Τμήμα Μηχανικών Πληροφορικής Τ.Ε

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

του φοιτητή Αντουάν Ντόρη
Α.Μ: 11765

Θέμα

Επεξεργασία οικονομικών δεδομένων με χρήση αλγορίθμων μηχανικής μάθησης

Επιβλέπων καθηγητής: Νικόλαος Γιαννακέας

Ιούλιος 2016

Μηχανικών Πληροφορικής Τ.Ε

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

του φοιτητή Αντουάν Ντόρη
Α.Μ: 11765

Θέμα

Επεξεργασία οικονομικών δεδομένων με χρήση αλγορίθμων μηχανικής μάθησης

Επιβλέπων καθηγητής: Νικόλαος Γιαννακέας

Ιούλιος 2016

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία επιχειρήθηκε η πρόβλεψη της συμπεριφοράς οικονομικών δεικτών και πιο συγκεκριμένα ισοτιμιών νομισμάτων. Για τον σκοπό αυτό περισσότερες από 70 ισοτιμίες συσχετίστηκαν μεταξύ τους ανά δύο, για χρονικό διάστημα 2 ετών, με σκοπό να αποκαλυφθούν ισοτιμίες, οι οποίες αντιδρούν παρόμοια σε οικονομικά γεγονότα ή οι οποίες επηρεάζουν η μια την άλλη. Αφού εντοπίστηκε ένα ζεύγος ισοτιμιών που πληροί ενδείξεις για κάτι τέτοιο, διενεργήθηκαν πειράματα ταξινόμησης με σκοπό την ποσοτικοποίηση της ακρίβειας μιας τέτοιας πρόβλεψης. Ο Μπευζιανός Ταξινομητής, το δένδρο Απόφασης C4.5, και οι Μηχανές Διανυσμάτων Υποστήριξης εφαρμόστηκαν σε ένα πρόβλημα 3 κατηγοριών. Τα αποτελέσματα τα οποία εξήχθησαν δεν ξεπέρασαν το 50% σε ακρίβεια, γεγονός που φανερώνει ότι η συμπεριφορά των δεικτών αυτών δεν επετεύχθη να μοντελοποιηθεί, τουλάχιστον με τα συγκεκριμένα χρονικά περιθώρια και τα συγκεκριμένα χαρακτηριστικά που χρησιμοποιήθηκαν.

ABSTRACT

This study attempts the prediction of behavior for financial indexes such as forex indexes. For this reason, over 70 forex indexes have been correlated, to indicate if there are cases that a forex index affect other ones. A correlation matrix is computed for all the pairs of forex indexes. Ideed, we have found that two indexes present high correlation. A classification problem is scheduled for the highest correlated pair of indexes, using three classes. Bayes Classifier, C4.5 Decision Tree, and Support Vector Machines have been employed in order to quantify the prediction accuracy for such a problem. The obtained results presents accuracy lower the 50%, which is not satisfactory value. Perhaps, either the employed features which have been used for classification, or the nature of forex indexes itself, could not be extracted by a prediction model.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

Εισαγωγή

Τις τελευταίες δεκαετίες υπήρχε μεγάλος όγκος δεδομένων που συσσωρευόταν και δεν υπήρχε η σωστή τεχνολογική υποδομή για να τον επεξεργαστούμε ανάλογα. Τα χρόνια όμως πέρασαν και βρήκαμε λύση ακόμα και σε αυτό το πρόβλημα χάρει στην εξόρυξη δεδομένων. Πλέον μπορούμε να εξάγουμε συγκεκριμένη πληροφορία από έναν όγκο δεδομένων με τη χρήση μερικών γραμμών κώδικα ή με το πάτημα ενός κουμπιού μέσω ενός προγράμματος.

Η εξόρυξη δεδομένων μας βοήθησε ουσιαστικά στην εργασία αυτή. Ο σκοπός αυτής της εργασίας είναι μέσα από τις αναλύσεις των τιμών νομισμάτων να καταλάβουμε κατά πόσο συσχετίζεται ένα νόμισμα με κάποιο άλλο. Επίσης αξιοσημείωτο είναι πως όταν κάποιο νόμισμα αλλάξει τιμή, θα αλλάξει και η τιμή κάποιου νομίσματος άλλης χώρας/οικονομίας που συνάπτουν εμπορικές δραστηριότητες.

Η εργασία ξεκινά με ιστορική αναδρομή των οικονομικών δεικτών, τον ορισμό για κάθε δείκτη καθώς κι από ποιους παράγοντες επηρεάζεται, τις μεθόδους και τα εργαλεία επεξεργασίας αυτών. Στην συνέχεια παρουσιάζεται η έννοια της Εξόρυξης Δεδομένων, ποια ήταν η ανάγκη που την γέννησε και μαζί με αυτήν οι μέθοδοι εξόρυξης δεδομένων. Την σειρά έχει το Κεφάλαιο Τρία, όπου γίνεται, εκτενής παρουσίαση και ανάλυση των μεθόδων εξόρυξης δεδομένων που χρησιμοποιήθηκαν στην εργασία αυτή. Κάθε μέθοδος όμως αποτελείται από τους δικούς της αλγόριθμους, οι οποίοι επεξηγούνται κι αυτοί. Έπειτα ακολουθούν τα αποτελέσματα αναλυτικά με βάση τις τιμές αλλά και ποιοι αλγόριθμοι χρησιμοποιήθηκαν. Τέλος, έχουμε μια περίληψη των ευρημάτων καθώς τα συμπεράσματα μας για αυτά.

Κεφάλαιο 1

Οικονομικοί Δείκτες και δεδομένα

Κεφάλαιο 1

1.1 Ιστορική αναδρομή Οικονομικών δεικτών

Το πρώτο οικονομικό κραχ της σύγχρονης ιστορίας λαμβάνει μέρος στις Η.Π.Α., γνωστό ως το πρώτο χρηματιστηριακό κραχ κι έπληξε αλυσιδωτά την παγκόσμια οικονομία. Κάποιες χώρες ανέκαμψαν σε διάστημα 2-3 ετών ενώ άλλες έφτασαν στο επίπεδο που ήταν ύστερα από 10-12 χρόνια, αν κι εφόσον δεν είχαν πάρει μέρος στον Α' Παγκόσμιο Πόλεμο. Λόγω της παγκόσμιας οικονομικής ύφεσης του προηγούμενου έτους (1929), το 1930 καταργείται ο χρυσός κανόνας που ίσχυε από το 1880.

Το 1944, μεταξύ των Μεγάλων Δυνάμεων της υφηλίου κι άλλων δυνατών οικονομιών (σύνολο γύρω στις 44 χώρες), αποφασίζεται στο Μπρέτον Γουντς των Η.Π.Α η δημιουργία του Διεθνούς Νομισματικού Ταμείου (ΔΝΤ), της Παγκόσμιας Τράπεζας, της GATT όπως επίσης η υιοθέτηση του συστήματος σταθερών συναλλαγματικών ισοτιμιών , γνωστό ως σύστημα Μπρέτον Γουντς. Το 1971, ο τότε πρόεδρος των Η.Π.Α Ρίτσαρντ Νίξον αποφασίζει η χώρα του να εγκαταλείψει τη μετατρεψιμότητα του δολαρίου σε χρυσό.

Εκείνη την περίοδο αποφασίζεται να πάψει να ισχύει το σύστημα Μπρέτον Γουντς λόγω προβλημάτων στο εμπόριο, τον πληθωρισμό, στον πόλεμο του Βιετνάμ και των εξόδων για την εξερεύνηση του διαστήματος.

Ξεκινάει η νεότερη εποχή των κυμαινόμενων συναλλαγματικών ισοτιμιών, όπου στις σύγχρονες οικονομίες επικρατούν τα παρακάτω οικονομικά καθεστώτα:

1. Σταθερών ισοτιμιών
2. Ελεύθερης διακύμανσης
3. Ελεγχόμενης διακύμανσης

Δείκτες οικονομίας των Η.Π.Α

Παρακάτω ακολουθεί μια αναφορά στους οικονομικούς δείκτες των Η.Π.Α χωρίς να σημαίνει πως δεν υπάρχουν άλλοι σημαντικοί δείκτες σε οικονομίες όπως της Γερμανίας, Ιαπωνίας, Μεγάλης Βρετανίας κλπ.

Οι μακροοικονομικοί αυτοί δείκτες παρακολουθούνται από επενδυτές σε όλον τον κόσμο. Η σημασία τους είναι υψηλή καθώς μέσα από τα δεδομένα τους μπορούμε εξάγουμε συμπεράσματα και να κάνουμε προβλέψεις για διάφορους κλάδους της οικονομίας σε επίπεδο χώρας αλλά και παγκόσμια.

Ο κάθε δείκτης έχει και τον αντίστοιχο βαθμό σημασίας και διαχωρίζονται σε:

- Υψηλός
- Μεσαίος
- Χαμηλός

- **CCI - Δείκτης Εμπιστοσύνης Καταναλωτών [Υψηλός]**
  Κέντρο Οικονομικών Ερευνών "Conference Board"
  Το CCI βασίζεται σε μια έρευνα σε δείγμα 5.000 νοικοκυριών των ΗΠΑ και θεωρείται ένας από τους ακριβέστερους δείκτες εμπιστοσύνης.
- **CPI - Δείκτης Τιμών Καταναλωτή / Δομικός δείκτης CPI [Υψηλός]**
  Bureau of Labor and Statistics (Γραφείο Εργασίας και Στατιστικής του Υπουργείου Εργασίας)
- **Employment Report (Έκθεση Εργασίας) [Υψηλός]**
  Department of Labor (Υπουργείο Εργασίας)
- **Employment Report (Έκθεση Προόδου Εργασίας) [Υψηλός]**
  Bureau of Labor and Statistics (Γραφείο Εργασίας και Στατιστικής του Υπουργείου Εργασίας)
- **Σύσκεψη FOMC (Νομισματικής Επιτροπής της Κεντρικής Τράπεζας): Ανακοίνωση επιτοκίου [Υψηλός]**
  Η σύσκεψη των αντιπροσώπων της Κεντρικής Τράπεζας των ΗΠΑ (Fed), λαμβάνει χώρα 8 φορές ετησίως. Η απόφαση σχετικά με το βασικό επιτόκιο ανακοινώνεται κατά τη διάρκεια της κάθε σύσκεψης (περίπου στις 14:15 EST).

- **GDP - Ακαθάριστό Εγχώριο Προϊόν [Υψηλός]**
  BEA (Υπηρεσία Οικονομικής Ανάλυσης)
  Το Υπουργείο Εμπορίου των ΗΠΑ δημοσιεύει το ΑΕΠ σε τρεις μετρήσεις:
  α' μέτρηση (advance),
  β' μέτρηση (preliminary) και
  γ' μέτρηση (final).
- **Μεταποιητικός Δείκτης ISM (Ινστιτούτου Διευθυντών Προμηθειών) [Υψηλός]**
  Ινστιτούτο Διευθυντών Προμηθειών
- **MCSI - Δείκτης Εμπιστοσύνης Καταναλωτών Πανεπιστημίου** Μίσιγκαν [Υψηλός]
  Πανεπιστήμιο του Μίσιγκαν - Πρώτη κάθε μήνα: κάλυψη στοιχείων προηγούμενου μήνα
- **NFP - Νέες θέσεις εργασίας εκτός του γεωργικού τομέα [Υψηλός]**
  Department of Labor (Υπουργείο Εργασίας)
  Τα δεδομένα αντιπροσωπεύουν τις αλλαγές στο σύνολο των μισθωτών των ΗΠΑ σε όλους τους κλάδους, με τις εξής εξαιρέσεις: α) γενικοί κυβερνητικοί υπάλληλοι, β) ιδιωτικοί οικιακοί υπάλληλοι, γ) υπάλληλοι των μη κερδοσκοπικών οργανώσεων που παρέχουν ατομική βοήθεια, δ)υπάλληλοι στον γεωργικό τομέα.
- **PMI - Δείκτης Υπευθύνων Προμηθευτών [Υψηλός]**
  Ινστιτούτο Διαχείρισης Αποθεμάτων
- **Στοιχεία Λιανικών Πωλήσεων - Λιανικές Πωλήσεις πλην Αυτοκινούμενων Οχημάτων [Υψηλός]**
  Υπηρεσία Απογραφής - Περίπου στις 12 κάθε μήνα, 8:30 π.μ. EST: κάλυψη στοιχείων προηγούμενου μήνα
- **Έρευνα Tankan [Υψηλός]**
  BoJ (Κεντρική Τράπεζα της Ιαπωνίας)
- **TIC (Treasury International Capital) Στοιχεία σχετικά με τις συναλλαγές μακροπρόθεσμων ομολόγων [Υψηλός]**
  Υπουργείο Οικονομικών
- **Εμπορικό Ισοζύγιο [Υψηλός]**
  Υπουργείο Εμπορίου

- **Beige Book [Χαμηλός]**
  Διοικητικό Συμβούλιο της Κεντρικής Τράπεζας (Fed)
- **ECI - Δείκτης Κόστους Εργασίας [Χαμηλός]**
  Bureau of Labor and Statistics (Γραφείο Εργασίας και Στατιστικής του Υπουργείου Εργασίας)
- **PCE - Δείκτης Προσωπικών Καταναλωτικών Δαπανών [Χαμηλός]**
  BEA (Υπηρεσία Οικονομικής Ανάλυσης)
- **Μηνιαία Ανακοίνωση Προϋπολογισμού [Μεσαίος]**
  Μια μηνιαία έκθεση που δημοσιεύει η κυβέρνηση των ΗΠΑ (Υπουργείο Οικονομικών) και που δείχνει το δημοσιονομικό έλλειμμα ή πλεόνασμα για κάθε μήνα.
- **Σύνθετοι Δείκτες Οικονομικής Δραστηριότητας (Leading Indicators) [Μεσαίος]**
  Κέντρο Οικονομικών Ερευνών "Conference Board"
  Ο δείκτης αυτός χρησιμοποιείται για την πρόβλεψη της κατεύθυνσης των κινήσεων της οικονομίας στους επόμενους μήνες. Αποτελείται από 10 επιμέρους οικονομικούς δείκτες, οι αλλαγές των οποίων συνήθως προηγούνται των αλλαγών στο σύνολο της οικονομίας. Οι 10 επιμέρους δείκτες είναι:

    - Μέσο επίπεδο ωρών απασχόλησης στον βιομηχανικό κλάδο,
    - Μέσος όρος των αρχικών αιτήσεων για επιδόματα ανεργίας,
    - Το ποσό των νέων βιομηχανικών παραγγελιών για καταναλωτικά αγαθά και υλικά,
    - Ταχύτητα παράδοσης νέων εμπορευμάτων από τους προμηθευτές,
    - Το ποσό των νέων παραγγελιών για κεφαλαιακά αγαθά μη αμυντικού χαρακτήρα,
    - Ποσότητα αδειών κατασκευής για νέες κατοικίες,
    - Δείκτης μετοχών 500 επιχειρήσεων της Standard & Poor's (S&P 500),
    - Προσφορά χρήματος κατόπιν προσαρμογής πληθωρισμού (M2),
    - Διαφορά μεταξύ μακροπρόθεσμων και βραχυπρόθεσμων επιτοκίων,
    - Καταναλωτική διάθεση.

- **Τρεχούμενος Λογαριασμός [Μεσαίος]**
  BEA (Υπηρεσία Οικονομικής Ανάλυσης)

- **Παραγγελίες Διαρκών Αγαθών [Μεσαίος]**
  Υπηρεσία Απογραφής
- **Αποπληθωριστής του ΑΕΠ [Μεσαίος]**
  BEA (Υπηρεσία Οικονομικής Ανάλυσης)
- **Ανέγερση Νέων Κατοικιών [Μεσαίος]**
  Υπηρεσία Απογραφής
- **Βιομηχανική Παραγωγή, Αξιοποίηση της Παραγωγής [Μεσαίος]**
  Κεντρική Τράπεζα των ΗΠΑ (Fed)
- **Αρχικές Αιτήσεις Επιδομάτων Ανεργίας [Μεσαίος]**
  Department of Labor (Υπουργείο Εργασίας)
  Μεταποιητικός δείκτης περιοχής Fed Φιλαδέλφειας (Έρευνα Επισκόπησης Επιχειρήσεων) [Μεσαίος]
  Κεντρική Τράπεζα των ΗΠΑ (Fed) της Φιλαδέλφειας
- **PPI - Δείκτης Τιμών Παραγωγού / Δομικός δείκτης PPI [Μεσαίος]**
  Bureau of Labor and Statistics (Γραφείο Εργασίας και Στατιστικής του Υπουργείου Εργασίας)

1.2 Οικονομικοί δείκτες

Τα CFD ή «Συμβόλαια επί της Διαφοράς» αποτελούν χρηματοοικονομικά προϊόντα, μια σύμβαση μεταξύ δύο μερών, αγοραστή και πωλητή, η οποία ορίζει ότι ο πωλητής θα καταβάλει στον αγοραστή τη διαφορά μεταξύ της τρέχουσας αξίας του περιουσιακού στοιχείου και της αξίας του κατά τη διάρκεια της σύμβασης. Σε περίπτωση όμως που η διαφορά είναι αρνητική, τότε ο αγοραστής πληρώνει τη διαφορά αντί του πωλητή. Στην πραγματικότητα τα CFD είναι «παράγωγα» που προσφέρουν στους επενδυτές τη δυνατότητα να εκμεταλλεύονται τις διακυμάνσεις της τιμής μετοχών, δεικτών κ.λ.π., μέσω των οποίων κερδοσκοπούν.

Τα CFD δεν επιφέρουν **φυσική παράδοση** του προϊόντος που υπόκειται στη σύμβαση. Για παράδειγμα, η αγορά CFDs χρυσού, σε καμία περίπτωση δεν συνεπάγεται με παράδοση μπάρων χρυσού στη διεύθυνση κατοικίας του επενδυτή.

Όταν ανοίγει μια θέση CFD, το spread (εννοώντας τη διαφορά μεταξύ της τιμής αγοράς (bid) και πώλησης (ask) το πληρώνει ο επενδυτής, το οποίο ορίζεται σαν το μικτό κέρδος της κάθε χρηματιστηριακής εταιρίας που παρέχει την υπηρεσία αυτή.

Όταν κλείσει μια θέση CFD, γίνεται μια εκκαθάριση της παρακάτω μορφής

Εκκαθάριση = (τιμή ανοίγματος – τιμή κλεισίματος) x αριθ.συμβολαίων επενδυτή

Το CFD χρησιμοποιεί μόχλευση (leverage), δηλαδή η έκθεση (exposure) που προσφέρει στην αγορά ισοδυναμεί με κεφάλαια **πολύ μεγαλύτερης αξίας** από αυτά που δεσμεύτηκαν για το άνοιγμα της θέσης. Το περιθώριο ασφάλισης (margin) είναι το ποσό (ενέχυρο) που παρακρατείται στον λογαριασμό ως εγγύηση όταν ο επενδυτής ανοίγει και διακρατά μια θέση CFD.

Το περιθώριο ασφάλισης που απαιτείται για το άνοιγμα μίας θέσης διαφέρει με βάση τα χαρακτηριστικά του υποκείμενου χρηματοοικονομικού προϊόντος (ρευστότητα, εύρος διακύμανσης), τις γενικότερες συνθήκες στην αγορά (κόστος χρήματος, τάση, μεταβλητότητα) και την πολιτική του εκάστοτε παρόχου.

Η χρήση μόχλευσης, που ισοδυναμεί **με χρήση επιπρόσθετων κεφαλαίων** από αυτά που βρίσκονται στο λογαριασμό του επενδυτή, συνεπάγεται ορισμένο κόστος χρηματοδότησης (rollover charge) για κάθε ημέρα που η μοχλευμένη θέση παραμένει ανοιχτή.

Για παράδειγμα, ένας πάροχος μπορεί να **απαιτεί περιθώριο ασφάλισης 1%** ια συμβόλαια CFD στον DAX. Αυτό σημαίνει ότι ο επενδυτής οφείλει να έχει στο λογαριασμό του και να δεσμεύει 200 Ευρώ παρά το γεγονός ότι η ονομαστική αξία ενός συμβολαίου CFD στον DAX ισούται με €20.000 (τιμή 16 Μαρτίου). Με άλλα λόγια, ο επενδυτής «ελέγχει» θέση αξίας 20.000 Ευρώ με μόλις 200 Ευρώ. Η μόχλευση στις συναλλαγές CFD απαιτεί μεγάλη προσοχή, καθώς λειτουργεί πολλαπλασιαστικά τόσο στα κέρδη όσο και στις ζημιές.*

Πλεονεκτήματα των CFD

Τα CFD έγιναν δημοφιλή λόγω πλεονεκτημάτων, όπως:

- Μικρό ποσό για αρχική επένδυση

- Οι επενδυτικές επιλογές ποικίλουν, επιλέγοντας μεταξύ άλλων, πρώτες ύλες, ομόλογα, χρηματιστηριακές μετοχές κτλπ.
- Με το άνοιγμα θέσεων short επιτυγχάνουμε κέρδη σε ανοδικές και καθοδικές αγορές
- Η δυνατότητα επίτευξης υπέρογκων κερδών ή αντίστοιχων ζημείων επιτυγχάνεται με τη χρήση υψηλής μόχλευσης
- Το κόστος των συναλλαγών (bid-ask spread) είναι χαμηλό λόγω του έντονου ανταγωνισμού μεταξύ των διαφορών παρόχων, κάτι που κάνει τα CFD ιδανικά για day trading και άλλες τακτικές βραχυπρόθεσμης διακράτησης
- Όλες οι κινήσεις των συναλλαγών μας γίνονται πλέον μέσω της εφαρμογής.πλατφόρμας που μας δίνει ο πάροχος
- **Η δυνατότητα αντιστάθμισης του κινδύνου αγοράς** (market risk) που υπάρχει σε επενδυτικό χαρτοφυλάκιο. Για παράδειγμα, επενδυτής που διατηρεί θέσεις σε ελληνικές μετοχές, μπορεί να «σορτάρει» το CFD του δείκτη του Χρηματιστηρίου Αθηνών πετυχαίνοντας έτσι μερική εξασφάλιση σε περίπτωση υποχώρησης της αγοράς.
- Το κόστος συμμετοχής στην αγορά ελαχιστοποιείται λόγω των CFD μετοχών καθώς δεν υπάρχει επιπρόσθετη επιβάρυνση μεταβιβαστικών τελών, εξόδων χρηματιστηρίου και φόρων επί χρηματιστηριακής συναλλαγής.
- Μπορείς να είσαι ενεργός στην αγορά με την απόκτηση CFD δεικτών/ισοτιμιών χωρίς να απαιτείται αγορά επιμέρους μετοχών.
- Τα CFD εμπορευμάτων προσφέρουν τη δυνατότητα ανοίγματος θέσεων που έχουν **πολύ μικρότερη ονομαστική αξία** σε σχέση με τα αντίστοιχα Συμβόλαια Μελλοντικής Εκπλήρωσης (Futures), τα οποία συνήθως αφορούν μια μεγάλη ποσότητα του υποκείμενου προϊόντος και απαιτούν μεγάλα περιθώρια ασφάλισης. Για παράδειγμα, ένα Σ.Μ.Ε. αργού πετρελαίου ποικιλίας WTI αφορά 1.000 βαρέλια, ενώ το CFD του ίδιου προϊόντος επιτρέπει συναλλαγές σε μικρότερες ποσότητες και συνεπώς μεγαλύτερη ευελιξία και δυνατότητα πρόσβασης σε επενδυτές οι οποίοι διαθέτουν λογαριασμούς περιορισμένου μεγέθους.

Χώρες που επιτρέπονται τα CFDs

Έλβετία	Σουηδία

Αυστρία	Ηνωμένο Βασίλειο
Όλλανδία	Ισπανία
Πολωνία	Ιρλανδία
Γαλλία	Γερμανία
Ιταλία	Σιγκαπούρη
Νορβηγία	Ιαπωνία
Λουξεμβούργο	Τουρκία
Ισραηλ	Καναδάς
Νέα Ζηλανδία	Νότια Αφρική
Αυστραλία	

Παρόλο που είναι δημοφιλή, είναι σημαντικό να αναφέρουμε πώς δεν επιτρέπονται σε ορισμένες χώρες όπως: Η.Π.Α και το Χονγκ Κονγκ

<u>Παράδειγμα Συναλλαγής σε CFD:</u> Έστω ότι έχουμε έναν επενδυτή ο οποίος μέσω της πλατφόρμας της εταιρίας της οποίας συνεργάζεται αγοράζει 5 συμβόλαια CFD επειδή προβλέπει πως ο NASDAQ θα κινηθεί ανοδικά. Αυτομάτως αυτο σημαίνει πως ο επενδυτής για κάθε μονάδου ανόδου θα κερδίζει 5 ευρώ και θα χάνει το ίδιο ποσό για κάθε μονάδα πτώσης.

