Σελίδα τίτλου

ΠΑΝΕΠΙΣΤΗΜΙΟ
ΙΩΑΝΝΙΝΩΝ

ΣΧΟΛΗ ΟΙΚΟΝΟΜΙΚΩΝ ΚΑΙ ΔΙΟΙΚΗΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΛΟΓΙΣΤΙΚΗΣ & ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗΣ

ΠΜΣ «ΛΟΓΙΣΤΙΚΗ, ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗ

ΚΑΙ ΔΙΟΙΚΗΤΙΚΗ ΤΩΝ ΕΠΙΧΕΙΡΗΣΕΩΝ»

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΑΝΊΧΝΕΥΣΗ ΨΕΥΔΏΝ ΟΙΚΟΝΟΜΙΚΏΝ ΚΑΤΑΣΤΆΣΕΩΝ ΜΕ ΜΕΘΌΔΟΥΣ ΕΞΌΡΥΞΗΣ ΔΕΔΟΜΈΝΩΝ

Βασιλική Διαμάντη

Επιβλέπων: Χρήστος Γκόγκος

Αναπληρωτής Καθηγητής

Πρέβεζα, Μάρτιος, 2020

DETECTION OF FRAUDULENT FINANCIAL STATEMENTS USING DATA MINING METHODS

Εγκρίθηκε από τριμελή εξεταστική επιτροπή

Πρέβεζα, 31/03/2020

ΕΠΙΤΡΟΠΗ ΑΞΙΟΛΟΓΗΣΗΣ

1. Επιβλέπων καθηγητής

   Χρήστος Γκόγκος,

   Αναπληρωτής Καθηγητής

2. Μέλος επιτροπής

   Ευάγγελος Χύτης,

   Επίκουρος Καθηγητής

3. Μέλος επιτροπής

   Κωνσταντίνος Καραμάνης,

   Αναπληρωτής Καθηγητής

© Διαμάντη, Βασιλική, 2020.

Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Δήλωση μη λογοκλοπής

Δηλώνω υπεύθυνα και γνωρίζοντας τις κυρώσεις του Ν. 2121/1993 περί Πνευματικής Ιδιοκτησίας, ότι η παρούσα μεταπτυχιακή εργασία είναι εκ ολοκλήρου αποτέλεσμα δικής μου ερευνητικής εργασίας, δεν αποτελεί προϊόν αντιγραφής ούτε προέρχεται από ανάθεση σε τρίτους. Όλες οι πηγές που χρησιμοποιήθηκαν (κάθε είδους, μορφής και προέλευσης) για τη συγγραφή της περιλαμβάνονται στη βιβλιογραφία.

Διαμάντη, Βασιλική

Υπογραφή

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή κ. Χρήστο Γκόγκο για την εμπιστοσύνη που μου επέδειξε και την πολύτιμη καθοδήγησή του κατά τη διάρκεια εκπόνησης της εργασίας. Ιδιαίτερα θα ήθελα να τον ευχαριστήσω για την εποικοδομητική συνεργασία μας καθώς και τις πολύτιμες υποδείξεις του, παράγοντες οι οποίοι συνετέλεσαν αποφασιστικά στην άρτια διεκπεραίωση της εργασίας. Θα ήθελα, επίσης, να ευχαριστήσω τον καθηγητή κ. Χύτη Ευάγγελο για τη συμβολή του κατά τη διάρκεια εκπόνησης της εργασίας.

ΠΕΡΙΛΗΨΗ

Εξαιτίας των γρήγορα αναπτυσσόμενων τεχνολογικών δραστηριοτήτων έχουν δημιουργηθεί τεράστια σύνολα δεδομένων (Big Data). Λόγω του μεγάλου όγκου δεδομένων και της πολυπλοκότητας τους είναι δύσκολο να επεξεργαστούν και να ανακαλυφθεί γνώση με παραδοσιακούς τρόπους. Πλέον, το κατάλληλο εργαλείο για την ανακάλυψη της κρυμμένης γνώσης (Knowledge Discovery in Databases, KDD) είναι η εξόρυξη δεδομένων (Data Mining). Η εύρεση χρήσιμων δεδομένων , η ταυτοποίηση κρυμμένων προτύπων και η σωστή εκμετάλλευση της γνώσης που εξάγεται κάνει τον κλάδο της εξόρυξης δεδομένων έναν από τους ταχύτερα αναπτυσσόμενους κλάδους.

Στην συγκεκριμένη εργασία γίνεται εξόρυξη δεδομένων για ανίχνευση ψευδών οικονομικών καταστάσεων σε Ελληνικές εταιρείες εισηγμένες στο χρηματιστήριο Αθηνών με μεθόδους μηχανικής μάθησης. Χρησιμοποιείται το λογισμικό Weka και τρεις μέθοδοι ταξινόμησης, τα δέντρα αποφάσεων, τα νευρωνικά δίκτυα και τα Bayesian Belief Networks.

Λέξεις-κλειδιά: Ελεγκτική, Απάτη, Οικονομικές Καταστάσεις, Εξόρυξη Δεδομένων, Επιστήμη των δεδομένων, Μηχανική Μάθηση

ABSTRACT

Due to the rapidly expanding technology activities, huge data sets have been created (Big Data). Due to the large volume of data and their complexity, it is difficult to work with them and discover knowledge in traditional ways. Now, the best tool for Knowledge Discovery in Databases (KDD) is Data Mining. Finding useful data, identifying hidden patterns and properly exploiting the knowledge discovered makes the data mining industry one of the fastest growing industries.

This study extracts data for the detection of fraudulent financial statements in Greek companies listed on the Athens Stock Exchange using machine learning methods. Weka software and three classification methods, decision trees, neural networks and Bayesian Belief Networks, are used.

Keywords: Auditing, Fraud, Financial Statements, Data Mining, Data Science, Machine Learning

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΚΑΤΑΛΟΓΟΣ ΔΙΑΓΡΑΜΜΑΤΩΝ/ΕΙΚΟΝΩΝ

1.ΕΙΣΑΓΩΓΗ

1.1 ΕΙΣΑΓΩΓΉ

Στο σύγχρονο επιχειρηματικό περιβάλλον, τα λογιστικά γεγονότα που αφορούν μια οικονομική μονάδα είναι άπειρα και πολυσύνθετα και έτσι ο έλεγχος είναι μια εξαιρετικά απαιτητική εργασία. Με την εξόρυξη δεδομένων αλλάζει ο τρόπος διενέργειας του ελέγχου και αναπτύσσονται καινούργιες μέθοδοι ελέγχου.

Η ανάλυση των οικονομικών δεικτών και το ύψος των λογαριασμών στις οικονομικές καταστάσεις είναι μια κλασική μέθοδος για την ανίχνευση απάτης. Ο λόγος είναι ότι οι εταιρείες επιθυμούν να παρουσιάζουν μια καλή εικόνα σχετικά με την ανάπτυξη της εταιρείας στους επενδυτές, στους πιστωτές και την επιτροπή κεφαλαιαγοράς.

Οι τεχνικές της ανακάλυψης γνώσης και εξόρυξης δεδομένων θα εφαρμοστούν σε Ελληνικές εταιρείες χρησιμοποιώντας τις χρηματοοικονομικές καταστάσεις τους. Μέσα από τις οικονομικές καταστάσεις αντλήθηκαν δεδομένα για το ύψος κάποιων λογαριασμών και υπολογίστηκαν αριθμοδείκτες από 5 κατηγορίες, αριθμοδείκτες Κερδοφορίας, Ρευστότητας, Βιωσιμότητας, Δραστηριότητας και Κεφαλαιακής Δομής. [Νιάρχου Ν., 2004]

Για το πρόβλημα των ψευδών οικονομικών καταστάσεων έχει γίνει έρευνα και από τους Kotsiantis S., Koumanakos E., Tzelepis D., Tampakas V. το 2005, τους Kirkos E., Spathis Ch., Manolopoulos Y. το 2007 a, τους Kirkos E., Spathis Ch., Manolopoulos Y. το 2007 b τους Dalnial H., Kamaluddin A., Sanusi Z.M., Khairuddin K.S. το 2014, τους Tarjo, Herawati N. το 2015, τον Coderre G.D. το 1999, τους Calderon T.G., Cheh J.J. το 2002, τους Kirkos S., Manolopoulos Y. το 2004, τον Koskivaara E. το 2004, τον Person O. το 1995, τους Spathis C., Doumpos M., Zopounidis C. το 2002, τον Gaganis C. το 2009, τον Perols J. το 2011, τους Ravisankar P., Ravi, Raghava R.G., Bose I. το 2011 και τους Kanapickiene R., Grundiene Z. το 2015 και στην εργασία ακολουθήθηκε η ίδια στρατηγική για την ανίχνευση ψευδών οικονομικών καταστάσεων. Αρχικά, γίνεται έρευνα και ανακαλύπτονται εταιρείες με ψευδείς και αληθείς οικονομικές καταστάσεις και για τις εταιρείες αυτές υπολογίζονται αριθμοδείκτες και το ύψος λογαριασμών των οικονομικών καταστάσεων. Έπειτα, γίνεται παλινδρόμηση για την ανίχνευση των σημαντικών

αριθμοδεικτών και λογαριασμών. Τέλος, επιλέγεται η μέθοδος εξόρυξης δεδομένων και αξιολογούνται τα αποτελέσματα της εξόρυξης.

1.2 ΣΤΌΧΟΙ

Κύριος στόχος της παρούσας εργασίας είναι να ανακαλυφθούν από ένα σύνολο οικονομικών καταστάσεων λογαριασμοί και αριθμοδείκτες και με βάση την μέθοδο εξόρυξης να ανιχνευθεί ποιες οικονομικές καταστάσεις προβλέπεται να είναι ψευδείς.

Άλλος ένας στόχος είναι από το σύνολο των οικονομικών καταστάσεων, που θα χρησιμοποιηθούν ως δεδομένα εκπαίδευσης, να ταξινομηθούν νέες οικονομικές καταστάσεις που δεν έχουν κατηγοριοποιηθεί εξ' αρχής.

1.3 ΠΡΑΚΤΙΚΉ ΕΦΑΡΜΟΓΉ

Συγκεντρώθηκαν αρχικά στοιχεία από 19 ελληνικές εταιρείες (159 οικονομικές χρήσεις) οι οποίες κατά την διάρκεια της εξεταζόμενης περιόδου (2005-2018) κατά ποσοστό 94,74% ήταν ή είναι ακόμα εισηγμένες στο Χρηματιστήριο Αθηνών και έχουν ελεγχτεί από εξωτερικό ελεγκτή και από την επιτροπή κεφαλαιαγοράς. Η κλαδική τους κατάταξη και το ποσοστό κατά κλάδο παρουσιάζεται στον Πίνακα 1.3.1 (Παράρτημα). Οι 32 οικονομικές καταστάσεις είχαν ενδείξεις ή αποδείξεις ότι ήταν ψευδείς [Kotsiantis S., Tselepis D., Tampakas V., 2005] [Kirkos E., Spathis Ch., Manolopoulos Y., 2007 a] [Kirkos E., Spathis Ch., Manolopoulos Y., 2007 b] και οι υπόλοιπες είναι ταξινομημένες ως αληθείς χωρίς αυτό να εγγυάται ότι δεν είναι ψευδείς. Η κατηγοριοποίηση μιας οικονομικής κατάστασης ως ψευδής βασίστηκε στις εξής παραμέτρους:

- Στην έκθεση ελέγχου, όπου η γνώμη του ελεγκτή διαφοροποιείται άρα έχει σοβαρές αμφιβολίες ως προς την ακρίβεια των οικονομικών καταστάσεων,
- Στην έκθεση ελέγχου, όπου υπάρχει έμφαση θέματος ή έγινε και αναμόρφωση κονδυλίων χωρίς να επιβάλλεται από κάποιο πρότυπο,
- Στον έλεγχο των φορολογικών αρχών που τροποποίησαν τον Ισολογισμό και την κατάσταση αποτελεσμάτων χρήσης,
- Στην εφαρμογή της νομοθεσίας που αφορά την αρνητική καθαρή θέση,
- Την συμπερίληψη της εταιρείας στις υπό επιτήρηση κατηγορίες μετοχών,
- Στην αναστολή διαπραγμάτευσης στο χρηματιστήριο λόγω ψευδών οικονομικών καταστάσεων ή σοβαρών φορολογικών παραβάσεων,

- Στην ύπαρξη εκκρεμών δικαστικών διαδικασιών για ψευδείς οικονομικές καταστάσεις ή σοβαρές φορολογικές παραβάσεις.

Για τις 159 οικονομικές καταστάσεις βρέθηκε το ύψος 16 λογαριασμών και υπολογίστηκαν 32 αριθμοδείκτες. [Kotsiantis S., Tselepis D., Tampakas V., 2005] [Kirkos E., Spathis Ch., Manolopoulos Y., 2007 a] [Kirkos E., Spathis Ch., Manolopoulos Y., 2007 b] [Dalnial H., at al, 2014] [Kanapickiene R., Grudiene Z., 2015] Προκειμένου να μειωθεί η διάσταση των συνολικά 48 μεταβλητών εφαρμόζεται λογιστική παλινδρόμηση (logistic regression) ώστε να βρεθούν ποιες είναι στατιστικά σημαντικές μεταβλητές. [Tarjo, Herawati N., 2015] [Kanapickiene R., Grudiene Z., 2015] Για την λογιστική παλινδρόμηση χρησιμοποιήθηκε το SPSS και ένα δείγμα 61 οικονομικών καταστάσεων, υποσύνολο των 159 οικονομικών καταστάσεων ώστε το δείγμα να αποτελείται από περίπου ίσο αριθμό περιπτώσεων, με 27 ψευδείς και 34 αληθείς οικονομικές καταστάσεις. Λόγω ότι κάποιες μεταβλητές είναι εξαρτημένες μεταξύ τους η λογιστική παλινδρόμηση έγινε σε 4 ομάδες για τους 16 λογαριασμούς και σε 6 ομάδες για τους 32 αριθμοδείκτες. Οι στατιστικά σημαντικές μεταβλητές που βρεθήκαν είναι:

(1) INVENTORIES: αποθέματα
(2) SALES: πωλήσεις ή κύκλος εργασιών
(3) TOTAL LIABILITIES: συνολικές υποχρεώσεις
(4) TOTAL DEBT: συνολικό χρέος
(5) CURRENT ASSETS: κυκλοφορούν ενεργητικό
(6) ACCOUNTS RECEIVABLE: εισπρακτέοι λογαριασμοί ή απαιτήσεις από πελάτες
(7) TOTAL LIABILITIES/TOTAL ASSETS: συνολικές υποχρεώσεις/σύνολο ενεργητικού
(8) CURRENT LIABILITIES/TOTAL ASSETS: βραχυπρόθεσμες υποχρεώσεις/σύνολο ενεργητικού
(9) INVENTORIES/TOTAL ASSETS: αποθέματα/σύνολο ενεργητικού
(10) LONG TERM DEBT/TOTAL ASSETS: μακροπρόθεσμο χρέος/σύνολο ενεργητικού
(11) SALES/FIXED ASSETS: πωλήσεις/πάγιο ενεργητικό
(12) INVENTORIES/CURRENT ASSETS: αποθέματα/κυκλοφορούν ενεργητικό
(13) EBT/FIXED ASSETS: κέρδη προ φόρων/πάγιο ενεργητικό
(14) INVENTORIES/SALES: αποθέματα/πωλήσεις.

Για την εξόρυξη δεδομένων χρησιμοποιείται το λογισμικό Weka και 3 μέθοδοι ταξινόμησης, τα δέντρα αποφάσεων, τα νευρωνικά δίκτυα και τα Bayesian Belief Networks αφού ο εντοπισμός ψευδών οικονομικών καταστάσεων μπορεί να θεωρηθεί τυπικό πρόβλημα ταξινόμησης.

Ο κύριος στόχος της εργασίας αυτής πραγματοποιείται χρησιμοποιώντας το 70% ως δείγμα εκπαίδευσης και το 30% ως δείγμα επαλήθευσης. Αφού έγινε η εξόρυξη δεδομένων και με τις τρεις μεθόδους και αξιολογήθηκαν τα αποτελέσματα βρέθηκε ότι η μέθοδος Bayesian Belief Networks έχει την καλύτερη εφαρμογή στα δεδομένα.

Για την επίτευξη του δεύτερου στόχου, βρέθηκαν νέα στοιχεία από 5 ελληνικές εταιρείες (10 οικονομικές χρήσεις) χωρίς να είναι κατηγοριοποιημένες ως ψευδείς ή αληθείς και η κλαδική τους κατάταξη όπως και το ποσοστό κατά κλάδο παρουσιάζεται στον Πίνακα 1.3.2 (Παράρτημα). Το αρχικό δείγμα με τις 159 οικονομικές καταστάσεις χρησιμοποιήθηκε ως δείγμα εκπαίδευσης, ένα υποσύνολο δεδομένων που χρησιμοποιείται για την αρχική διαμόρφωση του μοντέλου μηχανικής μάθησης, και με την μέθοδο των Bayesian Belief Network έγινε κατηγοριοποίηση του νέου δείγματος.

2.ΕΛΕΓΚΤΙΚΉ

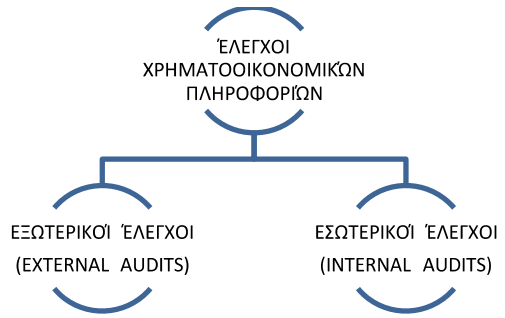
2.1 ΟΡΙΣΜΌΣ

Γεγονός αδιαμφισβήτητο είναι ότι οι εξελίξεις στις σύγχρονες οικονομίες είναι συνεχείς, τόσο σε τοπικό όσο και σε παγκόσμιο επίπεδο. Για το λόγο αυτό οι διοικήσεις έχουν να αντιμετωπίσουν νέες προκλήσεις σε παγκόσμιο επίπεδο που τις φέρουν αντιμέτωπες με νέους τρόπους άσκησης διοίκησης και αντιμετώπισης του επιχειρηματικού κινδύνου. Ουσιαστικό ρόλο στην αντιμετώπιση των κινδύνων παίζει η Ελεγκτική προσφέροντας υπηρεσίες μέσω της ελεγκτικής λειτουργίας. Η Ελεγκτική είναι ένας από τους πιο δυναμικούς και σημαντικούς κλάδους της Λογιστικής, κλάδος του οποίου οι συνεχείς εξελίξεις στο παγκόσμιο περιβάλλον επιβάλλουν διαρκεί εγρήγορση των μελετητών και κλάδος ο οποίος πραγματεύεται τους γενικούς κανόνες, τους όρους και τις προϋποθέσεις για την διενέργεια του ελέγχου.[Καραμάνης Κ., 2008]

Σύμφωνα με τα περισσότερα σύγχρονα εγχειρίδια, έλεγχος είναι η συγκέντρωση από τον ελεγκτή τεκμηρίων σχετικά με ορισμένες πληροφορίες για να διαπιστωθεί ο βαθμός συμφωνίας μεταξύ των πληροφοριών αυτών και ορισμένων κριτηρίων και η έκφραση σχετικής γνώμης μέσω της έκθεσης ελέγχου. Αποσκοπεί δηλαδή στο να διαπιστώσει αν οι προς έλεγχο πληροφορίες έχουν προκύψει με βάση καθορισμένα κριτήρια.

Σκοπός του ελέγχου είναι να βελτιωθεί ο βαθμός εμπιστοσύνης των χρηστών στις οικονομικές καταστάσεις. Αυτό επιτυγχάνεται με την έκφραση μίας γνώμης από τον ελεγκτή στην έκθεση ελέγχου για το εάν οι οικονομικές καταστάσεις έχουν καταρτιστεί σύμφωνα με το εφαρμοστέο πλαίσιο χρηματοοικονομικής αναφοράς. [Καραμάνης Κ., 2008][Κάντζος Κ., Χονδράκη Α., 2006]

Μια βασική διάκριση του ελέγχου είναι μεταξύ εξωτερικού και εσωτερικού.



Σχήμα 2.1: Διάκριση των ελέγχων

Ο εξωτερικός έλεγχος διενεργείται από ανεξάρτητους επαγγελματίες, εξωτερικούς ελεγκτές, για λογαριασμό των χρηστών, έχει δηλαδή το κριτήριο της ανεξαρτησίας. Ο εσωτερικός έλεγχος, αντίθετα, διενεργείται από υπάλληλους της διοίκησης, εσωτερικούς ελεγκτές για δικό της όφελος. [Καραμάνης Κ., 2008] [Κάντζος Κ., Χονδράκη Α., 2006] [Alexander Hamilton Institute, 2011]

2.2 ΙΣΤΟΡΙΚΉ ΕΞΈΛΙΞΗ

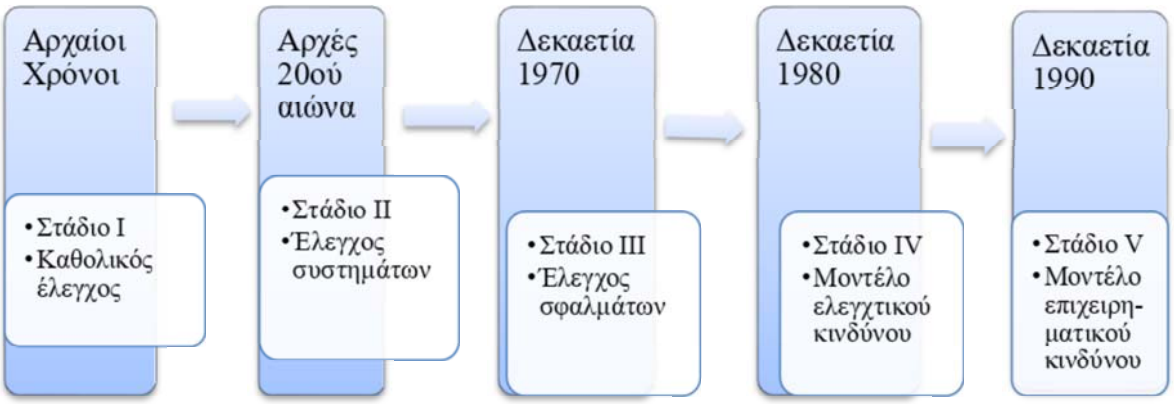
Η ελεγκτική έχει τις ρίζες της στους αρχαίους πολιτισμούς όταν ξεκίνησαν οι ανταλλαγές αγαθών μεταξύ ατόμων και κοινωνικών ομάδων. Οι πρώτες λογιστικές εκθέσεις εμφανίζονται στην Βαβυλώνα το 3000π.Χ. όπου ο γραμματέας νομιμοποιούσε τις μεγάλες συναλλαγές του κράτους. Έπειτα εμφανίστηκαν και στην αρχαία Αίγυπτο, Αθήνα, Ελλάδα και Ρώμη με διάφορες ονομασίες του ελεγκτή. Εκεί δημιουργείται το πρώτο στάδιο της εξέλιξης της ελεγκτικής μεθοδολογίας που είναι ο καθολικός έλεγχος, ο έλεγχος δηλαδή του συνόλου των συναλλαγών και των περιουσιακών στοιχείων. [Καραμάνης Κ., 2008] [Κάντζος Κ., Χονδράκη Α., 2006]

Μεγάλη ανάπτυξη, ωστόσο, άρχισε να γνωρίζει κατά την Βιομηχανική Επανάσταση όπου υπήρξε ραγδαία αύξηση των επιχειρήσεων που υποκαιόταν σε έλεγχο. Στις αρχές του 20ού αιώνα η μεγάλη τεχνολογική πρόοδος αύξησε τα μεγέθη και την πολυπλοκότητα των επιχειρήσεων κι ο καθολικός έλεγχος κατέστη ασύμφορος και αδύνατος γι’ αυτό και περνάει στο δεύτερο στάδιο, στον έλεγχο συστημάτων όπου δίνει έμφαση στον έλεγχο των διάφορων συστημάτων λειτουργίας της επιχείρησης.

Το τρίτο στάδιο , στη δεκαετία του 1970, ο έλεγχος σφαλμάτων εμφανίζεται λόγω της αδυναμίας του ελέγχου συστημάτων και του κόστους του ελέγχου αυτού. Με τον έλεγχο σφαλμάτων διενεργούνται ελεγκτικά τεστ για σφάλματα σε διάφορους επιμέρους λογαριασμούς των χρηματοοικονομικών καταστάσεων. [Καραμάνης Κ., 2008]

Το μοντέλο του ελεγκτικού κινδύνου που αποτελεί το τέταρτο στάδιο εμφανίζεται στη δεκαετία του 1980 λόγω της ανάγκης προστασίας των ελεγκτών από ελεγκτικά σκάνδαλα καθώς και την πίεση στις αμοιβές ελέγχου λόγω ανταγωνισμού. Βασικό χαρακτηριστικό του ελεγκτικού κινδύνου είναι η ανάλυση και αξιολόγηση των κινδύνων για σημαντικά σφάλματα στις χρηματοοικονομικές καταστάσεις.

Από την δεκαετία του 1990 έως και σήμερα υπάρχει το μοντέλο του επιχειρηματικού κινδύνου και αποτελεί το πέμπτο στάδιο της εξέλιξης της ελεγκτικής μεθοδολογίας. Το μοντέλο αυτό δίνει περισσότερη έμφαση στη μελέτη του περιβάλλοντος που λειτουργεί η επιχείρηση και την μελέτη των στρατηγικών κινδύνων που επηρεάζουν την πορεία της. [Καραμάνης Κ., 2008]



Σχήμα 2.2: Ιστορική εξέλιξη

Στην Ελλάδα η εξέλιξη του λογιστικού επαγγέλματος διακρίνεται σε τρία στάδια [Καραμάνης Κ., 2008]:

- 1ο στάδιο: περίοδος μέχρι την έναρξη λειτουργίας του Σώματος Ορκωτών Λογιστών (ΣΟΛ) όπου ο έλεγχος στις ετήσιες χρηματοοικονομικές καταστάσεις των Α.Ε. ήταν εντελώς τυπικός λόγω ανυπαρξίας οργανωμένου ελεγκτικού επαγγέλματος,
- 2ο στάδιο: περίοδος λειτουργίας του ΣΟΛ όπου υπήρξε οργάνωση και λειτουργία του ελεγκτικού επαγγέλματος μέσω υποχρεωτικής ένωσης προσώπων σε ένα Νομικό Πρόσωπο όπου τα μέλη ασκούσαν δημόσιο λειτούργημα,
- 3ο στάδιο: κατάργηση του ΣΟΛ και, σύσταση ελεγκτικών εταιρειών καθώς το επάγγελμα απελευθερώθηκε και ίδρυση μιας νέας επαγγελματικής οργάνωσης του Σώματος Ορκωτών Ελεγκτών Λογιστών (ΣΟΕΛ) όπου όλοι οι αναγνωρισμένοι ελεγκτές είναι μέλη της και μπορούν να παρέχουν ελεγκτικό έργο. [Σώμα Ορκωτών Ελεγκτών Λογιστών (ΣΟΕΛ)]

2.3 ΕΛΕΓΚΤΙΚΈΣ ΑΡΧΈΣ ΚΑΙ ΠΡΌΤΥΠΑ

Τα ελεγκτικά πρότυπα ρυθμίζουν τις αρχές και τις διαδικασίες εκτέλεσης του ελεγκτικού έργου και θέτουν τους κανόνες ώστε να υπάρχει άριστη ενημέρωση και συνεργασία τόσο σε εγχώριο όσο και σε παγκόσμιο επίπεδο.

