ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

Πανεπιστήμιο Ιωαννίνων

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ:

«ΜΕΛΕΤΗ ΛΟΓΙΣΜΙΚΟΥ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ WEKA»

Γίδαρη Ερμιόνη

ΑΜ: 1614 εξάμηνο :10ο

Email: thl1816014@uoi.gr

Επιβλέπων καθηγητής

Γλάβας Ευριπίδης

ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ:

«ΜΕΛΕΤΗ ΛΟΓΙΣΜΙΚΟΥ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ WEKA»

Γίδαρη Ερμιόνη

ΑΜ: 1614 εξάμηνο :10ο

Email: thl1816014@uoi.gr

Επιβλέπων καθηγητής

Γλάβας Ευριπίδης

Εγκρίθηκε από τριμελή εξεταστική επιτροπή

Άρτα, Αύγουστος 2023

ΕΠΙΤΡΟΠΗ ΑΞΙΟΛΟΓΗΣΗΣ

1. Επιβλέπων καθηγητής

Γλαβάς Ευριπίδης

2. Μέλος επιτροπής

Λιάγκου Βασιλική

3. Μέλος επιτροπής

Τζάλλας Αλέξανδρος

Ο Προϊστάμενος του Τμήματος

Δήλωση μη λογοκλοπής

Δηλώνω υπεύθυνα και γνωρίζοντας τις κυρώσεις του Ν. 2121/1993 περί Πνευματικής Ιδιοκτησίας, ότι η παρούσα πτυχιακή εργασία είναι εξ ολοκλήρου αποτέλεσμα δικής μου ερευνητικής εργασίας, δεν αποτελεί προϊόν αντιγραφής ούτε προέρχεται από ανάθεση σε τρίτους. Όλες οι πηγές που χρησιμοποιήθηκαν (κάθε είδους, μορφής και προέλευσης) για τη συγγραφή της περιλαμβάνονται στη βιβλιογραφία.

Γίδαρη Ερμιόνη

Υπογραφή

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω τον καθηγητή μου κ. Γλαβά Ευριπίδη για την ορθή καθοδήγηση, την βοήθειά του και τον χρόνο που αφιέρωσε για την πτυχιακή μου εργασία.

ΠΕΡΙΛΗΨΗ

Στην παρούσα πτυχιακή εργασία θα ασχοληθούμε με το λογισμικό Weka. Στο πρώτο κεφάλαιο θα ορίσουμε την εξόρυξη δεδομένων, τους στόχους της, την ιστορία καθώς επίσης και τα πλεονεκτήματα και τα μειονεκτήματα του Weka. Επίσης θα αναφερθούμε σε ορισμένες εφαρμογές σε διάφορους τομείς όπως η ιατρική, η οικονομία και οι τηλεπικοινωνίες.

Στο δεύτερο κεφάλαιο θα μιλήσουμε για τις εκδόσεις του Weka και άλλα ελεύθερα λογισμικά επιχειρηματικής Ευφυΐας και εξόρυξης δεδομένων.

Στο τρίτο κεφάλαιο θα κάνουμε μία ανάλυση στο γραφικό περιβάλλον του Weka και στον τρόπο λειτουργίας του καθώς επίσης θα παρουσιάσουμε και ορισμένα από τα χαρακτηριστικά του.

Το τέταρτο κεφάλαιο θα αφορά τα βήματα εγκατάστασης ένα προς ένα αναλυτικά.

Το πέμπτο θα αφορά τους αλγόριθμους μηχανικής μάθησης όπως ο K- means , SVM και δέντρα αποφάσεων.

Στο έκτο και τελευταίο κεφάλαιο θα βρούμε ένα αρχείο .arff και θα τρέξουμε στο weka τους : K- means, SMO, J48 και θα πάρουμε τα αντίστοιχα αποτελέσματα.

Λέξεις κλειδιά: K- means , SVM και δέντρα αποφάσεων, Weka, εξόρυξη δεδομένων.

ABSTRACT

In this thesis we will deal with the Weka software. In the first chapter we will define data mining, its goals, history as well as the advantages and disadvantages of Weka. We will also refer to some applications in various fields such as medicine, economy and telecommunications.

In the second chapter we will talk about the versions of Weka and other free software for business Intelligence and data mining.

In the third chapter we will analyze the graphical environment of Weka and its mode of operation as well as present some of its features.

The fourth chapter will cover the installation steps one by one in detail.

The fifth will be about machine learning algorithms such as K-means, SVM and decision trees.

In the sixth and last chapter we will find an .arff file and will run it in weka for: K-means, SMO, J48 and we will get the corresponding results.

Keywords: K-means, SVM and decision trees, Weka, data mining.

Πίνακας περιεχομένων

ΛΙΣΤΑ ΕΙΚΟΝΩΝ

Εικόνα 8: 4ο Βήμα εγκατάστασης του WEKA

https://www.softwaretestinghelp.com/weka-tutorial/?fbclid=IwAR3z6GaQJZQw5m0hjJRnuBKMB_NQfKV7g-PRWK2HDKLhNf5RUtXfI1oMNhE

Εικόνα 9: 5ο Βήμα εγκατάστασης του WEKA

https://www.softwaretestinghelp.com/weka-tutorial/?fbclid=IwAR3z6GaQJZQw5m0hjJRnuBKMB_NQfKV7g-PRWK2HDKLhNf5RUtXfI1oMNhE

Εικόνα 10: Βήμα εγκατάστασης του WEKA

https://www.softwaretestinghelp.com/weka-tutorial/?fbclid=IwAR3z6GaQJZQw5m0hjJRnuBKMB_NQfKV7g-PRWK2HDKLhNf5RUtXfI1oMNhE

Εικόνα 11 : 6ο Βήμα εγκατάστασης του WEKA

https://www.softwaretestinghelp.com/weka-tutorial/?fbclid=IwAR3z6GaQJZQw5m0hjJRnuBKMB_NQfKV7g-PRWK2HDKLhNf5RUtXfI1oMNhE

Εικόνα 12: 7ο Βήμα εγκατάστασης του WEKA

https://www.softwaretestinghelp.com/weka-tutorial/?fbclid=IwAR3z6GaQJZQw5m0hjJRnuBKMB_NQfKV7g-PRWK2HDKLhNf5RUtXfI1oMNhE

Εικόνα 13: 8ο Βήμα εγκατάστασης του WEKA

https://www.softwaretestinghelp.com/weka-tutorial/?fbclid=IwAR3z6GaQJZQw5m0hjJRnuBKMB_NQfKV7g-PRWK2HDKLhNf5RUtXfI1oMNhE

Εικόνα 14: Simple CLI

https://www.softwaretestinghelp.com/weka-tutorial/?fbclid=IwAR3z6GaQJZQw5m0hjJRnuBKMB_NQfKV7g-PRWK2HDKLhNf5RUtXfI1oMNhE

Εικόνα 15: Simple CLI

https://www.softwaretestinghelp.com/weka-tutorial/?fbclid=IwAR3z6GaQJZQw5m0hjJRnuBKMB_NQfKV7g-PRWK2HDKLhNf5RUtXfI1oMNhE

Εικόνα 16:Explorer

https://www.softwaretestinghelp.com/weka-tutorial/?fbclid=IwAR3z6GaQJZQw5m0hjJRnuBKMB_NQfKV7g-PRWK2HDKLhNf5RUtXfI1oMNhE

Εικόνα 17: Experimenter

https://www.softwaretestinghelp.com/weka-tutorial/?fbclid=IwAR3z6GaQJZQw5m0hjJRnuBKMB_NQfKV7g-PRWK2HDKLhNf5RUtXfI1oMNhE

Εικόνα 18: Experimenter

https://www.softwaretestinghelp.com/weka-tutorial/?fbclid=IwAR3z6GaQJZQw5m0hjJRnuBKMB_NQfKV7g-PRWK2HDKLhNf5RUtXfI1oMNhE

Εικόνα 19:Knowledge Flow

https://www.softwaretestinghelp.com/weka-tutorial/?fbclid=IwAR3z6GaQJZQw5m0hjJRnuBKMB_NQfKV7g-PRWK2HDKLhNf5RUtXfI1oMNhE

Εικόνα 20:Workbench

https://www.softwaretestinghelp.com/weka-tutorial/?fbclid=IwAR3z6GaQJZQw5m0hjJRnuBKMB_NQfKV7g-PRWK2HDKLhNf5RUtXfI1oMNhE

Εικόνα 21: Το αρχείο ARFF
https://www.softwaretestinghelp.com/weka-tutorial/?fbclid=IwAR3z6GaQJZQw5m0hjJRnuBKMB_NQfKV7g-PRWK2HDKLhNf5RUtXfI1oMNhE

Εικόνα 22: K-means

https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

Εικόνα 23 : K-means Παράδειγμα

https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

Εικόνα 24 : K-means Παράδειγμα

https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

Εικόνα 25: K-means Παράδειγμα

https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

Εικόνα 26: K-means Παράδειγμα

https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

Εικόνα 27: K-means Παράδειγμα

https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

Εικόνα 28: K-means Παράδειγμα

https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

Εικόνα 29: K-means Παράδειγμα

https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

Εικόνα 30: K-means Παράδειγμα

https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

Εικόνα 31: K-means Παράδειγμα

https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

Εικόνα 32: K-means Παράδειγμα

https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

Εικόνα 33: K-means Παράδειγμα

https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

Εικόνα 34: SVM

https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm?fbclid=IwAR2YLkKfadgdlKY0RrW1TZFcComY_U_9uFxHRiZTbYffo_seOgetVxaLIlI

Εικόνα 35: Παράδειγμα SVM

https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm?fbclid=IwAR2YLkKfadgdlKY0RrW1TZFcComY_U_9uFxHRiZTbYffo_seOgetVxaLIlI

Εικόνα 36: Παράδειγμα SVM

https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm?fbclid=IwAR2YLkKfadgdlKY0RrW1TZFcComY_U_9uFxHRiZTbYffo_seOgetVxaLIlI

Εικόνα 37: Παράδειγμα SVM

https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm?fbclid=IwAR2YLkKfadgdlKY0RrW1TZFcComY_U_9uFxHRiZTbYffo_seOgetVxaLIlI

Εικόνα 38: Παράδειγμα SVM

https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm?fbclid=IwAR2YLkKfadgdlKY0RrW1TZFcComY_U_9uFxHRiZTbYffo_seOgetVxaLIlI

Εικόνα 39: Παράδειγμα SVM

https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm?fbclid=IwAR2YLkKfadgdlKY0RrW1TZFcComY_U_9uFxHRiZTbYffo_seOgetVxaLIlI

Εικόνα 40: Παράδειγμα SVM

https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm?fbclid=IwAR2YLkKfadgdlKY0RrW1TZFcComY_U_9uFxHRiZTbYffo_seOgetVxaLIlI

Εικόνα 41: Παράδειγμα SVM

https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm?fbclid=IwAR2YLkKfadgdlKY0RrW1TZFcComY_U_9uFxHRiZTbYffo_seOgetVxaLIlI

Εικόνα 42: Παράδειγμα Decision Tree

https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm?fbclid=IwAR20uDOsXQ_wlHhKFZaVOFgICE_TUfGBU2LJIEmR74vl8z-ai0Hd5jy2BIo

Εικόνα 43: Παράδειγμα αρχείου .arff

Κεφάλαιο 1: Εξόρυξη δεδομένων

Εισαγωγή

Εξόρυξη δεδομένων είναι η ανακάλυψη μεγάλων βάσεων δεδομένων χρησιμοποιώντας αλγόριθμους ομαδοποίησης ή ταξινόμησης και τις ακόλουθες αρχές πληροφοριών ή μοτίβα:

- Στατιστικά στοιχεία
- Τεχνητή νοημοσύνη
- Μηχανική μάθηση και
- Συστήματα βάσεων δεδομένων.

Εικόνα 1 : Data mining

1.1 Ορισμός

Ο όρος εξόρυξη δεδομένων αναφέρεται γενικά σε κάθε τύπο μορφής ή επεξεργασίας δεδομένων αλλά και σε κάθε τύπο συστήματος υποστήριξης αποφάσεων, συμπεριλαμβανομένης της τεχνητής νοημοσύνης, της μηχανής μάθηση και επιχειρηματική ευφυΐα. Όταν ο όρος χρησιμοποιείται σωστά, η λέξη-κλειδί είναι ανακάλυψη, που ορίζεται ως ανακάλυψη κάτι νέου.

1.2 Στόχος

Ο μόνος πραγματικός στόχος της εξόρυξης δεδομένων είναι η αυτόματη ή ημιαυτόματη ανάλυση μεγάλων ποσοτήτων δεδομένων για την εξαγωγή ορισμένων προηγουμένως άγνωστων ενδιαφέροντων μοτίβων, όπως ομάδες εγγραφών δεδομένων (ομαδοποίηση), ασυνήθιστες εγγραφές (ανίχνευση ανωμαλιών) και εξαρτήσεις (κανόνες συσχέτισης). Αυτό συνήθως περιλαμβάνει τη χρήση μιας βάσης δεδομένων όπως ένα ευρετήριο. Αυτά τα μοτίβα μπορούν στη συνέχεια να θεωρηθούν ως περιγραφή των δεδομένων εισόδου και να χρησιμοποιηθούν για περαιτέρω ανάλυση ή για παράδειγμα μηχανική μάθηση και προγνωστική ανάλυση. Για παράδειγμα, η εξόρυξη δεδομένων μπορεί να αναγνωρίσει πολλαπλά σύνολα στα δεδομένα, τα οποία στη συνέχεια μπορούν να χρησιμοποιηθούν για να διασφαλιστούν πιο ακριβή αποτελέσματα για συστήματα υποστήριξης αποφάσεων. Η συλλογή δεδομένων, η προετοιμασία των δεδομένων και η ερμηνεία των αποτελεσμάτων και των εκθέσεων δεν αποτελούν μέρος της εξόρυξης δεδομένων, αλλά ανήκουν στην ανακάλυψη γνώσης από βάσεις δεδομένων ως κάποια πρόσθετα βήματα.

1.3 Ιστορία

Η πρώτη προσέγγιση για τον εντοπισμό προτύπων είναι η Bayesian θεωρία και η ανάλυση παλινδρόμησης. Η ευρεία χρήση και η ανάπτυξη της τεχνολογίας των υπολογιστών έχει αυξήσει τον όγκο των δεδομένων που συλλέγονται και την ανάγκη για αποτελεσματική επεξεργασία. Καθώς ο όγκος και η πολυπλοκότητα της συλλογής δεδομένων συνεχίζει να αυξάνεται, η μη αυτόματη ανάλυση δεδομένων έχει αντικατασταθεί από την αυτοματοποιημένη επεξεργασία δεδομένων. Άλλες ανακαλύψεις της επιστήμης των υπολογιστών συνέβαλαν επίσης σε αυτό, όπως τα νευρωνικά δίκτυα, η ομαδοποίηση, οι γενετικοί αλγόριθμοι (δεκαετία 1950), τα δέντρα αποφάσεων (δεκαετία 1960) και οι μηχανές διανυσμάτων υποστήριξης (δεκαετία του 1990). Η εξόρυξη δεδομένων είναι η διαδικασία εφαρμογής αυτών των μεθόδων σε δεδομένα για την ανακάλυψη άγνωστων μοτίβων σε μεγάλα σύνολα δεδομένων.

Αυτό γεφυρώνει το χάσμα μεταξύ εφαρμοσμένων στατιστικών και τεχνητής νοημοσύνης μέσω της διαχείρισης βάσεων δεδομένων, χρησιμοποιώντας τον τρόπο αποθήκευσης και οργάνωσης σε βάσεις δεδομένων για την αποτελεσματικότερη εκτέλεση θεωρητικών και διαθέσιμων αλγορίθμων. σε μεγάλα σύνολα δεδομένων.

