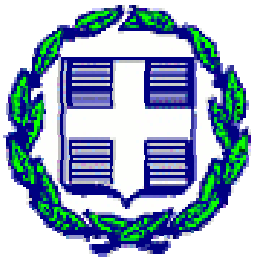
**ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ**

**ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ**
**ΤΜΗΜΑ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ (ΠΑΤΡΑ)**



**ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ**

# «Τεχνικές Προγνωστικής Μοντελοποίησης (Predictive Analytics) για την αντιμετώπιση σύνθετων επιχειρηματικών προβλημάτων»

**Πτυχιακή Εργασία** των

**Σωτηρίου Μαρία**

**Τουλάτου Χρυσούλα**

Επιβλέπων Καθηγητής: Γεράσιμος Ατζουλάτος

**ΠΑΤΡΑ, 2015**

# ΕΥΧΑΡΙΣΤΙΕΣ

Πριν από όλα θα θέλαμε να ευχαριστήσουμε τον επιβλέποντα καθηγητή μας κύριο Γεράσιμο Αντζουλάτο που μας έδωσε τη δυνατότητα να ασχοληθούμε με ένα τόσο ενδιαφέρον θέμα. Θα θέλαμε επίσης να τον ευχαριστήσουμε για τη βοήθεια, την καθοδήγηση και τις χρήσιμες συμβουλές που μας παρείχε. Ήταν πάντα διαθέσιμος να ασχοληθεί με κάθε απορία μας και να μας προσφέρει τις γνώσεις και την εμπειρία του.

Θα θέλαμε να ευχαριστήσουμε ακόμα, όλους τους καθηγητές του Ανώτατου Τεχνολογικού Εκπαιδευτικού Ιδρύματος Δυτικής Ελλάδας με έδρα την Πάτρα για τις πολύτιμες γνώσεις που μας προσέφεραν όλα αυτά τα χρόνια.

Τέλος, ένα μεγάλο ευχαριστώ οφείλουμε και στους γονείς μας, των οποίων η πίστη στις δυνατότητές μας αποτέλεσε αρωγό σε όλους τους στόχους και τα όνειρά μας, καθώς μας στήριξαν με κάθε τρόπο σε όλη τη διάρκεια των σπουδών μας.

# ΠΕΡΙΛΗΨΗ

Η παρούσα πτυχιακή εργασία πραγματοποιείται στα πλαίσια των προπτυχιακών σπουδών μας στο «ΑΝΩΤΑΤΟ ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ» και συγκεκριμένα στην σχολή «ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ», τμήμα «ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ (ΠΑΤΡΑ)». Ο τίτλος αυτής είναι *«Τεχνικές Προγνωστικής Μοντελοποίησης (Predictive Analytics) για την αντιμετώπιση σύνθετων επιχειρηματικών προβλημάτων»*.

Στα πλαίσια αυτής αναλύονται βασικές έννοιες όπως η εξόρυξη δεδομένων και η επιχειρηματική ευφυΐα. Η σπουδαιότητα της διαχείρισης μεγάλου όγκου δεδομένων είναι αναμφισβήτητα μεγάλη στις μέρες μας όπου όλα αλλάζουν άρδην και η πληροφορία διαδραματίζει ιδιαίτερα σημαντικό ρόλο στη λήψη επιχειρηματικών αποφάσεων και στην εξαγωγή συμπερασμάτων και ασφαλών προβλέψεων.

Υπάρχει πληθώρα τεχνικών πρόγνωσης και μοντελοποίησης των επιχειρηματικών προβλημάτων, παρόλα αυτά στη συγκεκριμένη εργασία αναλύονται οι τεχνικές της ανάλυσης παλινδρόμησης από την μια ($2^{ο}$ κεφάλαιο) και τα δένδρα απόφασης από την άλλη ($3^{ο}$ κεφάλαιο). Αναλυτικότερα, αναφορικά με την ανάλυση παλινδρόμησης χρησιμοποιείται το πολυμεταβλητό υπόδειγμα με την ύπαρξη μιας εξαρτημένης μεταβλητής και πολλών ανεξάρτητων. Πριν από αυτό προσδιορίζουμε τις βασικές υποθέσεις που πρέπει να ισχύουν προκειμένου να εφαρμοστεί η ανάλυση παλινδρόμηση, πρόκειται για την ετεροσκεδαστικότητα, την πολυσυγγραμμικότητα και την αυτοσυσχέτιση. Μέσα από την εφαρμογή των δεδομένων μπορούμε από το υπόδειγμα που προέκυψε να κάνουμε προβλέψεις για μελλοντικά δεδομένα.

Στη συνέχεια χρησιμοποιούμε τα δένδρα απόφασης προκειμένου να διαπιστώσουμε πως αυτά εφαρμόζονται για την λήψη αποφάσεων. Στην ερευνητική κοινότητα υπάρχουν διαθέσιμες πολλές υλοποιήσεις δένδρων αποφάσεων, όπως είναι το λογισμικό ανοιχτού κώδικα WEKA, το οποίο και χρησιμοποιήσαμε.

# ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

# ΕΙΣΑΓΩΓΗ

Σήμερα ζούμε σε ένα περιβάλλον διαρκώς διευρυνόμενο και μεταβαλλόμενο, με τις πληροφορίες και τα δεδομένα να κατέχουν εξέχουσα θέση. Οι μέθοδοι πρόβλεψης αποτελούν εργαλεία πρόβλεψης των δεδομένων. Η προγνωστική μοντελοποίηση μας επιτρέπει να γνωρίζουμε τι συνέβη στο παρελθόν, εστιάζοντας στο τι θα συμβεί στο μέλλον. Η ανάγκη μας να κατανοήσουμε γεγονότα του παρελθόντος έχει οδηγήσει σε μια σειρά διαδικασιών που εμείς σήμερα αποκαλούμε επιχειρηματική ευφυΐα. Η διαδικασία αυτή επιτρέπει στις επιχειρήσεις να λαμβάνουν αποφάσεις και με βάση τα στατιστικά στοιχεία που προέρχονται από τα ιστορικά δεδομένα.

Η διαδικασία της προγνωστικής μοντελοποίησης είναι σε θέση να παράγει αξιόπιστες στατιστικές, προβλέψεις και αποτελέσματα. Για παράδειγμα, μπορεί να χρησιμοποιηθεί μια σειρά από κανόνες που να ενεργοποιούν επιχειρηματικές αποφάσεις ανάλογα με την απόδοση που λαμβάνεται από ένα μοντέλο πρόβλεψης. Επιπλέον, εάν υπάρχει ένα μοντέλο για να προβλέψει τον κίνδυνο της απώλειας πελατών, θα μπορεί να θέσει κανόνες προκειμένου να καθορίσει συγκεκριμένες επιχειρηματικές αποφάσεις ανάλογα με τα διαφορετικά επίπεδα κινδύνου. Ως εκ τούτου, εάν ο κίνδυνος είναι υψηλός, μπορούμε να δώσουμε σε έναν πελάτη μια έκπτωση 20% στην επόμενη αγορά του, αλλά εάν ο κίνδυνος είναι πολύ υψηλός, μπορούμε να δώσουμε μια έκπτωση 50% αντ 'αυτού.

Ένα μοντέλο πρόβλεψης είναι απλά μια μαθηματική συνάρτηση η οποία είναι σε θέση να χαρτογραφήσει ένα σύνολο μεταβλητών δεδομένων εισόδου, τα οποία συνήθως ομαδοποιούνται σε ένα αρχείο, από το οποίο αντλούνται απαντήσεις για στοχευόμενες μεταβλητές. Ένα προγνωστικό μοντέλο μπορεί επίσης να χρησιμοποιήσει εκμάθηση χωρίς επίβλεψη, δηλαδή δεν δίνεται καμία εξωτερική επέμβαση σχετικά με τους στόχους που πρέπει να αναγνωρίζει ένα σύστημα. Στην περίπτωση αυτή, παρουσιάζεται μόνο με τα δεδομένα εισόδου, σκοπός της συγκεκριμένης διαδικασίας είναι η κατανόηση της συσχέτισης των διαφορετικών αρχείων μεταξύ τους. Η Ομαδοποίηση είναι η πιο συχνά χρησιμοποιούμενη μέθοδος πρόβλεψης, η οποία χρησιμοποιεί εκμάθηση χωρίς επίβλεψη.

Όπως έχει ήδη αναφερθεί η παρούσα εργασία αποτελείται από τρία επιμέρους κεφάλαια, την εισαγωγή και τα συμπεράσματα. Στο πρώτο κεφάλαιο προσεγγίζονται οι βασικές έννοιες που αφορούν την εξόρυξη δεδομένων και την επιχειρηματική ευφυΐα γενικότερα. Στο δεύτερο κεφάλαιο αναλύεται η τεχνική της παλινδρόμησης και στο τρίτο η ανάλυση των δένδρων απόφασης. Τα συμπεράσματα αλλά και οι βιβλιογραφικές αναφορές ολοκληρώνουν την εν λόγω συγγραφική προσπάθεια.

# ΚΕΦΑΛΑΙΟ 1$^{Ο}$ «ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ - ΕΠΙΧΕΙΡΗΜΑΤΙΚΗ ΕΥΦΥΙΑ»

## 1.1. ΕΠΙΧΕΙΡΗΜΑΤΙΚΗ ΕΥΦΥΙΑ

Η έννοια Επιχειρηματική Ευφυΐα (Business Intelligence) είναι ένας ευρύς όρος ο οποίος αναφέρεται και εστιάζει σε τεχνολογίες, σε εφαρμογές, σε ικανότητες αλλά και πρακτικές που χρησιμοποιούνται από τις επιχειρήσεις προκειμένου να κατανοήσουν καλύτερα την αγοραστική συμπεριφορά και να εντοπίσουν πιθανές επιχειρηματικές ευκαιρίες. Ουσιαστικά πρόκειται για ένα σύνολο εννοιών, μεθόδων και τεχνολογιών σχεδιασμένο για να μετατρέψει όλα τα δεδομένα μιας επιχείρησης σε χρήσιμη πληροφορία και τελικά γνώση. Με άλλα λόγια πρόκειται για την παροχή της σωστής πληροφορίας, στους άμεσα ενδιαφερόμενους χρήστες, την σωστή χρονική στιγμή. (www.kopanakis.info).

Κατά τη διάρκεια των προηγούμενων δεκαετιών, οι επιχειρηματικές ενέργειες και οι αντίστοιχες αποφάσεις στηρίζονταν σε περιορισμένα δεδομένα και στην επιχειρηματική αντίληψη ή ένστικτο κάθε διευθυντή ή προέδρου. Σήμερα, ο μεγάλος όγκος ηλεκτρονικών δεδομένων (η συλλογή των οποίων γίνεται αυτοματοποιημένα), σε συνδυασμό με τις εξελίξεις σε επίπεδο τεχνολογίας, έχει δημιουργήσει την ανάγκη για ειδικά εργαλεία διαχείρισης και εκμετάλλευσης της επιχειρηματικής πληροφορίας. Οι ανάγκες της σύγχρονης εποχής άρα και της σύγχρονης επιχείρησης χρήζουν γρήγορης, αποδοτικής και συνεπής διαχείρισης ενός συνόλου πληροφοριών και ενός αντίστοιχου όγκου δεδομένων ποικίλης ύλης.

Η χρησιμοποίηση δεδομένων στη λήψη σωστών, έγκυρων και έγκαιρων αποφάσεων έχει αναχθεί σε «εκ των ουκ άνευ» παράγοντα επιτυχίας για τις περισσότερες σύγχρονες επιχειρήσεις και οργανισμούς. Ταυτόχρονα, τα τελευταία χρόνια, με την ανάπτυξη νέων τεχνολογιών και εφαρμογών – όπως η εξάπλωση των κοινωνικών δικτύων, η εκτεταμένη χρήση smart phones, η εγκατάσταση αισθητήρων κ.α. – ο όγκος και η μορφή των δεδομένων έχει αλλάξει δραματικά, ενώ οι δυνατότητες ανάλυσης και επεξεργασίας αυτών είναι εντυπωσιακές. Οι όροι

Επιχειρηματική Ευφυΐα (Business Intelligence), Μεγάλης Κλάσης Δεδομένα (Big Data) και Επιστήμη των Δεδομένων (Data Science) βρίσκονται στην καθημερινή δραστηριότητα των μικρών και μεγάλων οργανισμών.

Η έννοια Επιχειρηματική Ευφυΐα αναφέρεται σε τεχνολογίες, εφαρμογές, ικανότητες και πρακτικές που χρησιμοποιούνται για να παρέχουν στους χρήστες τη δυνατότητα να λαμβάνουν αποτελεσματικές επιχειρησιακές αποφάσεις (informed decisions) και διοικητικά μέτρα, βασισμένες σε επίκαιρα στοιχεία και αναλύσεις. (Aronson, 2007).

Η Επιχειρηματική Ευφυΐα χρησιμοποιεί μεθόδους οι οποίες αναλύουν τα δεδομένα μιας Αποθήκης Δεδομένων (ως Αποθήκη Δεδομένων θεωρούμε μια βάση δεδομένων που χρησιμοποιείται για την αναφορά και ανάλυση) και, είτε προτείνουν είτε υποβοηθούν μια επιχειρηματική - επιχειρησιακή απόφαση. Πρόκειται δηλαδή για αναλυτικές μεθόδους που βοηθούν στην εξαγωγή συμπερασμάτων. Η εξόρυξη δεδομένων (data mining) και η ηλεκτρονική αναλυτική επεξεργασία (OLAP analysis) είναι δύο από τις πιο γνωστές κατηγορίες μεθόδων και τεχνικών. Εδώ θα πρέπει να τονίσουμε ότι μιλάμε πάντα για μεθόδους και όχι για συγκεκριμένα εργαλεία.

Έχοντας γρήγορη πρόσβαση σε πληροφορίες, που διαφορετικά θα ήταν κρυμμένες στο μεγάλο όγκο δεδομένων εμπλουτίζονται οι δυνατότητες των στελεχών για αξιοποίηση των παρουσιαζόμενων ευκαιριών και αντιμετώπιση των πιθανών δυσλειτουργιών στην ομαλή λειτουργία του οργανισμού. Συνεπώς, είναι εφικτή η αποδοτικότερη:

- αναγνώριση νέων επιχειρηματικών ευκαιριών
- αποκάλυψη των επιδράσεων των διαφόρων διαδικασιών της οργάνωσης και της επιρροής που ασκούν τελικά στην επιχείρηση
- ενίσχυση των σχέσεων με τους πελάτες και συνεργάτες ενώ ταυτόχρονα κερδίζεται ένα σημαντικό ανταγωνιστικό πλεονέκτημα στην αγορά.

Μια άλλη προσέγγιση της επιχειρηματικής ευφυΐας μπορεί να παρουσιασθεί ακολούθως: **Επιχειρηματική Αναλυτική (Business Analytics-B.A)[1]:** Ο όρος Επιχειρηματική Αναλυτική, αναφέρεται στις δεξιότητες, τις τεχνολογίες και πρακτικές που είναι απαραίτητες για τη συνεχή και επαναλαμβανόμενη εξερεύνηση και διερεύνηση των προηγούμενων επιδόσεων των

[1] http://eduportal.dmst.aueb.gr/html/det/Lecture7Chapter8SysthmataEpixeirhmatikhEyfyiasv2-gr__18002.pdf

επιχειρήσεων και οργανισμών, προκειμένου αυτές να ενσωματώνονται στη διαδικασία του επιχειρηματικού-επιχειρησιακού σχεδιασμού. Επίσης η Επιχειρηματική Αναλυτική, εστιάζει στην ανάπτυξη νέων ιδεών και την κατανόηση των επιδόσεων των οργανισμών κάνοντας εκτεταμένη χρήση όλων των δεδομένων, στατιστικής και ποσοτικής ανάλυσης και προγνωστικής μοντελοποίησης, με σκοπό την έγκαιρη και έγκυρη λήψη αποφάσεων. Τα προϊόντα αυτής της επεξεργασίας μπορούν να χρησιμοποιηθούν είτε για την υποβοήθηση ανθρώπινων αποφάσεων είτε για την εκκίνηση άλλων πλήρως αυτοματοποιημένων. Αποτελεί λοιπόν το πρώτο και ίσως σημαντικότερο στάδιο για κάθε οργανισμό ώστε αυτός να μπορέσει να λαμβάνει καλύτερες και οικονομικότερες αποφάσεις.

Συνεχίζοντας τον ορισμό της υπό εξέτασης έννοιας, της επιχειρηματικής ευφυΐας να υπογραμμίσουμε πως αναφέρεται σε ένα σύνολο τεχνικών και πρακτικών διαχείρισης, ανάλυσης και παρουσίασης δεδομένων, με σκοπό την ταχεία πρόσβαση σε μεγάλους όγκους αξιόπιστων πληροφοριών (πχ βασικοί επιχειρησιακοί δείκτες) για την έγκαιρη, έγκυρη και αποτελεσματική λήψη αποφάσεων.

Οι λύσεις επιχειρηματικής ευφυΐας προσφέρουν τη δυνατότητα συγκέντρωσης, οργάνωσης και επεξεργασίας δεδομένων[2] από διάφορα σημεία-πηγές της επιχείρησης, δημιουργώντας ένα ομοιογενές και ολοκληρωμένο σύνολο μεγάλου όγκου δεδομένων και συντελώντας στις επίκαιρες και τεκμηριωμένες επιχειρηματικές αποφάσεις. Χρησιμοποιώντας διάφορα στατιστικά μοντέλα τα εργαλεία Business Intelligence επιχειρούν να προβλέψουν συγκεκριμένες μελλοντικές συμπεριφορές, λαμβάνοντας υπόψη διάφορες υπάρχουσες και μελλοντικές συνθήκες, έτσι ώστε η επιχείρηση να οδηγηθεί στη λήψη κατάλληλων μέτρων και αποφάσεων με βάση τους επιχειρηματικούς της στόχους.

Επιτυχημένο θεωρείται ένα έργο επιχειρησιακής νοημοσύνης όταν οι συμπληρωματικές και βασικές εφαρμογές, υποδομές, αρχιτεκτονική και υπηρεσίες συνδυάζονται με τον βέλτιστο τρόπο με σκοπό την παροχή ποιοτικής πληροφόρησης προς εξυπηρέτηση της επιχείρησης[3]. Σημαντικά επίσης κριτήρια αξιολόγησης της αποτελεσματικής υλοποίησης των λύσεων Business Intelligence είναι η εκτενής αξιολόγηση και κατανόηση των επιχειρησιακών αναγκών, η εστίαση

[2] http://www.ethnodata.gr/
[3] http://www.unisystems.gr/el/solutions-services-inside/company-solutions/epixirisiakes-efarmoges/solutions-bi.html

στις ιδιαίτερες ανάγκες κάθε πελάτη, το επίπεδο επιχειρηματικής ωριμότητας της επιχείρησης, η ποιότητα των δεδομένων και η υπάρχουσα αυτοματοποίηση των διαδικασιών.

### 1.1.1. DATA WAREHOUSE & ΗΛΕΚΤΡΟΝΙΚΗ ΔΙΑΔΙΚΑΣΙΑ ΑΝΑΛΥΣΗΣ (ONLINE ANALYTICAL PROCESSING – OLAP)

Μέρος της επιχειρηματικής ευφυΐας αποτελεί και η OLAP ανάλυση, η οποία αφορά μια τεχνολογία που χρησιμοποιείται για την οργάνωση μεγάλων Βάσεων Δεδομένων και την υποστήριξη της επιχειρηματικής ευφυΐας. Έχει τελειοποιηθεί με σκοπό την αναζήτηση και την υποβολή εκθέσεων, αντί για την επεξεργασία συναλλαγών. Οι βάσεις δεδομένων OLAP χωρίζονται σε έναν ή περισσότερους κύβους και κάθε κύβος είναι οργανωμένος και σχεδιασμένος κατά τέτοιον τρόπο ώστε να ταιριάζει με τον τρόπο ανάκτησης και ανάλυσης δεδομένων που χρησιμοποιείται. Οι βάσεις δεδομένων OLAP περιέχουν δύο βασικούς τύπους δεδομένων: τα μέτρα, τα οποία είναι αριθμητικά δεδομένα που αντιπροσωπεύουν τις ποσότητες και τους μέσους όρους που χρησιμοποιείτε για τη λήψη τεκμηριωμένων επιχειρηματικών αποφάσεων και τις διαστάσεις, οι οποίες είναι κατηγορίες που χρησιμοποιούνται για την οργάνωση αυτών των μεγεθών.

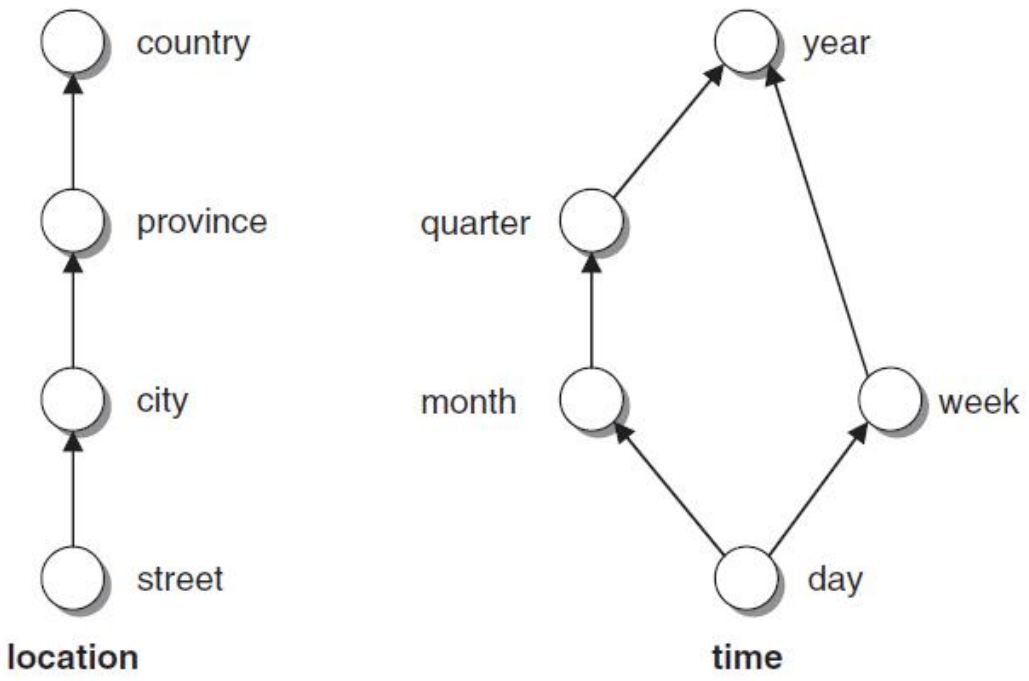
Σε πολλές περιπτώσεις, η OLAP[4] ανάλυση βασίζεται σε ιεράρχηση των εννοιών, την ενοποίηση των δεδομένων και τη δημιουργία λογικών τιμών κατά μήκος των διαστάσεων της αποθήκης δεδομένων. Κατά μια έννοια η ιεράρχηση ορίζεται από ένα σύνολο χαρτών από το χαμηλότερο επίπεδο εννοιών σε ένα υψηλότερο επίπεδο.

Μερικές από τις μεθόδους ιεράρχησης των εννοιών που χρησιμοποιούνται για να εκτελούν διάφορες λειτουργίες απεικόνισης, ασχολούνται με κύβους δεδομένων σε μια αποθήκη δεδομένων. Ενδεικτικά αναφέρονται οι ακόλουθοι μέθοδοι:

[4] http://kpe-kastor.kas.sch.gr/peekpe/proceedings/synedria_9_ereunes/Ravasopoulos_et_al.pdf

**Roll-up** [5] **.** Μια λειτουργία roll-up, που ονομάζεται επίσης drill-up, αποτελείται από μια συνάθροιση των στοιχείων στον κύβο, τα οποία εναλλακτικά μπορούν να ανακτηθούν με τους ακόλουθους δύο τρόπους:

- Προχωρώντας προς τα πάνω σε ένα υψηλότερο επίπεδο κατά μήκος μιας μόνο διάστασης που ορίζεται από περισσότερες από μια ιεραρχίες εννοιών.
- Μείωση κατά μία διάσταση. Για παράδειγμα, η απομάκρυνση του {χρόνου} διάσταση του χρόνου οδηγεί σε ενοποιημένα μέτρα από το άθροισμα πάνω από όλες τις χρονικές περιόδους που υπάρχουν στον κύβο δεδομένων.




**Roll-down:** Μια λειτουργία roll-down, αναφέρεται επίσης ως drill-down, είναι η αντίθετη λειτουργία του roll-up. Αυτό επιτρέπει την πλοήγηση μέσα από ένα κύβο δεδομένων από συγκεντρωτικές και ενοποιημένες πληροφορίες σε πιο λεπτομερείς πληροφορίες. Το αποτέλεσμα είναι να ανατρέψει το αποτέλεσμα που επιτεύχθηκε μέσω της λειτουργίας roll-up. Μια λειτουργία drill-down μπορεί να πραγματοποιηθεί με τους ακόλουθους δύο τρόπους.

[5] http://www.cs.uoi.gr/~pitoura/courses/dm09/warehouse09.pdf

- Η μετατόπιση προς τα κάτω σε ένα χαμηλότερο επίπεδο κατά μήκος μιας ενιαίας ιεραρχικής διάστασης.
- Η προσθήκη μιας διάστασης.

## 1.2. ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

### 1.2.1. ΕΙΣΑΓΩΓΗ

Στη σημερινή εποχή, οι περισσότερες εταιρείες και οργανισμοί, διατηρούν τεράστιες βάσεις δεδομένων στις οποίες καταγράφουν τόσο την κίνηση των πελατών τους ή στοιχεία για τους εργαζόμενους, όσο και διάφορες άλλες συναλλαγές για τα προϊόντα τους.

Αυτές οι βάσεις δεδομένων, μπορεί να φτάσουν να αποθηκεύουν τρισεκατομμύρια bytes δεδομένων. Ανάμεσα στην τεράστια μάζα δεδομένων, κρύβονται πληροφορίες οι οποίες μπορεί να είναι εξαιρετικής σημασίας. Είναι σα να έχει κανείς δεδομένα για τα δέντρα αλλά αυτά που χρειάζεται να είναι συμπεράσματα που να έχουν νόημα για το δάσος[6][7].

Μπορούν όμως αυτά τα δεδομένα να χρησιμοποιηθούν με τέτοιο τρόπο ώστε να είναι πραγματικά χρήσιμα για τους κατόχους τους; Την απάντηση σε αυτό το ερώτημα καλείται να τη δώσει ένας σχετικά νέος τομέας της τεχνολογίας, η Εξόρυξη Δεδομένων ή Data Mining[8]. Η εξόρυξη δεδομένων αφορά όλες τις τεχνικές και διαδικασίες, εύρεσης νέας και πιθανόν χρήσιμης γνώσης από μεγάλες Βάσεις Δεδομένων (1) [9][10].

[6] Sara Reese Hedberg, The Data Gold Rush, Byte.com magazine, October 1995
[7] Edmund X. DeJesus, Data Mining, Byte.com magazine, October 1995
[8] Berry, M., J., A., & Linoff, G., S., Mastering data mining. New York: Wiley, 2000
[9] Advances in Knowledge Discovery & Data Mining, U. Fayyad, et al.,editors; AAAI/MIT Press, 1995.
[10] Proceedings of the First International Conference on Knowledge Discovery and Data Mining, U. Fayyad, et al., editors; AAAI Press, 1995.

### 1.2.2. ΟΡΙΣΜΟΣ

Αρχικά θα πρέπει να κάνουμε τη διάκριση στους όρους δεδομένα και πληροφορική. Αναλυτικότερα, με τον όρο δεδομένα εννοούμε κάθε γεγονός, αριθμό ή κείμενο που μπορεί να αναπαρασταθεί και να επεξεργαστεί με έναν ηλεκτρονικό υπολογιστή. Τα δεδομένα μιας εταιρείας μπορεί να είναι λειτουργικά και συναλλακτικά δεδομένα, όπως πωλήσεις, κόστη, μισθοί κτλ, μη λειτουργικά δεδομένα, όπως δεδομένα προβλέψεων, μακρο-οικονομικά δεδομένα κτλ, και meta-δεδομένα, δεδομένα δηλαδή για τα ίδια τα δεδομένα (ορισμοί δεδομένων, λογικός σχεδιασμός βάσεων δεδομένων κτλ)[11].

Ο όρος Πληροφορία από την άλλη σημαίνει, τα πρότυπα και τις σχέσεις ανάμεσα στα δεδομένα. Για παράδειγμα πληροφορία, από ένα σύνολο δεδομένων πωλήσεων, είναι ποιο προϊόν πουλάει περισσότερο και πότε. Η έννοια της γνώσης, έχει άμεση σχέση με της πληροφορίας, αφού η γνώση προκύπτει από την καλύτερη ανάλυση της πληροφορίας. Η γνώση αποκτιέται με τις πληροφορίες που έχουμε σε ένα συγκεκριμένο χρονικό διάστημα και αφορά την εύρεση προτύπων και σχέσεων διαμέσου του χρόνου.

Η εξόρυξη δεδομένων είναι ένα από τα σύνολα πυρήνα των τεχνολογιών που βοηθά τους οργανισμούς να χρησιμοποιούν τεχνικές ανάλυσης προβλέψεων για την πρόβλεψη μελλοντικών αποτελεσμάτων, να βρουν νέες ευκαιρίες και να βελτιώσουν τις επιδόσεις των επιχειρήσεων. Ανάλυση Προβλέψεων είναι μια επαναληπτική διαδικασία στην οποία οι επιχειρήσεις μπορούν να:

- επιλέξουν, να εξερευνήσουν και να διαμορφώσουν μεγάλο όγκο δεδομένων.
- προσδιορίσουν τις τάσεις και τις σχέσεις μεταξύ των βασικών μεταβλητών.
- αναπτύξουν μοντέλα.
- αξιολογήσουν τα πλεονεκτήματα των διαφόρων μαθημάτων της δράσης.

Ως εξόρυξη δεδομένων ορίζουμε την ανακάλυψη γνώσης από βάσεις δεδομένων. Πρόκειται ουσιαστικά για την εξαγωγή ενδιαφέρουσας πληροφορίας ή προτύπων από μεγάλες βάσεις

[11] Edelstein, H., A. Introduction to data mining and knowledge discovery (3rd ed). Potomac, MD: Two Crows Corp. 1999

δεδομένων. Η εν λόγω διαδικασία αφορά πληροφορία μη τετριμμένη προηγούμενα άγνωστη και πιθανά χρήσιμη.

Προκειμένου να γίνει πιο σαφής η έννοια της εξόρυξης δεδομένων κρίνεται σκόπιμο στο σημείο αυτό να αναφέρουμε τι δεν είναι εξόρυξη δεδομένων. Πιο συγκεκριμένα, εξόρυξη δεδομένων δεν είναι η επεξεργασία ερωτημάτων ούτε τα στατιστικά προγράμματα μικρής κλίμακας.

Οι επιχειρήσεις, οι επιστήμονες και οι κυβερνήσεις έχουν χρησιμοποιήσει διαδικασίες εξόρυξης δεδομένων για πολλά χρόνια προκειμένου να μετατρέψουν τα δεδομένα σε προγνωστικά μοντέλα. Η διαδικασία εξόρυξης δεδομένων μπορεί να εφαρμοστεί σε μια ποικιλία θεμάτων, από τους πελάτες κάθε κλάδου για τις επιχειρήσεις, - την κατάτμηση και τη στόχευση για την ανίχνευση της απάτης και της βαθμολόγησης του πιστωτικού κινδύνου από επιχειρήσεις ή το κράτος, μέχρι για τον εντοπισμό δυσμενών επιπτώσεων του φαρμάκου κατά τη διάρκεια των κλινικών δοκιμών στην περίπτωση της ιατρικής επιστήμης.

«Πολλοί οργανισμοί χρησιμοποιούν τεχνικές εξόρυξης δεδομένων για τους πελάτες του τμήματος προκειμένου με βάση τη συμπεριφορά ή τη δημογραφία να κατανοήσουν ποια προϊόντα ή υπηρεσίες θα θέλουν ή έχουν ανάγκη στο μέλλον," δήλωσε ο Patel. "Μόλις έχετε ταυτοποιηθεί σωστά τα τμήματα, μπορείτε να δημιουργήσετε μοντέλα πρόβλεψης σχετικά με το ποιοι πελάτες είναι πιθανό να ανταποκριθούν". Η εξόρυξη δεδομένων ξεπερνά τα συστήματα που βασίζονται σε κανόνες για την ανίχνευση της απάτης, ακόμη και ως απατεώνες γίνονται πιο πολύπλοκα στην τακτική τους. "Τα μοντέλα μπορούν να κατασκευαστούν σε παραπομπή δεδομένα από μια ποικιλία πηγών, συσχετίζοντας μη προφανείς μεταβλητές με γνωστά δόλια χαρακτηριστικά για να εντοπίσουν νέες μορφές απάτης," ανέφερε ο Patel[12].

Γενικά, η Εξόρυξη Δεδομένων αναφέρεται στις διεργασίες για την ανάλυση δεδομένων από διαφορετικές πλευρές και η αποκόμιση χρήσιμων πληροφοριών από αυτά, πληροφορίες που μπορούν να χρησιμεύσουν στην πρόβλεψη μελλοντικών καταστάσεων και θα μπορέσουν να βοηθήσουν στη λήψη σωστών αποφάσεων. Ουσιαστικά η εξόρυξη δεδομένων αποτελείται από ένα πλήθος εργαλείων και μεθόδων για την ανάλυση των δεδομένων και επιτρέπει την ανάλυση αυτή από πολλές διαφορετικές γωνίες.

[12] D. Agrawal, P. Bernstein, E. Bertino, S. Davidson, U. Dayal, M. Franklin, J. Gehrke, L. Haas, A. Halevy, J. Han, H. V. Jagadish, A. Labrinidis, S. Madden, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, K. Ross, C. Shahabi, D. Suciu, S. Vaithyanathan, and J. Widom, Challenges and opportunities with big data» February 2012

Εξόρυξη δεδομένων είναι η εξεύρεση μιας (ενδιαφέρουσας, αυτονόητης, μη προφανής και πιθανόν χρήσιμης) πληροφορίας ή προτύπων από μεγάλες βάσεις δεδομένων με χρήση αλγορίθμων ομαδοποίησης ή κατηγοριοποίησης και των αρχών της στατιστικής, της τεχνητής νοημοσύνης, της μηχανικής μάθησης και των συστημάτων βάσεων δεδομένων. Στόχος της εξόρυξης δεδομένων είναι η πληροφορία που θα εξαχθεί και τα πρότυπα που θα προκύψουν να έχουν δομή κατανοητή προς τον άνθρωπο έτσι ώστε να τον βοηθήσουν να πάρει τις κατάλληλες αποφάσεις.

### 1.2.3. ΚΑΤΗΓΟΡΙΕΣ ΜΕΘΟΔΩΝ

Όταν εφαρμόζουμε μια τεχνολογία Εξόρυξης Δεδομένων, αυτό που ουσιαστικά κάνουμε είναι να χρησιμοποιούμε μια σειρά από στατιστικές μεθόδους ή μεθόδους τεχνητής νοημοσύνης με σκοπό τον προσδιορισμό πιθανών αρχέτυπων (patterns) και συσχετίσεων μεταξύ των δεδομένων. Μερικές από τις χρησιμοποιούμενες μεθόδους είναι οι Κανόνες Συσχετισμού (Association rules), η Ανάλυση Αλληλουχίας (Sequence analysis), η Ταξινόμηση (Classification), η Ομαδοποίηση (Clustering) και η Πρόβλεψη (Forecasting).

Πιο συγκεκριμένα, η εξόρυξη δεδομένων περιλαμβάνει κάποιες από τις ακόλουθες κατηγορίες μεθόδων (Fayyad, Usama (1996)):

**Ανίχνευση ανωμαλιών (Anomaly detection)** - Ο προσδιορισμός ασυνήθιστων εγγραφών δεδομένων, που μπορεί να παρουσιάζουν κάποιο ενδιαφέρον ή λάθη στα δεδομένα που απαιτούν περαιτέρω έρευνα.

**Κανόνες συσχέτισης (Μοντέλο αλληλεξάρτησης) -- (Association rules (interdependence Model))-** Αναζητήσεις για σχέσεις μεταξύ των μεταβλητών. Για παράδειγμα, ένα σούπερ μάρκετ μπορεί να συλλέξει δεδομένα που αφορούν τις αγοραστικές συνήθειες των καταναλωτών της εκάστοτε περιοχής. Χρησιμοποιώντας τους κανόνες συσχέτισης, το σούπερ μάρκετ μπορεί να υπολογίσει ποια προϊόντα αγοράζονται πιο συχνά μόνα ή συνδυαστικά με κάποια άλλα προϊόντα, τα οποία ενδεχομένως να βρίσκονται στο ίδιο ράφι.

**Συσταδοποίηση – (Clustering)** είναι η διαδικασία ανακάλυψης ομάδων και δομών στα δεδομένα που είναι "παρόμοια" κατά κάποιο τρόπο, χωρίς να χρησιμοποιούνται γνωστές δομές στα δεδομένα.

**Κατηγοριοποίηση – (Classification)** είναι η διαδικασία γενίκευσης γνωστών δομών για την εφαρμογή τους πάνω σε νέα δεδομένα. Παραδείγματος χάριν, ένα πρόγραμμα ηλεκτρονικού ταχυδρομείου ενδέχεται να προσπαθήσει να χαρακτηρίσει ένα μήνυμα ηλεκτρονικού ταχυδρομείου ως νόμιμο ή ανεπιθύμητη αλληλογραφία.

**Παλινδρόμηση (στατιστική) – (Regression)** - Προσπαθεί να βρει μία συνάρτηση που μοντελοποιεί τα δεδομένα με το λιγότερο λάθος.

Προκειμένου να ολοκληρωθεί η διαδικασία της ανακάλυψης γνώσης από δεδομένα απαιτείται η επικύρωση των προτύπων που εξήχθησαν από τους αλγορίθμους της εξόρυξης δεδομένων που απευθύνονται σε ευρύτερο σύνολο δεδομένων. Δεν είναι όλα τα πρότυπα που βρέθηκαν απαραίτητα έγκυρα. Για να ξεπεραστεί το πρόβλημα της υπερπροσαρμογής (overfitting), της υπερδιόγκωσης των αλγορίθμων, στην εκτίμηση χρησιμοποιείται ένα δοκιμαστικό σύνολο δεδομένων στο οποίο δεν έχουν εφαρμοστεί οι αλγόριθμοι της εξόρυξης δεδομένων. Τα πρότυπα, που έχουν προκύψει, εφαρμόζονται σε αυτό το δοκιμαστικό σύνολο και το προκύπτον αποτέλεσμα συγκρίνεται με το επιθυμητό. Για παράδειγμα, ένας αλγόριθμος της εξόρυξης δεδομένων που ξεχωρίζει τα ανεπιθύμητα μηνύματα με τα "επιθυμητά" θα εφαρμοζόταν σε ένα σύνολο εκπαίδευσης από δείγματα ηλεκτρονικών μηνυμάτων. Μόλις εφαρμοζόταν, τα εξαχθείσα πρότυπα θα εφαρμόζονταν στο δοκιμαστικό σύνολο μηνυμάτων στο οποίο δεν είχε εφαρμοστεί πριν. Η ευστοχία αυτών των προτύπων μπορεί τώρα να μετρηθεί από τα πόσα μηνύματα έχουν καταταχθεί-ταξινομηθεί σωστά. Ένας αριθμός από στατιστικές μεθόδους μπορεί να χρησιμοποιηθεί για την αξιολόγηση του αλγορίθμου, όπως το ROC curves.

Αν τα πρότυπα δεν ανταποκρίνονται με τα επιθυμητά κριτήρια, τότε είναι απαραίτητο να εκτιμηθεί ξανά και να αλλαχθεί η προ-επεξεργασία και η εξόρυξη δεδομένων. Στην αντίθετη περίπτωση που ανταποκρίνονται με τα επιθυμητά κριτήρια, το τελικό στάδιο είναι να ερμηνευτούν τα πρότυπα και να τα μετατρέψουμε σε γνώση.

Η εξόρυξη δεδομένων αποτελείται από διάφορα βήματα για την ανακάλυψη κανόνων, σχέσεων και προτύπων στα δεδομένα που επεξεργάζεται. Το πρώτο βήμα είναι η περιγραφή των

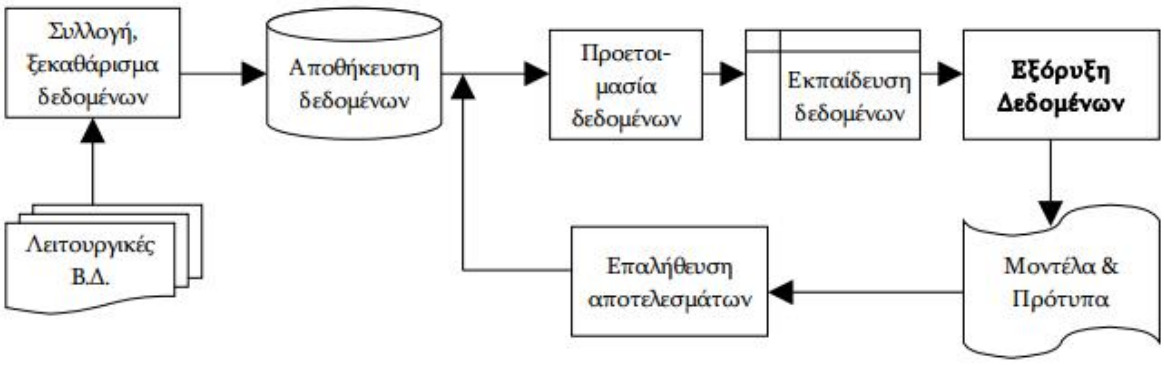
δεδομένων, η συλλογή περιληπτικών πληροφοριών και στατιστικών στοιχείων (μέσες τιμές και αποκλίσεις), η δημιουργία γραφημάτων και η εύρεση πιθανών σχέσεων ανάμεσα στις μεταβλητές (ειδικά σε τιμές που εμφανίζονται μαζί συχνά).

