ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΗΠΕΙΡΟΥ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ Τ.Ε.

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ ΔΙΕΘΝΩΝ ΧΡΗΜΑΤΙΣΤΗΡΙΑΚΩΝ ΔΕΙΚΤΩΝ ΚΑΙ ΝΟΜΙΣΜΑΤΙΚΩΝ ΙΣΟΤΙΜΙΩΝ

Νικόλαος Αγγελάκης - Α.Μ. 9975

Επιβλέπων καθηγητής

Νικόλαος Γιαννακέας

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ Τ.Ε.

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ ΔΙΕΘΝΩΝ ΧΡΗΜΑΤΙΣΤΗΡΙΑΚΩΝ ΔΕΙΚΤΩΝ ΚΑΙ ΝΟΜΙΣΜΑΤΙΚΩΝ ΙΣΟΤΙΜΙΩΝ

Νικόλαος Αγγελάκης - Α.Μ. 9975

Επιβλέπων καθηγητής

Νικόλαος Γιαννακέας

- Άρτα 2016 -

ΕΥΧΑΡΙΣΤΙΕΣ

Αισθάνομαι την ανάγκη να επισημάνω την απέραντη ευγνωμοσύνη στους γονείς μου για όλα όσα μου έχουν προσφέρει από τα πρώτα στάδια της ζωής μου έως και τα φοιτητικά μου χρόνια και την αμέριστη υποστήριξη τους σε κάθε μου επιλογή. Επίσης θα ήθελα να ευχαριστήσω για άλλη μια φορά τον επιβλέποντα καθηγητή μου κ. Νικόλαο Γιαννακέα για την συνεχόμενη καθοδήγηση και την υπερπολύτιμη συμβολή του σε ένα κλάδο όχι και τόσο γνωστό σε μένα (Εξόρυξη Δεδομένων) καθώς και για την αφιέρωση πολύτιμου χρόνου ώστε να ολοκληρωθεί η εργασία αυτή.

Περίληψη

Στην παρούσα εργασία επιχειρείται η μελέτη του χάσματος τιμών ισοτιμιών και συναλλάγματος. Με την βοήθεια τριών αλγορίθμων ταξινόμησης εφαρμόστηκε εξόρυξη δεδομένων πάνω στα στοιχεία τριών χρηματιστηριακών δεικτών, (Ευρώ σε δολάριο, Λίρα αγγλίας σε δολάριο, Τιμή χρυσού σε δολάρια)σε βάθος πενταετίας για την εξαγωγή συμπερασμάτων που ενδεχομένως θα προσδιορίζουν τη μελλοντική πορεία αυτών των δεικτών. Τα δεδομένα των δεικτών που επεξεργαστήκαμε στη εργασία αποτελούνται από βασικά στοιχεία όπως όνομα δείκτη, τιμή κλεισίματος, τιμή ανοίγματος, ώρα, ημ/νία έτσι ώστε να προκύψουν τα απαραίτητα χάσματα για περαιτέρω ανάλυση. Τα αποτελέσματα των πειραμάτων έδειξαν ότι με τα χαρακτηριστικά που χρησιμοποιήθηκαν δεν μπορούμε να προβλέψουμε με σωστή προσέγγιση τη πορεία ενός δείκτη καθώς τείνουν αλλά είναι πολύ χαμηλότερα από το τυχαίο που είναι το 50% στο πρόβλημα δύο κατηγοριών, ενώ στο πρόβλημα τριών κατηγοριών σημειώθηκαν αποτελέσματα κοντά στο 47% από τον αλγόριθμο Random Forest εξίσου χαμηλό ποσοστό παρολαυτά.

ABSTRACT

The current paper attempts to study the price gaps between foreign exchange & currency indicators. With the assistance of three sorting algorithms, data mining was applied on the data of three economic indicators, (Euro to Dollar,English pound to dollar, Price of gold in dollars) in a five year time span for the purpose of extracting crucial conclusions that potentially could define future trajectory of these indicators. The data of these indicators proccesed in this paper are concluded of basic elements such as name of indicator, opening price, closing price, time, date so that the necessary gaps could be aquired for further studies and conclusions. Results of the experiments used in this thesis showed that with the characteristics used, we cannot predict confidently the trajectory of an indicator because of results orbiting much lower than random which is 50% in the 2-class problem and while some of the results in the 3-class problem were near 47% by Random Forest algorithm, it is still a low percentage nevertheless.

Π Ε Ρ Ι Ε Χ Ο Μ Ε Ν Α

ΕΙΣΑΓΩΓΗ……...……..
..10

ΚΕΦΑΛΑΙΟ 1

Οικονομικοί Δείκτες και Δεδομένα

1.1 Ιστορική αναδρομή οικονομικών δεικτών.....11

1.2 Οικονομικοί Δείκτες.....……...12

1.3 Εργαλεία επεξεργασίας οικομικών δεικτών.....13

ΚΕΦΑΛΑΙΟ 2

Εξόρυξη Δεδομένων και οικονομικά δεδομένα

2.1 Τι είναι εξόρυξη δεδομένων…….....….....19

2.2 Μέθοδοι εξόρυξης δεδομένων…….....……...20

2.3 Εφαρμογές εξόρυξης δεδομένων σε οικονομικά δεδομένα.....23

ΚΕΦΑΛΑΙΟ 3

Μέθοδος πρόβλεψης

3.1 Εντοπισμός και μέτρηση χάσματος.....27

3.2 Μέθοδοι ταξινόμησης.....27

3.3 Μέτρα αξιολόγησης.....34

ΚΕΦΑΛΑΙΟ 4

Αποτελέσματα

4.1 Σύνολο Δεδομένων.....36

4.2 Περιγραφή πειραμάτων.....38

ΚΕΦΑΛΑΙΟ 5

Συμπεράσματα.....46

ΒΙΒΛΙΟΓΡΑΦΙΑ…….....48

ΕΙΣΑΓΩΓΗ

Η σύγχρονη εποχή επέβαλε την ανάπτυξη οικονομικών εργαλείων τα οποία ανταποκρίνονται στην πολυπλοκότητα του διεθνούς εμπορίου και σταθερότητας εθνικών οικονομιών. Οι οικονομικοί δείκτες το στατιστικό στοιχείο εκείνο που επιτρέπει την ανάλυση, πρόβλεψη και απόδοση της οικονομικής δραστηριότητας από κράτη και οργανισμούς αποτελεί αντικείμενο μελέτης στην παρούσα εργασία. Μέσω επεξεργασίας και εξαγωγής χαρακτηριστικών των χασμάτων τιμών διεθνών δεικτών και συναλλάγματος επιχειρείται η πρόγνωση και η πρόβλεψη της πορείας τους (δείκτες) με την εφαρμογή διαφόρων τεχνικών μάθησης. Έτσι ερευνήθηκαν δεδομένα τριών δεικτών σε βάθος πενταετίας ώστε να υπάρξει μια σχετικότερα ολοκληρωμένη εικόνα που να διαμορφώνει την πορεία τους,οι τρείς δείκτες είναι 1)Ισοτιμία ευρώ-δολάριο,2)Ισοτιμία λίρας αγγλίας-δολάριο,3)Τιμή χρυσού σε δολάρια. Για να επιτευχθεί αυτό χρησιμοποιήθηκαν τρείς διαφορετικοί αλγόριθμοι ταξινόμησης στα χάσματα τιμών που προέκυψαν απο τους δείκτες που επεξεργαστήκαμε : α) Μπευζιανός Ταξινομητής Naive Bayes ένας απλός πιθανοτικός ταξινομητής ο οποίος βασίζεται στο θεώρημα του Bayes, β)Τυχαία Δάση ή RandomForest ο αλγορίθμος ταξινόμησης που αποτελείται από μια συστάδα δένδρων αποφάσεων (decision trees) από τα οποία συλλέγει τις "αποφάσεις" που έκαναν και ταξινομεί με βάση την πλειοψηφία των αποφάσεων αυτών, γ) Ο αλγόριθμος kNN ή k-κοντινότερων γειτόνων που βασίζεται στην επιλογή των k εγγύτερων σημείων σε ένα σημείο ελέγχου που πραγματώνεται στο σε ένα σύνολο δεδομένων μάθησης. Ξεκινώντας στο Κεφάλαιο 1 γίνεται ιστορική αναδρομή των οικονομικών δεικτών, ανάλυση χρηματιστηριακών εννοιών και παρουσίαση εργαλείων επεξεργασίας δεικτών. Το Κεφάλαιο 2 πραγματεύεται την έννοια της εξόρυξης δεδομένων, ορισμένες μεθόδους εξόρυξης καθώς και διάφορες εφαρμογές της πάνω σε οικονομικά δεδομένα. Οι μέθοδοι πρόβλεψης ακολουθούν στο Κεφάλαιο 3 όπου γίνεται η μέτρηση του χάσματος στους δείκτες που επεξεργαστήκαμε, όπως επίσης γίνεται βαθύτερη ανάλυση των τριών αλγορίθμων ταξινόμησης και τέλος σε αυτήν την ενότητα γίνεται αναφορά στα μέτρα αξιολόγησης των αλγορίθμων. Στο Κεφάλαιο 4 παρουσιάζονται τα αποτελέσματα της επεξεργασίας των χασμάτων τιμών παράλληλα με συνοπτική περιγραφή για

κάθε πείραμα. Τέλος στο Κεφάλαιο 5 παρουσιάζονται τα συμπεράσματα που προέκυψαν.

ΚΕΦΑΛΑΙΟ 1

ΟΙΚΟΝΟΜΙΚΟΙ ΔΕΙΚΤΕΣ ΚΑΙ ΔΕΔΟΜΕΝΑ

1.1 Ιστορική Αναδρομή Οικονομικών Δεικτών

Με τον όρο οικονομικοί δείκτες αναφερόμαστε σε ένα στατιστικό στοιχείο που συνδέεται άμεσα με την οικονομική δραστηριότητα. Πιο συγκεκριμένα οι οικονομικοί δείκτες έχουν την ικανότητα να:

- Επιτρέπουν την ανάλυση οικονομικής απόδοσης
- Προβλέπουν την οικονομική απόδοση

Υπάρχει μια πληθώρα οικονομικών δεικτών παρόλα αυτά ορισμένοι μόνο εξ αυτών θα συζητηθούν στην παρούσα εργασία. Αξίζει να αναφέρουμε ότι η κύρια διερευνητική υπηρεσία στις ΗΠΑ στον τομέα της οικονομίας, της εργασίας και των στατιστικών είναι η Bureau of Labor Statistics.Το 1920 σύμφωνα με τον Μπράιαν Μάρκ, αντιπρόεδρο του τμήματος έρευνας της Ατλάντα η κυβέρνηση σύλλεξε οικονομικά στοιχεία μια χρονική περίοδο όπου υπήρχε βαθιά ύφεση. Λίγο αργότερα για να καλυφθεί η παρούσα ύφεση η κυβέρνηση των ΗΠΑ καταφεύγει στην πρόσληψη του οικονομολόγου Simon Kuznets με σκοπό να ενημερωθεί για το εισόδημα του έθνους. Μια δεκαετία αργότερα έκανε την εμφάνιση του το ακαθάριστο εγχώριο προϊόν. Καθοριστικό ρόλο έπαιξαν οι NIPAS στην οικονομική πορεία. Πιο συγκεκριμένα βοήθησαν στην χάραξη μιας πολιτικής σύμφωνα με την οποία τέθηκε ένα σημείο αναφορά για να ελέγχει την οικονομική κατάσταση. Το σπουδαιότερο έργο των NIPAS αποτέλεσε το ΑΕΠ (βλ. ακαθάριστο εγχώριο προϊόν). Η οικονομική κατάσταση μιας χώρας δεν είναι ποτέ σταθερή μεταβάλλεται διαρκώς. Οι διαρθρωτικές αλλαγές καθώς και οι αλλαγές κάποιον δεικτών οι οποίοι φθίνουν σε δημοτικότητα. Σύμφωνα με έρευνα διαπιστώθηκε ότι ορισμένοι δείκτες που

ενδεχομένως να είχαν ανοδική πορεία (βλ. άνθρακας) στις αρχές 20ου αιώνα με την πάροδο των χρόνων έχασαν την αξία τους. Οι οικονομικές αλλαγές κατά τη διάρκεια του 1980 αποδυναμώνουν την σχέση αύξησης της προσφοράς του χρήματος και της οικονομικής δραστηριότητας. Το 1993 ο πρώην πρόεδρος της Federal Reserve, Alan Greenspan είπε στο Κογκρέσο ότι: «τουλάχιστον προς το παρόν , M2 έχει υποβαθμιστεί ως αξιόπιστος δείκτης οικονομικών». Το 2000 η Fed σταμάτησε τον καθορισμό στόχων για την προσφορά χρήματος ενώ το 2006 σταμάτησε τη δημοσίευση του νομισματικού μεγέθους M3. Λίγο νωρίτερα μια ερευνητική ομάδα ανακοίνωσε την άρση του μέτρου M2 από τους κορυφαίους οικονομικούς δείκτες του LEI.

1.2 Οικονομικοί Δείκτες

Σε αυτό το υποκεφάλαιο της εργασίας θα περιγράψουμε τους όρους StockCFDs, Χρηματιστήριο, FOREX.

FOREX

Ο όρος FOREX (βλ. FOReign Exchange) αποτελεί μια παγκόσμια αγορά συναλλάγματος όπου τα νομίσματα διαπραγματεύονται το ένα το άλλο. Μερικοί από τους συμμετέχοντες φαίνονται παρακάτω:

- Κεντρικές τράπεζες , εμπορικές τράπεζες
- Εταιρίες
- επενδυτές
- Επιχειρηματίες
- Τουρίστες [3]

Ουσιαστικά περισσότεροι χρηματιστηριακοί δείκτες οι οποίοι προσφέρουν την δυνατότητα συναλλαγών στους επενδυτές στις ΗΠΑ, Ευρώπη , Ασία και Αυστραλία.[5]

COMMODITIES

Με τον όρο CCI -Commodity Channel Index (βλ. COMMODITIES) αναφερόμαστε σε έναν κυκλικό τεχνικό δείκτη ο οποίος έχει τη δυνατότητα να προσδιορίσει πότε μια αγορά είναι πολύ αγορασμένη ή πουλημένη αντίστοιχα.[2]

CFD

Ως CFD (βλ. Contract for Difference) αναφερόμαστε σε ένα πολύ σημαντικό εργαλείο επένδυσης. Ένας επενδυτής είτε κερδίζει είτε χάνει χρήματα αυτό προσδιορίζεται βάση της διαφοράς της τιμής ενός προϊόντος. Ουσιαστικά ενός διαπραγματεύσιμου χρηματικοοικονομικού προϊόντος (βλ. μετοχή, δείκτης).[3]

ΧΡΗΜΑΤΙΣΤΗΡΙΟ

Ως χρηματιστήριο αναφερόμαστε σε μια οργανωμένη αγορά επίσημη αναγνωρισμένη από το κράτος. Στην συγκεκριμένη αγορά συναντιόνται ενδιαφερόμενοι για την αγοραπωλησία κινητών αξιών ή εμπορευμάτων. Ως κινητές αξίες μπορούμε να αναφέρουμε: μετοχές, ομόλογα. [4]