1.4 Πλεονεκτήματα και μειονεκτήματα Weka

1.4.1 Πλεονεκτήματα

- Είναι λογισμικό ανοιχτού κώδικα
- Η ελευθερία χρήσης του προγράμματος χωρίς περιορισμούς
- Έχει εύκολη χρήση
- Η προσαρμοστικότητα του προγράμματος για τη δημιουργία νέων μεθόδων μηχανικής μάθησης
- Έχει πολλά εργαλεία
- Διαθέτει δωρεάν online μαθήματα
- Παρέχει πολλά βιβλία και δημοσιεύσεις
- Οι νέες ανακαλύψεις της τεχνητής νοημοσύνης εφαρμόζονται ταχύτατα
- Ποιοτικά αποτελέσματα
- Δεν χρειάζεται να γνωρίζει ο χρήστης κάποια γλώσσα προγραμματισμού

1.4.2 Μειονεκτήματα

- Καθυστερεί να φορτώσει τα δεδομένα
- Δεν μπορεί να επεξεργαστεί πολλά δεδομένα μαζί
- Δεν έχει ποιοτικό γραφικό περιβάλλον για τον χρήστη

1.5 Εφαρμογές

1.5.1 Ιατρική

Τα τελευταία χρόνια, η εξόρυξη δεδομένων έχει χρησιμοποιηθεί ευρέως στη βιοϊατρική, το DNA, τη γενετική, τα φαρμακευτικά προϊόντα και άλλους ιατρικούς τομείς. Στον τομέα της γενετικής, στόχος είναι η κατανόηση της χαρτογράφησης της σχέσης μεταξύ των αλλαγών στις αλληλουχίες του ανθρώπινου DNA και της προδιάθεσης για ασθένειες. Η εξόρυξη δεδομένων αποτελεί σημαντικό εργαλείο για τη βελτίωση της διάγνωσης, της πρόληψης και της θεραπείας των ασθενειών. Η αύξηση της βιοϊατρικής έρευνας απαιτεί την εξέταση γενετικών προτύπων και λειτουργιών μεγάλης κλίμακας. Τα εργαλεία εξόρυξης δεδομένων μπορούν να βοηθήσουν σημαντικά στη μελέτη της σύνθεσης του DNA και στην εύρεση των ποικίλων μοτίβων και λειτουργιών του. Ένας από τους κύριους στόχους που σχετίζονται με

την ανάλυση δεδομένων DNA είναι η σύγκριση διαφορετικών αλληλουχιών και η αναζήτηση ομοιοτήτων μεταξύ δεδομένων DNA. Η σύγκριση αυτή περιλαμβάνει κυρίως αλληλουχίες γονιδίων από υγιείς και κατεστραμμένους ιστούς και την εύρεση διαφορών μεταξύ αυτών των δύο τύπων. Αυτό μπορεί να επιτευχθεί με την αναζήτηση και των δύο κατηγοριών των αλληλουχιών υγιών και κατεστραμμένων γονιδίων και την εύρεση συχνών μορφών των δύο κατηγοριών. Η ανάλυση αυτή βοηθά στην εύρεση ομοιοτήτων και διαφορών στις γενετικές αλληλουχίες. Στη βιοϊατρική, οι περισσότερες ασθένειες μελετώνται για να διαπιστωθεί αν προκαλούνται από έναν συνδυασμό γονιδίων. Οι μέθοδοι συσχέτισης μπορούν να χρησιμοποιηθούν για τον προσδιορισμό της συνύπαρξης ομάδων γονιδίων και επίσης για τη μελέτη αλληλεπιδράσεων και σχέσεων μεταξύ γονιδίων. Τα εργαλεία οπτικοποίησης διαδραματίζουν επίσης σημαντικό ρόλο στην εξόρυξη δεδομένων στη βιοϊατρική. Τα εργαλεία αυτά μπορούν να εμφανίζουν πολύπλοκες δομές γονιδίων σε γραφήματα, δέντρα και αλυσίδες. Οι οπτικές απεικονίσεις βοηθούν στην καλύτερη κατανόηση αυτών των δομών για την ανακάλυψη γνώσης και την εξερεύνηση δεδομένων. Διαφορετικοί συνδυασμοί γονιδίων προκαλούν διαφορετικές ασθένειες, αλλά τα γονίδια αυτά ενεργοποιούνται σε διαφορετικά επίπεδα. Η ανάλυση μονοπατιών χρησιμοποιείται για τη σύνδεση διαφορετικών γονιδίων με διαφορετικά στάδια της εξέλιξης της νόσου. Η ανάλυση μονοπατιών διαδραματίζει σημαντικό ρόλο στη γενετική.

Εικόνα 2: Data mining στην υγεία

1.5.2 Οικονομία

Ένας άλλος τομέας στον οποίο εφαρμόζεται η εξόρυξη δεδομένων είναι η οικονομία. Τα οικονομικά δεδομένα συλλέγονται κυρίως από τράπεζες και άλλους οικονομικούς οργανισμούς. Τα δεδομένα αυτά είναι συνήθως αξιόπιστα, περιεκτικά και υψηλής ποιότητας

και ως εκ τούτου ,απαιτούν συστηματικές μεθόδους ανάλυσης .Η συμβολή της εξόρυξης δεδομένων στην οικονομία έγκειται στη συλλογή και κατανόηση των δεδομένων, στην επεξεργασία των δεδομένων ,στη δημιουργία και εκτίμηση μοντέλων και στην ανάπτυξή τους. Η ορθή ανάλυση των οικονομικών δεδομένων διευκολύνει την καλύτερη λήψη αποφάσεων, ενεργώντας σύμφωνα με την ανάλυση της αγοράς. Τα εργαλεία και οι τεχνικές εξόρυξης δεδομένων μπορούν να βοηθήσουν στην ανάλυση των οικονομικών δεδομένων με τους ακόλουθους τρόπους Τα δεδομένα που συλλέγονται από διάφορα χρηματοπιστωτικά ιδρύματα, όπως οι τράπεζες, συγκεντρώνονται αρχικά σε μια αποθήκη δεδομένων. Για την ανάλυση αυτών των δεδομένων που συλλέγονται σε αποθήκες δεδομένων χρησιμοποιούνται τεχνικές ανάλυσης δεδομένων λόγω των γενικών χαρακτηριστικών τους. Μια άλλη εφαρμογή της εξόρυξης δεδομένων αφορά την πρόβλεψη της αποπληρωμής δανείων και την πιστωτική πολιτική των πελατών. Μέθοδοι εξόρυξης, όπως η επιλογή χαρακτηριστικών,βοηθούνστον εντοπισμό διαφόρων χαρακτηριστικών,όπως τα επίπεδα εισοδήματος των πελατών, οι αποπληρωμές βάσει εισοδήματος και το πιστωτικό ιστορικό. Με την επεξεργασία αυτών των χαρακτηριστικών,οι τράπεζες μπορούν να λαμβάνουν αποφάσεις για την πολιτική δανεισμού με σχετικά χαμηλό κίνδυνο. Οι τεχνικές ομαδοποίησης και ταξινόμησης βοηθούν τα χρηματοπιστωτικά ιδρύματα να ομαδοποιήσουν διάφορους πελάτες με κοινά χαρακτηριστικά.Οι αποτελεσματικές τεχνικές ομαδοποίησης και φιλτραρίσματος μπορούν να βοηθήσουν τις τράπεζες να εντοπίσουν ομάδες πελατών, να συσχετίσουν νέους πελάτες με τις τρέχουσες ομάδες και να παρέχουν κοινά οφέλη. Τα εργαλεία εξόρυξης δεδομένων βοηθούν τα χρηματοπιστωτικά ιδρύματα να εντοπίζουν την απάτη και το έγκλημα από διάφορες βάσεις δεδομένων και το ιστορικό συναλλαγών των πελατών. Οι τεχνικές οπτικοποίησης βοηθούν στην παρουσίαση δεδομένων σε διάφορες μορφές, όπως γραφήματα με βάση συγκεκριμένα χαρακτηριστικά. Παρουσιάζοντας δεδομένα από διαφορετικές οπτικές γωνίες, οι τράπεζες μπορούν να διακρίνουν τους πελάτες που επιχειρούν παράνομες δραστηριότητες και στη συνέχεια να διερευνήσουν λεπτομερώς αυτές τις ύποπτες περιπτώσεις για τον εντοπισμό απάτης και εγκλήματος

Εικόνα 3 : Data mining στην οικονομία

1.5.3 Τηλεπικοινωνίες

Ο κλάδος των τηλεπικοινωνιών αναπτύσσεται ραγδαία, το ίδιο και η τεχνολογία. Οι πρόσφατες τηλεπικοινωνιακές υπηρεσίες έχουν επεκταθεί από τις τοπικές και υπεραστικές επικοινωνίες στη χρήση συσκευών φαξ, συσκευών τηλεειδοποίησης, κινητών τηλεφώνων και ηλεκτρονικού ταχυδρομείου.Οι εξελίξεις στην τεχνολογία των τηλεπικοινωνιών οδήγησαν στην ενσωμάτωση τεχνικών εξόρυξης δεδομένων σε αυτές τις τεχνολογίες ώστε να λειτουργούν αποτελεσματικά και να παράγουν αποδοτικά αποτελέσματα. Η εξόρυξη δεδομένων βοηθά στη διάκριση προτύπων στις τηλεπικοινωνίες, στην καταπολέμηση παράνομων δραστηριοτήτων, καθώς και στην καλύτερη χρήση των πόρων και στη βελτίωση της ποιότητας των υπηρεσιών. Η εξόρυξη δεδομένων βελτιώνει τις τηλεπικοινωνιακές υπηρεσίες. Τα τηλεπικοινωνιακά δεδομένα που συλλέγονται περιλαμβάνουν τύπους κλήσεων, τοποθεσία ,διάρκεια κλήσεων κ.λπ. Η ανάλυση βοηθά στον προσδιορισμό και τη σύγκριση του φόρτου του συστήματος, της κίνησης δεδομένων, των κερδών κ.λπ. Η ανάλυση μπορεί να χρησιμοποιεί εργαλεία οπτικοποίησης εξόρυξης δεδομένων για την εμφάνιση διαγραμμάτων όπως οι πόροι του συστήματος και οι προορισμοί. Εργαλεία όπως η οπτικοποίηση συσχέτισης και η ομαδοποίηση παρέχουν χρήσιμες υπηρεσίες για την ανάλυση τηλεπικοινωνιακών δεδομένων. Ένα σημαντικό πρόβλημα που αντιμετωπίζει ο κλάδος των τηλεπικοινωνιών είναι οι παράνομες πρακτικές. Οι δραστηριότητες αυτές μπορεί να σχετίζονται με σκόπιμες κλήσεις αιχμής, διαλείπουσες κλήσεις κ.λπ. οι οποίες με τη σειρά τους έχουν αρνητικό αντίκτυπο στην απόδοση του τηλεπικοινωνιακού δικτύου. Τεχνικές όπως η ομαδοποίηση και η ανάλυση ακραίων τιμών μπορούν να βοηθήσουν στον εντοπισμό παράνομων προτύπων βελτιώνοντας την αποδοτικότητα των τηλεπικοινωνιακών υπηρεσιών. Τα εργαλεία εξόρυξης δεδομένων μπορούν να χρησιμοποιηθούν για τη δημιουργία προφίλ πελατών και την ανίχνευση βλαβών δικτύου. Τέλος, η ανάλυση συσχετιζόμενων και διαδοχικών προτύπων διευκολύνει την προώθηση νέων και ποικίλων τηλεπικοινωνιακών υπηρεσιών.

Εικόνα 4 : Data mining στις τηλεπικοινωνίες

Κεφάλαιο 2: Εισαγωγή στο Weka

Εισαγωγή

Το WEKA είναι μια σουίτα λογισμικού για μηχανική μάθηση και εξόρυξη δεδομένων. Αναπτύχθηκε στο Πανεπιστήμιο Waikato της Νέας Ζηλανδίας και διατίθεται ως ελεύθερο λογισμικό. Η ποικιλία των μεθόδων εξόρυξης δεδομένων που περιλαμβάνονται είναι οι παρακάτω:

- Η συνεχής υποστήριξη και ανάπτυξη από μια διεθνή ομάδα προγραμματιστών
- Η δωρεάν διανομή του πηγαίου κώδικα
- Η δυνατότητα εγκατάστασης σε διάφορες πλατφόρμες λογισμικού.

Αυτοί είναι μερικοί από τους παράγοντες που οδήγησαν στην ευρεία αποδοχή του WEKA. Η γραφική διεπαφή επιτρέπει τη χρήση του από χρήστες χωρίς γνώσεις προγραμματισμού. Το πιο διαδεδομένο περιβάλλον εργασίας είναι το WEKA Explorer. Με το WEKA Explorer, οι χρήστες μπορούν να εκτελούν:

- Προεπεξεργασία δεδομένων
- Ταξινόμηση
- Ανάλυση συστάδων
- Ανάλυση κανόνων συσχέτισης
- Επιλογή χαρακτηριστικών
- Οπτικοποίηση δεδομένων

Όσον αφορά την προεπεξεργασία δεδομένων, γίνεται αναφορά σε διάφορες πηγές δεδομένων. Η γραφική διεπαφή στην καρτέλα "προεπεξεργασία" επιτρέπει την εύκολη πρόσβαση σε δεδομένα και αρχεία ARFF. Η διαγραφή πεδίων και εκτέλεση διαφόρων αλγορίθμων προεπεξεργασίας μπορούν εύκολα να διερευνηθούν. Η προσθήκη νέων υπολογιζόμενων πεδίων, κανονικοποίηση και διακριτοποίηση αριθμών, συγχώνευση ονομαστικών πεδίων δειγματοληψία, μείωση της διαστατικότητας με ανάλυση κύριων συνιστωσών, επιλογή χαρακτηριστικών, κ.λπ.

Οι αλγόριθμοι και τα εργαλεία ταξινόμησης που είναι διαθέσιμα στο WEKA είναι αξιοσημείωτα. Παρέχονται υλοποιήσεις όλων των σημαντικών μεθόδων ταξινόμησης, όπως:

- Δέντρα απόφασης
- Νευρωνικά δίκτυα

- Μηχανές διανυσμάτων υποστήριξης
- Μηχανές κατηγοριοποίησης Bayes
- Παλινδρόμηση
- K-Means

Κάθε μέθοδος προσφέρει πολλές δυνατότητες προσαρμογής. Επιπλέον, εκτός από τις βασικές μεθόδους, υπάρχουν εργαλεία για τη δημιουργία σύνθετων ταξινομητών bagging και boosting, ταξινομητές που χρησιμοποιούν ανάλυση συστάδων κ.λπ. χρήστες μπορούν να επικυρώσει τα μοντέλα χρησιμοποιώντας μεθόδους επικύρωσης, μεθόδους αναμονής ή ανεξάρτητα σύνολα δεδομένων. Λεπτομερή στοιχεία για κάθε μοντέλο που δείχνουν απόδοση και δομή(π.χ. βάρη συνδέσμων για δίκτυα perceptron πολλαπλών στρωμάτων). Το WEKA περιλαμβάνει k-Means, Cumulative Hierarchical AS και τους αλγορίθμους ανάλυσης συστάδων DBSCAN. Κάθε αλγόριθμος μπορεί να παραμετροποιηθεί. Είναι επίσης δυνατή η οπτική αναπαράσταση της κατανομής των παρατηρήσεων εντός μιας συστάδας. Η Καρτέλα Related περιέχει τους ακόλουθους αλγορίθμους: Ανάλυση κανόνων συσχέτισης, συμπεριλαμβανομένων των βασικών apriori αλγορίθμων. Δυνατότητα για εξόρυξη κανόνων συσχέτισης σε δεδομένα με πεδία κλάσεων. Αυτοί οι κανόνες μπορούν να αναλυθούν χρησιμοποιώντας τη δεξιά πλευρά της οθόνης τιμή της κλάσης. Στην καρτέλα "Select Attributes"μπορούν να δοκιμαστούν διαφορετικές μέθοδοι επιλογής. Μπορούν να επιλεγούν χαρακτηριστικά και να χρησιμοποιηθεί ένας συνδυασμός μεθόδων αναζήτησης και αξιολόγησης χαρακτηριστικών.
Η Καρτέλα "Οπτικοποίηση" περιέχει έναν πίνακα με διαγράμματα διασποράς.

2.1 Το WEKA

Το WEKA (Waikato Environment for Knowledge Analysis) όπως προαναφέραμε είναι ένα λογισμικό για μηχανική μάθηση και εξόρυξη δεδομένων. Το WEKA ανήκει στη λεγόμενη κατηγορία του ″ελεύθερου λογισμικού″ (freeware) και είναι γενικά διαθέσιμο με τους ακόλουθους όρους.
Η μεγάλη δημοτικότητα του οφείλεται σε ιδιαίτερα χαρακτηριστικά και τι δυνατότητές του. Περισσότερες πληροφορίες για το WEKA:

- Περιλαμβάνει ένα ευρύ φάσμα μεθόδων, όπως ταξινόμηση, παλινδρόμηση, ανάλυση συστάδων και κανόνες συσχέτισης.
- Προσφέρει επίσης δυνατότητες προεπεξεργασία δεδομένων, παρέχονται επίσης εργαλεία οπτικοποίησης.
- Είναι ένα Λογισμικό Ανοικτού κώδικα, δηλαδή ο πηγαίος κώδικας είναι διαθέσιμος στο κοινό και όσοι έχουν γνώσεις προγραμματισμού έχουν τη δυνατότητα να επέμβουν στον αλγόριθμο.
- Είναι σε γλώσσα Java οπότε μπορεί να εγκατασταθεί σε ένα ευρύ φάσμα πλατφορμών υλικού και λογισμικού.
- Διαθέτει ποιοτικό γραφικό περιβάλλον εργασίας.