Στη συνέχεια κάποιος μπορεί να δημιουργήσει ένα μοντέλο πρόβλεψης από τις σχέσεις που προέβλεψε και να το ελέγξει για το κατά πόσο ισχύει σε θεωρητικό επίπεδο. Το τελευταίο βήμα είναι να αναλυθεί εμπειρικά πλέον το μοντέλο, ώστε να βρεθεί κατά πόσο ισχύει στον πραγματικό κόσμο και με βάση αυτό να γίνει προγραμματισμός για το μέλλον.

Η εξόρυξη δεδομένων είναι βέβαια μία μεθοδολογία και όχι το μαγικό ραβδί. Η εξόρυξη δεδομένων είναι περισσότερο ένα βοηθητικό εργαλείο για την ανακάλυψη προτύπων και γνώσεων, αλλά δε μπορεί να εγγυηθεί και την αξία αυτών των προτύπων[13][14].

Τα πρότυπα και η γνώση που ανακαλύπτονται πρέπει να επαληθευτούν στον πραγματικό κόσμο. Για παράδειγμα, η εξόρυξη δεδομένων μπορεί να βγάλει σα συμπέρασμα ότι οι άντρες ηλικίας 25-30 ετών που ακούνε ένα συγκεκριμένο ραδιοφωνικό σταθμό, είναι πιθανοί αγοραστές ενός προϊόντος, που επιθυμούμε να πουλήσουμε. Μπορούμε να εκμεταλλευτούμε αυτή την πληροφορία, αγοράζοντας για παράδειγμα ραδιοφωνικό χρόνο για διαφημίσεις στο συγκεκριμένο σταθμό. Αυτό που δεν πρέπει να γίνει όμως είναι να υποθέσουμε ότι οι παράγοντες που περιγράφουμε παραπάνω θα είναι αυτοί που θα τους κάνουν να αγοράσουν το προϊόν.

Σχηματικά η διαδικασία της Εξόρυξης Δεδομένων παρουσιάζεται στο παρακάτω σχήμα:




Εικόνα 1: Διαδικασία εξόρυξης δεδομένων

[13] Han, J., Kamber, M. Data mining: Concepts and Techniques. New York: Morgan-Kaufman. 2000
[14] Weiss, S. M., & Indurkhya, N. Predictive data mining: A practical guide. New York: Morgan-Kaufman. 1997

Η ποιότητα των αποτελεσμάτων της εξόρυξης δεδομένων μπορεί να επηρεαστεί αρνητικά ή θετικά, από τιμές που δεν είναι συνηθισμένες στα δεδομένα μας, χαρακτηριστικά ή μεταβλητές που να εξαρτώνται η μία από την άλλη (πχ ημερομηνία γέννησης με ηλικία), ο τρόπος της κωδικοποίησης, οι στήλες που περιλήφθηκαν στην ανάλυση και αυτές που δεν χρησιμοποιήθηκαν. Οι διάφοροι αλγόριθμοι έχουν διαφορετική ευαισθησία στην ποιότητα των δεδομένων και η εξόρυξη δεδομένων δε μπορεί να παράγει αυτόματα λύσεις χωρίς την καθοδήγηση ενός ειδικού.

Ανάλογα με τις απαιτήσεις κάθε επιχείρησης ή κάθε περίπτωσης, χρησιμοποιείται και η κατάλληλη τεχνική. Η επιλογή της είναι καθοριστική τόσο για την ακρίβεια, όσο και για το χρόνο που θα χρειαστεί για να ολοκληρωθεί το μοντέλο.

Τα αποτελέσματα της Εξόρυξης Δεδομένων, παρουσιάζονται με διάφορους τρόπους. Μερικοί από αυτούς είναι οι παρακάτω – από τον ευκολότερο σε κατανόηση στο δυσκολότερο και από το λιγότερο καλό σε προβλέψεις στον καλύτερο:

- **Δέντρα Αποφάσεων.** Παρουσιάζονται οι πληροφορίες σε δεντρικές μορφές, που είναι πολύ εύκολο να διαβαστούν αλλά δεν παρέχουν και τόσο εξειδικευμένες πληροφορίες για το μέλλον.
- **Κανόνες.** Παρουσιάζονται τα αποτελέσματα σε μορφή κανόνων του τύπου «αν… τότε…».
- **Scorecards.** Παρουσιάζονται τα αποτελέσματα με τη μορφή συναρτήσεων, που εξαρτώνται από πολλούς παράγοντες. Καθώς αυξομειώνονται οι παράγοντες αυτοί αλλάζει και η τιμή της συνάρτησης.
- **Νευρωνικά Δίκτυα.** Τα αποτελέσματα δίνονται σε μη-γραμμικά δίκτυα που μοιάζουν με βιολογικά δίκτυα. Τα νευρωνικά δίκτυα προσφέρονται για δεδομένα που έχουν πολύ θόρυβο ή τις ελλιπείς τιμές.

Μέσα από τους αλγορίθμους εξόρυξης δεδομένων υπάρχει η δυνατότητα παραγωγής των ακόλουθων τεσσάρων τύπων σχέσεων:

- **Κλάσεις.** Τα δεδομένα είναι αποθηκευμένα σε προκαθορισμένες ομάδες. Για παράδειγμα κάποιο εστιατόριο θα ήθελε ανάλυση για το πότε επισκέπτονται οι πελάτες και τι

παραγγέλνουν. Έχουμε δηλαδή τις κλάσεις ανά ποιες ώρες έρχονται οι πελάτες και τι παραγγέλνουν.

- **Συστάδες.** Τα δεδομένα, έχουν ένα σύνολο χαρακτηριστικών και ανάλογα με το πόσο όμοια είναι τα κατατάσσουμε σε clusters – συστάδες.
- **Συσχετίσεις.** Οι συσχετίσεις είναι «κανόνες» ανάμεσα στα δεδομένα.
- **Σειριακά Πρότυπα.** Τα δεδομένα εξορύσσονται σε αναζήτηση προτύπων και μελλοντικών τάσεων.

### 1.2.4. Η ΤΥΠΟΠΟΙΗΜΕΝΗ ΔΙΑΔΙΚΑΣΙΑ CRISP

Για τη διεξαγωγή συστηματικής ανάλυσης των δεδομένων εξόρυξης, συνήθως ακολουθείται μία γενική διαδικασία. Υπάρχουν μερικές τυποποιημένες διαδικασίες, η πιο συνηθισμένη από αυτές είναι η CRISP (Cross-Industry Standard Process for Data Mining). Η CRISP είναι μια βιομηχανοποιημένη διαδικασία-πρότυπο, που αποτελείται από μια ακολουθία βημάτων που συνήθως εμπλέκονται σε μια μελέτη εξόρυξης δεδομένων[15].

Το μοντέλο αυτό αποτελείται από έξι στάδια που προορίζονται ως μια κυκλική διαδικασία. Πιο συγκεκριμένα[16]:

**Κατανόηση του προβλήματος (business understanding):** Περιλαμβάνει τον καθορισμό των επιχειρηματικών στόχων, αξιολόγηση της τρέχουσας κατάστασης, ίδρυση στόχου για την μέθοδο εξόρυξης, καθώς και την ανάπτυξη ενός σχεδίου (project).

**Κατανόηση των δεδομένων (data understanding):** Μόλις καθοριστούν οι επιχειρηματικοί στόχοι και το σχέδιο του έργου, με την κατανόηση των δεδομένων υπολογίζονται οι απαιτήσεις δεδομένων. Αυτό το βήμα μπορεί να περιλαμβάνει την αρχική συλλογή δεδομένων, περιγραφή, διερεύνηση δεδομένων, και την επαλήθευση της ποιότητας των δεδομένων. Δεδομένα εξερεύνησης, όπως η προβολή συνοπτικά των στατιστικών στοιχείων (η οποία περιλαμβάνει την οπτική απεικόνιση των μεταβλητών κατηγοριοποίησης) μπορεί να συμβεί στο τέλος αυτής της

[15] Advanced Data Mining Techniques Olson, D.L.; Delen, D. 2008 ISBN:978-3-540-76916-3
[16] Advanced Data Mining Techniques Olson, D.L.; Delen, D. 2008 ISBN:978-3- 540-76916-3

φάσης. Μοντέλα, όπως το σύμπλεγμα ανάλυση μπορούν επίσης να εφαρμοστούν κατά τη διάρκεια αυτής της φάσης, με την πρόθεση να προσδιοριστούν τα μοτίβα στα δεδομένα.

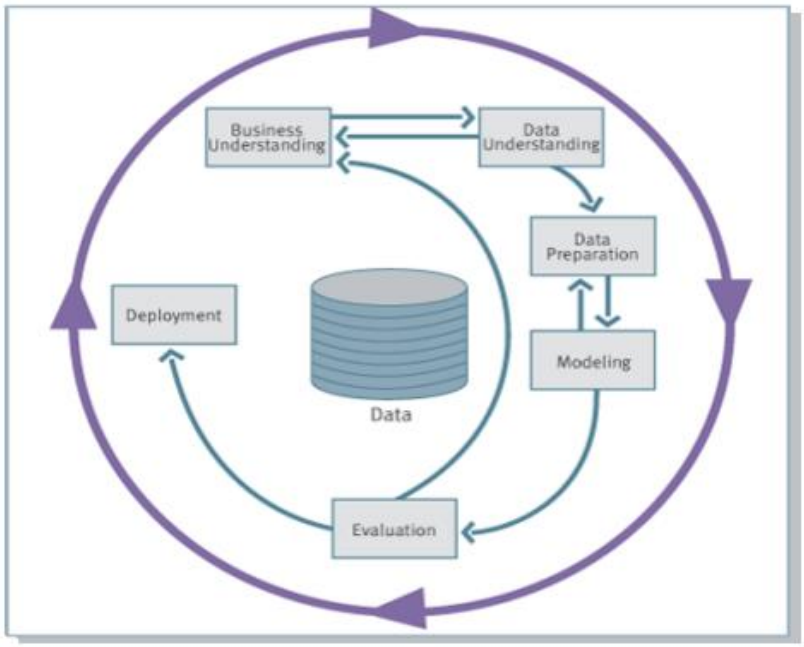
**Προετοιμασία των δεδομένων (data preparation):** Μόλις προσδιοριστούν τα δεδομένα και οι πόροι που είναι διαθέσιμοι, αυτά πρέπει να επιλέγουν, να καθαριστούν, να χτιστούν με βάση την επιθυμητή μορφή, και να μορφοποιηθούν. Ο καθαρισμός των δεδομένων και η μετατροπή των δεδομένων στο πλαίσιο της προετοιμασίας της μοντελοποίησης δεδομένων πρέπει να συμβεί σε αυτή την φάση. Δεδομένα εξερεύνησης σε μεγαλύτερο βάθος μπορεί να εφαρμόζονται κατά τη διάρκεια αυτής της φάσης, και μπορούν να χρησιμοποιηθούν επιπλέον μοντέλα, και πάλι παρέχοντας την ευκαιρία να δούμε τα πρότυπα που βασίζονται στην κατανόηση των επιχειρήσεων.

**Μοντελοποίηση δεδομένων (modeling)** εργαλείων λογισμικού εξόρυξης, όπως οπτικοποίηση (εκτυπώνοντας δεδομένα και τη θέσπιση σχέσεων) και ανάλυση διασποράς (για τον εντοπισμό μεταβλητών που πηγαίνουν καλά μαζί) είναι χρήσιμη για την αρχική ανάλυση. Εργαλεία όπως η γενικευμένη επαγωγή κανόνα μπορεί να αναπτύξει αρχικούς κανόνες συσχέτισης. Μόλις επιτευχθεί μεγαλύτερη κατανόηση των δεδομένων (συχνά μέσω μοτίβου αναγνώρισης, που προκλήθηκε από την προβολή της παραγωγής μοντέλο), μπορούν να εφαρμοστούν πιο λεπτομερή μοντέλα κατάλληλα για το είδος των δεδομένων. Η διαίρεση των δεδομένων σε σύνολα εκπαίδευσης (training) και test sets είναι επίσης απαραίτητη για τη μοντελοποίηση.

**Τα αποτελέσματα του Μοντέλου Αξιολόγησης (evaluation)** θα πρέπει να αξιολογηθούν στο πλαίσιο των στόχων που έχουν διατυπωθεί στο πρώτο στάδιο (κατανόηση του προβλήματος). Αυτό θα οδηγήσει στην ταυτοποίηση άλλων αναγκών (όπως για παράδειγμα μέσω της αναγνώρισης προτύπων), που συχνά επανέρχεται στην προηγούμενη φάση CRISP-DM. Επιτυγχάνοντας την κατανόηση του προβλήματος επιτυγχάνεται ταυτόχρονα μια επαναληπτική διαδικασία εξόρυξης δεδομένων, όπου τα αποτελέσματα των διαφόρων εργαλείων όπως οπτικοποίηση, στατιστικές, και τεχνητή νοημοσύνη δείχνουν στον χρήστη νέες σχέσεις που παρέχουν μια βαθύτερη κατανόηση των οργανωτικών εργασιών.

**Ανάπτυξη:** η εξόρυξη δεδομένων μπορεί να χρησιμοποιηθεί τόσο για την επαλήθευση προηγουμένων υποθέσεων, ή για την απόκτηση της γνώσης (εντοπισμός των απρόβλεπτων και χρήσιμων σχέσεων). Μέσω της απόκτησης της γνώσης στις προηγούμενες φάσεις της

διαδικασίας CRISP-DM , μοντέλα ήχου μπορούν να αποκτηθούν που μπορεί στη συνέχεια να εφαρμοστούν σε επιχειρηματικές δραστηριότητες για πολλούς σκοπούς, συμπεριλαμβανομένης της πρόβλεψης ή προσδιορισμού των βασικών καταστάσεων. Αυτά τα μοντέλα θα πρέπει να παρακολουθούνται για αλλαγές στις συνθήκες λειτουργίας, διότι αυτό, που θα μπορούσε να ισχύει σήμερα μπορεί να μην ισχύει σε ένα χρόνο από σήμερα. Εάν υπάρξουν σημαντικές αλλαγές, το μοντέλο θα πρέπει να επαναληφθεί. Είναι επίσης σκόπιμο να καταγράφονται τα αποτελέσματα των έργων (projects) δεδομένων εξόρυξης, ώστε τεκμηριωμένα αποδεικτικά στοιχεία να είναι διαθέσιμα για μελλοντικές μελέτες.



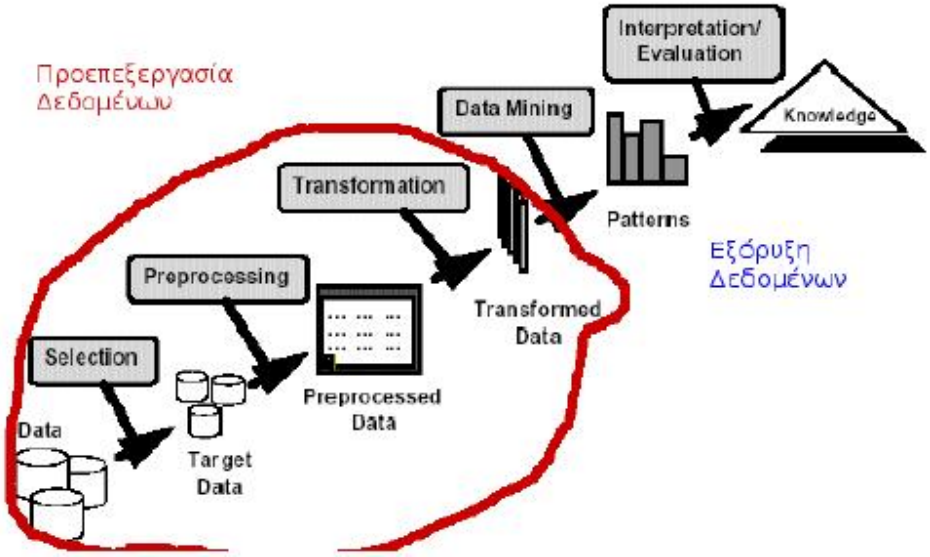

### 1.2.5. ΣΤΟΧΟΙ & ΧΡΗΣΙΜΟΤΗΤΑ

Στις περισσότερες επιχειρήσεις τα στοιχεία των πελατών συλλέγονται με κάθε συναλλαγή και συσσωρεύονται εξαιρετικά γρήγορα. Το μεγαλύτερο πλεονέκτημα της επιχειρησιακής νοημοσύνης (BI) είναι η δυνατότητά να απεικονίζει αυτά τα στοιχεία με έναν εύκολο και κατανοητό τρόπο για τα στελέχη της επιχείρησης.

Συχνά συγχέεται η δυνατότητα δημιουργίας αναφορών ενός προγράμματος-πελάτη μιας Αποθήκης Δεδομένων με τη μέθοδο του OLAP ή του Data Mining.

Τη σύγχυση επιτείνει ο ισχυρισμός αρκετών κατασκευαστών λογισμικού (software) ότι τα προϊόντα τους είναι λύσεις Επιχειρηματικής Ευφυΐας καθώς χρησιμοποιούν κάποια παρόμοια εργαλεία. Το μέγιστο επιχειρηματικό όφελος επιτυγχάνεται από την διάθεση των μοντέλων που προκύπτουν μέσω της διαδικασίας Data Mining στα σημεία επαφής της επιχείρησης με τον πελάτη, εσωτερικό ή εξωτερικό, (διαδίκτυο, καταστήματα πώλησης, τηλεφωνικό κέντρο, γραπτή επικοινωνία κ.λ.π), οπότε μπορούμε να απαντάμε ερωτήματα της μορφής «τι θα μπορούσαμε να προσφέρουμε στον συγκεκριμένο πελάτη σήμερα για να τον διατηρήσουμε ενεργό στο μέλλον».

Ο πραγματικός στόχος της εξόρυξης δεδομένων είναι η αυτόματη ή ημιαυτόματη ανάλυση μεγάλων ποσοτήτων δεδομένα για την εξαγωγή κάποιου ενδιαφέροντος προτύπου που ήταν άγνωστο μέχρι εκείνη τη στιγμή, όπως ομάδες από εγγραφές δεδομένων (συσταδοποίηση), ασυνήθιστες εγγραφές (anomaly detection) και εξαρτήσεις (κανόνες συσχετίσεων). Αυτά τα πρότυπα ύστερα μπορούν να θεωρηθούν ως μία περιγραφή των δεδομένων εισαγωγής και να χρησιμοποιηθούν για περαιτέρω ανάλυση ή για παράδειγμα στην εκμάθηση μηχανής και στην προγνωστική ανάλυση. Για παράδειγμα, η εξόρυξη δεδομένων θα μπορούσε να προσδιορίσει πολλαπλά σύνολα στα δεδομένα, τα οποία μπορούν να χρησιμοποιηθούν μετά για να εξασφαλίσουν περισσότερο ακριβή αποτελέσματα από ένα σύστημα υποστήριξης αποφάσεων. Παρότι η συλλογή δεδομένων, η προετοιμασία δεδομένων, αλλά και η ερμηνεία των αποτελεσμάτων και εκθέσεων δεν αποτελούν μέρος της εξόρυξης δεδομένων, παρ' όλα αυτά ανήκουν στην ανακάλυψη γνώσης από βάσεις δεδομένων σαν κάποια επιπρόσθετα βήματα. Αναλυτικότερα, τα παραπάνω συνοψίζονται στο σχήμα που ακολουθεί:




Άλλοι σχετικοί όροι της εξόρυξης δεδομένων είναι οι data dredging, data fishing και data snooping, που αναφέρονται στην χρήση μεθόδων της εξόρυξης δεδομένων για να πάρουν δείγματα από μεγαλύτερη συλλογή δεδομένων που είναι (ή μπορεί να είναι) πολύ μικρά για αξιόπιστα στατιστικά συμπεράσματα που έγιναν σχετικά με τη εγκυρότητα των προτύπων που ανακαλύφθηκαν. Αυτές οι μέθοδοι, επίσης, μπορούν να χρησιμοποιηθούν για την δημιουργία νέων υποθέσεων προς εξέταση έναντι μεγαλύτερων συλλογών δεδομένων.

Η χειροκίνητη εξαγωγή προτύπων από δεδομένα συμβαίνει εδώ και αιώνες. Οι πρώτοι μέθοδοι για τον προσδιορισμό προτύπων ήταν αυτοί της θεωρίας Bayes και της ανάλυσης της παλινδρόμησης. Ο πολλαπλασιασμός, η ευρεία διαθεσιμότητα και η εξέλιξη της τεχνολογίας υπολογιστών έχουν αυξήσει τον όγκο των συγκεντρωμένων δεδομένων και την ζήτηση για αποδοτικούς και αποτελεσματικούς χειρισμούς. Καθώς οι συλλογές δεδομένων αυξήθηκαν τόσο σε όγκο όσο και σε πολυπλοκότητα, η χειρωνακτική ανάλυση των δεδομένων έχει αντικατασταθεί από την αυτόματη επεξεργασία δεδομένων. Σε αυτό συνέβαλαν άλλες ανακαλύψεις της επιστήμης των υπολογιστών, όπως τα νευρωνικά δίκτυα, η συσταδοποίηση, οι γενετικοί αλγόριθμοι (1950), τα δέντρα απόφασης (1960) και η μηχανή υποστήριξης διανυσμάτων (1990). Η εξόρυξη δεδομένων είναι η διαδικασία εφαρμογής αυτών των μεθόδων στα δεδομένα με σκοπό την αποκάλυψη άγνωστων προτύπων (Kantardzic, Mehmed, (2003)) σε μεγάλα σύνολα δεδομένων. Αυτό γεφυρώνει το χάσμα της εφαρμοσμένης στατιστικής και της τεχνητής νοημοσύνης (τα οποία συνήθως παρέχουν το μαθηματικό υπόβαθρο) με την διαχείριση

βάσης δεδομένων κάνοντας χρήση του τρόπο με τον οποίο αποθηκεύονται και κατατάσσονται στη βάση δεδομένων για να εκτελέσουν την θεωρία και τους διαθέσιμους αλγορίθμους περισσότερο αποτελεσματικά, επιτρέποντας σε τέτοιες μεθόδους να εφαρμόζονται σε μεγάλα σύνολα δεδομένων.

### 1.2.6. ΕΦΑΡΜΟΓΕΣ

Οι λύσεις Επιχειρηματικής Ευφυΐας καλύπτουν ένα ευρύ φάσμα εφαρμογών που μπορούν να χρησιμοποιηθούν ώστε να αυξηθεί η αποδοτικότητα και αποτελεσματικότητα πολλών εσωτερικών διεργασιών. Για παράδειγμα:

- **Διαχείριση αλυσίδας προμηθειών:** Η ΕΕ παρέχει πλήρη πληροφόρηση σχετικά με τα επίπεδα αποθεμάτων και θέματα υπολογισμού κατά μήκος της αλυσίδας προμηθειών, εξασφαλίζοντας καλύτερη διαχείριση πάνω στις επιδράσεις της, στη ροή εσόδων, δαπανών και ικανοποίησης πελατών.
- **Διαχείριση και έλεγχος για τυχόν απάτες:** Παρέχοντας πρόσβαση σε εξαιρετικά μεγάλα μεγέθη λεπτομερών πληροφοριών, μια ΕΕ λύση καθιστά δυνατό τον εντοπισμό παράνομων ενεργειών αναλύοντας επικοινωνιακά αρχέτυπα (patterns).
- **Διαχείριση ρίσκου:** Όλες οι εταιρείες παίρνουν ρίσκα καθημερινά. Η ΕΕ παρέχει τα μέσα ώστε να εκτιμώνται καλύτερα αυτοί οι κίνδυνοι καθιστώντας εφικτή μέσω της ανάλυσης ιστορικών δεδομένων τη δημιουργία “προφίλ ρίσκου” των πελατών, σύμφωνα με το οποίο και θα αξιολογούνται οι νέοι πελάτες.
- **Διαχείριση προϊόντων:** Πολλοί οργανισμοί επιζητούν να μειώσουν τους επιμέρους χρόνους ανάπτυξης και να εξασφαλίσουν ότι τα προϊόντα τους θα συμβαδίζουν με τις ανάγκες της αγοράς, έτσι ώστε να διατηρήσουν και να επεκτείνουν την αγοραστική βάση και τα κέρδη τους. Η ΕΕ μπορεί να παρέχει γρήγορη και ακριβή ανατροφοδότηση σχετικά με το βαθμό επιτυχίας των αποφάσεων που πάρθηκαν σχετικά με τα προϊόντα.
- **Οικονομικοί έλεγχοι:** Επίσης, η ΕΕ μπορεί να χρησιμοποιηθεί για τη βελτίωση των οικονομικών περιθωρίων και την ελάττωση των δαπανών. Με λεπτομερή πληροφορία

για όλες τις δραστηριότητες της επιχείρησης είναι εφικτός ο καθορισμός των προϊόντων, των πελατών και των γεωγραφικών περιοχών που είναι οι πιο προσοδοφόρες για αυτήν.

Τώρα, περισσότερο από ποτέ, οι οργανισμοί χρειάζονται την Επιχειρηματική Ευφυΐα, προκειμένου να κατανοήσουν πώς μπορούν να ανταποκριθούν και να προσαρμοστούν γρήγορα και με τον καλύτερο δυνατό τρόπο στις υπάρχουσες συνθήκες και στην παρατηρούμενη δυναμική του περιβάλλοντος.

Τα τελευταία χρόνια, η εξόρυξη δεδομένων χρησιμοποιείται ευρέως στους τομείς της ιατρικής, όπως η βιοϊατρική, το DNA, η γενετική και η φαρμακευτική. Στον τομέα της γενετικής, ο σκοπός είναι να κατανοήσουμε την χαρτογράφηση της σχέσης μεταξύ της μεταβολής των ακολουθιών του ανθρώπινου DNA και την προδιάθεση στην αρρώστια. Η εξόρυξη δεδομένων είναι ένα σημαντικό εργαλείο που μπορεί να βοηθήσει στην βελτίωση της διάγνωσης, της πρόληψης και της θεραπείας των ασθενειών.

Ο χώρος της ιατρικής παράγει έναν συνεχώς αυξανόμενο όγκο δεδομένων και συνεπώς όλο και περισσότερη κρυμμένη πληροφορία υπάρχει σε αυτά. Είναι κοινά αποδεκτό από όλους τους φορείς υγείας ότι η αποτελεσματική διαχείριση των ιατρικών δεδομένων παίζει καθοριστικό ρόλο στην παροχή υγείας υψηλού επιπέδου[17]. Η ανάλυση των ιατρικών δεδομένων μπορεί να συνεισφέρει στη βελτίωση της αποτελεσματικότητας με ταυτόχρονη μείωση του χρόνου και του κόστους. Η αποκτώμενη γνώση με τις μεθόδους εξόρυξης δεδομένων μπορεί να αξιοποιηθεί από την ιατρική έρευνα, τόσο στο επίπεδο της διάγνωσης όσο και της θεραπείας[18].

Τα ιατρικά δεδομένα είναι από τη φύση τους ετερογενή, περιλαμβάνουν αποτελέσματα απεικονιστικά, γραφήματα, κείμενα από την κλινική εξέταση, εργαστηριακές μετρήσεις σε αριθμούς καθώς και δεδομένα σε άλλες μορφές. Έτσι, η εξόρυξη της κρυμμένης γνώσης από αυτά πρέπει να γίνει από συνδυασμό εικόνων, σχημάτων, κειμένων, αριθμών, το οποίο είναι δυσκολότερο από τις κλασσικές περιπτώσεις επεξεργασίας δεδομένων σε αριθμούς και κατηγορίες. Οι σύγχρονες τεχνολογίες εξόρυξης δεδομένων επιτρέπουν πλέον τη διαχείριση της ετερογενούς φύσης των ιατρικών δεδομένων. Πρώτιστα, η δυνατότητα επεξεργασίας της φυσικής γλώσσας και οι τεχνικές εξόρυξης δεδομένων από κείμενα επιτρέπουν την εξαγωγή

[17] Krzysztof j. Giosa, G. William Mooree, Uniqueness of medical data mining, Joornal of Artificial intelligence in medicine, 2002.

[18] Sackett, D. L., Rosenberg, W. M., Gray, J. A., Haynes, R B., Richardson, W. S., Evidence based medicine: what it is and what it isn't. BMJ, 312 (7023), 71-2, 2004.

πληροφορίας και γνώσης από τις ιατρικές σημειώσεις και τις κλινικές εξετάσεις[19][20]. Ιατρική οντολογία και ορολογία μπορούν να ανιχνευθούν με τη χρήση μεθόδων εξόρυξης δεδομένων του παγκόσμιου ιστού και με τεχνικές εκμάθησης οντολογίας[21][22].

Εξαιτίας της αύξησης των βιοϊατρικών ερευνών, η μεγάλη κλίμακα γονιδιακών προτύπων και λειτουργιών πρέπει να εξετασθεί. Τα εργαλεία της εξόρυξης δεδομένων μπορούν να βοηθήσουν σε μεγάλο βαθμό για να μελετήσουμε την σύσταση του DNA και να βρούμε ποικίλα πρότυπα και λειτουργίες αυτού.

Ένας από τους κύριους στόχους που σχετίζεται με την ανάλυση δεδομένων του DNA είναι η σύγκριση ποικίλων ακολουθιών και η αναζήτηση ομοιοτήτων μεταξύ των δεδομένων του DNA. Η σύγκριση κυρίως περιλαμβάνει την γονιδιακή ακολουθία υγιών και βλαβερών ιστών για να βρει την διαφορά ανάμεσα σε αυτούς τους δύο τύπους. Αυτό μπορεί να επιτευχθεί ανακτώντας τις τάξεις υγιών αλλά και βλαβερών γονιδιακών ακολουθιών και μετά βρίσκοντας τις συχνά εμφανιζόμενες μορφές των δύο τάξεων. Αυτή η ανάλυση βοηθάει στο να βρίσκουμε τις ομοιότητες και τις διαφορές στις γενετικές ακολουθίες.

Στην βιοϊατρική , ερευνάται αν οι περισσότερες ασθένειες προκαλούνται από ένα συνδυασμό των γονιδίων. Η μέθοδος της συσχέτισης χρησιμοποιείται για να καθορίσει την συνύπαρξη ομάδων των γονιδίων και επίσης μπορούμε να εξετάσουμε την αλληλεπίδραση και την σχέση μεταξύ των γονιδίων.

Τα εργαλεία της οπτικοποίησης παίζουν επίσης ένα σημαντικό ρόλο στην εξόρυξη δεδομένων στην βιοϊατρική. Τα εργαλεία αυτά μπορούν να παρουσιάσουν πολύπλοκες δομές γονιδίων σε γράφους, δένδρα και αλυσίδες. Η οπτική παρουσίαση βοηθάει στην καλύτερη κατανόηση αυτών των δομών για ανακάλυψη γνώσης και εξερεύνηση των δεδομένων.

---

[19] Savova, G. K., Ogren, P. V., Duffy, P. H., Buntrock, J. D., Chute, C. G., Mayo clinic NLP system for patient smoking status identification. J Am Med Inform Assoc, 15(1), 25-8, 2008.

[20] Cimiano, A., Hoto, A., Staab, S., Learning concept hierarchies from text corpora using formal concept analysis. Journal of Artificial Intelligence Research, 24, 305-339, 2005.

[21] Cimiano, A., Hoto, A., Staab, S., Learning concept hierarchies from text corpora using formal concept analysis. Journal of Artificial Intelligence Research, 24, 305-339, 2005.

[22] Montani, S., Portinale, L., Leonardi, G., Bellazzi, R., Case-based retrieval to support the treatment of end stage renal failure patients. Artif Intell Med, 37(1), 31-42, 2006

Υπάρχουν διάφοροι συνδυασμοί γονιδίων που συμβάλλουν στις ασθένειες, αλλά αυτά τα γονίδια ενεργοποιούνται σε διαφορετικά επίπεδα. Η ανάλυση μονοπατιού (path analysis) χρησιμοποιείται για να συνδέει διαφορετικά γονίδια με διαφορετικά στάδια κατά την εξέλιξη της ασθένειας. Η ανάλυση μονοπατιού διαδραματίζει ένα σπουδαίο ρόλο στην γενετική.

Ένας δεύτερος τομέας στον οποίο η εξόρυξη δεδομένων έχει έντονα αγκαλιαστεί είναι η βιο-πληροφορική. Βιο-πληροφορική είναι η επιστήμη που αφορά την διαχείριση, την εξόρυξη και την ερμηνεία των βιολογικών ακολουθιών και δομών.

Το Structural Genome Initiative έχει στόχο την καταγραφή της δομής-λειτουργίας πληροφοριών των πρωτεϊνών. Η πρόοδος στον τομέα των τεχνολογιών, όπως οι μικροσυστοιχίες, είχε ως αποτέλεσμα την δημιουργία των υποτομέων της γονιδιωματικής και πρωτεϊνωματικής. Αυτά τα πεδία αφορούν στην μελέτη των γονιδίων, των πρωτεϊνών και του κυκλώματος στο εσωτερικό του κυττάρου που ρυθμίζει την γονιδιακή έκφραση. Πολλά δεδομένα δημιουργούνται, δεδομένα που πρέπει να εξορυχτούν, αν η ανθρωπότητα θελήσει κάποτε να αντιληφθεί τα μυστήρια των κυττάρων[23].

Κατά τα τελευταία χρόνια, τεράστια πρόοδος έχει επιτευχθεί, αλλά εξακολουθούν να υπάρχουν μια σειρά από θεμελιώδη προβλήματα στην βιο-πληροφορική, όπως η δυσκολία στην πρόβλεψη πρωτεϊνικών δομών και στην εξεύρεση γονιδίων. Η εξόρυξη δεδομένων θα διαδραματίσει θεμελιώδη ρόλο στην κατανόηση της γονιδιακής έκφρασης, στην ανάπτυξη φαρμάκων και επίλυση άλλων προβλημάτων στον τομέα της γονιδιωματικής και πρωτεϊνωματικής. Επιπλέον, η εξόρυξη κειμένου θα είναι σημαντική καθώς θα φιλτράρει τις γνώσεις από την αυξανόμενη προσφορά της βιβλιογραφίας σχετικά με τη βιο-πληροφορική[24].

Άλλος τομέας που εφαρμόζεται η εξόρυξη δεδομένων είναι η οικονομία. Τα οικονομικά δεδομένα κυρίως συλλέγονται από τράπεζες και από άλλους οικονομικούς οργανισμούς. Τα δεδομένα αυτά συνήθως είναι αξιόπιστα, ολοκληρωμένα και έχουν υψηλή ποιότητα και απαιτούν συστηματική μέθοδο για την ανάλυση αυτών. Η συνεισφορά της εξόρυξης δεδομένων στην επιστήμη της οικονομίας συναντάται στην συλλογή και κατανόηση των δεδομένων, στην

[23] New Trends in Data Mining by J. HUYSMANS, B. BAESENS, D. MARTENS, K. DENYS and J. VANTHIENEN101 (http://www.econ.kuleuven.be/rebel//jaargangen/2001-2010/2005/TEM%202005-4/TEM_4_05_Huysmans.pdf)
[24] New Trends in Data Mining by J. HUYSMANS, B. BAESENS, D. MARTENS, K. DENYS and J. VANTHIENEN101 (http://www.econ.kuleuven.be/rebel//jaargangen/2001-2010/2005/TEM%202005-4/TEM_4_05_Huysmans.pdf )

βελτίωση δεδομένων (data refinement), στην δημιουργία και εκτίμηση ενός μοντέλου και στην ανάπτυξη αυτού. Η σωστή ανάλυση των οικονομικών δεδομένων μας διευκολύνει στο να λαμβάνουμε 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 διάφορα οικονομικά ινστιτούτα, όπως οι τράπεζες, συγκεντρώνονται αρχικά στην αποθήκη δεδομένων (data warehouse). Οι τεχνικές της πολυδιάστατης ανάλυσης δεδομένων χρησιμοποιούνται για την ανάλυση τέτοιων δεδομένων που συλλέγονται στην αποθήκη δεδομένων για τις γενικές ιδιότητές του.

Μία άλλη εφαρμογή της εξόρυξης δεδομένων σχετίζεται με την πρόβλεψη αποπληρωμής δανείου και πολιτικές πίστωσης του πελάτη. Μέθοδοι της εξόρυξης όπως η επιλογή χαρακτηριστικών (feature selection) βοηθάει στην ταυτοποίηση 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, 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 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 τα οικονομικά ινστιτούτα 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. Οι τεχνικές οπτικοποίησης βοηθούν στην παρουσίαση δεδομένων με διαφορετικές μορφές, όπως γράφοι που βασίζονται σε συγκεκριμένα γνωρίσματα. Προβάλλοντας τα δεδομένα από διάφορες οπτικές γωνίες, η 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.

Η τηλεπικοινωνιακή βιομηχανία αναπτύσσεται πολύ γρήγορα όπως και η τεχνολογία. Αυτές τις μέρες οι τηλεπικοινωνιακές υπηρεσίες έχουν επεκταθεί από τοπικές και μεγάλης απόστασης

τηλεπικοινωνίες, στην χρήση φαξ, συσκευές τηλεειδοποίησης, κινητό τηλέφωνο, και ηλεκτρονικό ταχυδρομείο. Εξαιτίας των εξελίξεων στις τηλεπικοινωνιακές τεχνολογίες και για να δουλέψουν αποτελεσματικά αυτές οι τεχνολογίες, οι τεχνικές της εξόρυξης δεδομένων ενσωματώνονται σε αυτές τις τεχνολογίες για να παράγουν αποδοτικά αποτελέσματα. Η εξόρυξη δεδομένων βοηθάει στην διάκριση τηλεπικοινωνιακών προτύπων, καταπολέμησης παράνομων δραστηριοτήτων, και επίσης βοηθάει στην καλύτερη χρήση των πόρων και στη βελτίωση της ποιότητας των υπηρεσιών. Η εξόρυξη δεδομένων βελτιώνει τις τηλεπικοινωνιακές υπηρεσίες με τους εξής τρόπους:

- Τα τηλεπικοινωνιακά δεδομένα που συλλέγονται, περιλαμβάνουν τον τύπο κλήσης, την τοποθεσία του καλούντος και του κληθέντος, τον χρόνο κλήσης, την διάρκεια κλήσης κλπ.
- Η πολυδιάστατη ανάλυση βοηθά στον προσδιορισμό και στην σύγκριση του φορτιού του συστήματος, κίνηση δεδομένων, και κέρδος κλπ.
- Η ανάλυση μπορεί να δείξει διαγράμματα και γράφους των πόρων του συστήματος, του προορισμού κλπ κάνοντας χρήση των εργαλείων οπτικοποίησης της εξόρυξης δεδομένων.
- Τέτοια εργαλεία όπως η συσχετισμένη οπτικοποίηση και η συσταδοποίηση παρέχουν χρήσιμες υπηρεσίες στην ανάλυση των δεδομένων τηλεπικοινωνίας.

Το κυρίως πρόβλημα που αντιμετωπίστηκε από την βιομηχανία τηλεπικοινωνιών είναι οι παράνομες δραστηριότητες. Αυτές οι δραστηριότητες μπορεί να έχουν να κάνουν με σκόπιμες κλήσεις κατά την ώρα αιχμής, περιοδικές κλήσεις κ.α. με αποτέλεσμα να επιδρούν αρνητικά στην επίδοση του δικτύου επικοινωνιών. Μέθοδοι όπως η συσταδοποίηση και η ανάλυση ακραίων τιμών, συνεισφέρει στην ανίχνευση παράνομων προτύπων βελτιώνοντας την αποτελεσματικότητα των υπηρεσιών τηλεπικοινωνίας.

Εκμεταλλευόμενοι τα εργαλεία της εξόρυξης δεδομένων είναι δυνατή η δημιουργία προφίλ των πελατών και ο εντοπισμός βλαβών στο δίκτυο.

Τέλος, η ανάλυση συσχετιζόμενων και ακολουθιακών προτύπων ενθαρρύνει την προώθηση νέων και ποικίλων υπηρεσιών τηλεπικοινωνίας.

Οι εκτεταμένες αλλαγές στην υιοθέτηση και χρησιμοποίηση των νέων τεχνολογιών στις μεγάλες αλλά και στις μικρές επιχειρήσεις έχει ως αποτέλεσμα την συγκέντρωση μεγάλου αριθμού δεδομένων από τις οικονομικές συναλλαγές. Είναι ευθύνη του αναλυτή να αναλύσει αυτές τις συναλλαγές και να εντοπίσει τις απάτες και τα λάθη μέσα σε αυτές. Λόγο των αλλαγών των τάσεων μέσα στην επιχείρηση, είναι δύσκολο να επεξεργαστείς και να αναλύσεις τα δεδομένα με παλαιές μεθόδους. Οι περιορισμοί που εμφανίζουν αυτές οι μέθοδοι μας έχουν οδηγήσει στην εκμετάλλευση των εργαλείων της εξόρυξης για καλύτερα και περισσότερο αξιόπιστα αποτελέσματα.

Τέλος, η εξόρυξη δεδομένων χρησιμοποιείται σήμερα από διάφορες εταιρείες στη σχεδίαση πολιτικών marketing. Χρησιμοποιείται, για παράδειγμα, στη βελτίωση ή προσωποποίηση των Δικτυακών τόπων, στην ανάλυση του προφίλ των πελατών τους. Τους επιτρέπει να ανακαλύψουν τις σχέσεις μεταξύ «εσωτερικών» παραγόντων, όπως η τιμή, η θέση του προϊόντος, ή η ικανότητα του προσωπικού, με «εξωτερικούς», όπως η οικονομία, ο ανταγωνισμός και τα προσωπικά στοιχεία του πελάτη (ηλικία, περιοχή, εισόδημα, μόρφωση κτλ). Μέσω αυτών των σχέσεων μπορεί να υπολογιστεί κατά πόσο είναι ευχαριστημένος ο πελάτης, οι πωλήσεις και τα κέρδη. Επίσης η εξόρυξη δεδομένων βοηθάει τη δημιουργία μιας περίληψης για τα δεδομένα των συναλλαγών.

