ΤΕΙ ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ

ΤΜΗΜΑ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ / ΜΕΣΟΛΟΓΓΙ

Πτυχιακή Εργασία

Πρόβλεψη Χρεοκοπίας Επιχειρήσεων με Χρήση Μεθόδων Μηχανικής Μάθησης

Κωστόπουλος Γεώργιος 16556

Επιβλέπων καθηγητής: Ντόβας Δημήτριος

Μεσολόγγι, Οκτώβριος 2018

ΤΕΙ ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ

ΤΜΗΜΑ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ / ΜΕΣΟΛΟΓΓΙ

Πτυχιακή Εργασία

Πρόβλεψη Χρεοκοπίας Επιχειρήσεων με Χρήση Μεθόδων Μηχανικής Μάθησης

Κωστόπουλος Γεώργιος 16556

Επιβλέπων καθηγητής: Ντόβας Δημήτριος

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΠΕΡΙΛΗΨΗ

Η πρόβλεψη της χρεοκοπίας επιχειρήσεων θεωρείται ένα από τα σημαντικότερα οικονομικά προβλήματα και έχει αναλυθεί εκτενώς στη σχετική οικονομική βιβλιογραφία. Πληθώρα μεθόδων και αλγορίθμων επιβλεπόμενης μηχανικής μάθησης έχουν εφαρμοστεί τα τελευταία χρόνια στο πεδίο της οικονομίας στοχεύοντας στη δημιουργία προγνωστικών μοντέλων για την ανίχνευση της επιχειρηματικής αποτυχίας με αξιοσημείωτα αποτελέσματα όπως, για παράδειγμα, τα νευρωνικά δίκτυα και οι σύνθετες μέθοδοι. Η παρούσα εργασία διερευνά την αποτελεσματικότητα γνωστών μεθόδων ημι-επιβλεπόμενης και ενεργής μηχανικής μάθησης για την πρόβλεψη της χρεοκοπίας επιχειρήσεων χρησιμοποιώντας οικονομικά στοιχεία παρελθόντων ετών που αφορούν ένα δείγμα ελληνικών επιχειρήσεων. Η ημι-επιβλεπόμενη και η ενεργή μηχανική μάθηση αποτελούν σύγχρονους κλάδους της μηχανικής μάθησης, οι οποίες εκμεταλλεύονται ένα μικρό αριθμό ετικετοποιημένων δεδομένων μαζί με ένα μεγάλο σύνολο μη ετικετοποιημένων δεδομένων για τη βελτίωση της ακρίβειας των μοντέλων. Μεγάλο πλήθος πειραμάτων πραγματοποιούνται στην έρευνά μας διερευνώντας την ακρίβεια γνωστών μεθόδων ημι-επιβλεπόμενης και ενεργής μηχανικής μάθησης και αποδεικνύοντας την αποτελεσματικότητά τους σε σχέση με αντιπροσωπευτικές μεθόδους επιβλεπόμενης μάθησης.

Λέξεις κλειδιά: Εξόρυξη γνώσης, Μηχανική μάθηση, Επιβλεπόμενη μάθηση, Ημι-επιβλεπόμενη μάθηση, Ενεργή μηχανική μάθηση, Χρεοκοπία επιχειρήσεων, Πρόβλεψη.

ABSTRACT

The prediction of corporate bankruptcy has been addressed as an increasingly important financial problem and has been extensively analyzed in the accounting literature. Over recent years, several machine learning methods have been effectively applied to build accurate predictive models for detecting business failure with remarkable results, such as neural networks and ensemble methods. The present study investigates the effectiveness of semi-supervised and active learning methods for predicting bankruptcy using financial data from a set of Greek firms. Semi-supervised learning and active learning constitute emerging subfields of machine learning exploiting a small amount of labeled data together with a large pool of unlabeled ones to improve learning accuracy. Several experiments take place in our study comparing the accuracy measures of familiar semi-supervised and active learning methods and demonstrating their efficiency in contrast to representative supervised methods.

Keywords: Data mining, Machine learning, Supervised learning, Semi-supervised learning, Active learning, Bankruptcy, Prediction.

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΣΥΝΤΟΜΟΓΡΑΦΙΕΣ

ΒΔ	Βάση Δεδομένων
DM	Data Mining
ML	Machine Learning
MDA	Multivariate Discriminant Analysis

ΑΠΟΔΟΣΗ ΟΡΩΝ

Accuracy	Ακρίβεια
Active Learning	Ενεργός Μάθηση
Association Rules	Κανόνες Συσχέτισης
Artificial Intelligence	Τεχνητή Νοημοσύνη
Attribute	Μεταβλητή
Bagging	Εμφωλίαση
Bankruptcy	Χρεοκοπία
Bayes Net	Δίκτυο Bayes
Binary	Δυαδικός
Categorical	Κατηγορικός
Classification	Κατηγοριοποίηση
Cluster	Συστάδα
Cluster Analysis	Ανάλυση Συσταδοποίησης
Clustering	Συσταδοποίηση
Confidence	Εμπιστοσύνη
Confusion matrix	Πίνακας Σύγχυσης
Continuous	Συνεχής
Corporate Bankruptcy	Εταιρική Χρεοκοπία
Cross Validation	Διασταυρωμένη Επικύρωση
Data	Δεδομένα
Database	Βάση Δεδομένων
Dataset	Σύνολο Δεδομένων
Data Mining	Εξόρυξη Γνώσης

Data Warehouse	Αποθήκη Δεδομένων
Decision Tree	Δέντρο Απόφασης
Description	Περιγραφή
Discrete	Διακριτός
Example	Παράδειγμα
Feature	Χαρακτηριστικό
Group	Ομάδα
Instance	Περίπτωση, Στιγμιότυπο
Learning	Μάθηση
Logistic Regression	Λογιστική Παλινδρόμηση
Machine Learning	Μηχανική Μάθηση
Model	Μοντέλο
Multivariate Discriminant Analysis	Πολυμεταβλητή Ανάλυση Διαφοροποίησης
Neuron	Νευρώνας
Ordinal	Διατακτικός
Qualitative	Ποιοτικός
Quantitative	Ποσοτικός
Prediction	Πρόβλεψη
Random Trees	Τυχαία Δάση
Regression	Παλινδρόμηση
Rules	Κανόνες
Scatter Plot	Διάγραμμα Διασποράς
Semi-supervised	Ημι-επιβλεπόμενη
Supervised	Επιβλεπόμενη
Support Vector Machines	Μηχανές Διανυσμάτων Υποστήριξης

Univariate Analysis	Μονομεταβλητή Ανάλυση
Unsupervised	Μη-επιβλεπόμενη
Variable	Μεταβλητή

Εισαγωγή

Εξόρυξη Γνώσης: Μια Τεχνολογία Αιχμής

Μεγάλο πλήθος δεδομένων αποθηκεύονται καθημερινά σε βάσεις δεδομένων (databases) και αποθήκες δεδομένων (data warehouses). Μέσα σε αυτές τις τεράστιες ποσότητες δεδομένων κρύβονται χρήσιμες πληροφορίες για ανθρώπους και οργανισμούς. Το ζητούμενο είναι η εξαγωγή ουσιαστικών πληροφοριών και συμπερασμάτων τα οποία να είναι χρήσιμα και άμεσα αξιοποιήσιμα (Zaki, Meira, & Meira, 2014). Η ανάλυση αυτών των δεδομένων και η άντληση χρήσιμων πληροφοριών από αυτά καθίσταται πρακτικά αδύνατη χωρίς τη χρήση εξειδικευμένης τεχνολογίας και μεθόδων. Η Εξόρυξη Γνώσης αποτελεί το κατάλληλο εργαλείο για την ανακάλυψη γνώσης μέσα από μεγάλες ποσότητες δεδομένων με χρήση μεθοδολογιών κυρίως της Στατιστικής, της Πληροφορικής, της Μηχανικής Μάθησης και της τεχνολογίας των Βάσεων Δεδομένων (ΒΔ).

Η Εξόρυξη Γνώσης αποτελεί μια νέα δυναμική τεχνολογία αιχμής για την αποτελεσματικότερη ανάλυση των δεδομένων με μεγάλες δυνατότητες για πολύτιμες εμπορικές και επιστημονικές ανακαλύψεις και πληθώρα εφαρμογών σε πολλά σύγχρονα πεδία της καθημερινής ζωής αλλά και της επιστήμης και των επιχειρήσεων.

- **Ιατρική**

  Ιδιαίτερα τα τελευταία χρόνια η Εξόρυξη Γνώσης έχει σημαντικές εφαρμογές στον τομέα της ιατρικής επιστήμης, και ιδιαίτερα στη βιοϊατρική, τη γενετική και τη φαρμακευτική. Η Εξόρυξη Γνώσης χρησιμοποιείται στη γενετική και τη βιοϊατρική για τη χαρτογράφηση της σχέσης των μεταβολών του ανθρώπινου DNA και της προδιάθεσης σε διάφορες ασθένειες. Επιπρόσθετα χρησιμοποιείται στην ανάλυση των δεδομένων του DNA και στην αναζήτηση ομοιοτήτων και διαφορών της γονιδιακής ακολουθίας υγειών και βλαβερών ιστών, προκειμένου να βρεθούν οι διαφορές μεταξύ των δύο αυτών τύπων. Τέλος, χρησιμοποιείται στην αναπαράσταση πολύπλοκων δομών γονιδίων αλλά και συνδυασμούς αυτών και τη συσχέτισή τους με ασθένειες (Zhu, 2007).

- **Ανεπιθύμητη Ηλεκτρονική Αλληλογραφία (spam mails)**

  Μια σημαντική εφαρμογή των μεθόδων Μηχανικής Μάθησης είναι το αυτόματο φιλτράρισμα της ανεπιθύμητης αλληλογραφίας, δηλαδή ο χαρακτηρισμός ενός μηνύματος του ηλεκτρονικού ταχυδρομείου ως ανεπιθύμητου ή όχι (Han, Pei, & Kamber, 2011). Τα ανεπιθύμητα μηνύματα του ηλεκτρονικού ταχυδρομείου είναι μηνύματα συνήθως εμπορικού ή διαφημιστικού περιεχομένου τα οποία στέλνονται μαζικά στους παραλήπτες. Τα μηνύματα αυτά δεν έχουν ζητηθεί από τους παραλήπτες και δημιουργούν πολλά «*σκουπίδια*» στην ηλεκτρονική λίστα με τα μηνύματα του κάθε χρήστη. Οι αλγόριθμοι κατηγοριοποίησης χρησιμοποιούνται για τον χαρακτηρισμό ενός μηνύματος ως ανεπιθύμητου με άμεσο επακόλουθο τη διαγραφή του ή την τοποθέτησή του σε ειδικό φάκελο.

- **Μάρκετινγκ**

  Οι εφαρμογές εξόρυξης γνώσης βρίσκουν μεγάλη απήχηση στον τομέα του μάρκετινγκ. Αυτό συμβαίνει διότι μεγάλες εταιρίες χρησιμοποιούν συστήματα διαχείρισης δεδομένων έτσι ώστε να διαχειρίζονται τον μεγάλο αριθμό πελατών και οικονομικών στοιχείων. Η τεχνολογία της Εξόρυξης Γνώσης συνεισφέρει σημαντικά στον τομέα του μάρκετινγκ με την ανάλυση δεδομένων μιας επιχείρησης και την εξαγωγή χρήσιμων πληροφοριών για τη συμπεριφορά του πελάτη (Linoff & Berry, 2011).

- **Τηλεπικοινωνίες**

  Το διαρκώς μεταβαλλόμενο επιχειρηματικό περιβάλλον σε συνδυασμό με τον τεράστιο ανταγωνισμό της εποχής αναγκάζουν τις εταιρίες να αναζητούν νέους τρόπους ενίσχυσης της θέσης τους έναντι άλλων. Εφαρμόζουν τεχνικές έγκαιρης πρόβλεψης διακοπής υπηρεσιών από πελάτες, κατηγοριοποιούν τις απαιτήσεις τους και ομαδοποιούν τις συνήθειες τους έτσι ώστε να διατηρούν τους ήδη υπάρχοντες πελάτες αλλά και για να προσελκύσουν νέους. Αρκετές εταιρίες τηλεπικοινωνιών χρησιμοποιούν εξόρυξη πληροφορίας για να καθορίζουν τις τάσεις και τις ανάγκες των πελατών με βάση χαρακτηριστικά όπως το μέγεθος της οικογένειας, το μέσο όρο ηλικίας των μελών της οικογένειας και την τοποθεσία (McCarthy, 1997).

- **Εκπαίδευση**

  Τα τελευταία χρόνια έχουν πραγματοποιηθεί πολλές έρευνες σχετικά με τη χρήση τεχνικών Εξόρυξης Γνώσης και Μηχανικής Μάθησης στον τομέα της εκπαίδευσης με σκοπό την βελτίωση της ποιότητας της εκπαίδευσης και την ενίσχυση της διαδικασίας της μάθησης. Αυτό οδήγησε στη δημιουργία ενός επιστημονικού τομέα ο οποίος ονομάζεται Εξόρυξη Γνώσης από Εκπαιδευτικά Δεδομένα (Educational Data Mining). Ο τομέας αυτός ασχολείται με την εφαρμογή των τεχνικών της Εξόρυξης Γνώσης στο πεδίο της εκπαίδευσης με στόχο την ενίσχυση της εκπαιδευτικής διαδικασίας και της βελτίωσης της ποιότητας αυτής (Romero & Ventura, 2007).

- **Χρηματοοικονομικός Τομέας**

  Τα τελευταία χρόνια, ο ανταγωνισμός μεταξύ των επιχειρήσεων στον χρηματοοικονομικό τομέα έχει αυξηθεί σε μεγάλο βαθμό εξαιτίας της παρουσίας αρκετών τραπεζικών και ασφαλιστικών ιδρυμάτων. Η Εξόρυξη Γνώσης παρέχει τα κατάλληλα εργαλεία ώστε να αντιμετωπίσει πολλά επιχειρηματικά ζητήματα όπως την προώθηση και την πώληση προϊόντων ή παροχή υπηρεσιών, την προσέλκυση νέων πελατών καθώς και τη διατήρηση των υπαρχόντων, τον εντοπισμό οικονομικής απάτης και την ανάλυση πιστωτικού κινδύνου (McCarthy, 1997).

- **Επιχειρήσεις**

  Οι εφαρμογές της Εξόρυξης Γνώσης έχουν μεγάλη εφαρμογή στον τομέα των επιχειρήσεων. Με την ανάλυση των δεδομένων των αγορών των πελατών, μια επιχείρηση μπορεί να πάρει όλα εκείνα τα στοιχεία τα οποία μπορεί να οδηγήσουν στην ανάπτυξή της ή και να εμποδίσουν την καταστροφή της. Στοιχεία όπως η σωστότερη τοποθέτηση των προϊόντων προς πώληση, οι καμπάνιες προώθησης συγκεκριμένων προϊόντων, ο συσχετισμός των προϊόντων αγοράς των πελατών (καλάθι αγορών) αποτελούν βασικά στοιχεία για μια εταιρία στην προσπάθεια βελτιστοποίησης του κέρδους της (Liao, Chu, & Hsiao, 2012).

Σκοπός Εργασίας

Η πρόβλεψη της χρεοκοπίας επιχειρήσεων (corporate bankruptcy) αποτέλεσε και συνεχίζει να αποτελεί ένα από τα σημαντικότερα θέματα της οικονομικής επιστήμης και έχει απασχολήσει πολλούς επιστήμονες και ερευνητές. Ιδιαίτερα στη σημερινή εποχή κατά την οποία επικρατεί παγκόσμια οικονομική και χρηματοπιστωτική κρίση, η χρήση μεθοδολογιών και μοντέλων πρόβλεψης της εξέλιξης μιας επιχείρησης είναι περισσότερο απαραίτητες από ποτέ. Δεν είναι άλλωστε λίγες οι περιπτώσεις χρεοκοπίας και πτώχευσης άλλοτε οικονομικά εύρωστων και ανταγωνιστικών επιχειρήσεων.

Η δυνατότητα πρόβλεψης της οικονομικής δυσχέρειας μιας επιχείρησης ενδιαφέρει όλα τα εμπλεκόμενα μέρη: τους υπαλλήλους, τις τράπεζες και τα διάφορα χρηματοπιστωτικά ιδρύματα δανεισμού και επενδύσεων, τα ασφαλιστικά ταμεία, το κράτος, το κοινωνικό σύνολο, αλλά την ίδια την επιχείρηση. Οι συνέπειες μιας χρεοκοπίας είναι επαχθείς για όλους τους πιθανούς αποδέκτες, οι οποίοι και θα κληθούν τελικά να τις αντιμετωπίσουν και να σηκώσουν το βάρος των δυσβάστακτων αποτελεσμάτων της.

Η έγκαιρη πρόβλεψη της πιθανής μιας χρεοκοπίας μιας επιχείρησης σε συνδυασμό με τη λήψη κατάλληλων μέτρων για την αντιμετώπισή της έχει αποτελέσει αντικείμενο έρευνας από πολλούς επιστήμονες, με πρωτεργάτες τους Beaver (1966) και Altman (1968). Βασιζόμενοι σε πληθώρα χρηματοοικονομικών δεικτών, ασχολήθηκαν εκτεταμένα με το πρόβλημα της έγκαιρης πρόβλεψης της χρεοκοπίας επιχειρήσεων προτείνοντας συγκεκριμένα προγνωστικά μοντέλα. Μονομεταβλητές και πολυμεταβλητές αναλύσεις όπως η ανάλυση διαφοροποίησης και τα πιθανοτικά υποδείγματα, αλλά και μη πολλές παραμετρικές μέθοδοι έχουν χρησιμοποιηθεί ευρύτατα στο παρελθόν για την πρόβλεψη της χρεοκοπίας επιχειρήσεων με αξιοσημείωτα αποτελέσματα.

Μέσα σε αυτό το πλαίσιο, η παρούσα εργασία διερευνά την αποτελεσματικότητα γνωστών μεθόδων μηχανικής μάθησης για την πρόβλεψη της χρεοκοπίας επιχειρήσεων χρησιμοποιώντας οικονομικά στοιχεία παρελθόντων ετών που αφορούν ένα δείγμα 145 ελληνικών επιχειρήσεων. Πιο συγκεκριμένα, εξετάζουμε τη χρήση μεθόδων ημι-επιβλεπόμενης και ενεργής μηχανικής μάθησης για αυτό τον σκοπό, ενώ παράλληλα συγκρίνουμε την αποτελεσματικότητα αυτών των μεθόδων με γνωστές μεθόδους επιβλεπόμενης μηχανικής μάθησης. Τα αποτελέσματα πληθώρας πειραμάτων καταδεικνύουν

την υπεροχή των μεθόδων ημι-επιβλεπόμενης και ενεργής μηχανικής μάθησης στην πρόβλεψη της χρεοκοπίας επιχειρήσεων.

Δομή Εργασίας

Η παρούσα εργασία αποτελείται από τέσσερα κεφάλαια.

Αρχικά, γίνεται μια πρώτη προσέγγιση της Εξόρυξης Γνώσης ως μια νέα δυναμική τεχνολογία αιχμής με πλήθος εφαρμογών σε πολλούς επιστημονικούς κλάδους, ένας εκ των οποίων είναι ο κλάδος της οικονομίας. Στη συνέχεια, παρουσιάζεται ο σκοπός της εργασίας, ο οποίος αφορά στη διερεύνηση της αποτελεσματικότητας γνωστών μεθόδων Ημι-επιβλεπόμενης και Ενεργής Μηχανικής Μάθησης για την πρόβλεψη της χρεοκοπίας επιχειρήσεων χρησιμοποιώντας διάφορους χρηματοοικονομικούς δείκτες που αφορούν ένα δείγμα ελληνικών επιχειρήσεων.

Στο 1ο κεφάλαιο παρουσιάζεται και αναλύεται ο όρος της χρεοκοπίας ή πτώχευσης επιχειρήσεων αναφέροντας χαρακτηριστικά σημεία του Πτωχευτικό Κώδικα. Στη συνέχεια παρουσιάζονται παράγοντες που οδηγούν στην εταιρική χρεοκοπία και σημειώνονται τα αποτελέσματα της. Τέλος, γίνεται ιδιαίτερη αναφορά, τόσο στις βασικές μεθοδολογίες οι οποίες έχουν χρησιμοποιηθεί για τη δημιουργία προγνωστικών μοντέλων της εταιρικής χρεοκοπίας, όσο και σε αξιοσημείωτες έρευνες όπως είναι, για παράδειγμα, αυτές των Beaver, Altman και Ohlson, οι οποίες αποτελούν σημείο αναφοράς για τις νεότερες εργασίες και έρευνες.

Το 2ο κεφάλαιο αναφέρεται στην Εξόρυξη Γνώσης (Data Mining). Αρχικά αναφέρονται βασικές έννοιες και στη συνέχεια παρουσιάζονται ορισμοί οι οποίοι έχουν δοθεί κατά καιρούς για να προσδιορίσουν τον όρο «Εξόρυξη Γνώσης». Στην ουσία πρόκειται για μια διεπιστημονική μεθοδολογία για την ανάλυση μεγάλων συνόλων δεδομένων και την ανακάλυψη χρήσιμων προτύπων και κανόνων. Στη συνέχεια, επιχειρείται μια σύνδεση της Εξόρυξης Γνώσης με άλλους επιστημονικούς τομείς και μεθοδολογίες όπως, για παράδειγμα, η Στατιστική, η Τεχνητή Νοημοσύνη (Russell & Norvig, 2003) και η Πληροφορική. Τέλος, παρουσιάζονται οι κυριότερες τεχνικές της Εξόρυξης Γνώσης, όπως είναι η κατηγοριοποίηση, η ανάλυση συσχετίσεων και η παλινδρόμηση. Η κατηγοριοποίηση

αποτελεί την μέθοδο η οποία εφαρμόζεται στην παρούσα εργασία για την πρόβλεψη της χρεοκοπίας επιχειρήσεων.

Το 3ο κεφάλαιο αναφέρεται στη Μηχανική Μάθηση (Machine Learning). Αρχικά αναφέρονται ορισμοί οι οποίοι έχουν δοθεί κατά καιρούς για να προσδιορίσουν τον όρο «Μηχανική Μάθηση». Στη συνέχεια παρουσιάζονται οι τέσσερις βασικές κατηγορίες της Μηχανικής Μάθησης: Επιβλεπόμενη, Μη επιβλεπόμενη, Ημι-επιβλεπόμενη και Ενεργή Μηχανική Μάθηση, ανάλογα με το είδος των δεδομένων εκπαίδευσης τα οποία χρησιμοποιούνται (ετικετοποιημένα ή/και μη ετικετοποιημένα). Παράλληλα, παρουσιάζεται το πρόβλημα της κατηγοριοποίησης και στη συνέχεια αναφέρονται και αναλύονται χαρακτηριστικές μέθοδοι επιβλεπόμενης μάθησης, οι οποίοι εφαρμόζονται κυρίως σε προβλήματα κατηγοριοποίησης.

Στο 4ο κεφάλαιο παρουσιάζεται η πειραματική μελέτη της εργασίας. Γίνεται αναφορά στο δείγμα που χρησιμοποιήθηκε στα πειράματα και περιγράφονται οι χρηματοοικονομικοί δείκτες (μεταβλητές). Η πειραματική μελέτη αποτελείται από δύο βασικά στάδια. Στο πρώτο στάδιο γίνεται εφαρμογή γνωστών μεθόδων Ημι-επιβλεπόμενης Μηχανικής Μάθησης για την πρόβλεψη της εταιρική χρεοκοπίας και συγκρίνεται η αποτελεσματικότητα των μεθόδων με τις αντίστοιχες μεθόδους Επιβλεπόμενης Μάθησης. Στο δεύτερο στάδιο πραγματοποιείται εφαρμογή γνωστών μεθόδων Ενεργής Μηχανικής Μάθησης και συγκρίνεται η αποτελεσματικότητα τους με τις αντίστοιχες μεθόδους Επιβλεπόμενης Μάθησης.

Τέλος, παρουσιάζονται συμπεράσματα και προβληματισμοί που προέκυψαν κατά τη διάρκεια της μελέτης και εκπόνησης της εργασίας.

1 Χρεοκοπία-Πτώχευση Επιχειρήσεων

Η χρεοκοπία επιχειρήσεων (corporate bankruptcy) αποτέλεσε και συνεχίζει να αποτελεί ένα από τα σημαντικότερα θέματα της οικονομικής επιστήμης και έχει απασχολήσει πολλούς επιστήμονες και ερευνητές. Ιδιαίτερα στη σημερινή εποχή κατά την οποία επικρατεί παγκόσμια οικονομική και χρηματοπιστωτική κρίση, η χρεοκοπία αποτελεί καθημερινό αντικείμενο έρευνας λόγω των καταστροφικών συνεπειών που μπορεί να έχει για όλους τους εμπλεκόμενους.