Πριν γίνουν τα παραπάνω, ο επενδυτής οφείλει να έχει στον λογαριασμό του ένα ποσό ίσο με το περιθώριο ασφάλισης, για παράδειγμα 1%, που χρειάζεται για την διεκπεραίωση της συναλλαγής.

Έστω ότι ο δείκτης NASDAQ διαπραγματεύεται στις 15.000 μονάδες, αυτό σημαίνει πως ο επενδυτής πρέπει να έχει στο λογαριασμό του το ποσό των 150 ευρώ = 1% του 15.000, για το «άνοιγμα» ενός συμβολαίου CFD. Για άνοιγμα 5 συμβολαίων όμως πρέπει να έχει 5 x 150 = 750 ευρώ στο λογαριασμό του.

Την συνολική αξία της θέσης του την βρίσκουμε πολλαπλασιάζοντας τις μονάδες με τον αριθμό των συμβολαίων 15.000 x 5 = 75.000 ευρώ. Αν λοιπόν υποθέσουμε πως ο NASDAQ κλείσει θετικά στις 15.500 μονάδες, δηλαδή 500 μονάδες παραπάνω από την τιμή που άνοιξε, το κέρδος λοιπόν αντιστοιχεί σε 5 x 500 = 2500 ευρώ μείον το ποσό spread που δαπανήθηκε για το άνοιγμα της θέσης μείον το κόστος χρηματοδότησης της θέσης (rollover).

Οι επενδυτές για να μην ζημιωθούν σε μεγάλο βαθμό και να χάσουν μεγάλο μέρος του χαρτοφυλακίου τους τοποθετούν εντολές διακοπής ζημιών (stop-loss). Αυτό συμβαίνει διότι υπάρχει μεγάλη μεταβλητότητα στις αγορές. Με αυτό το εργαλείο λοιπόν (stop-loss) αυξάνονται οι πιθανότητες επιτυχίας στις αγορές.

Οι ενδιαφερόμενοι πελάτες οφείλουν να αξιολογούν τον πάροχο για υπηρεσίες συναλλαγών σε CFD:

- Με βάση τις αρχές που τον εποπτεύουν
- Το ιστορικό και το μεγεθός του
- Την κάλυψη των κεφαλαίων των πελατών από εύρωστο Σύστημα Αποζημίωσης
- Την αξιοπιστία και το κόστος συναλλαγών
- Την ποικιλομορφία των προσφερόμενων προιόντων

Ο όρος Forex προέρχεται από τη συνένωση των όρων Foreign Exchange, δηλαδή, «ξένο συνάλλαγμα». Η Forex δραστηριοποιείται στην αγοραπωλησία νομισμάτων/ισοτιμιών, είναι παγκόσμιας εμβέλειας αγορά και δεν λαμβάνει χώρα σε κάποια συγκεκριμένη τοποθεσία στον πλανήτη. Τέθηκε σε λειτουργία όταν έπαψε να ισχύει το «σύστημα σταθερών ισοτιμιών Bretton Woods». Εξελίχθηκε στην μεγαλύτερη αγορά του κόσμου έχοντας τον μεγαλύτερο όγκο συναλλαγών που με τη σειρά τους επιφέρουν τον μεγαλύτερο ημερήσιο τζίρο. Αυτό οφείλεται στα ιδιαίτερα χαρακτηριστικά που την διαφοροποιούν από τις υπόλοιπες αγορές. Και όπως κάθετι αργεί να έρθει στην Ελλάδα, έτσι και η Forex αποτελεί κάτι σχετικά καινούργιο για τα ελληνικά δεδομένα.

Οι ισοτιμίες καθώς και οι παράγοντες που τις καθορίζουν είναι κάτι που πρέπει να απασχολεί τους περισσότερους από εμάς. Ο λόγος είναι διότι ανάλογα με την τιμή της ισοτιμίας μας με τις ξένες αλλάζουν και οι τιμές των αγαθών. Για παράδειγμα, για όποιο προιόν που εισάγουμε παίζει καθοριστικό ρόλο στην τιμή του, η τιμή ισοτιμίας της χώρας μας με αυτή της χώρας από την οποία το εισάγουμε.
Τα επίπεδα και οι τιμές των ισοτιμιών είναι κάτι που απασχολεί έντονα έμπορους(λιανικής/χονδρικής) που δραστηριοποιούνται σε τρίτες χώρες και εισάγουν προιόντα. Για παράδειγμα, αν κάποια ελληνική εταιρία εισάγει προιόντα από την Αφρική, το κέρδος της και η τιμολόγηση των προιόντων εξαρτώνται από τις ισοτιμίες

των χωρών εισαγωγής. Κάθε διακύμανση στην εγχώρια ισοτιμία (υποτίμηση/ανατίμηση) έχεις τις επιπτώσεις της, θετικές ή αρνητικές.

Τόπος και χρόνος διεξαγωγής των συναλλαγών: Οι συναλλαγές δεν λαμβάνουν μέρος σε κάποιο συγκεκριμένο μέρος της γης ή κάποια συγκεκριμένη ώρα, καθώς διεξάγονται όλο το 24-ώρο – εκτός από τα Σαββατοκύριακα - μέσω Διαδικτύου, διαμέσου ενός δικτύου τραπεζών (κεντρικών εθνικών και μη), εταιρειών και πωλητών νομισμάτων. Οι εργάσιμες ώρες/ ώρες συναλλαγών ανάλογα το γεωγραφικό μήκος και πλάτος του καθενός χωρίζονται ως παρακάτω:

Πίνακας 1: Εργάσιμες ώρες ανοιχτών αγορών

Περίοδος	Πόλη	Άνοιγμα	Κλείσιμο
Ευρωπαική	Λονδίνο	08:00:00	17:00:00
Αμερικάνικη	Νέα Υόρκη	13:00:00	22:00:00
Ασιάτικη	Τόκυο	23:00:00	09:00:00

Βασικά κέντρα αγοράς συναλλάγματος αποτελούν τα παρακάτω:

Ευρώπη	Σιγκαπούρη
Νέα Υόρκη	Χονγκ Κονγκ
Λος Άντζελες	Μπαχρέιν
Τόκυο	Σίδνευ

Η αγορά συναλλάγματος είναι μοναδική λόγω:

- Του τεράστιου όγκου συναλλαγών που αντιπροσωπεύουν τη μεγαλύτερη κατηγορία περιουσιακών στοιχείων στον κόσμο, πράγμα που οδηγεί σε υψηλή ρευστότητα
- της εξάπλωσης σε όλα τα μήκη και πλάτη του πλανήτη
- της ολοήμερης λειτουργίας του, 24 ώρες/μέρα εκτός από τα Σαββατοκύριακα
- των χαμηλών περιθωριών κέρδους σε σύγκριση με άλλες αγορές σταθερού εισοδήματος
- της ποικιλίας των παραγόντων που την επηρεάζουν
- της χρήσης μόχλευσης για την ενίσχυση των περιθωρίων κέρδους και ζημίας σε σχέση με το μέγεθος του λογαριασμού.

Αυτοί που συμμετέχουν σε μια αγορά Forex είναι:

- Ιδιώτες επενδυτές
- Εμπορικές επιχειρήσεις: οι οικονομικές δραστηριότητες των οποίων απαρτίζουν σημαντικό μερίδιο της αγοράς συναλλάγματος.
- Εμπορικές τράπεζες
- Κεντρικές τράπεζες: ο ρόλος των οποίων είναι αρκετά σημαντικός διότι προσπαθούν να ελέγξουν την προσφορά χρήματος, τον πληθωρισμό, την τιμή των επιτοκίων, έχοντας τις περισσότερες φορές προαποφασιμένες επίσημες (και μη) τιμές επιτοκίων για το νόμισμά τους.
- Χρηματοπιστωτικά ιδρύματα
- Κερδοσκοπικά κεφάλαια
- συναλλαγές στις μέρες μας – λόγω τεχνολογικής ανάπτυξης, Παγκόσμιος Ιστός (internet) – έχουν διευκολυνθεί σε μεγάλο βαθμό, δίνοντας έτσι πρόσβαση σε χρηματιστές και ιδιώτες επενδυτές σε διάφορα εργαλεία που αποφέρουν ακόμα πιο επικερδείς αγοραπωλησίες.

Πίνακας 2: 10 μεγαλύτεροι έμποροι συναλλάγματος

A/A	Χώρα	Έμπορος	Ποσοστό
1	Η.Π.Α	Citi	16,11%
2	Γερμανία	Deutsche Bank	14,54%
3	Ηνωμένο Βασίλειο	Barclays Investment Bank	8,11%
4	Η.Π.Α	JPMorgan	7,65%
5	Ελβετία	UBS AG	7,30%
6	Η.Π.Α	Bank of America Merrill Lynch	6,22%
7	Ηνωμένο Βασίλειο	HSBC	5,40%
8	Γαλλία	BNP Paribas	3,65%
9	Η.Π.Α	Goldman Sachs	3,40%
10	Ηνωμένο Βασίλειο	Royal Bank of Scotland	3,38%

Πίνακας 3: Περισσότερο διαπραγματεύσιμα νομίσματα με βάση την αξία τους

A/A	Νόμισμα	ISO 4217 κωδικός(σύμβολο)	Ποσοστό καθημερινής μετοχής (Απρίλιος 2013)
1	Δολλάριο Η.Π.Α	USD($)	87,0%
2	Ευρώ	Euro()	33,4%
3	Γιεν Ιαπωνίας	JPY()	23,0%
4	Λίρα Ην.Βασιλείου	GBP()	11,8%
5	Δολλάριο Αυστραλίας	AUD($)	8,6%
6	Φράγκο Ελβετίας	CHF()	5,2%
7	Δολλάριο Καναδά	CAD($)	4,6%
8	Πέσο Μεξικού	MXN()	2,5%
9	Γουάν Κίνας	CNY()	2,2%
10	Δολλάριο Νέας Ζηλανδίας	NZD($)	2,0%
11	Κορώνα Σουηδίας	SEK(KR)	1,8%
12	Ρούβλι Ρωσίας	RUB()	1,6%
13	Δολλάριο Χονγκ Κονγκ	HKD($)	1,4%
14	Κορώνα Νορβηγίας	NOK(KR)	1,4%
15	Δολλάριο Σιγκαπούρης	SGD($)	1,4%
16	Λίρα Τουρκίας	TRY()	1,3%
17	Γουν Νότιας Κορέας	KRW()	1,2%
18	Ραντ Νοτίου Αφρικής	ZAR(R)	1,1%
19	Ρεαλ Βραζιλίας	BRL(R $)	1,1%

20	Ρουπία Ινδίας	INR()	1,0%
		Άλλα	6,3%
		Σύνολο	200%

Εμπορεύματα (Commodities)

Η λέξη εμπόρευμα τέθηκε σε χρήση στην αγγλική γλώσσα τον 15ο αιώνα, από τη γαλλική commodité. Πηγαίνοντας πιο πίσω, η γαλλική λέξη προέρχεται από τη λατινική commoditas, που σημαίνει «καταλληλότητα, την ευκολία, πλεονέκτημα". Η λατινική λέξη Κόμμοδο, σήμαινε ποικιλοτρόπως "κατάλληλο", "ορθό μέτρο, χρόνο, ή την κατάσταση", και το "πλεονέκτημα".

Πρόκειται για αγαθά, μέταλλα, ή άλλη τυποποιημένη φυσική ουσία που είναι διαπραγματεύσιμη στο εμπόριο και την οποία οι επενδυτές αγοράζουν και πωλούν συνήθως μέσω συμβολαίων μελλοντικής εκπλήρωσης (futures). Άλλως, είναι οι φυσικοί πόροι, τα χημικά, και τα φυσικά προϊόντα που μπορεί ο καθένας να αγγίξει, γευθεί, εξορύξει, καταναλώσει ή να παραδώσει. Η διαπραγμάτευση και η αγοραπωλησία αυτών διαδραματίζουν ένα σημαντικό ρόλο στις διεθνείς αγορές. Τα πιο γνωστά συμβόλαια των commodities μπορούν να διακριθούν σε ευρύτερες κατηγορίες όπως ενέργεια (πετρέλαιο, φυσικό αέριο), μέταλλα (χρυσός, άργυρος, χαλκός, μόλυβδος, ψευδάργυρος, νίκελ κλπ), δημητριακά, τρόφιμα και κλωστοϋφαντουργικά. Ένα συμβόλαιο μελλοντικής εκπλήρωσης εμπορευμάτων (commodity future) είναι μια συμφωνία αγοράς ή πώλησης ενός ορισμένου τύπου εμπορεύματος για παράδοση σε συγκεκριμένη ημερομηνία στο μέλλον, σε συγκεκριμένο τόπο και σε προκαθορισμένη τιμή. Ο τόπος διαπραγμάτευσης είναι τα χρηματιστήρια εμπορευμάτων.

Το παλιότερο χρηματιστήριο εμπορευμάτων ήταν το Chicago Board of Trade που ιδρύθηκε το 1848 και ακολουθεί το Chicago Butter and Egg Board (1898) που εξελίχθηκε αργότερα με την ονομασία Chicago Mercantile Exchange σαν το μεγαλύτερο χρηματιστήριο παραγώγων και αγροτικών εμπορευμάτων του κόσμου. Τα δύο ανωτέρω χρηματιστήρια του Σικάγο συγχωνεύτηκαν το 2007 με την ονομασία CME Group (βλέπε όρο) που αποτελεί το μεγαλύτερο χρηματιστήριο παραγώγων και εμπορευμάτων στον κόσμο.

Χρηματιστήριο

Με τον όρο **Χρηματιστήριο** εννοούμε την οργανωμένη αγορά, η οποία συνήθως φέρει την επίσημη ανάγνωριση του εκάστοτε κράτους στο οποίο λειτουργούν. Αντικείμενο αγοραπωλησιών των ενδιαφερόμενων αποτελούν οι κινητές αξίες. Παραδείγματα κινητών αξιών είναι τα ακόλουθα: μερίδια κεφαλαίου ανωνύμων εταιρειών (μετοχές), τραπεζικά, κρατικά ή άλλα ομόλογα) ή/και εμπορευμάτων.

Η πλειονότητα των χρηματιστηρίων, με νομοθετικά και διοικητικά μέτρα καθορίζουν το πλαίσιο μέσα στο οποίο διαμορφώνονται οι αγορές και θεσπίζουν τις προϋποθέσεις και τους όρους λειτουργίας τους. Είναι ιδιόμορφες αγορές με την έννοια της ταυτόχρονης συνάντησης της προσφοράς και της ζήτησης. Οι παράγοντες που επηρεάζουν την προσφορά και τη ζήτηση στις χρηματιστηριακές αγορές επιφέρουν με τη σειρά τους διακυμάνσεις στις τιμές των μετοχών. Τα χρηματιστήρια διακρίνονται σε Αξιών, Εμπορευμάτων και Ναύλων.

Η τάση για κερδοσκοπία καθώς και, η ανάγκη πραγματοποίησης αγοραπωλησιών μεγάλων ποσοτήτων εμπορευμάτων που βρίσκονταν μακρυά από τον τόπο διαπραγμάτευσης τους ενώ απαιτούνταν σοβαρά βραχυπρόθεσμα αλλά κυρίως μακροπρόθεσμα κεφάλαια, αποτέλεσαν τους λόγους για τη δημιουργία του χρηματιστηρίου.

Η οργανωμένη μορφή τους δικαιολογείται από:

- Τη ταχύτητα διενέργειας των συναλλαγών
- Την αμεσότητα τους
- Τη δημοσιότητα των συναλλαγών όπου φαίνονται δημόσια όλα τα χαρακτηριστικά των συναλλαγών (προσφορά, ζήτηση, ποσότητα και αξία)
- Την καθαρότητα των συναλλαγών.

Σημασία και Ρόλος Χρηματιστηρίου

- Διευκολύνει τις συναλλαγές, γιατί επιτρέπουν στους εκπροσώπους να βρίσκονται ταυτόχρονα στον συγκεκριμένο τόπο διαπραγμάτευσης(της προσφοράς και της ζήτησης).

- Επιτρέπει την ελεύθερη διαμόρφωση τιμών των αγαθών με βάση τον θεμελιώδη νόμο της προσφοράς και της ζήτησης. Με αυτόν τον τρόπο περιορίζεται ο κίνδυνος της δημιουργίας τεχνητών τιμών.
- Δίνει την ευκαιρία στις επιχειρήσεις να αναζητήσουν κεφάλαια Ταυτόχρονα όμως , επιτρέπουν στους επενδυτές να διαθέσουν τα χρήματα που έχουν στην επένδυση τους σε τίτλους, με την προσδοκία του κέρδους. Κατά αυτόν τον τρόπο συμβάλλουν στην τόνωση της παραγωγικότητας και γενικότερα στην ανάπτυξη της χώρας που λειτουργεί το χρηματιστήριο.

Το πρώτο **Χρηματιστήριο Αξιών** ιδρύθηκε στην Αμβέρσα (σημερινό Βέλγιο) το 1460. Το 1602 εισήχθη στο Χρηματιστήριο του Άμστερνταμ και η πρώτη πολυμετοχική εταιρεία, η ολλανδική Εταιρεία των Ανατολικών Ινδιών. Η ύπαρξη του χρηματιστηρίου σήμερα είναι ένας σηματικός κι απαραίτητος θεσμός για την οικονομία της χώρας. Στις πλουσιότερες οικονομίες επόμενο είναι να συναντήσουμε και τις μεγαλύτερες κεφαλαιοαγορές, όπως του Λονδίνου, του Παρισίου, της Νέας Υόρκης, της Φρανκφούρτηςκ.ά.

Το Χρηματιστήριο Αθηνών ιδρύθηκε τον 19^{ο} αιώνα. Παρόλα αυτά όμως η οικονομία δεν ήταν ανταγωνιστική λόγω του ότι τον έλεγχο της ελληνικής οικονομίας τον είχε το κράτος. Την δεκαετία του 90’ δειλά δειλά άρχισε να απελευθερώνεται το τραπεζικό σύστημα της Ελλάδας, οι τράπεζες ιδιωτικοποιήθηκαν μερικώς αλλά και άλλες δημόσιες εταιρίες όπως ο Ο.Τ.Ε και λίγο αργότερα η Δ.Ε.Η. Λίγο πριν τον ερχομό της νέας χιλιετίας το χρηματιστήριο είχε περίπου 1.500.000 ενεργούς μετόχους. Αυτό που στιγμάτησε την ελληνική οικονομία ήταν το Κραχ (όπως ονομάστηκε) κατά το 1999-2003. Ο αριθμός των ενεργών επενδυτών εννοείτα πως μειώθηκε ριζικά. Κατά το 2007 έδειξαν ενδιαφέρον σε επενδύσεις επιχειρήσεων που είχαν δραστηριότητα σε βαλκανικά κράτη.

Οι Διεθνείς Οικονομικοί Δείκτες όπως είχαν διαμορφωθεί κατά τις 10 & 11/06/2016

ΕΥΡΩΠΗ

Δείκτης	Χώρα	ΤΙΜΗ	Διαφ. %	Διαφ.	Χαμηλό	Υψηλό	Ημ/νία
AEX	Ολλανδία	435,77	-2,32 %	-10,33	435,63	444,77	11/06
ATX	Αυστρία	2189,87	-2,02 %	-45,18	2179,12	2235,79	10/06
BEL 20	Βέλγιο	3441,60	-2,19 %	-77,19	3431,83	3509,56	11/06
BUX	Ουγγαρία	26519,61	-1,69 %	-456,30	26519,61	27029,60	10/06
CAC	Γαλλία	4306,72	-2,24 %	-98,89	4301,10	4390,81	11/06
DAX	Γερμανία	9834,62	-2,52 %	-254,25	9819,12	10026,49	11/06
FTSEMIB	Ιταλία	17120,16	-3,62 %	-643,72	17106,25	17705,68	11/06
Γενικός δείκτης	Ελλάδα	618,67	-4,18 %	-27,02	615,44	647,56	11/06
HEX	Σουηδία	3203,62	-1,80 %	-58,62	3201,31	3262,14	11/06
IBEX 35	Ισπανία	8490,50	-3,18 %	-279,00	8486,80	8732,60	10/06
OW 20	Πολωνία	1782,02	-1,77 %	-32,12	1777,73	1816,92	10/06
PSI 20	Πορτογαλία	4703,17	-2,15 %	-103,25	4703,17	4798,77	11/06
PX 50	Τσεχία	840,05	-3,20 %	-27,74	833,92	870,64	10/06
RTSI	Ρωσία	924,65	-2,76 %	-26,21	924,65	944,28	10/06
SMI	Ελβετία	7922,71	-1,90 %	-153,64	7909,65	8047,51	10/06

ΑΣΙΑ / ΕΙΡΗΝΙΚΟΣ

Δείκτης	Χώρα	ΤΙΜΗ	Διαφ. %	Διαφ.	Χαμηλό	Υψηλό	Ημ/νία
Hang Seng ▾	Χονκ Κογκ	21042,64	-1,20 %	-255,24	21017,98	21270,62	10/06
Jakarta comp. ▾	Ινδονησία	4848,06	-0,59 %	-28,73	4848,06	4887,59	10/06
Seoul Comp. ▾	Κορέα	2017,63	-0,32 %	-6,54	2014,15	2022,77	10/06
Nikkei ▾	Ιαπωνία	16601,36	-0,40 %	-67,05	16496,11	16643,36	11/06
NZ 50 ▲	Νέα Ζηλανδία	6688,04	0,29 %	19,17	6668,86	6692,08	23/03
Sensex ▾	Ινδία	26635,75	-0,48 %	-127,71	26620,50	26972,06	10/06
Straits Times ▾	Σιγκαπούρη	3177,06	-1,03 %	-33,12	3174,74	3201,84	30/09
Shanghai Comp. ▾	Κίνα	2927,16	-0,30 %	-8,89	2908,37	2937,99	10/06
Topix ▾	Ιαπωνία	1330,72	-0,50 %	-6,69	1321,90	1335,93	10/06
Taiwan Weighted ▲	Ταϊβάν	8715,48	0,41 %	35,58	8690,59	8754,82	10/06
All Ordinaries ▾	Αυστραλία	5391,60	-0,84 %	-45,80	5374,40	5437,40	10/06

ΑΜΕΡΙΚΗ							
Δείκτης	Χώρα	ΤΙΜΗ	Διαφ. %	Διαφ.	Χαμηλό	Υψηλό	Ημ/νία
Bovespa ▲	Βραζιλία	51629,29	2,26 %	1141,43	50490,11	51811,58	10/06
Dow Jones ▼	ΗΠΑ	17865,34	-0,67 %	-119,85	17812,34	17938,82	11/06
IPC ▼	Μεξικό	45177,50	-1,06 %	-485,21	45163,45	45655,94	10/06
ITSEC ▼	Καναδάς	14037,54	-1,42 %	-202,48	14003,74	14214,77	10/06
Merval 25 ▼	Αργεντινή	13324,43	-1,51 %	-203,94	13267,03	13528,37	10/06
Nasdaq 100 ▼	ΗΠΑ	4461,05	-1,14 %	-51,64	4447,33	4480,48	11/06

ΜΕΣΗ ΑΝΑΤΟΛΗ / ΑΦΡΙΚΗ							
Δείκτης	Χώρα	ΤΙΜΗ	Διαφ. %	Διαφ.	Χαμηλό	Υψηλό	Ημ/νία
TA100 ▼	Ισραήλ	1246,37	-0,43 %	-5,44	1245,46	1250,81	10/06

1.3 Εργαλεία επεξεργασίας Οικονομικών δεικτών

Η εξόρυξη δεδομένων εμπεριέχει ένα ευρύ φάσμα από αλγόριθμους και μεθόδους ανάλυση δεδομένων. Όσο κοντά ερχόμαστε προς της εποχή μας, τόσο πιο αποδοτικοί γίνονται οι αλγόριθμοι ώστε να έχουμε ακόμα πιο συγκεκριμένα αποτελέσματα.