Σε εθνικό επίπεδο η Επιτροπή Λογιστικής Τυποποίησης και Ελέγχων, ΕΛΤΕ (ΝΠΔΔ), είναι η εθνική εποπτική αρχή του ελεγκτικού και λογιστικού επαγγέλματος και είναι η αρμόδια αρχή για την θέσπιση και εποπτεία εφαρμογής των λογιστικών και ελεγκτικών προτύπων. Η ΕΛΤΕ αποτελείται από το Συμβούλιο Ποιοτικού Ελέγχου με το οποίο ελέγχεται η συμμόρφωση των ορκωτών ελεγκτών με τα Διεθνή Πρότυπα Ελέγχου (ΔΠΕ) και τον κώδικα δεοντολογίας της Διεθνούς Ομοσπονδίας Λογιστών (ΔΟΛ-IFAC) και από το Συμβούλιο Λογιστικής Τυποποίησης που λειτουργεί ως σύμβουλος του Υπουργείου Οικονομικών για τη θέσπιση Λογιστικών Προτύπων για την Ελληνική επικράτεια, και μέσω αυτών των συμβουλίων διενεργείται η εποπτεία της αγοράς. Επίσης η Επιτροπή Κεφαλαιαγοράς (ΝΠΔΔ) είναι η εποπτική αρχή που έχει σκοπό την προστασία των επενδυτών και την εύρυθμη λειτουργία της κεφαλαιαγοράς και εποπτεύει τους ημεδαπούς και αλλοδαπούς φορείς παροχής επενδυτικών υπηρεσιών, της εισηγμένες εταιρίες και τους μετόχους εισηγμένων εταιριών ως προς την χειραγώγηση της αγοράς και την κατάχρηση προνομιακών πληροφοριών. [Καραμάνης Κ., 2008]

Λόγω της παγκοσμιοποίησης και της αύξησης των συναλλαγών μεταξύ των μελών της Ε.Ε. υπήρξε απαραίτητη η θέσπιση Διεθνών Λογιστικών Προτύπων (ΔΛΠ) ώστε να υπάρχει άριστη ενημέρωση και ομαλή λειτουργία μεταξύ των χωρών. Τα ΔΛΠ εκδίδονται από την Διεθνή Ομοσπονδία Λογιστών (ΔΟΛ) που αποστολή της είναι η διασφάλιση του δημοσίου συμφέροντος. Σκοπός του ΔΟΛ είναι να εκδίδει πρότυπα και κανόνες για τον έλεγχο όπως και υπηρεσίες διασφάλισης, τον κώδικα δεοντολογίας, την εκπαίδευση των μελών και τα λογιστικά πρότυπα δημοσίου τομέα. Κάποια, ίσως τα πιο σημαντικά Διεθνή Πρότυπα Ελέγχου που εκδόθηκαν από την ΔΟΛ είναι τα ακόλουθα:

ΔΠΔΠ 1	ΔΙΚΛΙΔΕΣ ΠΟΙΟΤΗΤΑΣ ΓΙΑ ΛΟΓΙΣΤΙΚΕΣ ΕΠΙΧΕΙΡΗΣΕΙΣ ΠΟΥ ΔΙΕΝΕΡΓΟΥΝ ΕΛΕΓΧΟΥΣ ΚΑΙ ΕΠΙΣΚΟΠΗΣΕΙΣ ΟΙΚΟΝΟΜΙΚΩΝ ΚΑΤΑΣΤΑΣΕΩΝ, ΚΑΘΩΣ ΚΑΙ ΑΛΛΕΣ ΑΝΑΘΕΣΕΙΣ ΔΙΑΣΦΑΛΙΣΗΣ ΚΑΙ ΣΥΝΑΦΩΝ ΥΠΗΡΕΣΙΩΝ
ΔΠΕ 200	ΓΕΝΙΚΟΙ ΣΤΟΧΟΙ ΤΟΥ ΑΝΕΞΑΡΤΗΤΟΥ ΕΛΕΓΚΤΗ ΚΑΙ Η ΔΙΕΝΕΡΓΕΙΑ

	ΕΛΕΓΧΟΥ ΣΥΜΦΩΝΑ ΜΕ ΤΑ ΔΙΕΘΝΗ ΠΡΟΤΥΠΑ ΕΛΕΓΧΟΥ
ΔΠΕ 220	ΔΙΚΛΙΔΕΣ ΠΟΙΟΤΗΤΑΣ ΓΙΑ ΤΟΝ ΕΛΕΓΧΟ ΟΙΚΟΝΟΜΙΚΩΝ ΚΑΤΑΣΤΑΣΕΩΝ
ΔΠΕ 240	ΕΥΘΥΝΕΣ ΤΟΥ ΕΛΕΓΚΤΗ ΣΧΕΤΙΚΑ ΜΕ ΑΠΑΤΗ ΣΕ ΕΝΑΝ ΕΛΕΓΧΟ ΟΙΚΟΝΟΜΙΚΩΝ ΚΑΤΑΣΤΑΣΕΩΝ
ΔΠΕ 300	ΣΧΕΔΙΑΣΜΟΣ ΕΝΟΣ ΕΛΕΓΧΟΥ ΟΙΚΟΝΟΜΙΚΩΝ ΚΑΤΑΣΤΑΣΕΩΝ
ΔΠΕ 315	ΕΝΤΟΠΙΣΜΟΣ ΚΑΙ ΕΚΤΙΜΗΣΗ ΤΩΝ ΚΙΝΔΥΝΩΝ ΟΥΣΙΩΔΟΥΣ ΣΦΑΛΜΑΤΟΣ ΜΕΣΩ ΚΑΤΑΝΟΗΣΗΣ ΤΗΣ ΟΝΤΟΤΗΤΑΣ ΚΑΙ ΤΟΥ ΠΕΡΙΒΑΛΛΟΝΤΟΣ ΤΗΣ
ΔΠΕ 320	ΟΥΣΙΩΔΕΣ ΜΕΓΕΘΟΣ ΣΤΟ ΣΧΕΔΙΑΣΜΟ ΚΑΙ ΣΤΗΝ ΕΚΤΕΛΕΣΗ ΕΝΟΣ ΕΛΕΓΧΟΥ
ΔΠΕ 500	ΕΛΕΓΚΤΙΚΑ ΤΕΚΜΗΡΙΑ
ΔΠΕ 700	ΔΙΑΜΟΡΦΩΣΗ ΓΝΩΜΗΣ ΚΑΙ ΕΚΘΕΣΗ ΕΠΙ ΟΙΚΟΝΟΜΙΚΩΝ ΚΑΤΑΣΤΑΣΕΩΝ
ΔΠΕ 705	ΔΙΑΦΟΡΟΠΟΙΗΣΕΙΣ ΤΗΣ ΓΝΩΜΗΣ ΣΤΗΝ ΕΚΘΕΣΗ ΤΟΥ ΑΝΕΞΑΡΤΗΤΟΥ ΕΛΕΓΚΤΗ

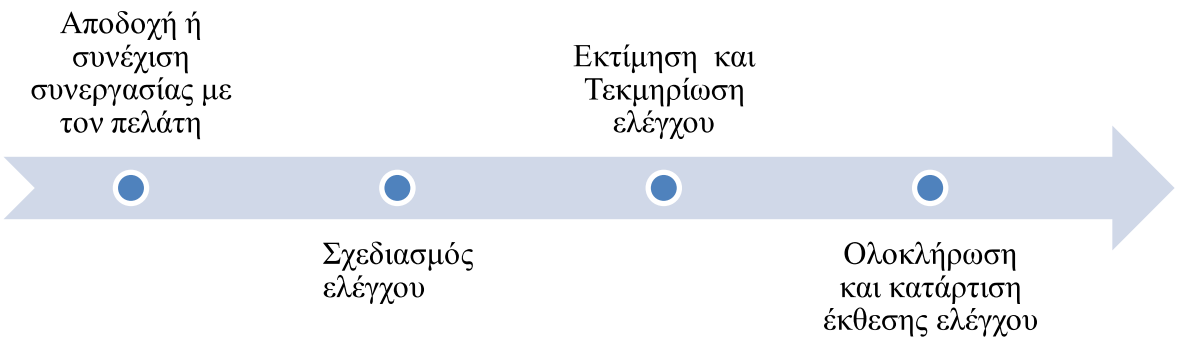
Πίνακας 2.3 : Διεθνή πρότυπα ελέγχου [Διεθνή Ομοσπονδία Λογιστών (IFAC),2010]

2.4 ΔΙΑΔΙΚΑΣΊΕΣ ΕΛΈΓΧΟΥ

Ανεξάρτητα αν είναι εσωτερικός ή εξωτερικός ,ο έλεγχος περιλαμβάνει τα παρακάτω τέσσερα σημαντικά στάδια [Καραμάνης Κ., 2008] [Κάντζος Κ., Χονδράκη Α., 2006]:

- Αποδοχή του πελάτη ή συνέχιση συνεργασίας,
- Σχεδιασμός ελέγχου,
- Εκτίμηση και τεκμηρίωση ελέγχου,
- Ολοκλήρωση και κατάρτιση έκθεσης αποτελεσμάτων ελέγχου.

Παράλληλα πρέπει να υπάρχει επαγγελματική κρίση και σκεπτικισμός.



Σχήμα 2.4: Στάδια ελέγχου

2.4.1 Αποδοχή του πελάτη ή συνέχιση συνεργασίας

Ο ελεγκτής πριν την αναδοχή του ελέγχου θα πρέπει να επικοινωνήσει με τον επερχόμενο ελεγκτή για να ενημερωθεί , να ελέγξει συνοπτικά τις χρηματοοικονομικές καταστάσεις και να εξετάσει το ποινικό μητρώο των διοικητικών στελεχών της επιχείρησης που θα αναλάβει. Επίσης θα πρέπει να εξετάσει ενδεχόμενους κινδύνους αναδοχής του πελάτη και να κάνει έρευνα γι' αυτόν σε τράπεζες, νομικούς συμβούλους κ.α. Αφού γίνει η αναδοχή του ελέγχου ακολουθεί η σύνταξη της επιστολής ανάθεσης ελέγχου που περιλαμβάνει τον σκοπό του ελέγχου, τις ευθύνες της διοίκησης, τις ευθύνες του ελεγκτή και τους ενδεχόμενους περιορισμούς του ελέγχου.

2.4.2 Σχεδιασμός ελέγχου

Στο στάδιο αυτό προσδιορίζονται οι διαδικασίες που πρέπει να εκτελέσει ο ελεγκτής ώστε να συγκεντρώσει τα κατάλληλα τεκμήρια προκειμένου να εκδώσει την έκθεση ελέγχου. Ο σχεδιασμός του ελέγχου περιλαμβάνει τα εξής βήματα [Καραμάνης Κ., 2008][Κάντζος Κ., Χονδράκη Α., 2006]:

1. Την κατανόηση του περιβάλλοντος της επιχείρησης δηλαδή τη φύση της επιχείρησης, τον σκοπό της, τις στρατηγικές και πολιτικές που ακολουθεί, ώστε να εκτιμήσει την πιθανότητα για ουσιώδη σφάλματα και να σχεδιάσει τις κατάλληλες διαδικασίες για τον εντοπισμό τους.
2. Την εκτίμηση των κινδύνων για σφάλματα στις χρηματοοικονομικές καταστάσεις μέσω της κατανόησης του συστήματος εσωτερικών δικλίδων και αξιολόγησης των εσωτερικών ελεγκτών ως προς την αντικειμενικότητα και την ικανότητά τους.
3. Τον καθορισμό του επιπέδου σημαντικότητας ,ή αλλιώς το ουσιώδες μέγεθος, το μέγεθος δηλαδή ενός σφάλματος που καθιστά πιθανό τον επηρεασμό των χρηματοοικονομικών καταστάσεων. Ο ελεγκτής προβαίνει στον σχεδιασμό και την εκτέλεση κάποιον δοκιμασιών για τον εντοπισμό σημαντικών σφαλμάτων. Δεν είναι επιδίωξη του ελεγκτή ο εντοπισμός ασήμαντων ανακριβειών αφού οι ασήμαντες ανακρίβειες δεν επηρεάζουν τις χρηματοοικονομικές καταστάσεις, το κόστος εντοπισμού και διόρθωσης είναι υψηλό και ο χρόνος εντοπισμού τους μπορεί να καθυστερήσουν την έκδοση των χρηματοοικονομικών καταστάσεων.
4. Τον καθορισμό επιπέδου ελεγκτικού κινδύνου, την πιθανότητα δηλαδή ο ελεγκτής να εκφέρει λανθασμένη γνώμη επί των χρηματοοικονομικών καταστάσεων όταν αυτές έχουν σφάλματα που δεν κατάφερε να εντοπίσει. Ο καθορισμός του επιπέδου του

κινδύνου αυτού είναι στην κρίση του ελεγκτή και στα χαρακτηριστικά της κάθε επιχείρησης. Σχεδιάζεται ένα προσωρινό επίπεδο σημαντικότητας και στο τέλος του ελέγχου αξιολογείται το συσωρευτικό άθροισμα των μη ουσιωδών λαθών και του ελεγκτικού κινδύνου.

2.4.3 Εκτίμηση και τεκμηρίωση ελέγχου

Τεκμηρίωση ελέγχου είναι το αρχείο των ελεγκτικών διαδικασιών που εκτελέσθηκαν, τα ελεγκτικά τεκμήρια που αποκτήθηκαν και τα συμπεράσματα που κατέληξε ο ελεγκτής. Έπειτα δημιουργείται ο φάκελος ελέγχου όπου περιέχει τα αρχεία που αποτελούν την τεκμηρίωση ελέγχου. Υπάρχουν 2 ειδών φάκελοι:

- Ο μόνιμος φάκελος που περιλαμβάνει:
    - Καταστατικό επιχείρησης
    - Ιστορικό επιχείρησης
    - Πρακτικά γενικής συνέλευσης και διοικητικού συμβουλίου
    - Οργανόγραμμα
    - Διάφορα συμφωνητικά
    - Οικονομικές καταστάσεις και εκθέσεις ελέγχου προηγούμενων ετών
- Ο προσωρινός φάκελος που περιλαμβάνει:
    - Επιστολή ανάθεσης ελέγχου
    - Συμπεράσματα από την κατανόηση του πελάτη
    - Αποτελέσματα αναλυτικών διαδικασιών
    - Εκτίμηση ελεγκτικού κινδύνου
    - Εκτίμηση επιπέδου σημαντικότητας
    - Πρόγραμμα ελέγχου
    - Αρχείο ελεγκτικών διαδικασιών και τεκμηρίων

2.4.4 Ολοκλήρωση και κατάρτιση έκθεσης αποτελεσμάτων ελέγχου

Κατά την ολοκλήρωση του ελέγχου ο ελεγκτής πρέπει να ολοκληρώσει τις ελεγκτικές διαδικασίες τεκμηρίωσης και εκτίμησης, να ενημερώσει τους νομικούς συμβούλους για ενδεχόμενες υποχρεώσεις, να παίρνει γραπτές βεβαιώσεις από την διοίκηση για μεταγενέστερα γεγονότα και συνδεδεμένα μέρη και να εξετάζει τη δυνατότητα συνέχισης της δραστηριότητας όπως και να γίνουν οι διορθωτικές εγγραφές σε περίπτωση ουσιωδών σφαλμάτων. Τέλος συντάσσεται η έκθεση ελέγχου όπου ο ελεγκτής εκφράζει την γνώμη

του επί των χρηματοοικονομικών καταστάσεων είτε αυτή είναι μη διαφοροποιημένη είτε είναι γνώμη με επιφύλαξη, αρνητική γνώμη ή έχει αδυναμία έκφρασης γνώμης.

2.5 ΤΕΧΝΙΚΈΣ ΕΛΈΓΧΟΥ

Στις μεγάλες επιχειρήσεις συμβαίνουν πολλές συναλλαγές καθημερινά γι΄αυτό ο ελεγκτής πρέπει να κατανέμει τους περιορισμένους πόρους που διαθέτει σε περιοχές υψηλού κινδύνου ,όπου η πιθανότητα να συμβούν ουσιώδη σφάλματα είναι αυξημένη, με τον καλύτερο δυνατό τρόπο. [Καραμάνης Κ., 2008] [Κάντζος Κ., Χονδράκη Α., 2006] Γι΄αυτό για την επιλογή μονάδων προς διενέργεια ελεγκτικών διαδικασιών αναγνωρίζει τρεις διαφορετικές τεχνικές:

1. Επιλογή του συνόλου των μονάδων.

   Ο ελεγκτής επιλέγει να ελέγξει το σύνολο των μονάδων ενός λογαριασμού ή τμήματος αυτού λόγω του επιπέδου σημαντικότητας των μονάδων αυτών ή επειδή αντιπροσωπεύουν αυξημένους κινδύνους σφάλματος. Αυτή η τεχνική είναι εφικτή όταν ο πληθυσμός αποτελείται από λίγες μονάδες μεγάλης αξίας, όταν υπάρχει υψηλός κίνδυνος σφάλματος και εναλλακτικές τεχνικές δεν είναι διαθέσιμες και όταν ο έλεγχος είναι ευχερής ή με ελάχιστο κόστος. Κλασικά παραδείγματα επιλογής του συνόλου μονάδων ενός πληθυσμού είναι οι συναλλαγές και υπόλοιπα που δεν δικαιολογούνται από την συνήθη δραστηριότητα της επιχείρησης, συναλλαγές που έχουν μεγάλη πιθανότητα σφάλματος ή εμπλέκεται η διοίκηση κ.α.

2. Κατευθυνόμενη επιλογή.

   Στην κατευθυνόμενη επιλογή μονάδων ο ελεγκτής μπορεί να επιλέξει συγκεκριμένες μονάδες με βάση την κρίση του και με βάση δύο βασικά κριτήρια:

   1. Κάλυψη μεγάλου ποσοστού της λογιστικής αξίας του πληθυσμού που στοχεύει σε σφάλματα υπερτίμησης. Στηρίζεται στο ότι όταν είναι εφικτή η εξέταση ενός μεγάλου ποσοστού λογιστικής αξίας οδηγεί σε υψηλή διασφάλιση.
   2. Κάλυψη των μονάδων με τον υψηλότερο κίνδυνο όπου χρειάζεται προσεκτική μελέτη των κριτηρίων και τεκμηρίων ώστε να εντοπίζεται το σύνολο του πληθυσμού που έχει τον υψηλότερο κίνδυνο.

Σε κάθε περίπτωση όμως ο ελεγκτής οφείλει να εξετάζει αν απαιτούνται πρόσθετες διαδικασίες για τις μονάδες που δεν ελέγχτηκαν εφόσον η λογιστική αξία αυτών είναι σημαντική και να επιλέξει και άλλες τεχνικές αν το κρίνει απαραίτητο.

3. Δειγματοληψία.

   Στην δειγματοληψία ο ελεγκτής εφαρμόζει διαδικασίες σε λιγότερο από το 100% των μονάδων ενός πληθυσμού και βγάζει συμπεράσματα για ολόκληρο τον πληθυσμό. Στην ελεγκτική δειγματοληψία υπάρχουν τέσσερα στάδια:

   1. Σχεδιασμός του δείγματος,
   2. Επιλογή του δείγματος,
   3. Εφαρμογή των ελεγκτικών διαδικασιών,
   4. Αξιολόγηση των ευρημάτων και εξαγωγή συμπερασμάτων για ολόκληρο τον πληθυσμό.

   Τα βασικά είδη της ελεγκτικής δειγματοληψίας είναι η στατιστική δειγματοληψία όπου ακολουθείται η στατιστική μεθοδολογία για την επιλογή του δείγματος και η μη στατιστική δειγματοληψία όπου είναι στην κρίση του ελεγκτή η μέθοδος επιλογής του δείγματος π.χ. τυχαία επιλογή, συστηματική επιλογή κ.α.

2.6 ΛΟΓΙΣΤΙΚΉ ΑΠΆΤΗ ΚΑΙ ΠΑΡΑΠΟΊΗΣΗ ΟΙΚΟΝΟΜΙΚΏΝ ΚΑΤΑΣΤΆΣΕΩΝ

Οι ελεγκτές αντιμετωπίζουν τον κίνδυνο να παρέχουν λανθασμένη πληροφόρηση στους χρήστες λόγω της ουσιωδώς εσφαλμένης ή παραπλανητικής πληροφόρησης από την επιχείρηση (κίνδυνος πληροφόρησης). Υπάρχει όμως και η πιθανότητα ο ελεγκτής να εκφέρει λανθασμένη γνώμη, να μην διαφοροποιήσει τη γνώμη του, όταν οι χρηματοοικονομικές καταστάσεις είναι ουσιωδώς εσφαλμένες λόγω σφαλμάτων τα οποία ο ελεγκτής δεν κατάφερε να εντοπίσει. Αυτό ονομάζεται ελεγκτικός κίνδυνος.

Τα ουσιώδη σφάλματα μπορεί να οφείλονται σε λάθη ή απάτη και το διακριτικό στοιχείο μεταξύ τους είναι αν έγιναν ακούσια ή εκούσια. Οι χρηματοοικονομικές καταστάσεις μπορεί να παραποιηθούν με υπερεκτίμηση των εσόδων, με υποεκτίμηση των εξόδων και των υποχρεώσεων, και με γνωστοποιήσεις οι οποίες είναι λανθασμένες ή παραλείπουν σημαντικές πληροφορίες. Αναλυτικότερα, πρακτικές παραποίησης των οικονομικών καταστάσεων αποτελούν [Διεθνή Ομοσπονδία Λογιστών (IFAC),2010]:

1. Η λογιστική κάθαρση όπου εφαρμόζεται κυρίως από τη διοίκηση και παρουσιάζει μειωμένα τα τρέχοντα αποτελέσματα και αυξημένα αυτά τον επομένων χρήσεων. Η συγκεκριμένη μέθοδος εφαρμόζεται από τις νέες διοικήσεις με σκοπό την συγκάλυψη των κινδύνων των προηγούμενων διοικήσεων.
2. Τα λογιστικά τεχνάσματα ώστε να παραποιηθούν:
   [1] Τα μεγέθη εκτός Ισολογισμού όπου δεν αποτυπώνονται στον Ισολογισμό όπως π.χ. ορισμένα μεγάλης αξίας πάγια ώστε να μην υπάρχουν μεγάλες αποσβέσεις, να υπάρχουν λιγότερα έσοδα ή δεν εμφανίζουν μακροπρόθεσμες υποχρεώσεις ώστε να μειώσουν τον βαθμό επικινδυνότητας και να βελτιώσουν τον βαθμό της οικονομικής απόδοσης.
   [2] Τα άυλα στοιχεία ενεργητικού και κεφαλαιοποίηση εξόδων όπου η αξία των άυλων στοιχείων δεν μπορεί να προσδιοριστεί με ακρίβεια και ο προσδιορισμός γίνεται από την διοίκηση, ή όπως η υπερβάλλουσα κεφαλαιοποίηση των εξόδων που δεν πληρούν κριτήρια κεφαλαιοποίησης και αντί για Κατάσταση Αποτελεσμάτων Χρήσης μεταφέρονται στον Ισολογισμό.
   [3] Πραγματικών εσόδων όπου αναγνωρίζεται ένα έσοδο πριν την πραγματοποίηση του ώστε να εμφανίζεται μια πλασματική αύξηση των εσόδων.
   [4] Η εικόνα των επενδυτών και οι προοπτικές της επιχείρησης. Αυτό εμφανίζεται στις λογιστικές πρακτικές επιχειρήσεων διαδικτύου όπου οι συγκεκριμένες επιχειρήσεις δεν συμπεριλαμβάνονται στα διεθνή πρότυπα ελέγχου και οι χρηματοοικονομικές τους καταστάσεις δεν μπορούν να συγκριθούν από περίοδο σε περίοδο ώστε να υπάρχει ξεκάθαρη εικόνα στους επενδυτές και στους ενδιαφερόμενους χρήστες σχετικά με τις πραγματικές προοπτικές της επιχείρησης.

Η διοίκηση αυξάνοντας ή μειώνοντας τα αποτελέσματα της επιχείρησης δεν στοχεύει μόνο στο να επηρεάσει το οικονομικό αποτέλεσμα αλλά και ολόκληρη την εικόνα της κατάστασης αποτελεσμάτων χρήσης. Με αυτόν τον τρόπο παραπλανούν τους ενδιαφερόμενους χρήστες για τις πραγματικές επιδόσεις της επιχείρησης και παρουσιάζουν μια βελτιωμένη εικόνα. Τα βασικότερα κίνητρα για να παραποιηθούν οι οικονομικές καταστάσεις είναι [Διεθνή Ομοσπονδία Λογιστών (IFAC),2010]:

- Κίνητρα που προέρχονται από την λειτουργία της κεφαλαιαγοράς όπως πίεση από τους χρηματοοικονομικούς αναλυτές, άντληση κεφαλαίων από το

Χρηματιστήριο Αξιών, ο δανεισμός από τράπεζες και άλλα πιστωτικά ιδρύματα, οι συγχωνεύσεις και οι εξαγορές επιχειρήσεων, η μερισματική πολιτική κ.α.

