπλοκή σε απάτη) βάσει έγκυρων οικονομικών δεδομένων που θα εισαχθούν.

ΚΕΦΑΛΑΙΟ 1: ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΑΝΑΣΚΟΠΗΣΗ

1.1. Γενικά στοιχεία

Το παράδειγμα των δεδομένων ελέγχου έχει τεράστιες επιπτώσεις τόσο στα τμήματα πληροφορικής όσο και στα ελεγκτικά τμήματα (Ghasemi *et al.,* 2011). Οι οικονομικές καταστάσεις παράγονται σε αυτοματοποιημένα Συστήματα Λογιστικής Πληροφορίας (AIS) και ο ελεγκτής αντιμετωπίζει αυξημένη πολυπλοκότητα και κινδύνους λόγω της ανάγκης επεξεργασίας συνεχώς αυξανόμενων δεδομένων (Vasarhelyi *et al.,* 2015; Cao *et al.,* 2015; Adamyk *et al.,* 2018). Τα τελευταία 30 χρόνια, τόσο τα πληροφοριακά όσο και τα ελεγκτικά συστήματα έχουν υποστεί ριζικές αλλαγές (Moffitt και Vasarhelyi, 2013). Τα πρότυπα και οι κανονισμοί έχουν γίνει επίσης σε απογοητευτικό βαθμό περίπλοκα. Αλλά υπάρχει μια ισχυρή θεραπεία για τους σημερινούς πονοκεφάλους ελέγχου: ο συνεχής έλεγχος και η αναφορά (Singleton και Singleton, 2005).

Οι οικονομικές καταστάσεις δεν είναι τόσο σημαντικές για τους επενδυτές όσο κάποτε ήταν, καθώς η τεχνολογία έχει αλλάξει τον τρόπο με τον οποίο οι εταιρείες δημιουργούν αξία σήμερα (Gallegos *et al.,* 2004). Ενώ αυτές οι αλλαγές αποτελούν σοβαρές απειλές για την οικονομική βιωσιμότητα του ελέγχου, δημιουργούν επίσης νέες ευκαιρίες και προκλήσεις για τους ελεγκτές ώστε να συνεχίσουν το έργο τους (Gangolly, 2016). Με την λογιστική και την ανταλλαγή δεδομένων σε πραγματικό χρόνο (real-time) να αποκτούν δημοφιλία, τα Εργαλεία Ελέγχου Υποβοηθούμενα από Υπολογιστές (CAATs) καθίστανται ολοένα και πιο απαραίτητα (Zhao *et al.,* 2004). Ενώ συνεχίζουν να αποκτούν τεχνικές γνώσεις και δεξιότητες πληροφορικής, πολλοί ελεγκτές δεν έχουν το χρόνο ή το ενδιαφέρον να γίνουν προγραμματιστές. Στην πιο βασική περίπτωση, οι ελεγκτές της νέας χιλιετίας πρέπει να κατανοήσουν τα βασικά των μηχανογραφημένων συστημάτων, συμπεριλαμβανομένων των βασικών στοιχείων υλικού ενός συστήματος υπολογιστή και της βασικής έννοιας για κάθε πρόγραμμα υπολογιστή (είσοδος-διαδικασία-έξοδος). Ταυτόχρονα, υπάρχουν πολλά περισσότερα για την κατανόηση της τεχνολογίας, συμπεριλαμβανομένων των βασικών στοιχείων της ανάπτυξης συστημάτων, κύκλων ζωής συστημάτων, διαγράμματος ροής διαδικασιών, λογικής προγραμματισμού και συγγραφής σεναρίων για αναλυτικά στοιχεία. Αυτές οι δεξιότητες πρέπει να υπάρχουν σε κάποια πτυχή του προσωπικού ή να ανατίθενται σε εξωτερικούς συνεργάτες (The Institute of Internal Auditors Research Foundation, 2015).

Οι Murphy και Groomer (2004) πρότειναν πως μπορούν να χρησιμοποιηθούν τα πλαίσια τεχνολογίας πληροφοριών (IT), όπως η επεκτάσιμη γλώσσα σήμανσης (XML) και οι υπηρεσίες Ιστού για να διευκολυνθεί ο έλεγχος για την επόμενη γενιά λογιστικών συστημάτων. Οι εναλλακτικές αρχιτεκτονικές για τον έλεγχο που έχουν προταθεί τόσο στα ερευνητικά όσο και στα πρακτικά περιβάλλοντα διερευνώνται από τους Kuhn και Sutton (2010). Συνδυάζουν την πρακτική πραγματικότητα των τεχνολογικών επιλογών και των δομών ERP με την αναδυόμενη θεωρία και έρευνα για μοντέλα συνεχούς διασφάλισης. Η εστίαση τους αφορά στον εντοπισμό των πλεονεκτημάτων και των αδυναμιών κάθε αρχιτεκτονικής μορφής ως βάση για τη διαμόρφωση μιας ερευνητικής ατζέντας που θα μπορούσε να επιτρέψει στους ερευνητές να συμβάλουν στη μελλοντική εξέλιξη τόσο των σχεδιασμών ERP συστημάτων όσο και των στρατηγικών εφαρμογής του ελεγκτή.

Ο Vasarhelyi *et al.,* (2012) συζήτησε την ανάγκη τα Λογιστικά Πληροφοριακά Συστήματα (AIS) να μπορούν να ανταποκριθούν στις επιχειρηματικές ανάγκες που δημιουργούνται από τις γρήγορες αλλαγές στην τεχνολογία. Υποστηρίχθηκε ότι η οικονομία σε πραγματικό χρόνο έχει δημιουργήσει ένα διαφορετικό περιβάλλον μέτρησης, διασφάλισης και επιχειρηματικής απόφασης. Συζητήθηκαν τρεις βασικοί ισχυρισμοί σχετικά με το περιβάλλον μέτρησης στη λογιστική, τη φύση των προτύπων δεδομένων για τη λογιστική που βασίζεται σε λογισμικό και τη φύση της παροχής πληροφοριών, μορφοποιημένης και σημασιολογικής.

Μια εφαρμογή παρακολούθησης του επιπέδου ελέγχου των επιχειρηματικών διαδικασιών (CMBPC) στο τμήμα εσωτερικού ελέγχου της Siemens Corporation των ΗΠΑ περιγράφεται από τους Alles *et al.,* (2018). Μεταξύ των βασικών συμπερασμάτων τους είναι ότι η «τυποποίηση» των ελεγκτικών διαδικασιών και η ελεγκτική κρίση είναι πολύ υποτιμημένη. Επιπλέον, ενώ η εξοικονόμηση κόστους και η σκοπιμότητα αναγκάζουν την εφαρμογή να παρακολουθεί στενά το υφιστάμενο και εγκεκριμένο πρόγραμμα εσωτερικού ελέγχου, ένα ορισμένο επίπεδο επανασχεδιασμού των διαδικασιών ελέγχου είναι αναπόφευκτο λόγω της ανάγκης διαχωρισμού τυποποιήσιμων και μη τυποποιήσιμων τμημάτων του προγράμματος.

Οι Lenz και Hahn (2015) βρίσκουν πρώτοι, τα κοινά θέματα στην εμπειρική βιβλιογραφία. Δεύτερον, συντίθενται τα κύρια θέματα σε ένα μοντέλο που περιλαμβάνει μάκρο και μίκρο παράγοντες που επηρεάζουν την αποτελεσματικότητα του ελέγχου. Τρίτον, δημιουργήθηκαν πολλά υποσχόμενα μελλοντικά ερευνητικά μονοπάτια που μπορεί να ενισχύσουν την αξία του ελέγχου. Η προοπτική «έξω-μέσα» υποδεικνύει τη διάθεση της

απογοήτευσης των ενδιαφερομένων στον έλεγχο: ο έλεγχος διατρέχει τον κίνδυνο περιθωριοποίησης ή πρέπει να αποδεχθεί την πρόκληση να αναδυθεί ως αναγνωρισμένο και ισχυρότερο επάγγελμα (PWC, 2013). Η προτεινόμενη ερευνητική ατζέντα προσδιορίζει εμπειρικά ερευνητικά θέματα που μπορούν να βοηθήσουν τους ελεγκτές να κάνουν τη διαφορά για τον οργανισμό τους, να αναγνωριστούν, να σεβαστούν και να εμπιστευθούν και να βοηθήσουν το επάγγελμα του ελεγκτή στην προσπάθειά του να δημιουργήσει μια μοναδική ταυτότητα.

Όπως αναφέρθηκε στην εισαγωγή, ο έλεγχος ορίζεται ως η διαδικασία εξέτασης των οικονομικών αρχείων κάθε επιχείρησης για να επιβεβαιωθεί ότι οι οικονομικές καταστάσεις τους συμμορφώνονται με τους πρότυπους λογιστικούς νόμους και αρχές (Cosserat και Rodda, 2004). Ταξινομείται σε δύο κατηγορίες, τον εσωτερικό και εξωτερικό έλεγχο (Cosserat, 2009). Ο εσωτερικός έλεγχος, αν και είναι ανεξάρτητο τμήμα ενός οργανισμού, αλλά κατοικεί εντός του οργανισμού. Πρόκειται για υπαλλήλους της επιχείρησης που είναι υπόλογοι για τη διενέργεια ελέγχων οικονομικών και μη χρηματοοικονομικών καταστάσεων σύμφωνα με το ετήσιο πρόγραμμα ελέγχου. Ο εξωτερικός έλεγχος είναι μια δίκαιη και ανεξάρτητη αρχή τακτικού ελέγχου, η οποία είναι υπεύθυνη για τον ετήσιο νόμιμο έλεγχο των οικονομικών αρχείων. Η εταιρεία εξωτερικού ελέγχου έχει καθήκον εμπιστευτικότητας και είναι ζωτικής σημασίας για την ορθή διεξαγωγή των εργασιών.

Υπάρχουν πολλά ζητήματα που σχετίζονται με τα συστήματα ελέγχου και υποστήριξης αποφάσεων (Socea, 2012; Schaltegger και Burritt, 2017). Δεδομένου ότι ο πρωταρχικός στόχος ενός ελεγκτή κατά τη φάση προγραμματισμού ελέγχου είναι να ακολουθήσει μια σωστή αναλυτική διαδικασία για να προσδιορίσει αμερόληπτα και κατάλληλα τις εταιρείες που καταφεύγουν σε υψηλού κινδύνου αθέμιτες πρακτικές, τα προγνωστικά αναλυτικά στοιχεία χρησιμοποιώντας τεχνικές εξόρυξης δεδομένων θα μπορούσαν να παράσχουν αξιόλογες πληροφορίες για τον έλεγχο. Σύμφωνα με μια έρευνα του Tysiac (2015), η ανάλυση δεδομένων έχει ωφελήσει τον εσωτερικό έλεγχο περισσότερο σε σύγκριση με την συνεισφορά της στον εξωτερικό έλεγχο. Μία από τις πιο κοινές εφαρμογές της προγνωστικής ανάλυσης στον έλεγχο είναι η ταξινόμηση μιας ύποπτης εταιρείας. Ο εντοπισμός δόλιων εταιρειών μπορεί να μελετηθεί ως πρόβλημα ταξινόμησης. Ο σκοπός της ταξινόμησης των εταιρειών κατά το προκαταρκτικό στάδιο ενός ελέγχου είναι η μεγιστοποίηση του πιθανού πεδίου δοκιμών των επιχειρήσεων υψηλού κινδύνου που απαιτούν σημαντική έρευνα.

1.2. Η χρήση τεχνικών εξόρυξης δεδομένων

Τεχνικές εξόρυξης δεδομένων έχουν ήδη εφαρμοστεί σε λογιστικά συστήματα πληροφοριών (Gelinas *et al.,* 2017). Οι τεχνικές εξόρυξης δεδομένων παρέχουν μεγάλη βοήθεια στον εντοπισμό της απάτης στη χρηματοοικονομική λογιστική, καθώς η αντιμετώπιση των μεγάλων όγκων δεδομένων και η πολυπλοκότητα των χρηματοοικονομικών δεδομένων είναι μεγάλες προκλήσεις για την ιατροδικαστική λογιστική (Sharma και Panigrahi, 2013). Οι συγγραφείς προτείνουν ένα πλαίσιο που βασίζεται σε τεχνικές εξόρυξης δεδομένων για τον εντοπισμό λογιστικής απάτης. Η αυτόματη ανίχνευση λογιστικής απάτης παρουσιάζεται επίσης από τον Wang (2010). Κατηγοριοποιεί, συγκρίνει και συνοψίζει το σύνολο των δεδομένων, τον αλγόριθμο και τη μέτρηση απόδοσης σε δημοσιευμένα άρθρα και σχετικά με τον εντοπισμό λογιστικής απάτης. Οι τεχνικές εξόρυξης δεδομένων επιτελούν το έργο της εντοπισμού απάτης διαχείρισης που θα μπορούσε να διευκολύνει τους ελεγκτές (Kirkos *et al.,* 2007). Οι εφαρμογές των τεχνικών εξόρυξης δεδομένων στη λογιστική και η πρόταση ενός οργανωτικού πλαισίου για αυτές τις εφαρμογές διερευνώνται από τους Amani και Fadlalla (2017). Δημιουργούν ένα πλαίσιο που συνδυάζει τις δύο γνωστές προοπτικές λογιστικής αναφοράς (αναδρομή και αναζήτηση) και τους τρεις καλά αποδεκτούς στόχους της εξόρυξης δεδομένων (περιγραφή, πρόβλεψη και συνταγή). Το προτεινόμενο πλαίσιο αποκάλυψε ότι ο τομέας της λογιστικής που ωφελήθηκε περισσότερο από την εξόρυξη δεδομένων είναι η διασφάλιση και η συμμόρφωση, συμπεριλαμβανομένης της ανίχνευσης απάτης, της υγείας των επιχειρήσεων και της δικαστικής λογιστικής. Η μέθοδος μηχανικής εκμάθησης ensemble εφαρμόζεται επίσης με επιτυχία για τη βελτίωση της ακρίβειας της ταξινόμησης των εργασιών ελέγχου (Kotsiantis *et al.,* 2006).

Ο στόχος είναι η χρήση της ανάλυσης δεδομένων να καταστεί μια βιώσιμη, αποτελεσματική και επαναλαμβανόμενη διαδικασία (Zhang *et al.,* 2015). Όπως με τις περισσότερες χρήσεις της τεχνολογίας λογισμικού, δεν είναι μια μαγική σφαίρα. Απαιτεί την προσοχή σε άτομα και σε ζητήματα επεξεργασίας, από τη δέσμευση και την υποστήριξη της διοίκησης μέσω της κατάρτισης και της ανάθεσης ρόλων (Lientz και Larssen, 2012).

Η βασική ανάλυση δεδομένων μπορεί να πραγματοποιηθεί χρησιμοποιώντας μια σειρά εργαλείων, όπως υπολογιστικά φύλλα και συστήματα ερωτημάτων και αναφορών βάσεων δεδομένων (Antipova και Rocha, 2018). Υπάρχουν σίγουρα κίνδυνοι από τη χρήση

υπολογιστικών φύλλων, προφανών σε οποιονδήποτε ελεγκτή λόγω της δυσκολίας διασφάλισης της ακεραιότητας των δεδομένων. Τα εργαλεία ανάλυσης γενικού σκοπού έχουν επίσης τους δικούς τους περιορισμούς (Henry και Robinson, 2009). Είναι σαφές ότι η διαδικασία ανάλυσης πρέπει να διαχειρίζεται με τρόπο τέτοιο ώστε να βασίζεται σε έλεγχο, γι 'αυτό το λογιστικό λογισμικό ανάλυσης πρέπει να περιλαμβάνει δυνατότητες όπως: (i) Διατήρηση ασφάλειας και ελέγχου δεδομένων, εφαρμογών και ευρημάτων, καταγραφή όλων των δραστηριοτήτων (iii) τεχνικές ανάλυσης σχεδιασμένες να υποστηρίζουν λογιστικούς στόχους και (iv) αυτοματοποιημένη δημιουργία και εκτέλεση δοκιμών (Bellino *et al.,* 2007).

Το λογισμικό ανοιχτού κώδικα R διαθέτει μία από τις μεγαλύτερες διαθέσιμες βιβλιοθήκες εφαρμογών. Ελεύθερο λογισμικό όπως το R και το Weka χρησιμοποιούνται σε εθνικό επίπεδο σε πανεπιστημιακά μαθήματα και σε ορισμένες εταιρείες έρευνας και τεχνολογίας, αλλά είναι κάπως περιφρονητικά από ελεγκτικές εταιρείες επειδή δεν έχουν επικυρωθεί (Appelbaum, 2017). Αυτές οι ανησυχίες δεν είναι χωρίς αξία, δεδομένου ότι το λογισμικό ανοιχτού κώδικα μπορεί να είναι πιο αδύνατο και λιγότερο φιλικό προς το χρήστη από το ιδιόκτητο λογισμικό, αλλά και πάλι η χρησιμότητά τους δεν πρέπει να αγνοηθεί. Επιπλέον, ενώ μια βασική γνώση στατιστικών και τεχνολογίας πληροφοριών καθίσταται απαραίτητη για όλους τους ελεγκτές, άλλες, πιο εξειδικευμένες λειτουργίες μπορούν να ανατεθούν σε άλλους ειδικούς, ίσως και online.

Ιδιόκτητα εργαλεία όπως η Γλώσσα Ελέγχου ACL και η Διαδραστική Εξαγωγή και Ανάλυση Δεδομένων (IDEA), καθώς και γενικό λογισμικό στατιστικής όπως το Σύστημα Στατιστικής Ανάλυσης (SAS) και το Στατιστικό Πακέτο για τις Κοινωνικές Επιστήμες (SPSS), χρησιμοποιούνται συχνά από μεγάλες επιχειρήσεις (Singleton, 2006; Tysiac, 2015). Επιπλέον, οι δυνατότητες και το εύρος αυτών των πακέτων εξελίσσονται συνεχώς, απαιτώντας οι λογιστές και οι ελεγκτές να έχουν επαρκή γνώση των αναλυτικών στοιχείων (Appelbaum *et al.,* 2016). Αυτή η σύγκλιση πιθανότατα μπορεί να πραγματοποιηθεί με την ανάπτυξη των αναδυόμενων εργαλείων στατιστικής και οπτικοποίησης.

ΚΕΦΑΛΑΙΟ 2: ΕΛΕΓΚΤΙΚΗ ΚΑΙ ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

2.1. Ορισμός και λειτουργίες της εξόρυξης δεδομένων

Η ανάπτυξη της Πληροφορικής έχει δημιουργήσει έναν μεγάλο αριθμό βάσεων δεδομένων, καθώς επίσης και μια πληθώρα δεδομένων, σε διάφορους τομείς. Η έρευνα σε διάφορες βάσεις και η τεχνολογία της πληροφορίας, έχουν με τη σειρά τους οδηγήσει σε μια προσέγγιση για την αποθήκευση και τον χειρισμό αυτών των πολύτιμων δεδομένων για περαιτέρω λήψη αποφάσεων (Ramageri, 2010). Η εξόρυξη δεδομένων (Data Mining) είναι μια διαδικασία εξαγωγής χρήσιμων πληροφοριών και σχεδίων, μέσα από έναν τεράστιο όγκο δεδομένων. Καλείται επίσης ως διαδικασία αποκάλυψης γνώσης, εξόρυξη γνώσης από δεδομένα, εξαγωγή γνώσης ή ανάλυση δεδομένων / προτύπου. Η εξόρυξη δεδομένων είναι μια λογική διαδικασία που χρησιμοποιείται για την αναζήτηση μεγάλου όγκου πληροφοριών για την εύρεση χρήσιμων δεδομένων. Η εξόρυξη δεδομένων (Data Mining) αποτελεί στην ουσία, τη διαδικασία ανεύρεσης «μοτίβων», σε μεγάλα σύνολα δεδομένων, που περιλαμβάνουν μεθόδους σχετικά με τη διασταύρωση της μηχανικής μάθησης, των στατιστικών και των συστημάτων βάσεων δεδομένων. Ο όρος «εξόρυξη δεδομένων» είναι στην πραγματικότητα μια εσφαλμένη ονομασία, αφού ο στόχος είναι η εξαγωγή προτύπων και γνώσεων μέσα από μεγάλες ποσότητες δεδομένων και όχι η εξαγωγή (εξόρυξη) των ίδιων των δεδομένων (Jiawei, και Micheline, 2006).

Ο στόχος αυτής της τεχνικής είναι να βρεθούν μοτίβα που ήταν προηγουμένως άγνωστα. Μόλις εντοπιστούν αυτά τα μοτίβα (πρότυπα), μπορούν να χρησιμοποιηθούν περαιτέρω για να ληφθούν ορισμένες αποφάσεις με σκοπό την ανάπτυξη των επιχειρήσεων τους. Υπάρχουν τρία βήματα που ακολουθούνται κατά τη διαδικασία (Ramageri, 2010):

- Εξερεύνηση
- Ταυτοποίηση μοτίβου (προτύπου)
- Ανάπτυξη

Αναλυτικότερα τα βήματα περιγράφονται ως εξής:

Εξερεύνηση: Στο πρώτο βήμα της εξερεύνησης δεδομένων, τα δεδομένα καθαρίζονται και μετατρέπονται σε άλλη μορφή, ενώ καθορίζονται σημαντικές μεταβλητές και στη συνέχεια η φύση των δεδομένων με βάση το πρόβλημα (Ramageri, 2010).

Αναγνώριση μοτίβων: Μόλις εξερευνηθούν, επεξεργαστούν και καθοριστούν τα δεδομένα για τις συγκεκριμένες μεταβλητές, το δεύτερο βήμα είναι να διαμορφωθεί η αναγνώριση προτύπου. Στη συνέχεια, προσδιορίζονται και επιλέγονται τα πρότυπα εκείνα, που εξασφαλίζουν την καλύτερη πρόβλεψη (Ramageri, 2010).

Ανάπτυξη: Τα μοτίβα αναπτύσσονται για το επιθυμητό αποτέλεσμα (Ramageri, 2010).

Όπως τονίστηκε και προηγουμένως, στα πλαίσια της παρούσας μεταπτυχιακής διατριβής θα πραγματοποιηθεί ανάλυση χρησιμοποιώντας εργαλεία εξόρυξης δεδομένων. Η εξόρυξη δεδομένων είναι μια επαναληπτική διαδικασία δημιουργίας προγνωστικών και περιγραφικών μοντέλων, αποκαλύπτοντας προηγουμένως άγνωστες τάσεις και μοτίβα σε τεράστιες ποσότητες δεδομένων, προκειμένου να εξαχθούν χρήσιμες πληροφορίες και να υποστηριχθεί η λήψη αποφάσεων (Kantardzic, 2003). Οι πιο δημοφιλείς τεχνικές για την εξόρυξη δεδομένων (DM) είναι η ομαδοποίηση, η ταξινόμηση και η εύρεση κανόνων συσχέτισης (Han *et al.,* 2011).

Οι μέθοδοι ταξινόμησης χρησιμοποιούν ένα σύνολο δεδομένων εκπαίδευσης για να εκτιμήσουν ορισμένες παραμέτρους ενός μαθηματικού μοντέλου που θα μπορούσαν θεωρητικά να εκχωρήσουν κάθε περίπτωση από ένα νέο σύνολο δεδομένων σε μια συγκεκριμένη κατηγορία. Με άλλα λόγια, το σύνολο δεδομένων εκπαίδευσης χρησιμοποιείται για να εκπαιδεύσει την τεχνική ταξινόμησης στο πως να εκτελέσει την ταξινόμησή της (Witten *et al.,* 2016). Υπάρχουν διάφορες μέθοδοι ταξινόμησης που εφαρμόζονται στο WEKA, όπως οι ZeroR, OneR, PART κ.λπ. Ο αλγόριθμος OneR χρησιμοποιεί το χαρακτηριστικό ελάχιστου σφάλματος για πρόβλεψη και διακριτοποίηση αριθμητικών ιδιοτήτων (Holte, 1993). Σε αυτήν την τεχνική, θα ανακαλυφθούν τα χαρακτηριστικά που περιγράφουν καλύτερα την ταξινόμηση.

Η ομαδοποίηση-συσταδοποίηση αναφέρεται σε μεθόδους όπου το εκπαιδευτικό σύνολο δεν είναι διαθέσιμο. Έτσι, δεν υπάρχει προηγούμενη γνώση σχετικά με τα δεδομένα ώστε να μπορούν να εκχωρηθούν σε συγκεκριμένες ομάδες. Σε αυτήν την περίπτωση, τεχνικές ομαδοποίησης μπορούν να χρησιμοποιηθούν για να χωρίσουν ένα σύνολο άγνωστων περιπτώσεων σε ομάδες. Το βήμα ομαδοποίησης περιέχει ομαδοποίηση ψηφιοποίησης με τη χρήση του αλγορίθμου k-means (MacQueen, 1967; Kaufmann και Rousseeuw, 2009) για

μη εποπτευόμενη μάθηση, που ονομάζεται SimpleKMeans στο WEKA. Το K-means είναι ένας αποτελεσματικός αλγόριθμος διαμέρισης που αποσυνθέτει το σύνολο δεδομένων σε ένα σύνολο συστάδων k disjoint. Είναι ένας επαναλαμβανόμενος αλγόριθμος στον οποίο τα αντικείμενα μετακινούνται μεταξύ των διαφόρων συστάδων μέχρι να φτάσουν στο επιθυμητό σύνολο συστάδων. Με αυτόν τον αλγόριθμο επιτυγχάνεται ένας μεγάλος βαθμός ομοιότητας για τα αντικείμενα του ίδιου συμπλέγματος και μια μεγάλη διαφορά αντικειμένων, τα οποία ανήκουν σε διαφορετικές συστάδες. Επιπλέον, ο αλγόριθμος αυτός ομαλοποιεί αυτόματα αριθμητικά χαρακτηριστικά όταν πραγματοποιεί υπολογισμούς απόστασης.

Σύμφωνα με τους Linoff και Berry (2011) η εξόρυξη σχέσεων (rule mining) είναι μια τεχνική που ανακαλύπτει σχέσεις μεταξύ μεταβλητών, σε ένα σύνολο δεδομένων με μεγάλο αριθμό μεταβλητών. Υπάρχουν τέσσερις τύποι εξόρυξης σχέσεων: εξόρυξη κανόνων συσχέτισης, εξόρυξη συσχέτισης, εξόρυξη διαδοχικών προτύπων και εξόρυξη δεδομένων αιτιώδους συνάφειας. Σε αυτή την εργασία επικεντρωνόμαστε στην εξόρυξη κανόνων σύνδεσης (Liu *et al.,* 1998). Η εξόρυξη κανόνων σύνδεσης είναι μια από τις πιο καλά μελετημένες εργασίες εξόρυξης δεδομένων. Ανακαλύπτει σχέσεις μεταξύ χαρακτηριστικών σε βάσεις δεδομένων, παράγοντας δηλώσεις εάν-τότε (if-then) σχετικά με τιμές χαρακτηριστικών (Agarwal *et al.,* 1993). Ένας κανόνας συσχέτισης X → Y εκφράζει μια στενή συσχέτιση μεταξύ των στοιχείων σε μια βάση δεδομένων, όπου οι συναλλαγές στη βάση δεδομένων όπου εμφανίζεται το X, υπάρχει επίσης μεγάλη πιθανότητα να εμφανίζεται και το Y επίσης. Σε έναν κανόνα συσχέτισης τα X και Y ονομάζονται το προηγούμενο και το συνακόλουθο του κανόνα αντίστοιχα. Η ισχύς ενός τέτοιου κανόνα μετράται από τις τιμές της υποστήριξης και της εμπιστοσύνης του. Η εμπιστοσύνη του κανόνα είναι το ποσοστό συναλλαγών με το προηγούμενο X στη βάση δεδομένων που περιέχουν επίσης το συνακόλουθο Y. Η υποστήριξη του κανόνα είναι το ποσοστό συναλλαγών στη βάση δεδομένων που περιέχει τόσο το προηγούμενο X όσο και το συνακόλουθο Y σε όλες τις συναλλαγές στη βάση δεδομένων. Υπάρχουν αρκετοί αλγόριθμοι ανακάλυψης κανόνων συσχέτισης, αλλά ο αλγόριθμος Apriori προτιμάται ως ο πιο δημοφιλής και αποτελεσματικός αλγόριθμος για την εξεύρεση κανόνων συσχέτισης έναντι του πίνακα διακριτών λογιστικών δεδομένων (Agrawal και Srikant, 1994). Ο Apriori είναι ο πιο γνωστός αλγόριθμος για τους κανόνες συσχέτισης. Χρησιμοποιεί μια στρατηγική αναζήτησης πρώτης έκτασης για τη μέτρηση της υποστήριξης των συνόλων στοιχείων. Επίσης, χρησιμοποιεί μια συνάρτηση δημιουργίας υποψηφίων, η οποία εκμεταλλεύεται την

ιδιότητα υποστήριξης του προς τα κάτω κλεισίματος. Επαναληπτικά μειώνει την ελάχιστη υποστήριξη μέχρι να βρει τον απαιτούμενο αριθμό κανόνων με τη δοθείσα ελάχιστη εμπιστοσύνη.

Υπάρχουν διαφορετικές τεχνικές κατηγοριοποίησης για την εξόρυξη κανόνων σύνδεσης. Οι περισσότερες από τις υποκειμενικές προσεγγίσεις περιλαμβάνουν τη συμμετοχή των χρηστών προκειμένου να εκφράσουν, σύμφωνα με τις προηγούμενες γνώσεις του, ποιοι κανόνες τον ενδιαφέρουν. Μία τεχνική βασίζεται στην απροσδόκητη εξέλιξη και στη δυνατότητα δράσης (Liu *et al.,* 1996; Liu *et al.,* 2000). Το απροσδόκητο εκφράζει ποιοι κανόνες είναι ενδιαφέροντες εάν είναι άγνωστοι στον χρήστη ή έρχονται σε αντίθεση με τις γνώσεις του χρήστη. Η δυνατότητα δράσης εκφράζει ότι οι κανόνες είναι ενδιαφέροντες εάν οι χρήστες μπορούν να κάνουν κάτι μαζί τους προς όφελός τους. Ο αριθμός των κανόνων μπορεί να μειωθεί μόνο σε απροσδόκητους και σε κανόνες που ήδη είναι σε ισχύ. Μια άλλη τεχνική προτείνει τη διαίρεση των ανακαλυφθέντων κανόνων σε τρεις κατηγορίες (Minaei-Bidgoli *et al.,* 2004). (1) Αναμενόμενο και προηγουμένως γνωστό: Αυτός ο τύπος κανόνα επιβεβαιώνει τις πεποιθήσεις των χρηστών και μπορεί να χρησιμοποιηθεί για την επικύρωση της προσέγγισης που χρησιμοποιείται. Αν και ίσως είναι ήδη γνωστοί, πολλοί από αυτούς τους κανόνες εξακολουθούν να είναι χρήσιμοι για τον χρήστη ως μια μορφή εμπειρικής επαλήθευσης των προσδοκιών. (2) Μη αναμενόμενο: Αυτός ο τύπος κανόνα έρχεται σε αντίθεση με τις πεποιθήσεις των χρηστών. Αυτή η ομάδα απροσδόκητων συσχετίσεων μπορεί να παρέχει ενδιαφέροντες κανόνες, ωστόσο το ενδιαφέρον τους και η πιθανή δυνατότητα δράσης τους απαιτούν περαιτέρω διερεύνηση. (3) Άγνωστος: Αυτός ο τύπος κανόνα δεν ανήκει σαφώς σε καμία κατηγορία και πρέπει να κατηγοριοποιείται από ειδικούς για συγκεκριμένους τομείς. Το σύστημα Weka διαθέτει διάφορους αλγόριθμους ανακάλυψης κανόνων συσχέτισης (Hipp *et al.,* 2000). Ο αλγόριθμος Apriori θα χρησιμοποιηθεί για την εύρεση κανόνων συσχέτισης σχετικά με τα διακριτά δεδομένα (Agrawal και Srikant, 1994).

