ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΛΟΠΟΝΝΗΣΟΥ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Νέες μέθοδοι εξόρυξης γνώσης από οικονομικά δεδομένα

Διπλωματική Εργασία

Βικάτος Αναστάσιος

Επιβλέπων καθηγητής:

Ταμπακάς Βασίλειος

Πάτρα, Μάρτιος 2019

Ευχαριστίες

Θα ήθελα να εκφράσω τις μου ευχαριστίες σε όλους όσοι με βοήθησαν και συνέβαλαν στην περάτωση της εργασίας αυτής. Πρώτον από όλους θέλω να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Ταμπακά Βασίλειο για την επιλογή του θέματος, για την άψογη συνεργασία που είχαμε και για τις γνώσεις που απέκτησα όλο αυτό το διάστημα.

Τέλος, θα ήθελα να πω ένα μεγάλο ευχαριστώ τον πατέρα μου Ηλία, την μητέρα μου Μαρία και τον αδερφό μου Ανδρέα καθώς και σε όλα τα κοντινά μου πρόσωπα για την υπομονή, τη βοήθεια και τη στήριξη που μου προσέφεραν κατά τη διάρκεια της εκπόνησης της παρούσας εργασίας, αλλά και καθ' όλη τη διάρκεια των σπουδών μου.

Αναστάσιος Η. Βικάτος,

Πάτρα, 2019

Περιεχόμενα

Ευρετήριο εικόνων

Ευρετήριο πινάκων

Κεφάλαιο 1. Εισαγωγή

1.1 Αντικείμενο της διπλωματικής εργασίας

Το αντικείμενο μελέτης της παρακάτω διπλωματικής εργασίας είναι η παρουσίαση νέων μεθόδων εξόρυξης δεδομένων από διάφορα οικονομικά συστήματα. Οι μέθοδοι αυτοί βασίζονται σε νέες τεχνικές που αναπτύσσονται στους κλάδους της μηχανικής μάθησης καθώς και της τεχνητής νοημοσύνης. Βασικοί στόχοι της εξόρυξης δεδομένων είναι η εφαρμογή τεχνικών πρόβλεψης, η αναγνώριση, η περιγραφή σε μεγάλες βάσεις δεδομένων και την ταξινόμηση-βελτιστοποίηση των πόρων της. Εμβαθύνοντας στους στόχους η εξόρυξη δεδομένων αποτελείτε από:

- Πρόβλεψη (*prediction*): Με τον όρο πρόβλεψη εννοούμε την προσπάθεια εκτίμησης συμπερασμάτων από τα διαθέσιμα δεδομενά. Η προσπάθεια αυτή έχει ως πρωταρχικό στόχο την λήψη σωστών αποφάσεων ώστε να μεγιστοποιηθεί το κέρδος και ταυτόχρονα την αποτροπή δυσάρεστων καταστάσεων.
- Αναγνώριση (*recognition*): Είναι η φάση όπου τυποποιημένες μορφές δεδομένων χρησιμοποιούνται για την ύπαρξη δραστιριοτήτων και γεγονότων.
- Περιγραφή (*description*): Οι περιγραφικές διαδικασίες της εξόρυξης δεδομένων, έχουν ως στόχο την περιγραφή των γενικών ιδιοτήτων στα υπάρχοντα διαθέσιμα δεδομένα. Η διαδικασία αυτή, επικεντρώνεται στην αποκάλυψη προτύπων.
- Ταξινόμηση (*classification*): Είναι η διαδικασία διαχωρισμού των στοιχείων, που καταλήγει σε διαφορετικές κατηγορίες ή κλάσεις.
- Βελτιστοποίηση (*optimization*): Ο χρόνος, ο χώρος, το χρήμα και η μεγιστοποίηση κάποιων μεγεθών είναι η βέλτιστη χρήση κάποιων πόρων για την εξόρυξη γνώσης.

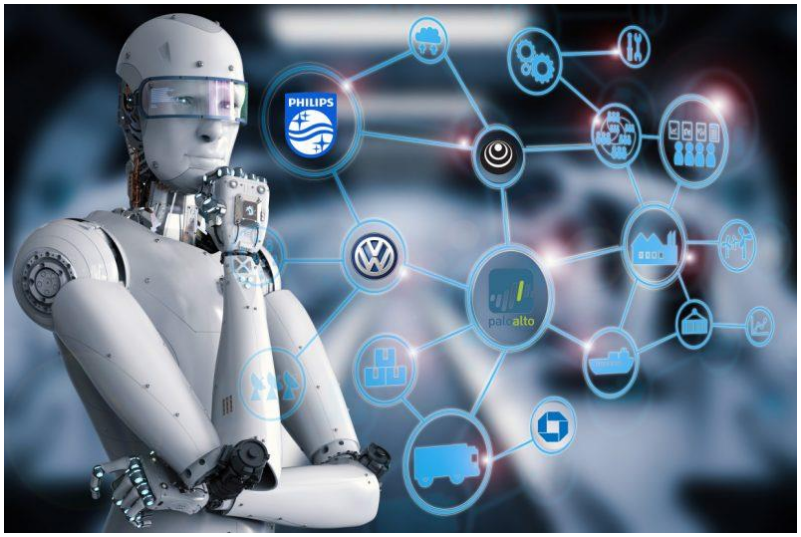
Η εξόρυξη γνώσης μπορεί να γίνει μέσω πολλών πηγών, μια εκ των οποίων είναι οι βάσεις δεδομένων. Για να επιτευχθεί αυτό η διαδικασία έχει ως εξής:

- Επιλογή (*selection*).
- Προεπεξεργασία (*preprocessing*).
- Μετασχηματισμός (*transformation*).
- Εξόρυξη (*data mining*).
- Ερμηνεία-Αξιολόγηση (*interpretation-evaluation*).

Εικόνα 1: Εξόρυξη δεδομένων

Η *τεχνητή νοημοσύνη* (Artificial Intelligence - AI) [1], μερικές φορές αποκαλείται μηχανική νοημοσύνη, είναι η νοημοσύνη που καταδεικνύεται από τις μηχανές, σε αντίθεση με τη φυσική νοημοσύνη που εμφανίζουν οι άνθρωποι και άλλα ζώα. Στην επιστήμη των υπολογιστών η έρευνα AI ορίζεται ως η μελέτη των «*έξυπνων πρακτόρων*»: κάθε συσκευή που αντιλαμβάνεται το περιβάλλον της και αναλαμβάνει δράσεις που μεγιστοποιούν την πιθανότητα επιτυχίας της επίτευξης των στόχων της. Ο όρος «τεχνητή νοημοσύνη» χρησιμοποιείται επιμελώς όταν ένα μηχάνημα μιμείται τις «*γνωστικές*» λειτουργίες που ο άνθρωπος συνδέεται με άλλα ανθρώπινα μυαλά, όπως «*μάθηση*» και «*επίλυση προβλημάτων*». Η τεχνητή νοημοσύνη ιδρύθηκε ως ακαδημαϊκή πειθαρχία το

1956 και τα τελευταία χρόνια έχει βιώσει αρκετά κύματα αισιοδοξίας, ακολουθούμενη από την απογοήτευση και την απώλεια της χρηματοδότησης (γνωστή ως «*Χειμώνας του AI*») ακολουθούμενη από νέες προσεγγίσεις, επιτυχία και ανανεωμένη χρηματοδότηση. Για το μεγαλύτερο μέρος της ιστορίας της, η έρευνα AI έχει διαιρεθεί σε υποπεδία που συχνά δεν επικοινωνούν μεταξύ τους. Αυτά τα υποπεδία βασίζονται σε τεχνικές εκτιμήσεις, όπως είναι οι συγκεκριμένοι στόχοι (π.χ. ρομποτική ή μηχανική μάθηση), η χρήση συγκεκριμένων εργαλείων (λογική ή τεχνητά νευρωνικά δίκτυα) ή βαθιές φιλοσοφικές διαφορές. Τα υποπεδία έχουν επίσης βασιστεί σε κοινωνικούς παράγοντες (συγκεκριμένα ιδρύματα ή έργα συγκεκριμένων ερευνητών). Στον εικοστό πρώτο αιώνα, οι τεχνικές του *AI* έχουν βιώσει μια αναζωπύρωση μετά από ταυτόχρονες εξελίξεις στην εξουσία του υπολογιστή, μεγάλα ποσά δεδομένων και θεωρητική κατανόηση. και οι τεχνικές *AI* έχουν καταστεί ουσιαστικό μέρος της βιομηχανίας τεχνολογίας, συμβάλλοντας στην επίλυση πολλών προκλητικών προβλημάτων στην επιστήμη των υπολογιστών, στη μηχανική λογισμικού και στην έρευνα λειτουργιών [2, 3, 4].



Εικόνα 2: Τεχνητή νοημοσύνη

1.2 Διάρθρωση διπλωματικής

Το κείμενο της εργασία χωρίζεται σε πέντε κεφάλαια. Το πρώτο κεφάλαιο αποτελεί την εισαγωγή στο αντικείμενο της διπλωματικής εργασίας, καθώς και να καθορίσει τους στόχους αυτής.

Στο δεύτερο κεφάλαιο της διπλωματικής γίνεται ανάλυση σε βασικές έννοιες που σχετίζονται με την μηχανική μάθηση καθώς και ορισμούς κατηγοριών και διαφόρων μεθόδων αυτής. Επιπλέον παρουσιάζονται οι αλγόριθμοι που θα χρησιμοποιηθούν στην συνέχεια της εργασίας.

Στο τρίτο κεφάλαιο αναλύεται ο ορισμός της εξόρυξης δεδομένων που βασίζεται σε οικονομικά συστήματα καθώς η ανάλυση του προαναφερθέντος προβλήματος.

Στο τέταρτο κεφάλαιο γίνεται παρουσιάση των αριθμητικών αποτελεσμάτων, των συνόλων των δεδομένων, οι μετρικές απόδοσης καθώς και η σύγκριση μεταξύ τον αλγορίθμων που χρησιμοποιήθηκαν ώστε να παρουσιαστούν οι πίνακες με των αποτελεσμάτων των δυο προηγούμενων εννοιών.

Τέλος, στο πέμπτο κεφάλαιο της διπλωματικής εργασίας αναγράφονται τα συμπεράσματα που προέκυψαν από τη μελέτη της, καθώς και από την ανάλυση των στοιχείων που παρουσιάστηκαν σε αυτή.

Κεφάλαιο 2.
Μηχανική μάθηση

2.1 Βασικές έννοιες

Η *μηχανική μάθηση* (Machine Learning (ML)) είναι ένα πεδίο τεχνητής νοημοσύνης (Artificial Intelligence (AI)) που χρησιμοποιεί στατιστικές τεχνικές για να δώσει στα συστήματα ηλεκτρονικών υπολογιστών τη δυνατότητα να «μαθαίνουν» (π.χ., να βελτιώνουν σταδιακά την απόδοση σε μια συγκεκριμένη εργασία) από τα δεδομένα, χωρίς να προγραμματίζονται ρητά. Στο πεδίο της ανάλυσης δεδομένων, η μηχανική μάθηση είναι μια μέθοδος που χρησιμοποιείται για την εκπόνηση σύνθετων μοντέλων και αλγορίθμων που προσφέρονται για την πρόβλεψη, κάτι που είναι γνωστό ως προγνωστική ανάλυση. Αυτά τα αναλυτικά μοντέλα επιτρέπουν στους ερευνητές, τους επιστήμονες δεδομένων, τους μηχανικούς και τους αναλυτές να «παράγουν αξιόπιστες, επαναλαμβανόμενες αποφάσεις και αποτελέσματα» και να αποκαλύπτουν «κρυφές πληροφορίες» μέσω της μάθησης από ιστορικές σχέσεις και τάσεις στα δεδομένα.

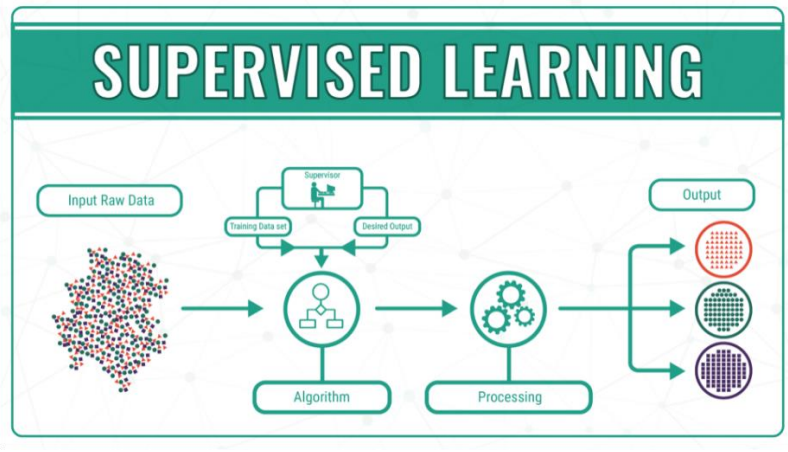
Η μηχανική μάθηση ταξινομείται συνήθως σε διάφορες ευρείες κατηγορίες:

- Μάθηση με πλήρη επίβλεψη (*Supervised learning*)
- Μάθηση με μερική επίβλεψη (*Semi-supervised learning*)
- Μάθηση χωρίς επίβλεψη (*Unsupervised learning*)
- Διαδραστική μάθηση (*Active learning*)
- Ενισχυμένη μάθηση (*Reinforcement learning*)

2.2 Μάθηση με πλήρη επίβλεψη

Η *μάθηση με πλήρη επίβλεψη* (Supervised learning) είναι μία από τις κατηγορίες μηχανικής μάθησης, στόχος της οποίας είναι ο χαρακτηρισμός δεδομένων με βάση κάποια δεδομένα εκπαίδευσης. Τα δεδομένα εκπαίδευσης αποτελούνται από ένα σύνολο παραδειγμάτων τα οποία χρησιμοποιούνται για

εκπαίδευση μοντέλων. Στην επιβλεπόμενη μάθηση, κάθε παράδειγμα αποτελείται από ένα σύνολο εισόδου και μιας τιμής εξόδου. Οι αλγόριθμοι μάθησης με πλήρη επίβλεψη κάνουν ανάλυση των δεδομένων εκπαίδευσης και παράγουν ένα μοντέλο το οποίο μπορεί να χρησιμοποιηθεί για να χαρακτηρίσει νέα παραδείγματα. Το καλύτερο σενάριο επιτρέπει στον αλγόριθμο να καθορίσει σωστά την ετικέτα της κατηγορίας για άγνωστα παραδείγματα. Για να επιτευχθεί αυτό, απαιτείται ο αλγόριθμος μάθησης να γενικεύει από τα δεδομένα εκπαίδευσης σε αόρατες συνθήκες με ένα «λογικό» τρόπο.