1.1 Η Έννοια της Χρεοκοπίας

Ο όρος χρεοκοπία είναι συχνά συνώνυμος της πτώχευσης και αναφέρεται σε εκείνη την κατάσταση κατά την οποία κάποιο νομικό ή φυσικό πρόσωπο αδυνατεί να εξυπηρετήσει τις δανειακές του οφειλές, δηλαδή να πληρώσει τους πιστωτές του. Συγκεκριμένα, σύμφωνα με το Άρθρο 2, Παρ. 1 του υπ'αριθμ. 3588/2007 του Πτωχευτικού Κώδικα:

«*Πτωχευτική ικανότητα έχουν οι έμποροι, καθώς και οι ενώσεις προσώπων με νομική προσωπικότητα που επιδιώκουν οικονομικό σκοπό*».

Σύμφωνα με το Άρθρο 3, Παρ. 1 του Πτωχευτικού Κώδικα:

«*Σε πτώχευση κηρύσσεται ο οφειλέτης που αδυνατεί να εκπληρώνει τις ληξιπρόθεσμες χρηματικές υποχρεώσεις του κατά τρόπο γενικό και μόνιμο (παύση πληρωμών). Δεν αποτελούν εκπλήρωση των υποχρεώσεων οι πληρωμές που γίνονται με δόλια ή καταστρεπτικά μέσα*».

Επίσης:

«*Επαπειλούμενη αδυναμία εκπλήρωσης αποτελεί λόγο κήρυξης της πτώχευσης, όταν την κήρυξή της ζητεί ο οφειλέτης*» (Άρθρο 3, Παρ. 2).

Στο Άρθρο 1 του Πτωχευτικού Κώδικα αναφέρεται ο σκοπός της πτώχευσης, σύμφωνα με τον οποίο:

«Η πτώχευση αποσκοπεί στη συλλογική ικανοποίηση των πιστωτών του οφειλέτη με τη ρευστοποίηση της περιουσίας του ή με άλλο τρόπο που προβλέπεται από σχέδιο αναδιοργάνωσης και ιδίως με τη διατήρηση της επιχείρησής του».

1.2 Παράγοντες και Αποτελέσματα της Χρεοκοπίας Επιχειρήσεων

Η χρεοκοπία μιας επιχείρησης αποτελεί αποτέλεσμα πολλών παραγόντων, με σημαντικότερους (Παπαδομιχελάκης, 2016):

- Αδυναμία κάλυψης βραχυπρόθεσμων υποχρεώσεων.
- Αναποτελεσματική διοίκηση.
- Εσκεμμένη παύση πληρωμών και ανακριβής εμφάνιση οικονομικών στοιχείων της επιχείρησης.
- Οικονομική κρίση.
- Συμβάντα τα οποία δύνανται να επηρεάσουν αρνητικά την πορεία ενός οικονομικού κλάδου.

Όπως προαναφέρθηκε, η χρεοκοπία μιας επιχείρησης αποτελεί σοβαρό ζήτημα για διάφορους λόγους, με κυριότερους τους εξής (Storey & Greene, 2010):

- Η χρεοκοπία μιας επιχείρησης έχει συχνά καταστροφικές κοινωνικές και οικονομικές συνέπειες.
- Το κλείσιμο μιας επιχείρησης οδηγεί σε απώλεια των θέσεων εργασίας πολλών ατόμων.
- Το κλείσιμο μιας επιχείρησης έχει κοινωνικές και οικονομικές συνέπειες σε όσους έχουν δανείσει ή επενδύσει χρήματα στην επιχείρηση.
- Η χρεοκοπία μιας επιχείρησης επηρεάζει αρνητικά το κράτος, αφού η εξέλιξη μιας επιχείρησης αποτελεί δομικό στοιχείο της εγχώριας οικονομίας.

Όπως αναφέρεται στο Άρθρο 17, Παρ. 1 του Πτωχευτικού Κώδικα για τις συνέπειες της πτώχευσης:

«*Ο οφειλέτης από την κήρυξη της πτώχευσης στερείται αυτοδικαίως της διοίκησης (διαχείρισης και διάθεσης) της περιουσίας του (πτωχευτική απαλλοτρίωση), την οποία ασκεί μόνος ο σύνδικος. Μετά την κήρυξη της πτώχευσης, πράξεις διαχείρισης ή διάθεσης στοιχείων της πτωχευτικής περιουσίας από τον οφειλέτη ή προς αυτόν, χωρίς τη σύμπραξη του συνδίκου, είναι ανενεργείς και απαγορεύεται να καταχωρηθούν σε δημόσια βιβλία οποιασδήποτε φύσεως, χωρίς τη γραπτή έγκριση του συνδίκου. Η πτώχευση θεωρείται ότι έχει κηρυχθεί από την έναρξη της ημέρας κατά την οποία δημοσιεύεται η απόφαση που κηρύσσει την πτώχευση στο ακροατήριο*».

Στον Πίνακα 1 παρουσιάζονται οι πτωχευτικές αποφάσεις επιχειρήσεων, οι οποίες εκδόθηκαν στην Ελλάδα κατά το χρονικό διάστημα 1998-2017[1].

Πίνακας 1: Πτωχευτικές αποφάσεις που εκδόθηκαν το διάστημα 1998-2017

Έτος	Πτωχευτικές Αποφάσεις
1998	921
1999	886
2000	805
2001	700
2002	576
2003	516
2004	569
2005	612
2006	532
2007	524
2008	342
2009	368
2010	380
2011	474
2012	455
2013	437
2014	335
2015	206
2016	111
2017	114

[1] ΕΛΣΤΑΤ

Στον Πίνακα 2 παρουσιάζονται οι κηρυχθείσες πτωχεύσεις επιχειρήσεων, κατά κλάδο οικονομικής δραστηριότητας το έτος 2017[2] σε φθίνουσα σειρά.

Πίνακας 2: Κηρυχθείσες πτωχεύσεις επιχειρήσεων κατά κλάδο οικ.δραστηριότητας (2017)

Κλάδος Οικονομικής Δραστηριότητας	Σύνολο
Χονδρικό και λιανικό εμπόριο, επισκευή μηχανοκίνητων οχημάτων και μοτοσυκλετών	37
Μεταποίηση	22
Δραστηριότητες υπηρεσιών παροχής καταλύματος και υπηρεσιών εστίασης	21
Κατασκευές	10
Μεταφορά και αποθήκευση	5
Διοικητικές και υποστηρικτικές δραστηριότητες	4
Παροχή ηλεκτρικού ρεύματος, φυσικού αερίου, ατμού και κλιματισμού	2
Ενημέρωση και επικοινωνία	2
Χρηματοπιστωτικές και ασφαλιστικές δραστηριότητες	2
Επαγγελματικές, επιστημονικές και τεχνικές δραστηριότητες	3
Δραστηριότητες σχετικές με την ανθρώπινη υγεία και την κοινωνική μέριμνα	2
Τέχνες, διασκέδαση και ψυχαγωγία	1
Άλλες δραστηριότητες παροχής υπηρεσιών	3
Σύνολο	**114**

[2] ΕΛΣΤΑΤ

1.3 Μεθοδολογίες για την Πρόβλεψη της Χρεοκοπίας

Η πρόβλεψη της χρεοκοπίας επιχειρήσεων αποτελεί τα τελευταία χρόνια ένα από τα σημαντικότερα ερευνητικά ζητήματα στο πεδίο της οικονομικής επιστήμης και έχει απασχολήσει πολλούς επιστήμονες και ερευνητές. Πολλές έρευνες έχουν δημοσιευθεί σχετικά με τη διερεύνηση της αποτελεσματικότητας διαφόρων προγνωστικών μοντέλων στηριζόμενα σε βασικούς χρηματοοικονομικούς δείκτες. Οι βασικότερες στατιστικές μεθοδολογίες που έχουν διατυπωθεί κατατάσσονται κυρίως στις παρακάτω κατηγορίες (Παπαγεωργίου, 2008):

- Μονομεταβλητή Ανάλυση (Univariate Analysis)
- Πολυμεταβλητή Ανάλυση Διαφοροποίησης (Multivariate Discriminant Analysis)
- Γραμμικά Υποδείγματα Πιθανότητας υπό συνθήκη (Linear Probability Models)
- Λογαριθμική Ανάλυση (Τεχνική Logit)

Ο William Beaver (1966) ήταν ο πρώτος που εφάρμοσε και διερεύνησε την αποτελεσματικότητα της μεθόδου της Μονομεταβλητής Ανάλυσης για την πρόβλεψη της εταιρικής αποτυχίας (failure) χρησιμοποιώντας διάφορους χρηματοοικονομικούς δείκτες. Συγκεκριμένα, όρισε ως αποτυχία την αδυναμία μιας επιχείρησης να καλύψει τις ληξιπρόθεσμες οφειλές της, ενώ εξέταζε την προβλεπτική ικανότητα ενός δείκτη κάθε φορά. Οι χρηματοοικονομικοί δείκτες τους οποίους χρησιμοποίησε ο Beaver ήταν:

- Δανειακής Επιβάρυνσης (Debt Ratio)
- Αποδοτικότητα Ενεργητικού (Return on Assets)
- Καθαρό Κεφάλαιο Κίνησης (Net Working Capital)
- Γενικής Ρευστότητας (Current Ratio)
- Αποδοτικότητα Ιδίων Κεφαλαίων (Return on Equity)
- Ταμειακές Ροές προς Σύνολο Υποχρεώσεων (Cash Flow/Total Liabilities)

Ο Edward Altman (1968) ήταν ο πρώτος που εφάρμοσε τη μέθοδο της Πολυμεταβλητής Ανάλυσης Διαφοροποίησης (MDA) για την ταξινόμηση επιχειρήσεων σε δύο κατηγορίες: υγιείς και πτωχευμένες. Η μέθοδος της Πολυμεταβλητής Ανάλυσης Διαφοροποίησης αποτελεί μια στατιστική τεχνική η οποία χρησιμοποιείται σε προβλήματα κατηγοριοποίησης δύο η περισσοτέρων κλάσεων. Οι κατηγορίες (κλάσεις) αυτές έχουν οριστεί εκ των προτέρων και εμφανίζουν κοινά χαρακτηριστικά και ιδιότητες. Στόχος της μεθόδου είναι να βρεθεί

ένας γραμμικός συνδυασμός των χρησιμοποιούμενων μεταβλητών ο οποίος να προσδιορίζει το συνολικό βαθμό μιας επιχείρησης. Με βάση αυτό τον βαθμό, μια επιχείρηση ταξινομείται τελικά ως υγιής ή πτωχευμένη. Στην έρευνά του, ο Altman χρησιμοποίησε ένα δείγμα 66 επιχειρήσεων (33 υγιείς και 33 πτωχευμένες) και πέντε χρηματοοικονομικούς δείκτες καταλήγοντας στη γραμμική σχέση:

$$Z = 0,\,021X_1 + 0,\,014X_2 + 0,\,033X_3 + 0,\,006X_4 + 0{,}999X_5$$

όπου:

X_1: Κεφάλαιο Κίνησης προς Σύνολο Ενεργητικού

X_2: Παρακρατηθέντα Κέρδη προς Σύνολο Ενεργητικού

X_3: Κέρδη προ τόκων και φόρων προς Σύνολο Ενεργητικού

X_4: Τρέχουσα Αξία Μετοχών προς Λογιστική Αξία Συνολικών Υποχρεώσεων

X_5: Πωλήσεις προς Σύνολο Ενεργητικού

Ο τελικός βαθμός (Z-score) αποτελεί σταθμικό μέσο των μεταβλητών X_1, X_2, X_3, X_4, X_5 (οι αντίστοιχοι συντελεστές στάθμισης αναγράφονται στην προηγούμενη εξίσωση) και χρησιμοποιείται για την κατηγοριοποίηση των επιχειρήσεων σε υγιείς και πτωχευμένες. Συγκεκριμένα:

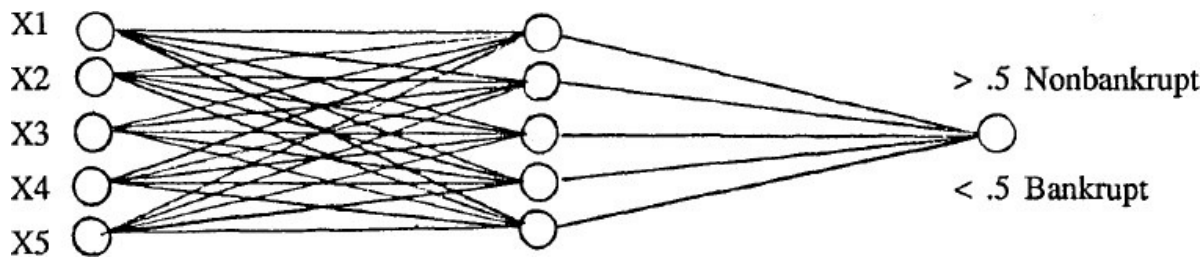
- Z>2.67 : Η επιχείρηση δεν κινδυνεύει άμεσα με αποτυχία εντός του τρέχοντος έτους
- Z<1.81 : Η επιχείρηση οδεύει προς αποτυχία εντός του τρέχοντος έτους
- 1.81<Z<2.67 : Δεν μπορούν να βγουν ασφαλή συμπεράσματα

Η εργασία του Altman και η προτεινόμενη εξίσωση αποτέλεσε και συνεχίζει να αποτελεί σημείο αναφοράς για εκατοντάδες εργασίες και έρευνες σχετικές με την πρόβλεψη της εταιρικής χρεοκοπίας.

Σε συνέχεια των προηγουμένων μελετών, ο James Ohlson (1980) παρουσίασε ορισμένα εμπειρικά αποτελέσματα από μια μελέτη πρόβλεψης της εταιρικής αποτυχίας με χρήση ενός πιθανοτικού μοντέλου (Ohlson, 1980). Τα αποτελέσματα της μελέτης αφορούσαν κυρίως στη χρήση τεσσάρων βασικών κατηγοριών δεικτών οι οποίοι επηρεάζουν στατιστικά σημαντικά την πιθανότητα πρόβλεψης και οι οποίοι είναι:

- Μέγεθος επιχείρησης
- Δείκτες ρευστότητας
- Δείκτες αποδοτικότητας
- Δείκτες αποδοτικότητας

Οι Odom και Sharda (1990) ανέπτυξαν ένα μοντέλο πρόγνωσης της εταιρικής χρεοκοπίας χρησιμοποιώντας νευρωνικά δίκτυα (Σχήμα 1) και σύγκριναν την αποτελεσματικότητα του μοντέλου με τη μέθοδο της Πολυμεταβλητής Ανάλυσης Διαφοροποίησης που πρότεινε ο Altman. Για τον σκοπό αυτό χρησιμοποίησαν τους ίδιους χρηματοοικονομικούς δείκτες και ένα δείγμα 129 επιχειρήσεων από τις οποίες 65 κατέληξαν σε χρεοκοπία. Τα αποτελέσματα της έρευνας έδειξαν την υπεροχή της προτεινόμενης μεθόδου έναντι της μεθόδου της Πολυμεταβλητής Ανάλυσης Διαφοροποίησης, η οποία αποτελεί ορόσημο και μέτρο σύγκρισης στις περισσότερες μελέτες.



Σχήμα 1: Νευρωνικό δίκτυο για την πρόβλεψη της χρεοκοπίας

Σε συνέχεια της προηγούμενης έρευνας, πληθώρα εργασιών έχουν εφαρμόσει προγνωστικά μοντέλα κάνοντας χρήση νευρωνικών δικτύων για την εταιρική χρεοκοπία με σημαντικά αποτελέσματα (Sharda & Wilson, 1996; Tam & Kiang, 1992; Wilson & Sharda, 1994).

Τα τελευταία χρόνια έχουν εφαρμοστεί γνωστοί μέθοδοι Μηχανικής Μάθησης για την πρόβλεψη της εταιρικής χρεοκοπίας δημιουργώντας προγνωστικά μοντέλα μεγάλης ακρίβειας και αποτελεσματικότητας. Σε μια πρόσφατη εργασία, προτάθηκε μια σύνθετη μηχανική μέθοδος αποτελούμενη από τους αλγόριθμους RIPPER και Naïve Bayes (Deligianni & Kotsiantis, 2012). Αξιοσημείωτο αποτελεί το γεγονός ότι η συγκεκριμένη μέθοδος ήταν αρκετά αποτελεσματική πολύ πριν το χρονικό σημείο της χρεοκοπίας, γεγονός που δίνει τη δυνατότητα λήψης κατάλληλων μέτρων για την αποφυγή της χρεοκοπίας.

Σε μια πρόσφατη εργασία επιχειρήθηκε η πρόβλεψη της οικονομικής δυσχέρειας Γαλλικών επιχειρήσεων χρησιμοποιώντας Μηχανές Διανυσμάτων Υποστήριξης, Νευρωνικά Δίκτυα, Λογαριθμική Ανάλυση και τη μέθοδο των Μερικών Ελαχίστων Τετραγώνων (Mselmi, Lahiani, & Hamza, 2017). Παράλληλα, προτάθηκε μια νέα υβριδική μέθοδος συνδυάζοντας τη μέθοδο των Μηχανών Διανυσμάτων Υποστήριξης με αυτή των Μερικών Ελαχίστων Τετραγώνων. Τα αποτελέσματα των πειραμάτων έδειξαν ακρίβεια 94.28% στην πρόβλεψη των «επικίνδυνων» επιχειρήσεων δύο χρόνια πριν το σημείο χρεοκοπίας.

2 Εξόρυξη Γνώσης

2.1 Βασικές Έννοιες

2.1.1 Μεταβλητές

Ως μεταβλητή εννοούμε ένα χαρακτηριστικό ή μια ιδιότητα των ατόμων ενός πληθυσμού το οποίο μεταβάλλεται, και το οποίο μπορεί να διαφοροποιείται από άτομο σε άτομο ή ακόμη και στο ίδιο άτομο ανάλογα με τη χρονική στιγμή μέτρησης της τιμής του (Tan, Steinbach, & Kumar, 2013). Για παράδειγμα, το χρώμα των ματιών διαφέρει από άτομο σε άτομο, με τιμές {μαύρα, καστανά, γαλάζια, πράσινα,…}, ενώ η θερμοκρασία σε μια περιοχή μεταβάλλεται κάθε χρονική στιγμή παίρνοντας μια συγκεκριμένη αριθμητική τιμή κάθε φορά. Οι μεταβλητές (attributes, variables) ή χαρακτηριστικά (features) διακρίνονται σε δύο βασικές κατηγορίες: ποσοτικές και ποιοτικές (Zaki, Meira, & Meira, 2014).

- **Ποσοτικές (quantitative)** ονομάζονται οι μεταβλητές οι οποίες παίρνουν μόνο αριθμητικές τιμές. Διακρίνονται σε:
    - **Διακριτές (discrete)**

      Είναι οι ποσοτικές μεταβλητές που μπορούν να πάρουν μόνο έναν ορισμένο αριθμό τιμών (συνήθως ακέραιους αριθμούς, χωρίς να μπορούν να πάρουν τις ενδιάμεσες τιμές), με το σύνολο των τιμών τους να είναι πεπερασμένο ή αριθμήσιμο. Για παράδειγμα, ο αριθμός των παιδιών μιας οικογένειας, το πλήθος συγκεκριμένων λέξεων που εμφανίζονται σε ένα έγγραφο, ο ταχυδρομικός κώδικας μιας περιοχής, ο προσωπικός αριθμός PIN μιας πιστωτικής κάρτας, κ.ά. Δεδομένα που αφορούν σε διακριτές μεταβλητές προέρχονται συνήθως από διαδικασίες καταμέτρησης ύστερα από παρατήρηση.

    - **Συνεχείς (continuous)**

      Είναι οι ποσοτικές μεταβλητές που μπορούν να πάρουν οποιαδήποτε τιμή σε ένα δεδομένο διάστημα πραγματικών αριθμών. Συνήθως δεν έχουν ελάχιστη μονάδα μέτρησης. Για παράδειγμα, το ύψος των μαθητών ενός σχολείου, το βάρος των αθλητών σε έναν αγώνα, ο χρόνος που αφιερώνει κάποιος μαθητής στο διάβασμα κάθε ημέρα, ο μισθός των υπαλλήλων μιας εταιρείας, η

ατμοσφαιρική πίεση, κ.ά. Δεδομένα που αφορούν σε συνεχείς μεταβλητές προέρχονται συνήθως από μετρήσεις.

- **Ποιοτικές (qualitative)** είναι οι μεταβλητές των οποίων οι τιμές τους δεν είναι αριθμοί, αλλά μπορούν απλώς να ταξινομούν αντικείμενα ή άτομα σε κατηγορίες ή ιδιότητες. Διακρίνονται σε:

    - **Κατηγορικές** ή **ονομαστικές (categorical)**

      Είναι οι ποιοτικές μεταβλητές που απλά ταξινομούν τα δεδομένα σε κατηγορίες, σαφώς διαχωρισμένες μεταξύ τους. Για παράδειγμα, η οικογενειακή κατάσταση με τιμές {*ελεύθερος, παντρεμένος, χωρισμένος, χήρος*}, η ομάδα αίματος με τιμές {A, B, AB, O}, το χρώμα των ματιών, κ.ά. Ειδική κατηγορία διακριτών μεταβλητών αποτελούν αυτές που παίρνουν μόνο δύο τιμές, για παράδειγμα: {0, 1}, {ναι, όχι}, {σωστό, λάθος}, {άνδρας, γυναίκα}, και οι οποίες χαρακτηρίζονται ως δυαδικές μεταβλητές (binary attributes).

    - **Διατακτικές (ordinal)**

      Είναι οι ποιοτικές μεταβλητές που καθορίζουν μια σειρά (ιεραρχία) μεταξύ των κατηγοριών. Για παράδειγμα, ο χαρακτηρισμός της επίδοσης στο σχολείο με τιμές {*άριστα, πολύ καλά, καλά, μέτρια, ανεπαρκώς*}, τα στάδια μιας ασθένειας με τιμές {I, II, III, IV}, ο χαρακτηρισμός του ύψους ατόμων με τιμές {κοντός, μέτριος, ψηλός}, κ.ά. Στις διατακτικές μεταβλητές υπάρχει η δυνατότητα σύγκρισης των τιμών τους (υπάρχει η έννοια της ισότητας εφόσον οι τιμές τους ταυτίζονται και της ανισότητας εφόσον οι τιμές τους είναι διαφορετικές). Για παράδειγμα, ένας μαθητής με χαρακτηρισμό επίδοσης άριστα είναι καλύτερης επίδοσης από κάποιον που χαρακτηρίζεται ως καλός (Zaki, Meira, & Meira, 2014).

Αξίζει να σημειωθεί εδώ ότι οι ιδιότητες των μεταβλητών δεν είναι απαραίτητο να ταυτίζονται με τις ιδιότητες των τιμών που παίρνουν αυτές οι μεταβλητές κάθε φορά (Tan, Steinbach, & Kumar, 2013). Για παράδειγμα, ας θεωρήσουμε δύο μεταβλητές που αφορούν τους υπαλλήλους μιας εταιρίας: «*Αριθμός Μητρώου*» και «*Ηλικία*» (σε έτη). Και οι δύο αυτές μεταβλητές είναι ποσοτικές και παριστάνονται με ακέραιους. Εντούτοις, είναι λογικό να μιλάμε για τη μέση ηλικία των υπαλλήλων της εταιρίας, αλλά όχι για τη μέση τιμή των αριθμών μητρώου των υπαλλήλων.

2.1.2 Σύνολα Δεδομένων

Τα σύνολα δεδομένων είναι συλλογές μετρήσεων (τιμών) προκαθορισμένων μεταβλητών (Zaki, Meira, & Meira, 2014). Συνήθως παριστάνονται με τη μορφή πινάκων διάστασης *nxm* (*n* γραμμών και *m* στηλών). Κάθε μια από τις *m* στήλες αντιστοιχεί σε μια μεταβλητή, ενώ κάθε γραμμή παριστάνει συγκεκριμένες τιμές μετρήσεων για τις μεταβλητές και αναφέρεται συνήθως ως στιγμιότυπο (επίσης χρησιμοποιούνται οι όροι παράδειγμα, περίπτωση ή εγγραφή). Γενικά, κάθε στιγμιότυπο ενός συνόλου δεδομένων παριστάνεται συνήθως με τη μορφή ενός διανύσματος διάστασης *m*, όπου *m* είναι το πλήθος των μεταβλητών:

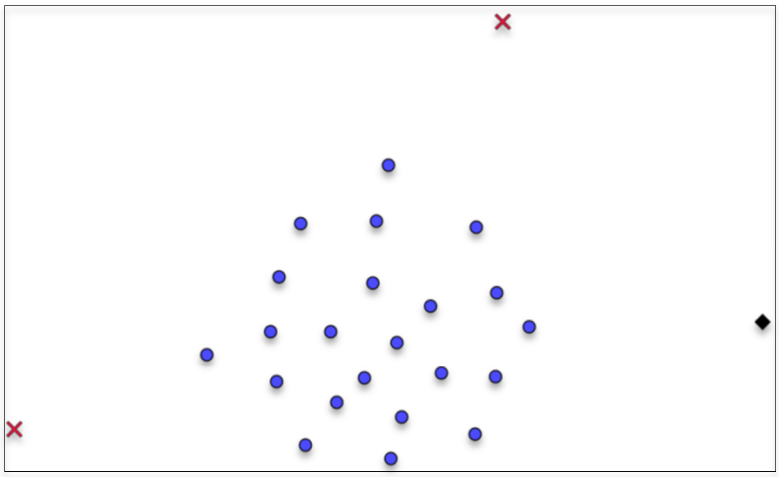
$$x = \left(x_1, x_2, \ldots, x_m\right)$$

Το πλήθος *m* των μεταβλητών που εμφανίζονται σε ένα σύνολο δεδομένων ορίζεται και ως διάσταση του συνόλου δεδομένων. Η μεγάλη διάσταση ενός συνόλου δεδομένων αποτελεί σημαντικό πρόβλημα κατά την ανάλυση αυτών. Έτσι, είναι σημαντικό κατά τη διαδικασία της προ-επεξεργασίας των δεδομένων να πραγματοποιηθεί μείωση της διάστασης με χρήση κατάλληλων τεχνικών (Tan, Steinbach, & Kumar, 2013). Τα αρχεία δεδομένων εμφανίζονται με διάφορες μορφές όπως, για παράδειγμα, σε μορφή αρχείων txt, xls, csv, dat, arff.