2.1. Στόχοι της εφαρμογής εξόρυξης δεδομένων

Ο στόχος της εξόρυξης δεδομένων (Data Mining) είναι να εντοπιστούν έγκυροι, νέοι, δυνητικά χρήσιμοι και κατανοητοί συσχετισμοί και πρότυπα, σε υπάρχοντα δεδομένα. Η εύρεση χρήσιμων μοτίβων στα δεδομένα, είναι γνωστή με διαφορετικά ονόματα (συμπεριλαμβανομένου της εξόρυξης δεδομένων) σε διάφορες κοινότητες (π.χ. εξαγωγή

γνώσης, ανακάλυψη πληροφοριών, συλλογή πληροφοριών, και επεξεργασία προτύπων δεδομένων) (Fayyad *et al.,* 1996). Ο όρος «εξόρυξη δεδομένων» χρησιμοποιείται κυρίως από τους στατιστικολόγους, τους ερευνητές βάσεων δεδομένων, τα πληροφοριακά συστήματα MIS και τις επιχειρηματικές κοινότητες. Ο όρος «Discovery Knowledge in Databases» (KDD) χρησιμοποιείται γενικά για να αναφερθεί στη συνολική διαδικασία ανεύρεσης χρήσιμων γνώσεων από δεδομένα, που συνιστά ένα συγκεκριμένο βήμα σε αυτή τη διαδικασία (Fayyad *et al.,* 1996; Han *et al*., 2000). Τα πρόσθετα βήματα στη διαδικασία KDD, όπως η προετοιμασία δεδομένων, η επιλογή δεδομένων, ο καθαρισμός των δεδομένων και η σωστή ερμηνεία των αποτελεσμάτων της διαδικασίας εξόρυξης δεδομένων, εξασφαλίζουν ότι οι χρήσιμες γνώσεις προέρχονται από τα δεδομένα.

Η εξόρυξη δεδομένων είναι μια επέκταση της παραδοσιακής ανάλυσης δεδομένων και των στατιστικών προσεγγίσεων, δεδομένου ότι ενσωματώνει αναλυτικές τεχνικές οι οποίες προέρχονται από μια σειρά επιστημονικών κλάδων που περιλαμβάνουν, αλλά δεν περιορίζονται για: α) αριθμητική ανάλυση, β) αντιστοίχιση προτύπων και περιοχές τεχνητής νοημοσύνης όπως μηχανική μάθηση, γ) νευρωνικά δίκτυα και γενετικοί αλγόριθμοι.

Ενώ πολλές εργασίες εξόρυξης δεδομένων ακολουθούν μια παραδοσιακή προσέγγιση υπολογισμού που βασίζεται στην υπόθεση, είναι συνηθισμένο να χρησιμοποιείται μια ευκαιριακή προσέγγιση που βασίζεται σε δεδομένα, η οποία ενθαρρύνει τους αλγορίθμους ανίχνευσης προτύπων να βρίσκουν χρήσιμες τάσεις, μοτίβα και σχέσεις. Ουσιαστικά, οι δύο τύποι προσεγγίσεων εξόρυξης δεδομένων διαφέρουν ως προς το α) αν επιδιώκουν να δημιουργήσουν μοντέλα ή, β) να βρουν μοτίβα.

Η πρώτη προσέγγιση, που αφορά τη δόμηση μοντέλων, είναι, εκτός από τα προβλήματα που είναι εγγενή από τα μεγάλα μεγέθη των συνόλων δεδομένων, παρόμοια με τις συμβατικές διερευνητικές στατιστικές μεθόδους. Ο στόχος είναι να παραχθεί μια συνολική σύνοψη ενός συνόλου δεδομένων για να προσδιοριστούν και να περιγραφούν τα κύρια χαρακτηριστικά του σχήματος της κατανομής. Παραδείγματα τέτοιων μοντέλων περιλαμβάνουν την κατάτμηση ανάλυσης συστάδων ενός συνόλου δεδομένων, ένα μοντέλο παλινδρόμησης για πρόβλεψη και ένας κανόνας ταξινόμησης που βασίζεται σε «δέντρα». Κατά τη δόμηση μοντέλων, γίνεται μερικές φορές διάκριση μεταξύ εμπειρικών και μηχανιστικών μοντέλων. Το πρότερο (επίσης μερικές φορές αποκαλούμενο «επιχειρησιακό») επιδιώκει να μοντελοποιήσει σχέσεις χωρίς να τους στηρίξει σε οποιαδήποτε υποκείμενη θεωρία. Το τελευταίο (μερικές φορές αποκαλούμενο «ουσιαστικό» ή «φαινομενολογικό») βασίζεται σε

κάποια θεωρία ή μηχανισμό για την υποκείμενη διαδικασία δημιουργίας δεδομένων. Η εξόρυξη δεδομένων, σχεδόν εξ ορισμού, ασχολείται κυρίως με τη λειτουργία.

Ο δεύτερος τύπος προσέγγισης εξόρυξης δεδομένων, η ανίχνευση προτύπων, επιδιώκει να εντοπίσει μικρές (αλλά ενδεχομένως σημαντικές) αναχωρήσεις από τον κανόνα, για να ανιχνεύσει ασυνήθιστα πρότυπα συμπεριφοράς. Παραδείγματα περιλαμβάνουν ασυνήθιστα μοντέλα δαπανών στη χρήση πιστωτικών καρτών (για ανίχνευση απάτης), σποραδικές κυματομορφές στα ίχνη EEG και αντικείμενα με μοτίβα χαρακτηριστικών σε αντίθεση με άλλα. Γενικά, οι βάσεις δεδομένων των επιχειρήσεων, αποτελούν ένα μοναδικό πρόβλημα για την εξόρυξη προτύπων λόγω της πολυπλοκότητας τους. Η πολυπλοκότητα προκύπτει από ανωμαλίες όπως την ασυνέχεια, το Θόρυβο, την αμφισημία και την ατέλεια (Fayyad *et al.,* 1996). Και ενώ οι περισσότεροι αλγόριθμοι εξόρυξης δεδομένων είναι σε θέση να διαχωρίσουν τα αποτελέσματα τέτοιων ασύνδετων χαρακτηριστικών για τον προσδιορισμό τον πραγματικού προτύπου, η προγνωστική δύναμη των αλγορίθμων εξόρυξης μπορεί να μειωθεί, καθώς ο αριθμός αυτών των ανωμαλιών αυξάνεται.

2.3. Ιστορική αναδρομή

Από τους αρχαίους χρόνους, οι άνθρωποι αναζητούσαν χρήσιμες πληροφορίες από τα δεδομένα με το χέρι. Ωστόσο, με τον ταχέως αυξανόμενο όγκο δεδομένων στη σύγχρονη εποχή, απαιτούνται πιο αυτόματες και αποτελεσματικές προσεγγίσεις εξόρυξης. Οι πρώιμες μέθοδοι όπως το Θεώρημα του Bayes στις δεκαετίες του 1700 και η ανάλυση παλινδρόμησης κατά τις δεκαετίες του 1800, ήταν μερικές από τις πρώτες τεχνικές πού χρησιμοποιήθηκαν για τον προσδιορισμό των μορφών στα δεδομένα.

Μετά από το 19° αιώνα, με την εξάπλωση, την έντονη παρουσία και τη συνεχώς αναπτυσσόμενη δύναμη της τεχνολογίας των υπολογιστών, η συλλογή δεδομένων και η αποθήκευση τους, αυξήθηκαν αξιοσημείωτα. Καθώς τα σύνολα δεδομένων είχαν αυξηθεί σε μέγεθος και πολυπλοκότητα, η άμεση ανάλυση δεδομένων εμφανιζόταν όλο και περισσότερο συνδεδεμένη με την αυτόματη επεξεργασία δεδομένων. Αυτό βοηθήθηκε από άλλες ανακαλύψεις στην επιστήμη των υπολογιστών, όπως τα νευρωνικά δίκτυα, την ομαδοποίηση, τους γενετικούς αλγόριθμούς στη δεκαετία τον 1950, αλλά και τα «δέντρα απόφασης» στη δεκαετία του 1960 και τα μηχανήματα φορέα υποστήριξης στη δεκαετία τον '80. Η εξόρυξη δεδομένων είναι η διαδικασία εφαρμογής αυτών των μεθόδων σε σύνολα δεδομένων, με σκοπό την αποκάλυψη κρυφών μοτίβων (Kantardzic, 2003). Η τεχνολογία

εξόρυξης δεδομένων, έχει χρησιμοποιηθεί εδώ και πολλά χρόνια από πολλούς τομείς όπως επιχειρήσεις, επιστημονικές κοινότητες, καθώς και κυβερνήσεις. Χρησιμοποιήθηκε επίσης για τη διερεύνηση όγκων δεδομένων, όπως πληροφορίες για τους ταξιδιώτες αεροπορικών εταιρειών, δεδομένα πληθυσμού και στοιχεία μάρκετινγκ, για τη δημιουργία αναφορών έρευνας αγοράς, παρόλο που η αναφορά αυτή μερικές φορές δεν θεωρείται ως εξόρυξη δεδομένων.

Κατά τη δεκαετία τον 1960, η τεχνολογία βάσεων δεδομένων και η τεχνολογία των πληροφοριών αναπτύχθηκαν σταδιακά από το βασικό σύστημα επεξεργασίας εγγράφων, σε ένα πιο περίπλοκο και πιο ισχυρό σύστημα βάσεων δεδομένων. Για παράδειγμα η ιεραρχική βάση δεδομένων και η βάση δεδομένων δικτύου, είναι τυπικά αντιπροσωπευτικά αυτής της εποχής, με ελάχιστη ανεξαρτησία και αφαίρεση δεδομένων. Κατά τη δεκαετία τον 1970, εμφανίζονται σχεσιακές βάσεις δεδομένων, επιτρέποντας στους χρήστες πρόσβαση σε μια ευέλικτη γλώσσα και διεπαφή πρόσβασης δεδομένων, ενώ η τεχνολογία OLTP καθιστά την εφαρμογή τεχνολογίας σχεσιακής βάσης δεδομένων, δημοφιλή. Μέσα στη δεκαετία του '80, η άνοδος ενός ισχυρού συστήματος βάσεων δεδομένων, έρχεται να προτείνει πολλά προηγμένα μοντέλα δεδομένων. Μετά το 2000, η δυνατότητα αποθήκευσης μεγάλων ποσοτήτων δεδομένων υπερβαίνει την ικανότητα ανάλυσης και κατανόησης τον ανθρώπου, ενώ δεν υπάρχει κατάλληλο εργαλείο για να βοηθήσει στην εξαγωγή πληροφοριών και γνώσεων από τα δεδομένα. Η ύπαρξη συγκεκριμένων προτύπων και κανόνων μπορεί να βρεθεί μέσω εργαλείων εξόρυξης δεδομένων σε μεγάλο αριθμό δεδομένων, τα οποία μπορούν να παράσχουν τις απαραίτητες πληροφορίες για την εμπορική δραστηριότητα, την επιστημονική έρευνα και την ιατρική έρευνα και πολλούς άλλους τομείς.

Η εξόρυξη δεδομένων περιλαμβάνει συνήθως τέσσερις κατηγορίες καθηκόντων: 1) ταξινόμηση, (κατανομή των δεδομένων σε προκαθορισμένες ομάδες, 2) ομαδοποίηση, (ταξινόμηση σε ομάδες που δεν είναι προκαθορισμένες, οπότε ο αλγόριθμος Θα προσπαθήσει να ομαδοποιήσει παρόμοια στοιχεία μαζί, 3) παλινδρόμηση, (εύρεση μιας συνάρτησης η οποία διαμορφώνει τα δεδομένα με το ελάχιστο σφάλμα, και, 4) σύνδεση κανόνα μάθησης και αναζήτησης σχέσεων μεταξύ μεταβλητών. Σύμφωνα με τον Han και τον Kamber (2001), οι λειτουργίες εξόρυξης δεδομένων περιλαμβάνουν τον χαρακτηρισμό δεδομένων, τη διάκριση δεδομένων, την ανάλυση συσχέτισης, την ταξινόμηση, τη ομαδοποίηση και την ανάλυση εξέλιξης δεδομένων.

Ο χαρακτηρισμός των δεδομένων είναι μια σύνοψη των γενικών χαρακτηριστικών ή χαρακτηριστικών μιας κατηγορίας στόχων δεδομένων. Η διάκριση δεδομένων είναι μια

σύγκριση των γενικών χαρακτηριστικών των αντικειμένων τάξης στόχου με τα γενικά χαρακτηριστικά των αντικειμένων από ένα σύνολο κατηγοριών αντίθεσης. Η ανάλυση της σύνδεσης είναι η ανακάλυψη κανόνων σύνδεσης πού εμφανίζουν συνθήκες χαρακτηριστικού-τιμής που συμβαίνουν συχνά μαζί σε ένα συγκεκριμένο σύνολο δεδομένων. Η ταξινόμηση είναι η διαδικασία εύρεσης ενός συνόλου μοντέλων ή λειτουργιών πού περιγράφουν και διακρίνουν τάξεις (κλάσεις) ή έννοιες δεδομένων, με σκοπό να είναι σε θέση να χρησιμοποιήσουν το μοντέλο για να προβλέψουν την κατηγορία αντικειμένων των οποίων η ετικέτα τάξεως είναι άγνωστη. Η ομαδοποίηση αναλύει αντικείμενα δεδομένων χωρίς να συμβουλευτεί ένα γνωστό μοντέλο τάξεως. Η ανάλυση εξέλιξης δεδομένων περιγράφει και μοντελοποιεί τάσεις για αντικείμενα των οποίων η συμπεριφορά μεταβάλλεται με το χρόνο (Han *et al.,* 2001).

Οι Singleton και Singleton (2005) υποστηρίζουν ότι με την εξέλιξη της τεχνολογίας ολοένα και πιο πολύπλοκα συστήματα δημιουργούνται με αποτέλεσμα το έργο των ελεγκτών να δυσχεραίνει. Η πάγια πρακτική του ανά καθορισμένα χρονικά διαστήματα ελέγχου, φαίνεται πως δεν αποδίδει σε πολλές περιπτώσεις, με αποτέλεσμα να είναι επιτακτική η εύρεση νέων μεθόδων ελέγχου προκειμένου οι ελεγκτές να καταλήγουν με μεγαλύτερη ακρίβεια και σε μικρό χρονικό διάστημα στο επιθυμητό αποτέλεσμα (Zhao *et al.,* 2004). Η δυσκολία της παραδοσιακής πρακτικής έγκειται στο γεγονός πως το μέγεθος των δεδομένων που καλείται ένας ελεγκτής να διαχειριστεί συνεχώς αυξάνεται, με αποτέλεσμα να απαιτείται περισσότερο προσωπικό, χρόνος και χρήμα για την ολοκλήρωση ενός ελέγχου (Global Technology Audit Guide, 2015).

Η ανάλυση δεδομένων με τεχνικές εξόρυξης πληροφορίας σε σύνολο δεδομένων οικονομικών καταστάσεων εταιριών οι οποίες ελέγχονται από Ορκωτούς Ελεγκτές Λογιστές έχει μελετηθεί και από την Ματιάκη (2007). Η ανάλυση αυτή οδήγησε στην δημιουργία ενός συστήματος υποστήριξης απόφασης για τη σύνταξη πιστοποιητικού και έκθεσης ελεγκτών με παρατηρήσεις ανάλογα με την μορφή οργάνωσης της εταιρίας. Η Εξόρυξη Δεδομένων παρέχει τεχνικές για την εξαγωγή γνώσης από μεγάλους όγκους δεδομένων όπως τα Νευρωνικά Δίκτυα και τα Δένδρα Αποφάσεων που επιτρέπουν τη δημιουργία μοντέλων πρόβλεψης (Κύρκος, 2007) για την αντιμετώπιση προβλημάτων που άπτονται της Λογιστικής και της Ελεγκτικής όπως των χρηματοοικονομικών καταστάσεων, όπου τα διοικητικά στελέχη των επιχειρήσεων μπορούν να παραποιούν οικονομικά στοιχεία με στόχο την εξαπάτηση των μετόχων, των πιστωτών, ή των φορολογικών αρχών. Αξίζει να σημειώσουμε ότι νευρωνικά δίκτυα χρησιμοποιούν και αλγόριθμοι που συνδυάζουν την

επιστήμη της ελεγκτικής και της μηχανικής μάθησης. Η μηχανική μάθηση έλαβε μεγάλη προσοχή στην ανάλυση δεδομένων καθώς προσφέρει νέες υπολογιστικές και επιστημολογικές τεχνικές για την παραγωγή καλύτερων αποτελεσμάτων. Η μηχανική μάθηση προτείνει διάφορους αλγορίθμους που προκύπτουν από τον τομέα της στατιστικής και της τεχνητής νοημοσύνης. Πολλοί ερευνητές έχουν χρησιμοποιήσει αλγόριθμους όπως το τεχνητό νευρωνικό δίκτυο, την υλικοτεχνική παλινδρόμηση, τα δέντρα αποφάσεων και τα δίκτυα πίστης Bayesian με σκοπό την ανίχνευση της απάτης στη διαχείριση οικονομικών καταστάσεων. Η μέθοδος της μηχανικής μάθησης εφαρμόζεται επίσης με επιτυχία για τη βελτίωση της ακρίβειας ταξινόμησης της διαδικασίας του ελέγχου. Αλγόριθμοι μηχανικής μάθησης όπως η λογιστική παλινδρόμηση, το πιθανοτικό νευρωνικό δίκτυο, ο γενετικός αλγόριθμος κ.λπ. έχουν συνδυαστεί επίσης με μεθόδους επιλογής χαρακτηριστικών προκειμένου να αποδειχθεί η χρησιμότητά τους στην ανίχνευση απάτης σε Κινεζικές επιχειρήσεις. Σε μια ανασκόπηση των εργαλείων ανάλυσης δεδομένων, όπως για την πρόβλεψη απάτης, οι ερευνητές απαρίθμησαν αλγόριθμους όπως το νευρωνικό δίκτυο, το δέντρο αποφάσεων, το Bayesian δίκτυο κ.λπ. ως συνηθέστερα χρησιμοποιούμενες μέθοδοι.

ΚΕΦΑΛΑΙΟ 3: ΕΡΕΥΝΗΤΙΚΗ ΜΕΘΟΔΟΛΟΓΙΑ ΚΑΙ ΕΡΓΑΛΕΙΑ

3.1. Χρήση δευτερογενών δεδομένων

Στην παρούσα εργασία, εφαρμόζουμε τις προαναφερθείσες τεχνικές εξόρυξης δεδομένων πάνω σε έτοιμα δεδομένα ελέγχου ενός υπάρχοντος ελεγκτικού οργανισμού κυβερνητικών εταιρειών της Ινδίας, χρησιμοποιώντας το πακέτο λογισμικού WEKA (Weka, 2018). Τα αποτελέσματα υποστηρίζουν τη διαδικασία λήψης αποφάσεων σχετικά με τις εταιρείες που ελέγχονται (Hooda *et al.*, 2018). Η εκπαίδευση και η δοκιμή ενός μοντέλου εντοπισμού και διαχείρισης κινδύνων θα συμβάλει στην κάλυψη ενός υφιστάμενου ερευνητικού χάσματος. Η αντιμετώπιση των παραπάνω προβλημάτων απαιτούσε τη χρήση είτε εξειδικευμένου λογισμικού όπως το ACL και το IDEA, είτε γενικών στατιστικών πακέτων όπως τα SAS και SPSS με δυσκολία στην προσαρμογή και παραμετροποίηση των δεδομένων ελέγχου. Αξίζει να σημειωθεί ότι όλα τα προαναφερθέντα πακέτα είναι εμπορικά, ενώ το WEKA είναι ελεύθερο λογισμικό.

Όπως τονίστηκε σε αρχική παράγραφο, η εταιρεία εξωτερικού ελέγχου έχει εμπιστευτικό καθήκον και κρίσιμο ρόλο για την ορθή διεξαγωγή των εργασιών. Η ορθή και έγκυρη εφαρμογή όμως του ερευνητικού σχεδιασμού που χρησιμοποιήθηκε στην παρούσα εργασία δεν θα ήταν δυνατή αν δεν υπήρχαν τα κατάλληλα δεδομένα ελέγχου. Στην παρούσα εργασία το εργαλείο συλλογής δεδομένων που επιλέχθηκε να χρησιμοποιηθεί για την έρευνα και την υλοποίηση του μοντέλου ελέγχου-ταξινόμησης των επιχειρήσεων είναι τα δευτερογενή δεδομένα. Σύμφωνα με τον Heaton (1998), η δευτερογενής ανάλυση αφορά στη χρήση υφιστάμενων δεδομένων, που έχουν συλλεχθεί για τον σκοπό μιας προγενέστερης μελέτης, προκειμένου να ασχοληθούμε με ένα ερευνητικό θέμα το οποίο είναι διαφορετικό από αυτό της αρχικής δουλειάς. Αυτό μπορεί να περιλαμβάνει ένα νέο ερευνητικό ερώτημα, ή μία εναλλακτική προσέγγιση του αρχικού ερωτήματος. Για τον Schutt (2007), ακόμα και η νέα ανάλυση των δεδομένων του ίδιου του ερευνητή για ένα νέο σκοπό αποτελεί δευτερογενή ανάλυση. Όποιος και αν είναι ο ορισμός, η δευτερογενής ανάλυση αφορά στη χρήση δεδομένων τα οποία έχουν ήδη συλλεγεί ή συνταχθεί και μπορεί να αφορούν στη χρήση νέων στατιστικών προσεγγίσεων ή θεωρητικών πλαισίων (Smith, 2008). Οι δευτερογενείς πηγές μπορούν να αναλυθούν σε συνδυασμό με πηγές πρωτογενών δεδομένων (όπως δημοσκοπήσεις, συνεντεύξεις, παρατηρήσεις κ.ά.), ή αντί για πρωτογενή

δεδομένα. Όσον αφορά στα είδη των πηγών, τα δευτερογενή δεδομένα μπορεί να αλιευτούν από απογραφές, κυβερνητικά αρχεία σε εθνικό ή τοπικό επίπεδο, από επιχειρήσεις (όπως οικονομικά αρχεία, ετήσιες εκθέσεις, πρακτικά συναντήσεων συμβουλίων, έγγραφα σχετικά με τις πολιτικές που ακολουθούνται πάνω σε διάφορα θέματα), επιστημονικά άρθρα, έγγραφα σχετικά με τους ανθρώπινους πόρους, άρθρα εφημερίδων, ιστοσελίδες ή κοινωνικά μέσα. Αν οι πηγές θα θεωρηθούν πρωτογενείς ή δευτερογενείς εξαρτάται από τη σχέση του ερευνητή με αυτές. Έτσι, αν τα δεδομένα συλλέχθηκαν από τον ερευνητή, μπορούν να θεωρηθούν πρωτογενή δεδομένα. Αν συλλέχθηκαν από έναν ερευνητή αλλά χρησιμοποιούνται από κάποιον άλλο, ο άλλος τα θεωρεί δευτερογενή δεδομένα.

Το Διαδίκτυο έδωσε στους επαγγελματίες ερευνητές και στους φοιτητές άμεση πρόσβαση σε πηγές και σε όγκους δεδομένων τα οποία οι προηγούμενες γενιές μόνο να φανταστούν μπορούσαν. Ωστόσο, αυτό είναι ένα δίκοπο μαχαίρι, αφού οι ερευνητές πρέπει πλέον να διακρίνουν μεταξύ του τι αποτελεί έγκυρα και χρήσιμα δεδομένα και τι είναι σκουπίδια. Έτσι, για την απόφαση χρησιμοποίησης δευτερογενών πηγών, οι ερευνητές πρέπει να ζυγίζουν τα υπέρ και τα κατά. Ενώ η ανάλυση δευτερογενών δεδομένων θεωρείται ολοένα και περισσότερο ένας αποδεκτός και πολύτιμος τρόπος διεξαγωγής έρευνας, τα πιθανά μειονεκτήματα και προβλήματα σημαίνουν πως αυτή πρέπει να γίνεται με προσοχή.

Παρακάτω αναλύονται τα *πλεονεκτήματα* της χρήσης δευτερογενών δεδομένων:

- Κόστος. Τα δεδομένα έχουν ήδη συλλεχθεί από κάποιο άλλο άτομο ή ερευνητική ομάδα. Ακόμα και αν τα δεδομένα πρέπει να αγοραστούν, το κόστος τους θα είναι μάλλον μικρότερο από αυτό που θα ήταν, αν τα δεδομένα μαζευόταν από την αρχή.
- Ο χρόνος. Ο ερευνητής μπορεί να συνεχίσει με την ανάλυση δεδομένων παρά με το να αναλάβει τις διαδικασίες του ερευνητικού σχεδιασμού, του σχεδιασμού των εργαλείων, της συλλογής των δεδομένων, της καταχώρησης των δεδομένων και του καθαρισμού των δεδομένων. Με το να γίνονται έγκαιρα, οι μελέτες που χρησιμοποιούν δευτερογενή δεδομένα μπορεί να έχουν μεγαλύτερο ενδιαφέρον για αυτούς που λαμβάνουν αποφάσεις, οι οποίοι χρειάζονται επίκαιρη πληροφόρηση (Hofferth, 2005).
- Εύρος και κλίμακα των διαθέσιμων συνόλων δεδομένων. Αυτά περιλαμβάνουν εθνικές δημοσκοπήσεις και δεδομένα τα οποία έχουν συλλεγεί σε διαχρονική βάση. Τα

δευτερογενή δεδομένα είναι επίσης διαθέσιμα σε μεγαλύτερες ποσότητες, επιτρέποντας τη χρήση περισσότερο ισχυρών στατιστικών ελέγχων (Rabinovich και Cheon, 2011).

- Αναπαραγωγή. Αν τα δεδομένα είναι δημόσια διαθέσιμα, θα δώσουν στους μελετητές την ευκαιρία να πραγματοποιήσουν μελέτες αναπαραγωγής για την τελειοποίηση ή την επαλήθευση των αρχικών ευρημάτων (Welch, 2000).
- Εξήγηση αλλαγών και εξέλιξης. Από τη στιγμή που τα δεδομένα καλύπτουν συχνά μακρές χρονικές περιόδους, μπορούν να είναι χρήσιμα για τη δημιουργία ολοκληρωμένων εξηγήσεων (τεκμηρίωση) όσον αφορά στην εξέλιξη ενός φαινομένου και τις αλλαγές που συνέβησαν σε αυτό.
- Αποστασιοποίηση. Η δευτερογενής ανάλυση μπορεί να επιτρέπει τη θεώρηση ενός συνόλου δεδομένων πιο αντικειμενικά, πράγμα που θα ήταν δύσκολο να επιτευχθεί για τον αρχικό ερευνητή (Szabo και Strang, 1997).
- Επαγγελματισμός. Τα δεδομένα προέρχονται συχνά από πηγές που έχουν αναπτυχθεί από ομάδες επαγγελματιών ερευνητών, οι οποίοι έχουν πολλά χρόνια εμπειρίας στον ερευνητικό σχεδιασμό και στη συλλογή δεδομένων (Boslaugh, 2007).
- Κοινωνικά οφέλη. Η δευτερογενής ανάλυση είναι μη-παρεμβατική, αφού δεν συλλέγονται πρόσθετα δεδομένα από τα άτομα και, έτσι, προστατεύεται η ιδιωτικότητά τους (υποθέτοντας πως διατηρείται η ανωνυμία). Σημαίνει, επίσης, πως ευαίσθητοι, ευάλωτοι ή δυσπρόσιτοι πληθυσμοί δεν χρειάζεται να προσεγγιστούν εκ νέου.
- Ευκολία για τους φοιτητές ερευνητές. Έχει υποστηριχτεί πως η χρήση δευτερογενών δεδομένων είναι μία ιδιαίτερα βολική προσέγγιση για τους φοιτητές ερευνητές (Szabo και Strang, 1997), με δεδομένο πως συχνά πρέπει να παραδώσουν τις διατριβές τους κάτω από πολύ απαιτητικά χρονοδιαγράμματα.

Σε ότι αφορά τα *μειονεκτήματα* της χρήσης δευτερογενών δεδομένων:

- Τα δεδομένα μπορεί να είναι ατελή, απαρχαιωμένα, ανακριβή ή με προκαταλήψεις. Στην τελευταία περίπτωση, για παράδειγμα, οι πηγές τείνουν να παρέχουν δημοσιευμένες μελέτες, όπου έχουν βρεθεί στατιστικά σημαντικά αποτελέσματα. Ένας τρόπος για να ξεπεραστεί αυτό είναι να χρησιμοποιήσουμε επίσης και μελέτες που δεν δημοσιεύτηκαν.

- Ασύμβατοι στόχοι / ερωτήματα. Τα δευτερογενή δεδομένα έχουν συλλεχθεί έχοντας υπόψη συγκεκριμένα ερευνητικά ερωτήματα, τα οποία μπορεί να μην είναι αυτά που επιζητά η ερευνητική ομάδα που κάνει τη δευτερογενή ανάλυση. Ή μπορεί να έχουν συλλεχθεί με βάση μία γεωγραφική περιοχή, ενώ η ομάδα να ενδιαφέρεται για μία άλλη. Το σύνολο των δεδομένων μπορεί, επίσης, να έχει αναπτυχθεί με βάση ένα σύνολο μεταβλητών, οι οποίες δεν αντιστοιχούν πλήρως σε αυτές που εξετάζει η νέα έρευνα. Για παράδειγμα, το σύνολο των δεδομένων μπορεί να περιέχει κατηγορικά (ιεραρχικά) δεδομένα, αλλά η έρευνα να απαιτεί αριθμητικά δεδομένα. Δεν υπάρχει, επίσης, η ευκαιρία να διατυπώσουν οι ερευνητές πρόσθετα ερωτήματα (Szabo και Strang, 1997). Αυτό μπορεί να ξεπεραστεί μερικώς κατά τη φάση της επαλήθευσης, βάζοντας κάποιον, ο οποίος ταιριάζει αρκετά στο δημογραφικό προφίλ του αρχικού δείγματος της έρευνας, να σχολιάσει για τη νέα ανάλυση.
- Ποιότητα δεδομένων. Ένα πιθανά σοβαρό μειονέκτημα είναι πως ο ερευνητής δεν γνωρίζει πώς ή πόσο καλά συλλέχθηκαν τα δεδομένα. Για παράδειγμα, τα ποσοστά απόκρισης μπορεί να ήταν χαμηλά, ή τα εργαλεία συλλογής δεδομένων μπορεί να περιείχαν λάθη ή ασυνέπειες που αμφισβητούν την εγκυρότητα και την αξιοπιστία των δεδομένων. Ο Thorne (1994) υποστηρίζει πως, όταν ο ερευνητής δεν αποτελεί μέλος της αρχικής ερευνητικής ομάδας, τα δευτερογενή σύνολα δεδομένων χρησιμοποιούνται καλύτερα μόνο από έμπειρους ερευνητές. Ωστόσο, οι Szabo και Strang (1997) συμβουλεύουν πως αυτά τα προβλήματα μπορούν να ξεπεραστούν, μερικώς, όταν οι ερευνητές διατηρούν αποτελεσματική επικοινωνία με τους αρχικούς ερευνητές και μπορούν να αλιεύσουν πληροφορίες για το πλαίσιο της έρευνας.
- Εξόρυξη δεδομένων. Η διαθεσιμότητα των δευτερογενών δεδομένων μπορεί να σημαίνει πως ο ερευνητής θα «σκαλίζει» τα δεδομένα αναζητώντας θέματα ενδιαφέροντος, παρά ότι θα ξεκινήσει με ένα σύνολο ερευνητικών ερωτημάτων ή υποθέσεων. Ο Hofferth (2005) συμβουλεύει πως αν ισχύει αυτό, ενώ τα δεδομένα μπορούν να χρησιμοποιηθούν με έναν διερευνητικό τρόπο για την ανάπτυξη μιας υπόθεσης, για τον έλεγχο της υπόθεσης πρέπει να αναζητηθούν συμπληρωματικά δεδομένα. Εναλλακτικά, τα δεδομένα μπορούν να χωριστούν σε δύο ξεχωριστά υποσύνολα, εκ των οποίων το πρώτο θα χρησιμοποιηθεί για διερεύνηση και το δεύτερο για τον έλεγχο υποθέσεων, αν υπάρχει αυτή η πρόθεση.