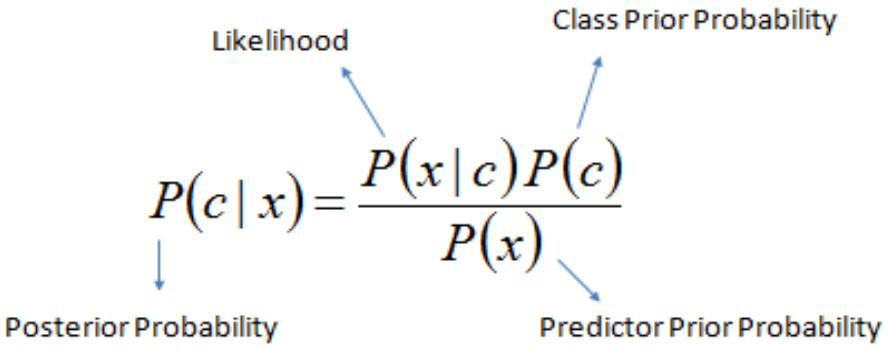
Εικόνα 3: Μάθηση με πλήρη επίβλεψη

2.2.1 Naive Bayes

Ο *αφελής ταξινομητής Bayes* (Naive Bayes) [5] είναι ένας απλός πιθανοτικός ταξινομητής που βασίζεται στην εφαρμογή του θεωρήματος του Bayes με ισχυρές υποθέσεις ανεξαρτησίας που υποθέτουν ότι όλα τα χαρακτηριστικά είναι εξίσου ανεξάρτητα. Χρησιμοποιεί έναν *Bayesian* αλγόριθμο για τη διαδικασία της συνολικής πιθανότητας, η αρχή είναι σύμφωνα με την πιθανότητα ότι το κείμενο ανήκει σε μια κατηγορία πιθανότητας εκ των προτέρων, το κείμενο θα ανατεθεί στην κατηγορία της πιθανότητας *posterior*. Με απλά λόγια, ένας αφελής ταξινομητής *Bayes* υποθέτει ότι η παρουσία (ή η απουσία) ενός συγκεκριμένου χαρακτηριστικού μιας τάξης δεν σχετίζεται με την παρουσία (ή απουσία) οποιουδήποτε άλλου χαρακτηριστικού.

Όπου:

- $P(c|x)$ είναι η πιθανότητα *posterior*
- $P(x|c)$ είναι η πιθανότητα
- $P(c)$ είναι η προηγούμενης τάξης πιθανότητα
- $P(x)$ είναι η πιθανότητα προγνωστικού



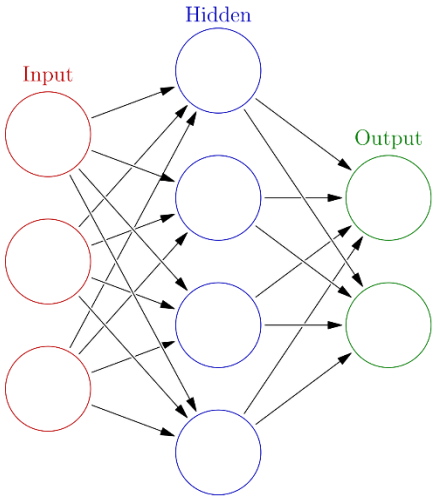
$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Εικόνα 4: Naive Bayes

2.2.2 Τεχνητά νευρωνικά δίκτυα

Τα *τεχνητά νευρικά δίκτυα* (ΤΝΔ) ή τα συστήματα σύνδεσης είναι υπολογιστικά συστήματα ασαφώς εμπνευσμένα από τα βιολογικά νευρωνικά δίκτυα που αποτελούν εγκέφαλο των ζώων [6] . Το ίδιο το νευρικό δίκτυο δεν είναι αλγόριθμος, αλλά ένα πλαίσιο για πολλούς διαφορετικούς αλγόριθμους μηχανικής μάθησης για να συνεργάζονται και να επεξεργάζονται πολύπλοκες εισόδους δεδομένων. Αυτά τα συστήματα «μαθαίνουν» να εκτελούν εργασίες εξετάζοντας παραδείγματα, γενικά χωρίς να προγραμματίζονται με ειδικούς κανόνες. Για παράδειγμα, στην αναγνώριση εικόνων, μπορεί να μάθουν να εντοπίζουν εικόνες που περιέχουν γάτες αναλύοντας παραδείγματα εικόνων που έχουν γίνει με χειροκίνητη επισήμανση ως «γάτα» ή «όχι γάτα» και χρησιμοποιώντας τα αποτελέσματα για τον εντοπισμό των γάτων σε άλλες εικόνες. Το κάνουν αυτό χωρίς προηγούμενη γνώση για τις γάτες, για

παράδειγμα, ότι έχουν γούνα, ουρές, μουστάκια και πρόσωπα που μοιάζουν με γάτες. Αντ 'αυτού, δημιουργούν αυτόματα χαρακτηριστικά ταυτοποίησης από το μαθησιακό υλικό που επεξεργάζονται. Ένα τεχνητό νευρωνικό δίκτυο βασίζεται σε μια συλλογή συνδεδεμένων μονάδων ή κόμβων που ονομάζονται τεχνητοί νευρώνες, οι οποίοι χαλαρά μοντελοποιούν τους νευρώνες σε έναν βιολογικό εγκέφαλο. Κάθε σύνδεση, όπως οι συνάψεις σε έναν βιολογικό εγκέφαλο, μπορεί να μεταδώσει ένα σήμα από έναν τεχνητό νευρώνα στον άλλο. Ένας τεχνητός νευρώνας που λαμβάνει ένα σήμα μπορεί να το επεξεργαστεί και στη συνέχεια να σηματοδοτήσει πρόσθετους τεχνητούς νευρώνες που συνδέονται με αυτό. Στις κοινές εφαρμογές των τεχνητών νευρωνικών δικτύων, το σήμα σε μια σύνδεση μεταξύ των τεχνητών νευρώνων είναι ένας πραγματικός αριθμός και η έξοδος κάθε τεχνητού νευρώνα υπολογίζεται από κάποια μη γραμμική συνάρτηση του αθροίσματος των εισροών του. Οι συνδέσεις μεταξύ τεχνητών νευρώνων ονομάζονται «άκρες». Οι τεχνητοί νευρώνες και οι άκρες έχουν συνήθως ένα βάρος που προσαρμόζεται ως έσοδα της μάθησης. Το βάρος αυξάνει ή μειώνει την ισχύ του σήματος σε μια σύνδεση. Οι τεχνητοί νευρώνες μπορεί να έχουν ένα κατώφλι τέτοιο ώστε το σήμα να αποστέλλεται μόνο αν το συνολικό σήμα διασχίσει αυτό το όριο. Τυπικά, οι τεχνητοί νευρώνες συσσωματώνονται σε στρώματα. Τα διαφορετικά στρώματα μπορούν να εκτελούν διαφορετικά είδη μετασχηματισμών στις εισόδους τους. Τα σήματα ταξιδεύουν από το πρώτο στρώμα (το στρώμα εισόδου), μέχρι το τελευταίο στρώμα (το στρώμα εξόδου), πιθανώς μετά από πολλαπλές διαδρομές. Ο αρχικός στόχος της προσέγγισης ΤΝΔ ήταν να λυθούν τα προβλήματα με τον ίδιο τρόπο με τον ανθρώπινο εγκέφαλο. Ωστόσο, με την πάροδο του χρόνου, η προσοχή μετακινήθηκε στην εκτέλεση συγκεκριμένων καθηκόντων, οδηγώντας σε αποκλίσεις από τη βιολογία. Τα τεχνητά νευρωνικά δίκτυα έχουν χρησιμοποιηθεί σε ποικίλες εργασίες, όπως η όραση στον υπολογιστή, η αναγνώριση ομιλίας, η μηχανική μετάφραση, το φιλτράρισμα κοινωνικών δικτύων, το παιχνίδι με παιχνίδια και τα βιντεοπαιχνίδια και η ιατρική διάγνωση.



Εικόνα 5: Τεχνητό νευρικό δίκτυο

2.2.3 Μηχανές διανυσμάτων υποστήριξης

Οι *μηχανές διανυσμάτων υποστήριξης* (support vector machines (SVM)) [7] είναι μαθησιακά μοντέλα με πλήρη επίβλεψη και με συναφείς αλγόριθμους εκμάθησης που αναλύουν δεδομένα που χρησιμοποιούνται για ανάλυση ταξινόμησης και παλινδρόμησης. Δεδομένου ότι ένα σύνολο εκπαιδευτικών παραδειγμάτων, κάθε ένα από τα οποία χαρακτηρίζεται ότι ανήκει σε μία ή την άλλη από δύο κατηγορίες, ένας αλγόριθμος κατάρτισης *SVM* δημιουργεί ένα μοντέλο που εκχωρεί νέα παραδείγματα σε μία ή την άλλη κατηγορία, καθιστώντας τον έναν μη πιθανοτικό δυαδικό γραμμικό ταξινομητή όπως η *Platt* κλίμακα υπάρχουν για να χρησιμοποιήσετε *SVM* σε μια πιθανοτική ρύθμιση ταξινόμησης). Ένα μοντέλο *SVM* είναι μια αναπαράσταση των παραδειγμάτων ως σημεία στο διάστημα, χαρτογραφημένα έτσι ώστε τα παραδείγματα των ξεχωριστών κατηγοριών διαιρούνται με ένα σαφές χάσμα όσο το δυνατόν ευρύτερο. Στη συνέχεια, νέα παραδείγματα χαρτογραφούνται στον ίδιο χώρο και προβλέπεται να ανήκουν σε μια κατηγορία με βάση την πλευρά του χάσματος που πέφτουν. Εκτός από την εκτέλεση γραμμικής ταξινόμησης, τα *SVM* μπορούν να εκτελούν αποτελεσματικά μια μη γραμμική ταξινόμηση χρησιμοποιώντας αυτό που ονομάζεται κόλπο του πυρήνα, χαρτογραφώντας σιωπηρά τις εισόδους τους σε χώρους μεγάλης διαστάσεως. Όταν τα δεδομένα δεν έχουν επισημανθεί, η εποπτευόμενη μάθηση δεν είναι δυνατή και απαιτείται μια μη εποπτευόμενη μαθησιακή προσέγγιση, η οποία προσπαθεί να βρει φυσική ομαδοποίηση των

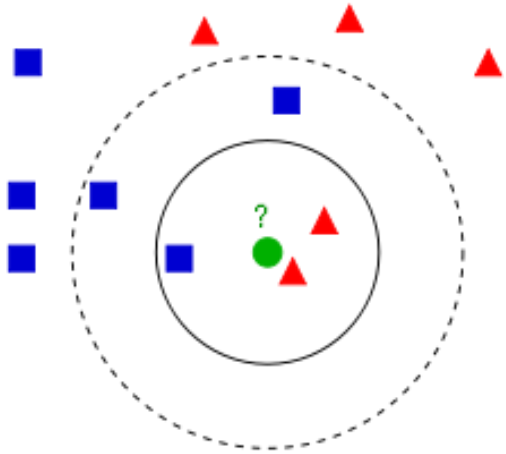
δεδομένων σε ομάδες και στη συνέχεια να χαρτογραφήσει νέα δεδομένα σε αυτές τις σχηματισμένες ομάδες. Ο αλγόριθμος ομαδοποίησης φορέα υποστήριξης, που δημιουργήθηκε από τους *Hava Siegelm* και *Vladimir Vapnik*, εφαρμόζει τα στατιστικά στοιχεία των φορέων υποστήριξης που αναπτύσσονται στον αλγόριθμο μηχανισμών υποστήριξης φορέων για την κατηγοριοποίηση μη επισημασμένων δεδομένων και είναι ένας από τους πιο ευρέως χρησιμοποιούμενους αλγόριθμους ομαδοποίησης σε βιομηχανικές εφαρμογές.