2.1.3 Προβλήματα σε Σύνολα Δεδομένων

Τα κυριότερα προβλήματα που εμφανίζονται κατά τη διαδικασία συλλογής δεδομένων είναι τα εξής (Han, Pei, & Kamber, 2011):

- **Ύπαρξη θορύβου και ακραίων περιπτώσεων**



Σχήμα 2: Ακραίες τιμές στα δεδομένα

Οι ακραίες περιπτώσεις αφορούν παρατηρήσεις οι οποίες διαφέρουν σημαντικά από το σύνολο των παρατηρήσεων (Σχήμα 2). Η εμφάνιση ακραίων περιπτώσεων σε ένα δείγμα μπορεί να προκύψει είτε τυχαία είτε μετά από εσφαλμένες μετρήσεις. Αυτές οι παρατηρήσεις είτε αφαιρούνται είτε χρησιμοποιούνται μέθοδοι οι οποίες δύνανται να τις χειριστούν. Πολλές φορές, τέτοιες περιπτώσεις αντιστοιχούν σε παρατηρήσεις οι οποίες είναι χρήσιμο να ανιχνευθούν, όπως είναι για παράδειγμα, ο εντοπισμός απάτης σε συστήματα ασφαλείας ή σε τραπεζικές συναλλαγές (Roiger & Geatz, 2003).

- **Μη καταχωρημένες τιμές μεταβλητών σε διάφορες εγγραφές**

  Εμφανίζονται συχνά περιπτώσεις εγγραφών στις οποίες απουσιάζουν οι τιμές κάποιων μεταβλητών. Αυτό μπορεί να συμβαίνει είτε λόγω μη μέτρησης της μεταβλητής κάποιας εγγραφής είτε λόγω καταστροφής είτε απώλειας λόγω λάθους. Εγγραφές με απουσία τιμών κάποιων μεταβλητών εμφανίζονται συχνά σε καθημερινά προβλήματα συγκέντρωσης δεδομένων. Τέτοιες περιπτώσεις αντιμετωπίζονται συχνά ως εξής (Han, Pei, & Kamber, 2011):

  - Αγνοώντας αυτές τις εγγραφές
  - Τοποθετώντας ως τιμή της μη καταχωρημένης μεταβλητής τη μέση τιμή ή τη διάμεσο των υπολοίπων εγγραφών που αφορούν στη συγκεκριμένη μεταβλητή
  - Συμπληρώνοντας με το χέρι τις ελλιπείς εγγραφές

- **Διπλότυπες εγγραφές**

  Σε πολλά σύνολα δεδομένων εμφανίζονται αρκετές φορές πανομοιότυπες εγγραφές. Σε αυτές τις περιπτώσεις γίνεται εκκαθάριση των διπλότυπων εγγραφών με χρήση κατάλληλων εργαλείων.

- **Μεγάλο πλήθος χαρακτηριστικών**

  Πολλά σύνολα δεδομένων εμφανίζουν μεγάλο πλήθος χαρακτηριστικών (μεγάλη διάσταση) τα οποία είναι δυνατόν να δημιουργήσουν προβλήματα κατά την εφαρμογή των μεθόδων Εξόρυξης Γνώσης και Μηχανικής Μάθησης. Έτσι, κρίνεται απαραίτητο να μειωθεί το πλήθος των χαρακτηριστικών του συνόλου επιλέγοντας ορισμένα από αυτά, ενώ παράλληλα να διατηρηθούν οι ουσιαστικές ιδιότητες των δεδομένων (Zaki, Meira, & Meira, 2014). Για τη μείωση του πλήθους των

χαρακτηριστικών ενός συνόλου δεδομένων έχουν αναπτυχθεί διάφορες μέθοδοι (Jolliffe, 2002; Schölkopf, Smola, & Müller, 1998).

2.2 Τι είναι Εξόρυξη Γνώσης;

Η Εξόρυξη Γνώσης (Data Mining) αποτελεί μια από τις πιο καινοτόμες και ριζοσπαστικές εξελίξεις στον τομέα της τεχνολογίας την τελευταία δεκαετία. Η εκρηκτική και ταχύτατη αύξηση του όγκου των αποθηκευμένων δεδομένων, κυρίως λόγω της αλματώδους εξέλιξης της τεχνολογίας, έχει κάνει επιτακτική την ανάγκη για επεξεργασία αυτών των δεδομένων και την εξαγωγή χρήσιμων πληροφοριών (Han, Pei, & Kamber, 2011). Η συνεχώς αυξανόμενη εξέλιξη στο πεδίο της εξόρυξης και ανακάλυψης γνώσης πυροδοτείται από τη συμβολή πολλών παραγόντων (Larose, 2014) με κυριότερους:

- Την εκρηκτική αύξηση δεδομένων καθημερινά
- Τη δυνατότητα αποθήκευσης αυτών των δεδομένων σε μεγάλες βάσεις δεδομένων
- Την ανάπτυξη εμπορικών και μη λογισμικών εξόρυξης γνώσης
- Την απίστευτη εξέλιξη της υπολογιστικής ισχύος των σύγχρονων υπολογιστών και τη δυνατότητας αποθήκευσης μεγάλου όγκου δεδομένων

Η Εξόρυξη Γνώσης αποτελεί το κατάλληλο εργαλείο για την ανάκτηση πληροφορίας από δεδομένα χρησιμοποιώντας μεθοδολογίες οι οποίες δεν είναι ικανές από μόνες τους να επεξεργαστούν αποτελεσματικά τα διαθέσιμα δεδομένα.

Αρκετοί ορισμοί έχουν δοθεί για τον όρο «*Εξόρυξη Γνώσης*». Οι περισσότεροι δίνουν έμφαση στην ανακάλυψη γνώσης με τη μορφή προτύπων μέσα από μεγάλο πλήθος δεδομένων, τα οποία είναι αποθηκευμένα σε βάσεις δεδομένων. Ο όρος «*πρότυπο*» αντανακλά τη μεθοδολογία της Μηχανικής Μάθησης, η οποία αποτελεί σημαντική συνιστώσα της Εξόρυξης Γνώσης. Στη συνέχεια, παρατίθενται με χρονολογική σειρά ορισμένοι από αυτούς τους ορισμούς:

> «*Εξόρυξη γνώσης είναι η διαδικασία ανάλυσης, συνήθως μεγάλων συνόλων δεδομένων, με σκοπό την εύρεση μη υπονοούμενων σχέσεων και τη σύνοψη των δεδομένων με καινοτόμους τρόπους έτσι ώστε να είναι κατανοητά και χρήσιμα*» (Hand, Mannila, & Smyth, 2001).

«Εξόρυξη Πληροφορίας (Data Mining) είναι η διαδικασία χρήσης μιας ή περισσότερων τεχνικών εκμάθησης υπολογιστών για την αυτόματη ανάλυση και εξαγωγή γνώσεων από δεδομένα που περιέχονται σε μια βάση δεδομένων. Ο σκοπός μιας συνεδρίας εξόρυξης πληροφορίας είναι να εντοπίσει τάσεις και πρότυπα (μοτίβα) στα δεδομένα.» (Roiger & Geatz, 2003).

«*Εξόρυξη γνώσης είναι η διαδικασία ανακάλυψης ενδιαφέρουσας γνώσης από μεγάλο πλήθος δεδομένων τα οποία είναι αποθηκευμένα σε βάσεις δεδομένων, αποθήκες δεδομένων ή άλλα αποθετήρια πληροφοριών*» (Han, Pei, & Kamber, 2011).

«*Εξόρυξη γνώσης είναι η διαδικασία εξερεύνησης και ανάλυσης μεγάλων ποσοτήτων δεδομένων, με στόχο την ανακάλυψη σημαντικών προτύπων και κανόνων*» (Linoff & Berry, 2011).

«*Εξόρυξη γνώσης είναι η διαδικασία αυτόματης ανακάλυψης χρήσιμης πληροφορίας σε μεγάλο πλήθος δεδομένων*» (Tan, Steinbach, & Kumar, 2013).

«*Εξόρυξη γνώσης είναι η μεθοδολογία εύρεσης προτύπων σε δεδομένα, πρότυπα τα οποία παρέχουν γνώση ή προσφέρουν δυνατότητες για γρήγορη και σωστή λήψη αποφάσεων*». (Witten, Frank, Hall, & Pal, 2016).

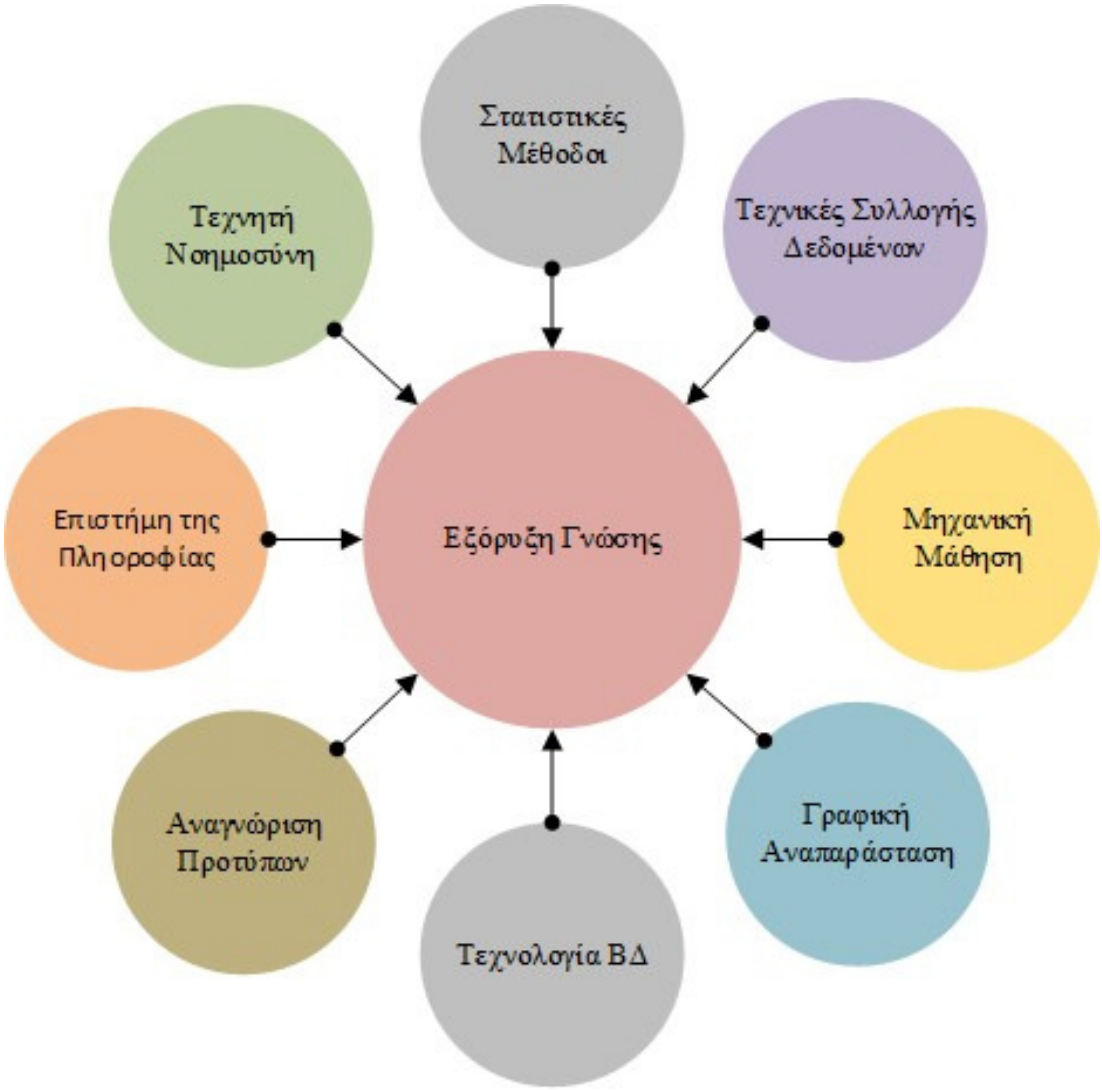
«*Εξόρυξη γνώσης είναι η διαδικασία εύρεσης ενδιαφερουσών δομών σε δεδομένα. Αυτές οι δομές μπορεί να είναι ένα σύνολο κανόνων, ένα γράφημα, ένα δίκτυο, ένα δέντρο απόφασης, εξισώσεις κ.ά.*» (Roiger, 2017).

Σύμφωνα με τους παραπάνω ορισμούς, η Εξόρυξη Γνώσης εφαρμόζεται σε δεδομένα τα οποία έχουν αποθηκευτεί σε κάποια βάση δεδομένων και τα οποία έχουν συλλεχθεί για διαφορετικό σκοπό. Αυτό σημαίνει ότι οι στόχοι της Εξόρυξης Γνώσης σε αυτά τα δεδομένα δεν σχετίζονται με τις μεθόδους και τις τεχνικές συλλογής αυτών, διαχωρίζοντάς τη σημαντικά από την Στατιστική, σύμφωνα με τις αρχές της οποίας τα δεδομένα συλλέγονται με τη χρήση συγκεκριμένων κάθε φορά μεθόδων για την απάντηση συγκεκριμένων ερωτημάτων (Hand, Mannila, & Smyth, 2001). Παρόλα αυτά, η Στατιστική αποτελεί βαρύνουσα και θεμελιώδη συνιστώσα της Εξόρυξης Γνώσης. Επομένως, η κεντρική ιδέα της Εξόρυξης Γνώσης είναι ότι υπάρχοντα δεδομένα περιέχουν πληροφορίες οι οποίες δύνανται

να αξιοποιηθούν μελλοντικά, ενώ ο βασικός στόχος της είναι η εύρεση προτύπων σε αυτά τα δεδομένα.

Με βάση όσα αναφέρθηκαν παραπάνω, είναι αναγκαίο να τονιστεί ότι Εξόρυξη Γνώσης **δεν είναι**:

- Οι στατιστικές μέθοδοι
- Η συγκέντρωση δεδομένων σε βάσεις δεδομένων ή αποθήκες δεδομένων
- Η απεικόνιση των δεδομένων σε διάφορες μορφές
- Οι αναφορές μετά από ερωτήματα σε βάσεις δεδομένων



Σχήμα 3: Συνιστώσες της Εξόρυξης Γνώσης

Η Εξόρυξη Γνώσης αποτελεί διεπιστημονική μεθοδολογία (Hand, Mannila, & Smyth, 2001). Πληροφορική, στατιστικές μέθοδοι, μηχανική μάθηση, τεχνητή νοημοσύνη, τεχνολογία βάσεων δεδομένων, αναγνώριση προτύπων, τεχνικές συλλογής δεδομένων, γραφικές αναπαραστάσεις, αποτελούν βασικές συνιστώσες της Εξόρυξης Γνώσης (Σχήμα 3), καθεμιά με τον δικό της διακριτό ρόλο. Επιπλέον, όσο δύσκολο είναι να οριστούν τα διαχωριστικά όρια μεταξύ αυτών των πεδίων, άλλο τόσο δύσκολο είναι να οριστούν και τα διαχωριστικά όρια μεταξύ αυτών και της Εξόρυξης Γνώσης.

Σύμφωνα με μια άλλη θεώρηση (Thearling, 1999) η Εξόρυξη Γνώσης απεικονίζεται ως η τομή τριών μεθοδολογιών (Σχήμα 4), οι οποίες είναι:

- Στατιστικές μέθοδοι και Αλγόριθμοι Μηχανικής Μάθησης
- Μέθοδοι Συλλογής, Αποθήκευσης και Διαχείρισης Δεδομένων
- Συστήματα μεγάλης Υπολογιστικής Ισχύος

Σχήμα 4: Η Εξόρυξη Γνώσης ως τομή πεδίων

2.3 Μέθοδοι Εξόρυξης Γνώσης

Η Εξόρυξη Γνώσης εφαρμόζεται για την επίλυση διαφόρων πρακτικών προβλημάτων στοχεύοντας κυρίως στην πρόβλεψη και την περιγραφή (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Με βάση αυτούς τους στόχους, η Εξόρυξη Γνώσης κατηγοριοποιείται αντίστοιχα σε δύο βασικές μεθόδους: Προβλεπτικές και Περιγραφικές (Σχήμα 5). Η κατηγοριοποίηση αυτή δεν είναι μοναδική, καθώς πολλές θεωρήσεις παρουσιάζουν και άλλες ταξινομήσεις, με κύριες συνιστώσες την πρόβλεψη και την περιγραφή.

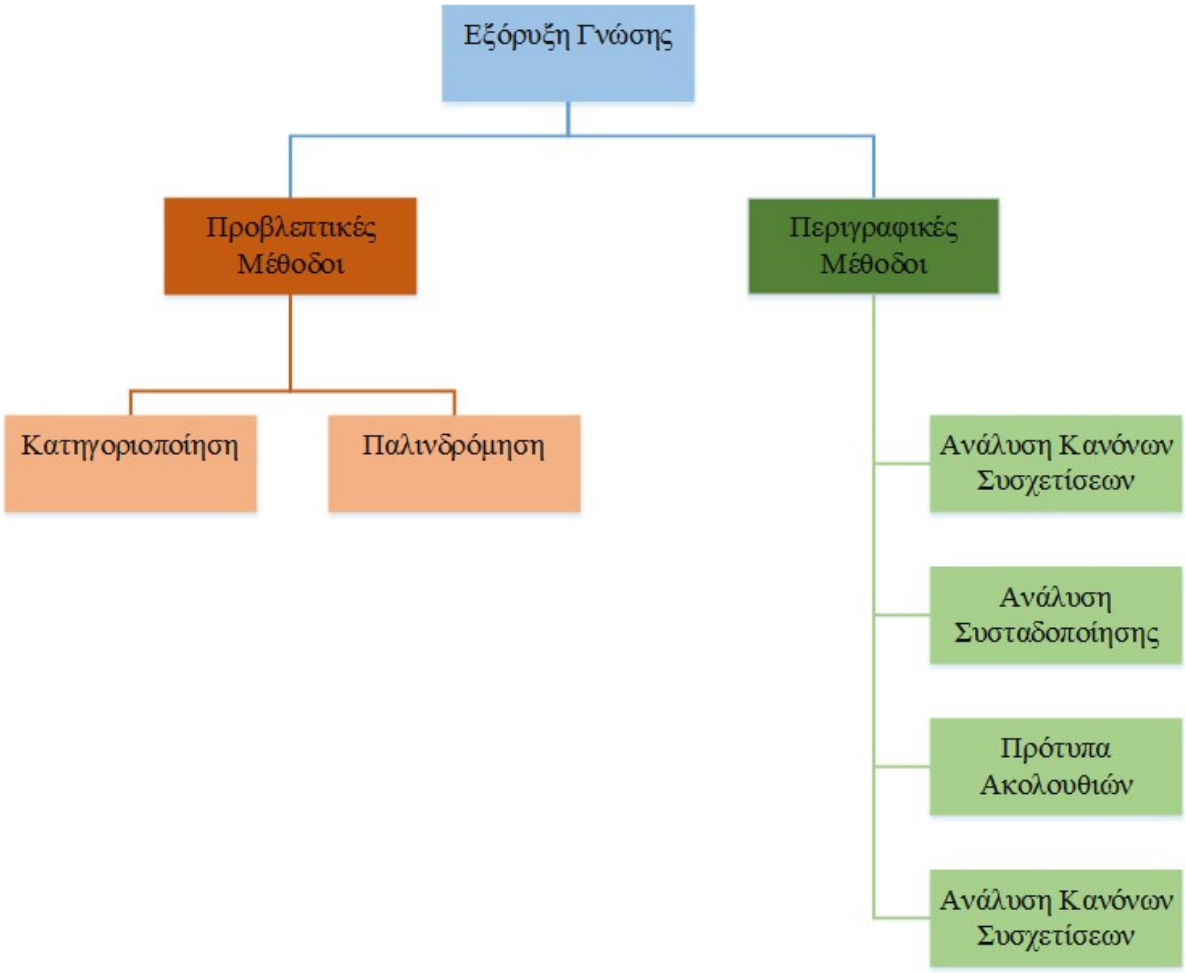
- **Προβλεπτικές μέθοδοι (Prediction methods)**

  Αφορούν στην εξαγωγή χρήσιμων συμπερασμάτων και εκτιμήσεων από διαθέσιμα δεδομένα. Πιο συγκεκριμένα, με χρήση διαφόρων μεταβλητών ή πεδίων μιας ΒΔ επιχειρείται η πρόβλεψη άγνωστων ή μελλοντικών τιμών κάποιας μεταβλητής, η οποία συνήθως ονομάζεται εξαρτημένη μεταβλητή ή μεταβλητή απόφασης (dependent-output variable). Πληθώρα μεθόδων Στατιστικής και Μηχανικής Μάθησης έχουν αναπτυχθεί για την επίλυση προβλημάτων που αφορούν στην πρόβλεψη των τιμών μιας μεταβλητής (Hand, Mannila, & Smyth, 2001) με κυριότερες αυτές της κατηγοριοποίησης ή ταξινόμησης και της παλινδρόμησης.

- **Περιγραφικές μέθοδοι (Description methods)**

  Επικεντρώνονται στην εύρεση προτύπων που αναπαριστούν τα δεδομένα μιας ΒΔ με όσο το δυνατό πιο κατανοητό και αξιοποιήσιμο τρόπο. Αντιπροσωπευτικές περιγραφικές μέθοδοι αποτελούν η συσταδοποίηση και η ανάλυση κανόνων συσχέτισης.

Τα όρια διαχωρισμού πρόβλεψης και περιγραφής δεν είναι απολύτως ξεκάθαρα και εξαρτώνται τόσο από τη φύση των δεδομένων, όσο και από το πεδίο εφαρμογής τους. Εντούτοις, αυτό που διαχωρίζει τις δύο μεθόδους είναι ότι οι προβλεπτικές μέθοδοι επικεντρώνονται σε μια μόνο μεταβλητή, τη μεταβλητή εξόδου, ενώ οι περιγραφικές μέθοδοι δεν επικεντρώνονται σε συγκεκριμένη μεταβλητή, αλλά προσπαθούν να περιγράψουν όλα τα δεδομένα (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).



Σχήμα 5: Ταξινόμηση μεθόδων Εξόρυξης Γνώσης

Οι διαδικασίες πρόβλεψης και περιγραφής της Εξόρυξης Γνώσης δύνανται να επιτευχθούν με τη χρήση διαφόρων τεχνικών, οι κυριότερες από τις οποίες είναι:

- Η Κατηγοριοποίηση (Classification)
- Η Παλινδρόμηση (Regression)
- Η Συσταδοποίηση (Clustering)
- Η Ανάλυση Κανόνων Συσχέτισης (Association Rules)
- Η Ανάλυση Εξαιρέσεων

2.3.1 Κατηγοριοποίηση (Classification)

Η κατηγοριοποίηση θεωρείται η δημοφιλέστερη τεχνική Εξόρυξης Γνώσης με εφαρμογές σε πολλά επιστημονικά πεδία. Γενικά, αφορά στη διαδικασία ταξινόμησης αντικειμένων σε προκαθορισμένες κατηγορίες, οι οποίες συχνά αναφέρονται και ως κλάσεις.

Ας υποθέσουμε ότι τα δεδομένα που αφορούν σε ένα πρόβλημα παρουσιάζονται με τη μορφή εγγραφών (*x*,*y*), όπου το διάνυσμα *x* αντιστοιχεί στο σύνολο των μεταβλητών του δείγματος, ενώ το y αντιστοιχεί στη μεταβλητή απόφασης (κλάση ή ετικέτα). Τότε ορίζουμε ως κατηγοριοποίηση:

> «*τη διαδικασία μάθησης μιας συνάρτησης στόχου f η οποία αντιστοιχεί κάθε διάνυσμα μεταβλητών x σε ένα προκαθορισμένο πλήθος κλάσεων y*» (Tan, Steinbach, & Kumar, 2013).