- Το κόστος μάθησης ενός νέου συνόλου δεδομένων. Χρειάζεται χρόνος για την εξοικείωση με ένα σύνολο δεδομένων, γιατί γίνεται μέσα από τη γνωριμία με τα αρχικά ερωτήματα, την τεκμηρίωση και τη δομή των αρχείων δεδομένων. Αυτό συμβαίνει επειδή για τα δεδομένα μεγάλης κλίμακας, οι οργανισμοί παρέχουν συχνά εκπαίδευση για την εξοικείωση των πιθανών χρηστών με τα δεδομένα.

Για την επιλογή αυτή πέρα από τη φύση της έρευνας και την ανάγκη για ύπαρξη οικονομικών δεδομένων προς έλεγχο, λήφθηκαν υπόψιν και τα κριτήρια καταλληλότητας όπως αυτά αποτυπώνονται παρακάτω. Μάλιστα, λόγω της μεγάλης σημασίας που πρέπει να δοθεί στην καταλληλότητα των δεδομένων, περισσότερη ανάλυση θα πραγματοποιηθεί σε ξεχωριστή ενότητα.

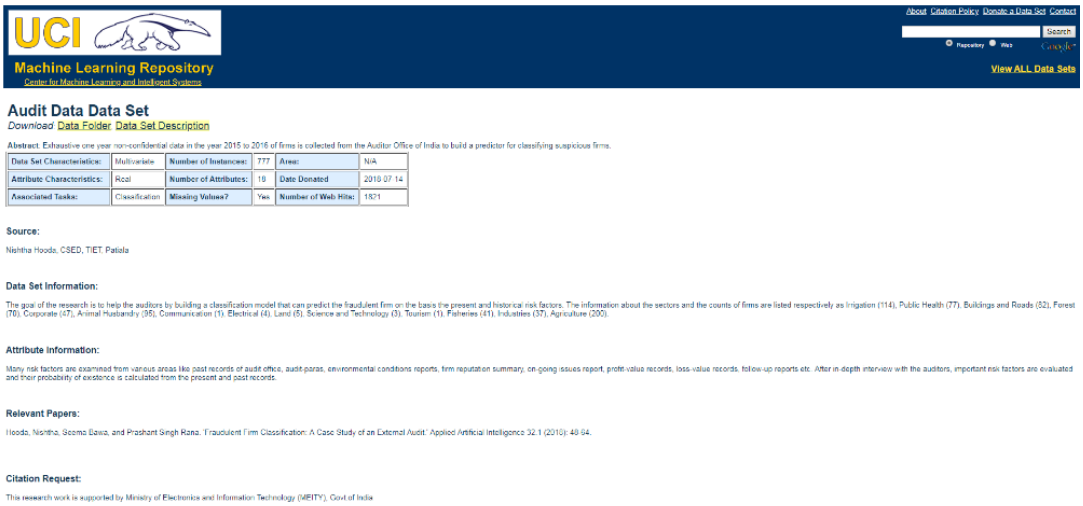
Ένα εργαλείο συλλογής δεδομένων πρέπει να είναι έγκυρο, αξιόπιστο, αντικειμενικό και χρηστικό. Ένα τέτοιο εργαλείο είναι (Καραγεώργος, 2002; Ψαρρού, 2004):

- Έγκυρο (valid) όταν μετρά αυτό για το οποίο έχει κατασκευαστεί.
- Αξιόπιστο (reliable) όταν δίνει συνεπή αποτελέσματα, δηλαδή όσες φορές και να χρησιμοποιηθεί στον ίδιο ή παρόμοιο πληθυσμό, κάτω από τις ίδιες συνθήκες, δίνει σχεδόν τα ίδια αποτελέσματα.
- Αντικειμενικό (objective) όταν δεν επηρεάζεται από υποκειμενική κρίση. Είναι αυτονόητο ότι η αντικειμενικότητα πιθανώς να μην επιτυγχάνεται πλήρως.
- Χρηστικό (usable) όταν χρησιμοποιείται εύκολα, απαιτείται λογικός χρόνος για τη χρήση του, δεν δημιουργεί ηθικά προβλήματα στα υποκείμενα, το κόστος της χρήσης είναι λογικό, υπάρχουν σαφείς ενδείξεις για την εγκυρότητα και την αξιοπιστία του, τα αποτελέσματά του ερμηνεύονται εύκολα και δεν έχουν αναφερθεί τυχόν προβλήματα από άλλους ερευνητές που το χρησιμοποίησαν.

Επειδή πρόκειται για ευαίσθητα δεδομένα και το αποτέλεσμα της εφαρμογής αφορά την κατάταξη μιας επιχείρησης ως ύποπτης για απάτη, είναι φανερό πως στην όλη διαδικασία πρέπει να δοθεί ιδιαίτερη σημασία αρχικά στην ορθή επιλογή του εργαλείου συλλογής δεδομένων όπως αναλύθηκε προηγουμένως και έπειτα στην εγκυρότητα των ίδιων των δεδομένων.

3.2. Το δείγμα δεδομένων ελέγχου

Το δείγμα δεδομένων στο οποίο θα εφαρμοστεί η μεθοδολογία προέρχεται από το παγκοσμίως γνωστό αποθετήριο για την μηχανική μάθηση, το UCI. Σ' αυτό περιέχονται 463 σύνολα δεδομένων σε μια μεγάλη γκάμα εφαρμογών (UCI1, 2018). Ειδικότερα για τον έλεγχο, υπάρχει ένα σύνολο δεδομένων που θα χρησιμοποιηθεί στην εργασία (UCI2, 2018). Οι γενικές πληροφορίες για το συγκεκριμένο σύνολο δεδομένων εμφανίζονται στο Σχεδιάγραμμα 1.


UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

About Citation Policy Donate a Data Set Contact

Search

View ALL Data Sets

Audit Data Data Set

Download: Data Folder, Data Set Description

Abstract: Exhaustive one year non-confidential data in the year 2015 to 2016 of firms is collected from the Auditor Office of India to build a predictor for classifying suspicious firms.

Data Set Characteristics:	Multivariate	Number of Instances:	777	Area:	N/A
Attribute Characteristics:	Real	Number of Attributes:	18	Date Donated	2018-07-14
Associated Tasks:	Classification	Missing Values?	Yes	Number of Web Hits:	1871

Source:

Nishtha Hooda, CSED, TIET, Patiala

Data Set Information:

The goal of the research is to help the auditors by building a classification model that can predict the fraudulent firm on the basis the present and historical risk factors. The information about the sectors and the counts of firms are listed respectively as Irrigation (114), Public Health (77), Buildings and Roads (52), Forest (76), Corporate (47), Animal Husbandry (56), Communication (1), Electrical (4), Land (5), Science and Technology (3), Tourism (1), Fisheries (41), Industries (37), Agriculture (200).

Attribute Information:

Many risk factors are examined from various areas like past records of audit office, audit-paras, environmental conditions reports, firm reputation summary, on-going issues report, profit-value records, loss-value records, follow-up reports etc. After in-depth interview with the auditors, important risk factors are evaluated and their probability of existence is calculated from the present and past records.

Relevant Papers:

Hooda, Nishtha, Seema Bawa, and Prashant Singh Rana. 'Fraudulent Firm Classification: A Case Study of an External Audit.' Applied Artificial Intelligence 32.1 (2018): 48-64.

Citation Request:

This research work is supported by Ministry of Electronics and Information Technology (MEITY), Govt of India

Σχεδιάγραμμα 1. Ελεγκτικά δεδομένα (Audit data) από το αποθετήριο UCI

Το αποθετήριο μηχανικής μάθησης UCI Machine Learning Repository είναι μια συλλογή από βάσεις δεδομένων και γεννητριών δεδομένων που χρησιμοποιούνται από την κοινότητα μηχανικής μάθησης για την εμπειρική ανάλυση αλγορίθμων μηχανικής μάθησης. Το αρχείο δημιουργήθηκε ως ένα αρχείο ftp το 1987 από τον David Aha και συναδέλφους του μεταπτυχιακούς φοιτητές στο UC Irvine. Από τότε, έχει χρησιμοποιηθεί ευρέως από φοιτητές, εκπαιδευτικούς και ερευνητές σε όλο τον κόσμο ως πρωταρχική πηγή των συνόλων δεδομένων μηχανικής μάθησης. Ως ένδειξη του αντίκτυπου και της αποδοχής του αρχείου, έχει αναφερθεί πάνω από 1000 φορές, καθιστώντας το ένα από τα 100 πιο αναφερόμενα paper στις επιστήμες υπολογιστών. Η τρέχουσα έκδοση του ιστότοπου σχεδιάστηκε το 2007 από τους Arthur Asuncion και David Newman και το έργο αυτό είναι σε συνεργασία με το Rexa.info στο Πανεπιστήμιο της Μασαχουσέτης Amherst. Σήμερα περιέχονται 463 σύνολα δεδομένων σε μια μεγάλη γκάμα εφαρμογών (https://archive.ics.uci.edu/ml/index.php). Ειδικότερα, τα δευτερογενή δεδομένα της παρούσας εργασίας έχουν συλλεχθεί από το Γενικό Ελεγκτικό Γραφείο (AGO) της CAG και

βρίσκονται διαθέσιμα για μεταφόρτωση στη διαδικτυακή διεύθυνση https://archive.ics.uci.edu/ml/datasets/Audit+Data#. Το προς μεταφόρτωση αρχείο το οποίο περιέχει το σύνολο των δεδομένων που χρησιμοποιήθηκαν και στην παρούσα εργασία είναι της μορφής .csv.

Ο σχεδιασμός της δειγματοληψίας των συγκεκριμένων δευτερογενών οικονομικών δεδομένων ελέγχου έχει πραγματοποιηθεί από τον Επιθεωρητή και Γενικό Ελεγκτή (Comptroller and Auditor General - CAG) της Ινδίας, ένα ανεξάρτητο συνταγματικό σώμα της χώρας της Ινδίας. Αποτελεί μια αρχή που ελέγχει τα έσοδα και τις δαπάνες του συνόλου των επιχειρήσεων που χρηματοδοτούνται από την κυβέρνηση της Ινδίας. Ενώ διατηρεί το απόρρητο των δεδομένων, εξαντλητικά μη εμπιστευτικά οικονομικά δεδομένα ενός έτους (2015-2016) διαφόρων επιχειρήσεων έχουν συλλεχθεί από το Γενικό Ελεγκτικό Γραφείο (AGO) της CAG. Τα δεδομένα αυτά έχουν συλλεχθεί από συνολικά 777 επιχειρήσεις από 46 διαφορετικές πόλεις ενός κράτους οι οποίες απαριθμούνται από τους ελεγκτές για τη στόχευση της επόμενης διεργασίας στον τομέα του ελέγχου. Οι στόχοι απαριθμούνται από 14 διαφορετικούς τομείς. Οι πληροφορίες σχετικά με τους τομείς και οι μετρήσεις τους συνοψίζονται στον Πίνακα 1.

ID Τομέα	**Στόχος τομέα**	**Πληροφορίες**	**Αριθμός εταιρειών-στόχων**
1	IR	Άρδευση	114
2	P	Δημόσια Υγεία	77
3	BR	Κτίρια και οδικά έργα	82
4	FO	Δάση	70
5	CO	Εταιρικές	47
6	AH	Κτηνοτροφία	95
7	C	Επικοινωνίες	1
8	E	Ηλεκτρολογικά	4
9	L	Γη	5
10	S	Επιστήμη και Τεχνολογία	3
11	T	Τουρισμός	1
12	F	Είδη αλιείας	41
13	I	Βιομηχανίες	37
14	A	Γεωργία	200

Πίνακας 1. Τομείς στόχων ελέγχου

Πολλοί παράγοντες κινδύνου εξετάζονται από διάφορους τομείς, όπως τα προηγούμενα αρχεία του ελεγκτικού γραφείου, οι εκθέσεις περιβαλλοντικών συνθηκών, η περίληψη της εταιρικής φήμης, η έκθεση τρεχόντων ζητημάτων, τα αρχεία κερδών-αξιών, τα αρχεία ζημιών-αξιών, οι αναφορές παρακολούθησης κ.λπ. Μέσω μιας σε βάθος συνέντευξη με τους ελεγκτές, αξιολογούνται σημαντικοί παράγοντες κινδύνου και η πιθανότητα ύπαρξής τους υπολογίζεται τόσο από τις προηγούμενες όσο και από τις παρούσες εγγραφές. Οι Πίνακες 2 και 3 περιγράφουν τους διάφορους παράγοντες κινδύνου που εξετάστηκαν και που εμπλέκονται στη παρούσα μελέτη. Οι διάφοροι παράγοντες κινδύνου κατηγοριοποιούνται, αλλά ο συνδυασμένος κίνδυνος ελέγχου εκφράζεται ως μία συνάρτηση που ονομάζεται Audit Risk Score (ARS) χρησιμοποιώντας μια αναλυτική διαδικασία ελέγχου. Στο τέλος της αξιολόγησης κινδύνου, οι εταιρείες με υψηλές βαθμολογίες ARS ταξινομούνται ως εταιρείες «Απάτης» και οι εταιρείες χαμηλού βαθμού ARS κατατάσσονται ως εταιρείες «Μη-απάτης».

Χαρακτηριστικό	Πληροφορία	Χαρακτηριστικό	Πληροφορία
Para A value	Διαφορά που διαπιστώθηκε στις προγραμματισμένες δαπάνες της επιθεώρησης και της περίληψης Αναφοράς A σε Rs	Sector score	Ιστορική τιμή βαθμολογίας κινδύνου της μονάδας στόχου του Πίνακα 1 χρησιμοποιώντας αναλυτική διαδικασία
Para B value	Διαφορά που διαπιστώθηκε στις μη προγραμματισμένες δαπάνες της επιθεώρησης και της περίληψης Αναφοράς A σε Rs	Loss	Ποσό της ζημίας που υπέστη η επιχείρηση το προηγούμενο έτος
Total	Συνολική διαφορά που βρέθηκε σε άλλες αναφορές Rs	History	Μέση ιστορική ζημία που υπέστη η επιχείρηση τα τελευταία 10 χρόνια
Number	Ιστορικό της βαθμολογίας διαφοράς	District score	Ιστορικό βαθμολογίας κινδύνου μιας περιφέρειας στην περιοχή τελευταία 10 χρόνια
Money value	Ποσό χρημάτων που εμπλέκονται σε ανακρίβειες κατά τους προηγούμενους ελέγχους		

Πίνακας 2. Ταξινόμηση παραγόντων κινδύνου

Χαρακτηριστικό	Πληροφορία	Χαρακτηριστικό	Πληροφορία
Sector ID	Μοναδικό αναγνωριστικό ID του τομέα στόχου	Location ID	Μοναδικό αναγνωριστικό ID της πόλης-επαρχίας
ARS	Συνολική βαθμολογία κινδύνου με χρήση αναλυτικής διαδικασίας	Audit ID	Μοναδικό αναγνωριστικό ID συσχετισμένο με μία υπόθεση ελέγχου
Risk class	Κλάση κινδύνου που έχει εκχωρηθεί την υπόθεση ελέγχου		

Πίνακας 3. Άλλα χαρακτηριστικά

3.3. Αξιολόγηση και έλεγχος της καταλληλότητας των δεδομένων ελέγχου

Το Διαδίκτυο έδωσε στους επαγγελματίες ερευνητές και στους φοιτητές άμεση πρόσβαση σε πηγές και σε όγκους δεδομένων τα οποία οι προηγούμενες γενιές μόνο να φανταστούν μπορούσαν. Ωστόσο, αυτό είναι ένα δίκοπο μαχαίρι, αφού οι ερευνητές πρέπει πλέον να διακρίνουν μεταξύ του τι αποτελεί έγκυρα και χρήσιμα δεδομένα και τι είναι σκουπίδια. Έτσι, για την απόφαση χρησιμοποίησης δευτερογενών πηγών, οι ερευνητές πρέπει να ζυγίζουν τα υπέρ και τα κατά. Ενώ η ανάλυση δευτερογενών δεδομένων θεωρείται ολοένα και περισσότερο ένας αποδεκτός και πολύτιμος τρόπος διεξαγωγής έρευνας, τα πιθανά μειονεκτήματα και προβλήματα σημαίνουν πως αυτή πρέπει να γίνεται με προσοχή.

Όπως αναφέρθηκε στην προηγούμενη ενότητα της εργασίας, το σύνολο των δεδομένων βρίσκεται αποθηκευμένο σε μορφή ενός αρχείου τύπου .csv. Η έρευνα των αρχείων περιγράφεται από τον Welch (2000) ως μία αρχαιολογική διαδικασία, αφού αφορά την ανακάλυψη κα την ερμηνεία αποσπασματικών στοιχείων. Τα αρχειακά δεδομένα μπορούν να δημιουργηθούν από άτομα για τους δικούς τους σκοπούς (για παράδειγμα, ημερολόγια, επιστολές, φωτογραφίες, ιστολογία και αναρτήσεις σε φόρουμ συζητήσεων, βλέπε Προσωπικά έγγραφα) ή από οργανισμούς (βλέπε Έγγραφα Οργανισμών). Οι Fischer και Parmentier (2010) υποστηρίζουν πως τα αρχειακά δεδομένα έχουν χρησιμοποιηθεί κυρίως για να υποστηρίζουν την ανάπτυξη και την κατανόηση του πλαισίου της έρευνας (για να προσθέσουν στις συνεντεύξεις και στα δεδομένα παρατηρήσεων), παρά ως η κύρια πηγή

δεδομένων. Ωστόσο, τα αρχειακά δεδομένα εξελίσσονται γοργά σε μία βιώσιμη πηγή, εν πολλοίς επειδή ένας ολοένα αυξανόμενος όγκος γίνεται διαθέσιμος μέσω του Διαδικτύου.

Κατά τη χρήση αρχειακών δεδομένων, ο Welch (2000) προτείνει μία διαδικασία πέντε σταδίων:

- Ανακάλυψη. Η εξακρίβωση του μέρους για μία κατάλληλη συλλογή δεδομένων δεν είναι πάντοτε μία απλή δουλειά. Για παράδειγμα, στην περίπτωση ενός οργανισμού, οι εναλλαγές του προσωπικού σημαίνουν πως η εταιρική μνήμη μπορεί συχνά να απολεσθεί οι τωρινοί υπάλληλοι δεν γνωρίζουν την ύπαρξη αρχειακών στοιχείων. Ή οι καταγραφές τις εταιρείας μπορεί να διατηρούνται κάπου αλλού μπορεί, για παράδειγμα, να βρίσκονται σε κάποιο εθνικό αρχείο.
- Πρόσβαση. Ακόμα και να εντοπιστεί μία συλλογή, μπορεί η πρόσβαση να είναι ελεγχόμενη, ή να είναι ιδιωτική. Για παράδειγμα, οι εταιρείες δεν έχουν καμία υποχρέωση να διαθέσουν δημόσια τα αρχεία τους. Μπορεί να χρειαστεί να προσπαθήσουμε να διαπραγματευτούμε την πρόσβασή μας σε αυτά.
- Εκτίμηση. Μετά την πρόσβαση, γίνεται απαραίτητο να αξιολογηθεί η ποιότητα των πηγών.
- Κοσκίνισμα. Αυτό σημαίνει ταξινόμηση των εγγράφων με ένα ουσιαστικό ή συστηματικό τρόπο, παραδείγματος χάρη, διατεταγμένα ανά χρονολογική σειρά ή κατά θεματική κατηγορία.
- Διασταύρωση. Αυτή χρησιμοποιείται για λόγους επαλήθευσης. Έτσι, ακολουθείται η τριγωνοποίηση των δεδομένων, διασταυρώνοντας πηγές από περισσότερες από μία συλλογές. Η μεθοδολογική τριγωνοποίηση σημαίνει διασταύρωση των πηγών με χρήση εναλλακτικής στρατηγικής για παράδειγμα, συνεντεύξεις.

Στην παρούσα μελέτη παρατηρούμε ότι το κριτήριο της Πρόσβασης καλύπτεται με εύκολο τρόπο καθώς τα δεδομένα είναι προσβάσιμα σε κάθε ενδιαφερόμενο μέσω του Διαδικτύου. Το κριτήριο της Εκτίμησης καλύπτεται από τη στιγμή που την ποιότητα της πηγής των δεδομένων εγγυάται ο Επιθεωρητής και Γενικός Ελεγκτής (Comptroller and Auditor General - CAG) της Ινδίας, ένα ανεξάρτητο συνταγματικό σώμα της χώρας της Ινδίας. Το Κοσκίνισμα των δεδομένων μπορεί να πραγματοποιηθεί με τη βοήθεια της πληροφορικής χρησιμοποιώντας κάποιο κατάλληλο λογισμικό όπως το Weka. Τέλος, η Διασταύρωση στη συγκεκριμένη περίπτωση εφαρμογής είναι εύκολο να πραγματοποιηθεί στην πράξη. Η

μεθοδολογία που προτείνεται έχει αρχικά συμβουλευτικό χαρακτήρα προς τους ελεγκτές. Σε πρώτη φάση μπορεί να διευκολύνει το έργο τους, να απομονώσουν τις επιχειρήσεις υψηλού κινδύνου για απάτη, στη συνέχεια όμως οι ελεγκτές λαμβάνουν δεδομένα και με άλλες μεθόδους όπως ο επιτόπιος έλεγχος.

Επιπρόσθετα, σχετικά με τα κριτήρια επιλογής ενός συνόλου δεδομένων, ο Hofferth (2005) κάνει αρκετές προτάσεις:
Μόλις εντοπιστεί ένα σύνολο δεδομένων, το επόμενο βήμα είναι να κριθεί αν ταιριάζει στον σκοπό, το οποίο σημαίνει αξιολόγηση της ποιότητας του δευτερογενούς υλικού σε σχέση με τους στόχους της προτεινόμενης μελέτης.

- Ταιριάζει ο σχεδιασμός της μελέτης στα ερευνητικά ερωτήματα; Είδαμε προηγουμένως πως μπορεί να υπάρξει μία αναντιστοιχία ανάμεσα στα ερευνητικά ερωτήματα της μελέτης και σε εκείνα της αρχικής έρευνας. Επιπλέον, συλλέχθηκαν τα δεδομένα από ένα δείγμα του πληθυσμού που ταιριάζει στην τρέχουσα μελέτη;
- Είναι τα μεγέθη των δειγμάτων για τις υπό-ομάδες που μας ενδιαφέρουν αρκετά μεγάλα; Ακόμα και αν το μέγεθος του αρχικού δείγματος είναι μεγάλο, ο αριθμός των περιπτώσεων μιας ειδικής κατηγορίας μπορεί να είναι πολύ μικρότερος.
- Είναι οι μέθοδοι και τα ερευνητικά εργαλεία που χρησιμοποιήθηκαν στη μελέτη τα κατάλληλα; Τα δεδομένα μπορεί να χρειάζονται στατιστικές προσαρμογές, για παράδειγμα, χρήση εργαλείων που λαμβάνουν υπόψη πολλαπλά επίπεδα ανάλυσης.
- Μπορούν να εξηγηθούν και να αντιμετωπιστούν τα ελλιπή δεδομένα; Ο ερευνητής χρειάζεται να γνωρίζει γιατί οι απαντήσεις σε μία ερώτηση δεν υπάρχουν για όλο το δείγμα. Συμβαίνει εξαιτίας των εργαλείων δημοσκόπησης που χρησιμοποιούν μοτίβα παράλειψης, τα οποία ταξινομούν τους ερωτώμενους σε διαφορετικές ομάδες και τους καθοδηγούν σε συγκεκριμένες ερωτήσεις, ή συμβαίνει επειδή ο ερωτώμενος δεν γνώριζε την απάντηση ή αρνήθηκε να απαντήσει την ερώτηση;
- Περιέχει το σύνολο δεδομένων τα απαιτούμενα μέτρα; Το σύνολο δεδομένων μπορεί να περιέχει πληροφορίες για το θέμα, αλλά οι ιδιότητες των κλιμάκων και των δεικτών που χρησιμοποιήθηκαν είναι αποδεκτές για το πεδίο; Τι πληροφορίες δίνονται για την εγκυρότητα και την αξιοπιστία των κλιμάκων που χρησιμοποιήθηκαν; Για παράδειγμα, η ανάλυση που βασίζεται σε αυτό-αναφορές μπορεί να μην είναι αποδεκτή από κάποιους χρηματοδότες ή από κάποιους εκδότες επιστημονικών άρθρων.

Στην περίπτωσή μας, παρατηρούμε ότι ο συνολικός σχεδιασμός της μελέτης αφορά στην υλοποίηση του κατάλληλου μοντέλου ελέγχου οικονομικών δεδομένων, δηλαδή είναι εξολοκλήρου δομημένος έτσι ώστε να απαντηθεί το αρχικό ερευνητικό ερώτημα. Έπειτα, σε ότι αφορά το δείγμα από το οποίο αντλήθηκαν τα προς έλεγχο οικονομικά δεδομένα, βλέπουμε ότι και εδώ καλύπτεται το κριτήριο του Hofferth (2005) που αφορά το μέγεθός του. Μάλιστα καλύπτεται και σε ότι αφορά το μέγεθος του δείγματος (μεγάλος αριθμός 777 επιχειρήσεων) αλλά και ως προς τη διασπορά των δεδομένων, τόσο γεωγραφική όσο και από την πλευρά των τομέων δραστηριότητας (46 πόλεις, 14 τομείς). Τέλος, σχετικά με το κριτήριο της χρήσης εργαλείων που λαμβάνουν υπόψη πολλαπλά επίπεδα ανάλυσης, έχουμε να παρατηρήσουμε ότι οι παράμετροι του μοντέλου ελέγχου περιλαμβάνουν και τις τωρινές αλλά και τις ιστορικές οικονομικές συνθήκες των επιχειρήσεων.

3.4. Δεοντολογικοί προβληματισμοί για τη χρήση των δεδομένων

Με την πρώτη ματιά, φαίνεται πως οι δεοντολογικοί προβληματισμοί δεν μας αφορούν, ή μας αφορούν ελάχιστα στη δευτερογενή ανάλυση, επειδή δεν υπάρχουν πρόσωπο-με-πρόσωπο συνεντεύξεις ή παρατηρήσεις ανθρώπινης συμπεριφοράς. Ωστόσο, όπως επισημαίνουν οι Gladstone *et al.,* (2007), τα ζητήματα της εν επιγνώσει συναίνεσης και της εμπιστευτικότητας δεν έχουν εξαλειφθεί, απλά ανακύπτουν με διαφορετικούς τρόπους. Ο Thorne (1998) διερωτάται αν η διατύπωση νέων ερωτημάτων σε δεδομένα που έχουν συλλεγεί σε προηγούμενες μελέτες παραβιάζει τη συναίνεση που είχε αποκτηθεί όταν διεξήχθη η αρχική μελέτη. Αυτός ίσως ένας λόγος για τον οποίο οι Hinds *et al.,* (1997) υποστηρίζουν πως οι ερευνητές πρέπει να αναζητούν την άδεια από τους συμμετέχοντες στην πρωταρχική μελέτη για τη δευτερογενή ανάλυση των δεδομένων τους. Ο Thorne (1998) συμβουλεύει, επίσης, πως οι δευτερογενείς αναλυτές πρέπει να εξοικειώνονται με τις πραγματικές και τις πιθανές ανάγκες για προστασία της ιδιωτικότητας των ατόμων και των πληθυσμών στις βάσεις δεδομένων που μεταχειρίζονται. Όπως συμβουλεύουν οι Gladstone *et al.,* (2007), οι ερευνητές πρέπει να είναι ικανοί να υποστηρίξουν την κρίση τους, ως προς την ισχύ της αρχικής συναίνεσης και τις συνθήκες κάτω από τις οποίες είναι κατάλληλη η δευτερογενής ανάλυση. Οι Επιτροπές Δεοντολογίας και Ηθικής της Έρευνας πλέον προβλέπουν πως οι συμμετέχοντες πρέπει να δίνουν τη συναίνεσή τους για τη μελλοντική χρήση των δεδομένων και για άλλους σκοπούς. Βέβαια, όταν πρόκειται για ανάλυση των

δεδομένων κοινωνικών μέσων, προκύπτουν πρόσθετες δεοντολογικές προκλήσεις. Όπως σχολιάζουν οι Boyd και Crawford (2012), απλά επειδή τα δεδομένα είναι προσβάσιμα, δεν σημαίνει πως η πρόσβαση είναι δεοντολογική.

Συμπερασματικά παρατηρούμε ότι στη συγκεκριμένη εργασία καλύπτονται οι απαιτήσεις για την προστασία της ιδιωτικότητας όπως την ορίζει ο Thorne (1998) διότι τα οικονομικά δεδομένα που έχουν συλλεχθεί από συνολικά 777 επιχειρήσεις καλύπτονται κάτω από την ομπρέλα της ανωνυμίας και της προστασίας που εγγυάται ο Επιθεωρητής και Γενικός Ελεγκτής (Comptroller and Auditor General - CAG) της Ινδίας. Ενώ συγκεντρώνει σε ένα αρχείο πληθώρα οικονομικών μεγεθών προς έλεγχο, σε καμία περίπτωση όπως αυτό είναι φυσικό δεν ονοματίζονται επιχειρήσεις παρά μόνο μεταβλητές και αριθμοί. Όπως αναφέρθηκε και σε προηγούμενη υπο-ενότητα, ο Επιθεωρητής και Γενικός Ελεγκτής διατηρεί το απόρρητο των δεδομένων έτσι ώστε τα δεδομένα αυτά να μη μπορούν να φτάσουν στον εκάστοτε ερευνητή χωρίς την απαραίτητη προστασία την ανωνυμίας. Τέλος, σύμφωνα με την ανάλυση των Boyd και Crawford (2012), μπορούμε να ισχυριστούμε ότι στον συγκεκριμένο ερευνητικό σχεδιασμό της παρούσας εργασίας τα δεδομένα είναι και εύκολα προσβάσιμα αλλά και η πρόσβαση αυτή χαρακτηρίζεται ως δεοντολογική. Σαν επίλογο στο κομμάτι της δεοντολογίας, αξίζει να τονιστεί ότι περισσότερη προσοχή πρέπει να δοθεί στο κομμάτι του ίδιου του ελέγχου και της ορθής χρήσης του εργαλείου έρευνας που προτείνεται καθώς πρόκειται για ηθικά λεπτό ζήτημα ο χαρακτηρισμός και η τυχόν στοχοποίηση μιας επιχείρησης ως ύποπτης για εμπλοκή σε απάτη.