## 1.3. ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ – ΕΠΙΧΕΙΡΗΜΑΤΙΚΗ ΕΥΦΥΪΑ – ΟΦΕΛΗ ΓΙΑ ΤΙΣ ΕΠΙΧΕΙΡΗΣΕΙΣ

Καθώς το περιβάλλον υποστήριξης λήψης αποφάσεων εξελίσσεται με το χρόνο σε περισσότερο σύνθετο, καθίσταται απαραίτητη η ύπαρξη γερών θεμελίων και υποδομών που θα το στηρίζουν. Πολλά δεδομένα πρέπει να ληφθούν υπόψη και πολλές ενέργειες που χρειάζονται αρκετό προσωπικό για να τις εκτελέσουν απαιτούνται για τη δημιουργία αυτών των υποδομών. Μια μικρή συλλογή εφαρμογών μπορεί να υλοποιηθεί από μια μικρή σχετικά ομάδα χρηστών χωρίς να είναι αναγκαία η ύπαρξη ενός προσεκτικά οργανωμένου συνόλου εκτελέσιμων ενεργειών, όμως μια εφαρμογή ΕΕ δε μπορεί. Η πρόχειρη δημιουργία ενός σχεδίου κατά τη διάρκεια της ανάπτυξης είναι ανεύθυνη τακτική καθώς θέτει σε κίνδυνο μεγάλες επενδύσεις του οργανισμού.

Προκειμένου να είναι ανταγωνιστικοί στο σημερινό κόσμο (όχι μόνο σε περιόδους ύφεσης αλλά και στο πλαίσιο του σύγχρονου ανταγωνιστικού κόσμου), οι οργανισμοί και οι επιχειρήσεις πρέπει να:

- Χειρίζονται τις πληροφορίες με ευφυείς τρόπους.
- Ανακαλύπτουν νέους τρόπους για να κατανοούν το περιβάλλον τους.
- Αναγνωρίζουν ταχύτερα τους κινδύνους και τις ευκαιρίες.
- Χαράσσουν αποτελεσματικές στρατηγικές αξιοποιώντας τα συμπεράσματα που εξάγουν βάσει της εμπειρίας των πελατών.
- Εκτελούν αποτελεσματικά τη στρατηγική τους.

Όπως έλεγε ο Abraham Maslow: "Όταν το μόνο εργαλείο που έχεις είναι ένα σφυρί, τείνεις να βλέπεις κάθε πρόβλημα σαν να ήταν καρφί" [25]. Επομένως, πρέπει να κατανοήσουν τις δυνατότητες και την απόδοσή τους. Αφού βεβαιωθούν ότι έχουν στη διάθεσή τους τις σωστές πληροφορίες, πρέπει να ενεργήσουν βάσει των πραγματικών δεδομένων και όχι των δεδομένων, αποφεύγοντας τις απότομες και απερίσκεπτες περικοπές.

Αξίζει να αναφέρουμε μερικά παραδείγματα εταιρειών που δραστηριοποιούνται στον τομέα της Επιχειρηματικής Ευφυΐας στον παγκόσμιο, ευρωπαϊκό και ελληνικό χώρο.

Παγκόσμια γνωστές εταιρείες όπως η SAS, η Oracle, η SAP και η IBM, έχουν αναπτύξει πακέτα εφαρμογών ΕΕ ενώ αναλαμβάνουν και την ενσωμάτωσή τους στις ανάγκες των πελατών τους. Άλλες εταιρείες που παρέχουν λύσεις ΕΕ είναι οι HZ Multimedia, Inc., Consillio, Inc., MicroStrategy, Inc., Cognos, Crystal Decisions, Hyperion Solutions Corporation, Brio Software Inc., αλλά και άλλες ακόμα που εύκολα μπορεί κανείς να αναζητήσει και να βρει πληροφορίες για αυτές και για το τι ακριβώς προσφέρουν μέσω του διαδικτύου.

Στην Ελλάδα οι επιχειρήσεις που επιθυμούν να αποκτήσουν κάποιου είδους σύστημα ΕΕ που θα βελτιώνει την παραγωγή ή θα βελτιώνει τις σχέσεις με τους πελάτες (CRM) και που γενικότερα θα βοηθάει στην καλύτερη εκμετάλλευση των δεδομένων, μπορούν να επιλέξουν ανάμεσα από διάφορες εταιρίες που είτε είναι θυγατρικές μεγάλων ξένων επιχειρήσεων που δραστηριοποιούνται στον τομέα αυτό είτε έχουν αναπτύξει δικιά τους μεθοδολογία

[25] http://www.neo2.gr/web/neo2.gr/searchpagebasedontags/-/asset_publisher/Ep0Q/content/13737?redirect=%2Fweb%2Fneo2.gr%2Fviews

ενσωμάτωσης συστημάτων BI για τον ελλαδικό χώρο. Μερικές από αυτές είναι η Unixfor, η ALTEC με το πακέτο Unisoft Xelixi για την μικρομεσαία αγορά, η Info-Quest, και άλλες που μπορούν να βρεθούν μέσω του διαδικτύου.

# ΚΕΦΑΛΑΙΟ 2$^{Ο}$ «ΑΝΑΛΥΣΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ»

Η ανάλυση παλινδρόμησης αποτελεί μια από τις πιο συχνά χρησιμοποιούμενες μεθόδους πρόβλεψης – πρόγνωσης, πρόκειται για εργαλείο στα χέρια όσων πραγματοποιούν οικονομετρικές αναλύσεις. Αναφερόμενοι στην οικονομετρική ανάλυση θα πρέπει να τονίσουμε και τα στάδια αυτής (Ανδρικόπουλος Α., 2003).

- Το πρώτο στάδιο αφορά την εξειδίκευση του υποδείγματος δηλαδή στον καθορισμό των μεταβλητών που θα το απαρτίζουν, στην καταγραφή αυτών σε εξωγενείς και ενδογενείς καθώς και στην μαθηματική διατύπωση του υποδείγματος.
- Το δεύτερο στάδιο αναφέρεται στην κατάλληλη επιλογή των οικονομετρικών τεχνικών για την εκτίμηση των συντελεστών των μεταβλητών μας και ονομάζεται εκτίμηση του υποδείγματος.
- Τέλος το τρίτο στάδιο αφορά των έλεγχο του υποδείγματος με την παράλληλη εφαρμογή οικονομικών, στατιστικών αλλά και οικονομετρικών κριτηρίων για τον έλεγχο των αποτελεσμάτων της εκτιμήσεως.

Πιο συγκεκριμένα, η διαδικασία της οικονομετρικής ανάλυσης περιλαμβάνει τα εξής στάδια (Ανδρικόπουλος Α., 2003):

- **Εξειδίκευση του υποδείγματος:**
  - Καθορισμός των μεταβλητών που θα περιληφθούν, διαχωρισμός σε ενδογενείς/ εξωγενείς, μαθηματική διατύπωση.
  - Η οικονομική θεωρία, αν το πεδίο εφαρμογής είναι η οικονομική θεωρία, είναι αυτή η οποία μπορεί να υποδείξει ποιες μεταβλητές είναι σημαντικές ή ίσως σχετικές αλλά δεν καθορίζει την μαθηματική μορφή που συνδέει τις μεταβλητές.
  - Συνήθως η επιλογή της μαθηματικής μορφής της συναρτησιακής σχέσεως είναι συνδυασμός των πληροφοριών από την οικονομική θεωρία και τα πραγματικά δεδομένα.

- **Εκτίμηση του υποδείγματος:**
  - Ø Εφαρμογή των κατάλληλων οικονομετρικών μεθόδων για την εκτίμηση των παραμέτρων του υποδείγματος.
  - Ø Οι στατιστικές παρατηρήσεις που μπορούν να χρησιμοποιηθούν, είναι: Χρονολογικές σειρές (time series data): Διαχρονικές παρατηρήσεις για μια σειρά ετών, μηνών κτλ για μια οικονομική μονάδα.
  - Ø Διαστρωματικά στοιχεία (cross-sectional data): Παρατηρήσεις για έναν αριθμό οικονομικών μονάδων για μια χρονική στιγμή.
  - Ø Δυναμικά διαστρωματικά στοιχεία (panel data): Διαχρονικές παρατηρήσεις για μια σειρά οικονομικών μονάδων.
- **Έλεγχος του υποδείγματος:**
  - Ø Αξιολόγηση και έλεγχος των αποτελεσμάτων της εκτιμήσεως.
  - Ø Επιβεβαιώνεται ή αμφισβητείται η οικονομική θεωρία;

Γραφικά όλα τα παραπάνω απεικονίζονται στο ακόλουθο διάγραμμα:


Εικόνα 1: Διαδικασία Οικονομετρικής Ανάλυσης (Ανδρικόπουλος Α., 2003)

## 2.1. ΑΝΑΛΥΣΗ ΣΥΣΧΕΤΙΣΗΣ ΚΑΙ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Ο προσδιορισμός της σχέσης ανάμεσα σε δυο ή περισσότερες μεταβλητές ονομάζεται ανάλυση παλινδρόμησης. Πρωταρχική επιδίωξη της ανάλυσης παλινδρόμησης είναι η εκτίμηση ενός οικονομετρικού υποδείγματος, προκειμένου να προσδιοριστεί με ακρίβεια η σχέση μεταξύ της εξαρτημένης και της ανεξάρτητης μεταβλητής (των ανεξάρτητων μεταβλητών).

Το πρόβλημα παλινδρόμησης βασίζεται στην προσπάθεια πρόβλεψης της συμπεριφοράς μιας μεταβλητής (εξαρτημένης), βασισμένη σε μια άλλη (ανεξάρτητη). Όταν αυτή η πρόβλεψη γίνεται σε δύο μόνο τυχαίες μεταβλητές τότε θα μιλάμε για την απλή παλινδρόμηση, ενώ όταν η πρόβλεψη για την εξαρτημένη μεταβλητή βασίζεται σε περισσότερες από μία μεταβλητές τότε θα ονομάζεται πολλαπλή παλινδρόμηση (Ανδρικόπουλος Α., 2003).

Η ανάλυση παλινδρόμησης δημιουργεί μια στοχαστική σχέση την οποία προσπαθεί να προσεγγίσει. Προκειμένου να κάνει κάτι τέτοιο βασίζεται σε ορισμένες υποθέσεις:

Έστω $Y_i = \beta_0 + \beta_1 * X_i + e_i$ η στοχαστική σχέση, όπου:

Y: ονομάζεται εξαρτημένη ή ερμηνευόμενη μεταβλητή,

X: είναι η ανεξάρτητη ή ερμηνευτική μεταβλητή και

i: είναι μία αντιπροσωπευτική από τις η παρατηρήσεις του δείγματος.

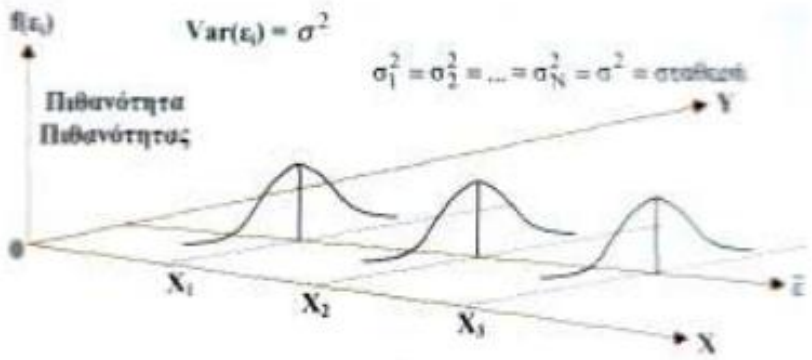
Η παραπάνω σχέση ονομάζεται εξίσωση γραμμικής παλινδρόμησης και οι υποθέσεις που συνιστούν το κλασσικό γραμμικό υπόδειγμα παλινδρόμησης είναι οι εξής (Greene W.H, 2011):

- Η συναρτησιακή μορφή του υποδείγματος είναι γραμμική: $Y_i = \beta_0 + \beta_1 * X_i + e_i$ , δηλαδή η μαθηματική μορφή η οποία και συνδέει την ανεξάρτητη με την εξαρτημένη μεταβλητή είναι γραμμικής μορφής. Η γραμμικότητα αυτή αναφέρεται στους συντελεστές παλινδρόμησης και όχι στις μεταβλητές του υποδείγματος.
- Ο μέσος του όρου σφάλματος είναι μηδέν: $E(\varepsilon_i / X_i) = 0$, δηλαδή η μεταβλητή $\varepsilon_i$ είναι μια τυχαία μεταβλητή η οποία μπορεί να παίρνει τόσο αρνητικές όσο και θετικές τιμές αλλά η μέση της τιμή (μαθηματική ελπίδα), υπό τον περιορισμό ότι η τιμή των ανεξάρτητων μεταβλητών είναι δεδομένες, είναι μηδέν. Η σημασία της υπόθεσης αυτής συνίσταται στο γεγονός ότι οι μη εμφανείς παράγοντες οι οποίοι και «υπολογίζονται»

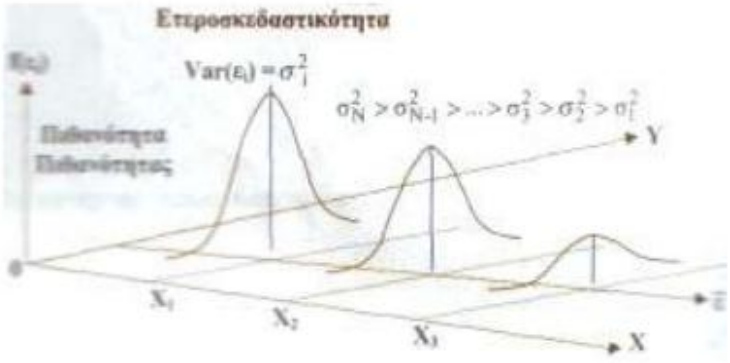
στον διαταρακτικό όρο δεν επηρεάζουν συστηματικά την μέση τιμή της εξαρτημένης μεταβλητής.

Ø Η διακύμανση όλων των όρων σφάλματος είναι η ίδια σταθερά $Var\left(\varepsilon_i/X_i\right) = \sigma^2$.

Η υπόθεση αυτή μας λέει ότι η διασπορά των τιμών της τυχαίας μεταβλητής γύρω από τον μέσο της δεν αλλάζει όταν μεταβάλλεται η τιμή της ανεξάρτητης μεταβλητής $X_i$. Όταν η διακύμανση παραμένει σταθερή ο διαταρακτικός όρος χαρακτηρίζεται ομοσκεδαστικός (*Εικόνα* ) ενώ όταν η διακύμανση δεν είναι σταθερή ετεροσκεδαστικός (Εικόνα 3).




*Εικόνα 2:* Ομοσκεδαστικότητα *(Ανδρικόπουλος Α., 2003).*

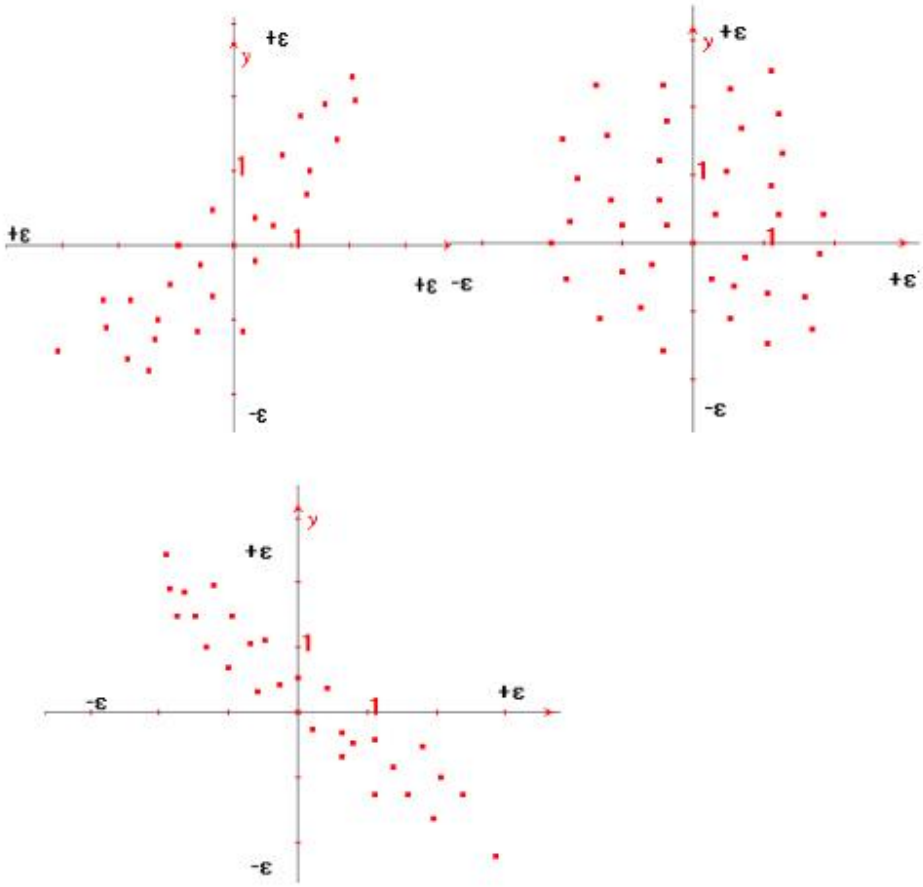



*Εικόνα 3:* Ετεροσκεδαστικότητα *(Ανδρικόπουλος Α., 2003).*

Ø Η συνδιακύμανση μεταξύ των όρων σφάλματος είναι μηδέν: $Cov\left(\varepsilon_{i,}\varepsilon_j\right) = 0$ .

Ø $Cov\left(\varepsilon_i, \varepsilon_j\right) = E[\varepsilon_i - E(\varepsilon_i)]\left[\varepsilon_j - E\left(\varepsilon_j\right)\right] = E\left(\varepsilon_i, \varepsilon_j\right) = 0$

Δηλαδή η σχέση αυτή μας λέει ότι οι διαταρακτικοί όροι χαρακτηρίζονται από την απουσία της αυτοσυσχέτισης καθώς και ότι για κάθε $X_i$ οι αποκλίσεις των κάθε τιμών Y από τις μέσες τιμές δεν μας δίνουν υποδείγματα των κάτωθι μορφών (Εικόνα 4).



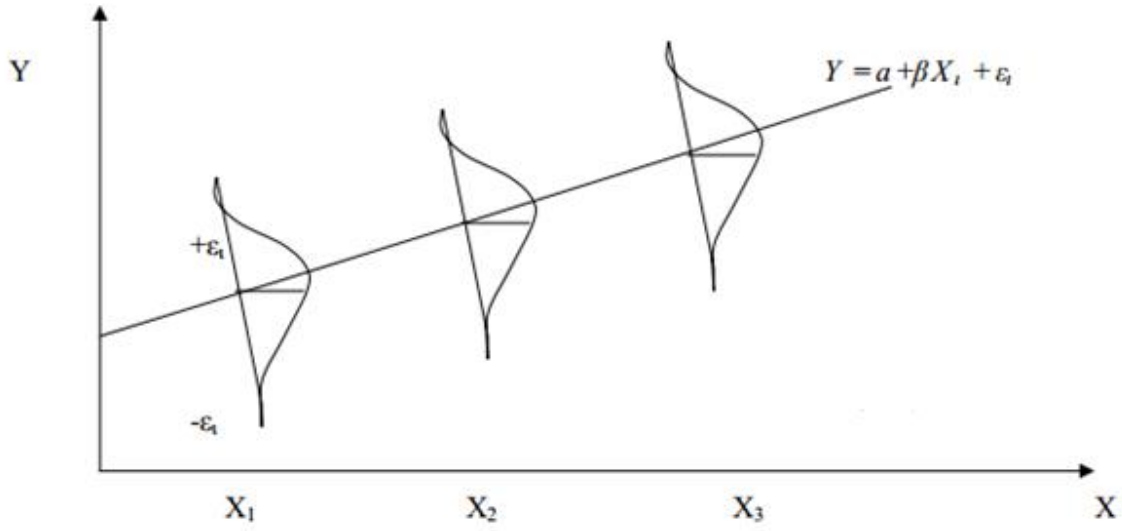

Εικόνα 4: Συντελεστής συσχέτισης (Ανδρικόπουλος Α., 2003).

Ø Η συνδιακύμανση των όρων σφάλματος και των παρατηρήσεων της ανεξάρτητης μεταβλητής είναι πάντα μηδέν: $Cov\left(\varepsilon_{i,}X_i\right) = 0$, για κάθε ί,j =1,2,.,.,η +ει. Η υπόθεση αυτή μας τονίζει η ανεξάρτητη μεταβλητή X δεν είναι στοχαστική και πως οι τιμές παραμένουν σταθερές σε μια επαναληπτική διαδικασία.

Ø Οι όροι σφάλματος, ανεξάρτητοι μεταξύ τους, ακολουθούν την κανονική κατανομή: $\varepsilon_i \sim N(0, \sigma^2)$ με μέσο μ=0 και διασπορά $\sigma^2$. Η τελευταία υπόθεση

τίθεται για μικρά δείγματα, όπου μικρά δείγματα στην οικονομετρία θεωρούνται αυτά με αριθμό παρατηρήσεων κάτω από 30 ή κάτω από 20. Η υπόθεση αυτή δεν είναι αναγκαία για μεγάλα δείγματα, αφού βάσει του κεντρικού οριακού θεωρήματος $\varepsilon_i \sim N(0, \sigma^2)$. Η τελευταία υπόθεση μας διευκολύνει στην στατιστική επαγωγή και την κατασκευή ελέγχων υποθέσεων σχετικά με την συμπεριφορά των εκτιμητριών. (Greene W.H, 2011)

Οι υποθέσεις αυτές μπορούν να συνοψισθούν στην εικόνα που ακολουθεί:




Εικόνα 5: Ανάλυση Παλινδρόμησης – Κατάλοιπα (Ανδρικόπουλος Α., 2003)

Η ανάλυση παλινδρόμησης είναι μια στατιστική μέθοδος που προσπαθεί να ερμηνεύσει και να ποσοτικοποιήσει τις μεταβολές μιας μεταβλητής (εξαρτημένη) σε σχέση με τις μεταβολές άλλων μεταβλητών (ανεξάρτητες). Ο ερευνητής με βάση την θεωρία και την εμπειρία επιλέγει: (Maddala, G.S., 2005)

- την εξαρτημένη μεταβλητή,
- τις ανεξάρτητες μεταβλητές,
- την μορφή της συνάρτησης,
- και εκτιμά τις παραμέτρους ($\beta_i$).

Μια από τις μεθόδους που χρησιμοποιούνται για την εκτίμηση της γραμμής παλινδρόμησης είναι η μέθοδος των ελαχίστων τετραγώνων (OLS).

Είναι η μέθοδος που χρησιμοποιείται περισσότερο επειδή:

- οι εκτιμητές έχουν πολλές επιθυμητές ιδιότητες,
- είναι εύκολη στην εφαρμογή της.

Ο αριθμός των εκτιμητών είναι ουσιαστικά άπειρος δηλαδή μπορούμε να κατασκευάσουμε και άπειρες γραμμές παλινδρομήσεως. Χρειαζόμαστε ένα κριτήριο. Το κριτήριο αυτό παρουσιάζεται ακολούθως (Ανδρικόπουλος Α., 2003):

Επιλογή των συντελεστών $\beta_0$ και $\beta_1$ που ελαχιστοποιούν τα τετράγωνα των αποκλίσεων της ευθείας παλινδρόμησης από τις πραγματικές τιμές.

Σε αυτό το σημείο πρέπει να αναφέρουμε ότι η ακρίβεια της εξίσωσης παλινδρόμησης εξαρτάται από τρεις κύριους παράγοντες. (Maddala, G.S., 2005)

- Από το μέγεθος του δείγματος. Αφού το n εμφανίζεται και στον παρανομαστή του κλάσματος όσο μεγαλύτερο είναι το μέγεθος τόσο και μικρότερη η τιμή του τυπικού σφάλματος (standard error), άρα τόσο πιο ακριβής η εκτίμηση αυτού.
- Επίσης, όσο μεγαλύτερη είναι η απόκλιση του X τόσο μικρότερη η τιμή του standard error, άρα τόσο πιο ακριβής η εκτίμηση αυτού.
- $Var\ (\varepsilon_i)$ Όσο μεγαλύτερη είναι η απόκλιση των $\varepsilon_i$ στην κατακόρυφη διεύθυνση τόσο το λιγότερο ακριβής είναι η εκτίμηση του standard error of the slope parameter.

Επιπλέον να αναφέρουμε πως στην εκτίμηση παλινδρόμησης σημαντικό ρόλο διαδραματίζει μεταξύ άλλων και ο συντελεστής προσδιορισμού. Η αναλογία της μεταβλητότητας της εξαρτημένης μεταβλητής που ερμηνεύεται από την παλινδρόμηση ονομάζεται συντελεστής προσδιορισμού (coefficient of determination). Ο συντελεστής προσδιορισμού $R^2$ μετράει το ποσοστό της μεταβατικότητας της μεταβλητής Y η οποία και ερμηνεύεται από την παλινδρόμηση του δείγματος. Ο συντελεστής προσδιορισμού $R^2$ συμπίπτει με το $r^2$, όπου r ο δειγματικός συντελεστής συσχέτισης στο απλό γραμμικό μοντέλο. Επίσης όταν το $R^2$ είναι κοντά στην μονάδα τότε λέμε ότι το μοντέλο μας είναι καλό στην ερμηνεία της απόκλισης του Y, υπάρχει τέλεια προσαρμογή της ευθείας παλινδρομήσεως (Fit Model). (Greene W.H, 2011)

## 2.2. ΜΕΘΟΔΟΣ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ

Σε πολλά πειράματα υπάρχει μία γραμμική σχέση ανάμεσα στα μετρούμενα μεγέθη. Για παράδειγμα, η ταχύτητα ενός σώματος το οποίο εκτελεί ελεύθερη πτώση, μεταβάλλεται γραμμικά με το χρόνο, εφόσον αγνοήσουμε την αντίσταση του αέρα. Τοποθετώντας τα σημεία σε ένα διάγραμμα, βλέπουμε ότι αυτά προσεγγίζουν μία ευθεία γραμμή. Το επόμενο βήμα είναι να βρούμε την κλίση της ευθείας η οποία προσεγγίζει περισσότερο αυτά τα σημεία, και το σημείο στο οποίο αυτή τέμνει τον άξονα y (τεταγμένη). Σε κάθε περίπτωση, δεν περιμένουμε η ευθεία να διέρχεται από όλα τα σημεία, λόγω της παρουσίας τυχαίων σφαλμάτων. Μπορούμε να βρούμε προσεγγιστικές τιμές τόσο για την κλίση όσο και για την τεταγμένη, εάν σχεδιάσουμε μία ευθεία η οποία να διέρχεται ανάμεσα από τα διεσπαρμένα σημεία. Η ακριβέστερη όμως μέθοδος για να το πετύχουμε αυτό είναι η μέθοδος των ελαχίστων τετραγώνων (Ανδρικόπουλος Α., 2003).

Αναφέραμε ότι γραμμική ευθεία παλινδρόμησης είναι α + β*x, και γνωρίζοντας ότι είναι αδύνατο να βρούμε ένα ακριβώς y, το οποίο να εισέρχεται από την ευθεία παλινδρόμησης, όταν γνωρίζουμε ένα x, τότε έχουμε κάποιο προβλεπόμενο y. Δηλαδή, την ευθεία πρόβλεψης ή παλινδρόμησης ή ελαχίστων τετραγώνων:

$$\hat{y} = \widehat{bo} + \widehat{b_1}X$$

Η εκτίμηση της πληθυσμιακής ευθείας παλινδρόμησης

$$E\,(Y/X) = \widehat{bo} + \widehat{b_1}X$$

ονομάζεται ευθεία ελαχίστων τετραγώνων από τη μέθοδο υπολογισμού των παραμέτρων της. Τα $\widehat{bo}$, $\widehat{b_1}$ είναι εκτιμητές των $b_0$ και $b_1$, που επιλέγονται με την μέθοδο των ελαχίστων τετραγώνων. Και επιλέγονται έτσι ώστε το άθροισμά των τετραγώνων των σφαλμάτων να είναι ελάχιστο. Δηλαδή: (Maddala, G.S., 2005)

$$\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2$$

Η γραμμή ελαχίστων τετραγώνων είναι αυτή που ελαχιστοποιεί το άθροισμα των τετραγώνων των σφαλμάτων σε σχέση με τις προσεγγίσεις των $b_0$ και $b_1$. Η ελαχιστοποίηση της τελευταίας

έκφρασης γίνεται σε σχέση με τα $b_0$ και $b_1$. Μετά από την παραγοντοποίηση αυτής αναφορικά με τα $b_0$ και $b_1$, αφού τεθούν οι πρώτες παράγωγοι ίσον με το μηδέν. (Greene W.H, 2011)

Και καταλήγουμε στους εξής τύπους:

$$\hat{b}_1 = \frac{\sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}$$

$$\widehat{b_0} = \bar{Y} - \widehat{b_1} x$$

Χρησιμοποιώντας τη μέθοδο ελαχίστων τετραγώνων σκοπός μας είναι να εκτιμήσουμε τις παραμέτρους του υποδείγματος της παλινδρόμησης, δηλαδή τους συντελεστές α και β κατά τέτοιο τρόπο, ώστε η ευθεία γραμμή που θα προκύψει να περιγράφει κατά τον καλύτερο δυνατό τρόπο τη σχέση μεταξύ των μεταβλητών X και Y. Η γραμμή της παλινδρόμησης πρέπει να περνάει κοντά από τα σημεία που αντιστοιχούν στα ζεύγη των παρατηρήσεων (Xi,Yi), έτσι ώστε να ελαχιστοποιούνται τα σφάλματα της πρόβλεψης. Για να γίνει κατανοητή η μέθοδος, πρέπει πρώτα να εισάγουμε μερικούς συμβολισμούς.

Έχει επικρατήσει στη διεθνή βιβλιογραφία να συμβολίζουμε με μικρούς ελληνικούς χαρακτήρες τις τιμές των παραμέτρων του πληθυσμού και με λατινικούς χαρακτήρες τις εκτιμήσεις τους από τα δεδομένα του δείγματος. Τον ίδιο συμβολισμό θα χρησιμοποιήσουμε και εδώ. Μόλις οι εκτιμήσεις αυτές γίνουν γνωστές, θα είμαστε σε θέση να προβλέπουμε τις τιμές της Y με την εξίσωση παλινδρόμησης:

$$\hat{Y} = b_0 + b_1 X$$

Δηλαδή, η $\hat{Y}$ είναι η εκτίμηση της E(Y). Έτσι, κατά αναλογία με την εξίσωση που αναφέρεται στη γραμμή παλινδρόμησης του πληθυσμού, οι αποκλίσεις μεταξύ των πραγματικών τιμών της Y και των τιμών $\hat{Y}$ με e (όπου e τα κατάλοιπα που προκύπτουν από τις προσεγγιστικές τιμές της γραμμής παλινδρόμησης $b_0 + b_1 * X$ στο σύνολο των n σημείων), δηλαδή:

$$e_i = Y_i - \hat{Y}_i$$

Ή

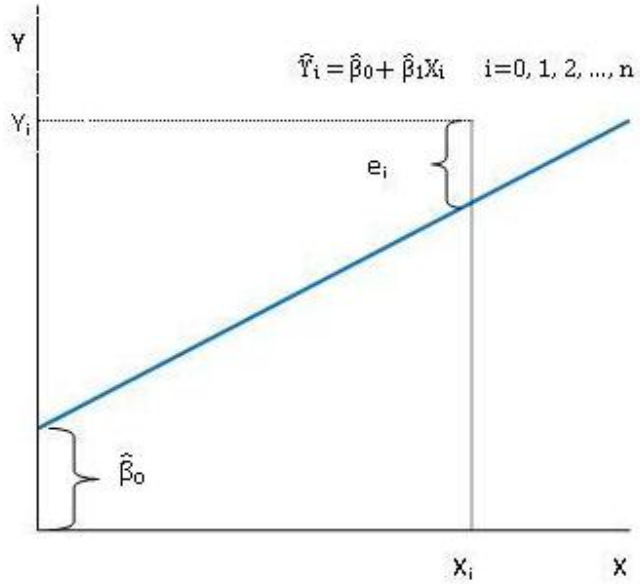
$$e_i = Y_i - (b_0 + b_1 X_i), \text{ για i} = 1, \ldots, \text{n}$$

Επομένως, αναζητούμε εκείνες τις τιμές των $b_0$ και $b_1$ που θα ελαχιστοποιούν τις αποκλίσεις (κατάλοιπα ή σφάλματα) $e_i$. (Maddala, G.S., 2005)

Επειδή τα σφάλματα έχουν και θετικό και αρνητικό πρόσημο, θα προσπαθήσουμε να ελαχιστοποιήσουμε τα τετράγωνα τους και μάλιστα το άθροισμά τους. Έτσι λοιπόν προκύπτει και η ονομασία της μεθόδου των ελαχίστων τετραγώνων. Το άθροισμα των τετραγώνων των αποκλίσεων για τα n ζεύγη των παρατηρήσεων ισούται με:

$$\sum_{i=1}^{n}(Y_i - \widehat{Y_i})^2 = \sum[Y_i - (b_0 + b_1X_i)]^2$$

Έστω ότι για την παραπάνω εξίσωση έχουν εκτιμηθεί οι συντελεστές παλινδρόμησης $\beta_0 = \widehat{\beta_0}$ και $\beta_1 = \widehat{\beta_1}$. Ο στόχος μας είναι οι τιμές αυτές να αποκλίνουν όσο το δυνατόν λιγότερο. Έτσι προκύπτει η γραμμή παλινδρόμησης (Εικόνα 6):




Εικόνα 6: Ευθεία Παλινδρόμησης (Ανδρικόπουλος Α., 2003)

Το Θεώρημα Gauss - Markov ισχύει μόνο για γραμμικά συστήματα παλινδρόμησης και λέει ότι δεδομένων των υποθέσεων, οι συντελεστές $\widehat{\beta_0}$ και $\widehat{\beta_1}$ είναι οι πιο αποτελεσματικοί από όλους τους πιθανούς αμερόληπτους εκτιμητές των $\beta_0$ και $\beta_1$, καθώς έχουν την μικρότερη διακύμανση μεταξύ των οποιονδήποτε αμερόληπτων εκτιμητών. Αν το θεώρημα δεν ισχύει τότε οι εκτιμητές $\widehat{\beta_0}$ και $\widehat{\beta_1}$, των ελαχίστων τετραγώνων δεν είναι BLUE (Best Linear Unbiased Estimator). Το θεώρημα Gauss - Markov μας λέει ότι αν ισχύουν οι συνθήκες παλινδρόμησης τότε οι εκτιμητές ελαχίστων τετραγώνων $\hat{\alpha}$ και $\hat{\beta}$: (Maddala, G.S., 2005)

- Είναι αμερόληπτοι (unbiased) εκτιμητές των συντελεστών παλινδρόμησης του πληθυσμού, δηλαδή $E\ (\hat{\alpha}) = \hat{\alpha}$ και $E\ (\hat{\beta}) = \hat{\beta}$ .
- Είναι συνεπείς (consistent) καθώς με την επ' άπειρον αύξηση του δείγματος, οι εκτιμητές συγκλίνουν προς τους συντελεστές παλινδρόμησης του πληθυσμού.
- Είναι αποτελεσματικοί (efficient) διότι έχουν τη μικρότερη διακύμανση άρα και το μικρότερο τυπικό σφάλμα (standarderror) εκτίμησης μεταξύ όλων των αμερόληπτων εκτιμητών.
- Το άθροισμα των καταλοίπων (ei) γύρω από την γραμμή παλινδρόμησης $Y_i$ ισούται με το μηδέν.

$$\sum e_i = \sum (Y_i - \bar{Y})$$

- Το κατάλοιπο (ei), μετρά την απόκλιση της εξαρτημένης μεταβλητής $Y_i$ από την γραμμή παλινδρόμησης.
- Ο διαταρακτικός όρος (ε), μετρά την απόκλιση της εξαρτημένης μεταβλητής $Y_i$ από την μέση τιμή της $\bar{Y}$.

Ένα σημαντικό ακόμη χαρακτηριστικό της ανάλυσης παλινδρόμησης είναι και ο συντελεστής παλινδρόμησης. Πιο συγκεκριμένα, μέρος της μεταβλητότητας που παρατηρείται στις τιμές της Y οφείλεται στις μεταβολές της X και μέρος στις επιδράσεις των τυχαίων παραγόντων. Πόση μεταβλητότητα της Y εξηγείται από την X και πόση από τυχαίους παράγοντες;

Μέτρο του βαθμού προσαρμογής της ευθείας παλινδρόμησης στις παρατηρήσεις του δείγματος. Συμβολίζεται με $R^2$. Ο αριθμητικός μέσος μιας μεταβλητής είναι η καλύτερη πρόβλεψη όταν η μόνη διαθέσιμη πληροφορία είναι οι τιμές της ίδιας της μεταβλητής. Η X μπορεί να θεωρηθεί ότι ερμηνεύει την Y στον βαθμό που συμβάλλει στην πρόβλεψή της πέρα από τον μέσο. (Greene W.H, 2011)

Οι τύποι για τον υπολογισμό του συντελεστή προσδιορισμού είναι:

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS}$$

Ή

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

Ο συντελεστής συσχέτισης είναι ένα ακόμη μέτρο που χαρακτηρίζει την ευθεία αλλά και την ανάλυση παλινδρόμησης και αντιπροσωπεύει τον βαθμό γραμμικής συσχέτισης ανάμεσα σε δύο μεταβλητές χωρίς να ενδιαφέρεται για την αιτιώδη σχέση που μπορεί να υπάρχει μεταξύ τους. Συμβολίζεται με το ρ και συνδέεται άμεσα με τον συντελεστή προσδιορισμού αφού είναι η τετραγωνική ρίζα αυτού:

$$\rho = \sqrt{R^2}$$

ή

$$\rho = \frac{cov\,(X,Y)}{\sqrt{var\,(X)}\sqrt{var(Y)}}$$

Στο σημείο αυτό αξίζει να αναφέρουμε και στο διορθωμένο συντελεστή προσδιορισμού ο υπολογισμός του οποίου βασίζεται στον ακόλουθο τύπο:

$$\bar{R}^2 = 1 - \frac{\sum \frac{\left(Y_i - \widehat{Y}_i\right)^2}{n-k}}{\sum \frac{(Y_i - \bar{Y})^2}{n-1}} = 1 - \frac{n-1}{n-k}\,(1 - R^2)$$

Όταν προστίθεται μια ανεξάρτητη μεταβλητή υπάρχει όφελος αλλά και κόστος. Αυξάνει η τιμή του πολλαπλού συντελεστή προσδιορισμού, χάνεται ένας βαθμός ελευθερίας (n - k). Λαμβάνει υπόψη και το κόστος και το όφελος και είναι μικρότερος από το συντελεστή προσδιορισμού. (Maddala, G.S., 2005)

## 2.3. ΕΛΕΓΧΟΙ ΥΠΟΘΕΣΕΩΝ στο ΑΠΛΟ ΓΡΑΜΜΙΚΟ ΜΟΝΤΕΛΟ

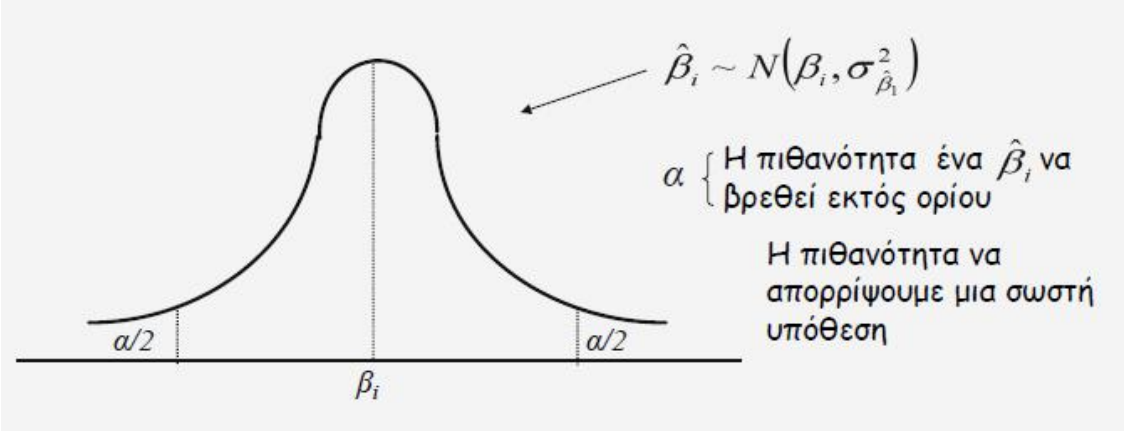
Πριν προσδιορίσουμε την έννοια των ελέγχων είναι χρήσιμο να αναφερθούμε στον προσδιορισμό των υποθέσεων, της μηδενικής και της εναλλακτικής. Πιο συγκεκριμένα (Ανδρικόπουλος Α., 2003):

- **η μηδενική υπόθεση ($H_0$)** αναφέρεται σε μια τιμή ή σε ένα διάστημα τιμών ενός συντελεστή που συνήθως δεν είναι οι αναμενόμενες.
- **η εναλλακτική υπόθεση ($H_1$)** αναφέρεται στην τιμή ή στο διάστημα τιμών του συντελεστή που ισχύουν αν δεν γίνει αποδεκτή η μηδενική υπόθεση.

Τα κριτήρια αποδοχής ή απόρριψης της μηδενικής υπόθεσης διαμορφώνονται ως εξής:

- **Περιοχή αποδοχής:** Το απαραίτητο εύρος τιμών της στατιστικής για να γίνει αποδεκτή η $H_0$.
- **Περιοχή απόρριψης:** Το απαραίτητο εύρος τιμών της στατιστικής για να απορριφθεί η $H_0$.