Στο διαδίκτυο διατίθεται μεγάλη ποικιλία βιβλιοθηκών μηχανικής μάθησης και εξόρυξης δεδομένων. Ωστόσο, για να τις χρησιμοποιήσετε πρέπει να γράψετε κώδικα. Αντίθετα, η γραφική διεπαφή του WEKA επιτρέπει στους τελικούς χρήστες χωρίς γνώσεις προγραμματισμού να χρησιμοποιούν το λογισμικό. Ακόμη και οι τελικοί χρήστες χωρίς γνώσεις προγραμματισμού μπορούν να χρησιμοποιήσουν το λογισμικό.

2.2 Οι εκδόσεις του WEKA

Το WEKA διατίθεται σε δύο διαφορετικές εκδόσεις:

- Η λεγόμενη ″σταθερή″ (stable) έκδοση είναι για τους τελικούς χρήστες και αντιστοιχεί στην τελευταία έκδοση του βιβλίου των Witten, Frank and Hall (2011).
- Στην έκδοση για προγραμματιστές. Αυτή η έκδοση χρησιμοποιείται από την κοινότητα προγραμματιστών του WEKA για να διορθωνουν τυχόν σφάλματα και για να διευρύνουν τις δυνατότητες του λογισμικού.

Το πανεπιστήμιο του Waikato διατηρεί ένα website αφιερωμένο στο WEKA (http://www.cs.waikato.ac.nz/ml/weka/). Εδώ οι χρήστες μπορούν:

- Να εγκαταστήσουν το WEKA (http://www.cs.waikato.ac.nz/ml/weka/downloading.html), σε λειτουργικά συστήματα Windows, Mac OS X και Linux.

- Να βρουν το manual του λογισμικού, καθοδήγηση για την αντιμετώπιση προβλημάτων και για σύνδεση με γλώσσες όπως το Matlab και η R, διαδικτυακά σεμινάρια κ.α.
- Να προμηθευτούν το Application Programming Interface (API)του λογισμικού, και άλλα πακέτα για εργασίες learning machine και data mining.
- Να αποκτήσουν δεδομένα και να τα χρησιμοποιήσουν για εξάσκηση.

2.3 Ελεύθερα λογισμικά Επιχειρηματικής Ευφυΐας και Εξόρυξης Δεδομένων

Το WEKA δεν είναι το μόνο ελεύθερο λογισμικό εξόρυξης δεδομένων. Άλλα λογισμικά επιχειρηματικής ευφυΐας και εξόρυξης δεδομένων. Στην κατηγορία των δωρεάν λογισμικών επιχειρηματικής ευφυΐας, οι πιο γνωστές περιπτώσεις, χωρίς ιδιαίτερη σειρά παρουσίασης, περιλαμβάνουν:

- IBM Watson Analytics.
- Microsoft Power BI.
- SAP Lumira Cloud.
- Pentaho Community Edition.
- Jaspersoft.
- Jedox Base Business Intelligence.
- SpagoBI.
- KNIME.
- Tableau Public.

Εκτός από το λογισμικό επιχειρηματικής ευφυΐας που έχει σχεδιαστεί ειδικά για επιχειρήσεις, υπάρχουν ακόμα διάφορα δωρεάν λογισμικά εξόρυξης δεδομένων,εκ των οποίων ορισμένα παρέχονται με τη μορφή κώδικα όπου χρήζει η γνώση προγραμματισμού και άλλα έχουν γραφικό περιβάλλον εργασίας και μπορούν να χρησιμοποιηθούν από τελικούς χρήστες
Το WEKA είναι ένα απ αυτά, αλλά υπάρχουν και άλλα:

- Orange: λογισμικό εξόρυξης δεδομένων και μηχανικής μάθησης. Χρησιμοποιείται και με οπτική γλώσσα προγραμματισμού, με ιδιαίτερη ικανότητα απεικόνισης δεδομένων. Μπορούν να το χρησιμοποιήσουν όλοι.
- Rattle GUI: υπάρχει γραφική διεπαφή όπου χρησιμοποιείται η γλωσσα R. H R είναι μια ισχυρή και ταχέως ανερχόμενη γλώσσα προγραμματισμού για Εξόρυξη Δεδομένων, στατιστική ανάλυση και δημιουργία γραφικών. Για άμεση χρήση είναι απαραίτητη η γνώση προγραμματισμού.
- KEEL: λογισμικό με ερευνητικούς και εκπαιδευτικούς στόχους το οποίο επικεντρώνεται στους εξελικτικούς αλγορίθμους. Έχει απλό γραφικό περιβάλλον εργασίας και περιλαμβάνει μεθόδους προεπεξεργασίας δεδομένων, στατιστικής ανάλυσης, μηχανικής μάθησης κλπ. και απλοποιεί τη διεξαγωγή πειραμάτων.
- TANAGRA: δωρεάν ακαδημαϊκό λογισμικό για διδασκαλία και έρευνα το οποίο παρέχει επιβλεπόμενη μάθηση, ανάλυση συστάδων, κανόνες συσχέτισης, στατιστική ανάλυση κλπ και το βασικό του πλεονέκτημα είναι το φιλικό του περιβάλλον.
- Alteryx Project Edition: Το λογισμικό Alteryx Designer περιλαμβάνει εργαλεία ολοκλήρωσης δεδομένων, καθώς και προγνωστικής και χωρικής ανάλυσης.
- CMSR Data Miner : λογισμικό το οποίο είναι δωρεάν για ακαδημαϊκή χρήση. Παρέχει ένα μεγάλο σύνολο μεθόδων εξόρυξης δεδομένων, όπως νευρωνικά δίκτυα, δέντρα αποφάσεων και μέθοδοι ανάλυσης συστάδων.

Κεφάλαιο 3 : Γραφική διεπαφή χρήστη του WEKA

Εισαγωγή

Το GUI του WEKA παρέχει πέντε επιλογές:

- Explorer
- Experimenter
- Knowledge flow
- Workbench
- Simple CLI.

Στη συνέχεια θα μιλήσουμε για καθένα από τα παραπάνω:

3.1 Simple CLI

```
java <classname> <args>
          Lists the capabilities of the specified class.
          If the class is a weka.core.OptionHandler then
          trailing options after the classname will be
          set as well.

kill
          Kills the running job, if any.

script <script_file>
          Executes commands from a script file.

set [name=value]
          Sets a variable.
          If no key=value pair is given all current variables are l

unset name
          Removes a variable.


Notes:
- Variables can be used anywhere using '${<name>}' with '<name>'
  being the name of the variable.
- Environment variables can be used with '${env.<name>}',
  e.g., '${env.PATH}' to retrieve the PATH variable.
```

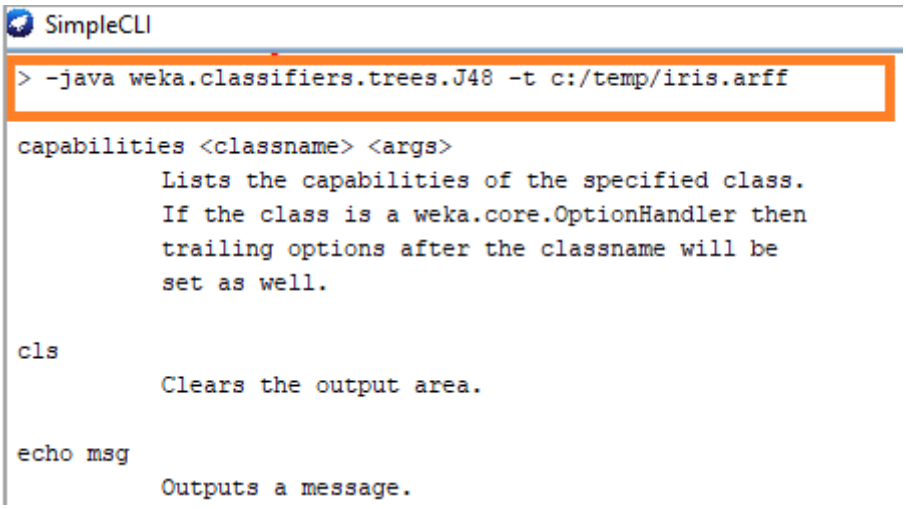
```
help
```

Εικόνα 5 : Simple CLI

Το Simple CLI είναι το Weka Shell με γραμμή εντολών και έξοδο. Το Simple CLI προσφέρει πρόσβαση σε όλες τις κλάσεις όπως ταξινομητές, συμπλέγματα, φίλτρα κ.λπ.

Μερικές από τις απλές εντολές CLI είναι:

- Διακοπή: Για να σταματήσετε το τρέχον νήμα
- Έξοδος: Έξοδος από το CLI
- Help[<command>] : Εξάγει τη βοήθεια για την καθορισμένη εντολή
- -java weka.classifiers.trees.J48 -t c:/temp/iris.arff : Για να καλέσετε μια κλάση WEKA, προσθέστε το πρόθεμά της με Java.



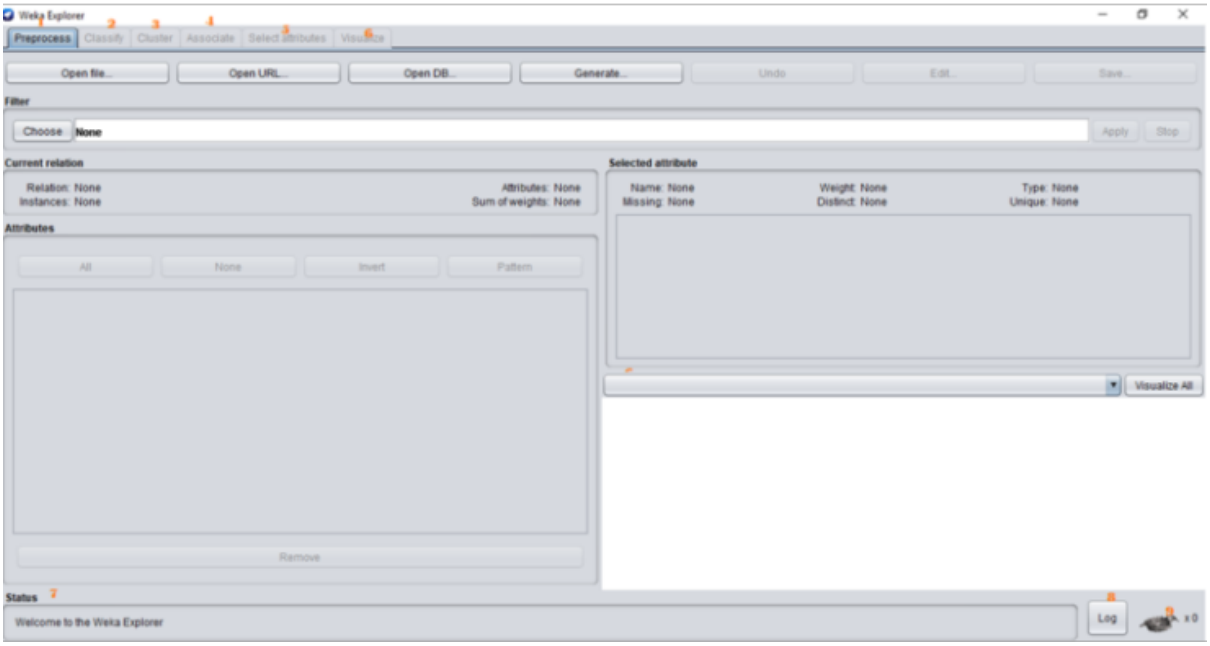
Εικόνα 6: Simple CLI

3.2 Explorer

Τα παράθυρα του WEKA Explorer εμφανίζουν διαφορετικές καρτέλες που ξεκινούν με την προεπεξεργασία. Αρχικά, η καρτέλα προεπεξεργασίας είναι ενεργή, καθώς πρώτα το σύνολο δεδομένων προεπεξεργάζεται προτού εφαρμοστούν αλγόριθμοι σε αυτό και εξερευνηθεί το σύνολο δεδομένων.

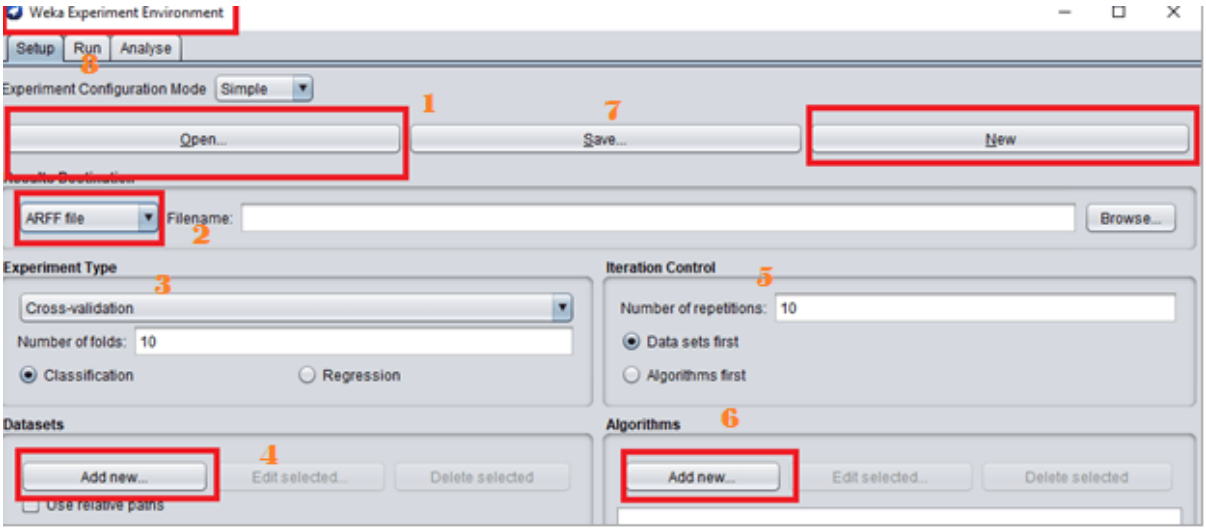
Οι καρτέλες είναι οι εξής:

- Προεπεξεργασία: Επιλέξτε και τροποποιήστε τα φορτωμένα δεδομένα.
- Ταξινόμηση: Εφαρμόστε αλγόριθμους εκπαίδευσης και δοκιμής στα δεδομένα που θα ταξινομήσουν και θα υποχωρήσουν τα δεδομένα.
- Συστάδα: Σχηματίστε συμπλέγματα από τα δεδομένα.
- Συνεργάτης: Εξορύξτε τον κανόνα συσχέτισης για τα δεδομένα.
- Επιλογή χαρακτηριστικών: Εφαρμόζονται μέτρα επιλογής χαρακτηριστικών.
- Οπτικοποίηση: εμφανίζεται 2D αναπαράσταση δεδομένων.
- Γραμμή κατάστασης: Το πιο κάτω τμήμα του παραθύρου εμφανίζει τη γραμμή κατάστασης.
- Κουμπί καταγραφής: Αποθηκεύει ένα αρχείο καταγραφής όλων των ενεργειών στο Weka με τη χρονική σήμανση..
- Εικονίδιο WEKA Bird: Το παρόν στην κάτω δεξιά γωνία δείχνει το πουλί WEKA με αντιπροσωπεύει τον αριθμό των διεργασιών που εκτελούνται ταυτόχρονα.