1.3 Εργαλεία επεξεργασίας Οικονομικών Δεικτών

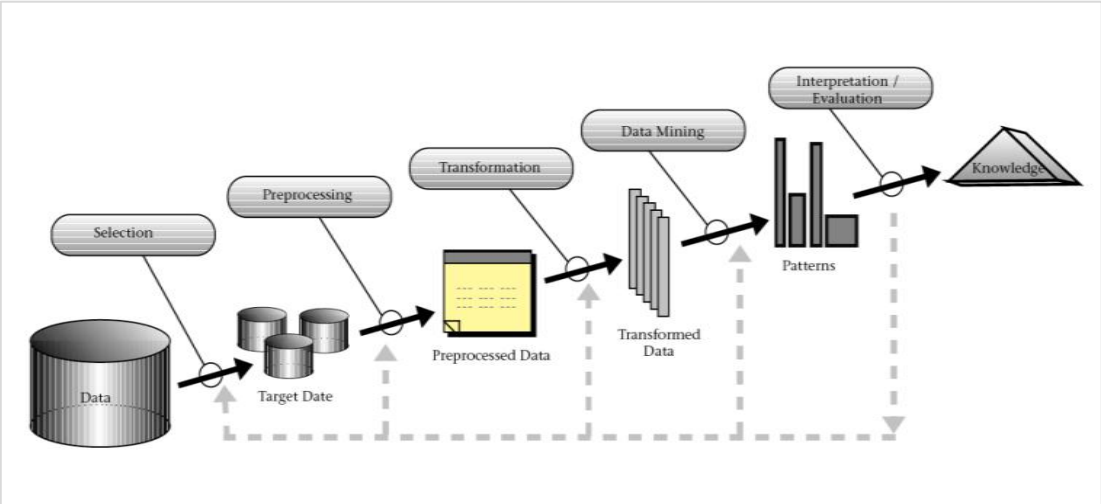
1.3.1 Μέθοδοι επεξεργασίας

Σε αυτή την παράγραφο θα αναφερθούμε στις μεθόδους επεξεργασίας που διακρίνονται σε :

- Στατιστική
- Εξόρυξη δεδομένων
- *Pattern Recognition*

Εικόνα 1.1 : Στατιστική

Η στατιστική δηλώνει αριθμητικές πληροφορίες τόσο πρωτογενείς όσο και επεξεργασμένες. Επίσης σημαντικό θα ήταν να αναφέρουμε ότι ο ίδιος όρος γραφόμενος με κεφαλαίο γράμμα αναφέρεται στην επιστήμη η οποία ασχολείται με αριθμητικές πράξεις. Η ετυμολογία της παραπάνω λέξης προέρχεται από τη λατινική λέξη Status αρχικά χρησιμοποιούνταν για να δηλώσει τη συλλογή στοιχείων για τις κρατικές ανάγκες λόγου χάρη του πληθυσμού.[6]

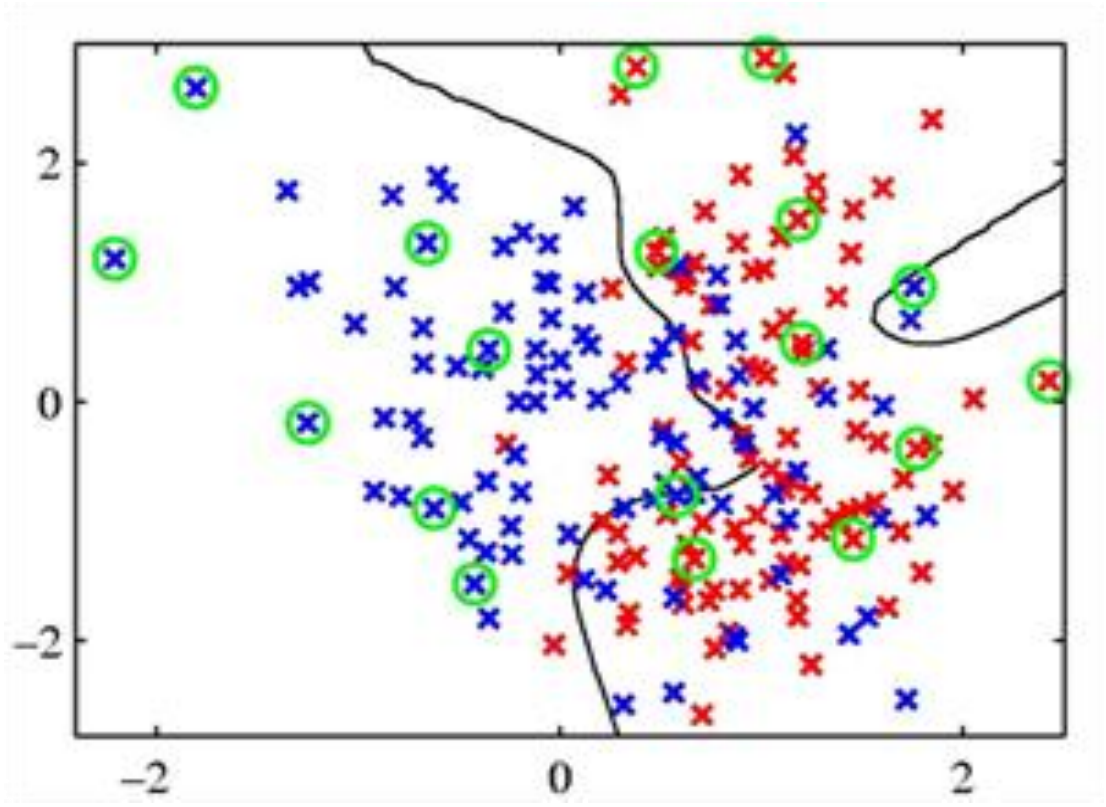


Εικόνα 1.2 : Εξόρυξη Δεδομένων

Με τον όρο εξόρυξη δεδομένων κατανοούμε ότι αναφερόμαστε σε μια φόρμα δεδομένων και πιο συγκεκριμένα σε μια πληθώρα δεδομένων ή επεξεργασία δεδομένων. Παρόλα αυτά θα πρέπει να επισημάνουμε ότι γενικεύεται σε κάθε

σύστημα υποστήριξης αποφάσεων.Λόγου χάρη στην τεχνητή νοημοσύνη, την εκμάθηση μηχανής και την επιχειρηματική ευφυΐα. Ολοκληρώνοντας την αναφορά μας η λέξη κλειδί για να ορίσουμε επακριβώς την εξόρυξη είναι η αποκάλυψη- η ανίχνευση του καινούριου. [7]

Η αναγνώριση προτύπων γνωστή τα τελευταία χρόνια με την ορολογία Pattern Recognition αποτελεί ένα εξαιρετικά σημαντικό επιστημονικό πεδίο που ως βασικό του μέλημα έχει ανάπτυξη αλγορίθμων καθώς και την αυτοποιημένη απόδοση κάποιας τιμής ή ακόμη και διακριτού στοιχείου σε εισαγόμενα δεδομένα. Κατά κύριο λόγο τα δεδομένα είναι κωδικοποιημένα σε αλληλουχίες αριθμών. Αφού τεθούν κάποια κριτήρια τα δεδομένα αυτά κατηγοριοποιούνται και ταξινομούνται σε ομάδες. Αυτό συμβαίνει ακόμη και όταν υπάρχει θόρυβος ο οποίος σε αρκετές περιπτώσεις δυσκολεύει την παραπάνω διαδικασία. Από ιστορικής πλευράς το ερευνητικό ενδιαφέρον έχει βαθιά τα θεμέλια του και πιο συγκεκριμένα κοντά στη δεκαετία του 1960. Δηλαδή την πρώτη περίοδο ανάπτυξης της πληροφορικής και συγκεκριμένα την τεχνητής νοημοσύνης.[8]



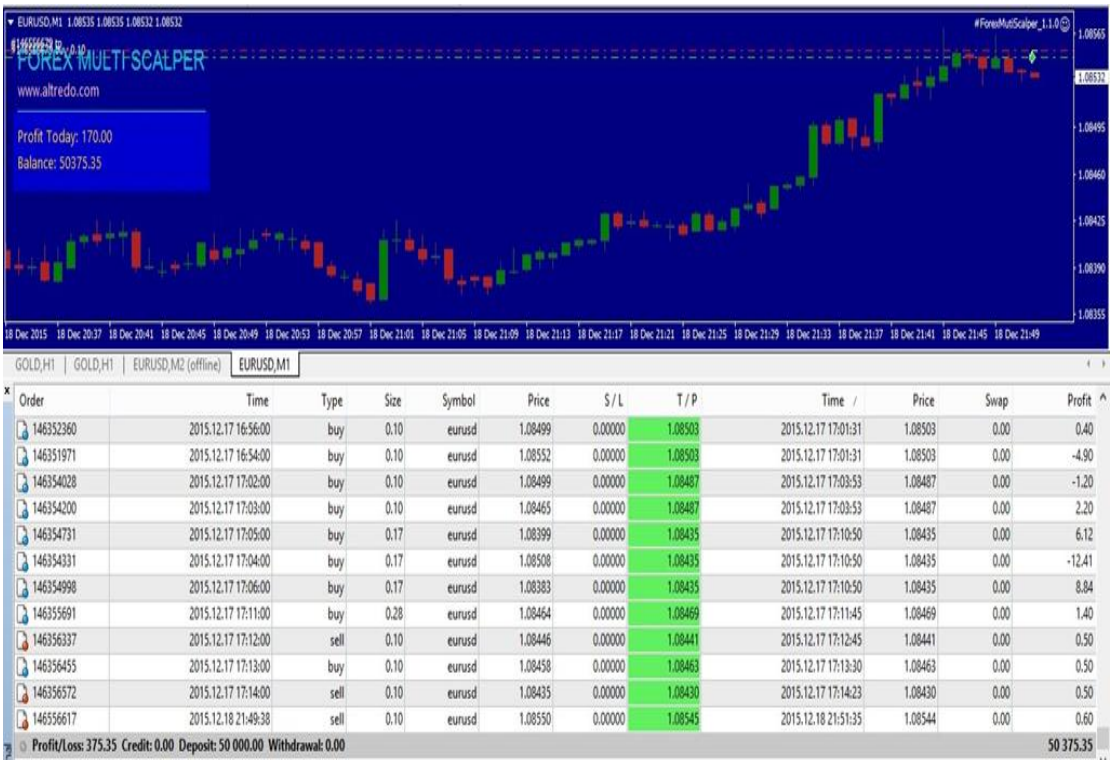
Εικόνα 1.3 : Pattern Recognition

1.3.2 Πακέτα επεξεργασίας (Softwares)

Ένα από τα σημαντικότερα πακέτα επεξεργασίας αποτελεί το Forex. Αναπτύχθηκε από μια ομάδα επαγγελματιών και προγραμματιστών λογισμικού και βασίζεται στα νευρωνικά δίκτυα και στους γενετικούς αλγορίθμους. Τόσο η αυτό-μάθηση όσο και η αυτό-ενημέρωση βασίζονται στις συνθήκες της αγοράς. Το παραπάνω πακέτο επεξεργασίας έχει ενσωματωμένο αυτόματο αλγόριθμό υπολογισμού κινδύνου . ανοίγει θέσεις κατά ημερήσιες κινήσεις σε περίπτωση και μόνο που αυτές οι κινήσεις έχουν πρόβλεψη επιτυχίας. Προστατεύει από κάθε θέση τις παραγγελίες και δεν υπάρχει καμία περίπτωση απώλειας λογαριασμού. Επιπροσθέτως έχει ενσωματωμένο λογισμικό προστασίας μεσίτη. Ακόμη ένα χαρακτηριστικό του εν λόγο λογισμικό είναι ότι έχει την ικανότητα να ανοίγει θέσεις με τα χαμηλότερα spreads καθώς και τις υψηλότερες ρευστότητες της αγοράς. Επιπλέον λειτουργεί τόσο σε βραχυπρόθεσμες όσο και σε μακροπρόθεσμες τάσεις και λειτουργεί με 4 ή 5 ψηφία μετά την υποδιαστολή. Παρέχει:

- 100% αυτοματοποιημένη διαπραγμάτευση
- 250% κέρδος ανά μήνα
- Μέγιστη ανάληψη 3,5%.

Ολοκληρώνοντας την αναφορά μας στο Forex θα προσθέσουμε ότι κάνει μετατροπή σε EURUSD , GBPUSD σε ένα λεπτό και μια ώρα και διακρίνεται για την ακρίβεια και ασφάλεια του.



Order	Time	Type	Size	Symbol	Price	S / L	T / P	Time	Price	Swap	Profit
146352360	2015.12.17 16:56:00	buy	0.10	eurusd	1.08499	0.00000	1.08503	2015.12.17 17:01:31	1.08503	0.00	0.40
146351971	2015.12.17 16:54:00	buy	0.10	eurusd	1.08552	0.00000	1.08503	2015.12.17 17:01:31	1.08503	0.00	-4.90
146354028	2015.12.17 17:02:00	buy	0.10	eurusd	1.08499	0.00000	1.08487	2015.12.17 17:03:53	1.08487	0.00	-1.20
146354200	2015.12.17 17:03:00	buy	0.10	eurusd	1.08465	0.00000	1.08487	2015.12.17 17:03:53	1.08487	0.00	2.20
146354731	2015.12.17 17:05:00	buy	0.17	eurusd	1.08399	0.00000	1.08435	2015.12.17 17:10:50	1.08435	0.00	6.12
146354331	2015.12.17 17:04:00	buy	0.17	eurusd	1.08508	0.00000	1.08435	2015.12.17 17:10:50	1.08435	0.00	-12.41
146354998	2015.12.17 17:06:00	buy	0.17	eurusd	1.08383	0.00000	1.08435	2015.12.17 17:10:50	1.08435	0.00	8.84
146355691	2015.12.17 17:11:00	buy	0.28	eurusd	1.08464	0.00000	1.08469	2015.12.17 17:11:45	1.08469	0.00	1.40
146356337	2015.12.17 17:12:00	sell	0.10	eurusd	1.08446	0.00000	1.08441	2015.12.17 17:12:45	1.08441	0.00	0.50
146356455	2015.12.17 17:13:00	buy	0.10	eurusd	1.08458	0.00000	1.08463	2015.12.17 17:13:30	1.08463	0.00	0.50
146356572	2015.12.17 17:14:00	sell	0.10	eurusd	1.08435	0.00000	1.08430	2015.12.17 17:14:23	1.08430	0.00	0.50
146556617	2015.12.18 21:49:38	sell	0.10	eurusd	1.08550	0.00000	1.08545	2015.12.18 21:51:35	1.08544	0.00	0.60
Profit/Loss: 375.35 Credit: 0.00 Deposit: 50 000.00 Withdrawal: 0.00											50 375.35

Εικόνα 1.4 : Forex Trading Robot

1.3.3 Στρατηγικές Πρόβλεψης Δεικτών

Κατά την ανάπτυξη ενός στρατηγικού σχεδιασμού για να επιτευχθούν καλύτερα αποτελέσματα στο μέλλον απαραίτητος είναι ο έλεγχος του παρελθόντος. Λόγου χάρη μπορείτε να βρείτε το που υστερούν σε έσοδα οι πωλήσεις της επιχείρησης σας. Χαρακτηριστικό παράδειγμα μπορεί να αποτελέσει μια εταιρία παροχής υπηρεσιών συσκευασίας η οποία μπορεί να παρατηρηθεί τόσο αύξηση όσο και μείωση των εσόδων της η οποία εξαρτάται από την αλλαγή του οικονομικού δείκτη παραγωγής διαρκών αγαθών κατά δέκα μήνες. Κατά συνέπεια η εταιρεία οφείλει να παρακολουθήσει το δείκτη σε μηνιαία βάση ώστε να υπάρξει βιωσιμότητα δείκτη ως προς την κατεύθυνση των δραστηριοτήτων της εκάστοτε επιχείρησης. Το γραφείο οικονομικών ερευνών έχει μια εκτενή λίστα δεδομένων. Επιπροσθέτως το σύστημα FRED στο Σαντ Λούις (βλ. Federal Reserve Bank) αποτελεί ένα καλό παράδειγμα συστήματος με εξαιρετικές δυνατότητες χαρτογράφησης μακρό-οικονομικών δεικτών. Αξίζει να επισημάνουμε ότι οι πληροφορίες είναι ελεύθερες για το κοινό.Μερικές από τις ιδέες που πρέπει να ληφθούν σοβαρά υπόψη για την προβολή καθώς και την πρόβλεψη των εσόδων περιλαμβάνουν τα ακόλουθα:

- Δεν προεξοφλούμε μια ένδειξη ενός δείκτη μόνο και μόνο από το παρελθόν ο έλεγχος πρέπει να είναι ακριβής και διεξοδικός στο μέλλον.
- Αν υπάρχει ένα ενδιαφέρον για ένα δείκτη δεν θα βασίζεστε μόνο σε πληροφορίες που έχετε σχετικά με το δείκτη αλλά θα πρέπει να γίνει ενδελεχής έλεγχος για επιπρόσθετες πληροφορίες ότι η πρόβλεψη είναι ορθή.
- Χρησιμοποιήστε το ιστορικό παρελθόν για να σας βοηθήσει να ενώσετε τις τελείες του μέλλοντος. Παρόλα αυτά δεν κοιτάζουμε πίσω από το πρόσφατο παρελθόν διότι αν το πρόσφατο παρελθόν θα αποτελούσε μια ένδειξη σταθερή για το μέλλον σε αυτή την περίπτωση όλοι πλούσιοι θα ήταν επικαλούμενοι τους δείκτες Dow ή NASDAQ.
- Θέσπιση ενός συνεχούς προγράμματος πρόβλεψης. Θα φροντίσετε να έχετε μια μηνιαία ή τριμηνιαία πρόβλεψη σαρώνοντας το οικονομικό τοπίο. Οι ακριβείς προβλέψεις είναι εξαιρετικά σημαντικές για την αποτελεσματική παρακολούθηση και εφαρμογή των στρατηγικών επιτυχίας μια επιχείρησης.
- Δημιουργία οικονομικών προβλέψεων
- Τέλος ελέγχουμε αν οι στόχοι ταιριάζουν με τους αριθμούς.[10]

ΚΕΦΑΛΑΙΟ 2

ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΟΙΚΟΝΟΜΙΚΑ ΔΕΔΟΜΕΝΑ

2.1 Τι είναι η εξόρυξη δεδομένων

Όπως προαναφέραμε στο κεφάλαιο 1 με τον όρο εξόρυξη δεδομένων αναφερόμαστε σε μια φόρμα δεδομένων και πιο συγκεκριμένα σε μια πληθώρα δεδομένων ή επεξεργασία δεδομένων. Σημαντικό θα ήταν να προσθέσουμε ότι ο βασικότερος στόχος την εξόρυξης δεδομένων είναι η διαδικασία της αυτοματοποίησης εξολοκλήρου ή κατά το ήμισυ έως ότου επιτευχθεί η ανάλυση μεγάλης ποσότητας δεδομένων για την εξαγωγή κάποιου ενδιαφέροντος προτύπου.