1.3.1 Μέθοδοι επεξεργασίας

Η παγκόσμια οικονομία σήμερα είναι μια αλυσίδα από μικρότερες οικονομίες, την οικονομία κάθε κράτους μαζί με την νομισματική τους πολιτική, που αλληλεξαρτώνται. Όταν αλλάζουν οι τιμές στους δείκτες τόσο σε τιμές αγαθών (διεθνές εμπόριο)αλλά και ισοτιμίας νομίσματος, επηρεάζουν και τις άλλες οικονομίες που συνδέονται, σε μεγαλύτερο ή μικρότερο βαθμό. Οι επιχειρήσεις καθώς και οι επενδυτές παρακολουθούν αυτές τις διακυμάνσεις στις τιμές εγχώριας και διεθνούς αγοράς χρηματιστηρίου, συναλλάγματος, ενεργειακών πόρων κτλπ, για να καταστρώσουν την πολιτική που θα ακολουθήσουν βραχυπρόθεσμα ή μακροπρόθεσμα. Όλα αυτά τα δεδομένα λοιπόν θα ήταν άχρηστα αν δεν μπορούσαμε με κάποιο τρόπο να τα επεξεργαστούμε και να ωφεληθούμε από αυτά.

Η πρόβλεψη των δεικτών συναλλάγματος, και της τιμής των ενεργειακών πόρων είναι κάτι που απασχολεί τους επενδυτές σε παγκόσμιο επίπεδο καθώς διακυβεύονται μεγάλα ποσά. Ποτέ δεν μπορείς να ξέρεις με ακρίβεια ποιά θα είναι η τιμή διότι υπάρχουν αρκετοί παράγοντες πολιτικο-οικονομικου-κοινωνικού περιεχομένου που την καθορίζουν και καθιστούν πολύ δύσκολη την πρόβλεψή της. Έχουν χρησιμοποιηθεί στο παρελθόν αρκετές μέθοδοι και μοντέλα για την πρόβλεψη διάφορων τιμών που με τα χρόνια αναβαθμίζονται και γίνονται πιο αποδοτικά. Στις μέρες μας, η χρήση των νέων αυτών μεθόδων και μοντέλων δίνει τη δυνατότητα στους επενδυτές και τους ερευνητές τους να ποντάρουν ασφαλέστερα τα χρήματά τους και να έχουν αποτελεσματικότερα κέρδη. Η ορθή ανάλυση των οικονομικών δεδομένων μέσω των διάφορων μεθόδων εξόρυξης δεδομένων μας οδηγεί σε καλύτερη κατανόηση της τάσης των αγορών. Εφόσον έχουμε κατανοήσει το πως «κινούνται» οι αγορές μπορούμε να καταστρώσουμε ανάλογα τα σχέδιά μας (βραχυπρόθεμσα ή/και μακροπρόθεσμα) έτσι ώστε να επιτύχουμε μια αποδοτικότερη επένδυση.

Η Αναγνώριση προτύπων αποτελεί κλάδο της μηχανικής μάθησης που επικεντρώνεται στην αναγνώριση προτύπων από δεδομένα. Τα συστήματα αναγνώρισης προτύπων τις περισσότερες φορές εκπαιδεύονται από εκπαιδευόμενα στοιχεία/δεδομένα (μάθηση υπό επίβλεψη), όταν δεν υπάρχουν αυτά με τη βοήθεια άλλων αλγορίθμων μπορούμε να ανακαλύψουμε διάφορα άγνωστα patterns (μάθηση χωρίς επίβλεψη). Στην αναγνώριση προτύπων, μηχανική μάθηση και εξόρυξη δεδομένων είναι δύσκολο να διαχωρίσουμε την ορολογία τους μιας και το πεδίο εφαρμογής συμπίμπτει.

Στη Μηχανική Μάθηση, αναγνώριση προτύπων είναι η ανάθεση μιας ετικέτας σε δεδομένη τιμή εισόδου. Η Διακριτική Ανάλυση στον τομέα της Στατιστικής εισήχθη ακριβώς για αυτόν το λόγο το 1936. Ένα παράδειγμα αναγνώρισης προτύπων είναι η Κατηγοριοποίηση, η οποία εκχωρεί κάθε τιμή εισόδου σε σε ένα προκαθορισμένο σύνολο κατηγοριών.

Αλγόριθμοι: Οι αλγόριθμοι για την αναγνώριση προτύπων εξαρτώνται από τον τύπο εξόδου, για το αν έχουμε μάθηση υπό επίβλεψη ή χωρίς κι επίσης για το αν ο αλγόριθμος είναι στατικής φύσεως ή όχι.

Πινακας 4: Αλγόριθμοι Κατηγοριοποίησης

Παραμετρικοί	Μη Παραμετρικοί
Γραμμική διακριτική ανάλυση	Δέντρα Απόφασης – Λίστες αποφάσεων
Τετραγωνική διακριτική ανάλυση	Εκτίμηση πυρήνα | Πλησιέστερου Κ Γείτονα
Ταξινομητής μέγιστης εντροπίας	Naïve Bayes Ταξινομητής
	Νευρωνικά Δίκτυα
	Perceptrons
	Μηχανές Διανυσμάτων Υποστήριξης
	Προγραμματισμός Γονιδιακής έκφρασης

1.3.2 Πακέτα επεξεργασίας (softwares)

Παλαιότερα η εξόρυξη γνώσης από δεδομένα ήταν χρονοβόρα και απαιτούσε πολλές εργατοώρες για να υλοποιηθεί. Πλέον, με τη χρήση των υπολογιστών η διαδικασία έχει γίνει ευκολότερη και πιο αποτελεσματική. Υπάρχει ποικιλομορφία χάρει στα διάφορα προγράμματα που κυκλοφορούν στο εμπόριο. Πρατίθενται παρακάτω μερικά από τα αποδοτικότερα data mining λογισμικά:

1. **IBM SPSS Modeler**: είναι ένα λογισμικό εξόρυξης δεδομένων από την εταιρία IBM. Η εξόρυξη δεδομένων και ο αναλυτής κειμένου του προγράμματος γίνονται σε εικονικό περιβάλλον κάτι που επιτρέπει τη χρήση αλγόριθμων στατιστικής και εξόρυξης δεδομένων δίχως να χρειαστεί να γράψουν γραμμές κώδικα (να προγραμματίσουν). Χαρακτηριστικά: αυτόματη ταξινόμηση, αυτόματη ομαδοποίηση, ανίχνευση ανωμαλιών, Apriori, Bayesian δίκτυα, CARMA, παλινδρόμηση Cox, λίστα αποφάσεων, υποστήριξη μηχανής διανυσμάτων, νευρωνικά δίκτυα, GenLin(GLM), KNN, γενικευμένη γραμμική – μικτά μοντέλα (GLMM): μοντελοποίηση γραμμικής εξίσωσης.
2. **SAS Data Mining**: παρέχει αλγόριθμους για τη δημιουργία προβλεπτικών και περιγραφικών μοντέλων. Πλαισιώνεται από εικονικό περιβάλλον χρήστη. Χαρακτηριστικά: Ισχυρό σύνολο προετοιμασίας δεδομένων, τεχνικές μείωσης διαστάσεων, διαδραστική οπτικοποίηση δεδομένων καθώς κι εξερεύνηση

αυτών, προηγμένη μοντελοποίηση, αυτοματοποιημένη διαδικασία βαθμολόγησης και κλιμακωτής επεξεργασίας.

3. **RapidMiner**: παρέχει ένα ολοκληρωμένο περιβάλλον μηχανικής μάθησης, εξόρυξη δεδομένων, εξόρυξης κειμένου, predictive analytics, business analytics. Χρησιμοποιείται για επιχειρήσεις και βιομηχανικές εφαρμογές, καθώς και για την έρευνα, την ανάπτυξη, την εκπαίδευση, την ταχεία προτυποποίηση και ανάπτυξη εφαρμογών.
4. **Angoss Knowledge STUDIO**: Δημμιουργήθηκε με βάση τις δυνατότητες ανάλυσης δεδομένων και predictive analytics που περιλαμβάνονται στο KnowledgeSEEKER. Το αποτέλεσμα που σου δίνει έχει πληθώρα χαρακτηριστικών προηγμένης μοντελοποίησης και predictive analytics για high-performance business χρήστες και ποσοτικούς αναλυτές.
5. **Microsoft Analysis Services**: η πλατφόρμα του SQL Server Analysis Services δημιουργεί υψηλής απόδοσης αναλυτικά μοντέλα, πολυδιάστατα αλλά και πίνακες, που μπορούν να χρησιμοποιηθούν για τη διαδραστική ανάλυση δεδομένων, υποβολή εκθέσεων και την οπτικοποίηση αυτών.
6. **Oracle Data Mining**: παρέχει ισχυρές και χρήσιμες λειτουργίες εξόρυξης δεδομένων. Επιτρέπει στους χρήστες να ανακαλύψουν νέες ιδέες που κρύβονται στα δεδομένα και να αξιοποιούν αποτελεσματικότερα τις επενδύσεις χάρη στην τεχνολογία που σου προσφέρουν οι βάσεις δεδομένων της Oracle.

Κεφάλαιο 2

Εξόρυξη δεδομένων και οικονομικά δεδομένα

Κεφάλαιο 2

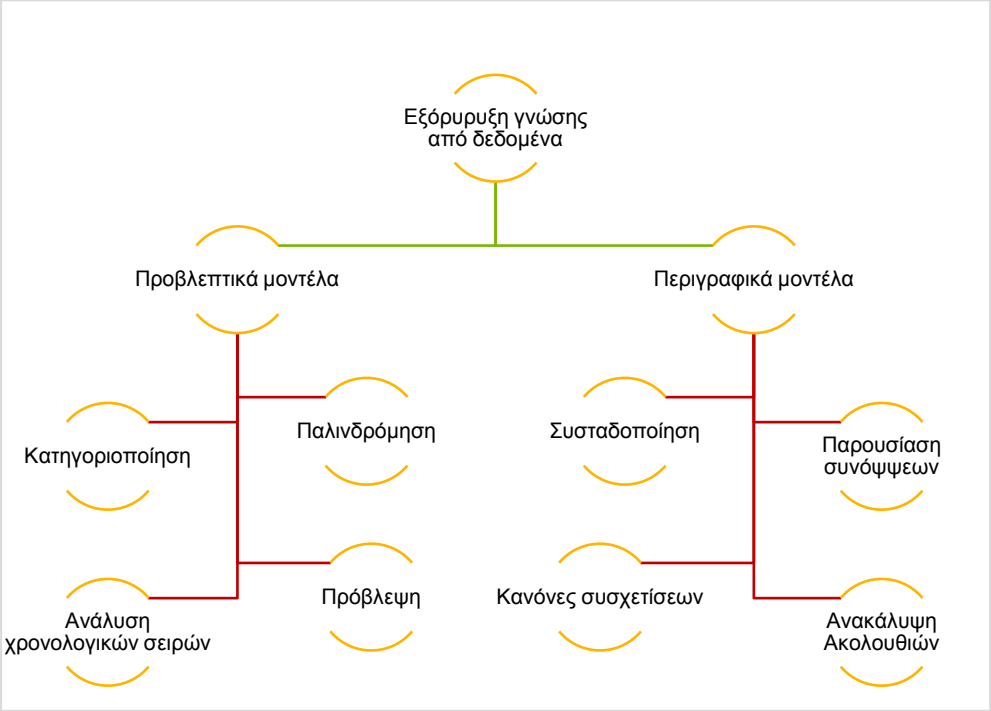
2.1 Τι είναι εξόρυξη δεδομένων

Ο όγκος των δεδομένων που φυλάσσονται στα αρχεία και στις βάσεις δεδομένων αυξάνεται με ένα εκπληκτικό ρυθμό. Την ίδια στιγμή, οι χρήστες αυτών των δεδομένων επιζητούν από αυτά πιο εξειδικευμένες πληροφορίες. Ένας διευθυντής πωλήσεων δεν είναι πια ικανοποιημένος με μία απλή λίστα από στοιχεία πελατών αλλά θέλει λεπτομερείς πληροφορίες σχετικά με τις προηγούμενες αγορές των πελατών καθώς επίσης και με τις προβλέψεις για τις μελλοντικές αγορές τους. Απλές ερωτήσεις, που μπορούν να εκφραστούν σε μία δομημένη γλώσσα ερωτήσεων (SQL), δεν αρκούν για να υποστηρίξουν τις αυξανόμενες αυτές τους απαιτήσεις για πληροφορίες. Η εξόρυξη γνώσης από δεδομένα παρεμβαίνει προκειμένου να ικανοποιήσει αυτές τις ανάγκες. Η εξόρυξη γνώσης από δεδομένα συχνά ορίζεται ως η εύρεση πληροφοριών που είναι κρυμμένες σε μια βάση δεδομένων. Εναλλακτικά, η εξόρυξη γνώσης από δεδομένα ονομάστηκε εξερευνητική ανάλυση δεδομένων, ανακάλυψη καθοδηγούμενη από δεδομένα και συμπερασματική μάθηση.

Οι παραδοσιακές ερωτήσεις σε βάσεις δεδομένων, έχουν πρόσβαση σε μια βάση δεδομένων χρησιμοποιώντας μια καλά ορισμένη ερώτηση, η οποία εκφράζεται σε μια γλώσσα όπως είναι η SQL. Το αποτέλεσμα της ερώτησης αποτελείται από δεδομένα που προέρχονται από τις βάσεις δεδομένων και που ικανοποιούν την ερώτηση. Η έξοδος συνήθως είναι ένα υποσύνολο της βάσης δεδομένων, αλλά μπορεί επίσης και να είναι και μια εξαγόμενη όψη ή να περιέχει συναθροίσεις. Η προσπέλαση σε μια βάση δεδομένων, μέσω της εξόρυξης γνώσης από δεδομένα διαφέρει από της παραδοσιακή προσπέλαση σε πολλά σημεία

2.2 Μέθοδοι εξόρυξης δεδομένων

Η εξόρυξη γνώσης από δεδομένα χωρίζονται σε δυο κατηγορίες, τα προβλεπτικά μοντέλα και τα περιγραφικά μοντέλα.



Εικόνα 1: Μέθοδοι εξόριξης γνώσης

Προβλεπτικά μοντέλα: Ο στόχος τους είναι να προβλέπουν την τιμή ενός συγκεκριμένου χαρακτηριστικού βασιζόμενες στις τιμές των άλλων χαρακτηριστικών. Το υπό πρόβλεψη χαρακτηριστικό είναι γνωστό ως στόχος (target) ή εξαρτημένη μεταβλητή (dependent variable), ενώ τα χαρακτηριστικά που χρησιμοποιούνται για να γίνει η πρόβλεψη είναι γνωστά ως επεξηγηματικές (explanatory) ή ανεξάρτητες μεταβλητές (independent variables).

- **Κατηγοριοποίηση** (classification): Σε αυτή τη μέθοδο τα δεδομένα απεικονίζονται σε προκαθορισμένες κατηγορίες/κλάσεις, οι οποίες έχουν καθοριστεί πριν ακόμα εξετάσουμε τα δεδομένα.

  Παράδειγμα είδους κατηγοριοποίησης αποτελεί η αναγνώριση προτύπου (pattern recognition). Δηλαδή ένα πρότυπο εισόδου κατηγοριοποιείται σε κάποια από τις ήδη υπάρχουσες κατηγορίες, με βάση την εγγύτητά του ως προς αυτές τις προκαθορισμένες κατηγορίες.

- **Παλινδρόμηση** (regression): χρησιμοποιείται για να απεικονιστεί ένα στοιχειώδες δεδομένο σε μία πραγματική μεταβλητή πρόβλεψης. Στην πραγματικότητα η παλινδρόμηση περιλαμβάνει την εκμάθηση της συνάρτησης που κάνει αυτήν την απεικόνιση. Η παλινδρόμηση προυποθέτει ότι τα σχετικά δεδομένα ταιριάζουν με μερικά γνωστά είδη συνάρτησης (για παράδειγμα γραμμική, λογαριθμική κ.ά) και μετά καθορίζει την καλύτερη συνάρτηση αυτού του είδους που μοντελοποιεί τα δεδομένα που έχουν δοθεί.
- **Ανάλυση χρονοσειρών** (time series analysis): μελετάται η τιμή γνωρίσματος καθώς μεταβάλλεται στο χρόνο. Οι τιμές συνήθως λαμβάνονται σε ίδα χρονικά διαστήματα (ανά ώρα, ημερησίως, εβδομαδιαίως κοκ). Για να παρασταθούν οπτικά οι χρονοσειρές χρησιμοποιείται ένα διάγραμμα χρονοσειρών.
- **Πρόβλεψη**: Πολλές από τις πρακτικές εφαρμογές εξόρυξης δεδομένων μπορούν να θεωρήσουν σαν πρόβλεψη μελλοντικών καταστάσεων με γνώση των προηγούμενων και των σημερινών δεδομένων. Η *πρόβλεψη* μπορεί να θεωρηθεί σαν ένα είδος κατηγοριοποίησης. (Σημείωση: Αυτή η εργασία εξόρυξης γνώσης είναι διαφορετική από το μοντέλο πρόβλεψης, παρόλο που η διαδικασία πρόβλεψης αποτελεί έναν τύπο μοντέλου πρόβλεψης.) Η διαφορά είναι ότι ως πρόβλεψη θεωρείται περισσότερο το να δίνεται τιμή σε μία μελλοντική κατάσταση παρά σε μια τρέχουσα. Οι εφαρμογές πρόβλεψης περιλαμβάνουν πρόγνωση πλημμύρων, αναγνώριση ομιλίας, μηχανική μάθηση και αναγνώριση προτύπου.

Περιγραφικά μοντέλα: Ο στόχος των οποίων είναι να αναγνωρίζουν πρότυπα ή συσχετίσεις στα δεδομένα. Σε αντίθεση με το προβλεπτικό, το περιγραφικό μοντέλο λειτουργεί σαν μέσο που διερευνά τις ιδιότητες των δεδομένων που εξετάζονται, όχι να προβλέπει νέες ιδιότητες. Η συσταδοποίηση, η παρουσίαση συνόψεων, οι κανόνες συσχετίσεων και η ανακάλυψη ακολουθιών συνήθως θεωρούνται περιγραφικές εργασίες από τη φύση τους.

- **Συσταδοποίηση** (clustering): Είναι παρόμοια με την *Κατηγοριοποίηση* εκτός από το γεγονός ότι οι συστάδες/ομάδες δεδομένων δεν είναι προκαθορισμένες αλλά ορίζονται κυρίως από τα ίδια τα δεδομένα που εισάγουμε κάθε φορά.

Στην ουσία είναι η διαμέριση/τμηματοποίηση σε ομάδες με κοινά και προκαθορισμένα γνωρίσματα των δεδομένων, με τα πιο σχετικά δεδομένα μεταξύ τους να εισάγονται στις ίδιες ομάδες. Παράδειγμα τέτοιου είδους αποτελεί μια διαφημιστική εταιρία που για λογαριασμό μιας πολυεθνικής πρέπει να δημιουργήσει ειδικούς καταλόγους με δημογραφικά στοιχεία των κατοίκων μιας πόλης, με κριτήρια όπως το εισόδημα, τόπος διαμονής, χαρακτηριστικά πελατών (όπως ύψος, βάρος, ηλικία).Η διαφημιστική εταιρία λοιπόν θα συσταδοποιήσει τους πιθανούς πελάτες της βασιζόμενη στις προκαθορισμένες τιμές γνωρισμάτων. Τα αποτελέσματα θα χρησιμοποιηθούν για τη δημιουργία και την αποστολή του ειδικού διαφημιστικού καταλόγου για το καταλληλότερο μέρος του πληθυσμού.

- **Παρουσίαση Συνόψεων** (summarization): απεικονίζει τα δεδομένα σε υποσύνολά τους με συνοδευτικές απλές περιγραφές. Η σύνοψη δεδομένων ονομάζεται επίσης και *χαρακτηρισμός* (characterization) ή *γενίκευση* (generalization). Εξάγει ή παράγει αντιπροσωπευτικές πληροφορίες σχετικά με τις βάσεις δεδομένων. Αυτό γίνεται ανακτώντας, στην πραγματικότητα, τμήματα από δεδομένα. Εναλλακτικά, μπορούν να εξαχθούν από τα δεδομένα συνοπτικές πληροφορίες (όπως είναι ο μέσος όρος κάποιου αριθμητικού γνωρίσματος). Εν ολίγοις, η παρουσίαση συνόψεων χαρακτηρίζει τα περιεχόμενα της βάσης δεδομένων.
- **Ανακάλυψη ακολουθιών** (sequential analysis): ή αλλιώς ανακάλυψη ακολουθιών (sequential discovery) χρησιμοποιείται για να καθοριστούν σειριακά πρότυπα στα δεδομένα.
- **Κανόνες Συσχέτισης***: Η ανάλυση συνδέσμων* (link analysis), που εναλλακτικά αναφέρεται και σαν *ανάλυση συγγένειας* (affinity analysis) ή *συσχέτιση*(association), αναφέρεται στη διαδικασία εκείνη της εξόρυξης γνώσης που αποκαλύπτει συσχετίσεις μεταξύ των δεδομένων. Το καλύτερο παράδειγμα αυτού του είδους της εφαρμογής είναι ο προσδιορισμός κανόνων συσχετίσεων. Ένας *κανόνας συσχέτισης* (association rule) είναι ένα μοντέλο που αναγνωρίζει ειδικούς τύπους συσχέτισης μεταξύ δεδομμένων. Αυτές οι συσχετίσεις συχνά χρησιμοποιούνται στις λιανικές πωλήσεις για να αναγνωριστούν προιόντα που συχνά αγοράζονται μαζί.

Στα πειράματα χρησιμοποιήθηκαν η Κατηγοριοποίηση και η Συσταδοποίηση.

2.3 Εφαρμογές εξόρυξης δεδομένων σε οικονομικά δεδομένα

Ένας τομέας που οι βάσεις δεδομένων χρησιμοποιούνται κατά κόρον είναι ο οικονομικός. Τα δεδομένα συλλέγονται από τράπεζες, διάφορους οργανισμούς, επενδυτές κι επιστήμονες με σκοπό μέσα από την επεξεργασία τους να πάρουμε καλύτερες αποφάσεις ενεργώντας σύμφωνα με την ανάλυση της αγοράς.

Η σειρά που ακολουθείται για την επεξεργασία των δεδομένων είναι η εξής:

- Συλλέγουμε τα δεδομένα στην αποθήκη δεδομένων (data bank)
- Τα βελτιώνουμε, γνωστό και ως data refinement
- Δημιουργία & ανάπτυξη ενός μοντέλου από διάφορες τεχνικές συσταδοποίησης και ταξινόμησης που βοηθούν την τράπεζα ή κάποιον οργανισμό να ομαδοποιεί τους πελάτες με κοινά χαρακτηριστικά.
- Μέσω των τεχνικών οπτικοποίησης η τράπεζα έχει τη δυνατότητα να καταλάβει αν κάποιος πελάτης επιχείρησε παράνομη πράξη προβάλλοντας τα δεδομένα από διάφορες οπτικές γωνίες.

Κεφάλαιο 3

Μέθοδος Πρόβλεψης

Κεφάλαιο 3

3.1 Μέτρηση συσχέτισης ισοτιμιών (περιγραφή ότι κάναμε με matlab)

Ας ξεκινήσουμε λοιπόν αναλύοντας την διαδικασία που ακολουθήσαμε για την επεξεργασία μέσω του προγράμματος Matlab (MathWorks).

Στην αρχή δημιουργήσαμε τον πίνακα συσχέτισης μεταξύ όλων των αρχείων (72 αρχεία) για να κατανοήσουμε ποιες χρονοσειρές δεικτών είναι παρόμοιες. Για να γίνουν πιο κατανοητά όλα αυτά, πρώτα θα δώσουμε τον ορισμό της Συσχέτισης και του Πίνακα συσχέτισης.