Η μεταβλητή *y* παίρνει συνήθως διακριτές ή κατηγορικές τιμές. Στην περίπτωση κατά την οποία η μεταβλητή *y* παίρνει οποιαδήποτε τιμή σε ένα διάστημα πραγματικών αριθμών [α, β] τότε έχουμε το πρόβλημα της παλινδρόμησης. Η συνάρτηση *f* αναφέρεται συχνά και ως μοντέλο ταξινόμησης ή κατηγοριοποίησης και χρησιμοποιείται συνήθως για περιγραφή ή πρόβλεψη. Για τη δημιουργία ενός μοντέλου ταξινόμησης απαιτείται ένα σύνολο δεδομένων (x,y) με γνωστές κλάσεις, το οποίο ονομάζεται σύνολο δεδομένων εκπαίδευσης.

- **Περιγραφή**

  Ένα μοντέλο ταξινόμησης μπορεί να χρησιμοποιηθεί ως επεξηγηματικό εργαλείο για τη διάκριση αντικειμένων που ανήκουν σε διαφορετικές κλάσεις. Για παράδειγμα, θα ήταν χρήσιμο ένα περιγραφικό μοντέλο το οποίο θα περιγράφει επακριβώς τα δεδομένα ενός πίνακα εγγραφών και το οποίο θα καθορίζει ποια χαρακτηριστικά (μεταβλητές) καθορίζουν τη μεταβλητή απόφασης (Tan, Steinbach, & Kumar, 2013).

- **Πρόβλεψη**

  Ένα μοντέλο ταξινόμησης μπορεί επίσης να χρησιμοποιηθεί για την πρόβλεψη των κλάσεων νέων εγγραφών (Tan, Steinbach, & Kumar, 2013). Συγκεκριμένα, δημιουργείται ένα μοντέλο ταξινόμησης και στη συνέχεια προσπαθούμε να προβλέψουμε την τιμή της μεταβλητής απόφασης. Επομένως, ένα προγνωστικό

μοντέλο για ανακάλυψη γνώσης από μια βάση δεδομένων, κάνει πρόβλεψη αγνώστων ή μελλοντικών τιμών κάποιων χαρακτηριστικών, βασιζόμενο στις άλλες τιμές που έχουν τα χαρακτηριστικά στην βάση δεδομένων. Παραδείγματα πρόβλεψης αποτελούν:

- Η πρόβλεψη επιτυχίας ή αποτυχίας ενός φοιτητή στις τελικές εξετάσεις ενός μαθήματος σύμφωνα με συγκεκριμένα ποιοτικά και ποσοτικά χαρακτηριστικά του φοιτητή κατά τη διάρκεια του πανεπιστημιακού έτους.
- Η πρόβλεψη χρεοκοπίας ή μη μιας εμπορικής επιχείρησης βασισμένη σε διάφορους χρηματοοικονομικούς δείκτες, όπως επιχειρείται στην παρούσα εργασία.

Συνοψίζοντας, η κατηγοριοποίηση αποτελεί τη βασικότερη μέθοδο Εξόρυξης Γνώσης και έχει τα εξής χαρακτηριστικά:

- Η εξαρτημένη μεταβλητή (μεταβλητή εξόδου ή απόφασης) είναι διακριτή ή κατηγορική.
- Η μάθηση είναι επιβλεπόμενη, με την έννοια ότι για κάθε εγγραφή του συνόλου εκπαίδευσης παρέχεται και η αντίστοιχη μεταβλητή απόφασης, η οποία συχνά ονομάζεται και κλάση.
- Στοχεύει στη δημιουργία μοντέλων τα οποία κατηγοριοποιούν νέες περιπτώσεις σε ένα σύνολο προκαθορισμένων κλάσεων.

Πολλές μέθοδοι έχουν χρησιμοποιηθεί για την κατηγοριοποίηση και την δημιουργία προγνωστικών μοντέλων όπως, για παράδειγμα (Tan, Steinbach, & Kumar, 2013):

- Δέντρα Αποφάσεων (Decision Trees-Classification Trees)
- Αλγόριθμοι Στατιστικής Κατηγοριοποίησης (Naïve Bayes Classifiers)
- Ταξινομητές του Πλησιέστερου Γείτονα (k-Nearest Neighbor Classifiers)
- Νευρωνικά Δίκτυα (Neural Networks)
- Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines)
- Παλινδρομικά Μοντέλα (Regression Models)
- Αλγόριθμοι Μάθησης Κανόνων (Rule-based Algorithms)

Στο Κεφάλαιο 3 αναλύονται όσες από τις προαναφερθείσες μεθοδολογίες και αλγόριθμους κατηγοριοποίησης εφαρμόζονται στην παρούσα εργασία.

2.3.2 Παλινδρόμηση (Regression)

Η Ανάλυση Παλινδρόμησης αποτελεί κλάδο της Στατιστικής που εξετάζει τη σχέση των τιμών μεταξύ δύο (Απλή Γραμμική Παλινδρόμηση-Linear Regression) ή περισσότερων μεταβλητών (Πολλαπλή Γραμμική Παλινδρόμηση-Multiple Linear Regression) στοχεύοντας στη πρόβλεψη μιας από αυτές (εξαρτημένη μεταβλητή) μέσω των άλλων (Gorunescu, 2011). Ο όρος «regression» προέρχεται από τον γενετιστή Francis Galton (1822-1911). Το πιο γνωστό παράδειγμα απλής παλινδρόμησης αποτελεί ίσως η εύρεση της σχέσης μεταξύ του ύψους και του βάρους ενός ατόμου και η πρόβλεψη του βάρους ενός ατόμου εφόσον δίνεται το ύψος του.

Σε αντίθεση με την απλή γραμμική παλινδρόμηση, στην οποία προσπαθούμε να εκφράσουμε την εξαρτημένη μεταβλητή σε σχέση με μια ανεξάρτητη μεταβλητή, στην πολλαπλή παλινδρόμηση υπάρχουν τουλάχιστον τρεις μεταβλητές, εκ των οποίων μια είναι η εξαρτημένη και οι υπόλοιπες οι ανεξάρτητες (Mendenhall, Sincich, & Boudreau, 2003). Ο όρος «multiple linear regression» πρωτοεμφανίστηκε το 1908 από τον Pearson. Στην περίπτωση αυτή, η εξαρτημένη μεταβλητή y εκφράζεται ως γραμμικός συνδυασμός των ανεξάρτητων μεταβλητών $x_1, x_2, \ldots, x_k$ (Downing & Clark, 2010):

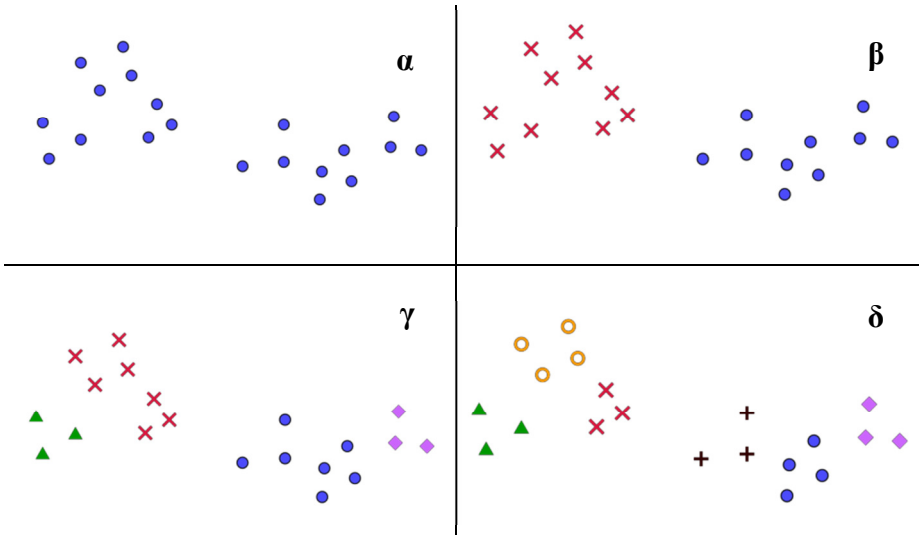
$$y=\alpha+b_1x_1+b_2x_2+\ldots+b_kx_k$$

Σημαντικές εφαρμογές της Ανάλυσης Παλινδρόμησης αποτελούν:

- Η πρόβλεψη των πωλήσεων ενός νέου προϊόντος με βάση της διαφημιστικές δαπάνες.
- Η πρόβλεψη της κατεύθυνσης και της ταχύτητας του ανέμου ως συνάρτηση της θερμοκρασίας, της υγρασία, της ατμοσφαιρικής πίεσης κ.ά.
- Η επίδραση του ύψους και του βάρους των γονέων στο ύψος και το βάρος των παιδιών.

2.3.3 Ανάλυση Συσταδοποίησης (Cluster Analysis)

Η Ανάλυση Συσταδοποίησης χωρίζει υπάρχοντα δεδομένα σε πεπερασμένο πλήθος ομάδων (groups) ή συστάδων (clusters) οι οποίες είναι σημαντικές ή χρήσιμες, βασιζόμενη σε πληροφορίες και σχέσεις που περιγράφουν αυτά τα δεδομένα. Η κεντρική ιδέα της μεθόδου αφορά στη δημιουργία ομάδων αντικειμένων έτσι ώστε τα αντικείμενα μιας ομάδας να παρουσιάζουν μεγάλη ομοιότητα, ενώ αντικείμενα διαφορετικών ομάδων να έχουν μεγάλη ανομοιότητα (Zaki, Meira, & Meira, 2014). Ο διαχωρισμός αντικειμένων σε ομάδες και η ταξινόμηση νέων αντικειμένων σε αυτά (κατηγοριοποίηση) αποτελεί σημαντική ανθρώπινη δραστηριότητα. Για παράδειγμα, τα μικρά παιδιά κατηγοριοποιούν αντικείμενα που βλέπουν σε εικόνες ή στην τηλεόραση ως ανθρώπους, ζώα, φυτά, οχήματα, κτίρια κ.ά. Σε αυτό το πλαίσιο, οι συστάδες αποτελούν ενδεχόμενες ομάδες και η συσταδοποίηση είναι η μελέτη και εφαρμογή τεχνικών για την αυτόματη αναζήτηση ομάδων (Tan, Steinbach, & Kumar, 2013).



Σχήμα 6: Συσταδοποίηση δεδομένων (περιπτώσεις)

Η εννοιολογική αποσαφήνιση του όρου συστάδα, καθώς και ο προσδιορισμός του πλήθους των δημιουργούμενων συστάδων σε ένα πρόβλημα δεν είναι καθόλου εύκολη διαδικασία. Ας θεωρήσουμε για παράδειγμα, ένα σύνολο δεδομένων το οποίο αποτυπώνεται γραφικά στο Σχήμα 6α και αποτελείται από συγκεκριμένα σημεία. Στο σχήμα αυτό παρουσιάζονται τρεις διαφορετικοί τρόποι διαχωρισμού αυτών των σημείων σε συστάδες. Συγκεκριμένα τα σημεία αυτά έχουν διαχωριστεί σε δύο, τέσσερις και έξι συστάδες (Σχήματα 6β, 6γ, 6δ). Είναι

φανερό ότι ο καθορισμός του πλήθους των συστάδων δεν αποτελεί εύκολη διαδικασία και εξαρτάται από τη φύση των δεδομένων που διαθέτουμε, το είδος του προβλήματος, το πεδίο εφαρμογής του, καθώς και τα επιθυμητά αποτελέσματα.

Η συσταδοποίηση μπορεί να θεωρηθεί σαν ένα είδος κατηγοριοποίησης χωρίς όμως να είναι γνωστές εκ των προτέρων οι κλάσεις (συστάδες) κατηγοριοποίησης και συχνά αναφέρεται ως «*μη επιβλεπόμενη κατηγοριοποίηση*».

Η μέθοδος της συσταδοποίησης έχει μεγάλες εφαρμογές σε ένα ευρύ φάσμα επιστημονικών πεδίων, όπως είναι η Βιολογία, η Ψυχολογία, η Μετεωρολογία και η Οικονομία. Για παράδειγμα:

- **Βιολογία**

  Οι βιολόγοι έχουν εφαρμόσει τεχνικές συσταδοποίησης για την ανάλυση μεγάλης ποσότητας γενετικού υλικού, όπως είναι για παράδειγμα η εύρεση ομάδων γονιδίων με παρόμοιες λειτουργίες ή η ιεραρχική κατηγοριοποίηση των ζωντανών οργανισμών.

- **Ψυχολογία**

  Μια ασθένεια παρουσιάζεται συχνά με διαφορετικές μορφές και η συσταδοποίηση μπορεί να χρησιμοποιηθεί για τον προσδιορισμό αυτών των μορφών, όπως είναι για παράδειγμα η κατάθλιψη.

- **Μετεωρολογία**

  Η συσταδοποίηση έχει εφαρμοστεί με μεγάλη επιτυχία για την εύρεση προτύπων στην ατμόσφαιρα και τους ωκεανούς με σκοπό την κατανόηση και την πρόβλεψη των κλιματικών αλλαγών στον πλανήτη.

- **Οικονομία**

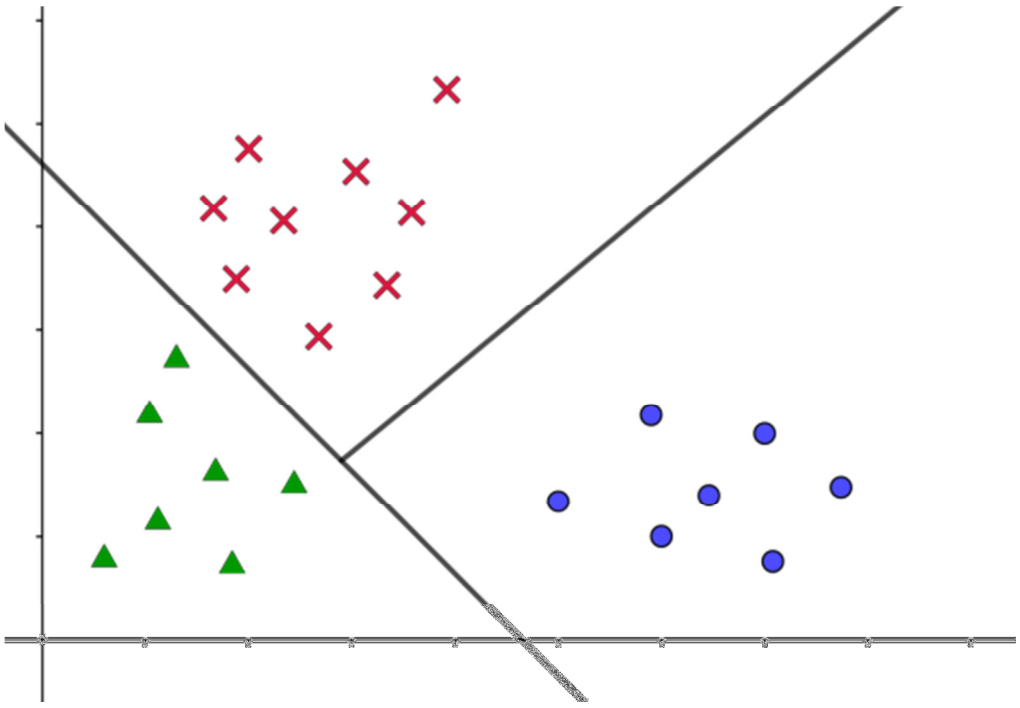
  Οι επιχειρήσεις συλλέγουν τεράστιες ποσότητες δεδομένων που αφορούν ενεργούς πελάτες τους. Η συσταδοποίηση χρησιμοποιείται για τον διαχωρισμό των πελατών σε μικρές υπο-ομάδες αγοραστών με παρόμοιες καταναλωτικές συνήθειες και αγοραστικές προτιμήσεις.

Μέθοδοι Συσταδοποίησης

Μεγάλο πλήθος μεθόδων έχουν χρησιμοποιηθεί στην Ανάλυση Συσταδοποίησης, καθεμιά με τα δικά της πλεονεκτήματα και μειονεκτήματα. Γενικά, διακρίνονται τρεις βασικές μέθοδοι συσταδοποίησης (Han, Pei, & Kamber, 2011): Διαχωρισμού (partitioning), ιεραρχικές (hierarchical) και μέθοδοι βασισμένες στην πυκνότητα (density-based).

- **Μέθοδοι Διαχωρισμού**

  Αφορούν στον απλό διαχωρισμό ενός συνόλου δεδομένων σε μη επικαλυπτόμενα υποσύνολα (συστάδες) έτσι ώστε κάθε αντικείμενο να περιέχεται ακριβώς σε μια συστάδα (Σχήμα 7). Συγκεκριμένα, n αντικείμενα διαχωρίζονται σε k ομάδες ($k \leq n$) και κάθε ομάδα περιέχει ένα τουλάχιστον αντικείμενο. Σημαντική παράμετρο στις μεθόδους διαχωρισμού αποτελεί ο αριθμός k των υπό σύσταση ομάδων, ο οποίος είναι γνωστός εκ των προτέρων.



Σχήμα 7: Συσταδοποίηση δεδομένων σε 3 ομάδες

- **Ιεραρχικές Μέθοδοι**

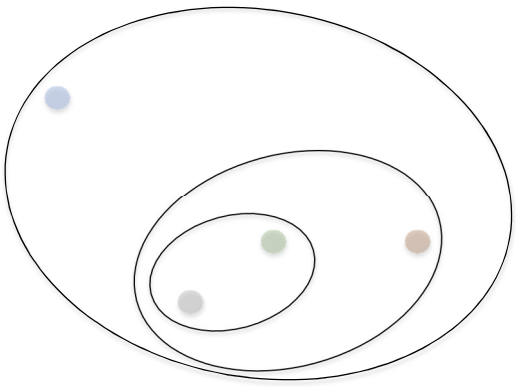
  Σε αυτές τις μεθόδους δημιουργείται μια ιεραρχική δομή από συστάδες, όπως είναι για παράδειγμα οργανωμένα τα αρχεία και οι φάκελοι στον σκληρό δίσκο κάποιου ηλεκτρονικού υπολογιστή (Jain & Dubes, 1988). Διακρίνονται σε:

- ***Διαμεριστικές (Divisive)***

  Σύμφωνα με αυτή τη μέθοδο όλες οι παρατηρήσεις τοποθετούνται αρχικά σε μια συστάδα, στη συνέχεια γίνεται διαχωρισμός της συστάδας σε δύο συστάδες, και η διαδικασία συνεχίζεται μέχρι κάθε συστάδα να αποτελείται από μια μόνο παρατήρηση. Σε αυτή τη μέθοδο είναι απαραίτητο να καθοριστεί το κριτήριο επιλογής της συστάδας που θα διαχωριστεί, καθώς και ο τρόπος διαχωρισμού της.

  Ένας από τους πιο αντιπροσωπευτικούς και ευρύτατα χρησιμοποιούμενους αλγόριθμους διαχωρισμού είναι αυτός των k-μέσων (MacKay, 2003). Σύμφωνα με αυτή τη μέθοδο, αρχικά επιλέγονται τυχαία k σημεία τα οποία αντιπροσωπεύουν τα κέντρα βάρους των συστάδων. Η παράμετρος k ορίζεται συνήθως από τον χρήστη και είναι ο αριθμός των επιθυμητών ομάδων. Καθένα από τα υπόλοιπα σημεία τοποθετείται στη συστάδα για την οποία γίνεται ελάχιστη η Ευκλείδεια απόσταση με το κέντρο βάρους των συστάδων. Στη συνέχεια, και για κάθε συστάδα, υπολογίζεται το νέο κέντρο βάρους και η διαδικασία επαναλαμβάνεται έως ότου οι συστάδες να παραμείνουν αμετάβλητες, δηλαδή τα κέντρα βάρους των συστάδων να είναι σταθερά. Αν και πρόκειται για μια απλή και αρκετά αποτελεσματική μέθοδο, πολλές φορές δεν είναι κατάλληλη για όλα τα είδη των δεδομένων (αφού η έννοια του κέντρου βάρους υφίσταται μόνο για ποσοτικά δεδομένα), ενώ επηρεάζεται και από την ύπαρξη ακραίων τιμών (Roiger & Geatz, 2003).

- ***Συσσωρευτικές (Agglomerative)***



Σχήμα 8: Συσσωρευτική ιεραρχική συσταδοποίηση

Σύμφωνα με αυτή τη μέθοδο, κάθε παρατήρηση τοποθετείται σε μια μοναδική συστάδα. Στη συνέχεια, ενώνονται οι δύο συστάδες με την ελάχιστη μεταξύ τους απόσταση και η διαδικασία συνεχίζεται μέχρι να δημιουργηθεί μια μόνο συστάδα, η οποία να περιέχει όλες τις παρατηρήσεις (Σχήμα 8). Σε αυτή τη μέθοδο είναι απαραίτητο να οριστεί ο τρόπος υπολογισμού της απόστασης μεταξύ των συστάδων.

- **Μέθοδοι Βασισμένες στην Πυκνότητα**

  Σύμφωνα με αυτές τις μεθόδους, κάθε συστάδα θεωρείται ως σφαιρική με ακτίνα r και εκτιμάται η πυκνότητα για ένα συγκεκριμένο σημείο του συνόλου δεδομένων η οποία αντιστοιχεί στο πλήθος των περιεχόμενων σημείων εντός μιας μεταβλητής ακτίνας. Αυτές οι μέθοδοι θεωρούνται κατάλληλες για την εύρεση οριακών σημείων ή σημείων θορύβου (Han, Pei, & Kamber, 2011).

Γενικά, η αποτελεσματικότητα μιας μεθόδου συσταδοποίησης εξαρτάται από πολλούς παράγοντες με βασικότερους τους εξής (Tan, Steinbach, & Kumar, 2013):

- Τον τρόπο καθορισμού του σωστού πλήθους συστάδων.
- Την ύπαρξη ακραίων τιμών και θορύβου στα δεδομένα.
- Τη διάσταση των δεδομένων.
- Τον τρόπο καθορισμού της ομοιότητας εντός μιας συστάδας, ο οποίος εξαρτάται κυρίως από τον τύπο των δεδομένων. Γνωστά μέτρα ομοιότητας που χρησιμοποιούνται είναι η Ευκλείδεια απόσταση, η απόσταση Manhattan και η μετρική Minkowski.

2.3.4 Κανόνες Συσχέτισης (Association Rules)

Η Ανάλυση Κανόνων Συσχέτισης ασχολείται με την εύρεση ενδιαφερουσών σχέσεων μεταξύ κάποιων χαρακτηριστικών ενός συνόλου δεδομένων (Gorunescu, 2011). Συνήθως, ένας κανόνας συσχέτισης εκφράζεται στη μορφή:

$$A \Rightarrow B$$

Τα Α, Β είναι σύνολα συνθηκών (λογικές εκφράσεις) που ικανοποιούν κάποια χαρακτηριστικά του συνόλου δεδομένων. Το Α χαρακτηρίζεται ως υπόθεση και το Β ως συμπέρασμα. Στην περίπτωση που το Β αναφέρεται στη μεταβλητή απόφασης (output variable) οι κανόνες συσχέτισης χαρακτηρίζονται ως κανόνες κατηγοριοποίησης. Γενικά, οι κανόνες συσχέτισης για ένα σύνολο δεδομένων δεν χρησιμοποιούνται όλοι μαζί ως ένα ενιαίο σύνολο κανόνων, απλά εκφράζουν συσχετίσεις μεταξύ των χαρακτηριστικών των δεδομένων και προβλέψεις μεταξύ μεταβλητών (Aggarwal & Yu, 1999).