- Κίνητρα που προέρχονται από συμβατικές υποχρεώσεις της επιχείρησης όπως δανειακές συμβάσεις, αμοιβές της διοίκησης κ.α.
- Κίνητρα που σχετίζονται με την συμπεριφορά των μελών της διοίκησης όπως διατήρηση των διοικητικών θέσεων μάνατζερ, προαγωγή στις ιεραρχίες κ.α.
- Κίνητρα που σχετίζονται με το ρυθμιστικό πλαίσιο λειτουργίας των επιχειρήσεων όπως το ρυθμιστικό πλαίσιο του κλάδου στον οποίον ανήκει, οι αντιμονοπωλιακές και άλλες ρυθμίσεις, η προσπάθεια φοροδιαφυγής κ.α.
- Κίνητρα που πηγάζουν από την επιχειρησιακή κουλτούρα όπως ο βραχυπρόθεσμος προσανατολισμός της επιχείρησης, μη ρεαλιστικοί προϋπολογισμοί και σχέδια δράσης κ.α.

3.ΕΞΌΡΥΞΗ ΔΕΔΟΜΈΝΩΝ

3.1 ΟΡΙΣΜΌΣ

Με την τεχνολογία να εξελίσσεται συνεχώς πλέον τα δεδομένα είναι διαθέσιμα στους χρήστες μέσα από την ροή πληροφοριών των πληροφοριακών συστημάτων. Όμως η οργανωτική πληροφόρηση που παρέχουν οι βάσεις δεδομένων είναι διασπασμένη και υπάρχει δυσκολία στην αποτελεσματική αξιοποίηση της πληροφόρησης ώστε οι χρήστες να εξάγουν συμπεράσματα και να αποκτήσουν γνώση. Επίσης, ο όγκος δεδομένων που παρέχουν οι βάσεις δεδομένων είναι πολύ μεγάλος και η ανάλυση των δεδομένων είναι χρονοβόρα και δύσκολη, και εκτός των άλλων είναι και επικίνδυνο να επιτρέπεται σε οποιονδήποτε χρήστη να χρησιμοποιεί τα δεδομένα για λήψη αποφάσεων αφού αφορούν πρωτογενή στοιχεία των επιχειρησιακών συστημάτων.

Υπήρξε ανάγκη για ένα σύστημα ενιαίο και ολοκληρωμένο ώστε οι χρήστες να μπορούν να αναλύουν μεγάλο αριθμό δεδομένων που υπάρχουν διαθέσιμα, να προσδιορίζουν σχέσεις μεταξύ των δεδομένων και να συντονίζουν αποτελεσματικά τα διάφορα τμήματα τόσο εσωτερικά μεταξύ τους όσο και με το εξωτερικό περιβάλλον. Αυτό έγινε με την τεχνολογία των Αποθετηρίων Δεδομένων (Data Warehouses, DW) που είναι μια συλλογή δεδομένων οργανωμένων γύρω από τις βασικές διαστάσεις της επιχείρησης, είναι ολοκληρωμένα, ενημερώνονται σε τακτές χρονικές στιγμές και χρησιμεύουν κυρίως στη λήψη αποφάσεων. Η ανάπτυξη των αποθετηρίων δεδομένων έδωσε μεγάλη ώθηση στην εξόρυξη δεδομένων. [Πραστάκος Γρ., 2006]

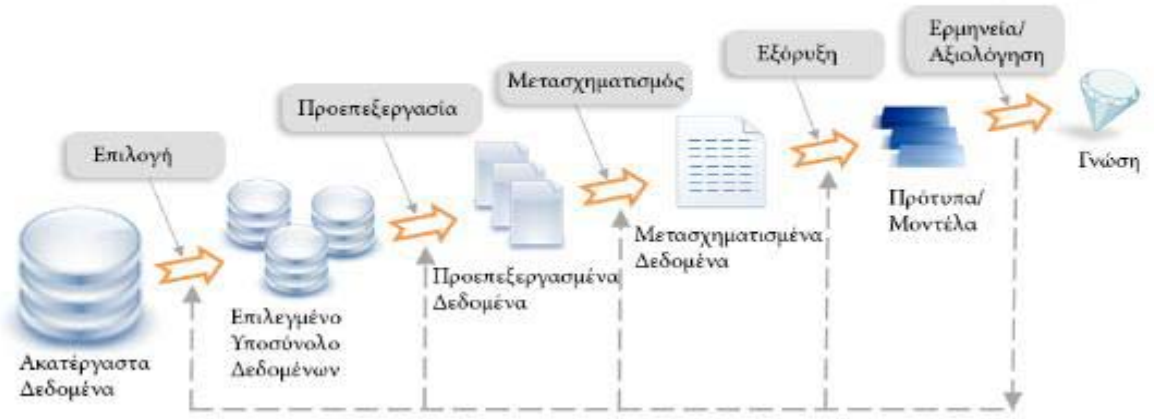
Η Εξόρυξη Δεδομένων (Data Mining, DM) είναι η εξερεύνηση και ανάλυση, με αυτοματοποιημένους ή μη αυτοματοποιημένους τρόπους, μεγάλου όγκου βάσεων δεδομένων προκειμένου να ανακαλυφθούν σχέσεις ή συμπεριφορές που έχουν κάποια σημασία. Αντλεί τα στοιχεία της από το αποθετήριο δεδομένων και χρησιμοποιεί διάφορες μεθόδους προκειμένου να αποκτήσει γνώσεις, μετατρέπει δηλαδή τα δεδομένα σε πληροφορία και οι πληροφορίες αυτές μετατρέπονται σε γνώσεις. Στόχος της εξόρυξης δεδομένων είναι να ανιχνεύσει μια σειρά από σχέσεις που τυχόν υπάρχουν μεταξύ μεγάλου όγκου δεδομένων προκειμένου οι χρήστες να έχουν μια καλή εικόνα και να σχεδιάσουν καλύτερα την πολιτική που θα ακολουθήσουν.

Όπως γίνεται αντιληπτό η εξόρυξη δεδομένων έχει ως βάση τα εξής βασικά επιστημονικά πεδία που αλληλεπιδρούν μεταξύ τους και δημιουργούν τέσσερις κατηγορίες αλγορίθμων, τους αλγόριθμους βάσεων δεδομένων, τους αλγόριθμους τεχνητής νοημοσύνης, τους αλγόριθμους στατιστικής και τους αλγόριθμους οπτικοποίησης. [Κάλλιπος]

Σχήμα 3.1: Βασικά επιστημονικά πεδία εξόρυξης δεδομένων

3.2 ΒΑΣΙΚΆ ΣΤΆΔΙΑ ΤΗΣ ΕΞΌΡΥΞΗΣ ΔΕΔΟΜΈΝΩΝ

Κατά την διαδικασία εξόρυξης γνώσης από τις βάσεις δεδομένων μπορεί να χρειαστεί επανάληψη κάποιου προηγούμενου σταδίου, αφού τα στάδια είναι υπό την μορφή αλυσίδας όπως στο παρακάτω σχήμα. [Κάλλιπος]



Σχήμα 3.2: Βασικά στάδια Εξόρυξης Δεδομένων

1. Συγκέντρωση δεδομένων και κατανόηση περιοχής εφαρμογής

   Γίνεται προετοιμασία για την κατανόηση της περιοχής εφαρμογής και του πλαισίου δράσης. Πρέπει να γίνουν σαφές από την αρχή οι στόχοι της εξόρυξης γνώσης και ώστε να ληφθούν οι σωστές αποφάσεις στα επόμενα βήματα.
2. Επιλογή και δημιουργία κατάλληλου υποσυνόλου δεδομένων

   Σε αυτό το βήμα, προσδιορίζονται τα δεδομένα στα οποία θα αναζητηθούν πρότυπα. Η συλλογή των δεδομένων γίνεται είτε αυτόματα είτε μη αυτόματα και είναι μία πολύ σημαντική διαδικασία αφού έμμεσα προσδιορίζει και το αποτέλεσμα.
3. Προ- επεξεργασία δεδομένων

   Είναι ίσως το πιο σημαντικό στάδιο, αφού προσδιορίζονται τα δεδομένα που είναι ποιοτικά και αξιόπιστα. Πραγματοποιείται καθαρισμός των δεδομένων, τακτοποιούνται δηλαδή εσφαλμένα, προβληματικά και ελλιπή δεδομένα.
4. Μετασχηματισμό δεδομένων

   Τα δεδομένα μετατρέπονται κάτω από ένα κοινό πλαίσιο ώστε να είναι κατάλληλα για εξόρυξη. Γίνεται κλιμάκωση των χαρακτηριστικών των δεδομένων σε ένα συγκεκριμένο και περιορισμένο εύρος τιμών ή ακόμα και δημιουργία νέων χαρακτηριστικών από τα ήδη υπάρχοντα δεδομένα.
5. Επιλογή κατάλληλης μεθόδου Εξόρυξης Δεδομένων και αλγόριθμου Εξόρυξης Δεδομένων

   Γίνεται επιλογή και εφαρμογή της κατάλληλης μεθόδου εξόρυξης δεδομένων . Η επιλογή του αλγορίθμου που θα εφαρμοστεί προσδιορίζεται από το είδος της γνώσης που θα αναζητηθεί.
6. Ερμηνεία και αξιολόγηση αποτελέσματος

   Στο τελευταίο βήμα, γίνεται ερμηνεία της αποτελεσματικότητας της σημαντικότητας των αποτελεσμάτων και αξιολόγησή τους. Οι σύγχρονες μέθοδοι εξόρυξης δεδομένων είναι δυνατό να αποδώσουν ακριβή αποτελέσματα αλλά η ερμηνεία τους αποτελεί ξεχωριστό έργο για την επικύρωση των αποτελεσμάτων.

3.3 ΜΈΘΟΔΟΙ ΕΞΌΡΥΞΗΣ ΔΕΔΟΜΈΝΩΝ

Λόγο των διαφόρων ειδών δεδομένων και ειδών γνώσεις που εξάγονται υπάρχει μεγάλος αριθμός μεθόδων εξόρυξης δεδομένων και διακρίνονται ανάλογα με την μοντελοποίηση που ακολουθούν.[Πραστάκος Γρ., 2006] [Κάλλιπος] Μερικές βασικές μέθοδοι εξόρυξης δεδομένων αναλύονται παρακάτω.

3.3.1 Ταξινόμηση ή κατηγοριοποίηση

Πρόκειται για προγνωστική μέθοδο (μοντέλο πρόβλεψης). Χρησιμοποιείται όταν η μεταβλητή απόκρισης είναι κατηγορική και στόχος της είναι η μάθηση μιας συνάρτησης η οποία απεικονίζει ένα αντικείμενο σε κλάση ή κατηγορία. Στο μοντέλο αυτό δίνονται προκαθορισμένες τάξεις για να εντοπίσει τις παραμέτρους ή τα χαρακτηριστικά ώστε να διαχωριστούν οι τάξεις βάση αυτών. Επίσης δίνονται και δεδομένα επαλήθευσης ώστε η εξόρυξη να έχει ικανοποιητικά αποτελέσματα.

3.3.2 Ομαδοποίηση ή συσταδοποίηση

Ανήκει στην περιγραφική μοντελοποίηση και δίνει λύση σε προβλήματα διαχωρισμού. Στόχος της είναι να διαχωριστούν τα δεδομένα με παρόμοια χαρακτηριστικά και να δημιουργηθούν συστάδες, δηλαδή ομάδες, οι οποίες παρέχουν κοινά ή παρεμφερή χαρακτηριστικά. Η διαφορά με την κατηγοριοποίηση είναι ότι δεν χρειάζονται προκαθορισμένες τάξεις και για το σκοπό της καλύτερης διαχείρισης των δεδομένων είναι συχνά ο πρώτος στόχος της εξόρυξης δεδομένων ώστε η έρευνα να γίνει στην ενδιαφέρουσα ομάδα δεδομένων.

3.3.3 Παλινδρόμηση ή πρόβλεψη

Πρόκειται για μέθοδο πρόβλεψης και είναι σχετική με την κατηγοριοποίηση. Στόχος της είναι η μάθηση μιας συνάρτησης η οποία απεικονίζει ένα αντικείμενο σε μία πραγματική μεταβλητή, δηλαδή με βάση κάποιες ανεξάρτητες μεταβλητές να προβλεφθούν οι τιμές μιας εξαρτημένης μεταβλητής. Χρησιμοποιείται στις περιπτώσεις που η μεταβλητή είναι συνεχής σε αντίθεση με την κατηγοριοποίηση που η μεταβλητή είναι διακριτή.

3.3.4 Ανακάλυψη συσχετίσεων

Ανήκει στην μοντελοποίηση ανάλυσης σύνδεσης ή συνάφειας και σκοπός της είναι να ψάξει την σημαντικότερη σχέση πέρα από τον αριθμό μεταβλητών ή ιδιοτήτων. Με τη μέθοδο αυτή ανακαλύπτονται κρυμμένες συσχετίσεις ενός συνόλου δεδομένων με την βοήθεια των μεγεθών υποστήριξης (support) και εμπιστοσύνης (confidence). Η μέτρηση υποστήριξης δηλώνει το ποσοστό των γεγονότων που εμφανίζονται μαζί συγκρίνοντάς τα με ολόκληρο το πλήθος και η μέτρηση εμπιστοσύνης βεβαιώνει το ποσοστό εμφάνισης του κύριου δεδομένου σε σχέση με δευτερεύοντα.

3.3.5 Ανίχνευση ανωμαλιών ή παρεκτροπών

Στόχος της μεθόδου αυτής είναι να ανακαλύψει αποκλίσεις στα δεδομένα σε σχέση με αντίστοιχα δεδομένα που έχουν συλλεχθεί στο παρελθόν. Τα δεδομένα που είναι ανόμοια με τα δεδομένα παρελθόντος ονομάζονται outlier και μπορούν να αντιμετωπιστούν ως θόρυβοι ή λάθη που πρέπει να διαχωριστούν από τα δεδομένα εισόδου στον αλγόριθμο εξόρυξης δεδομένων. Η μέθοδος αυτή χρησιμοποιείται συνήθως κατά την φάση της προ-επεξεργασίας δεδομένων όπου απαιτείται καθαρισμός των στοιχείων ώστε να υπάρξει ακριβές αποτέλεσμα.

3.4 ΑΠΑΙΤΉΣΕΙΣ ΤΗΣ ΕΞΌΡΥΞΗΣ ΔΕΔΟΜΈΝΩΝ

Είναι αναγκαίο να ελεγχθούν οι απαιτήσεις για την εφαρμογή των μεθόδων ώστε να υπάρξει ολοκληρωμένο αποτέλεσμα από την διαδικασία εξόρυξης δεδομένων. Τα κυριότερα ζητήματα που πρέπει να λαμβάνονται υπόψη κάθε φορά είναι:

- Ο χειρισμός διαφορετικών τύπων δεδομένων καθώς στην εξόρυξη δεδομένων χρησιμοποιούνται διαφορετικοί τύποι και βάσεις δεδομένων σε διαφορετικές εφαρμογές. Τα συστήματα εξόρυξης δεδομένων θα έπρεπε να λειτουργούν ανεξάρτητα από τους τύπους δεδομένων όμως λόγο της ποικιλίας τους είναι σπάνιο να μπορεί να διαχειριστεί όλα τα είδη δεδομένων γι’ αυτό είναι καλό να χρησιμοποιούνται εξειδικευμένα συστήματα για εξόρυξη δεδομένων πάνω σε συγκεκριμένους τύπους δεδομένων.
- Η απόδοση και η επιλεξιμότητα των αλγόριθμων εξόρυξης δεδομένων ώστε να υπάρξει αποτελεσματική εξόρυξη γνώσης με κατάλληλα προσαρμοσμένους αλγόριθμους παρόλα τα μεγάλα σύνολα δεδομένων και ο χρόνος εξόρυξης να είναι αποδεκτός και αναμενόμενος.
- Η χρησιμότητα, η βεβαιότητα και η εκφραστικότητα των αποτελεσμάτων της εξόρυξης δεδομένων καθώς η εξορυγμένη γνώση θα πρέπει να παρουσιάζει με ακριβή τρόπο τα δεδομένα και να είναι χρήσιμη για συγκεκριμένες εφαρμογές. Η ακρίβεια των αποτελεσμάτων εκφράζεται μέσω μέτρων βεβαιότητας , προσεγγιστικά ή ποσοτικά, με βάση τα διάφορα μοντέλα εξόρυξης δεδομένων.
- Οι εκφράσεις διαφορετικού τύπου για τα αποτελέσματα, αφού από μεγάλα σύνολα δεδομένων μπορούν να προκύψουν διαφορετικοί τύποι γνώσεων και θα ήταν χρήσιμο να ελεγχθεί η γνώση από διαφορετικές απόψεις και να εκφραστεί σε διαφορετικές μορφές. Απαιτείται από το σύστημα εξόρυξης δεδομένων να παρουσιάζει

αποτελεσματικά την γνώση χρησιμοποιώντας εκφραστικές τεχνικές αναπαράστασης της γνώσης.

- Η διαλογική ανακάλυψη γνώσης στα πολλαπλά εννοιολογικά επίπεδα μέσω μιας σειράς ερωτήσεων της εξόρυξης δεδομένων, προκειμένου να διαμορφωθεί η εστίαση στα δεδομένα από πολλαπλά επίπεδα και διαφορετικές πτυχές.
- Η εξόρυξη δεδομένων από διαφορετικές πηγές δεδομένων αφού λόγο της δημιουργίας μεγάλων βάσεων δεδομένων αναπτύχθηκαν παράλληλοι και κατανεμημένοι αλγόριθμοι.
- Η προστασία της ιδιωτικότητας και η ασφάλεια των δεδομένων αφού μπορούν να απειληθούν στην περίπτωση που παρατηρηθούν από διαφορετικές οπτικές γωνίες. Είναι σημαντικό να υπάρχουν μέτρα ασφαλείας για να εμποδιστεί η αποκάλυψη ευαίσθητων πληροφοριών όπως και όρια για την εισβολή στην ιδιωτικότητα.

4.ΑΛΓΌΡΙΘΜΟΙ ΕΞΌΡΥΞΗΣ ΔΕΔΟΜΈΝΩΝ

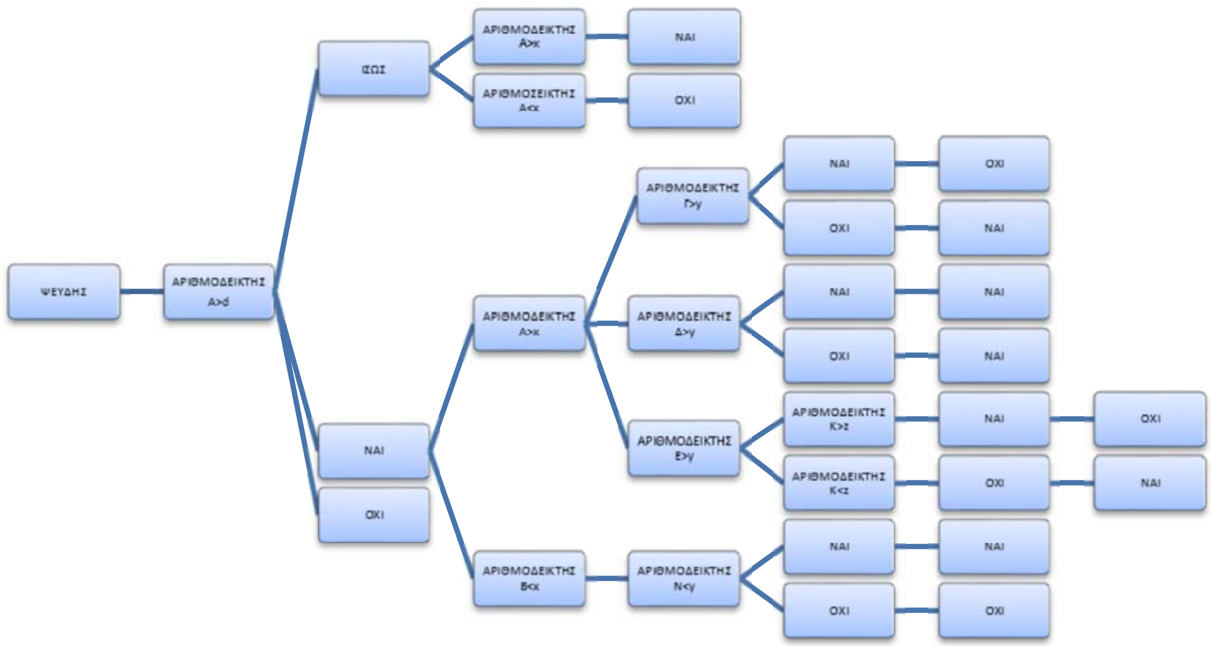
4.1 ΑΛΓΌΡΙΘΜΟΙ ΤΑΞΙΝΌΜΗΣΗΣ

Θεωρείται ότι οι αλγόριθμοι και οι βάσεις δεδομένων αποτελούν μία αδιάσπαστη ενότητα αφού τελικά αποτελούν τη βάση ενός προγράμματος που επιλύει ένα πρόβλημα. Ο εντοπισμός ψευδών οικονομικών καταστάσεων μπορεί να θεωρηθεί ως τυπικό πρόβλημα ταξινόμησης. Όπως αναλύθηκε στο προηγούμενο κεφάλαιο η ταξινόμηση είναι μια διαδικασία όπου ένα μοντέλο εκπαιδεύεται χρησιμοποιώντας ένα δείγμα εκπαίδευσης. ώστε να ταξινομηθούν τα χαρακτηριστικά στις προκαθορισμένη κλάση στην οποία ανήκουν. Αυτό το βήμα είναι επίσης γνωστό ως εποπτευόμενη μάθηση. Έπειτα, το μοντέλο προσπαθεί να ταξινομήσει αντικείμενα που δεν ανήκουν στο δείγμα εκπαίδευσης και να αποτελέσουν το δείγμα επικύρωσης. Η εξόρυξη δεδομένων προτείνει διάφορες μεθόδους ταξινόμησης που προέρχονται από τα πεδία των στατιστικών και της τεχνητής νοημοσύνης. Τρεις μέθοδοι, οι οποίες έχουν καλή φήμη για τις ικανότητες ταξινόμησης είναι οι μέθοδοι Δέντρα Αποφάσεων, Νευρικά Δίκτυα και Bayesian Belief Networks. [Kotsiantis S., Tselepis D., Tampakas V., 2005] [Kirkos E., Spathis Ch., Manolopoulos Y., 2007 a][Kirkos E., Spathis Ch., Manolopoulos Y., 2007 b]

4.2 ΔΈΝΔΡΑ ΑΠΟΦΆΣΕΩΝ

Ένα δένδρο απόφασης έχει μια δομή δέντρου, όπου κάθε κόμβος αντιπροσωπεύει μια δοκιμή σε ένα χαρακτηριστικό και κάθε κλαδί αντιπροσωπεύει ένα αποτέλεσμα της δοκιμής. Με αυτό τον τρόπο, το δέντρο επιχειρεί να διαιρέσει τις παρατηρήσεις σε αμοιβαία αποκλειόμενες υποομάδες. Το πόσο καλός θα είναι ένας διαχωρισμός βασίζεται στην επιλογή του χαρακτηριστικού που διαχωρίζει καλύτερα το δείγμα. Το δείγμα διαιρείται διαδοχικά σε υποσύνολα, έως ότου είτε δεν θα υπάρχει περαιτέρω διάσπαση που μπορεί να προκαλέσει στατιστικά σημαντικές διαφορές ή οι υποομάδες είναι πολύ μικρές για να υποβληθούν σε παρόμοια ουσιαστική διαίρεση. Η διαδοχική διαίρεση του δείγματος μπορεί να παράγει ένα μεγάλο δέντρο. Μερικά από τα κλαδιά του δέντρου μπορεί να αντανακλούν ανωμαλίες στο σύνολο εκπαίδευσης, όπως ψευδείς αξίες ή υπερβολικές τιμές. Για το λόγο αυτό απαιτείται κλάδεμα δέντρων. Το κλάδεμα των δέντρων περιλαμβάνει την αφαίρεση των κόμβων διάσπασης με τρόπο που δεν επηρεάζει σημαντικά το ποσοστό ακρίβειας του μοντέλου. Τα κυριότερα πλεονεκτήματα των δέντρων αποφάσεων είναι ότι παρέχουν έναν ουσιαστικό τρόπο αντιπροσώπευσης των

αποκτηθεισών γνώσεων και διευκολύνουν την εξαγωγή των κανόνων ταξινόμησης αντότε. [Πραστάκος Γρ., 2006]



Σχήμα 4.2: Γενική μορφή δέδρων αποφάσεων

Υπάρχουν πολλοί προτεινόμενοι αλγόριθμοι διαίρεσης με αυτόματη ανίχνευση αλληλεπίδρασης. Ένας βασικούς αλγόριθμους για δένδρα αποφάσεων με αυτόματη ανίχνευση αλληλεπίδρασης είναι ο ID3. Ο ID3 χρησιμοποιεί ένα μέτρο βασισμένο στην εντροπία (αταξία), γνωστό ως κέρδος πληροφοριών, για να επιλέξει το χαρακτηριστικό διάσπασης. Προκειμένου να ταξινομηθεί ένα προηγουμένως αόρατο αντικείμενο, οι τιμές χαρακτηριστικών του αντικειμένου εξετάζονται σε σχέση με τους κόμβους διαίρεσης του δέντρου αποφάσεων. Σύμφωνα με αυτό το τεστ, εντοπίζεται μια διαδρομή που θα ολοκληρωθεί με την πρόβλεψη της κλάσης του αντικειμένου. [Using ID3 Algorithm to build a Decision Tree to predict the weather] Τα βήματα που ακολουθούνται στον ID3 είναι τα εξής:

Βήμα 1: Γίνεται η επιλογή ενός μόνο κόμβου που αντιπροσωπεύει ολόκληρο το σύνολο των δεδομένων εκπαίδευσης, δημιουργείται δηλαδή η ρίζα του δένδρου.

Βήμα 2: Δημιουργούνται τα φύλλα και ορίζεται η ετικέτα της κατηγορίας όταν τα δείγματα είναι της ίδιας κατηγορίας και έχουν ταξινομηθεί σε κόμβους.