2.2.4 Μάθηση με στιγμιότυπα

Ο *k-πλησιέστερος αλγόριθμος των γειτόνων* (kNN) [8] είναι μια μη παραμετρική μέθοδος που χρησιμοποιείται για την ταξινόμηση και την παλινδρόμηση. Και στις δύο περιπτώσεις, η είσοδος αποτελείται από τα πλησιέστερα παραδείγματα εκπαίδευσης στο χώρο των χαρακτηριστικών. Η έξοδος εξαρτάται από το εάν το kNN χρησιμοποιείται για ταξινόμηση ή παλινδρόμηση: Στην ταξινόμηση kNN, η έξοδος είναι μέλος της τάξης. Ένα αντικείμενο ταξινομείται με πλειοψηφία των γειτόνων του, με το αντικείμενο να ανατίθεται στην τάξη που είναι πιο συνηθισμένη στους πλησιέστερους γείτονές του (k είναι θετικός ακέραιος, συνήθως μικρός). Αν $k = 1$, τότε το αντικείμενο απλώς αποδίδεται στην κλάση εκείνου του μοναδικού πλησιέστερου γείτονα. Στην κλίση kNN, η έξοδος είναι η τιμή ιδιότητας για το αντικείμενο. Αυτή η τιμή είναι ο μέσος όρος των τιμών των πλησιέστερων γειτόνων του k. Το kNN είναι ένας τύπος εκμάθησης βασισμένης σε στιγμές ή τεμπέλης μάθησης, όπου η λειτουργία προσεγγίζεται μόνο τοπικά και όλοι οι υπολογισμοί αναβάλλονται μέχρι την ταξινόμηση. Ο αλγόριθμος kNN είναι από τους απλούστερους όλων των αλγορίθμων μηχανικής μάθησης. Τόσο για ταξινόμηση όσο και για παλινδρόμηση, μια χρήσιμη τεχνική μπορεί να χρησιμοποιηθεί για να αποδώσει το βάρος στις συνεισφορές των γειτόνων, έτσι ώστε οι πλησιέστεροι γείτονες να συνεισφέρουν περισσότερο στον μέσο όρο από τους πιο μακρινούς. Για παράδειγμα, ένα κοινό σχήμα βαρύτητας συνίσταται στην παροχή σε κάθε γείτονα ενός βάρους $1/d$, όπου d είναι η απόσταση από τον γείτονα. Οι γείτονες λαμβάνονται από ένα σύνολο αντικειμένων για τα οποία είναι γνωστή η κλάση (για την ταξινόμηση kNN) ή η τιμή ιδιότητας αντικειμένου (για παλινδρόμηση kNN). Αυτό μπορεί να

θεωρηθεί ως το σετ κατάρτισης για τον αλγόριθμο, αν και δεν απαιτείται ρητό βήμα κατάρτισης. Μια ιδιαιτερότητα του αλγορίθμου kNN είναι ότι είναι ευαίσθητη στην τοπική δομή των δεδομένων. Τα παραδείγματα εκπαίδευσης είναι φορείς σε πολυδιάστατο χαρακτηριστικό χώρο, ο καθένας με ετικέτα κλάσης. Η φάση κατάρτισης του αλγορίθμου συνίσταται μόνο στην αποθήκευση των διανυσμάτων χαρακτηριστικών και των ετικετών κλάσης των δειγμάτων εκπαίδευσης. Στη φάση ταξινόμησης, k είναι μια σταθερά καθορισμένη από το χρήστη και ένας μη επισημασμένος φορέας (ερώτημα ή σημείο δοκιμής) ταξινομείται με την ανάθεση της ετικέτας η οποία είναι πιο συχνή μεταξύ των δειγμάτων κατάρτισης k πλησιέστερα στο εν λόγω σημείο αναζήτησης. Μια κοινώς χρησιμοποιούμενη απόσταση μέτρησης για συνεχείς μεταβλητές είναι η ευκλείδεια απόσταση. Για διακριτές μεταβλητές, όπως για την ταξινόμηση κειμένου, μπορεί να χρησιμοποιηθεί μια άλλη μέτρηση, όπως η μέτρηση επικάλυψης (ή απόσταση *Hamming*). Στο πλαίσιο των δεδομένων μικροσυστοιχίας γονιδιακής έκφρασης, για παράδειγμα, το kNN έχει επίσης χρησιμοποιηθεί με συντελεστές συσχέτισης όπως οι Pearson και Spearman. Συχνά, η ακρίβεια ταξινόμησης του kNN μπορεί να βελτιωθεί σημαντικά εάν η μέτρηση απόστασης αποκτηθεί με εξειδικευμένους αλγορίθμους όπως η ανάλυση των συστατικών του πλησιέστερου πλησιέστερου γείτονα ή της γειτονιάς μεγάλου περιθωρίου. Ένα μειονέκτημα της βασικής ταξινόμησης «ψηφοφορίας με πλειοψηφία» συμβαίνει όταν η κατανομή της τάξης είναι λοξή. Δηλαδή, παραδείγματα μιας πιο συχνής τάξης τείνουν να κυριαρχούν στην πρόβλεψη του νέου παραδείγματος, επειδή τείνουν να είναι κοινά μεταξύ των πλησιέστερων γ γειτόνων λόγω του μεγάλου αριθμού τους. Ένας τρόπος για να ξεπεραστεί αυτό το πρόβλημα είναι να σταθμίσετε την ταξινόμηση, λαμβάνοντας υπόψη την απόσταση από το σημείο δοκιμής σε καθέναν από τους πλησιέστερους γείτονές του k. Η κλάση (ή η τιμή, σε προβλήματα παλινδρόμησης) καθέτων από τα πλησιέστερα σημεία k πολλαπλασιάζεται με ένα βάρος ανάλογο προς το αντίστροφο της απόστασης από αυτό το σημείο στο σημείο δοκιμής. Ένας άλλος τρόπος για να ξεπεραστεί η παραμόρφωση είναι η αφαίρεση στην αναπαράσταση δεδομένων. Για παράδειγμα, σε έναν αυτο-οργανωμένο χάρτη (SOM), κάθε κόμβος είναι ένας εκπρόσωπος (κέντρο) μιας ομάδας παρόμοιων σημείων, ανεξάρτητα από την πυκνότητα του στα αρχικά δεδομένα εκπαίδευσης. Ο kNN

μπορεί στη συνέχεια να εφαρμοστεί στον αυτο-οργανωμένο χάρτη (*self-organizing map (SOM)*).



Εικόνα 6: k-πλησιέστερος αλγόριθμος των γειτόνων (kNN)

2.2.5 Κανόνες ταξινόμησης

Ο *κανόνας ταξινόμησης JRip* (*RIPPER*) είναι ένας από τους βασικούς και πιο δημοφιλείς αλγόριθμους. Οι κλάσεις εξετάζονται σε μέγεθος μεγέθυνσης και παράγεται ένα αρχικό σύνολο κανόνων για την κλάση με τη χρήση του αυξανόμενου μειωμένου σφάλματος *JRip* (*RIPPER*) με την επεξεργασία όλων των παραδειγμάτων μιας συγκεκριμένης απόφασης στα δεδομένα εκπαίδευσης ως τάξη και την εύρεση ενός συνόλου κανόνων καλύπτουν όλα τα μέλη αυτής της τάξης. Στη συνέχεια προχωρά στην επόμενη τάξη και κάνει το ίδιο, επαναλαμβάνοντας αυτό μέχρι να καλύπτονται όλες οι κλάσεις.

2.2.6 Δέντρα απόφασης

2.2.6.1 C4.5

Ένας από τους πιο γνωστούς για την κατασκευή δέντρων απόφασης αλγόριθμος που χρησιμοποιεί το λόγο του κέρδους πληροφορίας είναι ο *C4.5* [9]. Μια από τις πιο πρόσφατες έρευνες που έγιναν για να συγκρίνουν τα δέντρα απόφασης με άλλους αλγορίθμους μάθησης είχαν ως αποτέλεσμα πως ο *C4.5* έχει

έναν πολύ καλό συνδυασμό ακρίβειας και ταχύτητας στο να μαθαίνει. Η παρουσίαση του αλγορίθμου που ως τώρα αναπτύξαμε, προϋποθέτει τη χρήση κατηγορικών χαρακτηριστικών. Ο αλγόριθμος *C4.5* ωστόσο και οι διάφορες επεκτάσεις του, έχουν τη δυνατότητα να διαχειριστούν και συνεχή χαρακτηριστικά, εφαρμόζοντας στην αρχή κάθε αναδρομικού βήματος μια διαδικασία μετατροπής τους σε ένα σύνολο διακριτών λογικών (*boolean*) χαρακτηριστικών, γνωστή ως διακριτοποίηση (*discretization*).

2.2.6.2 LMT

Το *logistic model tree* (*LMT*) [10] είναι ένα μοντέλο ταξινόμησης με έναν συνδεδεμένο εποπτευόμενο αλγόριθμο εκπαίδευσης που συνδυάζει την υλικοτεχνική παλινδρόμηση (*LR*) και τη μάθηση των δέντρων αποφάσεων. Τα δέντρα λογικής μοντέλου βασίζονται στην παλαιότερη ιδέα ενός δέντρου μοντέλου: ένα δέντρο απόφασης που έχει μοντέλα γραμμικής παλινδρόμησης στα φύλλα του για να παρέχει ένα τετραγωνικό μοντέλο γραμμικής παλινδρόμησης (όπου τα κοινά δέντρα αποφάσεων με σταθερές στα φύλλα τους θα παράγουν ένα τετραγωνικό σταθερό μοντέλο). Στην *logistic* παραλλαγή, ο αλγόριθμος *LogitBoost* χρησιμοποιείται για την παραγωγή ενός μοντέλου *LR* σε κάθε κόμβο του δέντρου. ο κόμβος στη συνέχεια χωρίζεται χρησιμοποιώντας το κριτήριο *C4.5*. Κάθε κλήση του *LogitBoost* είναι από τα αποτελέσματα του στον γονικό κόμβο. Τέλος, το δέντρο κλαδεύεται. Ο βασικός αλγόριθμος επαγωγής *LMT* χρησιμοποιεί διασταυρούμενη επικύρωση για να βρει έναν αριθμό επαναλήψεων *LogitBoost* που δεν υπερκαλύπτει τα δεδομένα εκπαίδευσης. Έχει προταθεί μια ταχύτερη έκδοση που χρησιμοποιεί το κριτήριο πληροφοριών *Akaike* για τον έλεγχο της διακοπής του *LogitBoost*.

2.3 Μάθηση με μερική επίβλεψη

Η *μάθηση με μερική επίβλεψη* (Semi-Supervised Learning - SSL) είναι η δεύτερη κατά σείρα από τις κατηγορίες μηχανικής μάθησης που προαναφέρθηκαν στην αρχή αυτού του κεφαλαίου και παρουσιαζεί τις διακεκριμένες τεχνικές μηχανικής μάθησης με ισχυρή γενίκευση που προσπαθούν

να συνδυάσουν αποτελεσματικά τις πληροφορίες ταξινόμησης «*δεδομένων που έχουν ετικέτα*» (labeled data) και στα «*δεδομένα που δεν έχουν ετικέτα*» (unlabeled data). Το βασικό ζήτημα στη μάθηση με μερική επίβλεψη είναι πώς θα γίνει αποτελεσματικά η εκμετάλλευση των πληροφοριών που είναι κρυμμένες στα μη επισημασμένα δεδομένα. Η μάθηση με μερική επίβλεψη (*SSL*) έχει επίσης θεωρητικό ενδιαφέρον για τη μηχανική μάθηση (*ML*) και ως πρότυπο για την ανθρώπινη μάθηση.

2.3.1 Self-training

Ο *αλγόριθμος εκμάθησης Self-training* [11] είναι ένας άλγοριθμος μάθησης με μερική επίβλεψη που χαρακτηρίζεται από την απλότητα και την καλή του απόδοση ταξινόμησης. Στον αλγόριθμο αυτό, ένας ταξινομητής αρχικά εκπαιδεύεται χρησιμοποιώντας *labeled* δεδομένα και αυξάνει το σύνολο εκπαίδευσής του σταδιακά με τις πιο σίγουρες προβλέψεις σε *unlabeled* δεδομένα και επανεκπαιδεύεται. Το μειονέκτημα αυτής της μεθοδολογίας μπορεί να οδηγήσει σε λανθασμένες προβλέψεις εάν υπάρξει θόρυβος στα δεδομένα που ταξινομούνται.

Μια περιγραφή του αλγόριθμου Self-Training:

Input: L – Set of labeled instances
U – Set of unlabeled instances
ConLev – Confidence level
C – Base learner

Output: Trained classifier.

1: **repeat**
2: Train C on L.
3: Apply C on U .
4: Select instances with a predicted probability more than *ConLev*

per iteration (x_{MCP}).

5: Remove x_{MCP} from U and add to L.

6: **until** some stopping criterion is met or U is empty.

2.3.2 Co-training

Με τον όρο *Co-training* [12] αναφερόμαστε σε έναν semi-supervised αλγόριθμο εκμάθησης πολλαπλών όψεων. Στον αλγόριθμο αυτόν, δυο ταξινομητές εκπαιδευόνται από ένα ίδιο σύνολο δεδομένων, με τον κάθε ταξινομητή να χρησιμοποιεί ένα τελείως ανεξάρτητο σύνολο χαρακτηριστικών. Την ώρα αυτής της διαδικασίας κάθε ταξινομητής επαναληπτικά δίνει ετικέτες σε μερικά unlabeled δεδομένα, τα οποία έδειξαν μεγαλύτερη σιγουρία στην τιμή του αποτελεσματός τους από την «οπτική γωνία» κάθε ταξινομητή. Έτσι να νέα δεδομένα που προκύπτουν εντάσσονται στα labeled δεδομένα κάθε ταξινομητή, με αποτέλεσμα τα labeled δεδομένα να αυξηθούν . Ο Co-training χρησιμοποιεί επίσης τα *unlabeled* δεδομένα για να εκπαιδευεί τους ταξινομητές κάτω από δυο προϋπόθεσεις. Η πρώτη ότι οι οπτικές γωνίες δεν πρέπει να έχουν μεγάλη διαφορά μεταξύ τους. Η δεύτερη είναι ότι απαιτείται επαρκής αριθμός δεδομένων για την εκπαίδευση του ταξινομητή.