Πιο συγκεκριμένα τα παραπάνω αποτυπώνονται στο ακόλουθο διάγραμμα:




*Εικόνα 7: Έλεγχος Υπόθεσης (Ανδρικόπουλος Α., 2003)*

Όσον αφορά την οικονομετρία, η αξιολόγηση του υποδείγματος:

$$\widehat{Y_i} = \widehat{bo} + \widehat{b_1} X_i$$

Βασίζεται στην ακόλουθη μεθοδολογία:

- Οι συντελεστές $\beta_i$ ελέγχονται με βάση την t – κατανομή.
- Το σύνολο του υποδείγματος ελέγχεται με βάση την f – κατανομή.
- Η σημαντικότητα της διακύμανσης του διαταρακτικού όρου ελέγχεται με βάση την $X^2$ κατανομή.

Αν η πιθανότητα να συμβεί το ψευδές είναι α, τότε το ενδεχόμενο «αληθές» έχει (1-α) πιθανότητες να συμβεί. Το (1–α) ονομάζεται διάστημα εμπιστοσύνης (confidence interval) και είναι αυτό που προσδιορίζει την περιοχή εντός της οποίας είναι πιθανό να βρίσκονται οι τιμές

των συντελεστών. Το α καλείται επίπεδο σημαντικότητας (level significance) και είναι αυτό που ορίζει με ακρίβεια την περιοχή του διαστήματος εμπιστοσύνης.

Αρχικά πραγματοποιούμε ελέγχους για τη στατιστική σημαντικότητα των παραμέτρων του υποδείγματος, οι έλεγχο αυτοί διαμορφώνονται ως εξής (Ανδρικόπουλος Α., 2003):

**Δίπλευρος Έλεγχος:**

Στον δίπλευρο έλεγχο σκοπός μας είναι η διερεύνηση της παρακάτω υποθέσεως: Εάν υπάρχει στατιστικά σημαντική σχέση ανάμεσα στην εξαρτημένη και στην ανεξάρτητη μεταβλητή. Με αλλά λόγια εάν η κλίση της συνάρτησης παλινδρόμησης μας θα είναι θετική ή αρνητική με βάση την συναρτησιακή σχέση ανάμεσα στα X,Y. Ο έλεγχος διαμορφώνεται ως εξής:

$$H_0: \beta_1 = 0 \text{ (Μηδενική Υπόθεση)}$$

$$H_1: \beta_1 \neq 0 \text{ (Εναλλακτική Υπόθεση)}$$

**Δεξιός Μονόπλευρος Έλεγχος:**

Ας υποθέσουμε ότι με βάση την οικονομική θεωρία αναμένουμε ότι ο συντελεστής $\beta_1$ είναι θετικός και θέλουμε να διαπιστώσουμε εμπειρικά αν ο $\beta_1$ είναι, πρώτο, θετικός και, δεύτερο, στατιστικά σημαντικός.

Σε αυτή την περίπτωση, οι $H_0$ και H1, υποθέσεις εκφράζονται ως εξής:

$$H_0: \beta_1 = 0 \text{ (Μηδενική Υπόθεση)}$$

$$H_1: \beta_1 > 0 \text{ (Εναλλακτική Υπόθεση)}$$

**Αριστερός Μονόπλευρος Έλεγχος:**

Ας υποθέσουμε ότι με βάση την οικονομική θεωρία αναμένουμε ότι ο συντελεστής $\beta_1$ είναι αρνητικός και θέλουμε να διαπιστώσουμε εμπειρικά αν ο $\beta_1$ είναι, πρώτο, θετικός και, δεύτερο, στατιστικά σημαντικός.

Σε αυτή την περίπτωση, οι $H_0$ και H1, υποθέσεις εκφράζονται ως εξής:

$$H_0: \beta_1 = 0 \text{ (Μηδενική Υπόθεση)}$$

$$H_1: \beta_1 < 0 \text{ (Εναλλακτική Υπόθεση).}$$

Τα πιθανά λάθη που μπορούν να προκύψουν κατά την διαδικασία του στατιστικού ελέγχου είναι:

- **Λάθος Τύπου I (type I error):** Αφορά την πιθανότητα κατά την οποία μετά τη διαδικασία του στατιστικού ελέγχου απορρίπτεται μια αληθής μηδενική υπόθεση.
- **Λάθος Τύπου II (type II error):** Αφορά την πιθανότητα μετά την φάση του στατιστικού ελέγχου να γίνει αποδεκτή μια ψευδής μηδενική υπόθεση.

Στις υποενότητες που ακολουθούν περιγράφονται οι έλεγχοι υποθέσεων που γίνονται κάθε φορά προκειμένου να διαπιστωθεί αν οι παραπάνω υποθέσεις ισχύουν ή όχι. Πιο συγκεκριμένα:

### 2.3.1. ΕΤΕΡΟΣΚΕΔΑΣΤΙΚΟΤΗΤΑ

Η ετεροσκεδαστικότητα οφείλεται σε διάφορες αιτίες. Οι πιο σημαντικές από αυτές περιγράφονται ακολούθως.

Η ετεροσκεδαστικότητα μπορεί να είναι μια φυσική ιδιότητα των μεταβλητών του υποδείγματος. Έτσι αν κατασκευάσουμε ένα οικονομετρικό υπόδειγμα με ανεξάρτητη μεταβλητή τις ώρες εξάσκησης στη δακτυλογράφηση και με εξαρτημένη μεταβλητή τον αριθμό των τυπογραφικών λαθών τότε αναμένουμε ότι όσο αυξάνουν οι ώρες μειώνεται όχι μόνο ο αριθμός των λαθών αλλά και η μεταβλητότητα αυτών. Επίσης αν κατασκευάσουμε ένα οικονομετρικό υπόδειγμα με ανεξάρτητη μεταβλητή τα ετήσια κέρδη των επιχειρήσεων και εξαρτημένη τα μερίσματα που δίνει η μετοχή των εταιριών και πάλι αναμένουμε ότι όσο αυξάνουν τα κέρδη θα αυξάνει τόσο το μέγεθος του μερίσματος όσο και η μεταβλητότητα αυτού. Στα παραπάνω παραδείγματα η δεσμευμένη διασπορά της εξαρτημένης μεταβλητής, και άρα και του διαταρακτικού όρου, δεν είναι σταθερή κατά μήκος των τιμών της ανεξάρτητης μεταβλητής, δηλαδή τα υποδείγματα παρουσιάζουν δεσμευμένη ετεροσκεδαστικότητα (Ανδρικόπουλος Α., 2003).

Η ετεροσκεδαστικότητα μπορεί να οφείλεται και σε ακραίες παρατηρήσεις (outliers) των μεταβλητών. Για παράδειγμα οποιοδήποτε οικονομετρικό υπόδειγμα με εξαρτημένη μεταβλητή τη μεταβλητότητα (volatility) των αποδόσεων του δείκτη S&P 500 συμπεριλάβει στο δείγμα παρατηρήσεις από το φθινόπωρο του 2008 θα παρουσιάζει δεσμευμένη ετεροσκεδαστικότητα. Τις ημέρες εκείνες η μεταβλητότητα έφθασε στο πρωτοφανές 80% όταν η μέση μεταβλητότητα της προηγούμενης δεκαετίας ήταν 20%.

Τέλος η ετεροσκεδαστικότητα μπορεί να οφείλεται και στο ότι το υπόδειγμα είναι λάθος εξειδικευμένο. Αυτό σημαίνει ότι μπορεί να απουσιάζει από αυτό μια σημαντική ανεξάρτητη μεταβλητή ή ότι η συναρτησιακή μορφή των μεταβλητών δεν είναι σωστή. Για παράδειγμα αν το σωστά εξειδικευμένο υπόδειγμα περιλαμβάνει ως ανεξάρτητη μεταβλητή τη x2 ενώ εμείς χρησιμοποιήσουμε τη x τότε θα παρουσιαστεί ετεροσκεδαστικότητα στα κατάλοιπα της παλινδρόμησης.

Έτσι λοιπόν όταν το υπόδειγμα παρουσιάζει δεσμευμένη ετεροσκεδαστικότητα οι εκτιμητές ελαχίστων τετραγώνων δεν είναι αποτελεσματικοί και επίσης δεν μπορούμε να διεξάγουμε ελέγχους υποθέσεων βασισμένοι σε αυτούς. Χρειαζόμαστε λοιπόν στη περίπτωση αυτή να εισάγουμε κάποιους νέους εκτιμητές για το υπόδειγμα. Οι εκτιμητές αυτοί ονομάζονται εκτιμητές σταθμισμένων ελαχίστων τετραγώνων (weighted least squares) ή εκτιμητές γενικευμένων ελαχίστων τετραγώνων (generalized least squares). Η βασική ιδέα για τη κατασκευή αυτών των εκτιμητών είναι η μετατροπή του γραμμικού υποδείγματος σε ένα νέο το οποίο δεν παρουσιάζει ετεροσκεδασικότητα και η εκτίμηση του νέου υποδείγματος με τη μέθοδο ελαχίστων τετραγώνων. (Maddala, G.S., 2005)

Συνοψίζοντας τα αποτελέσματα μέχρι τώρα για ένα υπόδειγμα που παρουσιάζει ετεροσκεδαστικότητα έχουμε δύο συνεπείς και ασυμπωτικά κανονικούς εκτιμητές αυτού. Τους εκτιμητές ελαχίστων τετραγώνων και τους εκτιμητές εφικτών γενικευμένων ελαχίστων τετραγώνων. Γενικά ένας συνεπής και ασυμπτωτικά κανονικός εκτιμητής είναι ασυμπτωτικά πιο αποτελεσματικός από κάποιον άλλον επίσης συνεπή και ασυμπτωτικά κανονικό εκτιμητή αν η ασυμπτωτική διασπορά του δεύτερου είναι μεγαλύτερη ή ίση με αυτή του πρώτου. Μπορούμε να αποδείξουμε ότι οι εκτιμητές εφικτών γενικευμένων ελαχίστων τετραγώνων είναι ασυμπτωτικά πιο αποτελεσματικοί από τους εκτιμητές ελαχίστων τετραγώνων. Προσοχή όμως το συμπέρασμα αυτό βασίζεται στην υπόθεση ότι το μέγεθος του δείγματος είναι μεγάλο και ότι το υπόδειγμα της ετεροσκεδαστικότητας είναι σωστά εξειδικευμένο. Αν ένα από αυτά τα δύο δεν ικανοποιείται είναι πιθανό η ασυμπτωτική διασπορά των εκτιμητών εφικτών γενικευμένων ελαχίστων τετραγώνων να είναι μεγαλύτερη από αυτή των εκτιμητών ελαχίστων τετραγώνων.

Όλοι οι έλεγχοι ακολουθούν την ακόλουθη κοινή στρατηγική. Τα κατάλοιπα ελαχίστων τετραγώνων είναι συνεπείς εκτιμητές των τιμών του διαταρακτικού όρου ακόμα και όταν υπάρχει ετεροσκεδαστικότητα στο υπόδειγμα. Έτσι οι τιμές των καταλοίπων θα

αντικατοπτρίζουν σε κάποιο βαθμό την πιθανή ετεροσκεδαστικότητα του υποδείγματος ή διαφορετικά, δεδομένης της δεσμευμένης ομοσκεδαστικότητας για το πληθυσμό πόσο πιθανό είναι να έχουμε ένα δείγμα με κατάλοιπα αυτά που έχουμε εκτιμήσει; (Maddala, G.S., 2005)

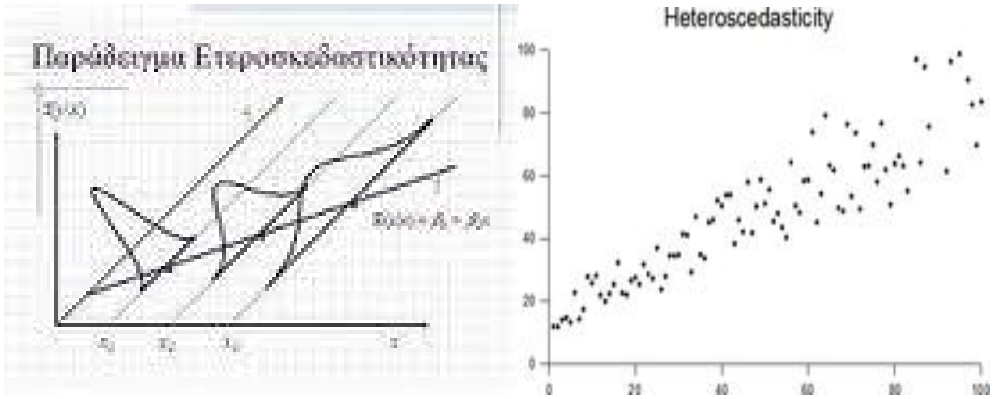
**Ο έλεγχος White** είναι ο γενικότερος έλεγχος δεσμευμένης ετεροσκεδαστικότητας που έχει παρουσιαστεί μέχρι στιγμής στη βιβλιογραφία. Ο έλεγχος White είναι πολύ γενικός μιας και δεν απαιτείται να κάνουμε κάποια συγκεκριμένη υπόθεση για τη μορφή της ετεροσκεδαστικότητας. Το χαρακτηριστικό αυτό αποτελεί ταυτόχρονα πλεονέκτημα και μειονέκτημα του ελέγχου. Αν ο έλεγχος απορρίψει τη μηδενική υπόθεση αυτό μπορεί πολύ απλά να οφείλεται, όχι στο ότι υπάρχει ετεροσκεδαστικότητα, αλλά στο ότι το υπόδειγμα δεν είναι σωστά εξειδικευμένο (για παράδειγμα η παράλειψη της μεταβλητής $x^2$ από τη παλινδρόμηση). Η ισχύς του ελέγχου White τείνει στη μονάδα όταν το N τείνει στο άπειρο, ενδεχομένως να απαιτηθεί ένα αρκετά μεγάλο δείγμα για να επιτευχθεί η σύγκλιση αυτή (Ανδρικόπουλος Α., 2003).

Αυτό σημαίνει ότι όταν το υπόδειγμα έχει ένα μεγάλο αριθμό ανεξάρτητων μεταβλητών και το δείγμα είναι πεπερασμένο η ισχύς του ελέγχου μπορεί να αποκλίνει σημαντικά από τη μονάδα.

Στον **έλεγχο Goldfeld - Quandt** χωρίζουμε τις παρατηρήσεις σε δύο σύνολα με τέτοιο τρόπο έτσι ώστε υπό τη μηδενική υπόθεση της ομοσκεδαστικότητας η διασπορά θα είναι ίδια στα δύο σύνολα ενώ υπό την εναλλακτική θα είναι στατιστικά διαφορετική. Ο έλεγχος Goldfeld - Quandt απαιτεί το διαχωρισμό του δείγματος σε δύο υποσύνολα για τα οποία μπορούμε να υποθέσουμε ότι έχουν διαφορετική διασπορά. Αυτό όμως δεν είναι πάντα δυνατό σε όλα τα υποδείγματα.

Ο έλεγχος **Breusch-Pagan-Godfrey** βασίζεται στην παραμετροποίηση της διασποράς. Ουσιαστικά ο έλεγχος Breusch-Pagan-Godfrey είναι μια ειδική περίπτωση του ελέγχου White.

Γραφικά η ετεροσκεδαστικότητα απεικονίζεται ως εξής:




Εικόνα 8: Ετεροσκεδαστικότητα (Ανδρικόπουλος Α., 2003)

Σύμφωνα με τα παραπάνω, το πρόβλημα της ετεροσκεδαστικότητας, παρουσιάζεται συνήθως σε διαστρωματικά στοιχεία και το πρόβλημα της αυτοσυσχέτισης σε διαχρονικά στοιχεία, παρόλα αυτά υπάρχουν περιπτώσεις που η ετεροσκεδαστικότητα απαντάται και σε διαχρονικά στοιχεία. (Maddala, G.S., 2005)

Ερευνητές στην προσπάθειά τους να κατασκευάσουν υποδείγματα πρόβλεψης για χρηματοοικονομικά στοιχεία παρατήρησαν ότι σε διάφορες χρονικές περιόδους οι μεταβλητές παρουσιάζουν μεγάλη μεταβατικότητα.

Αν προσπαθήσουμε να εκτιμήσουμε ένα τέτοιο υπόδειγμα πρόβλεψης θα καταλήξουμε ότι σε ορισμένες περιόδους τα σφάλματα προβλέψεων θα είναι μεγάλα (ασταθείς περίοδοι) και σε άλλες περιόδους μικρά (ήρεμοι περίοδοι), δηλαδή οι διακυμάνσεις των σφαλμάτων έτειναν να ομαδοποιούνται διαχρονικά κατά μεγέθη παρουσιάζοντας ένα είδος ετεροσκεδαστικότητας υπό συνθήκη.

Με τον τρόπο αυτό δημιουργήθηκαν ορισμένα υποδείγματα που λαμβάνουν υπόψιν τους τις διακυμάνσεις αυτές των διαταρακτικών όρων.

Τα υποδείγματα αυτά είναι τα παρακάτω:

- Το υπόδειγμα ARCH.
- Το υπόδειγμα GARCH.
- Το υπόδειγμα GARCH – M.

### 2.3.2. ΑΥΤΟΣΥΣΧΕΤΙΣΗ

Οι βασικότερες αιτίες που προκαλούν αυτοσυσχέτιση είναι οι ακόλουθες: (Maddala, G.S., 2005)

Οι περισσότερες οικονομικές χρονολογικές σειρές, όπως για παράδειγμα το ΑΕΠ, η κατανάλωση, η απασχόληση κ.α, παρουσιάζουν αδράνεια. Αυτό σημαίνει ότι κινούνται με βάση οικονομικούς κύκλους. Ξεκινώντας από τα χαμηλά μιας περιόδου ύφεσης οι τιμές αυτών των μεταβλητών μεταβάλλονται θετικά, δηλαδή η τιμή της επόμενης περιόδου είναι μεγαλύτερη από τη τιμή της προηγούμενης. Το φαινόμενο αυτό συνεχίζεται μέχρις ότου ξαναεμφανιστεί ύφεση και η μεταβολή επιβραδυνθεί ή και αναστραφεί.

Το φαινόμενο της αυτοσυσχέτισης μπορεί επίσης να οφείλεται στο ότι απουσιάζουν κάποιες ανεξάρτητες μεταβλητές από το υπόδειγμα οι οποίες επηρεάζουν την εξαρτημένη μεταβλητή ή ότι το υπόδειγμα δεν έχει τη σωστή συναρτησιακή μορφή.

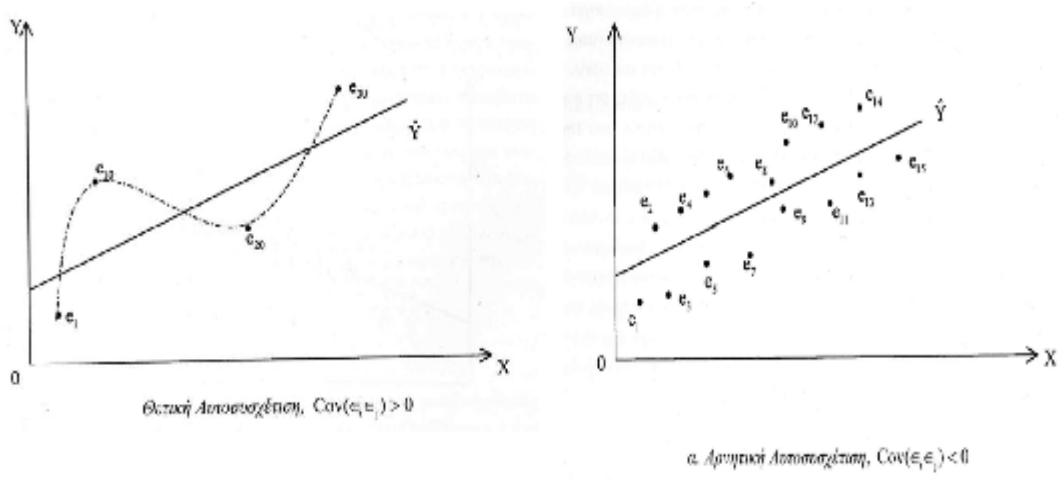
Η τρέχουσα τιμή της εξαρτημένης μεταβλητής μπορεί να εξαρτάται, περάν όλων των άλλων, και από τις παρελθούσες τιμές αυτής. Αν λοιπόν οι παρελθούσες τιμές απουσιάζουν από τις ανεξάρτητες μεταβλητές το υπόδειγμα θα παρουσιάσει αυτοσυσχέτιση (Ανδρικόπουλος Α., 2003).

Οι έλεγχοι αυτοσυσχέτισης βασίζονται στην ιδέα ότι αν ο διαταρακτικός όρος παρουσιάζει αυτοσυσχέτιση τότε αυτή θα εμφανιστεί και στα κατάλοιπα ελαχίστων τετραγώνων.

Ο **έλεγχος Durbin-Watson** ήταν ο πρώτος που παρουσιάστηκε για να εξετάσει πιθανή ύπαρξη αυτοσυσχέτισης στο διαταρακτικό όρο. Σε αντίθεση με όλους τους άλλους ελέγχους η κατανομή δειγματοληψίας της στατιστικής Durbin-Watson εξαρτάται από τις ανεξάρτητες μεταβλητές Χ. Αυτό συνεπάγεται ότι και οι κριτικές τιμές του ελέγχου εξαρτώνται από τις ανεξάρτητες μεταβλητές. Όμως οι Durbin και Watson έδειξαν ότι οι κριτικές τιμές φράσσονται αριστερά και δεξιά από δύο μεγέθη τα οποία εξαρτώνται μόνο από το μέγεθος του δείγματος, τον αριθμό των ανεξάρτητων μεταβλητών και φυσικά το επίπεδο σημαντικότητας α.

Ο **έλεγχος Durbin - Watson** απαιτεί ο διαταρακτικός όρος να ακολουθεί κανονική κατανομή και μπορεί να ελέγξει μόνο για αυτοσυσχέτιση 1ης τάξης. Αν θέλουμε να κατασκευάσουμε ελέγχους οι οποίοι θα γενικεύουν αυτά τα χαρακτηριστικά θα πρέπει να στραφούμε στην ασυμπτωτική θεωρία.

Οι έλεγχοι **Box - Pierce** και **Ljung - Box** υποθέτουν ότι Ε (ε/ Χ) = 0, το οποίο σημαίνει για παράδειγμα ότι ο πίνακας Χ δεν περιλαμβάνει παρελθούσες παρατηρήσεις της εξαρτημένης μεταβλητής. Αν όμως η τελευταία υπόθεση δεν ικανοποιείται από το γραμμικό υπόδειγμα τότε οι στατιστικές μπορεί να μην συγκλίνουν ασυμπτωτικά στην $X^2$ κατανομή. Ο έλεγχος **Breusch - Godfrey** έρχεται να καλύψει το κενό αυτό των παραπάνω ελέγχων.




Εικόνα 9: Αυτοσυσχέτιση (Ανδρικόπουλος Α., 2003)

Συμπερασματικά οι λόγοι που προκαλούν την αυτοσυσχέτιση:

- Η εσφαλμένη αλγεβρική εξειδίκευση του υποδείγματος.
- Η παράλειψη μιας ή περισσότερων ερμηνευτικών μεταβλητών.
- Η ύπαρξη συστηματικού σφάλματος μέτρησης στις μεταβλητές.
- Η εκτίμηση μέρους των παρατηρήσεων με παρεμβολή.
- Η κατανομή της επίδρασης ορισμένων τυχαίων γεγονότων σε περισσότερες από μια χρονικές περιόδους.

### 2.3.3. ΠΟΛΥΣΥΓΓΡΑΜΜΙΚΟΤΗΤΑ

Ο όρος "πολυσυγγραμμικότητα" εισήχθη για πρώτη φορά από τον Ragnar Frisch (1934) και εκφράζει την ύπαρξη μίας ή περισσοτέρων, ταυτοχρόνως, γραμμικών σχέσεων μεταξύ των ανεξαρτήτων μεταβλητών που συναπαρτίζουν ένα οικονομετρικό υπόδειγμα. Η ταυτόχρονη χρησιμοποίηση, σε ένα πολλαπλό γραμμικό υπόδειγμα, πολλών ανεξαρτήτων μεταβλητών που συσχετίζονται γραμμικά μεταξύ τους δεν οδηγεί σε βελτίωση της πληροφορίας που εμπερικλείεται στο υπόδειγμα. Ως άμεση συνέπεια του γεγονότος αυτού είναι το να μην επιτυγχάνεται ο βέλτιστος προσδιορισμός των τιμών της εξαρτημένης μεταβλητής. (Maddala, G.S., 2005)

Η κατάσταση η οποία δημιουργείται όταν υπάρχουν ισχυρές συσχετίσεις μεταξύ των ανεξάρτητων μεταβλητών στην πολλαπλή παλινδρόμηση ονομάζεται πολυσυγγραμικότητα (multicollinearity).

Στις περιπτώσεις που το πρόβλημα αυτό υφίσταται θα πρέπει κανείς να είναι ιδιαίτερα προσεκτικός στην ερμηνεία όλων των εκτιμητριών που προκύπτουν από το μοντέλο αυτό.

Υπάρχουν μια σειρά από προειδοποιητικές ενδείξεις που αν ο ερευνητής τις προσέξει είναι δυνατόν να αντιληφθεί ότι υπάρχει πολυσυγγραμμικότητα. Η πιο σημαντική από αυτές είναι ο πίνακας των συντελεστών συσχετίσεως (Correlation Matrix) των ανεξάρτητων μεταβλητών. Αν στον πίνακα αυτόν υπάρχουν μεγάλες θετικές ή αρνητικές τιμές θα έχουμε μια ένδειξη ότι οι αντίστοιχες ανεξάρτητες μεταβλητές που χρησιμοποιούνται στο μοντέλο έχουν μεταξύ τους ισχυρό βαθμό συσχέτισης. Το στατιστικό συμπέρασμα που προκύπτει στις περιπτώσεις αυτές είναι ότι κάποιες από τις μεταβλητές συνεισφέρουν ελάχιστα ή καθόλου στην πρόβλεψη της εξαρτημένης μεταβλητής οπότε και θα πρέπει να απομακρυνθούν από το μοντέλο.

Εάν, παρ' όλα αυτά, ο ερευνητής είναι βέβαιος ότι ο καθορισμός των ανεξάρτητων μεταβλητών έγινε σωστά θα πρέπει να εξετάσει δύο άλλες ενδείξεις για το κατά πόσον υπάρχει πολυσυγγραμμικότητα. (Ανδρικόπουλος Α., 2003)

- Εάν τα πρόσημα ορισμένων συντελεστών στην παλινδρόμηση είναι αντίθετα από αυτά που θα περίμενε κανείς λόγω της φύσης του προβλήματος, και
- εάν σημαντικοί συντελεστές της παλινδρόμησης εμφανίζονται να έχουν μεγάλες τιμές στις τυπικές αποκλίσεις τους.

Οποιαδήποτε από τις δύο αυτές ενδείξεις θα πρέπει να προβληματίσει τον ερευνητή και να τον οδηγήσει σε μια σοβαρή έρευνα για το κατά πόσον υφίσταται πολυσυγγραμμικότητα.

Ο καθορισμός εκείνων των γραμμικών συνδυασμών των παραμέτρων β που μπορούν να εκτιμηθούν με ακρίβεια είναι εξαιρετικά δύσκολο στην περίπτωση που υφίσταται πολυσυγγραμμικότητα. Παρότι, εν γένει, δεν είναι δυνατόν να εξαλειφθεί τελείως το πρόβλημα αυτό υπάρχει μια διαδικασία με την οποία ο ερευνητής μπορεί να εργασθεί με ένα μοντέλο που προκύπτει από το αρχικό με μετασχηματισμό των αρχικών μεταβλητών σε ένα σύνολο άλλων μεταβλητών που είναι ασυσχέτιστες μεταξύ τους. Η μεθοδολογία αυτή ονομάζεται ανάλυση κυρίων συνιστωσών (principal compenent analysis). Η τεχνική αυτή είναι μια παρά πολύ ισχυρή τεχνική τόσο στον εντοπισμό της πολυσυγγραμμικότητας όσο και ως μεθοδολογία που οδηγεί στον καθορισμό εκείνων των γραμμικών συνδυασμών των συντελεστών παλινδρόμησης που μπορούν να εκτιμηθούν με ακρίβεια. (Maddala, G.S., 2005)

Μια άλλη προσέγγιση είναι να υπολογισθούν οι k συντελεστές προσδιορισμού των παλινδρομήσεων κάθε μιας από τις ανεξάρτητες μεταβλητές στις υπόλοιπες k-1 ανεξάρτητες μεταβλητές. Εκείνες οι μεταβλητές που εμφανίζουν υψηλό συντελεστή προσδιορισμού θα πρέπει να θεωρηθεί ότι είναι συγγραμμικές με τουλάχιστον μια από τις υπόλοιπες μεταβλητές. Στη συνέχεια θα πρέπει να υπολογισθεί η ομάδα εκείνων των μεταβλητών που έχουν υψηλή πολυσυγγραμικότητα και μια ή περισσότερες από αυτές τις μεταβλητές μέσα στη συγκεκριμένη ομάδα θα πρέπει να απομακρυνθούν πριν προχωρήσει κανείς σε ανάλυση παλινδρόμησης για την αρχική εξαρτημένη μεταβλητή.

Συνοψίζοντας να αναφέρουμε πως το πρόβλημα της πολυσυγγραμμικότητας επηρεάζει το εύρος των τιμών του διαστήματος εμπιστοσύνης των παραμέτρων ενός υποδείγματος καθώς και την αξιοπιστία των στατιστικών ελέγχων που διενεργούνται επί των εν λόγω παραμέτρων. Επηρεάζει, επίσης, την ακρίβεια και τη σταθερότητα των λοιπών εκτιμήσεων που λαμβάνουν χώρα επί του θεωρουμένου υποδείγματος. Επιπροσθέτως, δημιουργεί προβλήματα στην ερμηνεία των προκυπτόντων αποτελεσμάτων καθώς επίσης και στον καθορισμό του υποδείγματος.

### 2.3.4. ΣΦΑΛΜΑ ΕΞΕΙΔΙΚΕΥΣΗΣ

Συνήθως ο όρος σφάλμα εξειδίκευσης (specification error) αναφέρεται στα σφάλματα που δημιουργούνται από λαθεμένη διατύπωση της εξίσωσης παλινδρόμησης (παραλείπεται από το υπόδειγμα μια σημαντική ερμηνευτική μεταβλητή) ή στη χρησιμοποίηση λαθεμένης μορφής συνάρτησης (π.χ. γραμμική αντί εκθετική).

Ο γενικός έλεγχος διερεύνησης των σφαλμάτων εξειδίκευσης ενός υποδείγματος είναι ο έλεγχος RESET (Regression Specification Error Test) που προτάθηκε από τον Ramsey (1969). Ο έλεγχος αυτός χρησιμοποιεί τα τετράγωνα των εκτιμημένων τιμών της εξαρτημένης μεταβλητής και ακολουθεί τα παρακάτω βήματα: (Ανδρικόπουλος Α., 2003)

**ΒΗΜΑ 1**

Γράφω τις δύο υποθέσεις για την κανονικότητα των καταλοίπων:

*Ηο: Το υπόδειγμα είναι σωστά εξειδικευμένο.*

*Ηα: Το υπόδειγμα δεν είναι σωστά εξειδικευμένο.*

Ο έλεγχος για την εξειδίκευση του υποδείγματος γίνεται με την F κατανομή (έλεγχος του Wald) καθώς και με $X^2$ κατανομή (έλεγχος του λόγου πιθανοφανειών).

**ΒΗΜΑ 2**

Σχηματίζοντας την F κατανομή βρίσκω το κρίσιμο σημείο για επίπεδο σημαντικότητας 5% και βαθμούς ελευθερίας $\nu_1 = h$ και $\nu_2 = \eta - (\kappa + 1 + h)$ ή την $X^2$ κατανομή και βρίσκω το κρίσιμο σημείο για επίπεδο σημαντικότητας 5% και βαθμούς ελευθερίας $\nu = h$.

**ΒΗΜΑ 3**

Εκτιμούμε το υπόδειγμα με τη μέθοδο των ελαχίστων τετραγώνων και σώζουμε τις εκτιμημένες τιμές καθώς και τον συντελεστή προσδιορισμού, όπως και την τιμή της πιθανοφάνειας $\lambda_1$.

**ΒΗΜΑ 4**

Εισάγουμε στο υπόδειγμα παλινδρόμησης h επιπλέον ερμηνευτικές μεταβλητές οι οποίες αποτελούν δυνάμεις της εκτιμημένης μεταβλητής, και εκτιμούμε το νέο υπόδειγμα και σημειώνουμε το νέο συντελεστή προσδιορισμού, καθώς και τη νέα τιμή της πιθανοφάνειας $\lambda_2$.

**ΒΗΜΑ 5**

Με τον έλεγχο του Wald ελέγχουμε αν οι επιπλέον ερμηνευτικές μεταβλητές είναι σημαντικές. Για τον έλεγχο αυτό χρησιμοποιούμε το στατιστικό F και την παρακάτω ποσότητα:

$$F = \frac{\frac{(R_2^2 - R_1^2)}{h}}{\frac{(1 - R_2^2)}{[n - (k + 1 + h)]}}$$

Ενώ για τον έλεγχο του λόγου πιθανοφανειών, υπολογίζουμε το στατιστικό LR από την παρακάτω ποσότητα: LR = -2( $\lambda_1 - \lambda_2$).

**ΒΗΜΑ 6**

Αν ισχύει:

$$F > F_{pin}[h, n - (k + 1 + h)]$$

Τότε απορρίπτω την $H_0$.

Ομοίως, απορρίπτω την $H_0$ και αν:

$$LR > X_{pin}^2(h)$$

## 2.3.5. ΟΛΟΚΛΗΡΩΣΗ - ΣΥΝΟΛΟΚΛΗΡΩΣΗ

Η παλινδρόμηση μη στάσιμων χρονολογικών σειρών μπορεί να οδηγήσει σε στατιστικά αξιόπιστα συμπεράσματα αν οι σειρές είναι συνολοκληρωμένες. (Maddala, G.S., 2005)

Ολοκλήρωση: η μετατροπή μίας μη-στάσιμης σειράς σε στάσιμη επιτυγχάνεται αν εκφράσουμε την σειρά σε διαφορές. Αν η σειρά πρέπει να εκφραστεί d φορές σε διαφορές για να γίνει στάσιμη λέμε ότι είναι ολοκληρωμένη σε d βαθμό και συμβολίζεται:

$$Y_t \rightarrow I(d)$$

Μία στάσιμη σειρά είναι I(0).

Αν σε ένα οικονομετρικό υπόδειγμα υπάρχουν δύο χρονολογικές σειρές ολοκληρωμένες με βαθμό ολοκλήρωσης 1, δηλαδή:

$$Y_t \sim I(1) \text{ και } X_t \sim I(1)$$

Τότε οι χρονολογικές σειρές είναι συνολοκληρωμένες αν τα κατάλοιπα από την παλινδρόμηση της μίας στην άλλη είναι στάσιμα δηλαδή αν:

$$e_t \sim I(0)$$

Ο έλεγχος για συνολοκλήρωση βασίζεται ακριβώς στην εξέταση της ύπαρξης μοναδιαίας ρίζας στα κατάλοιπα από την παλινδρόμηση των δύο σειρών (Engle-Granger test).

## 2.4. ΠΡΟΒΛΕΨΕΙΣ

Οι προβλέψεις αποτελούν αναπόσπαστο κομμάτι της οικονομετρικής ανάλυσης και αφορούν στον τρόπο με τον οποίο με τη χρήση του κατάλληλου υποδείγματος θα καταφέρει ο ερευνητής/ μελετητής να προβλέψει την τιμή του υποδείγματος στο μέλλον. (Ανδρικόπουλος Α., 2003)

Η πρόβλεψη των μελλοντικών τιμών μιας χρονολογικής σειράς βασίζεται στα αποτελέσματα του εκτιμώμενου υποδείγματος. Έτσι, μια πρόβλεψη που κάνουμε για την περίοδο T + 1 εκφράζεται ως:

$$\hat{Y}_{T+1} = E(y_{T+1} / y_{1,} \dots, y_T)$$

Επομένως, η αναμενόμενη τιμή την χρονολογικής μας σειράς για την περίοδο T+1 στηρίζεται στην πληροφόρηση μέχρι την περίοδο T.

Για την πραγματοποίηση προβλέψεων υπάρχουν και κάποιοι εκτιμητές τους οποίους θα πρέπει να λαμβάνουμε υπόψη μας, οι σημαντικότεροι εκ των οποίων είναι: (Greene W.H, 2011)

**Μέσο Σφάλμα Τετραγώνου (Mean Square Error):**

$$\boldsymbol{MSE} = \frac{\mathbf{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \hat{\boldsymbol{\varepsilon}}_{\boldsymbol{t}}^{\mathbf{2}}$$

**Τετραγωνική ρίζα του (Mean Square Error):**

$$\boldsymbol{RMSE} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \hat{\boldsymbol{\varepsilon}}_{\boldsymbol{t}}^{\mathbf{2}}}$$

**Μέσο Απόλυτο Σφάλμα (Mean Absolute Error):**

$$\boldsymbol{MSE} = \frac{\mathbf{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} |\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{t}}|$$

## 2.5. ΕΦΑΡΜΟΓΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Στην παρούσα ενότητα παρουσιάζονται τα αποτελέσματα από την ανάλυση παλινδρόμησης σε οικονομικά δεδομένα.

Προκειμένου να κάνουμε τη συγκεκριμένη ανάλυση αντλήσαμε δεδομένα από βάση δεδομένων της Google. Πιο συγκεκριμένα, το μονοπάτι που ακολουθήθηκε προκειμένου να αντλήσουμε τα δεδομένα είναι το εξής: UCI machine learning repository και η ηλεκτρονική διεύθυνση https://archive.ics.uci.edu/ml/datasets.html .

Στη συνέχεια επιλέξαμε τη βάση δεδομένων Istanbul Data Set, στη συγκεκριμένη βάση περιέχονται τιμές για 9 χρηματιστηριακούς δείκτες από τον Ιανουάριου του 2009 μέχρι τον Φεβρουάριου του 2011, τα δεδομένα είναι ημερήσια.

| **Ονομασία Μεταβλητής** | **Εξήγηση Μεταβλητής** |
|---|---|
| TL BASED ISE | Istanbul exchange national index |
| USD BASED ISE | Istanbul stock exchange national 100 index |
| SP | Standard & poor's 500 return index |

| | |
|---|---|
| DAX | Stock market return index of Germany |
| FTSE | Stock market return index of UK |
| NIKKEI | Stock market return index of Japan |
| BOVESPA | Stock market return index of Brazil |
| EU | MSCI European index |
| EM | MSCI emerging markets index |

Η ανάλυση που πραγματοποιήσαμε είναι ανάλυση πολλαπλής παλινδρόμησης, χρησιμοποιώντας κάθε φορά ως εξαρτημένη μεταβλητή έναν δείκτη και ανεξάρτητες τους υπόλοιπους δείκτες. Από τα αποτελέσματα που προκύπτουν, μπορούμε να βρούμε τη σχέση μεταξύ των μεταβλητών αλλά και το ποσοστό μεταβλητότητας του δείκτη που ερμηνεύεται από τους άλλους δείκτες.

### 2.5.1. ΕΞΑΡΤΗΜΕΝΗ ΜΕΤΑΒΛΗΤΗ: TL BASED ISE

| | *Συντελεστές* | *Τυπικό σφάλμα* | *t* | *τιμή-P* |
|---|---|---|---|---|
| Τεταγμένη επί την αρχή | 0,000535522 | 0,000215059 | 2,490117026 | 0,013077115 |
| USD BASED ISE | 0,790764958 | 0,015683248 | 50,42099473 | 8,8037E-204 |
| SP | -0,025442127 | 0,025431461 | -1,000419401 | 0,317566707 |
| DAX | 0,066880235 | 0,043542088 | 1,53599052 | 0,125140838 |
| FTSE | 0,111985228 | 0,054899278 | 2,03983061 | 0,041865315 |
| NIKKEI | -0,091821092 | 0,018426242 | -4,983169737 | 8,49692E-07 |
| BOVESPA | 0,082052188 | 0,02430329 | 3,376176093 | 0,00078924 |
| EU | -0,121577109 | 0,079067961 | -1,537627988 | 0,124739761 |
| EM | -0,229026004 | 0,043481343 | -5,267224728 | 2,02143E-07 |

Στον παραπάνω πίνακα μπορούμε να δούμε τις σχέσεις μεταξύ των μεταβλητών, δηλαδή τις τιμές των συντελεστών $b_i$. Πιο συγκεκριμένα, για θετικούς συντελεστές μπορούμε να αποφανθούμε πως η εξαρτημένη και η ανεξάρτητη μεταβλητή κινούνται προς την ίδια κατεύθυνση (αύξηση της μιας οδηγεί σε αύξηση της άλλης και αντίστροφα). Όταν η τιμή του συντελεστή είναι αρνητική οι μεταβλητές κινούνται προς αντίθετη κατεύθυνση (αύξηση της μιας

οδηγεί σε μείωση της άλλης και αντίστροφα). Η τιμή του συντελεστή $b_i$ δείχνει πόσο μεταβάλλεται η εξαρτημένη μεταβλητή όταν η ανεξάρτητη μεταβάλλεται κατά μια μονάδα.

Στην περίπτωσή μας, η εξαρτημένη μεταβλητή, TL BASED ISE κινείται προς την **ίδια** κατεύθυνση με τις εξής μεταβλητές:

- USD BASED ISE
- DAX
- FTSE
- BOVESPA

Ενώ οι μεταβλητές με τις οποίες κινείται προς **αντίθετη** κατεύθυνση είναι οι εξής:

- SP
- NIKKIE
- EU
- EM.