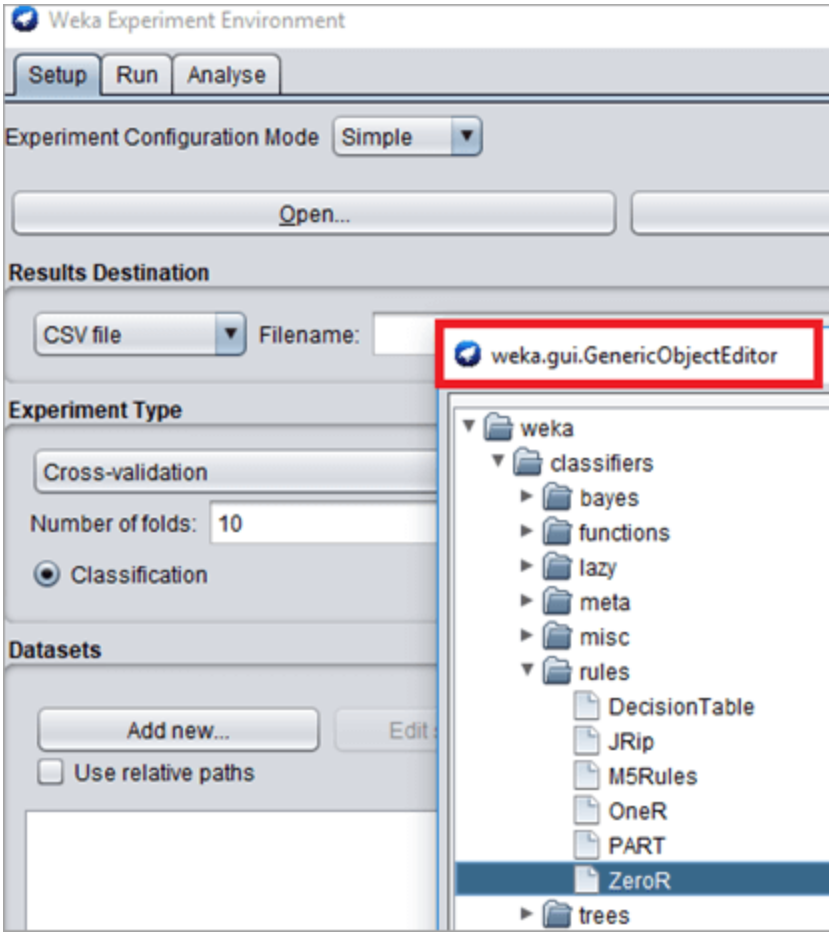
Εικόνα 7 :Explorer

3.3 Experimenter



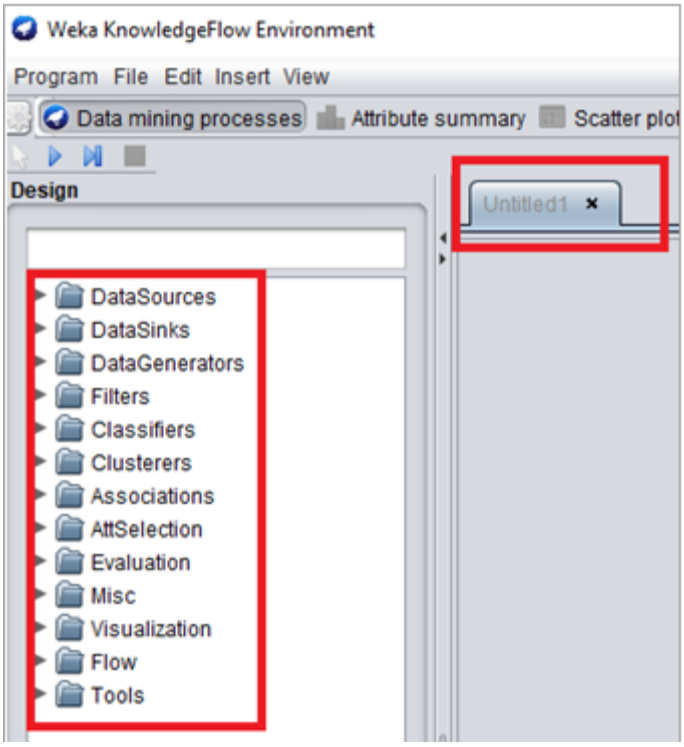
Εικόνα 8: Experimenter

- Τα κουμπιά "Άνοιγμα" και "Νέο" θα ανοίξουν ένα νέο παράθυρο πειράματος που μπορούν να κάνουν οι χρήστες.
- Αποτελέσματα: Ορίστε το αρχείο προορισμού αποτελέσματος από αρχεία ARFF, JDFC και CSV.
- Τύπος πειράματος: Ο χρήστης μπορεί να επιλέξει μεταξύ επικύρωσης και διαίρεσης ποσοστού εκπαίδευσης/δοκιμής.
- Σύνολα δεδομένων: Ο χρήστης μπορεί να περιηγηθεί και να επιλέξει σύνολα δεδομένων από εδώ.
- Επανάληψη: Ο προεπιλεγμένος αριθμός επανάληψης έχει οριστεί στο 10.
- Αλγόριθμοι: Νέοι αλγόριθμοι προστίθενται από το "Νέο Κουμπί". Ο χρήστης μπορεί να επιλέξει έναν ταξινομητή.



Εικόνα 9: Experimenter

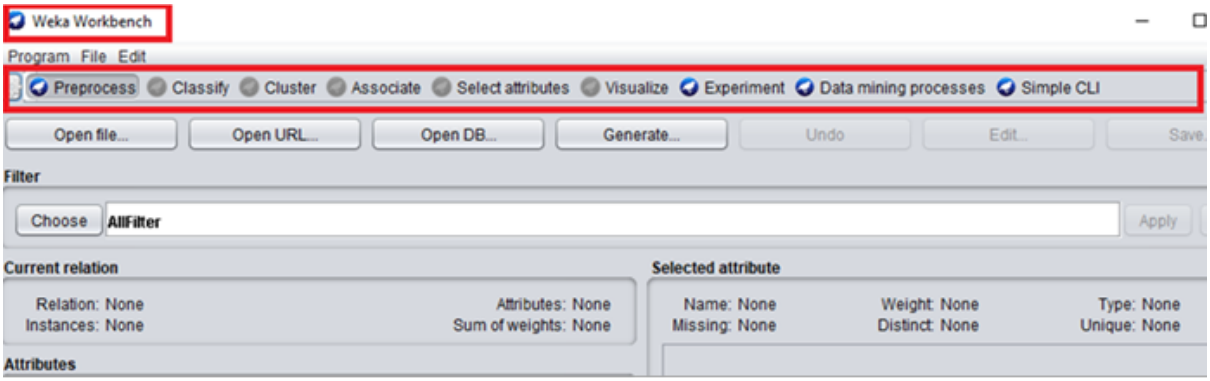
3.4 Knowledge Flow



Εικόνα 10:Knowledge Flow

Η ροή γνώσης δείχνει μια γραφική αναπαράσταση των αλγορίθμων WEKA. Ο χρήστης μπορεί να επιλέξει τα στοιχεία και να δημιουργήσει μια ροή εργασίας για να αναλύσει τα σύνολα δεδομένων.

3.5 Workbench



Εικόνα 11 :Workbench

3.6 Χαρακτηριστικά του WEKA Explorer

Ένα σύνολο δεδομένων αποτελείται από στοιχεία. Τα σύνολα δεδομένων περιγράφονται με χαρακτηριστικά. Το σύνολο δεδομένων περιέχει πλειάδες δεδομένων σε μια βάση δεδομένων. Ένα σύνολο δεδομένων έχει χαρακτηριστικά που μπορεί να είναι:

- Ονομαστικά χαρακτηριστικά
- Δυαδικά χαρακτηριστικά
- Τακτικές ιδιότητες
- Αριθμητικά χαρακτηριστικά

➢ ARFF Data format

Τα αρχεία ARFF λαμβάνουν χαρακτηριστικά:

- Ονομαστικό
- Αριθμητικό
- Συμβολοσειρά
- Ημερομηνία
- Σχεσιακά δεδομένα

Παράδειγμα αρχείου ARFF

```
@relation weather
@attribute outlook {sunny, overcast, rainy}:
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}       //class attribute: The class attribute represents the output.

@data
sunny, FALSE,85,85,no
sunny, TRUE,80,90,no
overcast, FALSE,83,86,yes
rainy, FALSE,70,96,yes
rainy, FALSE,68,80,yes
```

Εικόνα 12: Το αρχείο ARFF

- Ταξινομητές

Για την πρόβλεψη των δεδομένων εξόδου, το WEKA περιέχει ταξινομητές. Οι αλγόριθμοι ταξινόμησης που είναι διαθέσιμοι για εκμάθηση είναι τα δέντρα αποφάσεων, οι μηχανές διανυσμάτων υποστήριξης, οι ταξινομητές που βασίζονται σε στιγμιότυπα, η παλινδρόμηση και τα δίκτυα Bayes.

- Ομαδοποίηση

Το WEKA χρησιμοποιεί την καρτέλα Cluster για να προβλέψει τις ομοιότητες στο σύνολο δεδομένων.

- Οπτικοποίηση

Το WEKA υποστηρίζει τη δισδιάστατη αναπαράσταση δεδομένων, τις τρισδιάστατες απεικονίσεις με περιστροφή και την 1Δ αναπαράσταση ενός μόνο χαρακτηριστικού.

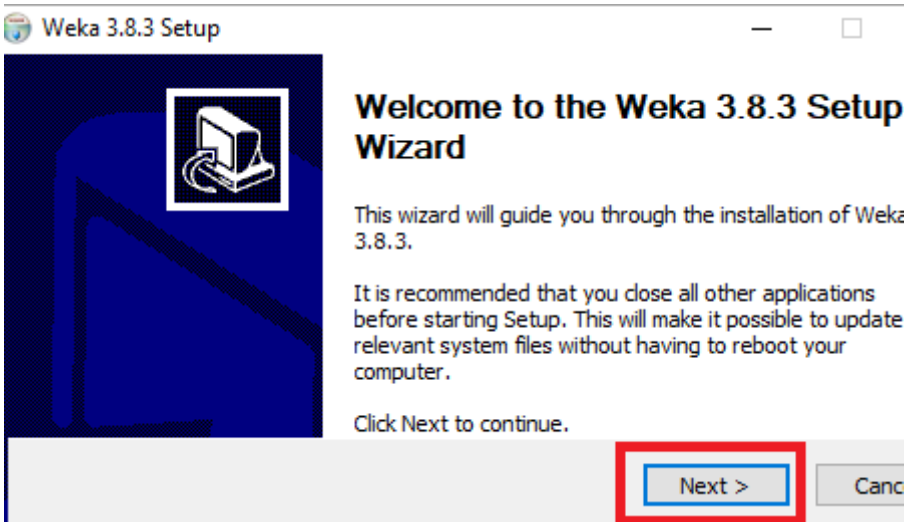
Κεφάλαιο 4: Βήματα εγκατάστασης του Weka

4.1 Βήματα εγκατάστασης Weka

Βήμα 1ο : Αρχικά κατεβάζουμε το λογισμικό και έπειτα ελέγχουμε τη διαμόρφωση του συστήματος υπολογιστή και κάνουμε λήψη της σταθερής έκδοσης του WEKA από αυτήν τη σελίδα.

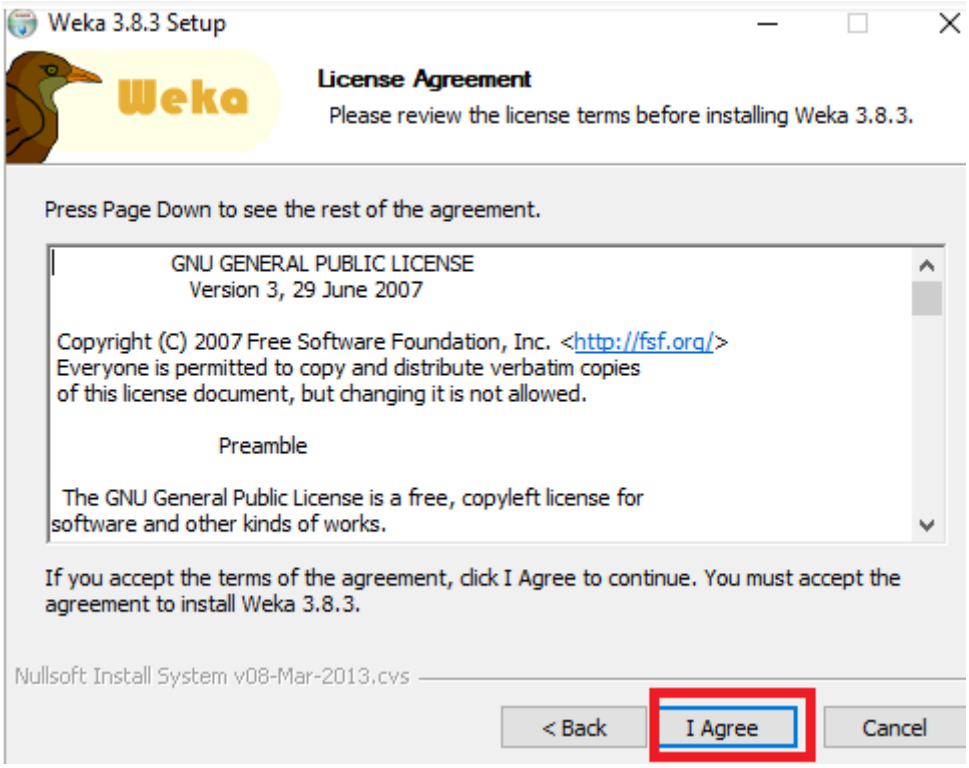
Εικόνα 13: 1ο Βήμα εγκατάστασης του WEKA

Βήμα 2ο : Μετά την επιτυχή λήψη, ανοίγουμε τη θέση του αρχείου και κάνουμε διπλό κλικ. Θα εμφανιστεί ο οδηγός Step Up. Κάνουμε κλικ στο Επόμενο.



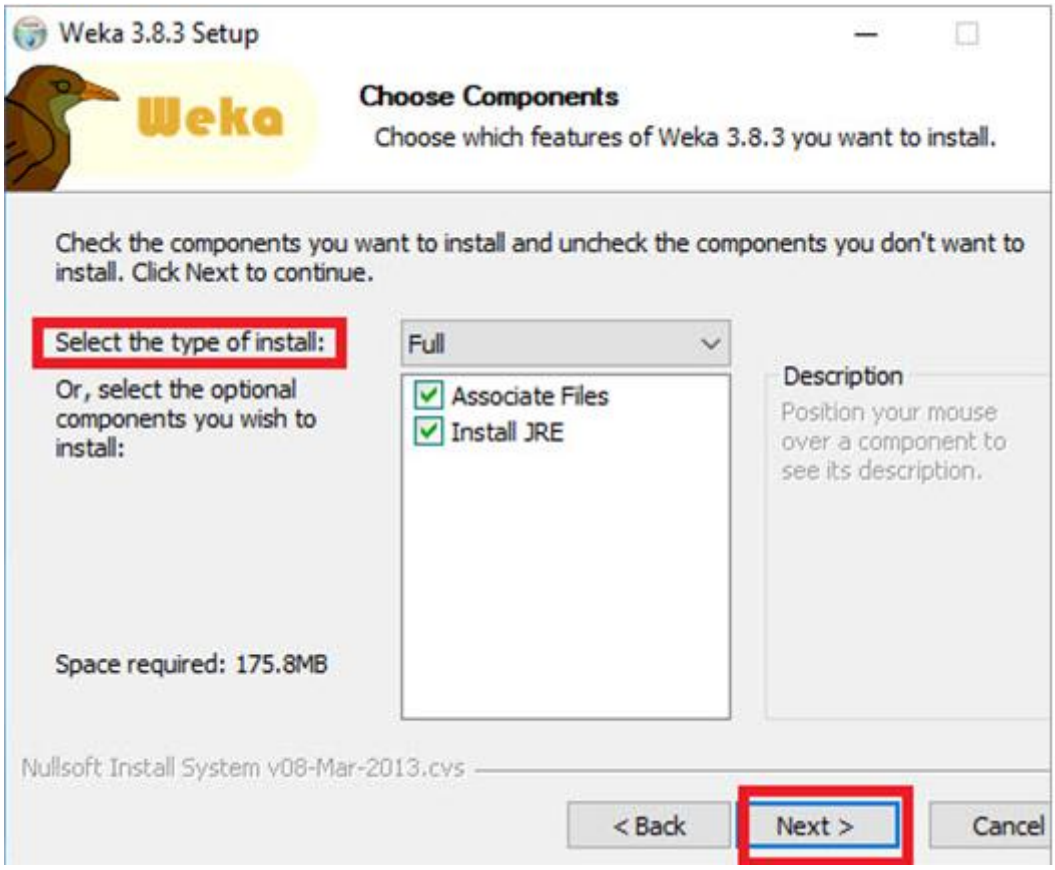
Εικόνα 14 : 2ο Βήμα εγκατάστασης του WEKA

Βήμα 3ο :Οι όροι της Άδειας Χρήσης θα ανοίξουν. Κάνουμε κλικ στο «Συμφωνώ».



