ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Πρόγραμμα Μεταπτυχιακών Σπουδών (Π.Μ.Σ.)
Ψηφιακά Συστήματα & Υπηρεσίες
Μεγάλα Δεδομένα και Αναλυτική (Big Data and Analytics)

ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΣΕ ΤΡΑΠΕΖΙΚΑ ΔΕΔΟΜΕΝΑ ΜΕ ΤΗ ΒΟΗΘΕΙΑ ΤΟΥ ΠΡΟΓΡΑΜΜΑΤΟΣ WEKA

ΜΑΥΡΟΜΑΝΩΛΑΚΗΣ ΙΩΑΝΝΗΣ

ΑΜ: ΜΕ1719

Διπλωματική Εργασία

ΕΠΙΒΛΕΠΩΝ: ΑΝΑΠΛΗΡΩΤΗΣ ΚΑΘΗΓΗΤΗΣ ΜΙΧΑΗΛ ΦΙΛΙΠΠΑΚΗΣ

ΠΕΙΡΑΙΑΣ

ΦΕΒΡΟΥΑΡΙΟΣ 2020

ΠΕΡΙΛΗΨΗ

Η οικονομική κρίση που ξεκίνησε στην Ευρώπη το 2009 και διαρκεί ως και σήμερα στην Ελλάδα αναδεικνύει το πόσο επιτακτική είναι η ανάγκη παρακολούθησης του ρίσκου που αναλαμβάνει μια τράπεζα όταν δίνει ένα καταναλωτικό δάνειο η μια πιστωτικη καρτα. Ένα μοντέλο που μπορεί να προβλέψει έγκαιρα και έγκυρα τους πελάτες που δεν θα μπορέσουν να εξοφλήσουν στο άμεσο μέλλον αποτελεί εργαλείο ζωτικής σημασίας για κάθε τραπεζικό οργανισμό.

Στόχος μας είναι η περιγραφή και σύγκριση τέτοιων μοντέλων μέσω της εφαρμογής τους σε δεδομένα από μεγάλη ελληνική τράπεζα που αφορούν πελάτες.

Εφαρμόσαμε στα δεδομένα μας την μέθοδο της Λογιστικής Παλινδρόμησης (Logistic Regression), τα Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks) και τα Δέντρα Ταξινόμησης (Classification Trees) κάνοντας χρήση είτε όλων των μεταβλητών είτε κάνοντας χρήση μόνο όσων χαρακτηρίστηκαν σημαντικές μετά από την κατάλληλη προεπεξεργασία. Συγκεκριμένα, εφαρμόστηκε η βηματική μέθοδος επιλογής μεταβλητών και η ανάλυση ευαισθησίας για τον εντοπισμό των μεταβλητών που παίζουν σημαντικό ρόλο στο εξαγόμενο αποτέλεσμα.

ABSTRACT

The economic crisis that began in Europe in 2009 and continues to this day in Greece highlights the urgent need to monitor the risk a bank takes when giving a consumer loan or a credit card. A model that can predict timely and validly customers who will not be able to repay in the near future is a vital tool for any banking organization.

Our goal is to describe and compare such models through their application to data from a Big4 Bank Department.

We applied Logistic Regression, Artificial Neural Networks and Classification Trees to our data using either all variables or using only those that were considered significant after proper pretreatment. In particular, the step-by-step method of selecting variables and sensitivity analysis were used to identify the variables that play an important role in the output.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον κ. Φιλιππακη Μιχαήλ, καθηγητή του τμήματος Ψηφιακών Συστημάτων του Πανεπιστημίου Πειραιώς, για την επίβλεψη και τις συμβουλές του καθ' όλη τη διάρκεια της διπλωματικής μου εργασίας μέχρι την ολοκλήρωση της. Επίσης την οικογένεια μου για την στήριξη τους.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

Στη σημερινή κοινωνία ο όγκος καινούργιων πληροφοριών ολοένα και μεγαλώνει με αποτέλεσμα η συγκέντρωση και καταγραφή τους να δημιουργεί μια τεράστια βάση δεδομένων. Οι οργανισμοί και οι επιχειρήσεις συγκεντρώνουν μεγάλο όγκο δεδομένων από διαφορετικές πηγές σε καθημερινή βάση. Παρόλα αυτά, η μετατροπή της μεγάλης ποσότητας δεδομένων σε γνώση και η δυνατότητα εκμετάλλευσης των δεδομένων ώστε να αποκτήσει κάποιος καλύτερη οπτική των καταστάσεων, παραμένει μία δυσκολία για τους περισσότερους οργανισμούς και επιχειρήσεις. Την ανάγκη καταγραφής, διαχείρισης των μεγάλων βάσεων δεδομένων αλλά και ανακάλυψης καινούργιας πληροφορίας μέσω αυτών έρχεται να ικανοποιήσει η εξόρυξη γνώσης από δεδομένα. Η εξόρυξη δεδομένων διευθετεί και επεξεργάζεται τα δεδομένα με την ανακάλυψη και αξιοποίηση προτύπων, δομών, μοντέλων, τάσεων και συσχετίσεων, με ένα αυτοματοποιημένο τρόπο. Η βέλτιστη απόκτηση γνώσης για μια επιχείρηση έχει ως αποτέλεσμα το σχεδιασμό και τη λήψη των βέλτιστων αποφάσεων, την αύξηση της παραγωγικότητας σε στρατηγικά και επιχειρησιακά επίπεδα και κατά συνέπεια την αύξηση της ανταγωνιστικότητας της έναντι άλλων επιχειρήσεων.

Η ιατρική, η δημογραφία, η στατιστική και η οικονομία αποτελούν κάποιους από τους κλάδους στους οποίους εφαρμόζεται η τεχνική της εξόρυξης γνώσης. Σίγουρα δεν εφαρμόζεται σε όλους τους κλάδους στον ίδιο βαθμό αλλά σε κάποιους κλάδους όπως της οικονομίας και του τραπεζικού συστήματος, η ομαλή και η αναπτυσσόμενη λειτουργία του επηρεάζεται και εξαρτάται από τον τρόπο εφαρμογής της εξόρυξης γνώσης. Συγκεκριμένα η τράπεζα μπορεί να πάρει απόφαση για το αν θα δανειοδοτήσει κάποιο φυσικό πρόσωπο ή μια επιχείρηση λαμβάνοντας υπόψιν την πληροφορία που θα λάβει μέσω της διαδικασίας της εξόρυξης γνώσης.

1.1 ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ ΚΑΙ ΔΕΔΟΜΕΝΩΝ (DATA MINING)

Η εξόρυξη γνώσης από δεδομένα (data mining) αποτελεί μια προηγμένη τεχνική που στόχος της είναι η εξαγωγή χρήσιμων πληροφοριών και προτύπων βοηθώντας τις επιχειρήσεις να εστιάσουν στην σημαντική πληροφορία που βρίσκεται μέσα σε βάσεις δεδομένων. Για την καλύτερη λειτουργία και αποτελεσματικότητα απαιτείται να εφαρμόζονται οι τεχνικές πρόβλεψης και περιγραφής σε μεγάλο όγκο δεδομένων. Η διαδικασία της ανεύρεσης γνώσης και ανάλυσης από βάσεις δεδομένων ονομάζεται Knowledge Discovery in Databases (KDD), ενώ ο όρος εξόρυξη δεδομένων αναφέρεται στις διάφορες μεθόδους που χρησιμοποιούνται για την ανάλυση αυτή. Η ανεύρεση γνώσης αποτελεί μια επαναληπτική διαδικασία μιας σειράς βημάτων τα οποία έχουν ως αποτέλεσμα την συλλογή δεδομένων και την ανακάλυψη και εξαγωγή χρήσιμης πληροφορίας από αυτά.

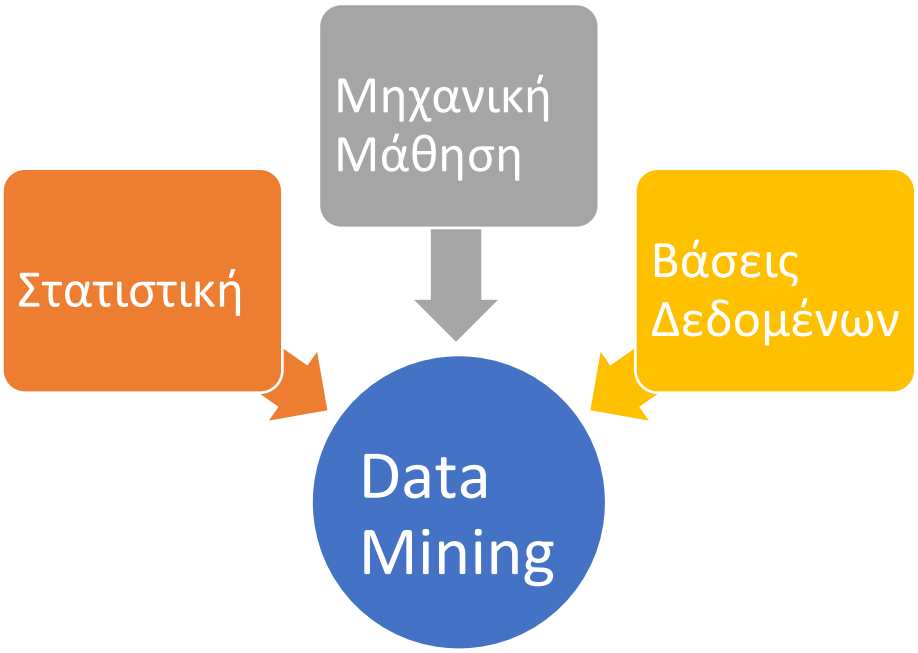
1.2 DATA MINING ΣΤΗΝ ΟΙΚΟΝΟΜΙΑ

Ο παγκόσμιος οικονομικό κλάδος , που έχει διαμορφωθεί στην σημερινή εποχή χαρακτηρίζεται από πολλές αλληλεξαρτήσεις που είναι πάρα πολύ σημαντικές. Η αλληλεξάρτηση των διαφόρων οικονομιών έχει δύο κυρίως συνιστώσες, το διεθνές εμπόριο αγαθών και υπηρεσιών και την παγκόσμια ολοκλήρωση των χρηματοπιστωτικών αγορών. Οι επιχειρήσεις πλέον παρακολουθούν τις τάσεις τόσο της εγχώριας όσο και της διεθνούς αγοράς ώστε να παράγουν αγαθά και υπηρεσίες. Επίσης σε συνεχή βάση παρακολουθούν τις διεθνείς τιμές συναλλάγματος και των ενεργειακών πόρων προκειμένου να καταστρώσουν την μελλοντική τους στρατηγική ανάπτυξης. Παράλληλα, η παραγωγικότητα παρουσιάζει σημαντική αύξηση, αφού η επέκταση του διεθνούς εμπορίου επιτρέπει στις οικονομίες να ειδικευτούν στην παραγωγή αγαθών και υπηρεσιών που ταιριάζουν στους ανθρώπινους και φυσικούς πόρους τους. Ωστόσο, η επέκταση αυτή υποδηλώνει και την ολοένα αυξανόμενη εξάρτηση των εθνικών οικονομιών από τις εξελίξεις στις υπόλοιπες χώρες. Σε ότι αφορά στην ολοκλήρωση των παγκόσμιων χρηματοπιστωτικών αγορών, αυτή οδηγεί σε αύξηση την παγκόσμια παραγωγικότητα, όπως συμβαίνει και στην περίπτωση ενός ολοκληρωμένου παγκόσμιου συστήματος εμπορίου, καθώς επιτρέπει στις αποταμιεύσεις να κατευθύνονται στις χρήσεις με την υψηλότερη απόδοση ανεξάρτητα από τον τόπο διαμονής των αποταμιευτών και επενδυτών. Η πρόβλεψη των τιμών συναλλάγματος , των διεθνών τιμών του πετρελαίου και η ρευστότητα των εταιριών ελκύει το ενδιαφέρον πολλών ερευνητών και επαγγελματιών που δραστηριοποιούνται στις αγορές. Αυτά τα ενδιαφέροντα στοιχεία, επηρεάζονται από πολλούς διαφορετικούς πολιτικούς, οικονομικούς, κοινωνικούς, αλλά και ψυχολογικούς παράγοντες, με αποτέλεσμα να καθίσταται αρκετά δύσκολη η πρόβλεψή τους.

Όμως λόγω των σημαντικών κερδών που μπορεί να προέλθουν από επενδύσεις στις αγορές συναλλάγματος και μετοχών , πολλοί ερευνητές έχουν δημιουργήσει διάφορα μοντέλα για να προβλέψουν τις διακυμάνσεις των ισοτιμιών των νομισμάτων και των τιμών των μετοχών. Σήμερα υπάρχουν αρκετά μοντέλα πρόβλεψης τα οποία δίνουν διάφορα επίπεδα επιτυχίας. Ένας άλλος σημαντικός τομέας που εφαρμόζεται η εξόρυξη δεδομένων είναι η οικονομία των επιχειρήσεων όπως είναι οι τράπεζες. Συνήθως τα δεδομένα τους είναι αξιόπιστα, ολοκληρωμένα και έχουν υψηλή ποιότητα και απαιτούν συστηματική μέθοδο για την ανάλυση αυτών. Η συνεισφορά της εξόρυξης δεδομένων στην επιστήμη της οικονομίας βοηθάει στην συλλογή και κατανόηση των δεδομένων, στην βελτίωση δεδομένων, στην δημιουργία και εκτίμηση ενός μοντέλου και στην ανάπτυξη αυτού για τυχόν προβλέψεις και αποφάσεις που θα θελήσουν να παρθούν .

Η σωστή ανάλυση των οικονομικών δεδομένων μας διευκολύνει στο να παίρνουμε καλύτερες αποφάσεις ενεργώντας σύμφωνα με την ανάλυση της αγοράς. Τα εργαλεία και οι τεχνικές της εξόρυξης δεδομένων βοηθούν στο να αναλύσουμε τα οικονομικά δεδομένα με τους παρακάτω τρόπους:

- Τα δεδομένα που συλλέγονται από διάφορες οικονομικά πήγες, όπως οι τράπεζες ,συγκεντρώνονται αρχικά στην αποθήκη δεδομένων (data warehouse).
- Οι Μέθοδοι της εξόρυξης όπως η επιλογή χαρακτηριστικών (feature selection) βοηθάει στην αυτοποίηση ποικίλων χαρακτηριστικών όπως το επίπεδο εισοδήματος του πελάτη, την εξόφληση ανάλογα με τα έσοδα, την πιστωτική του ιστορία κτλ. Με την επεξεργασία αυτών των χαρακτηριστικών, π.χ. η τράπεζα μπορεί να αποφασίσει για τις πολιτικές δανειοδότησης βάσει των σχετικά χαμηλών κινδύνων
- Οι τεχνικές της συσταδοποίησης και της ταξινόμησης βοηθούν τα οικονομικά ινστιτούτα να ομαδοποιούν διάφορους πελάτες που έχουν κοινά χαρακτηριστικά. Η αποτελεσματική συσταδοποίηση και οι μέθοδοι φιλτραρίσματος βοηθούν π.χ. τις τράπεζες να ταυτοποιούν μία ομάδα πελατών, να συσχετίζουν ένα νέο πελάτη με την παρούσα ομάδα και να τους παρέχουν κοινά οφέλη. Τα εργαλεία της εξόρυξης δεδομένων βοηθούν επίσης να αναγνωρίζουν τις απάτες και τα εγκλήματα από παραποιημένα δεδομένα από τις διάφορες βάσεις δεδομένων και από το ιστορικό συναλλαγών που έγιναν από τους πελάτες.



ΕΙΚΟΝΑ 1:ΟΙ 'ΡΙΖΕΣ' ΤΗΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

1.3 ΑΞΙΟΛΟΓΗΣΗ ΕΡΓΑΛΕΙΩΝ DATA MINING

Εφόσον μια επιχείρηση έχει πάρει την στρατηγική απόφαση να προχωρήσει στη χρησιμοποίηση των διαδικασιών του Data Mining, θα πρέπει στη συνέχεια να γίνει η επιλογή του κατάλληλου εργαλείου. Υπάρχουν τρία βασικά κριτήρια, τα οποία θα πρέπει κανείς να χρησιμοποιήσει οπωσδήποτε για να μπορέσει να αξιολογήσει ένα εργαλείο Data Mining.

- Εγκυρότητα. Το εργαλείο για το Data Mining είναι απαραίτητο να έχει την ικανότητα να παράγει όσο το δυνατόν πιο έγκυρα μοντέλα. Η εγκυρότητα των μοντέλων θα πρέπει να μην επηρεάζεται από τις εξωτερικές συνθήκες ή τον «θόρυβο» των δεδομένων.
- Ερμηνεία. Θα πρέπει το σύστημα, εκτός από την παραγωγή ορθών μοντέλων, να μπορεί να τα κάνει κατανοητά στον τελικό χρήστη. Θα πρέπει ο χρήστης να αισθάνεται ότι καταλαβαίνει πλήρως την μοντελοποίηση που έχει γίνει, ώστε να μπορεί να κατανοήσει και τα συμπεράσματα που προκύπτουν από την περαιτέρω επεξεργασία.

- Διασύνδεση. Είναι απαραίτητο να υπάρχει διασύνδεση της διαδικασίας του Data Mining με την επιχειρησιακή λειτουργία της εταιρίας ή γενικότερα του φορέα μου μελετάμε. Έτσι λοιπόν θα πρέπει και το σύστημα για το Data Mining να έχει την δυνατότητα να συνδέεται όσο το δυνατόν με περισσότερα στάδια της λειτουργίας της επιχείρησης ή του φορέα που μελετάμε κάθε φορά και να συλλέγει δεδομένα από διάφορα σημεία της ροής των πληροφοριών

Η ικανοποίηση αυτών των κριτηρίων από ένα σύστημα Data Mining είναι πολύ σημαντική για την παραγωγή αποτελεσματικών και ρεαλιστικών μοντέλων, τα οποία θα έχουν διαχρονικότητα και τη δυνατότητα να προσαρμόζονται στις νέες συνθήκες και στις μεταβολές των δεδομένων.