Προκειμένου να διαπιστώσουμε ποιες από τις συγκεκριμένες μεταβλητές έχουν στατιστικά σημαντική επίδραση στην εξαρτημένη μεταβλητή θα κάνουμε έλεγχο υπόθεσης. Η μηδενική και η εναλλακτική υπόθεση διαμορφώνονται ως εξής:

*$H_0$: Η επίδραση της εν λόγω μεταβλητής ΔΕΝ είναι στατιστικά σημαντική.*

*$H_1$: Η επίδραση της εν λόγω μεταβλητής είναι στατιστικά σημαντική.*

Για να αποδεχτούμε ή να απορρίψουμε τη μηδενική υπόθεση θα χρησιμοποιήσουμε την τιμή – P. Συγκρίνουμε τη συγκεκριμένη τιμή με το επίπεδο σημαντικότητας 5% (0,05). Αναλυτικότερα,

Αν τιμή – P < 0,05 απορρίπτουμε την $H_0$ δηλαδή η επίδραση της υπό εξέτασης μεταβλητής στην εξαρτημένη μεταβλητή είναι στατιστικά σημαντική.

Αν τιμή – P > 0,05 αποδεχόμαστε την $H_0$ δηλαδή η επίδραση της υπό εξέτασης μεταβλητής στην εξαρτημένη μεταβλητή ΔΕΝ είναι στατιστικά σημαντική.

Σύμφωνα με τα παραπάνω και με τον πίνακα που έχει προκύψει από την επεξεργασία στο excel μπορούμε να σημειώσουμε τα ακόλουθα.

Οι μεταβλητές που επηρεάζουν στατιστικά σημαντικά την εξαρτημένη μεταβλητή είναι οι εξής:

- USD BASED ISE
- FTSE
- NIKKEI
- BOVESPA
- EM.

Από την πίνακα που ακολουθεί και συγκεκριμένα χρησιμοποιώντας το συντελεστή προσδιορισμού $R^2$ μπορούμε να αποφανθούμε ότι η συνολική μεταβλητότητα της εξαρτημένης μεταβλητής που οφείλεται στις συγκεκριμένες ανεξάρτητες είναι περίπου 91%, (0.909).

| *Στατιστικά παλινδρόμησης* | |
|---|---|
| Πολλαπλό R | 0,953474647 |
| R Τετράγωνο | 0,909113902 |
| Προσαρμοσμένο R Τετράγωνο | 0,907734227 |
| Τυπικό σφάλμα | 0,004940318 |
| Μέγεθος δείγματος | 536 |

Τέλος, από τον πίνακα ANOVA, ανάλυσης διακύμανσης μπορούμε να κάνουμε έλεγχο στατιστικής σημαντικότητας για όλο το μοντέλο. Δηλαδή μέσα από τον πίνακα ANOVA μπορούμε να κάνουμε έλεγχο F, όπου η μηδενική και η εναλλακτική υπόθεση διαμορφώνονται ως εξής:

*$H_0$: Η ΑΠΟ ΚΟΙΝΟΥ επίδραση των συγκεκριμένων μεταβλητών στο υπόδειγμα ΔΕΝ είναι στατιστικά σημαντική.*

*$H_1$: Η ΑΠΟ ΚΟΙΝΟΥ επίδραση των συγκεκριμένων μεταβλητών στο υπόδειγμα είναι στατιστικά σημαντική.*

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ

| | *βαθμοί ελευθερίας* | *SS* | *MS* | *F* | *Σημαντικότητα F* |
|---|---|---|---|---|---|
| Παλινδρόμηση | 8 | 0,128659316 | 0,016082415 | 658,9333233 | 8,5993E-269 |
| Υπόλοιπο | 527 | 0,012862352 | 2,44067E-05 | | |
| Σύνολο | 535 | 0,141521668 | | | |

Και σε αυτή την περίπτωση η αποδοχή ή απόρριψη της μηδενικής υπόθεσης γίνεται με κριτήριο την τιμή «Σημαντικότητα F».

Αν Σημαντικότητα F < 0,05 απορρίπτουμε την μηδενική υπόθεση ($H_0$), επομένως η από κοινού επίδραση των μεταβλητών είναι στατιστικά σημαντική.

Αν Σημαντικότητα F > 0,05 αποδεχόμαστε την μηδενική υπόθεση ($H_0$), επομένως η από κοινού επίδραση των μεταβλητών ΔΕΝ είναι στατιστικά σημαντική.

### 2.5.2. ΕΞΑΡΤΗΜΕΝΗ ΜΕΤΑΒΛΗΤΗ: USD BASED ISE

| | *Συντελεστές* | *Τυπικό σφάλμα* | *t* | *τιμή-P* |
|---|---|---|---|---|
| Τεταγμένη επί την αρχή | -0,000466769 | 0,000248136 | -1,8811 | 0,060509 |
| SP | 0,036070962 | 0,029255247 | 1,232974 | 0,218135 |
| DAX | -0,106571442 | 0,050010601 | -2,13098 | 0,033553 |
| FTSE | -0,152659006 | 0,063084217 | -2,41992 | 0,015861 |
| NIKKEI | 0,103192938 | 0,021230389 | 4,860624 | 1,55E-06 |
| BOVESPA | -0,127844227 | 0,027718165 | -4,61229 | 5,01E-06 |
| EU | 0,314799587 | 0,090168199 | 3,491248 | 0,000521 |
| EM | 0,410282835 | 0,048133063 | 8,523929 | 1,63E-16 |
| TL BASED ISE | 1,047464611 | 0,020774374 | 50,42099 | 8,8E-204 |

Στον παραπάνω πίνακα μπορούμε να δούμε τις σχέσεις μεταξύ των μεταβλητών, δηλαδή τις τιμές των συντελεστών $b_i$. Πιο συγκεκριμένα, για θετικούς συντελεστές μπορούμε να αποφανθούμε πως η εξαρτημένη και η ανεξάρτητη μεταβλητή κινούνται προς την ίδια κατεύθυνση (αύξηση της μιας οδηγεί σε αύξηση της άλλης και αντίστροφα). Όταν η τιμή του συντελεστή είναι αρνητική οι μεταβλητές κινούνται προς αντίθετη κατεύθυνση (αύξηση της μιας οδηγεί σε μείωση της άλλης και αντίστροφα). Η τιμή του συντελεστή $b_i$ δείχνει πόσο μεταβάλλεται η εξαρτημένη μεταβλητή όταν η ανεξάρτητη μεταβάλλεται κατά μια μονάδα.

Στην περίπτωσή μας, η εξαρτημένη μεταβλητή, USD BASED ISE κινείται προς την **ίδια** κατεύθυνση με τις εξής μεταβλητές:

- SP

- NIKKIE
- EU
- EM
- TL BASED ISE

Ενώ οι μεταβλητές με τις οποίες κινείται προς **αντίθετη** κατεύθυνση είναι οι εξής:

- DAX
- FTSE
- BOVESPA

Προκειμένου να διαπιστώσουμε ποιες από τις συγκεκριμένες μεταβλητές έχουν στατιστικά σημαντική επίδραση στην εξαρτημένη μεταβλητή θα κάνουμε έλεγχο υπόθεσης. Η μηδενική και η εναλλακτική υπόθεση διαμορφώνονται ως εξής:

*$H_0$: Η επίδραση της εν λόγω μεταβλητής ΔΕΝ είναι στατιστικά σημαντική.*

*$H_1$: Η επίδραση της εν λόγω μεταβλητής είναι στατιστικά σημαντική.*

Για να αποδεχτούμε ή να απορρίψουμε τη μηδενική υπόθεση θα χρησιμοποιήσουμε την τιμή – P. Συγκρίνουμε τη συγκεκριμένη τιμή με το επίπεδο σημαντικότητας 5% (0,05). Αναλυτικότερα,

Αν τιμή – P < 0,05 απορρίπτουμε την $H_0$ δηλαδή η επίδραση της υπό εξέτασης μεταβλητής στην εξαρτημένη μεταβλητή είναι στατιστικά σημαντική.

Αν τιμή – P > 0,05 αποδεχόμαστε την $H_0$ δηλαδή η επίδραση της υπό εξέτασης μεταβλητής στην εξαρτημένη μεταβλητή ΔΕΝ είναι στατιστικά σημαντική.

Σύμφωνα με τα παραπάνω και με τον πίνακα που έχει προκύψει από την επεξεργασία στο excel μπορούμε να σημειώσουμε τα ακόλουθα.

Οι μεταβλητές που επηρεάζουν στατιστικά σημαντικά την εξαρτημένη μεταβλητή είναι οι εξής:

- DAX
- FTSE
- NIKKEI
- BOVESPA

- EU
- EM
- TL BASED ISE

Από την πίνακα που ακολουθεί και συγκεκριμένα χρησιμοποιώντας το συντελεστή προσδιορισμού $R^2$ μπορούμε να αποφανθούμε ότι η συνολική μεταβλητότητα της εξαρτημένης μεταβλητής που οφείλεται στις συγκεκριμένες ανεξάρτητες είναι περίπου 93%, (0.9286).

| *Στατιστικά παλινδρόμησης* | |
|---|---|
| Πολλαπλό R | 0,963648034 |
| R Τετράγωνο | 0,928617533 |
| Προσαρμοσμένο R Τετράγωνο | 0,927533928 |
| Τυπικό σφάλμα | 0,005685922 |
| Μέγεθος δείγματος | 536 |

Τέλος, από τον πίνακα ANOVA, ανάλυσης διακύμανσης μπορούμε να κάνουμε έλεγχο στατιστικής σημαντικότητας για όλο το μοντέλο. Δηλαδή μέσα από τον πίνακα ANOVA μπορούμε να κάνουμε έλεγχο F, όπου η μηδενική και η εναλλακτική υπόθεση διαμορφώνονται ως εξής:

*$H_0$: Η ΑΠΟ ΚΟΙΝΟΥ επίδραση των συγκεκριμένων μεταβλητών στο υπόδειγμα ΔΕΝ είναι στατιστικά σημαντική.*

*$H_1$: Η ΑΠΟ ΚΟΙΝΟΥ επίδραση των συγκεκριμένων μεταβλητών στο υπόδειγμα είναι στατιστικά σημαντική.*

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ

| | *βαθμοί ελευθερίας* | *SS* | *MS* | *F* | *Σημαντικότητα F* |
|---|---|---|---|---|---|
| Παλινδρόμηση | 8 | 0,221644861 | 0,02770 6 | 856,970 7 | 2,1E-296 |
| Υπόλοιπο | 527 | 0,017037754 | 3,23E-05 | | |
| Σύνολο | 535 | 0,238682615 | | | |

Και σε αυτή την περίπτωση η αποδοχή ή απόρριψη της μηδενικής υπόθεσης γίνεται με κριτήριο την τιμή «Σημαντικότητα F»

Αν Σημαντικότητα F < 0,05 απορρίπτουμε την μηδενική υπόθεση ($H_0$), επομένως η από κοινού επίδραση των μεταβλητών είναι στατιστικά σημαντική.

Αν Σημαντικότητα F > 0,05 αποδεχόμαστε την μηδενική υπόθεση ($H_0$), επομένως η από κοινού επίδραση των μεταβλητών ΔΕΝ είναι στατιστικά σημαντική.

# ΚΕΦΑΛΑΙΟ 3$^{ο}$: ΔΕΝΔΡΑ ΑΠΟΦΑΣΗΣ ως ΤΕΧΝΙΚΗ ΠΡΟΒΛΕΨΗΣ

## 3.1. ΕΙΣΑΓΩΓΙΚΕΣ ΕΝΝΟΙΕΣ – ΟΡΙΣΜΟΙ - ΧΡΗΣΙΜΟΤΗΤΑ

Στην Θεωρία Λήψης Αποφάσεων τα δεδομένα του κάθε προβλήματος ή περιπτώσεως που απαιτείται η επιλογή κάποιας απόφασης, μπορούν να αναπαρασταθούν γραφικά. (Πραστάκος Γρ., 2006) Αυτή η γραφική αναπαράσταση καθιερώθηκε να ονομάζεται Δέντρο Αποφάσεων (Decision Tree). Ένα δέντρο αποφάσεων ή decision tree είναι ένα υποστηρικτικό εργαλείο λήψης αποφάσεων που χρησιμοποιεί μια γραφική απεικόνιση όμοια της μορφής δέντρου, συμπεριλαμβάνοντας όλες τις πιθανές αποφάσεις, όλους τους παράγοντες επιρροής και όλα τα πιθανά αποτελέσματα.

Ουσιαστικά, τα δένδρα απόφασης αποτελούν μια μέθοδο κατάλληλη για λήψη ορθολογικών αποφάσεων σε συνθήκες αβέβαιου μέλλοντος. Οι βασικοί παράμετροι αυτών που θα πρέπει να λαμβάνονται υπόψη είναι οι στόχοι του αποφασίζοντα, τα τεχνικά και οικονομικά δεδομένα, οι καταστάσεις της φύσης και οι αντίστοιχες πληροφορίες.

Ο σχεδιασμός των δένδρων απόφασης είναι τέτοιος ώστε να βοηθούν του λήπτες αποφάσεων με έναν διαμήκη τρόπο λήψης αποφάσεων, δηλαδή μια απόφαση μπορεί να ληφθεί, ένα πλήθος γεγονότων μπορεί να συμβεί, μια μεταγενέστερη απόφαση μπορεί να κριθεί απαραίτητη με αποτέλεσμα να συμβούν ακόμη περισσότερα γεγονότα. Τα δένδρα απόφασης αποτελούν εργαλείο το οποίο θα πρέπει να βοηθά το λήπτη απόφασης, όταν διαγράφει μια αρχική απόφαση, να αναλογισθεί μια σειρά γεγονότων και επακόλουθων αποφάσεων που ίσως συμβούν. Συχνά τα δένδρα απόφασης μπορούν να χρησιμοποιηθούν και για προβλέψεις. Είναι κατανοητό ότι όλες οι προβλέψεις γίνονται κάτω από τον παράγοντα αβεβαιότητα ή χωρίς να είναι γνωστό το αποτέλεσμα. (Πραστάκος Γρ., 2006)

Στο σημείο αυτό θα αναφερθούμε στα κυριότερα πλεονεκτήματα που προκύπτουν από τη χρήση των δένδρων απόφασης και διαμορφώνονται ως εξής (Βλαχάβας, κ.α., 2002):

- Πρόκειται για μη παραμετρική προσέγγιση, η οποία δεν στηρίζεται σε υπόθεση εκ των προτέρων γνώσης σχετικά με τον τύπο της κατανομής πιθανότητας που ικανοποιεί η κλάση ή τα άλλα γνωρίσματα.
- Η κατασκευή του βέλτιστου δέντρου απόφασης είναι ένα NP - complete πρόβλημα.
- Αποδοτική κατασκευή ακόμα και στην περίπτωση πολύ μεγάλου συνόλου δεδομένων.
- Αφού το δέντρο κατασκευαστεί, η ταξινόμηση νέων εγγραφών πολύ γρήγορη O(h) όπου h το μέγιστο ύψος του δέντρου.
- Είναι εύκολα στην κατανόηση (ιδιαίτερα τα μικρά δέντρα).
- Η ακρίβεια τους είναι συγκρίσιμη με άλλες τεχνικές για μικρά σύνολα δεδομένων.

Από τα παραπάνω γίνεται σαφές πως η χρησιμότητα των δένδρων απόφασης είναι μεγάλη και μπορεί να οδηγήσει τον λήπτη αποφάσεων στην αντιμετώπιση κάποιου προβλήματος απόφασης ή ακόμη και στην επιλογή της ευνοϊκότερης λύσης. Παρόλα αυτά τα δένδρα απόφασης δεν γίνονται αποδεκτά από όλους τους αναλυτές αναφορικά με τη χρησιμότητά τους, χαρακτηρίζοντάς τα μικρής αξίας, ή δύσκολης εφαρμογής σε πραγματικές καταστάσεις. Συνήθως, η κριτική που τους ασκείται εστιάζει στα ακόλουθα σημεία (Πραστάκος Γρ., 2006):

- Δυσκολία εκτίμησης των πιθανοτήτων που αφορούν τα διάφορα πιθανά σενάρια για το μέλλον δεδομένου ότι η λήψη απόφασης γίνεται υπό καθεστώς αβεβαιότητας.
- Δυσκολία αναπαράστασης του προβλήματος με τρόπο που να ανταποκρίνεται και να απεικονίζει τις πραγματικές συνθήκες στις οποίες εμφανίζεται.
- Προβλήματα με πολλούς παράγοντες επιρροής και πολλές εναλλακτικές αποφάσεις χαρακτηρίζονται από έντονη πολυπλοκότητα.

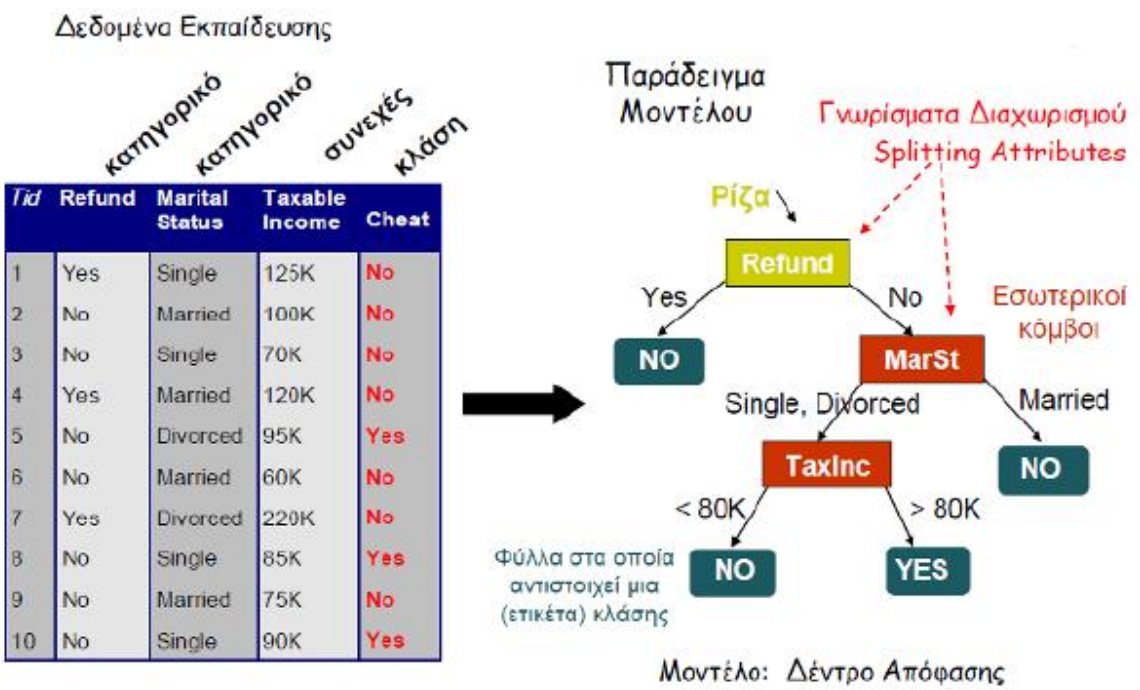
## 3.2. ΔΟΜΗ ΤΩΝ ΔΕΝΔΡΩΝ ΑΠΟΦΑΣΗΣ

Πριν αναλύσουμε τη δομή των δένδρων απόφασης κρίνουμε σκόπιμο να αναφέρουμε τα δομικά χαρακτηριστικά αυτών (Βλαχάβας, κ.α., 2002).

Έστω $D = \{t_1, \dots, t_n\}$, μια βάση δεδομένων, όπου $t_i = [t_{i1}, \dots, t_{ih}]$ είναι μια περίπτωση η οποία χαρακτηρίζεται από τα γνωρίσματα $\{A_1, A_{2,} \dots A_h\}$, επιπλέον δίνεται ένα σύνολο από κατηγορίες $C = \{C_1, \dots, C_m\}$. Ένα δένδρο απόφασης ή δένδρο κατηγοριοποίησης όπως αλλιώς ονομάζεται:

- *Κόμβος Ρίζα,* που δεν έχει εισερχόμενες ακμές και μηδέν ή περισσότερες εξερχόμενες.
- *Εσωτερικός Κόμβος,* παίρνει το όνομά του από ένα γνώρισμα $A_i$, και έχει ακριβώς μια εισερχόμενη ακμή και δύο ή πιο περισσότερες εξερχόμενες.
- *Ακμή,* παίρνει το όνομά της από ένα κατηγόρημα – τιμή, το οποίο εφαρμόζεται στο γνώρισμα που συνδέεται με τον πατέρα – κόμβο.
- *Φύλλο ή Τερματικός Κόμβος:* έχει ως όνομα μια κατηγορία $C_i$, έχει ακριβώς μια εισερχόμενη ακμή και καμία εξερχόμενη.

Τα γνωρίσματα που χρησιμοποιούνται για να ονοματίσουν τους κόμβους του δένδρου και γύρω από τα οποία θα λάβουν χώρα διαιρέσεις καλούνται γνωρίσματα διαχωρισμού (splitting attributes). Τα κατηγορήματα που χρησιμοποιούνται για να ονοματίσουν τις ακμές του δένδρου καλούνται κατηγορήματα διαχωρισμού (splitting predicates).




Εικόνα 1: Μοντέλο Δένδρο Απόφασης

Η διαδικασία κατασκευής του δένδρου απόφασης συνοψίζεται στα ακόλουθα βήματα (Πραστάκος Γρ., 2006):

- Έναρξη από έναν κόμβο που περιέχει όλες τις εγγραφές.
- Διάσπαση του κόμβου με βάση μια συνθήκη – διαχωρισμού σε κάποιο από τα γνωρίσματα. Επιλέγεται το «καλύτερο» γνώρισμα διαχωρισμού (να σημειώσουμε πως οι αλγόριθμοι διαφέρουν στην επιλογή του «καλύτερου» γνωρίσματος και των αντίστοιχων «καλύτερων» κατηγορημάτων).
- Αναδρομική κλήση του δεύτερου βήματος.
- Ολοκλήρωση της διαδικασίας όταν ικανοποιηθεί κάποιο κριτήριο τερματισμού (οι αλγόριθμοι διαφοροποιούνται στην επιλογή του εν λόγω κριτηρίου).
- Εκτέλεση κλαδέματος του δένδρου (tree pruining) για βελτίωση της επίδοσης.

## 3.3. ΑΛΓΟΡΙΘΜΟΙ ΒΑΣΙΣΜΕΝΟΙ ΣΤΑ ΔΕΝΔΡΑ ΑΠΟΦΑΣΗΣ

Επιθυμητό είναι να δημιουργούνται δέντρα που είναι ισορροπημένα και με τα λιγότερα επίπεδα (μικρότερο βάθος). Αυτό όμως δεν είναι πάντα εφικτό ούτε το υπολογιστικά φτηνότερο. Παρόλα αυτά να αναφέρουμε πως υπάρχουν αλγόριθμοι οι οποίοι δημιουργούν μόνο δυαδικά δέντρα.

Στο σημείο αυτό, πριν μελετήσουμε κάποιους αλγόριθμους εκτενέστερα, αξίζει να αναφέρουμε τους παράγοντες που επηρεάζουν την απόδοση ενός αλγορίθμου μάθησης αλλά και ορισμένα βασικά γνωρίσματα αυτών (J. Ross Quinlan, Morgan Kaufmann, 1993):

- **Επιλογή των γνωρισμάτων διαχωρισμού**
- **Διάταξη των γνωρισμάτων διαχωρισμού**
- **Δομή του δένδρου**
- **Κριτήρια τερματισμού**
- **Δεδομένα εκπαίδευσης:** Η δομή ενός ΔΑ εξαρτάται από τα δεδομένα εκπαίδευσης. Αν το σύνολο εκπαίδευσης είναι πολύ μικρό, τότε το δέντρο μπορεί να μην είναι τόσο λεπτομερές, ώστε να ταξινομεί γενικότερα δεδομένα. Αν είναι πολύ μεγάλο, το δέντρο πιθανόν να υπερπροσαρμόζεται (overfits). (Βλαχάβας, κ.α., 2002)

- **Κλάδεμα:** Μετά τη δημιουργία ενός ΔΑ μπορεί να χρειάζονται τροποποιήσεις για να βελτιώσουν την απόδοσή του, όπως π.χ. το κλάδεμα πλεοναζόντων συγκρίσεων ή υποδέντρων.

Εστιάζοντας στα κριτήρια τερματισμού μπορούμε να υπογραμμίσουμε τα ακόλουθα:

- Η δημιουργία ενός δέντρου σταματά οπωσδήποτε όταν όλα τα δεδομένα του (εναπομείναντος) συνόλου εκπαίδευσης κατηγοριοποιούνται πλήρως.
- Μπορεί όμως να είναι απαραίτητο να σταματήσει νωρίτερα για να αποφευχθούν π.χ. μεγάλα δέντρα. Το πότε ή πού θα σταματήσει είναι θέμα συναλλαγής (trade-off) μεταξύ ακρίβειας (accuracy) και απόδοσης (performance) του αλγορίθμου.
- Επίσης, πρώιμος τερματισμός μπορεί να γίνει για αποφυγή του φαινομένου της υπερπροσαρμογής (overfitting).
- Τέλος, μπορεί να προχωρήσει σε μεγαλύτερα δέντρα αν είναι γνωστό ότι υπάρχουν κατηγορίες δεδομένων που δεν αντιπροσωπεύονται στο σύνολο εκπαίδευσης.

Μια έννοια που συχνά σχετίζεται με τα δένδρα απόφασης είναι και αυτή της πολυπλοκότητας, τόσο σε επίπεδο χρόνου όσο και σε επίπεδο χώρου. Ειδικότερα, η πολυπλοκότητα χρόνου και χώρου των αλγορίθμων δένδρων απόφασης εξαρτώνται από το μέγεθος του συνόλου εκπαίδευσης k, τον αριθμό των χαρακτηριστικών διάσπασης n και το σχήμα του δένδρου απόφασης. Στη χειρότερη περίπτωση το δένδρο απόφασης είναι βαθύ και μη ισορροπημένο. Η πολυπλοκότητα χρόνου για τη δημιουργία ενός δένδρου απόφασης είναι O(n*k*logk).

Η πολυπλοκότητα χρόνου κατηγοριοποίησης μιας βάσης n παραδειγμάτων εξαρτάται από το ύψος του δένδρου απόφασης και είναι O(n*logk), υποθέτοντας πολυπλοκότητα για το ύψος O(logk). (Βλαχάβας, κ.α., 2002)

Παρακάτω αναλύονται τρεις από τους βασικότερους και πιο συχνά χρησιμοποιούμενους αλγορίθμους.

### 3.3.1. ΑΛΓΌΡΙΘΜΟΣ HUNT

Τα βασικότερα βήματα του συγκεκριμένου αλγορίθμου διαμορφώνονται ως εξής (J. Ross Quinlan, Morgan Kaufmann, 1993):

Κτίζει το δέντρο αναδρομικά, αρχικά όλες οι εγγραφές σε έναν κόμβο (ρίζα) Dt: το σύνολο των εγγραφών εκπαίδευσης που έχουν φτάσει στον κόμβο t.

Η Γενική Διαδικασία που ακολουθείται (αναδρομικά σε κάθε κόμβο) περιγράφεται με τα ακόλουθα βήματα:

- Αν το $D_t$ περιέχει εγγραφές που ανήκουν στην ίδια κλάση $y_t$, τότε ο κόμβος t είναι κόμβος φύλλο με ετικέτα $y_t$.
- Αν $D_t$ είναι το κενό σύνολο (αυτό σημαίνει ότι δεν υπάρχει εγγραφή στο σύνολο εκπαίδευσης με αυτό το συνδυασμό τιμών), τότε $D_t$ γίνεται φύλλο με κλάση αυτή της πλειοψηφίας των εγγραφών εκπαίδευσης ή ανάθεση κάποιας default κλάσης.
- Αν το $D_t$ περιέχει εγγραφές που ανήκουν σε περισσότερες από μία κλάσεις, τότε χρησιμοποίησε έναν έλεγχο-γνωρίσματος για το διαχωρισμό των δεδομένων σε μικρότερα υποσύνολα.

Ο καθορισμός των συνθηκών του ελέγχου για τα γνωρίσματα εξαρτάται από τον τύπο των γνωρισμάτων – μεταβλητών (Πραστάκος Γρ., 2006):

- Δ**ιακριτές – Nominal:** είναι μεταβλητές οι οποίες παίρνουν μόνο "μεμονωμένες" αριθμητικές τιμές, είναι δηλαδή στοιχεία ενός συνόλου τα οποία μπορούν να αντιστοιχηθούν ένα προς ένα με στοιχεία του συνόλου των θετικών ακέραιων αριθμών. Τέτοια δεδομένα είναι π.χ. ο αριθμός των παιδιών σε μία οικογένεια, ο αριθμός των δωματίων μιας κατοικίας, κ.ά.
- Δ**ιατεταγμένες – Ordinal:** μεταβλητές οι τιμές των οποίων διατάσσονται σε αύξουσα ή φθίνουσα σειρά

- **Συνεχείς – Continuous:** είναι μεταβλητές οι οποίες παίρνουν αριθμητικές τιμές που καλύπτουν ολόκληρο διάστημα τιμών των πραγματικών αριθμών (α, β), όπου

  $-\infty<\alpha<\beta<+\infty$ Π.χ. η ηλικία, η διάρκεια μιας τηλεφωνικής συνδιάλεξης, η θερμοκρασία κλπ.

Τα είδη διαχωρισμού που προκύπτουν είναι τα εξής:

- 2-αδικός διαχωρισμός - 2-way split
- Πολλαπλός διαχωρισμός - Multi-way split
- Διαχωρισμός βασισμένος σε διακριτές τιμές.

Στον πολλαπλό διαχωρισμό χρησιμοποιούνται τόσες διασπάσεις όσες οι διαφορετικές τιμές, στο Δυαδικό Διαχωρισμό χωρίζονται οι τιμές σε δύο υποσύνολα και ο αλγόριθμος πρέπει να βρει το βέλτιστο διαχωρισμό (partitioning).

Σε κάθε περίπτωση, όταν υπάρχει διάταξη, πρέπει οι διασπάσεις να μη την παραβιάζουν (Πραστάκος Γρ., 2006).

Δεδομένου ότι σε κάθε επίπεδο, υπάρχουν πολλές διαφορετικές δυνατότητες για την διάσπαση, προκειμένου να δούμε ποια θα επιλέξουμε, ορίζουμε ένα κριτήριο για την «ποιότητα» ενός κόμβου. Το κριτήριο αυτό διαμορφώνεται ως εξής:

- Έστω μια διάσπαση ενός κόμβου (parent) με N εγγραφές σε k παιδιά $u_i$.
- Έστω $N(u_i)$ ο αριθμός εγγραφών κάθε παιδιού ($\sum N(u_i) = N$).
- Κοιτάμε το κέρδος, δηλαδή τη διαφορά μεταξύ της ποιότητας του γονέα (πριν τη διάσπαση) και το «μέσο όρο» της ποιότητας των παιδιών του (μετά τη διάσπαση).
- Διαλέγουμε τη διάσπαση με το μεγαλύτερο κέρδος (μεγαλύτερο Δ).

Όπου Δ:

$$\Delta = I(parent) - \sum_{i=1}^{k} \frac{N(u_i)}{N} I(u_i)$$

Βάρος (εξαρτάται από τον αριθμό εγγραφών)

Να αναφέρουμε στο σημείο αυτό ότι το I αφορά στην καθαρότητα του κάθε κόμβου, δηλαδή πόσο ομογενοποιημένοι (εγγραφές από την ίδια κλάση) υπάρχουν στον κόμβο.

Επιπλέον, υπογραμμίζουμε την ύπαρξη μέτρων μη καθαρότητας (ή μέτρα ανομοιότητας όπως λέγονται) τα κυριότερα εκ των οποίων είναι τα ακόλουθα:

**1. Ευρετήριο Gini (Gini Index)**

$$GINI\ (t) = 1 - \sum_{j=1}^{c} [p(j/t)]^2$$

Όπου $[p(j/t)]$ η σχετική συχνότητα της κλάσης j στον κόμβο t (ποσοστό εγγραφών της κλάσης j στον κόμβο t και c ο αριθμός των κλάσεων).

**2. Εντροπία (Entropy)**

$$Entropy\ (t) = -\sum_{j=1}^{c} [p(j/t)] * \log_2 [p(j/t)]$$

Όπου $[p(j/t)]$ η σχετική συχνότητα της κλάσης j στον κόμβο t (ποσοστό εγγραφών της κλάσης j στον κόμβο t και c ο αριθμός των κλάσεων).

| N1 | |
|---|---|
| C1 | 0 |
| C2 | 6 |
| Entropy = 0.0000 | |
| Gini = 0.0000 | |

| N2 | |
|---|---|
| C1 | 1 |
| C2 | 5 |

| |
|---|
| Entropy = 0.650 |
| Gini = 0.278 |

| N3 | |
|---|---|
| C1 | 2 |
| C2 | 4 |
| Entropy = 0.92 | |
| Gini = 0.444 | |

| N4 | |
|---|---|
| C1 | 3 |
| C2 | 3 |
| Entropy = 1.0000 | |
| Gini = 0.500 | |

Ουσιαστικά η εντροπία μετράει την ομοιογένεια ενός κόμβου.

Μέγιστη τιμή log(c) όταν όλες οι εγγραφές είναι ομοιόμορφα κατανεμημένες στις κλάσεις. Ελάχιστη τιμή (0.0) όταν όλες οι εγγραφές ανήκουν σε μία κλάση. Όταν χρησιμοποιούμε την εντροπία για τη μέτρηση της μη καθαρότητας τότε η διαφορά καλείται κέρδος πληροφορίας (information gain).

Η εντροπία τείνει να ευνοεί διαχωρισμούς που καταλήγουν σε μεγάλο αριθμό από διασπάσεις που η κάθε μία είναι μικρή αλλά καθαρή.

**3. Λάθος ταξινομήσεις (Misclassification error)**

$$Error(t) = 1 - max_{class\ i} P(i/t)$$

Ο ψευδοκώδικας που χρησιμοποιείται στον αλγόριθμο Hunt έχει ως εξής:

1. Δημιουργήστε έναν κόμβο N.

2. Εάν όλα τα δείγματα είναι της ίδιας κατηγορίας C, στη συνέχεια την ονόμασε το N με C; Τερματισμός;

3. Αν το A είναι άδειο ονόμασε το N με την πιο κοινή κατηγορία C σε S (πλειοψηφία); Τερματισμός;

4. Επιλέξτε α ∈ A, με το υψηλότερο κέρδος; Ονόμασε το N με α;

5. Για κάθε τιμή του ν του α:

α. δημιούργησε ένα κλαδί από τον N με την συνθήκη a=v;

β. Έστω Sv είναι το υποσύνολο των δειγμάτων στο S με a=v;

γ. Αν Sv είναι άδειο μετά επισύναψε ένα φύλλο με την πιο κοινή κατηγορία στο S;

δ. Αλλιώς συνδέστε τον κόμβο που δημιουργείται από Gen Dec Tree (Sv, A-a)

Ο αλγόριθμος που είδαμε χρησιμοποιεί μια greedy, top-down, αναδρομική διάσπαση για να φτάσει σε μια αποδεκτή λύση.

### 3.3.2. ΑΛΓΟΡΙΘΜΟΣ ID3

Ο αλγόριθμος ID3 είναι γνωστός και σαν αλγόριθμος κατασκευής δένδρων απόφασης με επαγωγή (decision tree induction algorithm) από δεδομένα εκπαίδευσης. Το αποτέλεσμα είναι μία δενδροειδής δομή που με γραφικό τρόπο αναπαριστά τις συσχετίσεις στα δεδομένα εκπαίδευσης ή διαφορετικά, περιγράφει τα δεδομένα. Αρχικά, μία από τις παραμέτρους του συνόλου εκπαίδευσης ορίζεται ως παράμετρος - στόχος (εξαρτημένη μεταβλητή ή μεταβλητή που μοντελοποιείται). Οι υπόλοιπες παράμετροι θεωρούνται παράμετροι εισόδου (ανεξάρτητες μεταβλητές). (J. Ross Quinlan, Morgan Kaufmann, 1993)

Μια σύντομη περιγραφή του αλγορίθμου μπορεί να είναι η ακόλουθη:

1) Βρες την ανεξάρτητη μεταβλητή η οποία αν χρησιμοποιηθεί ως κριτήριο διαχωρισμού των δεδομένων εκπαίδευσης θα οδηγήσει σε κόμβους κατά το δυνατό διαφορετικούς σε σχέση με την εξαρτημένη μεταβλητή.

2) Κάνε το διαχωρισμό.

3) Επανέλαβε τη διαδικασία για κάθε έναν από τους κόμβους που προέκυψαν μέχρι να μην είναι δυνατός περαιτέρω διαχωρισμός.

Το βασικότερο στάδιο αυτού είναι η επιλογή της ανεξάρτητης μεταβλητής πάνω στην οποία θα συνεχιστεί η ανάπτυξη του δένδρου (βήμα 1). Ο ID3 απαιτεί τον ορισμό κάποιου ευριστικού μηχανισμού ο οποίος θα καθοδηγήσει την αναζήτηση προς το καλύτερο δένδρο (περιγραφή) μέσα στο σύνολο των δυνατών δένδρων. Ουσιαστικά ο ID3 είναι ένας αλγόριθμος αναρρίχησης λόφων καθώς σε κάθε κύκλο λειτουργίας επεκτείνει το τρέχον δένδρο με τον τοπικά καλύτερο τρόπο και συνεχίζει χωρίς δυνατότητα οπισθοδρόμησης.

Να σημειώσουμε επίσης ότι ο πιο διαδεδομένος ευριστικός μηχανισμούς διαχωρισμού είναι αυτός της εντροπίας της πληροφορίας (information entropy) ο οποίος επιλέγει εκείνη την ανεξάρτητη μεταβλητή που οδηγεί σε περισσότερο συμπαγές δένδρο.

Η εντροπία της πληροφορίας μετρά ουσιαστικά την ανομοιογένεια που υπάρχει στο σύνολο των δεδομένων εκπαίδευσης στο στάδιο (κόμβο) του διαχωρισμού αναφορικά με την υπό εξέταση εξαρτημένη μεταβλητή και έχει τις ρίζες της στη θεωρία πληροφοριών (information theory).

Ο συγκεκριμένος αλγόριθμος, προτιμά τα μικρότερα δέντρα από τα μεγαλύτερα και τοποθετεί χαρακτηριστικά με υψηλό κέρδος πληροφορίας κοντύτερα στη ρίζα. Είναι αλγόριθμος αναζήτησης τύπου Hill Climbing, που προχωρά από τα απλά στα σύνθετα ξεκινώντας από το κενό δέντρο.

Ψάχνει στον πλήρη χώρο των υποθέσεων (όλων των πιθανών δέντρων), διατηρεί μόνο μια υπόθεση κάθε φορά ενώ συνήθως δεν κάνει οπισθοδρόμηση (backtracking), δηλ. δεν αναθεωρεί προηγούμενη απόφαση / επιλογή (κίνδυνος τοπικού βέλτιστου). Ο εν λόγω αλγόριθμος χρησιμοποιεί όλα τα δεδομένα εκπαίδευσης (λιγότερο ευαίσθητος σε λάθη) και δεν φτάνει σε αποφάσεις αυξητικά, δηλ. βασιζόμενος σε ατομικά δεδομένα.

Τα παραδείγματα (δεδομένα) αναφέρονται σε ένα συγκεκριμένο σύνολο χαρακτηριστικών και τις τιμές τους, που είναι διακριτές και, κατά προτίμηση, λίγες. Ο χειρισμός των μεταβλητών με πραγματικές τιμές απαιτεί επέκταση του βασικού αλγορίθμου.

Τα δεδομένα εκπαίδευσης μπορεί να περιέχουν λάθη. Ο ID3 είναι ανεκτικός στα λάθη (και στην κατηγοριοποίηση και στις τιμές των χαρακτηριστικών). Από τα δεδομένα εκπαίδευσης μπορεί να λείπουν τιμές για ορισμένα χαρακτηριστικά. Αυτό μπορεί να αντιμετωπιστεί.

Επιπρόσθετα, ο αλγόριθμος ID3 απαιτεί οι τιμές των μεταβλητών να είναι διακριτές καθώς και τον ορισμό κατηγοριών αλλά και τη μετατροπή των συνεχών αριθμητικών τιμών σε διακριτές. Υπογραμμίζουμε στο σημείο αυτό πως ο ορισμός κατηγοριών εισάγει υποκειμενικότητα που επηρεάζει την τελική μορφή του δένδρου (υπάρχουν πολλοί τρόποι με τους οποίους μπορούν να οριστούν οι κατηγορίες).

Τέλος, παραλλαγές του ID3 περιλαμβάνουν τεχνικές κλαδέματος πριν την ολοκλήρωση της κατασκευής του δένδρου, διαχείριση πεδίων χωρίς τιμή, χρήση διαφόρων κριτηρίων διαχωρισμού, αυτόματη διαχείριση συνεχόμενων αριθμητικών τιμών, κλπ. Ο αλγόριθμος C4.5 που περιγράφεται στην ενότητα που ακολουθεί αποτελεί την περισσότερο διαδεδομένη βελτίωση του ID3.