Βήμα 3: Ο αλγόριθμος χρησιμοποιεί ένα μέτρο εντροπίας για την επιλογή των γνωρισμάτων που χωρίζουν καλύτερα τα δείγματα. Το γνώρισμα με το μεγαλύτερο κέρδος πληροφορίας επιλέγεται σαν γνώρισμα ελέγχου για την δημιουργία των κόμβων.
Βήμα 4: Δημιουργούνται οι κόμβοι που ονομάζονται γνώρισμα ελέγχου όσο δημιουργούνται κλαδιά για κάθε τιμή του κόμβου και παράλληλα το δείγμα χωρίζεται περεταίρω.
Βήμα 5: Ο αλγόριθμος συνεχώς ταξινομεί τα δείγματα σε κάθε προκαθορισμένη κατηγορία και σταματάει μόνο όταν όλα τα δείγματα του κάθε φύλλου ανήκουν στην ίδια κατηγορία ή δεν υπάρχουν άλλα γνωρίσματα που τα δείγματα θα μπορούσαν να ταξινομηθούν περεταίρω ή δεν υπάρχουν μη κατηγοριοποιημένα δείγματα για το κλαδί του γνωρίσματος ελέγχου.

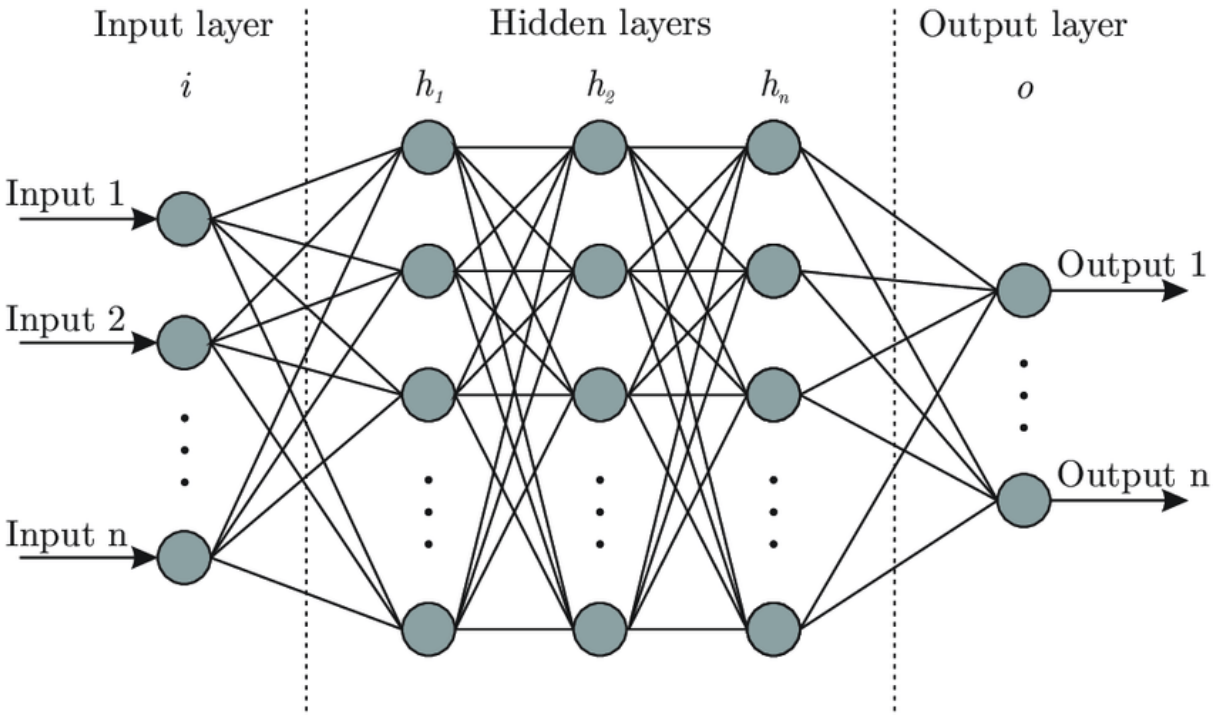
Ένας άλλος αλγόριθμος για την κατασκευή δένδρου αποφάσεων που αποτελεί εξέλιξη του ID3 είναι ο C4.5. Ο αλγόριθμος αυτός αναπτύσσει ένα σύνολο κανόνων αποφάσεων για την ταξινόμηση εναλλακτικών από το σύνολο των δειγμάτων που διαμορφώνουν το δείγμα εκμάθησης. Τα φύλλα υποδεικνύουν την κατηγορία μιας εναλλακτικής η οποία επαληθεύει τη συνθήκη του κλαδιού. Ακολουθούνται τα ίδια βήματα με τον ID3 όμως η ταξινόμηση δεν γίνεται με βάση την κατηγορία αλλά με βάση κριτήρια αξιολόγησης.

Έτσι, με το δένδρο αποφάσεων ταξινομείται το σύνολο του δείγματος και ελέγχεται η ακρίβεια της ταξινόμησης. Αν η ταξινόμηση έχει γίνει σωστά τότε ο αλγόριθμος τερματίζει αφού το δένδρο απόφασης είναι αποδεκτό, αλλιώς η διαδικασία επαναλαμβάνεται και δημιουργούνται και άλλα υποσύνολα μέχρι όλα το δείγμα να ταξινομηθεί σωστά.

4.3 ΝΕΥΡΩΝΙΚΆ ΔΊΚΤΥΑ

Τα νευρωνικά δίκτυα είναι αρκετά παρόμοια με τον ανθρώπινο εγκέφαλο. Αποτελούνται από τεχνητούς νευρώνες, λαμβάνουν πολλαπλές εισόδους και παράγουν μία μόνο έξοδο. Επειδή σχεδόν όλοι οι νευρώνες επηρεάζουν ο ένας τον άλλον, και συνεπώς είναι όλοι συνδεδεμένοι με κάποιο τρόπο, το δίκτυο είναι σε θέση να αναγνωρίσει και να παρατηρήσει όλες τις πτυχές των δεδομένων και τον τρόπο με τον οποίο αυτά τα διαφορετικά κομμάτια δεδομένων μπορούν ή δεν μπορούν να σχετίζονται μεταξύ τους. Μπορεί να βρει πολύ σύνθετα πρότυπα σε ένα μεγάλο όγκο δεδομένων που διαφορετικά θα ήταν αόρατα. Οι νευρώνες διατάσσονται σε στρώματα, ένα πολυεπίπεδο

δίκτυο αποτελείται από τουλάχιστον ένα (πρώτο) και ένα (τελευταίο) επίπεδο. Μεταξύ της στρώσης εισόδου και εξόδου μπορεί να υπάρχει ένα ή περισσότερα κρυμμένα στρώματα.



Σχήμα 4.3: Γενική μορφή νευρωνικών δικτύων

Αφού οριστεί η αρχιτεκτονική δικτύου, το δίκτυο πρέπει να εκπαιδευτεί. Στα δίκτυα οπίσθιας προώθησης, εφαρμόζεται ένα πρότυπο στο στρώμα εισόδου και υπολογίζεται μια τελική έξοδος στο στρώμα εξόδου. Η έξοδος συγκρίνεται με το επιθυμητό αποτέλεσμα και τα σφάλματα πολλαπλασιάζονται προς τα πίσω στο νευρωνικό δίκτυο ρυθμίζοντας τα βάρη των συνδέσεων. Αυτή η διαδικασία επαναλαμβάνεται μέχρι να επιτευχθεί ένα αποδεκτό ποσοστό σφάλματος. [5 algorithms to train a neural network]

Ένας από τους αλγόριθμους που χρησιμοποιούνται στα νευρωνικά δίκτυα είναι ο Gradient descent, γνωστός και ως απότομης πτώσης. Είναι ο απλούστερος αλγόριθμος νευρωνικών δικτύων και αποτελεί μέθοδο πρώτης τάξης. Τα βήματα που ακολουθούνται είναι τα εξής:
Βήμα 1: Υπολογίζεται η κλίση της ευθείας αξιολογώντας τον δείκτη ζημιών
Βήμα 2: Ορίζεται το κριτήριο αξιολόγησης και γίνεται έλεγχος αν είναι αποδεκτό.
Βήμα 3: Αν γίνεται αποδεκτό ο αλγόριθμος τερματίζει ενώ αν δεν γίνεται αποδεκτό επιστρέφει στο βήμα 1 βελτιώνοντας τις παραμέτρους του δικτύου.

Ο Gradient descent είναι ο συνιστώμενος αλγόριθμος όταν έχουμε πολύ μεγάλα νευρωνικά δίκτυα, με πολλές χιλιάδες παραμέτρους. Ένας άλλος αλγόριθμος, δεύτερης τάξης, είναι ο

Newton's method. Τα βήματα που ακολουθούνται είναι τα ίδια με του Gradient descent αλλά σε αυτή την περίπτωση, η βελτίωση των παραμέτρων γίνεται με την απόκτηση πρώτα της κατεύθυνσης εκπαίδευσης του Νεύτωνα, όπου είναι μια συνάρτηση γνωστή ως βήμα του Νεύτωνα και έπειτα ενός κατάλληλου ποσοστού κατάρτισης. Ο στόχος αυτής της μεθόδου είναι να βρεθούν καλύτερες κατευθύνσεις εκπαίδευσης με τη χρήση των δεύτερων παραγώγων της λειτουργίας απώλειας.

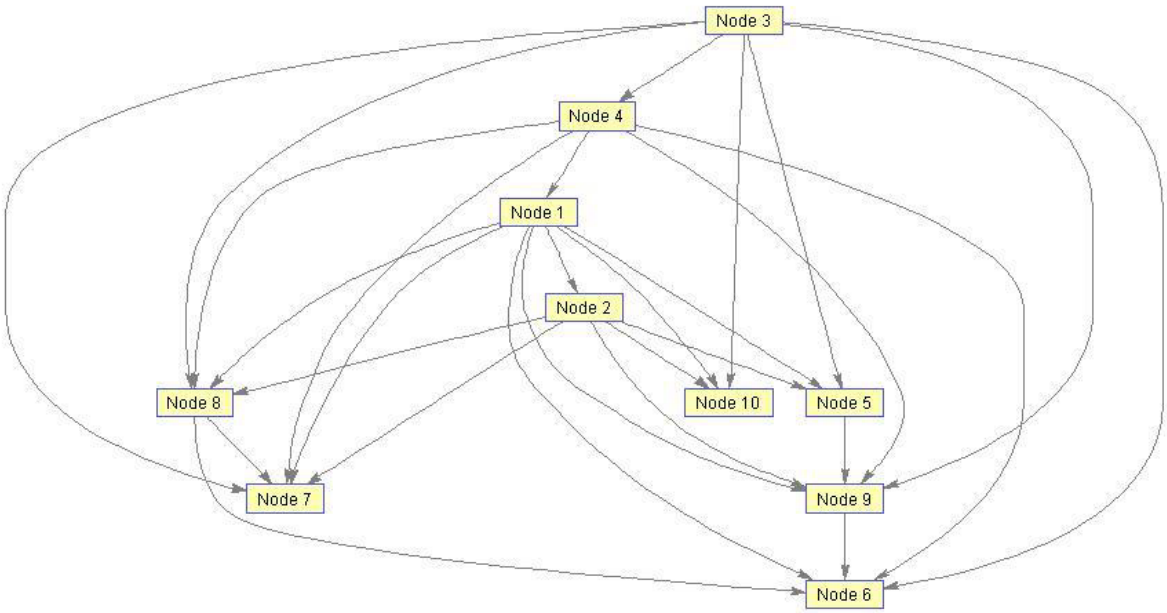
Τα νευρωνικά δίκτυα δεν κάνουν υποθέσεις σχετικά με την ανεξαρτησία των χαρακτηριστικών, είναι ικανά να χειρίζονται θορυβώδη ή ασυνεπή δεδομένα και αποτελούν κατάλληλη εναλλακτική λύση για προβλήματα όπου δεν υπάρχει αλγοριθμική λύση.

4.4 BAYESIAN BELIEF NETWORKS

Η Bayesian ταξινόμηση βασίζεται στο στατιστικό θεώρημα του Bayes. Το θεώρημα Bayes παρέχει έναν υπολογισμό για την εκ των υστέρων πιθανότητα. [Downing D., Clark J., 2010] Σύμφωνα με το θεώρημα Bayes η πιθανότητα να συμβεί το ενδεχόμενο A με δεδομένο ότι έχει συμβεί το ενδεχόμενο B είναι η εξής:

$$P(A/B) = \frac{P(B/A) * P(A)}{P(B)}$$

Οι απλοϊκοί Bayesian ταξινομητές καθιστούν τις κατηγορίες σε κατάσταση ανεξαρτησίας, η οποία δηλώνει ότι η επίδραση μιας τιμής χαρακτηριστικού σε μια δεδομένη κλάση είναι ανεξάρτητη από τις τιμές των άλλων χαρακτηριστικών. Αν αυτή η υπόθεση ισχύει, οι απλοϊκοί Bayesian ταξινομητές έχουν τα καλύτερα ποσοστά ακρίβειας σε σύγκριση με όλους τους άλλους ταξινομητές. Ωστόσο, σε πολλές περιπτώσεις, αυτή η υπόθεση δεν είναι έγκυρη, καθώς μπορούν να υπάρχουν εξαρτήσεις μεταξύ χαρακτηριστικών. Τα Bayesian Networks Belief επιτρέπουν την αντιπροσώπευση των εξαρτήσεων μεταξύ υποσυνόλων χαρακτηριστικών. Ένα Bayesian Networks Belief είναι ένα κατευθυνόμενο όχι σε κυκλικό γράφημα, όπου κάθε κόμβος αντιπροσωπεύει ένα χαρακτηριστικό και κάθε βέλος αντιπροσωπεύει μια πιθανολογική εξάρτηση.



Σχήμα 4.4: Γενική μορφή Bayesian Networks Belief

Αν ένα βέλος προέρχεται από έναν κόμβο Α σε έναν κόμβο Β, τότε το Α είναι γονέας του Β και το Β είναι απόγονος του Α. Σε ένα δίκτυο πεποιθήσεων, κάθε μεταβλητή εξαρτάται από τους μηδενικούς όρους, λόγω των γονέων της. Για κάθε κόμβο x υπάρχει ο Πίνακας Πιθανότητας Ποιότητας, ο οποίος προσδιορίζει την υποθετική πιθανότητα κάθε τιμής του x για κάθε πιθανό συνδυασμό των αξιών των γονέων του. Η δομή δικτύου μπορεί να οριστεί εκ των προτέρων ή μπορεί να συναχθεί από τα δεδομένα. Για λόγους ταξινόμησης ένας από τους κόμβους μπορεί να οριστεί ως ο κόμβος τάξης. Το δίκτυο μπορεί να υπολογίσει την πιθανότητα κάθε εναλλακτικής κλάσης. [Construction of Bayesian network structures from data: A brief survey and an efficient algorithm]

Ένας από τους αλγόριθμους που χρησιμοποιείται στη Bayesian Networks Belief είναι ο K2, των Cooper και Herskovits και δείχνει ότι αποτέλεσμα μπορεί να χρησιμοποιηθεί για να βρεθεί η πιο πιθανή δομή δικτύου, δεδομένης μιας βάσης δεδομένων. Ο αλγόριθμος αυτός μεγιστοποιεί την πιθανότητα P με την εύρεση του γονικού συνόλου κάθε μεταβλητής που μεγιστοποιεί τη συνάρτηση g με δεδομένου ότι υπάρχει ένα σύνολο τεσσάρων υποθέσεων, δηλαδή (1) οι μεταβλητές της βάσης δεδομένων είναι διακριτές, (2) οι περιπτώσεις συμβαίνουν ανεξάρτητα, δεδομένου ενός μοντέλου δικτύου πεποιθήσεων, (3) όλες οι μεταβλητές είναι τυποποιημένες σε κάποια αξία σε κάθε περίπτωση και τέλος (4) πριν από την παρατήρηση της βάσης δεδομένων, είμαστε αδιάφοροι όσον αφορά τις αριθμητικές πιθανότητες να τοποθετήσουμε στη δομή δικτύου πεποίθησης. Ωστόσο,

δεδομένου ότι ο αριθμός των πιθανών δομών αυξάνεται εκθετικά ως συνάρτηση του αριθμού των μεταβλητών, είναι υπολογιστικά ανέφικτο να βρεθεί η πιο πιθανή δομή του δικτύου πεποιθήσεων, λαμβάνοντας υπόψη τα δεδομένα, απαριθμώντας εξαντλητικά όλες τις πιθανές δομές δικτύου πεποιθήσεων.

Εκτός από τις τέσσερις υποθέσεις που αναφέρθηκαν παραπάνω, το K2 χρησιμοποιεί άλλες δύο παραδοχές, δηλαδή ότι υπάρχει διαθέσιμη παραγγελία για τις μεταβλητές και ότι όλες οι δομές είναι εξίσου πιθανές. Τότε, για να βρεθεί το γονικό σύνολο ενός κόμβου, υποθέτει πρώτα ότι ο κόμβος δεν έχει γονείς και στη συνέχεια προσθέτει διαδοχικά αυτόν τον κόμβο (μεταξύ των προκατόχων της παραγγελίας) στο γονικό σύνολο που αυξάνει την πιθανότητα της προκύπτουσας δομής από το μεγαλύτερο ποσό. Σταματά την προσθήκη γονέων στον κόμβο όταν κανένας πρόσθετος γονέας δεν μπορεί να αυξήσει την πιθανότητα της προκύπτουσας δομής.

Άλλος ένας αλγόριθμος Bayesian Networks Belief είναι ο Max-Min Hill-Climbing ο οποίος συνδυάζει τις ιδέες από τις κλασικές τεχνικές μάθησης βασισμένες στους περιορισμούς, και, αναζήτησης και βαθμολογίας, κατά τρόπο βασισμένο σε αρχές και αποτελεσματικό. Αρχικά ανοικοδομεί τον σκελετό ενός Bayesian δικτύου και στη συνέχεια εκτελεί μια Bayesian χάραξη άπληστης αναζήτησης λόφων-αναρρίχηση για τον προσανατολισμό των άκρων. Ο Max-Min Hill-Climbing μαθαίνει τον σκελετό (δηλαδή τις άκρες χωρίς τον προσανατολισμό του) ενός Bayesian δικτύου χρησιμοποιώντας έναν τοπικό αλγόριθμο ανακάλυψης που ονομάζεται Max-Min Parents and Children και στη συνέχεια, προσανατολίζει τον σκελετό χρησιμοποιώντας μια Bayesian χάραξη άπληστης αναζήτησης λόφων-αναρρίχηση. Η φάση αναγνώρισης σκελετού ο Max-Min Hill-Climbing είναι υγιής στο όριο του δείγματος, ενώ η φάση προσανατολισμού δεν παρέχει καμία θεωρητική εγγύηση. Με την εκμάθηση του σκελετού του Bayesian δικτύου, ο Max-Min Hill-Climbing εκτιμά τα υποψήφια γονικά σύνολα, δηλαδή ένας υποψήφιος γονέας του B είναι οποιαδήποτε άλλη μεταβλητή A που μοιράζεται μια άκρη με το B. [The max-min hill-climbing Bayesian network structure learning algorithm]

Το μέγεθος του χώρου αναζήτησης της άπληστης αναζήτησης είναι υπερθετικός του αριθμού των μεταβλητών. Έχουν προκύψει δύο διαφορετικές προσεγγίσεις για τη βελτίωση της αποτελεσματικότητας αυτών των μεθόδων. Η πρώτη προσέγγιση μειώνει την πολυπλοκότητα της αναζήτησης μετατρέποντας τον ίδιο τον χώρο αναζήτησης και η δεύτερη προσέγγιση βελτιώνει την αποτελεσματικότητα της αναζήτησης χρησιμοποιώντας

τους περιορισμούς που τίθενται στην αναζήτηση, όπως ο αλγόριθμος K2. Έτσι, ο αλγόριθμος μπορεί να κλιμακώνεται σε κατανομές με χιλιάδες μεταβλητές και ωθεί την ανακάλυψη αξιόπιστης εκμάθησης του δικτύου Bayesian τόσο σε χρόνο όσο και σε ποιότητα σε μια μεγάλη ποικιλία αντιπροσωπευτικών κλάδων.

Τα Bayesian δίκτυα είναι ιδανικά για την πραγματοποίηση ενός γεγονότος που συνέβη και προβλέποντας την πιθανότητα ότι ο ένας από τους πολλούς πιθανούς γνωστούς λόγους ήταν ο παράγοντας που συνέβαλε. Γενικεύσεις των Bayesian δικτύων μπορούν να αντιπροσωπεύουν και να επιλύουν προβλήματα αποφάσεων κάτω από αβεβαιότητα.

5.PYTHON

5.1 Η ΓΛΏΣΣΑ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΎ PYTHON

Η Python είναι διερμηνευόμενη, γενικού σκοπού και υψηλού επιπέδου, γλώσσα προγραμματισμού. Είναι δυναμική γλώσσα προγραμματισμού και υποστηρίζει τη συλλογή απορριμμάτων καθώς ανήκει στις γλώσσες προστακτικού προγραμματισμού και υποστηρίζει τόσο το διαδικαστικό όσο και το αντικειμενοστραφές προγραμματιστικό υπόδειγμα. Ο κύριος στόχος της είναι η αναγνωσιμότητα του κώδικά της και η ευκολία χρήσης της. Το συντακτικό της επιτρέπει στους προγραμματιστές να εκφράσουν έννοιες σε λιγότερες γραμμές κώδικα από ότι θα ήταν δυνατόν σε γλώσσες όπως η C++ ή η Java. Διακρίνεται λόγω του ότι έχει πολλές βιβλιοθήκες που διευκολύνουν ιδιαίτερα αρκετές συνηθισμένες εργασίες και για την ταχύτητα εκμάθησης της, αλλά μειονεκτεί στο ότι επειδή είναι διερμηνευόμενη είναι πιο αργή από τις μεταγλωττιζόμενες γλώσσες όπως η C και η C++ και γι΄αυτό το λόγο δεν είναι κατάλληλη για γραφή λειτουργικών συστημάτων. [Python]

Η γλώσσα προγραμματισμού Python δημιουργήθηκε από τον Ολλανδό Γκίντο βαν Ρόσσουμ (Guido van Rossum) στο ερευνητικό κέντρο Centrum Wiskunde & Informatica (CWI) το 1989 και κυκλοφόρησε για πρώτη φορά το 1991. Αρχικά, η Python ήταν γλώσσα σεναρίων που χρησιμοποιήθηκε στο κατανεμημένο λειτουργικό σύστημα Amoeba, ικανή και για κλήσεις συστήματος. Θεωρείται διάδοχος της γλώσσας προγραμματισμού ABC, μια και αυτή υπήρξε η βασική πηγή έμπνευσης για τον Γκίντο βαν Ρόσσουμ. [Wikipedia]

Λόγω της επεκτασιμότητας της Python και της φύσης γενικού σκοπού της συχνά χρησιμοποιούσε για την ανάλυση δεδομένων. Η Python είναι συχνά η επιλογή για προγραμματιστές που πρέπει να εφαρμόσουν στατιστικές τεχνικές ή ανάλυση δεδομένων στο έργο τους ή για επιστήμονες δεδομένων των οποίων τα καθήκοντα πρέπει να ενσωματωθούν με εφαρμογές ιστού ή περιβάλλοντα παραγωγής αφού ο συνδυασμός βιβλιοθηκών μηχανικής μάθησης και ευελιξίας την καθιστά κατάλληλη για την ανάπτυξη εξελιγμένων μοντέλων και μηχανισμών πρόβλεψης.

5.2 ΒΙΒΛΙΟΘΉΚΕΣ ΤΗΣ PYTHON

Η μεγάλη δύναμη της Python είναι το πλήθος των ελεύθερων third-party βιβλιοθηκών. [Python] Σχεδόν για κάθε περίπτωση, υπάρχει διαθέσιμη μια υψηλής ποιότητας

βιβλιοθήκη. Μεταξύ αυτών υπάρχουν μερικές εξαιρετικές βιβλιοθήκες για την επιστήμη των δεδομένων, καλύπτοντας κάθε στάδιο της ανάλυσης των δεδομένων, όπως:

- **Numpy**

  Η βιβλιοθήκη Numpy είναι το θεμέλιο πάνω στο οποίο κατασκευάζονται όλα τα εργαλεία υψηλότερου επιπέδου για την " επιστημονική πληροφορική" με την Python. Η βιβλιοθήκη NumPy είναι η πρώτη ύλη που δίνει στις βιβλιοθήκες εξαιρετική απόδοση. Χωρίς τη βιβλιοθήκη NumPy, πολλοί από τους υπολογισμούς στην επιστήμη δεδομένων θα ήταν ανέφικτο να πραγματοποιηθούν στην Python. Εκτός από τις προφανείς επιστημονικές του χρήσεις, το NumPy μπορεί επίσης να χρησιμοποιηθεί ως αποτελεσματικό πολυδιάστατο δοχείο γενικών δεδομένων και έτσι επιτρέπει στην NumPy να ενσωματώνεται χωρίς προβλήματα και ταχύτητα με μια μεγάλη ποικιλία βάσεων δεδομένων. Περιέχει μεταξύ άλλων λειτουργίες όπως ένα ισχυρό αντικείμενο N-διαστάσεων πίνακα, εξελιγμένες μεταδιδόμενες λειτουργίες και εργαλεία για την ενσωμάτωση του C / C ++ και του κώδικα Fortran.

- **Pandas**

  Μια πολύ χρήσιμη και διαδεδομένη βιβλιοθήκη για την διαδικασία της μηχανικής μάθησης είναι η Pandas, και είναι χτισμένη πάνω στη βιβλιοθήκη NumPy, μία από τις πρώτες βιβλιοθήκες πίσω από την ιστορία επιτυχίας της επιστήμης δεδομένων της Python. Οι λειτουργίες της βιβλιοθήκης NumPy εκτίθενται στη βιβλιοθήκη Pandas για προηγμένη αριθμητική ανάλυση. Είναι βιβλιοθήκη ανάλυσης δεδομένων, που χρησιμοποιείται για την εισαγωγή δεδομένων από υπολογιστικά φύλλα του Excel σε σύνολα επεξεργασίας για ανάλυση χρονοσειρών. Το Pandas τοποθετεί σχεδόν όλα τα κοινά εργαλεία για την ανάμειξη δεδομένων δίνοντας έτσι την δυνατότητα να γίνει η βασική εκκαθάριση και κάποια προηγμένη χειραγώγηση στα δεδομένων της Pandas. Επειδή η παροχή “καθαρών” δεδομένων με τη σωστή μορφή είναι ζωτικής σημασίας για την πραγματοποίηση σωστών προβλέψεων στη μηχανική μάθηση, και σίγουρα δεν πρέπει να παραβλεφθεί, η βιβλιοθήκη Pandas επιτρέπει να μορφοποιηθούν αποτελεσματικά τα δεδομένα, ακριβώς με τη μορφή που τα χρειαζονται, αποφεύγοντας την παγίδα να συμπεριληφθούν προβληματικά ή ελλειπή δεδομένα στις αναλύσεις.