ΚΕΦΑΛΑΙΟ 4: ΤΟ ΕΡΓΑΛΕΙΟ WEKA & ΑΠΟΤΕΛΕΣΜΑΤΑ

4.1. Weka: Χαρακτηριστικά και δυνατότητες

Το υπολογιστικό πακέτο WEKA (Waikato Environment for Knowledge Analysis) χρησιμοποιήθηκε για την εφαρμογή μεθόδων ταξινόμησης, ομαδοποίησης-συσταδοποίησης και εξόρυξης κανόνων συσχέτισης στο σύνολο των δεδομένων ελέγχου (Witten *et al.*, 2016). Το WEKA είναι λογισμικό ανοιχτού κώδικα που παρέχει μια συλλογή αλγορίθμων μηχανικής μάθησης και εξόρυξης δεδομένων.

Το WEKA (Waikato Environment for Knowledge Analysis) είναι μια σουίτα λογισμικού για μηχανική μάθηση και Εξόρυξη Δεδομένων. Αναπτύχθηκε στο Πανεπιστήμιο του Waikato της Ν. Ζηλανδίας και πήρε το όνομα του από το Weka, ένα μικρό και υπό εξαφάνιση πουλί της Ν. Ζηλανδίας. Το WEKA ανήκει στην κατηγορία του λεγόμενου ″ελεύθερου λογισμικού″ (freeware) και διατίθεται δημοσίως σύμφωνα με τους όρους της άδειας GNU General Public License, η οποία επιτρέπει στους χρήστες να χρησιμοποιούν, αλλά και να τροποποιούν ελεύθερα το λογισμικό.

Το WEKA είναι ένα από τα πιο διαδεδομένα λογισμικά Εξόρυξης Δεδομένων. Έχει χρησιμοποιηθεί σε μεγάλο αριθμό επιστημονικών εργασιών, και αρκετά βιβλία Εξόρυξης Δεδομένων αναφέρονται σε αυτό. Η μεγάλη δημοφιλία του οφείλεται στα ειδικά χαρακτηριστικά του και στις δυνατότητες που προσφέρει. Αναλυτικότερα το WEKA:

- Περιέχει αρκετά μεγάλη ποικιλία μεθόδων για κατηγοριοποίηση, παλινδρόμηση, ανάλυση συστάδων, και κανόνες συσχέτισης. Επίσης, παρέχει δυνατότητες για προεπεξεργασία των δεδομένων, καθώς και εργαλεία οπτικοποίησης.
- Είναι λογισμικό ανοικτού κώδικα. Αυτό σημαίνει ότι ο πηγαίος κώδικας είναι δημοσίως διαθέσιμος. Χρήστες με γνώσεις προγραμματισμού μπορούν να τροποποιούν και να εξελίσσουν τους αλγορίθμους.
- Είναι γραμμένο σε γλώσσα Java, γεγονός που το καθιστά ικανό να εγκαθίσταται σε διαφορετικές πλατφόρμες υλικού και λογισμικού.
- Διαθέτει γραφικό περιβάλλον εργασίας. Στο διαδίκτυο υπάρχει διαθέσιμη μεγάλη ποικιλία βιβλιοθηκών για μηχανική μάθηση και εξόρυξη δεδομένων. Ωστόσο, η χρήση τους απαιτεί τη συγγραφή κώδικα. Αντιθέτως, το γραφικό περιβάλλον του WEKA

επιτρέπει τη χρήση του λογισμικού από τελικούς χρήστες, οι οποίοι δεν διαθέτουν γνώσεις προγραμματισμού.

Το WEKA διατίθεται σε δύο διαφορετικές εκδόσεις:

- Στη λεγόμενη ″σταθερή″ (stable) έκδοση, η οποία απευθύνεται σε τελικούς χρήστες και αντιστοιχεί στην τελευταία έκδοση του βιβλίου των Witten, Frank και Hall (2011).
- Στην έκδοση η οποία απευθύνεται σε προγραμματιστές. Η έκδοση αυτή χρησιμοποιείται από την κοινότητα των προγραμματιστών του WEKA για τη διόρθωση σφαλμάτων και την επέκταση των δυνατοτήτων του λογισμικού.

Το πανεπιστήμιο του Waikato διατηρεί μια ιδιαίτερα πλούσια ιστοθέση αφιερωμένη στο WEKA (https://www.cs.waikato.ac.nz/ml/weka/). Στην ιστοθέση αυτή οι χρήστες μπορούν:

1. Να προμηθευτούν το WEKA (https://waikato.github.io/weka-wiki/downloading_weka/). Προσφέρονται διαφορετικές επιλογές για λειτουργικά συστήματα Windows, Mac OS X και Linux.
2. Να αναζητήσουν τεκμηρίωση σχετικά με το λογισμικό. Η τεκμηρίωση περιλαμβάνει το manual του λογισμικού, οδηγίες για την αντιμετώπιση προβλημάτων, απαντήσεις σε συχνές ερωτήσεις, οδηγίες για σύνδεση με γλώσσες προγραμματισμού όπως το Matlab και η R, παρουσιάσεις και ηλεκτρονικά σεμινάρια, καθώς και πολλά άλλα.
3. Να προμηθευτούν το Application Programming Interface (API) του λογισμικού, καθώς και μια μεγάλη λίστα πρόσθετων πακέτων για διάφορες εργασίες μηχανικής μάθησης και εξόρυξης δεδομένων.
4. Να προμηθευτούν ένα σημαντικό αριθμό συνόλων δεδομένων, τα οποία μπορούν να χρησιμοποιήσουν για εξάσκηση. Η πιο πρόσφατη έκδοση είναι η 3.8.4 και θα παρουσιαστεί στα πλαίσια της παρούσας πτυχιακής εργασίας.

Σχετικά με τη δομή τον Weka, το Weka είναι υλοποιημένο σε πακέτα (packages) που ακολουθούν ιεραρχία καταλόγου. Κάθε πρόγραμμα γραμμένο σε Java είναι υλοποιημένο σαν μια κλάση (class). Μια κλάση στον Αντικειμενοστραφή Προγραμματισμό είναι μια συλλογή από μεταβλητές (variables) και κάποιες μεθόδους (methods) που εκτελούνται πάνω σε αυτές.

Στα μεγάλα προγράμματα, ωστόσο, όπως είναι το Weka, υπάρχουν περισσότερες από μια κλάσεις που για να επικοινωνούν καλύτερα μεταξύ τους οργανώνονται σε πακέτα. Ένα πακέτο είναι μια συλλογή κλάσεων με παρόμοια λειτουργικότητα.

Στο Weka κάθε ένας υλοποιημένος αλγόριθμος είναι ενθυλακωμένος σε μια Java κλάση και κάθε φορά που η Εικονική Μηχανή Java (Java Virtual Machine) καλείται να εκτελέσει έναν αλγόριθμο, δημιουργεί ένα στιγμιότυπο (instance) της σχετικής κλάσης και διανέμει τη μνήμη που χρειάζεται για την εκτέλεση.

Παρακάτω γίνεται αναφορά στα σημαντικότερα πακέτα που αποτελούν τον κορμό τον προγράμματος.

Το *πακέτο* weka.core

Πρόκειται για το κυριότερο πακέτο τον προγράμματος που είναι κεντρικό στο σύστημα του Weka και οι κλάσεις τον είναι προσβάσιμες από σχεδόν όλες τις υπόλοιπες. Οι βασικές κλάσεις που περιέχονται σε αυτό το πακέτο είναι οι Attribute, Instance και Instances. 'Ενα αντικείμενο της κλάσης Attribute αναπαριστά κάποιο γνώρισμα. Περιέχει το όνομα τον γνωρίσματος, τον τύπο τον και τις πιθανές τιμές τον. Ένα αντικείμενο της κλάσης Instance περιέχει τις τιμές των γνωρισμάτων από ένα συγκεκριμένο στιγμιότυπο. Τέλος, ένα αντικείμενο της κλάσης Instances περιέχει ένα ταξινομημένο σύνολο στιγμιοτύπων, ουσιαστικά ένα σύνολο δεδομένων.

Το *πακέτο* weka.filters

Σε αυτό περιέχονται όλες οι κλάσεις που αφορούν στην προ-επεξεργασία των δεδομένων. Κάθε φίλτρο χωρίζεται σε supervised ή unsupervised ανάλογα, αν κάνει χρήση ή όχι τον γνωρίσματος κλάσης (class attribute) των δεδομένων. Έπειτα, υπάρχει διάκριση σε attribute ή instance ανάλογα με το αν χειρίζεται τα δεδομένα κατά στήλη ή κατά γραμμή.

Το *πακέτο* weka.classifiers

Σε αυτό το πακέτο περιέχονται όλες οι υλοποιήσεις που αφορούν την Κατηγοριοποίηση Δεδομένων. Βασική λειτουργία είναι η Classifier που ορίζει τη δομή κάθε αλγορίθμου για κατηγοριοποίηση.

Το *πακέτο* weka.associations

Σε αυτό το πακέτο, αντίστοιχα, περιέχονται όλες οι υλοποιήσεις που αφορούν στην εξαγωγή Κανόνων Συσχέτισης (Association Rules). Βασική κλάση είναι η AssociationRulesProducer.

Το *πακέτο* weka.clusterers

Σε αυτό το πακέτο, τέλος, περιέχονται όλες οι υλοποιήσεις που αφορούν στην Συσταδοποίηση Δεδομένων (Clustering). Βασική κλάση είναι η Clusterer.

Όσον αφορά τις διάφορες εκδόσεις του Weka, το Weka ξεκίνησε να αναπτύσσεται το 1999, όταν κυκλοφόρησε η πρώτη έκδοσή του (Weka 3.0). 'Εκτοτε, έχουν ακολουθήσει άλλες τρεις κυκλοφορίες σταθερών (stable) εκδόσεων και δύο πειραματικού σκοπού (development versions).

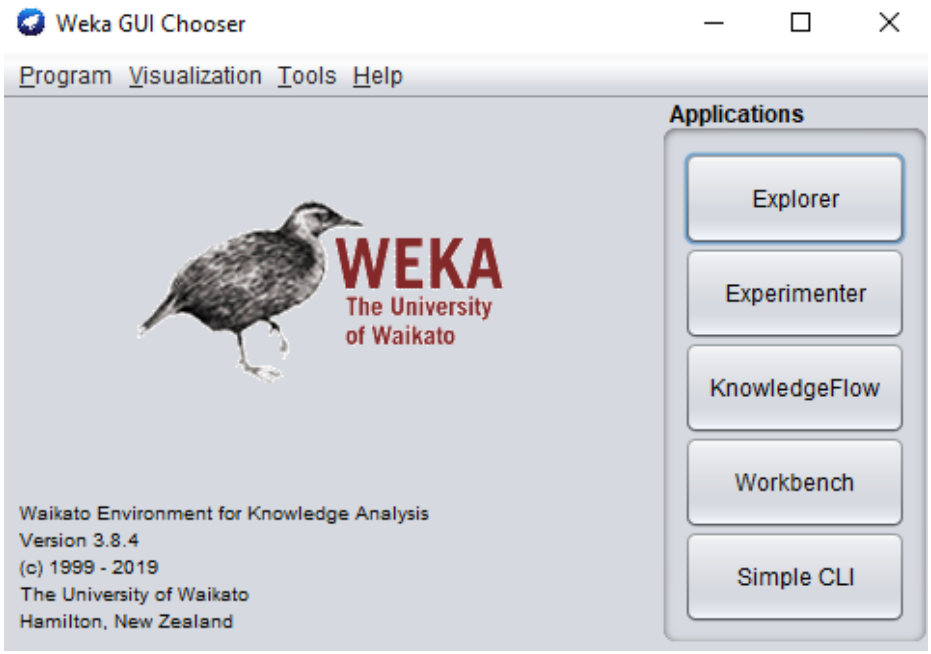
Στην ιεραρχία των εκδόσεων ακολουθείται το μοντέλο του Linux όπου κάθε έκδοση με δεύτερο ψηφίο άρτιο αριθμό αφορά σταθερή - αξιόπιστη έκδοση, ενώ αντίθετα όταν το δεύτερο ψηφίο είναι περιττός τότε πρόκειται για πειραματική έκδοση. Για παράδειγμα, όλες οι εκδόσεις 3.6.x του Weka με νεότερη την 3.6.13 είναι stable ενώ οι 3.7.x αφορούν δοκιμαστικές εκδόσεις. Στις δοκιμαστικές εκδόσεις μπορεί να υπάρχουν νέα ιδιαίτερα χαρακτηριστικά τα οποία όμως ελέγχονται για τη λειτουργικότητά τους και δεν είναι σίγουρο ότι θα ενταχθούν στην επόμενη έκδοση. Επομένως, όταν είναι ζητούμενο η σταθερότητα και η αξιοπιστία, όπως συμβαίνει για εκπαιδευτικούς σκοπούς, τότε προτιμώνται οι σταθερές εκδόσεις του λογισμικού.

Οι εκδόσεις 3.7.x τον λογισμικού εισήγαγαν ένα νέο, πολύ χρήσιμο χαρακτηριστικό, τον Διαχειριστή Πακέτων (Package Manager), με τη χρήση του οποίου δεν είναι αναγκαία η ενσωμάτωση των νέων πακέτων στο πρόγραμμα αλλά μπορούν να επιλέγονται αυτόνομα κατά την εκτέλεση τον προγράμματος.

Με την εκκίνηση του WEKA εμφανίζεται το παράθυρο του Σχεδιαγράμματος 2. Από το σημείο αυτό ο χρήστης μπορεί να εκκινήσει τις κύριες εφαρμογές του WEKA:

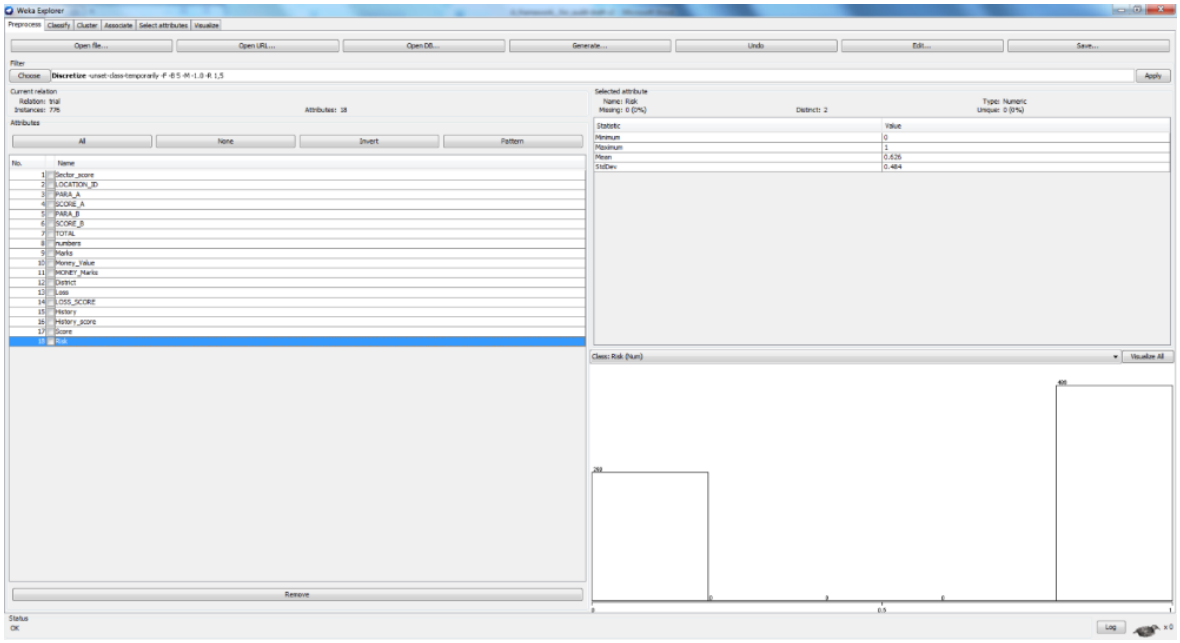
- Ο **Explorer** είναι η πιο δημοφιλής διεπαφή. Ο χρήσης μπορεί να εκτελέσει όλες τις κύριες εργασίες Εξόρυξης Δεδομένων, όπως κατηγοριοποίηση, παλινδρόμηση, ανάλυση συστάδων, ανακάλυψη κανόνων συσχέτισης, προεπεξεργασία των δεδομένων και οπτικοποίηση.

- Ο **Experimenter** είναι ένα περιβάλλον για διεξαγωγή πειραμάτων, όπου αξιολογούνται μέθοδοι κατηγοριοποίησης και παλινδρόμησης. Διευκολύνει τη σύγκριση της επίδοσης διαφορετικών μοντέλων και παρουσιάζει τα αποτελέσματα σε μορφή πίνακα.
- Το **Knowledge Flow** είναι ένα περιβάλλον που επιτρέπει τη διεξαγωγή των ιδίων εργασιών με τον Explorer, διαθέτει όμως διαφορετική διεπαφή (interface). Στο περιβάλλον αυτό χρησιμοποιούνται components, τα οποία συνδέονται μεταξύ τους με γραφικό τρόπο, ο οποίος ορίζει τη ροή εργασίας. Υπάρχουν components για τη φόρτωση των δεδομένων, την προεπεξεργασία τους, τη δημιουργία και εκπαίδευση μοντέλων, την οπτικοποίηση κλπ. Όπως αναφέρθηκε και προηγουμένως, ο Explorer είναι το πιο δημοφιλές περιβάλλον. Για τον λόγο αυτό, στον παρόντα μικρό οδηγό θα γίνει παρουσίαση του Explorer. Το περιβάλλον εργασίας του Explorer παρουσιάζεται στο Σχεδιάγραμμα 3. Το παράθυρο της εφαρμογής περιλαμβάνει 6 tabs για προεπεξεργασία των δεδομένων, κατηγοριοποίηση, ανάλυση συστάδων, επιλογή γνωρισμάτων και οπτικοποίηση.
- **Το Weka Simple CLI** είναι η τέταρτη επιλογή interface που παρέχεται από το εισαγωγικό παράθυρο και αφορά τη χρήση απλής Διασύνδεσης Γραμμής Εντολών (Command Line Interface). Λειτουργία γραμμής εντολών μπορεί να χρησιμοποιήσει κάποιος κατευθείαν από την command prompt του υπολογιστή εκτελώντας το Weka, αλλά δίνεται και αυτή η δυνατότητα.



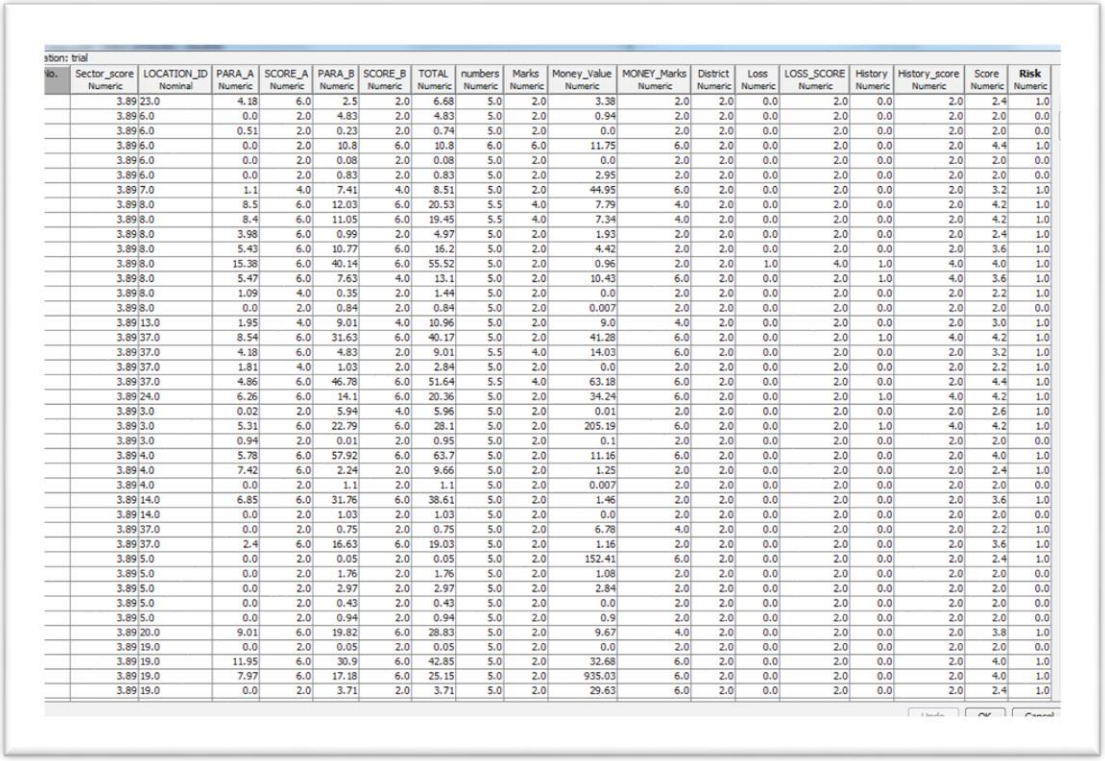
Σχεδιάγραμμα 2. Εκκίνηση του WEKA

Το Σχεδιάγραμμα 3 δείχνει το βασικό γραφικό περιβάλλον εργασίας χρήστη (GUI) του WEKA. Ένας από τους κύριους στόχους του WEKA είναι η εξόρυξη πληροφοριών από υπάρχοντα σύνολα δεδομένων. Φυσικά ο κύριος λόγος για την επιλογή του Weka στην παρούσα εργασία είναι ότι παρέχει μια συλλογή αλγορίθμων μηχανικής μάθησης και εξόρυξης δεδομένων για προεπεξεργασία δεδομένων, ταξινόμηση, παλινδρόμηση, ομαδοποίηση-συσταδοποίηση, κανόνες συσχέτισης και οπτικοποίηση (Hall *et al.*, 2009).



Σχεδιάγραμμα 3. Το γραφικό περιβάλλον (GUI) του WEKA

Όπως απεικονίστηκε προηγουμένως στο Σχεδιάγραμμα 1, το σύνολο δεδομένων περιέχει 777 παρουσίες. Δεν υπάρχουν τιμές που λείπουν για όλα τα χαρακτηριστικά. Στο περιβάλλον WEKA τα δεδομένα ελέγχου απεικονίζονται όπως στο Σχεδιάγραμμα 4.



ation: trial

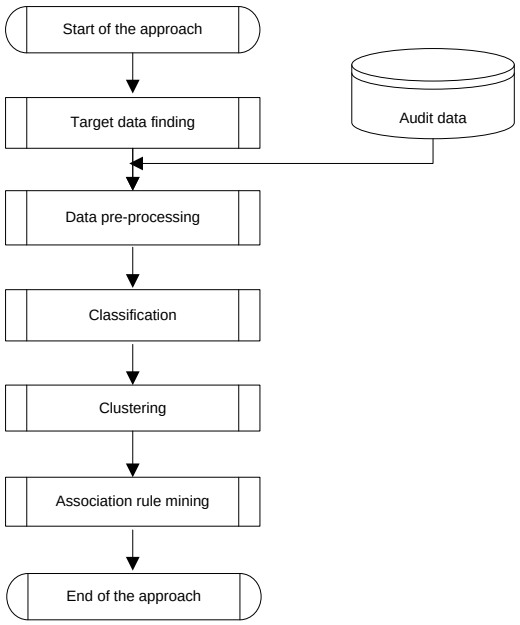
No.	Sector_score Numeric	LOCATION_ID Nominal	PARA_A Numeric	SCORE_A Numeric	PARA_B Numeric	SCORE_B Numeric	TOTAL Numeric	numbers Numeric	Marks Numeric	Money_Value Numeric	MONEY_Marks Numeric	District Numeric	Loss Numeric	LOSS_SCORE Numeric	History Numeric	History_score Numeric	Score Numeric	Risk Numeric
	3.89	23.0	4.18	6.0	2.5	2.0	6.68	5.0	2.0	3.38	2.0	2.0	0.0	2.0	0.0	2.0	2.4	1.0
	3.89	6.0	0.0	2.0	4.83	2.0	4.83	5.0	2.0	0.94	2.0	2.0	0.0	2.0	0.0	2.0	2.0	0.0
	3.89	6.0	0.51	2.0	0.23	2.0	0.74	5.0	2.0	0.0	2.0	2.0	0.0	2.0	0.0	2.0	2.0	0.0
	3.89	6.0	0.0	2.0	10.8	6.0	10.8	6.0	6.0	11.75	6.0	2.0	0.0	2.0	0.0	2.0	4.4	1.0
	3.89	6.0	0.0	2.0	0.08	2.0	0.08	5.0	2.0	0.0	2.0	2.0	0.0	2.0	0.0	2.0	2.0	0.0
	3.89	6.0	0.0	2.0	0.83	2.0	0.83	5.0	2.0	2.95	2.0	2.0	0.0	2.0	0.0	2.0	2.0	0.0
	3.89	7.0	1.1	4.0	7.41	4.0	8.51	5.0	2.0	44.95	6.0	2.0	0.0	2.0	0.0	2.0	3.2	1.0
	3.89	8.0	8.5	6.0	12.03	6.0	20.53	5.5	4.0	7.79	4.0	2.0	0.0	2.0	0.0	2.0	4.2	1.0
	3.89	8.0	8.4	6.0	11.05	6.0	19.45	5.5	4.0	7.34	4.0	2.0	0.0	2.0	0.0	2.0	4.2	1.0
	3.89	8.0	3.98	6.0	0.99	2.0	4.97	5.0	2.0	1.93	2.0	2.0	0.0	2.0	0.0	2.0	2.4	1.0
	3.89	8.0	5.43	6.0	10.77	6.0	16.2	5.0	2.0	4.42	2.0	2.0	0.0	2.0	0.0	2.0	3.6	1.0
	3.89	8.0	15.38	6.0	40.14	6.0	55.52	5.0	2.0	0.96	2.0	2.0	1.0	4.0	1.0	4.0	4.0	1.0
	3.89	8.0	5.47	6.0	7.63	4.0	13.1	5.0	2.0	10.43	6.0	2.0	0.0	2.0	1.0	4.0	3.6	1.0
	3.89	8.0	1.09	4.0	0.35	2.0	1.44	5.0	2.0	0.0	2.0	2.0	0.0	2.0	0.0	2.0	2.2	1.0
	3.89	8.0	0.0	2.0	0.84	2.0	0.84	5.0	2.0	0.007	2.0	2.0	0.0	2.0	0.0	2.0	2.0	0.0
	3.89	13.0	1.95	4.0	9.01	4.0	10.96	5.0	2.0	9.0	4.0	2.0	0.0	2.0	0.0	2.0	3.0	1.0
	3.89	37.0	8.54	6.0	31.63	6.0	40.17	5.0	2.0	41.28	6.0	2.0	0.0	2.0	1.0	4.0	4.2	1.0
	3.89	37.0	4.18	6.0	4.83	2.0	9.01	5.5	4.0	14.03	6.0	2.0	0.0	2.0	0.0	2.0	3.2	1.0
	3.89	37.0	1.81	4.0	1.03	2.0	2.84	5.0	2.0	0.0	2.0	2.0	0.0	2.0	0.0	2.0	2.2	1.0
	3.89	37.0	4.86	6.0	46.78	6.0	51.64	5.5	4.0	63.18	6.0	2.0	0.0	2.0	0.0	2.0	4.4	1.0
	3.89	24.0	6.26	6.0	14.1	6.0	20.36	5.0	2.0	34.24	6.0	2.0	0.0	2.0	1.0	4.0	4.2	1.0
	3.89	3.0	0.02	2.0	5.94	4.0	5.96	5.0	2.0	0.01	2.0	2.0	0.0	2.0	0.0	2.0	2.6	1.0
	3.89	3.0	5.31	6.0	22.79	6.0	28.1	5.0	2.0	205.19	6.0	2.0	0.0	2.0	1.0	4.0	4.2	1.0
	3.89	3.0	0.94	2.0	0.01	2.0	0.95	5.0	2.0	0.1	2.0	2.0	0.0	2.0	0.0	2.0	2.0	0.0
	3.89	4.0	5.78	6.0	57.92	6.0	63.7	5.0	2.0	11.16	6.0	2.0	0.0	2.0	0.0	2.0	4.0	1.0
	3.89	4.0	7.42	6.0	2.24	2.0	9.66	5.0	2.0	1.25	2.0	2.0	0.0	2.0	0.0	2.0	2.4	1.0
	3.89	4.0	0.0	2.0	1.1	2.0	1.1	5.0	2.0	0.007	2.0	2.0	0.0	2.0	0.0	2.0	2.0	0.0
	3.89	14.0	6.85	6.0	31.76	6.0	38.61	5.0	2.0	1.46	2.0	2.0	0.0	2.0	0.0	2.0	3.6	1.0
	3.89	14.0	0.0	2.0	1.03	2.0	1.03	5.0	2.0	0.0	2.0	2.0	0.0	2.0	0.0	2.0	2.0	0.0
	3.89	37.0	0.0	2.0	0.75	2.0	0.75	5.0	2.0	6.78	4.0	2.0	0.0	2.0	0.0	2.0	2.2	1.0
	3.89	37.0	2.4	6.0	16.63	6.0	19.03	5.0	2.0	1.16	2.0	2.0	0.0	2.0	0.0	2.0	3.6	1.0
	3.89	5.0	0.0	2.0	0.05	2.0	0.05	5.0	2.0	152.41	6.0	2.0	0.0	2.0	0.0	2.0	2.4	1.0
	3.89	5.0	0.0	2.0	1.76	2.0	1.76	5.0	2.0	1.08	2.0	2.0	0.0	2.0	0.0	2.0	2.0	0.0
	3.89	5.0	0.0	2.0	2.97	2.0	2.97	5.0	2.0	2.84	2.0	2.0	0.0	2.0	0.0	2.0	2.0	0.0
	3.89	5.0	0.0	2.0	0.43	2.0	0.43	5.0	2.0	0.0	2.0	2.0	0.0	2.0	0.0	2.0	2.0	0.0
	3.89	5.0	0.0	2.0	0.94	2.0	0.94	5.0	2.0	0.9	2.0	2.0	0.0	2.0	0.0	2.0	2.0	0.0
	3.89	20.0	9.01	6.0	19.82	6.0	28.83	5.0	2.0	9.67	4.0	2.0	0.0	2.0	0.0	2.0	3.8	1.0
	3.89	19.0	0.0	2.0	0.05	2.0	0.05	5.0	2.0	0.0	2.0	2.0	0.0	2.0	0.0	2.0	2.0	0.0
	3.89	19.0	11.95	6.0	30.9	6.0	42.85	5.0	2.0	32.68	6.0	2.0	0.0	2.0	0.0	2.0	4.0	1.0
	3.89	19.0	7.97	6.0	17.18	6.0	25.15	5.0	2.0	935.03	6.0	2.0	0.0	2.0	0.0	2.0	4.0	1.0
	3.89	19.0	0.0	2.0	3.71	2.0	3.71	5.0	2.0	29.63	6.0	2.0	0.0	2.0	0.0	2.0	2.4	1.0

Σχεδιάγραμμα 4. Το σύνολο των δεδομένων στο περιβάλλον του WEKA

4.2. Προσέγγιση

Η προτεινόμενη προσέγγιση που πρόκειται να εφαρμοστεί στο πρόγραμμα WEKA αποτελείται από τα πέντε παρακάτω βήματα:

1. Εύρεση δεδομένων στόχου (Target data finding)
2. Προ-επεξεργασία δεδομένων (Data pre-processing)
3. Ταξινόμηση (Classification)
4. Ομαδοποίηση-Συσταδοποίηση (Clustering)
5. Κανόνες συσχέτισης (Association rule mining)