Η αποτελεσματικότητα ενός κανόνα συσχέτισης καθορίζεται κυρίως από δύο μετρικές (Zaki, Meira, & Meira, 2014):

- Την υποστήριξη (support ή coverage), η οποία αντιστοιχεί στην πιθανότητα να πραγματοποιούνται συγχρόνως τα Α, Β, δηλαδή:

$$support = P(AB)$$

- Την εμπιστοσύνη (confidence ή accuracy), η οποία αντιστοιχεί στη δεσμευμένη πιθανότητα πραγματοποίησης του Β δεδομένου του Α (δηλαδή το ποσοστό των περιπτώσεων που προβλέπονται σωστά), δηλαδή:

$$confidence = P(B \mid A) = \frac{P(AB)}{P(A)}$$

Συνήθως καθορίζεται ένα ελάχιστο όριο υποστήριξης *minsup* (minimum support threshold) και ένα ελάχιστο όριο εμπιστοσύνης *minconf* (minimum confidence threshold), δηλαδή:

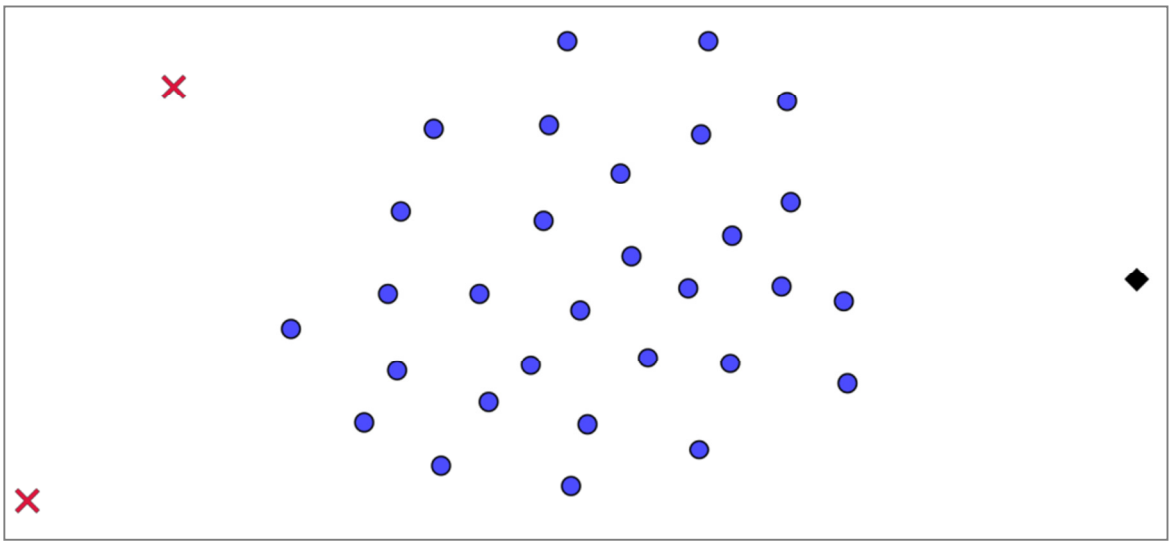
$$support \geq minsup$$

$$confidence \geq minconf$$

Οι κανόνες συσχέτισης που ικανοποιούν αυτές τις συνθήκες ταυτόχρονα χαρακτηρίζονται ως δυνατοί (Han, Pei, & Kamber, 2011). Δημοφιλείς αλγόριθμοι για την ανακάλυψη κανόνων συσχέτισης είναι ο Apriori (Agrawal & Srikant, 1994) και ο JRip (Cohen, 1995).

2.3.5 Ανάλυση Εξαιρέσεων

Η Ανάλυση Εξαιρέσεων ασχολείται με την εύρεση ακραίων περιπτώσεων, ανωμαλιών και παρεκκλίσεων στα δεδομένα ενός συνόλου. Στο Σχήμα 9 παρουσιάζεται ένα διάγραμμα διασποράς (scatter plot) που αφορά σε κάποιο σύνολο δεδομένων, στο οποίο σημειώνονται με διαφορετικό χρώμα (κόκκινο και μαύρο) οι ακραίες τιμές που εμφανίζονται σε αυτό.



Σχήμα 9: Διάγραμμα διασποράς για τον εντοπισμό ανωμαλιών σε δείγμα δεδομένων

Οι σημαντικότερες τεχνικές που χρησιμοποιεί η Ανάλυση Εξαιρέσεων είναι (Gorunescu, 2011):

- Γραφικές μέθοδοι, όπως είναι τα διαγράμματα διασποράς και τα θηκογράμματα.
- Στατιστικές μέθοδοι, οι οποίες εντοπίζουν αφύσικες παρατηρήσεις χρησιμοποιώντας διαφορετικά στατιστικά τεστ.
- Μέθοδοι βασισμένες σε μετρήσεις αποστάσεων, με βασικότερες αυτές της Συσταδοποίησης και της μεθόδου του κοντινότερου γείτονα.
- Μέθοδοι βασισμένες σε μοντέλα.

Σημαντικές εφαρμογές της Ανάλυσης Εξαιρέσεων είναι:

- Ο εντοπισμός απάτης σχετικά με πιστωτικές κάρτες με βάση το προφίλ των χρηστών
- Ο εντοπισμός εισβολής σε κάποιο σύστημα ασφαλείας
- Αποτροπή επίθεσης από ιούς
- Πρόβλεψη σφαλμάτων σε πληροφοριακά συστήματα

3 Μηχανική Μάθηση

Για την επίλυση ενός προβλήματος με τη χρήση υπολογιστή χρειαζόμαστε έναν αλγόριθμο, μια πεπερασμένη αλληλουχία βημάτων τα οποία θα πρέπει να εφαρμοστούν προκειμένου τα δεδομένα εισόδου να μετασχηματιστούν στην επιθυμητή έξοδο. Ας υποθέσουμε για παράδειγμα ότι θέλουμε να διατάξουμε σε αύξουσα σειρά ένα πλήθος αριθμών. Η είσοδος του προβλήματος είναι το σύνολο των αριθμών και η έξοδος είναι η διάταξη αυτών των αριθμών σε αύξουσα σειρά. Για την επίλυση του συγκεκριμένου προβλήματος μπορούν να εφαρμοστούν αρκετοί αλγόριθμοι, αλλά ίσως τελικά χρειαζόμαστε τον πιο αποδοτικό αλγόριθμο τόσο σε ταχύτητα, όσο και σε κατανάλωση μνήμης.

Υπάρχουν ωστόσο προβλήματα, στα οποία δεν είναι διαθέσιμος κάποιος αλγόριθμος, όπως για παράδειγμα στην περίπτωση που θέλουμε να διακρίνουμε τις χρεοκοπημένες από τις μη χρεοκοπημένες επιχειρήσεις. Στην περίπτωση αυτή έχουμε ως είσοδο ένα πλήθος επιχειρήσεων και γνωρίζουμε το αποτέλεσμα της εξόδου: χρεοκοπημένη/μη χρεοκοπημένη. Αυτό που δεν γνωρίζουμε είναι ο τρόπος (αλγόριθμος) σύμφωνα με τον οποίο τα δεδομένα εισόδου θα μετασχηματιστούν στην έξοδο. Στην πραγματικότητα, δεν γνωρίζουμε τον τρόπο με τον οποίο θα χαρακτηρίσουμε μια επιχείρηση ως χρεοκοπημένη ή μη.

Με την εξέλιξη στην επιστήμη και τεχνολογία των υπολογιστών, η έλλειψη γνώσης μπορεί να αντισταθμιστεί από την ύπαρξη πληθώρας δεδομένων, δηλαδή πληροφοριών που αφορούν καθεμιά επιχείρηση. Η χρήση των υπαρχουσών περιπτώσεων επιχειρήσεων οι οποίες έχουν χαρακτηριστεί ως χρεοκοπημένες μπορεί να βοηθήσει στην απόκτηση νέας γνώσης με τη βοήθεια ηλεκτρονικών υπολογιστών. Αν και δεν είμαστε σε θέση να κατανοήσουμε πλήρως τη διεργασία πίσω από αυτό τον μετασχηματισμό, μπορούμε να καταλήξουμε συχνά σε μια πολύ ακριβή και χρήσιμη προσέγγιση εντοπίζοντας συγκεκριμένα πρότυπα. Αυτό χαρακτηρίζει τελικά αυτό που αναφέρεται ως Μηχανική Μάθηση (Alpaydin, 2009).

Για τον προσδιορισμό του όρου «Μηχανική Μάθηση» έχουν δοθεί αρκετοί ορισμοί, όπως για παράδειγμα:

> «*Μηχανική Μάθηση είναι η μελέτη υπολογιστικών μεθόδων για την απόκτηση νέας γνώσης, νέων δεξιοτήτων και νέων τρόπων οργάνωσης της υπάρχουσας γνώσης*» (Carbonell & Gil, 1987).

«*Ένα πρόγραμμα υπολογιστή θεωρείται ότι μαθαίνει από την εμπειρία Ε ως προς κάποια κλάση εργασιών Τ και μέτρο απόδοσης Ρ, αν η απόδοσή του σε εργασίες από το Τ, όπως μετριέται από το Ρ, βελτιώνεται μέσω της εμπειρίας Ε*» (Mitchell, 1997).

«*Κάτι μαθαίνει όταν αλλάζει τη συμπεριφορά του κατά τέτοιο τρόπο ώστε να αποδίδει καλύτερα στο μέλλον*» (Witten, Frank, Hall, & Pal, 2016).

Η Μηχανική Μάθηση δεν είναι ένα απλό πρόβλημα διαχείρισης μεγάλου όγκου δεδομένων. Αποτελεί μέρος της Τεχνητής Νοημοσύνης. Για να είναι νοήμων ένα σύστημα, θα πρέπει να έχει την ικανότητα να μαθαίνει μέσα σε ένα διαρκώς μεταβαλλόμενο περιβάλλον. Εάν το σύστημα έχει την ικανότητα να μαθαίνει και να προσαρμόζεται σε αυτές τις μεταβολές, τότε ο σχεδιαστής του συστήματος δεν χρειάζεται να προβλέπει και να παρέχει λύσεις για όλες τις πιθανές καταστάσεις.

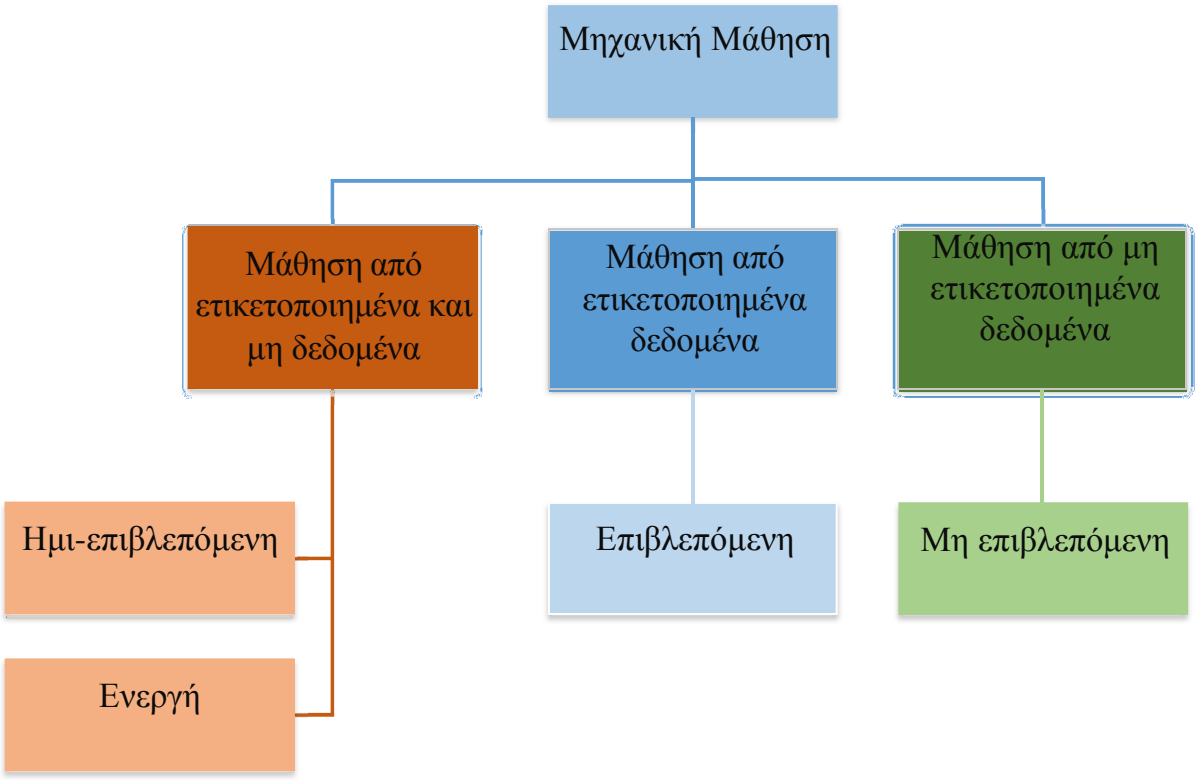
Σύμφωνα με τα παραπάνω, η Μηχανική Μάθηση είναι η εκτέλεση ενός προγράμματος με τη χρήση υπολογιστή με σκοπό τη μεγιστοποίηση των παραμέτρων ενός μοντέλου χρησιμοποιώντας δεδομένα εκπαίδευσης ή προηγούμενη εμπειρία (Alpaydin, 2009). Το μοντέλο μπορεί να χρησιμοποιηθεί για πρόβλεψη στο μέλλον ή/και για περιγραφή των δεδομένων του προβλήματος. Ο ρόλος της επιστήμης της Πληροφορικής στη Μηχανική Μάθηση είναι διττός:

- Κατά τη διάρκεια της εκπαίδευσης, χρειάζονται αποτελεσματικοί αλγόριθμοι για την μεγιστοποίηση των παραμέτρων που σχετίζονται με το πρόβλημα, καθώς και η απαραίτητη τεχνολογία για την αποθήκευση και επεξεργασία μεγάλου όγκου δεδομένων.
- Το μοντέλο θα πρέπει να είναι αρκετά αποτελεσματικό όσον αφορά την ακρίβεια και τον χρόνο της πρόβλεψης.

Η Μηχανική Μάθηση διαχωρίζεται ανάλογα με το είδος των δεδομένων του συνόλου εκπαίδευσης (ετικετοποιημένα ή/και μη ετικετοποιημένα) στις παρακάτω τέσσερις κατηγορίες (Σχήμα 10):

- Επιβλεπόμενη Μάθηση
- Μη Επιβλεπόμενη Μάθηση

- Ημι-επιβλεπόμενη Μάθηση
- Ενεργή Μηχανική Μάθηση



Σχήμα 10: Είδη Μηχανικής Μάθησης

3.1 Επιβλεπόμενη Μάθηση

Στην Επιβλεπόμενη Μάθηση (Supervised Learning) ή Μάθηση με Παραδείγματα (Learning from Examples) το σύστημα μάθησης λαμβάνει ένα σύνολο από δεδομένα εκπαίδευσης (training data), τα οποία αποτελούνται από ζεύγη της μορφής (x_i, y_i), i=1,2,…,n. Κάθε ζεύγος αποτελείται από ένα διάνυσμα τιμών x_i των χαρακτηριστικών εισόδου και την αντίστοιχη τιμή για την μεταβλητή απόφασης y_i (Zhu & Goldberg, 2009). Με βάση αυτά τα δεδομένα εκπαίδευσης δημιουργείται ένα μοντέλο κατηγοριοποίησης με σκοπό την πρόβλεψη της τιμής της μεταβλητής απόφασης y σε μελλοντικά δεδομένα x. Επομένως, η δημιουργία του μοντέλου είναι επαγωγική (inductive), ενώ η εφαρμογή του στην πρόβλεψη των τιμών νέων περιπτώσεων είναι συμπερασματική (deductive) (Tan, Steinbach, & Kumar, 2013).

Ανάλογα με τη μεταβλητή απόφασης y, τα προβλήματα επιβλεπόμενης μάθησης διαχωρίζεται σε προβλήματα κατηγοριοποίησης (η μεταβλητή y παίρνει διακριτές τιμές) και προβλήματα παλινδρόμησης (η μεταβλητή y παίρνει συνεχείς τιμές μέσα σε κάποιο διάστημα πραγματικών αριθμών). Με βάση τα προαναφερόμενα, η κατηγοριοποίηση και η παλινδρόμηση ορίζονται ως εξής:

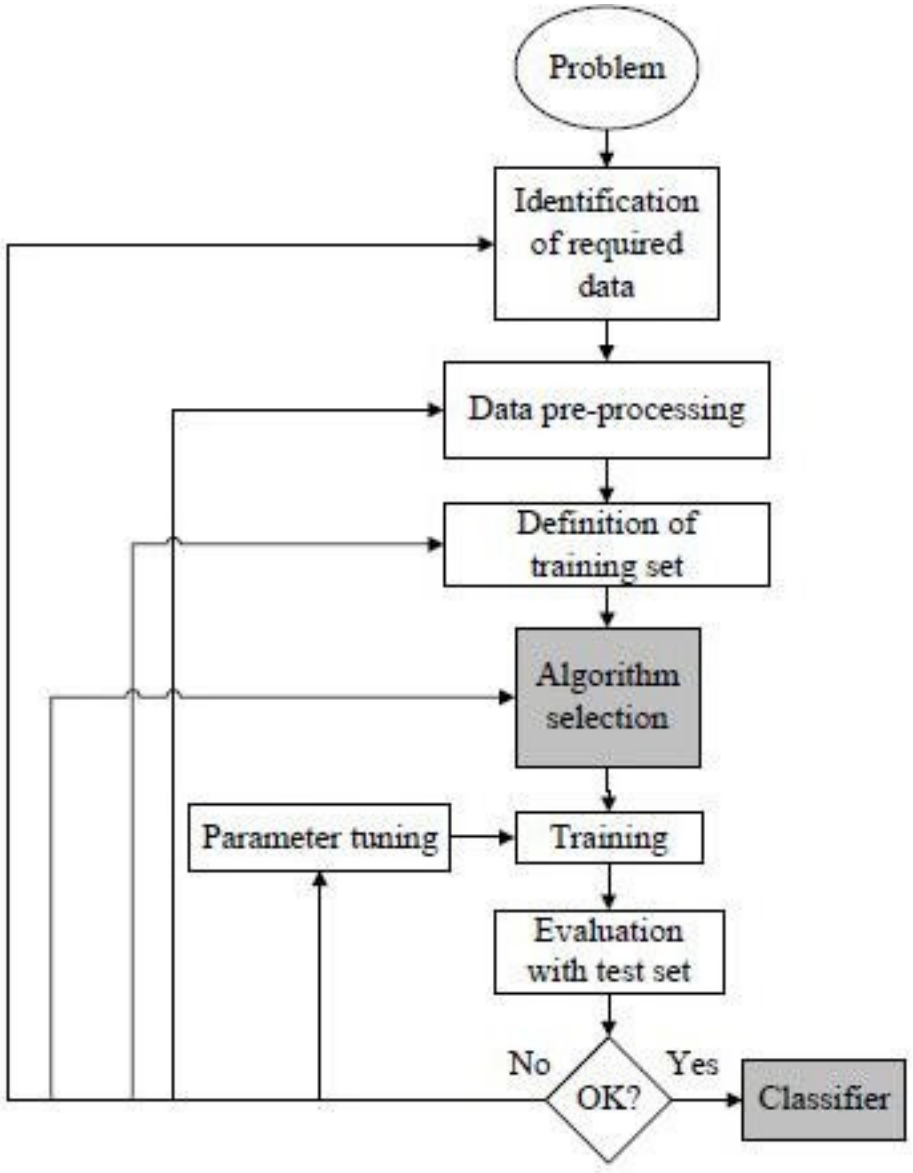
> *Κατηγοριοποίηση* είναι το πρόβλημα επιβλεπόμενης μάθησης με διακριτές κλάσεις y.
>
> *Παλινδρόμηση* είναι το πρόβλημα επιβλεπόμενης μάθησης με συνεχείς κλάσεις y.

Η εκτίμηση της αποτελεσματικότητας-ακρίβειας του μοντέλου σε ένα πρόβλημα επιβλεπόμενης μάθησης γίνεται χρησιμοποιώντας ένα ξεχωριστό σύνολο δεδομένων, το οποίο ονομάζεται δείγμα ελέγχου (test set). Για το δείγμα ελέγχου είναι γνωστή η κατηγοριοποίηση των περιπτώσεων που περιέχει, όμως δεν χρησιμοποιείται κατά τη διάρκεια της εκπαίδευσης. Έτσι, η ορθότητα της κατηγοριοποίησης των περιπτώσεων του δείγματος ελέγχου αποτελεί μια καλή εκτίμηση για την αποτελεσματικότητα του μοντέλου. Η διαδικασία της επιβλεπόμενης μηχανικής μάθησης φαίνεται στο επόμενο Σχήμα 11 (Kotsiantis, Zaharakis, & Pintelas, 2007). Σύμφωνα με το σχήμα, το πρόβλημα της επιβλεπόμενης Μηχανικής Μάθησης αποτελείται από τα εξής βήματα:

- Καθορισμός του προβλήματος.
- Αναγνώριση των δεδομένων του προβλήματος.
- Προ-επεξεργασία των δεδομένων.

- Καθορισμός του συνόλου εκπαίδευσης (ένα σύνολο από εγγραφές των οποίων οι κλάσεις είναι γνωστές).
- Επιλογή του κατάλληλου ταξινομητή (classifier).
- Εκπαίδευση του ταξινομητή με τη χρήση των δεδομένων του συνόλου εκπαίδευσης για την κατασκευή ενός μοντέλου κατηγοριοποίησης.

- Αξιολόγηση των αποτελεσμάτων στο σύνολο ελέγχου.

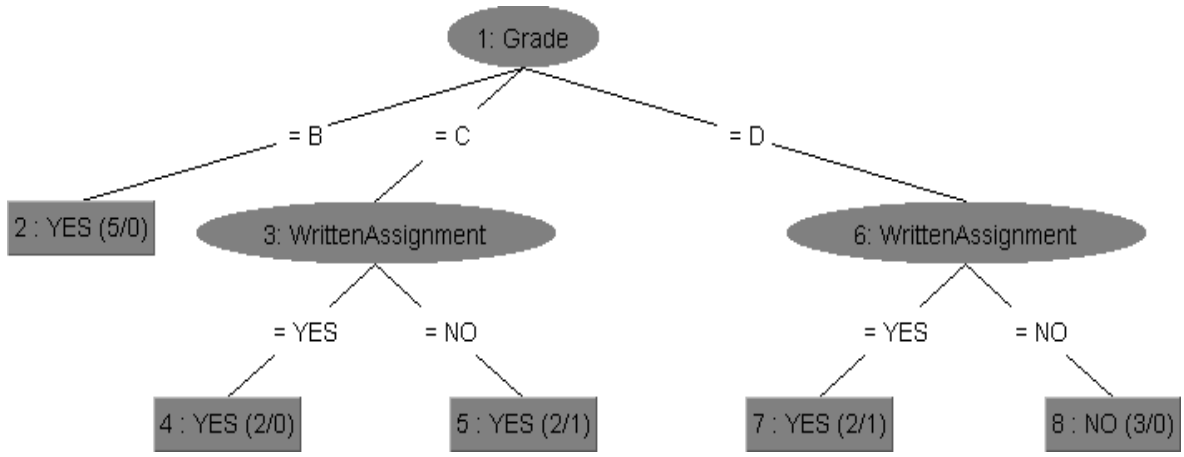


Σχήμα 11: Διαδικασία επιβλεπόμενης μηχανικής μάθησης

Χαρακτηριστικές μέθοδοι επιβλεπόμενης μάθησης είναι:

3.1.1 Δέντρα Αποφάσεων

Τα δένδρα αποφάσεων (Decision Trees) αποτελούν μία από τις πιο σημαντικές και διαδεδομένες μεθόδους για την κατηγοριοποίηση δεδομένων, στην οποία επιχειρείται η προσέγγιση μιας τιμής μιας κατηγορικής συνάρτησης απόφασης ακολουθώντας την τεχνική του «διαίρει και βασίλευε» (Roiger & Geatz, 2003). Ένα δέντρο απόφασης είναι μία γραφική απεικόνιση όλων των πιθανών διαδρομών που οδηγούν στο τελικό αποτέλεσμα (Σχήμα 12).



Σχήμα 12: Δέντρο απόφασης για εκπαιδευτικά δεδομένα

Ένα δένδρο απόφασης έχει τους εξής τύπους κόμβων (Tan, Steinbach, & Kumar, 2013):

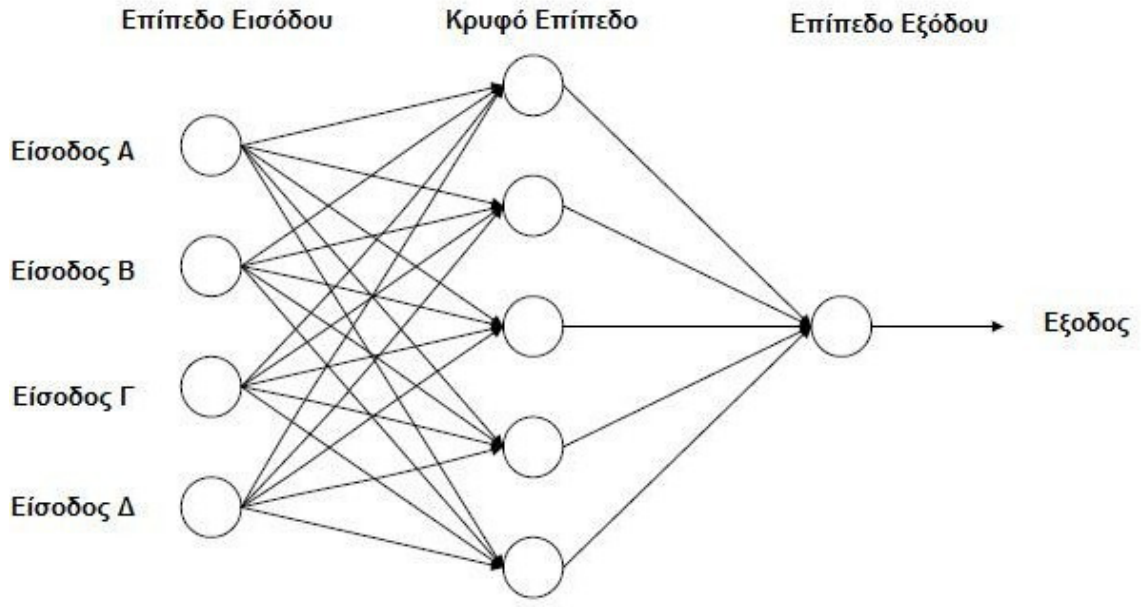
- Τον αρχικό κόμβο (ρίζα), ο οποίος δεν έχει εισερχόμενες ακμές.
- Τους εσωτερικούς κόμβους, οι οποίοι αντιστοιχούν σε μια μεταβλητή που χρησιμοποιείται για περαιτέρω διαχωρισμό του δένδρου. Στις εξερχόμενες ακμές από τον αρχικό ή κάθε εσωτερικό κόμβο, αντιστοιχεί μία συνθήκη ελέγχου με βάση την τιμή της μεταβλητής.
- Τους εξωτερικούς κόμβους (φύλλα), οι οποίοι αντιστοιχούν στα αποτελέσματα.