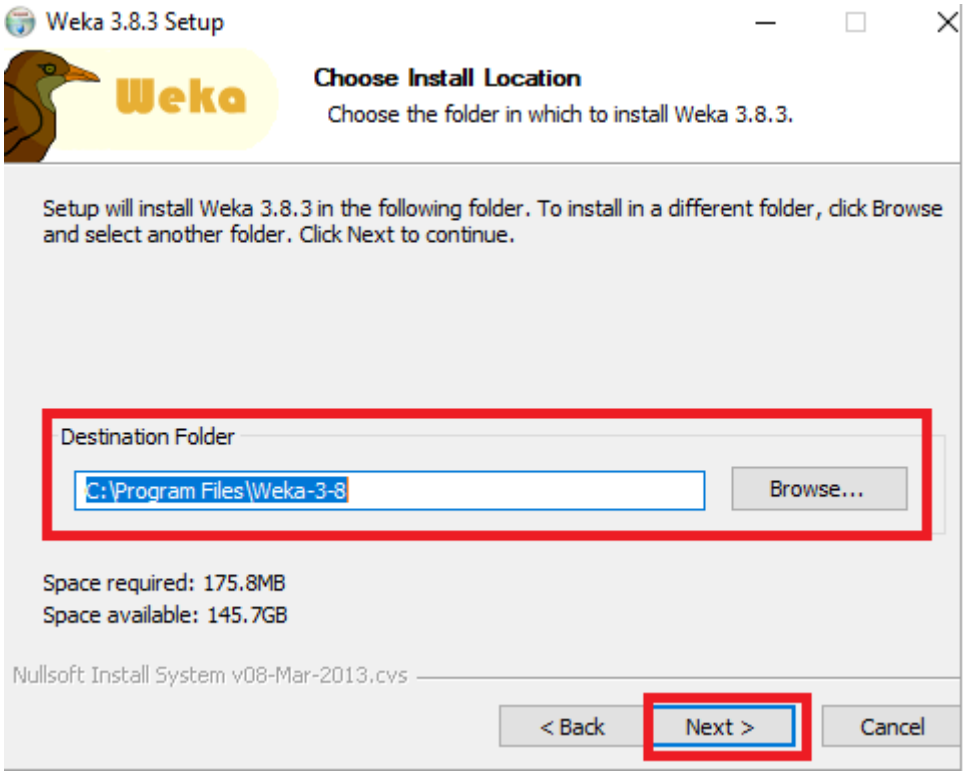
Εικόνα 15: 3ο Βήμα εγκατάστασης του WEKA

Βήμα 4ο : Σύμφωνα με τις απαιτήσεις σας, επιλέγουμε τα εξαρτήματα που θα εγκατασταθούν. Κάνουμε κλικ στο Επόμενο.



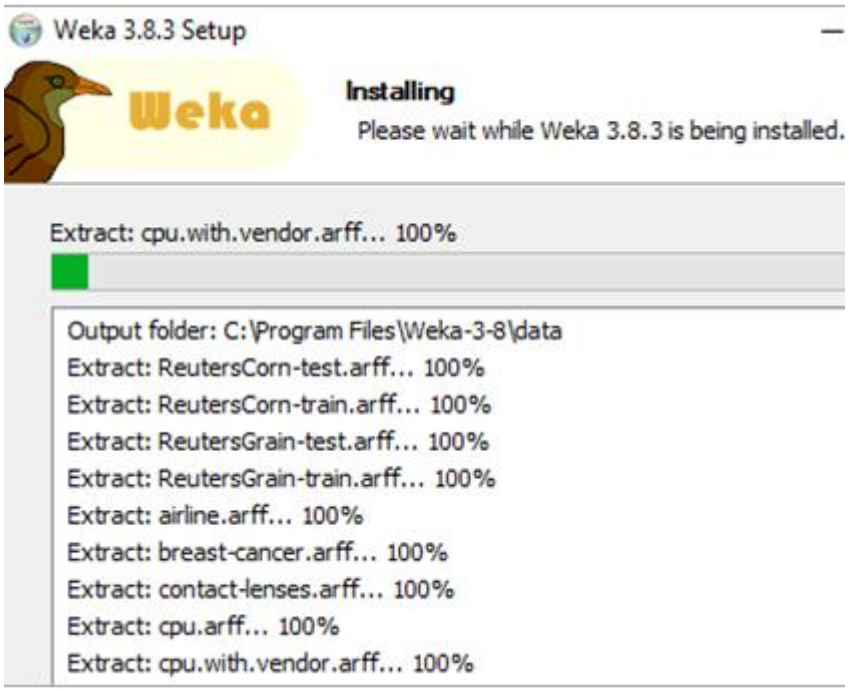
Εικόνα 16: 4ο Βήμα εγκατάστασης του WEKA

Βήμα 5ο : Επιλέγουμε το φάκελο προορισμού και κάνουμε κλικ στο Επόμενο.



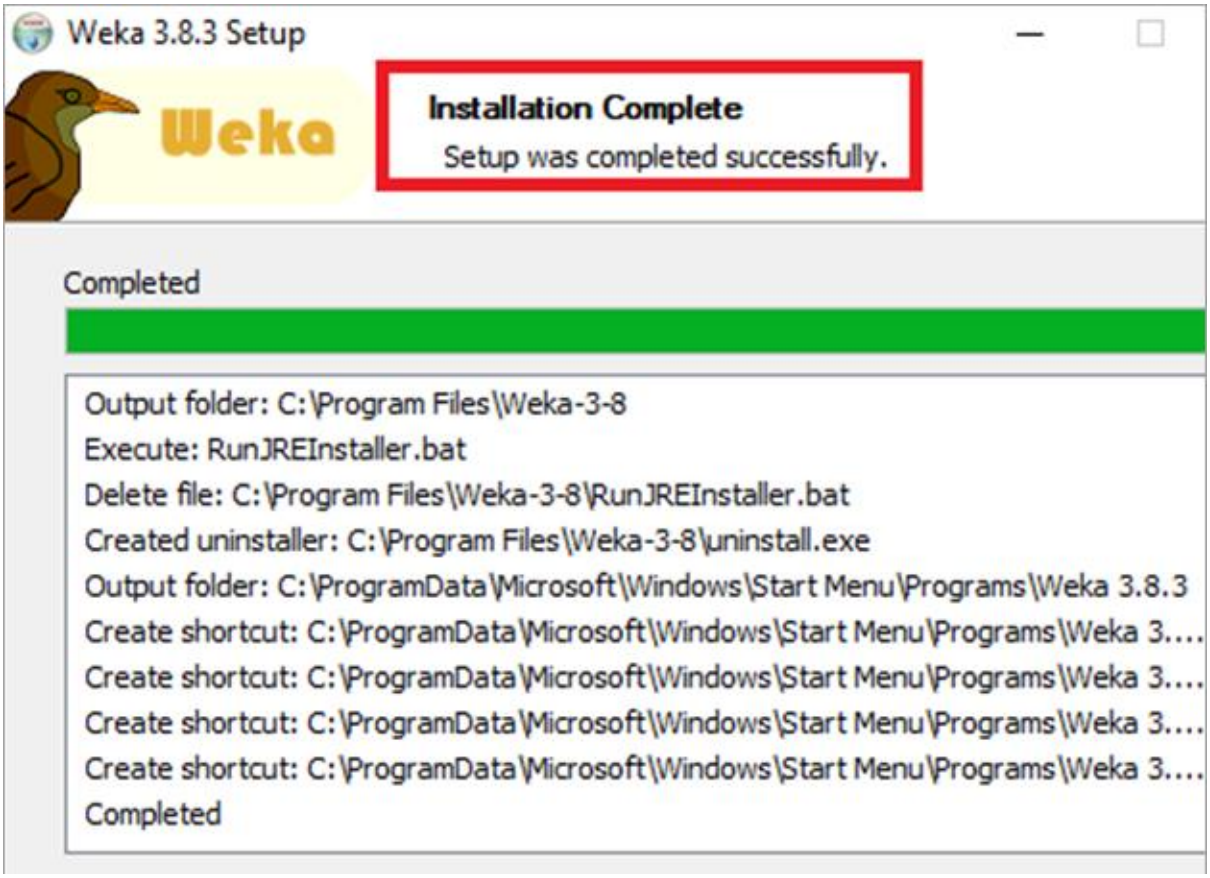
Εικόνα 17: 5ο Βήμα εγκατάστασης του WEKA

Γίνεται σταδιακά η εγκατάσταση



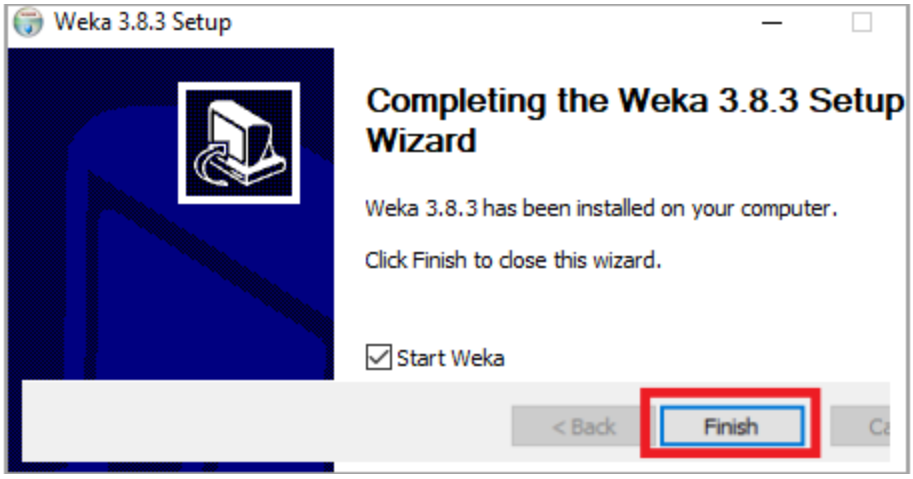
Εικόνα 18: Βήμα εγκατάστασης του WEKA

Βήμα 6ο : Αφού ολοκληρωθεί η εγκατάσταση, θα εμφανιστεί το ακόλουθο παράθυρο. Κάνουμε κλικ στο Επόμενο.



Εικόνα 19: 6ο Βήμα εγκατάστασης του WEKA

Βήμα 7ο : Επιλέγουμε το πλαίσιο ελέγχου Έναρξη Weka. Κάνουμε κλικ στο Τέλος.



Εικόνα 20: 7ο Βήμα εγκατάστασης του WEKA

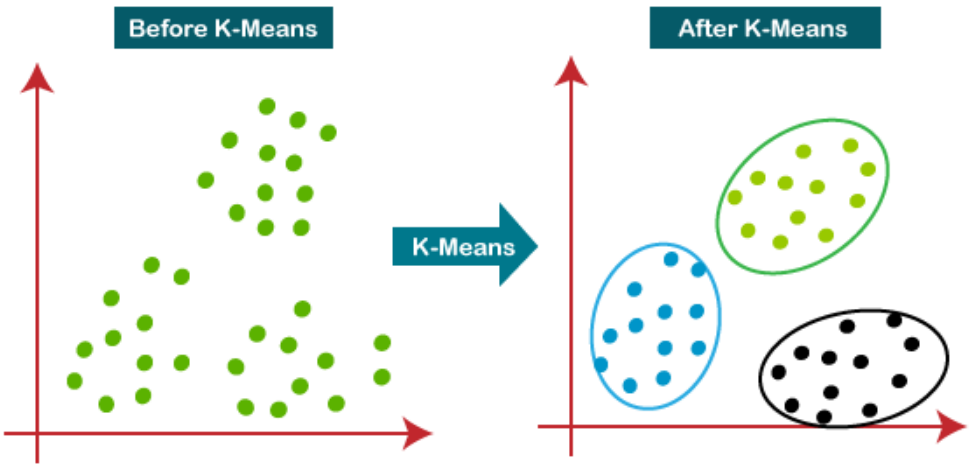
Βήμα 8ο : Ανοίγει το παράθυρο WEKA Tool and Explorer

Εικόνα 21: 8ο Βήμα εγκατάστασης του WEKA

Κεφάλαιο 5: Αλγόριθμοι μηχανικής μάθησης

5.1 K-means

Η χρήση του K-Means Clustering ως αλγόριθμος χωρίς επίβλεψη στη μηχανική μάθηση ή στην επιστήμη δεδομένων μπορεί να λύσει προβλήματα σχετικά με την ομαδοποίηση. Ένα σημαντικό πλεονέκτημα είναι η δυνατότητα ομαδοποίησης δεδομένων σε διακριτές κατηγορίες χωρίς προηγούμενη εκπαίδευση. Αυτός ο αλγόριθμος επικεντρώνεται στη συσχέτιση κάθε συστάδας με ένα κέντρο, στοχεύοντας τελικά στην ελαχιστοποίηση των συνολικών αποστάσεων μεταξύ των σημείων δεδομένων και των αντίστοιχων συστάδων τους. Αυτή η τεχνική βοηθά στην ανακάλυψη ομαδικών ταξινομήσεων μέσα σε ένα σύνολο δεδομένων. Σε αυτήν την επαναληπτική διαδικασία, ο αλγόριθμος ομαδοποίησης k-means λαμβάνει ως είσοδο ένα σύνολο δεδομένων, το διαιρεί σε συστάδες και συνεχίζει μέχρι να μην βρεθούν οι καλύτερες συστάδες. Ο αλγόριθμος απαιτεί μια προκαθορισμένη τιμή k. Εδώ το K ορίζει τον αριθμό των προκαθορισμένων συστάδων που πρέπει να δημιουργηθούν στη διαδικασία, καθώς εάν K=2, θα υπάρχουν δύο συστάδες, και για K=3, θα υπάρχουν τρία συμπλέγματα κ.ο.κ.Τα κύρια καθήκοντά του περιλαμβάνουν τον προσδιορισμό της βέλτιστης τιμής για τα κεντροειδή σε μια επαναληπτική διαδικασία και τη συσχέτιση κάθε σημείου δεδομένων με το πλησιέστερο k-κέντρο δημιουργώντας ένα συμπλεγμα. Τα συμπλέγματα έχουν απόσταση μεταξύ τους και έχουν κάποια κοινά σημεία δεδομένων με άλλα.



Εικόνα 22: K-means

5.2 Πώς λειτουργεί ο αλγόριθμος K-Means;

Η λειτουργία του αλγόριθμου K-Means εξηγείται από τα εξής στάδια:

1ο Στάδιο: Επιλογή αριθμού K για το πλήθος των συμπλεγμάτων.

2ο Στάδιο: Επιλογή K τυχαίων σημείων ή κεντροειδών. (Μπορεί να διαφέρει από το σύνολο δεδομένων εισόδου).

3ο Στάδιο: Αντιστοίχιση του κάθε σημείου δεδομένων με το πλησιέστερο κεντροειδή, ώστε να σχηματιστούν τα προκαθορισμένα συμπλέγματα K.

4ο Στάδιο: Υπολογισμός διακύμανσης και τοποθέτηση ενός νέου κεντροειδούς για κάθε σύμπλεγμα.

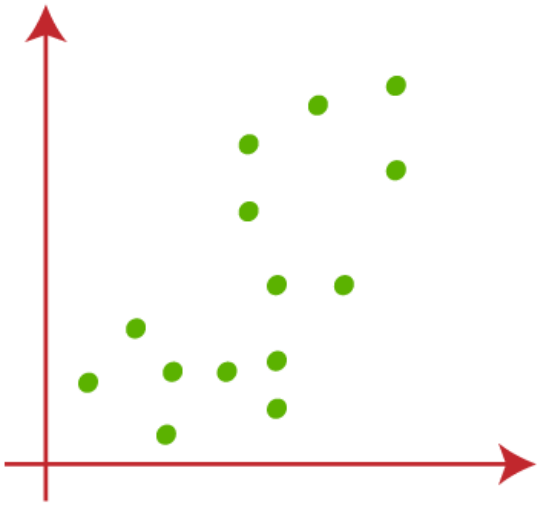
5ο Στάδιο: Επανάληψη 3ου σταδίου, δηλαδή αντιστοίχιση κάθε σημείου δεδομένων στο νέο πλησιέστερο κεντροειδούς κάθε συμπλέγματος.

6ο Στάδιο:Αν γίνει κάποια αντιστοίχιση, τότε γίνεται το στάδιο 4 αλλιώς FINISH.

7ο Στάδιο : Το μοντέλο είναι έτοιμο.