Λόγου χάρη ομάδες από εγγραφές δεδομένων, ασυνήθιστες εγγραφές καθώς και εξαρτήσεις. Απαραίτητα σε αυτή την περίπτωση είναι οι βάσεις δεδομένων (βλ. ευρετήρια). Έπειτα τα παραπάνω πρότυπα ενδεχομένως να υποστούν περεταίρω επεξεργασία και ανάλυση όπως για παράδειγμα εκμάθηση μηχανής καθώς και προγνωστική ανάλυση. Στην εξόρυξη δεδομένων σύνηθες είναι η χρήση ορολογιών όπως : data dredging , data fishing και data snooping. Οι παραπάνω όροι σχετίζονται κυρίως με τη διαδικασία εξόρυξης μεγάλων ποσοτήτων δεδομένων που διακρίνονται από μικρή αξιόπιστα στατιστικά συμπεράσματα καθώς και μεγάλων υποθέσεων που έγιναν για μεγάλες συλλογές δεδομένων. Από ιστορικής πλευράς η διαδικασία της εξόρυξης διεξάγεται εδώ και αιώνες. Πιο συγκεκριμένα οι πρώτες μέθοδοι για τον προσδιορισμό προτύπων ήταν η Bayes καθώς και η ανάλυση της παλινδρόμησης. Με την πάροδο των χρόνων και την εισχώρηση της τεχνολογίας ολοένα και περισσότερο στην καθημερινότητα μας επέφεραν ραγδαίους ρυθμούς ανάπτυξης των συγκεντρωμένων δεδομένων καθώς και την ανεύρεση αποτελεσματικότερων και ευκολότερων τρόπων χειρισμού δεδομένων. Συνεπώς με την πάροδο των χρόνων ο χειροκίνητος τρόπος επεξεργασίας δεδομένων αντικαταστάθηκε από

αυτοματοποιημένες μεθόδους που χαρακτηρίστηκαν περισσότερο αποτελεσματικές στην πληθώρα δεδομένων που κλήθηκαν να επεξεργαστούν. Σε αυτό καθοριστικό ρόλο έπαιξαν και άλλες πολύ σημαντικές επιστήμες τις πληροφορικής όπως:

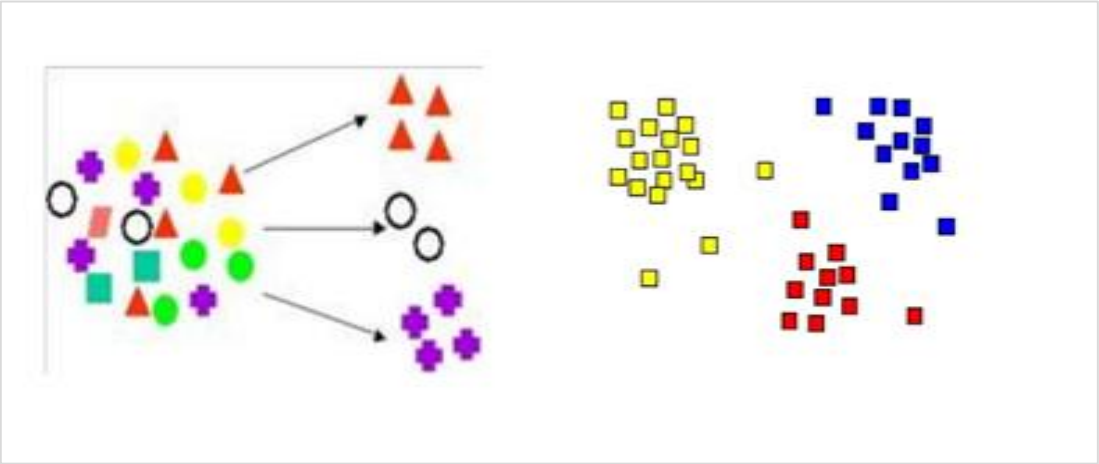
- Τα νευρωνικά δίκτυα
- Η συσταδοποίηση
- 1950 : Οι γενετικοί αλγόριθμοι
- 1960 : Τα δέντρα αποφάσεων
- 1990 : Οι μηχανές υποστήριξης διανυσμάτων

Συμπερασματικά προκύπτει λοιπόν ότι η εξόρυξη δεδομένων αποτελεί μια διαδικασία επεξεργασίας όλων των προαναφερθέντων μεθόδων με απώτερο σκοπό την εύρεση άγνωστων προτύπων σε μεγάλα σύνολα δεδομένων. Με αυτό τον τρόπο δημιουργείται μια γέφυρα που έρχεται να ενώσει την εφαρμοσμένη στατιστική και την τεχνίτη νοημοσύνη με τις βάσεις δεδομένων. Ουσιαστικά κάνοντας χρήση των βάσεων δεδομένων με τον τρόπο που αποθηκεύονται και κατατάσσονται τα δεδομένα με σκοπό να εκτελέσουν την θεωρία των αλγορίθμων όσο το δυνατό με τον καλύτερο τρόπο. Και να επιτρέπουν στις μεθόδους που προαναφέραμε να επεξεργαστούν μεγάλες ποσότητες δεδομένων.[11]

2.2 Μέθοδοι Εξόρυξης Δεδομένων

Η συσταδοποιήση γνωστή και ως clustering αποτελεί μια διαδικασία που μοιάζει με την κατηγοριοποίηση με τη βασική διαφορά ότι σε αυτή τη μέθοδο οι συστάδες δηλαδή οι ομάδες των δεδομένων δεν είναι προκαθορισμένες αλλά κατά κύριο λόγο ορίζονται από τα δεδομένα. Η παραπάνω μέθοδος είναι γνωστή και με την ορολογία μη εποπτευόμενη μάθηση ή τμηματοποίηση. Επιπλέον θα προσθέσουμε ότι σε ορισμένες περιπτώσεις μπορεί να μην αναφερθεί ως διαμέριση ή τμηματοποίηση δεδομένων τα οποία δύναται να είναι ή και να μην είναι χωρισμένα σε διακριτές ομάδες μεταξύ τους. Για να επιτευχθεί η συσταδοποίηση χρειάζεται ο καθορισμός της ομοιότητας ως προς συγκεκριμένα γνωρίσματα τα οποία είναι προκαθορισμένα ανάμεσα στα δεδομένα. Αξίζει να αναφέρουμε ότι τα πιο σχετικά δεδομένα είναι ομαδοποιημένα στις ίδιες ομάδες. Ακολουθεί ένα παράδειγμα συσταδοποίησης στο οποίο θα κατανοήσουμε καλύτερα την προαναφερθείσα μέθοδο. Όσον αφορά το παράδειγμα θα επισημάνουμε ότι αφού οι ομάδες δεν είναι προκαθορισμένες

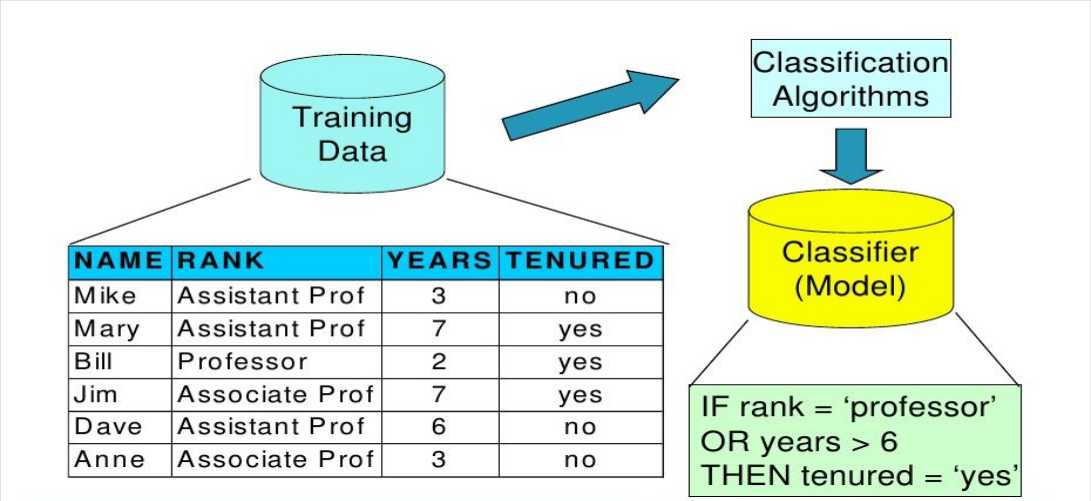
απαραίτητη προϋπόθεση είναι ύπαρξη ενός ειδικού ο οποίος θα ερμηνεύσει την σημασία των συστάδων που δημιουργούνται.



Εικόνα 2.1 : Συσταδοποίηση (Clustering)

Παράδειγμα

Μια αλυσίδα πολυκαταστημάτων έχει τη δυνατότητα να δημιουργήσει ειδικούς καταλόγους οι οποίοι κατά κύριο λόγο έχουν ως στόχο διάφορες δημογραφικές ομάδες, με βάση τα γνωρίσματα ως προς το εισόδημα, τον τόπο διαμονής καθώς και τα φυσικά χαρακτηριστικά των δυνητικών πελατών (βλ. ηλικία, ύψος, βάρος). Για να καθοριστεί σε ποιους από τους παραπάνω πελάτες θα αποσταλεί μέσω ταχυδρομείου διαφημιστικό υλικό με στόχο την δημιουργία πιο συγκεκριμένων νέων καταλόγων η εταιρία καταφεύγει στην μέθοδο της συσταδοποίησης των πιθανών πελατών με γνώμονα τις τιμές γνωρισμάτων τους. Στη συνέχεια τα αποτελέσματα της παραπάνω διαδικασίας χρησιμοποιούνται από τη διεύθυνση προκειμένου να δημιουργηθούν κατάλληλοι κατάλογοι και μετέπειτα να διανεμηθούν στο κατάλληλο τμήμα του πληθυσμού με βάση τις ομάδες που αντιστοιχούν στον κατάλογο αυτόν. Ολοκληρώνοντας την αναφορά μας στο εν λόγο παράδειγμα αξίζει να επισημάνουμε ότι μια ειδική κατηγορία συσταδοποίησης καλείται κατάτμηση ή αλλιώς έχει γίνει γνωστή και ως segmentation. Με την συγκεκριμένη διαδικασία μια βάση δεδομένων χωρίζεται σε διακριτές ομάδες παρόμοιων εγγραφών που ονομάζονται τμήματα. Οι διαδικασία της κατάτμησης είναι παρόμοια της συσταδοποίησης ενώ για πολλούς θεωρείται ως ένας ειδικός τύπος συσταδοποίησης που εφαρμόζεται στην ίδια βάση δεδομένων.



Εικόνα 2.2 : Κατηγοριοποίηση (Classification)

Με τον όρο κατηγοριοποίηση (βλ. classification) αναφερόμαστε σε απεικόνιση δεδομένων σε προκαθορισμένες ομάδες ή κατηγορίες- κλάσεις. Ουσιαστικά έχει άμεση σχέση με την εποπτευόμενη μάθηση για το λόγο ότι οι προαναφερθείσες κατηγορίες- κλάσεις έχουν την ιδιότητα να καθορίζουν πριν καλά εξετάσουν τα δεδομένα. Για να κατανοήσουμε καλύτερα την μέθοδο αυτή θα αναφέρουμε δύο παραδείγματα:

- Ο καθορισμός δηλαδή το ένα θα δοθεί τραπεζικό δάνειο
- Ο προσδιορισμός του πιστωτικού ρίσκου

Βασικός στόχος της κατηγοριοποίησης είναι η απαίτηση οι κατηγορίες να καθορίζονται με βάση τις τιμές των γνωρισμάτων των δεδομένων. Είναι σύνηθες να περιγράφουν τις κατηγορίες κοιτάζοντας τα χαρακτηριστικά των δεδομένων που είναι γνωστό εξ αρχής ότι ανήκουν στην κατηγορία. Χαρακτηριστικό παράδειγμα κατηγοριοποίησης αποτελεί η αναγνώριση προτύπων όπου στην προκειμένη περίπτωση ένα πρότυπο εισόδου κατηγοριοποιείται σε διάφορες τεχνολογίες. Ο βασικός γνώμονας κατηγοριοποίησης του είναι η εγκυρότητα του ως προς αυτές τις προκαθορισμένες κατηγορίες.

Παράδειγμα

Οι εταιρίες πιστωτικών καρτών θα πρέπει να καθορίζουν εάν θα εγκρίνουν αγορές μέσω πιστωτικών καρτών. Ας υποθέσουμε ότι με βάση το αγοραστικό ιστορικό ενός

πελάτη κάθε αγορά τοποθετείται σε μια από τις παρακάτω τέσσερις κατηγορίες:

- Να εγκριθεί
- Να ζητηθούν επιπλέον τα στοιχεία ταυτότητας πριν από την έγκριση
- Να μην εγκριθεί
- Να μην εγκριθεί και να ενημερωθεί η αστυνομία

Οι λειτουργίες της εξόρυξης γνώσης από δεδομένα εξυπηρετούν δύο σκοπούς:

Α) εξέταση των δεδομένων του ιστορικού των πελατών και καθορισμός του τρόπου που ταιριάζουν στις κατηγορίες ,

Β) το πρόβλημα είναι ότι θα εφαρμοστεί αυτό το μοντέλο σε κάθε μια από τις νέες αγορές. Εάν και μπορεί να θεωρηθεί και ότι το δεύτερο μέρος είναι πραγματικά μια απλή ερώτηση βάσεων δεδομένων, το πρώτο μέρος δεν μπορεί να θεωρηθεί σαν τέτοια.