Η συσχέτιση είναι μια τεχνική που χρησιμοποιείται στην Στατιστική μέσω της οποίας καταλαβαίνουμε το εάν και κατά πόσο έντονα κάποια ζεύγη μεταβλητών. Ας δώσουμε ένα παράδειγμα, το ύψος και το βάρος ενός ανθρώπου είναι σχετικά έχοντας κατά νου πως ψηλότεροι άνθρωποι έχουν μεγαλύτερο βάρος από άλλους κοντύτερούς τους. Η σχέση αυτή δεν είναι τέλεια, διότι, άνθρωποι πάλι του ίδιου ύψους έχουν διαφορετικό βάρος. Η συσχέτιση απλά μας δείχνει πόση μεταβολή του βάρους των ανθρώπων σχετίζεται με το ύψος τους. Επιστημονικά, ο συντελεστής συσχέτισης είναι ένας αριθμός που ποσοτικοποιεί κάποιος είδος συσχέτισης/εξάρτησης, δηλαδή μιας στατιστικής σχέσης μεταξύ δύο ή περισσότερων τυχαίων μεταβλητών.

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{n s_x s_y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\,\sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

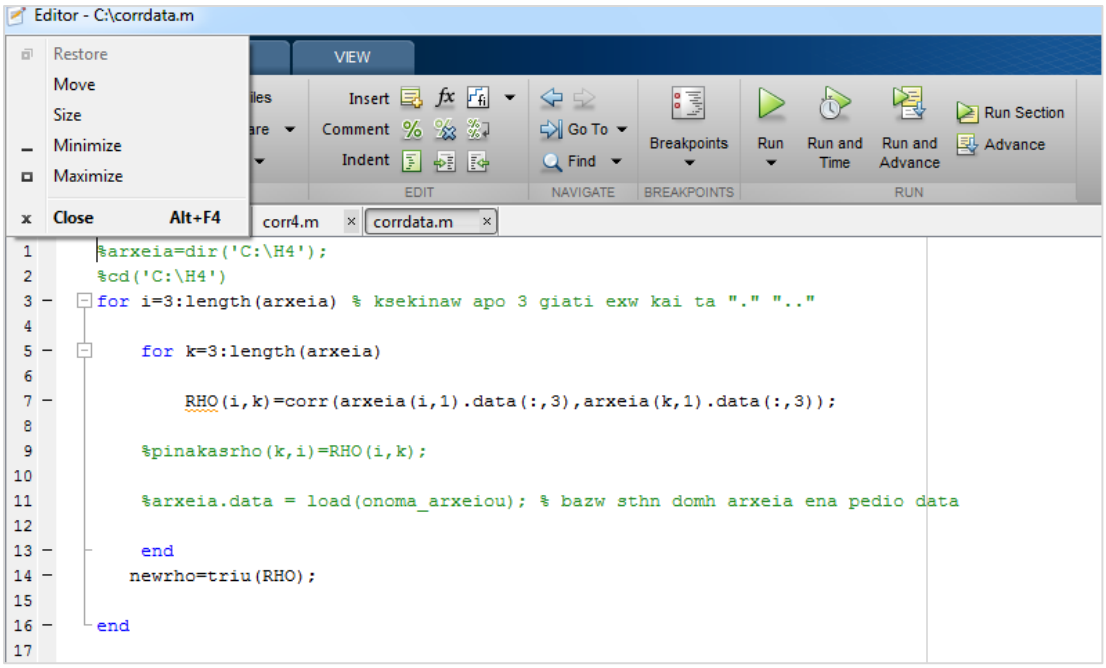
Τύποι συντελεστών συσχέτισης (ονομαστικά):

- Pearson product-moment correlation coefficient
- Intraclass correlation
- Rank correlation
    - Spearman’s rank correlation coefficient

- Kendall tau tank correlation coefficient
- Goodman & Kruskal's gamma

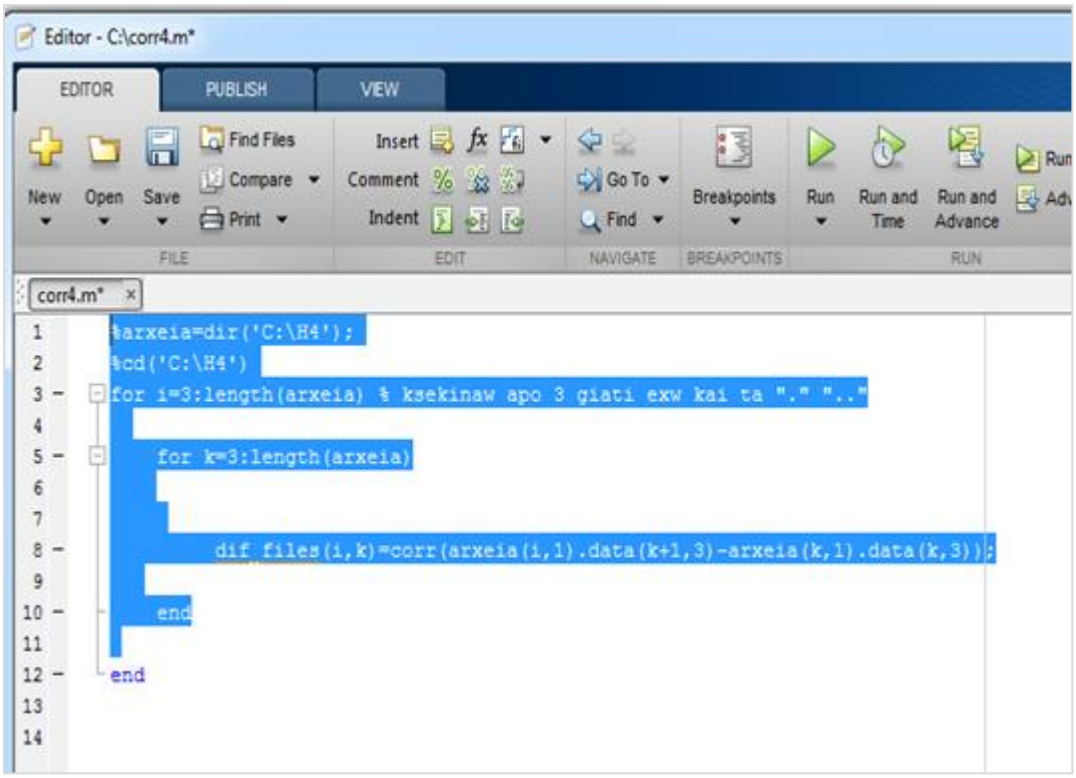
Ο πίνακας συσχέτισης χρησιμοποιείται για να διερευνηθεί η εξάρτηση μεταξύ πολλαπλών μεταβλητών την ίδια στιγμή. Περιέχει τους συντελεστές συσχέτισης μεταξύ κάθε μεταβλητής με τις υπόλοιπες.

Ακολουθεί ο κώδικας που χρησιμοποιηθήκε για τη δημιουργία του πίνακα συσχέτισης :



Εικόνα 2: Περιβάλλον MATLAB για την εξαγωγή του πίνακα συσχέτισης

Δεύτερο βήμα, ήταν η δημιουργία πίνακα συσχέτισης με ολισθημένο το ένα σήμα από το άλλο κατά 4 ώρες. Η συχνότητα είναι ανά 4 ώρες διότι κατά αυτή τη συχνότητα εναλλάσσονται οι τιμές των αρχείων. Ο κώδικας για να υλοποιηθεί είναι ο παρακάτω.



Εικόνα 3:

Στην συνέχεια επιλέχθηκαν οι σημαντικότερες συσχετίσεις, όπως παρατίθενται στον παρακάτω πίνακα.

- $1^{η}$ & $2^{η}$ Στήλη: Περιέχονται όλα τα ονόματα των αρχείων ισοτιμιών που έχουν τις 30 σημαντικότερες συσχετίσεις
- $3^{η}$ Στήλη: Περιέχει την ολισθημένη κατά 4 ώρες τιμή των συσχετίσεων
- $4^{η}$ Στήλη: Περιέχει την τιμή του συμμετρικού πίνακα των συσχετίσεων

Πίνακας 30 σημαντικότερων συσχετίσεων				
	isotimia1	**isotimia2**	**corr4_095**	**timi_sym**
1	AUDUSD240.csv	CADHKD240.csv	0.956319734943518	0.956391477818213
2	AUDUSD240.csv	GBPHKD240.csv	0.954455752380940	0.955025636312074
3	CADHKD240.csv	AUDUSD240.csv	0.955973756426466	0.956391477818213
4	CADHKD240.csv	USDDKK240.csv	-0.951243874288652	-0.951654634718004
5	CADHKD240.csv	USDSGD240.csv	-0.972957329876814	-0.973343356294383
6	EURHKD240.csv	GBPCZK240.csv	-0.952975738028461	-0.953441446233711
7	EURHKD240.csv	SGDPLN240.csv	-0.968172433300755	-0.968752912220143
8	EURHKD240.csv	USDDKK240.csv	-0.996124733931665	-0.996408073948328
9	EURSGD240.csv	GBPCZK240.csv	-0.972188088208813	-0.972729897956049
10	EURSGD240.csv	GBPDKK240.csv	-0.952566109234926	-0.953116791300428
11	EURSGD240.csv	GBPPLN240.csv	-0.953280668861556	-0.953864912134954
12	EURSGD240.csv	GBPSEK240.csv	-0.953548955349280	-0.953915276020958
13	EURSGD240.csv	USDDKK240.csv	-0.971128638695127	-0.971428557335850
14	GBPCZK240.csv	EURHKD240.csv	-0.953202962653627	-0.953441446233711
15	GBPCZK240.csv	EURSGD240.csv	-0.972456649413355	-0.972729897956049
16	GBPCZK240.csv	USDDKK240.csv	0.951406353985655	0.951616477209311
17	GBPDKK240.csv	EURSGD240.csv	-0.952741435546821	-0.953116791300428
18	GBPHKD240.csv	AUDUSD240.csv	0.955002651391012	0.710478517050077
19	GBPHKD240.csv	USDSGD240.csv	-0.951987343359319	-0.952074742314815
20	GBPPLN240.csv	EURSGD240.csv	-0.953992072119667	-0.953864912134954
21	GBPSEK240.csv	EURSGD240.csv	-0.953855352454865	-0.953915276020958
22	SGDPLN240.csv	EURHKD240.csv	-0.968671525083776	-0.968752912220143
23	SGDPLN240.csv	USDDKK240.csv	0.952978594978895	0.953038771939840
24	USDDKK240.csv	CADHKD240.csv	-0.951676705661791	-0.951654634718004
25	USDDKK240.csv	EURHKD240.csv	-0.996064521736240	-0.996408073948328
26	USDDKK240.csv	EURSGD240.csv	-0.970944688302773	-0.971428557335850
27	USDDKK240.csv	GBPCZK240.csv	0.951118340990971	0.951616477209311
28	USDDKK240.csv	SGDPLN240.csv	0.952425943794237	0.953038771939840
29	USDSGD240.csv	CADHKD240.csv	-0.973302194561152	-0.973343356294383
30	USDSGD240.csv	GBPHKD240.csv	-0.951451131492844	-0.952074742314815

Από τις παραπάνω τιμές καταλάβαμε πως 2 φαίνεται να έχουν μεγαλύτερη συσχέτιση στην ολίσθηση τεσσάρων ωρών από ότι στην συσχέτιση των δύο ισοτιμιών την ίδια χρονική στιγμή. Αυτό πρακτικά σημαίνει ότι οι δύο χρονοσειρές είναι περισσότερο όμοιες όταν η μία είναι όλισθημένη σε σχέση με την άλλη. Αυτό είναι μια ένδειξη είτε ότι η μια επηρεάζει την άλλη είτε ότι η μία αντιδρά στις εξελίξεις λίγο ετεροχρονισμένα σε σχέση με την άλλη. Από τις δύο αυτές συσχετίσεις μελετήσαμε την συσχέτιση των ισοτιμιών που βρίσκεται στην $17^{η}$ γραμμή του παραπάνω πίνακα.

Αφού καταλήξαμε στην συσχέτιση αυτών των ισοτιμιών, εξάγαμε τα χαρακτηριστικά από μια χρονική στιγμή, και δημιουργήσαμε πρόβλημα ταξηνόμισης γνωρίζοντας από τα δεδομένα μας πως κινήθηκαν οι ισοτιμίες το επόμενο τετράωρο. Τα χαρακτηρισικά που χρησιμοοποιήθηκαν ήταν η τιμή του δείκτη αλλά και η κλίση του δείκτη του τελευταίο 4ώρο. Θέωρήσαμε πιο ορθή την δημιουργία προβλήματος 3 κατηγοριών, η οποίες να αντιστοιχούν α) σε πτώση του δείκτη στο επόμενο 4ωρο, β) σε άνοδο του δείκτη στο επόμενο 4ωρο, γ) σε σταθεροποίηση του δείκτη στο επόμενο τετράωρο.

Στο σήμειο αυτό όμως προκύπτει το ζήτημα πότε μπορούμε να θεωρήσουμε ότι είναι σταθερός ο δείκτης. Από τις τιμές των ισοτιμιών παρατηρήσαμε ότι ούτως ή άλλως οι μεταβολές στους δείκτες είναι της τάξης μεγέθους το 0.1‰ έως 1‰ περίπου. Όποτε μελετήσαμε διάφορες τιμές του κατωφλιού αυτού. Για να γίνει κατανοητό αν για παράδειγμα χρησιμοποιούσαμε κατώφλι μεταβολής 0.5‰ τότε:

- **Α) εάν στο 4ωρο έχουμε άνοδο πάνω από 0.5‰ τότε δίνουμε επισημίωση ανόδου (ετικέτα κλάσης 1)**
- **Β) εάν στο 4ωρο έχουμε πτώση πάνω από -0.5‰ τότε δίνουμε επισημίωση καθόδου (ετικέτα κλάσης -1)**
- **Γ) εάν δεν έχουμε κανένα από τα δύο παραπάνω τότε θεωρούμε ότι ο δείκτης έμεινε σταθερός (ετικέτα κλάσης 0)**

Με το παραπάνω σκεπτικό δημιουργήσαμε αντίστοιχα προβλήματα τριών κατηγοριών για τιμές κατωφλιού 0.1‰, 0.3‰, 0.5‰, 0.7‰. 0.9‰. Τέλος μηδενίσαμε το κατώφλι αυτό και ως εκ τούτου το πρόβλημα εκφυλίστηκε σε πρόβλημα μόνο ανόδου και καθόδου, δηλαδή πρόβλημα 2 κατηγοριών. Παρακάτω παραθέτουμε τους κώδικες MATLAB την εξαγωγή των χαρακτηριστικών, σε όλες αυτές τις περιπτώσεις.

Editor - C:\ora12\ora12.m

EDITOR | PUBLISH | VIEW

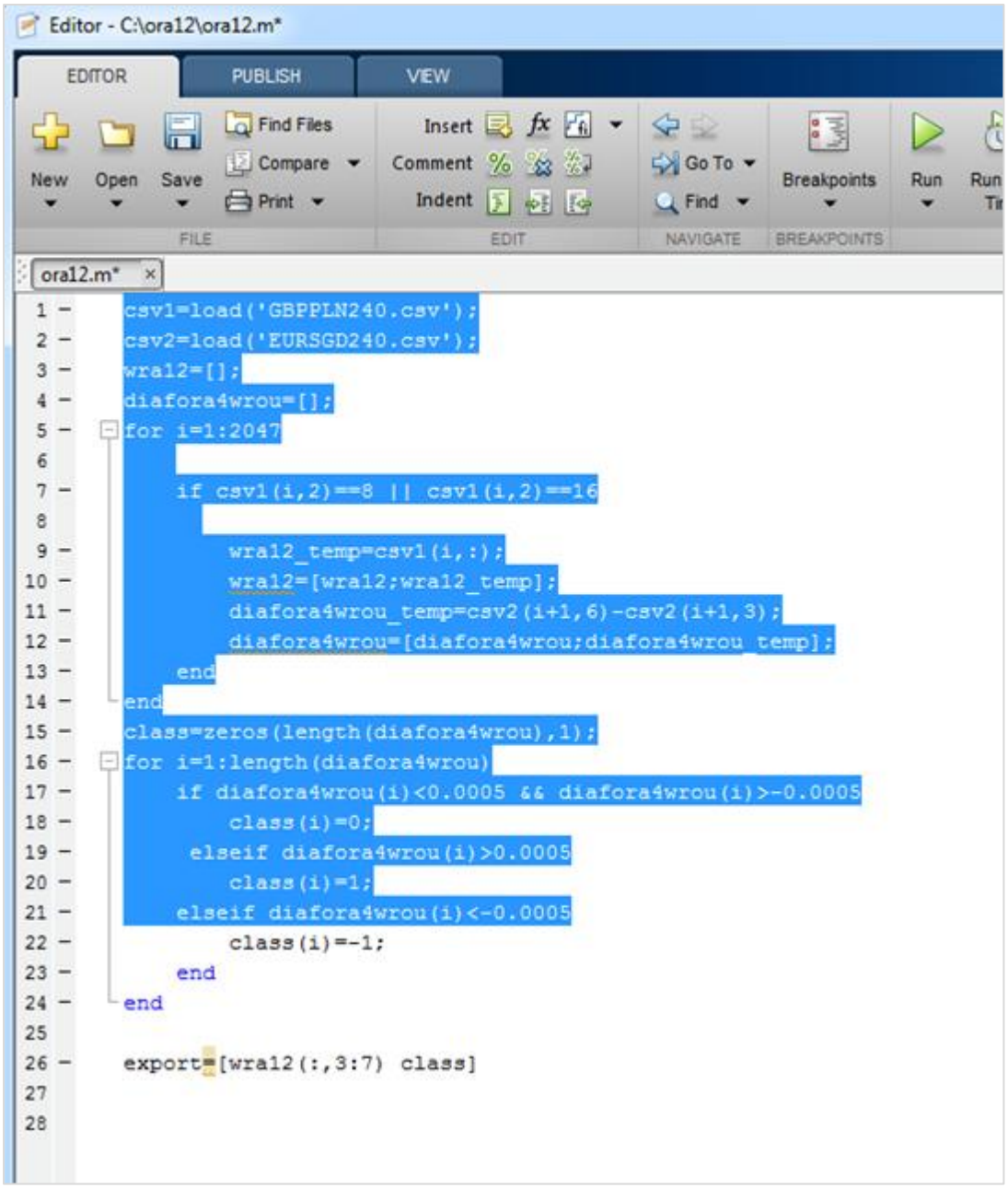
New | Open | Save | Find Files | Compare | Print | Insert | Comment | Indent | Go To | Find | Breakpoints | Run | Run a Time

FILE | EDIT | NAVIGATE | BREAKPOINTS

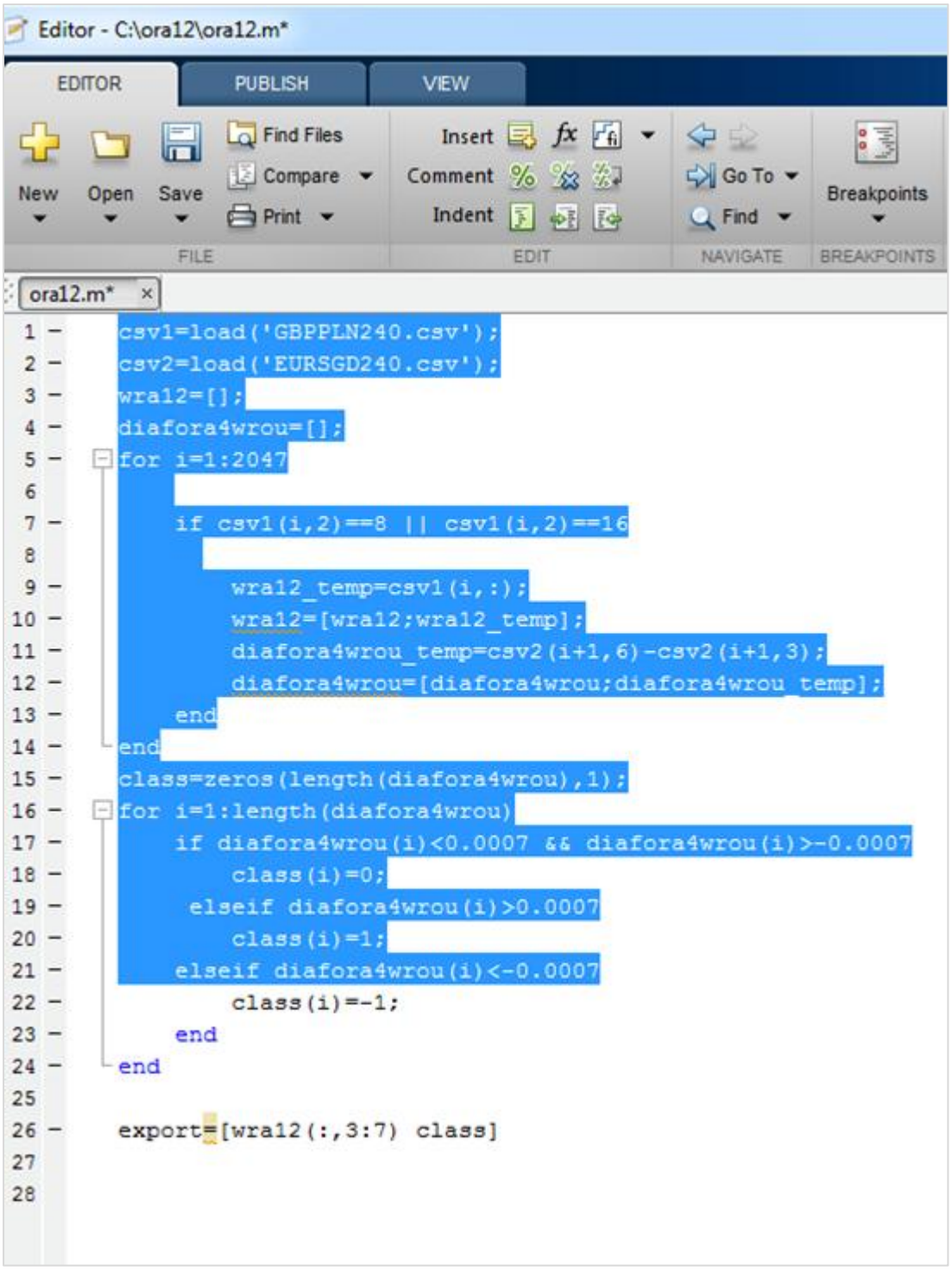
ora12.m

```
1  csv1=load('GBPPLN240.csv');
2  csv2=load('EURSGD240.csv');
3  wra12=[];
4  diafora4wrou=[];
5  for i=1:2047
6
7      if csv1(i,2)==8 || csv1(i,2)==16
8
9          wra12_temp=csv1(i,:);
10         wra12=[wra12;wra12_temp];
11         diafora4wrou_temp=csv2(i+1,6)-csv2(i+1,3);
12         diafora4wrou=[diafora4wrou;diafora4wrou_temp];
13     end
14 end
15 class=zeros(length(diafora4wrou),1);
16 for i=1:length(diafora4wrou)
17     if diafora4wrou(i)<0.0003 && diafora4wrou(i)>-0.0003
18         class(i)=0;
19      elseif diafora4wrou(i)>0.0003
20         class(i)=1;
21     elseif diafora4wrou(i)<-0.0003
22         class(i)=-1;
23     end
24 end
25
26 export=[wra12(:,3:7) class]
27
28
```