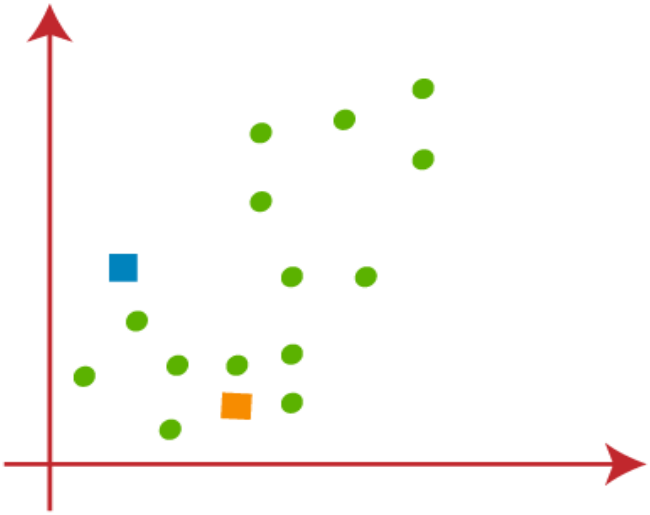
Ας κατανοήσουμε τα παραπάνω:

Έστω ότι έχουμε δύο μεταβλητές M1 και M2. Το διάγραμμα διασποράς άξονα xy αυτών των δύο μεταβλητών δίνεται παρακάτω:



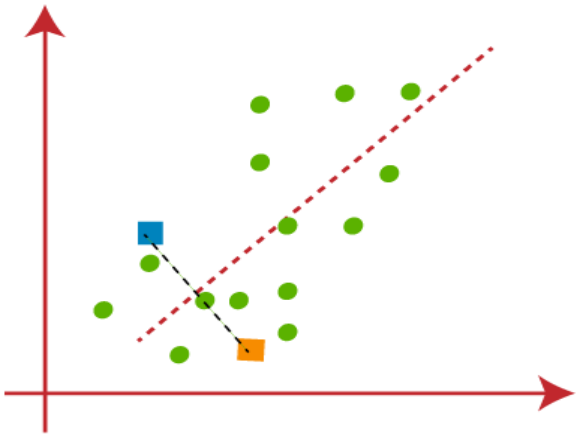
Εικόνα 23: K-means Παράδειγμα

- Έστω K=2, για το σύνολο δεδομένων τα οποία τοποθετούμε σε διαφορετικά συμπλέγματα. Δηλαδή ομαδοποιούμε αυτά τα σύνολα δεδομένων σε δύο διαφορετικά συμπλέγματα.
- Επιλέγουμε τυχαία σημεία k ή κεντρα έτσι ώστε να σχηματίσουμε το σύμπλεγμα. Αυτά μπορεί να είναι σημεία από το σύνολο δεδομένων ή οποιοδήποτε άλλο σημείο. Οπότε, επιλέγουμε τα παρακάτω δύο σημεία ως k σημεία, τα οποία δεν αποτελούν μέρος του συνόλου δεδομένων μας. Δηλαδή:



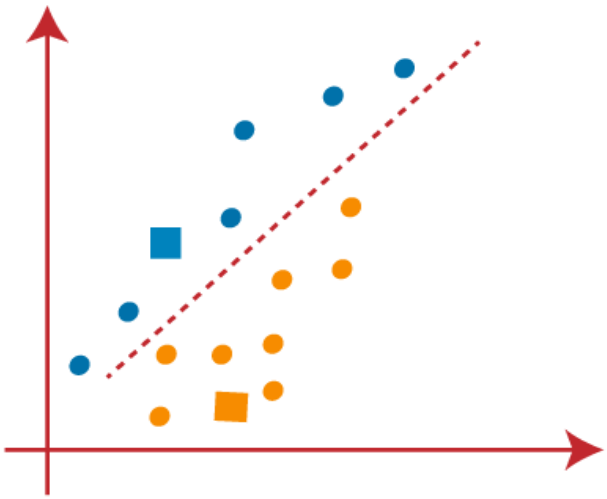
Εικόνα 24 : K-means Παράδειγμα

Αντιστοιχίζουμε κάθε σημείο δεδομένων της γραφικής παράστασης διασποράς στο πλησιέστερο σημείο Κ ή κεντρο, υπολογίζουμε την απόστασή τους και σχεδιάζουμε μια διάμεσο μεταξύ των δύο κεντρων. Δηλαδή:



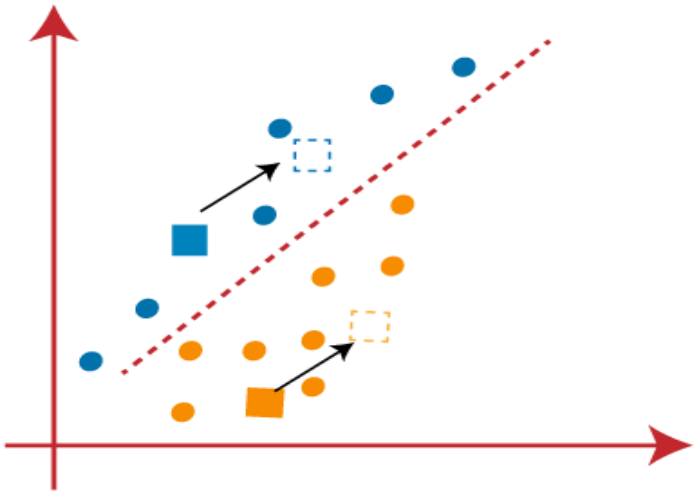
Εικόνα 25: K-means Παράδειγμα

Βλέπουμε ότι τα σημεία αριστερά από τη γραμμή είναι κοντά στο Κ1 ή το μπλε κέντρο ενώ από δεξιά της είναι κοντά στο κίτρινο κέντρο. Τα χρωματίζουμε μπλε και κίτρινα αντίστοιχα για καλύτερη απεικόνιση .



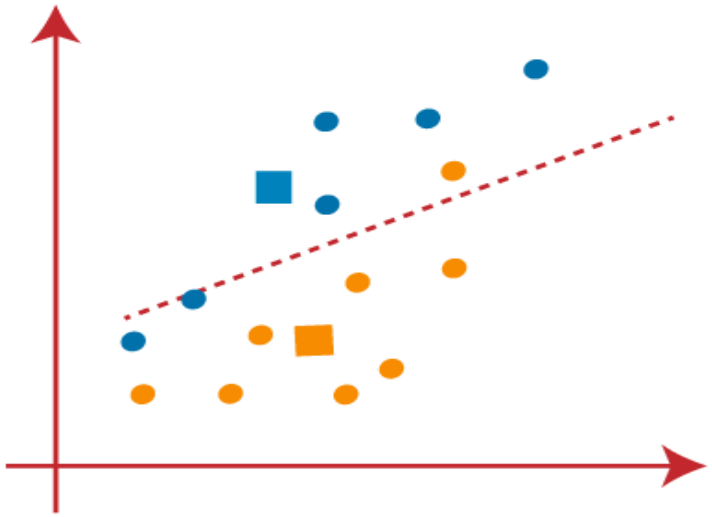
Εικόνα 26: K-means Παράδειγμα

Πρέπει να βρούμε το πλησιέστερο σύμπλεγμα, οπότε επαναλαμβάνουμε τη διαδικασία και επιλέγουμε ένα νέο κέντρο . Για να επιλέξουμε τα νέα κεντρα, υπολογίζουμε το κέντρο βάρους αυτών των κεντροειδών και βρίσκουμε νέα:



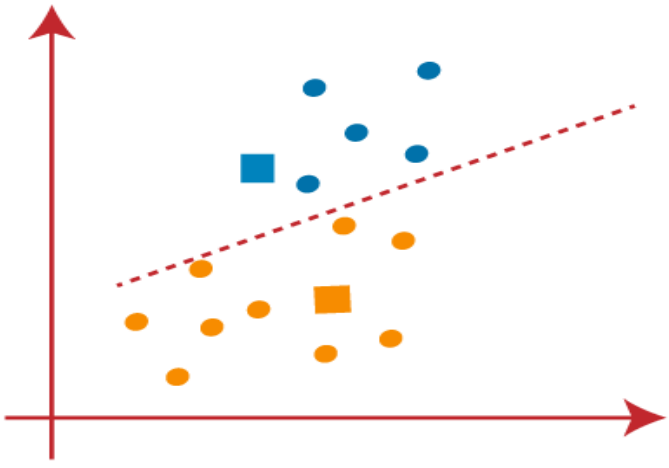
Εικόνα 27: K-means Παράδειγμα

Επαναλαμβάνουμε την ίδια διαδικασία εύρεσης μέσης γραμμής και η διάμεσος θα είναι:



Εικόνα 28: K-means Παράδειγμα

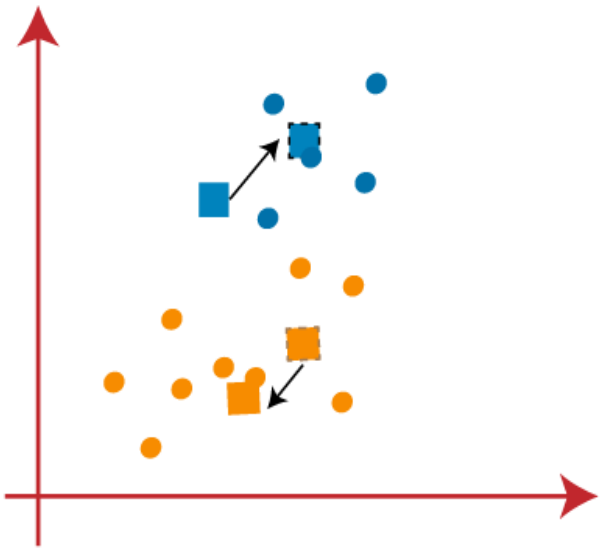
Βλέπουμε ότι δύο μπλε σημεία είναι πάνω στη γραμμή και ένα κίτρινο σημείο βρίσκεται στην αριστερή πλευρά της γραμμής οπότε, θα εκχωρηθούν σε νέα κέντρα.Δηλαδή:



Εικόνα 29: K-means Παράδειγμα

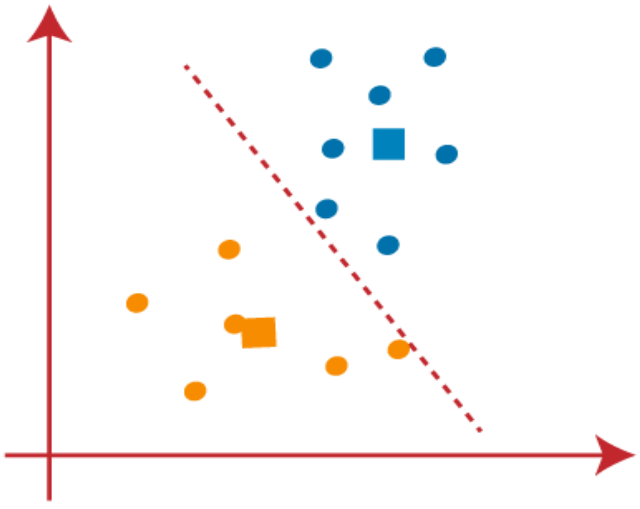
Εύρεση νέων κεντρων ή σημείων K.(4ο Στάδιο)

- Βρίσκουμε το κέντρο βάρους των κεντρων, τα νέα κεντρα θα είναι:



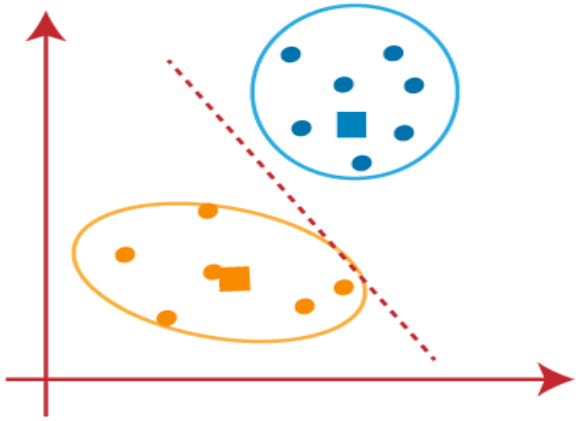
Εικόνα 30: K-means Παράδειγμα

Φτιάχνουμε τη διάμεση γραμμή και βάζουμε εκ νέου τα σημεία δεδομένων:



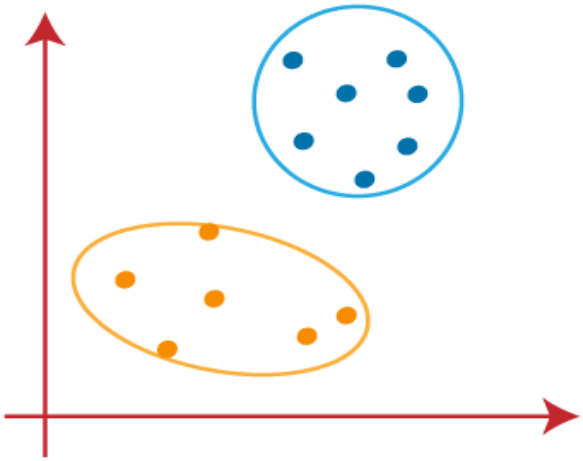
Εικόνα 31: K-means Παράδειγμα

Δεν υπάρχουν ανόμοια σημεία δεδομένων σε καμία πλευρά, άρα το μοντέλο μας έχει διαμορφωθεί. Δηλαδή:



Εικόνα 32: K-means Παράδειγμα

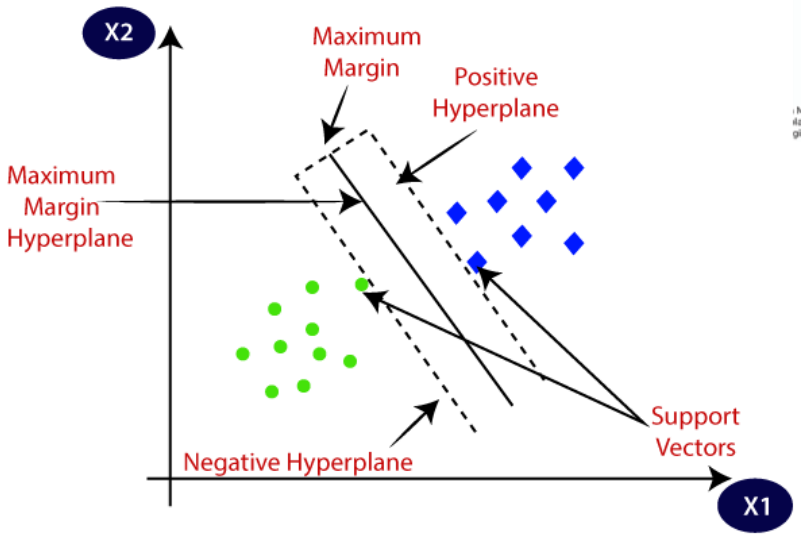
Οπότε τα δύο τελικά συμπλέγματα, θα είναι:



Εικόνα 33: K-means Παράδειγμα

5.3 Support Vector Machine Algorithm

Η Μηχανική Εκμάθηση χρησιμοποιεί κυρίως Support Vector Machine (SVM), έναν από τους δημοφιλέστερους αλγόριθμους για εποπτευόμενη μάθηση. Το SVM χρησιμοποιείται συνήθως για προβλήματα παλινδρόμησης αλλά κυρίως σε προβλήματα ταξινόμησης στη Μηχανική Μάθηση. Ο αλγόριθμος Support Vector Machine στοχεύει στη δημιουργία ενός υπερεπίπεδου, το οποίο αποτελεί το καλύτερο όριο απόφασης που μπορεί να διαιρέσει τον n-διάστατο χώρο σε κλάσεις για εύκολη ταξινόμηση μελλοντικών σημείων δεδομένων. Για να δημιουργήσει αυτό το υπερεπίπεδο, το SVM επιλέγει διανύσματα υποστήριξης, τα οποία είναι ακραία σημεία. (ακραιες περιπτωσεις που ονομαζονται Διάνυσμα υποστήριξης)Επομένως, ο αλγόριθμος είναι γνωστός ως Support Vector Machine.Ο αλγόριθμος SVM χρησιμοποιείται για ανίχνευση προσώπου, ταξινόμηση εικόνων καθώς και για κατηγοριοποίηση κειμένου και άλλα. Στο συγκεκριμένο διάγραμμα, δύο διακριτές κατηγορίες ταξινομούνται χρησιμοποιώντας ένα υπερεπίπεδο ή ένα όριο απόφασης.



Εικόνα 34: SVM

Ο Γραμμικός SVM :Για δεδομένα που μπορούν να διαχωριστούν γραμμικά, πράγμα που σημαίνει ότι εάν μια ομάδα σημείων δεδομένων μπορεί να χωριστεί σε δύο κατηγορίες χρησιμοποιώντας μια ευθεία γραμμή, τότε τα δεδομένα ονομάζονται γραμμικά διαχωρισμένα και ο ταξινομητής χρησιμοποιείται ως Γραμμικός ταξινομητής SVM..