**Παράδειγμα Εφαρμογής ID3**

- Θέλουμε δέντρο απόφασης για διάρκεια άθλησης στην ύπαιθρο ανάλογα με τον καιρό.
- Χαρακτηριστικά (FS):
  - ουρανός = {καθαρός, συννεφιά, βροχή}
  - θερμοκρασία = {υψηλή, μέτρια, χαμηλή}
  - υγρασία = {υψηλή, κανονική}
  - άνεμος = {δυνατός, αδύναμος}
- Κατηγορίες (C):
  - διάρκεια άθλησης = {μικρή, κανονική, καμία}

| Δείγμα | Ουρανός | Θερμοκρασία | Υγρασία | Άνεμος | Διάρκεια |
|---|---|---|---|---|---|
| T1 | καθαρός | υψηλή | υψηλή | αδύναμος | μικρή |
| T2 | καθαρός | υψηλή | υψηλή | δυνατός | μικρή |
| T3 | συννεφιά | υψηλή | υψηλή | αδύναμος | κανονική |
| T4 | βροχή | μέτρια | υψηλή | αδύναμος | καμία |
| T5 | βροχή | χαμηλή | κανονική | αδύναμος | καμία |
| T6 | βροχή | χαμηλή | κανονική | δυνατός | καμία |
| T7 | συννεφιά | χαμηλή | κανονική | δυνατός | κανονική |
| T8 | καθαρός | μέτρια | υψηλή | αδύναμος | μικρή |
| T9 | καθαρός | χαμηλή | κανονική | αδύναμος | κανονική |
| T10 | βροχή | μέτρια | κανονική | δυνατός | καμία |
| T11 | καθαρός | μέτρια | κανονική | δυνατός | κανονική |
| T12 | συννεφιά | μέτρια | υψηλή | αδύναμος | κανονική |
| T13 | συννεφιά | υψηλή | κανονική | αδύναμος | κανονική |
| T14 | βροχή | μέτρια | υψηλή | δυνατός | καμία |

- Παραδείγματα υπολογισμών τιμών:
  - Εντροπία του συνόλου δεδομένων:

$$E(S) = \sum_{i=1}^{n} p_i * log_2 \frac{1}{p_i} = \frac{3}{14} log_2 \frac{14}{3} + \frac{6}{14} log_2 \frac{14}{6} + \frac{5}{14} log_2 \frac{14}{5}$$

Όπου:

- 3/14: η πιθανότητα να είναι η διάρκεια μικρή
- 6/14: η πιθανότητα να είναι κανονική
- 5/14: η πιθανότητα να είναι μηδενική

- Παραδείγματα υπολογισμών τιμών:
  - Εντροπία του χαρακτηριστικού «Ουρανός»

$$E(Ουρανός) = \frac{5}{14} E(καθαρός) + \frac{4}{14} E(συννεφιά) + \frac{5}{14} E(βροχή)$$

όπου:

- 5/14: το ποσοστό των τιμών «καθαρό» στο S
- 4/14: το ποσοστό των τιμών «συννεφιά» στο S
- 5/14: το ποσοστό των τιμών «βροχή» στο S

- Παραδείγματα υπολογισμών τιμών:
  - Εντροπία της τιμής «καθαρός» του χαρακτηριστικού «Ουρανός»

$$E(καθαρός) = \frac{3}{5} log_2 \frac{5}{3} + \frac{2}{5} log_2 \frac{5}{2} + 0$$

όπου:

  - 3/5: πιθανότητα όταν η τιμή είναι «καθαρός» η διάρκεια να είναι μικρή
  - 2/5: πιθανότητα η διάρκεια να είναι κανονική
  - 0: η πιθανότητα η διάρκεια να είναι μηδενική
  - Κέρδος πληροφορίας χαρακτηριστικού «Ουρανός»

$$Gain\ (S, Ουρανός) = E\ (S) - E\ (Ουρανός)$$

1. Με χρήση των παραπάνω τύπων υπολογίζουμε:
   - Gain(Ουρανός)=1.183851
   - Gain(Θερμοκρασία)=0.333841
   - Gain(Υγρασία)=0.259677
   - Gain(Άνεμος)=0.04812703
2. Το χαρακτηριστικό «Ουρανός» με το μεγαλύτερο κέρδος ανατίθεται σε κόμβο
3. Αφαιρούμε από το σύνολο των χαρακτηριστικών το συγκεκριμένο => ***FS΄={Θερμοκρασία, Υγρασία, Άνεμος}***
4. για κάθε τιμή του:

- τρέχουμε τον αλγόριθμο για τα υποπροβλήματα που δημιουργούνται λαμβάνοντας υπόψη το υποσύλο του S για το οποίο έχει τη συγκεκριμένη τιμή:
  - για «καθαρός»: ***TS΄={T1,T2,T8,T9,T11}***
  - για «συννεφιά»: ***TS΄={T3,T7,T12,T13}***
  - για «βροχή»: ***TS΄={T4,T5,T6,T10,T14}***

### 3.3.3. ΑΛΓΌΡΙΘΜΟΣ C4.5

Ο C4.5 είναι ένας αλγόριθμος, (όπως και ο ID3) που αναπτύχθηκε από τον Ross Quinlan και χρησιμοποιείται για να δημιουργήσει ένα δέντρο απόφασης, όπως έχει ήδη αναφερθεί αποτελεί μια επέκταση του προηγούμενου αλγόριθμο ID3. Τα δέντρα απόφασης που παράγονται από τον εν λόγω αλγόριθμο μπορούν να χρησιμοποιηθούν για ταξινόμηση, και για το λόγο αυτό, ο C4.5 συχνά αναφέρεται ως ένας στατιστικός ταξινομητής (statistical classifier). (Βλαχάβας, κ.α., 2002)

Ο C4.5 χτίζει δέντρα απόφασης από ένα σύνολο δεδομένων εκπαίδευσης με τον ίδιο τρόπο όπως ο ID3, χρησιμοποιώντας το Gain Ratio κριτήριο. Τα δεδομένα εκπαίδευσης είναι ένα σύνολο S = $\{s_1, s_2, ...\}$ από ήδη ταξινομημένα δείγματα. Κάθε $s_i$ δείγμα αποτελείται από έναν p-διάστατο διάνυσμα ($\{x_1, i\}, \{x_2, \theta\}, ..., \{x_\rho, \theta\}$), όπου οι $x_j$ αντιπροσωπεύουν χαρακτηριστικά του δείγματος, καθώς και την κατηγορία στην οποία υπάγεται το $s_i$.

Σε κάθε κόμβο του δέντρου, ο C4.5 επιλέγει το χαρακτηριστικό των δεδομένων που χωρίζει με τον πλέον αποτελεσματικό τρόπο το σύνολο των δειγμάτων σε υποσύνολα εμπλουτισμένα σε μία κλάση ή σε μια άλλη. Ο διαχωρισμός του κριτηρίου είναι το κανονικοποιημένο κέρδος πληροφορίας (διαφορά της εντροπίας). Το χαρακτηριστικό με το υψηλότερο κανονικοποιημένο όφελος πληροφοριών είναι αυτό που θα επιλεγεί για τη λήψη της απόφασης.. Ο αλγόριθμος C4.5, στη συνέχεια, επαναλαμβάνεται και στα επόμενα, χαμηλότερα επίπεδα.

Ο αλγόριθμος αυτός έχει μερικά βασικά χαρακτηριστικά (Βλαχάβας, κ.α., 2002):

- Όλα τα δείγματα στον πίνακα ανήκουν στην ίδια κατηγορία. Όταν συμβαίνει αυτό, απλώς δημιουργεί ένα κόμβο-φύλλο για το δέντρο απόφασης λέγοντας να επιλέξουν την εν λόγω κατηγορία.
- Κανένα από τα χαρακτηριστικά δεν παρέχει κάποιο όφελος πληροφοριών. Στην περίπτωση αυτή, ο C4.5 δημιουργεί ένα κόμβο αποφάσεων χρησιμοποιώντας την αναμενόμενη τιμή της τάξης.

Σε ψευδοκώδικα, ο γενικός αλγόριθμος για την κατασκευή δέντρων απόφασης είναι:

Ελέγξτε τις περιπτώσεις βάσεων

Για κάθε γνώρισμα α

Βρείτε τον κανονικοποιημένο δείκτη κέρδους πληροφορίες από διάσπαση του α

Έστω «a_best» είναι το χαρακτηριστικό με το υψηλότερο κανονικοποιημένο κέρδος πληροφορίας

Δημιουργήστε έναν κόμβο απόφασης που χωρίζει σε a_best

Επαναλάβετε στους υποκαταλόγους που λαμβάνεται με διαχωρισμό σε a_best, και προσθέστε αυτούς τους κόμβους ως παιδιά του κόμβου.

Να αναφέρουμε επιπλέον ότι ο C4.5 έκανε μια σειρά από βελτιώσεις στον αλγόριθμο ID3, οι σπουδαιότερες εκ των οποίων είναι οι ακόλουθες:

- Χρησιμοποιείται τόσο σε συνεχή όσο και διακριτά χαρακτηριστικά. Για να χειριστεί συνεχή χαρακτηριστικά, ο C4.5 δημιουργεί ένα όριο και, στη συνέχεια, χωρίζει τη λίστα σε εκείνα των οποίων το χαρακτηριστικό (η αξία) είναι πάνω από το όριο και εκείνα που είναι μικρότερη ή ίση με το όριο.
- Δυνατότητα χρήσης σε περιπτώσεις όπου λείπουν κάποιες τιμές. Οι τιμές γνωρισμάτων που λείπουν απλά δεν χρησιμοποιούνται στους υπολογισμούς κέρδους και εντροπίας.
- Δυνατότητα αντιμετώπισης χαρακτηριστικών με διαφορετικό κόστος.
- Κλάδεμα δέντρων μετά τη δημιουργία – ο C4.5 πηγαίνει πίσω από το δέντρο αφού έχουν δημιουργηθεί και επιχειρεί να αφαιρέσει κλαδιά που δεν βοηθούν με την αντικατάσταση τους με κόμβους.

Συνοψίζοντας να αναφέρουμε πως τα κυριότερα πλεονεκτήματα του εν λόγω αλγορίθμου είναι τα εξής:

- Λογικός χρόνος εκπαίδευσης.
- Γρήγορη εφαρμογή.
- Ευκολία στην κατανόηση.
- Εύκολη υλοποίηση.
- Μπορεί να χειριστεί μεγάλο αριθμό γνωρισμάτων.

Η χρήση του εν λόγω αλγορίθμου συχνά συνοδεύεται και από ορισμένα μειονεκτήματα, τα σημαντικότερα των οποίων παρουσιάζονται ως εξής:

- Δεν μπορεί να χειριστεί περίπλοκες σχέσεις μεταξύ των γνωρισμάτων.
- Απλά όρια απόφασης (decision boundaries).
- Παρουσιάζονται προβλήματα όταν λείπουν πολλά δεδομένα.

## 3.4. ΕΦΑΡΜΟΓΗ ΔΕΝΔΡΩΝ ΑΠΟΦΑΣΗΣ ΣΕ ΠΡΑΓΜΑΤΙΚΑ ΔΕΔΟΜΕΝΑ (με το πρόγραμμα WEKA)

### 3.4.1. Περιγραφή Συνόλου Δεδομένων

Στην παρούσα ενότητα παρουσιάζονται τα αποτελέσματα από την εφαρμογή των δένδρων απόφασης σε πραγματικά δεδομένα. Πρόκειται για δεδομένα που σχετίζονται με τις άμεσες εκστρατείες μάρκετινγκ του πορτογαλικού τραπεζικού ιδρύματος. Τα δεδομένα προέρχονται από τον ιστότοπο UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/). Οι εκστρατείες μάρκετινγκ βασίστηκαν σε τηλεφωνικές κλήσεις. Μας ενδιαφέρει να εντοπίσουμε αν υπάρχει ή όχι προθεσμιακή κατάθεση, συχνά όμως παρατηρείται το φαινόμενο ένα πελάτης να συμμετέχει σε περισσότερες από μια εγγραφές.

Πριν προβούμε σε ανάλυση των αποτελεσμάτων, κρίναμε σκόπιμο να αναλύσουμε συνοπτικά τα 17 χαρακτηριστικά του συνόλου δεδομένων Bank που χρησιμοποιήθηκαν. Αναλυτικότερα:

| ΟΝΟΜΑΣΙΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΟΥ | ΠΕΡΙΓΡΑΦΗ ΜΕΤΑΒΛΗΤΗΣ | ΤΙΜΕΣ/ ΠΕΡΙΓΡΑΦΗ |
|---|---|---|
| Age | Η ηλικία του ερωτώμενου | Αριθμητική(εύρος τιμών….) |
| Job | Η επαγγελματική κατάσταση του ερωτώμενου | Κατηγορική, με διακριτές τιμές: admin.", "unknown", "unemployed", "management", "housemaid", "entrepreneur", "student", "blue-collar", "self-employed", "retired", "technician", "services" |
| Marital | Η οικογενειακή κατάσταση του ερωτώμενου | Κατηγορική, με διακριτές τιμές: "married", "divorced", "single" |
| Education | Το εκπαιδευτικό επίπεδο του ερωτώμενου | Κατηγορική, με διακριτές τιμές: "unknown", "secondary", "primary", "tertiary" |
| Default | «Έχει ο πελάτης πιστωτική κάρτα by default» ; | Δυαδική με διακριτές τιμές: "yes", "no" |
| Balance | Μέσος όρος εισοδήματος ανά χρόνο σε Ευρώ | Αριθμητική |
| Housing | «Έχει ο πελάτης στεγαστικό δάνειο;» | Δυαδική με διακριτές τιμές: "yes", "no" |
| Loan | «Έχει ο πελάτης προσωπικό δάνειο;» | Δυαδική με διακριτές τιμές: "yes", "no" |
| Contact | Ο τύπος της επικοινωνίας | Κατηγορική, με διακριτές τιμές: "unknown", "telephone", "cellular" |
| Day | Η τελευταία μέρα που έγινε κάποιας μορφής επικοινωνία | Αριθμητική |

| | | |
|---|---|---|
| Month | Ο τελευταίος μήνας επικοινωνίας | Κατηγορική, με διακριτές τιμές: "jan", "feb", "mar", ..., "nov", "dec" |
| Duration | Η διάρκεια (σε δευτερόλεπτα) της τελευταίας επικοινωνίας | Αριθμητική |
| Campaign | Ο αριθμός των επαφών που έχει γίνει για το συγκεκριμένο πελάτη στην συγκεκριμένη εκστρατεία | Αριθμητική |
| pdays | Πόσες μέρες έχουν μεσολαβήσει από την τελευταία επικοινωνία με τον πελάτη για προηγούμενη καμπάνια | Αριθμητική |
| Previous | Πόσες φορές υπήρξε επικοινωνία με το συγκεκριμένο πελάτη πριν από τη συγκεκριμένη καμπάνια | Κατηγορική, με διακριτές τιμές: |
| poutcome | Το αποτέλεσμα της προηγούμενης καμπάνιας μάρκετινγκ | Κατηγορική, με διακριτές τιμές: "unknown", "other", "failure", "success" |
| Y | Αν υπάρχει ήδη προθεσμιακή κατάθεση ή όχι | Δυαδική, με διακριτές τιμές: "yes", "no"<br>Χαρακτηριστικό **κλάση** |

Στο σημείο αυτό να αναφέρουμε πως το πλήθος των εγγραφών στο σύνολο training είναι 45.211 ενώ στην περίπτωση του test είναι σαφώς λιγότερες, ανέρχονται σε 4.521.

### 3.4.2. Ανάλυση

Αρχικά αναφέρουμε μερικά βασικά λόγια για το πρόγραμμα που χρησιμοποιήθηκε προκειμένου να γίνει η ανάλυση. Πιο συγκεκριμένα, πρόκειται για το Weka: **W**ekato **E**nvironment for **K**nowledge **A**nalysis, το οποίο είναι ένα software για εξόρυξη δεδομένων γραμμένο σε JAVA (http://www.cs.waikato.ac.nz/ml/weka/). Αποτελεί μια συλλογή από αλγορίθμους μηχανικής μάθησης με σκοπό την εξόρυξη δεδομένων (data mining). Οι αλγόριθμοι που είναι

ενσωματωμένοι στο πρόγραμμα μπορούν να εφαρμοστούν σε ένα σύνολο δεδομένων άμεσα. Τα εργαλεία που είναι διαθέσιμα στο εν λόγω εργαλεία είναι:

- Προεπεξεργασία Δεδομένων
- Ταξινόμηση (classification): δημιουργία «μοντέλων» από τα δεδομένα με κάποια διαδικασία εκπαίδευσης
- Συσταδοποίηση (clustering): ομαδοποίηση 'όμοιων' δεδομένων
- Εύρεση Κανόνων Συσχέτισης (association rules): παραγωγή κανόνων συσχετίσεων και προτύπων.
- Οπτικοποίηση (visualization): απεικόνιση τόσο των αρχικών δεδομένων όσο και των αποτελεσμάτων μετά τη διαδικασία της εκπαίδευσης.

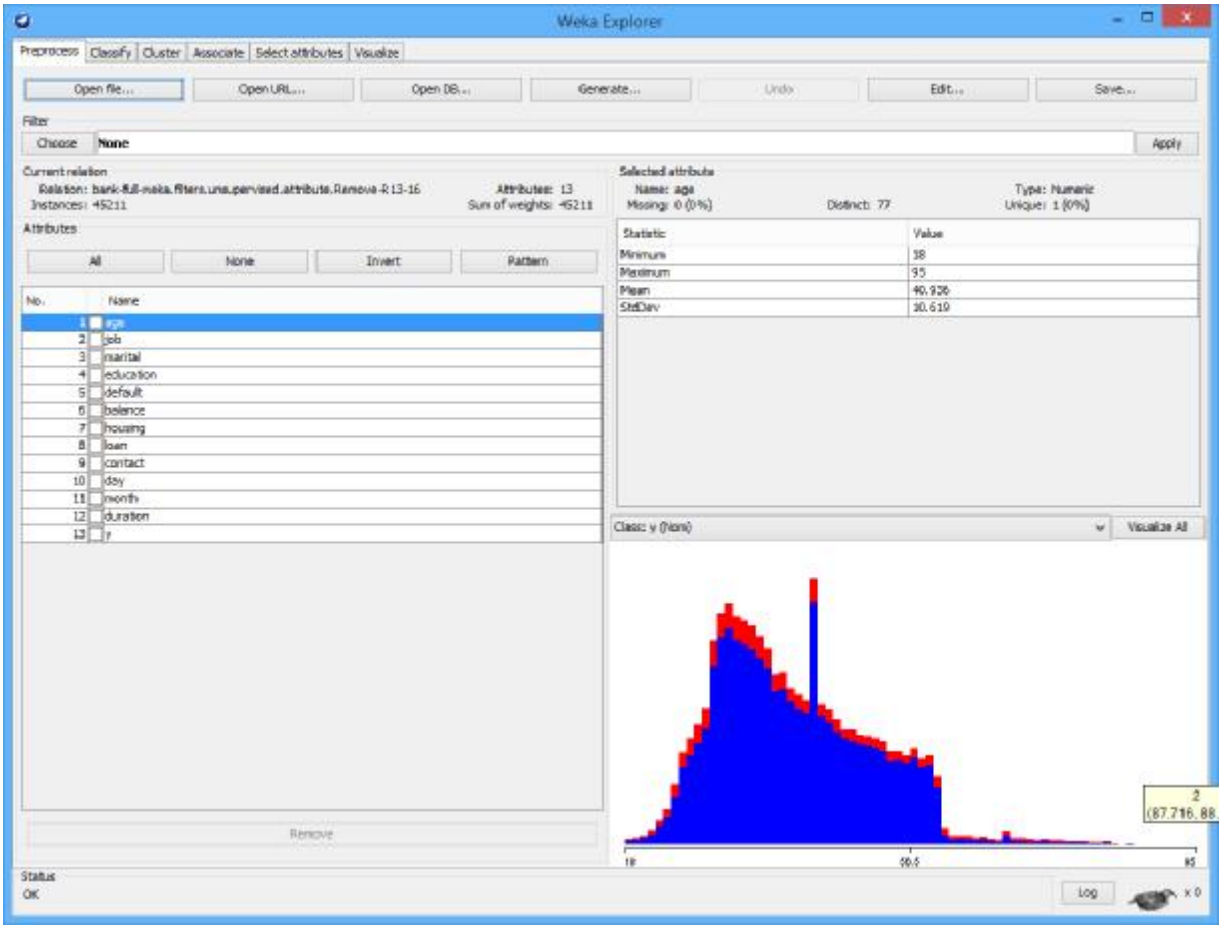
Όλα τα παραπάνω πραγματοποιούνται σε ένα γραφικό περιβάλλον το οποίο ονομάζεται «Explorer».


Αναφορικά με το περιβάλλον WEKA μπορούμε να αναφέρουμε τα ακόλουθα:

Ανοίγοντας το πρόγραμμα, μέσω του μενού Application→Explorer→Open file δίνεται η δυνατότητα να επιλεγεί ένα σύνολο δεδομένων στο οποίο μπορούν να εφαρμοστούν τεχνικές που αναφέρθηκαν παραπάνω.

Επιλέγοντας ένα σύνολο δεδομένων (αρχείο .arff), εμφανίζονται γραφικά τα δεδομένα για καθένα από τα γνωρίσματα ξεχωριστά καθώς και στατιστικές πληροφορίες για αυτά. Εάν στο

σύνολο δεδομένων δίνεται και κάποια κλάση στην οποία ταξινομούνται, τα δεδομένα που ανήκουν στην ίδια κλάση εμφανίζονται με το ίδιο χρώμα.




Πριν αναλύσουμε τα αποτελέσματα που προέκυψαν να σημειώσουμε ότι θα πραγματοποιήσουμε τα ακόλουθα 4 πειράματα:

1. χρησιμοποίηση όλων των χαρακτηριστικών για τον αλγόριθμο ID3, δηλαδή των 17 χαρακτηριστικών που υπάρχουν στο σύνολο των δεδομένων, όπως περιγράφηκαν παραπάνω,

2. χρησιμοποίηση όλων των χαρακτηριστικών για τον αλγόριθμο J48,

3. χρησιμοποίηση ορισμένων χαρακτηριστικών για τον αλγόριθμο ID3, δηλαδή των 13 πρώτων χαρακτηριστικών (age, job, marital, education, default, balance, housing, loan, contact, day, month, duration, y) με σκοπό τη δημιουργία ενός σεναρίου με σκοπό να μελετήσουμε την αξία της επικοινωνίας μεταξύ τράπεζας και πελάτη,

4. χρησιμοποίηση ορισμένων χαρακτηριστικών για τον αλγόριθμο J48, όπως περιγράφεται στη προηγούμενη εκτέλεση.

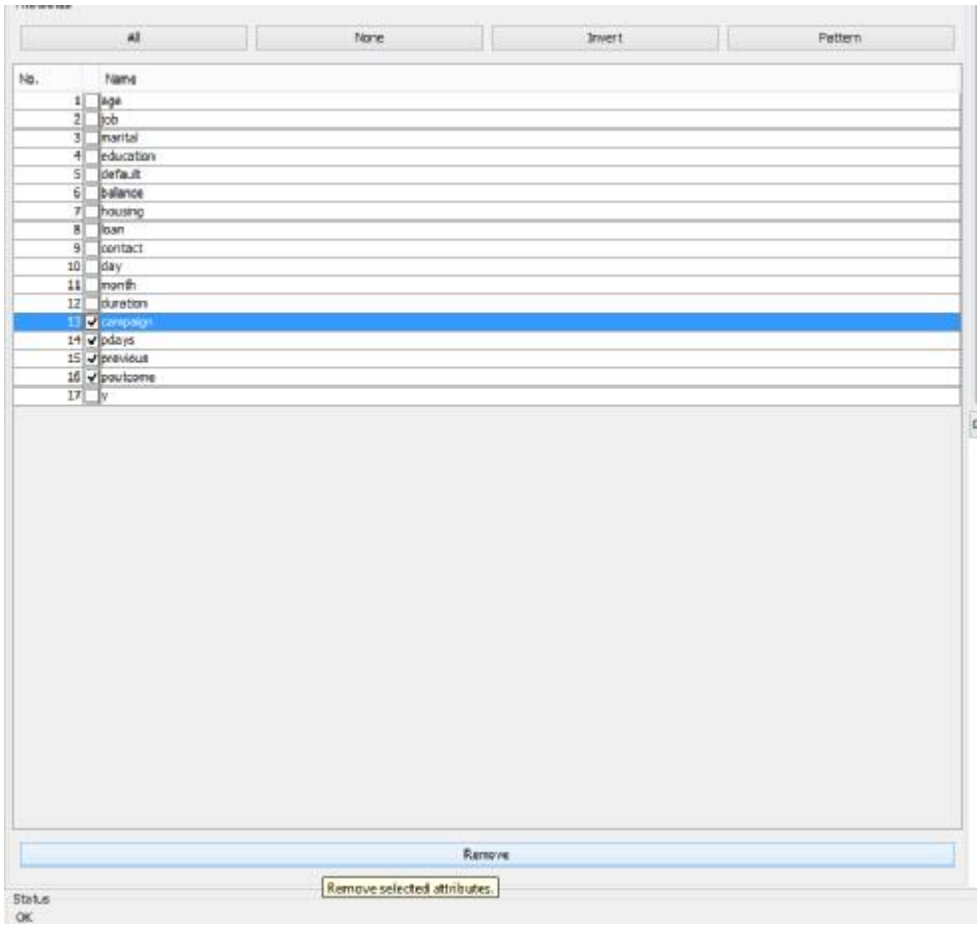
Προκειμένου να πραγματοποιήσουμε τις παραπάνω αναλύσεις οφείλουμε να διευκρινίσουμε τα ακόλουθα:

Για τις εκτελέσεις 1 και 2 των αντίστοιχων αλγορίθμων, τα δεδομένα μας δεν έχουν υποστεί καμία προεπεξεργασία. Αντίστοιχα για τις εκτελέσεις 3 και 4 απαλείφουμε τα τελευταία 4 χαρακτηριστικά που υπάρχουν στο dataset, καθώς σκοπός μας είναι να μελετήσουμε την αξία της επικοινωνίας μεταξύ τράπεζας και πελάτη ως προς το χρονικό διάστημα που εκείνη γίνεται (ημέρα, μήνα (καλοκαίρι-χειμώνας)) αλλά και τη διάρκεια της επικοινωνίας. Τα 4 χαρακτηριστικά τα οποία δεν χρησιμοποιούμε είναι το ***campaign***, ***pdays***, ***previous***, ***poutcome***. Οι συγκεκριμένοι παράγοντες δεν εξαιρέθηκαν τυχαία, αλλά ύστερα από μελέτη των δεδομένων παρατηρήσαμε πως τα χαρακτηριστικά αυτά δεν προσφέρουν επιπλέον πληροφόρηση, ούτε αυξάνουν την ακρίβεια του δένδρου απόφασης, οπότε μπορούν να εξαιρεθούν από το υπόδειγμα.

Το σενάριο το οποίο χρησιμοποιούμε αφορά το βαθμό στον οποίο η τελική απόφαση σχετικά με τις προθεσμιακές καταθέσεις επηρεάζεται από τη διάρκεια αλλά και το χρόνο στον οποίο πραγματοποιείται η επικοινωνία. Πιο συγκεκριμένα μας ενδιαφέρει αν η ημέρα, ο μήνας αλλά και ο χρόνος επηρεάζει την τελική έκβαση του αποτελέσματος.

Ακολούθως προσπαθούμε να προσδιορίσουμε τα βήματα που ακολουθούνται προκειμένου να προκύψει το τελικό αποτέλεσμα. Πιο συγκεκριμένα, να αναφέρουμε ότι μπορούμε να ακολουθήσουμε την εξής διαδικασία προκειμένου να εφαρμόσουμε τους αλγορίθμους:

**Επιλογή προς ανάλυση χαρακτηριστικών** (Feature Selection): Επιλογή ορισμένων χαρακτηριστικών με βάση το πρόβλημα που θέλουμε να μελετήσουμε. Μέσω του λογισμικού WEKA μπορούμε να αγνοήσουμε τα χαρακτηριστικά (attributes) που δε μας ενδιαφέρουν, επιλέγοντας τα και απομακρύνοντας τα (remove). Παραθέτουμε ακολούθως μια εικόνα σχετικά με την απαλοιφή των τεσσάρων χαρακτηριστικών (13 – campaign, 14 – pdays, 15 – previous, 16 – poutcome) από το δείγμα μας:

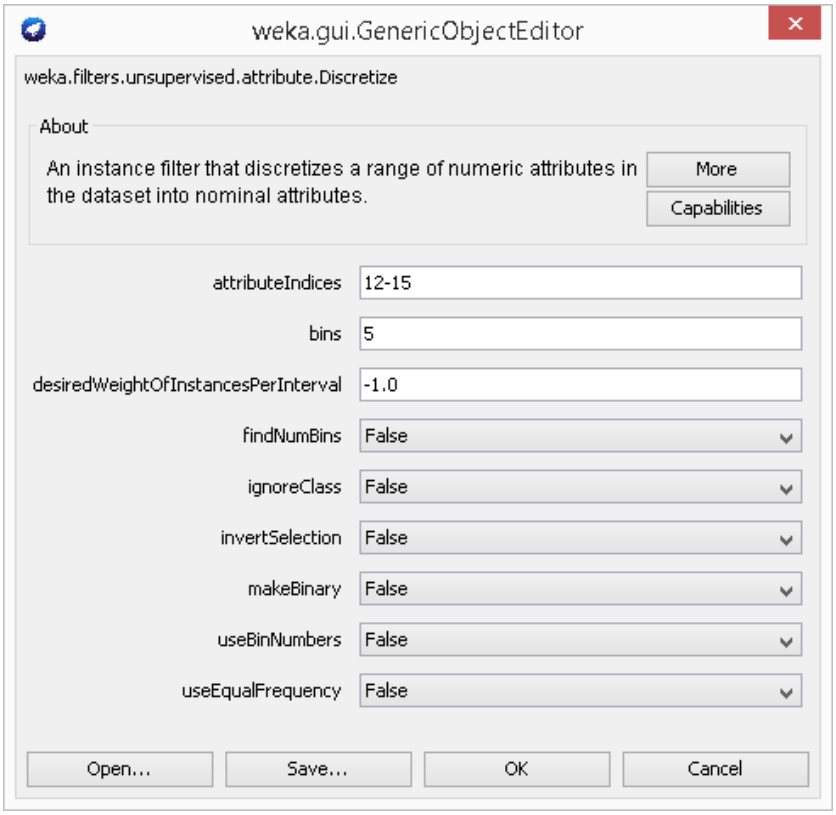



Σε συνέχεια των παραπάνω, αξίζει να αναφερθούμε στη διαδικασία μετασχηματισμού δεδομένων.

**Μετασχηματισμοί Δεδομένων** (Data transformation):

Διακριτοποίηση δεδομένων: απαιτείται μόνο στη περίπτωση εκτέλεσης του αλγορίθμου ID3 ο οποίος δεν μπορεί να τρέξει με Numeric δεδομένα όπως είναι η ηλικία. Έτσι σε αυτή τη περίπτωση πρέπει να γίνει προεπεξεργασία δεδομένων και μάλιστα εφαρμογή κατάλληλου φίλτρου με σκοπό τη διακριτοποίηση.

Στη καρτέλα preprocess για όλα τα αριθμητικά χαρακτηριστικά (π.χ. age) εφαρμόσαμε το φίλτρο Discretize, ομαδοποιώντας τα δεδομένα μας σε 5 bins, μετατρέποντας κατά αυτόν τον τρόπο σε κατηγορικά.




Παρατηρήσαμε πως αν εφαρμόσουμε την ίδια μέθοδο δεν ομαδοποιούνται τα δεδομένα με τον ίδιο τρόπο οπότε και δε μπορούμε να κάνουμε αυτή τη σύγκριση. Να υπογραμμίσουμε πως κάτι τέτοιο συμβαίνει γιατί οι μεταβλητές στο μεγάλο σύνολο έχουν διαφορετικό εύρος τιμών από το εύρος των αντίστοιχων μεταβλητών στο μικρό, με αποτέλεσμα να αλλάζει η διακριτοποίηση.

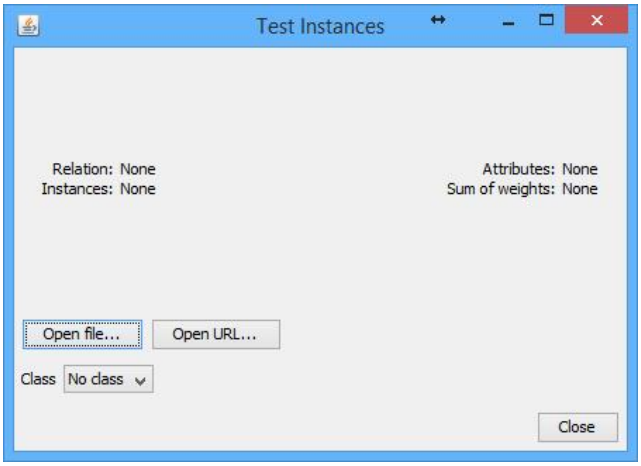
Επειδή θέλουμε να εφαρμόσουμε την ίδια διακριτοποίηση αλλά και να έχουμε το ίδιο test set με αυτό που μας δίνεται εξ αρχής, ενοποιήσαμε τα 2 αρχεία (bank-full+bank.arff), εφαρμόσαμε τη διακριτοποίηση (bank-full+bank_discretize_5.arff) και έπειτα τα διχοτομήσαμε εκ νέου(bank-full_discretize_5.arff και bank_discretize_5.arff). Για να μπορέσουμε να εκτελέσουμε σωστά τον αλόγριθμο ID3 για τις περιπτώσεις 1 και 3 εφαρμόσαμε το φίλτρο remove όπως έχει περιγραφεί παραπάνω.(bank-full_new_discretize_5.arff και bank_new_discretize_5.arff).

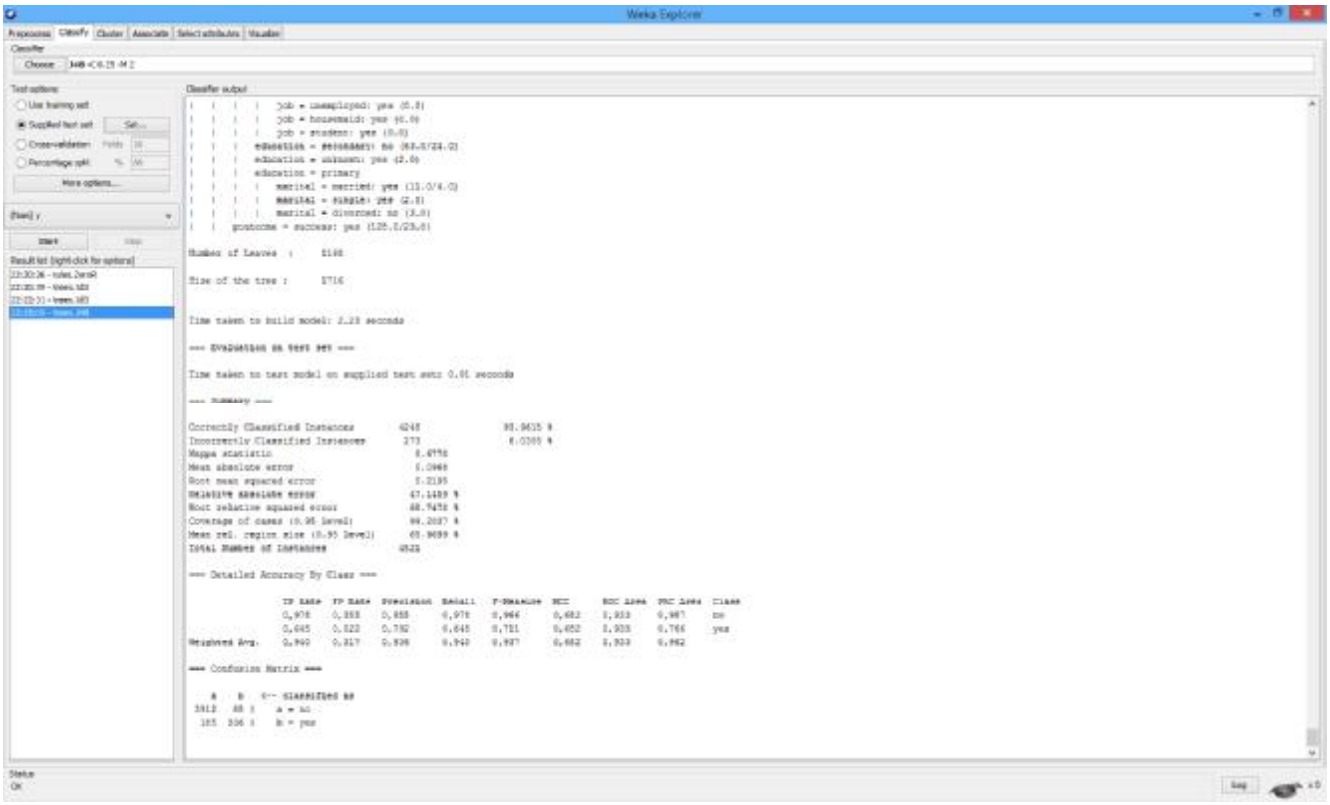
### 3.4.3. Περιγραφή των τρεξιμάτων

Στην ενότητα αυτή παρουσιάζεται ο τρόπος με τον οποίο γίνεται η εκτέλεση των τεσσάρων αλγορίθμων.

**Εκτέλεση J48**

Πιο συγκεκριμένα, εκτελέσαμε τον J48 με το data set bank-full.arff χρησιμοποιώντας ως test set το αρχείο bank.arff. Η επιλογή που κάναμε είναι η «Supplied test set». Η επιλογή αυτή μας επιτρέπει να καθορίσουμε ένα αρχείο με τα δεδομένα δοκιμής. Για να γίνει αυτό, επιλέγουμε την επιλογή και κάνουμε κλικ στο Set.

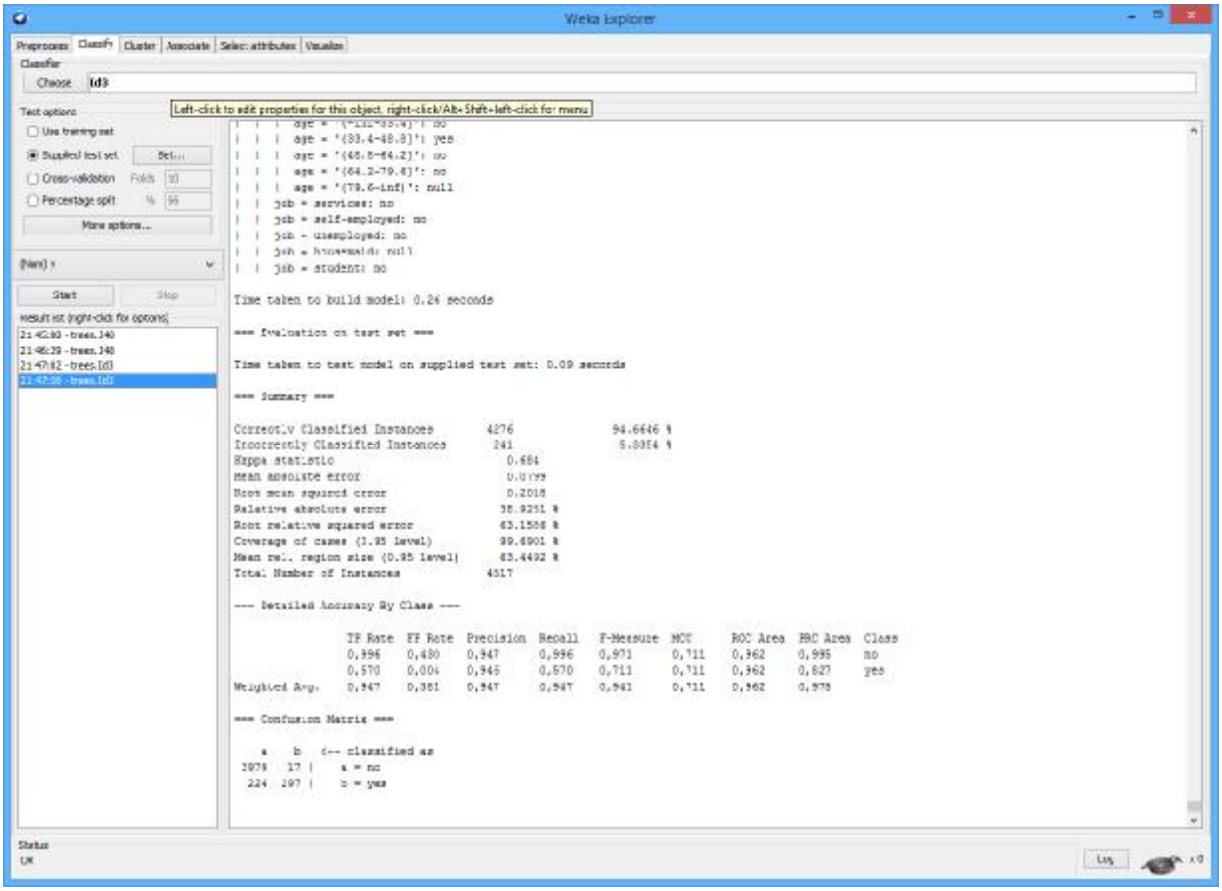




Στη συνέχεια, εκτελέσαμε τον J48 με το data set bank-full_new.arff χρησιμοποιώντας ως test set το αρχείο bank.arff έπειτα από προεπεξεργασία και αφαίρεση των 4 χαρακτηριστικών που δε θα μελετήσουμε. Η επιλογή που κάναμε και πάλι είναι η «Supplied test set».

**Εκτέλεση ID3**

Με αντίστοιχο τρόπο, εκτελούμε τον αλγόριθμο ID3 με το data set bank-full_discretize_5.arff και χρησιμοποιώντας ως test set το αρχείο bank_discretize_5.arff. Η επιλογή που κάναμε είναι η «Supplied test set», όπως αντίστοιχα κάναμε για τον J48.




Τέλος, εκτελούμε τον αλγόριθμο ID3 με το data set bank-full_new_discretize_5.arff και χρησιμοποιώντας ως test set το αρχείο bank_new_discretize_5.arff. Τα παραπάνω αρχεία έχουν υποστεί προεπεξεργασία, όπως περιγράφεται και παραπάνω.