ΚΕΦΑΛΑΙΟ 2

2.1 ΤΑ ΒΗΜΑΤΑ ΤΗΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Η ανεύρεση γνώσης είναι μια επαναληπτική διαδικασία που αποτελείται από μια σειρά βημάτων, τα οποία οδηγούν από τη συλλογή των δεδομένων στην ανακάλυψη και εξαγωγή χρήσιμης πληροφορίας από αυτά. Τα βήματα από τα οποία αποτελείται η διαδικασία ανεύρεσης γνώσης είναι τα ακόλουθα:

<u>1ο ΒΗΜΑ</u>: Ανάπτυξη και κατανόηση της περιοχής της εφαρμογής

Σε αυτό το προκαταρκτικό στάδιο γίνεται προετοιμασία για την κατανόηση του πλαισίου δράσης. Πρέπει να γίνει σαφές, δηλαδή, ποιες αποφάσεις θα ληφθούν σχετικά με μετασχηματισμούς, αλγορίθμους, αναπαράσταση κ.λπ. Το Βήμα αυτό βοηθά στην κατανόηση των στόχων από τον τελικό χρήστη, καθώς και στην εύρεση του περιβάλλοντος όπου θα δράσει η διαδικασία ανακάλυψης της γνώσης. Στα πλαίσια αυτά περιλαμβάνεται και η προγενέστερη γνώση του υπό εξέταση τομέα. Είναι πιθανό να απαιτηθεί επανάληψη αυτού του Βήματος στην πορεία

<u>2ο ΒΗΜΑ</u>
δεδομένων : Επιλογή και δημιουργία ενός κατάλληλου συνόλου

Έχοντας ορίσει τους στόχους, θα έπρεπε να έχουν προσδιοριστεί και τα δεδομένα που θα χρησιμοποιηθούν. Το Βήμα αυτό περιλαμβάνει τον εντοπισμό των δεδομένων που είναι διαθέσιμα, την απόκτηση επιπρόσθετων αναγκαίων δεδομένων και την ενσωμάτωση όλων αυτών σε ένα σύνολο δεδομένων το οποίο θα περιλαμβάνει τα χαρακτηριστικά (attributes) που θα ληφθούν υπόψη. Το βήμα αυτό είναι πολύ σημαντικό, καθώς η Εξόρυξη Δεδομένων μαθαίνει και ανακαλύπτει από τα δεδομένα που έχει εκείνη τη στιγμή στη διάθεσή της. Σε αυτή τη βάση κατασκευάζονται και τα μοντέλα. Είναι πιθανό, όμως, να προκύψουν προβλήματα στην περίπτωση όπου λείπουν χαρακτηριστικά από κάποιες παρατηρήσεις, καθώς μπορεί να δημιουργηθούν σφάλματα στη μελέτη. Άρα, χρειάζεται η μέγιστη δυνατή συλλογή χαρακτηριστικών. Από την άλλη πλευρά, όμως, αυτή η ανάγκη ανεβάζει το κόστος διεξαγωγής της ανάλυσης. Για το λόγο αυτό, η διαδικασία της KDD αναλαμβάνει να αξιοποιήσει αρχικά το βέλτιστο διαθέσιμο σύνολο δεδομένων και στη συνέχεια επεκτείνεται και παρατηρεί τα αποτελέσματα στα πλαίσια της ανακάλυψης γνώσης και μοντελοποίησης.

<u>3ο ΒΗΜΑ</u>: Προ-επεξεργασία και καθαρισμός δεδομένων.

Ένα πολύ σημαντικό σημείο που μας απασχολεί είναι η αξιοπιστία των δεδομένων, η οποία μελετάται μέσα από αυτό το απαραίτητο Βήμα της διαδικασίας. Στα πλαίσια της αναζήτησης ενός αξιόπιστου συνόλου δεδομένων, οφείλουμε να πραγματοποιήσουμε καθαρισμό δεδομένων (data cleaning). Με τη χρήση του όρου αυτού εννοούμε τη διαχείριση ελλειπουσών τιμών (missing values) και την απομάκρυνση θορύβου (noise) ή έκτροπων παρατηρήσεων (outliers). Οι διαδικασίες καθαρισμού των δεδομένων μπορούν να επιτευχθούν μέσω σύνθετων στατιστικών μεθόδων ή χρησιμοποιώντας έναν αλγόριθμο Εξόρυξης Δεδομένων

<u>4ο ΒΗΜΑ</u>: Επιλογή της κατάλληλης μεθόδου εξόρυξης δεδομένων

Ύστερα από όσα βήματα έχουμε εκτελέσει, είμαστε σε θέση να αποφασίσουμε ποιον τύπο Εξόρυξης Δεδομένων θα χρησιμοποιήσουμε (ταξινόμηση, παλινδρόμηση, συσταδοποίηση). Αυτή η επιλογή βασίζεται περισσότερο στους στόχους της KDD, αλλά και στα βήματα που έχουν ήδη προηγηθεί. Όπως έχουμε ήδη αναφέρει και θα σχολιάσουμε και παρακάτω, οι δύο βασικοί στόχοι της Εξόρυξης Δεδομένων είναι η περιγραφή και η πρόβλεψη. Οι τεχνικές Εξόρυξης Δεδομένων βασίζονται στην πλειοψηφία τους στην επαγωγική εκμάθηση (inductive learning), όπου κατασκευάζεται ένα σαφές ή εννοούμενο μοντέλο μέσω γενίκευσης ενός επαρκούς αριθμού εκπαιδευτικών παραδειγμάτων (training examples). Βασική προϋπόθεση είναι ότι αυτό το μοντέλο εκπαίδευσης (trained model) θα μπορεί να εφαρμοστεί σε μελλοντικές περιπτώσεις. Επίσης, η στρατηγική αυτή λαμβάνει υπόψη την περίπτωση μετά-εκμάθησης (metalearning) για το συγκεκριμένο σύνολο των διαθέσιμων δεδομένων

<u>5ο ΒΗΜΑ</u>: Επιλογή και εκτέλεση αλγορίθμου εξόρυξης δεδομένων.

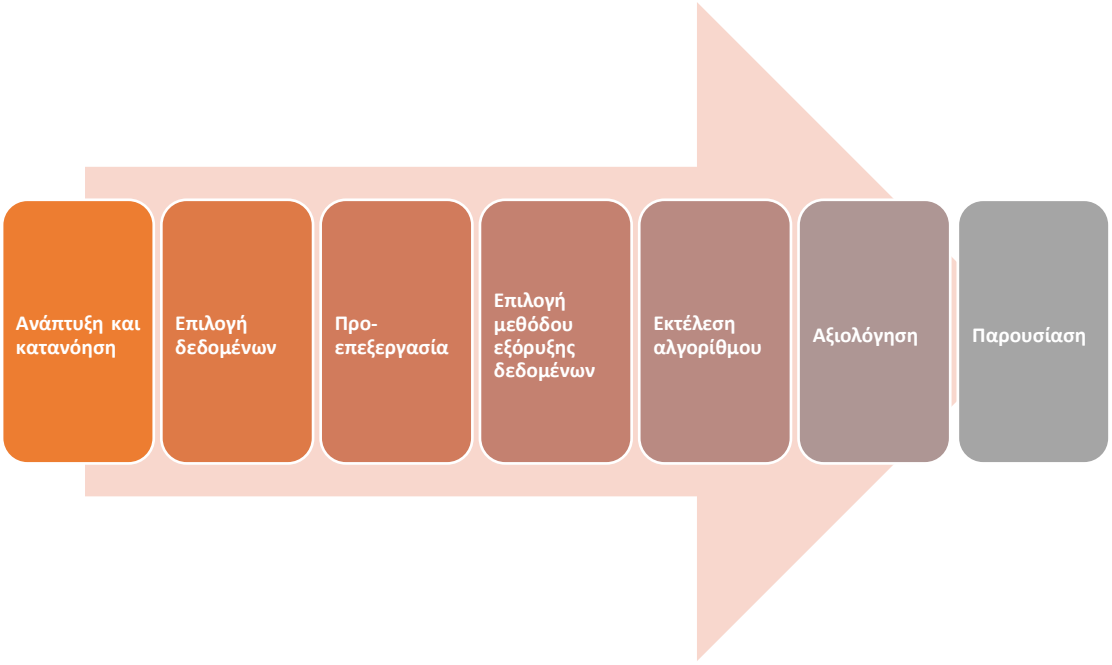
Έχοντας ορίσει τη στρατηγική, μπορούμε να επιλέξουμε τον τρόπο επίτευξης του στόχου. Στο στάδιο αυτό εφαρμόζονται ευφυείς μέθοδοι με σκοπό την αναζήτηση ενδιαφερόντων προτύπων γνώσης. Για παράδειγμα, ένας έλεγχος ακρίβειας θα ήταν καλύτερα να γίνει μέσω νευρωνικών δικτύων, ενώ για την κατανόηση της δομής (understandability) θα επιλέγονταν τα δέντρα αποφάσεων. Τα πρότυπα που αναζητούνται θα μπορούσαν να είναι μια συγκεκριμένης αντιπροσωπευτικής μορφής ή ενός συνόλου αντιπροσωπεύσεων, όπως κανόνες ταξινόμησης, δέντρα, παλινδρόμηση, συσταδοποίηση κ.λπ. Η απόδοση και τα αποτελέσματα της μεθόδου εξόρυξης δεδομένων εξαρτώνται από τα προηγούμενα Βήματα.

<u>6ο ΒΗΜΑ</u>: Αξιολόγηση

Σε αυτό το στάδιο γίνεται εκτίμηση και ερμηνεία των εξορυχθέντων προτύπων (κανόνες, αξιοπιστία κ.λπ.), λαμβάνοντας υπόψη τους στόχους που είχαν τεθεί στο πρώτο Βήμα. Επίσης, παρατηρούμε την επίδραση των Βημάτων 2, 3 και 4 (προεπεξεργασία δεδομένων) στον αλγόριθμο Εξόρυξης Δεδομένων που έχει επιλεγεί μέσα από τα βήματα 5, 6 και 7 (εξόρυξη δεδομένων). Για παράδειγμα, μπορεί να κριθεί αναγκαία η προσθήκη χαρακτηριστικών (μεταβλητών) στο βήμα 4, ώστε να επαναληφθεί η εφαρμογή της αλυσίδας KDD από εκεί. Το Βήμα της αξιολόγησης επικεντρώνεται στην κρίση εάν το προκύπτων μοντέλο είναι κατανοητό και χρήσιμο, καθώς και στην επιλογή των πιο ενδιαφερόντων εξαγόμενων προτύπων. Επιπλέον, στο βήμα αυτό, τεκμηριώνεται η ανακαλυφθείσα γνώση και είναι πλέον διαθέσιμη για περαιτέρω χρήση.

7ο ΒΗΜΑ: Παρουσίαση και χρήση της ανακαλυφθείσας γνώσης

Στο τελευταίο Βήμα, η εξορυγμένη γνώση ενσωματώνεται στο σύστημα για περαιτέρω δράση (πραγματοποίηση αλλαγών στο σύστημα, μέτρηση επιδράσεων). Η επιτυχία αυτού του βήματος αποδεικνύει την αποτελεσματικότητα χρήσης της αλυσίδας KDD. Επιπλέον, μέσα από αυτό το βήμα γίνεται έλεγχος για επίλυση τυχών συγκρούσεων με προηγούμενη εξορυγμένη γνώση. Είναι πιθανό να αλλάξουν ορισμένες δομές δεδομένων, καθώς κάποιες μεταβλητές μπορεί να μην είναι πλέον διαθέσιμες. Επίσης, μπορεί να αλλάξει η περιοχή δράσης των δεδομένων, καθώς μπορεί να προκύψει για μια μεταβλητή μια τιμή η οποία να μην είχε υποτεθεί πριν.



ΕΙΚΟΝΑ 2-ΤΑ ΒΗΜΑΤΑ

Πολλές φορές κάποια από τα παραπάνω βήματα μπορούν να συνδυαστούν μεταξύ τους για το καλύτερο δυνατό αποτέλεσμα. Για παράδειγμα, τα βήματα του καθαρισμού και της ενσωμάτωσης των δεδομένων, μπορούν να υλοποιηθούν μαζί με στόχο την δημιουργία μια αποθήκης δεδομένων. Με την ίδια λογική μπορούν να συνδυαστούν και τα βήματα της επιλογής και τροποποίησης των δεδομένων

2.2 ΣΤΟΧΟΙ ΤΗΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Οι μέθοδοι εξόρυξης γνώσης στοχεύουν στην ανακάλυψη στοιχειών που θα είναι χρήσιμα για τους οργανισμούς και τις επιχειρήσεις. Πληροφορίες για τυποποιημένες μορφές όπως για παράδειγμα, ότι υπάρχουν πελάτες που θα ψωνίσουν περισσότερο από δύο φορές σε περίοδο εκπτώσεων ή προσφορών, ή είναι πιθανό να αγοράσουν τουλάχιστον μια φορά κατά την διάρκεια των εορταστικών ημερών, Πάσχα και Χριστουγέννων, είτε για συσχετίσεις όπως όταν ένας πελάτης αγοράζει dvd player τότε πιθανότατα να αγοράσει και κάποια άλλη ηλεκτρονική συσκευή, μπορεί να αποτελέσουν καθοριστικούς παράγοντες για την λήψη αποφάσεων όσον αφορά τη λειτουργία μιας εμπορικής επιχείρησης. Αυτό συμβαίνει επειδή μπορεί να ληφθούν αποφάσεις σχετικά με το ωράριο, το ύψος και τη διάρκεια των εκπτώσεων, ακόμη και για την τοποθέτηση των προϊόντων μέσα στα καταστήματα.

Παράλληλα τέτοιου είδους πληροφορίες χρησιμοποιούνται για τον προγραμματισμό χρήσης πρόσθετων αποθηκευτικών χώρων ή και για τον σχεδιασμό διαφορετικών στρατηγικών μάρκετινγκ. Τα στελέχη της επιχείρησης, που είναι υπεύθυνα για την λήψη των αποφάσεων εκμεταλλεύονται τις δυνατότητες της εξόρυξης γνώσης και μετατρέπουν τις γνώσεις σε επιτυχή αποτελέσματα. Παρακάτω περιγράφονται και αναλύονται οι στόχοι της εξόρυξης δεδομένων.

- Πρόβλεψη: Περιλαμβάνει την χρήση μερικών μεταβλητών ή χαρακτηριστικών μιας βάσης δεδομένων για την πρόβλεψη άγνωστων ή μελλοντικών τιμών χρήσιμων μεταβλητών. Με άλλα λόγια, οι διαδικασίες πρόβλεψης της εξόρυξης δεδομένων (predictive data mining tasks), προσπαθούν να κάνουν εκτιμήσεις βγάζοντας συμπεράσματα από τα διαθέσιμα δεδομένα. Η προσπάθεια πρόβλεψης μελλοντικών συμπεριφορών έχει ως στόχο να ληφθούν αποφάσεις που να μεγιστοποιούν το κέρδος και να προλαμβάνουν δυσάρεστες καταστάσεις. Τα αποτελέσματα της εξόρυξης μπορεί να είναι πληροφορίες σχετικές με το ύψος των πωλήσεων ενός καταστήματος για μια συγκεκριμένη χρονική περίοδο, αλλά και αν το κλείσιμο μιας γραμμής παραγωγής θα είχε θετική επίδραση στις πωλήσεις. Συγχρόνως σε επιστημονικό επίπεδο,η μελέτη παλαιότερων

σεισμικών φαινόμενων ίσως να οδηγούσε στην πρόβλεψη σεισμικής δραστηριότητας.

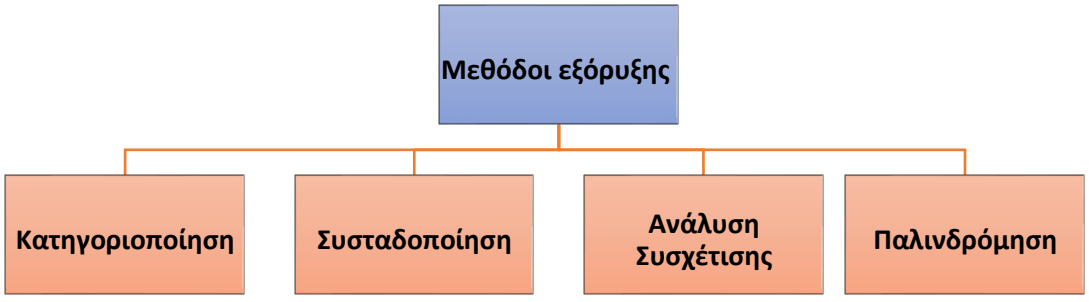
- Περιγραφή: Είναι η διαδικασία η οποία επικεντρώνεται στην ανακάλυψη προτύπων και αναπαριστά τα δεδομένα μιας πολύπλοκης βάσης δεδομένων με όσο το δυνατό πιο κατανοητό και αξιοποιήσιμο τρόπο. Με άλλα λόγια, οι περιγραφικές διαδικασίες της εξόρυξης δεδομένων (descriptive data mining tasks) περιγράφουν τις γενικές ιδιότητες των υπαρχόντων διαθέσιμων δεδομένων.

Αν και τα όρια μεταξύ της πρόβλεψης και της περιγραφής δεν είναι απολύτως ξεκάθαρα (μερικά από τα πρότυπα πρόβλεψης μπορούν να είναι περιγραφικά, στο βαθμό που είναι κατανοητά και αντίστροφα), η διάκριση είναι χρήσιμη για την κατανόηση του γενικού στόχου ανακάλυψης. Η σχετική σημασία της πρόβλεψης και της περιγραφής για συγκεκριμένες εφαρμογές εξόρυξης, μπορεί να ποικίλει αρκετά.

2.3 ΜΕΘΟΔΟΙ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ ΚΑΙ ΔΕΔΟΜΕΝΩΝ

Η εξόρυξη γνώσης επιτυγχάνεται μέσα από ένα ευρύ φάσμα αλγόριθμων οι οποίοι χρησιμοποιούν τεχνικές από διαφορετικούς τομείς όπως είναι η στατιστική, η μηχανική μάθηση και η αναγνώριση προτύπων. Υπάρχει μια πληθώρα υπολογιστικών μεθόδων εξόρυξης γνώσης , οι βασικότερες από τις μεθόδους της εξόρυξης, μέσω των οποίων επιτυγχάνονται οι στόχοι που αναφέραμε προηγουμένως, είναι οι εξής:

- Κατηγοριοποίηση (Classification)
- Συσταδοποίηση (Clustering)
- Ανάλυση Συσχέτισης
- Παλινδρόμηση



ΕΙΚΟΝΑ 3 -ΜΕΘΟΔΟΙ ΕΞΟΡΥΞΗΣ

2.3.1 ΣΥΣΤΑΔΟΠΟΙΗΣΗ (CLUSTERING)

Η συσταδοποίηση ή αλλιώς ομαδοποίηση (clustering) αφορά τον διαχωρισμό (partition) των αντικειμένων μιας βάσης δεδομένων σε μη συνδεδεμένες μεταξύ τους και ομοιογενείς ομάδες, κατά τέτοιο τρόπο ώστε τα αντικείμενα του συνόλου που ανήκουν σε μια ομάδα, να είναι πιο όμοια μεταξύ τους, παρά με τα αντικείμενα που ανήκουν σε διαφορετικές ομάδες. Ένα ιδιαίτερο χαρακτηριστικό της ομαδοποίησης, σε αντίθεση με την κατηγοριοποίηση, είναι ότι η δομή και το πλήθος των ομάδων είναι καταρχάς άγνωστα και καθορίζονται δε από τον εκάστοτε αλγόριθμο συσταδοποίησης. Αυτοί οι αλγόριθμοι βασίζονται στο σύνολο τους στην αρχή της μεγιστοποίησης της ομοιότητας ανάμεσα στα αντικείμενα την ίδιας ομάδας (intra-class similarity) και την ταυτόχρονη αρχή της ελαχιστοποίησης της ομοιότητας μεταξύ των αντικειμένων διαφορετικών ομάδων (inter-class similarity). Αξίζει να σημειωθεί ότι η ερμηνεία των ομάδων που προκύπτουν από την ανωτέρω διαδικασία καθορίζεται από τον εκάστοτε χρήστη.