Μη γραμμικό SVM:Για περιπτώσεις όπου τα δεδομένα δεν μπορούν να ταξινομηθούν χρησιμοποιώντας μια ευθεία γραμμή,τα δεδομένα ονομάζονται μη γραμμικά και ο ταξινομητής χρησιμοποιείται ως Μη Γραμμικός ταξινομητής SVM

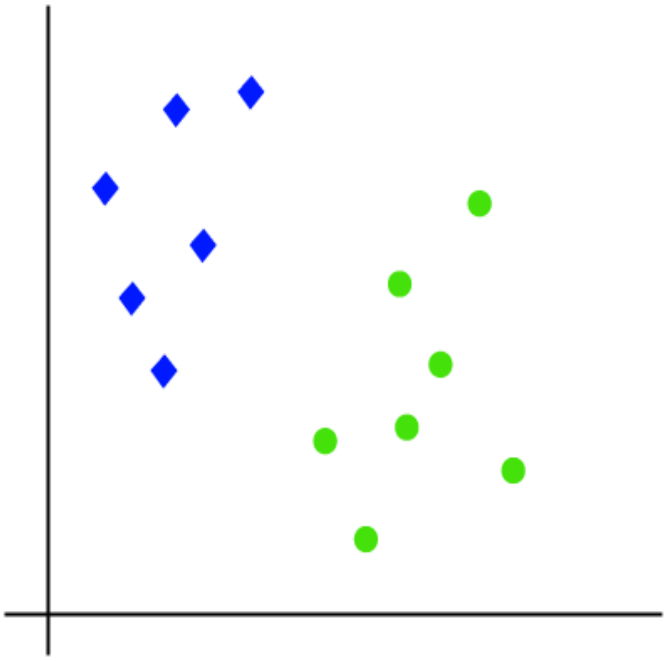
Υπερεπίπεδο: Στον n-διάστατο χώρο, υπάρχουν πολλαπλές γραμμές απόφασης που μπορούν να διαχωρίσουν τις κλάσεις. Ωστόσο, στόχος μας είναι να βρούμε το βέλτιστο όριο απόφασης για την ταξινόμηση των σημείων δεδομένων. Αυτό το βέλτιστο όριο αναφέρεται ως υπερεπίπεδο του SVM. Οι διαστάσεις του υπερεπίπεδου καθορίζονται από τα χαρακτηριστικά του συνόλου δεδομένων. Για παράδειγμα, εάν υπάρχουν δύο χαρακτηριστικά, το υπερεπίπεδο θα είναι μια ευθεία γραμμή,εάν υπάρχουν 3 θα είναι ένα δισδιάστατο επίπεδο. Στόχος μας είναι να δημιουργήσουμε ένα υπερεπίπεδο με μέγιστο περιθώριο, το οποίο είναι η μεγαλύτερη απόσταση μεταξύ των σημείων δεδομένων.

Διάνυσμα υποστήριξης: Τα πλησιέστερα σημεία δεδομένων / διανύσματα στο υπερεπίπεδο λέγονται διανύσματα υποστήριξης. Αυτά υποστηρίζουν το υπερεπίδο οπότε λέγονται διάνυσμα υποστήριξης.

5.4 Πώς λειτουργεί το SVM;

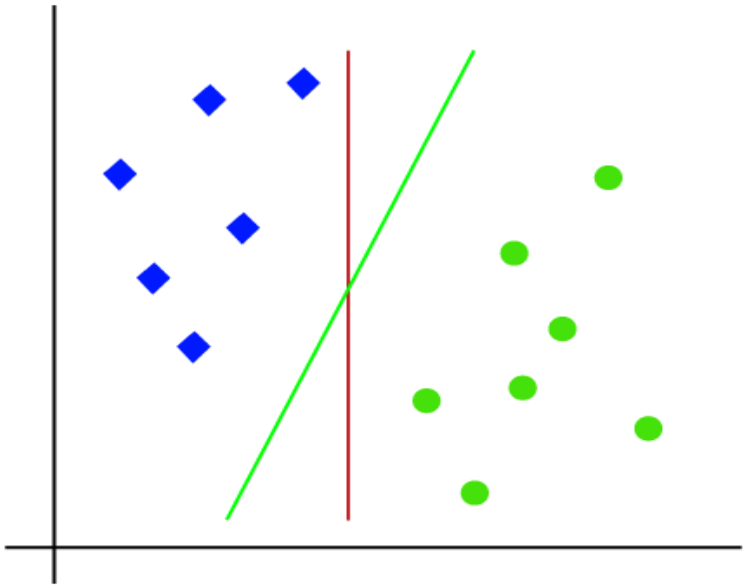
γραμμικο svm

Έστω ότι έχουμε ένα σύνολο δεδομένων το οποίο έχει δύο ετικέτες, πράσινη και μπλε και το σύνολο δεδομένων έχει δύο χαρακτηριστικά το x1 και το x2. Θέλουμε έναν ταξινομητή που να μπορεί να ταξινομήσει τις συντεταγμένες (x1, x2) σε πράσινο ή σε μπλε.Δηλαδή:



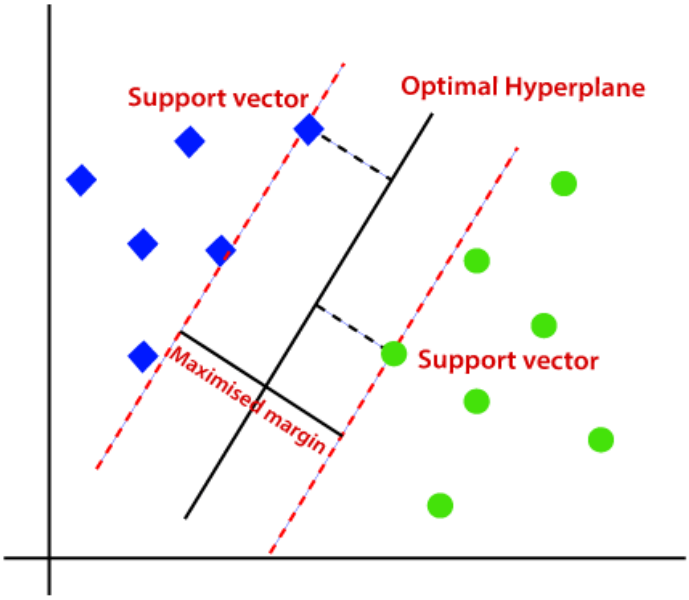
Εικόνα 35: Παράδειγμα SVM

Είναι δισδιαστατο , οπότε χρησιμοποιώντας μια ευθεία γραμμή, μπορούμε να διαχωρίσουμε τις δύο κατηγορίες. Υπάρχουν όμως πολλές γραμμές που διαχωρίζουν αυτές τις κλάσεις. Δηλαδή:



Εικόνα 36: Παράδειγμα SVM

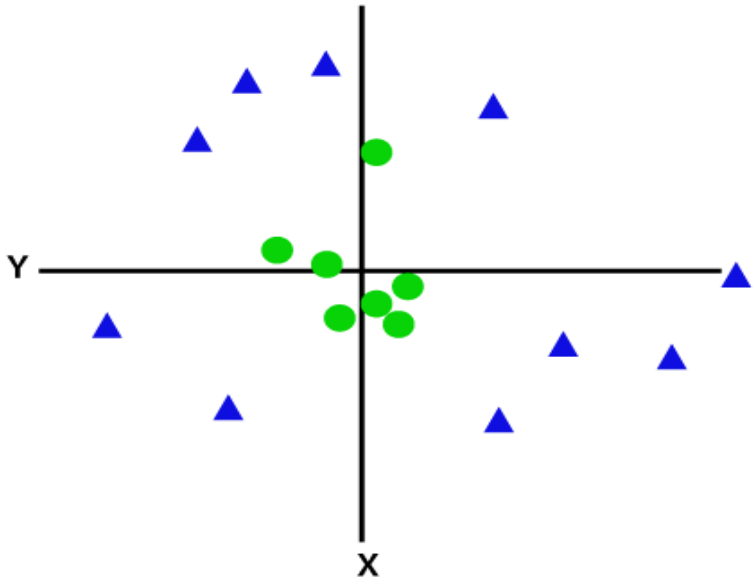
Συνεπώς, ο αλγόριθμος SVM βρίσκει την καλύτερη γραμμή ή όριο απόφασης (δηλ.υπερεπίπεδο) και το πλησιέστερο σημείο των γραμμών και από τις δύο κλάσεις (δηλ. διανύσματα υποστήριξης). Η απόσταση μεταξύ διανυσμάτων και υπερεπίπεδου λέγεται περιθώρι και στόχος του SVM είναι η μεγιστοποίηση του περιθωριου. Το υπερεπίπεδο με μέγιστο περιθώριο ονομάζεται βέλτιστο υπερεπίπεδο .



Εικόνα 37: Παράδειγμα SVM

Μη Γραμμικό SVM:

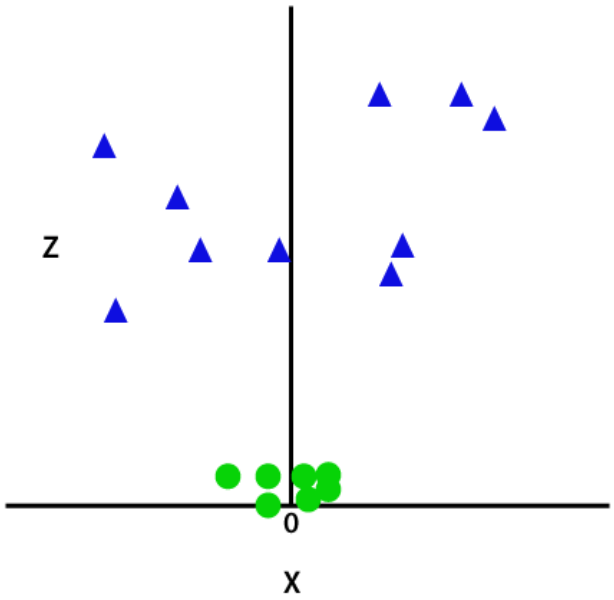
Αν τα δεδομένα είναι γραμμικά διατεταγμένα, τότε τα διαχωρίζουμε με μια ευθεία γραμμή, αλλά για τα μη γραμμικά δεδομένα, δεν μπορούμε. Δηλαδή:



Εικόνα 38: Παράδειγμα SVM

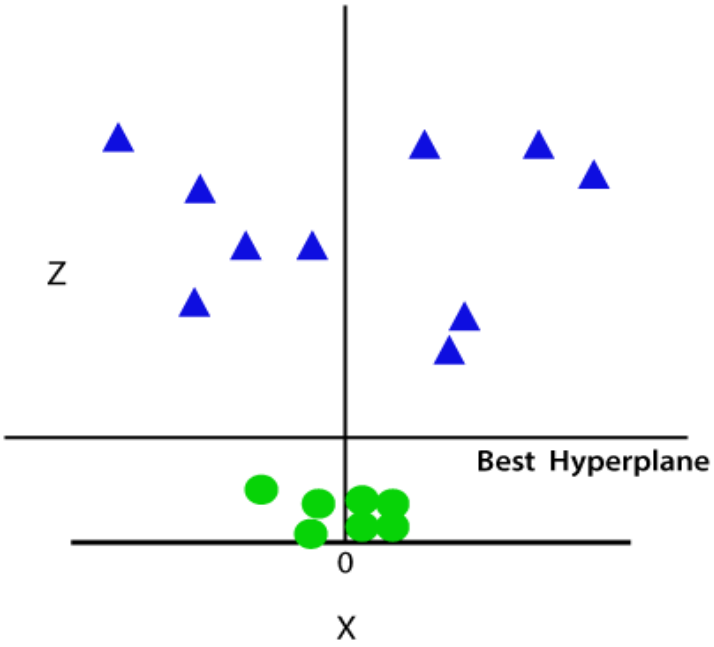
Για να τα διαχωρίσουμε, πρέπει να προσθέσουμε άλλη μια διάσταση. Για τα γραμμικά δεδομένα,είχαμε τις διαστάσεις x και y, άρα για μη γραμμικά δεδομένα, θα προσθέσουμε και τη διάσταση z. Υπολογίζεται:z=x 2 +y 2

Ο χώρος του δείγματος γίνεται:



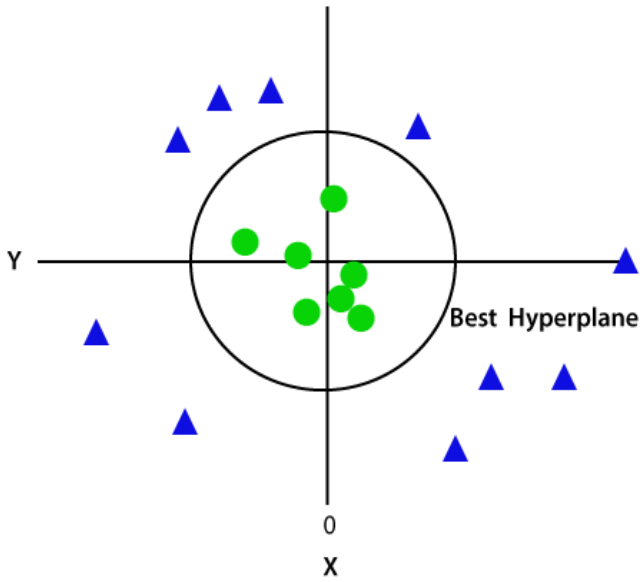
Εικόνα 39: Παράδειγμα SVM

Το SVM θα διαιρεί τα σύνολα δεδομένων σε κλάσεις με. Δηλαδή:



Εικόνα 40: Παράδειγμα SVM

Είναι τρισδιάστατο άρα, μοιάζει με ένα επίπεδο παράλληλο στον άξονα x. Το μετατρέπουμε σε δισδιαστατο με z=1 και θα γίνει:



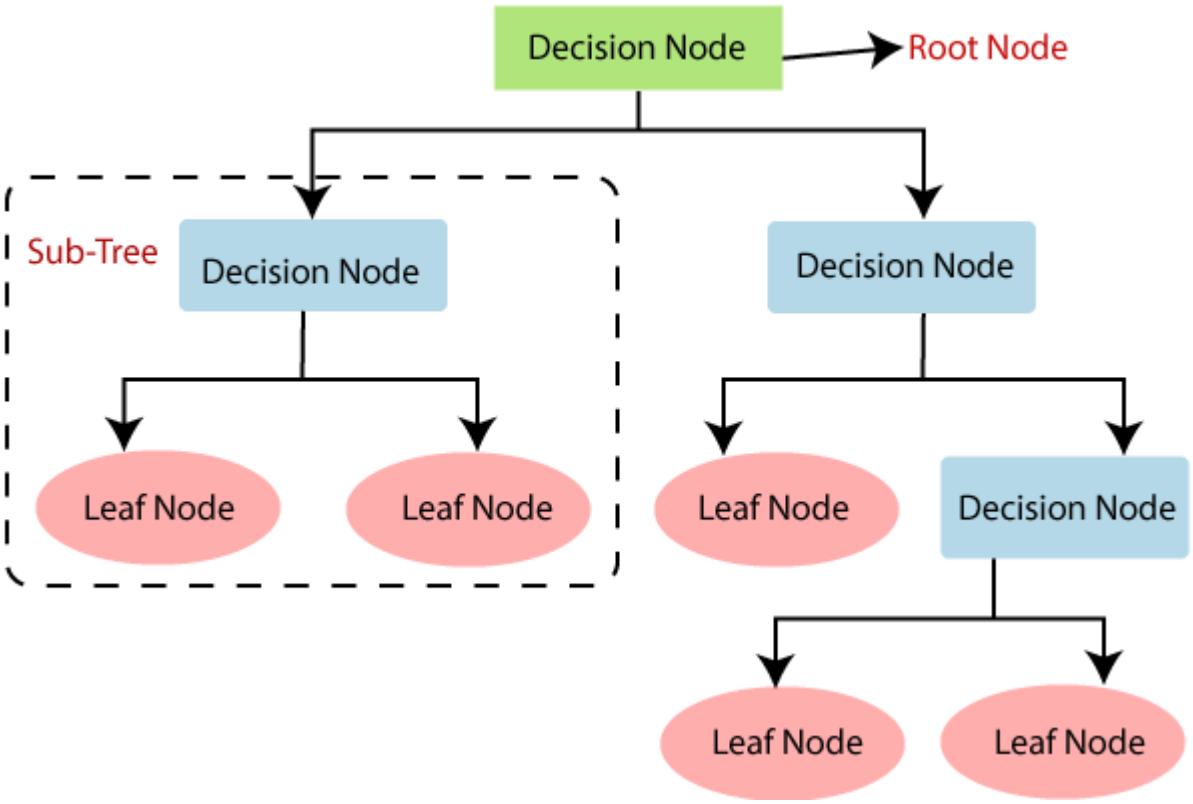
Εικόνα 41: Παράδειγμα SVM

Συνεπώς, σε περίπτωση μη γραμμικών δεδομένων παίρνουμε την περιφέρεια της ακτίνας 1.