3.1.2 Νευρωνικά Δίκτυα

Τα Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks) αποτελούν μία χαρακτηριστική μέθοδο μοντελοποίησης σύνθετων προβλημάτων πρόβλεψης που αφορούν μεγάλο αριθμό εξαρτημένων μεταβλητών και μοντελοποιούνται σύμφωνα με τις λειτουργίες του ανθρώπινου εγκεφάλου (Gorunescu, 2011). Ο ανθρώπινος εγκέφαλος αποτελείται από μεγάλο πλήθος μονάδων επεξεργασίας (περίπου 10^{11}), οι οποίες ονομάζονται νευρώνες και λειτουργούν παράλληλα, ενώ κάθε νευρώνας διασυνδέεται με περίπου 10^4 άλλους νευρώνες.

Ένα νευρωνικό δίκτυο (Σχήμα 13) αποτελείται από τρείς βασικές κατηγορίες νευρώνων:

- Τους νευρώνες εισόδου (input neurons), οι οποίοι δέχονται τις πληροφορίες που θα υποστούν επεξεργασία
- Τους νευρώνες εξόδου (output neurons), στους οποίους καταλήγουν τα αποτελέσματα της παραπάνω επεξεργασίας.
- Τους ενδιάμεσους νευρώνες, οι οποίοι βρίσκονται μεταξύ των νευρώνων εισόδου και εξόδου και οι οποίοι ονομάζονται κρυφοί νευρώνες (hidden neurons). Το πλήθος των κρυφών νευρώνων καθορίζονται από τον χρήστη, όπως επίσης και το πλήθος των κόμβων τους.



Σχήμα 13: Παράδειγμα νευρωνικού δικτύου

3.1.3 Αλγόριθμος Bayes (Στατιστικής Κατηγοριοποίησης)

Αποτελεί μια απλή, γρήγορη και αρκετά αποτελεσματική μέθοδο ταξινόμησης η οποία χρησιμοποιεί πιθανοτικά μοντέλα τα οποία στηρίζονται στο θεώρημα του Bayes (Hanson, Stutz, & Cheeseman, 1991) σύμφωνα με το οποίο ισχύει ότι:

$$P(A \mid B) = \frac{P(A)\, P(B \mid A)}{P(B)}$$

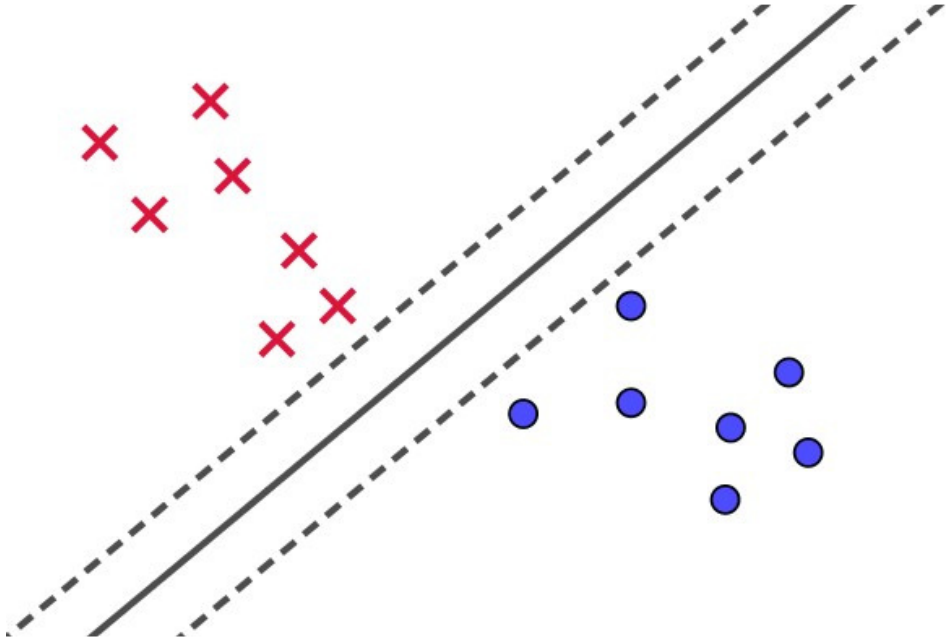
Όπου :

- $P(A|B)$ η δεσμευμένη πιθανότητα (a-posteriori probability) του ενδεχομένου A, δεδομένου του ενδεχομένου B.
- $P(A)$ είναι η πιθανότητα πραγματοποίησης του ενδεχομένου A και είναι γνωστή ως «*εκ των προτέρων πιθανότητα του A*» (a-priori probability).
- $P(B|A)$ είναι η δεσμευμένη πιθανότητα του ενδεχομένου B, δεδομένου του A. Η πιθανότητα αυτή είναι δυνατόν να υπολογιστεί από τη γνώση που διαθέτουμε για το συγκεκριμένο πρόβλημα.
- $P(B)$ είναι η πιθανότητα πραγματοποίησης του ενδεχομένου B.

Ο κατηγοριοποιητής Bayes χρησιμοποιείται για την εκτίμηση της πιθανότητας ενός στιγμιότυπου να ανήκει σε μια από τις προκαθορισμένες κλάσεις υπό την υπόθεση ότι τα χαρακτηριστικά είναι μεταξύ τους ανεξάρτητα. Η υπόθεση της ανεξαρτησίας των χαρακτηριστικών δεν ισχύει πάντοτε, όμως απλοποιεί κατά πολύ τους υπολογισμούς οδηγώντας σε καλή εκτίμηση της πιθανότητας χωρίς να απαιτεί μεγάλο σύνολο εκπαίδευσης.

3.1.4 Μηχανές Διανυσμάτων Υποστήριξης

Η χρήση αλγορίθμων βασισμένων στις Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines) αποτελεί μια ευρέως χρησιμοποιούμενη μέθοδο κατηγοριοποίησης, η οποία είναι κατάλληλη για δεδομένα πολλών διαστάσεων, ενώ μπορεί να χρησιμοποιηθεί και για κατηγορικές μεταβλητές (Cortes & Vapnik, 1995). Σύμφωνα με αυτή, επιλέγεται ένα μικρό σύνολο δεδομένων εκπαίδευσης από κάθε κλάση, τα οποία ονομάζονται διανύσματα υποστήριξης. Τα διανύσματα υποστήριξης συνορεύουν με αυτά άλλων κλάσεων (Σχήμα 14) και χρησιμοποιούνται για την εύρεση ενός υπερ-επιπέδου με το μεγαλύτερο περιθώριο γραμμικού διαχωρισμού τους (Gorunescu, 2011).



Σχήμα 14: Παράδειγμα ενός γραμμικά διαχωρίσιμου συνόλου δεδομένων

3.1.5 Μάθηση Βασισμένη σε Στιγμιότυπα

Οι μέθοδοι μάθησης βασισμένες σε στιγμιότυπα (Instance Based methods) στηρίζονται στην ομοιότητα μεταξύ των δεδομένων, η οποία υπολογίζεται με τη χρήση κατάλληλης μετρικής. Από τους γνωστότερους αλγόριθμους κατηγοριοποίησης βάσει στιγμιότυπων είναι αυτός των πλησιέστερων γειτόνων (Aha, 1997), ο οποίος χρησιμοποιείται και για προβλήματα παλινδρόμησης (Altman, 1992).

Ο αλγόριθμος των k πλησιέστερων γειτόνων (kNN) βρίσκει k στιγμιότυπα στο σύνολο εκπαίδευσης τα οποία βρίσκονται πλησιέστερα σε ένα συγκεκριμένο στιγμιότυπο. Στη συνέχεια αποδίδει το στιγμιότυπο σε εκείνη την κλάση η οποία υπερέχει μεταξύ των k στιγμιότυπων.

Σημαντικές παράμετροι οι οποίες καθορίζουν την αποτελεσματικότητα της μεθόδου αποτελούν:

- Το σύνολο των δεδομένων εκπαίδευσης.
- Η τιμή του k, η οποία αντιστοιχεί στο πλήθος των πλησιέστερων γειτόνων (στιγμιότυπων).
- Η μετρική που θα χρησιμοποιηθεί για τον υπολογισμό της απόστασης μεταξύ των στιγμιότυπων.

Αξίζει να σημειωθεί ότι η μέθοδος των k πλησιέστερων γειτόνων δύναται να χρησιμοποιηθεί για δύσκολα προβλήματα κατηγοριοποίησης με κατάλληλη τροποποίηση (Wu & et al.,, 2008).

3.2 Μη-Επιβλεπόμενη Μάθηση

Στην μη-επιβλεπόμενη μάθηση (Unsupervised Learning) το σύστημα μάθησης χρησιμοποιεί δεδομένα εκπαίδευσης για τα οποία δεν είναι γνωστές οι κλάσεις τους και προσπαθεί να τα χειριστεί με βάση ομοιότητες ή ανομοιότητες που μπορεί να παρουσιάζουν μεταξύ τους (Zhu & Goldberg, 2009). Επειδή δεν υπάρχουν δεδομένα με γνωστές κλάσεις είναι δύσκολο να γίνει ποσοτική αξιολόγηση της απόδοσης του συστήματος. Βασικά παραδείγματα προβλημάτων μάθησης χωρίς επίβλεψη αποτελούν:

- Η συσταδοποίηση (clustering), στην οποία τα δεδομένα διαχωρίζονται σε *n* ομάδες (συστάδες).
- Η ελάττωση διαστάσεων (dimensionality reduction), η οποία προσπαθεί να αναπαραστήσει κάθε περίπτωση των δεδομένων εκπαίδευσης με μικρότερο πλήθος χαρακτηριστικών, διατηρώντας όμως παράλληλα τις χαρακτηριστικές ιδιότητες των δεδομένων.
- Ο εντοπισμός καινοτομιών (novelty detection), στην οποία αναγνωρίζονται κάποιες περιπτώσεις (λίγες), οι οποίες διαφέρουν από την πλειοψηφία των περιπτώσεων.

3.3 Ημι-Επιβλεπόμενη Μάθηση

Στην ημι-επιβλεπόμενη μάθηση (Semi-Supervised Learning), το σύστημα μάθησης λαμβάνει ένα σύνολο δεδομένων εκπαίδευσης που αποτελείται από μικρό πλήθος δεδομένων με γνωστές τις κλάσεις τους και μεγάλο πλήθος δεδομένων χωρίς γνωστές κλάσεις και στη συνέχεια παράγει προβλέψεις για νέα δεδομένα (Zhu & Goldberg, 2009). Συγκεκριμένα, το πρόβλημα της ημι-επιβλεπόμενης μάθησης μπορεί να διατυπωθεί ως εξής:

> Έστω ένα σύνολο δεδομένων εκπαίδευσης, το οποίο αποτελείται από ένα μικρό σύνολο δεδομένων $L_d=\{(x_1,y_1),(x_2,y_2),\ldots,(x_l,y_l)\}$ με γνωστές τις κλάσεις τους και ένα μεγάλο σύνολο δεδομένων $U_d=\{x_{l+1},x_{l+2},\ldots,x_{l+u}\}$ με άγνωστες κλάσεις, με $x_i \in \mathrm{R}^N$. Ο σκοπός της ημι-επιβλεπόμενης μάθησης είναι διττός. Αρχικά, θέλουμε να δημιουργήσουμε ένα μοντέλο μάθησης βασισμένο στα δεδομένα εκπαίδευσης και στη συνέχεια να χρησιμοποιήσουμε το μοντέλο για την πρόβλεψη των τιμών μελλοντικών δεδομένων.

Η ημι-επιβλεπόμενη μάθηση εφαρμόζεται συχνά σε προβλήματα όπου είναι εύκολη η συλλογή δεδομένων χωρίς να είναι γνωστές οι κλάσεις, ενώ αντίθετα τα δεδομένα με γνωστές κλάσεις είναι δύσκολο να αποκτηθούν, είτε λόγω διότι απαιτείται πολύς χρόνος, είτε λόγου μεγάλου κόστους. Διάφοροι τύποι προβλημάτων όπως η ταξινόμηση, η πρόβλεψη τιμής, η ταξινόμηση με βάση κριτήριο μπορούν να αντιμετωπιστούν ως προβλήματα ημι-επιβλεπόμενης μάθησης. Πολλές μελέτες έχουν δείξει ότι ο συνδυασμός επιβλεπόμενης και μη επιβλεπόμενης μάθησης μπορεί να οδηγήσει στην αξιοποίηση δεδομένων με άγνωστες κλάσεις για τη δημιουργία μοντέλων μάθησης με καλύτερη απόδοση από αυτά που δημιουργούνται μέσω της επιβλεπόμενης μάθησης (Witten, Frank, Hall, & Pal, 2016).

Γνωστές μέθοδοι ημι-επιβλεπόμενης μάθησης είναι:

- **Self-training**.

  Η μέθοδος self-training θεωρείται από τις πιο απλές και αποτελεσματικές μεθόδους Ημι-επιβλεπόμενης Μάθησης (Yarowsky, 1995). Η συγκεκριμένη μέθοδος βασίζεται στις δικές της προβλέψεις σε μη ετικετοποιημένα δεδομένα για να εκπαιδευτεί. Αρχικά, ένας ταξινομητής εκπαιδεύεται σε μικρό πλήθος ετικετοποιημένων δεδομένων του συνόλου εκπαίδευσης και στη συνέχεια χρησιμοποιείται για να προβλέψει τις κλάσεις μη ετικετοποιημένων δεδομένων. Οι

περισσότερο σίγουρες προβλέψεις προστίθενται στο σύνολο των ετικετοποιημένων δεδομένων και η διαδικασία επαναλαμβάνεται για συγκεκριμένο πλήθος επαναλήψεων (Zhu and Goldberg, 2009).

- **Co-training**

  Η μέθοδος co-training έχει χρησιμοποιηθεί ευρέως σε πολλά προβλήματα Ημι-επιβλεπόμενης Μηχανικής Μάθησης (Blum & Mitchell, 1998). Η μέθοδος στηρίζεται στην υπόθεση ότι κάθε στιγμιότυπο του συνόλου δεδομένων μπορεί να διαχωριστεί σε δύο διακριτά σύνολα χαρακτηριστικών, τα οποία ονομάζονται πεδία. Καθένα από τα πεδία αυτά είναι επαρκές για σωστή κατηγοριοποίηση , ενώ είναι μεταξύ τους ανεξάρτητα. Σε αυτή τη βάση, δύο αλγόριθμοι μάθησης εκπαιδεύονται ξεχωριστά σε κάθε πεδίο με βάση τα ετικετοποιημένα δεδομένα εκπαίδευσης, οι πιο σίγουρες προβλέψεις καθενός στα μη ετικετοποιημένα δεδομένα προστίθενται στο σύνολο εκπαίδευσης του άλλου και η διαδικασία επαναλαμβάνεται για συγκεκριμένο πλήθος επαναλήψεων (Blum & Mitchell, 1998).

- **Tri-training**

  Η μέθοδος tri-training αποτελεί παραλλαγή της μεθόδου co-training, αλλά δεν απαιτεί την ύπαρξη δύο ανεξάρτητων πεδίων χαρακτηριστικών (Zhou & Li, 2005). Αντίθετα, στηρίζεται στη μέθοδο εμφωλίασης (bagging) σύμφωνα με την οποία δημιουργούνται αυτοδύναμα υποσύνολα του αρχικού συνόλου δεδομένων ίδιου μεγέθους μέσω μιας δειγματοληπτικής επαναληπτικής διαδικασίας (Breiman, 1996). Στη συνέχεια χρησιμοποιεί τρεις αλγόριθμους μάθησης οι οποίοι εκπαιδεύονται στα υποσύνολα αυτά. Στην ουσία πρόκειται για έναν αλγόριθμο εμφωλίασης τριών ταξινομητών (Hady & Schwenker, 2010). Αν δύο από τους αλγόριθμους συμφωνούν στην πρόβλεψη της κλάσης ενός μη ετικετοποιημένου στιγμιότυπου, τότε αυτό χρησιμοποιείται για την εκπαίδευση του τρίτου αλγόριθμου.

- **Democratic Co-training**

  Αποτελεί μια επίσης παραλλαγή της μεθόδου co-training (Zhou & Goldman, 2004). Σε αυτή την μέθοδο, τρεις αλγόριθμοι μάθησης εκπαιδεύονται στο ίδιο σύνολο ετικετοποιημένων δεδομένων. Αν δύο από τους αλγόριθμους συμφωνούν στην πρόβλεψη της κλάσης ενός μη ετικετοποιημένου στιγμιότυπου, τότε αυτό χρησιμοποιείται για την εκπαίδευση του τρίτου αλγόριθμου.

- **Tri-training with Editing** (Deng & Guo, 2006).
- **RASCO** (Wang, Luo, & Zeng, 2008).
- **Rel-RASCO** (Yaslan & Cataltepe, 2009).

  Οι μέθοδοι αυτές αποτελούν επίσης τροποποιήσεις της μεθόδου co-training και έχουν εφαρμοστεί με επιτυχία για την επίλυση προβλημάτων κατηγοριοποίησης Ημι-επιβλεπόμενης Μηχανικής Μάθησης.

3.4 Ενεργή Μηχανική Μάθηση

Όπως και στην ημι-επιβλεπόμενη μάθηση, έτσι και στην Ενεργή Μηχανική Μάθηση (Active Learning), το σύστημα μάθησης λαμβάνει ένα σύνολο δεδομένων εκπαίδευσης που αποτελείται από μικρό πλήθος δεδομένων με γνωστές τις κλάσεις τους και μεγάλο πλήθος δεδομένων χωρίς γνωστές κλάσεις και στη συνέχεια παράγει προβλέψεις για νέα δεδομένα. Σε αυτή την περίπτωση, το μοντέλο επιλέγει με προσοχή εκείνες τις περιπτώσεις για τις οποίες είναι περισσότερο αβέβαιο να προβλέψει την τιμή της μεταβλητής απόφασης (ετικέτα) και στη συνέχεια θέτει ερωτήματα και ζητά από έναν ειδικό (σύστημα ή άνθρωπο) τις ετικέτες αυτών των περιπτώσεων (Dasgupta, 2011). Το βασικό σημείο της ενεργής μηχανικής μάθησης είναι η δημιουργία ενός ταξινομητή υψηλής ακρίβειας χωρίς να γίνουν πάρα πολλά ερωτήματα χρησιμοποιώντας ένα μικρό σύνολο δεδομένων εκπαίδευσης.

4 Πειραματική Μελέτη-Αποτελέσματα

Στο κεφάλαιο αυτό παρουσιάζονται τα πειράματα που πραγματοποιήθηκαν στην παρούσα εργασία, καθώς και τα αντίστοιχα αποτελέσματα. Όπως προαναφέρθηκε, ο σκοπός της εργασίας είναι η διερεύνηση της αποτελεσματικότητας διαφόρων μεθόδων μηχανικής μάθησης για την πρόβλεψη της χρεοκοπίας επιχειρήσεων χρησιμοποιώντας οικονομικά στοιχεία (χρηματοοικονομικούς δείκτες) παρελθόντων ετών που αφορούν ένα δείγμα ελληνικών επιχειρήσεων. Συγκεκριμένα, θέλουμε να συγκρίνουμε την αποτελεσματικότητα διαφόρων μεθόδων ημι-επιβλεπόμενης και ενεργής μηχανικής μάθησης σε σχέση με τις αντίστοιχες μεθόδους επιβλεπόμενης μάθησης. Για το σκοπό αυτό, η πειραματική μελέτη αποτελείται από τρία μέρη, καθένα από τα οποία αντιστοιχεί στην εφαρμογή των αντίστοιχων μεθόδων.

4.1 Το Δείγμα και οι Μεταβλητές

Το σύνολο των δεδομένων που χρησιμοποιήθηκε στη συγκεκριμένη μελέτη προέρχεται από την Εθνική Τράπεζα της Ελλάδος. Για ένα χρονικό διάστημα τριών ετών (2003-2005) συγκεντρώθηκαν οικονομικά στοιχεία που αφορούσαν 145 Ελληνικές μικρομεσαίες επιχειρήσεις και τα οποία καλύπτουν τις περιόδους ενός έως τριών ετών πριν από την πτώχευση. Από αυτές τις επιχειρήσεις, 49 αντιστοιχούν σε περιπτώσεις πτώχευσης, ενώ οι υπόλοιπες 96 αντιστοιχούν σε περιπτώσεις μη πτώχευσης. Έτσι, δημιουργήθηκαν τρία σύνολα δεδομένων, ένα για κάθε έτος, αποτελούμενα αθροιστικά από 435 περιπτώσεις (Πίνακας 3).

Πίνακας 3: Περιγραφή των Συνόλων Δεδομένων

Σύνολο	Περιγραφή
Year -1	Δεδομένα που συγκεντρώθηκαν 1 έτος πριν την χρεοκοπία
Year -2	Δεδομένα που συγκεντρώθηκαν 2 έτη πριν την χρεοκοπία
Year -3	Δεδομένα που συγκεντρώθηκαν 3 έτη πριν την χρεοκοπία

Κάθε περίπτωση (στιγμιότυπο) του συνόλου δεδομένων χαρακτηρίζεται από τις τιμές δεκατριών ποσοτικών μεταβλητών (Πίνακας 4) και αντιστοιχεί σε μια μεμονωμένη επιχείρηση. Οι χρησιμοποιούμενες μεταβλητές εισόδου αντιστοιχούν σε γνωστούς

χρηματοοικονομικούς δείκτες, οι οποίοι περιγράφουν τη ρευστότητα, την αποδοτικότητα, την ανάπτυξη, την κερδοφορία και τη δανειοληπτική ικανότητα μιας επιχείρησης (Groppelli & Nikbakht, 2000).

Πίνακας 4: Περιγραφή των μεταβλητών

Μεταβλητή	Περιγραφή
GRTA	Ρυθμός αύξησης των συνολικών περιουσιακών στοιχείων
GRNI	Ρυθμός αύξησης των καθαρών εσόδων
SIZE	Μέγεθος της επιχείρησης
GIMAR	Ακαθάριστα έσοδα προς πωλήσεις
S/CE	Πωλήσεις προς απασχολούμενα κεφάλαια
S/EQ	Πωλήσεις προς ίδια κεφάλαια μετόχων
CE/NFA	Απασχολούμενα κεφάλαια προς καθαρά πάγια στοιχεία του ενεργητικού
TD/EQ	Συνολικό χρέος προς ίδια κεφάλαια μετόχων
EQ/CE	Ίδια κεφάλαια προς το απασχολούμενο κεφάλαιο
WC/TA	Κεφάλαιο κίνησης προς το συνολικό ενεργητικό
COLPER	Μέση περίοδος είσπραξης των απαιτήσεων
PAYPER	Μέση περίοδος πληρωμής των πληρωτέων λογαριασμών προς τους πιστωτές
INVTURN	Δείκτης ανακύκλωσης αποθεμάτων: Μέση περίοδος του κύκλου εργασιών για τα αποθέματα

Ο πίνακας 5 παρουσιάζει την κατανομή των 49 χρεοκοπημένων επιχειρήσεων στους αντίστοιχους βιομηχανικούς κλάδους.

Στο Σχήμα 15 παρουσιάζεται ένα μέρος του συνόλου δεδομένων ‘Year -1’ σε μορφή arff. Σε αυτό διακρίνονται οι προαναφερόμενοι χρηματοοικονομικοί δείκτες (attributes), η δυαδική μεταβλητή απόφασης ‘Final’ με τιμές {Bankrupt, Non-Bankrupt}, καθώς και πέντε περιπτώσεις με τις αντίστοιχες τιμές των μεταβλητών (χρηματοοικονομικοί δείκτες).