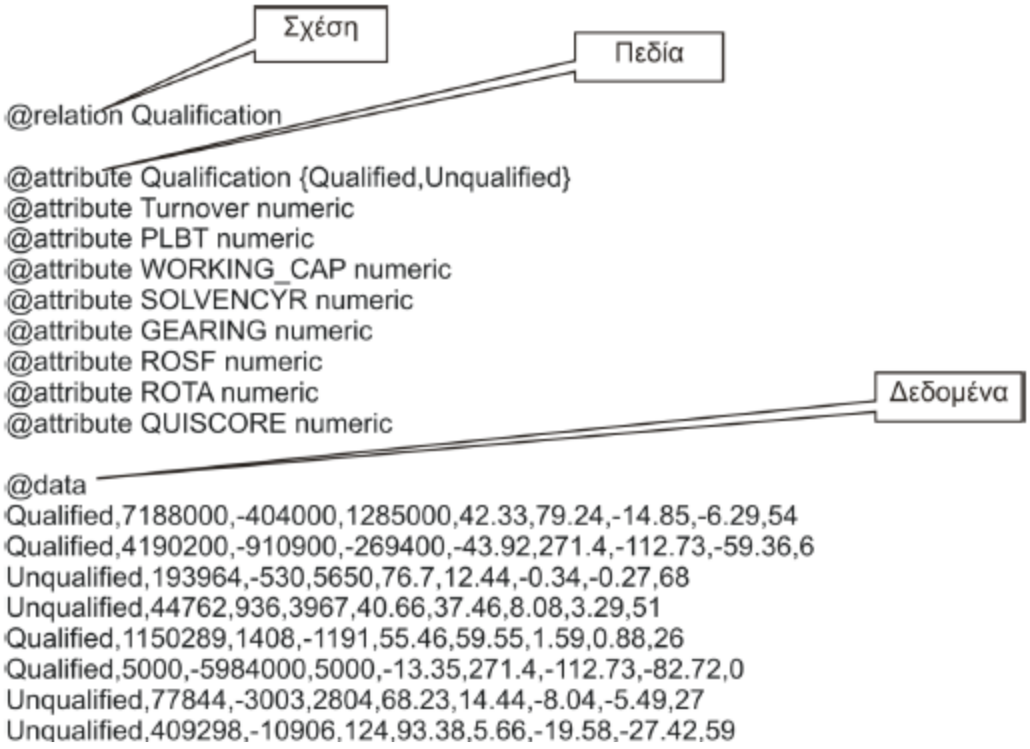
Σχεδιάγραμμα 5. Η προσέγγιση σε 5 βήματα

4.3. Προ-επεξεργασία των δεδομένων

Ένας επίδοξος αναλυτής, ο οποίος αναλαμβάνει εργασίες ανάλυσης πραγματικών δεδομένων ενός οργανισμού, πολύ σύντομα θα διαπιστώσει ότι τα δεδομένα που τηρούνται στα διάφορα πληροφοριακά συστήματα πάσχουν από πολλά και διαφορετικά προβλήματα (Kyrkos, 2015). Ένας αρχάριος αναλυτής πιθανότατα θα δυσφορούσε με όλα αυτά τα προβλήματα που καλείται να αντιμετωπίσει, πριν ακόμα αρχίσει την καθαυτό εργασία του. Ένας πιο έμπειρος αναλυτής όμως γνωρίζει ότι η ύπαρξη προβλημάτων είναι ο κανόνας στα δεδομένα του πραγματικού κόσμου. Έτσι, αμέσως αντιλαμβάνεται κανείς την αναγκαιότητα της προ-επεξεργασίας των δεδομένων, δηλαδή των εργασιών εκείνων για την προετοιμασία των δεδομένων, οι οποίες εκτελούνται πριν την καθαυτό εξόρυξη γνώσης. Η προεπεξεργασία των δεδομένων είναι απαραίτητη, καθώς τα αρχικά δεδομένα πάσχουν από διαφόρων ειδών προβλήματα. Σε αυτά συγκαταλέγονται η ύπαρξη αλληλοσυγκρουόμενων πληροφοριών, η ύπαρξη ασυνεπειών ως προς την κωδικοποίηση, την ονοματοδοσία πεδίων και τις μονάδες μέτρησης, καθώς και η ύπαρξη χαμένων τιμών και θορύβου, τυχαία δηλαδή κυμαινόμενων δεδομένων χωρίς ουσιαστικό περιεχόμενο. Τα προβληματικά αυτά δεδομένα καλούνται «ακάθαρτα» και η διαδικασία αντιμετώπισης των προβλημάτων τους καλείται «καθαρισμός δεδομένων». Η προεπεξεργασία των δεδομένων περιλαμβάνει τον καθαρισμό τους, αλλά δεν περιορίζεται σε αυτόν. Ειδικές απαιτήσεις των μεθόδων επεξεργασίας συχνά

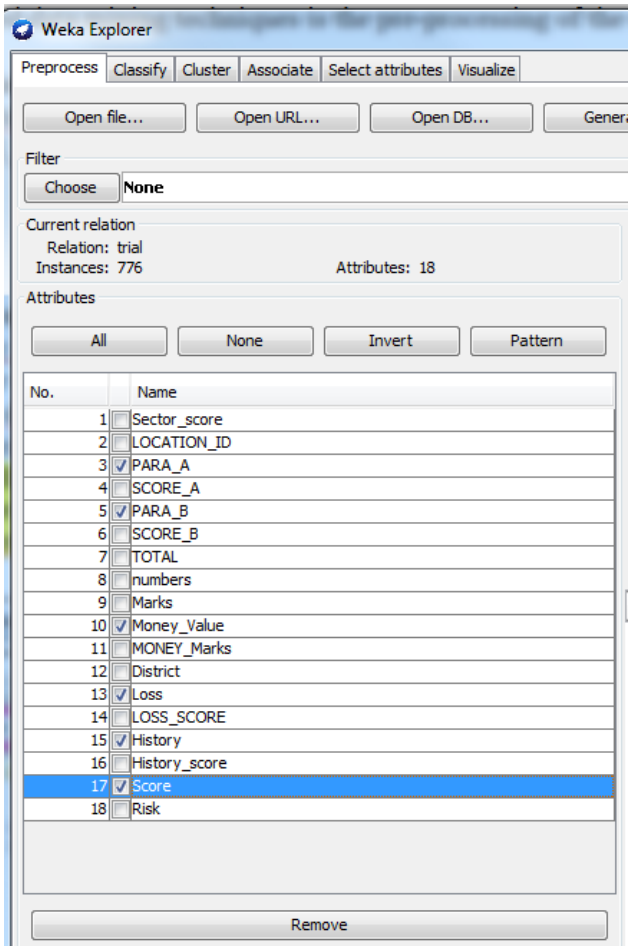
επιβάλλουν τον μετασχηματισμό των δεδομένων. Δύο συνήθεις εργασίες μετασχηματισμού είναι η διακριτοποίηση και η κανονικοποίηση. Ο όρος διακριτοποίηση αναφέρεται στον μετασχηματισμό αριθμητικών τιμών σε ονομαστικές τιμές. Η κανονικοποίηση είναι η μετατροπή αριθμητικών τιμών σε άλλες, πιο «κατάλληλες», αριθμητικές τιμές. Ένα επιπλέον θέμα που εμπίπτει στην προεπεξεργασία των δεδομένων είναι η μείωση του όγκου τους. Ειδική περίπτωση μείωσης των δεδομένων, με βαρύνουσα σημασία, είναι η επιλογή σημαντικών χαρακτηριστικών, η επιλογή δηλαδή εκείνων των μεταβλητών ή πεδίων που είναι απαραίτητες για την εξόρυξη της γνώσης (Kyrkos, 2015).

Κατά τη χρήση του WEKA Explorer, το πρώτο βήμα είναι η εισαγωγή των δεδομένων. Δεδομένα μπορούν να εισαχθούν από μια SQL βάση δεδομένων (με χρήση του Java Data Base Connectivity (JDBC)) ή από μια διεύθυνση URL. Ο πιο συνηθισμένος όμως τρόπος φόρτωσης δεδομένων είναι μέσω ενός αρχείου ARFF. Τα αρχεία ARFF είναι απλά αρχεία κειμένου, όπου οι τιμές διαχωρίζονται με κόμμα (Coma Separated Values (CSV)). Επιπλέον, το αρχείο περιέχει μια επικεφαλίδα, στην οποία ορίζονται το όνομα της σχέσης (πίνακα δεδομένων) και τα πεδία. Παράδειγμα αρχείου ARFF παρουσιάζεται στο Σχεδιάγραμμα 6. Στην αρχή του αρχείου αναφέρεται η λέξη ″@relation″ και ακολουθεί το όνομα του πίνακα δεδομένων (Qualification). Στη συνέχεια γίνεται η δήλωση των πεδίων. Για κάθε πεδίο χρειάζεται μια γραμμή, στην αρχή της οποίας υπάρχει η λέξη ″@attribute″, ακολουθεί το όνομα του πεδίου (πχ Turnover), και κατόπιν δηλώνεται ο τύπος του πεδίου. Αν το πεδίο είναι αριθμητικό, χρησιμοποιείται η λέξη ″numeric″. Αν το πεδίο είναι ονομαστικό, δηλώνονται οι δυνατές τιμές μέσα σε αγκύλες. Για παράδειγμα, το πρώτο πεδίο έχει όνομα ″Qualification″, είναι ονομαστικό και μπορεί να πάρει δύο τιμές, την τιμή ″Qualified″ και την τιμή ″Unqualified″. Μετά τη δήλωση των πεδίων ακολουθούν τα δεδομένα. Πριν από τα καθαυτό δεδομένα υπάρχει μια γραμμή με τη λέξη ″@data″. Τα δεδομένα είναι τιμές, οι οποίες χωρίζονται με κόμμα. Σημειώνεται ότι τα δεδομένα του παραδείγματος αφορούν επιχειρήσεις και τον εξωτερικό τους έλεγχο. Κάθε γραμμή των δεδομένων αντιστοιχεί σε μια επιχείρηση. Στο πρώτο πεδίο καταγράφεται το αποτέλεσμα του εξωτερικού ελέγχου. Επιχειρήσεις οι οποίες πήραν δυσμενή σχόλια από τους εξωτερικούς ελεγκτές χαρακτηρίζονται ″Qualified″, ενώ επιχειρήσεις οι οποίες δεν πήραν δυσμενή σχόλια από τους εξωτερικούς ελεγκτές χαρακτηρίζονται ″Unqualified″. Τα υπόλοιπα πεδία είναι διάφοροι αριθμοδείκτες. Αξίζει να σημειωθεί ότι στο Διαδίκτυο διατίθεται εφαρμογή μετατροπής αρχείων Excel σε αρχεία ARFF.



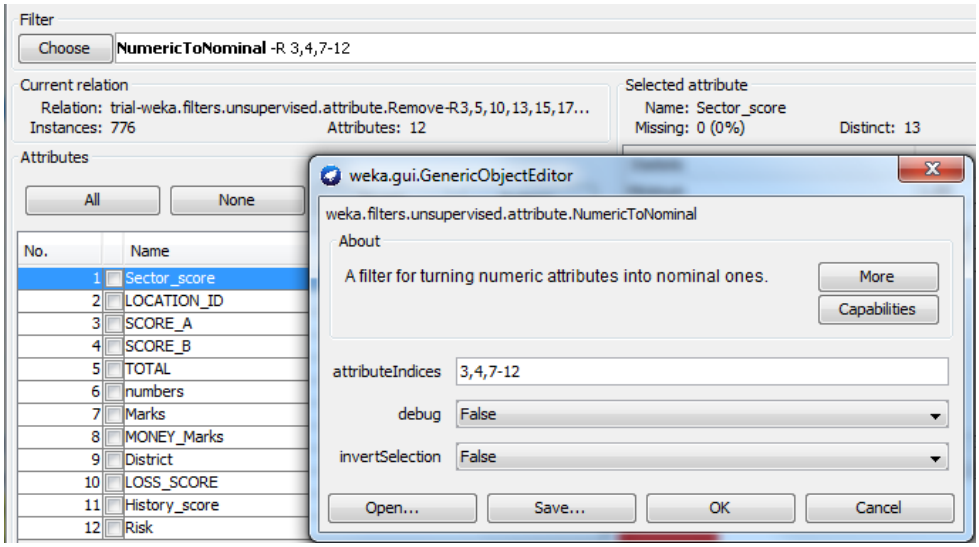
Σχεδιάγραμμα 6. Αρχείο ARFF

Το πρώτο βήμα πριν από την εφαρμογή των περιγραφόμενων τεχνικών εξόρυξης δεδομένων είναι η προεπεξεργασία των δεδομένων για την προετοιμασία τους για ανάλυση δεδομένων. Ορισμένα φίλτρα εφαρμόστηκαν στα δεδομένα: πρώτον, το φίλτρο Remove εφαρμόστηκε στα χαρακτηριστικά PARA_A, PARA_B, Money_Value, Loss, History and Score, καθώς, προφανώς, εξαρτώνται από τα χαρακτηριστικά SCORE_A, SCORE_B, Money_Marks Loss_Score, History_Score και Risk αντίστοιχα.



Σχεδιάγραμμα 7. Το φίλτρο Αφαίρεση (Remove)

Το φίλτρο NumericalToNominal εφαρμόστηκε στα χαρακτηριστικά SCORE_A, SCORE_B, Marks, MONEY_Marks, District, LOSS_SCORE, History_score και Risk προκειμένου να μετατραπούν οι αριθμητικές μεταβλητές και οι τιμές τους σε ονομαστικές. Τα χαρακτηριστικά με αριθμό 3, 4, 7-12 μετατρέπονται σε ονομαστικά.

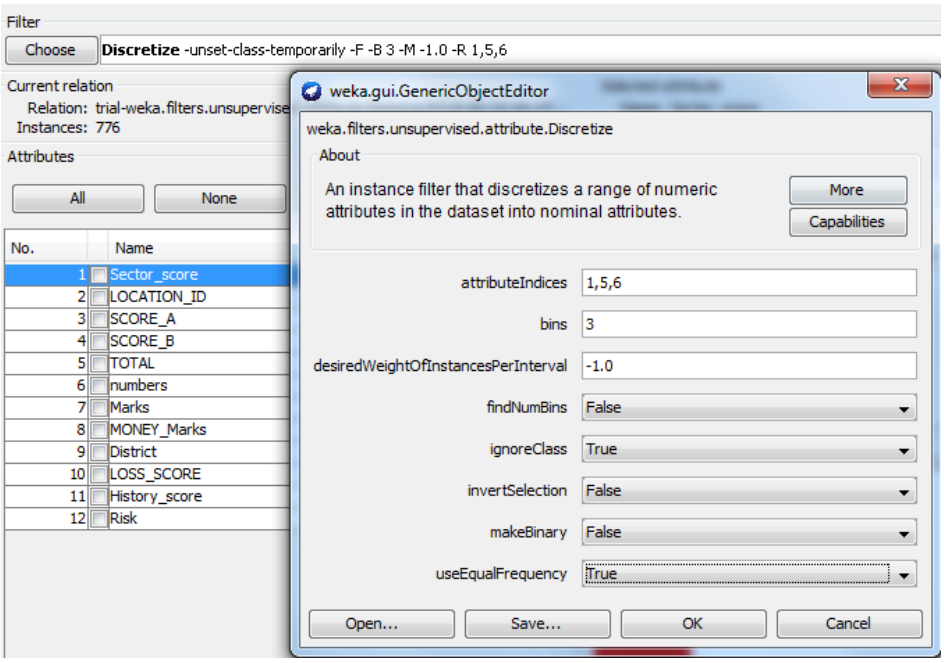


Σχεδιάγραμμα 8. Το φίλτρο NumericalToNominal

Επιπλέον, εφαρμόστηκε το φίλτρο Discretize (διακριτοποίηση) προκειμένου να διακριθούν οι αριθμητικές μεταβλητές Sector_score και TOTAL και να γίνουν ονομαστικές.

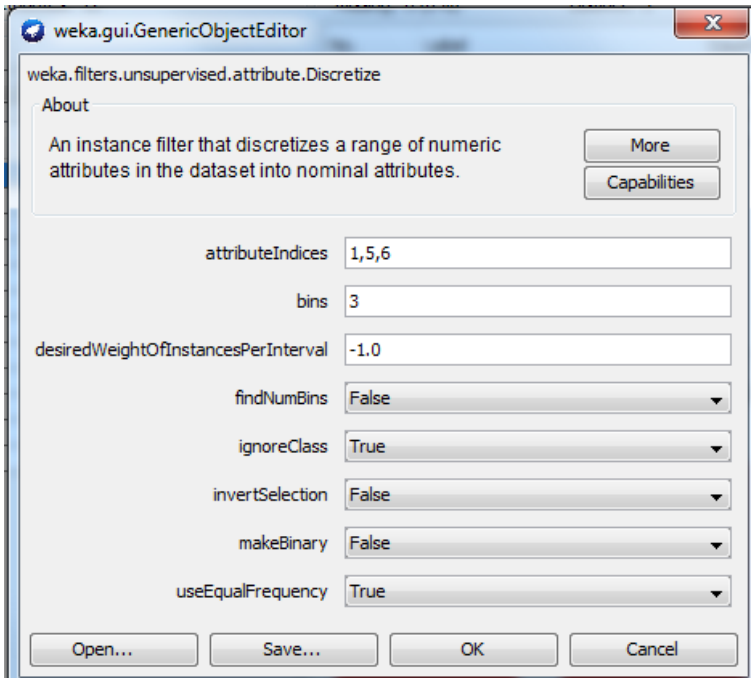
Η διακριτοποίηση (discretization) είναι μια διαδικασία μετασχηματισμού των δεδομένων. Για την ακρίβεια, είναι η διαδικασία μετατροπής αριθμητικών δεδομένων σε ονομαστικά δεδομένα, δεδομένα δηλαδή που οι τιμές τους αποτελούνται από ονομαστικές τιμές - λέξεις (Kyrkos, 2015). Εναλλακτικά μπορούμε να πούμε ότι η διακριτοποίηση είναι η μετατροπή ποσοτικών σε ποιοτικά δεδομένα. Κατά κανόνα, τα αριθμητικά δεδομένα χωρίζονται σε περιοχές τιμών και δημιουργούνται νέες στήλες, όπου στη θέση της αριθμητικής τιμής εισάγεται το όνομα της περιοχής της τιμής. Υπάρχουν πολλοί λόγοι για να διακριτοποιήσει κανείς τα δεδομένα του. Καταρχάς, ορισμένες μέθοδοι εξόρυξης δέχονται σαν είσοδο μόνο διακριτά δεδομένα. Σε περίπτωση που ο χρήστης θέλει να εφαρμόσει αυτές τις μεθόδους, είναι υποχρεωμένος να κάνει διακριτοποίηση. Επιπλέον, η διακριτοποίηση των δεδομένων μπορεί να επιταχύνει τη διαδικασία εκπαίδευσης των μοντέλων και να βελτιώσει τις επιδόσεις τους, αυξάνοντας έτσι την αποτελεσματικότητα και την αποδοτικότητα (Frank et al., 1999). Τέλος, η διακριτοποίηση μπορεί να οδηγήσει σε αποτελέσματα που είναι πιο κατανοητά. Για όλους αυτούς τους λόγους, η διακριτοποίηση έχει αποτελέσει αντικείμενο έρευνας, η οποία απέδωσε διάφορες τεχνικές.

Το Σχεδιάγραμμα 9 απεικονίζει όλες τις μεταβλητές που χρησιμοποιήθηκαν στην ανάλυσή μας.



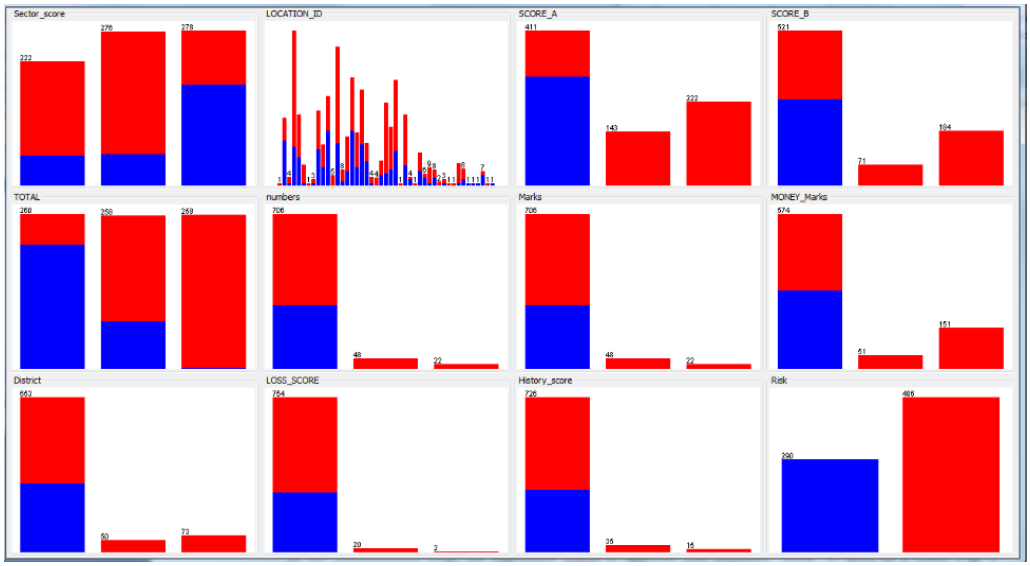
Σχεδιάγραμμα 9. Το φίλτρο Discretize

Οι επιλογές διακριτοποίησης απεικονίζονται στο Σχεδιάγραμμα 10.



Σχεδιάγραμμα 10. Οι επιλογές διακριτοποίησης (Discretize)

Οπτικοποιώντας όλα τα παραπάνω, είναι δυνατή η εμφάνιση των γραφικών απεικονίσεων κάθε χαρακτηριστικού σε σχέση με οποιοδήποτε άλλο χαρακτηριστικό όπως απεικονίζεται παρακάτω. Στο τελευταίο tab του WEKA Explorer παρέχονται εργαλεία οπτικοποίησης των δεδομένων. Η οπτικοποίηση είναι πολύ χρήσιμη στην πράξη, καθώς επιτρέπει στον χρήστη να κατανοήσει με εύκολο και γρήγορο τρόπο τη διασπορά των παρατηρήσεων.



Σχεδιάγραμμα 11. Οπτικοποίηση των χαρακτηριστικών με μεταβλητή κλάσης "Risk"

4.4. Ταξινόμηση (Classification)

Η ταξινόμηση ή αλλιώς όπως μπορούμε να τη βρούμε στη βιβλιογραφία κατηγοριοποίηση (classification) είναι μια από τις βασικότερες εργασίες Εξόρυξης Δεδομένων, με μεγάλο αριθμό εφαρμογών στον χώρο των οικονομικών. Η πρόβλεψη χρεοκοπίας, η έγκριση δανείων, η αναγνώριση απάτης είναι τυπικά προβλήματα κατηγοριοποίησης (Kyrkos, 2015). Η κατηγοριοποίηση είναι εργασία επιβλεπόμενης μάθησης. Στόχος της επιβλεπόμενης μάθησης είναι η ανακάλυψη της σχέσης ανάμεσα σε ένα γνώρισμα στόχο και σε ένα σύνολο άλλων γνωρισμάτων. Το γνώρισμα στόχος αναφέρεται και ως εξαρτημένη μεταβλητή, ενώ τα υπόλοιπα γνωρίσματα αναφέρονται και ως ανεξάρτητες μεταβλητές. Με την επιβλεπόμενη μάθηση επιτυγχάνεται η δημιουργία ενός μηχανισμού λήψης αποφάσεων ή υπολογισμών, ο οποίος είναι ικανός να προβλέπει τις τιμές της εξαρτημένης μεταβλητής χρησιμοποιώντας τις ανεξάρτητες μεταβλητές. Ο μηχανισμός λήψης απόφασης καλείται και μοντέλο και μπορεί να έχει διάφορες μορφές, όπως πχ να είναι ένα σύνολο κανόνων ή μια

εξίσωση ή το πλέγμα των νευρώνων και συνδέσεων ενός Νευρωνικού Δικτύου. Στην επιβλεπόμενη μάθηση ανήκουν η Κατηγοριοποίηση (Classification) και η Παλινδρόμηση (Regression). Η κατηγοριοποίηση και η παλινδρόμηση έχουν πολλές ομοιότητες. Και στις δύο περιπτώσεις στόχος είναι η πρόβλεψη των τιμών ενός γνωρίσματος, με χρήση άλλων γνωρισμάτων. Επίσης, και στις δύο περιπτώσεις χρησιμοποιείται ένα σύνολο δεδομένων εκπαίδευσης, με την επεξεργασία του οποίου κατασκευάζεται το μοντέλο. Η διαφορά ανάμεσα στην κατηγοριοποίηση και στην παλινδρόμηση έχει σχέση με τον τύπο της εξαρτημένης μεταβλητής. Στόχος της παλινδρόμησης είναι η πρόβλεψη μιας εξαρτημένης μεταβλητής, η οποία περιέχει συνεχόμενες (αριθμητικές) τιμές. Αντιθέτως, κατηγοριοποίηση είναι η πρόβλεψη διακριτών ονομαστικών τιμών. Οι τιμές αυτές είναι συγκεκριμένες, γνωστές εκ των προτέρων και ορίζουν την κλάση (κατηγορία) στην οποία ανήκει κάθε αντικείμενο. Για τον λόγο αυτό, η εξαρτημένη μεταβλητή σε προβλήματα κατηγοριοποίησης καλείται και γνώρισμα κλάσης. Με την ένταξη αντικειμένων σε ομάδες ασχολείται και μια άλλη εργασία Εξόρυξης Δεδομένων, η Ανάλυση Συστάδων (Clustering) (Kyrkos, 2015). Οι διαφορές ανάμεσα στην Ανάλυση Συστάδων και στην Κατηγοριοποίηση είναι μεγάλες. Η Ανάλυση Συστάδων επιμερίζει τα αντικείμενα σε ομάδες βάσει της ομοιότητας τους. Οι συστάδες και το πλήθος τους δεν είναι εκ των προτέρων γνωστές. Επίσης, δεν υπάρχει στα δεδομένα κάποιο πεδίο που να καθορίζει την ομάδα στην οποία ανήκει το κάθε αντικείμενο. Αντιθέτως, στην κατηγοριοποίηση οι κατηγορίες είναι εκ των προτέρων γνωστές. Οι τιμές του γνωρίσματος κλάσης ορίζουν την κατηγορία στην οποία ανήκει κάθε αντικείμενο. Ένα παράδειγμα προβλήματος κατηγοριοποίησης είναι η έγκριση των τραπεζικών δανείων. Το σύνολο δεδομένων περιλαμβάνει στοιχεία για τον υποψήφιο δανειολήπτη, στοιχεία σχετικά με το δάνειο, καθώς επίσης και την τελική απόφαση για την έγκριση ή την απόρριψη του δανείου. Κάθε γραμμή του συνόλου δεδομένων αντιστοιχεί σε μια αίτηση. Οι γραμμές καλούνται και αντικείμενα, παραδείγματα ή παρατηρήσεις. Οι στήλες αναφέρονται σε μια ιδιότητα των αντικειμένων, όπως πχ το επάγγελμα του δανειολήπτη ή το είδος του δανείου (στεγαστικό, καταναλωτικό κλπ.). Οι στήλες καλούνται και πεδία (fields), μεταβλητές (variables), γνωρίσματα (attributes) ή χαρακτηριστικά (features). Το γνώρισμα το οποίο περιέχει την απόφαση της έγκρισης ή απόρριψης του δανείου είναι το γνώρισμα της κλάσης. Η έγκριση του δανείου εξαρτάται από τα στοιχεία της αίτησης, όπως η ηλικία, το επάγγελμα και η οικονομική κατάσταση του δανειολήπτη, το ποσό και ο τύπος του δανείου κλπ. Η δημιουργία ενός μοντέλου, το οποίο θα μπορεί να

προβλέπει την έγκριση ή απόρριψη του δανείου χρησιμοποιώντας τα υπόλοιπα στοιχεία της αίτησης, είναι ένα πρόβλημα κατηγοριοποίησης.

Η ταξινόμηση-κατηγοριοποίηση περιλαμβάνει τρία στάδια, το στάδιο της επιβλεπόμενης μάθησης, το στάδιο της επικύρωσης του μοντέλου και το στάδιο της χρήσης του μοντέλου. Αναλυτικότερα, οι εργασίες που λαμβάνουν χώρα σε κάθε στάδιο είναι οι ακόλουθες (Kyrkos, 2015):

1. Επιβλεπόμενη μάθηση. Στο στάδιο αυτό, μια μέθοδος κατηγοριοποίησης αναλύει ένα σύνολο δεδομένων. Η μέθοδος θα ανακαλύψει σχέσεις μεταξύ της εξαρτημένης μεταβλητής και των ανεξάρτητων μεταβλητών. Το αποτέλεσμα αυτής της επεξεργασίας είναι η κατασκευή ενός μοντέλου. Η κατασκευή ή εκπαίδευση του μοντέλου καθοδηγείται από τις τιμές του γνωρίσματος της κλάσης και για τον λόγο αυτό η διαδικασία ονομάζεται επιβλεπόμενη μάθηση. Το σύνολο δεδομένων, το οποίο χρησιμοποιείται για την εκπαίδευση του μοντέλου, ονομάζεται σύνολο εκπαίδευσης (training data set). Η επιλογή του συνόλου εκπαίδευσης είναι καθοριστικής σημασίας, γιατί το μοντέλο που θα προκύψει θα αποτυπώνει σχέσεις που υπάρχουν στο σύνολο εκπαίδευσης. Μεροληπτικά σύνολα εκπαίδευσης θα οδηγήσουν στην κατασκευή μεροληπτικών μοντέλων (Kyrkos, 2015).
2. Επικύρωση μοντέλου. Στο στάδιο αυτό δοκιμάζεται η ακρίβεια του μοντέλου, η ικανότητα του δηλαδή να προβλέπει σωστά την κλάση των παρατηρήσεων. Το μοντέλο τροφοδοτείται με παρατηρήσεις, των οποίων η κλάση είναι γνωστή. Αναλύοντας τα στοιχεία των ανεξάρτητων μεταβλητών κάθε παρατήρησης, το μοντέλο προβλέπει την κλάση της παρατήρησης και στη συνέχεια συγκρίνεται η πρόβλεψη του μοντέλου με την πραγματική τιμή της κλάσης. Αν το μοντέλο επιδείξει ικανοποιητική ακρίβεια προβλέψεων, εάν δηλαδή προβλέψει σωστά την κλάση ενός ικανοποιητικού ποσοστού παρατηρήσεων, τότε θεωρείται επιτυχημένο και μπορεί να χρησιμοποιηθεί για τη διατύπωση προβλέψεων. Η διαδικασία δοκιμής του μοντέλου καλείται επικύρωση (validation) και το σύνολο δεδομένων που χρησιμοποιείται για τη δοκιμή καλείται σύνολο επικύρωσης (validation set). Σκοπός ενός μοντέλου είναι να χρησιμοποιηθεί για τη διατύπωση προβλέψεων στην πραγματική ζωή και όχι να αναλύσει ένα συγκεκριμένο σύνολο δεδομένων. Το μοντέλο πρέπει να αποδείξει την ικανότητα του να προβλέπει την κλάση άγνωστων παρατηρήσεων, παρατηρήσεων δηλαδή διαφορετικών από αυτές που

χρησιμοποιήθηκαν για την εκπαίδευση του. Για τον λόγο αυτό, το σύνολο εκπαίδευσης και το σύνολο επικύρωσης πρέπει να περιέχουν διαφορετικές παρατηρήσεις (Kyrkos, 2015).