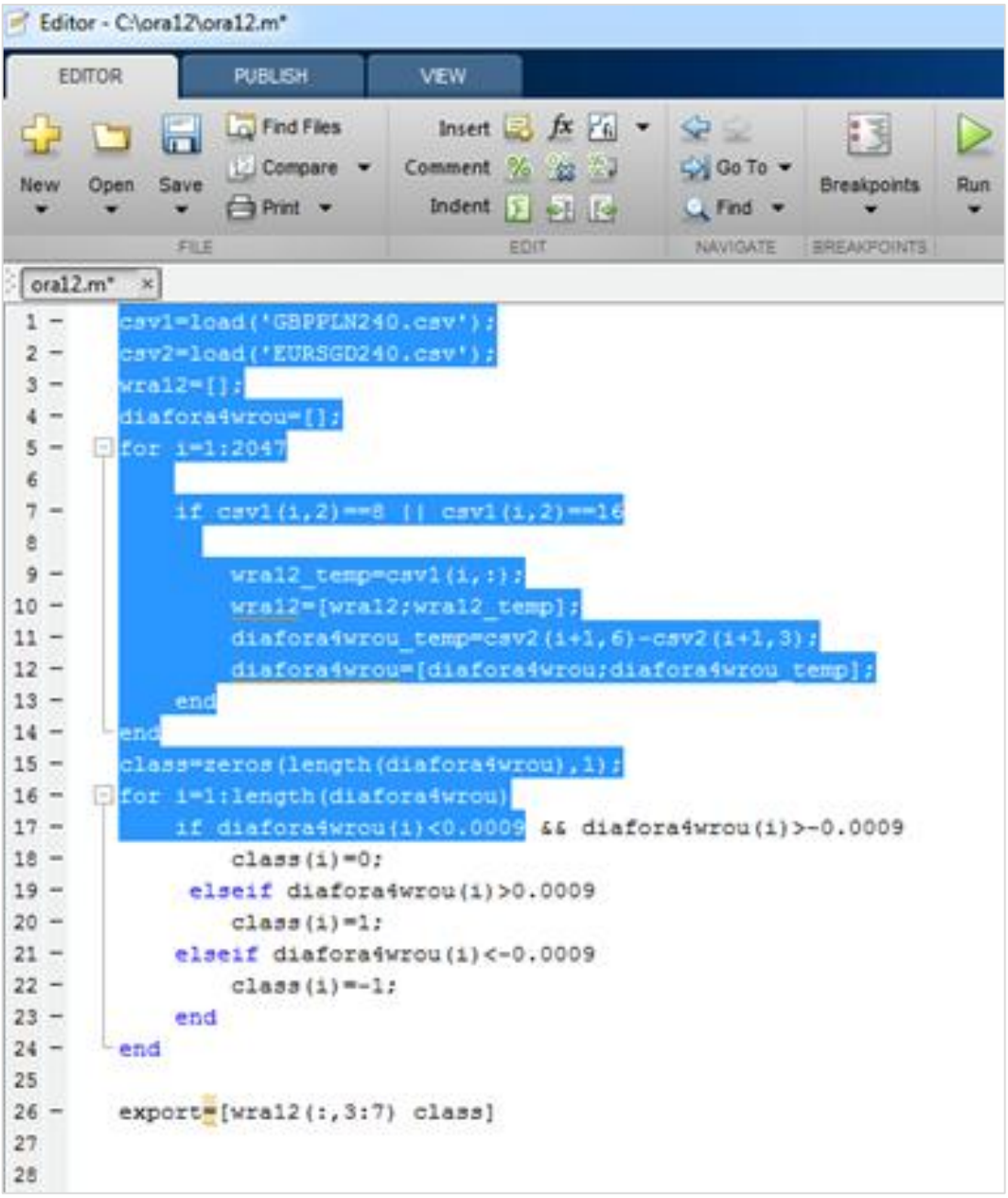
Εικόνα 4: Ο κώδικας για την εξαγωγή χαρακτηριστικών με τιμή 0.3‰



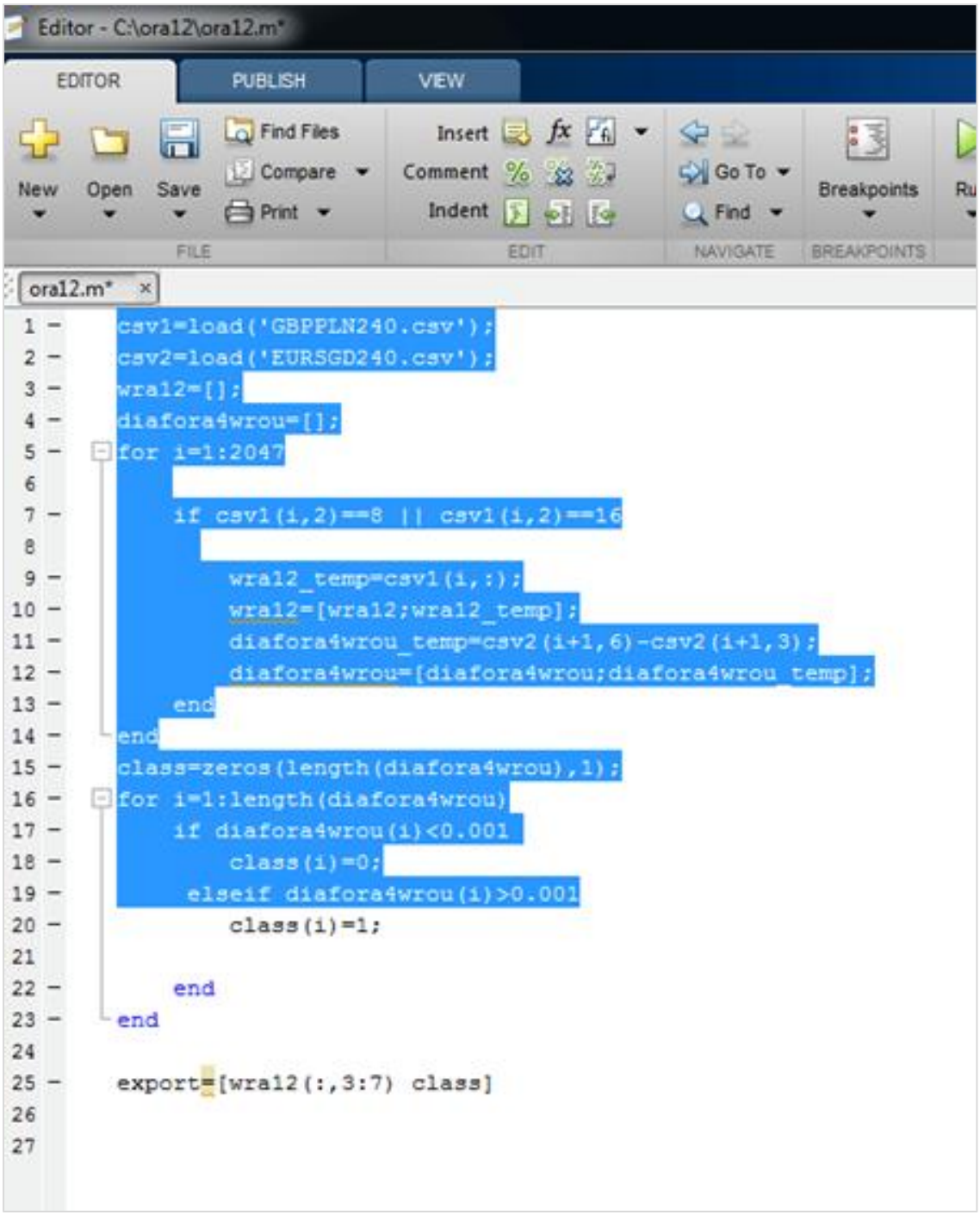
Εικόνα 5: Ο κώδικας για την εξαγωγή χαρακτηριστικών με τιμή 0.5‰



Εικόνα 6: Ο κώδικας για την εξαγωγή χαρακτηριστικών με τιμή 0.7‰



Εικόνα 7: Ο κώδικας για την εξαγωγή χαρακτηριστικών με τιμή 0.9‰



Εικόνα 8: Κώδικας για την εξαγωγή των χαρακτηριστικών του προβληματος 2 κατηγοριών

Μετα από τα παραπανω πειραματιστίκαμε με τρεις διαφορετικούς αλγορίθμους ταξηνόμησης, με την βοήθεια του λογισμικού Weka 3.7. Συγκεκριμένα, χρησιμοποιήσαμε των μπευζιανό ταξινομητη, τις Μηχανές Διανυσμάτων Υποστήριξης, και το Δέντρο Απόφασης C4.5, τα οποία περιγράφονται αναλυτικά παρακάτων.

3.2 Μέθοδοι Ταξινόμησης

3.2.1 Μπευζιανός Ταξινομητής Naïve Bayes (Περιγραφή)

Στη Μηχανική Μάθηση(Machine Learning), ο Naïve Bayes ταξινομητής ανήκει σε μια οικογένεια απλοικών πιθανολογικών ταξινομητών που βασίζονται στην εφαρμογή του θεωρήματος του Bayes μέσω ισχυρών υποθέσεων ανεξαρτησίας μεταξύ των χαρακτηριστικών.

Ένας απλοικός κατά Bayes ταξινομητής, εκτιμά την εξαρτώμενη από την κατηγορία πιθανότητα υποθέτοντας, ότι τα χαρακτηριστικά είναι υπο συνθήκη ανεξάρτητα, δεδομένης μίας ετικέτας κατηγορίας y. Η υπόθεση της υπο συνθήκ ανεξαρτησίας μπορεί να εκφραστεί τυπικά όπως φαίνεται παρακάτω:

$$P(X \mid Y = y) = \prod_{i=1}^{d} P(X_i \mid Y = y) \qquad (\text{εξ.1})$$

όπου κάθε σύνολο χαρακτηριστικών X = $\{X_1, X_2, \ldots, X_d\}$ αποτελέιται από d χαρακτηριστικά.

Πριν προχωρίσουμε λεπτομερώς στον τρόπο λειτουργίας του απλοικού κατά Bayes ταξινομητή, ας εξετάσουμε την έννοια της Υπό Συνθήκη Ανεξαρτησίας. Εστω ότι X, Y και Z είναι τρία σύνολα από τυχαίες μεταβλητές. Οι μεταβλητές στο X είναι υπό συνθήκη ανεξάρτητες του Y δοθέντος του Z, αν ισχύει η παρακάτω συνθήκη:

$$P(X \mid Y, Z) = P(X \mid Z) \qquad (\text{εξ.2})$$

Ένα παράδειγμα υπό συνθήκη ανεξαρτησίας είναι η σχέση μεταξύ του μήκους του χεριού ενός ατόμου και των δεξιοτήτων διαβάσματός του. Κάποιος μπορεί να παρατηρήσει ότι τα άτομα με μακρύτερα χέρια τείνουν να έχουν υψηλότερη δεξιότητα στο διάβασμα. Αυτή η σχέση εξηγείται από την παρουσία ενός παράγοντα, ο οποίος προκαλεί σύγχυση, όπως είναι η ηλικία. Ένα μικρό παιδί λοιπόν που έχει μικρά χέρια του λείπει η δεξιότητα διαβάσματος ενός ενήλικα. Αν η ηλικία ενός ατόμου είναι σταθερή, τότε η παρατηρούμενη σχέση μεταξύ του μήκους του χεριού και των δεξιοτήτων ανάγνωσης εξαφανίζεται. Επομένως, προκύπτει το συμπέρασμα ότι το μήκος του χεριού και οι δεξιότητες διαβάσματος είναι υπό συνθήκη ανεξάρτητα, όταν η ηλικία είναι σταθερή.

Η υπό συνθήκη ανεξαρτησία μεταξύ των X και Y μπορεί επίσης να γραφεί σε μια μορφή η οποία να μοιάζει με την εξίσωση 1:

$$P(X \mid Y,Z) = \frac{P(X,Y,Z)}{P(Z)} = \frac{P(X,Y,Z)}{P(Y,Z)} \cdot \frac{P(Y,Z)}{P(Z)} = P(X \mid Y,Z) \cdot P(Y \mid Z)$$
$$= P(X \mid Z) \cdot P(Y \mid Z) \qquad \text{(εξ.3)}$$

όπου η εξίσωση 2 χρησιμοποιήθηκε για να ληφθεί η τελυταία γραμμή της Εξίσωσης 3.

Με την υπόθεση της υπό συνθήκη ανεξαρτησίας, αντί να υπολογίζεται η εξαρτώμενη από την κατηγορία πιθανότητα για κάθε συνδυασμό του X, αρκεί να εκτιμηθεί η υπό συνθήκη πιθανότητα για κάθε X, δοθέντος του Y. Η τελευταία προσέγγιση είναι πιο πρακτική επειδή δεν απαιτεί ένα πολύ μεγάλο σύνολο εκπαίδευσης για να λάβουμε μια καλή εκτίμηση της πιθανότητας.

Για να κατηγοριοποιήσει μία εγγραφή ελέγχου, ο απλοικός κατά Bayes ταξινομητής υπολογίζει την εκ των υστέρων για κάθε κατηγορία Y:

$$P(Y \mid X) = \frac{P(Y)\Pi_{i=1}^{d} P(X_i \mid Y)}{P(X)} \qquad \text{(εξ.4)}$$

Δεδομένου ότι η τιμή P(X) είναι μια σταθερή για κάθε Y, αρκεί να επιλεγεί η κατηγορία που μεγιστοποιεί τον αριθμητή

Οι απλοικοί κατά Bayes ταξινομητές έχουν γενικά τα παρακάτω χαρακτηριστικά:

Είναι εύρωστοι σε απομονωμένα σημεία θορύβου, επειδή τέτοια σημεία υπολογίζονται στο μέσο όρο όταν εκτιμώνται οι υπό συνθήκη πιθανότητες από τα δεδομένα.Οι απλοικοί κατά Bayes ταξινομητές μπορούν επίσης να διαχειριστούν ελλειπής τιμές μη λαμβάνοντας υπόψη το δείγμα κατά τη δημιουργία του μοντέλου και της κατηγοριοποίησης.

Είναι εύρωστοι στην ύπαρξη μη σχετικών χαρακτηριστικών. Αν Xi είναι ένα τέτοιο χαρακτηριστικό, τότε η P(Xi|Y) κατανέμεται σχεδόν ομοιόμορφα. Η εξαρτώμενη από την κατηγορία πιθανότητα για το Xi δεν έχει καμία επίδραση στο συνολικό υπολογισμό της εκ των υστέρων πιθανότητας.

Τα συσχετιζόμενα χαρακτηριστικά μπορούν να μειώσουν την απόδοση των απλοικών κατά Bayes ταξινομητών επειδή η υπόθεση της υπό συνθήκη ανεξαρτησίας δεν ισχύει πλέοον για αυτά τα χαρακτηριστικά. Για παράδειγμα, θεωρείστε τις ακόλουθες πιθανότητες:

P(A=0|Y=0)=0.4, P(A=1|Y=0)=0.6,

P(A=0|Y=1)=0.6, P(A=1|Y=0)=0.4,

Όπου το Α είναι ένα δυαδικό χαρακτηριστικό και το Υ είναι μια δυαδική μεταβλητή. Ας υποτεθεί ότι υπάρχει ένα άλλο δυαδικό χαρακτηριστικό Β το οποίο είναι τέλεια συσχετιζόμενο με το Α, όταν Υ=0, αλλά είναι ανεξάρτητο του Α όταν Υ=1. Για απλούστευση, ας υποτεθεί ότι οι εξαρτώμενες από την κατηγορία πιθανότητες για το Β είναι ίδιες, όπως του Α. Δοθείσης μιας εγγραφής με χαρακκτηριστικά Α=0,Β=0, οι εκ των υστέρων πιθανότητες υπολογίζονται όπως φαίνεται παρακάτω.

$$P(Y=0 \mid A=0,B=0)=\frac{P(A=0 \mid Y=0)P(B=0 \mid Y=0)P(Y=0)}{P(A=0,B=0)}=$$
$$=\frac{0.16\times P(Y=0)}{P(A=0,B=0)}$$

$$P(Y=1 \mid A=0,B=0)=\frac{P(A=0 \mid Y=1)P(B=0 \mid Y=1)P(Y=1)}{P(A=0,B=0)}=$$
$$=\frac{0.36\times P(Y=1)}{P(A=0,B=0)}$$

Αν P(y=0)=P(Y=1), τότε ο απλοικός κατά Bayes ταξινομητής θα αποδόσει την εγγραφή στην κατηγορία 1. Ωστόσο μ η αλήθεια είναι ότι

P(A=0,B=0|Y=0) = P(A=0|Y=0)=0.4

Επειδή το Α και Β είναι τέλεια συσχετιζόμενα όταν Υ=0. Ως αποτέλεσμα, η εκ των υστέρων πιθανότητα για το Υ = 0 θα είναι:

$$P(Y=0 \mid A=0,B=0)=\frac{P(A=0,\mathrm{B}=0 \mid \mathrm{Y}=0)P(\mathrm{Y}=0)}{P(A=0,B=0)}=$$
$$=\frac{0.4\times P(Y=0)}{P(A=0,B=0)}$$

η οποία είναι μεγαλύτερη από την αντίστοιχη για Y = 1. Η εγγραφή θα έπρεπ να είχε κατηγοριοποιηθεί ως κατηγορία 0.

3.2.2 Δένδρα απόφασης - Decision Trees (Αλγόριθμος C4.5)

Ο C4.5 είναι ένας αλγόριθμος ο οποίος δημιουργήθηκε από τον Ross Quinian. Αποτελεί προέκταση του προηγούμενου αλγόριθμου ID3. Χρησιμοποιείται για τη δημιουργία ενός δέντρου απόφασης καθώς είναι κι ένα πολύ δυνατό εργαλείο σαν στατιστικός ταξινομητής. Η δημοφιλία του οφείλεται στην ανάδειξή του σαν Nr.1 των Top 10 Αλγόριθμοι σε Data Mining σε έγγραφο που δημοσιεύθηκε από την Springer LNCS το 2008.

Σαν αλγόριθμος, η δουλειά του C4.5 είναι να δημιουργεί δέντρα απόφασης από ένα σύνολο δεδομένων που τα «εκπαιδεύουμε» χρησιμοποιώντας τη μέθοδο εντροπίας πληροφοριών. Τα δεδομένα που εκπαιδεύονται αποτελούν ένα σύνολο $S = s_1,s_2, \ldots$ Των δειγμάτων που έχουν ήδη ταξινομηθεί. Κάθε δείγμα si αποτελείται από ένα ρ-διάστατο διάνυσμα $(x_{1,i}, x_{2,I}, \ldots, x_{p,i})$, όπου ο x_j αντιπροσωπεύει τις χαρακτηριστικές αξίες ή τα χαρακτηριστικά του δείγματος, καθώς και την κατηγορία την οποία πέφτει το si.

Σε κάθε κόμβο του δέντρου ο C4.5 επιλέγει το χαρακτηριστικό των δεδομένων που χωρίζει αποτελεσματικότερα το σύνολο των δειγμάτων σε διάφορα υποσύνολα. Το κριτήριο διαχωρισμού είναι το κανονικοποιημένο κέρδος πληροφοριών. Το χαρακτηριστικό λοιπόν που αποφέρει το υψηλότερο όφελος πληροφοριών θα επιλεχτεί για να πάρει την απόφαση. Στη συνέχεια ο αλγόριθμος επαναλμβάνεται στα μικρότερα υποσύνολα.

Ο συγκεκριμένος αλγόριθμος έχει μερικές βασικές περιπτώσεις:

- Όλα τα δείγματα του πίνακα/λίστας ανήκουν στην ίδια κατηγορία. Όταν και αν συμβεί αυτό, δημιουργεί έναν κόμβο για το δέντρο απόφασης λέγοντάς του να επιλέξει αυτήν την κατηγορία.
- Κανένα από τα χαρακτηριστικά δεν περιέχει κάποια ωφέλιμη πληροφορία. Σε αυτήν την περίπτωση ο αλγόριθμος δημιουργεί έναν κόμβο υψηλότερα από το δέντρο απόφασης χρησιμοποιώντας την αναμενόμενη τιμή της κατηγορίας.

- Αντιμετώπιση προηγούμενης αθέατης/κρυφής κατηγορίας. Θα πράξει όπως παραπάνω.

Πρακτικά ο αλγόριθμος μπορεί να υλοποιηθεί μέσω του προγράμματος εξόρυξης δεδομένων Weka με την επιλογή/όνομα J48.

3.2.3 Μηχανές Διανυσμάτων υποστήριξης (Support Vector Machines - SVM)

Μια μηχανή διανυσμάτων υποστήριξης δημιουργεί ένα υπερεπίπεδο ή το σύνολο των υπερπιπέδων σε έναν μεγάλων/άπειρων διαστάσεων χώρο, που μπορεί να χρησιμοποιηθεί για την ταξινόμηση, οπισθοδρόμηση ή άλλες εργασίες. Διαισθητικά, καταλαβαίνεις πως έχει επιτευχθεί καλός διαχωρισμός από το υπερεπίπεδο που έχει τη μεγαλύτερη απόσταση από το πλησιέστερο σημείο εκπαιδευόμενων δεδομένων οποιασδήποτε κλάσης , γνωστά ως functional margin, δεδομένου ότι σε γενικές γραμμές όσο μεγαλύτερο είναι το περιθώριο τόσο χαμηλότερη είναι η γενίκευση λάθους του ταξινομητή.

Στη Μηχανική Μάθηση, οι Μηχανές Διανυσμάτων Υποστήριξης –Support Vector Machines- εποπτευόμενα μοντέλα μάθησης που συνδέονται με αλγόριθμους μάθησης που αναλύουν δεδομένα που χρησιμοποιούνται για Κατηγοριοποίηση και Ανάλυση Παλινδρόμησης. Λαμβάνοντας υπόψη ένα σύνολο από παραδείγματα εκπαίδευσης, που το καθένα σημειώνεται να ανήκει σε μια από τις δυο κατηγορίες, ένας αλγόριθμος εκπαίδευσης SVM χτίζει ένα μοντέλο που εκχωρεί τα νέα παραδείγματα στη μία κατηγορία ή στην άλλη.καθιστώντας τον ως απίθανο δυαδικό γραμμικό ταξινομητή. Ένα μοντέλο SVM είναι μια αναπαράσταση των παραδειγμάτων σαν σημεία στο χώρο, χαρτογραφημένα έτσι ώστε τα παραδείγματα διαφορετικών κατηγοριών να διαχωρίζονται από μια σαφής απόσταση. Τα νέα παραδείγματα εισάγονται στον ίδιο χώρο και ανήκουν από τη μεριά της κατηγορίας της οποίας έχουν εισαχθεί.

Εκτός από την εκτέλεση γραμμικής ταξινόμησης, οι SVM μπορούν να εκτελέσουν αποτελεσματικά μη-γραμμική κατηγοριοποίηση χρησιμοποιώντας κάτι που είναι

γνωστό ως kernel trick, έμμεση χαρτογράφηση των δεδομένων εισόδου σε χαρακτηριστικούς χώρους πολλών διαστάσεων.

3.3 Μέτρα αξιολόγησης (True Positives, True Negatives …. Confusion Matrix)

3.3.1 True Positives & True Negatives

Η sensitivity(ευαισθησία) και η specificity(εξειδίκευση) είναι στατιστικά μέτρα της απόδοσης ενός δυαδικού classification test, γνωστό επίσης ως συνάρτηση ταξινόμησης:

Ευασθησία (γνωστη επίσης σαν true positive rate): μετρά το ποσοστό των θετικών που έχουν αναγνωριστεί σωστά

$$\begin{aligned}\text{sensitivity} &= \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}} \\ &= \frac{\text{number of true positives}}{\text{total number of sick individuals in population}} \\ &= \text{probability of a positive test given that the patient has the disease}\end{aligned}$$

Εξειδίκευση (γνωστή επίσης σαν true negative rate): μετρά το ποσοστό των αρνητικών που έχουν αναγνωριστεί σωστά

$$\begin{aligned}\text{specificity} &= \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}} \\ &= \frac{\text{number of true negatives}}{\text{total number of well individuals in population}} \\ &= \text{probability of a negative test given that the patient is well}\end{aligned}$$

Για να γίνουν καλύτερα κατανοητές οι παραπάνω έννοιες θα δωθεί το παρακάτω παράδειγμα. Έστω ότι διεξάγεται κάποια εξέταση για να δούμε αν κάποιος άνθρωπος έχει μια ασθένεια. Το αποτέλεσμα της εξέτασης μπορεί να είναι θετικό (που ταξινομεί

το πρόσωπο ως έχοντα την ασθένεια) ή αρνητικό(που ταξινομεί το πρόσωπο ως μη έχοντα την ασθένεια). Τα αποτελέσματα των εξετάσεων του κάθε προσώπου μπορεί να μην ταιριάζουν με την πραγματική του κατάσταση.

- Αληθώς θετικό: Άρρωστοι άνθρωποι που αναγνωρίστηκαν σωστά ως άρρωστοι
- Ψευδώς θετικό: Υγιείς άνθρωποι που λανθασμένα αναγνωρίστηκαν ως άρρωστοι
- Αληθώς αρνητικό: Υγιείς άνθρωποι που αναγνωρίστηκαν σωστά ως υγιείς
- Ψευδώς αρνητικό: Άρρωστοι άνθρωποι που λανθασμένα αναγνωρίστηκαν ως υγιείς

3.3.2 Confusion Matrix

Στον τομέα της μηχανικής μάθησης(machine learning) confusion matrix (πίνακας σύγχυσης) - επίσης γνωστό και ως error matrix - είναι ένας ειδικός πίνακας που απεικονίζει την απόδοση ενός αλγόριθμου (εποπτευόμενης μάθησης). Κάθε στήλη του πίνακα αντιπροσωπεύει τις περιπτώσεις σε μια προβλεπόμενη κατηγορία ενώ κάθε σειρά τις περιπτώσεις σε μια πραγματική κατηγορία (ή αντίστροφα). Το όνομα προέρχεται από το γεγονός ότι είναι εύκολο να δούμε αν το σύστημα συγχέει τη μια κατηγορία με την άλλη.