2.3 Εφαρμογές Εξόρυξης Δεδομένων σε Οικονομικά Δεδομένα

Η πρόβλεψη της χρηματιστηριακής αγοράς, τα οικονομικά καθήκοντα, η συναλλαγματική ισοτιμία, η τραπεζική πτώχευση, οι χρηματικοοικονομικοί κίνδυνοι, η διαχείριση δανείων , το τραπεζικό προφίλ τον πελατών αποτελούν μερικά βασικά θέματα μετά οποία ασχολείται η εξόρυξη δεδομένων. Μερικά από τα παραπάνω για παράδειγμα το τραπεζικό προφίλ των πελατών έχουν μεγάλες ομοιότητες με την εξόρυξη δεδομένων για τη σκιαγράφηση των πελατών σε άλλους τομείς. Όσον αφορά την χρηματικοοικονομική πρόβλεψη της αγοράς αξίζει να αναφέρουμε ότι περιλαμβάνει τον προγραμματισμό των στρατηγικών επενδύσεων και έχει την ικανότητα να προσδιορίζει το πότε είναι η καταλληλότερη στιγμή για την αγορά αποθεμάτων αλλά και τι αποθέματα πρέπει να αγοράσει. Αξίζει να αναφέρουμε ότι τα χρηματοπιστωτικά ιδρύματα παράγουν ένα πολύ μεγάλο σύνολο δεδομένων και χτίζουν ένα σημαντικό θεμέλιο με στόχο να προσεγγίσουν τα δυναμικά και πολύπλοκα προβλήματα με τα κατάλληλα εργαλεία της εξόρυξης δεδομένων. Με την πάροδο των χρόνων έρχονται και σημαντικές εξελίξεις στην παραδοσιακή ανάλυση των καμπυλών του χρηματιστηρίου που χρησιμοποιούνται από τα χρηματοπιστωτικά ιδρύματα. Η εξόρυξη δεδομένων έχει την δική της θέση στην οικονομική

μοντελοποίηση. Υπάρχει μια ποικιλία γραμμικών και μη γραμμικών μοντέλων μερικά εκ των οποίων φαίνονται παρακάτω:

- Νευρωνικά δίκτυα
- K-means
- Δέντρο αποφάσεων
- Παλινδρόμηση
- Ανάλυση κυρίων συνιστωσών
- Μπευζιανή Μάθηση

Οι παραδοσιακότερες μέθοδοι περιλαμβάνουν ακατέργαστα σετ, σχεσιακή μέθοδο εξόρυξης δεδομένων (βλ. ντετερμινιστική επαγωγή), ανεξάρτητη συνιστώσα ανάλυσης, μοντέλα Markov και κρυμμένα μοντέλα Markov και τεχνικές αξιολόγησης Bootstrapping που έχουν χρησιμοποιηθεί πολύ για να βελτιώσουν τα αποτελέσματα της εξόρυξης δεδομένων. Ο πραγματικός τρόπος προσέγγισης είναι η χρήση μεθόδων σύγκρισης που να δείχνουν τόσο τα δυνατά όσο και τα αδύνατα σημεία του προβλήματος. Βέβαια ο χρήστης σε κάθε περίπτωση είναι αυτός που θα καταλήξει σε μια μέθοδο θα χρησιμοποιήσει ανάλογη με το πρόβλημα που καλείται να επιλύσει σε κάθε περίπτωση. Πειράματα που έχουν γίνει κατά καιρούς αποδεικνύουν την επιτυχία της εξόρυξης δεδομένων στον τομέα των οικονομικών καθώς και σε άλλους σημαντικούς τομείς λόγου χάρη τη γεωλογία και την ιατρική. Μπορούμε να κατανοήσουμε πως μια πρόβλεψη μπορεί να θέση σε σοβαρή δοκιμασία χωρίς το κόστος και τα κεφάλαια επιχειρηματικού κινδύνου να εμπλέκονται στην πραγματική διαπραγμάτευση. Εξέχουσα θέση στις οικονομικές εφαρμογές της εξόρυξης δεδομένων έχουν τα δέντρα αποφάσεων, ο k-means και τα νευρωνικά δίκτυα. Οι παραπάνω μέθοδοι είναι αρκετά απλοί και αποτελεσματικοί και δύναται να επεξεργαστούν και δεδομένα με πολύ θόρυβο. Οι μέθοδοι αυτοί έχουν ένα σημαντικό μειονέκτημα δεν μπορούν να εκπροσωπήσουν γνωστικό υπόβαθρο και δεν μπορούν να διαχειριστούν τις περίπλοκες σχέσεις. Αρχικά οι μέθοδοι αυτοί ήταν αρκετά αναποτελεσματικοί στο να υπολογίσουν αριθμητικά δεδομένα με την πάροδο των χρόνων ενισχύθηκαν και παρέχουν πολλές δυνατότητες. Την σημερινή εποχή οι προαναφερθείσες μέθοδοι έχουν θεωρηθεί κατάλληλες για χρηματικοοικονομική ανάλυση. Μέθοδοι όπως : ασαφής λογική, υβριδική αρχιτεκτονική σε νευρωνικά δίκτυα, θεωρία του χάους και οι γενετικοί αλγόριθμοι για την άμεση διαχείριση των

μοντέλων εμπορίας νευρωνικών δικτύων σύμφωνα με έρευνα έχει διατεθεί περίπου 5 με 10 δισεκατομμύρια δολάρια.

Οι ιδιαιτερότητες στην εξόρυξη δεδομένων όσον αφορά τη χρηματοδότηση προκύπτουν από:

- Πρόβλεψη πολυδιάστατων χρονοσειρών
- Χρονοσειρές με υψηλό επίπεδο θορύβου
- Φιλοξενία συγκεκριμένων κριτηρίων
- Συντονισμένη πρόβλεψη πολλαπλών αναλύσεων
- Εξήγηση πρόγνωσης και του μοντέλου πρόβλεψης
- Να είναι σε θέση να μπορούν να επωφεληθούν σε ένα σύντομο χρονικό διάστημα από τα πολύ λεπτά μοτίβα
- Να μπορούν να ενσωματώσουν την επίδραση των παραγόντων της αγοράς στις κανονικότητες της αγοράς.

Η σημερινή αποδοτική θεωρία αγοράς /υπόθεσης αποθαρρύνει ανακάλυψης μακροπρόθεσμων κανόνων σταθερές εμπορικές/ κανονικότητες με σημαντικό κέρδος. Η παραπάνω θεωρία βασίζεται στο ότι υπάρχουν τέτοιες κανονικότητες που θα χρησιμοποιούνται από πλειοψηφία παραγόντων της αγοράς. Αυτό θα καταστήσει τις αγορές λιγότερο κερδοφόρες, άχρηστες ή και επιβλαβείς. Οι δείκτες του λογισμικού ηλεκτρονικών υπολογιστών και συστημάτων πληροφορικής αντικατοπτρίσουν σημαντικές αποκλίσεις χρησιμοποιώντας το πρότυπο t-test. Κατέληξαν στο συμπέρασμα ότι και για τους Wall Street αναλυτές ισχύει ο κανόνας το καλοκαίρι λιποθυμούν. Δηλαδή τον Μάιο πωλούν για να ξαναγοράσουν το τέλος του καλοκαιριού. Η θεωρία της αποτελεσματικότητας της αγοράς δεν αποκλείει ότι κρύβονται βραχυπρόθεσμες τοπικές και ύπαρξη κανονικότητες όρων οι οποίες διαρθρώνονται συχνά. Η εξόρυξη δεδομένων δεν δύναται να αποδεχτεί ή απορρίπτει την αποτελεσματικότητα της θεωρίας της αγοράς . για τον λόγο αυτό δημιουργεί εργαλεία εξόρυξης δεδομένων τα οποία να ανακαλύψουν λεπτά βραχυπρόθεσμα μοτίβα όρων και τις τάσεις σε ευρύ φάσμα των οικονομικών δεδομένων. Όσον αφορά την εξόρυξη δεδομένων για τις οικονομικές εφαρμογές είναι γνωστή και ως προσέγγιση της χρηματοδότησης και χρησιμοποιεί την ορολογία "econophysic" (βλ.

φυσική της χρηματοδότησης). Η κύρια διαφορά της από την εξόρυξη δεδομένων είναι η προσέγγιση ότι η εξόρυξη δεδομένων δεν είναι κατάλληλη για την ανάπτυξη μεθόδων για οικονομικά καθήκοντα αλλά η προσέγγιση της φυσικής. Μπορούμε να διαπιστώσουμε ότι το μέλλον της εξόρυξης δεδομένων στον τομέα των οικονομικών θα δημιουργήσει περισσότερες εμπειρικές κανονικότητες συνδυαστικά με τις γνώσεις του τομέα μέσο της γενικής ανάλυσης δεδομένων. Όσον αφορά τις εφαρμογές των μοντέλων εξόρυξης δεδομένων τα καθήκοντα πρόβλεψης της χρηματοδότησης έχουν μια από τις δύο παρακάτω μορφές: Α)Ευθεία πρόβλεψη της αγοράς με αριθμητικά χαρακτηριστικά (βλ. απόδοση, χρ. Ισοτιμία), Β)Πρόβλεψη αν η χρηματιστηριακή βούληση της αγοράς αυξάνεται ή μειώνεται

Γνωρίζοντας ότι πρέπει να υπολογίσουμε τού κόστος των συναλλαγών καθώς και τη σημασία της επιστροφή των συναλλαγών στη δεύτερη περίπτωση χρειαζόμαστε μια πρόβλεψη για το αν το χαρακτηριστικό της αγοράς θα μειωθεί ή θα αυξηθεί αντίστοιχα όχι λιγότερο από ένα συγκεκριμένο όριο. Συνεπώς η διαφορά ανάμεσα στις δυο παραπάνω μεθόδους είναι λιγότερο εμφανής διότι στην δεύτερη περίπτωση μπορεί να απαιτείται κάποιο είδος αριθμητικής πρόβλεψης.

Ένα παράδειγμα άλλου τύπου παρουσιάζεται στο έργο Becerra-Fernandez et. Al., 2002. Η μελέτη έχει να κάνει με την αξιολόγηση του επενδυτικού κινδύνου. Χρησιμοποιείται ένα δέντρο απόφασης και πιο συγκεκριμένα η τεχνική C5.0 και των νευρωνικών δικτύων σε ένα σύνολο από 52 χώρες όπου ο επενδυτικός κίνδυνος αξιολογήθηκε σε μια έρευνα της Wall Street Journal από διεθνείς εμπειρογνώμονες . στο σύνολο των δεδομένων περιλαμβάνονταν 27 μεταβλητές.[12]

ΚΕΦΑΛΑΙΟ 3

3.1 Εντοπισμός και μέτρηση χάσματος

Για τον εντοπισμό και και την μέτρηση χάσματος χρησιμοποιήσαμε αναλυτικές τιμές τριών δεικτών όπως θα δούμε αναλυτικά στο επόμενο κεφάλαιο, σε ένα φύλλο excel με στόχο να μελετήσουμε το χάσμα (gap) των δεικτών που πιθανώς να εμφανίζουν από το κλείσιμο των χρηματιστηρίων της παρασκευής έως το άνοιγμα της Κυριακής. Επεξεργαστήκαμε τα απαραίτητα δεδομένα όπως όνομα δείκτη, ημερομηνία, ώρα, τιμή δείκτη στο κλείσιμο της παρασκευής, τιμή δείκτη στο άνοιγμα της Κυριακής με σκοπό να μελετήσουμε το χάσμα σε όποια περίπτωση υπάρχει. Έτσι για την μελέτη του χάσματος δημιουργήσαμε μια binary κλάση τριών τιμών :

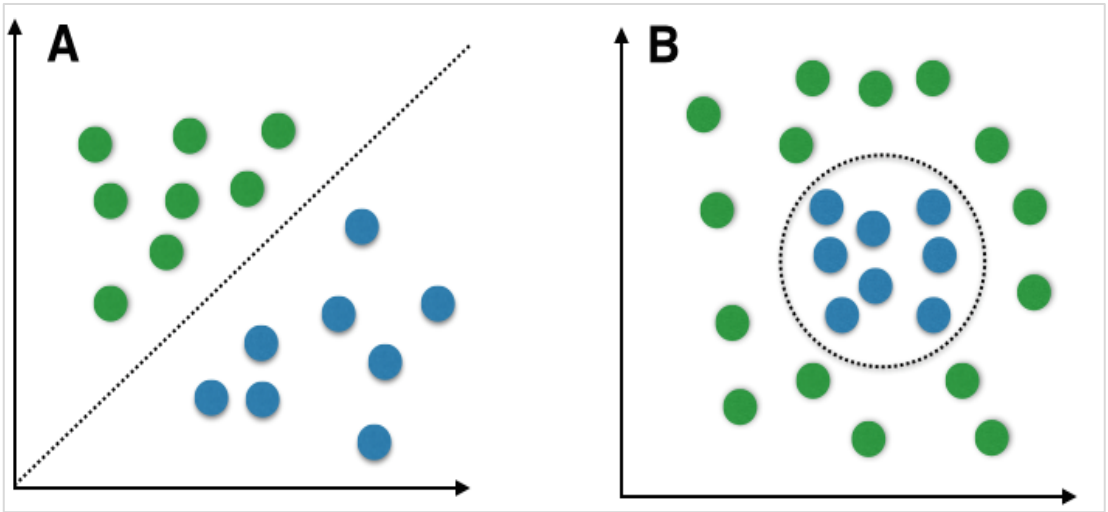
- "0", Εάν το πρώτο μισάωρο της Κυριακής συνεχίστηκε η φορά του χάσματος
- "1", Εάν η φορά του χάσματος υπέστη αντιστροφή είτε προς τα επάνω είτε προς τα κάτω αλλά όχι πολύ (<50%)
- "2", Εάν η φορά του χάσματος υπέστη αντιστροφή μεγαλύτερη του 50%

3.2 Μέθοδοι Ταξινόμησης

3.2.1 Μπευζιανός Ταξινομιτής Naive-Bayes

Ο Μπευζιανός Ταξινομιτής- Naïve Bayes σε μηχανική μάθηση αποτελεί μια ομάδα πιθανολογικών ταξινομητών η οποία κυρίως βασίζεται στην εφαρμογή του θεωρήματος Bayes . Αξίζει να αναφέρουμε ότι το παραπάνω θεώρημα στηρίζεται στις ισχυρές υποθέσεις ανεξαρτησίας μεταξύ των χαρακτηριστικών. Ο ταξινομητής Naïve Bayes έκανε τα πρώτα βήματα εμφάνισης του το 1950. Αρχικά το όνομα που του δόθηκε ήταν διαφορετικό. Πιο συγκεκριμένα από την κοινότητα ανάκτησης κειμένου δόθηκε η ονομασία 1:488 η οποία παραμένει ακόμη και σήμερα σαν μια εξαιρετικά γνωστή μέθοδος κατηγοριοποίησης κειμένου, δηλαδή για το εάν κρίνουμε έγγραφα τα οποία βρίσκονται στην μια ή την άλλη κατηγορία.. Θα πρέπει να επισημάνουμε ότι με την κατάλληλη προεπεξεργασία αποτελεί μια πλέον

ανταγωνιστική μέθοδο στην κατηγορία των προηγμένων μεθόδων όπου ανήκει και για τον λόγο ότι συμπεριλαμβάνει και μηχανές υποστήριξης. Έχει βοηθήσει εξαιρετικά στην αυτόματη διάγνωση στον τομέα της ιατρικής. Όσον αφορά την επεκτασιμότητα χαρακτηρίζονται ως πολύ επεκτάσιμοι και για την ακρίβεια απαιτούν μια σειρά από παραμέτρους γραμμικής του αριθμού των μεταβλητών σε ένα πρόβλημα εκμάθησης. Χαρακτηριστικό παράδειγμα αποτελεί η εκπαίδευση μέγιστης πιθανότητας η οποία δύναται να πραγματοποιηθεί με μια έκφραση κλειστής μορφής [1]:718 η οποία λαμβάνει γραμμικό χρόνο και σε καμία περίπτωση με επαναληπτική προσέγγιση η οποία λαμβάνει χώρα σε πολλούς τύπους ταξινομητών. Στην παραπάνω εικόνα διακρίνουμε γραμμικά έναντι μη γραμμικών προβλημάτων, παρουσιάζονται τυχαία δείγματα των δύο κατηγοριών σαν χρωματιστές σφαίρες και η διακεκομμένη γραμμή καθορίζει τα όρια τάξης του ταξινομητή που καλείται να προσεγγίσει υπολογίζοντας τα όρια αποφάσεων. Στην δεύτερη περίπτωση βλέπουμε ένα μη γραμμικό πρόβλημα όπου οι γραμμικοί ταξινομητές λόγου χάρη ο Μπευζιανός Ταξινομητής- Naïve Bayes δεν είναι κατάλληλοι για το λόγο ότι οι κατηγορίες δεν χωρίζονται γραμμικά. Σε αυτή την περίπτωση προτιμάται η χρήση μη γραμμικών ταξινομητών όπως για παράδειγμα ο ταξινομητής με βάση τον πλησιέστερο γείτονα.[13].