Μια περιγραφή του αλγόριθμου Co-training:

Input: L – Set of labeled instances.

U – Set of unlabeled instances.

C_i – Base learner ($i = 1, 2$).

Output: Trained classifier.

1: Create a pool U' of u examples by randomly choosing from U.

2: repeat

3: Train C_1 on $L(V1)$.

4: Train C_2 on $L(V2)$.

5: **for each** classifier C_i **do** $(i = 1, 2)$

6: C_i chooses p samples (P) that it most confidently labels as positive and n instances (N) that it most confidently labels as negative from U.

7: Remove P and N from U'.

8: Add P and N to L.

9: **end for**

10: Refill U' with examples from U to keep U' at constant size of u examples.

11: until some stopping criterion is met or U is empty.

2.3.3 Tri-training

Μια άλλη προσέγγιση που βασίζεται επίσης σε μια μεθοδολογία του συνόλου είναι ο αλγόριθμος *Tri-training* [13], ο οποίος αποτελεί μια βελτιωμένη επέκταση μιας προβολής του αλγορίθμου Co-training. Αυτός ο αλγόριθμος χρησιμοποιεί ένα labeled σύνολο δεδομένων για να εκπαιδεύσει αρχικά τρεις βασικούς ταξινομητές οι οποίοι χρησιμοποιούνται για να κάνουν προβλέψεις για τις περιπτώσεις του unlabeled συνόλου δεδομένων. Στη συνέχεια, εάν δύο βασικοί ταξινομητές συμφωνούν στο ίδιο αποτέλεσμα που προκύπτει από τα δεδομένα, δηλαδή το σύνολο να γίνει labeled, τότε και για τον τρίτο βασικό ταξινομητή το αποτέλεσμα θα γίνει labeled. Γίνεται χρήση της στρατηγικής η πλειοψηφία καθοδηγεί την μειονότητα.

Μια περιγραφή του αλγόριθμου Tri-training:

Input: L – Set of labeled instances.

U – Set of unlabeled instances.

Ci – Base learner (i = 1, 2, 3).

Output: Trained classifier.

1: for $i = 1, 2, 3$ **do**

2: S_i = BootstrapSample(L).

3: Train C_i on S_i.

4: end for

5: repeat

6: **for** $i = 1, 2, 3$ **do**

7: $L_i = \emptyset$.

8: **for** $u \in U$ **do**

9: **if** $C_j(u) = C_k(u)$ **then** $(j, k \neq i)$

10: $L_i = L_i \cup (u, C_j(u))$.

11: **end if**

12: **end for**

13: **end for**

14: **for** i = 1, 2, 3 **do**

15: Train C_i on S_i.

17: **end for**

18: until some stopping criterion is met or U is empty.

2.3.4 Democratic Co-learning Ο *Democratic Co-learning* (*Demo-Co*) [14] είναι ένας νεός semi-supervised αλγόριθμος μιας-όψης, ο οποίος χρησιμοποιείται για εφαρμογές χωρίς δύο ανεξάρτητα και υπεράριθμα σύνολα δυνατοτήτων αλλά για μια μικρή ομάδα *labeled* δεδομένων. Στον Demo-Co, μια σειρά διαφορετικών αλγορίθμων μάθησης χρησιμοποιούνται για την εκπαίδευση ενός συνόλου ταξινομητών σε labeled σύνολο δεδομένων, ξεχωριστά.

2.3.5 Co-Forest

Στον *αλγόριθμο Co-Forest* [15], ένας αριθμός *τυχαίων δέντρων* (Random Tree) εκπαιδευόνται πάνω σε αυτοδύναμα δεδομένα, από το σύνολο δεδομένων, και η έξοδος ορίζεται σαν συνδιασμός ξεχωριστών προβλέψεων για κάθε δέντρο. Η βασική ιδεά πίσω από τον αλγόριθμο είναι ότι κατά την διάρκεια της

εκπαιδευτικής διαδικασίας, ο αλγόριθμος αναθέτει μερικά unlabeled παραδείγματα σε κάθε τυχαίο δέντρο ξεχωριστά. Επίσης, παρατηρούμε ότι η αποδοτικότητα του Co-Forest βασίζεται στην χρήση τυχαίων δέντρων, παρόλο που ο αριθμός των διαθέσιμων labeled παραδειγμάτων μειώνονται σημαντικά.

2.3.6 SETRED

Η μέθοδος *SETRED* (*Self-trained with editing*) [16], ο αλγόριθμος αυτός μας παρουσιάζει μια τεχνική επεξεργασίας δεδομένων διαδικασία αυτο-εκπαίδευσης με σκοπό να φιλτράρει τον θόρυβο στα self-labeled παραδείγματα. Συγκεκριμένα, μετά την επισλυμανση των κάποιων παραδειγμάτων που έχουν επιλεχθεί από *unlabeled* σύνολο. Ο SETRED αναγωνρίζει τα πιθάνα παραδείγματα που έχουν πάρει λάθος ετικέτα μέσω της βοήθειας που παίρνει από τα δεδομένα του γειτονικού γραφήματος, έτσι κανει κράτηση των παραδειγμάτων αυτών για να αποφύγει ότι θα προστεθούν στο σύνολο εκπαίδευσης κάθε ταξινομητή, έχοντας έτσι λιγοτέρο θόρυβο στην εκπαίδευση.

2.3.7 Co-Bagging

Ο αλγόριθμος Co-Bagging [17] δημιουργεί μερικούς βασικούς ταξινομητές χρησιμοποιώντας τον ίδιο αλγόριθμο μάθησης σε αυτοδύναμο δείγμα, που δημιουργήθηκε με τυχαία αναδειγματοληψία με αντικατάσταση από το αρχικό σετ εκπαίδευσης. Κάθε αυτοδύναμο δείγμα περιέχει περίπου τα 2/3 του αρχικού σετ εκπαίδευσης, όπου κάθε παράδειγμα μπορεί να εμφανιστεί πολλές φορές.

2.3.8 CST-Voting

Ο *CST-Voting* [18] αλγόριθμος αποτελείται από τον συνδυασμό τριών SSL αλγορίθμων και συγκεκριμένα κάνει χρήση των Co-training, Self-training και Tri-training. Η αποτελεσματικότητα του CST-Voting αξιολογείται με τον αριθμό που προκύπτει από την δοκιμασία επιδόσεων στο σύνολο δεδομένων, σε σχέση με την

ακρίβεια ταξινόμησης με την χρήση των τρίων supervised ταξινομητών ως βασικοί μαθητές.

Μια περιγραφή του αλγόριθμου CST-Voting:

Input: L - Set of labeled training instances.
U - Set of unlabeled training instances.

Output: The labels of instances in the testing set.

/* Phase I: Training*/
1: Self-training*(L,U)*
2: Co-training*(L,U)*
3: Tri-training*(L,U)*

/* Phase II: Voting-Fusion*/
4: for each $x \in T$ **do**
5: Apply Self-training, Co-training, and Tri-training on x.
6: Use majority vote to predict the label y^* of x.
7: end for

Κεφάλαιο 3.
Εξόρυξη γνώσης από οικονομικά δεδομένα

3.1 Εισαγωγή

Η εξόρυξη γνώσης (*Data Mining*) είναι μια τεχνολογία που βοηθάει τις επιχειρήσεις καθώς και διάφορους άλλους τομείς που έχουν σαν κύριο χαρακτηριστικό και βασίζονται στην οικονομία, να εστιάσουν στην πληροφορία που βρίσκεται μέσα στις αποθήκες δεδομένων τους (*Data Warehouses*). Οι τεχνικές της είναι σε θέση να αναζητήσουν και να βρουν γρήγορα καθώς και με λεπτομέρια βάσεις δεδομένων για την αναζήτηση κρυμμένων προτύπων (*patterns*). Για τον λόγο αυτό η εξόρυξη γνώσης έχει χαρακτηριστεί σαν μια διαδικασία εξαγωγής κρυμμένης πληροφορίας που κάνει την αναζήτηση της μέσα σε μεγάλες βάσεις δεδομένων. «Εξόρυξη δεδομένων είναι η διαδικασία εξαγωγής υπονοούμενης και εν πολλοίς άγνωστης αλλά ενδεχομένως χρήσιμης γνώσης υπό την μορφή συσχετίσεων προτύπων και τάσεων, μέσω της εξέτασης ανάλυσης και επεξεργασίας βάσεων δεδομένων, συνδυάζοντας και χρησιμοποιώντας τεχνικές από την μηχανική μάθηση, την αναγνώριση προτύπων, την στατιστική, τις βάσεις δεδομένων και την οπτικοποίηση.» (*Piatetsky-Shapiro & Frawley*). Παρά το γεγονός ότι υπάρχει μια γενικότερη συμφωνία ότι ο στόχος της εξόρυξης δεδομένων είναι η ανακάλυψη νέας και χρήσιμης πληροφορίας σε βάσεις δεδομένων, τα μέσα για την επίτευξη του στόχου αυτού ποικίλουν σε πολύ υψηλό βαθμό. Η εξόρυξη γνώσης περιλαμβάνει ένα ευρύ πεδίο υπολογιστικών μεθόδων που μεταξύ άλλων περιλαμβάνουν, την στατιστική ανάλυση (statistical analysis), τα δένδρα αποφάσεων (*decision trees*), τα νευρωνικά δίκτυα (*neural networks*), την εξαγωγή κανόνων (*rule induction*) και την γραφική οπτικοποίηση (*graphic visualization*). Τέτοιες μέθοδοι χρησιμοποιούνται για την εύρεση συσχετίσεων, προτύπων και δομών σε μεγάλες και διαρκώς αυξανόμενες βάσεις δεδομένων. Ειδικά η εύρεση εργαλείων είναι ένα ιδιαίτερα σημαντικό εξαγόμενο της εξόρυξης δεδομένων μέσω σχέσεων μεταξύ των χαρακτηριστικών των βάσεων δεδομένων.

3.2 Δείκτης Dow Jones

Η τεράστια διαθέσιμη ποσότητα δεδομένων από τις χρηματιστηριακές αγορές καθώς και από τις ταχείες εξελίξεις στην τεχνολογία, επέτρεψε την ανάπτυξη συστημάτων υποστήριξης αποφάσεων για να βοηθήσουν σε περίπλοκα περιβάλλοντα λήψης αποφάσεων. Ως εκ τούτου, τις τελευταίες δεκαετίες, οι ερευνητές άρχισαν να εφαρμόζουν τεχνικές και μηχανικής μάθησης και εξόρυξης γνώσης για την ανάπτυξη έξυπνων συστημάτων για την πρόβλεψη της κίνησης των αποθεμάτων (*Forecasting stocks movement*) και την τιμή των μετοχών (*Stock's price index*) [19].

Παρόλα αυτά, παρά την προσπάθεια αυτή, οι Hajizadeh et al. [20] μας υπέδειξαν σε ερευνά τους ότι δεν υπάρχει ακόμα ακριβής μέθοδος πρόβλεψης καθώς το χρηματιστήριο είναι ένα πολύπλοκο, όχι σταθερό, χαοτικό και μη γραμμικό δυναμικό σύστημα όπου δεν υπάρχουν ακόμη συστήματα που να μπορούν να προβλέψουν με ακρίβεια την κινησή του.

Οι Enke και Thawornwong [21] διερεύνησαν την προβλεπτική ισχύ πολλών οικονομικών μεταβλητών υιοθετώντας την τεχνική μεταβλητής ανάλυσης συνάφειας για την πρόβλεψη των αποδόσεων των χρηματιστηρίων. Δηλώνουν ότι η προτεινόμενη τεχνική τους φαίνεται ελκυστική στην επιλογή των μεταβλητών όταν η χρησιμότητα των δεδομένων είναι άγνωστη, ειδικά όταν υπάρχει μη γραμμικότητα. Επιπλέον, αξιολόγησαν την αποτελεσματικότητα των μοντέλων νευρωνικών δικτύων για την εκτίμηση και την ταξινόμηση επιπέδων και παρουσίασαν μια τεχνική διασταυρούμενης επικύρωσης και πρόωρης διακοπής, η οποία στοχεύει στη βελτίωση της ικανότητας γενίκευσης των μοντέλων πρόβλεψης. Τέλος, τα αποτελέσματα που παρουσιάστηκαν έδειξαν ότι οι εμπορικές στρατηγικές που καθοδηγούνται από τα μοντέλα ταξινόμησης νευρωνικών δικτύων δημιουργούν υψηλότερα κέρδη κάτω από την ίδια έκθεση κινδύνου από εκείνα που προτείνονται από τις άλλες στρατηγικές, συμπεριλαμβανομένης της στρατηγικής buy-and hold, καθώς και των προβλέψεων δίκτυα και μοντέλα γραμμικής παλινδρόμησης.

Οι Senthamarai Kannan et al. [22] αξιολόγησαν διάφορες τεχνικές εξόρυξης δεδομένων για την πρόβλεψη της κίνησης των μετοχών. Η προτεινόμενη

μέθοδος βασίζεται στον συνδυασμό πέντε αλγορίθμων και παράγει μια πρόβλεψη για το εάν οι τιμές των μετοχών θα αυξηθούν ή θα μειωθούν την επόμενη ημέρα. Οι συγγραφείς διενήργησαν μια πειραματική ανάλυση που έδειξε ότι η μέθοδος τους ήταν σε θέση να προβλέψει εάν η τιμή κλεισίματος της επόμενης ημέρας θα αυξηθεί ή θα μειωθεί καλύτερα από τυχαία (50%) με υψηλό επίπεδο σπουδαιότητας. Επιπλέον, δήλωσαν ότι η προτεινόμενη μέθοδος θα μπορούσε να χρησιμοποιηθεί ως σύστημα στήριξης αποφάσεων αγοράς ή πώλησης ή θα μπορούσε να χρησιμοποιηθεί για να δώσει εμπιστοσύνη στην πρόβλεψη των τιμών των μετοχών από έναν έμπορο.