3. Χρήση του μοντέλου. Το μοντέλο, αφού εκπαιδευτεί και επικυρωθεί, χρησιμοποιείται για τη διατύπωση προβλέψεων. Μια νέα παρατήρηση, της οποίας η κλάση είναι άγνωστη, εισάγεται στο μοντέλο. Το μοντέλο χρησιμοποιώντας τις τιμές των ανεξάρτητων μεταβλητών υπολογίζει την τιμή της κλάσης (Kyrkos, 2015).

Σε ότι αφορά *τα κριτήρια αξιολόγησης μεθόδων κατηγοριοποίησης* τώρα, η έρευνα σχετικά με την κατηγοριοποίηση έχει αποδώσει πλούσιους καρπούς και σήμερα υπάρχουν διαθέσιμες αρκετές και πολύ διαφορετικές μέθοδοι κατηγοριοποίησης (Kyrkos, 2015). Ορισμένες από αυτές, όπως πχ τα Νευρωνικά Δίκτυα, θεωρούνται ιδιαίτερα ικανές να προβλέπουν την κλάση άγνωστων παρατηρήσεων. Οι μέθοδοι αυτές μπορούν να θεωρηθούν «καλύτερες» από άλλες, όμως η ακρίβεια δεν είναι το μοναδικό κριτήριο αξιολόγησης των μεθόδων κατηγοριοποίησης. Αναλυτικότερα, οι μέθοδοι κατηγοριοποίησης μπορούν να αξιολογηθούν με βάση τα παρακάτω κριτήρια:

- Ακρίβεια πρόβλεψης (accuracy). Είναι η ικανότητα των μοντέλων να προβλέπουν την κλάση άγνωστων παρατηρήσεων. Προφανώς πρόκειται για ένα πολύ σημαντικό κριτήριο και μεγάλο μέρος της έρευνας προσανατολίζεται στην ανακάλυψη μεθόδων υψηλών επιδόσεων (Kyrkos, 2015).
- Ταχύτητα (speed). Σχετίζεται με την πολυπλοκότητα της μεθόδου και το υπολογιστικό κόστος που αυτή συνεπάγεται. Η εκτέλεση περίπλοκων αλγορίθμων, οι οποίοι απαιτούν εκτεταμένους υπολογισμούς, προκαλούν καθυστερήσεις. Καθυστερήσεις μπορεί να υπάρχουν στη διαδικασία κατασκευής, αλλά και στη χρήση των μοντέλων, στην εφαρμογή τους δηλαδή για την κατηγοριοποίηση μιας νέας παρατήρησης. Ορισμένες μέθοδοι, όπως τα Δένδρα Αποφάσεων, διαθέτουν γρήγορους αλγορίθμους και ο χρόνος κατασκευής των μοντέλων είναι μικρός. Άλλες μέθοδοι, όπως τα Νευρωνικά Δίκτυα, χρειάζονται πολύ περισσότερο χρόνο για την εκπαίδευση των μοντέλων. Κατά κανόνα ο χρόνος χρήσης των μοντέλων είναι πολύ μικρός. Ωστόσο, υπάρχουν μέθοδοι, όπως οι k-Πλησιέστεροι Γείτονες, οι οποίες δεν εκπαιδεύουν κάποιο μοντέλο, όμως ο χρόνος για την κατηγοριοποίηση νέων παρατηρήσεων είναι μεγάλος (Kyrkos, 2015).
- Ερμηνευσιμότητα (interpretability). Είναι η ικανότητα της μεθόδου να παράγει μοντέλα, τα οποία είναι κατανοητά από τον άνθρωπο. Για παράδειγμα, στα Δένδρα Αποφάσεων

ο τρόπος λήψης της απόφασης κατηγοριοποίησης είναι απολύτως κατανοητός και το μοντέλο μπορεί εύκολα να μετατραπεί σε ένα σύνολο κανόνων της μορφής EAN-ΤΟΤΕ. Αντιθέτως, τα μοντέλα άλλων μεθόδων, όπως τα Νευρωνικά Δίκτυα και οι Μηχανές Διανυσμάτων Υποστήριξης, λειτουργούν ως «μαύρα κουτιά». Στα μοντέλα αυτά παρέχονται οι τιμές των μεταβλητών εισόδου και υπολογίζεται η απόφαση κατηγοριοποίησης στην έξοδο. Ο τρόπος λήψης της απόφασης όμως δεν είναι κατανοητός στον άνθρωπο. Η ερμηνευσιμότητα είναι μια σημαντική ιδιότητα των μεθόδων κατηγοριοποίησης. Σε πολλές περιπτώσεις οι χρήστες των μοντέλων επιθυμούν να γνωρίζουν τον τρόπο λήψης της απόφασης, ώστε να είναι πιο σίγουροι για το αποτέλεσμα. Επίσης, στο μοντέλο καταγράφονται σχέσεις μεταξύ των δεδομένων. Ορισμένες από τις σχέσεις αυτές μπορεί να είναι νέες και άγνωστες. Αν το μοντέλο είναι ερμηνεύσιμο θα αποκαλυφθούν οι νέες σχέσεις και η μέθοδος κατηγοριοποίησης θα χρησιμοποιηθεί ως εργαλείο ανάλυσης, ικανό να προσφέρει καινοτόμα γνώση (Kyrkos, 2015).

- Επεκτασιμότητα (scalability). Αναφέρεται στην ικανότητα των μεθόδων να χειριστούν πολύ μεγάλα σύνολα δεδομένων. Η Μηχανική Μάθηση και η Στατιστική προσφέρουν μεθόδους κατηγοριοποίησης. Ωστόσο, η εφαρμογή αυτών των μεθόδων για την επεξεργασία δεδομένων μεγάλου όγκου δεν είναι πάντα εύκολη. Σε αρκετές περιπτώσεις η υπολογιστική πολυπλοκότητα των μεθόδων είναι συνάρτηση του πλήθους των παρατηρήσεων και μάλιστα με σχέση περισσότερο από γραμμική. Επίσης, οι περισσότερες μέθοδοι απαιτούν την εγκατάσταση του συνόλου εκπαίδευσης στην κύρια μνήμη του υπολογιστή. Τα ζητήματα αυτά θέτουν όρια στη δυνατότητα εφαρμογής των μεθόδων. Όμως αντικείμενο της Εξόρυξης Δεδομένων είναι η ανακάλυψη γνώσης από δεδομένα μεγάλου όγκου. Ειδικά στη σημερινή εποχή, η παραγωγή και καταγραφή δεδομένων είναι μαζικότατη. Σε ότι αφορά την εφαρμογή των μεθόδων αυτών για επιχειρηματικούς σκοπούς, η τάση που παρουσιάστηκε στα τέλη της δεκαετίας του 90’ για δημιουργία Αποθηκών Δεδομένων, έχει οδηγήσει στην αποθήκευση δεδομένων, που ο όγκος τους είναι της τάξης μεγέθους terabyte. Για να έχουν πρακτική χρησιμότητα οι μέθοδοι Εξόρυξης Δεδομένων πρέπει να είναι ικανές να χειριστούν αυτά τα πολύ μεγάλου όγκου δεδομένα. Όπως χαρακτηριστικά επισημαίνουν οι Fayyad, Piatetsky και Smyth (1996), η πρόκληση για την κοινότητα των ερευνητών Εξόρυξης Δεδομένων είναι η κατασκευή μεθόδων που διευκολύνουν τη χρήση αλγορίθμων εξόρυξης δεδομένων σε βάσεις δεδομένων του πραγματικού κόσμου (Kyrkos, 2015).

- Ανθεκτικότητα (robustness). Αναφέρεται στην ικανότητα των μεθόδων να πραγματοποιήσουν ορθές προβλέψεις, όταν τα δεδομένα χαρακτηρίζονται από προβλήματα, όπως ο θόρυβος και οι χαμένες τιμές (Kyrkos, 2015).

Λαμβάνοντας υπόψιν όλα τα παραπάνω κριτήρια σχετικά με τις μεθόδους ταξινόμησης, στα πλαίσια της πρακτικής εφαρμογής που μελετάται, επιλέχθηκε να εφαρμοστεί ο αλγόριθμος *OneR*. Ως κλάση χρησιμοποιείται το χαρακτηριστικό "Risk". Παρακάτω αναλύεται ο τρόπος λειτουργίας του καθώς και για ποιο λόγο προτιμήθηκε έναντι άλλων αλγορίθμων.

Ο OneR ή “One Rule” είναι ένας απλός αλγόριθμός που προτάθηκε από τον Holt. Ο OneR κατασκευάζει έναν κανόνα για κάθε μεταβλητή στα δεδομένα εκπαίδευσης και μετά επιλέγει τον κανόνα με το μικρότερο ποσοστό σφάλματος. Για να δημιουργηθεί ένας κανόνας για μια μεταβλητή, η πιο συχνή κλάση για κάθε τιμή της μεταβλητής πρέπει να προσδιοριστεί. Η πιο συχνή κλάση είναι απλά η κλάση που εμφανίζεται πιο συχνά για αυτή την τιμή της μεταβλητής. Ένας κανόνας είναι απλά ένα σύνολο από τιμές μεταβλητών που δεσμεύεται στην κλάση πλειοψηφίας τους. Ο OneR επιλέγει τον κανόνα με το χαμηλότερο ποσοστό σφάλματος. Σε περίπτωση που δύο ή περισσότεροι κανόνες έχουν το ίδιο ποσοστό σφάλματος, ο κανόνας επιλέγεται τυχαία.

Με άλλα λόγια, ένας εύκολος τρόπος παραγωγής πολύ απλών κανόνων ταξινόμησης από ένα σύνολο από instances είναι ο αλγόριθμος OneR, ο οποίος παράγει ένα δέντρο αποφάσεων μονού επιπέδου, εκφρασμένο ως ένα σύνολο κανόνων οι οποίοι εξετάζουν το ίδιο συγκεκριμένο attribute. Προκύπτει συχνά, ότι η απόδοση τον OneR είναι πολύ υψηλή, ίσως επειδή τα πραγματικά datasets συχνά κρύβουν στοιχειώδη μοτίβα, και ένα attribute αρκεί για να βρεθεί η σωστή κλάση με υψηλή ακρίβεια. Ξεκινώντας με τις ονομαστικές τιμές, για κάθε attribute τον dataset κτίζουμε ένα σύνολο κανόνων. Οι κανόνες ανά κάθε τέτοιο σύνολο είναι ένας ανά τιμή του attribute. Για κάθε κανόνα, κοιτάμε για αυτήν την τιμή ποια κλάση εμφανίζεται στα instances εισόδου πιο συχνά, και την προσάπτουμε ως συμπέρασμα τον κανόνα αυτού. Έπειτα μετράμε για κάθε εμφάνιση αυτής της τιμής τον συγκεκριμένου attribute στο dataset, πόσες φορές δεν εμφανίζεται η κλάση πού έχει επιλεχθεί για τον κανόνα, και διαιρούμε με το πλήθος εμφανίσεων της τιμής αυτής. Έτσι προκύπτει το ποσοστό λάθους του κανόνα. Επαναλαμβάνουμε για κάθε attribute και τιμή attribute. Έπειτα για κάθε attribute επιλέγουμε τον κανόνα (άρα και τιμή) με το μικρότερο

ποσοστό λάθους. Στη συνέχεια από όλα τα attribute επιλέγουμε αυτό με τον κανόνα με το μικρότερο ποσοστό λάθους, και προκύπτει έτσι το τελικό ζητούμενο attribute. Ισοπαλίες στα ποσοστά λαθών αντιμετωπίζονται με τυχαία επιλογή ενός συγκρινόμενού από όλους τούς ισόπαλους. Οι τιμές που λείπουν σε ένα dataset αντιμετωπίζονται συνολικά ανά attribute: έστω και μια φορά να λείπει τιμή σε ένα attribute, αυτό αποκτά και μία νέα τιμή στο πεδίο ορισμού του, την "missing". Ο χειρισμός δε των αριθμητικών τιμών , υλοποιείται μέσω της διαδικασίας της διακριτοποίησης: πρώτα ταξινομούνται τα instances ως προς το επιλεχθέν από τον αλγόριθμο attribute, και έπειτα χωρίζεται το εύρος των τιμών σε διακριτές περιοχές, οι οποίες μπορούν να ονοματιστούν, έτσι ώστε η διαδικασία του OneR που Θα ακολουθήσει να είναι ολόιδια με όταν έχω ονομαστικές τιμές. Ένας τρόπος διαχωρισμού είναι όποτε αλλάζει στην ταξινομημένη ακολουθία η τιμή της κλάσης. Το πότε αλλάζει όμως δεν είναι σαφές. Επιλέγουμε συνήθως λοιπόν να Θέσουμε το σημείο διαχωρισμού στον μέσο όρο των τιμών του επιλεχθέντος attribute, των συνοριακών instances των δύο γειτονικών διαστημάτων. Όταν όμως οι τιμές είναι ίδιες, μετακινούμε το σημείο διαχωρισμού ξεπερνώντας το πρόβλημα, αλλά αποκτώντας ένα μικτό διάστημα ως προς την κλάση instances του.

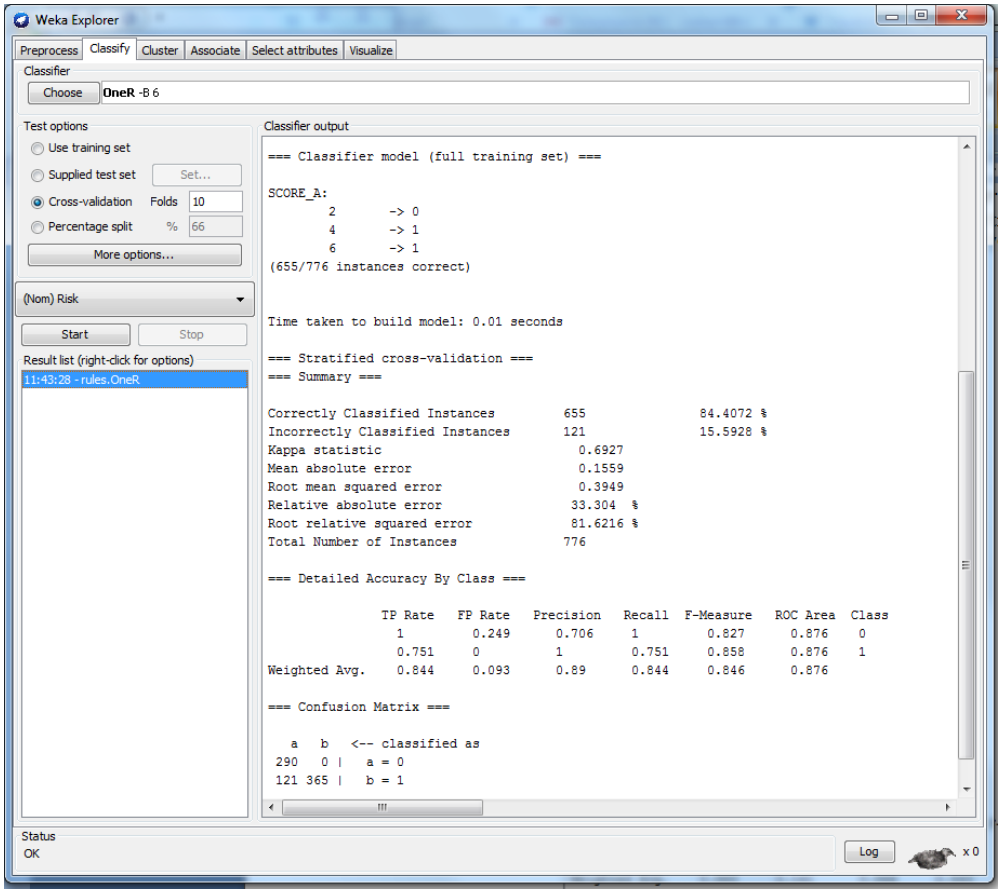
Το σοβαρότερο πρόβλημα που έχει να αντιμετωπίσει ο OneR, είναι το overfitting. Όταν το επιλεχθέν attribute έχει πολλές πιθανές τιμές, τότε δημιουργούνται πολλοί κανόνες για το attribute, όμως στην περίπτωση που οι τιμές του attribute είναι ομοιόμορφα κατανεμημένες στο dataset, τότε κάθε κανόνας Θα έχει μικρό ποσοστό λάθους, γιατί διευκολύνεται η επιλογή στην κλάση να αποτελεί την μεγάλη πλειοψηφία. Στην οριακή περίπτωση, ένα attribute αποτελείται από τιμές-δείκτες-αναγνωριστικά για τα instances, με αποτέλεσμα να προκύπτουν τόσοι κανόνες όσα και instances. Προφανώς, ο τελικός κανόνας που θα δώσει ο OneR θα είναι ακραία εξαρτώμενος από το training dataset, άρα μη-αξιοποιήσιμος. Το παραπάνω πρόβλημα overfitting προφανώς προκύπτει για αριθμητικές τιμές. Η λύση είναι να δίνεται ένα ελάχιστο όριο στο πλήθος των instances με κλάση αυτήν τον κανόνα(κλάση πλειοψηφίας), στα διαστήματα της διακριτοποίησης. Σύμφωνα με τη δημοσίευση του Holte (1993), μία προτεινόμενη τιμή για το όριο, η οποία προέκυψε πειραματικά, είναι το έξι. Ακόμα και με την εφαρμογή αυτού τον ορίου, αλλά και γενικότερα, όταν ένα instance έχει κλάση αυτή της πλειοψηφίας του γειτονικού της διαστήματος, μεταφέρεται αυτό το instance στο γειτονικό διάστημα, καθώς δε θα επηρεάσει το ποια είναι η κλάση πλειοψηφίας. Αυτό γίνεται, γιατί επιθυμούμε όσο το δυνατόν λιγότερα διαστήματα. Στο ίδιο σκεπτικό, δύο γειτονικά διαστήματα με ίδια κλάση πλειοψηφίας συγχωνεύονται σε μία ολική με ίδια πάλι

κλάση πλειοψηφίας. Τέλος, εάν προκύψουν αριθμητικές τιμές που λείπουν, τότε δημιουργείται η τιμή "missing" , αλλά η διακριτοποίηση εφαρμόζεται μόνο στα instances που έχουν τιμή. Εναλλακτικά, ένας πιο εκφραστικός τρόπος υλοποίησης τον OneR, είναι να δημιουργείται ένας κα-νόνας ανά κλάση. Κάθε κανόνας είναι μια σύζευξη από ελέγχους, ένας για κάθε attribute. Στις αριθμητικές τιμές, κάθε έλεγχος κοιτά εάν η τιμή τον αντίστοιχού attribute σε ένα νέο instance είναι εντός ενός διαστήματος, ενώ στις ονομαστικές εάν είναι εντός ενός υποσυνόλου τιμών, για το attribute αυτό. Κάθε αριθμητικό διάστημα αντιστοιχεί σε μία κλάση και έχει στα άκρα τον το μέγιστο και το ελάχιστο των τιμών του συγκεκριμένου attribute πού εμφανίζονται για την κλάση αυτήν. Ομοίως το ονομαστικό υποσύνολο δεν έχει μεν ακρότατα, αλλά περιέχει όλες τις τιμές τον attribute για αυτήν την κλάση.

Όπως εξηγήθηκε αναλυτικά στην Ενότητα 4.3, η προεπεξεργασία των δεδομένων είναι ένα απαραίτητο στάδιο, το οποίο προηγείται της καθαυτό εξόρυξης δεδομένων. Για την περίπτωση της κατηγοριοποίησης, η προεπεξεργασία μπορεί να βελτιώσει την αποτελεσματικότητα, την αποδοτικότητα και την επεκτασιμότητα των μεθόδων. Στο στάδιο της προεπεξεργασίας αντιμετωπίζεται το πρόβλημα του θορύβου και των χαμένων τιμών. Επίσης, τα δεδομένα μπορούν να αναχθούν σε υψηλότερα επίπεδα γενίκευσης, να διακριτοποιηθούν ώστε να μετατραπούν τα αριθμητικά πεδία σε ονομαστικά και τέλος, να κανονικοποιηθούν, να αντικατασταθούν δηλαδή οι αριθμητικές τιμές με άλλες, πιο «κατάλληλες», αριθμητικές τιμές. Σε πολλές περιπτώσεις η διακριτοποίηση και η κανονικοποίηση είναι απαραίτητες, ώστε να προσαρμοστούν τα δεδομένα σε ιδιαιτερότητες των μεθόδων κατηγοριοποίησης. Για παράδειγμα, η μέθοδος των k-Πλησιέστερων Γειτόνων είναι ιδιαίτερα ευπαθής σε δεδομένα που περιέχουν πεδία με πολύ μεγάλες τιμές και πεδία με πολύ μικρές τιμές. Ιδιαίτερης σημασίας είναι το ζήτημα του πλήθους των διαστάσεων και της επιλογής χαρακτηριστικών. Ερευνητικές εργασίες έχουν αποδείξει ότι το πλήθος των διαστάσεων είναι άμεσα συναρτημένο με το πλήθος των παρατηρήσεων, οι οποίες είναι απαραίτητες για την κατασκευή των μοντέλων. Ανάλογα με το είδος του κατηγοριοποιητή, το πλήθος των παρατηρήσεων μπορεί να είναι γραμμική ή και εκθετική συνάρτηση του πλήθους των διαστάσεων (Fukunaga, 1990; Hwang και Lippman, 1994). Ωστόσο, η επιλογή χαρακτηριστικών δεν αποτελεί πανάκεια. Σε ορισμένες περιπτώσεις, είναι πιθανόν να επιλεχθεί ένας μεγάλος αριθμός μεταβλητών εισόδου. Αυτό συμβαίνει όταν το γνώρισμα της κλάσης εξαρτάται ουσιαστικά από πολλά άλλα γνωρίσματα. Επίσης, ορισμένες μέθοδοι

συναρτούν το πλήθος των επιλεγμένων γνωρισμάτων με το πλήθος των παρατηρήσεων που χρησιμοποιούνται. Αν οι παρατηρήσεις που θα χρησιμοποιηθούν για την επιλογή χαρακτηριστικών είναι λίγες, τότε και τα επιλεγμένα χαρακτηριστικά θα είναι λίγα. Το αποτέλεσμα είναι η απόρριψη σημαντικών χαρακτηριστικών και ο αποκλεισμός τους από τη διαδικασία της κατηγοριοποίησης. Σε κάθε περίπτωση, ο αναλυτής θα πρέπει να έχει επίγνωση του προβλήματος των διαστάσεων και της επιλογής σημαντικών χαρακτηριστικών σε εργασίες κατηγοριοποίησης. Ο αναλυτής θα πρέπει πιθανώς να πειραματιστεί με διαφορετικές μεθόδους επιλογής χαρακτηριστικών και να μην επαφίεται άκριτα σε μία μόνο μέθοδο.

Το Σχεδιάγραμμα 12 παρουσιάζει τη συνολική ακρίβεια του μοντέλου που υπολογίστηκε από το σύνολο δεδομένων εκπαίδευσης και ισούται με 84,4072%. Η χειρότερη απόδοση για το Precision στην κατηγορία 0 ισούται με 70,6%, ενώ η καλύτερη απόδοση είναι επίσης για το Precision αλλά στην κατηγορία 1 και ισούται με 100%. Ο πίνακας confusion matrix επιβεβαιώνει ότι η ακρίβεια για την κλάση 1 (μεταβλητή b) είναι 100%. Από την άλλη πλευρά, 121 περιπτώσεις ήταν προβληματικές και δεν ταξινομήθηκαν στην κλάση 0.



Σχεδιάγραμμα 12. Αποτελέσματα ταξινόμησης χρησιμοποιώντας ως κλάση τη μεταβλητή "Risk"

Τα αποτελέσματα δείχνουν ότι το χαρακτηριστικό που περιγράφει την ταξινόμηση είναι η μεταβλητή SCORE_A.
Αυτό σημαίνει ότι η μεταβλητή Risk σχετίζεται στενότερα με τη μεταβλητή SCORE_A σε σχέση με τις άλλες μεταβλητές.

4.5. Συσταδοποίηση (Clustering)

Στην επιβλεπόμενη μάθηση (Supervised Learning) μας δίνεται ένα σύνολο δεδομένων με τις αντίστοιχες κλάσεις-ετικέτες κάθε εγγραφής. Στόχος είναι η δημιουργία ενός μοντέλου, το οποίο να μπορεί να κατηγοριοποιήσει νέα δεδομένα σε κάποια από τις προϋπάρχουσες κλάσεις. Αντίθετα, στη μη επιβλεπομένη μάθηση (Unsupervised Learning) μας δίνεται ένα σύνολο δεδομένων, χωρίς τις αντίστοιχες κλάσεις-ετικέτες κάθε εγγραφής και στόχος είναι η χρήση κάποιου αλγορίθμου, ώστε αυτόματα να ανακαλύψουμε κάποια ενδεχομένως ενδιαφέρουσα δομή των δεδομένων. Για παράδειγμα, η συσταδοποίηση είναι μια από τις τεχνικές μη επιβλεπόμενης μάθησης. Δοθέντων κάποιων δεδομένων χωρίς κλάσεις, οι αλγόριθμοι συσταδοποίησης ομαδοποιούν τα δεδομένα σε συστάδες, έτσι ώστε εγγραφές, οι οποίες ανήκουν στην ίδια συστάδα, να έχουν όμοια ή παραπλήσια χαρακτηριστικά.
Σχετικά με την έννοια της συστάδας τώρα, στο πρόβλημα της συσταδοποίησης μας δίνεται ένα σύνολο δεδομένων, χωρίς τις αντίστοιχες κλάσεις ή ετικέτες και χρειαζόμαστε κάποιον αλγόριθμο, ο οποίος θα ομαδοποιήσει αυτόματα τα δεδομένα σε συστάδες. Οι συστάδες που δημιουργούνται θέλουμε να διαχωρίζουν ορθά τα δεδομένα. Αυτό πρακτικά σημαίνει ότι μια συστάδα θέλουμε να απαρτίζεται από αντικείμενα, όπου κάθε αντικείμενο είναι πιο κοντά σε κάθε άλλο αντικείμενο της ίδιας συστάδας απ’ ότι σε κάποιο άλλο αντικείμενο διαφορετικής συστάδας.
Η Ανάλυση Συστάδων (ΑΣ) (Clustering) είναι μια από τις βασικότερες εργασίες Εξόρυξης Δεδομένων. Σε γενικές γραμμές, η ΑΣ αφορά την ένταξη οντοτήτων σε ομοειδείς ομάδες. Η δραστηριότητα αυτή είναι εγγενής στους ανθρώπους, και εκτελείται αυθόρμητα από την παιδική τους ηλικία. Ένας άνθρωπος σε πρωτόγονες συνθήκες, αλλά με σχετική εμπειρία, κατανοεί αυθόρμητα ομάδες, όπως δένδρα, πουλιά κ.λπ. (Kruskal, 1977). Στην επιστημονική ΑΣ, οι ομάδες εξάγονται βάσει αλγορίθμων από τα δεδομένα.

Στόχος της Ανάλυσης Συστάδων είναι ο επιμερισμός ενός συνόλου παραδειγμάτων σε υποσύνολα. Τα υποσύνολα καλούνται συστάδες. Για τον επιμερισμό, καθοριστικό ρόλο παίζει η ομοιότητα. Τα παραδείγματα μιας συστάδας «μοιάζουν» μεταξύ τους, ενώ «δεν μοιάζουν» με τα παραδείγματα των άλλων συστάδων. Ένα σχετικά συγγενές αντικείμενο είναι η Κατηγοριοποίηση, η οποία στοχεύει στην πρόβλεψη της κατηγορίας κάθε παρατήρησης. Όμως οι διαφορές ανάμεσα στην ΑΣ και την Κατηγοριοποίηση είναι πολλές. Στην Κατηγοριοποίηση, οι κατηγορίες είναι γνωστές εκ των προτέρων. Στα δεδομένα υπάρχει ένα γνώρισμα, το γνώρισμα της κλάσης, στο οποίο καταγράφεται η κατηγορία της εκάστοτε παρατήρησης. Οι αλγόριθμοι μοντελοποιούν τις σχέσεις ανάμεσα στο γνώρισμα της κλάσης και στα υπόλοιπα γνωρίσματα. Η Κατηγοριοποίηση είναι μια μορφή εκπαίδευσης μέσω παραδειγμάτων (learning by examples). Το γεγονός ότι υπάρχει εκ των προτέρων γνώση σχετικά με τις κατηγορίες, και ότι η γνώση αυτή καθοδηγεί τη διαδικασία εκπαίδευσης, χαρακτηρίζει την Κατηγοριοποίηση ως επιβλεπόμενη μάθηση (supervised learning). Στην ΑΣ δεν υπάρχει κάποιο γνώρισμα στο οποίο να καταγράφεται η κλάση των παραδειγμάτων, και οι συστάδες δεν είναι γνωστές εκ των προτέρων. Αντιθέτως, το ζητούμενο είναι να εντοπιστούν συστάδες και να ενταχθούν τα παραδείγματα στην κατάλληλη συστάδα. Οι συστάδες συγκροτούνται στη βάση της ομοιότητας των μελών τους. Για τον λόγο αυτό, η ΑΣ θεωρείται μια μορφή εκπαίδευσης μέσω παρατήρησης (learning by observation). Επίσης, το γεγονός ότι δεν υπάρχει εκ των προτέρων γνώση χαρακτηρίζει την ΑΣ ως μη επιβλεπόμενη μάθηση (unsupervised learning). Ο κύριος σκοπός της Κατηγοριοποίησης είναι η διατύπωση προβλέψεων (predictive), ενώ ο κύριος σκοπός της ΑΣ είναι περιγραφικός (descriptive). Σε διαδικαστικό επίπεδο, η ΑΣ αντιμετωπίζει όλα τα γνωρίσματα ισότιμα και τα χρησιμοποιεί για τον υπολογισμό της ομοιότητας των παρατηρήσεων. Αντιθέτως, η Κατηγοριοποίηση χρησιμοποιεί τα υπόλοιπα γνωρίσματα για να προβλέψει τις τιμές του γνωρίσματος της κλάσης.