Για να γίνει καλύτερα κατανοητό, ας δώσουμε ένα παράδειγμα. Ένας αλγόριθμος/σ'υστημα ταξινόμησης έχει εκπαιδευτεί στο να διακρίνει δελφίνια, καρχαρίες και φάλαινες. Σε ένα πίνακα σύγχυσης θα καταχωρηθούν τα αποτελέσματα του αλγόριθμου για επιθεώρηση. Το δείγμα μας αποτελείται από 20 ζώα (9 δελφίνια, 3 καρχαρίες και 8 φάλαινες).

		Προβλεπόμενη		
		Δελφίνι	Καρχαρίας	Φάλαινα
Πραγματική τάξη	Δελφίνι	6	2	0
	Καρχαρίας	1	2	0
	Φάλαινα	0	1	7

Ο παραπάνω πίνακας μας δείχνει πως από

- τα 9 πραγματικά δελφίνια, τα 2 προέβλεψε πως είναι καρχαρίες
- τους 3 πραγματικούς καρχαρίες, προέβλεψε πως το ένα ήταν δελφίνι
- τις 8 πραγματικές φάλαινες, προέβλεψε πως ο ένας είναι καρχαρίας

Καταλαβαίνουμε λοιπόν πως υπάρχει πρόβλημα στο να διακρίνει τα δελφίνια από τους καρχαρίες, αλλά μπορεί να κάνει τη διάκριση μεταξύ φαλαινών σε πολύ καλύτερα επίπεδα. Οι σωστές προβλέψεις με τις σωστές τιμές βρίσκονται στη διαγώνιο του πίνακα και όλες οι λανθασμένες εκτός αυτής.

Στην **προγνωστική ανάλυση (predictive analytics)**, πίνακας σύγχυσης είναι ένας πίνακας που αποτελείται από δύο σειρές και δύο στήλες όπου εκεί αναφέρεται ο αριθμός των αληθώς θετικών, ψευδώς θετικών, αληθώς αρνητικών, ψευδώς αρνητικών. Κατά αυτόν τον τρόπο έχουμε μια πιο λεπτομερή ανάλυση από την απλή αναλογία σωστών προβλέψεων. Για παράδειγμα, έστω ότι είχαμε 94 δελφίνια και 6 καρχαρίες στο σύνολο δεδομένων. Η συνολική ακρίβεια θα ήταν 94% αλλά ο ταξινομητής θα είχε ποσοστό 100% αναγνώρισης για την κατηγορία των δελφινιών και ποστό 0% για την αναγνώριση των καρχαριών.

Για το παραπάνω παράδειγμα προκύπτει ο πίνακας σύγχυσης για τα δελφίνια ως έχει παρακάτω:

6 αληθώς θετικά (πραγματικά δελφίνια που είχαν αναγνωρισθεί σωστά ως δελφίνια	2 ψευδώς αρνητικά (δελφίνια που είχαν αναγνωρισθεί ως καρχαρίες)
1 ψευδώς θετικά (καρχαρίας που λανθασμένα αναγνωρίστηκε ως δελφίνι)	10 αληθώς αρνητικά (όλα τα υπόλοιπα ζώα, σωστά αναγνωρισμένα ως μη-δελφίνια)

Κεφάλαιο 4

Αποτελέσματα

Κεφάλαιο 4

4.1 Σύνολο δεδομένων

Τα δεδομένα τα οποία χρησιμοποιηθήκαν για τους σκοπούς των πειραμάτων μας ήταν δεδομένα από αντλήθηκαν από ιστορικών δεικτών, για 2 όλοκληρα χρόνια, και με δειγματοληψία 4 ωρών. Επομένως υπήρχαν 6 τιμές των ισοτιμιών ανά ημέρα. Για κάθε τέτοια χρονική στιγμή δειγματολειψίας υπάρχουν οι τιμές το ανοίγματος του 4ώρου του κλεισίματος του 4ώρου, ο όγκος των συναλλαγών, η μέγιστη και η ελάχιστη τιμή του 4ώρου, και η μέση τιμή του δείκτη για αυτό 4ωρο. Όλες αυτές οι τιμές χρησιμοποιήθηκαν για χαρακτηριστικά στο πρόβλημα ταξηνόμησης.

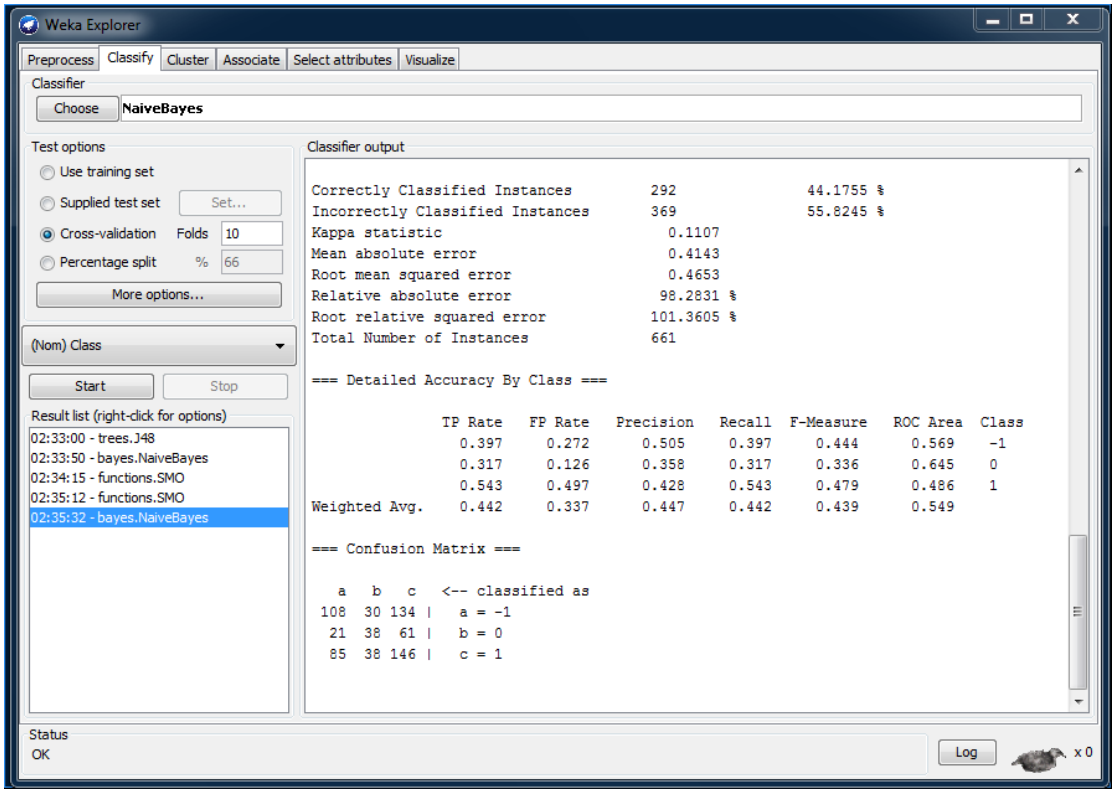
Για να συγκεντρωθούν δεδομένα για το πρόβλημα ταξηνόμησης που έχει οριστεί, αρκεί να λαμβάνουμε χαρακτηριστικά οποιαδήποτε χρονική στιγμή, και να γνωρίζουμε στο επόμενο 4ωρο πως τελικά κινήθηκε ο δείκτης ώστε να του δώσουμε την κατάλληλη επισημίωση κατηγορίας (-1, 0, 1) όπως ακτιβώς συζητήσαμε στο προηγούμενο κεφάλαιο. Έτσι επιλέξαμε να λαμβάνουμε δειγματοληπτικα μια τιμή την ημερα στις 12:00 το μεσημέρι και να χαρακτηρίζουμε την κλάση μας από την τιμή του δείκτη στις 16:00 το απόγευμα. Με αυτόν τον τρόπο συγκεντρώσαμε ένα σύνολο 661 δειγματων για το πρόβλημα μας.

Παρακάτω παραθέτουμε τα αποτελέσματα των 18 συολικά πειραμάτων όπως αυτά προέκυψαν από την επεξεργασία στο πρόγραμμα εξόρυξης δεδομένων Weka.

4.2 Περιγραφή πειραμάτων

4.2.1 Πρόβλημα τριών κλάσεων (3-class problem)

1) Πείραμα με Μπευζιανό ταξινομητή και τιμή κατωφλιού για την μεταβολή του δείκτη στο 4ωρο 0.3‰



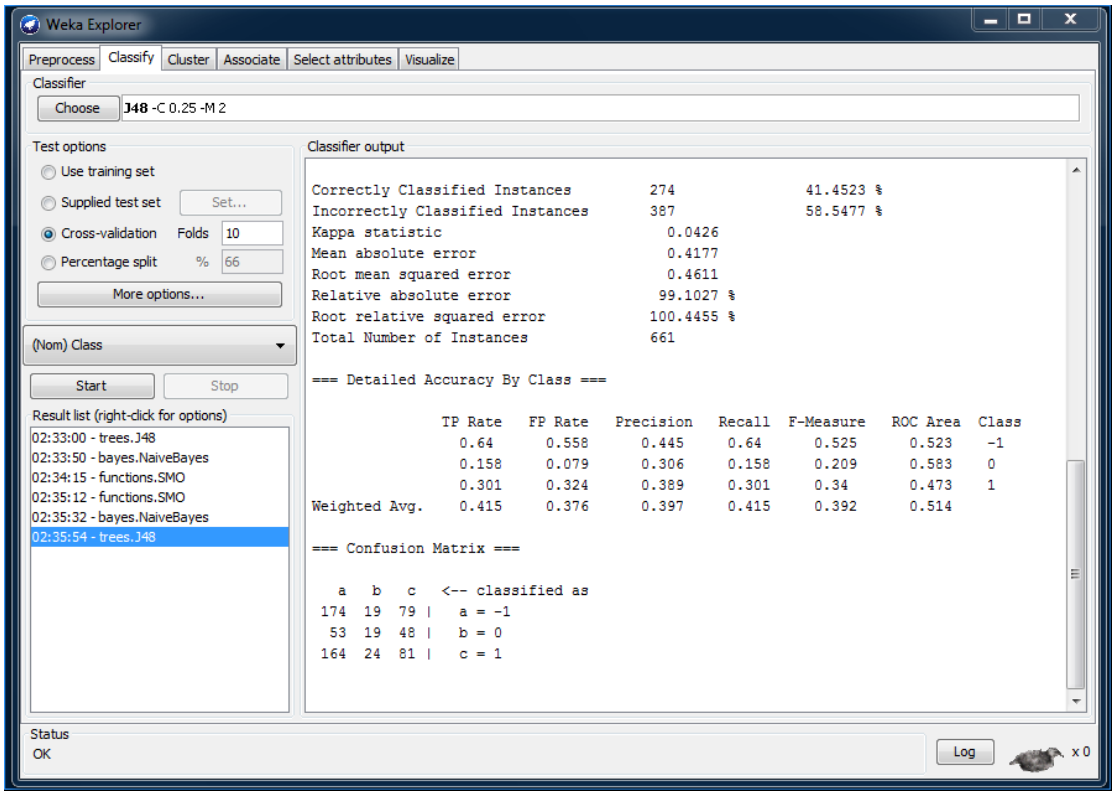
Σωστά κατηγοριοποιημένες περιπτώσεις → 292 (44.1755%)

Λανθασμένα κατηγοριοποιημένες περιπτώσεις → 369 (55.6245%)

Σύνολο περιπτώσεων → 661 (100%)

Πίνακας Σύγχυσης – Confusion Matrix		
A(-1)	B(0)	C(1)
108	30	134
21	38	61
85	38	146

2) Πείραμα με δέντρο απόφασης και τιμή κατωφλιού για την μεταβολή του δείκτη στο 4ωρο 0.3‰



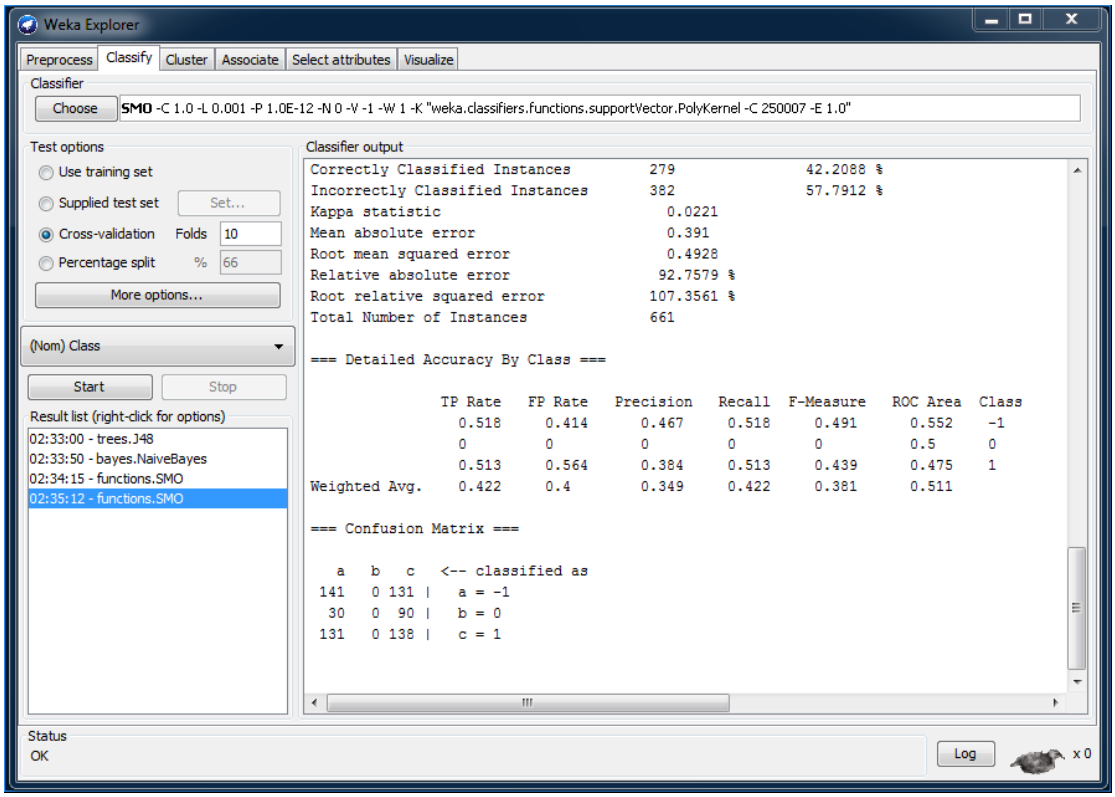
Σωστά κατηγοριοποιημένες περιπτώσεις → 274 (41.4523%)

Λανθασμένα κατηγοριοποιημένες περιπτώσεις → 387 (58.5477%)

Σύνολο περιπτώσεων → 661 (100%)

Πίνακας Σύγχυσης – Confusion Matrix		
A(-1)	B(0)	C(1)
174	19	79
53	19	48
164	24	81

3) Πείραμα με Μηχανές Διανυσμάτων Υποστήριξης και τιμή κατωφλιού για την μεταβολή του δείκτη στο 4ωρο 0.3‰



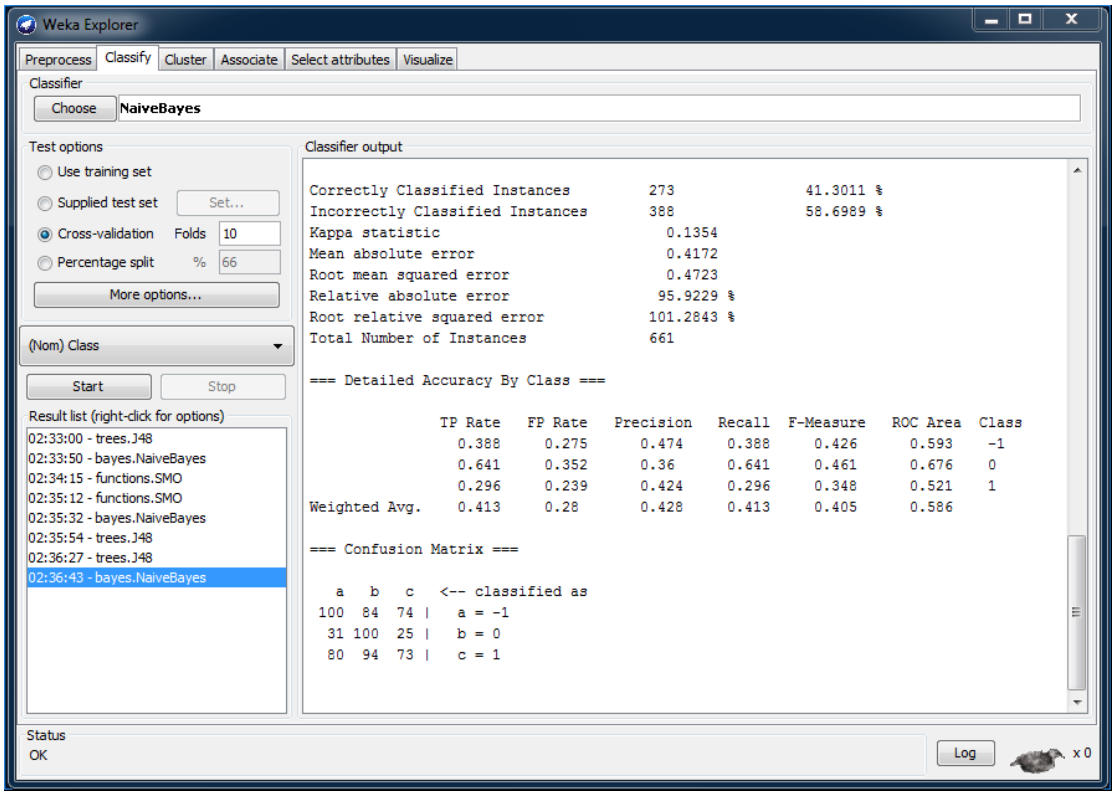
Σωστά κατηγοριοποιημένες περιπτώσεις → 279 (42.2088%)

Λανθασμένα κατηγοριοποιημένες περιπτώσεις → 382 (57.7912%)

Σύνολο περιπτώσεων → 661 (100%)

Πίνακας Σύγχυσης – Confusion Matrix		
A(-1)	B(0)	C(1)
141	0	131
30	0	90
131	0	138

4) Πείραμα με Μπευζιανό ταξινομητή και τιμή κατωφλιού για την μεταβολή του δείκτη στο 4ωρο 0.4‰



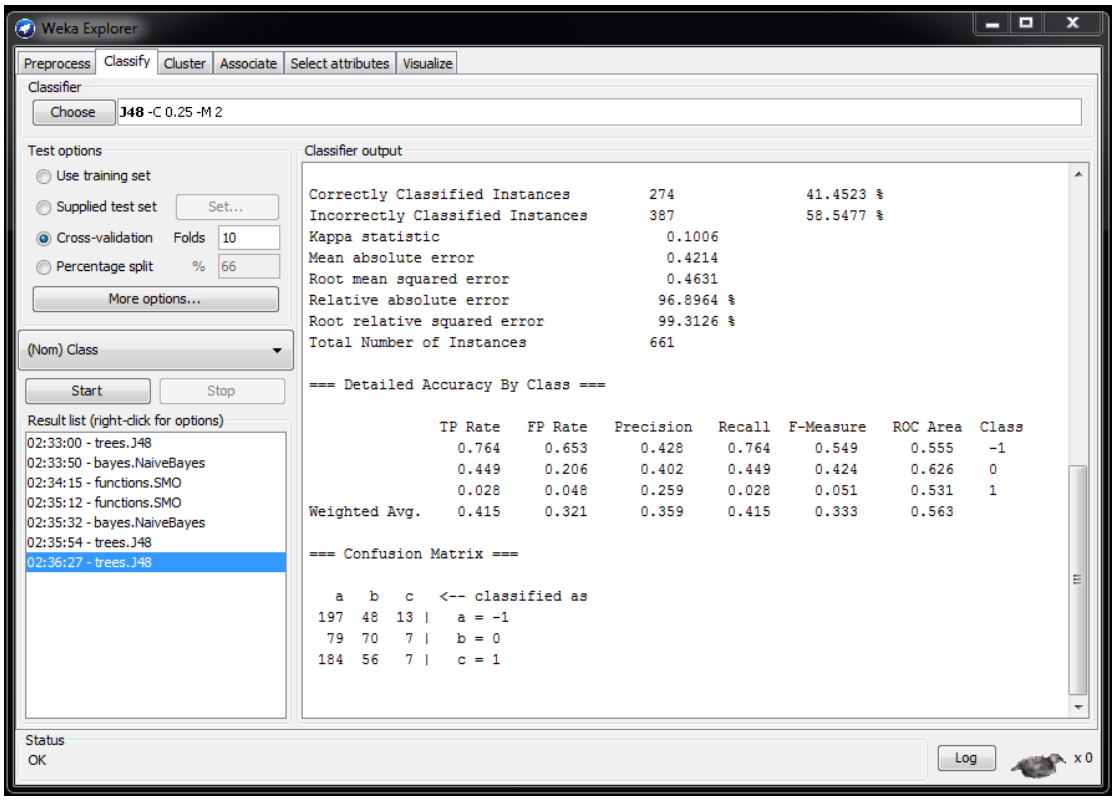
Σωστά κατηγοριοποιημένες περιπτώσεις → 273 (41.3011%)

Λανθασμένα κατηγοριοποιημένες περιπτώσεις → 388 (58.6989%)

Σύνολο περιπτώσεων → 661 (100%)

Πίνακας Σύγχυσης – Confusion Matrix		
A(-1)	B(0)	C(1)
100	84	74
31	100	25
80	94	73

5) Πείραμα με Δένδρο απόφασης και τιμή κατωφλιού για την μεταβολή του δείκτη στο 4ωρο 0.4‰



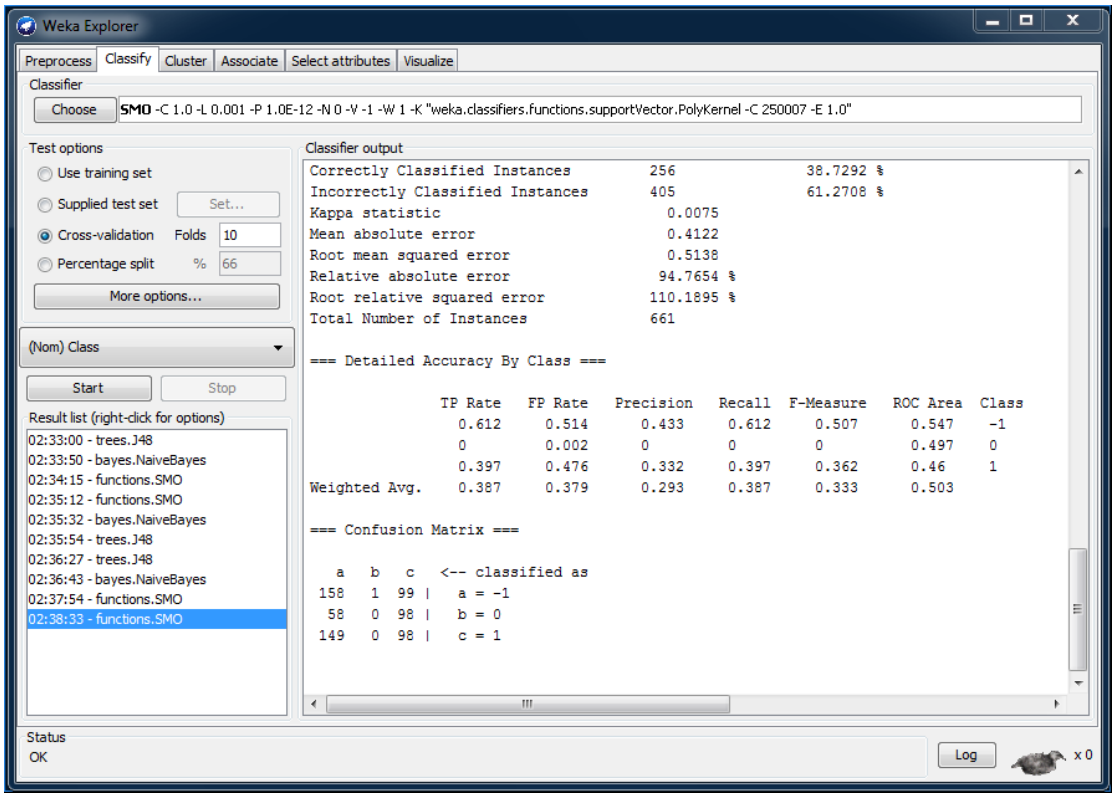
Σωστά κατηγοριοποιημένες περιπτώσεις → 274 (41.4523%)