Από τον παραπάνω ορισμό προκύπτει άμεσα και η βασική διαφορά μεταξύ κατηγοριοποίησης και συσταδοποίησης. Στην κατηγοριοποίηση ο αριθμός και η ουσία των συστάδων αποτελεί πληροφορία εκ των προτέρων γνωστή. Εξαιτίας αυτού, στη συσταδοποίηση εφαρμόζεται πάντα μη εποπτευόμενη μάθηση, εν αντιθέση με την κατηγοριοποίηση όπου λόγω της πρότερης γνώσης των κλάσεων κάνουμε χρήση της εποπτευόμενης μάθησης. Στην συσταδοποίηση δεν υπάρχουν προκαθορισμένες κατηγορίες ομαδοποίησης αλλά οι εγγραφές συγκεντρώνονται σε ομάδες με βάση το κριτήριο που θέτει ο χρήστης για κάθε συστάδα όπως για παράδειγμα, η ομαδοποίηση πελατών που αγοράζουν παρόμοια αγαθά. Σκοπός είναι η δημιουργία συστάδων με όσο το δυνατόν περισσότερα κοινά χαρακτηριστικά εντός της εκάστοτε ομάδας, ενώ ταυτόχρονα η μία ομάδα από την άλλη θα πρέπει να διαφοροποιείται ικανοποιητικά ώστε να μη συγχέονται. Δηλαδή θα πρέπει να δημιουργηθούν διακριτές ομάδες με βάση ξεκάθαρα χαρακτηριστικά που περιγράφουν την κάθε ομάδα και την κάνουν να ξεχωρίζει από τις υπόλοιπες.

Στον οικονομικό τομέα είναι ιδιαίτερα σημαντικό για της επιχειρήσεις να μπορούν να ομαδοποιούν τους πελάτες τους σε συγκεκριμένες κατηγορίες. Με βάση αυτές τις κατηγορίες μπορούν να αξιολογούν έναν νέο πελάτη με βάση την ομάδα στην οποία κατατάσσεται ή ακόμα να προσδιορίσουν τα χαρακτηριστικά των πελατών που αποφέρουν μεγάλα κέρδη στην εταιρεία. Με βάση αυτόν τον διαχωρισμό των πελατών μπορούν να προσανατολίσουν την στρατηγική της

εταιρείας στην εξειδικευμένη εξυπηρέτηση ορισμένων πελατειακών ομάδων. Για παράδειγμα από την ανάλυση ενός πολύ μεγάλου συνόλου πελατών, μπορεί να μειωθεί το κόστος μίας διαφημιστικής εκστρατείας που βασίζεται στην αποστολή διαφημιστικών φυλλαδίων. Αυτό γίνεται περιορίζοντας το πλήθος των πελατών στους οποίους απευθύνεται, επιλέγοντας αυτούς με μεγαλύτερη πιθανότητα να αντιδράσουν θετικά. Για να μπορέσει να γίνει η επιλογή του κατάλληλου αλγορίθμου απαραίτητη προϋπόθεση είναι η μελέτη των δεδομένων που θα χρησιμοποιηθούν για τον προσδιορισμό κυρίως του κριτηρίου ομοιότητας των εγγραφών μίας ομάδας.

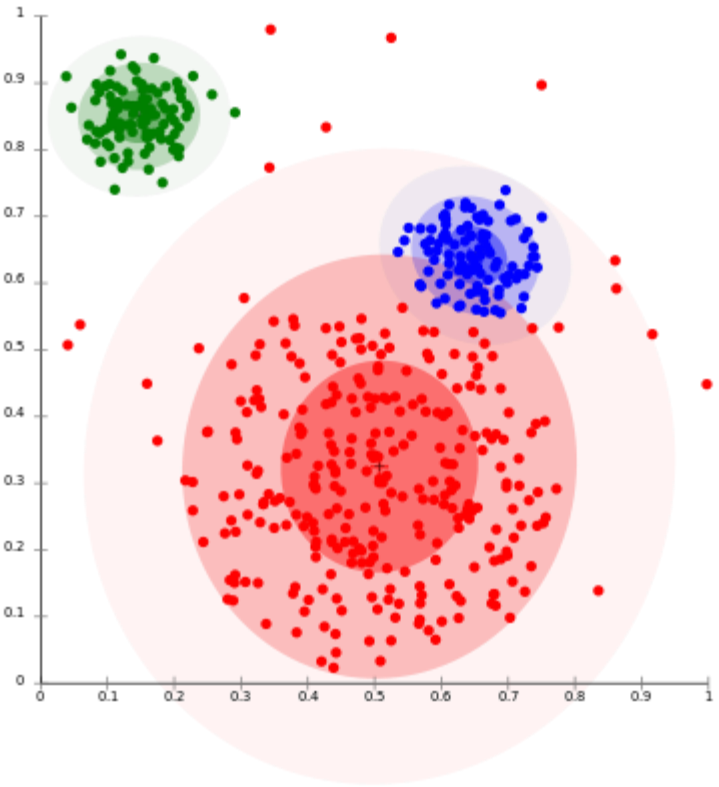
Η Συσταδοποίηση διακρίνεται σε τρεις βασικές μεθόδους:

1. Μέθοδοι διαχωρισμού (partitioning methods): Δημιουργούν ομάδες από ένα δεδομένο αρχικό σύνολο αντικειμένων με κάθε ομάδα να αντιπροσωπεύει ένα cluster και να ικανοποιούνται οι εξής δύο συνθήκες: (α) κάθε cluster περιέχει τουλάχιστον ένα αντικείμενο και (β) κάθε αντικείμενο ανήκει σε ένα μόνο cluster.

2. Ιεραρχικές μέθοδοι (hierarchical methods): Διασπούν το αρχικό σύνολο δεδομένων δημιουργώντας μια ιεραρχική δομή από clusters και διακρίνονται σε agglomerative (bottom-up) ή divisive (top-down) ανάλογα με τον τρόπο που γίνεται η διάσπαση.

3. Μέθοδοι βασισμένες σε μοντέλα (model-based methods): Υποθέτουν ότι καθένα από τα clusters περιγράφεται από ένα μαθηματικό μοντέλο και εντοπίζουν τα αντικείμενα που ανήκουν σε κάθε cluster, ώστε να ικανοποιούν το αντίστοιχο μοντέλο.

Βέβαια σε ένα πρόβλημα συσταδοποίησης δεν υπάρχει μόνο μια λύση. Επίσης βασικό ζήτημα είναι και η επιλογή του πλήθους των συστάδων. Είναι ένα θέμα να υπολογιστεί το ακριβές πλήθος των συστάδων που απαιτείται. Μια καλή αρχικοποίηση των κεντροειδών των συστάδων είναι πάρα πολύ σημαντική. Τέλος υπάρχει πιθανότητα κάποιες συστάδες να είναι άδειες (κανένα στοιχείο μέσα τους) εάν τα κεντροειδή του βρίσκονται αρχικά μακριά από τα δεδομένα. Ένα είναι σίγουρο ότι η συσταδοποίηση δημιουργεί κανόνες για ανάθεση νέων εγγραφών σε κλάσεις και χρησιμεύει για διάγνωση και αναγνώριση.

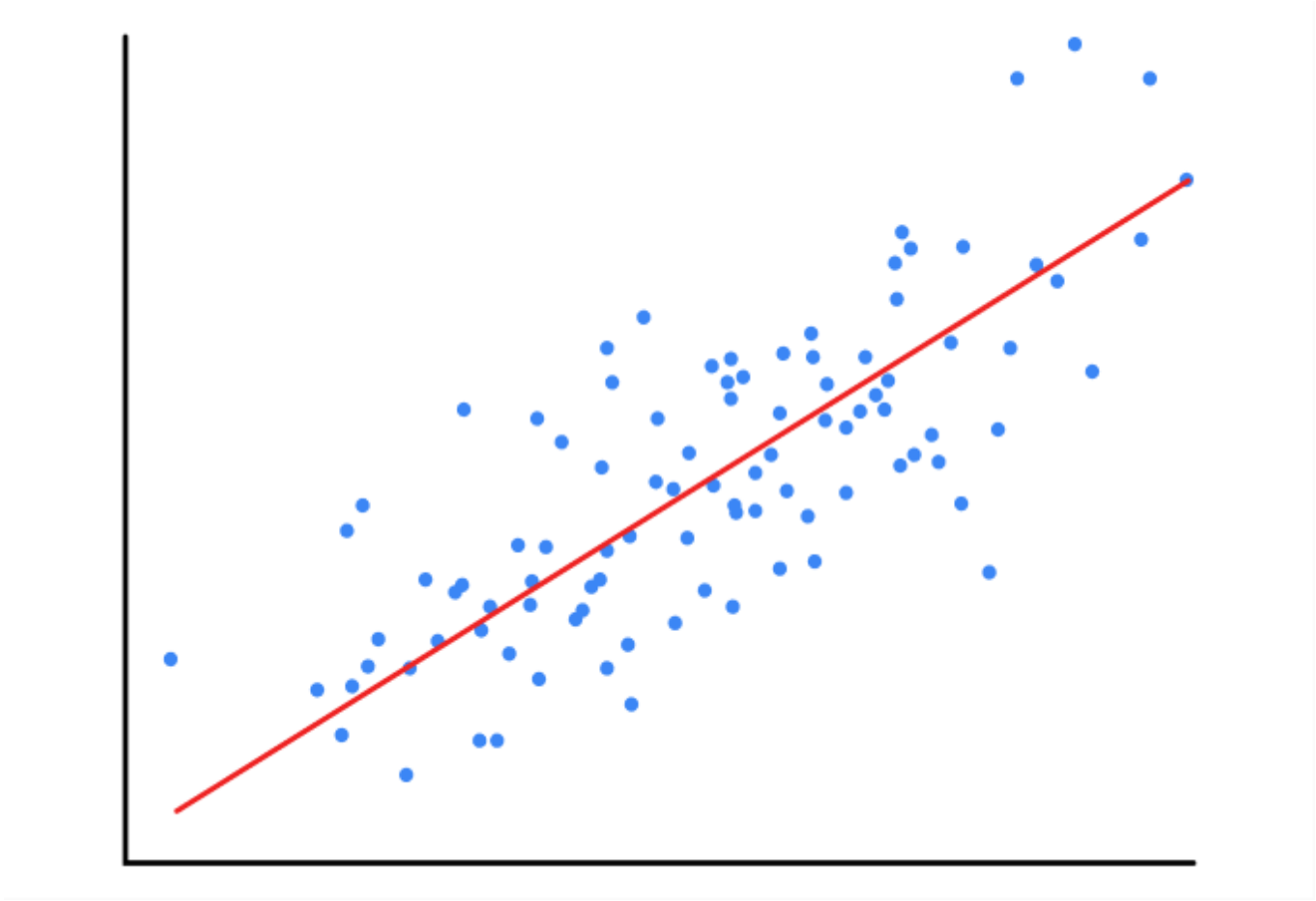


ΕΙΚΟΝΑ 4-ΣΥΣΤΑΔΟΠΟΙΗΣΗ ΣΕ 3 CLUSTERS

2.3.2 ΠΑΛΙΝΔΡΟΜΗΣΗ

Η πιο διαδεδομένη και γνωστή μέθοδος που χρησιμοποιείται συχνότερα για τις αριθμητικές προβλέψεις είναι η παλινδρόμηση μια στατιστική μέθοδος που αναπτύχθηκε από τον μαθηματικό Sir Frances Galton (1822-1911). Με την ανάλυση παλινδρόμησης επιτυγχάνουμε να διακρίνουμε την σχέση που υπάρχει ανάμεσα σε μια ή περισσότερες ανεξάρτητες μεταβλητές πρόβλεψης και σε εξαρτημένες μεταβλητές απόκρισης. Στο πλαίσιο της εξόρυξης δεδομένων οι μεταβλητές πρόβλεψης αποτελούν τα χαρακτηριστικά του ενδιαφέροντος που περιγράφουν την πλειάδα. Σε γενικές γραμμές οι τιμές των μεταβλητών πρόβλεψης είναι γνωστές. Οι μεταβλητές απόκρισης (έξοδος) είναι αυτές που θέλουμε να προβλέψουμε.Η ανάλυση παλινδρόμησης είναι μία καλή επιλογή και όταν όλες οι μεταβλητές πρόβλεψης είναι συνεχείς. Πολλά προβλήματα μπορούν να λυθούν με γραμμική παλινδρόμηση και ακόμα περισσότερο μπορεί να αντιμετωπίσουν προβλήματα υιοθετώντας μετασχηματισμούς μεταβλητών έτσι ώστε ένα μη γραμμικό πρόβλημα να μετασχηματιστεί σε γραμμικό. Πολλά

πακέτα λογισμικού υπάρχουν που μπορούν να λύσουν το πρόβλημα της παλινδρόμησης όπως είναι π.χ. το SPSS,SPLUS



ΕΙΚΟΝΑ 5- ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

2.3.3 ΑΝΑΛΥΣΗ ΣΥΣΧΕΤΙΣΗΣ

Η ανάλυση συσχέτισης (association analysis) έχει σαν βασικό της στόχο την ανακάλυψη κρυμμένων συσχετίσεων μεταξύ των χαρακτηριστικών μιας βάσης δεδομένων. Με άλλα λόγια, η παραπάνω ανάλυση ψάχνει να βρει κανόνες για την ποσοτικοποίηση των σχέσεων μεταξύ δύο ή περισσοτέρων χαρακτηριστικών μιας βάσης δεδομένων. Οι κανόνες αυτοί ονομάζονται κανόνες συσχέτισης (association rules), και έχουν την μορφή « If A then B ».Οι κανόνες συσχέτισης χρησιμοποιούνται για τον υπολογισμό της πιθανότητας να συμβεί το B, με δεδομένο το ότι συνέβη το A. Η επιλογή ενός κανόνα συσχέτισης και η αποτίμηση του ως ενδιαφέροντα εξαρτάται από τις τιμές των μμεγεθών support (συχνότητα εμφάνισης του itemset AUB στην αρχική συλλογή) και confidence (την υπο-συνθήκη προβλεψιμότητα του B με δεδομένο το A). Ο πλέον δημοφιλής αλγόριθμος για την ανακάλυψη κανόνων συσχέτισης είναι ο Apriori .Αξίζει να σημειωθεί ότι η ανάλυση συσχέτισης είναι γνωστή στον επιχειρηματικό κόσμο σαν ανάλυση συνάφειας (affinity analysis) με πολλές εφαρμογές.

2.3.4 ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ (CLASSIFICATION)

Η διαδικασία της κατηγοριοποίησης, ή αλλιώς ταξινόμησης (classification) περιλαμβάνει την οργάνωση ενός συνόλου αντικειμένων (objects) που περιγράφονται από ένα σύνολο χαρακτηριστικών (attributes), σε μια σειρά από προκαθορισμένες κλάσεις (classes), χρησιμοποιώντας μεθόδους μάθησης με επίβλεψη (supervised learning methods). Οι τεχνικές της ταξινόμησης ή αλλιώς κατηγοριοποίησης χρησιμοποιούν κατά κανόνα ένα σύνολο εκπαίδευσης (training set), όπου όλα τα αντικείμενα είναι ήδη συνδεδεμένα με γνωστές κλάσεις. Ο αλγόριθμος ταξινόμησης μαθαίνει από αυτό το σύνολο, χρησιμοποιώντας την μάθηση αυτή για την κατασκευή ενός μοντέλου και το μοντέλο αυτό στην συνέχεια ταξινομεί νέα αντικείμενα στις κατάλληλες κλάσεις. Άρα μπορούμε να πούμε ότι η κατηγοριοποίηση μαθαίνει σε μία λειτουργία να χαρτογραφεί ή πιο απλά να ταξινομεί ένα στοιχείο δεδομένων σε μία από τις διάφορες προκαθορισμένες κατηγορίες. Η κατηγοριοποίηση πρόκειται ίσως για την πιο δημοφιλή τεχνική με πλήθος εφαρμογών στην αναγνώριση προτύπων και εικόνας σε διάφορους κλάδους.

Στην πράξη μια διαδικασία κατηγοριοποίησης μπορεί να οριστεί ως η εκτέλεση δύο συγκεκριμένων βημάτων:

1. Δημιουργία μοντέλου βασιζόμενου σε δεδομένα εκπαίδευσης
2. Εφαρμογή του μοντέλου στο σύνολο των δεδομένων

Αν και βάσει του επιστημονικού ορισμού το δεύτερο από τα παραπάνω βήματα είναι αυτό της κατηγοριοποίησης, το πρώτο είναι το βήμα που απαιτεί και την μεγαλύτερη προσπάθεια. Η εργασία της κατηγοριοποίησης χαρακτηρίζεται από έναν καλά καθορισμένο ορισμό των κατηγοριών και το σύνολο που χρησιμοποιείται για την εκπαίδευση του μοντέλου αποτελείται από προ κατηγοριοποιημένα παραδείγματα.

Η αξιολόγηση της απόδοσης ενός μοντέλου ταξινόμησης βασίζεται στον αριθμό των εγγραφών του test set που προβλέφθηκαν σωστά ή λάθος από τον ταξινομητή. Για να είναι ευκολότερη η σύγκριση των αποδόσεων διαφορετικών μοντέλων χρησιμοποιούνται δύο δείκτες επίδοσης, η ακρίβεια (accuracy) και η αποτίμηση του σφάλματος (error rate).Έτσι τελικά ο ταξινομητής με τη μεγαλύτερη ακρίβεια και το μικρότερη αποτίμηση σφάλματος είναι ορθότερος και πιο αποτελεσματικός, δηλαδή μπορεί και κάνει καλύτερες προβλέψεις. Αυτοί οι δείκτες θα μας βοηθήσουν και με τον σκοπό της παρούσας διπλωματικής ,ο οποίος είναι να προβλέψουμε με όσο το δυνατόν μεγαλύτερη ακρίβεια το αν μια επιχείρηση ή ένας ιδιώτης πελάτης θα μπορέσει να εξυπηρετήσει το δάνειο που θα πάρει από την τράπεζα.