- **Scikit-learn**

  Η βιβλιοθήκη Scikit-learn είναι βιβλιοθήκη μηχανικής μάθησης που παρέχει ενότητες για την οικοδόμηση νευρωνικών δικτύων και προεπεξεργασίας δεδομένων και βασίζεται στη NumPy και τη SciPy. Η βιβλιοθήκη Scikit-learn περιλαμβάνει εργαλεία για πολλές τυπικές εργασίες εκμάθησης μηχανών και εξόρυξης δεδομένων όπως η ομαδοποίηση, ταξινόμηση και η παλινδρόμηση. Η βιβλιοθήκη Scikit-learn παρέχει επίσης εργαλεία για την επικύρωση των αποτελεσμάτων και τη διασφάλιση της βέλτιστης επιλογής των μοντέλων.

- **Scipy**

  Η βιβλιοθήκη SciPy εξαρτάται από τη NumPy, η οποία παρέχει εύκολο και γρήγορο χειρισμό πίνακα N-διαστάσεων. Η βιβλιοθήκη SciPy είναι κατασκευασμένη για να λειτουργεί με τους αριθμητικούς πίνακες NumPy και παρέχει πολλές φιλικές προς το χρήστη και αποτελεσματικές αριθμητικές ρουτίνες, για αριθμητική ενσωμάτωση, βελτιστοποίηση κ.α. Η SciPy διαθέτει ενότητες βελτιστοποίησης, γραμμικής άλγεβρας, ολοκλήρωσης και άλλων κοινών καθηκόντων στην επιστήμη των δεδομένων.

5.3 ΤΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΆ ΤΗΣ PYTHON

5.3.1 Εγκατάσταση της Python

Είναι δυνατό να γίνει προγραμματισμός με την Python στον υπολογιστή αφού πρώτα εγκατασταθεί η Python και διαφορετικά πακέτα (βιβλιοθήκες) Python στον υπολογιστή. Παρόλο που είναι δυνατή η εγκατάσταση της Python , συνιστάται ιδιαίτερα η χρήση της Anaconda, η οποία είναι μια διανομή ανοιχτού κώδικα των γλωσσών προγραμματισμού Python και R για επεξεργασία δεδομένων μεγάλης κλίμακας, προβλέψιμες αναλύσεις και επιστημονική πληροφορική με στόχο την απλοποίηση της διαχείρισης πακέτων και την ανάπτυξη κάνοντας έτσι πολύ πιο εύκολη την εγκατάσταση νέων εργαλείων στη Python. [Basic elements of Python]

5.3.2 Βασικά στοιχεία

- Αναγνωριστικά: το όνομα που προσδιορίζει μια μεταβλητή, μια συνάρτηση, μια κλάση ή ένα αντικείμενο.

- Λέξεις κλειδιά: είναι δεσμευμένες λέξεις που δεν μπορούν να χρησιμοποιηθούν ως αναγνωριστικό.
- Εσοχή: δεν περιέχεται εντολή κλεισίματος γι΄αυτό κάθε σύνολο εντολών υποδηλώνεται με εσοχή.
- Εντολή help: χρησιμοποιείται για την λειτουργία και σύνταξη κάποιας εντολής.
- Δήλωση πολλαπλών γραμμών: υπάρχει η δυνατότητα συνέχισης της εντολής με την χρήση του / όταν η εντολή συνεχίζεται σε παραπάνω από μια γραμμή.
- Σχόλια: για την προσθήκη σχολίων χρησιμοποιείται η δίεση (#).
- Εντολή print: εμφανίζει αποτελέσματα στην οθόνη.
- Εντολή input: χρησιμοποιείται για την εισαγωγή τιμής κατά την εκτέλεση του κώδικα από το πληκτρολόγιο.
- Πολλαπλές δηλώσεις σε μια γραμμή: γίνεται με την χρήση του ερωτηματικού (;).
- Πολλαπλές ομάδες δηλώσεων: αφού δηλωθούν λέξεις κλειδιά χρησιμοποιείται το σύμβολο της άνω κάτω τελείας (:) και συνεχίζονται οι δηλώσεις ώστε να συνθέσουν έναν κώδικα.
- Μεταβλητές: είναι ένα αναγνωριστικό που περιέχει ένα αντικείμενο και με τον ορισμό μιας μεταβλητής καταλαμβάνεται μνήμη. Υπάρχουν διάφοροι τύποι μεταβλητών με τους πιο συνηθισμένους τους ακέραιους (int) και πραγματικούς (str) αριθμούς και αληθής (is) και ψευδής (is not), και οι τύποι αυτοί μπορούν να μετατραπούν χρησιμοποιώντας το όνομα του τύπου ως συνάρτηση. Επίσης το εύρος των μεταβλητών αλλάζει, υπάρχουν οι τοπικές μεταβλητές που είναι προσβάσιμες μόνο σε τοπικό εύρος, δηλαδή στη συνάρτηση που ορίζεται και οι καθολικές που μπορούν να χρησιμοποιηθούν οπουδήποτε μέσα στον κώδικα αφού πρώτα δηλωθούν στον κύριο κώδικα.
- Σταθερές: οι ορισμένες σταθερές είναι η pi (π=3,14) και η e=2,718.
- Τελεστές: Υπάρχουν διάφοροι τελεστές όπως οι αριθμητικοί (+,-,/,*), οι συγκριτικοί (<,>,=,<=), σύνθετοι (α+=β⇔α=α+β) κ.α. Κάποιο τελεστές έχουν υψηλότερη προτεραιότητα από άλλους με τους πρώτους να είναι η ύψωση σε δύναμη, το συμπλήρωμα, ο πολλαπλασιασμός, διαίρεση και πρόσθεση, αφαίρεση.

5.3.3 Δομές ελέγχου και επανάληψης

Τρείς βασικές δομές ελέγχου ροής προγράμματος είναι η δομή ακολουθίας εντολών, η δομή απόφασης και η δομή επανάληψης.

Στη δομή ακολουθίας εντολών οι εντολές εκτελούνται σειριακά, η μια μετά την άλλη. Για κάθε εντολή που εκτελείται ακολουθεί η εκτέλεση της επόμενης και κάθε εντολή εκτελείται ακριβώς μια φορά, χωρίς να επαναληφθεί καμία από αυτές τις εντολές ή να ξαναεκτελεστεί κάποια.

Στη δομή προγράμματος απόφασης ο έλεγχος του προγράμματος καλείται να επιλέξει ανάμεσα σε δύο ή περισσότερες διαφορετικές διαδρομές ανάλογα με το αν ισχύει ή όχι κάποια συνθήκη. Η πιο συνηθισμένη δομή απόφασης στη Python είναι η if-elif-else. Με το που θα περάσει ο έλεγχος στη δομή if θα ελεγχθεί η συνθήκη και θα αποτιμηθεί σε αληθής ή ψευδής. Αν η συνθήκη της if είναι αληθής εκτελείται η εντολή, ενώ αν είναι ψευδής μπορεί να χρησιμοποιηθεί ένας άλλος έλεγχος συνθήκης της δομής elif. Ελέγχονται όλες οι συνθήκες της δομής elif αυτές μέχρι να βρεθεί μια αληθής, και εάν καμία συνθήκη elif δεν είναι αληθής τότε εκτελούνται οι εντολές της δομής else.

Οι δομές επανάληψης χρησιμοποιούνται όταν κάποιες εργασίες πρέπει να γίνουν περισσότερες από μια φορές. Δύο χαρακτηριστικές δομές επανάληψης είναι η For και η while. Οι δομές που υλοποιούν κάποιας μορφής επανάληψης λέγονται βρόγχοι και μέσα σε έναν βρόγχο μπορούν να υπάρχουν ένας οι περισσότεροι βρόγχοι. Η δομή for ορίζει μια ακολουθία εντολών που επαναλαμβάνεται όσες φορές καθορίζεται από τις παραμέτρους της εντολής. Τη δομή for τη χαρακτηρίζει μια μεταβλητή η οποία έχει κάποια τιμή και σε κάθε επανάληψη αυξάνει τη τιμή της σταθερά όσο δηλώνεται σε ένα προαιρετικό βήμα της εντολής που ονομάζεται βήμα (αν δεν δηλωθεί αυξάνεται ίσο με 1). Η δομή while μοιάζει πολύ με την δομή for. Η δομή while εκτελεί μια κάποιες εντολές για όσο διάστημα μια συνθήκη είναι αληθής. Αν η συνθήκη ισχύει, τότε εκτελούνται οι εντολές και μετά ο έλεγχος επαναφέρεται στην αρχή της δομής while όπου και ξανά ελέγχεται η συνθήκη. Έτσι οι εντολές εκτελούνται για όσο καιρό η συνθήκη ισχύει. Οτιδήποτε μπορεί να γίνει με την δομή for μπορεί να γίνει και με την δομή while, αρκεί στη δομή for να αρχικοποιηθεί η μεταβλητή του βρόγχου και να αυξάνεται το βήμα.

5.3.4 Συναρτήσεις

Στην Python μια συνάρτηση ορίζεται με την λέξη κλειδί def και έπειτα ακολουθεί το όνομα της συνάρτησης ώστε να χρησιμοποιηθεί μέσα στον κώδικα σε οποιοδήποτε σημείο και πάνω από μια φορές. Έπειτα μέσα σε δυο παρενθέσεις δηλώνονται οι παράμετροι της συνάρτησης. Η συνάρτηση τελειώνει με την εντολή return η οποία καθορίζει ποια είναι η τιμή της συνάρτησης, έτσι όταν βρεθεί η εντολή return ο έλεγχος επιστρέφει στο κύριο

πρόγραμμα. Εκτός από τις συναρτήσεις που ορίζονται με την εντολή def υπάρχουν και οι ενσωματωμένες συναρτήσεις οι οποίες παρέχονται με τα πακέτα όπως π.χ. το πακέτο math με τις συναρτήσεις math.sin(x), math.cos(x) κ.α.

5.3.5 Βασικές δομές

Οι βασικές δομές της Python είναι αλφαριθμητικά (strings), οι λίστες (list), τα σύνολα (sets), οι πλειάδες (tuples) και το λεξικό (dictionary). Τα αλφαριθμητικά είναι μια ακολουθία χαρακτήρων σε αριθμημένη σειρά και δεν μπορεί να αλλάξει η τιμή τους, είναι αμετάβλητα, αλλά μπορούν να δημιουργηθούν καινούργια. Για να οριστούν θα πρέπει να βρίσκονται μέσα σε μονούς ή διπλούς αποστρόφους. Η μορφοποίηση ενός αλφαριθμητικού μπορεί να γίνει μέσω συναρτήσεων όπως π.χ. mystring.find('a') που βρίσκει την θέση του a μέσα στο αλφαριθμητικό, mystring.rjust(x) που κάνει δεξιά στοίχιση χρησιμοποιώντας x χαρακτήρες κ.α.

Οι λίστες είναι μια δομή δεδομένων η οποία περιέχει σε μια συγκεκριμένη σειρά μια συλλογή τιμών και η ίδια τιμή μπορεί να υπάρχει περισσότερες από μια φορές. Οι λίστες ορίζονται με αγκύλες και τα στοιχεία της χωρίζονται με κόμμα. Όπως και τα αλφαριθμητικά, οι λίστες μπορούν να μορφοποιηθούν μέσω συναρτήσεων όπως cmp(mylist1,mylist2) που συγκρίνονται οι δύο λίστες, max(mylist) ή min(mylist) που επιστρέφει το μέγιστο και το ελάχιστο της λίστας αντίστοιχα κ.α.

Τα σύνολα διευκολύνουν την ομαδοποίηση πολλών αντικειμένων ώστε να εφαρμοστούν στη συνέχεια πράξεις με αποδοτικό τρόπο εξασφαλίζοντας πως κάθε στοιχείο αν περιέχεται σε πάνω από ένα σύνολα θα βρεθεί μόνο μια φορά στο τελικό αποτέλεσμα. Τα σύνολα ορίζονται με άγκιστρο ή με την συνάρτηση set που περικλείεται από αγκύλες και παρενθέσεις. Κάποιες από τις λειτουργίες των συνόλων είναι η len(myset) η οποία εμφανίζει το μέγεθος του συνόλου, το myset.add("x") που χρησιμοποιείται για την προσθήκη του στοιχείου x σε ένα σύνολο κ.α.

Η πλειάδα είναι ένα στιγμιότυπο της λίστας μόνο για ανάγνωση, περιέχει αντικείμενα που σχετίζονται μεταξύ τους και δεν μπορεί να αλλάξει ούτε μέγεθος ούτε στοιχεία όπως γίνεται στις λίστες. Άλλη μια διαφορά που διακρίνει τις πλειάδες από τις λίστες είναι ότι οι πλειάδες ορίζονται με παρενθέσεις ενώ οι λίστες από αγκύλες. Όπως και στις λίστες, οι πλειάδες υποστηρίζουν πολλές συναρτήσεις όπως cmp(t1,t2), max(t1) ή min(t1) κ.α.

Τέλος, τα λεξικά είναι δομές που μπορούν να περιέχουν πολλούς τίμους δεδομένων και μια λίστα αντιστοιχεί σε λέξεις κλειδιά κάποιες τιμές. Τα λεξικά είναι μη ταξινομημένα, επαναληπτικά και μεταβλητά και ακόμη και όταν έχει δημιουργηθεί ένα λεξικό μπορεί εύκολα να προστεθούν τιμές σε αυτό. Το λεξικό ορίζεται με άγκιστρα και οι τιμές με αγκύλες ή με την συνάρτηση dict. Υπάρχουν διάφορες λειτουργίες του λεξικού όπως mydict.clear() που διαγράφει όλα τα στοιχεία του λεξικού, mydict.keys() η οποία εμφανίζει τα κλειδιά του λεξικού κ.α.

5.3.6 Αρχείο

Τα αρχεία αποθηκεύουν και ανακτούν πληροφορίες σε οποιοδήποτε πρόγραμμα τρέχει. Για να ανοίξει ένα αρχείο και να εκτελεστεί οποιαδήποτε λειτουργία πάνω σε αυτό χρησιμοποιείται η συνάρτηση open με όρισμα το όνομα του αρχείου. Υπάρχουν διάφορες λειτουργίες που μπορούν να εφαρμοστούν πάνω στο αρχείο και υλοποιούνται μέσω της open όπως το “r” για ανάγνωση, το “a” για προσθήκη κ.α. Η εντολή για διάβασμα από αρχείο είναι η read και για εγγραφή χρησιμοποιείται η write, ενώ για το κλείσιμο ενός αρχείου χρησιμοποιείται η εντολή f.close.

5.4 Η ΕΠΙΣΤΉΜΗ ΤΩΝ ΔΕΔΟΜΈΝΩΝ ΜΕ ΤΗΝ PYTHON

Για τον λόγο ότι η Python υποστηρίζει αντικειμενοστραφή, δομημένο και με λειτουργικά πρότυπα προγραμματισμό, την κάνει μια από πιο δημοφιλείς γλώσσες προγραμματισμού και χρησιμοποιείται συχνά για την εξόρυξη δεδομένων στην επιστήμη των δεδομένων. Μεγάλες εταιρείες όπως η Google,η Nasa και το Cern χρησιμοποιούν την Python για κάθε σκοπό προγραμματισμού, συμπεριλαμβανομένης όλο και περισσότερο της μηχανικής μάθησης. [5 Huge Tech Companies That Use Python]

Τρείς διαδεδομένοι τομείς εφαρμογής της μηχανικής μάθησης είναι η χρηματοδότηση, η υγειονομική περίθαλψη και η λιανικές πωλήσεις. Και στους τρείς τομείς μπορεί να χρησιμοποιηθεί η Python για την εφαρμογή μηχανικής μάθησης. Αν και κάθε επιχείρηση του κάθε τομέα είναι διαφορετική και οι λύσεις τους είναι διαφορετικές, η προσέγγιση της μηχανικής μάθησης δεν διαφέρει πολύ. [Puneet Mathur, 2019]

5.4.1 Τομέας χρηματοοικονομικής

Στον τομέα της χρηματοδότησης εφαρμόζεται η μηχανική μάθηση για την ενίσχυση των κανονισμών και την επιβολή των νόμων αφού αποτελεί την ραχοκοκαλιά της οικονομίας. Κάποιες από τις βασικές κατηγορίες της χρηματοδότησης που η μηχανική μάθηση έχει

εφαρμογή είναι οι επενδύσεις στο χρηματιστήριο, η τραπεζική, η λογιστική και οι υπηρεσίες ηλεκτρονικών πληρωμών.

Οι επενδύσεις στην χρηματιστηριακή αγορά καθίστανται ολοένα και πιο αυτοματοποιημένες γι΄αυτό τον λόγο οι εταιρίες αναβαθμίζουν τις διαδικασίες για καλύτερη ταχύτητα. Έχοντας γρήγορες ταχύτητες έχουν μεγαλύτερο όφελος σε μια συναλλαγή αφού μπορούν να προλάβουν να επενδύσουν σε μια καλή συναλλαγή από κάποια εταιρεία που έχει χαμηλές ταχύτητες. Δεδομένου ότι τα συστήματα συνεχώς βελτιώνονται για ταχύτητα και αξιοπιστία, οι μέθοδοι μηχανικής μάθησης είναι εξαιρετικά βασικές για πετύχουν τον στόχο τους.

Η τραπεζική είναι ένας τομέας όπου η χρήση διαδικασιών μηχανικής μάθησης εφαρμόζεται σε πλούσια ενημερωμένες βάσεις δεδομένων και μπορεί να βοηθήσει στη μάχη με οικονομικές απάτες, να εξοικονομήσει χρόνο και μετρητά για πελάτες και να μηχανογραφεί τις διοικητικές ικανότητες. Μια από τις βασικές χρήσεις της μηχανικής μάθησης στη διαχείριση ενός λογαριασμού πελατών στον τραπεζικό κόσμο είναι η προσπάθεια καταπολέμησης της απάτης και της πειρατείας, όπως και η ενίσχυση της συνέπειας. Οι υπολογισμοί της μηχανικής μάθησης μπορούν να γίνουν από αναζητήσεις σε τεράστιες βάσεις δεδομένων και συλλογές πληροφοριών ώστε να ανιχνευθούν ασυνήθιστες οικονομικές συναλλαγές.

Η λογιστική και ο λογιστικός προγραμματισμός γίνονται ολοένα και πιο "έξυπνα" αφού εκτελούνται πολλές εργασίες μέσω προγραμμάτων που πριν απαιτούσαν την ανθρώπινη διαμεσολάβηση. Πλέον οι εταιρείες επενδύουν λιγότερη ενέργεια στις χειρονακτικές εργασίες και έχουν περισσότερο χρόνο για άλλα βασικά καθήκοντα. Για παράδειγμα, οι εφαρμογές λογιστικής καταγράφουν τις πρακτικές κωδικοποίησης των εισπράξεων και προτείνουν την κατανομή των συναλλαγών. Επίσης, έχει εφαρμογή και σε απευθείας συνδέσεις ελέγχου σε πραγματικό χρόνο ώστε οι ελεγκτές να μην χάνουν χρόνο για την διεκπεραίωση του ελέγχου σε ένα συγκεκριμένο χρονικό πλαίσιο αλλά και χρήμα γιατί χρειάζεται μια μεγάλη ομάδα να κάνει τον έλεγχο. Πλέον κατά την εφαρμογή της μηχανικής μάθησης επεξεργάζονται και ερευνώνται πληροφορίες, διακρίνονται οι ιδιαιτερότητες και συγκεντρώνονται μια σειρά εξαιρέσεις από τον έλεγχο. Για να διασφαλιστεί μια σωστή επένδυση στη χρηματιστηριακή αγορά μπορεί να γίνει έλεγχος για ψευδείς οικονομικές καταστάσεις των εταιρειών για επένδυση. Μέσω της μηχανικής

μάθησης ελέγχονται οι οικονομικές καταστάσεις, τα σημάδια στην αγορά, οι ειδήσεις για την εταιρεία, οι ειδήσεις για τους ιδρυτές της, σύγκριση με άλλες περιπτώσεις κ.α.

Οι εταιρείες ηλεκτρονικών πληρωμών παγκοσμίως κατέχουν την μηχανική μάθηση αφού κατά τον έλεγχο συναλλαγής καρτών η μηχανική μάθηση αναλαμβάνει τμήματα σε στενή και συνεχή έγκριση συναλλαγών. Μέσω της μηχανικής μάθησης αξιοποιούνται πληροφορίες που προέρχονται από τον ιστότοπο για ακριβέστερη πρόβλεψη παρατυπιών του συναλλασσομένου, τη χρήση εικονικών συνεργατών ή την ενίσχυση της εκτέλεσης των ωφελειών των πελατών.

5.4.1 Τομέας υγειονομικής περίθαλψης

Στον τομέα της υγειονομικής περίθαλψης η μηχανική μάθηση έχει εφαρμογή σε βασικές κατηγορίες της όπως η ταυτοποίηση ασθενών σε ψηφιακά αρχεία υγείας, στη διάγνωση ασθενειών, στην ακτινολογία και στην χειρουργική ρομποτική.

Η ταυτοποίηση ασθενών μέσω ψηφιακών αρχείων υγείας δίνει τεράστιες δυνατότητες στους γιατρούς καθώς ο αριθμός των ασθενών αυξάνεται, τα δεδομένα και το ιστορικό ανά ασθενή αυξάνονται. Έτσι, με την μηχανική μάθηση επιτρέπεται η ανάλυση δεδομένων με γρήγορο και αποδοτικό τρόπο. Όταν ο ασθενής βρεθεί σε κάποια μονάδα υγείας περίθαλψης, δίνοντας την εξουσιοδότησή του για τα προσωπικά δεδομένα, θα μπορεί η μονάδα υγείας να έχει πρόσβαση στα ψηφιακά αρχεία υγείας σε οποιοδήποτε τμήμα της. Η αντίστοιχη διαδικασία μπορεί να γίνει και σε διαγνωστικά κέντρα για τις εργαστηριακές εξετάσεις ασθενών.

Οποιοδήποτε σύστημα διάγνωσης ασθενειών που αναπτύσσεται χρησιμοποιώντας μηχανική μάθηση πρέπει να είναι το ίδιο αποδοτικό με ενός ανθρώπου γιατρού. Μέσω της μηχανικής μάθησης, μεγάλος όγκος ασθενειών, συμπτωμάτων, αναφορών εργαστηρίων και μελετών στο ημερολόγιο του γιατρού βοηθάνε τα λογισμικά προκειμένου να αποκτήσουν εμπειρία στη διάγνωση ασθενειών και να αυξήσουν την αξία της υγειονομικής περίθαλψης.

Η χρήση της μηχανικής μάθησης στην ακτινολογία γίνεται στην τμηματοποίηση της ιατρικής εικόνας, την εγγραφή, την ανίχνευση και την διάγνωση μέσω υπολογιστή της λειτουργίας του εγκεφάλου, την ανάλυση δραστηριότητας του εγκεφάλου, τη διάγνωση νευρολογικών ασθενειών κ.α. Ο πρωταρχικός σκοπός της μηχανικής μάθησης είναι να

βοηθήσει να ληφθούν σωστές αποφάσεις σε σχέση με τα δεδομένα ακτινολογίας όμως μπορεί να συνδυαστεί και με τα συστήματα ανίχνευσης και διάγνωσης ασθενειών.

Στη χειρουργική ρομποτική χρησιμοποιούνται χαρτογραφημένες πληροφορίες για την μηχανική μάθηση και την τεχνολογική πρόοδο της ρομποτικής. Έτσι η ρομποτική γίνεται βοηθός των χειρούργων και ρομπότ μαθαίνουν δεξιότητες για διάγνωση και χειρουργική επέμβαση ώστε να εκτελούν πολύπλοκες και με βάση την ακρίβεια χειρουργικές επεμβάσεις.

5.4.1 Τομέας λιανικών πωλήσεων

Το λιανικό εμπόριο επηρεάζει τη ζωή όλων σε αυτόν τον πλανήτη και η μηχανική μάθηση επηρεάζει τον τομέα αυτόν ολοένα και περισσότερο. Κάποιες κατηγορίες στον τομέα των λιανικών πωλήσεων που η μηχανική μάθηση έχει εφαρμογή είναι τα συστήματα διαχείρισης καταστημάτων, το ψηφιακό εμπόριο, την αλυσίδα εφοδιασμού και τη διαχείριση πελατών.

Τα συστήματα διαχείρισης καταστημάτων περιλαμβάνουν τη διαχείριση της εσωτερικής διαρρύθμισης των διαφόρων προϊόντων στα πατώματα των καταστημάτων. Επίσης, περιλαμβάνουν τη χρήση δεξιοτήτων για την εμφάνιση των πιο ελκυστικών προσφορών κοντά στις περιοχές που οι πελάτες συχνάζουν. Αυτές οι δραστηριότητες βασιζόταν στις δεξιότητες και στην εμπειρία του ανθρώπου, αλλά πλέον βασίζονται στην μηχανική μάθηση αφού οποιοδήποτε μεγάλο κατάστημα έχει πολλά διαφορετικά προϊόντα σε διαφορετικές κατηγορίες να εμφανίσει προς τους πελάτες.

Το ψηφιακό εμπόριο είναι ταχέως αναπτυσσόμενο και όλο και περισσότεροι πελάτες στρέφονται προς αυτό. Το ψηφιακό εμπόριο μπορεί να γίνει με φυσική παρουσία στο κατάστημα αφού γίνει η παραγγελία για την παραλαβή αλλά και χωρίς καμία φυσική παρουσία. Με την χρήση της μηχανικής μάθησης οι εταιρείες στις ιστοσελίδες τους δίνουν καλύτερες συστάσεις προϊόντων σε έναν πελάτη με βάση τα προηγούμενα πρότυπα αγοράς του.

Η διαχείριση της αλυσίδας εφοδιασμού είναι μια απαραίτητη λειτουργία των καταστημάτων που ο πελάτης δεν βλέπει άμεσα και περιλαμβάνει τη διατήρηση της σχέσης των προμηθευτών όλων των προϊόντων και την ενημέρωση με τις επερχόμενες απαιτήσεις. Με την μηχανική μάθηση η διαχείριση γίνεται πιο αποδοτική. Παίζει σημαντικό ρόλο να μην υπάρχουν ούτε υπερβάλλον ούτε υπολειπόμενα αποθέματα αφού

οδηγούν στην αύξηση του κόστους αποθήκευσης και στη δυσαρέσκεια των πελατών αντίστοιχα.