Πίνακας 5: Κατανομή των χρεοκοπημένων επιχειρήσεων

Βιομηχανικός κλάδος	Πλήθος επιχειρήσεων
Χονδρικό εμπόριο	6
Εκδόσεις και εκτυπώσεις	1
Πληροφορική	1
Διαφήμιση	3
Πλαστικά	1
Ενδύματα	4
Βιομηχανικά ορυκτά	1
Μηχανήματα	2
Τρόφιμα	2
Κλωστοϋφαντουργία	4
Ιδιωτική εκπαίδευση	1
Super Market	1
Λιανικό εμπόριο	10
Μεταφορές	1
Κατασκευές	2
Γεωργία και κτηνοτροφία	1
Ηλεκτρολογικός εξοπλισμός	1
Υπηρεσίες υγείας	1
Εστιατόρια	1
Μεταλλικά προϊόντα	1
Εμπόριο και συντήρηση οχημάτων	1
Logistics	1
Τηλεπικοινωνίες	1
Άλλες υπηρεσίες	1

```
@relation bankruptcypredictiondata(-1year)
@attribute SCALEDCHANGEASSETS {[-0.006033-0.07423],<-0.006033,>0.07423}
@attribute SIZE {>15.648881,<14.455296,[14.455296-15.648881]}
@attribute CHNETINCOME {<-0.23153,[-0.23153-0.145452],>0.145452}
@attribute GRPROFITMARGIN {<15.933695,[15.933695-27.435],>27.435}
@attribute CAPITALEMPLOYEDTURNOVER {[1.39524-4.705],>4.705,<1.39524}
@attribute STOCKHOLDERSEQUITYTURNOVER {[1.47-5.743759],>5.743759,<1.47}
@attribute CAPITALEMPLOYEDNETFIXEDASSETS {[1.025-2.655],<1.025,>2.655}
@attribute DEBTEQUITY {>5.241106,[1.165-5.241106],<1.165}
@attribute EQUITYCAPITALEMPLOYED {<0.905,>0.999996,[0.905-0.999996]}
@attribute WORKINGCAPITALTOTALASSETS {[-0.05664-0.155183],<-0.05664,>0.155183}
@attribute AVCOLLECTIONPERIODFORRECIEVABLES {>196.5,<103.787449,[103.787449-196.5]}
@attribute AVPAYMENTPERIOD {>218.553137,[105.534528-218.553137],<105.534528}
@attribute AVTURNOVERPERIODFORINVENTORIES {<52.465584,[52.465584-199.020884],>199.020884}
@attribute Final {Bankrupt,Non-Bankrupt}
@data
'[-0.006033-0.07423]','>15.648881','<-0.23153','<15.933695','[1.39524-4.705]','[1.47-5.743759]','[1.025-2.655]','>5.241106','<0.905','[-0.05664-0.155183]','>196.5','>218.553137','<52.465584',Bankrupt
'<-0.006033','>15.648881','<-0.23153','<15.933695','[1.39524-4.705]','[1.47-5.743759]','<1.025','[1.165-5.241106]','<0.905','<-0.05664','<103.787449','[105.534528-218.553137]','<52.465584',Bankrupt
'>0.07423','<14.455296','<-0.23153','<15.933695','[1.39524-4.705]','[1.47-5.743759]','>2.655','[1.165-5.241106]','>0.999996','>0.155183','>196.5','>218.553137','<52.465584',Bankrupt
'[-0.006033-0.07423]','<14.455296','[-0.23153-0.145452]','[15.933695-27.435]','>4.705','>5.743759','>2.655','>5.241106','>0.999996','[-0.05664-0.155183]','>196.5','[105.534528-218.553137]','<52.465584',Bankrupt
'<-0.006033','[14.455296-15.648881]','<-0.23153','[15.933695-27.435]','>4.705','>5.743759','<1.025','>5.241106','>0.999996','<-0.05664','<103.787449','[105.534528-218.553137]','<52.465584',Bankrupt
```

Σχήμα 15: Αρχείο δεδομένων 'Year -1' σε μορφή arff

4.2 Ημι-επιβλεπόμενη Μάθηση

Για την εκτίμηση της απόδοσης των κατηγοριοποιητών, χρησιμοποιήθηκε η μέθοδος της διασταυρωμένης επικύρωσης (cross validation) 10-αναδιπλώσεων. Σύμφωνα με τη μέθοδο αυτή, κάθε δείγμα χωρίζεται σε 10 τμήματα ίσου μεγέθους. Κατά τη διάρκεια κάθε εκτέλεσης, ένα τμήμα επιλέγεται για έλεγχο και τα υπόλοιπα για την εκπαίδευση. Η διαδικασία αυτή επαναλαμβάνεται 10 φορές, ώστε κάθε ένα από τα 10 τμήματα να χρησιμοποιηθεί για έλεγχο ακριβώς μια φορά. Το συνολικό σφάλμα υπολογίζεται αθροίζοντας τα σφάλματα των 10 εκτελέσεων.

Κατά τη διαδικασία της Ημι-επιβλεπόμενης Μηχανικής Μάθησης το σύνολο δεδομένων εκπαίδευσης διαμερίζεται σε ένα μικρό σύνολο ετικετοποιημένων δεδομένων L και ένα μεγάλο σύνολο μη ετικετοποιημένων δεδομένων U. Ο τρόπος διαχωρισμού του συνόλου δεδομένων καθορίζεται από τον λόγο:

$$R = \frac{N(L)}{N(L) + N(U)}$$

όπου $N(L)$, $N(U)$ το πλήθος των συνόλων L, U αντίστοιχα.

Για τον σχηματισμό των ταξινομητών Ημι-επιβλεπόμενης Μάθησης χρησιμοποιήθηκαν τρεις χαρακτηριστικές μέθοδοι και δύο ταξινομητές ως βάση. Οι τρεις μέθοδοι που χρησιμοποιήθηκαν είναι:

- Co-training (Blum & Mitchell, 1998).
- RASCO (Wang, Luo & Zeng, 2008).
- Rel-RASCO (Yaslan & Cataltepe, 2009).

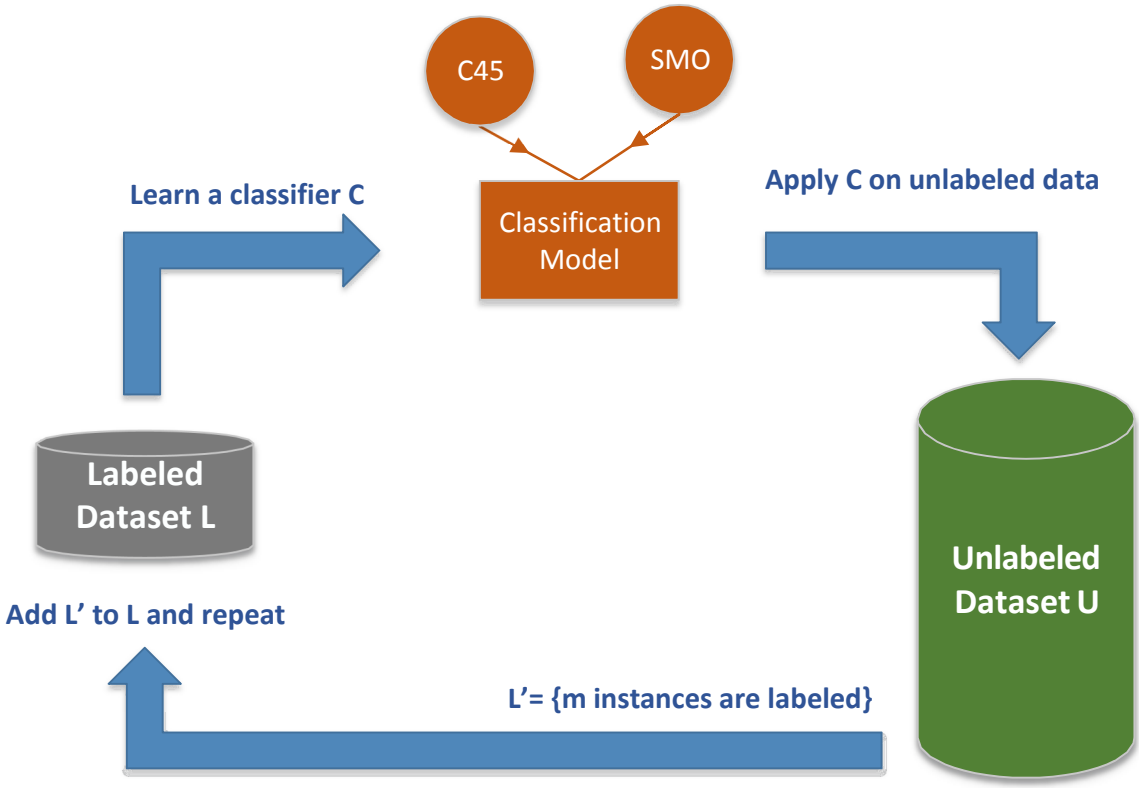
Οι ταξινομητές που χρησιμοποιήθηκαν ως βάση στις προαναφερόμενες μεθόδους είναι:

- Το δέντρο απόφασης C4.5 (Quinlan, 2014).
- Sequential Minimal Optimization (SMO) (Platt, 1998).

Σε αυτή τη βάση δημιουργήθηκαν τελικά έξι κατηγοριοποιητές Ημι-επιβλεπόμενης Μηχανικής Μάθησης. Για τη διενέργεια των πειραμάτων χρησιμοποιήθηκε το ελεύθερο λογισμικό KEEL (Knowledge Extraction based on Evolutionary Learning), το οποίο αποτελεί ένα από τα πιο δημοφιλή προγράμματα για Ημι-επιβλεπόμενη μηχανική μάθηση (Triguero et al., 2017) διατίθεται ελεύθερα υπό τους όρους της άδειας GPLv3 License.

Στο Σχήμα 16 απεικονίζεται η λειτουργία των προαναφερόμενων κατηγοριοποιητών Ημι-επιβλεπόμενης Μηχανικής Μάθησης, οι οποίοι χρησιμοποιήθηκαν κατά τη διεξαγωγή των πειραμάτων. Σύμφωνα με το σχήμα:

- Αρχικά, το σύνολο *L* των ετικετοποιημένων δεδομένων χρησιμοποιείται για την κατασκευή ενός μοντέλου κατηγοριοποίησης χρησιμοποιώντας τον αντίστοιχο αλγόριθμο μάθησης (ταξινομητή).
- Επιλέγονται από το σύνολο *U* των μη ετικετοποιημένων δεδομένων *m* εγγραφές.
- Οι εγγραφές αυτές ετικετοποιούνται και προστίθεται στο σύνολο *L*.
- Η διαδικασία επαναλαμβάνεται μέχρι το κριτήριο τερματισμού.



Σχήμα 16: Μηχανισμός Ημι-επιβλεπόμενης Μηχανικής Μάθησης

Σε κάθε βήμα μετράται η ακρίβεια του μοντέλου, η οποία αντιστοιχεί στο ποσοστό των σωστά ταξινομημένων περιπτώσεων. Κατά την πειραματική διαδικασία χρησιμοποιήθηκαν τρεις διαφορετικές τιμές για τον λόγο *R* διαχωρισμού του συνόλου δεδομένων σε δύο σύνολα δεδομένων: ετικετοποιημένων και μη ετικετοποιημένων. Οι τιμές του *R*, οι οποίες χρησιμοποιήθηκαν στα πειράματα είναι: 20%, 30% και 40%. Τα αντίστοιχα αποτελέσματα

για τους κατηγοριοποιητές Ημι-επιβλεπόμενης Μηχανικής Μάθησης παρουσιάζονται στους Πίνακες 6-8.

Πίνακας 6: Ακρίβεια (%) κατηγοριοποιητών Ημι-επιβλεπόμενης Μάθησης (*R*=20%)

	Σύνολο Δεδομένων		
Ταξινομητής	**Year -3**	**Year -2**	**Year -1**
Co-training (C4.5)	56,29	57,24	64,90
Co-training (SMO)	49,76	55,19	59,19
RASCO (C4.5)	62,10	57,24	66,19
RASCO (SMO)	51,81	55,19	65,62
Rel-RASCO (C4.5)	59,38	58,24	66,19
Rel-RASCO (SMO)	54,48	61,24	64,00

Στην πρώτη περίπτωση (*R*=20%) παρατηρούμε ότι:

- Ένα έτος πριν από το σημείο χρεοκοπίας, οι καλύτεροι κατηγοριοποιητές Ημι-επιβλεπόμενης Μηχανικής Μάθησης είναι οι RASCO(C4.5) και Rel-RASCO(SMO) με ακρίβεια 66.19% ακολουθούμενοι από τον RASCO(SMO), του οποίου η ακρίβεια είναι 65.62%.
- Τρία έτη πριν από το σημείο χρεοκοπίας, ο καλύτερος κατηγοριοποιητής Ημι-επιβλεπόμενης Μηχανικής Μάθησης είναι ο RASCO(C4.5) με ακρίβεια 62.10%.

Πίνακας 7: Ακρίβεια (%) κατηγοριοποιητών Ημι-επιβλεπόμενης Μάθησης (R=30%)

	Σύνολο Δεδομένων		
Ταξινομητής	**Year -3**	**Year -2**	**Year -1**
Co-training (C4.5)	59,81	55,19	58,62
Co-training (SMO)	58,67	53,38	62,53
RASCO (C4.5)	58,67	60,38	61,90
RASCO (SMO)	52,33	53,76	63,95
Rel-RASCO (C4.5)	62,91	61,14	62,62
Rel-RASCO (SMO)	54,43	56,86	63,38

Στη δεύτερη περίπτωση (*R*=30%) παρατηρούμε ότι:

- Ένα έτος πριν από το σημείο χρεοκοπίας, οι καλύτεροι κατηγοριοποιητές Ημι-επιβλεπόμενης Μηχανικής Μάθησης είναι οι RASCO(SMO), Rel-RASCO(SMO) και Rel-RASCO(C4.5) με ακρίβεια 63.95%, 63.38% και 62.62% αντίστοιχα.
- Τρία έτη πριν από το σημείο χρεοκοπίας, ο καλύτερος κατηγοριοποιητής Ημι-επιβλεπόμενης Μηχανικής Μάθησης είναι ο Rel-RASCO(C4.5) με ακρίβεια 62.91% ακολουθούμενος από τον Co-training(C4.5) με ακρίβεια 59.81%.

Πίνακας 8: Ακρίβεια (%) κατηγοριοποιητών Ημι-επιβλεπόμενης Μάθησης (R=40%)

	Σύνολο Δεδομένων		
Ταξινομητής	**Year -3**	**Year -2**	**Year -1**
Co-training (C4.5)	68,10	63,48	64,81
Co-training (SMO)	57,19	52,43	62,67
RASCO (C4.5)	64,76	57,90	59,95
RASCO (SMO)	56,62	56,00	69,76
Rel-RASCO (C4.5)	67,47	62,10	66,86
Rel-RASCO (SMO)	53,62	60,10	64,62

Στην τρίτη περίπτωση (*R*=40%) παρατηρούμε ότι:

- Ένα έτος πριν από το σημείο χρεοκοπίας, οι καλύτεροι κατηγοριοποιητές Ημι-επιβλεπόμενης Μηχανικής Μάθησης είναι οι RASCO(SMO), Rel-RASCO(C4.5) και Rel-RASCO(SMO) με ακρίβεια 69.76%, 66.86% και 64.81% αντίστοιχα.
- Τρία έτη πριν από το σημείο χρεοκοπίας, ο καλύτερος κατηγοριοποιητής Ημι-επιβλεπόμενης Μηχανικής Μάθησης είναι ο Co-training(C4.5) με ακρίβεια 68.10%.

Στη συνέχεια, συγκρίνονται οι κατηγοριοποιητές Ημι-επιβλεπόμενης Μηχανικής Μάθησης με τους αντίστοιχους κατηγοριοποιητές Επιβλεπόμενης Μάθησης. Τα αντίστοιχα αποτελέσματα παρουσιάζονται στους Πίνακες 9-11.

Πίνακας 9: Ακρίβεια (%) κατηγοριοποιητών Επιβλεπόμενης Μάθησης (R=20%)

	Σύνολο Δεδομένων		
Ταξινομητής	**Year -3**	**Year -2**	**Year -1**
C4.5	53,66	58,67	60,00
SMO	49,52	59,19	60,71

Στην πρώτη περίπτωση (*R*=20%) παρατηρούμε ότι:

- Ένα έτος πριν από το σημείο χρεοκοπίας, ο καλύτερος αλγόριθμος κατηγοριοποίησης είναι ο SMO με ακρίβεια 60.71%.
- Τρία έτη πριν από το σημείο χρεοκοπίας, ο καλύτερος αλγόριθμος κατηγοριοποίησης είναι ο C4.5 με ακρίβεια 53.66%.

Πίνακας 10: Ακρίβεια (%) κατηγοριοποιητών Επιβλεπόμενης Μάθησης (*R*=30%)

	Σύνολο Δεδομένων		
Ταξινομητής	**Year -3**	**Year -2**	**Year -1**
C4.5	60,71	56,67	58,62
SMO	54,48	48,38	58,43

Στη δεύτερη περίπτωση (*R*=30%) παρατηρούμε ότι:

- Ο καλύτερος αλγόριθμος κατηγοριοποίησης είναι ο C4.5 με ακρίβεια. η οποία κυμαίνεται από 56.67% έως 60.71%

Πίνακας 11: Ακρίβεια (%) κατηγοριοποιητών Επιβλεπόμενης Μάθησης (*R*=40%)

	Σύνολο Δεδομένων		
Ταξινομητής	**Year -3**	**Year -2**	**Year -1**
C4.5	65,62	60,09	58,04
SMO	53,10	51,05	61,02

Στην τρίτη περίπτωση (*R*=40%) παρατηρούμε ότι:

- Ένα έτος πριν από το σημείο χρεοκοπίας, ο καλύτερος αλγόριθμος κατηγοριοποίησης είναι ο SMO με ακρίβεια 61.02%.
- Τρία έτη πριν από το σημείο χρεοκοπίας, ο καλύτερος αλγόριθμος κατηγοριοποίησης είναι ο C4.5 με ακρίβεια 65.62%.

Είναι φανερό από τα αποτελέσματα της ακρίβειας των χρησιμοποιούμενων μεθόδων Ημι-επιβλεπόμενης και Επιβλεπόμενης Μηχανικής Μάθησης, ότι οι μέθοδοι της Ημι-επιβλεπόμενης Μηχανικής Μάθησης παρουσιάζουν μεγαλύτερη ακρίβεια σε όλες τις περιπτώσεις.

4.3 Ενεργή Μηχανική Μάθηση

Στην περίπτωση της Ενεργής Μηχανικής Μάθησης χρησιμοποιήθηκε επίσης η μέθοδος της διασταυρωμένης επικύρωσης 10-αναδιπλώσεων. Σε κάθε ένα από τα δέκα τμήματα, το σύνολο εκπαίδευσης αποτελείται αρχικά από 6 ετικετοποιημένες και 124 μη ετικετοποιημένες περιπτώσεις. Παράλληλα, ορίστηκε ως κριτήριο τερματισμού της επαναληπτικής διαδικασίας ένα πλήθος 30 επαναλήψεων, ενώ επιλέχθηκε μια μόνο εγγραφή για ετικετοποίηση σε καθεμιά από τις επαναλήψεις. Σε αυτή τη βάση, στο τέλος της διαδικασίας μάθησης υπήρχαν 36 ετικετοποιημένες περιπτώσεις.

Για τον σχηματισμό των κατηγοριοποιητών Ενεργής Μηχανικής Μάθησης χρησιμοποιήθηκαν πέντε χαρακτηριστικοί ταξινομητές ως βάση, οι οποίοι περιγράφονται συνοπτικά παρακάτω:

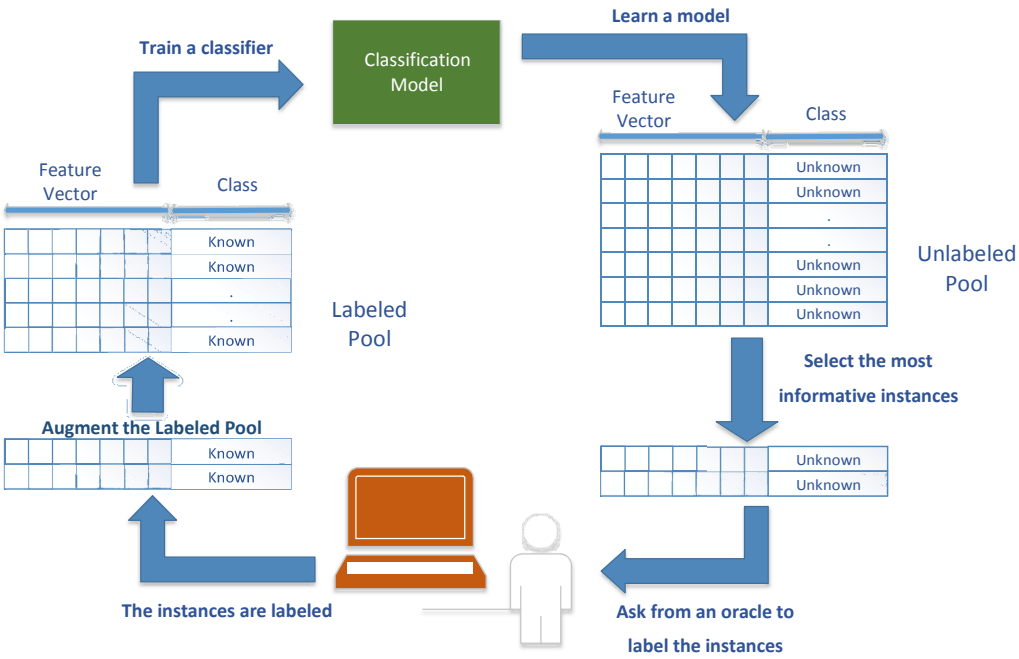
- Multilayer Perceptrons (MLPs) (Gardner & Dorling, 1998). Αποτελούν τεχνητά νευρωνικά δίκτυα πολλών επιπέδων νευρώνων perceptron και χρησιμοποιούνται συχνά για την επίλυση προβλημάτων κατηγοριοποίησης.
- Bayes Net (Pearl, 2014). Πρόκειται για ένα αντιπροσωπευτικό δίκτυο του Bayes, το οποίο παρέχει μια γραφική αναπαράσταση των πιθανοφανών σχέσεων μεταξύ των τυχαίων μεταβλητών ενός συνόλου.
- Logistic Regression (LR) (Ng & Jordan, 2002). Η Λογιστική Παλινδρόμηση δημιουργεί ένα μοντέλο ταξινόμησης στηριζόμενη στη θεωρία των πιθανοτήτων σε περιπτώσεις δυαδικής μεταβλητής εξόδου, όπως στην παρούσα εργασία.
- Random Forests (RF) (Breiman, 2001). Τα Τυχαία Δάση αποτελούν μια σύνθετη μέθοδο η οποία συνδυάζει τις προβλέψεις οι οποίες προέρχονται από πολλά δέντρα απόφασης και εφαρμόζονται ευρέως σε προβλήματα κατηγοριοποίησης με αξιοσημείωτα αποτελέσματα.
- Sequential Minimal Optimization (SMO) (Platt, 1998). Αποτελεί χαρακτηριστικό παράδειγμα ταξινομητή ο οποίος ανήκει στην κατηγορία των μηχανών διανυσμάτων υποστήριξης (support vector machines).

Για τη διενέργεια των πειραμάτων χρησιμοποιήθηκε το ελεύθερο λογισμικό Weka (Waikato Environment for Knowledge Analysis), το οποίο αποτελεί ένα από τα πιο δημοφιλή προγράμματα για μηχανική μάθηση και εξόρυξη δεδομένων (Frank et al., 2009). Έχει αναπτυχθεί στο πανεπιστήμιο Waikato της Νέας Ζηλανδίας, είναι γραμμένο σε γλώσσα Java

και διατίθεται ελεύθερα υπό τους όρους της άδειας GNU General Public License. Ενσωματώνει πληθώρα αλγορίθμων για την υλοποίηση όλων των βασικών μεθόδων κατηγοριοποίησης, παλινδρόμησης και συσταδοποίησης με μεγάλες δυνατότητες ρύθμισης των παραμέτρων τους. Επιπλέον, περιλαμβάνει εργαλεία για τη δημιουργία σύνθετων κατηγοριοποιητών, καθώς και εργαλεία παρουσίασης, οπτικοποίησης και προ-επεξεργασίας των δεδομένων σε ένα φιλικό και ευέλικτο περιβάλλον εργασίας.

Στο Σχήμα 17 απεικονίζεται η λειτουργία των προαναφερόμενων κατηγοριοποιητών Ενεργής Μηχανικής Μάθησης, οι οποίοι χρησιμοποιήθηκαν κατά τη διεξαγωγή των πειραμάτων. Σύμφωνα με το σχήμα:

- Αρχικά, το σύνολο *L* (Labeled Pool) των ετικετοποιημένων δεδομένων χρησιμοποιείται για την κατασκευή ενός μοντέλου κατηγοριοποίησης χρησιμοποιώντας τον αντίστοιχο αλγόριθμο μάθησης (ταξινομητή).
- Επιλέγεται από το σύνολο *U* (Unlabeled Pool) των μη ετικετοποιημένων δεδομένων μια εγγραφή (η πιο χαρακτηριστική).
- Ζητείται η ετικέτα της εγγραφής από τον ειδικό (oracle).
- Η εγγραφή προστίθεται στο σύνολο *L* και η διαδικασία επαναλαμβάνεται μέχρι το κριτήριο τερματισμού (στη συγκεκριμένη περίπτωση είναι 30 επαναλήψεις).