Η εργασία της κατηγοριοποίησης χαρακτηρίζεται από έναν καλά καθορισμένο ορισμό των κατηγοριών και το σύνολο που χρησιμοποιείται για την εκπαίδευση

του μοντέλου αποτελείται από προ κατηγοριοποιημένα παραδείγματα. Η βασική εργασία είναι να δημιουργηθεί ένα μοντέλο το οποίο θα μπορούσε να εφαρμοστεί για να οργανώσει δεδομένα που δεν έχουν ακόμα κατηγοριοποιηθεί. Στις περισσότερες περιπτώσεις, υπάρχει ένα περιορισμένος αριθμός κατηγοριών και εμείς θα πρέπει να αναθέσουμε κάθε εγγραφή στην κατάλληλη κατηγορία. Για αυτό το σκοπό χρησιμοποιούνται κάποιες τεχνικές, τις οποίες μπορούμε να κατατάξουμε σε δύο βασικές κατηγορίες. Η πρώτη χρησιμοποιεί τα λεγόμενα δέντρα απόφασης (decision trees) ενώ η δεύτερη τα νευρωνικά δίκτυα (neural networks).

2.3.4.1 ΔΕΝΔΡΑ ΑΠΟΦΑΣΗΣ

Τα δέντρα απόφασης (decision trees) είναι μια από τις πιο σημαντικές και ευρύτατα διαδεδομένες μεθόδους για την ταξινόμηση δεδομένων. Σύμφωνα με τους Quinlan και Murphy, τα δέντρα απόφασης είναι δομές που ταξινομούν τα αντικείμενα μιας βάσης δεδομένων βάσει των τιμών των χαρακτηριστικών αυτών. Η κατασκευή του από την άλλη βασίζεται σε ένα σύνολο εκπαίδευσης, το οποίο περιλαμβάνει προ-ταξινομημένα δεδομένα.

Η ταξινόμηση ενός νέου αντικειμένου μέσω ενός δέντρου απόφασης ακολουθεί κάποια βήματα. Ξεκινώντας από την ρίζα του δέντρου (αρχικός κόμβος) και εξετάζοντας τα χαρακτηριστικά που καθορίζονται από τον κόμβο αυτό, προσδιορίζονται διαδοχικά οι εσωτερικοί κόμβοι του δέντρου που πρέπει να ακολουθηθούν, έως ότου καταλήξουμε σε ένα συγκεκριμένο φύλλο. Πολλά διαφορετικά φύλλα μπορούν να οδηγούν στην ίδια ταξινόμηση, αλλά κάθε φύλλο κάνει την ταξινόμηση αυτή για διαφορετικό λόγο. Σε κάθε εσωτερικό κόμβο, εξετάζεται αν το προς ταξινόμηση αντικείμενο ικανοποιεί τον συγκεκριμένο κόμβο. Η έκβαση της εξέτασης αυτής καθορίζει το κλαδί που θα ακολουθηθεί στην συνέχεια, καθώς και τον επόμενο κόμβο. Η κλάση στην οποία θα ταξινομηθεί το νέο αντικείμενο αντιστοιχεί σε ένα από τα φύλλα του δέντρου απόφασης, είναι δε αυτή του τελικού κόμβου.

Τα πλεονεκτήματα από τη χρήση δένδρων αποφάσεων κατηγοριοποίησης είναι πολλά και παρατίθενται παρακάτω:

1. Τα δένδρα απόφασης είναι εύκολα στη χρήση και αποτελεσματικά, με κανόνες κατανοητούς και βατούς ως προς την ερμηνεία τους.
2. Δένδρα απόφασης μπορούν να κατασκευαστούν και για τα δεδομένα με πολλά γνωρίσματα.
3. Λειτουργούν πάρα πολύ καλά σε μεγάλες βάσεις δεδομένων λόγω του γεγονότος ότι το μέγεθος του δένδρου είναι ανεξάρτητο από το μέγεθος της βάσης.
4. Η ευρωστία που επιδεικνύουν αναφορικά με το θόρυβο που ενδέχεται να παρουσιαστεί στα δεδομένα που απαρτίζουν το χώρο του προβλήματος.

5. Η ανοχή στην απουσία τιμών (missing values), σε κάποια χαρακτηριστικά του σώματος εκπαίδευσης.
6. Η χρήση ακόμα και συνεχών (μη διακριτών) χαρακτηριστικών και η προσέγγιση μη διακριτών συναρτήσεων στόχου, μέσω εξειδικευμένων τεχνικών που αναλαμβάνουν τη διακριτοποίηση τους (discretization), τη διαδικασία δηλαδή της μετατροπής συνεχών αριθμητικών χαρακτηριστικών σε κατηγορικά.
7. Η δυνατότητα μεταφοράς του παραγόμενου μοντέλου από δένδρο απόφασης σε ένα σύνολο κανόνων, προς διευκόλυνση της κατανόησής του.

Δε λείπουν ωστόσο και τα μειονεκτήματα από τη χρήση δένδρων απόφασης μερικά από τα οποία είναι:

1. Δε χειρίζονται εύκολα δεδομένα, τα γνωρίσματα των οποίων αποτελούνται από συνεχείς τιμές.
2. Υπάρχει η πιθανότητα υπερ-προσπαρμογής ενός δένδρου στα σύνολα δεδομένων εκπαίδευσης.

2.3.4.2 ΤΕΧΝΗΤΑ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ

Πέρα από τις μεθόδους ταξινόμησης που βασίζονται στα δέντρα και τους κανόνες απόφασης, τα τεχνητά νευρωνικά δίκτυα (artificial neural networks) είναι επίσης μια διαδεδομένη μέθοδος ταξινόμησης.

Συγκεκριμένα, είναι μια δομή που αποτελείται από ένα δίκτυο νευρώνων (neurons) οι οποίοι συνδέονται μεταξύ τους και αποτελούν τα δομικά στοιχεία του δικτύου. Κάθε τέτοιος κόμβος δέχεται ένα σύνολο αριθμητικών εισόδων από διαφορετικές πηγές (είτε από άλλους νευρώνες, είτε από το περιβάλλον), επιτελεί έναν υπολογισμό με βάση αυτές τις εισόδους και παράγει μία έξοδο. Η εν λόγω έξοδος είτε κατευθύνεται στο περιβάλλον, είτε τροφοδοτείται ως είσοδος σε άλλους νευρώνες του δικτύου. Η πιο διαδεδομένη κατηγορία νευρωνικών δικτύων είναι τα λεγόμενα δίκτυα πρόσθιας τροφοδότησης (feed-forward neural networks), τα οποία επιτρέπουν την κίνηση των δεδομένων μόνο προς μια κατεύθυνση, δηλαδή από μια είσοδο προς μια έξοδο και έχουμε και τα δίκτυα που σχηματίζουν κυκλικές δομές τα οποία ονομάζονται ανατροφοδοτούμενα νευρωνικά δίκτυα (recurrent neural networks).

Τα νευρωνικά δίκτυα είναι μία προσέγγιση ανάπτυξης και εκτίμησης μαθηματικών δομών. Οι μέθοδοι αυτοί είναι αποτελέσματα ακαδημαϊκών ερευνών με στόχο την μοντελοποίηση συστημάτων μάθησης. Τα νευρωνικά δίκτυα έχουν την ικανότητα να εξάγουν κάποιο συμπέρασμα από πολύπλοκα ή μη ακριβή δεδομένα και μπορούν να χρησιμοποιηθούν για να εξάγουν πρότυπα και να προσδιορίζουν τάσεις οι οποίες είναι πολύπλοκες για να προσδιοριστούν από ανθρώπους ή από άλλες υπολογιστικές τεχνικές. Ένα εκπαιδευμένο νευρωνικό δίκτυο μπορεί να αντιμετωπιστεί ως ένας ειδικός για την κατηγορία

της πληροφορίας που του δόθηκε να αναλύσει. Έτσι μπορεί να χρησιμοποιηθεί για να κάνει κάποιες προβλέψεις, όταν προκύψουν κάποιες νέες περιπτώσεις. Τα νευρωνικά δίκτυα χρησιμοποιούν ένα σύνολο από στοιχεία επεξεργασίας (κόμβους) ανάλογους με τους νευρώνες στο ανθρώπινο μυαλό. Τα στοιχεία αυτά διασυνδέονται μεταξύ τους σε ένα δίκτυο το οποίο μπορεί να αναγνωρίζει πρότυπα μέσα σε ένα σύνολο δεδομένων μόλις αυτά παρουσιαστούν μέσα στα δεδομένα, δηλαδή το δίκτυο μπορεί να μαθαίνει από την εμπειρία όπως ακριβώς κάνουν και οι άνθρωποι. Αυτό διακρίνει τα νευρωνικά δίκτυα από τα παραδοσιακά προγράμματα υπολογιστών, τα οποία απλά ακολουθούν οδηγίες σύμφωνα με μία καλά ορισμένη σειρά.

Το κύριο χαρακτηριστικό των νευρωνικών δικτύων είναι η εγγενής ικανότητα μάθησης. Ως μάθηση μπορεί να οριστεί η σταδιακή βελτίωση της ικανότητας του δικτύου να επιλύει κάποιο πρόβλημα όπως για παράδειγμα η σταδιακή προσέγγιση μίας συνάρτησης. Η μάθηση επιτυγχάνεται μέσω της εκπαίδευσης μιας επαναληπτικής διαδικασίας σταδιακής προσαρμογής των παραμέτρων του δικτύου, σε τιμές κατάλληλες ώστε να επιλύεται με επαρκή επιτυχία το προς εξέταση πρόβλημα. Αφού ένα δίκτυο εκπαιδευτεί, οι παράμετροί του συνήθως παγώνουν στις κατάλληλες τιμές και έπειτα είναι σε λειτουργική κατάσταση. Το ζητούμενο είναι το λειτουργικό δίκτυο να χαρακτηρίζεται από μία ικανότητα γενίκευσης. Αυτό σημαίνει ότι πρέπει να δίνει ορθές εξόδους για εισόδους καινοφανείς και διαφορετικές από αυτές με τις οποίες εκπαιδεύτηκε.

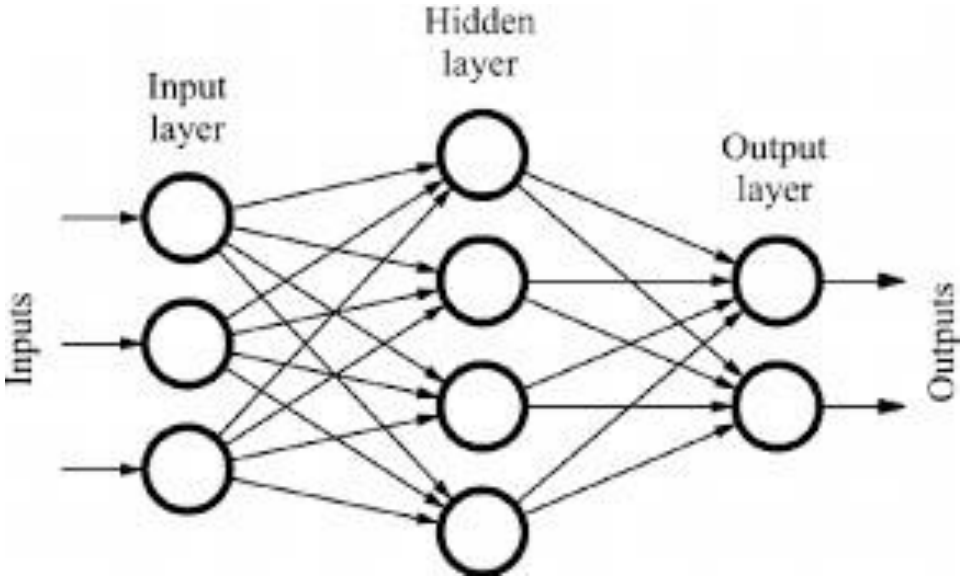
Οι νευρώνες ενός δικτύου χωρίζονται σε τρείς βασικές κατηγορίες:

1) Τους νευρώνες εισόδου (input neurons): οι οποίοι δέχονται τις πληροφορίες που θα υποστούν επεξεργασία

2) Τους νευρώνες εξόδου (output neurons): στους οποίους καταλήγουν τα αποτελέσματα της παραπάνω επεξεργασίας

3) Τους ενδιάμεσους νευρώνες: οι οποίοι βρίσκονται μεταξύ των νευρώνων εισόδου και εξόδου. Οι τελευταίοι εναλλακτικά ονομάζονται και κρυφοί νευρώνες (hidden neurons).

Ουσιαστικά, οι νευρώνες σε ένα δίκτυο είναι αφενός ένα σύνολο εισερχόμενων τιμών και των αντίστοιχων βαρών τους και αφετέρου μια συνάρτηση που αθροίζει τα παραπάνω βάρη, αντιστοιχώντας τα αποτελέσματα σε ένα νευρώνα εξόδου.



Καταλήγοντας αξίζει να σημειώσουμε ότι εν πολλοίς η εκπαίδευση ενός νευρωνικού δικτύου βασίζεται στον υπολογισμό των τιμών των βαρών που προ αναφέρθηκαν. Ο πιο γνωστός αλγόριθμος, μεταξύ άλλων, στον οποίο βασίζεται ο παραπάνω υπολογισμός, είναι ο αλγόριθμος ανάστροφης μετάδοσης (back propagation algorithm).Άλλες προσεγγίσεις που χρησιμοποιούνται για την εκπαίδευση των νευρωνικών δικτύων, με κύριο στόχο την βελτίωση των χρονικών τους επιδόσεων, είναι αυτές των Weigend et al και Yam & Chow . Επιπλέον, για την εκπαίδευση των νευρωνικών δικτύων μπορούν να χρησιμοποιηθούν τόσο γενετικοί αλγόριθμοι όσο και στατιστικές μέθοδοι.

ΚΕΦΑΛΑΙΟ 3

Ο ΤΟΜΕΑΣ ΤΗΣ ΟΙΚΟΝΟΜΙΑΣ ΣΗΜΕΡΑ

ΕΙΣΑΓΩΓΗ

Ένας από τους τομείς που εφαρμόζεται κατά κόρον η εξόρυξη δεδομένων είναι αυτός της οικονομίας. Τα οικονομικά δεδομένα συλλέγονται κυρίως από τράπεζες, σουπερμάρκετ και από άλλους οικονομικούς οργανισμούς. Τα δεδομένα αυτά συνήθως είναι αξιόπιστα, ολοκληρωμένα, έχουν υψηλή ποιότητα και απαιτούν συστηματική μέθοδο για την ανάλυση τους. Η συνεισφορά της εξόρυξης δεδομένων στην επιστήμη της οικονομίας συναντάται στην συλλογή, κατανόηση και βελτίωση των δεδομένων, στην δημιουργία και εκτίμηση ενός μοντέλου και στην ανάπτυξη αυτού. Η σωστή ανάλυση των οικονομικών δεδομένων διευκολύνει στο να παρθούν καλύτερες αποφάσεις ενεργώντας σύμφωνα με την ανάλυση της αγοράς. Τα εργαλεία και οι τεχνικές της εξόρυξης δεδομένων βοηθούν στο να αναλύσουμε τα οικονομικά δεδομένα και είναι τέτοια η συμβολή τους έτσι ώστε για παράδειγμα, τα οικονομικά ινστιτούτα να αναγνωρίζουν τις απάτες από παραποιημένα δεδομένα από τις διάφορες βάσεις δεδομένων και από το ιστορικό συναλλαγών που έγιναν από τους πελάτες. Οι τεχνικές οπτικοποίησης βοηθούν στην παρουσίαση δεδομένων με διαφορετικές μορφές, όπως γράφοι που βασίζονται σε συγκεκριμένα γνωρίσματα. Παραδείγματος χάρη προβάλλοντας τα δεδομένα από διάφορες οπτικές γωνίες, μία τράπεζα δύναται να διακρίνει τους πελάτες που έχουν επιχειρήσει παράνομες πράξεις και μετά μια λεπτομερής έρευνα αυτών των ύποπτων περιπτώσεων βοηθάει στην εξιχνίαση των απατών και των εγκλημάτων

3.1 ΟΙ ΚΙΝΔΥΝΟΙ ΤΟΥ ΤΡΑΠΕΖΙΚΟΥ ΣΥΣΤΗΜΑΤΟΣ

Οι τράπεζες θα πρέπει με κατάλληλες μεθόδους και διαδικασίες να εκτιμήσουν και να διαχειριστούν τους κινδύνους που μπορεί να υπάρξουν. Με αυτό τον τρόπο θα μπορούν να εξασφαλίσουν την βιωσιμότητά τους.

Οι κυριότεροι κίνδυνοι είναι οι εξής :

- Κίνδυνος Αγοράς (Market Risk): αφορά τον κίνδυνο μείωσης του επιπέδου τιμών της αγοράς στο σύνολο της ή σε κάποια από τα στοιχεία του ενεργητικού κάποιου επενδυτικού προϊόντος. Για παράδειγμα η μεταβολή αυτή μπορεί να οφείλεται στην αυξομείωση των επιτοκίων ή των τιμών των επενδυτικών τίτλων.
- Πιστωτικός Κίνδυνος (Credit Risk): είναι ο κίνδυνος που διατρέχει μια επιχείρηση ή ένας οικονομικός οργανισμός να μην εισπράξει τις απαιτήσεις του.
- Επιτοκιακός Κίνδυνος (Interest Rate Risk) : είναι ο κίνδυνος να μεταβληθεί η αξία μιας επένδυσης κάτι το οποίο οφείλεται σε μεταβολές στο επίπεδο των επιτοκίων.
- Κίνδυνος Ρευστότητας (Liquidity Risk) : οφείλεται στην αβεβαιότητα που δημιουργείται όταν κάποια επένδυση δεν μπορεί να ρευστοποιηθεί έγκαιρα.