Η διαχείριση των πελατών είναι επίσης σημαντική ώστε να δημιουργούνται νέοι πελάτες. Μέσω της μηχανικής μάθησης παρακολουθείται η νέα ροή πελατών αλλά και παλαιών πελατών ώστε να εξαπλώνεται το εμπορικό σήμα. Έτσι οι διαχειριστές του καταστήματος έχουν καλύτερη και πιο γρήγορη εικόνα για την υγεία του καταστήματος, τόσο μέσω των παλαιών όσο και μέσω των νέων πελατών, ώστε αν χρειαστεί να εφαρμόσουν άλλες πρακτικές.

6. ΛΟΓΙΣΜΙΚΌ ΕΞΌΡΥΞΗΣ ΔΕΔΟΜΈΝΩΝ WEKA

6.1 ΤΟ ΠΡΌΓΡΑΜΜΑ WEKA

Το Weka (Waikato Environment for Knowledge Analysis) είναι ένα δημοφιλής λογισμικού μηχανικής μάθησης, ελεύθερο υπό την άδεια GNU General Public License, γραμμένο σε Java και αναπτύχθηκε στο Πανεπιστήμιο του Waikato της Νέα Ζηλανδία. Το Weka είναι ένα λογισμικό το οποίο περιέχει μια συλλογή από εργαλεία οπτικοποίησης και αλγορίθμους για την ανάλυση δεδομένων και την προγνωστική μοντελοποίηση, μαζί με γραφικές διεπαφές χρήστη για εύκολη πρόσβαση σε αυτές τις λειτουργίες. Η αρχική μη-Java έκδοση του Weka ήταν ένα Tcl /Tk front -end για μοντελοποίηση αλγορίθμων που εφαρμόζονται σε άλλες γλώσσες προγραμματισμού, περιέχοντας δυνατότητες προ-επεξεργασίας δεδομένων σε C, και ένα σύστημα βασισμένο σε Makefile για τη πραγματοποίηση πειραμάτων μηχανικής μάθησης. Αυτή η αρχική έκδοση είχε σχεδιαστεί ως ένα εργαλείο για την ανάλυση των δεδομένων από γεωργικούς τομείς, αλλά η πιο πρόσφατη πλήρης έκδοση βασισμένη σε Java (Weka 3), η ανάπτυξη της οποίας άρχισε το 1997, έχει πλέον πολλούς τομείς εφαρμογής, κυρίως εκπαιδευτικούς σκοπούς και έρευνες. [WEKA-The workbench for machine learning] [Wikipedia]

Το Weka υποστηρίζει διάφορες βασικές διεργασίες εξόρυξης δεδομένων και την δυνατότητα επιλογής τους όπως η προεπεξεργασία δεδομένων, η ομαδοποίηση, η ταξινόμηση, η παλινδρόμηση, η απεικόνιση. Όλες οι τεχνικές του Weka στηρίζονται στην υπόθεση ότι τα δεδομένα είναι διαθέσιμα ως ένα ενιαίο αρχείο ή σχέση, όπου κάθε σημείο δεδομένων περιγράφεται από ένα σταθερό αριθμό των χαρακτηριστικών (κανονικά, αριθμητικά ή ονομαστικά χαρακτηριστικά, αλλά και κάποιοι άλλοι τύποι χαρακτηριστικών υποστηρίζονται επίσης). Το Weka παρέχει πρόσβαση σε SQL βάσεις δεδομένων χρησιμοποιώντας Java Database Connectivity και μπορεί να επεξεργαστεί το αποτέλεσμα που επιστρέφονται από ένα ερώτημα βάσης δεδομένων. [Weka tutorial]

Το Weka υποστηρίζει αρχεία Attribute-Relation File Format (.arff), μορφή αρχείου συσχετίσεων, όπου ένα αρχείο κειμένου ASCII περιγράφει μια λίστα εμφανίσεων που μοιράζονται ένα σύνολο χαρακτηριστικών. Τα αρχεία .arff έχουν δυο διαφορετικές ενότητες. Η πρώτη ενότητα είναι οι πληροφορίες κεφαλίδας που ακολουθούν τις πληροφορίες δεδομένων. Η επικεφαλίδα του αρχείου περιέχει το όνομα της σχέσης, μια λίστα με τα χαρακτηριστικά και τους τύπους τους. Η δεύτερη ενότητα είναι τα δεδομένα

όπου στην αρχή τους έχουν το σύμβολο @data και ακριβώς μετά ακολουθούν οι τιμές ανάλογα με την μεταβλητή.

```
@relation ARITHMODEIKTES_WEKA3

@attribute 'NET PROFIT/FIXED ASSETS' numeric
@attribute 'TOTAL LIABILITIES/TOTAL ASSETS' numeric
@attribute 'LONG TERM DEBT/TOTAL ASSETS' numeric
@attribute 'TOTAL DEBT/EQUITY' numeric
@attribute INVENTORIES/SALES numeric
@attribute SALES/EQUITY numeric
@attribute 'INVENTORIES/TOTAL ASSETS' numeric
@attribute 'INVENTORIES/CURRENT ASSETS' numeric

@data
-0.012063,0.622912,0.373215,1.36388,0.039338,1.353445,0.020077,0.
109796
-0.171016,0.761688,0.462471,2.540524,0.03547,2.304738,0.019482,0.
104622
-0.062076,0.75697,0.519944,2.461811,0.032304,2.275015,0.017861,0.
113007
-0.186702,0.881863,0.002301,5.948783,0.031491,4.306075,0.01602,0.
10124
```

Σχήμα 6.1: Μορφή αρχείου .arff

Στα πλεονεκτήματα του Weka περιλαμβάνονται η δωρεάν διαθεσιμότητα υπό την GNU Γενική Άδεια Δημόσιας χρήσης, η φορητότητα, δεδομένου ότι έχει υλοποιηθεί πλήρως στην γλώσσα προγραμματισμού Java και έτσι τρέχει σε σχεδόν κάθε σύγχρονη υπολογιστική πλατφόρμα, η ολοκληρωμένη συλλογή δεδομένων προεπεξεργασίας και τεχνικές μοντελοποίησης και τέλος η ευκολία στη χρήση λόγω των γραφικών διεπαφών χρήστη. Βέβαια δεν έχει μόνο πλεονεκτήματα, κάποια από τα μειονεκτήματα του Weka είναι ότι δεν είναι ικανό για εξόρυξη από πολύ-σχεσιακές βάσεις δεδομένων, αλλά υπάρχει ξεχωριστό λογισμικό για τη μετατροπή μιας συλλογής συνδεδεμένων πινάκων της βάσης δεδομένων σε έναν πίνακα που είναι κατάλληλος για επεξεργασία χρησιμοποιώντας το Weka και δεν καλύπτεται προς το παρόν από τους αλγορίθμους που περιλαμβάνονται στο Weka είναι η μοντελοποίηση αλληλουχιών.

6.2 ΤΟ ΜΕΝΟΥ ΤΟΥ WEKA

Το Weka διατίθεται για κάθε λογισμικό (Windows, Mac, Linux) και υπάρχει και η επιλογή Java Virtual Machine που δίνει την επιλογή στον χρήστη να γράψει κώδικα και να ενσωματώσει το Weka στις δικές του ανάγκες.

Στο αρχικό μενού του Weka υπάρχουν πέντε επιλογές:

- Explorer, είναι το γραφικό περιβάλλον που χρησιμοποιείται για την επεξεργασία δεδομένων τα οποία δεν έχουν υποστεί κάποια άλλη επεξεργασία.
- Experimenter, είναι ένα περιβάλλον όπου πραγματοποιούνται εργασίες που έχουν να κάνουν με την στατιστική.
- Knowledge Flow, η χρήση του είναι ίδια με του explorer με την διαφορά ότι υπάρχει η δυνατότητες να μεταφερθούν άλλα δεδομένα από άλλα αρχεία με μια κίνηση του ποντικιού και επίσης υποστηρίζει incremental learning όπου συνδέει διάφορες τεχνικές.
- Workbench, με την επιλογή του αλλάζει το περιβάλλον εργασίας και μέσα από την ίδια οθόνη υπάρχει η δυνατότητα χρήσης όλων των υπόλοιπων επιλογών.
- Simple CLI, παρέχει στους χρήστες την δυνατότητα να χρησιμοποιήσουν το γραφικό περιβάλλον μέσα από την γραμμή εντολών.

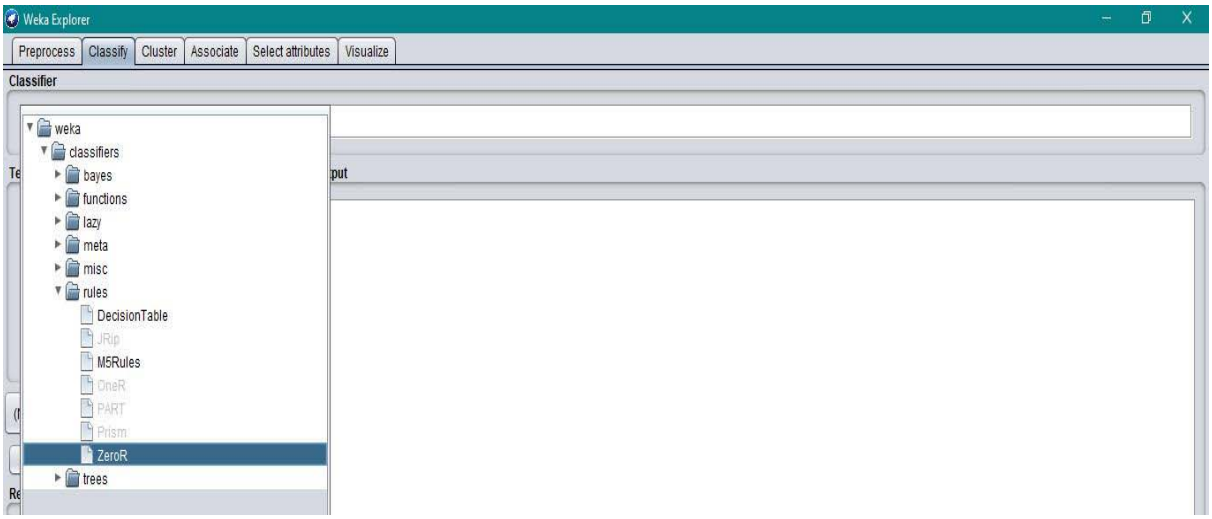
Σχήμα 6.2: Αρχικό μενού WEKA

Επίσης, στην αρχική οθόνη υπάρχουν οι επιλογές Program, Visualization, Tools και Help όπου αφορούν τις ρυθμίσεις του λογισμικού ώστε να προσαρμοστεί στις ανάγκες του κάθε χρήστη.

6.3 ΔΥΝΑΤΌΤΗΤΕΣ ΤΟΥ WEKA

Ο explorer έχει σχεδιαστεί για την επεξεργασία δεδομένων βασισμένη σε παρτίδες όπου τα δεδομένα εκπαίδευσης φορτώνονται στην μνήμη και στην συνέχεια επεξεργάζονται. Στην αρχική σελίδα του explorer υπάρχουν οι εξής επιλογές:

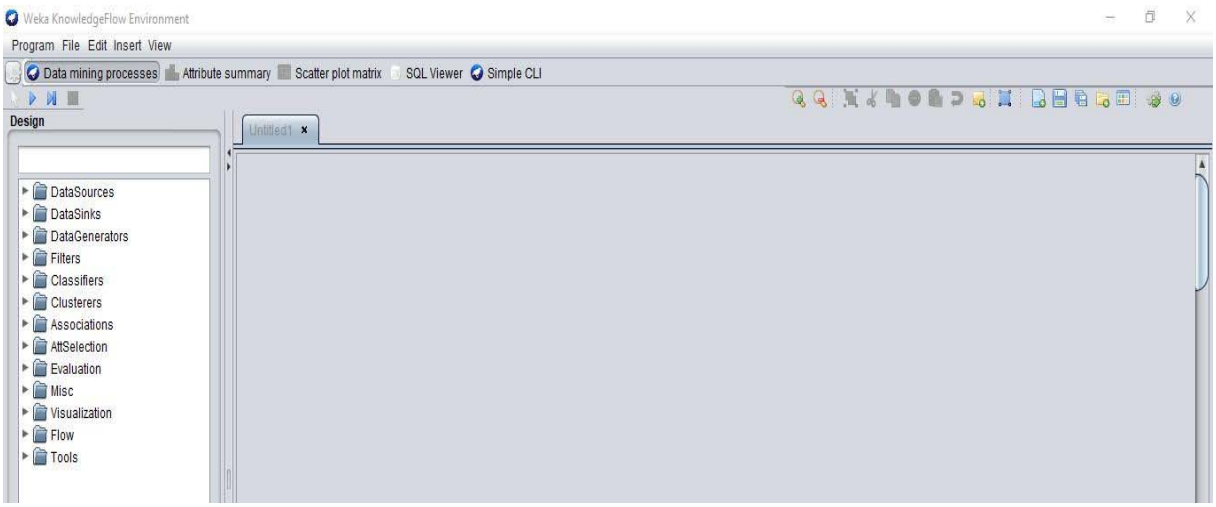
- Preprocess που επιτρέπει να επιλεχθεί το αρχείο που θα χρησιμοποιηθεί στο πρόγραμμα είτε είναι αρχείο στον υπολογιστή, είτε είναι αρχείο σε διαφορετική τοποθεσία, είτε είναι βάση δεδομένων από την οποία θα αντληθούν δεδομένα.
- Classify όπου εφαρμόζονται στα δεδομένα διάφοροι αλγόριθμοι κατηγοριοποίησης ώστε να συγκεντρωθούν πληροφορίες και ρυθμίζονται και οι παράμετροι που θα γίνει το τεστ όπως π.χ. να χρησιμοποιηθεί εκπαιδευτικό σετ για να εξάγει αποτελέσματα.
- Cluster που γίνεται κατάταξη με την χρήση αλγορίθμων των παρατηρήσεων σε ομάδες χρησιμοποιώντας πληροφορίες που υπάρχουν ώστε οι ομάδες που δημιουργούνται να έχουν όμοιες παρατηρήσεις.
- Associate όπου χρησιμοποιούνται αλγόριθμοι για να βρει σχέσεις ανάμεσα στα δεδομένα αλλά οι μόνες τιμές που αναγνωρίζει είναι ονομαστικές.
- Select attributes που δίνεται η επιλογή να ψαχθούν όλοι οι δυνατοί συνδυασμοί των γνωρισμάτων στο αρχείο ώστε να βρεθεί το καλύτερο για να γίνει κάποια πρόγνωση.
- Visualization που παρέχει την δυνατότητα να απεικονιστούν τα διαγράμματα σε 1-d, 2-d και 3-d με την δυνατότητα επιλογής της μεταβλητής στον άξονα x και y.



Σχήμα 6.3.1: Μενού Explorer

Η επιλογή experimenter δίνει την δυνατότητα να τρέξει, να δημιουργηθεί και να αναλυθεί ένα πείραμα με έναν βολικό τρόπο ώστε να βρεθεί αυτό που είναι πιο στατιστικά σημαντικό. Το πείραμα αποθηκεύεται σε .arff μορφή ώστε αν χρειαστεί να ξαναχρησιμοποιηθεί. Στον experimenter υπάρχει η δυνατότητα επιλογής για ομαδοποίηση ή παλινδρόμηση ανάλογα με τον αλγόριθμο ή τα αρχεία που χρησιμοποιούνται. Τέλος, αφού επιλεχθούν οι τεχνικές, τα αρχεία και οι αλγόριθμοι που θα χρησιμοποιηθούν γίνεται ανάλυση του αποτελέσματος. Συγκρίνονται οι αλγόριθμοι που έχουν επιλεγεί μεταξύ τους και για τους στατιστικά σημαντικούς εμφανίζεται το (v) και για τους όχι στατιστικά σημαντικούς το (*).

Το knowledge flow είναι μια διαφορετική έκδοση του explorer αλλά δεν υποστηρίζονται όλες οι επιλογές του όμως υποστηρίζει επιλογές που δεν υποστηρίζει ο explorer. Στο knowledge flow υπάρχει η επιλογή για αρχεία σε παρτίδες ή σταδιακά και γι΄αυτό υπάρχουν ειδικοί αλγόριθμοι για την επεξεργασία.



Σχήμα 6.3.2: Μενού Knowledge Flow

6.4 ΕΠΕΚΤΆΣΕΙΣ ΤΟΥ WEKA

Το λογισμικό Weka επιτρέπει την εύκολη ενσωμάτωση σε άλλες εφαρμογές Java καθώς και την επέκταση με νέα χαρακτηριστικά που μπορεί να είναι είτε πρόσθετοι αλγόριθμοι μηχανικής μάθησης και εργαλεία για οπτικοποίηση δεδομένων είτε επεκτάσεις του γραφικού περιβάλλοντος χρήστη. Εκτός από τις επεκτάσεις που υποστηρίζει το Weka μπορούν να χρησιμοποιηθούν και άλλα εξωτερικά εργαλεία προγραμματισμού όπως και βιβλιοθήκες τρίτων. Σκοπός αυτών των εργαλείων είναι η αυτοματοποίηση πολλών κοινών εργαλείων προγραμματισμού.

7.ΕΞΌΡΥΞΗ ΔΕΔΟΜΈΝΩΝ ΓΙΑ ΑΝΊΧΝΕΥΣΗ ΨΕΥΔΏΝ ΟΙΚΟΝΟΜΙΚΏΝ ΚΑΤΑΣΤΆΣΕΩΝ

7.1 ΕΙΣΑΓΩΓΉ ΔΕΔΟΜΈΝΩΝ

Το δείγμα που χρησιμοποιείται περιέχει 159 οικονομικές καταστάσεις, από 19 εταιρείες που το 94,74% είναι ή ήταν εισηγμένες στο Χρηματιστήριο Αθηνών. Οι 32 οικονομικές καταστάσεις είχαν ενδείξεις ή αποδείξεις ότι ήταν ψευδείς. Η κατηγοριοποίηση μιας οικονομικής κατάστασης ως ψευδείς βασίστηκε στις εξής παραμέτρους:

- Στην έκθεση ελέγχου, όπου η γνώμη του ελεγκτή διαφοροποιείται άρα έχει σοβαρές αμφιβολίες ως προς την ακρίβεια των οικονομικών καταστάσεων,
- Στην έκθεση ελέγχου, όπου υπάρχει έμφαση θέματος ή έγινε και αναμόρφωση κονδυλίων χωρίς να επιβάλλεται από κάποιο πρότυπο,
- Στον έλεγχο των φορολογικών αρχών που τροποποίησαν τον Ισολογισμό και την κατάσταση αποτελεσμάτων χρήσης,
- Στην εφαρμογή της νομοθεσίας που αφορά την αρνητική καθαρή θέση,
- Την συμπερίληψη της εταιρείας στις υπό επιτήρηση κατηγορίες μετοχών,
- Στην αναστολή διαπραγμάτευσης στο χρηματιστήριο λόγω ψευδών οικονομικών καταστάσεων ή σοβαρών φορολογικών παραβάσεων,
- Στην ύπαρξη εκκρεμών δικαστικών διαδικασιών για ψευδείς οικονομικές καταστάσεις ή σοβαρές φορολογικές παραβάσεις.

Αρχικά βρέθηκαν 16 λογαριασμοί του Ισολογισμού και της κατάστασης αποτελεσμάτων χρήσης και έπειτα υπολογίστηκαν και 32 αριθμοδείκτες. [Kotsiantis S., Tselepis D., Tampakas V., 2005] [Kirkos E., Spathis Ch., Manolopoulos Y., 2007 a] [Kirkos E., Spathis Ch., Manolopoulos Y., 2007 b] [Dalnial H., at al, 2014] [Kanapickiene R., Grudiene Z., 2015] Προκειμένου να μειωθεί η διάσταση των 48 χρηματοοικονομικών δεικτών εφαρμόζεται λογιστική παλινδρόμηση (logistic regression) ώστε να βρεθούν ποιοί από τους 48 χρηματοοικονομικούς δείκτες είναι στατιστικά σημαντικοί άρα επηρεάζουν το αν μια οικονομική κατάσταση είναι ψευδής ή όχι.

7.2 ΠΡΟ-ΕΠΕΞΕΡΓΑΣΊΑ ΔΕΔΟΜΈΝΩΝ

Πριν από την λογιστική παλινδρόμηση, είναι απαραίτητο να επαληθευτεί αν τα δεδομένα προέρχονται από κανονικά κατανεμημένο πληθυσμό. Η κύρια δοκιμασία για την ανάλυση

της κανονικότητας είναι το Kolmogorov-Smirnov τεστ. [Kanapickiene R., Grudiene Z., 2015]

Τόσο για την λογιστική παλινδρόμηση όσο και για το Kolmogorov-Smirnov τεστ χρησιμοποιείται το SPSS και επιλέχθηκε ένα δείγμα 61 οικονομικών καταστάσεων, με 27 ψευδείς οικονομικές καταστάσεις και 34 αληθείς χωρίς να εγγυάται ότι είναι δεν είναι ψευδείς ή ότι δεν έχουν παραβατική συμπεριφορά ως προς τις οικονομικές καταστάσεις. [Downing D., Clark J., 2010]

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
EBIT	,257	61	,000	,840	61	,000
EBT	,366	61	,000	,471	61	,000
SALES	,267	61	,000	,584	61	,000
NET PROFIT	,352	61	,000	,473	61	,000
GROSS PROFIT	,309	61	,000	,727	61	,000
TOTAL ASSETS	,302	61	,000	,569	61	,000
FIXED ASSETS	,417	61	,000	,501	61	,000
CURRENT ASSETS	,264	61	,000	,740	61	,000
EQUITY	,309	61	,000	,603	61	,000
CURRENT LIABILITIES	,276	61	,000	,648	61	,000
TOTAL LIABILITIES	,346	61	,000	,526	61	,000
INVENTORIES	,313	61	,000	,625	61	,000
WORKING CAPITAL	,272	61	,000	,818	61	,000

TOTAL DEBT	,310	61	,000	,610	61	,000
LONG TERM DEBT	,316	61	,000	,514	61	,000
ACCOUNTS RECIEVABLE	,240	61	,000	,679	61	,000
EBIT/SALES	,484	61	,000	,126	61	,000
EBT/SALES	,476	61	,000	,129	61	,000
NET PROFIT/SALES	,486	61	,000	,124	61	,000
NET PROFIT/GROSS PROFIT	,411	61	,000	,185	61	,000
EBIT/TOTAL ASSETS	,137	61	,006	,953	61	,021
EBT/TOTAL ASSETS	,204	61	,000	,834	61	,000
NET PROFIT/TOTAL ASSETS	,165	61	,000	,829	61	,000
EBT/FIXED ASSETS	,181	61	,000	,875	61	,000
NET PROFIT/FIXED ASSETS	,177	61	,000	,899	61	,000
EBT/EQUITY	,294	61	,000	,601	61	,000
NET PROFIT/EQUITY	,306	61	,000	,626	61	,000
CURRENT ASSETS/CURRENT LIABILITIES	,234	61	,000	,789	61	,000
(CURRENT ASSETS-INVENTORIES)/CURRENT LIABILITIES	,261	61	,000	,834	61	,000
WORKING CAPITAL/TOTAL ASSETS	,132	61	,010	,946	61	,009

TOTAL LIABILITIES/TOTAL ASSETS	,137	61	,006	,935	61	,003
TOTAL DEBT/TOTAL ASSETS	,319	61	,000	,339	61	,000
LONG TERM DEBT/TOTAL ASSETS	,136	61	,007	,923	61	,001
CURRENT LIABILITIES/TOTAL ASSETS	,191	61	,000	,913	61	,000
TOTAL DEBT/EQUITY	,415	61	,000	,447	61	,000
LONG TERM DEBT/EQUITY	,372	61	,000	,427	61	,000
INVENTORIES/SALES	,530	61	,000	,111	61	,000
ACCOUNTS RECIEVABLE/SALES	,170	61	,000	,826	61	,000
SALES/FIXED ASSSETS	,368	61	,000	,403	61	,000
SALES/TOTAL ASSETS	,350	61	,000	,453	61	,000
SALES/EQUITY	,311	61	,000	,693	61	,000
SALES/TOTAL DEBT	,196	61	,000	,744	61	,000
FIXED ASSETS/TOTAL ASSETS	,148	61	,002	,909	61	,000
CURRENT ASSETS/TOTAL ASSETS	,150	61	,002	,899	61	,000
(INVENTORIES+ACCOUNTS RECEIVABLE)/TOTAL ASSETS	,145	61	,003	,925	61	,001
INVENTORIES/TOTAL ASSETS	,239	61	,000	,781	61	,000
INVENTORIES/CURRENT ASSETS	,174	61	,000	,795	61	,000

CURRENT LIABILITIES/TOTAL LIABILITIES	,129	61	,013	,952	61	,018

Πίνακας 7.2: Output Kolmogorov-Smirnov τεστ

Κατά γενικό κανόνα όταν p<a, όπου a=0,05 το επίπεδο σημαντικότητας, απορρίπτεται η μηδενική υπόθεση, στη συγκεκριμένη περίπτωση ότι ακολουθεί την κανονική κατανομή, άρα η μεταβλητή δεν ακολουθεί την κανονική κατανομή. Στο SPSS ο έλεγχος του p<a γίνεται με το "Sig.", άρα όπως φαίνεται στον παραπάνω πίνακα καμία μεταβλητή δεν ακολουθεί την κανονική κατανομή.

Λόγω ότι δεν ικανοποιούνται οι υποθέσεις του μοντέλου της γραμμικής παλινδρόμησης, ομοσκεδαστικότητα- κανονικότητα- γραμμικότητα, χρησιμοποιείται η λογιστική παλινδρόμηση και συγκεκριμένα η δυαδική λογιστική παλινδρόμηση (binary logistic regression) με την κατηγορική μεταβλητή FFS να παίρνει τιμές 0= non fraud και 1=fraud. Η παλινδρόμηση έγινε ξεχωριστά για τους 16 λογαριασμούς και τους 32 αριθμοδείκτες. Επιπλέον, λόγω του ότι κάποιες μεταβλητές είναι εξαρτημένες μεταξύ τους, η λογιστική παλινδρόμηση έγινε για τους 16 λογαριασμούς σε 4 ομάδες και για τους 32 αριθμοδείκτες σε 6 ομάδες.