### 3.4.4. Ανάλυση των αποτελεσμάτων

Παρακάτω παρουσιάζονται τα αποτελέσματα για κάθε ένα «τρέξιμο» αλγορίθμου:

**Αποτελέσματα από την εκτέλεση του αλγορίθμου J48 με όλα τα χαρακτηριστικά**

```
=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.07 seconds

=== Summary ===

Correctly Classified Instances        4248               93.9615 %
Incorrectly Classified Instances       273                6.0385 %
Kappa statistic                          0.6778
Mean absolute error                      0.0968
Root mean squared error                  0.2195
Relative absolute error                 47.1459 %
Root relative squared error             68.7478 %
Coverage of cases (0.95 level)          99.2037 %
Mean rel. region size (0.95 level)      65.9699 %
Total Number of Instances             4521

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,978     0,355     0,955       0,978   0,966      0,682    0,933     0,987     no
               0,645     0,022     0,792       0,645   0,711      0,682    0,933     0,766     yes
Weighted Avg.  0,940     0,317     0,936       0,940   0,937      0,682    0,933     0,962

=== Confusion Matrix ===

    a    b   <-- classified as
 3912   88 |    a = no
  185  336 |    b = yes
```

Σύμφωνα με τα παραπάνω δεδομένα το ποσοστό επιτυχίας του αλγορίθμου είναι **93,9615%.**

Επίσης από τη Μήτρα Σύγχυσης (Confusion Matrix) συμπεραίνουμε ότι οι 3912 πελάτες «τοποθετούνται» σωστά στην κλάση “no” που σημαίνει ότι δεν έχουν προθεσμιακό λογαριασμό. Ενώ οι 88 τοποθετούνται λανθασμένα στην κλάση “yes”. Αντίστοιχα οι 185 πελάτες τοποθετήθηκαν σωστά στη κλάση “yes” , που σημαίνει ότι έχουν προθεσμιακό λογαριασμό, ενώ οι 336 τοποθετήθηκαν λανθασμένα στην κλάση “no”.

Συμπεραίνουμε ότι από την εκπαίδευση του αλγορίθμου, περισσότερες εγγραφές του test έχουν τοποθετηθεί σωστά στη κλάση “no” ενώ το λάθος τοποθέτησης για τη κλάση “yes” είναι μεγαλύτερο.

**Αποτελέσματα από την εκτέλεση του αλγορίθμου J48 με λιγότερα χαρακτηριστικά σύμφωνα με το σενάριο μας**

```
=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.04 seconds

=== Summary ===

Correctly Classified Instances        4206               93.0325 %
Incorrectly Classified Instances       315                6.9675 %
Kappa statistic                          0.6213
Mean absolute error                      0.1091
Root mean squared error                  0.2336
Relative absolute error                 53.1592 %
Root relative squared error             73.142  %
Coverage of cases (0.95 level)          98.7171 %
Mean rel. region size (0.95 level)      66.1469 %
Total Number of Instances             4521

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,975    0,415    0,948      0,975    0,961      0,628    0,914     0,983     no
               0,585    0,025    0,755      0,585    0,659      0,628    0,914     0,718     yes
Weighted Avg.  0,930    0,370    0,925      0,930    0,926      0,628    0,914     0,953

=== Confusion Matrix ===

    a    b   <-- classified as
 3901   99 |    a = no
  216  305 |    b = yes
```

Σύμφωνα με τα παραπάνω δεδομένα το ποσοστό επιτυχίας του αλγορίθμου είναι 93,0325%.

Επίσης από το παραπάνω confusion matrix συμπεραίνουμε ότι οι 3901 πελάτες «τοποθετούνται» σωστά στην κλάση “no”, δηλαδή στην κλάση “yes”, που σημαίνει ότι δεν έχουν προθεσμιακό λογαριασμό. Ενώ οι 99 τοποθετούνται λανθασμένα στην κλάση “yes”. Αντίστοιχα οι 305 πελάτες τοποθετήθηκαν σωστά στην κλάση “yes” που σημαίνει ότι έχουν προθεσμιακό λογαριασμό, ενώ οι 216 τοποθετήθηκαν λανθασμένα στην κλάση “no”.

Και σε αυτή την εκτέλεση συμπεραίνουμε ότι το λάθος τοποθέτησης για τη κλάση “no” είναι περίπου είκοσι φορές μεγαλύτερο από ότι αυτό για τη κλάση “yes”. Αυτό μας δείχνει ότι τα χαρακτηριστικά που αφαιρέσαμε είναι καθοριστικά για την απάντηση που πρέπει να δώσει ο αλγόριθμος όταν η πλειάδα ανήκει στη κλάση “yes”. Ωστόσο οι περισσότερες πλειάδες ανήκουν στη κλάση “no” οπότε το ποσοστό επιτυχίας του αλγορίθμου δεν επηρεάζεται από την αφαίρεση

των 4 χαρακτηριστικών σύμφωνα με το σενάριο που εφαρμόσαμε και φαίνεται και από τη παραπάνω εικόνα.

**Αποτελέσματα από την εκτέλεση του αλγορίθμου ID3 με όλα τα χαρακτηριστικά**

```
=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.02 seconds

=== Summary ===

Correctly Classified Instances        4350               96.2177 %
Incorrectly Classified Instances       171                3.7823 %
Kappa statistic                          0.7878
Mean absolute error                      0.0555
Root mean squared error                  0.1693
Relative absolute error                 27.0389 %
Root relative squared error             53.0349 %
Coverage of cases (0.95 level)          99.8009 %
Mean rel. region size (0.95 level)      59.3453 %
Total Number of Instances             4521

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,998     0,309     0,961       0,998   0,979      0,802    0,981     0,997     no
               0,691     0,003     0,973       0,691   0,808      0,802    0,981     0,903     yes
Weighted Avg.  0,962     0,274     0,963       0,962   0,959      0,802    0,981     0,987

=== Confusion Matrix ===

    a    b   <-- classified as
 3990   10 |    a = no
  161  360 |    b = yes
```

Σύμφωνα με τα παραπάνω δεδομένα το ποσοστό επιτυχίας του αλγορίθμου είναι 96,2177%.

Επίσης από το παραπάνω confusion matrix συμπεραίνουμε ότι οι 3990 πελάτες «τοποθετούνται» σωστά στην κλάση “no”, που σημαίνει ότι δεν έχουν προθεσμιακό λογαριασμό. Ενώ οι 10 τοποθετούνται λανθασμένα στην κλάση “yes”. Αντίστοιχα οι 360 πελάτες τοποθετήθηκαν σωστά στην κατηγορία “yes”, που σημαίνει ότι έχουν προθεσμιακό λογαριασμό, ενώ οι 161 τοποθετήθηκαν λανθασμένα στην κλάση “no”.

**Αποτελέσματα από την εκτέλεση του αλγορίθμου ID3 με λιγότερα χαρακτηριστικά σύμφωνα με το σενάριο μας.**

```
=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.02 seconds

=== Summary ===

Correctly Classified Instances        4276               94.6646 %
Incorrectly Classified Instances       241                5.3354 %
Kappa statistic                          0.684
Mean absolute error                      0.0799
Root mean squared error                  0.2018
Relative absolute error                 38.9251 %
Root relative squared error             63.1586 %
Coverage of cases (0.95 level)          99.6901 %
Mean rel. region size (0.95 level)      63.4492 %
Total Number of Instances             4517

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,996    0,430    0,947      0,996    0,971      0,711    0,962     0,995     no
               0,570    0,004    0,946      0,570    0,711      0,711    0,962     0,827     yes
Weighted Avg.  0,947    0,381    0,947      0,947    0,941      0,711    0,962     0,975

=== Confusion Matrix ===

    a    b   <-- classified as
 3979   17 |    a = no
  224  297 |    b = yes
```

Σύμφωνα με τα παραπάνω δεδομένα το ποσοστό επιτυχίας του αλγορίθμου είναι 94,6646%.

Επίσης από το παραπάνω confusion matrix συμπεραίνουμε ότι οι 3979 πελάτες «τοποθετούνται» σωστά στην κλάση “no”, που σημαίνει ότι δεν έχουν προθεσμιακό λογαριασμό. Ενώ οι 17 τοποθετούνται λανθασμένα στην κλάση “yes”.

Αντίστοιχα οι 297 πελάτες τοποθετήθηκαν σωστά στην κλάση “yes”, που σημαίνει ότι έχουν προθεσμιακό λογαριασμό, ενώ οι 224 τοποθετήθηκαν λανθασμένα στην κλάση “no”. Από τα παραπάνω συμπεραίνουμε πως ο αλγόριθμος ID3 έχει οριακά υψηλότερο ποσοστό επιτυχίας από τον αντίστοιχο αλγόριθμο J48.

Αυτό το οποίο οφείλουμε να συγκρίνουμε είναι το ποσοστό των δεύτερων εκτελέσεων καθώς μας ενδιαφέρει να αξιολογήσουμε το σενάριο το οποίο δημιουργήσαμε για να μελετήσουμε την επιρροή του χρόνου επικοινωνίας στην απόφαση των πελατών. Αξίζει επίσης να σημειώσουμε

στο σημείο αυτό ότι για να «τρέξουμε» τον αλγόριθμο ID3 απαιτείται μετατροπή των αριθμητικών δεδομένων σε κατηγορικά, γεγονός το οποίο αποτελεί κόστος για την εκτέλεση.

**Συγκεντρωτική παρουσίαση αποτελεσμάτων**

| | ID3 Όλα τα χαρακτηριστικά | J48 Όλα τα χαρακτηριστικά | ID3 επιλεγμένα χαρακτηριστικά | J48 επιλεγμένα χαρακτηριστικά |
|---|---|---|---|---|
| F-Measure | 0,959 | 0,937 | 0,941 | 0,926 |
| ROC Area | 0,981 | 0,933 | 0,962 | 0,914 |

Από τον παραπάνω πίνακα παρατηρούμε πως το F- Measure είναι υψηλότερο για τον αλγόριθμο ID3, ομοίως και το κριτήριο ROC Area.

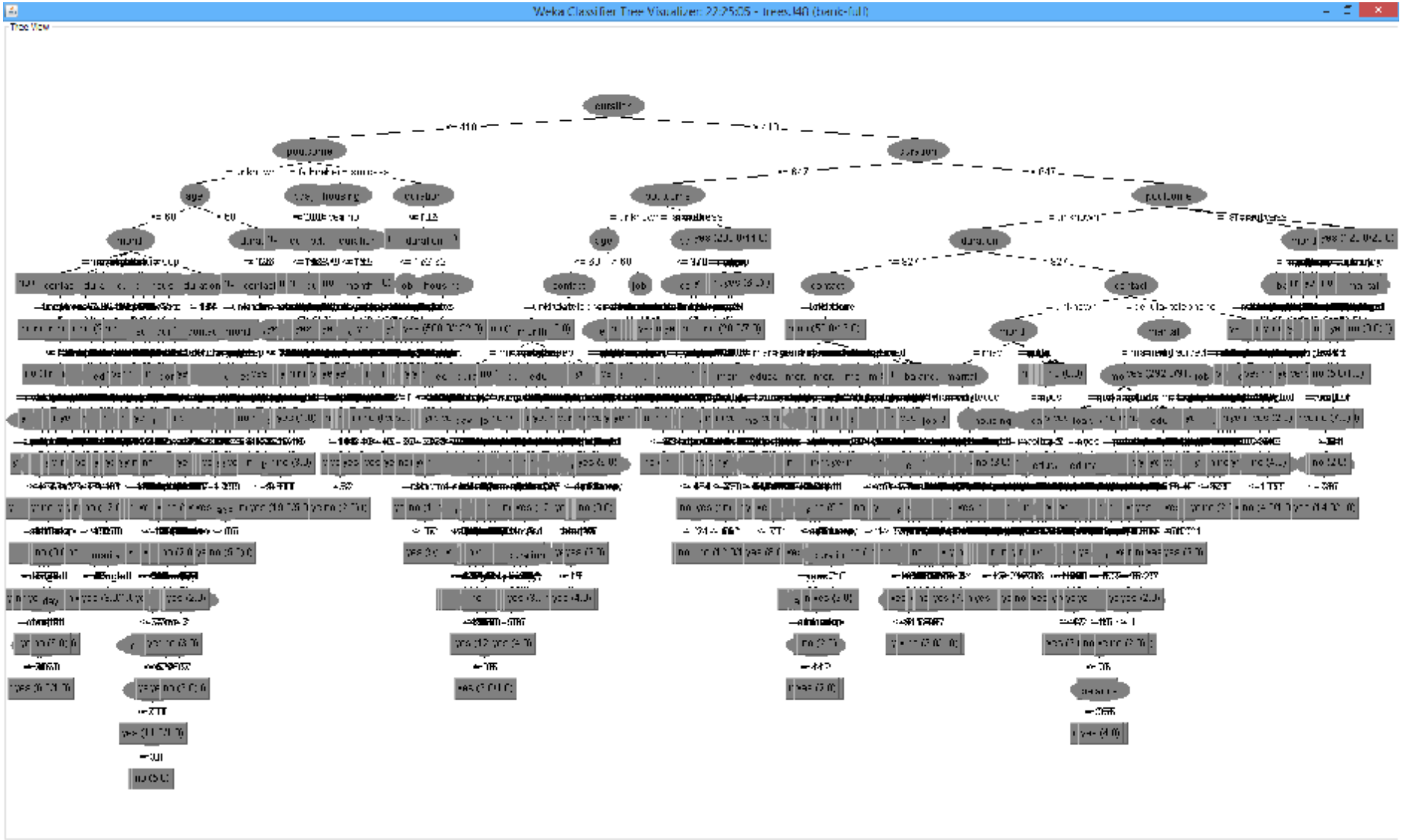
### 3.4.5. Δέντρα Απόφασης

**Εκτέλεση J48 με όλα τα χαρακτηριστικά**

Η κατασκευή του δένδρου από την εκτέλεση του αλγορίθμου J48 με όλα τα χαρακτηριστικά, που περιέχει το σύνολο δεδομένων, παρατηρώντας το δέντρο της Εικόνα 2 στηρίζεται στη διάκριση των πρώτων φύλων ως προς τη διάρκεια της συνομιλίας. Στο επόμενο επίπεδο η διάκριση γίνεται με βάση το αποτέλεσμα της προηγούμενης εκστρατείας. Σε κάθε όμως μια από τις περιπτώσεις υπάρχει διαφορετικός διαχωρισμός στο επόμενο επίπεδο. Στο αριστερό μέρος του δέντρου παρατηρούμε ότι δεν υπάρχει ένας ενιαίος κανόνας δημιουργίας του δέντρου απόφασης και δε μπορούμε να καταλήξουμε σε ένα συνολικό συμπέρασμα. Αντίθετα όμως στο δεξί μέρος του δέντρου παρατηρούμε ότι ο διαχωρισμός γίνεται και στο δεύτερο επίπεδο βάση της διάρκειας της συνομιλίας και έπειτα βάση του αποτελέσματος της προηγούμενης εκστρατείας. Σε αυτό όμως το επίπεδο παρατηρούμε ότι ο διαχωρισμός γίνεται με βάση το μήνα επικοινωνίας κατά τον οποίο έγινε η επικοινωνία στη προηγούμενη εκστρατεία. Στη δημιουργία λοιπόν του δέντρου απόφασης χρησιμοποιώντας το σύνολο των χαρακτηριστικών βλέπουμε ότι καθοριστικό ρόλο παίζει το αποτέλεσμα της προηγούμενης εκστρατείας και έπειτα ο χρόνος στον οποίο πραγματοποιήθηκε η επικοινωνία στη παρούσα εκστρατεία. Αυτό λοιπόν το οποίο διαπιστώνουμε είναι ότι δε μπορούμε να βγάλουμε ένα σαφές αποτέλεσμα ως προς τη συσχέτιση του αποτελέσματος της εκστρατείας με το χρόνο στον οποίο έγινε η επικοινωνία.

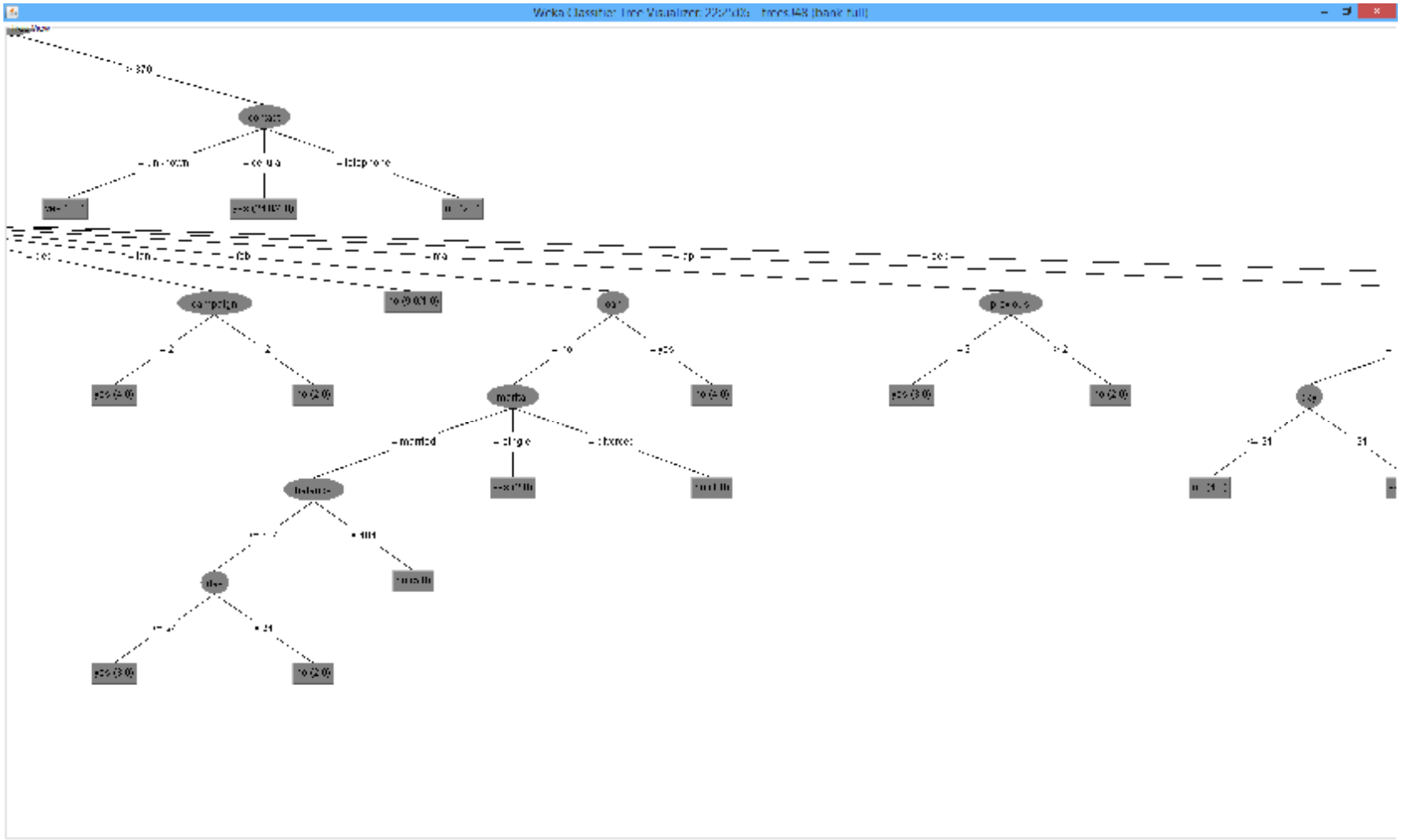
Αντίστοιχα στο δέντρο απόφασης που δημιουργήθηκε από την εκτέλεση του αλγορίθμου με τα χαρακτηριστικά που επιλέξαμε στο σενάριο μας, η κατασκευή του στηρίζεται σε πρώτο επίπεδο στη διάρκεια της συνομιλίας. Το επόμενο φύλο διακρίνεται με βάση το είδος της επικοινωνίας, στη συνέχεια ο διαχωρισμός γίνεται κατά κύριο λόγο βάσει του μήνα. Μελετώντας το δένδρο που ακολουθεί (Εικόνα 3) μπορούμε να συμπεράνουμε πως τους άκρως καλοκαιρινούς ή άκρως χειμερινούς μήνες οι πιθανότητες για δημιουργία προθεσμιακού λογαριασμού είναι σαφώς μικρότερες από τις αντίστοιχες για τους υπόλοιπους μήνες, γεγονός το οποίο οφείλεται στη φύση των δεδομένων και όχι στη λειτουργία του αλγορίθμου, καθώς ο αλγόριθμος αποτελεί οπτικοποίηση των δεδομένων.

Η χρήση του ID3 θα ήταν αποδοτική σε περίπτωση όπου όλα τα δεδομένα είναι κατηγορικά.

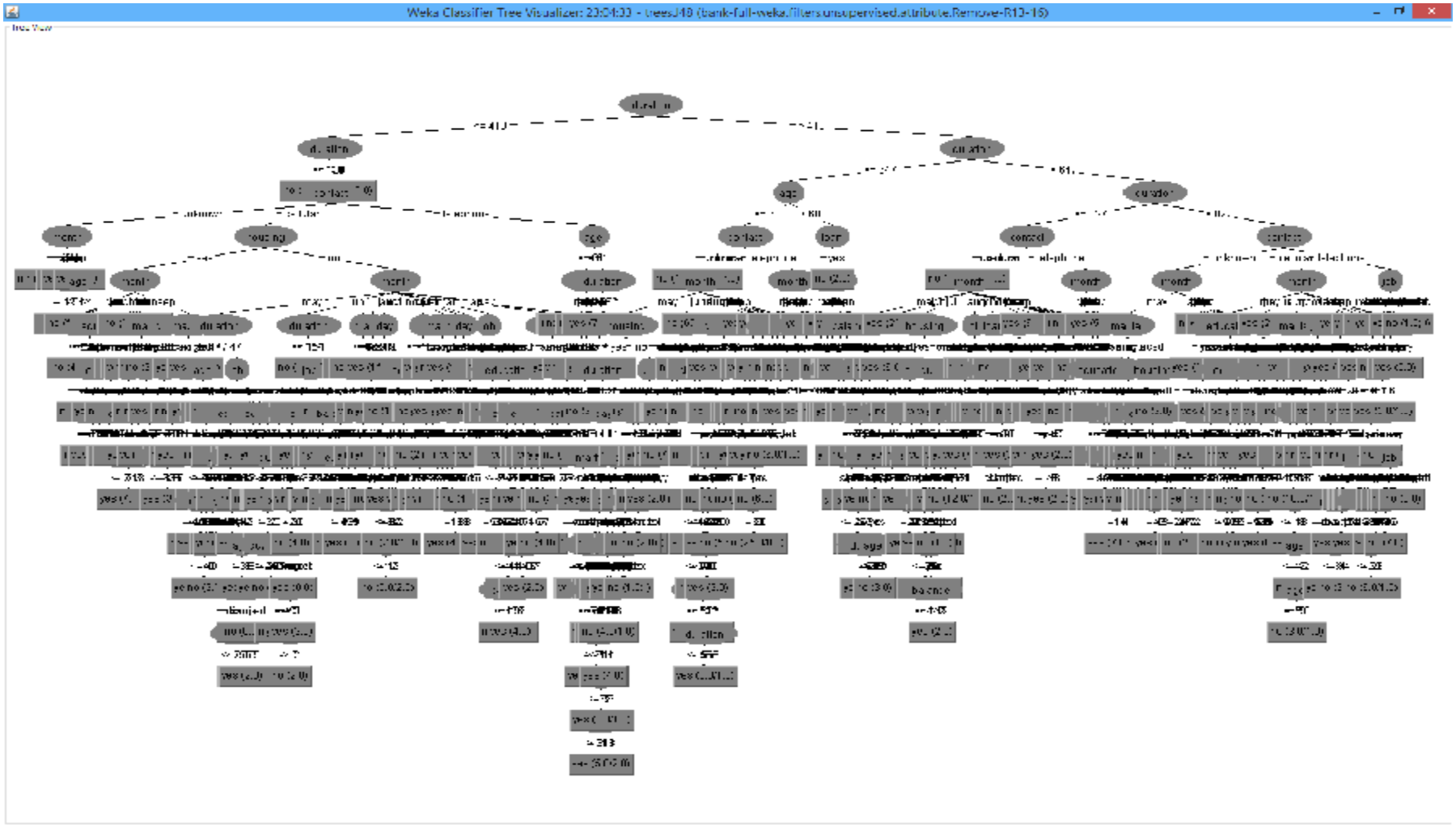



*Εικόνα 2. Δένδρο απόφασης με χρήση του αλγόριθμου J48 για όλα τα χαρακτηριστικά*

Στη παρακάτω εικόνα, φαίνεται με σαφήνεια ένα μέρος του παραπάνω δέντρου.



**Εκτέλεση J48 με ορισμένα χαρακτηριστικά σύμφωνα με το σενάριο μας**




*Εικόνα 3 Δένδρο απόφασης με χρήση του αλγόριθμου J48 για τα 13 χαρακτηριστικά*

# ΚΕΦΑΛΑΙΟ 4$^{ο}$: ΣΥΜΠΕΡΑΣΜΑΤΑ

Τα Predictive analytics είναι ένας ευρύς όρος, που περιγράφει μια σειρά από στατιστικές και αναλυτικές τεχνικές που χρησιμοποιούνται για την ανάπτυξη μοντέλων που προβλέπουν γεγονότα ή πρότυπα συμπεριφοράς. Η προγνωστική μοντελοποίηση βασίζεται στο "σκορ", με την υψηλότερη βαθμολογία να δείχνει μια ισχυρότερη πιθανότητα συγκεκριμένου γεγονότος ή ένα πρότυπο συμπεριφοράς που θα επαναλαμβάνεται στο μέλλον.

Η εξόρυξη δεδομένων περιλαμβάνει κάποιες από τις ακόλουθες κατηγορίες μεθόδων:

- **Ανίχνευση ανωμαλιών (Anomaly detection)**
- **Κανόνες συσχέτισης (Μοντέλο αλληλεξάρτησης)**
- **Συσταδοποίηση**
- **Κατηγοριοποίηση**
- **Παλινδρόμηση (στατιστική)**.

Η άντληση δεδομένων είναι ένα κρίσιμο συστατικό των predictive analytics. Αυστηρή ανάλυση δεδομένων γίνεται προκειμένου να εντοπιστούν τάσεις, πρότυπα ή οι σχέσεις μεταξύ των διαφόρων συνόλων δεδομένων. Αυτές οι πληροφορίες χρησιμοποιούνται περαιτέρω για την ανάπτυξη του μοντέλου πρόβλεψης.

Ένα σημαντικό στοιχείο για την προγνωστική μοντελοποίηση είναι επίσης η εξεύρεση νέων χαρακτηριστικών, στοιχείων και προτύπων, που θα προσφέρουν αυξημένη διορατικότητα ή ακρίβεια στην πρόβλεψη μελλοντικών παρόμοιων περιστατικών. Προκειμένου να εξηγήσουν τη σχέση μεταξύ διαφόρων μεταβλητών, οι ερευνητές εξετάζουν το σύνολο του πληθυσμού. Η γραμμική παλινδρόμηση, οι καμπύλες παλινδρόμησης και άλλα προηγμένα μοντέλα είναι κάποιες από τις στατιστικές μεθόδους που χρησιμοποιούνται στην προγνωστική μοντελοποίηση.

Οι μέθοδοι πρόβλεψης βασίζονται είτε σε μαθηματικά μοντέλα χρησιμοποιώντας ιστορικά δεδομένα προηγουμένων περιόδων είτε σε ποιοτικές μεθόδους χρησιμοποιώντας την εμπειρία

των στελεχών της επιχείρησης είτε σε συνδυασμούς αυτών των δυο. Σε κάθε περίπτωση τα στοιχεία και οι πληροφορίες που χρησιμοποιούνται για τις προβλέψεις θα πρέπει να ανανεώνονται σε συνεχή βάση (π.χ. κάθε μήνα). Με αυτόν τον τρόπο, εξασφαλίζεται αφενός η επικαιροποίηση των προβλέψεων και αφετέρου μειώνονται τα σφάλματα και αυξάνεται η ακρίβεια των προβλέψεων τουλάχιστον βραχυπρόθεσμα.
Η διαδικασία προβλέψεων περιλαμβάνει συνήθως τρεις κυρίες φάσεις ως ακολούθως:

- Συλλογή και ανάλυση ιστορικών στοιχείων και πληροφοριών (Collect & analyzing data).
- Αξιολόγηση παραγόντων που επηρεάζουν (Adding deterministic overrides).
- Παρακολούθηση των προβλέψεων (Management Action).

Οι προβλέψεις πολλές φορές εμπεριέχουν σφάλματα. Τα σφάλματα των προβλέψεων διακρίνονται σε στατιστικά και τυχαία. Τα τυχαία σφάλματα οφείλονται σε μη προβλέψιμους παράγοντες. Αντίθετα, τα στατιστικά σφάλματα αφορούν στο μοντέλο πρόβλεψης και οφείλονται στην κακή εκτίμηση ή παράλειψη παραγόντων που επηρεάζουν το υπό εξέταση μέγεθος, για παράδειγμα εποχικότητα. Το σφάλμα των προβλέψεων μπορεί να μετρηθεί συγκρίνοντας τις προβλέψεις με τις πραγματικές τιμές.

Αρνητικές τιμές υποδηλώνουν υπερεκτίμηση του μεγέθους, ενώ θετικές τιμές δείχνουν υποεκτίμηση αυτού. Ωστόσο, αξίζει επίσης να σημειωθεί ότι μεγάλες θετικές τιμές του σφάλματος πρόβλεψης αντισταθμίζονται από μεγάλες αρνητικές. Για αυτό το λόγο χρησιμοποιούνται κυρίως μετρήσεις με απόλυτες τιμές σφάλματος και μέσες τιμές.

Όπως έχει ήδη παρατηρηθεί από τα προηγούμενα κεφάλαια, εκτός από βιβλιογραφική προσέγγιση πραγματοποιήθηκε και τη δευτερογενή ανάλυση μέσα από την επεξεργασία τόσο ως προς την μέθοδο της παλινδρόμησης όσο και ως προς την μέθοδο των δένδρων απόφασης.

Εκείνο που αξίζει να υπογραμμίσουμε με βάση την παλινδρόμηση είναι ότι τα μοντέλα τα οποία κατασκευάσαμε είχαν ιδιαίτερα σημαντική προβλεπτική ικανότητα καθώς ο συντελεστής προσδιορισμού αυτών ήταν ιδιαίτερα υψηλός. Γεγονός που με τη σειρά του σημαίνει ότι τα

υποδείγματα που δημιουργήσαμε μπορούν να χρησιμοποιηθούν με αξιοπιστία για μελλοντικές προβλέψεις. Από τα παραπάνω συμπεραίνουμε πως η ανάλυση παλινδρόμησης αποτελεί ένα σημαντικό εργαλείο στα χέρια των ιθυνόντων μιας επιχείρησης καθώς μπορεί να συμβάλει σημαντικά στον τρόπο λήψης απόφασης.

Αντίστοιχης σπουδαιότητας είναι και τα δένδρα απόφασης που κατασκευάστηκαν με τη βοήθεια διαφόρων αλγορίθμων. Και στην περίπτωση αυτή, το συμπέρασμα που προκύπτει είναι πως με την κατάλληλη εκπαίδευση και την κατάλληλη μορφή δεδομένων υπάρχει η δυνατότητα ορθής πρόβλεψης για ένα μικρότερο σύνολο δεδομένων μελλοντικά.

Εν κατακλείδι, από όσα προηγήθηκαν συμπεραίνουμε πως η κατάλληλη χρήση των κατάλληλων προβλεπτικών μεθόδων έχει την δυνατότητα να δημιουργήσει πλεονέκτημα για την επιχείρηση, γεγονός ιδιαίτερα σημαντικό για το δυναμικά και συνεχώς μεταβαλλόμενο περιβάλλον μέσα στο οποίο εξελίσσονται οι επιχειρηματικές δραστηριότητες.

Όπως είναι γνωστό από την βιβλιογραφία οι τεχνικές πρόβλεψης αποτελούν «εργαλείο» για όσους διαμορφώνουν πολιτική και λαμβάνουν αποφάσεις καίριας σημασίας.

# ΒΙΒΛΙΟΓΡΑΦΙΑ

D. Agrawal, P. Bernstein, E. Bertino, S. Davidson, U. Dayal, M. Franklin, J. Gehrke, L. Haas, A. Halevy, J. Han, H. V. Jagadish, A. Labrinidis, S. Madden, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, K. Ross, C. Shahabi, D. Suciu, S. Vaithyanathan, and J. Widom, Challenges and opportunities with big data» February 2012

Berry, M., J., A., & Linoff, G., S., Mastering data mining. New York: Wiley, 2000

Cimiano, A., Hoto, A., Staab, S., Learning concept hierarchies from text corpora using formal concept analysis. Journal of Artificial Intelligence Research, 24, 305-339, 2005

Delen, D, Olson, D.L Advanced Data Mining Techniques;. 2008 ISBN:978-3-540-76916-3

Dunham M.H., (2004,): Data Mining introductory and advanced topics", Prentice Hall

Edelstein, H., A. Introduction to data mining and knowledge discovery (3rd ed). Potomac, MD: Two Crows Corp. 1999

Edmund X. DeJesus, Data Mining, Byte.com magazine, October 1995

Greene W.H, "Econometric Methods", Pearson Education, 2011.

Han, J., Kamber, M. Data mining: Concepts and Techniques. New York: Morgan-Kaufman. 2000

U. Fayyad, et al., editors Advances in Knowledge Discovery & Data Mining; AAAI/MIT Press, 1995

U. Fayyad, et al., editors Proceedings of the First International Conference on Knowledge Discovery and Data Mining,; AAAI Press, 1995

Hall A. M., Frank E., Witten H. I (2011): Data Mining, Practical Machine Learning Tools and Techniques

Jeffrey M. Wooldridge, "Εισαγωγή στην οικονομετρία", Εκδόσεις Παπαζήση, Αθήνα, 2006

Johnston, J. and J. Dinardo, "Econometric Methods", McGraw-Hill, 1997.

Krzysztof j. Giosa, G. William Mooree, Uniqueness of medical data mining, Joornal of Artificial intelligence in medicine, 2002.

Kumar V., Steinbach M. (2006): Introduction to Data Mining, Addison Wesley

Lazarevic A., (2008): Data Mining for Anomaly Detection, Tutorial at the European

Maddala, G.S., "Introduction to Econometrics", Wiley, 2002

Montani, S., Portinale, L., Leonardi, G., Bellazzi, R., Case-based retrieval to support the treatment of end stage renal failure patients. Artif Intell Med, 37(1), 31-42, 2006

New Trends in Data Mining by J. Huysmans, B. baesens, d. martens, k. denys and j. vanthienen101 (http://www.econ.kuleuven.be/rebel//jaargangen/2001-2010/2005/TEM%202005-4/TEM_4_05_Huysmans.pdf)

Prodromidis A. and Chan P., (2000): Meta-learning in distributed data mining systems, Issues and Approaches, in Advances of Distributed Data Mining, AAAI Press

J. Ross Quinlan. Morgan Kaufmann Publishers Programs for Machine Learning, Inc., 1993

Sackett, D. L., Rosenberg, W. M., Gray, J. A., Haynes, R B., Richardson, W. S., Evidence based medicine: what it is and what it isn't. BMJ, 312 (7023), 71-2, 2004

Sara Reese Hedberg, The Data Gold Rush, Byte.com magazine, October 1995

Savova, G. K., Ogren, P. V., Duffy, P. H., Buntrock, J. D., Chute, C. G., Mayo clinic NLP system for patient smoking status identification. J Am Med Inform Assoc, 15(1), 25-8, 2008.

Spanos A., "Statistical Foundation of Econometric Modelling", Cambridge University Press, 1986.

Weiss, S. M., & Indurkhya, N. Predictive data mining: A practical guide. New York: Morgan-Kaufman. 1997

Ανδρικόπουλος, Ανδρέας Α, Οικονομετρία, (2003), Εκδόσεις Μπένου Ε.

Βαζιργιάννης Μ., Χαλκίδη Μ., "Εξόρυξη Γνώσης από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό", Εκδ. Gutenberg

Βλαχάβας Ι., Κεφάλας Π., Βασιλειάδης Ν. Τεχνητή Νοημοσύνη, Εκδόσεις Γαργατάνη, 2002

Θεοδωρίδης Γ., Πελέκης Ν. (2011): Εξόρυξη Γνώσης από Δεδομένα - Συσταδοποίηση, Ομάδα Διαχείρισης Δεδομένων Πανεπιστήμιο Πειραιώς

Κάτος, Α., "Οικονομετρία, Θεωρία και Εφαρμογές", Εκδόσεις Ζυγός, 2004

Δερματάς Κ.. (2007): Τεχνητά Νευρωνικά Δίκτυα., Εκδόσεις Κλειδάριθμος, Αθήνα

Πραστάκος Γρηγόρης Π. (2006), Διοικητική Επιστήμη, Λήψη Επιχειρησιακών Αποφάσεων στην κοινωνία της πληροφορίας, Β' έκδοση, Εκδόσεις Σταμούλης, Αθήνα

Σαλατάς Ι. (2011): Υλοποίηση και εφαρμογή Τεχνητών Νευρωνικών Δικτύων για την πρόβλεψη χρονοσειρών συναλλαγματικών ισοτιμιών, Ελληνικό Ανοιχτό Πανεπιστήμιο.

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

Χρήστου Γ. Κ., "Εισαγωγή στην οικονομετρία", Gutenberg, Αθήνα, 2002

http://www.ssc.wisc.edu/~bhansen/

http://www.economicsnetwork.ac.uk/

http://eduportal.dmst.aueb.gr/html/det/Lecture7Chapter8SysthmataEpixeirhmatikhEyfyiasv2-gr__18002.pdf

http://www.ethnodata.gr/

http://www.unisystems.gr/el/solutions-services-inside/company-solutions/epixirisiakes-efarmoges/solutions-bi.html

http://kpe-kastor.kas.sch.gr/peekpe/proceedings/synedria_9_ereunes/Ravasopoulos_et_al.pdf

http://www.cs.uoi.gr/~pitoura/courses/dm09/warehouse09.pdf

http://www.neo2.gr/web/neo2.gr/searchpagebasedontags/-/asset_publisher/Ep0Q/content/13737?redirect=%2Fweb%2Fneo2.gr%2Fviews

# ΠΑΡΑΡΤΗΜΑ 2ου ΚΕΦΑΛΑΙΟΥ

## ΕΞΑΡΤΗΜΕΝΗ ΜΕΤΑΒΛΗΤΗ: SP

| | *Συντελεστές* | *Τυπικό σφάλμα* | *t* | *τιμή-P* |
|---|---|---|---|---|
| Τεταγμένη επί την αρχή | 0,000137 | 0,00037 | 0,371227 | 0,710618 |
| DAX | 0,331717 | 0,073266 | 4,527546 | 7,38E-06 |
| FTSE | 0,13182 | 0,094141 | 1,400238 | 0,162031 |
| NIKKEI | 0,010326 | 0,032263 | 0,320056 | 0,749052 |
| BOVESPA | 0,514259 | 0,03557 | 14,45774 | 3,83E-40 |
| EU | 0,048246 | 0,135591 | 0,355819 | 0,722119 |
| EM | -0,33098 | 0,074967 | -4,41499 | 1,23E-05 |
| TL BASED ISE | -0,0745 | 0,074472 | -1,00042 | 0,317567 |
| USD BASED ISE | 0,079742 | 0,064675 | 1,232974 | 0,218135 |

Στον παραπάνω πίνακα μπορούμε να δούμε τις σχέσεις μεταξύ των μεταβλητών, δηλαδή τις τιμές των συντελεστών $b_i$. Πιο συγκεκριμένα, για θετικούς συντελεστές μπορούμε να αποφανθούμε πως η εξαρτημένη και η ανεξάρτητη μεταβλητή κινούνται προς την ίδια κατεύθυνση (αύξηση της μιας οδηγεί σε αύξηση της άλλης και αντίστροφα). Όταν η τιμή του συντελεστή είναι αρνητική οι μεταβλητές κινούνται προς αντίθετη κατεύθυνση (αύξηση της μιας οδηγεί σε μείωση της άλλης και αντίστροφα). Η τιμή του συντελεστή $b_i$ δείχνει πόσο μεταβάλλεται η εξαρτημένη μεταβλητή όταν η ανεξάρτητη μεταβάλλεται κατά μια μονάδα.

Στην περίπτωσή μας, η εξαρτημένη μεταβλητή, SP κινείται προς την **ίδια** κατεύθυνση με τις εξής μεταβλητές:

- DAX
- FTSE
- NIKKEI
- BOVESPA
- EU
- USD BASED ISE

Ενώ οι μεταβλητές με τις οποίες κινείται προς **αντίθετη** κατεύθυνση είναι οι εξής:

- EM
- TL BASED ISE

Προκειμένου να διαπιστώσουμε ποιες από τις συγκεκριμένες μεταβλητές έχουν στατιστικά σημαντική επίδραση στην εξαρτημένη μεταβλητή θα κάνουμε έλεγχο υπόθεσης. Η μηδενική και η εναλλακτική υπόθεση διαμορφώνονται ως εξής:

*$H_0$: Η επίδραση της εν λόγω μεταβλητής ΔΕΝ είναι στατιστικά σημαντική.*

*$H_1$: Η επίδραση της εν λόγω μεταβλητής είναι στατιστικά σημαντική.*

Για να αποδεχτούμε ή να απορρίψουμε τη μηδενική υπόθεση θα χρησιμοποιήσουμε την τιμή – P. Συγκρίνουμε τη συγκεκριμένη τιμή με το επίπεδο σημαντικότητας 5% (0,05). Αναλυτικότερα,

Αν τιμή – P < 0,05 απορρίπτουμε την $H_0$ δηλαδή η επίδραση της υπό εξέτασης μεταβλητής στην εξαρτημένη μεταβλητή είναι στατιστικά σημαντική.

Αν τιμή – P > 0,05 αποδεχόμαστε την $H_0$ δηλαδή η επίδραση της υπό εξέτασης μεταβλητής στην εξαρτημένη μεταβλητή ΔΕΝ είναι στατιστικά σημαντική.

Σύμφωνα με τα παραπάνω και με τον πίνακα που έχει προκύψει από την επεξεργασία στο excel μπορούμε να σημειώσουμε τα ακόλουθα.