3.2 ΠΙΣΤΩΤΙΚΕΣ ΚΑΡΤΕΣ ΣΕ ΙΔΙΩΤΕΣ

Οι πιστωτικές κάρτες χρησιμοποιούνται στην αγορά, ως ένα συμπληρωματικό μέσο, για την ολοκλήρωση ενός μεγάλου μέρους των συναλλαγών των καταναλωτών. Αλλιώς μπορούν να χαρακτηριστούν και ως πλαστικό χρήμα. Αν και στις ηλεκτρονικές δραστηριότητες μπορεί να εμφανιστούν ποικίλα προβλήματα, οι πιστωτικές κάρτες παρέχουν ασφάλεια και επιπλέον δεν απαιτούν τη μεταφορά μετρητών. Η χρήση πιστωτικής κάρτας χρησιμεύει σε αγορές αγαθών, των οποίων η εξόφληση μπορεί να πραγματοποιηθεί με δόσεις ενώ ταυτόχρονα μπορούν να καλύψουν και ανάγκες ανάληψης μετρητών. Αντιλαμβανόμαστε λοιπόν ότι οι τράπεζες, για να τις χορηγήσουν, εξετάζουν κάποια κριτήρια αξιολόγησης. Πιο συγκεκριμένα λαμβάνουν υπόψη κυρίως το ατομικό εισόδημα, όπως αυτό εμφανίζεται στο εκκαθαριστικό σημείωμα της εφορίας και ταυτόχρονα αν υπάρχουν ή μη δυσμενή στοιχεία στο σύστημα γνωστό και ως «Τειρεσίας». Μεταξύ των άλλων εξετάζουν το επάγγελμα των αιτούντων, αν είναι μόνιμοι κάτοικοι της Ελλάδας, καθώς και αν η μορφή απασχόλησης τους είναι ολική ή μερική. Σημαντικό ρόλο κατέχει και το χρονικό διάστημα που εργάζονται, δηλαδή αν οι αιτούντες εργάζονται μόνο λίγους μήνες ή αρκετά έτη. Ο υποψήφιος κάτοχος πιστωτικής κάρτας είναι αναγκαίο να συμπληρώσει την ειδική αίτηση για την χορήγηση της, η οποία αποτελεί και την σύμβαση του με την τράπεζα, στην περίπτωση που εγκριθεί το αίτημα του.

Αξιοσημείωτο είναι ότι η τράπεζα ανάλογα με τα κριτήρια της, μπορεί να χορηγήσει απλή ή χρυσή κάρτα. Καταλαβαίνουμε πως η χρυσή κάρτα απευθύνεται σε άτομα με υψηλά εισοδήματα και υψηλή πιστοληπτική ικανότητα. Παράλληλα είναι φυσιολογικό στις περιπτώσεις καλών πελατών, δηλαδή σε πελάτες που δεν παρουσιάζουν αρνητικά στοιχεία στο σύστημα «Τειρεσίας», με ικανοποιητικό δηλωθέν εισόδημα, και που εξοφλούν έγκαιρα τις δόσεις τους, οι τράπεζες να αναπροσαρμόζουν άμεσα την αύξηση των πιστωτικών ορίων των καρτών. Πάντως είναι απαραίτητο να αναφέρουμε, σύμφωνα με έρευνα, ότι ένας από τους σημαντικότερους λόγους που δεν έχει αναπτυχθεί ιδιαίτερα στην χώρα μας η χρήση των πιστωτικών καρτών οφείλεται στο τραπεζικό σύστημα, που δεν έχει τελειοποιήσει τους μηχανισμούς εξακρίβωσης της πιστοληπτικής ικανότητας των δανειζόμενων. Αυτό το κενό μπορεί να καλύψει η εξόρυξη γνώσης.

3.3 ΔΙΑΔΙΚΑΣΙΑ ΕΚΔΟΣΗΣ ΠΙΣΤΩΤΙΚΗΣ ΚΑΡΤΑΣ

Ο υποψήφιος κάτοχος πιστωτικής κάρτας ζητάει από την εκδότρια τράπεζα, που έχει λογαριασμό την έκδοση μιας κάρτας. Αυτό συμβαίνει διότι η κάρτα που εκδίδεται από την ελληνική τράπεζα πρέπει να είναι συνδεδεμένη με κάποιον από τους παγκόσμιους οργανισμούς πιστωτικών καρτών, για να δίνεται ταυτόχρονα

η δυνατότητα στον κάτοχο της να μπορεί να την χρησιμοποιεί παντού σε παγκόσμια κλίμακα Η κάρτα εκδίδεται στο όνομα του πελάτη και οι συναλλαγές χρεώνονται σε έναν ανοιχτό λογαριασμό, στον οποίο έχει καθοριστεί το πιστωτικό όριο. Αντιλαμβανόμαστε ότι το πιστωτικό όριο εξαρτάται από την οικονομική επιφάνεια του πελάτη, καθώς και από την πολιτική της τράπεζας που εκδίδει την πιστωτική κάρτα. Ο τρόπος που θα την χρησιμοποιεί ο κάτοχος της μπορεί να την μετατρέψει σε χρήσιμο εργαλείο. Άρα ο χρήστης της δεν πρέπει να υπερβαίνει το πιστωτικό όριο που του δίνει η τράπεζα και συγχρόνως θα πρέπει να εξοφλεί ολόκληρο το οφειλόμενο ποσό έτσι, ώστε να μην χρεώνεται με υπέρμετρους τόκους καθυστέρησης. Υπάρχουν πολλά εμπορικά καταστήματα που προσφέρουν πολλές διευκολύνσεις στους κατόχους των πιστωτικών καρτών, όπως τη δυνατότητα εξόφλησης σε αρκετές άτοκες δόσεις. Επομένως όποιος προτιμά αυτόν τον τρόπο για να κάνει τις αγορές του εξασφαλίζει μια άτοκη πίστωση. Εκτός των άλλων είναι απαραίτητο να γνωρίζουμε ότι οι πιστωτικές κάρτες έχουν όλες τις συναλλακτικές δυνατότητες, ενώ οι αναλήψεις ορίζονται έως κάποιο συγκεκριμένο όριο πάντα μετά από συμφωνία με την τράπεζα. Συνήθως υπάρχει και άτοκη περίοδος χάριτος για αναλήψεις, μέχρι 60 ημέρες. Βεβαίως ορισμένες τράπεζες δίνουν τη δυνατότητα στον πελάτη τους για ανάληψη μετρητών προκαταβολικά, σε περιπτώσεις εκτάκτου ανάγκης, αλλά και την έκδοση κάρτας και σε μέλη της οικογένειας εντός του εγκεκριμένου ποσού. Επίσης διαφοροποιούνται από τα προσωπικά καταναλωτικά δάνεια ως προς τα επιτόκια δανεισμού, ενώ συνοδεύονται με διάφορες παροχές όπως ειδικά προγράμματα εκπτώσεων, ταξιδιωτική ασφάλιση και δώρα. Στην περίπτωση που ο χρήστης της πιστωτικής κάρτας δεν μπορεί να εργαστεί λόγω ατυχήματος είτε ασθένειας τότε διευκολύνεται στην αποπληρωμή της ελάχιστης μηνιαίας καταβολής του. Ο λογαριασμός του κατόχου της πιστώνεται με το 10% της συνολικής οφειλής κάθε μήνα ενώ παράλληλα η πίστωση της ελάχιστης μηνιαίας καταβολής συνεχίζεται για το χρονικό διάστημα που αυτός δεν μπορεί να εργαστεί, με μέγιστο χρονικό όριο τους 10 μήνες

3.4 ΧΡΗΜΑΤΟΔΟΤΗΣΗ ΣΕ ΕΠΙΧΕΙΡΗΣΕΙΣ

Υπολογίζεται ότι τουλάχιστον μια στις τρεις μικρομεσαίες επιχειρήσεις καταφεύγει σήμερα στο τραπεζικό σύστημα για χρηματοδότηση, αν και έχουν αναπτυχθεί σύγχρονοι χρηματοδοτικοί μηχανισμοί για νέες επιχειρήσεις τα τελευταία χρόνια.

Οι τράπεζες δίνουν δύο ειδών δάνεια προς τις επιχειρήσεις:

- Τα μακροπρόθεσμα: Στα μακροπρόθεσμα περιλαμβάνονται τα δάνεια επαγγελματικού εξοπλισμού, αυτά δηλαδή που χρειάζονται για να καλυφθούν οι ανάγκες εξοπλισμού μιας επιχείρησης και τα δάνεια εγκατάστασης, τα οποία καλύπτουν την ανάγκη απόκτησης επαγγελματικής στέγης.
- Τα κεφαλαίου κίνησης: Τα συγκεκριμένα δάνεια στοχεύουν στη βελτίωση της ρευστότητας της επιχείρησης και είναι μικρής διάρκειας.

Εκτός των άλλων πρέπει να τονίσουμε ένα σημαντικό χαρακτηριστικό των τραπεζικών δανείων το οποίο είναι η λογική των εμπραγμάτων ασφαλειών. Αυτό σημαίνει ότι οι τράπεζες δανείζουν μόνο σε όσους έχουν κάποιο περιουσιακό στοιχείο, το οποίο θα χρησιμοποιηθεί ως εγγύηση για την εξόφληση ολόκληρου ή μέρους του δανείου στην περίπτωση που ο δανειολήπτης δεν ανταποκριθεί στις υποχρεώσεις του. Σε περίπτωση λοιπόν που ένας επιχειρηματίας θέλει να ξεκινήσει μια προσπάθεια δίχως να έχει προσωπική περιουσία είναι αναγκαία η ύπαρξη ενός τρίτου προσώπου που να εγγυηθεί την δική του περιουσία. Συνειδητοποιούμε ότι οι διαφορετικές κατηγορίες πίστωσης είναι αρκετές και δύσκολα κατηγοριοποιούνται σε μικρότερες ομάδες δεδομένου της πολυπλοκότητας τους. Η διαφορετικότητα των τραπεζικών προϊόντων μπορεί να οφείλεται είτε στο χρόνο διάρκειας και στην σταθερότητα των δόσεων, είτε στο επιτόκιο και στα ενέχυρα ανταλλάγματα, είτε στον σκοπό του δανείου. Ωστόσο είναι απαραίτητο να αναφέρουμε ότι ο ενδιαφερόμενος δανειολήπτης είναι ορθό να εξετάσει τις δυνατότητες χρηματοδότησης του δανείου του με σταθερό ή κυμαινόμενο επιτόκιο, ανάλογα με την εξέλιξη του πληθωρισμού, των επιτοκίων και των υπόλοιπων οικονομικών μεγεθών. Πάντως το επιτόκιο αποπληρωμής του μακροπρόθεσμου δανείου είναι στις περισσότερες περιπτώσεις χαμηλότερο, από το αντίστοιχο επιτόκιο αποπληρωμής του δανείου κεφαλαίου κίνησης, εξαιτίας της μεγαλύτερης διάρκειας αποπληρωμής του.

ΚΕΦΑΛΑΙΟ 4

ΕΦΑΡΜΟΓΕΣ ΤΗΣ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ ΚΑΙ ΔΕΔΟΜΕΝΩΝ ΣΤΗΝ ΟΙΚΟΝΟΜΙΑ

Σε αυτό το σημείο θα επικεντρωθούμε στο κατά πόσο οι τεχνικές εξόρυξης δεδομένων μπορούν να εφαρμοστούν πάνω σε οικονομικής φύσης προβλήματα και ζητήματα. Θα παρουσιάσουμε στοιχεία και πληροφορίες σχετικά με το πώς δημιουργείται ένα μοντέλο πρόβλεψης και εκτίμησης καθώς και τα στάδια προεργασίας που απαιτούνται.

4.1 ΟΡΙΣΜΟΣ ΠΡΟΒΛΗΜΑΤΟΣ ΚΑΙ ΠΙΣΤΩΤΙΚΟΣ ΚΙΝΔΥΝΟΣ

Αρχικά πρέπει να αναλογιστούμε πως μπορούμε να κατασκευάσουμε διάφορα υπολογιστικά μοντέλα τα οποία θα παίρνουν σαν είσοδο οικονομικά δεδομένα χρησιμοποιώντας τεχνικές εξόρυξης δεδομένων και θα επιστρέφουν σαν έξοδο μία εκτίμηση πρόβλεψης για το ανάλογο μέγεθος που εξετάζουμε. Η τεχνική εξόρυξης δεδομένων όπως προαναφέραμε απορρέει από τα πεδία της μηχανικής μάθησης και τεχνολογιών που αφορούν ανάλυση βάσεων δεδομένων . Ο σκοπός της μηχανικής μάθησης είναι να κατασκευαστούν υπολογιστικά προγράμματα τα οποία αυτόματα θα βελτιώνονται όσο αποκτούν πείρα και να παράγουν χρήσιμες εκτιμήσεις. Γενικά οι τεχνικές εξόρυξης δεδομένων όπως είναι τα δέντρα αποφάσεων και τα νευρωνικά δίκτυα που αναφέραμε στο προηγούμενο κεφάλαιο, μπορούν να αποδειχθούν ιδιαίτερα χρήσιμες στον κάθε ενδιαφερόμενο οικονομικό αναλυτή.

Το σύστημα εκτίμησης πιστωτικού κινδύνου συμβάλλει στην απόφαση που θα λάβει ένα πιστωτικό ίδρυμα για το αν θα δανειοδοτήσει ή όχι μια επιχείρηση ή έναν πελάτη, υπολογίζοντας τον κίνδυνο που θα αναλάβει σε περίπτωση που προχωρήσει με την δανειοδότηση. Η πιστοληπτική ικανότητα δηλαδή, η ικανότητα να μπορούν να ανταπεξέλθουν στις δανειακές υποχρεώσεις τους προκύπτει από την έρευνα και αξιολόγηση ποιοτικών και ποσοτικών τους στοιχείων και συνδέεται άμεσα με τον πιστωτικό κίνδυνο. Η αξιολόγηση των παραπάνω στοιχείων γίνεται σήμερα στο πλαίσιο των μοντέλων που προκύπτουν μέσω των συστημάτων εκτίμησης πιστωτικού κινδύνου.

4.2 ΔΗΜΙΟΥΡΓΙΑ ΜΟΝΤΕΛΟΥ ΠΡΟΒΛΕΨΗΣ

Είναι αναγκαίο να αναφέρουμε ότι η αξιολόγηση των αλγορίθμων θα γίνει με το ελεύθερο εργαλείο ανοικτού κώδικα WEKA (Waikato Environment for Knowledge Analysis). Το Weka είναι μια συλλογή εργαλείων και τεχνικών μάθησης για εφαρμογές εξόρυξης γνώσης, και όχι μόνο, το οποίο έχει αναπτυχθεί σε γλώσσα Java και ο κώδικας είναι ανοικτός στο κοινό. Συμπεριλαμβάνει δε υλοποιήσεις από αρκετούς γνωστούς αλγορίθμους εξόρυξης γνώσης, και ταυτόχρονα οι ενσωματωμένες επιλογές του μας βοηθούν στη μέτρηση της ακρίβειας αυτών. Το γραφικό περιβάλλον του διευκολύνει την πρόσβαση σε αυτές τις δυνατότητες. Εκτός των άλλων το Weka προσφέρει και προ επεξεργασία των δεδομένων, κατηγοριοποίηση, οπτικοποίηση και επιλογή χαρακτηριστικών. Οι τεχνικές αυτού

του εργαλείου βασίζονται στην προϋπόθεση ότι τα δεδομένα βρίσκονται σε κάποιο αρχείο, όπου η κάθε εγγραφή προσδιορίζεται από συγκεκριμένο αριθμό γνωρισμάτων, δηλαδή ονομαστικές μμεταβλητές, αριθμητικές μμεταβλητές, ενώ συγχρόνως μπορεί να έχει τη δυνατότητα πρόσβασης και σε SQL databases.

Στην συνέχεια θα κατασκευάσουμε ένα πρωτότυπο λογισμικό εργαλείο, το οποίο θα υλοποιεί τον αλγόριθμο που δίνει τα καλυτέρα αποτελέσματα στα πειράματα μας και θα κατηγοριοποιεί τους υποψήφιους χρήστες πιστωτικών καρτών σε << Καλούς>> ή << Κακούς>>. Στην συνέχεια θα κατασκευάσουμε ένα πρότυπο εργαλείο για την έγκριση δανειοδότησης.

4.3 ΠΡΟ-ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

Προτού τα διάφορα δεδομένα δοθούν σαν είσοδο σε έναν αλγόριθμο, πρέπει να συλλεχθούν, να ερευνηθούν και να καθαριστούν. Ακόμα και το καλύτερο σύστημα πρόβλεψης θα αποτύχει στην σωστή εξαγωγή συμπερασμάτων αν τα δεδομένα που πραγματεύεται είναι κακής ποιότητας. Επίσης είναι σκόπιμο να εξεταστούν ποια δεδομένα κρίνονται σαν καταλληλότερα για αξιοποίηση ώστε να έχουμε το καλύτερο δυνατό αποτέλεσμα. Πολλές φορές κάποια τιμή απουσιάζει από την βάση δεδομένων που διαθέτουμε για εκμετάλλευση. Είναι προτιμότερο να αντικατασταθεί με μία άλλη τιμή παρά να διαγραφεί. Σε περίπτωση όπου έχουμε μεγάλο αριθμό τέτοιων εγγραφών τότε είναι καλό να τις αντικαταστήσουμε με εκείνη την τιμή που θα επηρεάσει λιγότερο την εγκυρότητα της πρόβλεψής μας. Επίσης μπορούν να χρησιμοποιηθούν κάποιοι δείκτες (indicators) των οποίων η χρησιμότητα είναι πολύ μεγάλη καθώς μειώνουν τον θόρυβο (από την στιγμή που εκφράζουν έναν μέσο όρο) και παρέχουν όψεις από δεδομένα τα οποία είναι κατάλληλα για επεξεργασία.

Συγκεντρωτικά μπορούμε να αναφέρουμε ότι οι δυσκολίες που προέρχονται στην προ επεξεργασία των δεδομένων είναι :

- Τα ελλιπή δεδομένα: δηλαδή οι τιμές που λείπουν από γνωρίσματα του dataset ακόμη και έλλειψη ολόκληρων γνωρισμάτων που είναι σημαντικά για την έρευνα. Αιτίες των ελλιπών τιμών μπορεί να είναι πολλά πεδία τα οποία δεν συμπληρώνονται ως μη σημαντικά ή η διαγραφή των παλαιότερων δεδομένων για την εξοικονόμηση χώρου.

- Τα δεδομένα που περιέχουν θόρυβο: δηλαδή να υπάρχουν λάθη και ακραίες τιμές στα dataset και γενικότερα αν υπάρχουν διακυμάνσεις στις τιμές διαφόρων μεταβλητών. Αιτίες θορύβου μπορεί να είναι τα τυχαία λάθη, προβλήματα στη μετάδοση των δεδομένων ακόμη και οι περιορισμοί της τεχνολογίας.