Variables in the Equation

		B	S.E.	Wald	df	Sig.
Step 1[a]	EBIT	,000	,000	,262	1	,609
	TOTAL ASSETS	,000	,000	,000	1	,993
	CURRENT LIABILITIES	,000	,000	,236	1	,627
	WORKING CAPITAL	,000	,000	1,587	1	,208
	Constant	-,619	,351	3,116	1	,078

Πίνακας 7.2.1: Output binary logistic regression λογαριασμών ομάδα 1

Variables in the Equation

		B	S.E.	Wald	df	Sig.
Step 1[a]	EBT	,000	,000	2,970	1	,085
	FIXED ASSETS	,000	,000	1,292	1	,256
	INVENTORIES	,000	,000	7,218	1	,007
	LONG TERM DEBT	,000	,000	3,557	1	,059
	Constant	-,855	,342	6,255	1	,012

Πίνακας 7.2.2: Output binary logistic regression λογαριασμών ομάδα 2

Variables in the Equation

		B	S.E.	Wald	df	Sig.
Step 1[a]	SALES	,000	,000	4,096	1	,043
	NET PROFIT	,000	,000	,155	1	,694
	TOTAL LIABILITIES	,000	,000	5,729	1	,017
	TOTAL DEBT	,000	,000	5,437	1	,020
	Constant	-,949	,383	6,122	1	,013

Πίνακας 7.2.3: Output binary logistic regression λογαριασμών ομάδα 3

Variables in the Equation

		B	S.E.	Wald	df	Sig.
Step 1[a]	GROSS PROFIT	,000	,000	,030	1	,863
	CURRENT ASSETS	,000	,000	4,751	1	,029
	EQUITY	,000	,000	,713	1	,398
	ACCOUNTS RECIEVABLE	,000	,000	6,737	1	,009
	Constant	-,990	,396	6,251	1	,012

Πίνακας 7.2.4: Output binary logistic regression λογαριασμών ομάδα 4

Οι στατιστικά σημαντικοί δείκτες, με Sig<0,05, είναι οι INVENTORIES, SALES, TOTAL LIABILITIES, TOTAL DEBT, CURRENT ASSETS, ACCOUNTS RECEIVABLE.

Variables in the Equation

		B	S.E.	Wald	df	Sig.
Step 1[a]	EBIT/SALES	-,759	1,479	,263	1	,608
	NET PROFIT/GROSS PROFIT	-,359	,253	2,010	1	,156
	EBT/EQUITY	1,010	,572	3,122	1	,077
	TOTAL LIABILITIES/TOTAL ASSETS	-5,877	1,968	8,916	1	,003
	CURRENT ASSETS/TOTAL ASSETS	-2,175	1,516	2,058	1	,151
	Constant	3,974	1,562	6,472	1	,011

Πίνακας 7.2.5: Output binary logistic regression αριθμοδεικτών ομάδα 1

Variables in the Equation

		B	S.E.	Wald	df	Sig.
Step 1[a]	EBT/TOTAL ASSETS	-1,193	7,883	,023	1	,880
	NET PROFIT/FIXED ASSETS	-5,663	4,847	1,365	1	,243
	CURRENT LIABILITIES/TOTAL ASSETS	-7,034	2,482	8,030	1	,005
	TOTAL DEBT/EQUITY	-,132	,078	2,837	1	,092
	INVENTORIES/TOTAL ASSETS	29,858	9,624	9,625	1	,002
	Constant	,346	,683	,256	1	,613

Πίνακας 7.2.6: Output binary logistic regression αριθμοδεικτών ομάδα 2

Variables in the Equation

		B	S.E.	Wald	df	Sig.
Step 1[a]	EBT/SALES	-3,058	2,173	1,981	1	,159
	NET PROFIT/TOTAL ASSETS	4,905	7,102	,477	1	,490
	(CURRENT ASSETS-INVENTORIES)/CURRENT LIABILITIES	,008	,462	,000	1	,986
	SALES/TOTAL DEBT	-,358	,184	3,801	1	,051
	WORKING CAPITAL/TOTAL ASSETS	2,600	2,170	1,435	1	,231
	Constant	,232	,666	,121	1	,728

Πίνακας 7.2.7: Output binary logistic regression αριθμοδεικτών ομάδα 3

Variables in the Equation

		B	S.E.	Wald	df	Sig.
Step 1[a]	NET PROFIT/EQUITY	,344	,424	,660	1	,417
	LONG TERM DEBT/TOTAL ASSETS	-7,788	3,473	5,029	1	,025
	SALES/FIXED ASSSETS	-2,854	1,135	6,318	1	,012
	(INVENTORIES+ACCOUNTS RECEIVABLE)/TOTAL ASSETS	9,153	5,376	2,899	1	,089
	INVENTORIES/CURRENT ASSETS	12,553	4,192	8,967	1	,003
	Constant	,368	,929	,157	1	,692

Πίνακας 7.2.8: Output binary logistic regression αριθμοδεικτών ομάδα 4

Variables in the Equation

		B	S.E.	Wald	df	Sig.
Step 1[a]	NET PROFIT/SALES	-3,209	1,699	3,566	1	,059
	EBIT/TOTAL ASSETS	16,703	9,012	3,435	1	,064
	CURRENT ASSETS/CURRENT LIABILITIES	,263	,308	,731	1	,392
	LONG TERM DEBT/EQUITY	-1,478	,861	2,943	1	,086
	SALES/TOTAL ASSETS	-1,127	,999	1,273	1	,259
	ACCOUNTS RECIEVABLE/SALES	-,611	1,066	,328	1	,567
	Constant	,611	,951	,413	1	,520

Πίνακας 7.2.9: Output binary logistic regression αριθμοδεικτών ομάδα 5

Variables in the Equation

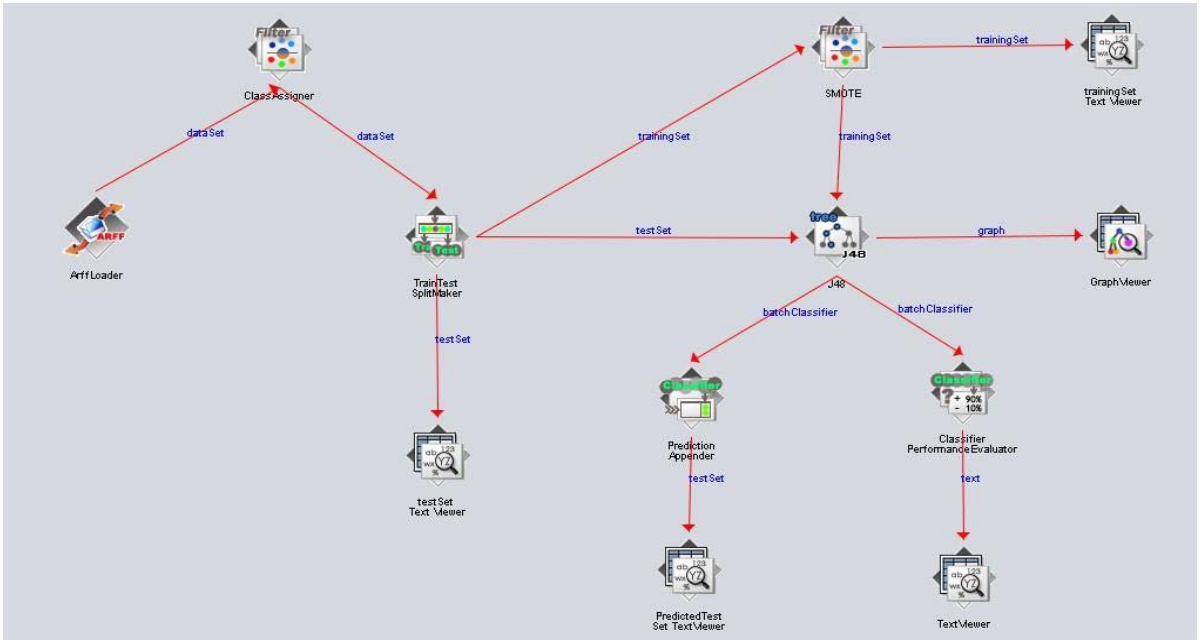
		B	S.E.	Wald	df	Sig.
Step 1[a]	EBT/FIXED ASSETS	-7,167	3,431	4,364	1	,037
	TOTAL DEBT/TOTAL ASSETS	-,589	,642	,842	1	,359
	SALES/EQUITY	-,122	,082	2,241	1	,134
	INVENTORIES/SALES	35,539	10,473	11,515	1	,001
	FIXED ASSETS/TOTAL ASSETS	-1,962	2,706	,526	1	,468
	CURRENT LIABILITIES/TOTAL LIABILITIES	2,279	2,555	,795	1	,372
	Constant	-3,216	2,867	1,258	1	,262

Πίνακας 7.2.10: Output binary logistic regression αριθμοδεικτών ομάδα 6

Οι στατιστικά σημαντικοί δείκτες, με Sig<0,05, είναι οι TOTAL LIABILITIES/TOTAL ASSETS, CURRENT LIABILITIES/TOTAL ASSETS, INVENTORIES/TOTAL ASSETS, LONG TERM DEBT/TOTAL ASSETS, SALES/FIXED ASSETS, INVENTORIES/CURRENT ASSETS, EBT/FIXED ASSETS, INVENTORIES/SALES.

7.3 ΕΞΌΡΥΞΗ ΔΕΔΟΜΈΝΩΝ ΓΙΑ ΑΝΊΧΝΕΥΣΗ ΨΕΥΔΏΝ ΟΙΚΟΝΟΜΙΚΏΝ ΚΑΤΑΣΤΆΣΕΩΝ

Ο εντοπισμός ψευδών οικονομικών καταστάσεων μπορεί να θεωρηθεί ως τυπικό πρόβλημα ταξινόμησης και η κατηγορική μεταβλητή FFS να παίρνει τιμές a=NO (non fraud) και b=YES (fraud). Μέσα από το Knowledge Flow του Weka ακολουθούνται τα εξής βήματα για να γίνει η ταξινόμηση. Πρώτων, ανεβαίνει το αρχείο του δείγματος (Arff Loader) και έπειτα επιλέγεται η εξαρτημένη μεταβλητή, η μεταβλητή δηλαδή που θα γίνει η ταξινόμηση (Class Assigner). Δεύτερον, το δείγμα χωρίζεται σε δείγμα εκπαίδευσης και σε δείγμα που θα χρησιμοποιηθεί για να γίνει το τεστ. Ο διαχωρισμός έγινε κατά 70% σε δείγμα εκπαίδευσης, 111 οικονομικές καταστάσεις, και 30% σε δείγμα για το τεστ, 48 οικονομικές καταστάσεις από τις οποίες οι 37 είναι χαρακτηρισμένες ως αληθείς και 11 ως ψευδείς. Τρίτον, προκείμενου το τεστ εκπαίδευσης να αποτελείται από περίπου ίσες οικονομικές καταστάσεις χαρακτηρισμένες ως a=NO (non fraud) και b=YES (fraud) χρησιμοποιείται το φίλτρο SMOTE και το τεστ εκπαίδευσης αποτελείται πλέον από 174 οικονομικές καταστάσεις, 90 αληθείς και 84 ψευδείς. Το φίλτρο SMOTE με την μέθοδο των πλησιέστερων γειτόνων συνθέτει νέες φανταστικές περιπτώσεις μειοψηφίας χρησιμοποιώντας το πιο κοντινό και το δεύτερο πιο κοντινό γείτονα και επιλέγοντας τυχαία ένα νέο μεταξύ τους.



Σχήμα 7.3.1: Βήματα στο Knowledge Flow του Weka

Αφού γίνουν όλα τα παραπάνω βήματα επιλέγεται η μέθοδος και ο αλγόριθμος που θα χρησιμοποιηθούν, δέντρα αποφάσεων, νευρωνικά δίκτυα και Bayesian Belief Network. Έπειτα, εμφανίζονται τα αποτελέσματα της πρόβλεψης (Prediction Appender) και ο πίνακας αποτελεσμάτων ώστε να αξιολογηθεί η απόδοση της κάθε μεθόδου (Classifier Performance Evaluator). Υπάρχει, η δυνατότητα να εμφανιστούν το δείγμα εκπαίδευσης και το δείγμα που θα χρησιμοποιηθεί για επαλήθευση, όπως επίσης και το γράφημα που κάθε αλγόριθμου και ο πίνακας αποτελεσμάτων (Text Viewer).

7.3.1 Δέντρα αποφάσεων

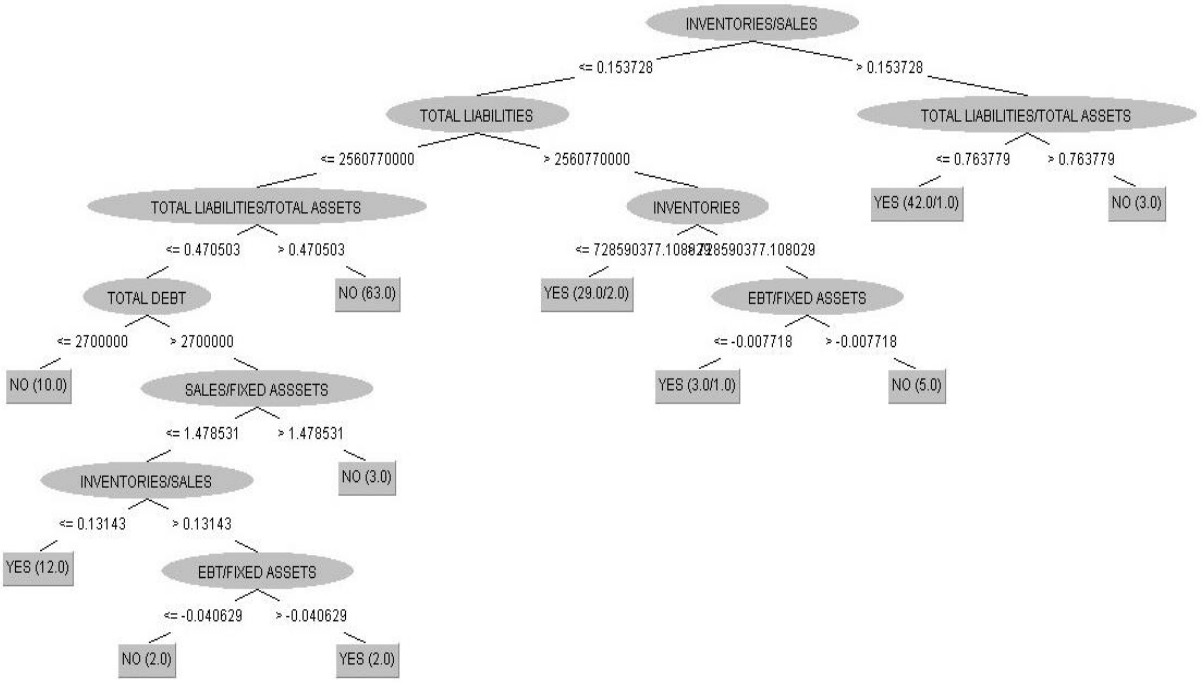
Για εξόρυξη δεδομένων με την μέθοδο των δέντρων αποφάσεων θα χρησιμοποιηθεί ο αλγόριθμος J48 (μια εφαρμογή ανοιχτού κώδικα Java του αλγορίθμου C4.5 στο Weka). Τα αποτελέσματα φαίνονται παρακάτω.

```
Scheme: J48
Options: -C 0.25 -M 2
Relation: WEKA_6_4-weka.filters.unsupervised.attribute.Remove-R1-2-weka.filters.unsupervised.attribute

=== Summary ===

Correctly Classified Instances          38               79.1667 %
Incorrectly Classified Instances        10               20.8333 %
Kappa statistic                          0.4771
Mean absolute error                      0.2083
Root mean squared error                  0.4462
Relative absolute error                 42.4506 %
Root relative squared error             90.8835 %
Total Number of Instances               48

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,811    0,273    0,909      0,811    0,857      0,488    0,776     0,891     NO
                 0,727    0,189    0,533      0,727    0,615      0,488    0,776     0,479     YES
Weighted Avg.    0,792    0,254    0,823      0,792    0,802      0,488    0,776     0,797

=== Confusion Matrix ===

  a  b   <-- classified as
 30  7 |  a = NO
  3  8 |  b = YES
```

Πίνακας 7.3.1.1: Output J48

Ο αλγόριθμος J48 προβλέπει ότι έχουν ταξινομηθεί σωστά το 79,1667% των περιπτώσεων, 38 οικονομικές καταστάσεις, και λάθος το 20,8333% των περιπτώσεων, 10 οικονομικές καταστάσεις. Πιο συγκεκριμένα, 30 οικονομικές καταστάσεις είχαν ταξινομηθεί σωστά ως αληθείς και 8 οικονομικές καταστάσεις είχαν ταξινομηθεί σωστά ως ψευδείς άλλα 3 οικονομικές καταστάσεις είχαν ταξινομηθεί λάθος ως ψευδείς και 7 οικονομικές καταστάσεις είχαν ταξινομηθεί λάθος ως αληθείς.

Οι μεταβλητές που χρησιμοποιήθηκαν στον J48 και οι τιμές διαχωρισμού τους φαίνονται στο παρακάτω δέντρο απόφασης. Το δένδρο απόφασης έχει γίνει για τα δεδομένα εκπαίδευσης και προβλέπει ότι 86 οικονομικές καταστάσεις είχαν ταξινομηθεί σωστά ως αληθείς και 84 οικονομικές καταστάσεις είχαν ταξινομηθεί σωστά ως ψευδείς ενώ 4 οικονομικές καταστάσεις είχαν ταξινομηθεί λάθος ως αληθείς και τις προβλέπει ψευδείς.



Σχήμα 7.3.1.2: Δέντρο απόφασης του αλγόριθμου J48

7.3.2 Νευρωνικά δίκτυα

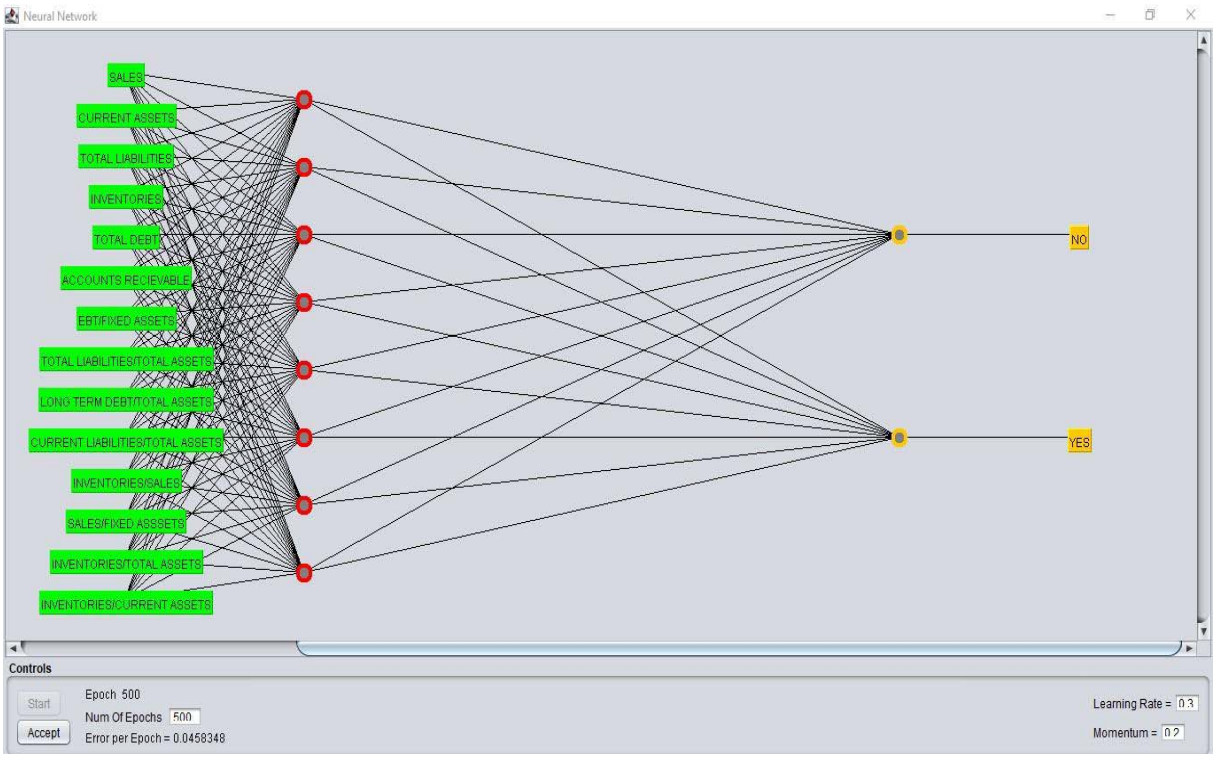
Στο Weka για την εξόρυξη δεδομένων με την μέθοδο των νευρωνικών δικτύων θα χρησιμοποιηθεί ο MultilayerPerceptron που είναι μια μέθοδος τεχνητού νευρωνικού δικτύου και χρησιμοποιεί τον αλγόριθμο perceptron, έναν αλγόριθμο για την εποπτευόμενη μάθηση των δυαδικών ταξινομητών. Τα αποτελέσματα είναι τα εξής.

```
Scheme: MultilayerPerceptron
Options: -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a -G -R
Relation: WEKA_6_4-weka.filters.unsupervised.attribute.Remove-R1-2-weka.filters.unsupervised.attribute.

=== Summary ===

Correctly Classified Instances          40               83.3333 %
Incorrectly Classified Instances         8               16.6667 %
Kappa statistic                          0.5817
Mean absolute error                      0.1839
Root mean squared error                  0.3816
Relative absolute error                 37.4704 %
Root relative squared error             77.7173 %
Total Number of Instances               48

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,838     0,182     0,939      0,838    0,886      0,595    0,887     0,965     NO
               0,818     0,162     0,600      0,818    0,692      0,595    0,887     0,744     YES
Weighted Avg.  0,833     0,177     0,862      0,833    0,841      0,595    0,887     0,915

=== Confusion Matrix ===

  a  b   <-- classified as
 31  6 |  a = NO
  2  9 |  b = YES
```

Πίνακας 7.3.2.1: Output MultilayerPerceptron

Ο MultilayerPerceptron προβλέπει ότι είχαν ταξινομηθεί σωστά το 83,3333% των περιπτώσεων, 40 οικονομικές καταστάσεις, και λάθος το 16,6667% των περιπτώσεων, 8οικονομικές καταστάσεις. Πιο συγκεκριμένα, 31 οικονομικές καταστάσεις είχαν ταξινομηθεί σωστά ως αληθείς και 9 οικονομικές καταστάσεις είχαν ταξινομηθεί σωστά ως ψευδείς ενώ 2 οικονομικές καταστάσεις είχαν ταξινομηθεί λάθος ως ψευδείς και 6 οικονομικές καταστάσεις είχαν ταξινομηθεί λάθος ως αληθείς.

Όλες οι μεταβλητές χρησιμοποιήθηκαν στον MultilayerPerceptron και φαίνονται στο παρακάτω γράφημα που μοιάζει με βιολογικό νευρώνα του εγκεφάλου.



Σχήμα 7.3.2.2: Γράφημα του MultilayerPerceptron

7.3.3 Bayesian Belief Network

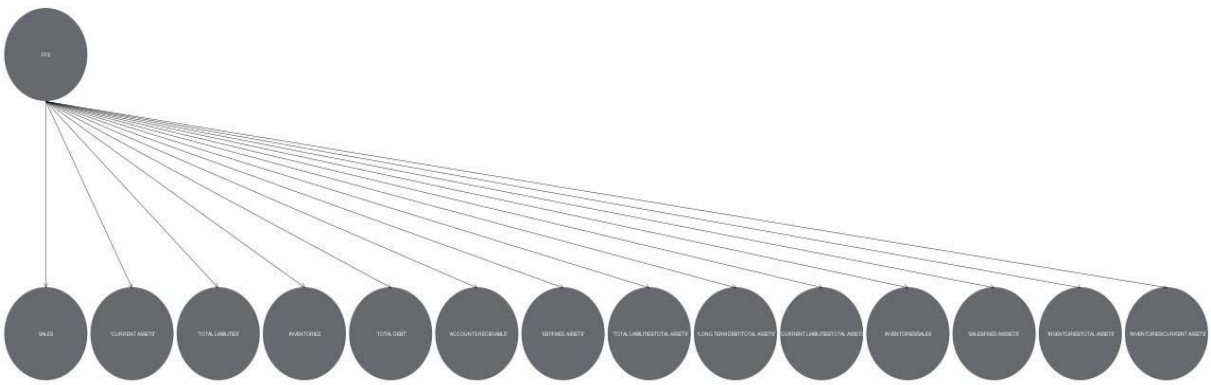
Για εξόρυξη δεδομένων με την μέθοδο του Bayesian Belief Network μέσα από το λογισμικό Weka θα χρησιμοποιηθεί ο αλγόριθμος K2 με την επιλογή του BayesNet. Τα αποτελέσματα είναι του K2 είναι τα εξής.

```
Scheme: BayesNet
Options: -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.ne
Relation: WEKA_6_4-weka.filters.unsupervised.attribute.Remove-R1-2-weka.filters.unsupervised.attribute.

=== Summary ===

Correctly Classified Instances          41               85.4167 %
Incorrectly Classified Instances         7               14.5833 %
Kappa statistic                          0.6233
Mean absolute error                      0.1514
Root mean squared error                  0.3784
Relative absolute error                 30.8512 %
Root relative squared error             77.0681 %
Total Number of Instances               48

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,865     0,182     0,941       0,865   0,901      0,632    0,904     0,969     NO
               0,818     0,135     0,643       0,818   0,720      0,632    0,904     0,769     YES
Weighted Avg.  0,854     0,171     0,873       0,854   0,860      0,632    0,904     0,923

=== Confusion Matrix ===

  a  b   <-- classified as
 32  5 |  a = NO
  2  9 |  b = YES
```

Πίνακας 7.3.3.1: Output K2

Ο αλγόριθμος K2 προβλέπει ότι είχαν ταξινομηθεί σωστά το 85,4167% των περιπτώσεων, 41 οικονομικές καταστάσεις, και λάθος το 14,5833% των περιπτώσεων, 7 οικονομικές καταστάσεις. Πιο συγκεκριμένα, 32 οικονομικές καταστάσεις είχαν ταξινομηθεί σωστά ως αληθείς και 9 οικονομικές καταστάσεις είχαν ταξινομηθεί σωστά ως ψευδείς άλλα 2 οικονομικές καταστάσεις είχαν ταξινομηθεί λάθος ως ψευδείς και 5 οικονομικές καταστάσεις είχαν ταξινομηθεί λάθος ως αληθείς.