Οι Nanda et al. [23] παρουσίασαν μια μεθοδολογία για την ενσωμάτωση μιας ποικιλίας τεχνικών ομαδοποίησης στη διαχείριση χαρτοφυλακίου και την οικοδόμηση ενός υβριδικού συστήματος για τη δημιουργία αποδοτικών χαρτοφυλακίων. Όλες οι μέθοδοι ομαδοποίησης χρησιμοποιήθηκαν για τη συγκέντρωση χρηματιστηριακών στοιχείων από το Χρηματιστήριο της Βομβάης, το οποίο αποτελείται από αποδόσεις για μεταβλητά μήκη περιόδου μαζί με τους δείκτες αποτίμησης. Τα αποτελέσματά τους έδειξαν ότι η προτεινόμενη τεχνική τους μπορεί να μειώσει σημαντικά τον χρόνο κατά την επιλογή των αποθεμάτων, δεδομένου ότι τα αποθέματα παρόμοιων κατηγοριών μπορούν εύκολα να ομαδοποιηθούν σε ένα σύμπλεγμα. επομένως, μπορούν να επιλεγούν τα αποθέματα με τις καλύτερες επιδόσεις από αυτές τις ομάδες.

Οι Patel el al. [24] μελέτησαν το πρόβλημα της πρόβλεψης της κατεύθυνσης της κίνησης του δείκτη τιμών μετοχών και μετοχών για τις ινδικές χρηματιστηριακές αγορές. Αξιολόγησαν την απόδοση διαφόρων αλγορίθμων μηχανικής μάθησης χρησιμοποιώντας δύο προσεγγίσεις για τα δεδομένα εισόδου. Συγκεκριμένα, η πρώτη προσέγγιση περιλαμβάνει τον υπολογισμό δέκα τεχνικών παραμέτρων με τη χρήση στοιχείων εμπορικών συναλλαγών, ενώ ο δεύτερος εστιάζει στην εκπροσώπηση αυτών των τεχνικών παραμέτρων ως θεωρητικών δεδομένων τάσεων. Η εκτεταμένη πειραματική ανάλυση τους έδειξε ότι η απόδοση όλων των προτύπων πρόβλεψης βελτιώθηκε όταν αυτές οι τεχνικές παράμετροι αντιπροσωπεύονται ως δεδομένα αιτιολογικών τάσεων.

Οι Ng και Khor [25] μέσα από την δημιουργία ενός προφίλ μετοχών, το οποίο ονομάστηκε *StockProF*, το οποίο μπορεί να βοηθήσει τους επενδυτές ώστε

να δημιουργήσουν ένα χαρτοφυλάκιο μετοχών με βάση τις επενδυτικές στρατηγικές τους. Το *StockProF* εντοπίζει τα αποθέματα από μια ομάδα αποθεμάτων χρησιμοποιώντας έναν αλγόριθμο ανίχνευσης εξωστρέφειας προκειμένου να εντοπίσει αποθέματα με καλές ή κακές οικονομικές επιδόσεις. Επιπλέον, χρησιμοποιεί έναν αλγόριθμο ομαδοποίησης για την ομαδοποίηση των υπόλοιπων αποθεμάτων, επιτρέποντας τον προσδιορισμό των αποθεμάτων με διάφορες οικονομικές επιδόσεις. Χρησιμοποίησαν μεταβολές των τιμών των μετοχών κατά ένα (1) έτος για να αξιολογήσουν την απόδοση των αποθεμάτων καθώς και των ομάδων και τα αποτελέσματά τους έδειξαν ότι το *StockProF* είναι αποτελεσματικό καθώς το προφίλ αντιστοιχεί στο μέσο κέρδος ή απώλεια κεφαλαίου των μετοχών.

Σε πιο πρόσφατη μελέτη, οι Kia et al. [26] πρότειναν ένα υβριδικό μοντέλο που κάνει χρήση μάθησης με πλήρη επίβλεψη καθώς και με μερική επίβλεψη, που ονομάζεται *HyS3*, για την πρόβλεψη καθημερινής κατεύθυνσης κίνησης για τις καθημερινές αγορές σε ολόκληρο τον κόσμο. Το τμήμα που χρησιμοποιεί μερική επίβλεψη της *HyS3* που βασίζεται σε γραφήματα διαμορφώνει τις παγκόσμιες αλληλεπιδράσεις των αγορών μέσω ενός δικτύου σχεδιασμένου με έναν νέο συνεχή αλγόριθμο κατασκευής γραφημάτων με βάση το Kruskal. Ακόμα, το κομμάτι με μάθηση με πλήρη επίβλεψη του μοντέλου εισάγει τα αποτελέσματα που προέρχονται από τα ιστορικά δεδομένα κάθε αγοράς στο δίκτυο όποτε το επιτρέπει το υβριδικό μοντέλο με έναν καινοτόμο μηχανισμό υπό όρους. Με βάση τα αριθμητικά τους πειράματα, οι συγγραφείς κατέληξαν στο συμπέρασμα ότι το προτεινόμενο μοντέλο που χρησιμοποιεί ιστορικά δεδομένα αγοράς για κάθε αγορά μαζί με δεδομένα από άλλες παγκόσμιες αγορές θα μπορούσε να παράσχει μεγαλύτερη ακρίβεια από άλλα υπάρχοντα πρότυπα πρόβλεψης.

3.3 Πρόβλημα έγκρισης πίστωσης

Τις τελευταίες δεκαετίες, οι εξελίξεις των συστημάτων μηχανικής μάθησης στη λήψη αποφάσεων σε πιστώσεις έχουν αποκτήσει δημοτικότητα, αντιμετωπίζοντας πολλά θέματα στον τραπεζικό και χρηματοπιστωτικό τομέα. Οι Louzada et al. [27] παρουσίασαν μια εκτεταμένη ανασκόπηση, συζητώντας τα

χρονικά της πρόσφατης πιστοληπτικής αξιολόγησης της οικονομικής ανάλυσης και των εξελίξεων και αναλύοντας τα αποτελέσματα μιας προσέγγισης μηχανικής μάθησης. Επιπλέον, περιέγραψαν λεπτομερώς τα πιο ακριβή μοντέλα πρόγνωσης που χρησιμοποιήθηκαν για να αποκτήσουν σημαντικές πληροφορίες σχετικά με το πρόβλημα βαθμολόγησης της πιστοληπτικής ικανότητας και διεξήγαγαν μια ποικιλία πειραμάτων, χρησιμοποιώντας τρία σύνολα δεδομένων πραγματικού κόσμου (Αυστραλιανή βαθμολογία πιστοληπτικής ικανότητας, ιαπωνική βαθμολογία πιστοληπτικής ικανότητας και γερμανική βαθμολογία πιστοληπτικής ικανότητας). Ορισμένες μελέτες επιβράβευσης έχουν πραγματοποιηθεί τα τελευταία χρόνια. ορισμένα χρήσιμα αποτελέσματα αυτών παρουσιάζονται συνοπτικά παρακάτω.

Οι Kennedy et al. [28] αξιολόγησε την καταλληλότητα των αλγορίθμων με μερική επίβλεψη κατηγοριοποίησης μιας κατηγορίας έναντι των αλγορίθμων με πλήρη επίλεψη δύο κατηγοριών για το πρόβλημα του χαρτοφυλακίου χαμηλής προεπιλογής. Χρησιμοποιήθηκαν εννέα τραπεζικά σύνολα δεδομένων και δημιουργήθηκε τεχνητά ανισορροπία κλάσης, αφαιρώντας το 10% των παραληρηματικών παρατηρήσεων από το σύνολο εκπαίδευσης μετά από κάθε εκτέλεση. Επιπλέον, διερεύνησαν επίσης την καταλληλότητα της υπερδειγματοληψίας, η οποία συνιστά κοινή προσέγγιση για την αντιμετώπιση χαρτοφυλακίων χαμηλής προεπιλογής. Τα πειραματικά αποτελέσματά τους κατέδειξαν ότι οι τεχνικές με αλγορίθμους με μερική επίβλεψη δεν πρέπει να αναμένεται να ξεπεράσουν τις τεχνικές πλήρης επίλεψης ταξινόμησης δύο κατηγοριών και θα πρέπει να χρησιμοποιούνται μόνο στην πλησιέστερη ή πλήρη απουσία παραβατών. Επιπλέον, παρά το γεγονός ότι η υπερβολική δειγματοληψία βελτίωσε την απόδοση ορισμένων ταξινομητών δύο κατηγοριών, δεν οδηγεί σε συνολική βελτίωση των ταξινομητών με τις καλύτερες επιδόσεις.

Οι Alaraj και Abbod [29] εισήγαγαν ένα μοντέλο βασισμένο στον συνδυασμό υβριδικών και συνθετικών μεθόδων για βαθμολόγηση πιστώσεων. Πρώτον, συνδυάζουν μεθόδους φιλτραρίσματος και επιλογής χαρακτηριστικών για την ανάπτυξη ενός αποτελεσματικού προ-επεξεργαστή για μοντέλα μηχανικής μάθησης. Επιπλέον, πρότειναν έναν νέο κανόνα συνδυασμού ταξινομητή βασισμένο στην προσέγγιση συναίνεσης διαφορετικών αλγορίθμων

ταξινόμησης κατά τη διάρκεια της φάσης μοντελοποίησης του συνόλου. Η πειραματική ανάλυσή τους σε επτά σύνολα δεδομένων πραγματικού κόσμου δείχνει ότι το προτεινόμενο μοντέλο παρουσίαζε καλύτερη πρόβλεψη σε σχέση με τους μεμονωμένους ταξινομητές.

Οι Abellán και Castellano [30] πραγματοποίησαν μια συγκριτική μελέτη σχετικά με αρκετούς βασικούς ταξινομητές που χρησιμοποιήθηκαν σε διαφορετικά σύνολα για τα καθήκοντα αξιολόγησης πιστώσεων. Επιπλέον, αξιολόγησαν την απόδοση του Δέντρου Αποφάσεων Πιστότητας (*Credal Decision Tree (CDT)*), το οποίο χρησιμοποιεί ασαφείς πιθανότητες και μέτρα αβεβαιότητας για τη δημιουργία ενός δέντρου αποφάσεων. Μέσω μιας πειραματικής μελέτης, κατέληξαν στο συμπέρασμα ότι όλα τα εξεταζόμενα σύνολα παρουσιάζουν καλύτερες επιδόσεις όταν χρησιμοποιούν μοντέλο CDT ως βασικό μαθητή σε προβλήματα βαθμολόγησης της πιστοληπτικής ικανότητας.

Σε πιο πρόσφατη μελέτη, οι Tripathi et al. [31] πρότειναν ένα υβριδικό μοντέλο πιστοληπτικής αξιολόγησης με βάση τη μείωση των διαστάσεων από τον αλγόριθμο της γειτονιάς *Rough Set* για την επιλογή χαρακτηριστικών και την ταξινόμηση των ομάδων με σταθμισμένη προσέγγιση ψηφοφορίας για την ενίσχυση της απόδοσης ταξινόμησης. Έχουν προτείνει έναν νέο αλγόριθμο ταξινόμησης ταξινομητή ως υποκείμενο μοντέλο για την αναπαράσταση των τάξεων των ταξινομητών με βάση την ακρίβεια ταξινομητή. Τα πειραματικά αποτελέσματα αποκάλυψαν την αποτελεσματικότητα και την ευρωστία της προτεινόμενης μεθόδου σε δύο σύνολα δεδομένων που βαθμολογούσαν τα κριτήρια αξιολόγησης.

Οι Zhang et al. [32] πρότειναν ένα νέο προγνωστικό μοντέλο το οποίο βασίζεται σε μια νέα τεχνική για την επιλογή των ταξινομητών χρησιμοποιώντας έναν γενετικό αλγόριθμο, λαμβάνοντας υπόψη τόσο την ακρίβεια όσο και την ποικιλομορφία του συνόλου. Πραγματοποίησαν ποικιλία πειραμάτων, χρησιμοποιώντας τρία σύνολα δεδομένων πραγματικού κόσμου (Αυστραλιανή βαθμολογία πιστοληπτικής ικανότητας, ιαπωνική βαθμολογία πιστοληπτικής ικανότητας και γερμανική βαθμολογία πιστοληπτικής ικανότητας) για να διερευνήσουν την αποτελεσματικότητα του προτεινόμενου μοντέλου τους. Με βάση τα αριθμητικά τους πειράματα, οι συγγραφείς κατέληξαν στο συμπέρασμα

ότι η προτεινόμενη μέθοδος σύνολων τους υπερέχει των κλασσικών ταξινομητών όσον αφορά την ακρίβεια της πρόβλεψης.