- Οι ασυνέπειες στα δεδομένα: δηλαδή να υπάρχουν διαφορές σε κωδικούς που αντιπροσωπεύουν ονόματα. Αιτίες αυτού του προβλήματος μπορεί να είναι η έλλειψη μοναδικού αναγνωριστικού ή τα χαρακτηριστικά που προκύπτουν από τις τιμές των άλλων.

Για αυτό και οι ερευνητές προβαίνουν σε μία σειρά από διεργασίες οι οποίες θα εξασφαλίσουν την ποιότητα των δεδομένων που θα χρησιμοποιηθούν στην έρευνα. Τέτοιες διεργασίες είναι οι εξής:

- Ο καθαρισμός των δεδομένων: δηλαδή η συμπλήρωση τιμών που λείπουν, η διαχείριση του θορύβου που είναι πιθανό να δημιουργείται και η επίλυση των ασυνεπειών

- Η ολοκλήρωση των δεδομένων: δηλαδή η ολοκλήρωση των βάσεων δεδομένων ή των αρχείων έτσι ώστε να επιλυθεί ο πλεονασμός των δεδομένων.

- Η μείωση των δεδομένων : δηλαδή η μείωση του συνόλου των δεδομένων με τρόπο έτσι ώστε να παραχθούν αποδοτικότερα αποτελέσματα.

- Η διακριτοποίηση των δεδομένων (discretization): δηλαδή η μείωση μέρους των αριθμητικών δεδομένων με απόδοση ιδιαίτερων τιμών

Όλες αυτές οι διεργασίες που πρέπει να πραγματοποιηθούν για τη βελτίωση του συνόλου των δεδομένων αποτελούν το στάδιο της προ-επεξεργασίας των δεδομένων στην εξόρυξη γνώσης

4.4 ΕΠΙΛΟΓΗ ΚΑΤΑΛΛΗΛΩΝ ΑΛΓΟΡΙΘΜΩΝ ΚΑΙ ΕΠΙΛΟΓΗ ΜΟΝΤΕΛΟΥ

Όσον αφορά τις μεθόδους και τους αλγορίθμους που χρησιμοποιούνται για την πρόβλεψη χρηματοοικονομικών μεγεθών υπάρχει πληθώρα τέτοιων τεχνικών οι σημαντικότερες εκ των οποίων είναι οι γραμμικές μέθοδοι και τα νευρωνικά δίκτυα. Μελετώντας αυτές τις τεχνικές το συμπέρασμα που βγαίνει είναι ότι οι καταλληλότερες είναι οι υβριδικές, δηλαδή συνδυασμός πολλών τεχνικών για την αξιοποίηση των δυνατών σημείων της κάθε μίας. Μία άλλη προσέγγιση που σχετίζεται, με την μέθοδο εξόρυξης δεδομένων, είναι να υιοθετηθεί ένα μοντέλο το οποίο είναι ευέλικτο στο ότι μπορεί να συνδυάσει έναν μεγάλο αριθμό από συναρτήσεις με μεγάλη ακρίβεια. Τέτοια μοντέλα είναι μη παραμετρικά καθώς δεν χρειάζεται να υπάρχει άμεση σχέση μεταξύ των τιμών των παραμέτρων ενός μοντέλου με δεδομένα.

Τα πλεονεκτήματα αυτού του είδους των μοντέλων είναι τα παρακάτω:

1. Η δυνατότητα που παρέχουν στο να μοντελοποιούν υψηλής πολυπλοκότητας συναρτήσεις

2. Η δυνατότητα να χρησιμοποιούν έναν υψηλό αριθμό μεταβλητών στο μοντέλο, και παρ' όλα αυτά να περιέχουν και άλλα δεδομένα όπως θεμελιώδεις και τεχνικούς παράγοντες.

Σαν μειονέκτημα για αυτά τα μοντέλα μπορεί να θεωρηθεί ότι δεν μπορούν να ερμηνευθούν εύκολα

4.5 ΣΤΑΔΙΑ ΣΧΕΔΙΑΣΗΣ ΜΟΝΤΕΛΟΥ

Η σχεδίαση του μοντέλου για την εξόρυξης γνώσης από τα δεδομένα πρέπει να γίνει πολύ προσεκτικά έτσι ώστε να υπάρχουν αποδοτικά αποτελέσματα. Στη συγκεκριμένη έρευνα η σχεδίαση του μοντέλου αποτελείται από έντεκα συνολικά στάδια τα οποία αναφέρονται παρακάτω αναλυτικά:

1. Αρχικά γίνεται η επιλογή του συνόλου των δεδομένων
2. Έπειτα ξεκινά το στάδιο της προ-επεξεργασίας το οποίο περιέχει και τα στάδια 3, 4, 5 και 6.
3. Στο τρίτο στάδιο μετατρέπεται το αρχικό .txt αρχείο σε αρχείο .csv.
4. Το τέταρτο στάδιο περιλαμβάνει τη φόρτωση του αρχείου στο πρόγραμμα Weka.
5. Στο πέμπτο στάδιο μετατρέπεται το .csv αρχείο σε αρχείο .arff.
6. Κατά το έκτο στάδιο γίνονται οι τελικές διεργασίες της προεπεξεργασίας πάνω στο .arff αρχείο.
7. Στο έβδομο στάδιο, το οποίο πραγματοποιείται παράλληλα με το έκτο, γίνεται η επιλογή του κατάλληλου ταξινομητή.
8. Στο όγδοο στάδιο έχουμε την παραγωγή του τελικού αρχείου προς κατηγοριοποίηση.
9. Κατά το ένατο στάδιο εφαρμόζονται οι διάφορες τεχνικές κατηγοριοποίησης και γίνεται η εκπαίδευση των δεδομένων για την εξαγωγή της γνώσης.
10. Στο δέκατο στάδιο γίνεται η αξιολόγηση των αποτελεσμάτων
11. Στο τελευταίο στάδιο εξετάζεται η γνώση που προέκυψε

4.6 ΕΡΜΗΝΕΙΑ ΔΕΔΟΜΕΝΩΝ

Τα δεδομένα (credit.card) αντιπροσωπεύουν τα στοιχεία 45211 αιτούντων στο παρελθόν για έγκριση έκδοσης πιστωτικής κάρτας. Πρόκειται για διαθέσιμα

δεδομένα που προέρχονται από μια από τις μεγαλύτερες τράπεζες στην Ελλάδα. Ο υποψήφιος χρήστης της έχει χαρακτηριστεί ως καλός ή κακός πιστωτής. Από τις 50 μεταβλητες του αρχικου σετ, επιλεχθηκαν τελικα τα 11 πιο σημαντικα γνωρισματα για να μπορεσουμε να εχουμε καλυτερη εφαρμογη του μοντελου που κατασκευαστηκε. Τα χαρακτηριστικά των δεδομένων μας φαίνονται στον παρακάτω πίνακα.

1. Age-Ηλικία (Numeric)

2. Job-Επάγγελμα (Nominal)

- Unknown : Άγνωστο
- Unemployed : Άνεργος
- Management : Διοίκηση
- Housemaid : Οικιακά
- Entrepreneur : Επιχειρηματίας
- Student : Φοιτητής/Μαθητής
- Self-employed : Αυτοαπασχολούμενος
- Retired : Συνταξιούχος
- Services : Υπηρεσίες

3. Marital-Οικογενειακή κατάσταση (Nominal)

- Married-Παντρεμένος
- Divorced-Διαζευγμένος
- Single-Ελεύθερος

4. Education-Μορφωτικό επίπεδο (Εκπαίδευση) (Nominal)

- Unknown-Άγνωστο
- Secondary-Δευτεροβάθμια εκπαίδευση
- Primary-Πρωτοβάθμια εκπαίδευση
- Tertiary-Τριτοβάθμια εκπαίδευση

5. Previous Credit-Ύπαρξη προηγουμενης πίστωσης (Nominal)

- Yes-Ναι
- No-Όχι

6. Balance-Μέσο ετήσιο υπόλοιπο σε ευρώ (Numeric)

7. Housing-Ύπαρξη στεγαστικού δανείου (Nominal)

- Yes-Ναι
- No-Όχι

8. Loan-Ύπαρξη καταναλωτικού δανείου (Nominal)

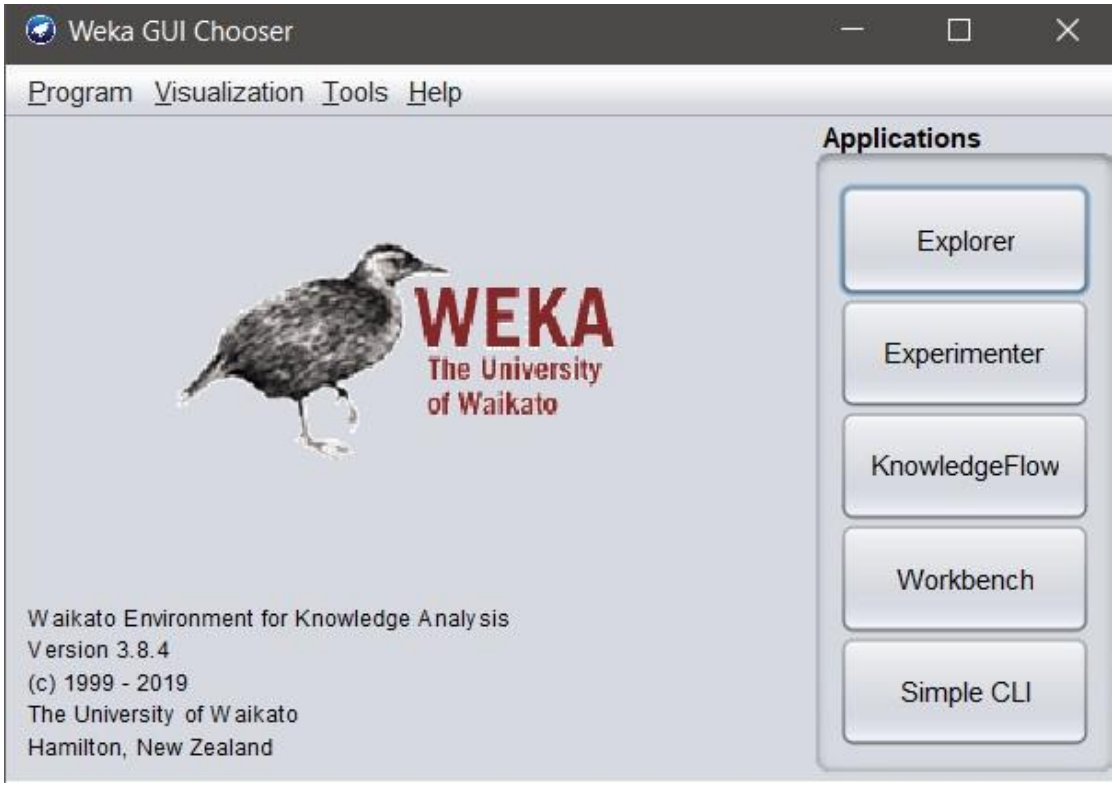
- Yes-Ναι
- No-Όχι

9.Number of Credit Card – Αριθμός Ενεργών Πιστωτικών καρτών (Numeric)

10. Περίοδος Κατοχής (σε μήνες) ενεργών πιστωτικών καρτών (Numeric)

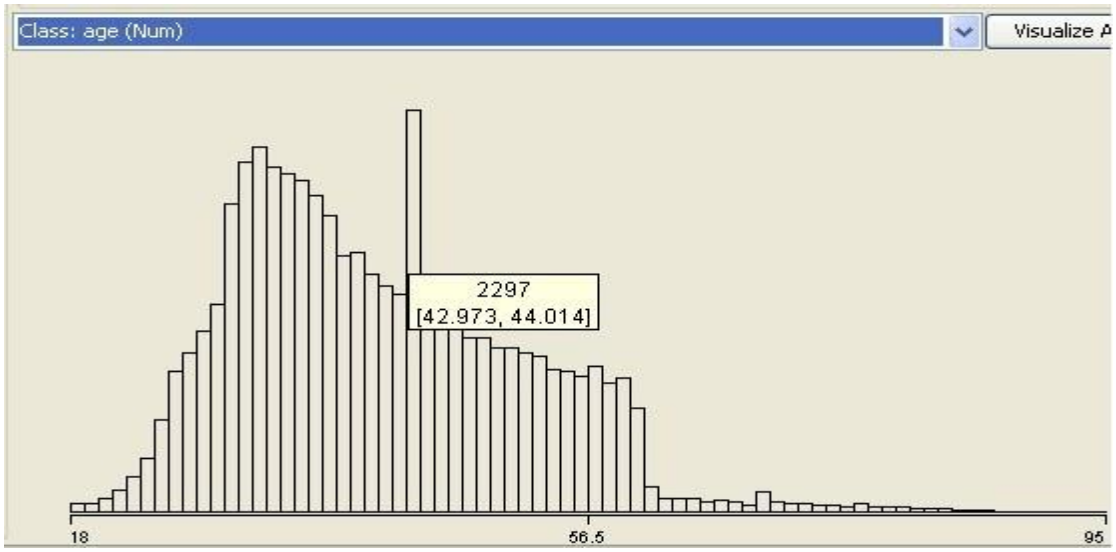
11.Credit Card Status-(Nominal)

- Yes-Καλός
- No-Κακός



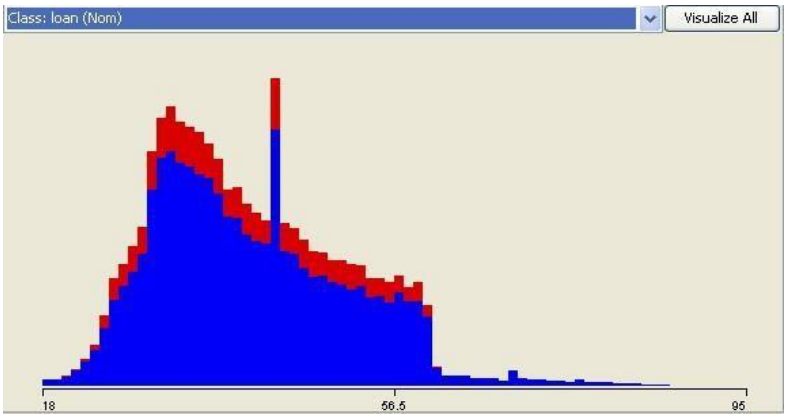
ΕΙΚΟΝΑ 6-ΤΟ ΕΡΓΑΛΕΙΟ WEKA

Η συντριπτική πλειοψηφία της έρευνας κάλυψε τα άτομα από 25 έως 60 ετών, όπως φαίνεται στον παρακάτω πίνακα. Μάλιστα, διαπιστώνει κανείς ότι συγκεκριμένα τα άτομα ηλικίας 43 ετών αποτελούσαν την πλειονότητα του δείγματος, καθώς 2.297 πελάτες είχαν αυτήν την ηλικία

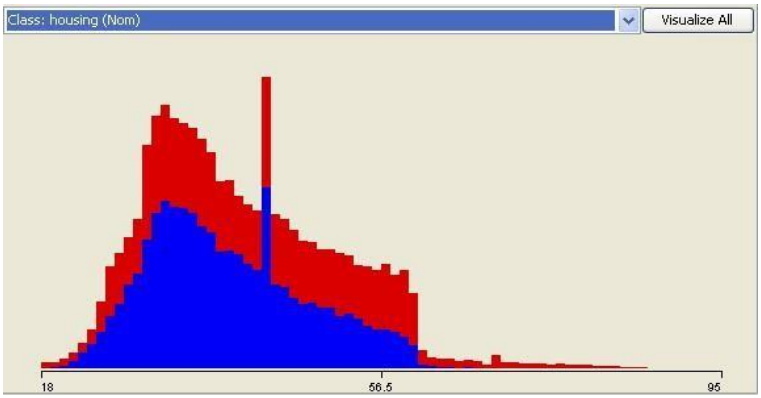


ΕΙΚΟΝΑ 7- ΣΧΕΔΙΑΓΡΑΜΜΑ ΤΙΜΩΝ ΤΗΣ ΜΕΤΑΒΛΗΤΗΣ ΗΛΙΚΙΑΣ

Επιπλέον δίνεται και η συσχέτιση της ηλικίας με τα δύο είδη δανείων (καταναλωτικά και στεγαστικά). Στην πρώτη εικόνα με μπλε χρώμα απεικονίζονται οι πελάτες που δεν έχουν λάβει καταναλωτικό δάνειο και με κόκκινο αυτοί που έχουν λάβει. Αντίθετα, στη δεύτερη εικόνα με μπλε χρώμα απεικονίζονται οι πελάτες που έχουν λάβει στεγαστικό δάνειο, ενώ με κόκκινο χρώμα αυτοί που δεν έχουν λάβει. Από αυτούς τους δύο πίνακες φαίνεται, ξεκάθαρα, ότι είναι πολλοί περισσότεροι αυτοί οι οποίοι έχουν λάβει στεγαστικά δάνεια παρά καταναλωτικά. Επιπρόσθετα, παρατηρείται πως από τους πελάτες που έχουν ηλικία, περίπου, 60 ετών ή μεγαλύτερη των 60 ετών υπάρχει, σχεδόν, μηδενική λήψη δανείων είτε καταναλωτικών είτε στεγαστικών (με ελάχιστες εξαιρέσεις στην περίπτωση των στεγαστικών δανείων).