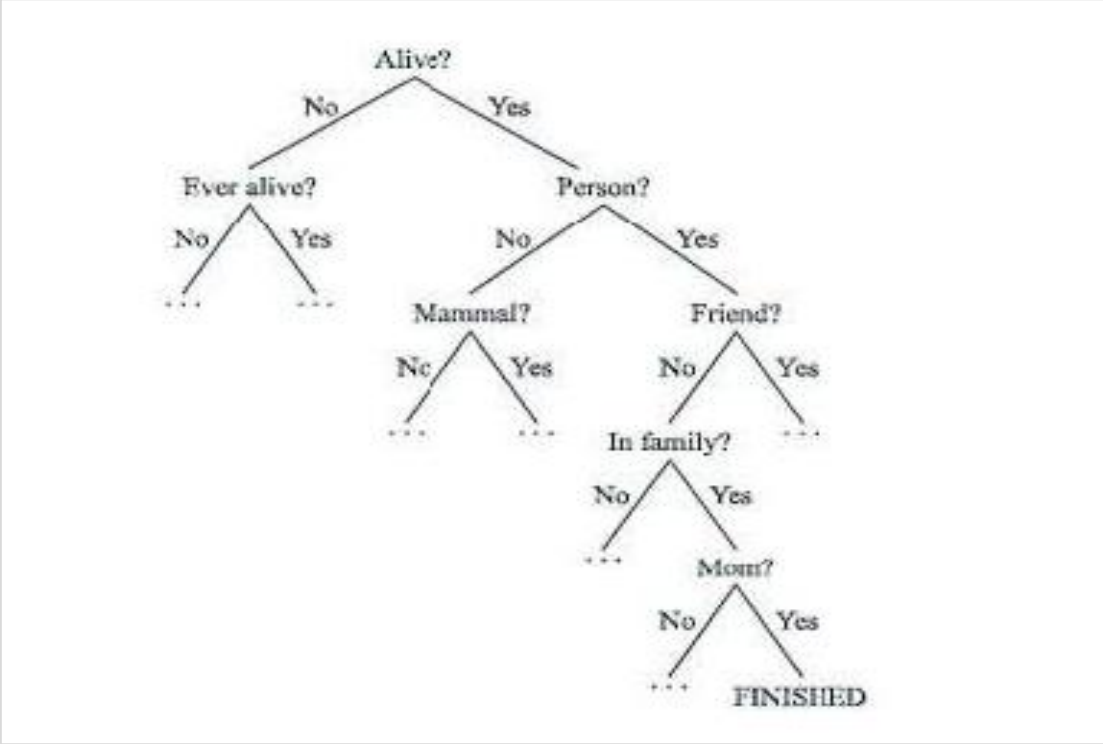


Εικόνα 3.1 : Μπευζιανός Ταξινομητής (Naive-Bayes)

3.2.2 Δένδρα Απόφασης-Decision Trees

Με τον όρο δέντρο αποφάσεων αναφερόμαστε σε ένα εργαλείο λήψης αποφάσεων το οποίο κατά κύριο λόγο χρησιμοποιεί μια απεικόνιση- γραφική αναπαράσταση που έχει τη μορφή δέντρου. Το συγκεκριμένο δέντρο περιλαμβάνει:

- Όλες τις πιθανές αποφάσεις
- Όλους τους παράγοντες επιρροής
- Όλα τα πιθανά αποτελέσματα

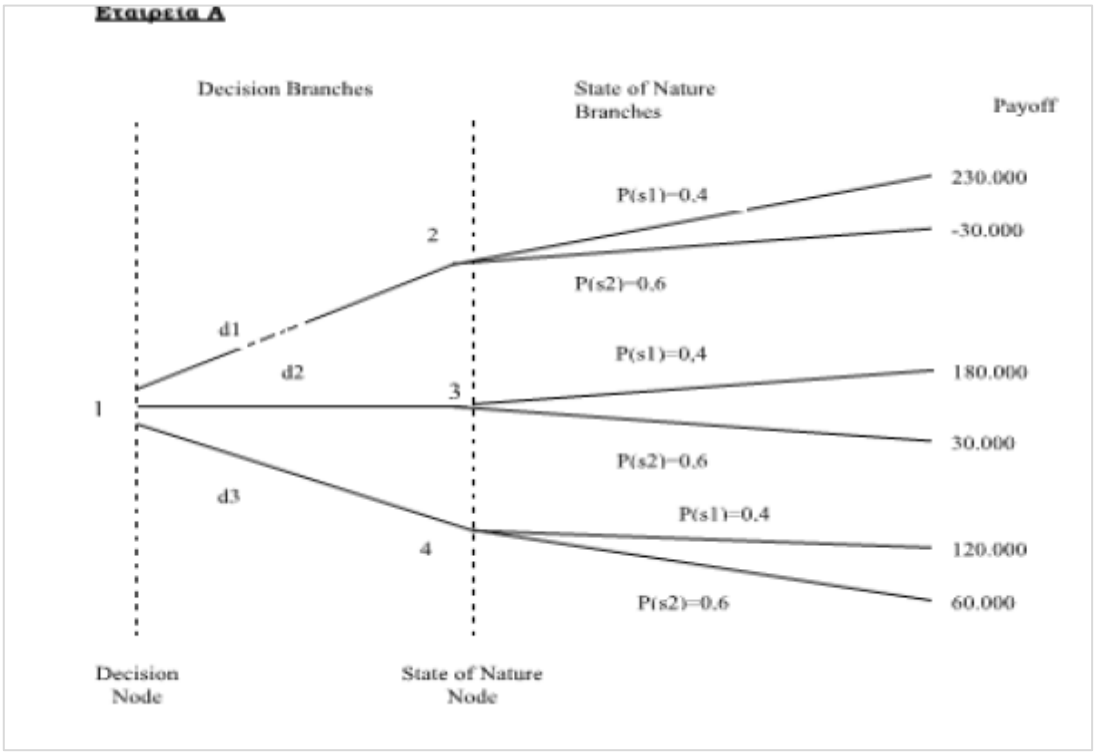


Εικόνα 3.2 : Παράδειγμα ενός Decision Tree

Τα δέντρα αποφάσεων δομούνται με μια διαδικασία που αποτελείται από δύο βήματα:

- Προς τα εμπρός πορεία
- Προς τα πίσω πορεία

Όσον αφορά το πρώτο βήμα δηλαδή το προς τα εμπρός πορεία έχει να κάνει με τον λήπτη αποφάσεων στην διαδικασία ταυτοποίησης αποφάσεων, των γεγονότων που πρέπει να συμβούν και τη σειρά με την οποία θα γίνουν. Πιο συγκεκριμένα το πρώτο βήμα έχει να κάνει τόσο με τη δομή όσο και με τον υπολογισμό των απολαβών και των πιθανοτήτων που αφορούν τα μελλοντικά γεγονότα. Όσον αφορά το δεύτερο βήμα δηλαδή την πορεία προς τα πίσω έχει άμεση σχέση με την ανάλυση του προβλήματος. Κατά συνέπεια θα πρέπει να υπολογιστούν οι προστιθέμενες αξίες ώστε να οδηγηθούμε στην τελική απόφαση.



Εικόνα 3.3 : Παράδειγμα Εταιρείας Α

Παράδειγμα

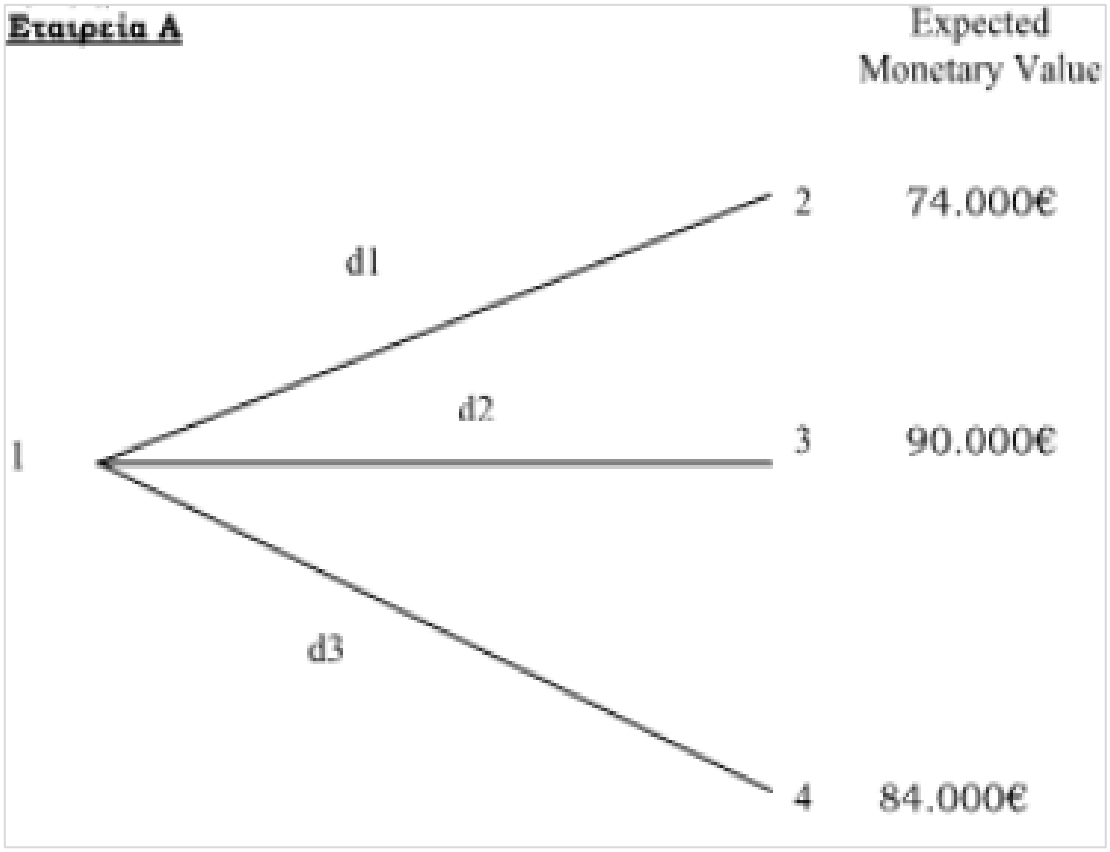
Στο παράδειγμα μας έχουμε μια υποθετική εταιρεία Α όπου οι εναλλακτικές αποφάσεις για την Α είναι d1, d2, d3 και οι παράγοντες επιρροής (states of nature). s1, s2. Επίσης θα προσθέσουμε ότι οι πιθανότητες που αφορούν τους παράγοντες επιρροής είναι: P (s1)=0,4 P(s2)=0,6. Σκοπός μας είναι να καταλήξουμε στη βέλτιστη απόφαση. Αυτό θα το επιτύχουμε τις πιθανότητες κλαδιών καθώς και τα προσδοκώμενες αξίες. Όπως αναφέραμε και σε άλλο σημείο της παρούσας εργασίας στα δέντρα αποφάσεων πάντα δουλεύουμε από πίσω προς τα μπροστά κατά συνέπεια θα ξεκινήσουμε με την προσδοκώμενη αξία για καθέναν από τους κόμβους επιρροής.

Κάνοντας χρήση των απολαβών που καταλήγει κάθε κλαδί παράγοντα επιρροής και τις πιθανότητες που αντιστοιχούν σε αυτά οι προσδοκώμενες αξίες που αντιστοιχούν στους κόμβους 2,3,4 φαίνονται στην εικόνα που ακολουθεί:

EMV(node 2)= 0,4(230.000) + 0,6(-30.000) = 92.000 – 18000= 74.000€
EMV(node 3)= 0,4(180.000) + 0,6(30.000) = 72.000 + 18000= 90.000€
EMV(node 4)= 0,4(120.000) + 0,6(60.000) = 48.000 + 36.000= 84.000€

Εικόνα 3.4 : Προσδοκώμενες αξίες των κόμβων 2,3,4

Στην συνέχεια εργαζόμαστε προς τα πίσω κατά μήκος του δέντρου και συγκεκριμένα στον κόμβο αποφάσεων. Υπό την προϋπόθεση ότι οι προστιθέμενες αξίες των κόμβων 2,3,4 είναι γνωστές τότε ο αναλυτής ή αλλιώς εκείνος ο οποίος καλείται να λάβει την απόφαση οφείλει να παρατηρήσει τον κόμβο απόφασης 1 όπως φαίνεται την παρακάτω εικόνα:



Εικόνα 3.5 : Κόμβος Απόφασης 1

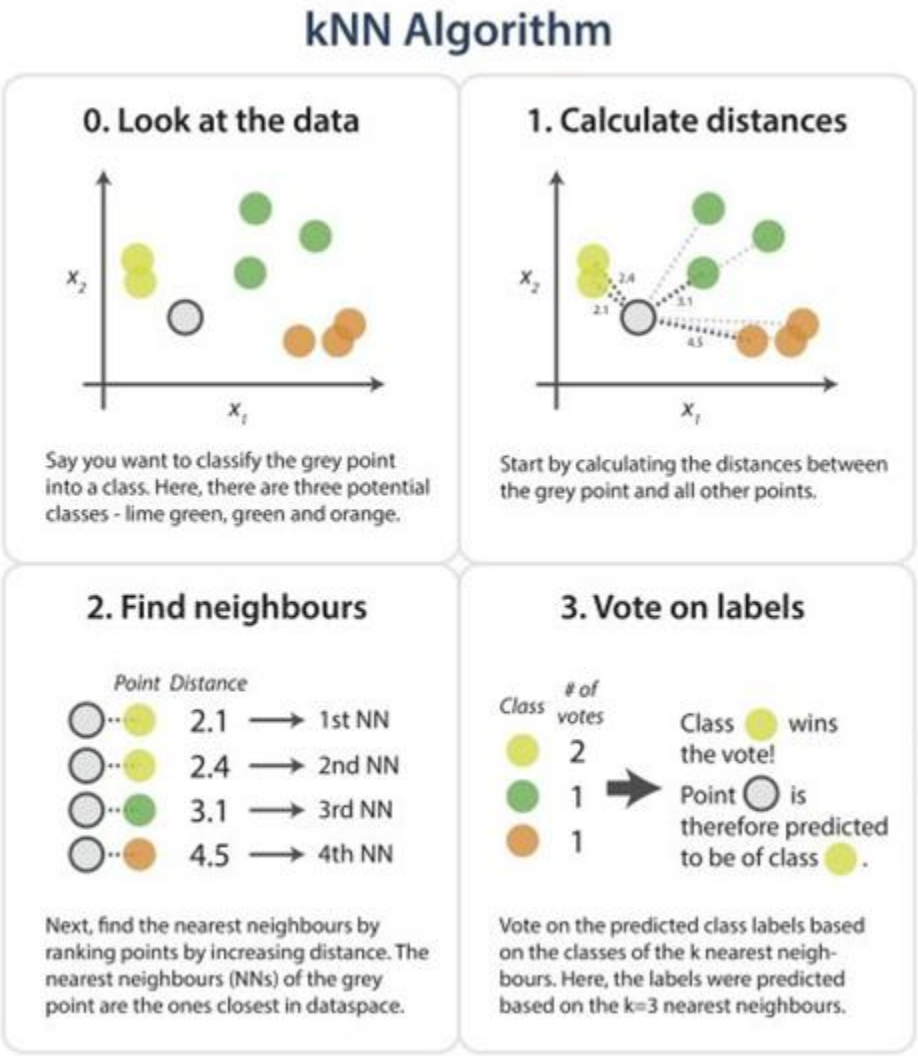
Συνεπώς από τη στιγμή που έχουν υπολογιστεί οι προστιθέμενες αξίες της κάθε απόφασης και στόχος είναι η μεγιστοποίηση κέδρους της εταιρείας Α είναι εμφανές ότι η εταιρεία Α θα καταφύγει στην απόφαση d2 δηλαδή στην επιλογή του κόμβου 3 όπως φαίνεται και στην παραπάνω εικόνα η προσδοκώμενη αξία είναι 90000 ευρώ .[14] [15]

3.2.3 Αλγόριθμος kNN (k-nearest neighbors)

Ο αλγόριθμος kNN (k-Nearest Neighbors) ανήκει στην κατηγορία των "lazy learning algorithms",καθότι χρειάζεται ένα σετ στοιχείων χρησιμοποιείται μονάχα για να γεμίσει ένα δείγμα του διαστήματος αναζήτησης που του ζητείται με αντικείμενα που η κλάση τους είναι γνωστή και για τον λόγο ότι δεν χρησιμοποιείται κάποιο μοντέλο ή γνώση κατά τη διάρκεια αυτής της διαδικασίας και έτσι κατηγοριοποιείται στους "τεμπέληδες" αλγορίθμους.