Λανθασμένα κατηγοριοποιημένες περιπτώσεις → 387 (58.5477%)

Σύνολο περιπτώσεων → 661 (100%)

Πίνακας Σύγχυσης – Confusion Matrix		
A(-1)	B(0)	C(1)
197	48	13
79	70	7
184	56	7

6) Πείραμα με Μηχανές Διανυσμάτων Υποστήριξης και τιμή κατωφλιού για την μεταβολή του δείκτη στο 4ωρο 0.4‰



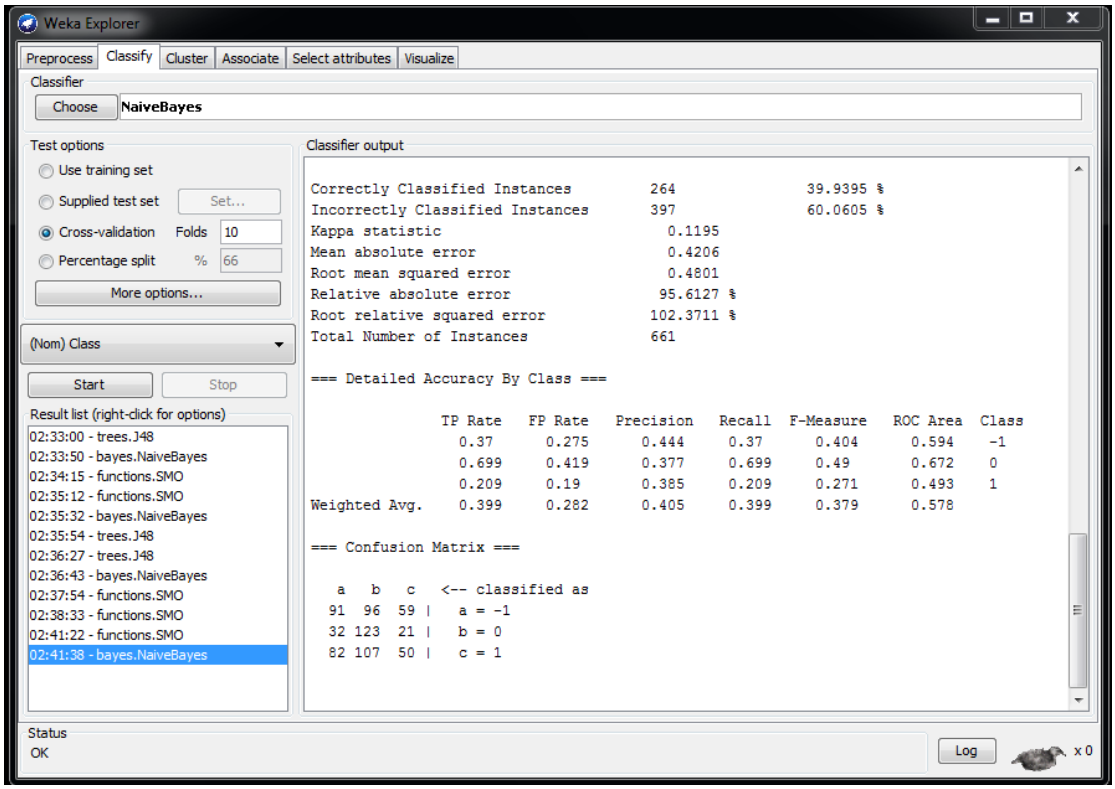
Σωστά κατηγοριοποιημένες περιπτώσεις → 256 (38.7292%)

Λανθασμένα κατηγοριοποιημένες περιπτώσεις → 405 (61.2708%)

Σύνολο περιπτώσεων → 661 (100%)

Πίνακας Σύγχυσης – Confusion Matrix		
A(-1)	B(0)	C(1)
158	1	99
58	0	98
149	0	98

7) Πείραμα με Μπευζιανό Ταξινομητή και τιμή κατωφλιού για την μεταβολή του δείκτη στο 4ωρο 0.5‰



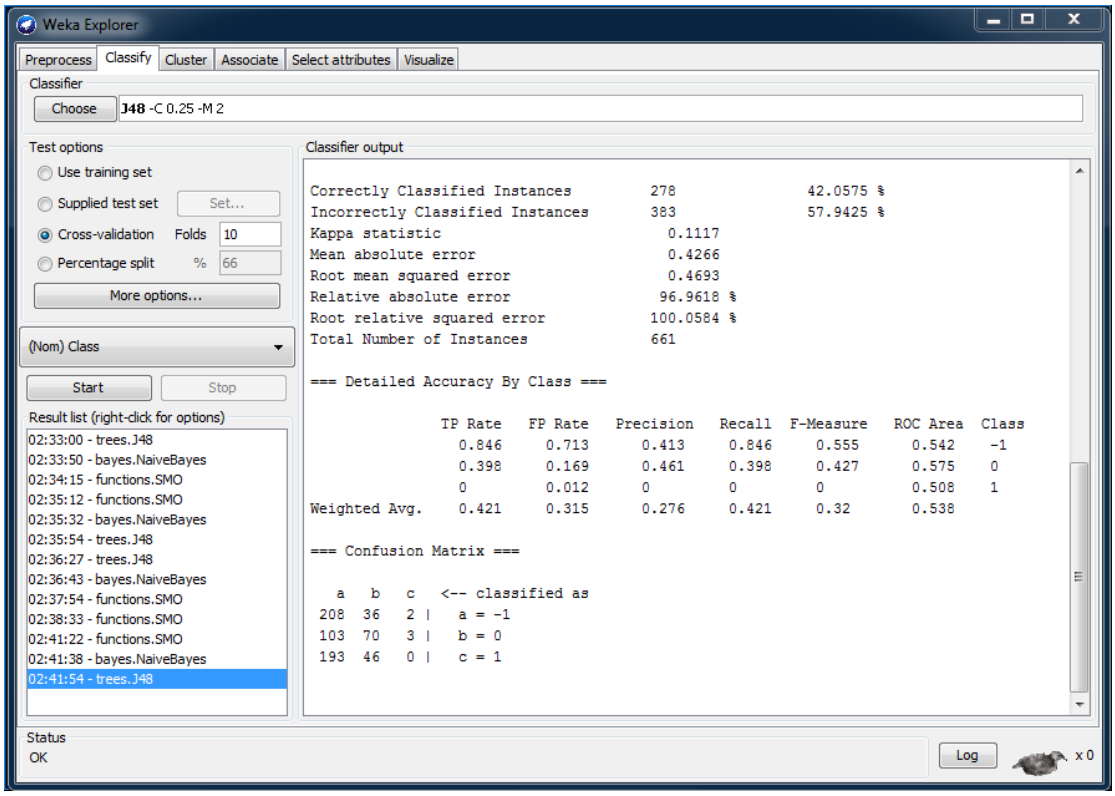
Σωστά κατηγοριοποιημένες περιπτώσεις → 264 (39.9395%)

Λανθασμένα κατηγοριοποιημένες περιπτώσεις → 397 (60.0605%)

Σύνολο περιπτώσεων → 661 (100%)

Πίνακας Σύγχυσης – Confusion Matrix		
A(-1)	B(0)	C(1)
91	96	59
32	123	21
82	107	50

8) Πείραμα με Δένδρο απόφασης και τιμή κατωφλιού για την μεταβολή του δείκτη στο 4ωρο 0.5‰



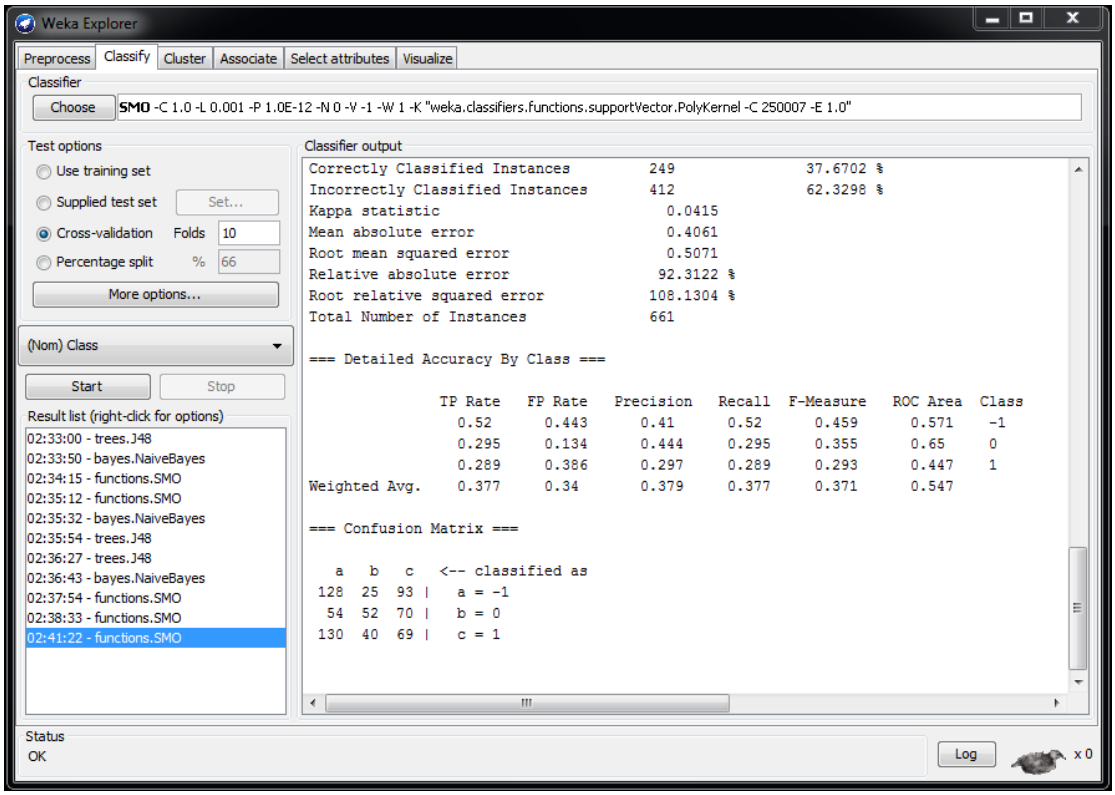
Σωστά κατηγοριοποιημένες περιπτώσεις → 278 (42.0575%)

Λανθασμένα κατηγοριοποιημένες περιπτώσεις → 383 (57.9425%)

Σύνολο περιπτώσεων → 661 (100%)

Πίνακας Σύγχυσης – Confusion Matrix		
A(-1)	B(0)	C(1)
208	36	2
103	70	3
193	46	0

9) Πείραμα με Μηχανές Διανυσματων Υποστήριξης και τιμή κατωφλιού για την μεταβολή του δείκτη στο 4ωρο 0.5‰



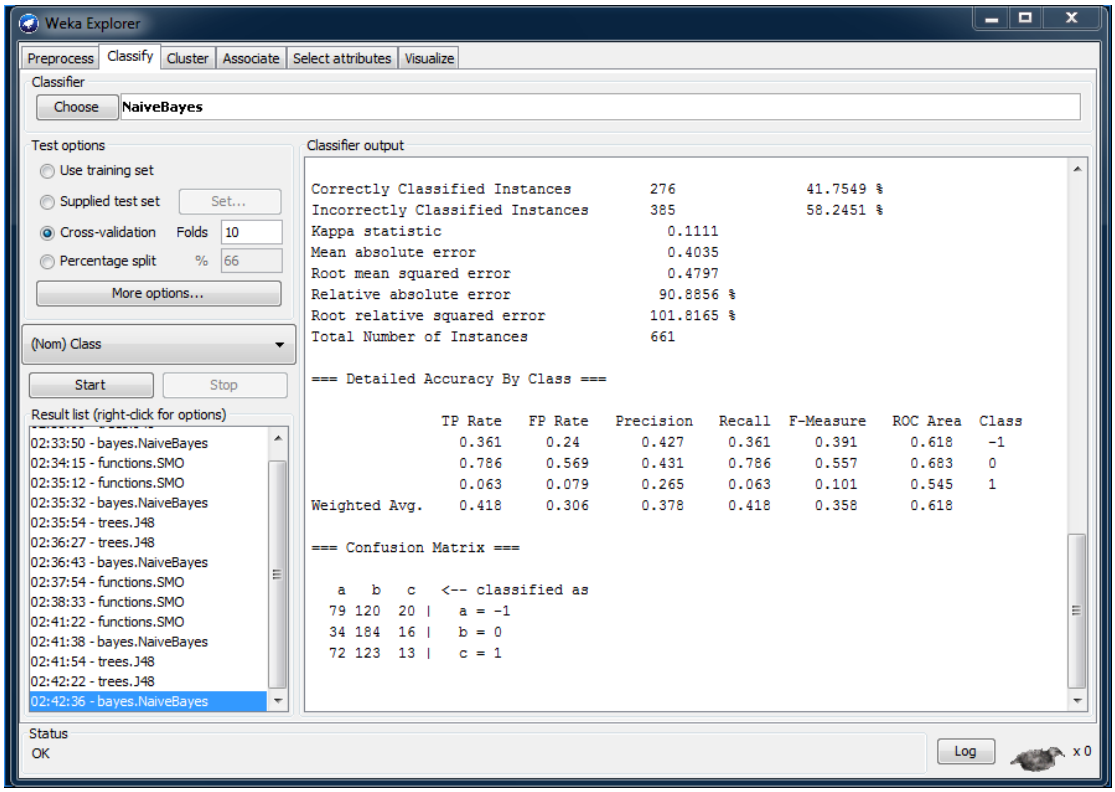
Σωστά κατηγοριοποιημένες περιπτώσεις → 249 (37.6702%)

Λανθασμένα κατηγοριοποιημένες περιπτώσεις → 412 (62.3298%)

Σύνολο περιπτώσεων → 661 (100%)

Πίνακας Σύγχυσης – Confusion Matrix		
A(-1)	B(0)	C(1)
128	25	93
54	52	70
130	40	69

10) Πείραμα με Μπευζιανός Ταξινομητής και τιμή κατωφλιού για την μεταβολή του δείκτη στο 4ωρο 0.7‰



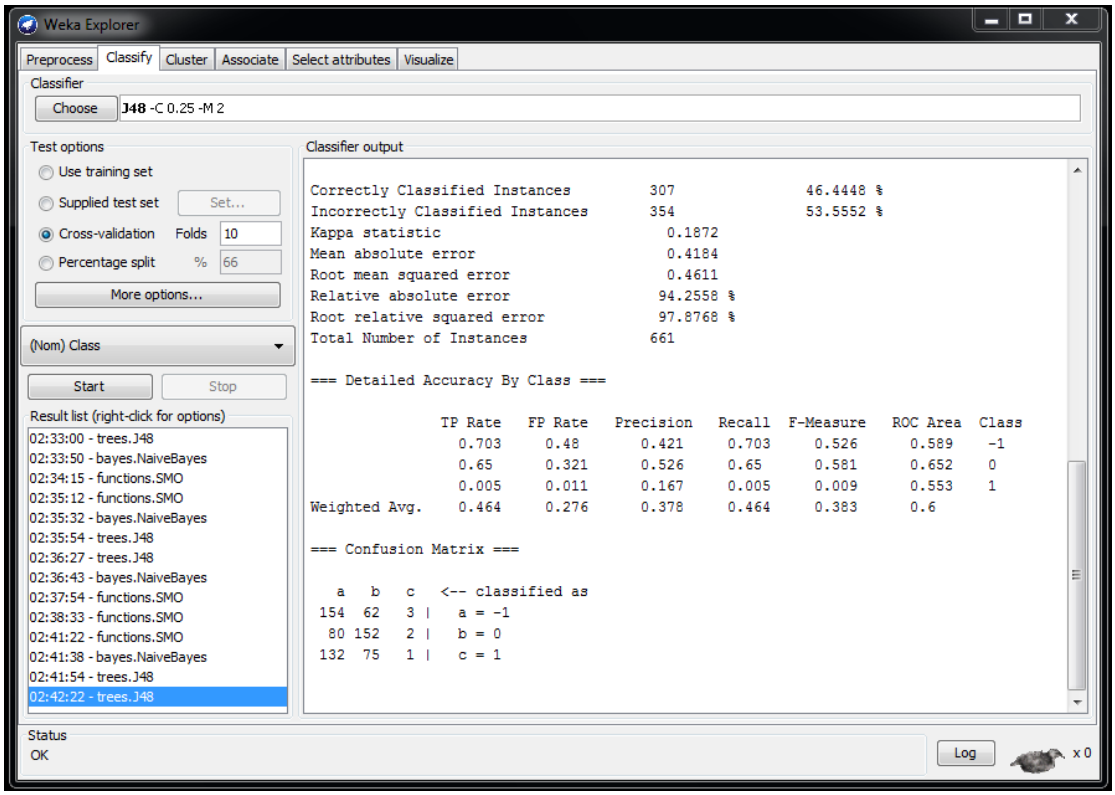
Σωστά κατηγοριοποιημένες περιπτώσεις → 276 (41.7549%)

Λανθασμένα κατηγοριοποιημένες περιπτώσεις → 385 (58.2451%)

Σύνολο περιπτώσεων → 661 (100%)

Πίνακας Σύγχυσης – Confusion Matrix		
A(-1)	B(0)	C(1)
79	120	20
34	184	16
72	123	13

11) Πείραμα με Δένδρο απόφασης και τιμή κατωφλιού για την μεταβολή του δείκτη στο 4ωρο 0.7‰



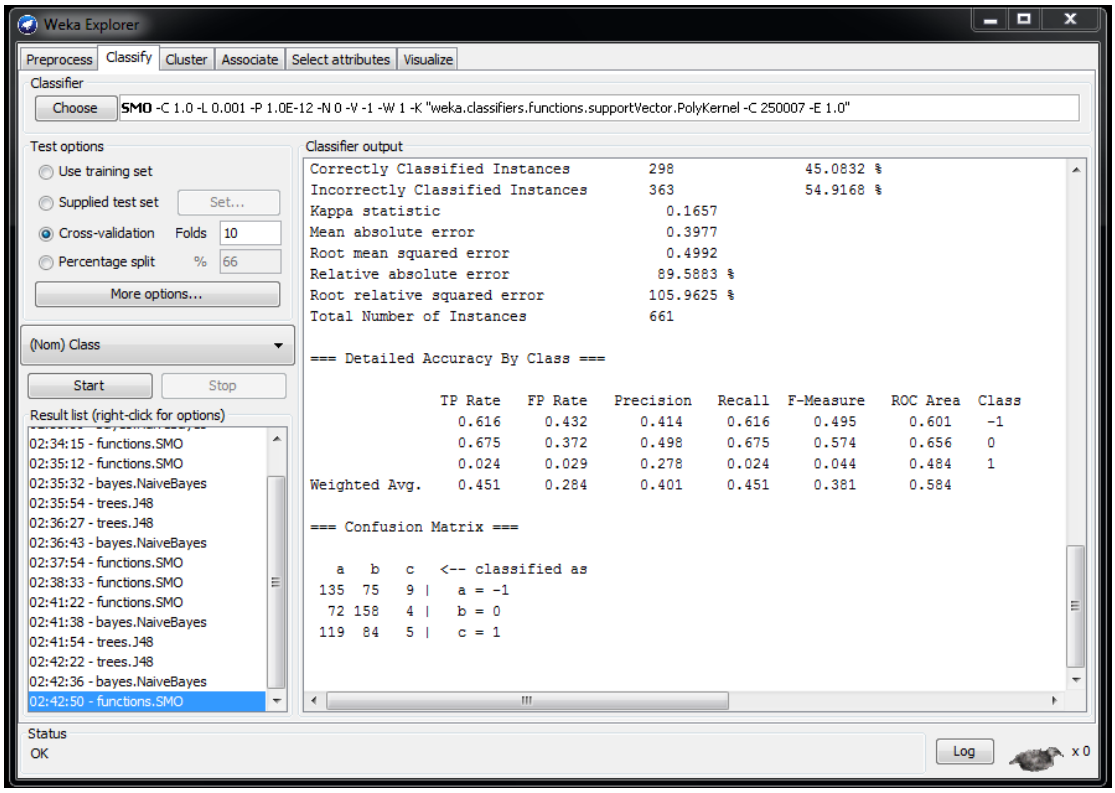
Σωστά κατηγοριοποιημένες περιπτώσεις → 307 (46.448%)

Λανθασμένα κατηγοριοποιημένες περιπτώσεις → 354 (53.5552%)

Σύνολο περιπτώσεων → 661 (100%)

Πίνακας Σύγχυσης – Confusion Matrix		
A(-1)	B(0)	C(1)
154	62	3
80	152	2
132	75	1

12) Πείραμα με Μηχανές Διανυσμάτων Υποστήριξης και τιμή κατωφλιού για την μεταβολή του δείκτη στο 4ωρο 0.7‰



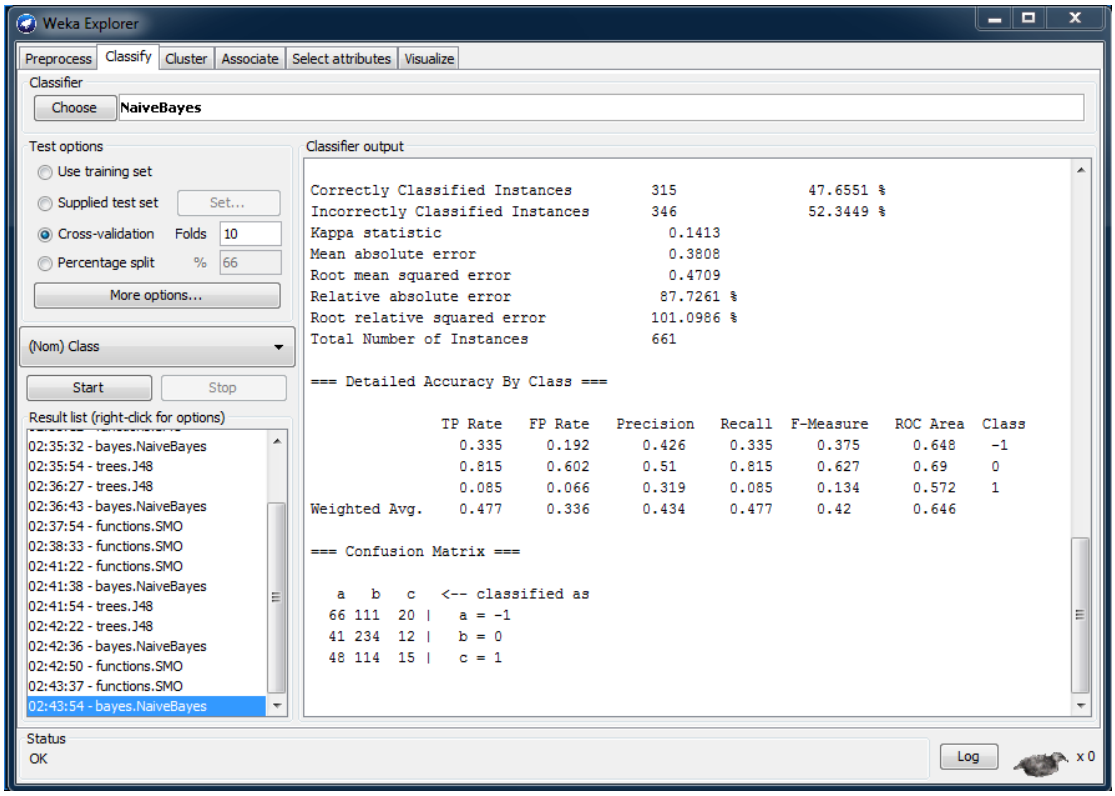
Σωστά κατηγοριοποιημένες περιπτώσεις → 298 (45.0832%)

Λανθασμένα κατηγοριοποιημένες περιπτώσεις → 363 (54.9168%)