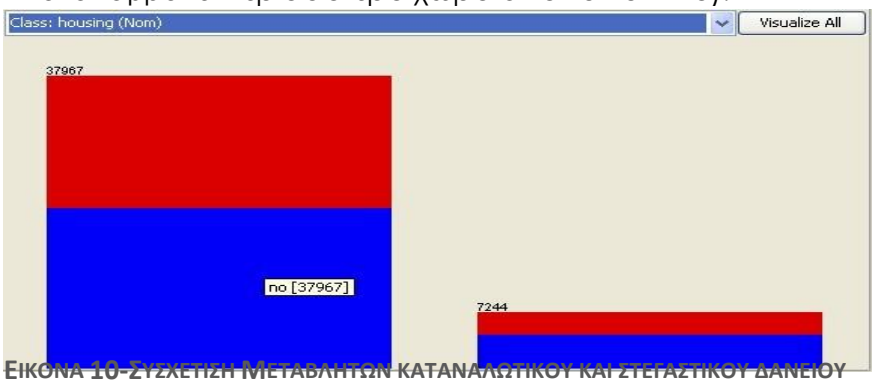


ΕΙΚΟΝΑ 8- ΣΥΣΧΕΤΙΣΗ ΜΕΤΑΒΛΗΤΩΝ ΗΛΙΚΙΑΣ ΚΑΙ ΚΑΤΑΝΑΛΩΤΙΚΟΥ ΔΑΝΕΙΟΥ

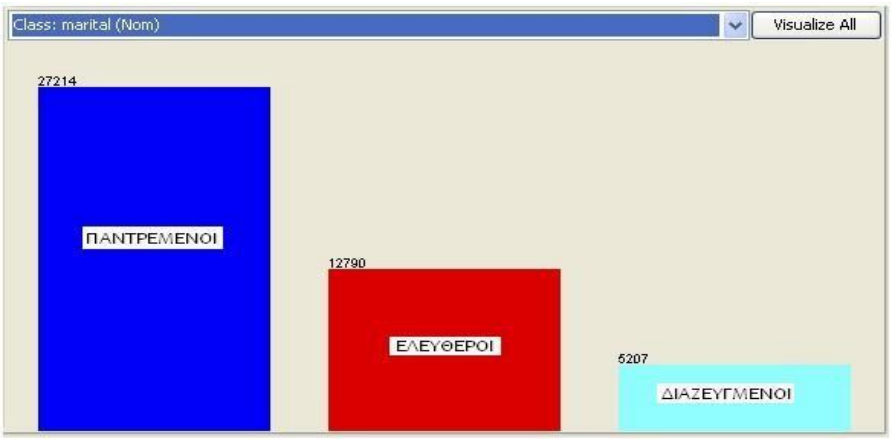


ΕΙΚΟΝΑ 9- ΣΥΣΧΕΤΙΣΗ ΜΕΤΑΒΛΗΤΩΝ ΗΛΙΚΙΑΣ ΚΑΙ ΣΤΕΓΑΣΤΙΚΟΥ ΔΑΝΕΙΟΥ

Ενδιαφέρον παρουσιάζει και ο πίνακας συσχέτισης μεταξύ καταναλωτικών και στεγαστικών δανείων που έχουν λάβει οι πελάτες. Αριστερά διακρίνονται όσοι δεν έχουν λάβει καταναλωτικό δάνειο και δεξιά όσοι έχουν λάβει. Γίνεται αντιληπτό από την εικόνα ότι, και στις δύο περιπτώσεις, παραπάνω από τους μισούς έχουν λάβει στεγαστικό δάνειο (πρακτικά δηλαδή το μπλε και στα δύο γραφήματα καταλαμβάνει περισσότερο χώρο από το κόκκινο).



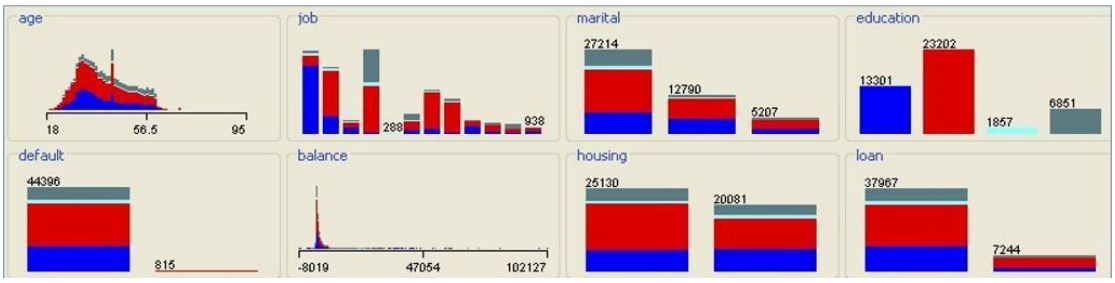
ΕΙΚΟΝΑ 10-ΣΥΣΧΕΤΙΣΗ ΜΕΤΑΒΛΗΤΩΝ ΚΑΤΑΝΑΛΩΤΙΚΟΥ ΚΑΙ ΣΤΕΓΑΣΤΙΚΟΥ ΔΑΝΕΙΟΥ



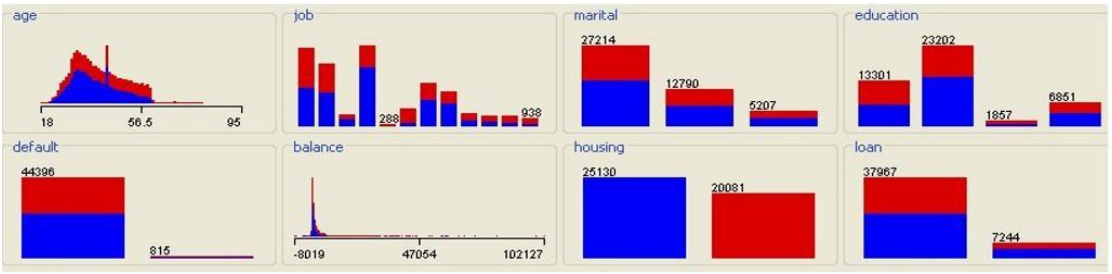
ΕΙΚΟΝΑ 11-ΔΙΑΜΟΙΡΑΣΜΟΣ ΜΕΤΑΒΛΗΤΗΣ ΟΙΚΟΓΕΝΕΙΑΚΗΣ ΚΑΤΑΣΤΑΣΗΣ ΣΤΟ ΔΕΙΓΜΑ

Διαπιστώνεται ότι οι παντρεμένοι υπερισχύουν σε αριθμό των διαζευγμένων και των ελεύθερων μαζί και μάλιστα κατά περίπου δέκα χιλιάδες

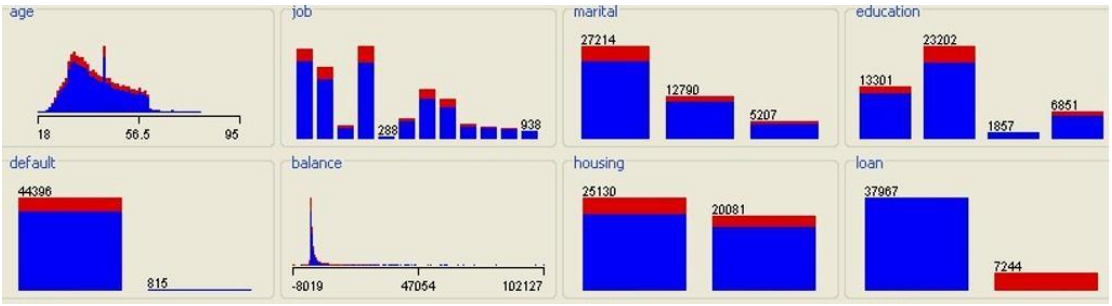
Τέλος δίνονται τρεις χρήσιμες ομάδες πινάκων, μέσω της επιλογής Visualize All, που παρουσιάζουν σχηματικά τη συσχέτιση των μεταβλητών του μορφωτικού επιπέδου, του στεγαστικού δανείου και του καταναλωτικού δανείου με όλες τις μεταβλητές του δείγματος. Στην πρώτη ομάδα χρησιμοποιείται από το Weka μπλε χρώμα για την τριτοβάθμια εκπαίδευση, κόκκινο για τη δευτεροβάθμια, γκρι για την πρωτοβάθμια και γαλάζιο για άγνωστο μορφωτικό επίπεδο, στη δεύτερη μπλε χρώμα για ύπαρξη στεγαστικού δανείου και κόκκινο για απουσία στεγαστικού δανείου και στην τρίτη μπλε χρώμα για απουσία καταναλωτικού δανείου και κόκκινο για ύπαρξη καταναλωτικού δανείου.



ΕΙΚΟΝΑ 12- ΣΥΣΧΕΤΙΣΗ ΜΟΡΦΩΤΙΚΟΥ ΕΠΙΠΕΔΟΥ ΜΕ ΟΛΕΣ ΤΙΣ ΜΕΤΑΒΛΗΤΕΣ ΤΟΥ ΔΕΙΓΜΑΤΟΣ



ΕΙΚΟΝΑ 13- ΣΥΣΧΕΤΙΣΗ ΥΠΑΡΞΗΣ ΣΤΕΓΑΣΤΙΚΟΥ ΔΑΝΕΙΟΥ ΜΕ ΟΛΕΣ ΤΙΣ ΜΕΤΑΒΛΗΤΕΣ ΤΟΥ ΔΕΙΓΜΑΤΟΣ



ΕΙΚΟΝΑ 13- ΣΥΣΧΕΤΙΣΗ ΥΠΑΡΞΗΣ ΚΑΤΑΝΑΛΩΤΙΚΟΥ ΔΑΝΕΙΟΥ ΜΕ ΟΛΕΣ ΤΙΣ ΜΕΤΑΒΛΗΤΕΣ ΤΟΥ ΔΕΙΓΜΑΤΟΣ

ΚΕΦΑΛΑΙΟ 5

ΕΞΑΓΩΓΗ ΓΝΩΣΗΣ ΑΠΟ ΤΑ ΔΕΔΟΜΕΝΑ

Οι αλγόριθμοι που θα συγκρίνουμε και έπειτα θα χρησιμοποιήσουμε για την κατασκευή του εργαλείου μας ανήκουν στις πιο διαδεδομένες τεχνικές. Αυτές είναι τα δέντρα απόφασης (Decision Trees), οι κανόνες ταξινόμησης (Rulebased Classification), και τα τεχνητά νευρωνικά δίκτυα (Neural Networks).Σε αυτή τη φάση της εργασίας έγιναν οι μετρήσεις και τα πειράματα. Πιο συγκεκριμένα

κάνουμε σύγκριση μεταξύ διάφορων αλγορίθμων της κάθε κατηγορίας, για να δούμε ποιος αλγόριθμός δίνει τη μεγαλύτερη ακρίβεια. Οι αλγόριθμοι που θα χρησιμοποιήσουμε αναλύονται παρακάτω.

5.1 ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΛΓΟΡΙΘΜΟΥ LOGISTIC

Το λογιστικό μοντέλο είναι ένα μη γραμμικό μοντέλο στο οποίο όμως τα σφάλματα δεν ακολουθούν κανονική κατανομή και η μεταβλητή απόκρισης είναι διακριτή. Η λογιστική παλινδρόμηση χρησιμοποιείται σε περιπτώσεις στις οποίες επιθυμούμε να προβλέψουμε την απουσία ή την παρουσία ενός χαρακτηριστικού, ή ενός συμβάντος. Είναι μια γενίκευση της απλής γραμμικής παλινδρόμησης για την περίπτωση όπου η εξαρτημένη μεταβλητή (Y) είναι δίτιμη (δηλαδή παίρνει την τιμή 0 όταν απουσιάζει το χαρακτηριστικό ή την τιμή 1 όταν υπάρχει το χαρακτηριστικό)

Τα μοντέλα που εξετάστηκαν με τον αλγόριθμο Logistic στο Weka ξεπέρασαν τα 10 σε αριθμό, κάποια εκ των οποίων παρουσιάζονται ακολούθως. Οι πληροφορίες εφαρμογής του αλγορίθμου, όπως δίνονται από το Weka, φαίνονται παρακάτω. Αρχικά, γίνεται δοκιμή του αλγορίθμου με όλες τις μεταβλητές που υπάρχουν στο σύνολο δεδομένων. Αρχικα δείχνει ότι το μοντέλο πέτυχε μια ικανοποιητική πρόβλεψη με 40.758 σωστά κατανεμημένες περιπτώσεις στο δείγμα (ποσοστό επιτυχίας 90,1506% και αντίστοιχα αποτυχίας 9,8494%, όπως φαίνεται πάνω κεντρικά της εικόνας [Π.1]), με το confusion matrix να δίνει πιο συγκεκριμένες πληροφορίες, κατά τις οποίες προβλέφθηκαν σωστά οι 38.927 από τους 39.922 (οι υπόλοιποι 995 προβλέφθηκε, λανθασμένα, ότι θα απαντήσουν θετικά) και οι 1.831 από τους 5.289 (οι υπόλοιποι 3.458 προβλέφθηκε, λανθασμένα, ότι θα απαντήσουν αρνητικά).

Με τον ίδιο τρόπο σχολιάζονται όλα τα μοντέλα πρόβλεψης όλων των αλγορίθμων από εδώ και στο εξής, δίνοντας έμφαση σε εκείνα στα οποία αυτό κρίνεται απαραίτητο να συμβεί.

Στην προσπάθεια να ανακαλυφθεί ένα ακόμα καλύτερο σε απόδοση μοντέλο πρόβλεψης, ακολουθήθηκε η διαδικασία αφαίρεσης των μεταβλητών μία προς μία και αντίστοιχη καταγραφή των αποτελεσμάτων που έδωσε ο αλγόριθμος.

Τα βελτιωμένα μοντέλα προήλθαν από την αφαίρεση, ξεχωριστά, των μεταβλητών της οικογενειακής κατάστασης [Π.2] ,του μορφωτικού επιπέδου[Π.3] και του μέσου ετησίου υπολοίπου σε ευρώ [Π.4].

Μεταβλητή που αφαιρέθηκε	Ποσοστό επιτυχούς πρόβλεψης
Οικογενειακής κατάστασης	90,1727%
Μορφωτικού επιπέδου	90,175%

Μέσου ετησίου υπολοίπου	90,1617%

5.2 ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΛΓΟΡΙΘΜΟΥ J48

Ο J48 χρησιμοποιεί το γεγονός ότι κάθε χαρακτηριστικό των δεδομένων μπορεί να χρησιμοποιηθεί για να λάβει μια απόφαση, η οποία χωρίζει τα δεδομένα σε μικρότερα υποσύνολα. Ο J48 εξετάζει το ομαλοποιημένο κέρδος πληροφοριών (information gain - διαφορά στην εντροπία) που προκύπτει από την επιλογή ενός χαρακτηριστικού για το διαχωρισμό των δεδομένων. Το χαρακτηριστικό με το υψηλότερο ομαλοποιημένο κέρδος πληροφοριών είναι αυτό που χρησιμοποιείται για να ληφθεί μια απόφαση

Οποτε για τον αλγόριθμο J48 η στρατηγική που ακολουθείται στην εξαγωγή μοντέλων είναι ακριβώς η ίδια με τον προηγούμενο. Αυτό που γίνεται, δηλαδή, είναι η παρουσίαση των πληροφοριών της εφαρμογής του αλγορίθμου και του μοντέλου πρόβλεψής του με όλες τις μεταβλητές και στη συνέχεια η αφαίρεση κάθε μίας μεταβλητής ξεχωριστά για την εξαγωγή καλύτερων μοντέλων.

Επίσης, δοκιμάζονται τα μοντέλα εκείνα που δοκιμάστηκαν και στο Logistic για να διαπιστωθεί ποια ομάδα μεταβλητών παίζει το σημαντικότερο ρόλο στην τελική απόφαση αν ενας πελατης μπορει να προμηθευτεί πιστωτικη καρτα.

Το πρώτο συμπέρασμα που μπορεί να βγει είναι το γεγονός πως ο αλγόριθμος J48, εκ πρώτης όψεως τουλάχιστον, είναι πιο αποδοτικός, ποιοτικά, από τον αλγόριθμο Logistic, καθώς δίνει ποσοστό πρόβλεψης ίσο με 90,3121% που είναι μεγαλύτερο από το 90,1506% που έδινε ο δεύτερος [Π.5]. Μια διαφορά που όπως υπογραμμίστηκε στο προηγουμενως μπορεί να μοιάζει μικρή, αλλά ενδεχομένως να είναι πολύ σημαντική σε μεγαλύτερα δείγματα από αυτό το οποίο εξετάζεται.

Το μοντέλο βελτιώνεται σε περισσότερες περιπτώσεις αφαίρεσης μεμονωμένων μεταβλητών σε σχέση με αυτό που συνέβαινε με τον προηγούμενο αλγόριθμο. Συγκεκριμένα, τα μοντέλα πρόβλεψης που ανεβάζουν την απόδοση του αρχικού μοντέλου J48, με όλες τις μεταβλητές, δηλαδή, προκύπτουν από την αφαίρεση, ξεχωριστά, των μεταβλητών της ηλικίας [Π.6], του μέσου ετησίου υπολοίπου σε ευρώ [Π.7], της ύπαρξης στεγαστικού δανείου [Π.8] και της ύπαρξης καταναλωτικού δανείου [Π.9]

Μεταβλητή που αφαιρέθηκε	Ποσοστό επιτυχούς πρόβλεψης
Ηλικιας	90,3298 %
Μεσου Ετησιου υπολοιπου	90,3165%
Υπαρξη Καταναλωτικου	90,3674%

Υπαρξη Στεγαστικου Δανειου	90,332%

5.3 ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΛΓΟΡΙΘΜΟΥ SMO

Ο αλγόριθμος της διαδοχικής ελάχιστης βελτιστοποίησης-SMO (sequential minimal optimization) εφαρμόζει τον διαδοχικό αλγόριθμο ελάχιστης βελτιστοποίησης για την εκπαίδευση ενός "*support vector classifier*" χρησιμοποιώντας πολυωνυμικούς ή γκαουσιανούς .Αυτή η εφαρμογή αντικαθιστά συνολικά όλες τις ελλιπείς τιμές και μετασχηματίζει τα ονομαστικά χαρακτηριστικά σε δυαδικά. Εξομαλύνει επίσης όλα τα χαρακτηριστικά, σ' αυτή την περίπτωση οι συντελεστές των αποτελεσμάτων είναι βασισμένοι στα ομαλοποιημένα στοιχεία και όχι στα αρχικά στοιχεία - αυτό είναι σημαντικό για την ερμηνεία του ταξινομητή. Τα πολλαπλής κλάσης προβλήματα λύνονται χρησιμοποιώντας ταξινόμηση ανά ζευγάρια, δηλαδή 1-εναντίον-1 (pairwise).

Ολοκληρωνοντας το κεφάλαιο, παρασουσιαζουμε τα αποτελεσματα του αλγορίθμου SMO. Η διαδικασία που ακολουθείται είναι ακριβώς η ίδια με τους προηγούμενους δύο αλγορίθμους. Σε αυτό το σημείο πρέπει να τονιστεί ότι παρουσιάστηκε ένα πολύ σημαντικό πρόβλημα με το συγκεκριμένο αλγόριθμο. Τόσο το γεγονός ότι ο αλγόριθμος αυτός είναι από τη φύση του πολύ αργός, όσο και το γεγονός ότι το δείγμα το οποίο εξετάζεται έχει πολύ μεγάλο μέγεθος είχαν ως αποτέλεσμα ο χρόνος που απαιτείται από το Weka για να εξάγει ένα μοντέλο πρόβλεψης να είναι πολύ μεγάλος. Ουσιαστικά ο αλγόριθμος SMO δεν είναι, χρονικά, αποδοτικός στη συγκεκριμένη περίπτωση, και σίγουρα πολύ χειρότερος σε σύγκριση με τις προηγούμενες δύο περιπτώσεις.