Είναι μη παραμετρικός αλγόριθμος και με αυτό εννοούμε πως δεν κάνει υποθέσεις σε υποβοσκον διανομή δεδομένων, αυτό είναι χρήσιμο γιατί όπως και στον πραγματικό κόσμο τα περισσότερα από τα πρακτικά δεδομένα που έχουμε δεν υπακούν τις τυπικές θεωρητικές εικασίες (πχ Gaussian mixture model).

Στην πράξη ο kNN υποθέτει πως τα δεδομένα βρίσκονται σε ένα "feature space" η αλλιώς η συλλογή των χαρακτηριστικών που χρησιμοποιούνται για να χαρακτηρίσουν τα δεδομένα μας όπως για παράδειγμα ύψος, φύλο, χρώμα. Από τη στιγμή που τα σημεία βρίσκονται στο feature space έχουν μια αίσθηση της απόστασης χωρίς απαραίτητα να κάνουμε λόγο μονάχα για Ευκλίδεια απόσταση αν και είναι η περισσότερο συχνά χρησιμοποιούμενη. Έτσι έχουμε ένα αριθμό "k" ο οποίος υπολογίζει πόσοι γείτονες επηρεάζουν την ταξινόμηση (όπου οι γείτονες επηρεάζονται βάση της απόστασης στο μετρικό σύστημα).Πρακτικά αυτό που συμβαίνει είναι ότι στο σύνολο δεδομένων εκμάθησης ο αλγόριθμος kNN επιλέγει τα k κοντινότερα σημεία από ένα σημείο ελέγχου.[17]



Εικόνα 3.6 : Βήματα λειτουργίας του kNN

3.2.4 Τυχάια δένδρα (Random Forests)

Μία μέθοδος ταξινόμησης συνδυαστικού τύπου είναι τα τυχαία δάση τα οποία στην ουσία αποτελούνται από μια συστάδα δένδρων απόφασης/decision trees όπως είδαμε σε προηγούμενη ενότητα. Ο τρόπος με τον οποίο λειτουργούν τα τυχαία δάση βασίζεται στην ανάπτυξη πολλών δένδρων αποφάσεων, το κάθε δένδρο πραγματοποιεί μια ταξινόμηση η οποία καταλήγει στην προτίμηση μιας κλάσης, οι κλάσεις επομένως έχουν ένα αριθμό "προτιμήσεων" από τα δένδρα αποφάσεων. Η τελική ταξινόμηση γίνεται με το τυχαίο δάσος να επιλέγει την κλάση εκείνη η οποία είχε τις περισσότερες "προτιμήσεις" , έτσι εύλογα καταλήγουμε στο συμπέρασμα πως η ανάπτυξη των decision trees παίζει καταλυτικό ρόλο στη λειτουργία των τυχαίων

δασών.

Υπάρχουν κάποιες μεταβλητές που επηρεάζουν την λειτουργία των τυχαίων δασών :

- Συσχέτιση μεταξύ των δένδρων
- Η δυναμική του κάθε δένδρου

Όταν δένδρα παρουσιάζουν μεταξύ τους αρκετές ομοιότητες άρα συσχετίζονται επίσης αρκετά, ο ρυθμός λάθους που μπορεί να κάνει το δάσος στην ταξινόμηση αυξάνεται. Βέβαια βάση της τεχνικής που εφαρμόζεται (bootstrap sampling) η οποία φροντίζει να κρατάει χαμηλό τον ρυθμό λάθους καθώς χρησιμοποιεί συχνά διαφορετικά διανύσματα εισαγωγής των δεδομένων με στόχο την διαφοροποίηση των δένδρων, ο ρυθμός εμφάνισης δένδρων που παρουσιάζουν μεγάλο βαθμό συσχέτισης περιορίζεται στο ελάχιστο. Η δυναμική του κάθε δένδρου σημαίνει πρακτικά το πόσο καλός ταξινομητής είναι το εκάστοτε δένδρο, δένδρα τα οποία έχουν μικρό ρυθμού λάθους στην ταξινόμηση θεωρούνται καλοί ταξινομητές επομένως ένα τυχαίο δάσος όσο περισσότερα δένδρα "δυναμικά" διαθέτει τόσο μικρότερη η πιθανότητα για λάθος υπολογισμό [15][16].

3.3 Μέτρα Αξιολόγησης

Σύμφωνα με τους Kohavi και Provost το 1998 ένας πίνακας σύγχυσης όπως αλλιώς καλείται Confusion Matrix είναι ένας πίνακας που περιέχει πληροφορίες σχετικά με την απόδοση της ταξινόμησης σε ένα σύστημα ταξινόμησης σε ένα σύστημα δεδομένων δοκιμής για τις οποίες είναι γνωστές οι πραγματικές αξίες. Ακολουθεί ένα παράδειγμα (μήτρα σύγχυσης) για έναν δυαδικό ταξινομητή που μπορεί να επεκταθεί για δύο ή και περισσότερες κατηγορίες.

n=165	**Predicted: NO**	**Predicted: YES**
Actual: NO	50	10
Actual: YES	5	100

Εικόνα 3.7 : Confusion Matrix-Μήτρα Σύγχυσης(1)

Όπως μπορούμε να παρατηρήσουμε στην παραπάνω εικόνα υπάρχουν δύο πιθανές

προβλεπόμενες κατηγορίες αυτή του Yes και αυτή του No. ο ταξινομητής κάνει 165 προβλέψεις (βλ. n=165). Όπως μπορούμε να παρατηρήσουμε Yes 110 και No μόνο 55.

Στη συνέχεια θα ορίσουμε κάποιους όρους που είναι θετικοί αριθμοί και όχι ποσοστά.

- Αληθώς θετικά (TP): πρόκειται για περιπτώσεις με πρόβλεψη Yes
- Αληθώς αρνητικά (TN): πρόκειται για περιπτώσεις με πρόβλεψη No
- False positives (FP): πρόκειται για περιπτώσεις με πρόβλεψη Yes (γνωστό ως σφάλμα τύπου |).
- False negatives (FN): πρόκειται για περιπτώσεις με πρόβλεψη «κανείς» (γνωστό και ως σφάλμα τύπου ||).

n=165	**Predicted: NO**	**Predicted: YES**	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Εικόνα 3.8 : Confusion Matrix-Μήτρα Σύγχυσης (2)

Συχνά από τη μήτρα σύγχυσης υπολογίζεται και μια λίστα ποσοστών όπως:

- Η ακρίβεια (TP+TN/ σύνολο)
- Η λανθασμένη ταξινόμηση βαθμολογίας (Error Rate) (FP+FN/σύνολο)
- Αληθινός θετικός ρυθμός (TP/πραγματικά Yes)
- Ποσοστό ψευδώς θετικών(FP/ πραγματικά κανένα)
- Ιδιαιτερότητα (TN/ πραγματικά κανένα)
- Ακρίβεια (TP)
- Επικράτηση

ΚΕΦΑΛΑΙΟ 4

4.1 Σύνολο Δεδομένων

Στην εργασία χρησιμοποιήθηκαν 3 αρχεία δεικτών με τιμές ανά μισάωρο τα τελευταία 5 χρόνια :

- Ισοτιμία Ευρώ-Δολάριο (EURUSD)
- Ισοτιμία Λίρα Αγγλίας-Δολάριο (GBPUSD)
- Τιμή χρυσού σε δολάρια (XAAUSD)

Ο κάθε δείκτης εμπεριέχει τιμές όπως ημερομηνία, ώρα, τιμή ανοίγματος του δείκτη το τρέχων μισάωρο, χαμηλότερη τιμή στο μισάωρο, υψηλότερη τιμή στο μισάωρο και τιμή κλεισίματος του μισάωρο.

Καταγράψαμε στο φύλλο excel σε στήλες τα :

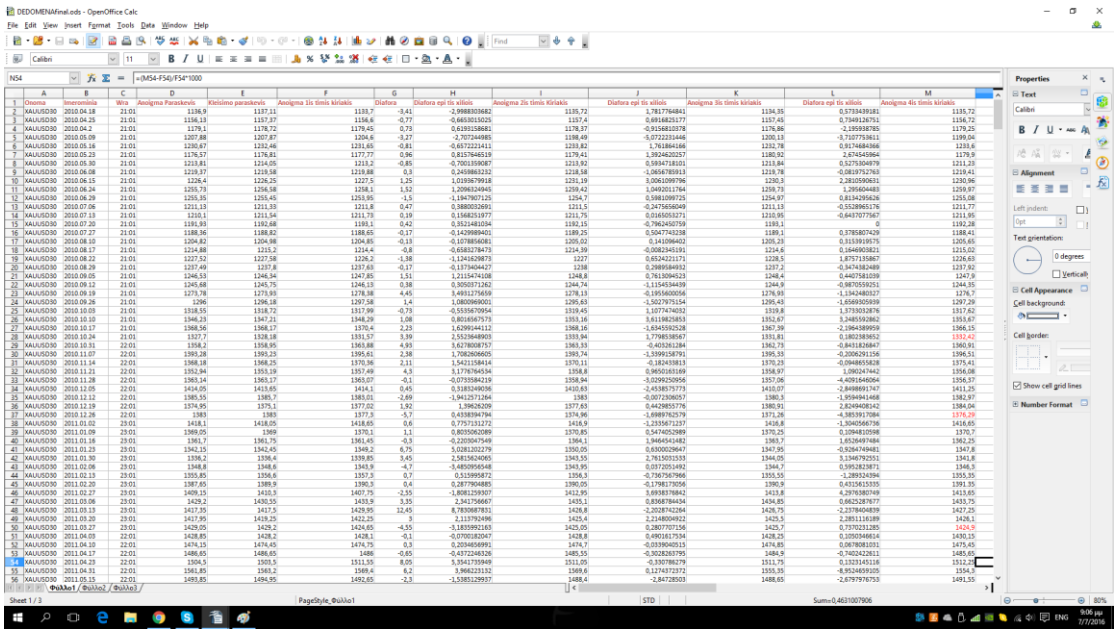
- 'Ονομα δείκτη
- Ώρα
- Ημερομηνία
- Άνοιγμα Παρασκευής
- Κλείσιμο Παρασκευής
- Άνοιγμα πρώτης, Δεύτερης, τρίτης, τέταρτης και πέμπτης τιμής της Κυριακής
- Χάσμα (μεταξύ κλεισίματος Παρασκευής και άνοιγμα Κυριακής)
- Χάσμα επι της χιλίοις (Χάσμα/Κλείσιμο Παρασκευής *1000)
- Χάσμα επι της χιλίοις για τις μετέπειτα τιμές της Κυριακής (δεύτερης, τρίτης, τέταρτης και πέμπτης)
- Class

Για την διαμόρφωση της κλάσης δημιουργήθηκαν 5 στήλες :

- ABS
- Max ABS

- Διαφορά %
- Διατηρείται
- Διαφορά > 50%

Αυτό που πραγματώνουν αυτές οι στήλες είναι ο έλεγχος δύο τιμών (χάσμα επι της χιλίοις της πρώτης τιμής Κυριακής και χάσμα επι της χιλίοις της δεύτερης τιμής της Κυριακής) εάν είναι και οι δύο θετικές ή αρνητικές, τις προσθέσαμε πήραμε την απόλυτη τιμή τους και ελέγξαμε αν το αποτέλεσμα είναι μεγαλύτερο απο την απόλυτη τιμή από τις δύο τιμές. Αυτό κάνουν οι στήλες "ABS","MAXABS" & "Διατηρείται" βάζοντας "0" εάν δεν αντιστράφηκε η πτώση ή η άνοδος και "1" εάν αντιστράφηκε. Στην συνέχεια ελέγχουμε που αντιστράφηκε η πτώση ή η άνοδος και σε τι ποσοστό έγινε αυτό. Για παράδειγμα εάν δούμε "1" στην στήλη "Διατηρείται" υπολογίζουμε το ποσοστό της αλλαγής στη στήλη "Διαφορά %". Στην στήλη "Διαφορά > 50%" βάζουμε "1" εάν η ποσοστιαία διαφορά είναι πάνω από 50%. Και τέλος προσθέτοντας τις στήλες "Διατηρείται" και " Διαφορά >50%" παίρνουμε την τελική κλάση.



Πίνακας 4.1 : Excel Table

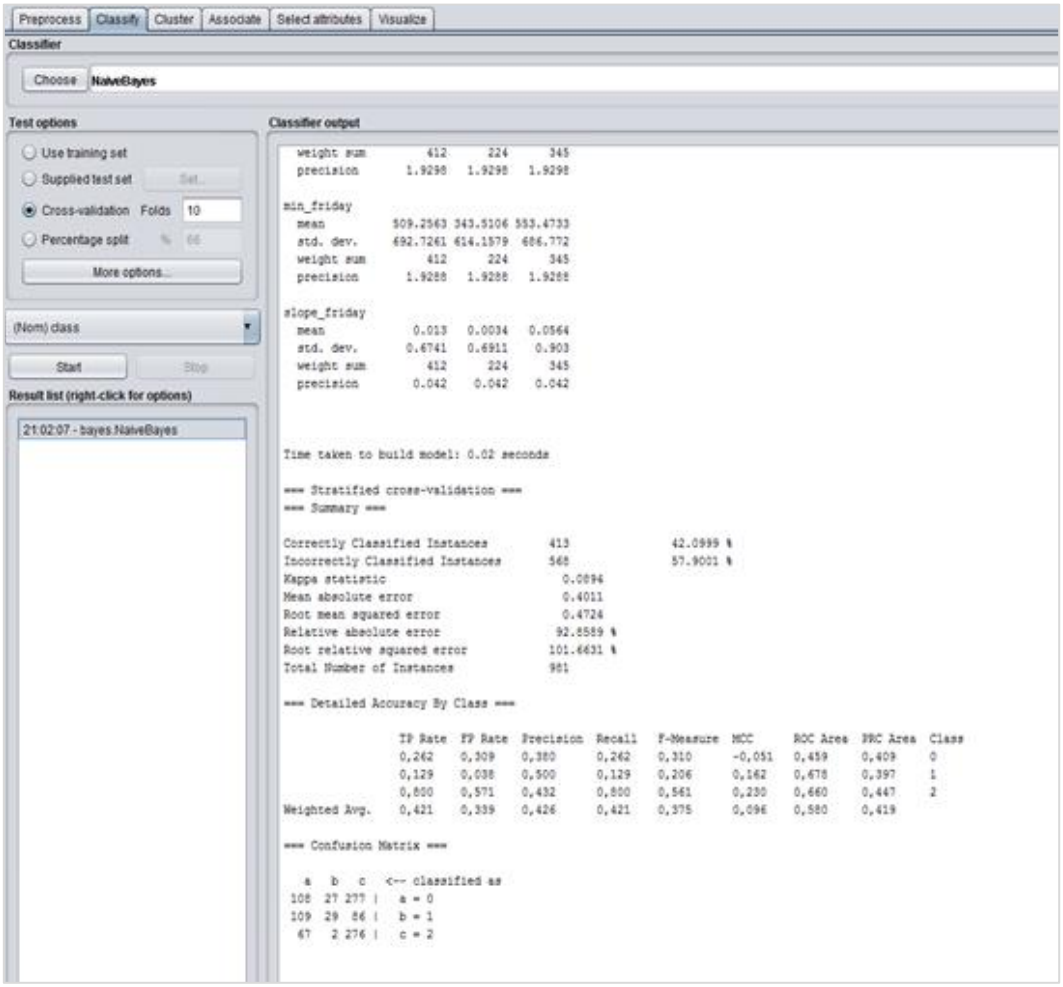
4.2 Περιγραφή Πειραμάτων

Σε αυτή την ενότητα θα δούμε τα αποτελέσματα της ταξινόμησης που μας έδωσαν οι τρείς αλγόριθμοι που συζητήσαμε σε περιβάλλον Weka,software κατάλληλο για data mining.Δημιουργήσαμε δύο ειδών προβλήματα, ένα πρόβλημα 2 κατηγοριών, σύμφωνα με το οποίο μελετάμε αν αναστράφηκε απλώς η πορεία του χάσματος και ένα πρόβλημα 3 κατηγοριών στο οποίοι έχουμε εισάγει και μια κατηγορία παραπάνω όπου η αναστροφή του χάσματος ξεπερνά το 50% του ίδιου χάσματος. Για να γίνει κατανοητή η επιτυχία ή μη των πειραμάτων αρκεί να αναλογιστούμε ποια θα ήταν η επιτυχία μας σε περίπτωση όπου τυχαίως προσπαθούσαμε να προβλέψουμε την πορεία του δείκτη. Είναι εύκολα κατανοητό ότι σε ένα πρόβλημα με δύο κλάσεις η τυχαία πρόβλεψη μας θα είχε κατά 50% επιτυχία, ενώ σε ένα πρόβλημα 3 κατηγοριών το τυχαίο ποσοστό επιτυχίας θα ήταν 33,3%.