Οι μεταβλητές που επηρεάζουν στατιστικά σημαντικά την εξαρτημένη μεταβλητή είναι οι εξής:

- DAX
- BOVESPA
- EM

Από την πίνακα που ακολουθεί και συγκεκριμένα χρησιμοποιώντας το συντελεστή προσδιορισμού $R^2$ μπορούμε να αποφανθούμε ότι η συνολική μεταβλητότητα της εξαρτημένης μεταβλητής που οφείλεται στις συγκεκριμένες ανεξάρτητες είναι περίπου 65%, (0.6455).

| *Στατιστικά παλινδρόμησης* | |
|---|---|
| Πολλαπλό R | 0,803446 |
| R Τετράγωνο | 0,645526 |
| Προσαρμοσμένο R Τετράγωνο | 0,640144 |
| Τυπικό σφάλμα | 0,008454 |

Μέγεθος δείγματος 536

Τέλος, από τον πίνακα ANOVA, ανάλυσης διακύμανσης μπορούμε να κάνουμε έλεγχο στατιστικής σημαντικότητας για όλο το μοντέλο. Δηλαδή μέσα από τον πίνακα ANOVA μπορούμε να κάνουμε έλεγχο F, όπου η μηδενική και η εναλλακτική υπόθεση διαμορφώνονται ως εξής:

*$H_0$: Η ΑΠΟ ΚΟΙΝΟΥ επίδραση των συγκεκριμένων μεταβλητών στο υπόδειγμα ΔΕΝ είναι στατιστικά σημαντική.*

*$H_1$: Η ΑΠΟ ΚΟΙΝΟΥ επίδραση των συγκεκριμένων μεταβλητών στο υπόδειγμα είναι στατιστικά σημαντική.*

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ

| | *βαθμοί ελευθερίας* | *SS* | *MS* | *F* | *Σημαντικότητα F* |
|---|---|---|---|---|---|
| Παλινδρόμηση | 8 | 0,068592 | 0,008574 | 119,9635 | 1,7E-113 |
| Υπόλοιπο | 527 | 0,037665 | 7,15E-05 | | |
| Σύνολο | 535 | 0,106257 | | | |

Και σε αυτή την περίπτωση η αποδοχή ή απόρριψη της μηδενικής υπόθεσης γίνεται με κριτήριο την τιμή «Σημαντικότητα F»

Αν Σημαντικότητα F < 0,05 απορρίπτουμε την μηδενική υπόθεση ($H_0$), επομένως η από κοινού επίδραση των μεταβλητών είναι στατιστικά σημαντική.

Αν Σημαντικότητα F > 0,05 αποδεχόμαστε την μηδενική υπόθεση ($H_0$), επομένως η από κοινού επίδραση των μεταβλητών ΔΕΝ είναι στατιστικά σημαντική.

## ΕΞΑΡΤΗΜΕΝΗ ΜΕΤΑΒΛΗΤΗ: DAX

| | *Συντελεστές* | *Τυπικό σφάλμα* | *t* | *τιμή-P* |
|---|---|---|---|---|
| Τεταγμένη επί την αρχή | 0,000205 | 0,000216 | 0,948471 | 0,343325 |
| FTSE | -0,2618 | 0,053821 | -4,86429 | 1,52E-06 |
| NIKKEI | -0,00198 | 0,018821 | -0,10499 | 0,916424 |
| BOVESPA | -0,05732 | 0,024393 | -2,3497 | 0,019157 |
| EU | 1,249826 | 0,057385 | 21,77957 | 1,76E-75 |
| EM | 0,063157 | 0,044446 | 1,420988 | 0,155912 |
| TL BASED ISE | 0,066639 | 0,043385 | 1,535991 | 0,125141 |
| USD BASED ISE | -0,08016 | 0,037618 | -2,13098 | 0,033553 |
| SP | 0,112869 | 0,024929 | 4,527546 | 7,38E-06 |

Στον παραπάνω πίνακα μπορούμε να δούμε τις σχέσεις μεταξύ των μεταβλητών, δηλαδή τις τιμές των συντελεστών $b_i$. Πιο συγκεκριμένα, για θετικούς συντελεστές μπορούμε να αποφανθούμε πως η εξαρτημένη και η ανεξάρτητη μεταβλητή κινούνται προς την ίδια κατεύθυνση (αύξηση της μιας οδηγεί σε αύξηση της άλλης και αντίστροφα). Όταν η τιμή του συντελεστή είναι αρνητική οι μεταβλητές κινούνται προς αντίθετη κατεύθυνση (αύξηση της μιας οδηγεί σε μείωση της άλλης και αντίστροφα). Η τιμή του συντελεστή $b_i$ δείχνει πόσο μεταβάλλεται η εξαρτημένη μεταβλητή όταν η ανεξάρτητη μεταβάλλεται κατά μια μονάδα.

Στην περίπτωσή μας, η εξαρτημένη μεταβλητή, DAX κινείται προς την **ίδια** κατεύθυνση με τις εξής μεταβλητές:

- EU
- EM
- TL BASED ISE
- SP

Ενώ οι μεταβλητές με τις οποίες κινείται προς **αντίθετη** κατεύθυνση είναι οι εξής:

- FTSE
- NIKKEI
- BOVESPA

- USD BASED ISE

Προκειμένου να διαπιστώσουμε ποιες από τις συγκεκριμένες μεταβλητές έχουν στατιστικά σημαντική επίδραση στην εξαρτημένη μεταβλητή θα κάνουμε έλεγχο υπόθεσης. Η μηδενική και η εναλλακτική υπόθεση διαμορφώνονται ως εξής:

*$H_0$: Η επίδραση της εν λόγω μεταβλητής ΔΕΝ είναι στατιστικά σημαντική.*

*$H_1$: Η επίδραση της εν λόγω μεταβλητής είναι στατιστικά σημαντική.*

Για να αποδεχτούμε ή να απορρίψουμε τη μηδενική υπόθεση θα χρησιμοποιήσουμε την τιμή – P. Συγκρίνουμε τη συγκεκριμένη τιμή με το επίπεδο σημαντικότητας 5% (0,05). Αναλυτικότερα,

Αν τιμή – P < 0,05 απορρίπτουμε την $H_0$ δηλαδή η επίδραση της υπό εξέτασης μεταβλητής στην εξαρτημένη μεταβλητή είναι στατιστικά σημαντική.

Αν τιμή – P > 0,05 αποδεχόμαστε την $H_0$ δηλαδή η επίδραση της υπό εξέτασης μεταβλητής στην εξαρτημένη μεταβλητή ΔΕΝ είναι στατιστικά σημαντική.

Σύμφωνα με τα παραπάνω και με τον πίνακα που έχει προκύψει από την επεξεργασία στο excel μπορούμε να σημειώσουμε τα ακόλουθα.

Οι μεταβλητές που επηρεάζουν στατιστικά σημαντικά την εξαρτημένη μεταβλητή είναι οι εξής:

- FTSE
- EU
- SP

Από την πίνακα που ακολουθεί και συγκεκριμένα χρησιμοποιώντας το συντελεστή προσδιορισμού $R^2$ μπορούμε να αποφανθούμε ότι η συνολική μεταβλητότητα της εξαρτημένης μεταβλητής που οφείλεται στις συγκεκριμένες ανεξάρτητες είναι περίπου 89%, (0.886957).

| *Στατιστικά παλινδρόμησης* | |
|---|---|
| Πολλαπλό R | 0,941784 |
| R Τετράγωνο | 0,886957 |
| Προσαρμοσμένο R Τετράγωνο | 0,885241 |
| Τυπικό σφάλμα | 0,004931 |
| Μέγεθος δείγματος | 536 |

Τέλος, από τον πίνακα ANOVA, ανάλυσης διακύμανσης μπορούμε να κάνουμε έλεγχο στατιστικής σημαντικότητας για όλο το μοντέλο. Δηλαδή μέσα από τον πίνακα ANOVA μπορούμε να κάνουμε έλεγχο F, όπου η μηδενική και η εναλλακτική υπόθεση διαμορφώνονται ως εξής:

*$H_0$: Η ΑΠΟ ΚΟΙΝΟΥ επίδραση των συγκεκριμένων μεταβλητών στο υπόδειγμα ΔΕΝ είναι στατιστικά σημαντική.*

*$H_1$: Η ΑΠΟ ΚΟΙΝΟΥ επίδραση των συγκεκριμένων μεταβλητών στο υπόδειγμα είναι στατιστικά σημαντική.*

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ

| | *βαθμοί ελευθερίας* | *SS* | *MS* | *F* | *Σημαντικότητα F* |
|---|---|---|---|---|---|
| Παλινδρόμηση | 8 | 0,100557 | 0,01257 | 516,868 | 7,4E-244 |
| Υπόλοιπο | 527 | 0,012816 | 2,43E-05 | | |
| Σύνολο | 535 | 0,113373 | | | |

Και σε αυτή την περίπτωση η αποδοχή ή απόρριψη της μηδενικής υπόθεσης γίνεται με κριτήριο την τιμή «Σημαντικότητα F».

Αν Σημαντικότητα F < 0,05 απορρίπτουμε την μηδενική υπόθεση ($H_0$), επομένως η από κοινού επίδραση των μεταβλητών είναι στατιστικά σημαντική.

Αν Σημαντικότητα F > 0,05 αποδεχόμαστε την μηδενική υπόθεση ($H_0$), επομένως η από κοινού επίδραση των μεταβλητών ΔΕΝ είναι στατιστικά σημαντική.

## ΕΞΑΡΤΗΜΕΝΗ ΜΕΤΑΒΛΗΤΗ: FTSE

| | *Συντελεστές* | *Τυπικό σφάλμα* | *t* | *τιμή-P* |
|---|---|---|---|---|
| Τεταγμένη επί την αρχή | 5,31E-05 | 0,000171 | 0,310684 | 0,756164 |
| NIKKEI | -0,01887 | 0,01488 | -1,26802 | 0,205349 |
| BOVESPA | -0,02534 | 0,019383 | -1,30725 | 0,191699 |
| EU | 1,075968 | 0,041544 | 25,89939 | 5,14E-96 |
| EM | 0,083719 | 0,035069 | 2,387243 | 0,017327 |
| TL BASED ISE | 0,069952 | 0,034293 | 2,039831 | 0,041865 |
| USD BASED ISE | -0,07199 | 0,029749 | -2,41992 | 0,015861 |
| SP | 0,028119 | 0,020082 | 1,400238 | 0,162031 |
| DAX | -0,16413 | 0,033741 | -4,86429 | 1,52E-06 |

Στον παραπάνω πίνακα μπορούμε να δούμε τις σχέσεις μεταξύ των μεταβλητών, δηλαδή τις τιμές των συντελεστών $b_i$. Πιο συγκεκριμένα, για θετικούς συντελεστές μπορούμε να αποφανθούμε πως η εξαρτημένη και η ανεξάρτητη μεταβλητή κινούνται προς την ίδια κατεύθυνση (αύξηση της μιας οδηγεί σε αύξηση της άλλης και αντίστροφα). Όταν η τιμή του συντελεστή είναι αρνητική οι μεταβλητές κινούνται προς αντίθετη κατεύθυνση (αύξηση της μιας οδηγεί σε μείωση της άλλης και αντίστροφα). Η τιμή του συντελεστή $b_i$ δείχνει πόσο μεταβάλλεται η εξαρτημένη μεταβλητή όταν η ανεξάρτητη μεταβάλλεται κατά μια μονάδα.

Στην περίπτωσή μας, η εξαρτημένη μεταβλητή, FTSE κινείται προς την **ίδια** κατεύθυνση με τις εξής μεταβλητές:

- EU
- EM
- TL BASED ISE
- SP

Ενώ οι μεταβλητές με τις οποίες κινείται προς **αντίθετη** κατεύθυνση είναι οι εξής:

- NIKKEI
- BOVESPA
- DAX

- USD BASED ISE

Προκειμένου να διαπιστώσουμε ποιες από τις συγκεκριμένες μεταβλητές έχουν στατιστικά σημαντική επίδραση στην εξαρτημένη μεταβλητή θα κάνουμε έλεγχο υπόθεσης. Η μηδενική και η εναλλακτική υπόθεση διαμορφώνονται ως εξής:

*$H_0$: Η επίδραση της εν λόγω μεταβλητής ΔΕΝ είναι στατιστικά σημαντική.*

*$H_1$: Η επίδραση της εν λόγω μεταβλητής είναι στατιστικά σημαντική.*

Για να αποδεχτούμε ή να απορρίψουμε τη μηδενική υπόθεση θα χρησιμοποιήσουμε την τιμή – P. Συγκρίνουμε τη συγκεκριμένη τιμή με το επίπεδο σημαντικότητας 5% (0,05). Αναλυτικότερα,

Αν τιμή – P < 0,05 απορρίπτουμε την $H_0$ δηλαδή η επίδραση της υπό εξέτασης μεταβλητής στην εξαρτημένη μεταβλητή είναι στατιστικά σημαντική.

Αν τιμή – P > 0,05 αποδεχόμαστε την $H_0$ δηλαδή η επίδραση της υπό εξέτασης μεταβλητής στην εξαρτημένη μεταβλητή ΔΕΝ είναι στατιστικά σημαντική.

Σύμφωνα με τα παραπάνω και με τον πίνακα που έχει προκύψει από την επεξεργασία στο excel μπορούμε να σημειώσουμε τα ακόλουθα.

Οι μεταβλητές που επηρεάζουν στατιστικά σημαντικά την εξαρτημένη μεταβλητή είναι οι εξής:

- EU
- DAX
- EM
- TL BASED ISE
- USD BASED ISE

Από την πίνακα που ακολουθεί και συγκεκριμένα χρησιμοποιώντας το συντελεστή προσδιορισμού $R^2$ μπορούμε να αποφανθούμε ότι η συνολική μεταβλητότητα της εξαρτημένης μεταβλητής που οφείλεται στις συγκεκριμένες ανεξάρτητες είναι περίπου 91%, (0.906236).

| *Στατιστικά παλινδρόμησης* | |
|---|---|
| Πολλαπλό R | 0,951964 |
| R Τετράγωνο | 0,906236 |
| Προσαρμοσμένο R Τετράγωνο | 0,904813 |

| | |
|---|---|
| Τυπικό σφάλμα | 0,003905 |
| Μέγεθος δείγματος | 536 |

Τέλος, από τον πίνακα ANOVA, ανάλυσης διακύμανσης μπορούμε να κάνουμε έλεγχο στατιστικής σημαντικότητας για όλο το μοντέλο. Δηλαδή μέσα από τον πίνακα ANOVA μπορούμε να κάνουμε έλεγχο F, όπου η μηδενική και η εναλλακτική υπόθεση διαμορφώνονται ως εξής:

*$H_0$: Η ΑΠΟ ΚΟΙΝΟΥ επίδραση των συγκεκριμένων μεταβλητών στο υπόδειγμα ΔΕΝ είναι στατιστικά σημαντική.*

*$H_1$: Η ΑΠΟ ΚΟΙΝΟΥ επίδραση των συγκεκριμένων μεταβλητών στο υπόδειγμα είναι στατιστικά σημαντική.*

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ

| | *βαθμοί ελευθερίας* | *SS* | *MS* | *F* | *Σημαντικότητα F* |
|---|---|---|---|---|---|
| Παλινδρόμηση | 8 | 0,077655 | 0,009707 | 636,6875 | 3,1E-265 |
| Υπόλοιπο | 527 | 0,008035 | 1,52E-05 | | |
| Σύνολο | 535 | 0,085689 | | | |

Και σε αυτή την περίπτωση η αποδοχή ή απόρριψη της μηδενικής υπόθεσης γίνεται με κριτήριο την τιμή «Σημαντικότητα F».

Αν Σημαντικότητα F < 0,05 απορρίπτουμε την μηδενική υπόθεση ($H_0$), επομένως η από κοινού επίδραση των μεταβλητών είναι στατιστικά σημαντική.

Αν Σημαντικότητα F > 0,05 αποδεχόμαστε την μηδενική υπόθεση ($H_0$), επομένως η από κοινού επίδραση των μεταβλητών ΔΕΝ είναι στατιστικά σημαντική.

## ΕΞΑΡΤΗΜΕΝΗ ΜΕΤΑΒΛΗΤΗ: NIKKIE

| | *Συντελεστές* | *Τυπικό σφάλμα* | *t* | *τιμή-P* |
|---|---|---|---|---|
| Τεταγμένη επί την αρχή | -0,00019 | 0,0005 | -0,38253 | 0,702221 |
| BOVESPA | -0,28773 | 0,055349 | -5,19837 | 2,88E-07 |
| EU | 0,00647 | 0,183076 | 0,035342 | 0,971821 |
| EM | 1,063735 | 0,092059 | 11,5549 | 1,1E-27 |
| TL BASED ISE | -0,49007 | 0,098346 | -4,98317 | 8,5E-07 |
| USD BASED ISE | 0,415793 | 0,085543 | 4,860624 | 1,55E-06 |
| SP | 0,01882 | 0,058803 | 0,320056 | 0,749052 |
| DAX | -0,01058 | 0,100817 | -0,10499 | 0,916424 |
| FTSE | -0,16121 | 0,127137 | -1,26802 | 0,205349 |

Στον παραπάνω πίνακα μπορούμε να δούμε τις σχέσεις μεταξύ των μεταβλητών, δηλαδή τις τιμές των συντελεστών $b_i$. Πιο συγκεκριμένα, για θετικούς συντελεστές μπορούμε να αποφανθούμε πως η εξαρτημένη και η ανεξάρτητη μεταβλητή κινούνται προς την ίδια κατεύθυνση (αύξηση της μιας οδηγεί σε αύξηση της άλλης και αντίστροφα). Όταν η τιμή του συντελεστή είναι αρνητική οι μεταβλητές κινούνται προς αντίθετη κατεύθυνση (αύξηση της μιας οδηγεί σε μείωση της άλλης και αντίστροφα). Η τιμή του συντελεστή $b_i$ δείχνει πόσο μεταβάλλεται η εξαρτημένη μεταβλητή όταν η ανεξάρτητη μεταβάλλεται κατά μια μονάδα.

Στην περίπτωσή μας, η εξαρτημένη μεταβλητή, NIKKIE κινείται προς την **ίδια** κατεύθυνση με τις εξής μεταβλητές:

- EU
- EM
- USD BASED ISE
- SP

Ενώ οι μεταβλητές με τις οποίες κινείται προς **αντίθετη** κατεύθυνση είναι οι εξής:

- BOVESPA
- TL BASED ISE
- DAX
- FTSE

Προκειμένου να διαπιστώσουμε ποιες από τις συγκεκριμένες μεταβλητές έχουν στατιστικά σημαντική επίδραση στην εξαρτημένη μεταβλητή θα κάνουμε έλεγχο υπόθεσης. Η μηδενική και η εναλλακτική υπόθεση διαμορφώνονται ως εξής:

*$H_0$: Η επίδραση της εν λόγω μεταβλητής ΔΕΝ είναι στατιστικά σημαντική.*

*$H_1$: Η επίδραση της εν λόγω μεταβλητής είναι στατιστικά σημαντική.*

Για να αποδεχτούμε ή να απορρίψουμε τη μηδενική υπόθεση θα χρησιμοποιήσουμε την τιμή – P. Συγκρίνουμε τη συγκεκριμένη τιμή με το επίπεδο σημαντικότητας 5% (0,05). Αναλυτικότερα,

Αν τιμή – P < 0,05 απορρίπτουμε την $H_0$ δηλαδή η επίδραση της υπό εξέτασης μεταβλητής στην εξαρτημένη μεταβλητή είναι στατιστικά σημαντική.

Αν τιμή – P > 0,05 αποδεχόμαστε την $H_0$ δηλαδή η επίδραση της υπό εξέτασης μεταβλητής στην εξαρτημένη μεταβλητή ΔΕΝ είναι στατιστικά σημαντική.

Σύμφωνα με τα παραπάνω και με τον πίνακα που έχει προκύψει από την επεξεργασία στο excel μπορούμε να σημειώσουμε τα ακόλουθα.

Οι μεταβλητές που επηρεάζουν στατιστικά σημαντικά την εξαρτημένη μεταβλητή είναι οι εξής:

- BOVESPA
- EM
- TL BASED ISE
- USD BASED ISE

Από την πίνακα που ακολουθεί και συγκεκριμένα χρησιμοποιώντας το συντελεστή προσδιορισμού $R^2$ μπορούμε να αποφανθούμε ότι η συνολική μεταβλητότητα της εξαρτημένης μεταβλητής που οφείλεται στις συγκεκριμένες ανεξάρτητες είναι περίπου 42%, (0.41813).

| *Στατιστικά παλινδρόμησης* | |
|---|---|
| Πολλαπλό R | 0,64663 |
| R Τετράγωνο | 0,41813 |
| Προσαρμοσμένο R Τετράγωνο | 0,409297 |
| Τυπικό σφάλμα | 0,011413 |
| Μέγεθος δείγματος | 536 |

Τέλος, από τον πίνακα ANOVA, ανάλυσης διακύμανσης μπορούμε να κάνουμε έλεγχο στατιστικής σημαντικότητας για όλο το μοντέλο. Δηλαδή μέσα από τον πίνακα ANOVA μπορούμε να κάνουμε έλεγχο F, όπου η μηδενική και η εναλλακτική υπόθεση διαμορφώνονται ως εξής:

*$H_0$: Η ΑΠΟ ΚΟΙΝΟΥ επίδραση των συγκεκριμένων μεταβλητών στο υπόδειγμα ΔΕΝ είναι στατιστικά σημαντική.*

*$H_1$: Η ΑΠΟ ΚΟΙΝΟΥ επίδραση των συγκεκριμένων μεταβλητών στο υπόδειγμα είναι στατιστικά σημαντική.*

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ

| | *βαθμοί ελευθερίας* | *SS* | *MS* | *F* | *Σημαντικότητα F* |
|---|---|---|---|---|---|
| Παλινδρόμηση | 8 | 0,049332 | 0,006166 | 47,33758 | 2,49E-57 |
| Υπόλοιπο | 527 | 0,06865 | 0,00013 | | |
| Σύνολο | 535 | 0,117982 | | | |

Και σε αυτή την περίπτωση η αποδοχή ή απόρριψη της μηδενικής υπόθεσης γίνεται με κριτήριο την τιμή «Σημαντικότητα F».

Αν Σημαντικότητα F < 0,05 απορρίπτουμε την μηδενική υπόθεση ($H_0$), επομένως η από κοινού επίδραση των μεταβλητών είναι στατιστικά σημαντική.

Αν Σημαντικότητα F > 0,05 αποδεχόμαστε την μηδενική υπόθεση ($H_0$), επομένως η από κοινού επίδραση των μεταβλητών ΔΕΝ είναι στατιστικά σημαντική.

## ΕΞΑΡΤΗΜΕΝΗ ΜΕΤΑΒΛΗΤΗ: BOVESPA

| | *Συντελεστές* | *Τυπικό σφάλμα* | *t* | *τιμή-P* |
|---|---|---|---|---|
| Τεταγμένη επί την αρχή | -0,00019 | 0,000384 | -0,48777 | 0,625913 |
| EU | 0,245895 | 0,140117 | 1,754931 | 0,079852 |
| EM | 1,014904 | 0,065603 | 15,4704 | 8,81E-45 |
| TL BASED ISE | 0,258022 | 0,076424 | 3,376176 | 0,000789 |
| USD BASED ISE | -0,3035 | 0,065802 | -4,61229 | 5,01E-06 |
| SP | 0,552237 | 0,038197 | 14,45774 | 3,83E-40 |
| DAX | -0,18089 | 0,076984 | -2,3497 | 0,019157 |
| FTSE | -0,12756 | 0,097579 | -1,30725 | 0,191699 |
| NIKKEI | -0,16952 | 0,032611 | -5,19837 | 2,88E-07 |

Στον παραπάνω πίνακα μπορούμε να δούμε τις σχέσεις μεταξύ των μεταβλητών, δηλαδή τις τιμές των συντελεστών $b_i$. Πιο συγκεκριμένα, για θετικούς συντελεστές μπορούμε να αποφανθούμε πως η εξαρτημένη και η ανεξάρτητη μεταβλητή κινούνται προς την ίδια κατεύθυνση (αύξηση της μιας οδηγεί σε αύξηση της άλλης και αντίστροφα). Όταν η τιμή του συντελεστή είναι αρνητική οι μεταβλητές κινούνται προς αντίθετη κατεύθυνση (αύξηση της μιας οδηγεί σε μείωση της άλλης και αντίστροφα). Η τιμή του συντελεστή $b_i$ δείχνει πόσο μεταβάλλεται η εξαρτημένη μεταβλητή όταν η ανεξάρτητη μεταβάλλεται κατά μια μονάδα.

Στην περίπτωσή μας, η εξαρτημένη μεταβλητή, BOVESPA κινείται προς την **ίδια** κατεύθυνση με τις εξής μεταβλητές:

- SP
- EU
- EM
- TL BASED ISE

Ενώ οι μεταβλητές με τις οποίες κινείται προς **αντίθετη** κατεύθυνση είναι οι εξής:

- USD BASED ISE
- DAX
- FTSE
- NIKKEI

Προκειμένου να διαπιστώσουμε ποιες από τις συγκεκριμένες μεταβλητές έχουν στατιστικά σημαντική επίδραση στην εξαρτημένη μεταβλητή θα κάνουμε έλεγχο υπόθεσης. Η μηδενική και η εναλλακτική υπόθεση διαμορφώνονται ως εξής:

*$H_0$: Η επίδραση της εν λόγω μεταβλητής ΔΕΝ είναι στατιστικά σημαντική.*

*$H_1$: Η επίδραση της εν λόγω μεταβλητής είναι στατιστικά σημαντική.*

Για να αποδεχτούμε ή να απορρίψουμε τη μηδενική υπόθεση θα χρησιμοποιήσουμε την τιμή – P. Συγκρίνουμε τη συγκεκριμένη τιμή με το επίπεδο σημαντικότητας 5% (0,05). Αναλυτικότερα,

Αν τιμή – P < 0,05 απορρίπτουμε την $H_0$ δηλαδή η επίδραση της υπό εξέτασης μεταβλητής στην εξαρτημένη μεταβλητή είναι στατιστικά σημαντική.

Αν τιμή – P > 0,05 αποδεχόμαστε την $H_0$ δηλαδή η επίδραση της υπό εξέτασης μεταβλητής στην εξαρτημένη μεταβλητή ΔΕΝ είναι στατιστικά σημαντική.

Σύμφωνα με τα παραπάνω και με τον πίνακα που έχει προκύψει από την επεξεργασία στο excel μπορούμε να σημειώσουμε τα ακόλουθα.

Οι μεταβλητές που επηρεάζουν στατιστικά σημαντικά την εξαρτημένη μεταβλητή είναι οι εξής:

- EU
- EM
- TL BASED ISE
- USD BASED ISE
- SP
- DAX
- NIKKEI

Από την πίνακα που ακολουθεί και συγκεκριμένα χρησιμοποιώντας το συντελεστή προσδιορισμού $R^2$ μπορούμε να αποφανθούμε ότι η συνολική μεταβλητότητα της εξαρτημένης μεταβλητής που οφείλεται στις συγκεκριμένες ανεξάρτητες είναι περίπου 70%, (0.695).

| *Στατιστικά παλινδρόμησης* | |
|---|---|
| Πολλαπλό R | 0,833817 |
| R Τετράγωνο | 0,695251 |

| | |
|---|---|
| Προσαρμοσμένο R Τετράγωνο | 0,690625 |
| Τυπικό σφάλμα | 0,008761 |
| Μέγεθος δείγματος | 536 |

Τέλος, από τον πίνακα ANOVA, ανάλυσης διακύμανσης μπορούμε να κάνουμε έλεγχο στατιστικής σημαντικότητας για όλο το μοντέλο. Δηλαδή μέσα από τον πίνακα ANOVA μπορούμε να κάνουμε έλεγχο F, όπου η μηδενική και η εναλλακτική υπόθεση διαμορφώνονται ως εξής:

*$H_0$: Η ΑΠΟ ΚΟΙΝΟΥ επίδραση των συγκεκριμένων μεταβλητών στο υπόδειγμα ΔΕΝ είναι στατιστικά σημαντική.*

*$H_1$: Η ΑΠΟ ΚΟΙΝΟΥ επίδραση των συγκεκριμένων μεταβλητών στο υπόδειγμα είναι στατιστικά σημαντική.*

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ

| | *βαθμοί ελευθερίας* | *SS* | *MS* | *F* | *Σημαντικότητα F* |
|---|---|---|---|---|---|
| Παλινδρόμηση | 8 | 0,092276 | 0,011534 | 150,2868 | 1,1E-130 |
| Υπόλοιπο | 527 | 0,040447 | 7,67E-05 | | |
| Σύνολο | 535 | 0,132723 | | | |

Και σε αυτή την περίπτωση η αποδοχή ή απόρριψη της μηδενικής υπόθεσης γίνεται με κριτήριο την τιμή «Σημαντικότητα F».

Αν Σημαντικότητα F < 0,05 απορρίπτουμε την μηδενική υπόθεση ($H_0$), επομένως η από κοινού επίδραση των μεταβλητών είναι στατιστικά σημαντική.

Αν Σημαντικότητα F > 0,05 αποδεχόμαστε την μηδενική υπόθεση ($H_0$), επομένως η από κοινού επίδραση των μεταβλητών ΔΕΝ είναι στατιστικά σημαντική.

## ΕΞΑΡΤΗΜΕΝΗ ΜΕΤΑΒΛΗΤΗ: EU

| | *Συντελεστές* | *Τυπικό σφάλμα* | *t* | *τιμή-P* |
|---|---|---|---|---|
| Τεταγμένη επί την αρχή | -0,00015 | 0,000119 | -1,30329 | 0,193043 |
| EM | 0,010128 | 0,024519 | 0,413067 | 0,679725 |
| TL BASED ISE | -0,03674 | 0,023892 | -1,53763 | 0,12474 |
| USD BASED ISE | 0,07181 | 0,020569 | 3,491248 | 0,000521 |
| SP | 0,004978 | 0,013991 | 0,355819 | 0,722119 |
| DAX | 0,379021 | 0,017403 | 21,77957 | 1,76E-75 |
| FTSE | 0,520479 | 0,020096 | 25,89939 | 5,14E-96 |
| NIKKEI | 0,000366 | 0,010365 | 0,035342 | 0,971821 |
| BOVESPA | 0,023628 | 0,013464 | 1,754931 | 0,079852 |

Στον παραπάνω πίνακα μπορούμε να δούμε τις σχέσεις μεταξύ των μεταβλητών, δηλαδή τις τιμές των συντελεστών $b_i$. Πιο συγκεκριμένα, για θετικούς συντελεστές μπορούμε να αποφανθούμε πως η εξαρτημένη και η ανεξάρτητη μεταβλητή κινούνται προς την ίδια κατεύθυνση (αύξηση της μιας οδηγεί σε αύξηση της άλλης και αντίστροφα). Όταν η τιμή του συντελεστή είναι αρνητική οι μεταβλητές κινούνται προς αντίθετη κατεύθυνση (αύξηση της μιας οδηγεί σε μείωση της άλλης και αντίστροφα). Η τιμή του συντελεστή $b_i$ δείχνει πόσο μεταβάλλεται η εξαρτημένη μεταβλητή όταν η ανεξάρτητη μεταβάλλεται κατά μια μονάδα.

Στην περίπτωσή μας, η εξαρτημένη μεταβλητή, EU κινείται προς την **ίδια** κατεύθυνση με τις εξής μεταβλητές:

- EM
- USD BASED ISE
- SP
- DAX
- FTSE
- NIKKEI
- BOVESPA

Ενώ η μεταβλητή με την οποία κινείται προς **αντίθετη** κατεύθυνση είναι η TL BASED ISE.

Προκειμένου να διαπιστώσουμε ποιες από τις συγκεκριμένες μεταβλητές έχουν στατιστικά σημαντική επίδραση στην εξαρτημένη μεταβλητή θα κάνουμε έλεγχο υπόθεσης. Η μηδενική και η εναλλακτική υπόθεση διαμορφώνονται ως εξής:

*$H_0$: Η επίδραση της εν λόγω μεταβλητής ΔΕΝ είναι στατιστικά σημαντική.*

*$H_1$: Η επίδραση της εν λόγω μεταβλητής είναι στατιστικά σημαντική.*

Για να αποδεχτούμε ή να απορρίψουμε τη μηδενική υπόθεση θα χρησιμοποιήσουμε την τιμή – P. Συγκρίνουμε τη συγκεκριμένη τιμή με το επίπεδο σημαντικότητας 5% (0,05). Αναλυτικότερα,

Αν τιμή – P < 0,05 απορρίπτουμε την $H_0$ δηλαδή η επίδραση της υπό εξέτασης μεταβλητής στην εξαρτημένη μεταβλητή είναι στατιστικά σημαντική.

Αν τιμή – P > 0,05 αποδεχόμαστε την $H_0$ δηλαδή η επίδραση της υπό εξέτασης μεταβλητής στην εξαρτημένη μεταβλητή ΔΕΝ είναι στατιστικά σημαντική.

Σύμφωνα με τα παραπάνω και με τον πίνακα που έχει προκύψει από την επεξεργασία στο excel μπορούμε να σημειώσουμε τα ακόλουθα.

Οι μεταβλητές που επηρεάζουν στατιστικά σημαντικά την εξαρτημένη μεταβλητή είναι οι εξής:

- DAX
- FTSE
- USD BASED ISE

Από την πίνακα που ακολουθεί και συγκεκριμένα χρησιμοποιώντας το συντελεστή προσδιορισμού $R^2$ μπορούμε να αποφανθούμε ότι η συνολική μεταβλητότητα της εξαρτημένης μεταβλητής που οφείλεται στις συγκεκριμένες ανεξάρτητες είναι περίπου 96%, (0.9569).

| *Στατιστικά παλινδρόμησης* | |
|---|---|
| Πολλαπλό R | 0,978237 |
| R Τετράγωνο | 0,956948 |
| Προσαρμοσμένο R Τετράγωνο | 0,956295 |
| Τυπικό σφάλμα | 0,002716 |
| Μέγεθος δείγματος | 536 |

Τέλος, από τον πίνακα ANOVA, ανάλυσης διακύμανσης μπορούμε να κάνουμε έλεγχο στατιστικής σημαντικότητας για όλο το μοντέλο. Δηλαδή μέσα από τον πίνακα ANOVA μπορούμε να κάνουμε έλεγχο F, όπου η μηδενική και η εναλλακτική υπόθεση διαμορφώνονται ως εξής:

*$H_0$: Η ΑΠΟ ΚΟΙΝΟΥ επίδραση των συγκεκριμένων μεταβλητών στο υπόδειγμα ΔΕΝ είναι στατιστικά σημαντική.*

*$H_1$: Η ΑΠΟ ΚΟΙΝΟΥ επίδραση των συγκεκριμένων μεταβλητών στο υπόδειγμα είναι στατιστικά σημαντική.*

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ

| | *βαθμοί ελευθερίας* | *SS* | *MS* | *F* | *Σημαντικότητα F* |
|---|---|---|---|---|---|
| Παλινδρόμηση | 8 | 0,08639 | 0,010799 | 1464,259 | 0 |
| Υπόλοιπο | 527 | 0,003887 | 7,37E-06 | | |
| Σύνολο | 535 | 0,090276 | | | |

Και σε αυτή την περίπτωση η αποδοχή ή απόρριψη της μηδενικής υπόθεσης γίνεται με κριτήριο την τιμή «Σημαντικότητα F»

Αν Σημαντικότητα F < 0,05 απορρίπτουμε την μηδενική υπόθεση ($H_0$), επομένως η από κοινού επίδραση των μεταβλητών είναι στατιστικά σημαντική.

Αν Σημαντικότητα F > 0,05 αποδεχόμαστε την μηδενική υπόθεση ($H_0$), επομένως η από κοινού επίδραση των μεταβλητών ΔΕΝ είναι στατιστικά σημαντική.

## ΕΞΑΡΤΗΜΕΝΗ ΜΕΤΑΒΛΗΤΗ: ΕΜ

| | *Συντελεστές* | *Τυπικό σφάλμα* | *t* | *τιμή-P* |
|---|---|---|---|---|
| Τεταγμένη επί την αρχή | 0,000432 | 0,00021 | 2,055494 | 0,040324 |
| TL BASED ISE | -0,21837 | 0,041458 | -5,26722 | 2,02E-07 |
| USD BASED ISE | 0,29532 | 0,034646 | 8,523929 | 1,63E-16 |
| SP | -0,10776 | 0,024409 | -4,41499 | 1,23E-05 |
| DAX | 0,060435 | 0,04253 | 1,420988 | 0,155912 |
| FTSE | 0,127787 | 0,053529 | 2,387243 | 0,017327 |
| NIKKEI | 0,190027 | 0,016446 | 11,5549 | 1,1E-27 |
| BOVESPA | 0,307723 | 0,019891 | 15,4704 | 8,81E-45 |
| EU | 0,031958 | 0,077367 | 0,413067 | 0,679725 |

Στον παραπάνω πίνακα μπορούμε να δούμε τις σχέσεις μεταξύ των μεταβλητών, δηλαδή τις τιμές των συντελεστών $b_i$. Πιο συγκεκριμένα, για θετικούς συντελεστές μπορούμε να αποφανθούμε πως η εξαρτημένη και η ανεξάρτητη μεταβλητή κινούνται προς την ίδια κατεύθυνση (αύξηση της μιας οδηγεί σε αύξηση της άλλης και αντίστροφα). Όταν η τιμή του συντελεστή είναι αρνητική οι μεταβλητές κινούνται προς αντίθετη κατεύθυνση (αύξηση της μιας οδηγεί σε μείωση της άλλης και αντίστροφα). Η τιμή του συντελεστή $b_i$ δείχνει πόσο μεταβάλλεται η εξαρτημένη μεταβλητή όταν η ανεξάρτητη μεταβάλλεται κατά μια μονάδα.

Στην περίπτωσή μας, η εξαρτημένη μεταβλητή, ΕΜ κινείται προς την **ίδια** κατεύθυνση με τις εξής μεταβλητές:

- USD BASED ISE
- DAX
- FTSE
- NIKKEI
- BOVESPA
- EU

Ενώ οι μεταβλητές με τις οποίες κινείται προς **αντίθετη** κατεύθυνση είναι οι εξής:

- TL BASED ISE

- SP

Προκειμένου να διαπιστώσουμε ποιες από τις συγκεκριμένες μεταβλητές έχουν στατιστικά σημαντική επίδραση στην εξαρτημένη μεταβλητή θα κάνουμε έλεγχο υπόθεσης. Η μηδενική και η εναλλακτική υπόθεση διαμορφώνονται ως εξής:

*$H_0$: Η επίδραση της εν λόγω μεταβλητής ΔΕΝ είναι στατιστικά σημαντική.*

*$H_1$: Η επίδραση της εν λόγω μεταβλητής είναι στατιστικά σημαντική.*

Για να αποδεχτούμε ή να απορρίψουμε τη μηδενική υπόθεση θα χρησιμοποιήσουμε την τιμή – P. Συγκρίνουμε τη συγκεκριμένη τιμή με το επίπεδο σημαντικότητας 5% (0,05). Αναλυτικότερα,

Αν τιμή – P < 0,05 απορρίπτουμε την $H_0$ δηλαδή η επίδραση της υπό εξέτασης μεταβλητής στην εξαρτημένη μεταβλητή είναι στατιστικά σημαντική.

Αν τιμή – P > 0,05 αποδεχόμαστε την $H_0$ δηλαδή η επίδραση της υπό εξέτασης μεταβλητής στην εξαρτημένη μεταβλητή ΔΕΝ είναι στατιστικά σημαντική.

Σύμφωνα με τα παραπάνω και με τον πίνακα που έχει προκύψει από την επεξεργασία στο excel μπορούμε να σημειώσουμε τα ακόλουθα.

Οι μεταβλητές που επηρεάζουν στατιστικά σημαντικά την εξαρτημένη μεταβλητή είναι οι εξής:

- TL BASED ISE
- USD BASED ISE
- SP
- FTSE
- NIKKEI
- BOVESPA.

Από την πίνακα που ακολουθεί και συγκεκριμένα χρησιμοποιώντας το συντελεστή προσδιορισμού $R^2$ μπορούμε να αποφανθούμε ότι η συνολική μεταβλητότητα της εξαρτημένης μεταβλητής που οφείλεται στις συγκεκριμένες ανεξάρτητες είναι περίπου 79%, (0.7921).

| *Στατιστικά παλινδρόμησης* | |
|---|---|
| Πολλαπλό R | 0,890016 |
| R Τετράγωνο | 0,792129 |

| Προσαρμοσμένο R | |
|---|---|
| Τετράγωνο | 0,788973 |
| Τυπικό σφάλμα | 0,004824 |
| Μέγεθος δείγματος | 536 |

Τέλος, από τον πίνακα ANOVA, ανάλυσης διακύμανσης μπορούμε να κάνουμε έλεγχο στατιστικής σημαντικότητας για όλο το μοντέλο. Δηλαδή μέσα από τον πίνακα ANOVA μπορούμε να κάνουμε έλεγχο F, όπου η μηδενική και η εναλλακτική υπόθεση διαμορφώνονται ως εξής:

*$H_0$: Η ΑΠΟ ΚΟΙΝΟΥ επίδραση των συγκεκριμένων μεταβλητών στο υπόδειγμα ΔΕΝ είναι στατιστικά σημαντική.*

*$H_1$: Η ΑΠΟ ΚΟΙΝΟΥ επίδραση των συγκεκριμένων μεταβλητών στο υπόδειγμα είναι στατιστικά σημαντική.*

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ

| | *βαθμοί ελευθερίας* | *SS* | *MS* | *F* | *Σημαντικότητα F* |
|---|---|---|---|---|---|
| Παλινδρόμηση | 8 | 0,046733 | 0,005842 | 251,0278 | 2,7E-174 |
| Υπόλοιπο | 527 | 0,012264 | 2,33E-05 | | |
| Σύνολο | 535 | 0,058997 | | | |

Και σε αυτή την περίπτωση η αποδοχή ή απόρριψη της μηδενικής υπόθεσης γίνεται με κριτήριο την τιμή «Σημαντικότητα F».

Αν Σημαντικότητα F < 0,05 απορρίπτουμε την μηδενική υπόθεση ($H_0$), επομένως η από κοινού επίδραση των μεταβλητών είναι στατιστικά σημαντική.

Αν Σημαντικότητα F > 0,05 αποδεχόμαστε την μηδενική υπόθεση ($H_0$), επομένως η από κοινού επίδραση των μεταβλητών ΔΕΝ είναι στατιστικά σημαντική.