Για την αντιμετώπιση του άνωθεν προβλήματος προτιμήθηκε η επιλογή να εξαχθούν μοντέλα πρόβλεψης που θα αφορούσαν μονάχα ένα ποσοστό του συνολικού δείγματος του συνόλου των δεδομένων, το οποίο πρέπει να έχει δύο κύρια χαρακτηριστικά. Αφενός, να είναι ένα ποσοστό ικανοποιητικό σε μέγεθος, ώστε να μπορούν να βγουν ασφαλή και αξιόπιστα αποτελέσματα και αφετέρου, να είναι ένα τέτοιο ποσοστό το οποίο μπορεί γρήγορα και εύκολα να επεξεργαστεί το Weka χωρίς να χάνει πολύτιμο χρόνο. Αποφασίστηκε, λοιπόν, το ποσοστό αυτό να οριστεί στο 10% των μεταβλητών, ήτοι 4.521 περιπτώσεις.

Όσον αφορά τα μοντέλα πρόβλεψης που εξήχθησαν, το αρχικό ποσοστό πρόβλεψης κυμάνθηκε στο 88,9847% [Π.10]. Δεν μπορεί να υπάρξει ποσοτική σύγκριση με τους άλλους αλγορίθμους διότι παρήγαγαν μοντέλα με περισσότερα δεδομένα. Αντίθετα, μπορεί να υπάρξει ποιοτική σύγκριση μεταξύ τους, στο επίπεδο των μεταβλητών των οποίων η αφαίρεση βελτιώνει ή χειροτερεύει το αρχικό μοντέλο.

Παρατηρήθηκε βελτίωση του αρχικού μοντέλου με την αφαίρεση μίας προς μία των μεταβλητών του επιπέδου μόρφωσης[Π.11] και του μέσου ετησίου υπολοίπου σε ευρω [Π.12].

Μεταβλητή που αφαιρέθηκε	Ποσοστό επιτυχούς πρόβλεψης
Καταναλωτικου Δανειου	88.8963%
Μορφωτικού επιπέδου	89,029%

Αυτό, λοιπόν, που διαπιστώνεται είναι πως και οι τρεις αλγόριθμοι συμφωνούν ότι η μεταβλητή του μορφωτικού επιπέδου επιβαρύνει τα μοντέλα και, συνεπώς, δεν βοηθάει την πρόβλεψη. Επιπλέον, ο SMO συμφωνεί με τον J48 για τις μεταβλητές καταναλωτικού δανείου ενώ ταυτόχρονα συμφωνεί και με το Logistic όσον αφορά τη μεταβλητή του μορφωτικού επιπέδου.

ΚΕΦΑΛΑΙΟ 6

ΣΥΜΠΕΡΑΣΜΑΤΑ

Στη παρούσα εργασία ασχοληθήκαμε με την επιστήμη της εξόρυξης γνώσης και των δεδομένων, κάναμε μια εισαγωγή στις βασικές της έννοιες, αναφερθήκαμε στους στόχους, τα στάδια, τις κατηγορίες και τις μεθόδους της επιστήμης αυτής, ενώ απαραίτητη ήταν και η αναφορά στα πλεονεκτήματα και τα μειονεκτήματα που απορρέουν από τη συγκεκριμένη επιστήμη. Σημαντική ήταν επίσης και η αναφορά στα εργαλεία και τα μοντέλα που χρησιμοποιεί η εξόρυξη γνώσης, την αποτελεσματικότητα και χρησιμότητα τους και το κατά πόσο αυτά μπορούν να υλοποιηθούν από τις τραπεζικες επιχειρήσεις της σύγχρονης εποχή.

Καταλήγοντας λοιπόν, μπορούμε να πούμε ότι η εξόρυξη γνώσης από βάσεις δεδομένων, αποτελεί μία από τις πλέον σύγχρονες τεχνικές εξόρυξης πληροφορίας από ακατέργαστα δεδομένα. Πιο αναλυτικά, ψάχνει δομές και πληροφορίες σε ένα τεράστιο όγκο δεδομένων με σκοπό την μετατροπή αυτών σε χρήσιμη πληροφορία. Οι εφαρμογές της εξόρυξης δεδομένων εκτείνονται σε όλο το φάσμα των επιστημών από Πληροφορική, Οικονομία, Εκπαίδευση μέχρι Βιολογία και Αστρονομία.

Όλες τις πληροφορίες για την παρούσα εργασία τις αντλήσαμε από ελληνική και ξένη βιβλιογραφία, ενώ σημαντική ήταν και η συνεισφορά πηγών μέσω διαδικτύου. Οι συνεχείς έρευνες και οι όλο και πιο εξονυχιστικές μελέτες πάνω στην επιστήμη της εξόρυξης γνώσης, έχουν σαν αποτέλεσμα να την καθιστούν σαν ένα από τα πλέον απαραίτητα εργαλεία στη σημερινή κοινωνία.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. **Ben-Gal, I. (2007), "Bayesian Networks", Ruggeri F., Faltin F. & Kenett R., Encyclopedia of Statistics in Quality & Reliability, Wiley & Sons**

2. **Hastie, T., Tibshirani, R., Friedman, J. (2009), The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, Springer.**

3. **MacKay, D. (2005), Information Theory, Inference, and Learning Algorithms, Version 7.2 (fourth printing), Cambridge University Press.**

4. **Anand, S., Rajesh, K. (2012). Analysis of SEER Dataset for Breast Cancer Diagnosis using C4.5 Classification Algorithm. International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 2, April 2012**

5. **S. B. Kotsiantis, (2007). Supervised machine learning: A review of classification techniques, Informatica 31, 249-268**

6. **Sotiris B. Kotsiantis, Ioannis D. Zaharakis, Panayiotis E. Pintelas: Machine learning: a review of classification and combining techniques. Artif. Intell. Rev. 26(3): 159-190 (2006)**

7. **I. E. Livieris, P. Pintelas (2008). A survey on algorithms for training artificial neural networks, Technical Report TR08-01, Department of Mathematics, University of Patras**

8. **W. J. Boyes, D. L. Hoffman & S. A. Low (1989). An econometric analysis of the bank credit scoring problem. Journal of Econometrics, 40, 3-14**

9. **Y. Liu & M. Schumann, (2005). Data mining feature selection for credit scoring models, Journal of the Operational Research Society, 1099-1108**

10. **M. H. Dunham, (2004). Data Mining: Εισαγωγικά και Προηγμένα Θέματα Εξόρυξης Γνώσης από Δεδομένα, Εκδόσεις Νέων Τεχνολογιών**

11. **K. Murthy, (1998). Automatic construction of decision trees from data: A multidisciplinary survey, Data Mining and Knowledge Discovery, 2(4), 345389**

12. **C. S. Ong, J. J. Huang & G. H. Tzeng, (2005). Building credit scoring models using genetic programming, Expert Systems with Applications: An International Journal, 29, 41-47**

13. **D. West, (2000). Neural network credit scoring models. Computers and Operations Research, 27(11-12), 1131-1152**

14. **Σταυλιώτης Ε. Γεράσιμος (2009): Εξόρυξη Δεδομένων και Αναγνώριση προτύπων σε κατηγορικά δεδομένα μέσω συσταδοποίησης, Ελληνικό Στατιστικό Ινστιτούτο**

15. **Kumar, Steinbach, Tan (2004): Introduction to Data Mining, University of Stanford**

16. https://en.wikipedia.org/wiki/Decision_tree_learning

17. https://en.wikipedia.org/wiki/Weka_(machine_learning)

ΠΑΡΑΡΤΗΜΑ

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       40758               90.1506 %
Incorrectly Classified Instances      4453                9.8494 %
Kappa statistic                          0.4026
Mean absolute error                      0.1391
Root mean squared error                  0.2667
Relative absolute error                 67.3306 %
Root relative squared error             82.9921 %
Coverage of cases (0.95 level)          99.3055 %
Mean rel. region size (0.95 level)      72.3718 %
Total Number of Instances            45211

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.975    0.654    0.918      0.975   0.946      0.427    0.907     0.985     no
               0.346    0.025    0.648      0.346   0.451      0.427    0.907     0.551     yes
Weighted Avg.  0.902    0.58     0.887      0.902   0.888      0.427    0.907     0.934

=== Confusion Matrix ===

     a     b   <-- classified as
 38927   995 |     a = no
  3458  1831 |     b = yes
```

[1]-ΜΟΝΤΕΛΟ ΠΡΟΒΛΕΨΗΣ LOGISTIC ΜΕ ΧΡΗΣΗ ΟΛΩΝ ΤΩΝ ΜΕΤΑΒΛΗΤΩΝ

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       40768               90.1727 %
Incorrectly Classified Instances      4443                9.8273 %
Kappa statistic                          0.4022
Mean absolute error                      0.1393
Root mean squared error                  0.2669
Relative absolute error                 67.4217 %
Root relative squared error             83.0375 %
Coverage of cases (0.95 level)          99.2922 %
Mean rel. region size (0.95 level)      72.5211 %
Total Number of Instances            45211

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.976    0.656    0.918      0.976   0.946      0.427    0.906     0.985     no
               0.344    0.024    0.651      0.344   0.451      0.427    0.906     0.55      yes
Weighted Avg.  0.902    0.582    0.887      0.902   0.888      0.427    0.906     0.934

=== Confusion Matrix ===

     a     b   <-- classified as
 38946   976 |     a = no
  3467  1822 |     b = yes
```

[2]-ΜΟΝΤΕΛΟ ΠΡΟΒΛΕΨΗΣ LOGISTIC

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       40769               90.175  %
Incorrectly Classified Instances      4442                9.825  %
Kappa statistic                          0.4045
Mean absolute error                      0.1392
Root mean squared error                  0.2668
Relative absolute error                 67.3915 %
Root relative squared error             83.0253 %
Coverage of cases (0.95 level)          99.3055 %
Mean rel. region size (0.95 level)      72.5454 %
Total Number of Instances            45211

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.975    0.652    0.919      0.975   0.946      0.428    0.907     0.985     no
               0.348    0.025    0.649      0.348   0.453      0.428    0.907     0.551     yes
Weighted Avg.  0.902    0.579    0.887      0.902   0.888      0.428    0.907     0.934

=== Confusion Matrix ===

     a     b   <-- classified as
 38929   993 |     a = no
  3449  1840 |     b = yes
```

[3]-ΜΟΝΤΕΛΟ ΠΡΟΒΛΕΨΗΣ LOGISTIC

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       40763               90.1617 %
Incorrectly Classified Instances      4448                9.8383 %
Kappa statistic                          0.4032
Mean absolute error                      0.1391
Root mean squared error                  0.2667
Relative absolute error                 67.33   %
Root relative squared error             82.9813 %
Coverage of cases (0.95 level)          99.2988 %
Mean rel. region size (0.95 level)      72.3906 %
Total Number of Instances            45211

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.975    0.653    0.918      0.975   0.946      0.427    0.907     0.985     no
               0.347    0.025    0.649      0.347   0.452      0.427    0.907     0.551     yes
Weighted Avg.  0.902    0.58     0.887      0.902   0.888      0.427    0.907     0.934

=== Confusion Matrix ===

     a     b   <-- classified as
 38930   992 |     a = no
  3456  1833 |     b = yes
```

[4]-ΜΟΝΤΕΛΟ ΠΡΟΒΛΕΨΗΣ LOGISTIC

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       40831               90.3121 %
Incorrectly Classified Instances      4380                9.6879 %
Kappa statistic                          0.4839
Mean absolute error                      0.1269
Root mean squared error                  0.2773
Relative absolute error                 61.4259 %
Root relative squared error             86.2833 %
Coverage of cases (0.95 level)          97.3657 %
Mean rel. region size (0.95 level)      66.0072 %
Total Number of Instances            45211

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC    ROC Area  PRC Area  Class
               0.959    0.519    0.933      0.959   0.946      0.488  0.843     0.947     no
               0.481    0.041    0.609      0.481   0.537      0.488  0.843     0.486     yes
Weighted Avg.  0.903    0.463    0.895      0.903   0.898      0.488  0.843     0.893

=== Confusion Matrix ===

     a     b   <-- classified as
 38289  1633 |     a = no
  2747  2542 |     b = yes
```

[5]-ΜΟΝΤΕΛΟ ΠΡΟΒΛΕΨΗΣ J48 ΜΕ ΧΡΗΣΗ ΟΛΩΝ ΤΩΝ ΜΕΤΑΒΛΗΤΩΝ

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       40839               90.3298 %
Incorrectly Classified Instances      4372                9.6702 %
Kappa statistic                          0.4852
Mean absolute error                      0.1281
Root mean squared error                  0.2754
Relative absolute error                 62.005  %
Root relative squared error             85.6895 %
Coverage of cases (0.95 level)          97.5028 %
Mean rel. region size (0.95 level)      66.2405 %
Total Number of Instances            45211

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC    ROC Area  PRC Area  Class
               0.959    0.518    0.933      0.959   0.946      0.489  0.847     0.951     no
               0.482    0.041    0.61       0.482   0.538      0.489  0.847     0.492     yes
Weighted Avg.  0.903    0.462    0.895      0.903   0.898      0.489  0.847     0.897

=== Confusion Matrix ===

     a     b   <-- classified as
 38289  1633 |     a = no
  2739  2550 |     b = yes
```

[6]-ΜΟΝΤΕΛΟ ΠΡΟΒΛΕΨΗΣ J48

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       40833               90.3165 %
Incorrectly Classified Instances      4378                9.6835 %
Kappa statistic                          0.4835
Mean absolute error                      0.1278
Root mean squared error                  0.2749
Relative absolute error                 61.8372 %
Root relative squared error             85.5338 %
Coverage of cases (0.95 level)          97.5935 %
Mean rel. region size (0.95 level)      66.2571 %
Total Number of Instances            45211

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC    ROC Area  PRC Area  Class
               0.959    0.521    0.933      0.959   0.946      0.488  0.852     0.952     no
               0.479    0.041    0.609      0.479   0.537      0.488  0.852     0.493     yes
Weighted Avg.  0.903    0.464    0.895      0.903   0.898      0.488  0.852     0.899

=== Confusion Matrix ===

     a     b   <-- classified as
 38297  1625 |     a = no
  2753  2536 |     b = yes
```

[7]-ΜΟΝΤΕΛΟ ΠΡΟΒΛΕΨΗΣ J48

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       40856               90.3674 %
Incorrectly Classified Instances      4355                9.6326 %
Kappa statistic                          0.4822
Mean absolute error                      0.1278
Root mean squared error                  0.2775
Relative absolute error                 61.8525 %
Root relative squared error             86.342  %
Coverage of cases (0.95 level)          97.3745 %
Mean rel. region size (0.95 level)      66.4186 %
Total Number of Instances            45211

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC    ROC Area  PRC Area  Class
               0.961    0.527    0.932      0.961   0.946      0.487  0.837     0.945     no
               0.473    0.039    0.615      0.473   0.535      0.487  0.837     0.484     yes
Weighted Avg.  0.904    0.47     0.895      0.904   0.898      0.487  0.837     0.891

=== Confusion Matrix ===

     a     b   <-- classified as
 38352  1570 |     a = no
  2785  2504 |     b = yes
```

[8]-ΜΟΝΤΕΛΟ ΠΡΟΒΛΕΨΗΣ J48

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       40840               90.332  %
Incorrectly Classified Instances      4371                9.668  %
Kappa statistic                          0.4855
Mean absolute error                      0.1267
Root mean squared error                  0.2775
Relative absolute error                 61.3379 %
Root relative squared error             86.3492 %
Coverage of cases (0.95 level)          97.3192 %
Mean rel. region size (0.95 level)      65.9364 %
Total Number of Instances            45211

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.959    0.517    0.933      0.959   0.946      0.49     0.843     0.948     no
               0.483    0.041    0.61       0.483   0.539      0.49     0.843     0.485     yes
Weighted Avg.  0.903    0.462    0.895      0.903   0.898      0.49     0.843     0.894

=== Confusion Matrix ===

    a     b   <-- classified as
 38288  1634 |     a = no
  2737  2552 |     b = yes
```

[9]-ΜΟΝΤΕΛΟ ΠΡΟΒΛΕΨΗΣ J48

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        4023               88.9847 %
Incorrectly Classified Instances       498               11.0153 %
Kappa statistic                          0.2937
Mean absolute error                      0.1102
Root mean squared error                  0.3319
Relative absolute error                 51.9159 %
Root relative squared error            101.9318 %
Total Number of Instances             4521

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.979     0.763      0.903     0.979     0.94       0.608    no
                 0.237     0.021      0.611     0.237     0.341      0.608    yes
Weighted Avg.    0.89      0.674      0.868     0.89      0.868      0.608

=== Confusion Matrix ===

    a    b   <-- classified as
 3894   82 |     a = no
  416  129 |     b = yes
```

[10]-ΜΟΝΤΕΛΟ ΠΡΟΒΛΕΨΗΣ SMO ΜΕ ΧΡΗΣΗ ΟΛΩΝ ΤΩΝ ΜΕΤΑΒΛΗΤΩΝ

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        4019               88.8963 %
Incorrectly Classified Instances       502               11.1037 %
Kappa statistic                          0.2283
Mean absolute error                      0.111
Root mean squared error                  0.3332
Relative absolute error                 52.3329 %
Root relative squared error            102.3404 %
Total Number of Instances             4521

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.988     0.833      0.896     0.988     0.94       0.577    no
                 0.167     0.012      0.655     0.167     0.266      0.577    yes
Weighted Avg.    0.889     0.734      0.867     0.889     0.859      0.577

=== Confusion Matrix ===

    a    b   <-- classified as
 3928   48 |    a = no
  454   91 |    b = yes
```

[11]-ΜΟΝΤΕΛΟ ΠΡΟΒΛΕΨΗΣ SMO

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        4025               89.029  %
Incorrectly Classified Instances       496               10.971  %
Kappa statistic                          0.2858
Mean absolute error                      0.1097
Root mean squared error                  0.3312
Relative absolute error                 51.7075 %
Root relative squared error            101.7269 %
Total Number of Instances              4521

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.981     0.774      0.902     0.981     0.94       0.604    no
                 0.226     0.019      0.624     0.226     0.332      0.604    yes
Weighted Avg.    0.89      0.683      0.869     0.89      0.867      0.604

=== Confusion Matrix ===

    a    b   <-- classified as
 3902   74 |    a = no
  422  123 |    b = yes
```

[12]-ΜΟΝΤΕΛΟ ΠΡΟΒΛΕΨΗΣ SMO