4.2.1 Πρόβλημα τριών κλάσεων (3-class problem)

1) Naive Baiyes

Στην θεωρία οι baiyesian ταξινομητές έχουν το μικρότερο ρυθμό λάθους σε σχέση με άλλους ταξινομητές, παρολαυτά στην πράξη αυτό δεν ισχύει πάντοτε για τον λόγο ότι υπάρχουν σφάλματα που γίνονται στις υποθέσεις. Από την άλλη οι baiyesian ταξινομητές είναι αρκετά χρήσιμοι διότι παράγουν μεγάλη απόδοση και ταχύτητα όταν εφαρμόζονται σε μεγάλες βάσεις δεδομένων.

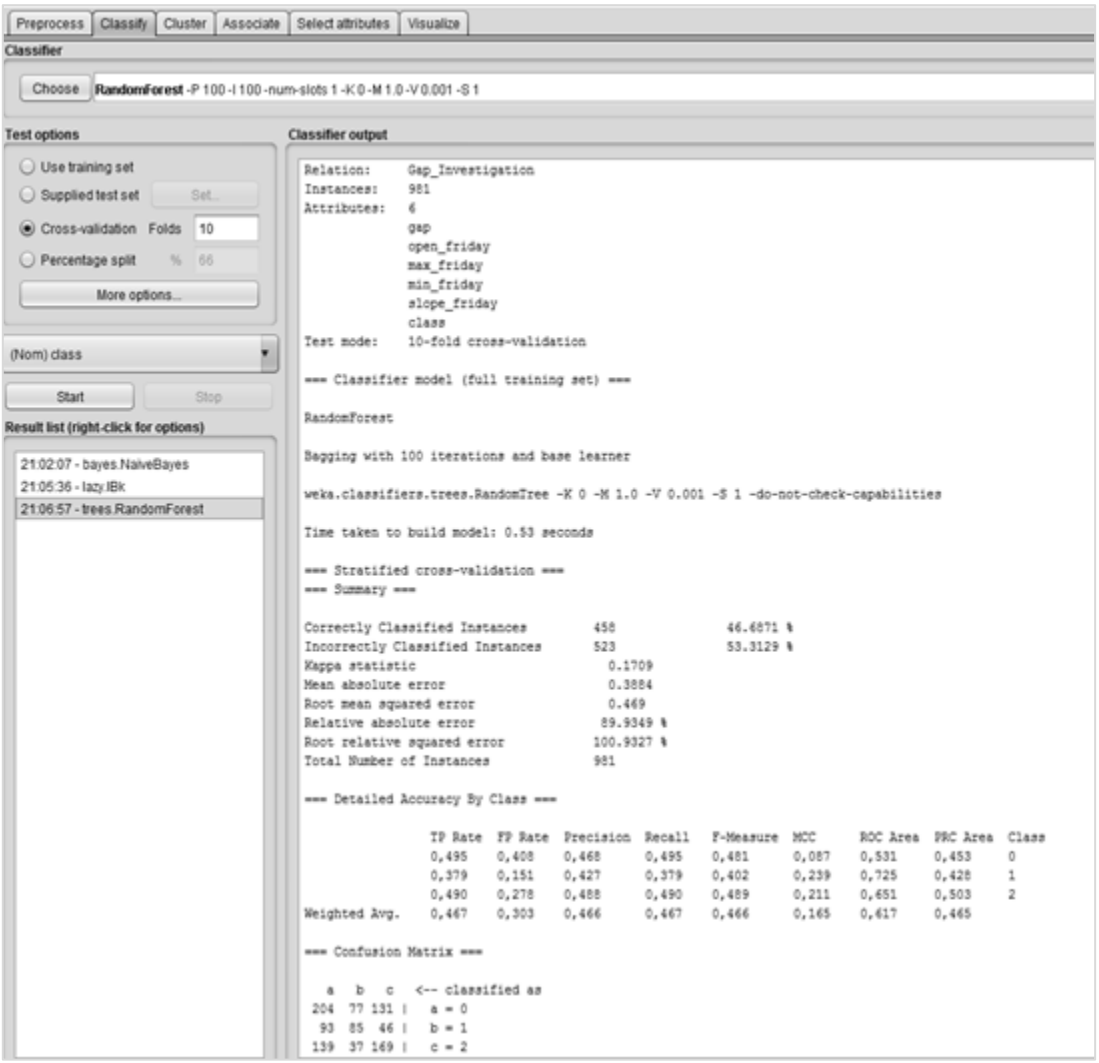


Εικόνα 4.1 : Αποτελέσματα Naive-Bayes

Correctly Classified Instances			**42.1%**
	Pred. Class 1	**Pred. Class 2**	**Pred. Class 3**
Class 1	108	27	277
Class2	109	29	86
Class3	47	2	276

2) Random Forest (Τυχαία Δάση)

Τα τυχαία δάση όπως είδαμε σε προηγούμενη ενότητα χρησιμοποιούν πολλά διαφορετικά decision trees,δημιουργώντας πληθώρα training sets (κάθε δένδρο έχει το δικό του) ώστε να παραχθεί μια οριστική απόφαση/ταξινόμηση με βάση τη κλάση που προτιμάται από τα περισσότερα trees.

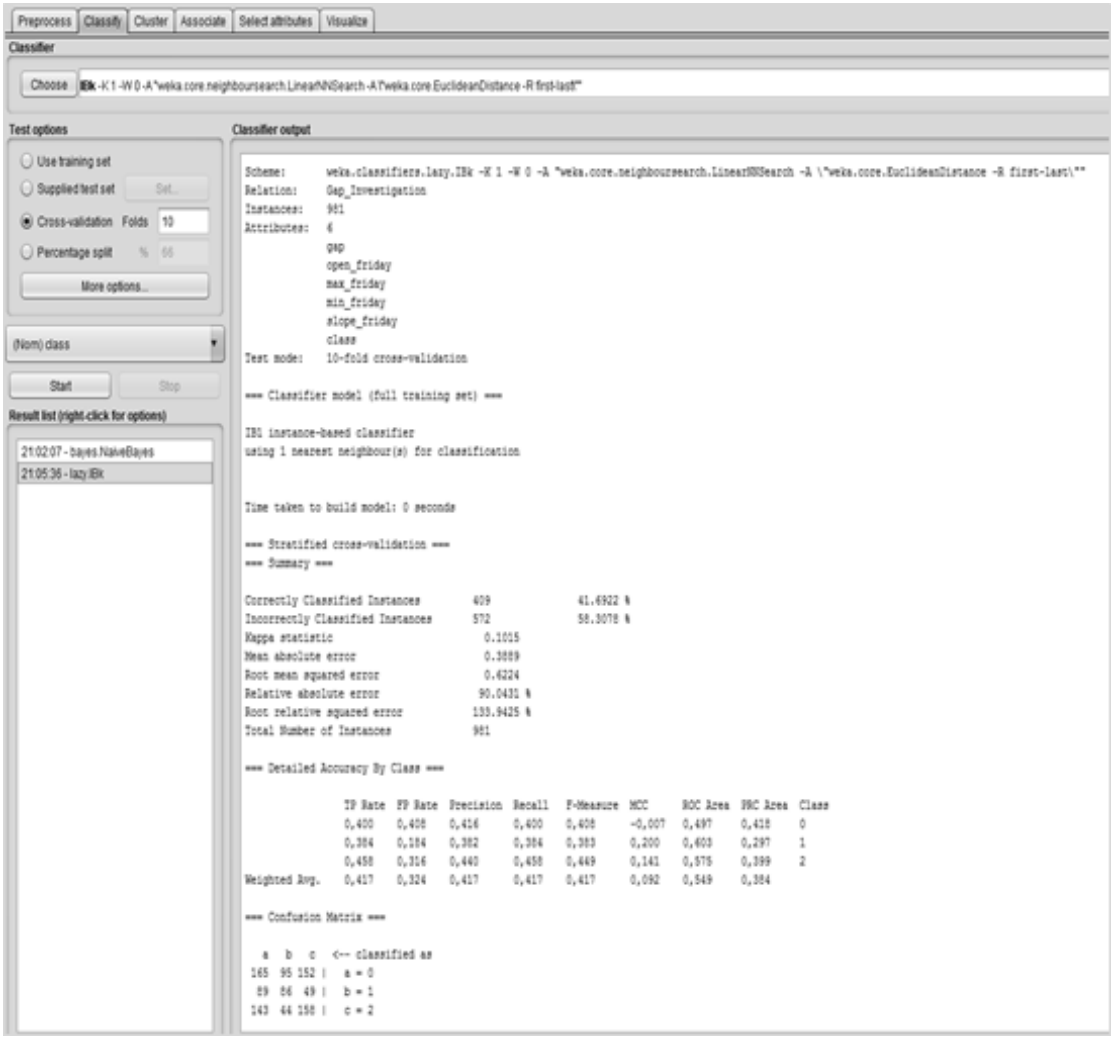


Εικόνα 4.2 : Αποτελέσματα Random Forest

Correctly Classified Instances			**46.6%**
	Pred. Class 1	**Pred. Class 2**	**Pred. Class 3**
Class 1	204	7	131
Class2	93	85	46
Class3	139	37	169

3) kNN (k-Nearest Neighbor)

Η ταξινόμηση βάση των k-κοντινότερων γειτόνων :



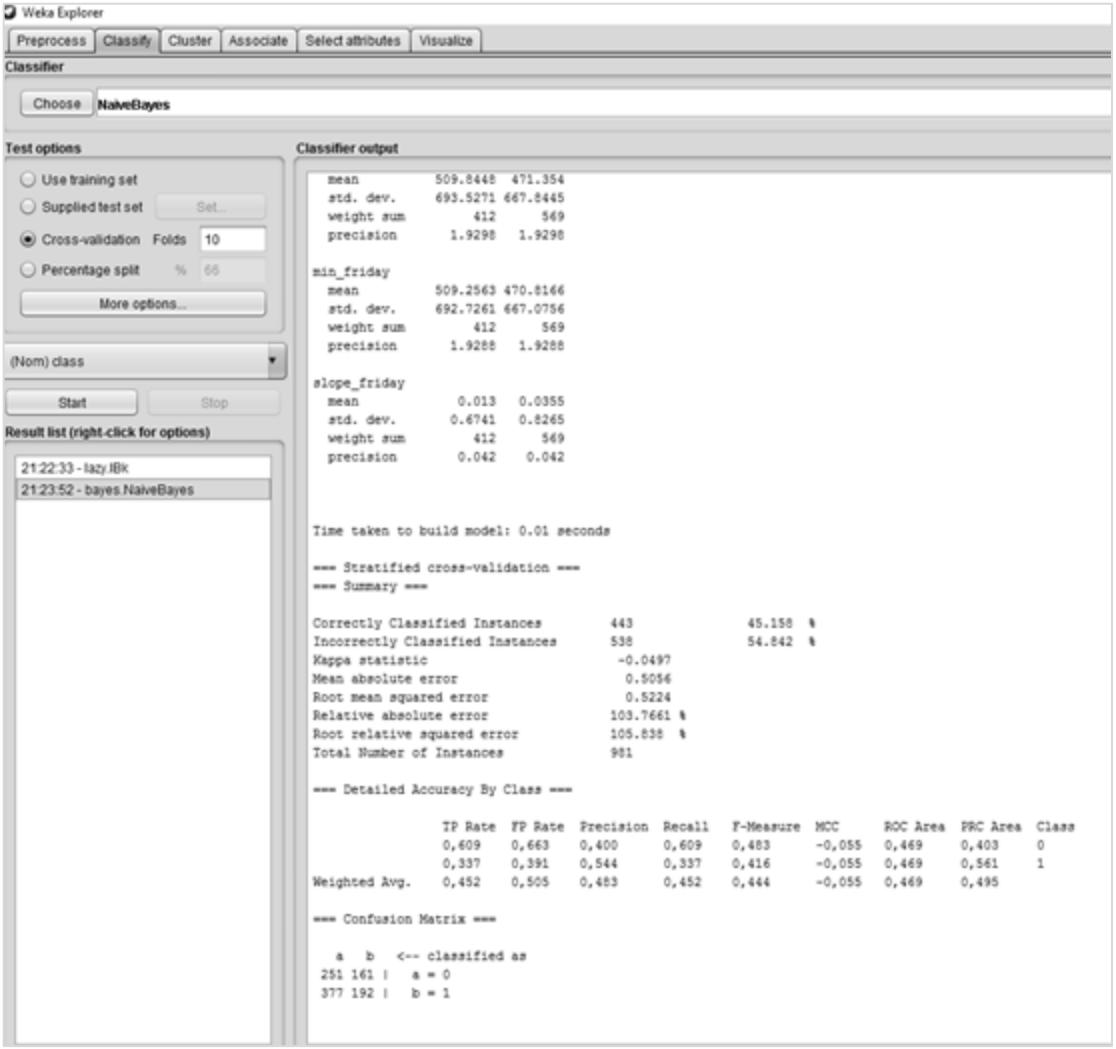
Εικόνα 4.3 : Αποτελέσματα kNN

Correctly Classified Instances			**41.7%**
	Pred. Class 1	**Pred. Class 2**	**Pred. Class 3**
Class 1	165	95	152
Class2	89	86	49
Class3	143	44	158

4.2.1 Πρόβλημα δύο-κλάσεων (2-class problem)

1) Naive Bayes

Στο πρόβλημα 2 κλάσεων η διαφοροποίηση των αποτελεσμάτων από το πρόβλημα τριών κλάσεων διαφαίνεται στην εικόνα 4.4 παρακάτω :