Όλες οι μεταβλητές χρησιμοποιήθηκαν στον K2 και ο γονέας όλων των μεταβλητών είναι η εξαρτημένη μεταβλητή FFS όπως φαίνεται και στο παρακάτω γράφημα.



Σχήμα 7.3.3.2: Γράφημα του K2

7.4 ΑΠΟΤΕΛΈΣΜΑΤΑ ΕΞΌΡΥΞΗΣ

Αφού έγινε η εξόρυξη δεδομένων θα γίνει μια σύγκριση των αποτελεσμάτων καθώς και ανάλυση αποτελεσμάτων. Επίσης, θα αξιολογηθούν το μέσο απόλυτο σφάλμα (Mean Absolute Error) και η ρίζα του μέσου τετραγωνικού σφάλματος (Root Mean Square Error). Το μέσο απόλυτο σφάλμα είναι ένα μέτρο διασποράς μεταξύ δύο μεταβλητών και μετρά την μέση οριζόντια απόσταση της κάθε μεταβλητής από την γραμμή παλινδρόμησης. Το μέσο απόλυτο σφάλμα είναι ένας τρόπος σύγκρισης των προβλέψεων με τα τελικά αποτελέσματα, και όσο πιο μικρή η τιμή του τόσο το καλύτερο. Η ρίζα του μέσου τετραγωνικού σφάλματος είναι ένα μέτρο ακριβείας ώστε να συγκρίνονται οι μεταβλητές των μεταβλητών του δείγματος με τις τιμές των μεταβλητών του προβλεπόμενου δείγματος. Είναι πάντα θετικός αριθμός και μια τιμή 0 θα έδειχνε μια καλή εφαρμογή στα δεδομένα, επειδή όμως στην πράξη αυτό σπάνια επιτυγχάνεται όσο πιο κοντά στο 0 τόσο καλύτερα.

Με την μέθοδο των δέντρων αποφάσεων η απόδοση του αλγόριθμου J48 είναι 76,91% και δείχνει το ανώτερο όριο απόδοσης του μοντέλου σε νέα δεδομένα. Το μέσο απόλυτο σφάλμα (MAE) είναι 0,2083 και η ρίζα του μέσου τετραγωνικού σφάλματος (RMSE) είναι 0,4462.

J48	ΠΟΣΟΣΤΌ ΣΩΣΤΉΣ ΤΑΞΙΝΌΜΗΣΗΣ	ΠΟΣΟΣΤΌ ΛΆΘΟΣ ΤΑΞΙΝΌΜΗΣΗΣ
a=NO	30/37 =81,08%	7/37 =18,92%
b=YES	8/11 =72,73%	3/11 =27,27%
ΜΈΣΟΣ ΌΡΟΣ	76,91%	23,09%
MAE	0,2083	
RMSE	0,4462	

Πίνακας 7.4.1: Αποτελέσματα J48

Με την μέθοδο των νευρωνικών δικτύων και του MultilayerPerceptron η απόδοση είναι 82,8% και δείχνει το ανώτερο όριο απόδοσης του μοντέλου σε νέα δεδομένα. Το μέσο απόλυτο σφάλμα (MAE) είναι 0,1839 και η ρίζα του μέσου τετραγωνικού σφάλματος (RMSE) είναι 0,3816.

Multilayer Perceptron	ΠΟΣΟΣΤΌ ΣΩΣΤΉΣ ΤΑΞΙΝΌΜΗΣΗΣ	ΠΟΣΟΣΤΌ ΛΆΘΟΣ ΤΑΞΙΝΌΜΗΣΗΣ
a=NO	31/37 =83,78%	6/37 =16,22%
b=YES	9/11 =81,82%	2/11 =18,18%
ΜΈΣΟΣ ΌΡΟΣ	82,8%	17,2%
MAE	0,1839	
RMSE	0,3816	

Πίνακας 7.4.2: Αποτελέσματα MultilayerPerceptron

Με την μέθοδο του Bayesian Belief Network του αλγόριθμου K2 η απόδοση είναι 84,16% και δείχνει το ανώτερο όριο απόδοσης του μοντέλου σε νέα δεδομένα. Το μέσο απόλυτο σφάλμα (MAE) είναι 0,1514 και η ρίζα του μέσου τετραγωνικού σφάλματος (RMSE) είναι 0,3784.

K2	ΠΟΣΟΣΤΌ ΣΩΣΤΉΣ ΤΑΞΙΝΌΜΗΣΗΣ	ΠΟΣΟΣΤΌ ΛΆΘΟΣ ΤΑΞΙΝΌΜΗΣΗΣ
a=NO	32/37 =86,49%	5/37 =13,51%
b=YES	9/11 =81,82%	2/11 =18,18%
ΜΈΣΟΣ ΌΡΟΣ	84,16%	15,84%
MAE	0,1514	
RMSE	0,3784	

Πίνακας 7.4.3: Αποτελέσματα Bayesian Belief Network

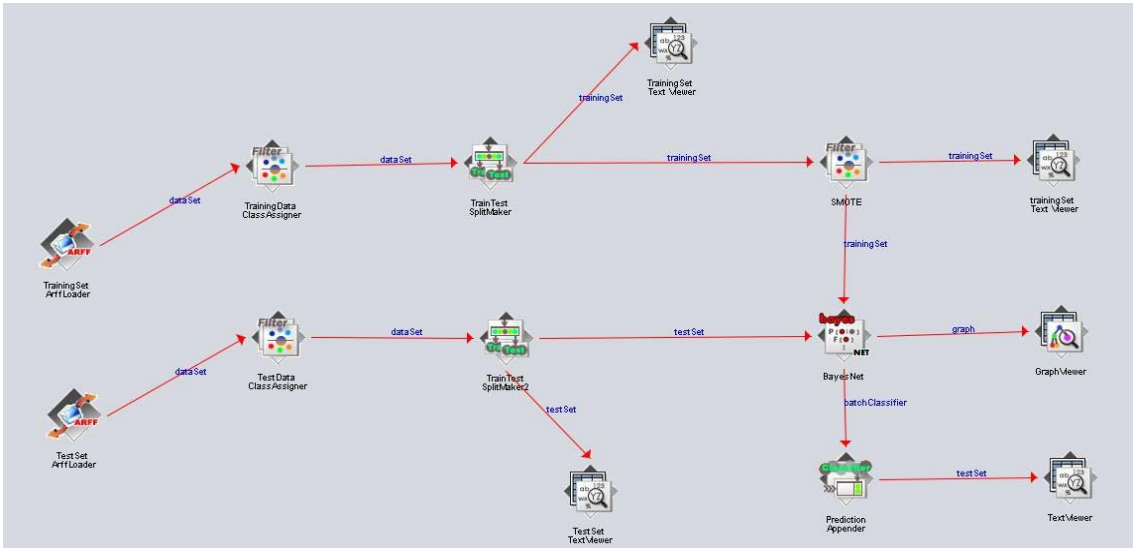
Συγκριτικά, η μέθοδος του Bayesian Belief Network έχει την καλύτερη απόδοση και τόσο το μέσο απόλυτο σφάλμα όσο και η ρίζα του μέσου τετραγωνικού σφάλματος είναι πιο κοντά στο 0 σε σχέση με τις άλλες μεθόδους άρα έχει και την καλύτερη εφαρμογή στα δεδομένα.

	Μ.Ο. ΠΟΣΟΣΤΌ ΣΩΣΤΉΣ ΤΑΞΙΝΌΜΗΣΗΣ	MAE	RMSE
J48	76,91%	0,2083	0,4462
Multilayer Perceptron	82,8%	0,1839	0,3816
K2	84,16%	0,1514	0,3784

Πίνακας 7.4.4: Συγκριτικά αποτελέσματα

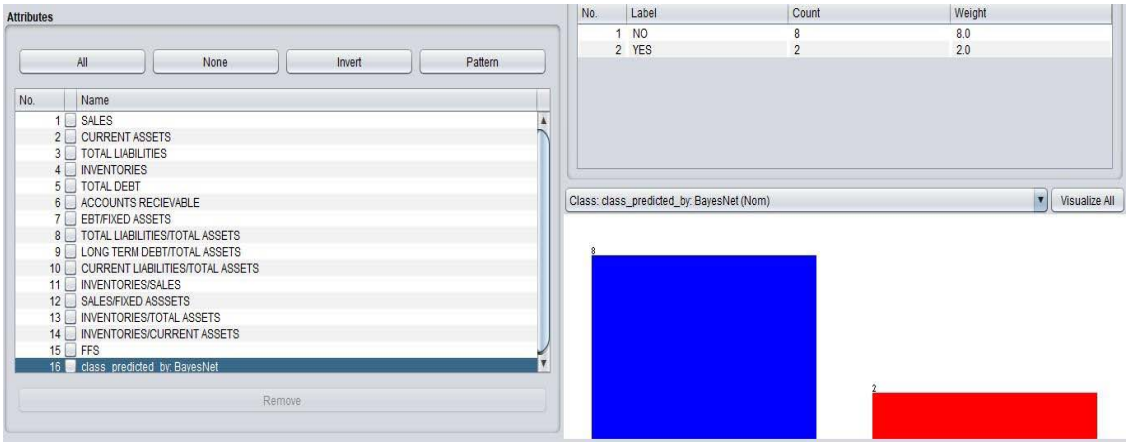
7.5 ΕΞΌΡΥΞΗ ΔΕΔΟΜΈΝΩΝ ΓΙΑ ΑΝΊΧΝΕΥΣΗ ΨΕΥΔΏΝ ΟΙΚΟΝΟΜΙΚΏΝ ΚΑΤΑΣΤΆΣΕΩΝ ΣΕ ΝΈΟ ΔΕΊΓΜΑ

Με την μέθοδο Bayesian Belief Network και το λογισμικό Weka θα γίνει κατηγοριοποίηση σε νέο δείγμα. Το νέο δείγμα αποτελείται από 10 οικονομικές καταστάσεις, μη κατηγοριοποιημένο, και ως δείγμα εκπαίδευσης θα χρησιμοποιηθούν και οι 159 οικονομικές καταστάσεις. Τα βήματα που θα ακολουθηθούν στο Knowledge Flow του Weka είναι περίπου ίδια με τα προηγούμενα. Οι διαφορές τους είναι ότι δεν γίνεται διαχωρισμός κατά 70%-30%, αλλά χρησιμοποιείται το 100% ως τεστ εκπαίδευσης (training set) και ανεβαίνει δεύτερο αρχείο, το νέο δείγμα, που χρησιμοποιείται 100% ως δείγμα για τεστ (test set).



Σχήμα 7.5.1: Βήματα στο Knowledge Flow του Weka_Νέο δείγμα

Το γράφημα είναι ακριβώς ίδιο με το Σχήμα 7.3.2.3, όλες οι μεταβλητές χρησιμοποιήθηκαν στον K2 και ο γονέας όλων των μεταβλητών είναι η εξαρτημένη μεταβλητή FFS.



Σχήμα 7.5.2: Αποτελέσματα του K2_Νέο δείγμα

Τα αποτελέσματα του K2 για το νέα δείγμα είναι ότι 8 οικονομικές καταστάσεις είναι αληθείς και 2 ψευδείς, και συγκεκριμένα η κατηγοριοποίηση της κάθε μιας φαίνεται παρακάτω.

```
@relation 'TEST SET-weka.filters.unsupervised.attribute.StringToNominal-Rlast-weka.filters.unsupervised.attribute.Remove-R1-2-weka.filters.unsupervised.

@attribute SALES numeric
@attribute 'CURRENT ASSETS' numeric
@attribute 'TOTAL LIABILITIES' numeric
@attribute INVENTORIES numeric
@attribute 'TOTAL DEBT' numeric
@attribute 'ACCOUNTS RECIEVABLE' numeric
@attribute 'EBT/FIXED ASSETS' numeric
@attribute 'TOTAL LIABILITIES/TOTAL ASSETS' numeric
@attribute 'LONG TERM DEBT/TOTAL ASSETS' numeric
@attribute 'CURRENT LIABILITIES/TOTAL ASSETS' numeric
@attribute INVENTORIES/SALES numeric
@attribute 'SALES/FIXED ASSSETS' numeric
@attribute 'INVENTORIES/TOTAL ASSETS' numeric
@attribute 'INVENTORIES/CURRENT ASSETS' numeric
@attribute FFS {NO,YES}
@attribute 'class_predicted_by: BayesNet' {NO,YES}

@data
114884325.8,53975484.88,56857206.91,26335608.35,38466053.93,24443773.69,0.138612,0.594955,0.228116,0.328922,0.229236,2.697441,0.275576,0.487918,?,NO
78297459,101559084,177056000,418873,120860000,56341851,-0.160534,0.872151,0.334918,0.498724,0.00535,0.771771,0.002063,0.004124,?,NO
73201876.91,42082251.21,34940621.4,9350249.32,21451074.07,14487032.63,0.138312,0.536867,0.161718,0.352988,0.127732,3.182662,0.143668,0.22219,?,NO
77264511,102644799,175520000,427519,123684000,54436140,-0.205314,0.817625,0.383294,0.400169,0.005533,0.689703,0.001992,0.004165,?,NO
1001351000,1044951000,1770186000,111245000,694899000,523561000,0.086403,0.676006,0.214639,0.329953,0.111095,0.660418,0.042483,0.10646,?,YES
88050923,49332133,69939346,9916499,42324395,27757112,0.046095,0.45488,0.137883,0.222325,0.112622,0.843229,0.064496,0.201015,?,NO
255965152,172550086,195010615,66380158,1413762270,72960564,-0.069951,0.633833,0.011806,0.573532,0.259333,1.894375,0.215752,0.384701,?,NO
101330254,68280665,74860028,20179125,34267977,33664319,0.047589,0.451671,0.003218,0.338717,0.199142,1.039717,0.121752,0.295532,?,NO
115710618,43972898,227831472,22240268,152236969,10997625,-0.406979,1.553482,0.00007,1.510614,0.192206,1.126843,0.151647,0.505772,?,NO
1526154000,1482637000,1779945000,184377000,562940000,799307000,0.08982,0.532759,0.159841,0.260685,0.120812,0.821238,0.055186,0.124357,?,YES
```

Σχήμα 7.5.3: Κατηγοριοποίηση νέου δείγματος

8.ΣΥΜΠΕΡΆΣΜΑΤΑ

Όσο οι βάσεις δεδομένων αναπτύσσονται τόσο καλύτερη εφαρμογή θα έχει η μέθοδος της εξόρυξης δεδομένων όχι μόνο στη λογιστική και την ελεγκτική αλλά σε πολλούς ακόμα τομείς. Σημαντικό ρόλο γι' αυτό έχει η τεχνητή νοημοσύνη. Η τεχνητή νοημοσύνη ασχολείται με την σχεδίαση ευφυών υπολογιστικών συστημάτων που επιδεικνύουν χαρακτηριστικά που σχετίζονται με την νοημοσύνη στην ανθρώπινη συμπεριφορά. Όσο η τεχνητή νοημοσύνη αναπτύσσεται τόσο καλύτερα θα προσαρμόζονται οι μηχανές στις ανάγκες του κάθε χρήστη, θα μαθαίνει από τα λάθη και τόσο καλύτερα θα επιλύουν προβλήματα. Η τεχνητή νοημοσύνη δεν είναι σκοπός αλλά μέσο αποτελεσματικότητας, αποδοτικότητας και καινοτομίας. Η εξόρυξη δεδομένων και η ανάλυση των αποτελεσμάτων με παραδοσιακούς τρόπους μπορεί σε κάποιες περιπτώσεις να απαιτούν πολλές ώρες. Η τεχνητή νοημοσύνη μπορεί να συμβάλει στη μείωση μέσω αξιόπιστων λειτουργιών αναζήτησης και αναλύσεων. Οι διαδικασίες δηλαδή κατά την λήψη αποφάσεων θα παίρνονται αυτοματοποιημένα και με μεγαλύτερη ακρίβεια. [Behrouz Forouzan, Firouz Mesharraf, 2010]

Η εξόρυξη δεδομένων με μεθόδους μηχανικής μάθησης είναι μια πρακτική που συνεχώς εξελίσσεται και στο μέλλον θα χρησιμοποιείται από όλο και περισσότερους τομείς. Η χρήση των πρακτικών που χρησιμοποιήθηκαν στην εργασία θα μπορούσαν να χρησιμοποιηθούν σαν εργαλεία από εσωτερικούς ελεγκτές, εξωτερικούς ελεγκτές, φορολογικές αρχές, επενδυτικές αρχές, τους επενδυτές, το χρηματιστήριο, τους οικονομικούς αναλυτές, το τραπεζικό σύστημα, τους πιστωτές, τα δικηγορικά γραφεία. Για όλους αυτούς, θα ήταν σημαντικό να κάνουν μια πρόβλεψη μηχανικά με την καλύτερη δυνατή απόδοση. Ανάλογα, η εξόρυξη δεδομένων με μεθόδους μηχανικής μάθησης έχει εφαρμογή τόσο σε δεδομένα που έχουμε πληροφορίες του αποτελέσματος και γίνεται επαλήθευση και πρόβλεψη για το αν π.χ. είναι σωστά ταξινομημένο όπως στην συγκεκριμένη εργασία, αλλά και σε νέα δεδομένα ώστε να προβλεφθεί που ανήκουν.

Με την ίδια διαδικασία που χρησιμοποιήθηκε στην συγκεκριμένη εργασία θα μπορούσε να γίνει πρόβλεψη χρεοκοπίας, όπως έχει αναφερθεί και έχει υλοποιηθεί σε άλλες εργασίες. Τα βήματα είναι ακριβώς τα ίδια αλλά αλλάζει το αρχείο εκπαίδευσης. Στην στήλη που έχουμε τώρα την κατηγοριοποίηση ως ψευδής ή αληθής, θα γίνει κατηγοριοποίηση ως χρεοκοπημένη ή όχι. Έπειτα κατά την λογιστική παλινδρόμηση ίσως οι μεταβλητές που

είναι σημαντικές θα αλλάξουν, έτσι στο αρχείο εκπαίδευσης εκτός από το την στήλη της κατηγορικής μεταβλητής θα αλλάξουν και οι στήλες των ανεξάρτητων μεταβλητών.

Το πρακτικό κομμάτι κατά την εκτέλεση μια εφαρμογής εξόρυξης δεδομένων είναι το πιο ουσιαστικό όμως για την επιτυχημένη εκτέλεση μιας τεχνικής απαιτείται η ύπαρξη ενός καλά θεμελιωμένου θεωρητικού υποβάθρου. Γι' αυτό τον λόγο αρχικά επικεντρωνόμαστε στη θεωρητική προσέγγιση της εξόρυξης δεδομένων και έπειτα στην εκτέλεση με το λογισμικό Weka. Στη συγκεκριμένη εργασία καλύτερη εφαρμογή έχει η μέθοδος Bayesian Belief Network και η ίδια μέθοδος έχει βρεθεί σαν αυτή με την καλύτερη εφαρμογή και από τους E.Kirkos, Ch.Spathis, Y.Manolopoulos (2007 b), ενώ οι E.Kirkos, Ch.Spathis, Y.Manolopoulos (2007 a) και Kotsiantis S., Tselepis D., Tampakas V. (2005) βρήκαν τις μεθόδους των δέντρων αποφάσεων και των νευρωνικών δικτύων σαν αυτές με την καλύτερη εφαρμογή.

ΠΑΡΑΡΤΗΜΑ

ΚΛΆΔΟΣ	ΑΡΙΘΜΌΣ	ΠΟΣΟΣΤΌ
Ταξίδια & Αναψυχή	10	6.3%
Κατασκευές & Υλικά κατασκευών	10	6.3%
Τρόφιμα & Ποτά	20	12.6%
Υπηρεσίες Κοινής Ωφέλειας	22	13.8%
Εμπόριο	12	7.5%
Βιομηχανικά Προϊόντα & Υπηρεσίες	20	12.6%
Υγεία	10	6.3%
Πετρέλαια & Αέριο	10	6.3%
Χρηματοοικονομικές Υπηρεσίες	12	7.5%
Προσωπικά & Οικιακά Αγαθά	16	10.1%
Τεχνολογία	4	2.5%
Ακίνητη Περιουσία	3	1.9%
Άλλο	10	6.3%
Σύνολο	159	100%

Πίνακας 1.3.1: Κλαδικά κατάταξη 159 εταιρειών

ΚΛΆΔΟΣ	ΑΡΙΘΜΌΣ	ΠΟΣΟΣΤΌ
Τρόφιμα & Ποτά	4	40%
Βιομηχανικά Προϊόντα & Υπηρεσίες	4	40%
Μέσα Ενημέρωσης	2	20%
Σύνολο	10	100%

Πίνακας 1.3.2: Κλαδικά κατάταξη νέου δείγματος 10 εταιρειών

ΒΙΒΛΙΟΓΡΑΦΙΑ

5 algorithms to train a neural network. https://www.neuraldesigner.com/blog/5_algorithms_to_train_a_neural_network

5 Huge Tech Companies That Use Python, https://www.cleveroad.com/blog/discover-5-leading-companies-that-use-python-and-learn-does-it-fit-your-project#top

Alexander Hamilton Institute, (2011), *Εσωτερικός Έλεγχος*, Αθήνα: Εκδόσεις Κριτήριον.

Basic elements of Python, https://geo-python.github.io/site/2018/notebooks/L2/Python-basic-elements.html#

Behrouz F., Firouz M., (2010), *Εισαγωγή στην επιστήμη των υπολογιστών*, Αθήνα: Εκδόσεις Κλειδάριθμος.

Calderon T.G., Cheh J.J., (2002), *A roadmap for future neural networks research in auditing and risk assessment,* International Journal of accounting international system,3(4), 203-236.

Coderre G.D., (1999), *Fraud detection. Using data analysis techniques to detect fraud*, Global Audit Publications.

Dalnial H., Kammaluddin A., Sanusi Z.M., Khairuddin K.S., (2014), *Accountability in financial reporting: detecting fraudulent firms,* Procedia - Social and Behavioral Sciences, 145(25), Pages 61-69.

Data science graduate programs, /https://www.datasciencegraduateprograms.com/

Downing D., Clark J., (2010), *Στατιστική των επιχειρήσεων*, Αθήνα: Εκδόσεις Κλειδάριθμος.

Gaganis C., (2009), *Classification techniques for the identification of falsified financial statements: a comparative analysi., Intelligent systems in accounting*, Finance &Management, 16, 2007-229.

IBM SPSS software, https://www.ibm.com/analytics/spss-statistics-software

Kanapickiene R., Grudiene Z., (2015), *The Model of Fraud Detection in Financial Statements by Means of Financial Ratios*, Procedia - Social and Behavioral Sciences, 213, 321 – 327.

Kirkos E., Spathis Ch., Manolopoulos Y., (2007 a), *Applying data mining methodologies for auditor selection*, 11th Panhellenic Conference in Informatics, 165-177.

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

Kirkos S., Manolopoulos Y., (2004), *Data mining in financial and accounting: A review of current research trends*, Proceedings of the 1st International Conference of Enterprise System and Accounting, 63-78.

Koskivaara E., (2004), *Artificial neural networks in analytical review procedures,* Managerial Auditing Journal, 19(2), 191-223.

Kotsiantis S., Tselepis D., Tampakas V., (2005), *Forecasting fraudulent financial statements using data mining*, International Journal of Computational Intelligence, 3(2), 1304-2386.

Logistic Regression in SPSS, http://ecourse.uoi.gr/pluginfile.php/105443/mod_resource/content/1/Logistic%20Regression%20in%20SPSS.pdf

Perols J., (2011), *Financial Statement Fraud Detection: An analysis of statistical and machine learning algorithms,* American Accounting Association, 30(2), 19-50.

Person O., (1995), *Using financial statement data to identify factors associated with fraudulent financial reporting,* Journal of Applied Business Research, 11(3), 38-46.

Puneet Mathur , (2019), *Machine Learning Applications Using Python*

Python, https://www.python.org/

Ravisankar P., Ravi, Raghava R.G., Bose I., (2011), *Detection of financial statement fraud and feature selection using data mining techniques,* Decision Support Systems, 50(2), 491-500.

Singh M., Marco V., (1995), *Construction of Bayesian network structures from data: A brief survey and an efficient algorithm*, International Journal of Approximate Reasoning, 12(2), 111-131.
Spathis C., Doumpos M., Zopounidis C., (2002), *Detecting falsified financial statement: A comparative study using multicriteria analysis and multivariate statistical techniques,* The European Accounting Review, 11(3), 509–535.
Tarjo, Herawati N., (2015), *Application of Beneish M-Score Models and Data Mining to Detect Financial Fraud*, Procedia - Social and Behavioral Sciences, 211, 924-930.
Tsamardinos I., Brown E. L., Aliferis F. C.,(2006), *The max-min hill-climbing Bayesian network structure learning algorithm*, Kluwer Academic Publishers, 1-64.
Using ID3 Algorithm to build a Decision Tree to predict the weather, https://iq.opengenus.org/id3-algorithm/
Weka tutorial, https://wekatutorial.com/
WEKA-The workbench for machine learning, https://www.cs.waikato.ac.nz/ml/weka/
Βικιπαίδεια ή Wikipedia, Η ελεύθερη εγκυκλοπαίδεια, https://el.wikipedia.org/wiki/
Ελληνικά Ακαδημαϊκά Ηλεκτρονικά Συγγράμματα και Βοηθήματα/Κάλλιπος, https://repository.kallipos.gr/
Ελληνικό κείμενο των Διεθνών Πρότυπων Ελέγχου και Διεθνών Πρότυπων Δικλίδων Ποιότητας,(2010), *Διεθνή Ομοσπονδία Λογιστών (IFAC)*
Κάντζος Κ., Χονδράκη Α., (2006), *Ελεγκτική-Θεωρία και Πρακτική*, Αθήνα: Εκδόσεις Σταμούλη
Καραμάνης Κ., (2008), *Σύγχρονη Ελεγκτική*, Αθήνα: Εκδόσεις ΟΠΑ
Λογιστικός- Φορολογικός κόμβος ενημέρωσης, https://www.taxheaven.gr/ias
Νιάρχου Ν., (2004), *Χρηματοοικονομική ανάλυση λογιστικών καταστάσεων*, Αθήνα: Εκδόσεις Σταμούλη
Πραστάκος Γρ., (2006), *Διοικητική Επιστήμη- Λήψη επιχειρησιακών αποφάσεων στην κοινωνία της πληροφορίας*, Αθήνα: Εκδόσεις Σταμούλη
Σώμα Ορκωτών Ελεγκτών Λογιστών (ΣΟΕΛ), https://www.soel.gr/el/
Χρηματιστήριο Αθηνών, https://www.athexgroup.gr/el/