Στα πλαίσια της Εξόρυξης Δεδομένων, η ΑΣ έχει πολλαπλή χρησιμότητα (Kyrkos, 2015). Ως αυτόνομη αναλυτική εργασία, επιτρέπει στον αναλυτή να επιμερίσει τα δεδομένα σε ομάδες ομοειδών παρατηρήσεων. Ακολούθως, ο αναλυτής μπορεί να επικεντρωθεί στην εκάστοτε ομάδα, να αναγνωρίσει τα κοινά χαρακτηριστικά της, και να εξάγει γνώση χρήσιμη για τη λήψη αποφάσεων. Η πιο γνωστή εφαρμογή της ΑΣ στις επιχειρηματικές διαδικασίες είναι στη διαφήμιση, και ειδικότερα για την τμηματοποίηση της αγοράς. Ο όρος τμηματοποίηση της αγοράς περιγράφει τον επιμερισμό των καταναλωτών σε ομάδες με

όμοια καταναλωτική συμπεριφορά. Η τμηματοποίηση της αγοράς είναι κεφαλαιώδους σημασίας για το μάρκετινγκ. Οι διαφημίσεις μαζικής απεύθυνσης έχουν υψηλό κόστος και μικρό ποσοστό ανταπόκρισης. Με τον εντοπισμό ομάδων όμοιων καταναλωτών μπορούν να σχεδιαστούν διαφημιστικές εκστρατείες προσαρμοσμένες στα ιδιαίτερα χαρακτηριστικά της κάθε ομάδας. Η στοχευμένη σε συγκεκριμένες ομάδες διαφήμιση κοστίζει λιγότερο και επιτυγχάνει σημαντικά υψηλότερα ποσοστά ανταπόκρισης των καταναλωτών. Πέρα από την αξία της ως αυτόνομο εργαλείο ανάλυσης, η ΑΣ μπορεί να συνδυαστεί με άλλες εργασίες Εξόρυξης Δεδομένων και να αποτελέσει στάδιο προεπεξεργασίας. Χάρη στην ικανότητα των αλγορίθμων της να ομαδοποιούν τις παρατηρήσεις σύμφωνα με την ομοιότητα τους, μπορεί να χρησιμοποιηθεί για τον εντοπισμό παρατηρήσεων με ακραίες τιμές (outliers) (Ng και Han, 1994; Shekhar και Chawla, 2003). Οι ακραίες παρατηρήσεις θα απομακρυνθούν από το σύνολο δεδομένων, ώστε να προκύψει ένα βελτιωμένο σύνολο εκπαίδευσης. Επίσης, οι συστάδες, οι οποίες θα προκύψουν, μπορούν να θεωρηθούν κατηγορίες. Σε ακόλουθο στάδιο, μπορεί να εκτελεστεί κατηγοριοποίηση για την ανάπτυξη μοντέλων ικανών να προβλέπουν την κατηγορία. Συνδυασμός μεθόδων ΑΣ και Κατηγοριοποίησης μπορεί να αποφέρει υβριδικούς κατηγοριοποιητές.

Σε ότι αφορά τις κατηγορίες μεθόδων Ανάλυσης Συστάδων, η επιστημονική βιβλιογραφία περιλαμβάνει έναν μεγάλο αριθμό διαφορετικών μεθόδων. Οι μέθοδοι αυτές παρουσιάζουν σημαντικές διαφορές στις επαγωγικές αρχές τους και στον τρόπο σχηματισμού των συστάδων. Ένας από τους λόγους ύπαρξης αυτής της ποικιλίας μεθόδων είναι το γεγονός ότι δεν υπάρχει ένας αυστηρός ορισμός της έννοιας της συστάδας (Estivill και Yang, 2000). Οι Han *et al* (2011) ορίζουν πέντε κατηγορίες μεθόδων ΑΣ:

- Ιεραρχικές μέθοδοι. Οι ιεραρχικές μέθοδοι (hierarchical methods) δημιουργούν μια ιεραρχία από συστάδες. Στο κατώτατο επίπεδο της ιεραρχίας βρίσκονται τα μεμονωμένα αντικείμενα. Στο ανώτατο επίπεδο βρίσκεται μια υπερσυστάδα, η οποία περιλαμβάνει όλα τα αντικείμενα. Κάθε ενδιάμεσο επίπεδο ορίζει ένα σύνολο συστάδων. Η ιεραρχία προκύπτει από μια διαδικασία διαδοχικών διασπάσεων ή συγχωνεύσεων συστάδων. Η επιλογή του κατάλληλου συνόλου συστάδων εναπόκειται στον χρήστη.
- Διαχωριστικές μέθοδοι. Οι διαχωριστικές μέθοδοι (partitioning methods) επιμερίζουν τα αντικείμενα σε k συστάδες. Τυπικά το πλήθος των συστάδων προκαθορίζεται από τον χρήστη. Στις μεθόδους αυτής της κατηγορίας εφαρμόζεται μια επαναληπτική

διαδικασία, κατά την οποία τα αντικείμενα μετακινούνται από μια συστάδα σε μια άλλη. Η ποιότητα της κάθε λύσης ενδεχόμενων συστάδων μετράται με τη βοήθεια ενός κριτηρίου. Σε κάθε επανάληψη και με τη μετακίνηση των σημείων, η τιμή του κριτηρίου μειώνεται. *Ο πιο γνωστός αλγόριθμος διαχωριστικής ΑΣ είναι ο k-Means.*

- Μέθοδοι βασισμένες στην πυκνότητα. Στις βασισμένες στην πυκνότητα μεθόδους (density based methods) ελέγχεται η πυκνότητα των αντικειμένων στον χώρο και δημιουργούνται συστάδες, οι οποίες καλύπτουν τις πυκνές περιοχές. Για κάθε παρατήρηση που ανήκει σε μια συστάδα, η γειτονιά της, η οποία είναι καθορισμένης διαμέτρου, πρέπει να περιλαμβάνει έναν ελάχιστο αριθμό παρατηρήσεων. Η συστάδα συνεχίζει να επεκτείνεται όσο η γειτονιά των παρακείμενων σημείων διαθέτει την απαιτούμενη πυκνότητα. Οι μέθοδοι αυτές μπορούν να δημιουργήσουν συστάδες με μη κυρτά και περίπλοκα σχήματα. Επίσης, είναι ιδιαίτερα ικανές να απομονώνουν τις εξαιρέσεις.

- Μέθοδοι πλέγματος. Οι μέθοδοι πλέγματος (grid based methods) επιμερίζουν τον χώρο των δεδομένων σε διακριτά κελιά, τα οποία συγκροτούν ένα πλέγμα. Τα αντικείμενα πλέον αντιπροσωπεύονται από τα κελιά στα οποία ανήκουν. Η αναζήτηση των συστάδων γίνεται στα κελιά του πλέγματος και όχι στα αντικείμενα. Στις μεθόδους πλέγματος ο χρόνος επεξεργασίας εξαρτάται από το πλήθος των κελιών και όχι από το πλήθος των αντικειμένων. Επειδή κατά κανόνα ο αριθμός των κελιών είναι πολύ μικρότερος από τον αριθμό των αντικειμένων, οι μέθοδοι αυτές είναι σημαντικά ταχύτερες. Ένα σημαντικό ζήτημα είναι ο καθορισμός κελιών κατάλληλου μεγέθους.
- Μέθοδοι βασισμένες σε μοντέλα. Στις βασισμένες σε μοντέλα μεθόδους (model based methods), όπως υπονοεί το όνομα τους, γίνεται χρήση μοντέλων. Στόχος τους είναι να βελτιστοποιηθεί η προσαρμογή ανάμεσα στα δεδομένα και στα μοντέλα. Το μοντέλο εκπαιδεύεται με μη επιβλεπόμενη μάθηση σχετικά με τη συμμετοχή των παρατηρήσεων σε συστάδες. Μια πολύ διαδεδομένη μέθοδος αυτής της κατηγορίας είναι ένα ειδικός τύπος νευρωνικών δικτύων, που ονομάζονται Αυτοοργανούμενοι Χάρτες (Self Organizing Maps).

Αμέσως παρακάτω θα αναλυθεί η μέθοδος της Διαχωριστικής Ανάλυσης Συστάδων. Οι διαχωριστικές μέθοδοι θεωρούν ένα πλήθος N σημείων και ένα πλήθος k συστάδων, και διαμερίζουν τα σημεία στις συστάδες. Τυπικά, το πλήθος των συστάδων k προκαθορίζεται

από τον χρήστη. Ξεκινώντας από έναν αρχικό διαχωρισμό, με μια επαναληπτική διαδικασία, τα σημεία μετακινούνται από μια συστάδα σε μια άλλη. Ο σχηματισμός των συστάδων γίνεται με τρόπο τέτοιο, ώστε να βελτιστοποιείται ένα κριτήριο διαχωρισμού. Στόχος είναι να δημιουργηθούν συστάδες, οι οποίες να περιέχουν όμοια αντικείμενα, ενώ τα αντικείμενα διαφορετικών συστάδων να είναι ανόμοια.

Οι διαχωριστικές μέθοδοι παρουσιάζουν ευαισθησία στις αρχικές τους συνθήκες. Ένα σημαντικό πρόβλημα είναι το πλήθος των συστάδων k. Η εργασία του Dubes (1987) παρέχει καθοδήγηση για τον καθορισμό του πλήθους των συστάδων. Επίσης, για την εύρεση της καθολικά βέλτιστης λύσης θα έπρεπε να δοκιμαστούν όλοι οι δυνατοί διαχωρισμοί. Ωστόσο, λόγω υπολογιστικού κόστους, αυτό δεν είναι εφικτό. Στην πράξη εφαρμόζεται μια διαδικασία αρχικοποίησης του διαχωρισμού, και στη συνέχεια, μετακίνησης των σημείων.

Οι διαχωριστικές μέθοδοι δημιουργούν ένα σύνολο συστάδων, σε αντίθεση με τις ιεραρχικές μεθόδους, οι οποίες δημιουργούν μια ιεραρχική δομή διαδοχικών επιπέδων, όπου κάθε επίπεδο ορίζει ένα σύνολο συστάδων. Επίσης, είναι υπολογιστικά λιγότερο ακριβές από τις ιεραρχικές μεθόδους, και για τον λόγο αυτό μπορούν να εφαρμοστούν σε μεγαλύτερα σύνολα δεδομένων. *Η πιο γνωστή μέθοδος διαχωριστικής ανάλυσης συστάδων είναι ο αλγόριθμός k-Means.*

Η μέθοδος k-Means προτάθηκε από τον MacQueen (1967), και είναι η πιο γνωστή και διαδεδομένη διαιρετική μέθοδος ΑΣ. Στόχος της είναι να κατανείμει ένα σύνολο αντικειμένων σε έναν προκαθορισμένο αριθμό συστάδων, με τρόπο τέτοιο που να αυξάνει την ομοιότητα εντός των συστάδων. Ο αλγόριθμος περιλαμβάνει μια επαναληπτική διαδικασία, όπου σε κάθε επανάληψη υπολογίζεται το κέντρο της συστάδας (centroid). Τα αντικείμενα εντάσσονται στη συστάδα με το πλησιέστερο κέντρο.

Αναλυτικότερα, ο αλγόριθμος της μεθόδου k-Means έχει ως ακολούθως:

1. Αρχικά επιλέγονται τυχαία k αντικείμενα. Ο αριθμός k είναι το πλήθος των συστάδων που θα προκύψουν και προκαθορίζεται από τον χρήστη. Τα επιλεγμένα σημεία θεωρούνται κέντρα συστάδων.
2. Κάθε αντικείμενο κατατάσσεται στη συστάδα, της οποίας το κέντρο είναι πλησιέστερα του. Για τον υπολογισμό της απόστασης συνήθως χρησιμοποιείται η Ευκλείδεια απόσταση.

3. Τα κέντρα της κάθε συστάδας επαναϋπολογίζονται. Για κάθε διάσταση το κέντρο έχει τιμή ίση με τη μέση τιμή όλων των αντικειμένων, τα οποία ανήκουν στη συστάδα.

$$m_i = \frac{1}{M_i} \sum_{j=1}^{M_i} x_j$$

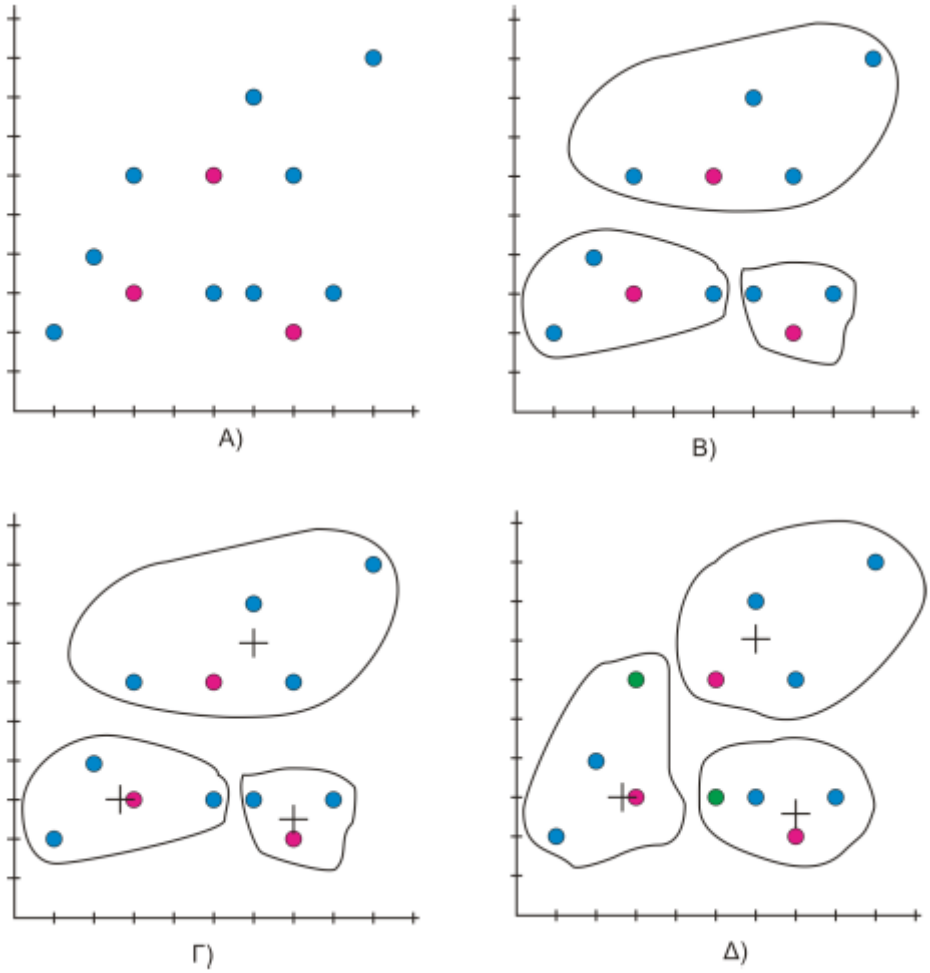
όπου M_i είναι το πλήθος των αντικειμένων της συστάδας i, και mi είναι το υπολογιζόμενο κέντρο.

4. Τα προηγούμενα δύο βήματα επαναλαμβάνονται μέχρι να ικανοποιηθεί η συνθήκη εξόδου. Τυπικά, συνθήκη εξόδου είναι η ελαχιστοποίηση του τετραγωνικού σφάλματος, το οποίο ορίζεται από την παρακάτω εξίσωση

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} (x - m_i)^2$$

όπου C_i είναι οι συστάδες, x είναι τα αντικείμενα και mi είναι το κέντρο της συστάδας C_i.

Στο Σχεδιάγραμμα 13 παρουσιάζεται ο σχηματισμός των συστάδων με τη μέθοδο k-Means. Στο τμήμα Α) παρουσιάζονται τα σημεία. Τα κόκκινα σημεία συμβολίζουν τα αρχικώς επιλεγμένα κέντρα. Στο τμήμα Β) σχηματίζονται οι συστάδες. Κάθε σημείο εντάσσεται στη συστάδα, στης οποίας το κέντρο βρίσκεται πλησιέστερα. Στο τμήμα Γ) υπολογίζονται τα νέα κέντρα των υφιστάμενων συστάδων. Τα νέα κέντρα συμβολίζονται με το σχήμα του σταυρού. Στο τμήμα Δ) επαναϋπολογίζεται η απόσταση των σημείων από τα νέα κέντρα, και τα σημεία επανεντάσσονται στις συστάδες. Τα δύο πράσινα σημεία αλλάζουν συστάδα.



Σχεδιάγραμμα 13. Δημιουργία συστάδων με k-Means

Ο αλγόριθμος k-Means διαθέτει τα παρακάτω **πλεονεκτήματα**:

- Είναι απλός και κατανοητός.
- Τα αντικείμενα μοιράζονται σε συστάδες με αυτόματο τρόπο.
- Είναι αρκετά γρήγορος, τουλάχιστον σε σχέση με τις ιεραρχικές μεθόδους. Ο χρόνος εκτέλεσης του αλγορίθμου εξαρτάται γραμμικά από τα στοιχεία του προβλήματος, όπως το πλήθος των συστάδων k, το πλήθος των αντικειμένων n και το πλήθος των επαναλήψεων l. Η υπολογιστική πολυπλοκότητα του αλγορίθμου είναι O(nkl). Για τον λόγο αυτό, είναι πιο κατάλληλος από άλλες μεθόδους για την ομαδοποίηση μεγάλων συνόλων αντικειμένων.

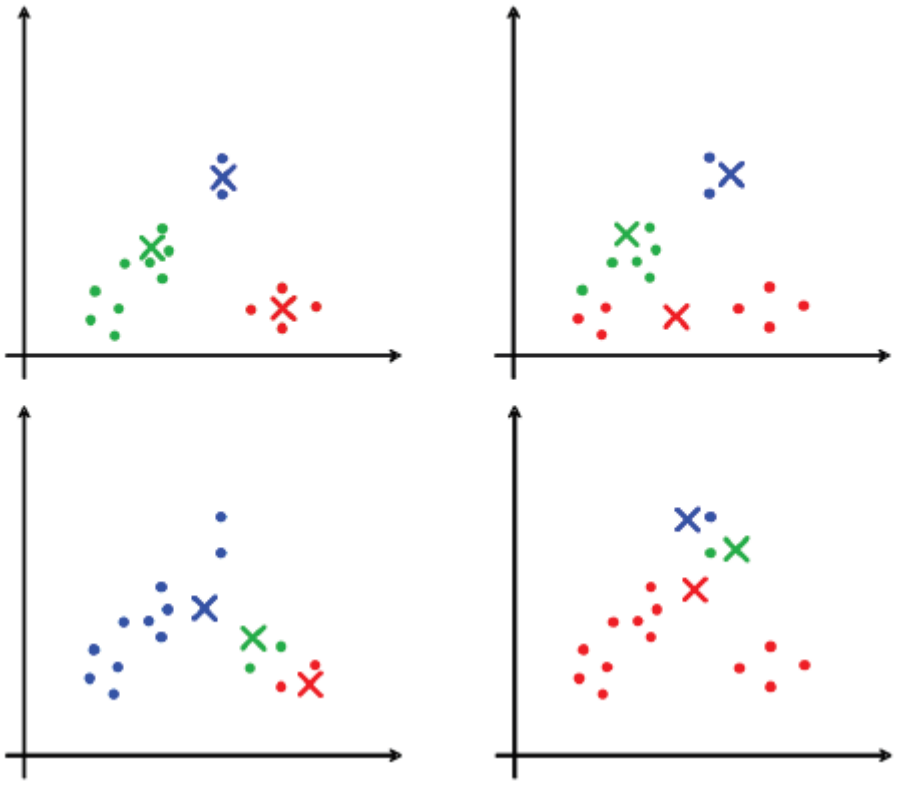
Τα βασικά **μειονεκτήματα** του k-Means είναι τα ακόλουθα:

- Ο αριθμός των συστάδων πρέπει να προκαθοριστεί από τον χρήστη.

- Το τελικό αποτέλεσμα εξαρτάται σε σημαντικό βαθμό από την επιλογή των αρχικών κέντρων. Επιλογή διαφορετικών κέντρων μπορεί να οδηγήσει σε σημαντικά διαφορετικές συστάδες.
- Είναι πολύ ευαίσθητος στην ύπαρξη αντικειμένων με ακραίες τιμές (outliers). Λίγα αντικείμενα με πολύ μεγάλες τιμές μπορούν να επηρεάσουν σημαντικά τον υπολογισμό των νέων κέντρων και κατά συνέπεια τη διαμόρφωση των τελικών συστάδων.
- Έχει την τάση να δημιουργεί σφαιρικές και ίσου μεγέθους συστάδες. Για τον λόγο αυτό, δεν είναι κατάλληλος για συστάδες με περίπλοκα σχήματα ή με πολύ διαφορετικά μεγέθη. Για την αντιμετώπιση των προβλημάτων του k-Means έχουν προταθεί διάφορες λύσεις. Ένα βασικό πρόβλημα είναι ο προκαθορισμός του αριθμού των συστάδων. Μια δυνατή λύση σε αυτό το πρόβλημα είναι να εφαρμοστεί αρχικά ιεραρχική ΑΣ. Η ιεραρχική ΑΣ συνίσταται σε μια διαδικασία διαδοχικών συνενώσεων ή διασπάσεων των συστάδων. Με τον τρόπο αυτόν, ο χρήστης μπορεί να εκτιμήσει το πλήθος των συστάδων, και στη συνέχεια να εκτελέσει τον k-Means. Ένα άλλο σημαντικό πρόβλημα είναι ότι ο αλγόριθμος μπορεί να συγκλίνει σε τοπικά βέλτιστα, και δεν υπάρχει εγγύηση για την εύρεση ενός καθολικού βέλτιστου. Το τελικό αποτέλεσμα επηρεάζεται σημαντικά από την επιλογή των αρχικών κέντρων. Ένας τρόπος αντιμετώπισης αυτού του προβλήματος είναι οι διαδοχικές, πολλαπλές εκτελέσεις του αλγορίθμου, με διαφορετικά αρχικά κέντρα κάθε φορά. Πρόσθετες τεχνικές επιδιώκουν τη σύγκλιση σε καθολικό βέλτιστο. Οι Likas *et al* (2003) εφαρμόζουν μια αιτιοκρατική διαδικασία καθολικής αναζήτησης. Στη διαδικασία αυτή εκτελούνται πολλαπλές τοπικές αναζητήσεις με τον k-Means για διαρκώς αυξανόμενο πλήθος συστάδων, μέχρι το τελικό επιθυμητό πλήθος συστάδων M.

Εν’ συνεχεία, αξίζει περαιτέρω ανάλυσης η Τυχαία Αρχικοποίηση Κεντροειδών. Όπως τονίστηκε, το πρώτο βήμα του αλγορίθμου k-means είναι η τυχαία αρχικοποίηση των k κεντροειδών των συστάδων. Παρόλο που το συγκεκριμένο βήμα φαίνεται απλό και ασήμαντο, αρκετές φορές μια «κακή» αρχικοποίηση μπορεί να οδηγήσει σε κακής ποιότητας συστάδες στην πορεία. Στο Σχεδιάγραμμα 14 βλέπουμε ένα παράδειγμα τεσσάρων τυχαίων αρχικοποιήσεων των κεντροειδών, ενώ με χρώμα υποδεικνύεται το πώς τελικά καταλήγουν να είναι οι συστάδες που δημιουργεί ο αλγόριθμος. Πάνω αριστερά έχουμε την καλύτερη περίπτωση. Ακολουθεί μια λιγότερο ποιοτικά καλή συσταδοποίηση πάνω δεξιά. Στις δυο τελευταίες περιπτώσεις είναι προφανές ότι η αρχικοποίηση επηρεάζει αρνητικά τη

διαδικασία συσταδοποίησης. Σε αυτές τις δυο περιπτώσεις, οι δύο συστάδες περιέχουν πολύ λίγα δείγματα, ενώ η μία περιέχει όλα τα υπόλοιπα δείγματα.

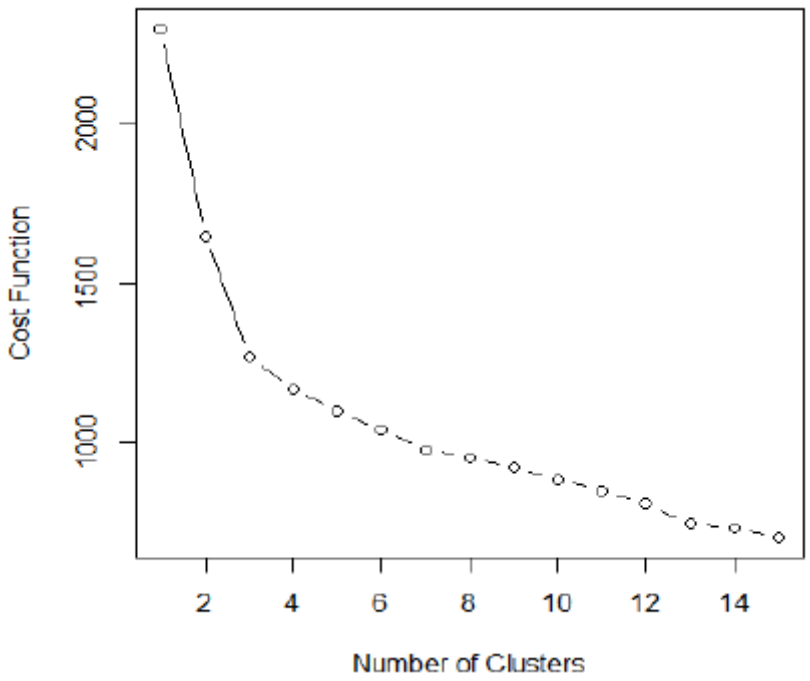


Σχεδιάγραμμα 14. Τυχαία αρχικοποίηση κεντροειδών

Όπως αναφέρθηκε και προηγουμένως, ένα από τα μειονεκτήματα του αλγορίθμου k-means είναι το γεγονός ότι δεν υπάρχει κάποιος αυτοματοποιημένος τρόπος επιλογής του k, δηλαδή του αριθμού των συστάδων. Ο αριθμός των συστάδων δίνεται ως είσοδος από τον χρήστη και η επιλογή του σωστού αριθμού επαφίεται στη δική του γνώση και εμπειρία. Να υπενθυμίσουμε ότι κατά τη συσταδοποίηση δεν δίνεται το επιπλέον χαρακτηριστικό κλάσης των δειγμάτων. Συνεπώς, η διαδικασία επιλογής του αριθμού συστάδων, ενδεχομένως, να απαιτήσει την εξερεύνηση και μελέτη των δεδομένων, για παράδειγμα, μέσα από οπτικοποιήσεις, προκειμένου να καταλήξουμε στον σωστό αριθμό συστάδων.

Δυστυχώς, για την επιλογή του αριθμού των συστάδων δεν υπάρχει κάποιος γενικός κανόνας, ο οποίος να λειτουργεί εγγυημένα και για όλες τις περιπτώσεις. Ένα απλό και πρακτικό τέχνασμα, το οποίο μπορεί να βοηθήσει σε ορισμένες περιπτώσεις, είναι «ο κανόνας του αγκώνα» (the elbow rule). Στο Σχεδιάγραμμα 15 ο κανόνας του αγκώνα

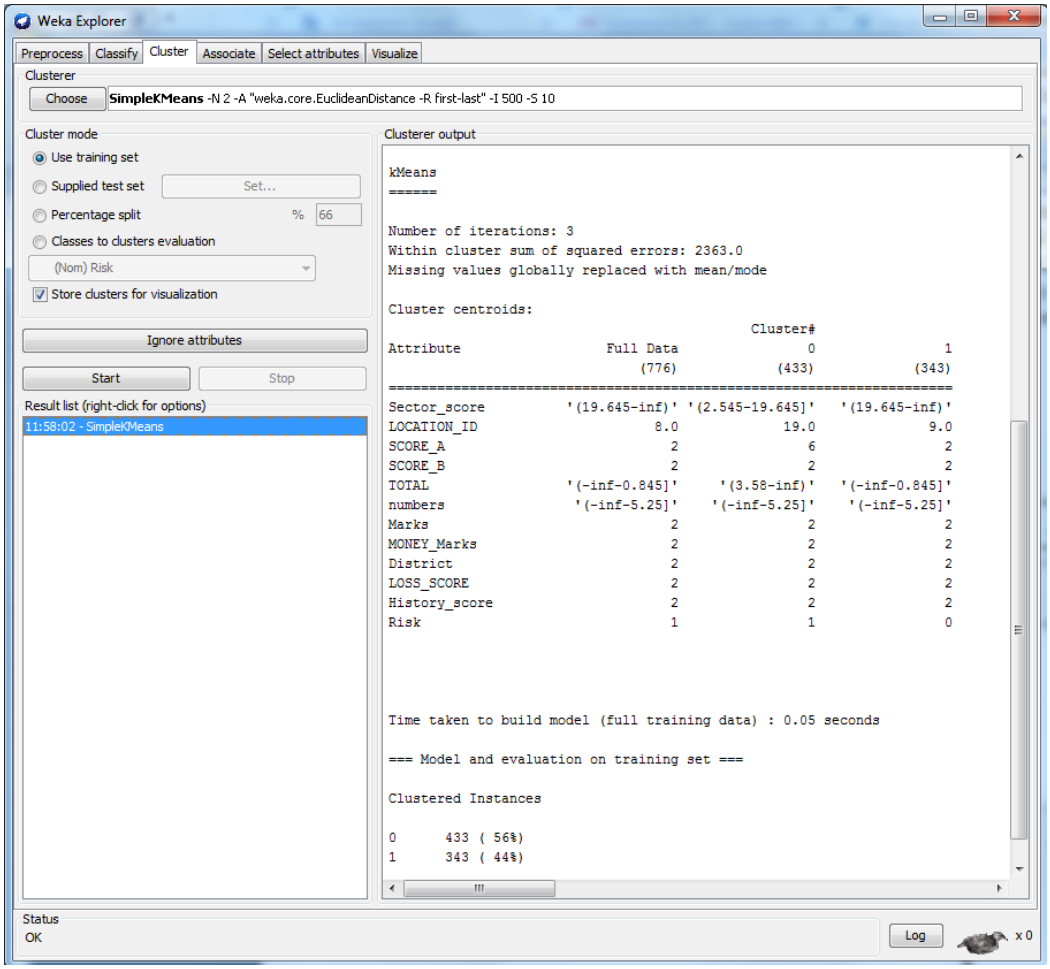
υποδεικνύει ότι η επιλογή k=3 είναι αρκετά καλή. Ωστόσο, υπάρχουν περιπτώσεις, όπου η γραφική είναι πιο ομαλή και δεν έχει τον τύπο σχήματος του αγκώνα, με αποτέλεσμα η επιλογή και πάλι να μην είναι ξεκάθαρη.



Σχεδιάγραμμα 15. Ο κανόνας του αγκώνα

Λαμβάνοντας υπόψιν το παραπάνω θεωρητικό πλαίσιο, για το βήμα της συσταδοποίησης επιλέχθηκε να χρησιμοποιηθεί ως καταλληλότερος ο αλγόριθμος k-means (SimpleKmeans στο πλαίσιο του WEKA) λόγω του «χειροκίνητου» τρόπου σχηματισμού των συστάδων.

Ο αριθμός των συστάδων ορίζεται σε 2, δεδομένου ότι η μεταβλητή Risk χρησιμοποιήθηκε για τον υπολογισμό της ακρίβειας της ομαδοποίησης και την επιθεώρηση των δεδομένων ελέγχου. Το Σχεδιάγραμμα 16 δείχνει τα αποτελέσματα της ομαδοποίησης με βάση τη μεταβλητή Risk. Οι ομαδοποιημένες εμφανίσεις είναι 433 (56%) και 343 (44%) αντίστοιχα. Είναι επίσης προφανές ότι από το κέντρο των συστάδων η μεταβλητή Risk έχει τιμή 0 στην πρώτη συστάδα και τιμή 1 στη δεύτερη συστάδα.



Σχεδιάγραμμα 16. Αποτελέσματα συσταδοποίησης. Η μεταβλητή "Risk" χρησιμοποιείται για την αξιολόγησή της

Οι διαφορές μεταξύ των δύο συστάδων εστιάζονται στα χαρακτηριστικά: Sector_score, LOCATION_ID, SCORE_A, TOTAL and Risk.

4.6. Κανόνες συσχέτισης (Association rule mining)

Η ανακάλυψη Κανόνων Συσχέτισης είναι μια από τις βασικότερες εργασίες Εξόρυξης Δεδομένων. Από πολλούς μάλιστα θεωρείται ως το πιο γνήσιο τέκνο της Εξόρυξης Δεδομένων, καθώς άλλες εργασίες εξόρυξης, μεθοδολογίες και τεχνικές προέρχονται κυρίως από τη Μηχανική Μάθηση, τη Στατιστική, τις Βάσεις Δεδομένων κλπ. Οι Κανόνες Συσχέτισης αφορούν την ανακάλυψη και διατύπωση σχέσεων που υπάρχουν στα δεδομένα.