Σύνολο περιπτώσεων → 661 (100%)

Πίνακας Σύγχυσης – Confusion Matrix		
A(-1)	B(0)	C(1)
135	75	9
72	158	4
119	84	5

13) Πείραμα με Μπευζιανός Ταξινομητής και τιμή κατωφλιού για την μεταβολή του δείκτη στο 4ωρο 0.9‰



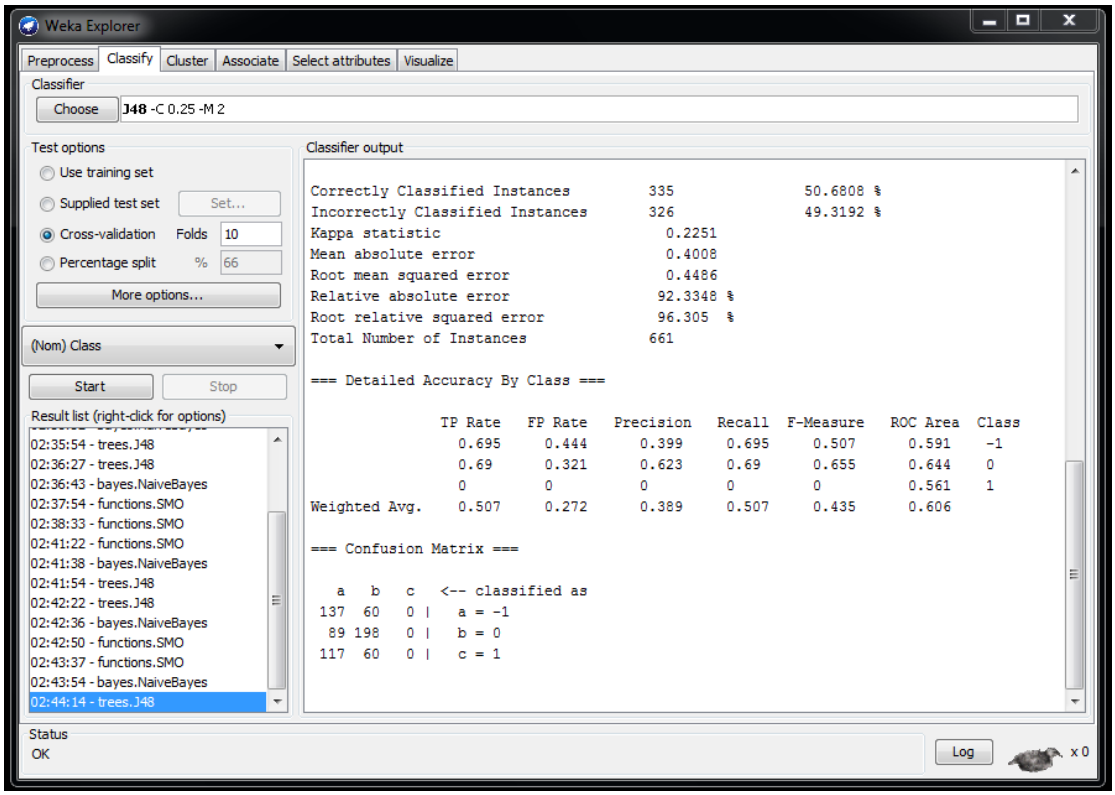
Σωστά κατηγοριοποιημένες περιπτώσεις → 315 (47.6551%)

Λανθασμένα κατηγοριοποιημένες περιπτώσεις → 346 (52.3449%)

Σύνολο περιπτώσεων → 661 (100%)

Πίνακας Σύγχυσης – Confusion Matrix		
A(-1)	B(0)	C(1)
66	111	20
41	234	12
48	114	15

14) Πείραμα με Δένδρο απόφασης και τιμή κατωφλιού για την μεταβολή του δείκτη στο 4ωρο 0.9‰



Σωστά κατηγοριοποιημένες περιπτώσεις → 335 (50.6808%)

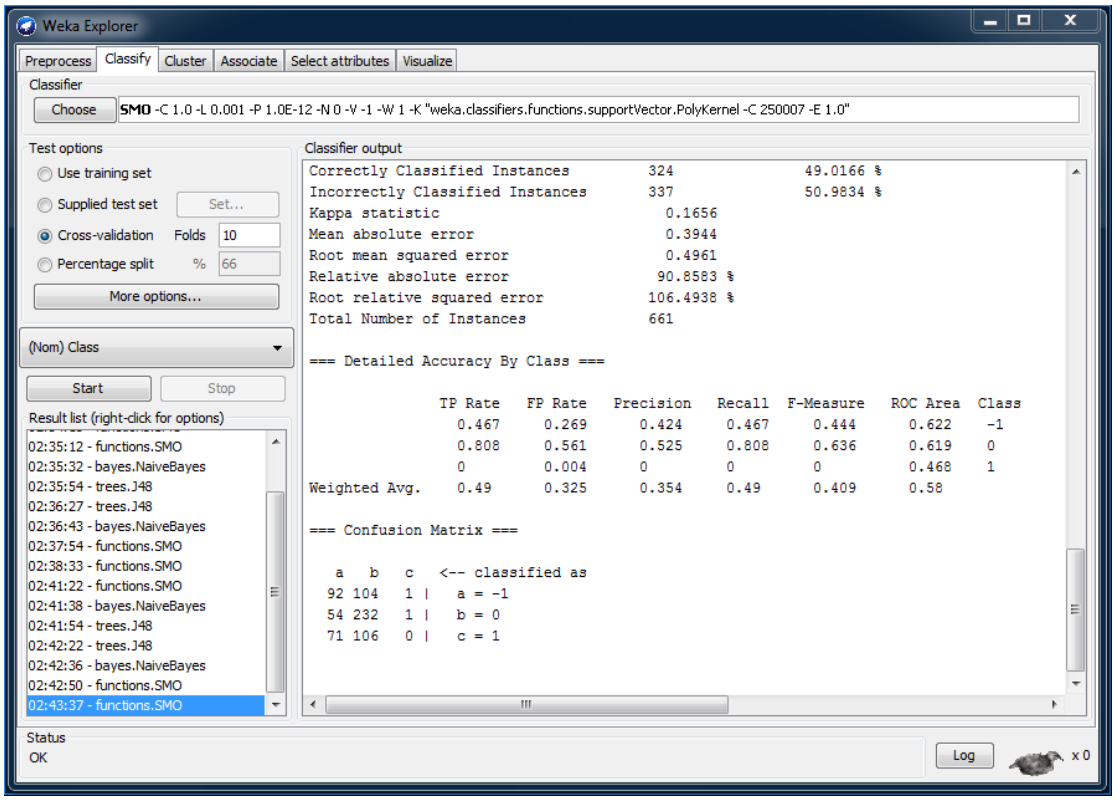
Λανθασμένα κατηγοριοποιημένες περιπτώσεις → 326 (49.3192%)

Σύνολο περιπτώσεων → 661 (100%)

Πίνακας Σύγχυσης – Confusion Matrix		
A(-1)	B(0)	C(1)
137	60	0
89	198	0
117	60	0

15) Πείραμα με Μηχανές Διανυσμάτων Υποστήριξης και τιμή κατωφλιού για την μεταβολή του δείκτη στο 4ωρο 0.9‰

iii) SVM



Σωστά κατηγοριοποιημένες περιπτώσεις → 292 (44.1755%)

Λανθασμένα κατηγοριοποιημένες περιπτώσεις → 369 (55.6245%)

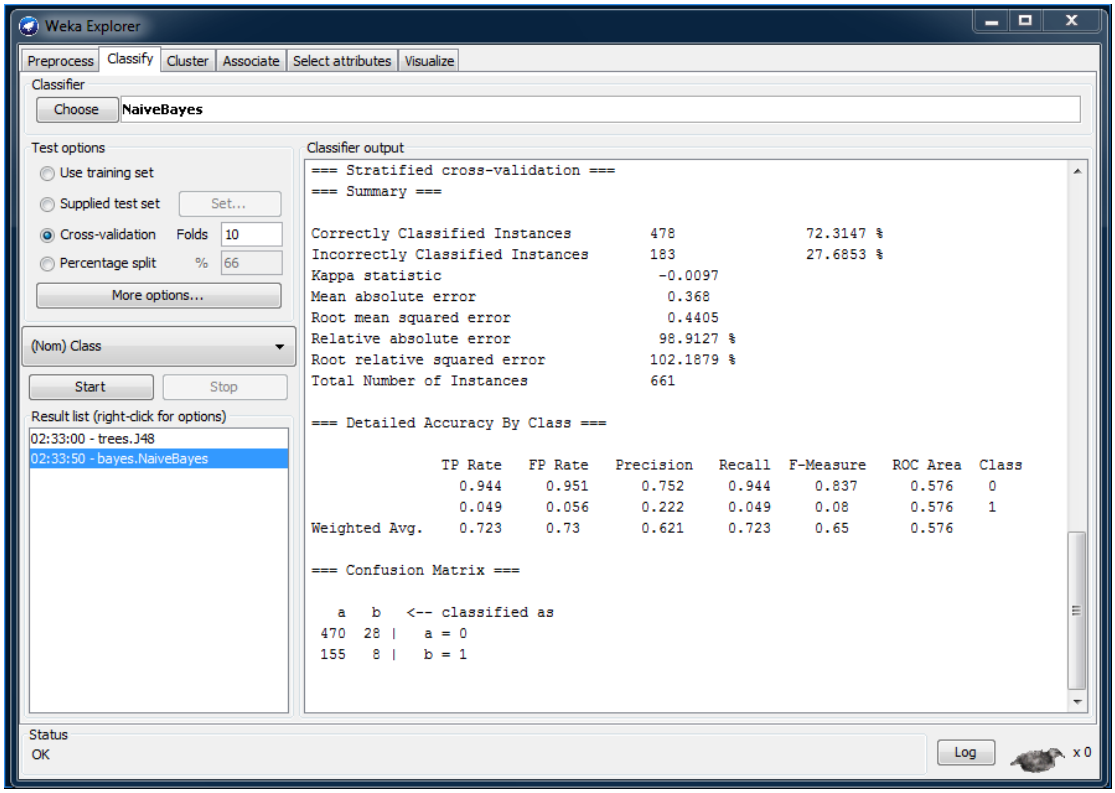
Σύνολο περιπτώσεων → 661 (100%)

Πίνακας Σύγχυσης – Confusion Matrix		
A(-1)	B(0)	C(1)
92	104	1
54	232	1
71	106	0

4.2.2 Πρόβλημα δύο κλάσεων (2-class problem)

I) Naïve Bayes

16) Πείραμα με 2 κατηγοριών με Μπευζιανο Ταξινομητή



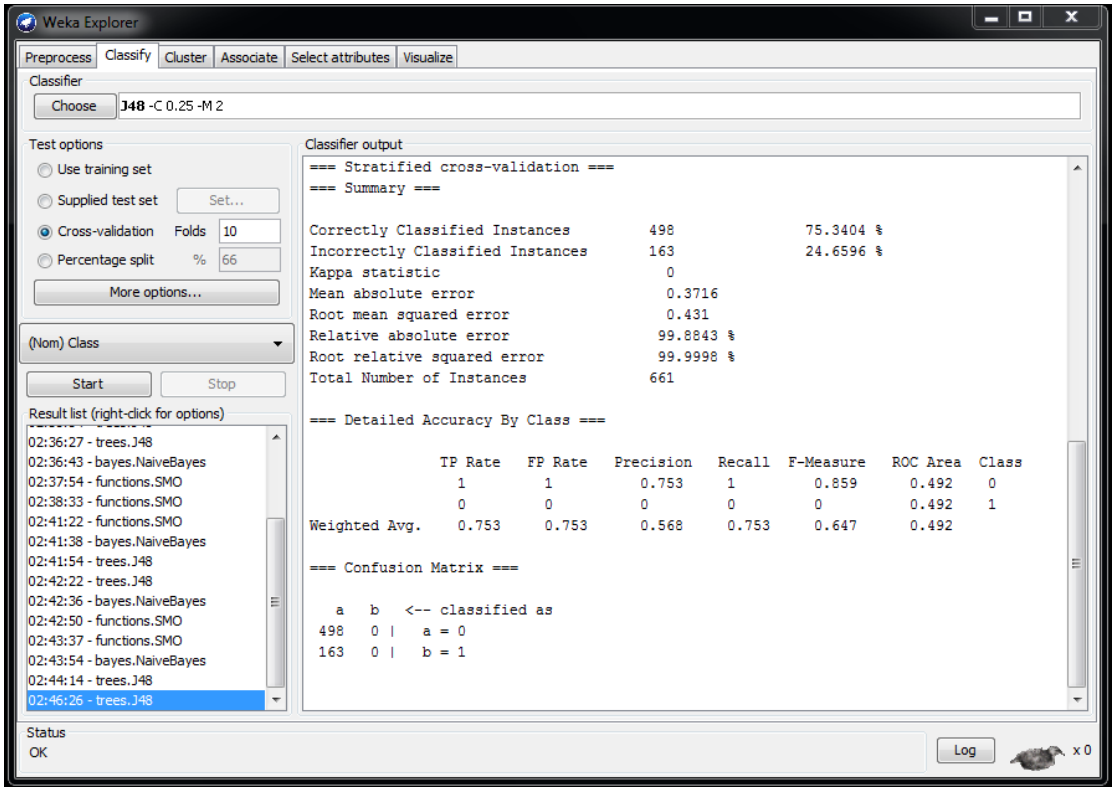
Σωστά κατηγοριοποιημένες περιπτώσεις → 478 (72.3147%)

Λανθασμένα κατηγοριοποιημένες περιπτώσεις → 183 (27.6853%)

Σύνολο περιπτώσεων → 661 (100%)

Πίνακας Σύγχυσης – Confusion Matrix	
A(0)	B(1)
470	28
155	8

17) Πείραμα με 2 κατηγοριών με Δένδρο Απόφασης



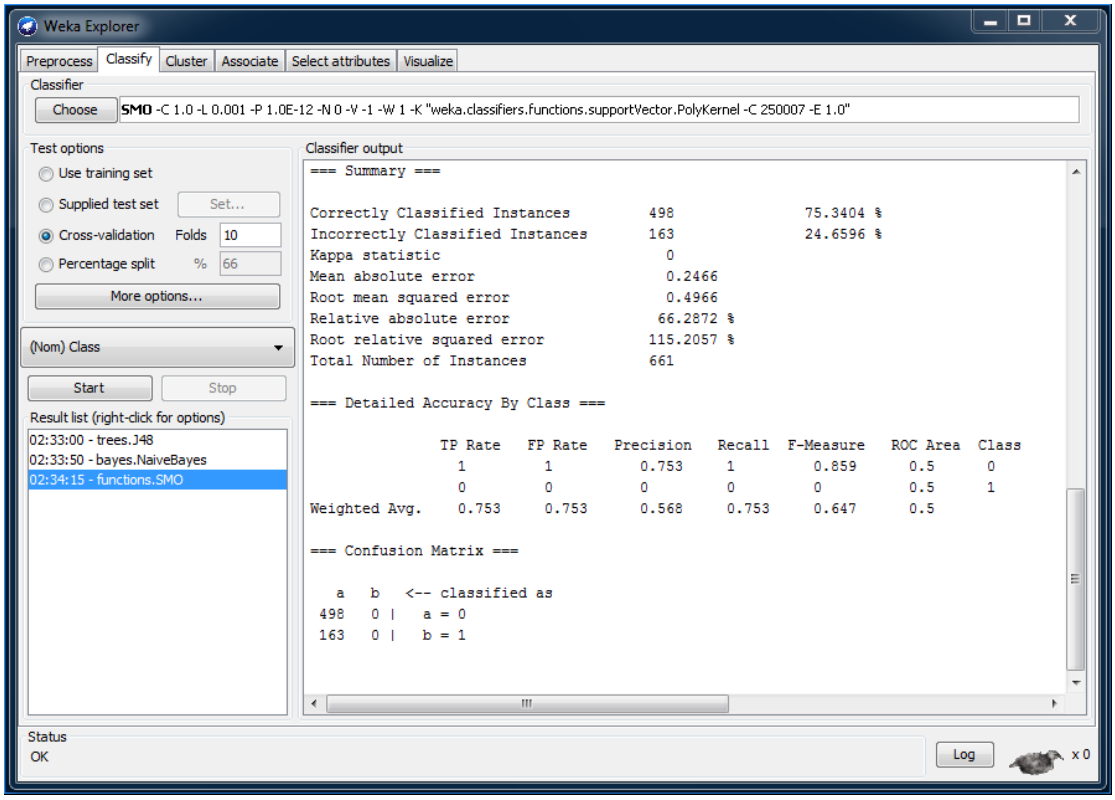
Σωστά κατηγοριοποιημένες περιπτώσεις → 498 (74.3404%)

Λανθασμένα κατηγοριοποιημένες περιπτώσεις → 163 (24.6596%)

Σύνολο περιπτώσεων → 661 (100%)

Πίνακας Σύγχυσης – Confusion Matrix	
A(0)	B(1)
498	0
163	0

18) Πείραμα με 2 κατηγοριών με Μηχανές Διανυσμάτων Υποστήριξης



Σωστά κατηγοριοποιημένες περιπτώσεις → 498 (74.3404%)

Λανθασμένα κατηγοριοποιημένες περιπτώσεις → 163 (24.6596%)

Σύνολο περιπτώσεων → 661 (100%)

Πίνακας Σύγχυσης – Confusion Matrix	
A(0)	B(1)
498	0
163	0

Τέλος παραθέτουμε τον Συγκεντρωτικό πίνακα και των 18 πειραμάτων, στον οποίο περιέχεται το ποσοστώ των σωστά ταξινομημένων δειγμάτων.

Συγκεντρωτικός πίνακας

Κατηγορία (εύρος ματαβολής του δείκτη)			**Αλγόριθμος** (Acc %)		
-1	**0**	**1**	**Naïve Bayes**	**C4.5**	**SVM**
Πτώση	**Σταθερός**	**Άνοδος**			
δ < -0.3‰	-0.3‰< δ <0.3‰	δ > 0.3‰	44.2	41.5	42.2
δ < -0.4 ‰	-0.4‰< δ <0.4‰	δ > 0.4‰	41.3	41.5	38,7
δ < -0.5‰	-0.5‰< δ <0.5‰	δ > 0.5‰	39.9	42.0	37.7
δ < -0.7‰	-0.7‰< δ <0.7‰	δ > 0.7‰	41.8	46.4	45.1
δ < -0.9‰	-0.9‰< δ <0.9‰	δ > 0.9‰	47.6	**50.7**	44.2
2-class	δ < 0‰	δ > 0‰	**72.3**	**74.3**	**74.3**

Κεφάλαιο 5

Συμπεράσματα (1-2 σελ)

Κεφάλαιο 5

Συμπεράσματα

Στην παρούσα εργασία επιχειρήθηκε ο εντοπισμός συσχετιζόμενων στο χρόνο οικονομικών δεικτών με σκοπό την πρόβλεψη της συμπεριφοράς τους βραχυπρόθεσμα. Συγκεκριμένα μελετήθηκαν 72 ισοτιμίες νομισμάτων ως προς την ομοιότητας του, δημιουργώντας έναν μεγάλο πίνακα συσχέτισης των χρονοσειρών αυτών. Αρχικά ο πίνακας συσχέτισης υπολογίστηκε για τις ίδιες ακριβώς χρονικές στιγμές ανά δύο των ισοτιμιών, ενώ στην συνέχεια ολισθήσαμε την μια εκ των δύο στον χρόνο κατά μια χρονική στιγμή΄(που αντιστοιχεί σε ένα 4ωρο) και υπολογίσαμε ξανά τον πίνακα συσχέτισης. Σκοπός ήταν να εντοπιστούν ισοτιμίες οι οποίες είτε επηρεάζει η μία την άλλη είτε αντιδρούν με χρονοκαθυστέριση σε αντίστοιχα γεγονότα. Εάν πράγματι ισχύει κάτι τέτοιο θα ήμαστε σε θέση παρατηρώντας την συμπεριφορά της μιας ισοτιμίας να προβλέψουμε την μελλοντική συμπεριφορά της άλλης.

Πράγματι εντοπίσαμε 2 ισοτιμίες οι οποίες αυξάνουν την συσχέτιση μετά την ολίσθηση στο χρόνο, επιλέξαμε να μελετήσουμε αυτή η οποία παρουσίαση την μεγαλύτερη αύξηση συσχέτισης από της 2. Για τον σκοπό αυτό εξήγαμε χαρακτηριστικά από την μια, για κάποιες χρονικές στιγμές και δημιουργήσαμε πρόβλημα ταξινόμησης τριών κατηγοριών με κλάση την συμπεριφορά την άλλης μετά από 4 ώρες. Για να αντιληφθούμε την επιτυχία η μη των πειραμάτων αρκεί να σκεφτούμε ότι εάν τυχαία επιχειρούσαμε να ταξινομήσουμε ένα δείγμα σε 3 διαφορετικές κλάσεις, τότε θα είχαμε πιθανότητα 33,3% να το επιτύχουμε. Εάν τα πειράματα παρουσιάσουν αξιόλογη αύξηση σε σχέση με αυτό το ποσοστό τότε αυτό είναι μια ένδειξη ότι το πείραμα μας έχει βάση. Στην συγκεκριμένη περίπτωση παρατηρούμε από τον συγκεντρωτικό πίνακα ότι τα ποσοστά ακρίβειας της πρόβλεψης είναι περίπου 40%, γεγονός που αποδεικνύει ότι η αύξηση του ποσοστού σε σχέση με το τυχαίο δεν είναι σημαντική. Υπάρχουν τιμές οι οποίες προσεγγίζουν το 50% για τιμές κατωφλιού μεταβολής κοντά στο 1 τοις χιλίοις, ένα όμως παρατηρηθεί με μεγαλύτερη προσοχή εξάγεται το συμπέρασμα ότι αυτό το ποσοστό είναι επίπλαστο καθώς η μια εκ τρων κλάσεων είναι άδεια.

Το γεγονός ότι δεν παρατηρούνται υψηλά ποσοστά πρόβλεψης είχε αρχίσει να διαφαίνεται από την στιγμή που εξήγαμε την συσχέτιση με ολίσθηση στον χρόνο, όπου εκεί μόνο σε 2

περιπτώσεις από τις 30 παρατηρήσαμε ότι αυξάνεται η συσχέτιση, αλλά και σε αυτές η αύξηση ήταν πάρα πολύ μικρή.

Βιβλιογραφία

1. Dunham, M. H. (2004). Data Mining. (Θ. Γ. Βρύκιος Βασίλης, Επιμ.) Αθήνα: Εκδόσεις Νέων Τεχνολογιών.
2. forexpros.gr. (n.d.). Forexpros. Ανάκτηση April 16, 2016, από Forexpros: http://www.forexpros.gr/ti-einai-to-forex/
3. Imanuel. (2016). Predictive Analytics Today. Ανάκτηση April 15, 2016, από Predictive Analytics Today: http://www.predictiveanalyticstoday.com/top-data-mining-software/
4. M., C. T. (2001). Συστήματα βάσεων δεδομένων : Μια πρακτική προσέγγιση στο σχεδιασμό, την υλοποίηση και τη διαχείριση βάσεων δεδομένων. (B. C. E., Επιμ., & P. Στέργιος, Μεταφρ.) Αθήνα: Ίων.
5. marketbet. (n.d.). marketbet.gr. Ανάκτηση April 19, 2016, από http://www.marketbet.gr/ti-einai/ti-einai-ta-cfd/
6. markets.com. (n.d.). Markets.com. Ανάκτηση April 16, 2016, από MARKETS.COM: http://mt4.markets.com/el/education/forex-education/what-is-forex.html
7. Periergos. (2015, August 31). periergos.gr. Ανάκτηση April 2017, 2016, από http://www.periergos.gr/erotiseis/ti-einai-commodities
8. Uknown. (n.d.). wikipedia.org. Ανάκτηση April 18, 2016, από https://en.wikipedia.org/wiki/Contract_for_difference
9. Unknown. (n.d.). wikipedia.org. Ανάκτηση April 17, 2016, από https://en.wikipedia.org/wiki/Commodity
10. Παν, M. I. (2006). Συστήματα βάσεων δεδομένων : Θεωρία και πρακτική εφαρμογή. (Π. A. N., Επιμ.) Αθήνα: Εκδόσεις Νέων Τεχνολογιών.