Οι J. Levatić et al. [33] πρότειναν μια μέθοδο για μάθηση με μερική επίβλεψη των ταξινομικών δέντρων. Τα δέντρα μπορούν να ταξινομηθούν με ονομαστικά και αριθμητικά χαρακτηριστικά σε σύνολα δεδομένων δυαδικής και πολυκλασσικής ταξινόμησης. Επιπλέον, πραγματοποίησαν μια εκτενή εμπειρική αξιολόγηση του πλαισίου τους χρησιμοποιώντας ένα σύνολο από δέντρα αποφάσεων ως ταξινομητές βάσης λαμβάνοντας κάποια ενδιαφέροντα αποτελέσματα. Κατά τη διάρκεια αυτής της γραμμής, επέκτειναν τη δουλειά τους, παρουσιάζοντας ορισμένους αλγόριθμους που βασίζονται σε σύνολο για προβλήματα παλινδρόμησης πολλαπλών στόχων [33, 34].

Κεφάλαιο 4.
Αριθμητικά αποτελέσματα

4.1 Σύνολα δεδομένων

Ένα *σύνολο δεδομένων* (data set ή dataset) είναι μια συλλογή από δεδομένα. Συνηθέστερα ένα σύνολο δεδομένων αντιστοιχεί στο περιεχόμενο ενός μόνο πίνακα βάσης δεδομένων ή ενός ενιαίου πίνακα στατιστικών δεδομένων όπου κάθε στήλη του πίνακα αντιπροσωπεύει μια συγκεκριμένη μεταβλητή και κάθε σειρά αντιστοιχεί σε ένα δεδομένο μέλος του εν λόγω συνόλου δεδομένων. Το σύνολο δεδομένων παραθέτει τιμές για κάθε μια από τις μεταβλητές, όπως το ύψος και το βάρος ενός αντικειμένου, για κάθε μέλος του συνόλου δεδομένων. Κάθε τιμή είναι γνωστή ως δεδομένο. Το σύνολο δεδομένων μπορεί να περιλαμβάνει δεδομένα για ένα ή περισσότερα μέλη, που αντιστοιχούν στον αριθμό των σειρών. Για να αξιολογήσουμε των επιρροή των unlabeled data στους αλγορίθμους, χρησιμοποιήσαμε το R=10%, R=20% και R=30% των δεδομένων ως labeled δεδομένα και υπόλοιπο ως unlabeled δεδομένα

4.1.1 Σύνολο δεδομένων του δείκτη Dow Jones

Το *σύνολο δεδομένων του δείκτη Dow Jones* περιλαμβάνει 750 παρουσίες από το *UCI Machine Learning Repository* σχετικά με τις εβδομαδιαίες μετρήσεις κάθε αποθέματος *DJIA* στα πρώτα και στα δεύτερα οικονομικά τρίμηνα του 2011. Αποτελείται απο 10 επεξηγηματικές μεταβλητές που χωρίζονται σε 4 χαρακτηριστικά αφορούν την τιμή του, 4 χαρακτηριστικά αφορούν τον αριθμό

των μετοχών ανα βδομάδα και 2 χαρακτηριστικά σχετίζονται με τον αριθμό των ημερών μέχρι το επόμενο μέρισμα [18].

4.1.2 Σύνολο δεδομένων Australian credit card

Το συγκεκριμένο σύνολο δεδομένων έχει 690 περιπτώσεις (*Instances*), με 14 επεξηγηματικές μεταβλητές που χωρίζονται σε 6 συνεχείς και 8 κατηγορηματικές βάση του *UCI Machine Learning Repository*. Επίσης, στο αυστραλιανό σύνολο δεδομένων υπάρχει μια μικρή ανισορροπία απόρριψης και αποδοχής περιπτώσεις, δηλαδή 383 και 307, αντίστοιχα [18].

4.1.3 Σύνολο δεδομένων Japanese credit card

Το συγκεκριμένο σύνολο δεδομένων έχει 653 περιπτώσεις (*Instances*), με 14 επεξηγηματικές μεταβλητές που χωρίζονται σε 3 συνεχείς, 3 ακέραιες και 9 κατηγορικές, βάση του *UCI Machine Learning Repository*. Επίσης, στο αυστραλιανό σύνολο δεδομένων υπάρχει μια μικρή ανισορροπία απόρριψης και αποδοχής περιπτώσεις, δηλαδή 357 και 296, αντίστοιχα [18].

4.1.4 Σύνολο δεδομένων German credit card

Το συγκεκριμένο σύνολο δεδομένων έχει 1000 περιπτώσεις (*Instances*), με 20 επεξηγηματικές μεταβλητές που χωρίζονται σε 7 συνεχείς και 13 κατηγορηματικές, βάση του *UCI Machine Learning Repository*. Επίσης, στο γερμανικό σύνολο δεδομένων μια έντονη ανισορροπία παρατηρείται, με 300 αρνητικές αποφάσεις έναντι 700 θετικών [18].

4.2 Μετρικές απόδοσης

Μετρική απόδοσης [35] ονομάζεται το κριτήριο ποσοτικοποίησης της απόδοσης ενός συστήματος. Η απόδοση των αλγορίθμων ταξινόμησης αξιολογήθηκε χρησιμοποιώντας τα ακόλουθα τέσσερα μετρήσεις απόδοσης: Sensitivity (Sen), Specificity (Spe), F_1 και Accuracy (Acc), οι οποίες ορίζονται αντίστοιχα από:

$$Sen = \frac{T_P}{T_P + F_N}$$

$$Spe = \frac{T_N}{T_N + F_P}$$

$$F_1 = \frac{2T_P}{2T_P + F_N + F_P}$$

$$Acc = \frac{T_P + T_N}{T_P + T_N + F_P + F_N}$$

όπου το T_P αντιπροσωπεύει τον αριθμό των περιπτώσεων που έχουν ταξινομηθεί σωστά ως θετικά, το TN είναι σταθερό για τον αριθμό των περιπτώσεων που έχουν ταξινομηθεί σωστά ως αρνητικές, το F_P βρίσκεται για τον αριθμό των περιπτώσεων που έχουν ταξινομηθεί εσφαλμένα ως θετικές, βρίσκεται το F_N για τον αριθμό των περιπτώσεων που έχουν ταξινομηθεί εσφαλμένα ως αρνητικές. Αξίζει να σημειωθεί ότι η ευαισθησία της ταξινόμησης είναι η αναλογία των πραγματικών θετικών που θεωρούνται θετικά. Η εξειδίκευση αντιπροσωπεύει το ποσοστό των πραγματικών αρνητικών που προβλέπονται ως αρνητικό, το F_1 αποτελείται από έναν αρμονικό μέσο ακρίβειας και ανάκλησης ενώ η ακρίβεια είναι η αναλογία σωστών προβλέψεις ενός μοντέλου ταξινόμησης [18].

4.3 Σύγκριση αλγορίθμων

Βασισμένοι στα *σύνολα δεδομένων* (datasets) που αναφέρθηκαν στη παράγραφο 4.1 έγιναν μελέτες και τα αποτελέσματα καθώς και το ποιος αλγόριθμος είχε το καλύτερο αποτέλεσμα παρουσιάζεται στους Πίνακες 1-15.

4.3.1 Σύγκριση του αλγορίθμου CST-Voting με τους αλγορίθμους Self-training, Co-training και Tri-training

Στη συνέχεια, αξιολογήσαμε την απόδοση ταξινόμησης του αλγορίθμου που παρουσιάστηκε, CST-Voting, ενάντια σε κάποιους άλλους υψηλής τεχνολογίας self-labeled αλγορίθμους όπως ο SETRED, το Co-Forest και η Democratic-Co learning. Παρατηρήστε ότι ο CST-Voting χρησιμοποιεί το NB και το LMT ως βασικούς ταξινομητές, που και αυτοί αντίστοιχα παρουσίασαν την καλύτερη απόδοση, σε σχέση με όλες τις μετρικές αποδόσεις.

Στους Πίνακες 1-12 παρουσιάζουμε τα αποτελέσματα της αξιολόγησης των αλγορίθμων Self-training, Co-training, Tri-training και CST-Voting. Ο αλγόριθμος CST-Voting υπερτερεί έχοντας την καλύτερη συνολική απόδοση, καθώς όπως μπορούμε να δούμε ξεπερνά τους υπόλοιπους self-labeled αλγόριθμους στα στατισικά αποτελέσματα.

Base learner	Algorithm	Ratio = 10%			
		Sen	**Spe**	**F1**	**Acc**
Naïve Bayes	**Self-Training**	73.9%	88.3%	78.4%	81.9%
	Co-Training	**78.2%**	83.6%	78.7%	81.2%
	Tri-Training	61.2%	**91.6%**	71.3%	78.1%
	CST-Voting	75.6%	90.3%	**80.6%**	**83.8%**
SMO	**Self-Training**	88.9%	79.1%	82.7%	83.5%
	Co-Training	**92.2%**	79.1%	84.5%	84.9%
	Tri-Training	77.5%	**86.7%**	79.9%	82.6%
	CST-Voting	89.9%	84.9%	**86.1%**	**87.1%**
MLP	**Self-Training**	82.1%	87.7%	83.2%	85.2%
	Co-Training	80.8%	87.7%	82.4%	84.6%
	Tri-Training	71.3%	**89.0%**	77.1%	81.2%
	CST-Voting	**82.4%**	88.0%	**83.5%**	**85.5%**
kNN	**Self-Training**	73.9%	88.3%	78.4%	81.9%
	Co-Training	**78.2%**	83.6%	78.7%	81.2%
	Tri-Training	61.2%	**91.6%**	71.3%	78.1%
	CST-Voting	75.6%	90.3%	**80.6%**	**83.8%**

Πίνακας 1: Αξιολόγηση των επιδόσεων των Self-training, Co-training, Tri-training και CST-Voting στο Australian credit dataset.

Base learner	Algorithm	Ratio = 20%			
		Sen	**Spe**	**F1**	**Acc**
Naive Bayes	**Self-Training**	**78.2%**	91.4%	82.8%	85.5%
	Co-Training	77.5%	**92.7%**	**83.1%**	85.9%
	Tri-Training	76.2%	86.7%	79.1%	82.0%
	CST-Voting	**78.2%**	91.9%	83.0%	**85.8%**
SMO	**Self-Training**	85.7%	**83.3%**	83.0%	84.3%
	Co-Training	**94.1%**	79.1%	**85.5%**	**85.8%**
	Tri-Training	89.3%	83.0%	84.8%	**85.8%**
	CST-Voting	90.6%	80.4%	84.2%	84.9%
MLP	**Self-Training**	80.1%	88.0%	82.1%	84.5%
	Co-Training	79.8%	**91.4%**	**83.8%**	**86.2%**
	Tri-Training	**83.1%**	82.2%	81.0%	82.6%
	CST-Voting	82.4%	88.0%	83.5%	85.5%
kNN	**Self-Training**	73.3%	88.3%	78.0%	81.6%
	Co-Training	**77.5%**	84.6%	78.8%	81.4%
	Tri-Training	67.8%	**91.6%**	76.1%	81.0%
	CST-Voting	74.6%	90.9%	**80.2%**	**83.6%**

Πίνακας 2: Αξιολόγηση των επιδόσεων των Self-training, Co-training, Tri-training και CST-Voting στο Australian credit dataset.

Base learner	Algorithm	Ratio = 30%			
		Sen	**Spe**	**F1**	**Acc**
Naive Bayes	**Self-Training**	78.2%	90.3%	82.2%	84.9%
	Co-Training	78.8%	91.4%	83.2%	85.8%
	Tri-Training	**79.8%**	86.7%	81.3%	83.6%
	CST-Voting	79.2%	**91.6%**	**83.5%**	**86.1%**
SMO	**Self-Training**	88.9%	81.7%	84.0%	84.9%
	Co-Training	**94.1%**	79.1%	85.5%	85.8%
	Tri-Training	89.3%	80.9%	83.8%	84.6%
	CST-Voting	93.8%	**82.0%**	**86.7%**	**87.2%**
MLP	**Self-Training**	82.7%	86.9%	83.1%	85.1%
	Co-Training	79.5%	**91.1%**	83.4%	85.9%
	Tri-Training	**89.3%**	83.0%	**84.8%**	85.8%
	CST-Voting	85.0%	87.2%	84.6%	**86.2%**
kNN	**Self-Training**	73.3%	91.4%	79.6%	79.6%
	Co-Training	**78.8%**	87.5%	81.1%	81.1%
	Tri-Training	74.9%	89.3%	79.6%	79.6%
	CST-Voting	78.5%	**92.2%**	**83.4%**	**83.4%**

Πίνακας 3: Αξιολόγηση των επιδόσεων των Self-training, Co-training, Tri-training και CST-Voting στο Australian credit dataset.

Base learner	Algorithm	Ratio = 10%			
		Sen	**Spe**	**F1**	**Acc**
Naive Bayes	**Self-Training**	75.3%	88.2%	79.5%	82.4%
	Co-Training	**83.1%**	86.8%	**83.5%**	85.1%
	Tri-Training	74.3%	88.2%	78.9%	81.9%
	CST-Voting	79.4%	**90.8%**	83.3%	**85.6%**
SMO	**Self-Training**	92.2%	81.0%	85.7%	86.1%
	Co-Training	**93.9%**	79.8%	86.1%	86.2%
	Tri-Training	86.5%	**86.3%**	85.2%	86.4%
	CST-Voting	93.6%	80.7%	**86.3%**	**86.5%**
MLP	**Self-Training**	**84.1%**	87.4%	**84.4%**	**85.9%**
	Co-Training	81.1%	88.8%	83.3%	85.3%
	Tri-Training	69.3%	88.0%	75.4%	79.5%
	CST-Voting	80.7%	**90.2%**	83.9%	**85.9%**
kNN	**Self-Training**	75.7%	88.2%	79.7%	82.5%
	Co-Training	**79.4%**	85.4%	80.6%	82.7%
	Tri-Training	56.1%	88.2%	65.9%	73.7%
	CST-Voting	76.0%	**90.8%**	**81.2%**	**84.1%**

Πίνακας 4: Αξιολόγηση των επιδόσεων των Self-training, Co-training, Tri-training και CST-Voting στο Japanese credit dataset.