Οι σχέσεις αυτές προκύπτουν από τη συχνή ταυτόχρονη εμφάνιση τιμών δεδομένων (Kyrkos, 2015).

Η εύρεση συχνών στοιχειοσυνόλων είναι ένα ενδιαφέρον πρόβλημα. Ο πρώτος σχετικός αλγόριθμος που προτάθηκε ονομάζεται *Apriori* (Agrawal *et al.*, 1993). Το όνομα του οφείλεται στο γεγονός ότι χρησιμοποιεί προηγούμενη (prior) γνώση σχετικά με τη συχνότητα k-Στοιχειοσυνόλων, για να βρει συχνά (k+1) Στοιχειοσύνολα. Η προηγούμενη γνώση που χρησιμοποιείται αφορά την αντιμονότονη ιδιότητα της υποστήριξης. Η ιδιότητα αυτή ορίζει ότι η υποστήριξη ενός στοιχειοσυνόλου είναι ίση ή μικρότερη από την υποστήριξη κάθε δυνατού υποσυνόλου του. Η αντιμονότονη ιδιότητα της υποστήριξης μαθηματικά ορίζεται με τη Σχέση:

$$\forall X, Y: (X \subseteq Y) \Longrightarrow s\text{upp}(X) \geq s\text{upp}\,(Y)$$

Από τη σχέση αυτή προκύπτει ότι για να είναι ένα στοιχειοσύνολο συχνό πρέπει όλα τα μη κενά υποσύνολα του να είναι επίσης συχνά. Αντιστρόφως, εάν ένα στοιχειοσύνολο είναι μη συχνό, τότε η πρόσθεση σε αυτό ενός νέου στοιχείου δεν μπορεί να δημιουργήσει ένα νέο συχνό στοιχειοσύνολο. Τα υπερσύνολα ενός μη συχνού στοιχειοσυνόλου είναι μη συχνά.

Η υποστήριξη και η εμπιστοσύνη είναι δύο ισχυρά μέτρα της ισχύος ενός κανόνα. Η υποστήριξη εξασφαλίζει ότι ο κανόνας αφορά ένα ικανοποιητικό ποσοστό των συναλλαγών. Κανόνες με μικρή υποστήριξη μπορεί να θεωρηθεί ότι εκφράζουν ένα τυχαίο γεγονός. Επίσης, η εμπιστοσύνη αποτελεί μέτρο του κατά πόσο η εμφάνιση του αριστερού μέρους πριμοδοτεί την εμφάνιση του δεξιού μέρους του κανόνα. Η υποστήριξη και η εμπιστοσύνη χρησιμοποιούνται για την εύρεση των κανόνων. Ωστόσο, πρέπει να σημειωθεί ότι υψηλά ποσοστά υποστήριξης και εμπιστοσύνης δεν εξασφαλίζουν ότι ο κανόνας αναδεικνύει μια πραγματική σχέση. Βέβαια, έχουν προταθεί και άλλα μέτρα για την αξιολόγηση των Κανόνων Συσχέτισης (Kyrkos, 2015).

Η ανακάλυψη Κανόνων Συσχέτισης δεν είναι ένα εύκολο καθήκον εξαιτίας του όγκου των δεδομένων. Η απλούστερη εκδοχή εύρεσης κανόνων είναι να δημιουργηθούν όλοι οι δυνατοί κανόνες, στη συνέχεια να υπολογιστεί η υποστήριξη και η εμπιστοσύνη για κάθε έναν από αυτούς και τέλος να διατηρηθούν μόνο όσοι κανόνες έχουν για αυτά τα δύο μέτρα

τιμές μεγαλύτερες ή ίσες από τις καθορισμένες τιμές κατωφλίου. Ένας τέτοιος τρόπος επίλυσης του προβλήματος είναι πρακτικά αδύνατος. Για k στοιχεία το πλήθος των δυνατών στοιχειοσυνόλων m δίνεται από τη σχέση $m=2^k -1$. Αυτό σημαίνει ότι για 20 μόλις στοιχεία, το πλήθος των δυνατών στοιχειοσυνόλων είναι περίπου 1.000.000. Σε ένα πραγματικό πρόβλημα, το k μπορεί να ανέρχεται σε εκατοντάδες ή και χιλιάδες στοιχεία. Σε τέτοιες περιπτώσεις ο έλεγχος όλων των δυνατών στοιχειοσυνόλων είναι απλώς αδύνατος. Έχουν προταθεί πιο αποτελεσματικές μέθοδοι για την ανακάλυψη Κανόνων Συσχέτισης, των οποίων η υποστήριξη και η εμπιστοσύνη υπερβαίνουν ένα προκαθορισμένο κατώφλι. Η διαδικασία ανακάλυψης Κανόνων Συσχέτισης ολοκληρώνεται σε δύο στάδια (Kyrkos, 2015):

1. Στο πρώτο στάδιο εντοπίζονται τα συχνά στοιχειοσύνολα. Τα στοιχειοσύνολα αυτά έχουν υποστήριξη μεγαλύτερη ή ίση από την τιμή κατωφλιού.
2. Στο δεύτερο στάδιο δημιουργούνται οι κανόνες σχέσης από τα συχνά στοιχειοσύνολα. Οι κανόνες ικανοποιούν τη συνθήκη της εμπιστοσύνης

Η συγκεκριμένη εφαρμογή στο WEKA αποτελεί την πιο σημαντική τεχνική εξόρυξης δεδομένων στα πλαίσια της παρούσας εργασίας καθώς φανερώνει τον τρόπο με τον οποίο τα ελεγκτικά δεδομένα συνδυάζονται μεταξύ τους. Για παράδειγμα μπορεί να φανερώσει, για επιχειρήσεις με υψηλό χαρακτηριστικό Risk, με ποιες άλλες μεταβλητές συνδέεται αυτή η αυξημένη τιμή.

Ο αλγόριθμος Apriori (Agrawal *et al.*, 1993) χρησιμοποιήθηκε για την εύρεση κανόνων συσχέτισης για το σύνολο των δεδομένων. Το WEKA δημιούργησε μια λίστα με 15 κανόνες (Πίνακας 4) με την υποστήριξη του προηγούμενου και του επακόλουθου (συνολικός αριθμός των αντικειμένων) στο 0,1 ελάχιστο, και την εμπιστοσύνη του κανόνα στο 0,9 ελάχιστο (ποσοστό των αντικειμένων σε μια κλίμακα από 0 έως 1). Η εφαρμογή του αλγορίθμου Apriori για τη συσχέτιση παρείχε πολύ χρήσιμες πληροφορίες για τα δεδομένα ελέγχου. Ο Πίνακας 4 δείχνει πώς μπορεί να ανακαλυφθεί ένας μεγάλος αριθμός κανόνων σύνδεσης.

Best rules found:
1. Marks=2 706 ==> numbers='(-inf-5.25]' 706 conf:(1)
2. numbers='(-inf-5.25]' 706 ==> Marks=2 706 conf:(1)
3. Marks=2 LOSS_SCORE=2 688 ==> numbers='(-inf-5.25]' 688 conf:(1)
4. numbers='(-inf-5.25]' LOSS_SCORE=2 688 ==> Marks=2 688 conf:(1)
5. Marks=2 History_score=2 673 ==> numbers='(-inf-5.25]' 673 conf:(1)
6. numbers='(-inf-5.25]' History_score=2 673 ==> Marks=2 673 conf:(1)
7. History_score=2 726 ==> LOSS_SCORE=2 710 conf:(0.98)
8. numbers='(-inf-5.25]' 706 ==> LOSS_SCORE=2 688 conf:(0.97)
9. Marks=2 706 ==> LOSS_SCORE=2 688 conf:(0.97)
10. numbers='(-inf-5.25]' Marks=2 706 ==> LOSS_SCORE=2 688 conf:(0.97)
11. Marks=2 706 ==> numbers='(-inf-5.25]' LOSS_SCORE=2 688 conf:(0.97)
12. numbers='(-inf-5.25]' 706 ==> Marks=2 LOSS_SCORE=2 688 conf:(0.97)
13. numbers='(-inf-5.25]' 706 ==> History_score=2 673 conf:(0.95)
14. Marks=2 706 ==> History_score=2 673 conf:(0.95)
15. numbers='(-inf-5.25]' Marks=2 706 ==> History_score=2 673 conf:(0.95)

Πίνακας 4. Οι καλύτεροι κανόνες που βρέθηκαν με τον αλγόριθμο Apriori βάσει της μέτρησης εμπιστοσύνης

Υπάρχουν μερικοί μη ενδιαφέροντες κανόνες αναφορικά με τον στόχο της έρευνας, όπως οι παρόμοιοι κανόνες 1 και 2 που δείχνουν τις αναμενόμενες ή συμμορφούμενες συσχετίσεις. Εάν Marks = 2 τότε η numbers είναι μεταξύ 0 και 5,25 και αντίστροφα. Παρατηρούνται επίσης συμμετρικοί κανόνες, καθώς το προηγούμενο στοιχείο και το συνακόλουθο στοιχείο εναλλάσσονται μεταξύ τους.

Υπάρχουν ορισμένοι παρόμοιοι κανόνες, κανόνες με το ίδιο στοιχείο προηγούμενοι και συνακόλουθοι αλλά εναλλάξιμοι (3 και 4, και 5 και 6). Οι μεταβλητές Marks και numbers εμφανίζονται σε προηγούμενα και συνακόλουθα στοιχεία, αλλά είναι εναλλάξιμα. Υπάρχει επίσης μια συμμετρική τριάδα κανόνων (10, 11 και 12) όπου τα Marks και numbers εμφανίζονται επίσης σε προηγούμενα και συνακόλουθα στοιχεία και εναλλάσσονται. Παρατηρούμε ότι υπάρχει και ένας μη ενδιαφέρων ή περιττός κανόνας (κανόνας με γενίκευση των σχέσεων άλλων κανόνων) όπως ο κανόνας 15 με τους κανόνες 13 και 14.

Υπάρχουν όμως και ενδιαφέροντες κανόνες όπως οι 7, 8 και 9 που προσφέρουν δυνατότητα δράσης για έναν ελεγκτή. Αυτοί οι τρεις κανόνες είναι χρήσιμοι για έναν ελεγκτή, καθώς μπορεί να δώσει μεγαλύτερη προσοχή στις εταιρείες με το History_score = 2, numbers μεταξύ 0 και 5,25 και Marks = 2.

Συνοψίζοντας τα αποτελέσματα από τις μεθόδους ταξινόμησης, συσταδοποίησης και κανόνων συσχέτισης, μπορούν να συναχθούν τα συμπεράσματα ότι:

1. Το χαρακτηριστικό που περιγράφει καλύτερα την ταξινόμηση είναι η μεταβλητή SCORE_A. Το χαρακτηριστικό "Risk" (Απάτη / Μη απάτη) χρησιμοποιείται ως κλάση.
2. Χρησιμοποιώντας το "Risk" ως χαρακτηριστικό κλάσης στην συσταδοποίηση, τα αποτελέσματα δείχνουν ότι οι εταιρείες που ανήκουν στο δεύτερο σύμπλεγμα έχουν καλύτερες τιμές στις παραμέτρους σχετικά με τον Κίνδυνο "Risk".
3. Για εταιρείες με History_score = 2, numbers μεταξύ 0 και 5,25 και Marks = 2, ο ελεγκτής πρέπει να δώσει μεγαλύτερη προσοχή.

ΚΕΦΑΛΑΙΟ 5: ΣΥΜΠΕΡΑΣΜΑΤΑ

Σκοπός της παρούσας εργασίας είναι η παρουσίαση και ανάλυση ενός ερευνητικού πλαισίου κατάλληλο για στελέχη ελέγχου, λογιστικής, χρηματοοικονομικής και διαχείρισης κινδύνων. Πιο συγκεκριμένα, το ανεπτυγμένο αυτό πλαίσιο είναι σε θέση να επιλύσει τα προβλήματα που παρουσιάζονται κατά τη διαδικασία ελέγχου επιχειρήσεων ύποπτων για οικονομική απάτη χρησιμοποιώντας τεχνικές εξόρυξης δεδομένων. Τα δεδομένα ελέγχου που χρησιμοποιήθηκαν για τη δοκιμή του πλαισίου αυτού προέρχονται από έναν αξιόπιστο ελεγκτικό οργανισμό αποθηκευμένο σε ένα πολύ γνωστό αποθετήριο δεδομένων ενώ το πακέτο λογισμικού που χρησιμοποιήθηκε ήταν το WEKA. Ο στόχος της πιλοτικής εφαρμογή των δεδομένων ελέγχου είναι να εξαχθεί ένα προτεινόμενο πλαίσιο υποστήριξης αποφάσεων ικανό να βοηθήσει έναν ελεγκτή να αποφασίσει σχετικά με το μέγεθος της εργασίας που απαιτείται για μια συγκεκριμένη εταιρεία ή οργανισμό, ή ακόμη και να παραλείψει να επισκεφθεί εταιρείες χαμηλού κινδύνου. Η πρόβλεψη απάτης σε μια εταιρεία είναι ένα σημαντικό βήμα από το προκαταρκτικό κιόλας στάδιο του ελέγχου καθώς μπορούν να στοχευθούν εξαρχής οι εταιρείες υψηλού κινδύνου και σε αυτές να μεγιστοποιηθεί η έρευνα του ελέγχου.

Δεδομένου ότι, ο εμπλουτισμός και η βελτίωση του ελέγχου είναι μια αναγνωρισμένη πρόκληση μεταξύ ερευνητών και επαγγελματιών και τα παραδοσιακά εργαλεία και οι τεχνικές ελέγχου παραμελούν το δυναμικό της ανάλυσης δεδομένων, η ανάπτυξη ενός κατάλληλου πλαισίου ελέγχου που βασίζεται σε εργαλεία και τεχνικές εξόρυξης δεδομένων καθίσταται επιτακτική ανάγκη. Αναλύθηκαν εδραιωμένα δεδομένα ελέγχου τα οποία οδήγησαν σε μια πρόταση εννοιολογικής αρχιτεκτονικής για μια ολοκληρωμένη προσέγγιση ελέγχου. Αξίζει να τονιστεί ότι το προτεινόμενο πλαίσιο είναι ανεξάρτητο από το συγκεκριμένο σύνολο δεδομένων και μπορεί να εφαρμοστεί σε άλλα παρόμοια σύνολα δεδομένων χρησιμοποιώντας απλώς τις ίδιες τεχνικές εξόρυξης δεδομένων. Τα αποτελέσματα που εξάχθηκαν υποστηρίζουν τη διαδικασία λήψης αποφάσεων σχετικά με τις εταιρείες που πρόκειται να ελεγχθούν. Βάσει των ευρημάτων της εργασίας η εκπαίδευση και η δοκιμή του παρόντος μοντέλου εντοπισμού και διαχείρισης κινδύνου μπορεί να συμβάλλει στην κάλυψη ενός υφιστάμενου ερευνητικού χάσματος. Με τον αυξανόμενο αριθμό περιπτώσεων οικονομικής απάτης, η εφαρμογή τεχνικών εξόρυξης δεδομένων θα

μπορούσε να διαδραματίσει σημαντικό ρόλο στη βελτίωση της ποιότητας της διενέργειας ελέγχου στο μέλλον.

Το ερώτημα εάν το προτεινόμενο πλαίσιο μπορεί να εφαρμοστεί και σε άλλες χρηματοοικονομικές και διοικητικές εφαρμογές μπορεί να απαντηθεί ικανοποιητικά μόνο όταν θα δοκιμαστεί και σε αυτές οπότε και αφήνεται σαν πιθανή μελλοντική εργασία. Επίσης, με το παρόν ερευνητικό πλαίσιο τίθενται και κάποιοι περιορισμοί καθώς η προτεινόμενη μέθοδος απαιτεί χρήστες με συγκεκριμένες δυνατότητες και γνώσεις. Αυτό σημαίνει ότι θα πρέπει να γνωρίζουν να χρησιμοποιούν σε βάθος τις τεχνικές ελέγχου και εξόρυξης δεδομένων. Προς την κατεύθυνση αυτή μια ακόμη δυνατή προσθήκη θα μπορούσε να συμπεριλάβει την υλοποίηση ενός πιο απλοποιημένου περιβάλλοντος χρήστη (GUI) χωρίς την εμπλοκή εξειδικευμένων προγραμμάτων όπως το Weka.

ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΑΝΑΦΟΡΕΣ

Adamyk, O., Adamyk, B. and Khorunzhak, N. (2018), "Auditing of the Software of Computer Accounting System", *ICTERI Workshops*, pp. 251-262.

Alles, M., Brennan, G., Kogan, A. and Vasarhelyi, M. A. (2018), "Continuous monitoring of business process controls: A pilot implementation of a continuous auditing system at Siemens", *Continuous Auditing: Theory and Application*, Emerald Publishing Limited Bingley, West Yorkshire, England, pp. 219-246.

Amani, F.A. and Fadlalla, A.M. (2017), "Data mining applications in accounting: A review of the literature and organizing framework", *International Journal of Accounting Information Systems*, **24**, pp. 32-58.

Antipova, T., Rocha, Á., (2018), *Information Technology Science*, Springer, Cham.

Appelbaum, D. (2017), "Introduction to Data Analysis for Auditors and Accountants", *The CPA Journal*, **7**.

Agrawal, R., Imieliński, T., Swami, A., (1993), "Mining association rules between sets of items in large databases", *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pp. 207-216.

Agrawal, R., Srikant, R., (1994), "Fast algorithms for mining association rules", *Proc. 20th int. conf. very large data bases*, VLDB, Vol. 1215, pp. 487-499.

Appelbaum, D., Kogan, A., Vasarhelyi, MA., (2016), "Analytics in External Auditing: A Literature Review", *Rutgers University CARLab Newark, NJ*, USA Working Paper.

Bellino, C., Wells, J., Hunt, S., (2007), "Global Technology Audit Guide (GTAG)", *Auditing Application Controls*.

Boslaugh, S. (2007), "Secondary analysis for public health: A practical guide", New York, NY: Cambridge

Boyd, D. and Crawford, K. (2012), "Critical Questions for Big Data: Provocations for a Cultural", *Technological and Scholarly Phenomenon, Information, Communication, & Society*, **15**, pp. 662-679.

Cao, M. Chychyla, R. and Stewart, T. (2015), "Big Data analytics in financial statement audits", *Accounting Horizons,* **29(2)**, pp. 423-429.

Cosserat GW., Rodda, N., (2004), "Modern auditing", *John Wiley & Sons. Hoboken*, New Jersey.

Cosserat, G. (2009), "Accepting the engagement and planning the audit", *Modern auditing, ed. G. Cosserat and N. Rodda, 3rd ed.*, John Wiley & Sons, pp. 734–36.

Dubes, R. C. (1987), "How many Clusters are Best: An Experiment. Pattern Recognition", **20(6)**, pp. 645-663.

Dull, R. B. Tegarden, D. P. and Schleifer, L. L. F. (2006), "ACTVE: A Proposal for an Automated Continuous Transaction Verification Environment", *Journal of Emerging Technologies in Accounting, Vol. 3*, pp. 81-96.

Estivill-Castro, V. and Yang, J.A. (2000), "Fast and Robust General Purpose Clustering Algorithm", *Proceedings of the Pacific Rim International Conference on Artificial Intelligence*, pp. 208-218.

Fanning, K. and Cogger, K. (1998), "Neural network detection of management fraud using published financial data", *International Journal of Intelligent Systems in Accounting, Finance & Management*, **7(1)**, pp. 21–41.

Fayyad, U.M. Piatetsky-Shapiro, G. and Smyth, P. (1996), "From Data Mining to Knowledge Discovery: An Overview in Advances in Knowledge Discovery and Data Mining", *AAAI Press*, pp. 1-34.

Fischer, E. and Parmentier, M.-A. (2010), "Doing qualitative research with archival data: Making secondary data a primary resource", *North American Conference Proceedings*

Frank, E. and Witten, I. H. (1999), "Making Better Use of global Discretization", *Proceedings of the 16th International Conference on Machine Learning*, San Francisco, CA: Morgan Kaufmann, pp. 115-123.

Fukunaga, K. (1990), "Introduction to Statistical Pattern Recognition", *Boston: Academic Press*

Gallegos, FS., Manson, DP., Gonzales, C., (2004), "Information Technology Control and Audit,", *Auerbach Publications.*

Gangolly, JS. (2016), "Audit Analytics and Continuous Audit: Looking towards the Future", *Journal of Emerging Technologies in Accounting, American Institute of Certified Public Accountants, INC.*, **13(1)**, pp. 187-188.

Gelinas, UJ., Dull, RB., Wheeler, P., (2011), "Accounting information systems", *Cengage learning.*

Ghasemi, M., Shafeiepour, V., Aslani, M., Barvayeh, E., (2011), "The impact of Information Technology (IT) on modern accounting systems", *Procedia-Social and Behavioral Sciences*, **28**, pp. 112-116.

Gladstone, B.M.., Volpe, T. and Boydell, K.M. (2007), "Issues encountered in a qualitative secondary analysis of help seeking in the prodrome to psychosis", *The Journal of Behavioural Health Sciences and Research*, **34 (4)**, pp. 431- 442.

Global Technology Audit Guide - GTAG ® (2015), Coordinating Continuous Auditing and Monitoring to Provide Continuous Assurance, Global Technology Audit Guide 3, 2nd Edition, pp. 11-15.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, IH., (2009), "The WEKA data mining software: an update", *ACM SIGKDD explorations newsletter,* **11(1)**, pp. 10-18.

Han, J., Pei, J., Kamber, M., (2001), "Data mining: concepts and techniques", *Morgan Kaufmann*, Burlington, Massachusetts.

Henry, E., Robinson, T., R., (2009), "Financial Statement Analysis: An Introduction", *International Financial Statement Analysis*, John Wiley & Sons, Inc., Hoboken, New Jersey.

Heaton, J. (1998), "Secondary analysis of qualitative data", *Social Research Update*, **22**, University of Surrey.

Hinds, P.S. Vogel, R.J. and Clarke-Steffen, L. (1997), "The possibilities and pitfalls of doing a secondary analysis of a qualitative data set", *Qualitative Health Research*, **7 (3)**, pp. 408-24.

Hipp, J., Güntzer, U., Nakhaeizadeh, G., (2000), "Algorithms for association rule mining-a general survey and comparison", *ACM sigkdd explorations newsletter,* **2(1),** pp. 58-64.

Hofferth, S. (2005), "Secondary Data Analysis in Family Research. Journal of Marriage and the Family", **67(4)**, pp. 891-907.

Holte, R. C. (1993), "Very simple classification rules perform well on most commonly used datasets", *Machine learning*, **11(1)**, pp. 63-90.

Hooda, N., Bawa, S., Rana, PS., (2018), "Fraudulent Firm Classification: A Case Study of an External Audit", *Applied Artificial Intelligence*, **32(1)**, pp. 48-64.

Hooda, N. Seema, B. and Prashant, S. R. (2018), "Fraudulent Firm Classification: A Case Study of an External Audit", *Applied Artificial Intelligence,* **32.1**.

Jiawei, Han. and Micheline, Kamber. (2006), "Data Mining Concepts and Techniques, 2nd ed.", *Morgan Kaufmann publishers,* SanFrancisco.

Kantardzic, M. (2003), "Data Mining: Concepts, Models, Methods, and Algorithms", *New York, NY: John Wiley & Sons.*

Kaufmann, L., Rousseeuw, PJ., (1990), "Finding Groups in Data: An Introduction to Cluster Analysis", *New York, John Wiley & Sons.*

Kirkos, E., Spathis, C., Manolopoulos, Y., (2007), "Data mining techniques for the detection of fraudulent financial statements", *Expert systems with applications,* **32(4)**, pp. 995-1003.

Koskivaara, E. (2004), "Artificial Neural Networks in Analytical Review Procedures", *Managerial Auditing Journal*, **19(2)**, pp. 191–223.

Kotsiantis, S., Koumanakos, E., Tzelepis, D., Tampakas, V., (2006), "Forecasting fraudulent financial statements using data mining", *International journal of computational intelligence*, **3(2)**, pp. 104-110.

Kruskal, J. (1977), "The Relationship between Multidimensional Scaling and Clustering", In J. Van Ryzin (Ed.), *Classification and Clustering*, New York, NY: Academic Press Inc., pp. 17-45.

Kyrkos, E. (2015), "Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων", ebook *Athens:Hellenic Academic Libraries Link,* Available Online at: http://hdl.handle.net/11419/1226

Lenz, R., Hahn, U., (2015), "A synthesis of empirical internal audit effectiveness literature pointing to new research opportunities", *Managerial Auditing Journal,* **30(1)**, pp. 5-33.

Lientz, B., Larssen, L., (2012), "Manage IT as a Business", London, Routledge.

Likas, A. Vlassis, N. and Verbeek, J. J. (2003), "The Global k-Means Clustering Algorithm, Pattern Recognition", **36(2)**, pp. 451-461.

Linoff, GS., Berry, MJ., (2011), "Data mining techniques: for marketing, sales, and customer relationship management", *John Wiley & Sons*, Indianapolis, Indiana

Liu, B., Hsu, W., (1996), "Post-analysis of learned rules", AAAI/IAAI, Vol. 1, pp. 828-834.

Liu, B., Hsu, W., Chen, S., Ma, Y., (2000), "Analyzing the Subjective Interestingness of Association Rules", *IEEE Intelligent Systems,* **15(5)**, pp. 47–55.

Liu, B., Hsu, W., Ma, Y., (1998), "Integrating classification and association rule mining", *KDD*, Vol. **98**, pp. 80-86.

MacQueen, J., (1967), "Some methods for classification and analysis of multivariate observations", *Proceedings of the fifth berkeley symposium on mathematical statistics and probability,* California, USA, pp. 281–297.

Minaei-Bidgoli, B., Tan, PN., Punch, WF., (2004), "Mining Interesting Contrast Rules for a Web-based Educational System", *Proceedings of Int. Conf. on Machine Learning Applications*, Louisville, USA 2004, pp. 320- 327.

Moffitt, K C., Vasarhelyi, M A., (2013), "AIS in an age of Big Data. Journal of Information Systems", **27(2)**, pp. 1-19.

Murthy, U S., Groomer, SM., (2004), "A continuous auditing web services model for XML-based accounting systems", *International Journal of Accounting Information Systems,* **5(2)**, pp. 139-163.

Ng, R. T. and Han, J. (1994), "Efficient and Effective Clustering Methods for Spatial Data Mining", *Proceedings of the 20th International conference on Very Large Data Bases*, San Francisco, CA: Morgan Kaufmann Publishers, pp. 144-155.

PwC, (2013), "Maximising internal audit value: 2013 state of the internal audit profession survey - Russia supplement", PwC Russia, Moscow, Ανακτήθηκε στις 20-10-2019 από https://www.pwc.ru/ru/riskassurance/assets/russian-ia-survey-2013-en.pdf.

Rabinovich, E. and Cheon, S.-H. (2011), 'Expanding horizons and deepening understanding via the use of secondary data sources', Journal of Business Logistics, **32(4)**, pp. 303–316.

Ramageri, B. M. (2010), "DATA MINING TECHNIQUES AND APPLICATIONS", *Indian Journal of Computer Science and Engineering*, **1(4)**, pp. 301-305.

Schaltegger, S., Burritt, R., (2017), "Contemporary environmental accounting: issues, concepts and practice", London, Routledge.

Schutt, R. K. (2007), "Secondary Data Analysis, In The Blackwell Encyclopedia of Sociology", G. Ritzer (Ed.)

Sharma, A., Panigrahi, PK., (2013), "A review of financial accounting fraud detection based on data mining techniques", arXiv preprint arXiv:1309.3944.

Singleton, T. (2006), "Generalized Audit Software: Effective and Efficient Tool for Today's IT Audits", *IS ACA*.

Singleton, T., Singleton, AJ., (2005), "Auditing headaches? Relieve them with CAR", *Journal of Corporate Accounting & Finance*, **16(4)**, pp. 17-27.

Smith, E. (2008), "Pitfalls and promises: the use of secondary data analysis in educational research", *British Journal of Educational Studies*, **56:3**, pp. 323-339.

Socea, A D. (2012), "Managerial decision-making and financial accounting information", *Procedia-Social and Behavioral Sciences,* **58**, pp. 47-55.

Staff, A. (2014), "Reimagining auditing in a wired world, Technical report", *University of Zurich, Department of Informatics*, Zurich: Citeseer.

Szabo, V. and Strang, V. R. (1997), "Secondary analysis of qualitative data. Advances in Nursing Science", **20(2)**, pp. 66–74.

The Institute of Internal Auditors Research Foundation (2015), *Staying a Step Ahead Internal Audit's Use of Technology.*

Tysiac, K. (2015), "Data analytics helps auditors gain deep insight. Journal of Accountancy", **219(4)**, pp. 52.

Thorne, S. (1998), "Ethical and representational issues in qualitative secondary analysis", *Qualitative Health Research*, **8 (4)**, pp. 547-55.

UCI1, (2018), Ανακτήθηκε στις 14-11-2019 από https://archive.ics.uci.edu/ml/index.php.

UCI2, (2018), Ανακτήθηκε στις 14-11-2019 από https://archive.ics.uci.edu/ml/datasets/Audit+Data#.

Vasarhelyi, M., Chan, DY., (2011), "Innovation and Practice of Continuous Auditing", *International Journal of Accounting Information Systems,* **12 (2)**, pp. 152-160.

Vasarhelyi, M., Kogan, A., Tuttle, BM., (2015), "Big Data in accounting: An overview", *Accounting Horizons,* **29(2)**, pp. 381-396.

Vasarhelyi, M., Romero, S., Kuenkaikaew, S., Littley, J. (2012), "Adopting Continuous Audit/Continuous Monitoring in Internal Audit", *ISACA Journal,* pp. 3-31.

Wang, S. (2010), "A comprehensive survey of data mining-based accounting-fraud detection research", *International Conference on Intelligent Computation Technology and Automation*, IEEE, Vol. 1, pp. 50-53.

Weka, (2018), Ανακτήθηκε στις 20-11-2019 από https://www.cs.waikato.ac.nz/ml/weka/.

Welch, C. (2000), "The archaeology of business networks: The use of archival records in case study research", *Journal of Strategic Marketing*, **8**, pp. 197-208.

Wells, J. T. (1997), "Occupational Fraud and Abuse", *Austin, TX: Obsidian Publishing*.

Witten, IH., Frank, E., Hall, MA., Pal, CJ., (2016), "Data Mining: Practical machine learning tools and techniques", *Morgan Kaufmann*, Burlington, Massachusetts.

Zhao, N., Yen, D., Chang, I., (2004), "Auditing in the e-commerce era", *Information Management & Computer Security*, **12(5)**, pp. 389-400.

Καραγεώργος, Δ.Λ. (2002), "Μεθοδολογία Έρευνας στις Επιστήμες Αγωγής", *Εκδόσεις Σαββάλας*.

Κύρκος, Ε. (2007), "Εξόρυξη χρηματοοικονομικών δεδομένων με εφαρμογή στη λογιστική", *Doctoral dissertation, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης (ΑΠΘ)*, Σχολή Θετικών Επιστημών. Τμήμα Πληροφορικής.

Ματιάκη, Α. (2007), "Η Εξόρυξη Δεδομένων (Data Mining) στην Λογιστική και Ελεγκτική", *Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης (ΑΠΘ)*, ΜΠΣ Τμήμα Πληροφορικής.

Ψαρρού, Μ. και Ζαφειρόπουλος, Κ. (2004), "Επιστημονική Έρευνα", *Εκδόσεις Τυπωθείτω*.