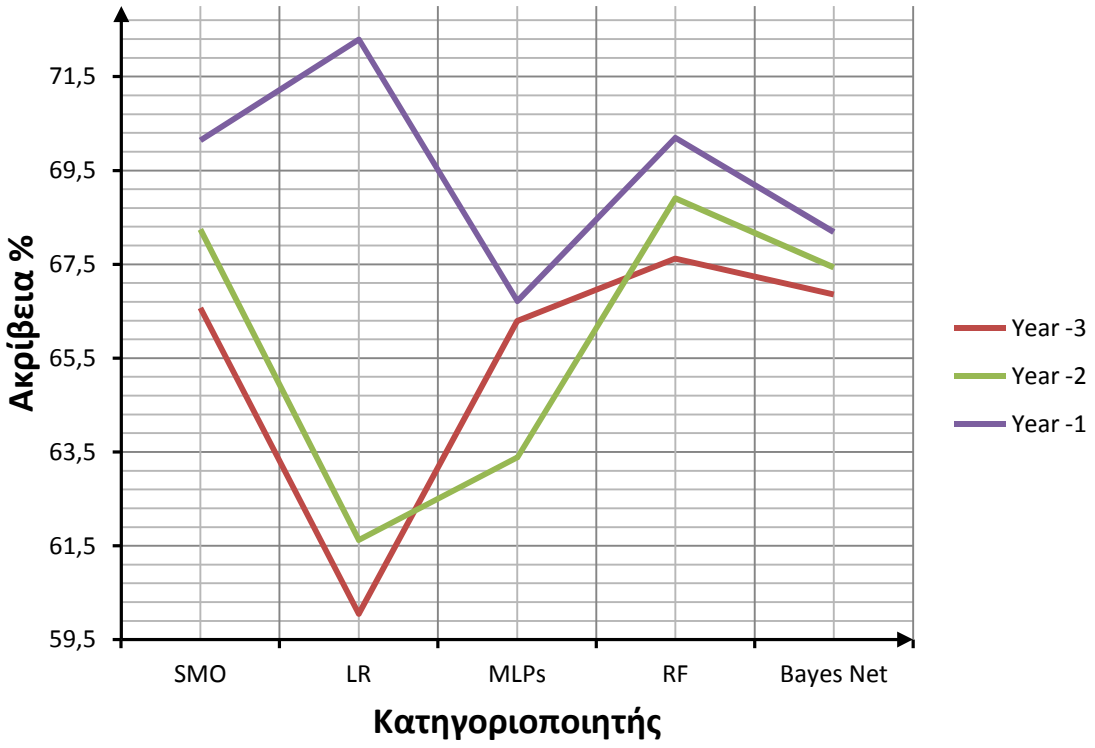


Σχήμα 17: Μηχανισμός Ενεργής Μηχανικής Μάθησης

Η πειραματική διαδικασία αποτελείται από τρία διακριτά βήματα καθένα από τα οποία αντιστοιχεί χρονικά στα έτη που απομένουν μέχρι το φαινόμενο της χρεοκοπίας (τρία, δύο και ένα έτος αντίστοιχα). Σε κάθε βήμα μετράται η ακρίβεια του μοντέλου, η οποία αντιστοιχεί στο ποσοστό των σωστά ταξινομημένων περιπτώσεων (Πίνακας 12). Στο Σχήμα 18 απεικονίζεται η μεταβολή της ακρίβειας των κατηγοριοποιητών Ενεργής Μηχανικής Μάθησης με την πάροδο του χρόνου και την ανανέωση των τιμών των χρηματοοικονομικών δεικτών κάθε επιχείρησης.

Πίνακας 12: Ακρίβεια (%) των κατηγοριοποιητών Ενεργής Μηχανικής Μάθησης

	Ταξινομητής				
Σύνολο Δεδομένων	**SMO**	**LR**	**MLPs**	**RF**	**Bayes Net**
Year -3	66.57	60.05	66.29	67.62	66.86
Year -2	68.24	61.62	63.38	68.90	67.43
Year -1	70.14	72.29	66.71	70.19	68.19



Σχήμα 18: Μεταβολή ακρίβειας κατηγοριοποιητών Ενεργής Μηχανικής Μάθησης

Παρατηρούμε ότι:

- Τρία έτη και δύο έτη πριν από το χρονικό σημείο χρεοκοπίας, ο καλύτερος κατηγοριοποιητής Ενεργής Μηχανικής Μάθησης είναι αυτός που χρησιμοποιεί ως βάση τα Τυχαία Δάση (RF) με ακρίβεια 67.62% και 68.90% αντίστοιχα.
- Ένα έτος πριν από το σημείο χρεοκοπίας, ο καλύτερος κατηγοριοποιητής Ενεργής Μηχανικής Μάθησης είναι αυτός που χρησιμοποιεί ως βάση τον αλγόριθμο Λογιστικής Παλινδρόμησης (LR) με ακρίβεια 72.29%.
- Ένα έτος πριν από το σημείο χρεοκοπίας, η ακρίβεια τριών κατηγοριοποιητών Ενεργής Μηχανικής Μάθησης ξεπερνά το 70%, το οποίο μας δίνει τη δυνατότητα να προβλέψουμε την χρεοκοπία επιχειρήσεων με σχετικά μεγάλη ακρίβεια.
- Η ακρίβεια των κατηγοριοποιητών αυξάνεται από βήμα σε βήμα καθώς πλησιάζουμε χρονικά στο σημείο χρεοκοπίας, ενώ τα καλύτερα αποτελέσματα στην ακρίβεια των κατηγοριοποιητών Ενεργής Μηχανικής Μάθησης εμφανίζονται ένα έτος πριν από το σημείο χρεοκοπίας.

Στη συνέχεια, συγκρίνονται οι κατηγοριοποιητές Ενεργής Μηχανικής Μάθησης με τους αντίστοιχους κατηγοριοποιητές Επιβλεπόμενης Μάθησης. Τα αντίστοιχα αποτελέσματα παρουσιάζονται στον Πίνακα 13.

Πίνακας 13: Ακρίβεια (%) κατηγοριοποιητών Επιβλεπόμενης Μάθησης

	Ταξινομητής				
Σύνολο Δεδομένων	**SMO**	**LR**	**MLPs**	**RF**	**Bayes Net**
Year -3	62.80	66.20	60.00	65.50	62.10
Year -2	60.00	63.40	59.30	62.10	63.40
Year -1	62.80	61.40	56.60	67.60	62.10

Σύμφωνα με τα αποτελέσματα του Πίνακα 13 παρατηρούμε ότι:

- Τρία έτη πριν από το σημείο χρεοκοπίας, ο καλύτερος κατηγοριοποιητής Επιβλεπόμενης Μάθησης είναι ο αλγόριθμος Λογιστικής Παλινδρόμησης (LR) με ακρίβεια 66.20%.

- Δύο έτη πριν από το σημείο χρεοκοπίας, οι καλύτεροι κατηγοριοποιητές Επιβλεπόμενης Μάθησης είναι οι αλγόριθμοι Λογιστικής Παλινδρόμησης (LR) και Bayes Net με ακρίβεια 63.40%.
- Ένα έτος πριν από το σημείο χρεοκοπίας, ο καλύτερος κατηγοριοποιητής Επιβλεπόμενης Μάθησης είναι τα Τυχαία Δάση (RF) με ακρίβεια 67.60%.

Για τη σύγκριση της ακρίβειας των κατηγοριοποιητών Ενεργής Μηχανικής Μάθησης και Επιβλεπόμενης Μάθησης πραγματοποιήθηκε μη παραμετρικός έλεγχος με χρήση του Friedman Aligned Ranks nonparametric test (Hodges & Lehmann, 1962). Σύμφωνα με τα αποτελέσματα του παραμετρικού ελέγχου (Πίνακας 14), οι κατηγοριοποιητές ταξινομούνται σε αύξουσα σειρά απόδοσης (ο κατηγοριοποιητής με τη μικρότερη σειρά είναι αυτός με την μεγαλύτερη απόδοση). Παρατηρούμε ότι:

- Οι κατηγοριοποιητές Ενεργής Μηχανικής Μάθησης υπερέχουν αυτών της Επιβλεπόμενης Μάθησης, το οποίο καταδεικνύει την αποτελεσματικότητα των αντίστοιχων μεθόδων για την πρόβλεψη της χρεοκοπίας επιχειρήσεων.
- Ο κατηγοριοποιητής Ενεργής Μηχανικής Μάθησης με βάση τα Τυχαία Δάση (RF) είναι ο καλύτερος ταξινομητής.

Πίνακας 14: Αποτελέσματα μη παραμετρικού ελέγχου

Κατηγοριοποιητής	Σειρά
Ενεργή Μηχανική Μάθηση (RF)	4.33
Ενεργή Μηχανική Μάθηση (SMO)	7.00
Ενεργή Μηχανική Μάθηση (Bayes Net)	7.67
Ενεργή Μηχανική Μάθηση (MLPs)	14.33
RF	15.33
Ενεργή Μηχανική Μάθηση (LR)	16.00
LR	18.50
Bayes Net	20.83
SMO	22.33
MLPs	28.67

Συμπερασματικά, οι κατηγοριοποιητές Ενεργής Μηχανικής Μάθησης παρουσιάζουν μεγαλύτερη ακρίβεια στην πρόβλεψη της χρεοκοπίας επιχειρήσεων χρησιμοποιώντας ένα μικρό πλήθος ετικετοποιημένων δεδομένων σε αντίθεση με τους αντίστοιχους κατηγοριοποιητές Επιβλεπόμενης Μηχανικής Μάθησης οι οποίοι εκπαιδεύονται σε όλο το δείγμα των δεδομένων. Αυτό αποτελεί και το συγκριτικό πλεονέκτημα της Ενεργής Μηχανικής Μάθησης έναντι της Επιβλεπόμενης Μηχανικής Μάθησης.

Συμπεράσματα

Η χρεοκοπία επιχειρήσεων (corporate bankruptcy) αποτέλεσε και συνεχίζει να αποτελεί ένα από τα σημαντικότερα θέματα της οικονομικής επιστήμης έχοντας απασχολήσει πολλούς επιστήμονες και ερευνητές. Ιδιαίτερα στη σημερινή εποχή κατά την οποία επικρατεί παγκόσμια οικονομική και χρηματοπιστωτική κρίση, η χρεοκοπία αποτελεί καθημερινό αντικείμενο έρευνας λόγω των καταστροφικών συνεπειών που μπορεί να έχει για όλους τους εμπλεκόμενους.

Πληθώρα στατιστικών μεθόδων και αλγορίθμων μηχανικής μάθησης έχουν εφαρμοστεί τα τελευταία χρόνια στο πεδίο της οικονομίας στοχεύοντας στη δημιουργία προγνωστικών μοντέλων για την ανίχνευση της επιχειρηματικής αποτυχίας όπως είναι, για παράδειγμα, η Πολυμεταβλητή Ανάλυση Διαφοροποίησης, η Λογαριθμική Ανάλυση και τα Νευρωνικά Δίκτυα. Στην παρούσα εργασία διερευνήθηκε η αποτελεσματικότητα γνωστών μεθόδων Ημι-επιβλεπόμενης και Ενεργής Μηχανικής Μάθησης για την πρόβλεψη της χρεοκοπίας επιχειρήσεων χρησιμοποιώντας βασικούς χρηματοοικονομικούς δείκτες και οικονομικά στοιχεία σχετικά με ένα δείγμα 145 ελληνικών επιχειρήσεων κατά τη διάρκεια τριών ετών (2003-2005).

Η πειραματική ανάλυση δείχνει την αποτελεσματικότητα των συγκεκριμένων μεθόδων για την πρόβλεψη της εταιρική χρεοκοπίας με σχετικά μεγάλη ακρίβεια, αλλά και την υπεροχή τους σε σχέση με τις αντίστοιχες μεθόδους Επιβλεπόμενης Μάθησης. Όσον αφορά στις μεθόδους Ημι-επιβλεπόμενης Μηχανικής Μάθησης παρατηρήθηκε ότι:

- Τρία έτη πριν από το σημείο χρεοκοπίας, ο καλύτερος κατηγοριοποιητής Ημι-επιβλεπόμενης Μηχανικής Μάθησης είναι ο Co-training(C4.5) με ακρίβεια 59.81%.
- Ένα έτος πριν από το σημείο χρεοκοπίας, ο καλύτερος κατηγοριοποιητής Ημι-επιβλεπόμενης Μηχανικής Μάθησης είναι οι RASCO(SMO) με ακρίβεια 69.76%.

Όσον αφορά στις μεθόδους Ενεργής Μηχανικής Μάθησης παρατηρήθηκε ότι:

- Τρία έτη και δύο έτη πριν από το χρονικό σημείο χρεοκοπίας, ο καλύτερος κατηγοριοποιητής Ενεργής Μηχανικής Μάθησης είναι αυτός που χρησιμοποιεί ως βάση τα Τυχαία Δάση (RF) με ακρίβεια 67.62% και 68.90% αντίστοιχα.

- Ένα έτος πριν από το σημείο χρεοκοπίας, ο καλύτερος κατηγοριοποιητής Ενεργής Μηχανικής Μάθησης είναι αυτός που χρησιμοποιεί ως βάση τον αλγόριθμο Λογιστικής Παλινδρόμησης (LR) με ακρίβεια 72.29%.

Σε σύγκριση με τις μεθόδους Επιβλεπόμενης Μηχανικής Μάθησης, η ακρίβεια των χρησιμοποιούμενων μεθόδων Ημι-επιβλεπόμενης και Επιβλεπόμενης Μηχανικής Μάθησης είναι συγκριτικά μεγαλύτερη, όπως επιβεβαιώνεται στατιστικά με χρήση του μη παραμετρικού τεστ Friedman Aligned Ranks.

Οι μέθοδοι Ημι-επιβλεπόμενης και Ενεργής Μηχανικής Μάθησης αποτελούν το κατάλληλο εργαλείο για την επίλυση διαφόρων προβλημάτων κατηγοριοποίησης και έχουν εφαρμοστεί με επιτυχία σε πολλούς επιστημονικούς τομείς, μεταξύ των οποίων και ο τομέας της οικονομίας. Η χρήση μικρού πλήθους ετικετοποιημένων δεδομένων οδηγεί σε αξιοσημείωτα αποτελέσματα όπως αποδεικνύεται στην πειραματική μελέτη της παρούσας εργασίας. Ταυτόχρονα, οι συγκεκριμένες μέθοδοι δίνουν τη δυνατότητα της έγκαιρης πρόγνωσης για τη λήψη μέτρων περιορισμού και αντιμετώπισης της οικονομικής δυσχέρειας μιας επιχείρησης.

Είναι αξιοσημείωτο το γεγονός ότι όλες οι μελέτες σχετικά με την πρόβλεψη της εταιρικής χρεοκοπίας με τη χρήση διαφόρων στατιστικών μοντέλων και μοντέλων μηχανικής μάθησης τονίζουν την αναγκαιότητα χρήσης βασικών χρηματοοικονομικών δεικτών, οι οποίοι μετρούν την αποδοτικότητα, τη ρευστότητα και τη φερεγγυότητα μιας επιχείρησης. Παρόλα αυτά, η βαρύτητα αυτών των δεικτών δεν είναι ξεκάθαρη, αφού σε κάθε μελέτη διαφορετικοί δείκτες έχουν βαρύνουσα επίδραση στην αποτελεσματικότητα του αντίστοιχου προγνωστικού μοντέλου.

Συμπερασματικά, παρά την ακρίβεια των προβλέψεων των χρησιμοποιούμενων μοντέλων Μηχανικής Μάθησης για την πρόβλεψη της εμπορικής χρεοκοπίας, απαιτείται συστηματική και μεθοδευμένη χρήση των χρηματοοικονομικών δεικτών που αφορούν την πορεία μιας επιχείρησης καθόλη τη διάρκεια του κύκλου ζωής της. Σημαντικό εργαλείο προς αυτή την κατεύθυνση θα συνεχίσουν να αποτελούν διάφορα προβλεπτικά μοντέλα χρεοκοπίας βασισμένα σε παραμετρικές και μη παραμετρικές στατιστικές μεθόδους, καθώς και σε σύγχρονες μεθόδους Μηχανικής Μάθησης. Οι ραγδαίες εξελίξεις στην επιστήμη της Μηχανικής Μάθησης αποτελούν τον πλέον σημαντικό παράγοντα για την επιτυχία αυτών των μεθόδων.

Όπως αναφέρεται και στο πρόσφατο βιβλίο των Altman & Hotchkiss (2010): "…the bankruptcy business is big-business".

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ξένη Βιβλιογραφία

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

[2] Aggarwal, C., & Yu, P. (1999). Data mining techniques for associations, clustering and classification. *Methodologies for Knowledge Discovery and Data Mining*, 13-23.

[3] Aha, D. (1997). Lazy Learning. Dordrecht: Kluwer Academic Publishers

[4] Alpaydin, E. (2009). Introduction to machine learning. MIT press.

[5] Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3), 175-185.

[6] Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The journal of finance, 23(4), 589-609.

[7] Altman, E. I., & Hotchkiss, E. (2010). Corporate financial distress and bankruptcy: Predict and avoid bankruptcy, analyze and invest in distressed debt (Vol. 289). John Wiley & Sons.

[8] Beaver, W. H. (1966). Financial ratios as predictors of failure. Journal of accounting research, 71-111.

[9] Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory* (pp. 92-100). ACM.

[10] Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140.

[11] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[12] Carbonell, J. G., & Gil, Y. (1987). 'Learning by Experimentation. *Machine Learning*, 256-266.

[13] Cohen, W. W. (1995). Fast effective rule induction. In *Proceedings of the twelfth international conference on machine learning* (pp. 115-123).

[14] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297.

[15] Dasgupta, S. (2011). Two faces of active learning. Theoretical computer science, 412(19), 1767-1781.

[16] Deligianni, D., & Kotsiantis, S. (2012). Forecasting corporate bankruptcy with an ensemble of classifiers. In Hellenic Conference on Artificial Intelligence (pp. 65-72). Springer, Berlin, Heidelberg.

[17] Deng, C., & Guo, M. (2006). Tri-training and data editing based semi-supervised clustering algorithm. *MICAI 2006: Advances in Artificial Intelligence*, 641-651.

[18] Downing, D., & Clark, J. (2010). *Business statistics*. Barron's Educational Series.

[19] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3), 37.

[20] Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I. H., & Trigg, L. (2009). Weka-a machine learning workbench for data mining. In *Data mining and knowledge discovery handbook* (pp. 1269-1277). Springer, Boston, MA.

[21] Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron)-a review of applications in the atmospheric sciences. Atmospheric environment, 32(14-15), 2627-2636.

[22] Gorunescu, F. (2011). *Data Mining: Concepts, models and techniques* (Vol. 12). Springer Science & Business Media.

[23] Groppelli, A. A., & Nikbakht, E. (2000). Finance . Barron's Educational Series.

[24] Hady, M. F. A., & Schwenker, F. (2010). Combining committee-based semi-supervised learning and active learning. Journal of Computer Science and Technology, 25(4), 681-698.

[25] Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

[26] Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. MIT press.

[27] Hanson, R., Stutz, J., & Cheeseman, P. (1991). Bayesian classification theory.

[28] Hodges, J. L., & Lehmann, E. L. (1962). Rank methods for combination of independent experiments in analysis of variance. The Annals of Mathematical Statistics, 33(2), 482-497.

[29] Jain, A. K. and Dubes, R. C. (1988). *Algorithms for clustering data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

[30] Jolliffe, I. T. (2002). Principal component analysis and factor analysis. *Principal component analysis*, 150-166.

[31] Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques.

[32] Larose, D. T. (2014). *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons.

[33] Liao, S. H., Chu, P. H., & Hsiao, P. Y. (2012). Data mining techniques and applications–A decade review from 2000 to 2011. *Expert systems with applications*, 39(12), 11303-11311.

[34] Linoff, G. S., & Berry, M. J. (2011). *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons.

[35] MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

[36] McCarthy, V. 1997. "Strike It Rich" *Datamation*, 43, 2: 44-50.

[37] Mendenhall, W., Sincich, T., & Boudreau, N. S. (2003). *A second course in statistics: regression analysis* (Vol. 6). Upper Saddle River, NJ: Prentice Hall.

[38] Mitchell, T. M. (1997). Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, 45(37), 870-877.

[39] Mselmi, N., Lahiani, A., & Hamza, T. (2017). Financial distress prediction: The case of French small and medium-sized firms. International Review of Financial Analysis, 50, 67-80.

[40] Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In Advances in neural information processing systems, 841-848.

[41] Odom, M. D., & Sharda, R. (1990). A neural network model for bankruptcy prediction. In Neural Networks, 1990 IJCNN International Joint Conference on, IEEE, 163-168.

[42] Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. Journal of accounting research, 109-131.

[43] Pearl, J. (2014). Probabilistic reasoning in intelligent systems: networks of plausible inference. Elsevier.

[44] Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines.

[45] Quinlan, J. R. (2014). C4. 5: programs for machine learning. Elsevier.

[46] Roiger, R. J. (2017). *Data mining: a tutorial-based primer*. CRC Press.

[47] Roiger, J. R., & Geatz, M. W. (2003). Data Mining: A tutorial-based primer. *NY: Pearson Education*.

[48] Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. Expert systems with applications, 33(1), 135-146.

[49] Russell, S. J., & Norvig, P. (2003). Artificial Intelligence: A Modern Approach. *Pearson Education*.

[50] Sharda, R., & Wilson, R. L. (1996). Neural network experiments in business-failure forecasting: Predictive performance measurement issues. International Journal of Computational Intelligence and Organizations, 1(2), 107-117.

[51] Schölkopf, B., Smola, A., & Müller, K. R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5), 1299-1319.

[52] Storey, D. J., & Greene, F. J. (2010). Small business and entrepreneurship. Financial Times Prentice Hall.

[53] Tam, K. Y., & Kiang, M. Y. (1992). Managerial applications of neural networks: the case of bank failure predictions. Management science, 38(7), 926-947.

[54] Tan, P. N., Steinbach, M., & Kumar, V. (2013). Data mining cluster analysis: basic concepts and algorithms. *Introduction to data mining*.

[55] Thearling, K. (1999). An introduction to data mining. *Direct Marketing Magazine*, 28-31.

[56] Triguero, I., González, S., Moyano, J. M., García, S., Alcalá-Fdez, J., Luengo, J., ... & Herrera, F. (2017). KEEL 3.0: an open source software for multi-stage analysis in data mining. International Journal of Computational Intelligence Systems, 10(1), 1238-1249.

[57] Wang, J., Luo, S. W., & Zeng, X. H. (2008). A random subspace method for co-training. In *Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on* (pp. 195-200). IEEE.

[58] Wilson, R. L., & Sharda, R. (1994). Bankruptcy prediction using neural networks. Decision support systems, 11(5), 545-557.

[59] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

[60] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Zhou, Z. H. (2008). Top 10 algorithms in data mining. Knowledge and information systems, 14(1), 1-37.

[61] Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics* (pp. 189-196). Association for Computational Linguistics.

[62] Yaslan, Y., & Cataltepe, Z. (2009). Random relevant and non-redundant feature subspaces for co-training. In *International Conference on Intelligent Data Engineering and Automated Learning* (pp. 679-686). Springer, Berlin, Heidelberg.

[63] Zaki, M. J., Meira Jr, W., & Meira, W. (2014). *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press.

[64] Zhu, X. (Ed.). (2007). *Knowledge Discovery and Data Mining: Challenges and Realities: Challenges and Realities*. Igi Global.

[65] Zhu, X., & Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1), 1-130.

[66] Zhou, Y., & Goldman, S. (2004). Democratic co-learning. In *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on* (pp. 594-602). IEEE.

[67] Zhou, Z. H., & Li, M. (2005). Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on knowledge and Data Engineering*, 17(11), 1529-1541.

Ελληνική Βιβλιογραφία

[68] Παπαγεωργίου, Κ. (2008). Μοντέλα πρόβλεψης της πτώχευσης: κατασκευή υποδείγματος λογιστικής παλινδρόμησης στις εταιρείες εμπορίας ιατροτεχνολογικών προϊόντων. Διπλωματική εργασία, Πάντειο Πανεπιστήμιο Κοινωνικών και Πολιτικών Επιστημών, Αθήνα.

[69] Παπαδομιχελάκης, Θ. (2016). Εφαρμογή μεθόδων μηχανικής μάθησης με μερική επίβλεψη στην πρόβλεψη πτωχεύσεων επιχειρήσεων. Πτυχιακή εργασία, ΕΑΠ, Αθήνα.