Base learner	Algorithm	Ratio = 20%			
		Sen	**Spe**	**F1**	**Acc**
	Self-Training	**79.1%**	90.2%	82.8%	85.1%

Naive Bayes	**Co-Training**	78.7%	90.8%	**82.9%**	**85.3%**
	Tri-Training	73.6%	**91.6%**	80.1%	83.5%
	CST-Voting	78.0%	91.3%	82.8%	**85.3%**
SMO	**Self-Training**	91.9%	81.0%	85.5%	85.9%
	Co-Training	**93.9%**	79.8%	**86.1%**	**86.2%**
	Tri-Training	79.7%	**86.0%**	81.1%	83.2%
	CST-Voting	93.2%	80.1%	85.8%	86.1%
MLP	**Self-Training**	**86.1%**	85.7%	84.7%	85.9%
	Co-Training	82.8%	89.9%	84.9%	86.7%
	Tri-Training	65.2%	**91.6%**	74.4%	79.6%
	CST-Voting	84.1%	89.9%	**85.7%**	**87.3%**
kNN	**Self-Training**	76.4%	88.5%	80.3%	83.0%
	Co-Training	**79.1%**	86.6%	81.0%	83.2%
	Tri-Training	59.8%	**93.0%**	71.1%	77.9%
	CST-Voting	76.7%	90.2%	**81.4%**	**84.1%**

Πίνακας 5: Αξιολόγηση των επιδόσεων των Self-training, Co-training, Tri-training και CST-Voting στο Japanese credit dataset.

Base learner	**Algorithm**	**Ratio = 30%**			
		Sen	**Spe**	**F1**	**Acc**
Naive Bayes	**Self-Training**	**79.1%**	90.8%	83.1%	85.5%
	Co-Training	79.7%	91.6%	**84.0%**	**86.2%**
	Tri-Training	73.0%	90.5%	79.1%	82.5%
	CST-Voting	**79.1%**	**92.2%**	83.9%	**86.2%**
SMO	**Self-Training**	92.9%	80.7%	85.9%	86.2%
	Co-Training	**93.9%**	80.1%	86.2%	86.4%
	Tri-Training	74.7%	84.0%	77.0%	79.8%
	CST-Voting	93.2%	**86.6%**	**89.0%**	**89.6%**
MLP	**Self-Training**	**86.1%**	87.7%	85.7%	87.0%
	Co-Training	82.8%	89.6%	84.8%	86.5%
	Tri-Training	65.2%	**93.3%**	75.2%	80.6%
	CST-Voting	84.1%	90.2%	**85.9%**	**87.4%**
kNN	**Self-Training**	75.3%	89.6%	80.2%	83.2%
	Co-Training	**83.1%**	85.4%	82.8%	84.4%
	Tri-Training	74.3%	**95.2%**	82.6%	85.8%
	CST-Voting	79.4%	92.2%	**84.1%**	**86.4%**

Πίνακας 6: Αξιολόγηση των επιδόσεων των Self-training, Co-training, Tri-training και CST-Voting στο Japanese credit dataset.

Base learner	Algorithm	Ratio = 10%			
		Sen	**Spe**	**F1**	**Acc**
Naive Bayes	**Self-Training**	80.7%	45.3%	79.1%	70.1%
	Co-Training	80.0%	45.0%	78.6%	69.5%
	Tri-Training	81.6%	45.7%	79.6%	70.8%
	CST-Voting	**81.7%**	**46.0%**	**79.8%**	**71.0%**
SMO	**Self-Training**	84.6%	44.7%	81.2%	72.6%
	Co-Training	84.3%	45.0%	81.1%	72.5%
	Tri-Training	84.4%	45.7%	81.3%	72.8%
	CST-Voting	**86.4%**	**46.0%**	**82.5%**	**74.3%**
MLP	**Self-Training**	84.6%	**47.0%**	81.6%	73.3%
	Co-Training	85.4%	43.3%	81.5%	72.8%
	Tri-Training	**87.4%**	45.0%	**82.9%**	**74.7%**
	CST-Voting	87.0%	46.0%	82.8%	**74.7%**
kNN	**Self-Training**	84.6%	40.0%	80.4%	71.2%
	Co-Training	85.4%	40.7%	81.0%	72.0%
	Tri-Training	**87.4%**	40.7%	**82.1%**	73.4%
	CST-Voting	85.0%	**47.7%**	82.0%	**73.8%**

Πίνακας 7: Αξιολόγηση των επιδόσεων των Self-training, Co-training, Tri-training και CST-Voting στο German credit dataset.

Base learner	Algorithm	Ratio = 20%			
		Sen	**Spe**	**F1**	**Acc**
Naive Bayes	**Self-Training**	84.6%	**48.7%**	81.9%	73.8%
	Co-Training	85.7%	46.7%	82.2%	74.0%
	Tri-Training	86.0%	46.0%	82.2%	74.0%
	CST-Voting	**86.4%**	47.7%	**82.8%**	**74.8%**
SMO	**Self-Training**	86.4%	45.0%	82.3%	74.0%
	Co-Training	86.0%	**47.3%**	82.5%	74.4%
	Tri-Training	86.7%	46.7%	82.8%	74.7%
	CST-Voting	**87.0%**	47.0%	**83.0%**	**75.0%**
MLP	**Self-Training**	86.4%	**47.3%**	**82.7%**	**74.7%**
	Co-Training	86.0%	44.0%	81.9%	73.4%
	Tri-Training	86.7%	44.0%	82.3%	73.9%
	CST-Voting	**87.1%**	45.0%	**82.7%**	74.5%
kNN	**Self-Training**	86.4%	42.3%	81.9%	73.2%
	Co-Training	86.0%	41.7%	81.5%	72.7%
	Tri-Training	86.7%	42.7%	82.1%	73.5%
	CST-Voting	**87.0%**	**46.7%**	**82.9%**	**74.9%**

Πίνακας 8: Αξιολόγηση των επιδόσεων των Self-training, Co-training, Tri-training και CST-Voting στο German credit dataset.

Base learner	Algorithm	Ratio = 30%			
		Sen	**Spe**	**F1**	**Acc**
Naive Bayes	**Self-Training**	84.6%	50.3%	82.2%	74.3%
	Co-Training	86.4%	**51.7%**	83.4%	76.0%
	Tri-Training	87.1%	51.0%	83.7%	76.3%
	CST-Voting	**87.9%**	**51.7%**	**84.2%**	**77.0%**
SMO	**Self-Training**	87.1%	46.0%	82.9%	74.8%
	Co-Training	87.0%	**48.3%**	83.2%	75.4%
	Tri-Training	**87.4%**	47.3%	83.3%	75.4%
	CST-Voting	**87.4%**	48.0%	**83.4%**	**75.6%**
MLP	**Self-Training**	87.1%	**48.3%**	83.3%	75.5%
	Co-Training	87.4%	44.3%	82.8%	74.5%
	Tri-Training	87.9%	45.0%	83.1%	75.0%
	CST-Voting	**88.3%**	47.0%	**83.7%**	**75.9%**
kNN	**Self-Training**	86.1%	43.3%	81.9%	73.3%
	Co-Training	87.4%	43.7%	82.6%	74.3%
	Tri-Training	**87.9%**	44.0%	**82.9%**	**74.7%**
	CST-Voting	86.4%	**46.0%**	82.5%	74.3%

Πίνακας 9: Αξιολόγηση των επιδόσεων των Self-training, Co-training, Tri-training και CST-Voting στο German credit dataset.

Base learner	Algorithm	Ratio = 10%		
		Sen	**Spe**	**Acc**
Naive Bayes	**Self-Training**	61.4%	**45.7%**	50.5%
	Co-Training	**72.3%**	37.0%	51.5%
	Tri-Training	70.7%	38.0%	51.3%
	CST-Voting	71.2%	38.6%	**51.8%**
SMO	**Self-Training**	61.4%	**32.6%**	44.4%
	Co-Training	58.2%	21.7%	37.7%
	Tri-Training	64.1%	27.2%	43.1%
	CST-Voting	**65.8%**	31.0%	**45.6%**
MLP	**Self-Training**	77.2%	24.5%	47.9%
	Co-Training	66.3%	**32.1%**	46.4%
	Tri-Training	73.4%	27.7%	47.7%
	CST-Voting	**78.8%**	27.2%	**50.0%**
3*NN*	**Self-Training**	56.0%	47.3%	48.7%
	Co-Training	55.4%	51.6%	50.5%
	Tri-Training	53.3%	**58.7%**	52.8%
	CST-Voting	**58.7%**	54.3%	**53.3%**
LMT	**Self-Training**	78.8%	17.9%	45.6%
	Co-Training	63.6%	**22.8%**	40.8%
	Tri-Training	73.4%	**22.8%**	45.4%
	CST-Voting	**81.5%**	21.7%	**48.7%**
JRip	**Self-Training**	74.5%	21.2%	45.1%
	Co-Training	74.5%	**26.6%**	**47.7%**
	Tri-Training	75.0%	21.7%	45.6%
	CST-Voting	**77.2%**	22.8%	47.2%

Πίνακας 10: Αξιολόγηση των επιδόσεων των Self-training, Co-training, Tri-training και CST-Voting στο Dow Jones dataset.

Base learner	Algorithm	Ratio = 20%		
		Sen	**Spe**	**Acc**
Naive Bayes	**Self-Training**	71.7%	35.3%	50.5%
	Co-Training	**72.3%**	37.0%	51.5%
	Tri-Training	66.8%	**38.6%**	49.7%
	CST-Voting	**72.3%**	38.0%	**52.1%**
SMO	**Self-Training**	71.2%	**35.3%**	50.3%
	Co-Training	75.5%	22.8%	46.4%
	Tri-Training	74.5%	32.1%	50.3%
	CST-Voting	**78.8%**	28.8%	**50.8%**
MLP	**Self-Training**	76.1%	24.5%	47.4%
	Co-Training	72.3%	**31.0%**	48.7%
	Tri-Training	69.6%	**31.0%**	47.4%
	CST-Voting	**78.8%**	28.3%	**50.5%**
3*NN*	**Self-Training**	51.6%	51.1%	48.5%
	Co-Training	52.2%	51.1%	48.7%
	Tri-Training	53.3%	**58.7%**	52.8%
	CST-Voting	**56.5%**	58.2%	**54.1%**
LMT	**Self-Training**	80.4%	21.2%	47.9%
	Co-Training	74.5%	19.0%	44.1%
	Tri-Training	82.1%	**25.5%**	50.8%
	CST-Voting	**84.2%**	23.9%	**51.0%**
JRip	**Self-Training**	78.3%	**29.3%**	50.8%
	Co-Training	81.0%	27.2%	**51.0%**
	Tri-Training	79.3%	23.9%	48.7%
	CST-Voting	**81.5%**	26.1%	50.8%

Πίνακας 11: Αξιολόγηση των επιδόσεων των Self-training, Co-training, Tri-training και CST-Voting στο Dow Jones dataset.

Base learner	Algorithm	Ratio = 30%		
		Sen	Spe	Acc
Naive Bayes	Self-Training	66.3%	**41.3%**	50.8%
	Co-Training	72.3%	37.0%	51.5%
	Tri-Training	**72.8%**	37.0%	51.8%
	CST-Voting	71.7%	39.7%	**52.6%**
SMO	Self-Training	71.7%	**37.0%**	51.3%
	Co-Training	78.3%	36.4%	**54.1%**
	Tri-Training	79.9%	34.8%	**54.1%**
	CST-Voting	**81.0%**	33.7%	**54.1%**
MLP	Self-Training	77.2%	22.3%	46.9%
	Co-Training	72.8%	**31.0%**	49.0%
	Tri-Training	75.0%	29.3%	49.2%
	CST-Voting	**85.9%**	29.3%	**54.4%**
3*NN*	Self-Training	55.1%	22.3%	51.3%
	Co-Training	54.3%	**31.0%**	51.3%
	Tri-Training	53.4%	29.3%	52.6%
	CST-Voting	**60.3%**	29.3%	**56.2%**
LMT	Self-Training	82.6%	22.8%	49.7%
	Co-Training	77.2%	25.0%	48.2%
	Tri-Training	**83.7%**	22.8%	50.3%
	CST-Voting	83.2%	**26.1%**	**51.5%**
JRip	Self-Training	70.7%	**31.0%**	47.9%
	Co-Training	79.9%	28.8%	51.3%
	Tri-Training	78.8%	25.0%	49.0%
	CST-Voting	**83.7%**	27.2%	**52.3%**

Πίνακας 12: Αξιολόγηση των επιδόσεων των Self-training, Co-training, Tri-training και CST-Voting στο Dow Jones dataset.

4.3.2 Σύγκριση του αλγορίθμου CST-Voting με τους κλασικούς self-labeled αλγόριθμους

Στη συνέχεια, αξιολογήσαμε την απόδοση ταξινόμησης του αλγορίθμου που παρουσιάστηκε, CST-Voting, ενάντια σε κάποιους άλλους υψηλής τεχνολογίας self-labeled αλγορίθμους όπως ο SETRED, το Co-Forest και η Democratic-Co learning. Παρατηρήστε ότι ο CST-Voting χρησιμοποιεί το SMO και το MLP ως βασικούς ταξινομητές, που και αυτοί αντίστοιχα παρουσίασαν την καλύτερη απόδοση, σε σχέση με όλες τις μετρικές αποδόσεις.