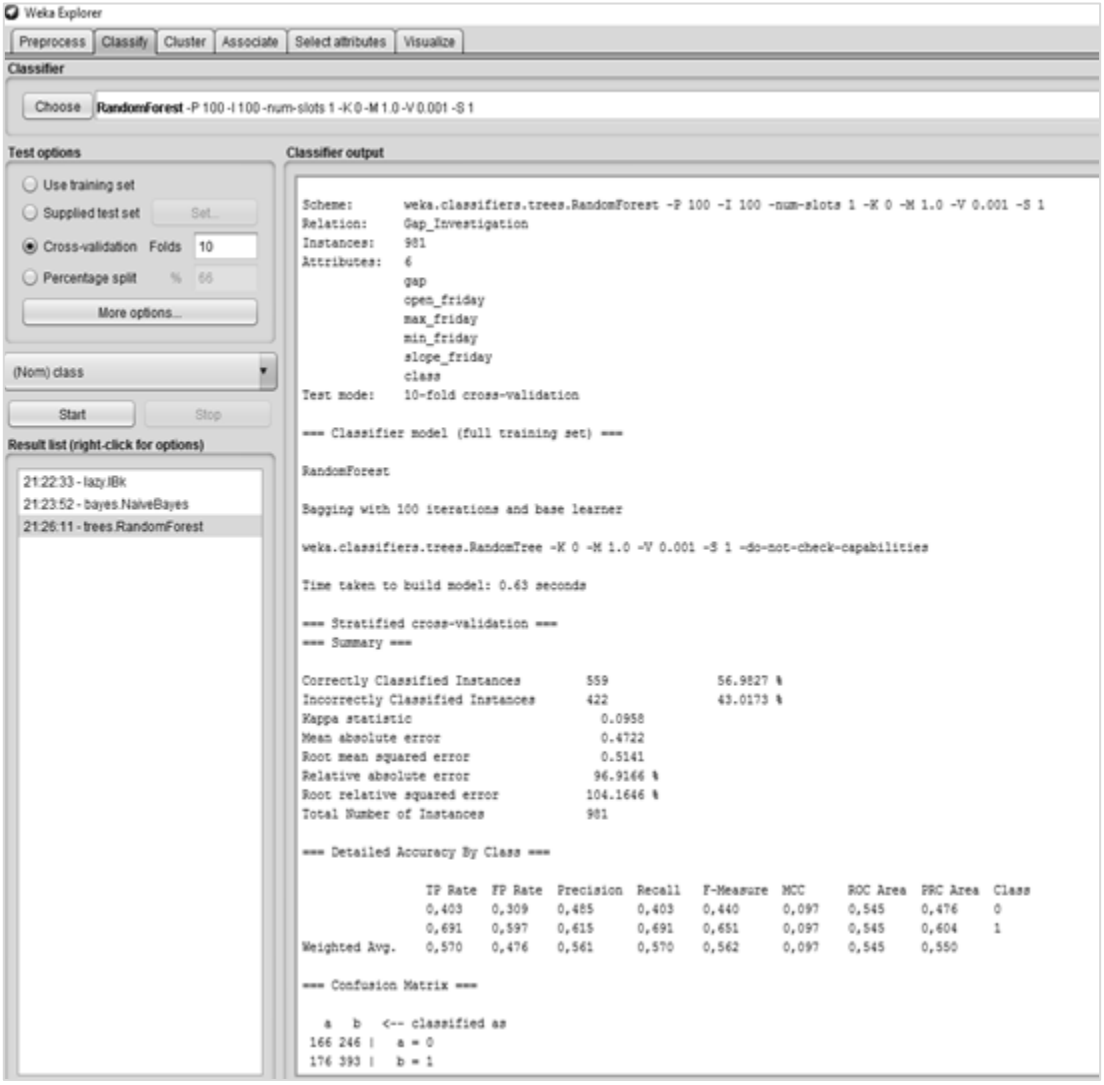


Εικόνα 4.4 : Ααποτελέσματα Naive-Bayes 2-class problem

Correctly Classified Instances		**45.2%**
	Pred. Class 1	**Pred. Class 2**
Class 1	253	141
Class2	377	192

2) Random Forest (Τυχαία Δάση)

Παρομοίως και εδώ :

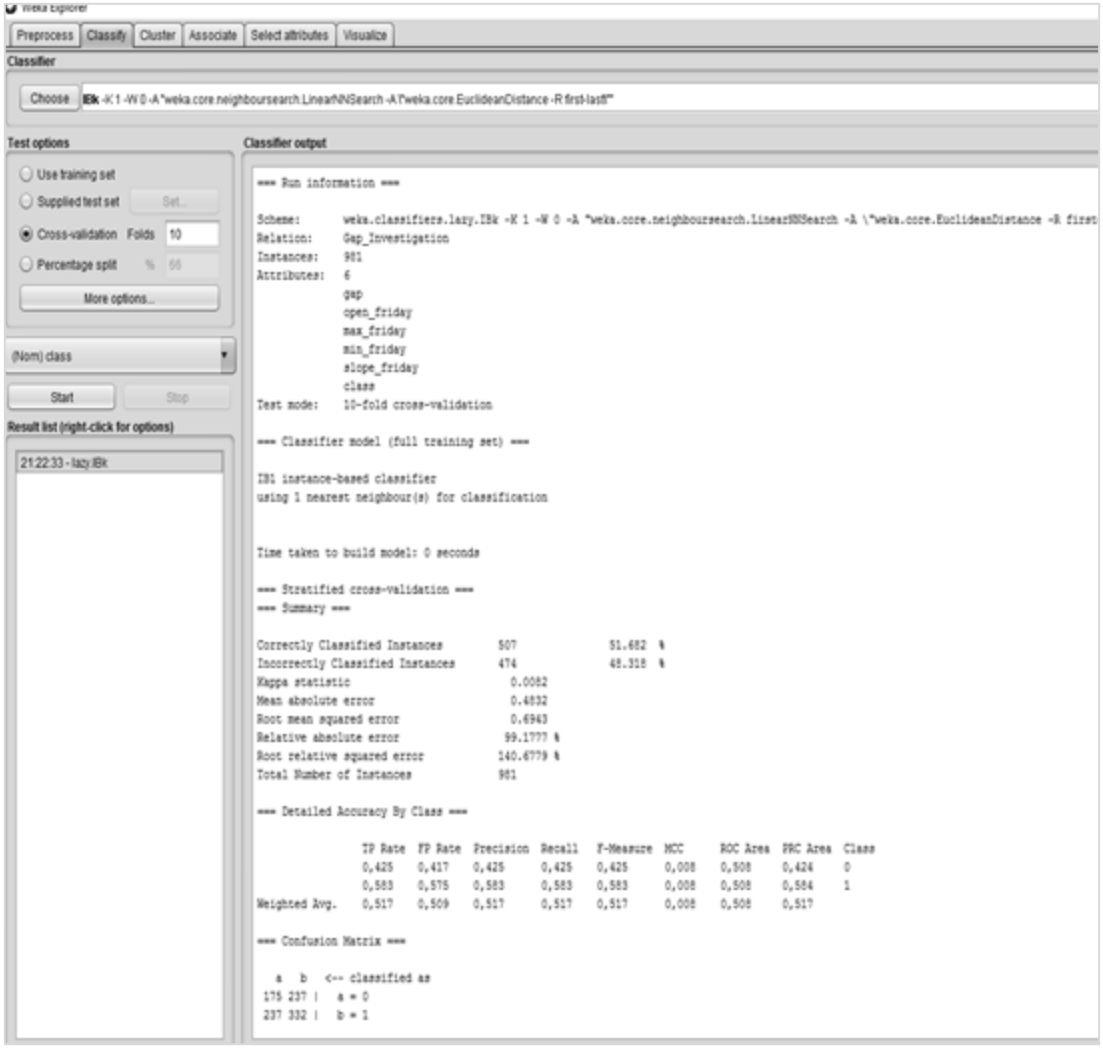


Εικόνα 4.5 : Αποτελέσματα Random Forest 2-class problem

Correctly Classified Instances		**57.0%**
	Pred. Class 1	**Pred. Class 2**
Class 1	166	246
Class2	176	393

3)kNN (k-Nearest Neighbor)

Τέλος ο kNN:



Εικόνα 4.6 : Αποτελέσματα kNN 2-class problem

Correctly Classified Instances		**51.5%**
	Pred. Class 1	**Pred. Class 2**
Class 1	175	237
Class 2	297	332

Παρακάτω παραθέτουμε ένα συγκεντρωτικό πίνακα ο οποίος παρουσιάζει τα ποσοστά των σωστά ταξινομημένων δειγμάτων και για τα 6 πειράματα που διεξήχθησαν.

Συγκεντρωτικός Πίνακας Αποτελεσμάτων

	Naïve Bayes	**KNN**	**Random Forests**
3-class	42.1	41.7	**46.6**
2-class	45.2	51.5	**57.0**

ΚΕΦΑΛΑΙΟ 5

Συμπεράσματα

Στην παρούσα εργασία διεξήχθη μελέτη του χάσματος τιμών οικονομιών δεικτών. Η μελέτη του χάσματος τιμών αποτελεί μια από τις πιο απλές αλλά και διαδεδομένες στρατηγικές που χρησιμοποιούν επενδυτές με σκοπό να έχουν βραχυπρόθεσμο κέρδος. Η ιδέα στηρίζεται στην παρατήρηση ότι όταν ένας δείκτης έχει την τάση να μεταβληθεί απότομα σε ώρες και μέρες που είναι κλειστή η αγορά, τότε συνήθως το επόμενο χρονικό διάστημα διορθώνει το χάσμα που έχει προκύψει από το κλείσιμο της αγοράς στο επόμενο άνοιγμα της. Για τον σκοπό αυτό εντοπίσαμε όλα τα χάσματα σε βάθος πενταετίας για 3 δείκτες συναλλάγματος και εφαρμόσαμε τεχνικές ταξινόμησης για να προβλέψουμε τι μετά το χάσμα πορεία των δεικτών αυτών. Συγκρίνοντας τα αποτελέσματα των πειραμάτων που έγιναν στο πλαίσιο της εργασίας αυτής, γίνεται φανερό ότι δεν μπορούμε να προβλέψουμε σωστά την πορεία ενός δείκτη, τουλάχιστον όχι με τα χαρακτηριστικά που χρησιμοποιήθηκαν. Ιδιαίτερα στο πρόβλημα δύο κατηγορίων, τα αποτελέσματα είναι παραπλήσια του τυχαίο, δηλαδή περίπου κοντά στο 50%. Στο πρόβλημα 3 κατηγοριών παρατηρούμε ότι έχουμε αυξήσει το ποσοστό της πρόβλεψης προσεγγίζοντας το 47%. Ωστόσο και αυτό είναι ένα χαμηλό ποσοστό επιτυχίας.

Πρέπει να επισημάνουμε ότι από την στιγμή που εξήχθησαν τα χάσματα των σαββατοκύριακων για τους τρεις δείκτες που επιχειρήσαμε να μελετήσουμε, άρχισε να διαφαίνεται ότι τα πειράματα δεν θα έχουν ιδιαίτερη επιτυχία, και αυτό διότι δεν φάνηκε να υπάρχει επαρκής αριθμός μεγάλων χασμάτων στην πενταετία που μελετήσαμε για τους δείκτες αυτούς. Ίσως και η φύση των δεικτών συναλλάγματος να μην ενδείκνυται ιδιαίτερα για τέτοιοι είδους προβλέψεις καθώς δεν φαίνεται να έχουν τόσο απότομες διακυμάνσεις στον χρόνο. Αυτός ο συλλογισμός ενισχύεται ακόμα περισσότερο εφόσον μιλάμε για ισοτιμίες ιδιαίτερα ισχυρών νομισμάτων, τα οποία δεν επηρεάζονται τόσο εύκολα. Επειδή οι τιμές των χασμάτων δεν εξήχθησαν

με αυτόματο τρόπο, δεν υπήρχε η δυνατότητα να αντικατασταθούν οι δείκτες που εξαρχής είχαν αποφασιστεί να μελετηθούν

Τέλος άξιο σχολιασμού είναι το γεγονός ότι τα καλύτερα αποτελέσματα εξήχθησαν με χρήση του αλγόριθμου Random Forest στο πείραμα με 3 κατηγορίες, ο οποίος παρουσίασε ακρίβεια 46.6%.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1]. Wikipedia , " *Economic_indicator"*. Διαθέσιμο στον ιστότοπο: https://en.wikipedia.org/wiki/Economic_indicator

[2]. Easy-forex , "*Οικονομικοί δείκτες"*. Διαθέσιμο στον ιστότοπο: http://www.easy-forex.com/au/el/economicindicators/

[3]. Fxstreet.gr , "*Βασική Ορολογία του Forex Trading"*. Διαθέσιμο στον ιστότοπο: http://www.fxstreet.gr/index.php/learning/forex-trading-orologia

[4]. Wikipedia, "*Χρηματιστήριο"*. Διαθέσιμο στον ιστότοπο: https://el.wikipedia.org/wiki/%CE%A7%CF%81%CE%B7%CE%BC%CE%B1%CF%84%CE%B9%CF%83%CF%84%CE%AE%CF%81%CE%B9%CE%BF

[5]. Fxcm, "*CFDs"*.Διαθέσιμο στον ιστότοπο: http://www.fxcm.gr/markets/cfds/stock-indices/

[6]. *"1^{ο} Κεφάλαιο εισαγωγή στην στατιστική επιστήμη*" .Διαθέσιμο στον ιστότοπο:http://eclass.gunet.gr/modules/document/file.php/LAWGU115/doc1.pdf

[7]. Wikipedia, *"Εξόρυξη δεδομένων"*. Διαθέσιμο στον ιστότοπο: https://el.wikipedia.org/wiki/%CE%95%CE%BE%CF%8C%CF%81%CF%85%CE%BE%CE%B7_%CE%B4%CE%B5%CE%B4%CE%BF%CE%BC%CE%AD%CE%BD%CF%89%CE%BD#.CE.99.CF.83.CF.84.CE.BF.CF.81.CE.AF.CE.B1_.CE.BA.CE.B1.CE.B9_.CE.95.CE.BE.CE.AD.CE.BB.CE.B9.CE.BE.CE.B7

[8]. Wikipedia, "Αναγνώριση προτύπων". Διαθέσιμο στον ιστότοπο: https://el.wikipedia.org/wiki/%CE%91%CE%BD%CE%B1%CE%B3%CE%BD%CF%8E%CF%81%CE%B9%CF%83%CE%B7_%CF%80%CF%81%CE%BF%CF%84%CF%8D%CF%80%CF%89%CE%BD

[9]. *Altredo,* "Binary Options Auto Trader". *Διαθέσιμο στον ιστότοπο: http://www.altredo.com/metro_binary_options_autotrader.aspx*

[10]. Dummies, "Strategic Planning: Forecasting with Indicators".

[11]. Διαθέσιμο στον ιστότοπο: http://www.dummies.com/how-to/content/strategic-planning-forecasting-with-indicators.html

[12]. Wikipedia, "*Εξόρυξη δεδομένων".* Διαθέσιμο στον ιστότοπο: https://el.wikipedia.org/wiki/%CE%95%CE%BE%CF%8C%CF%81%CF%85%CE%BE%CE%B7_%CE%B4%CE%B5%CE%B4%CE%BF%CE%BC%CE%AD%CE%BD%CF%89%CE%BD#.CE.99.CF.83.CF.84.CE.BF.CF.81.CE.AF.CE.B1_.CE.BA.CE.B1.CE.B9_.CE.95.CE.BE.CE.AD.CE.BB.CE.B9.CE.BE.CE.B7

[13]. Καθηγητής Χρ. Στυλιος, ΤΕΙ ΗΠΕΙΡΟΥ "Σημειώσεις μαθήματος «Εξόρυξη Δεδομένων» .

[14]. Wikipedia, "*Naive_Bayes_classifier".* Διαθέσιμο στον ιστότοπο: https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[15]. Δρ. Σταύρος Καμινάρης , Πειραιάς 2012, «*Συστήματα αποφάσεων»* .Διαθέσιμο στον ιστότοπο http://electrical.dep.teipir.gr/LotusQuickr/ads/Main.nsf/$defaultview/BAEFA3A5F7943991C2257C3800491AF2/$File/2%CE%BF-%CE%A4%CE%95%CE%A5%CE%A7%CE%9F%CE%A3-%CE%94%CE%95%CE%9D%CE%94%CE%A1%CE%91%20%CE%91%CE%A0%CE%9F%CE%A6%CE%91%CE%A3%CE%95%CE%A9%CE%9D%20-%20final%20version.pdf?OpenElement

[16]. Wikipedia,*"Random Forest",* Διαθέσιμο στον ιστότοπο :

[17]. https://en.wikipedia.org/wiki/Random_forest

[18]. [Mendenhall, W. & Beaver, R. G. & Beaver, B. M. «Introduction to Probability and Statistics. Duxbury Press».

[19]. Wikipedia,*"k-nearest neighbors algorithm",*Διαθέσιμο στον ιστότοπο :

[20]. https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

[21]. «Simple guide to confusion matrix terminology», March 2014

[22]. Διαθέσιμο στον ιστότοπο: http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/

[23]. Κατηγοριοποίηση ΙΙ, Έτος 2010-2011, *« Εξόρυξη Δεδομένων».* Διαθέσιμο στον ιστότοπο: http://www.cs.uoi.gr/~pitoura/courses/dm/classification11b.pdf