Στους Πίνακες 13-15 παρουσιάζουμε τα αποτελέσματα της αξιολόγησης των αλγορίθμων SETRED, Co-Forest, Demo-Co και CST-Voting. Ο αλγόριθμος CST-

Voting υπερτερεί έχοντας την καλύτερη συνολική απόδοση, καθώς όπως μπορούμε να δούμε ξεπερνά τους υπόλοιπους κλασικούς self-labeled αλγόριθμους στα στατισικά αποτελέσματα.

Algorithm	Ratio = 10%			
	Sen	**Spe**	**F1**	**Acc**
SETRED	87.9%	78.3%	81.8%	82.6%
Co-Forest	81.4%	**87.5%**	82.6%	84.8%
Demo-Co	82.7%	82.0%	80.6%	82.3%
CST- Voting	**89.9%**	84.9%	**86.1%**	**87.1%**

Algorithm	Ratio = 20%			
	Sen	**Spe**	**F1**	**Acc**
SETRED	87.6%	82.2%	83.5%	84.6%
Co-Forest	80.5%	**89.0%**	82.9%	**85.2%**
Demo-Co	83.1%	85.4%	82.5%	84.3%
CST- Voting	**90.6%**	80.4%	**84.2%**	84.9%

Algorithm	Ratio = 30%			
	Sen	**Spe**	**F1**	**Acc**
SETRED	91.2%	82.8%	85.8%	86.5%
Co-Forest	81.4%	**91.4%**	84.7%	87.0%
Demo-Co	84.0%	86.9%	83.9%	85.7%
CST- Voting	**93.8%**	82.0%	**86.7%**	**87.2%**

Πίνακας 13: Αξιολόγηση των επιδόσεων των SETRED, Co-Forest, Democratic-Co learning, CST-Voting στο Australian credit dataset.

Algorithm	Ratio = 10%			
	Sen	**Spe**	**F1**	**Acc**
SETRED	91.2%	81.2%	85.3%	85.8%
Co-Forest	84.5%	88.5%	85.2%	86.7%
Demo-Co	85.5%	84.6%	83.8%	85.0%
CST- Voting	93.6%	80.7%	86.3%	86.5%

Algorithm	Ratio = 20%			
	Sen	**Spe**	**F1**	**Acc**
SETRED	92.2%	81.2%	85.8%	86.2%
Co-Forest	85.1%	**89.4%**	**86.0%**	**87.4%**
Demo-Co	84.5%	85.7%	83.8%	85.1%
CST- Voting	**93.2%**	80.1%	85.8%	86.1%

Algorithm	Ratio = 30%			
	Sen	**Spe**	**F1**	**Acc**
SETRED	92.9%	81.5%	86.3%	86.7%
Co-Forest	85.1%	**89.9%**	86.3%	87.7%
Demo-Co	84.8%	85.7%	83.9%	85.3%
CST- Voting	**93.2%**	86.6%	**89.0%**	**89.6%**

Πίνακας 14: Αξιολόγηση των επιδόσεων των SETRED, Co-Forest, Democratic-Co learning, CST-Voting Japanese credit dataset.

Algorithm	Ratio = 10%			
	Sen	**Spe**	**F1**	**Acc**
SETRED	84.3%	44.7%	81.0%	72.4%
Co-Forest	85.7%	45.0%	81.9%	73.5%
Demo-Co	83.6%	43.7%	80.5%	71.6%
CST- Voting	**86.4%**	**46.0%**	**82.5%**	**74.3%**

Algorithm	Ratio = 20%			
	Sen	**Spe**	**F1**	**Acc**
SETRED	86.7%	45.0%	82.5%	74.2%
Co-Forest	**87.1%**	45.0%	82.7%	74.5%
Demo-Co	86.0%	45.3%	82.1%	73.8%
CST- Voting	87.0%	**47.0%**	**83.0%**	**75.0%**

Algorithm	Ratio = 30%			
	Sen	**Spe**	**F1**	**Acc**
SETRED	**87.4%**	46.7%	83.2%	75.2%
Co-Forest	87.3%	46.7%	83.1%	75.1%
Demo-Co	87.0%	**48.0%**	83.1%	75.3%
CST- Voting	**87.4%**	**48.0%**	**83.4%**	**75.6%**

Πίνακας 15: Αξιολόγηση των επιδόσεων των SETRED, Co-Forest, Democratic-Co learning, CST-Voting German credit dataset.

Κεφάλαιο 5. Συμπεράσματα

Στην παρούσα εργασία έγινε μια εισαγωγή στις βασικές της έννοιες, αναφορά στους στόχους, τα στάδια, τις κατηγορίες και τις μεθόδους της τεχνητής νοημοσύνης και της μηχανικής μάθησης.

Η εργασία στηρίχθηκε στην εξόρυξης γνώσης οικονομικών δεδομένων. Όσα αναφέρθηκαν πιο πάνω καθώς και οι αλγόριθμοι μπορούν να εφαρμοστούν σε σε οποιοδήποτε σύνολο δεδομένων (dataset) τους δοθεί, είτε είναι από τον κλάδο της ιατρικής, είτε τον επιχειρήσεων μέχρι ακόμα και τον κλάδο της εκπαίδευσης.

Κατά την διάρκεια συγγραφής της πτυχιακής αυτής εργασίας έγινε δημοσίευση άρθρου με τίτλο «Performance evaluation of a SSL algorithm for forecasting the Dow Jones index» [19]. Στην έρευνα μας αυτή, αξιολογούμε την απόδοση ενός semi-supervised αλγορίθμου, CST-Voting, για την πρόβλεψη της κίνησης του δείκτη Dow Jones. Τα πειραματικά μας αποτελέσματα υποδηλώνουν ότι ο προτεινόμενος αλγόριθμος υπερετεί ένατι άλλων semi-supervised αλγόριθμων, που δείχνει ότι θα μπορούσαν να αναπτυχθούν αξιόπιστα και αξιόπιστα μοντέλα πρόβλεψης χρησιμοποιώντας μερικά ετικέτα και πολλά μη επισημασμένα δεδομένα, που δείχνει ότι θα μπορούσαν να αναπτυχθούν αξιόπιστα μοντέλα πρόβλεψης, βασισμένα σε αυτόν, χρησιμοποιώντας μερικά labeled και πολλά unlabeled data.

Τέλος, για όσα μελετήθηκαν στην εργασία, έγιναν μέσα από την έρευνα υλικού πάνω σε ελληνικά, ξένα συγγράμματα ενώ υπήρξαν φόρες που η χρήση του διαδικτύου ήταν απαραίτητη.

Βιβλιογραφία

[1] N. Nilsson, Artificial Intelligence: A New Synthesis, 1998.

[2] Council και NRC, Developments in Artificial Intelligence, 1999.

[3] S. J. Russell και P. Norvig, Artificial Intelligence: A Modern Approach, 2003.

[4] R. Kurzweil, The Singularity is Near, 2005.

[5] P. Domingos και M. Pazzani, «On the optimality of the simple Bayesian classifier under zero-one loss,» *Machine Learning 29,* pp. 103-130, 1997.

[6] P. Churchland, «Neurophilosophy: Toward a Unified Science of the Mind/Brain, MIT Press,» *Synapse 1,* pp. 221-222, 1986.

[7] J. Platt, «Using sparseness and analytic QP to speed training of support vector machines,» *Proceedings of the 1998 conference on Advances in neural information processing systems II,* pp. 557-563, 1999.

[8] D. Aha, «Lazy Learning,» 1997.

[9] S. L. Salzberg, C4.5: Programs for Machine Learning, 1993.

[10] N. Landwehr, M. Hall και E. Frank, «Logistic model trees,» *Machine Learning,* τόμ. 50, αρ. 1-2, pp. 161-205, 2005.

[11] D. Yarowsky, «Unsupervised word sense disambiguation rivaling supervised methods,» *ACL '95 Proceedings of the 33rd annual meeting on Association for Computational Linguistics,* pp. 189-196 , 1995.

[12] A. Blum και T. Mitchell, «Combining labeled and unlabeled data with co-training.,» *COLT' 98 Proceedings of the eleventh annual conference on Computational learning theory,* pp. 92-100 , 1998.

[13] Z. Zhou και M. Li, «Tri-training: Exploiting unlabeled data using three classifiers,» *IEEE Transactions on Knowledge and Data Engineering,* τόμ. 17, αρ. 11, pp. 1529 - 1541, 2005.

[14] Y. Zhou και S. Goldman, «Democratic co-learning,» *ICTAI '04 Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence,* pp. 594-202 , 2004.

[15] M. Li και Z. Zhou, «Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples,» *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans,* τόμ. 37, αρ. 6, pp. 1088 - 1098, 2007.

[16] M. Li και Z. Zhou, «SETRED: Self-training with editing,» *PAKDD 2005: Advances in Knowledge Discovery and Data Mining ,* pp. 611-621, 2005.

[17] M. F. A. Hady και F. Schwenker, «Combining Committee-Based Semi-Supervised Learning and Active Learning,» *Journal of Computer Science and Technology,* τόμ. 25, αρ. 4, p. 681–698, 2010.

[18] I. E. Livieris, N. Kiriakidou, A. Kanavos, V. Tampakas και P. Pintelas, «On Ensemble SLL Algorithms for Credit Scoring Problem,» *Informatics,* pp. 1-16, 2018.

[19] I. Livieris, A. Kanavos, G. Vonitsanos, N. Kiriakidou, A. Vikatos, K. Giotopoulos και V. Tampakas, «Performance evaluation of a SSL algorithm for forecasting the Dow Jones index,» *IEEE 9th International Conference on Information, Intelligence, Systems and Applications (IISA 2018),* pp. 1-8, 2018.

[20] E. Hajizadeh, H. D. Ardakani και J. Shahrabi, «Application of data mining techniques in stock markets: A survey,» *Journal of Economics and International Finance,* τόμ. 2, pp. 109-118, 2010.

[21] S. Thawornwong και D. Enke, «The use of data mining and neural networks for forecasting stock market returns,» *Expert Systems with Applications: An International Journal,* τόμ. 29, αρ. 4, pp. 927-940 , 2005.

[22] K. S. Kannan, P. S. Sekar, M. M. Sathik και P. Arumugam, «Financial stock market forecast using data mining techniques.,» *International MultiConference of Engineers and Computer Scientists,* τόμ. I, pp. 1-5, 2010.

[23] S. R. Nanda, B. Mahanty και a. M. Tiwari, «Clustering indian stock market data for portfolio management,» *Expert Systems with Applications,* τόμ. 37, p. 8793–8798, 2010.

[24] J. Patel, S. Shah, P. Thakkar και a. K. Kotecha, «Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques,» *Expert Systems with Applications,* τόμ. 42, αρ. 1, pp. 259-268, 2015.

[25] K. Khor και K. H. Ng, «StockProF: A stock profiling framework using data mining approaches,» *Information Systems and e-Business Management,* τόμ. 15, αρ. 1, p. 139–158, 2016.

[26] A. N. Kia, S. Haratizadeh και S. B. Shouraki, «A hybrid supervised semi-supervised graph-based model to predict one-day ahead movement of global stock markets and commodity prices,» *Expert Systems with Applications,* τόμ. 105, pp. 159-173, 2018.

[27] F. Louzada, A. Ara και G. Fernandes, «Classification methods applied to credit scoring: Systematic review and overall comparison,» *Surveys in Operations Research and Management Science,* τόμ. 21, αρ. 2, pp. 117-134, 2016.

[28] K. Kennedy, B. Namee και S. Delany, «Using semi-supervised classifiers for credit scoring,» *Dublin Institute of Technology ARROW@DIT,* pp. 1-20, 2013.

[29] M. Alaraj και M. Abbod, «A new hybrid ensemble credit scoring model based on classifiers consensus system,» *Expert Systems with Applications,* τόμ. 64, pp. 36-55, 2016.

[30] J. Abellán και J. Castellano, «A comparative study on base classifiers in ensemble methods for credit scoring,» *Expert Systems with Applications,* τόμ. 73, pp. 1-10, 2017.

[31] D. Tripathi, D. Edla και R. Cheruku, «Hybrid credit scoring model using neighborhood rough set and multi-layer ensemble classification,» *Journal of Intelligent & Fuzzy Systems,* τόμ. 34, αρ. 3, pp. 1543-1549, 2018.

[32] H. Zhang, H. He και W. Zhang, «Classifier selection and clustering with fuzzy assignment in ensemble model for credit scoring,» *Neurocomputing,* τόμ. 316, pp. 210-221, 2018.

[33] J. Levatić, M. Ceci, D. Kocev και S. Džeroski, «Self-training for multi-target regression with tree ensembles,» *Knowledge-Based Systems,* τόμ. 123, pp. 41-60, 2017.

[34] J. Levatić, D. Kocev, M. Ceci και S. Džeroski, «Semi-supervised trees for multi-target regression,» *Information Sciences,* τόμ. 450, pp. 109-127, 2018.

[35] D. M. W. Powers, « Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation,» *Bioinfo Publications,* 2011.

[36] Ε. Παρασύρη, «Εξόρυξη γνώσης και δεδομένων σε επιχειρήσεις,» 2014.

[37] P. Murphy και D. Aha, «UCI repository of machine learning databases.,» 1994.

[38] I. E. Livieris, «A new ensemble semi-supervised self-labeled algorithm,» *Informatica,* αρ. 49, pp. 1-14, 2018.

[39] F. S. Mohamed Farouk Abdel Hady, «Co-Training by Committee: A New Semi-Supervised Learning Framework,» *2008 IEEE International Conference on Data Mining Workshops,* pp. 563-572, 2008.