5.5 Αλγόριθμος ταξινόμησης δένδρων αποφάσεων

. Το Decision Tree, μια τεχνική που χρησιμοποιείται για την επίλυση προβλημάτων ταξινόμησης, είναι ένας ταξινομητής δομημένου δέντρου που μπορεί επίσης να χειριστεί προβλήματα παλινδρόμησης. Είναι μια εποπτευόμενη μέθοδος εκμάθησης όπου τα χαρακτηριστικά ενός συνόλου δεδομένων αντιπροσωπεύονται από εσωτερικούς κόμβους, οι κανόνες απόφασης αντιπροσωπεύονται από κλάδους και το αποτέλεσμα αντιπροσωπεύεται από κάθε κόμβο φύλλου. Οι αποφάσεις που χρησιμοποιούν πολλαπλές επιλογές λαμβάνονται σε κόμβους αποφάσεων σε ένα δέντρο αποφάσεων και το προκύπτον αποτέλεσμα χωρίς πρόσθετες εναλλακτικές αντιπροσωπεύεται από τον κόμβο φύλλου. Οι δοκιμές ή οι αποφάσεις λαμβάνονται ανάλογα με τα ειδικά χαρακτηριστικά του συνόλου δεδομένων που δίνεται. Όλες οι πιθανές λύσεις σε ένα πρόβλημα ή μια απόφαση μπορούν να ληφθούν με μια γραφική αναπαράσταση, με βάση τις δεδομένες συνθήκες.Ονομάζεται δέντρο απόφασης αφού, όπως και με ένα δέντρο, ξεκινά με τον κόμβο ρίζας, ο οποίος επεκτείνεται σε περαιτέρω κλάδους, κατασκευάζοντας μια δομή που μοιάζει με δέντρο.Για τη δημιουργία δέντρου, χρησιμοποιείται ο αλγόριθμος CART(Classification and Regression Tree).Ένα δέντρο απόφασης θέτει μια ερώτηση και βάση της απάντησης : Yes/No (κατηγορικά δεδομένα αλλά μπορεί να περιέχει και αριθμητικά), χωρίζει το δέντρο σε υποδέντρα.

Γενική δομή δέντρου αποφάσεων:



Εικόνα 42: Παράδειγμα Decision Tree

5.6 Γιατί να χρησιμοποιήσετε τα Δέντρα απόφασης;

Τα δέντρα αποφάσεων είναι σχεδιασμένα να μοιάζουν με την ανθρώπινη σκέψη και επομένως είναι εύκολα κατανοητά.

Η κατανόηση της λογικής του δέντρου αποφάσεων γίνεται απλή αφού η εμφάνισή του είναι όμοια σε μια δομή δέντρου.

5.7 Ορολογίες Δένδρου Αποφάσεων

Root Node: Η ρίζα κόμβος είναι η αρχή του δέντρου αποφάσεων. Αντιπροσωπεύει ολο το σύνολο δεδομένων, το οποίο χωρίζεται σε δύο και ανομοιογενή σύνολα.

Leaf Node: Οι κόμβοι φύλλων είναι ο τελικός κόμβος εξόδου και το δέντρο δεν μπορεί να διαχωριστεί περαιτέρω μετά τη λήψη ενός κόμβου φύλλου.

Splitting: Ο διαχωρισμός είναι η διαδικασία διαίρεσης του κόμβου απόφασης/κόμβου ρίζας σε υποκόμβους αναλόγως των δεδομένων συνθηκών.

Branch/Sub Tree: Το υποδένδρο είναι ένα δέντρο το οποίο σχηματίζεται από το διαχωρισμό του δέντρου.

Pruning: Το κλάδεμα είναι η διαδικασία αφαίρεσης των ανεπιθύμητων κλαδιών από το δέντρο.

Parent/Child node: Η ρίζα κόμβος του δέντρου ονομάζεται κόμβος γονιός και άλλοι κόμβοι ονομάζονται κόμβοι παιδί.

Κεφάλαιο 6: Εφαρμογή αλγορίθμων

Εισαγωγή

Σε αυτό το σημείο χρησιμοποιήσαμε ένα αρχείο .arff που αφορά τον καιρό και τρέξαμε όλους τους αλγορίθμους. Πιο συγκεκριμένα τους παρακάτω:

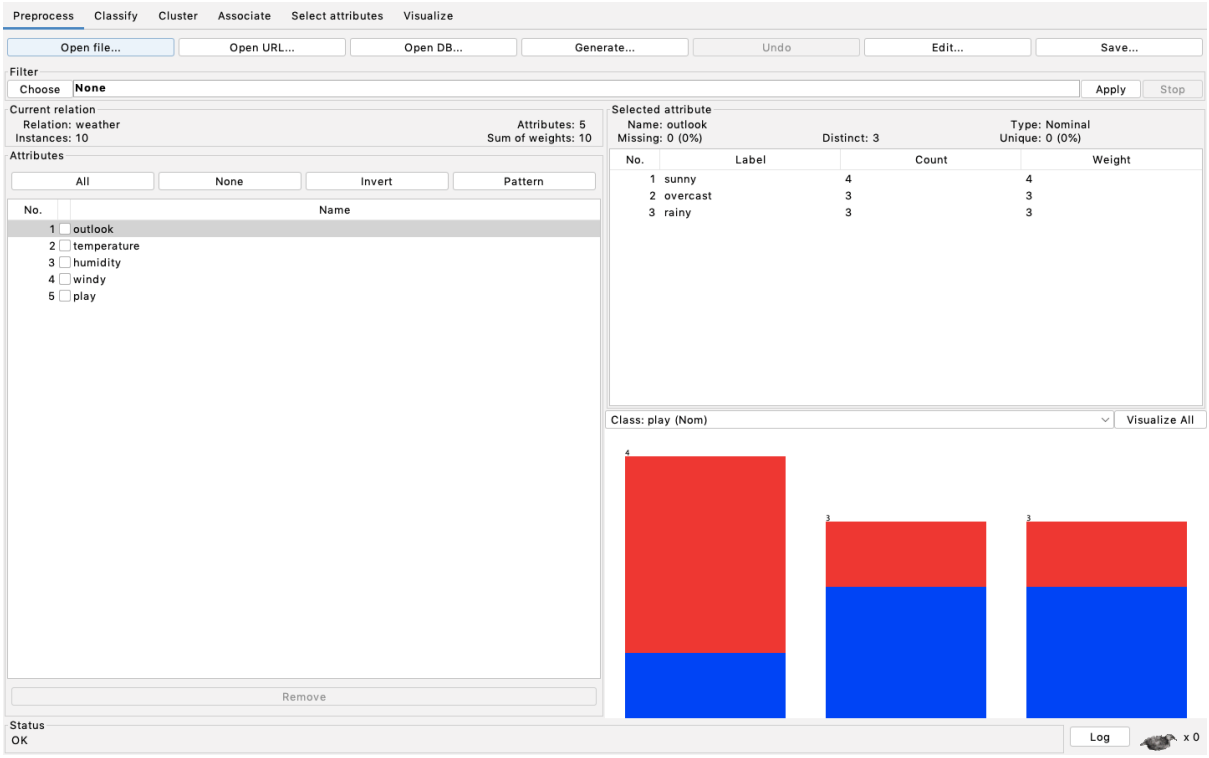
- K-means
- SMO
- J48

Στη συνέχεια παρουσιάζεται το αρχείο και τα γραφήματα του αρχείου:

```
@relation weather
@attribute outlook {sunny,overcast,rainy}
@attribute temperature numeric
@attribute humidity numeric
@attribute windy {true,false}
@attribute play {yes,no}

@data
sunny,85,85,false,no
sunny,80,90,true,no
overcast,83,86,false,yes
rainy,70,96,false,yes
rainy,68,80,false,yes
rainy,65,70,true,no
overcast,64,65,true,yes
sunny,72,95,false,no
sunny,69,70,false,yes
overcast,75,90,true,no
```

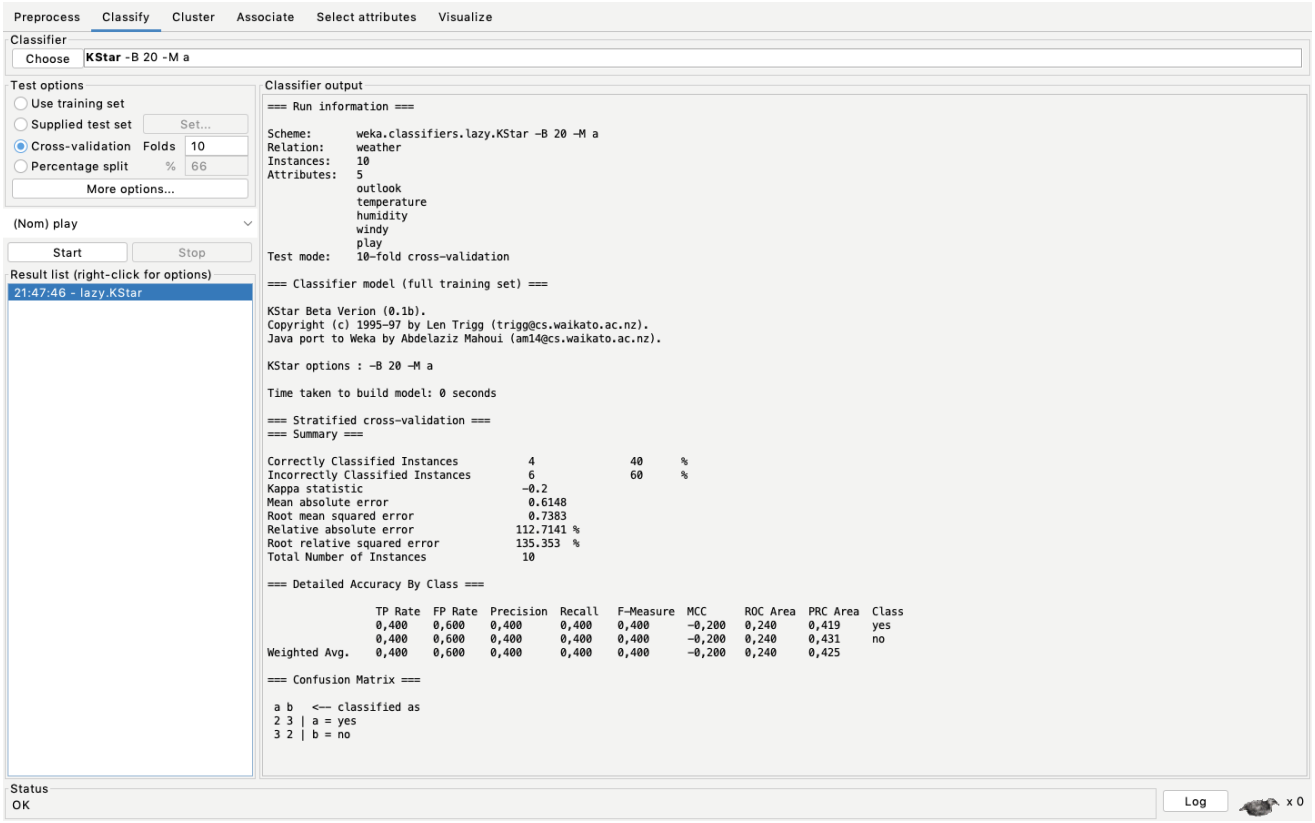
Εικόνα 43: Παράδειγμα αρχείου .arff



Εικόνα 44: Παράδειγμα weka

6.2 K-means

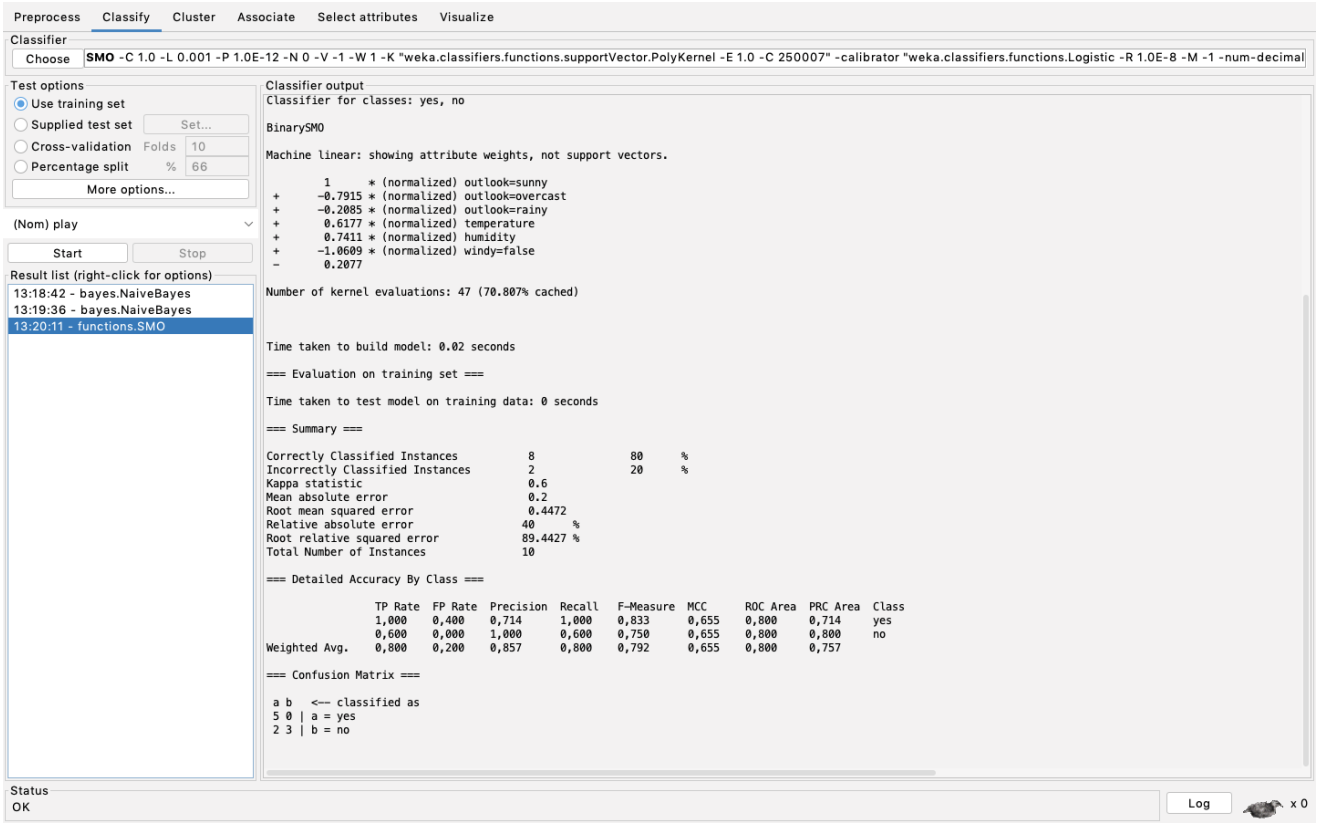
Σε αυτό το σημείο τρέξαμε το αρχείο .arff και στο classify επιλέξαμε K-means όπως βλέπουμε και στην εικόνα που ακολουθεί:



Εικόνα 45: Παράδειγμα K-means

6.3 SMO

Σε αυτό το σημείο τρέξαμε το αρχείο .arff και στο classify επιλέξαμε SMO όπως βλέπουμε και στην εικόνα που ακολουθεί:



Εικόνα 46: Παράδειγμα SMO

6.4 J48

Σε αυτό το σημείο τρέξαμε το αρχείο .arff και στο classify επιλέξαμε J48 όπως βλέπουμε και στην εικόνα που ακολουθεί:

Εικόνα 47: Παράδειγμα j48

ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην παρούσα πτυχιακή εργασία μελετήσαμε το λογισμικό Weka. Στο πρώτο κεφάλαιο ορίσαμε την εξόρυξη δεδομένων ως την ανακάλυψη μεγάλων βάσεων δεδομένων χρησιμοποιώντας αλγόριθμους ομαδοποίησης ή ταξινόμησης. Έπειτα αναφερθήκαμε στους στόχους και την ιστορία της. Η πρώτη προσέγγιση για τον εντοπισμό προτύπων είναι η Bayesian θεωρία και η ανάλυση παλινδρόμησης. Η ευρεία χρήση και η ανάπτυξη της τεχνολογίας των υπολογιστών έχει αυξήσει τον όγκο των δεδομένων που συλλέγονται και την ανάγκη για αποτελεσματική επεξεργασία. Επίσης αναφερθήκαμε σε ορισμένες εφαρμογές σε διάφορους τομείς όπως η ιατρική, η οικονομία και οι τηλεπικοινωνίες.

Υγεία: Τα τελευταία χρόνια, η εξόρυξη δεδομένων έχει χρησιμοποιηθεί ευρέως στη βιοϊατρική, το DNA, τη γενετική, τα φαρμακευτικά προϊόντα και άλλους ιατρικούς τομείς.

Οικονομία: Τα οικονομικά δεδομένα συλλέγονται κυρίως από τράπεζες και άλλους οικονομικούς οργανισμούς. Τα δεδομένα αυτά είναι συνήθως αξιόπιστα, περιεκτικά και υψηλής ποιότητας και ως εκ τούτου ,απαιτούν συστηματικές μεθόδους ανάλυσης .

Τηλεπικοινωνίες: Οι εξελίξεις στην τεχνολογία των τηλεπικοινωνιών οδήγησαν στην ενσωμάτωση τεχνικών εξόρυξης δεδομένων σε αυτές τις τεχνολογίες ώστε να λειτουργούν αποτελεσματικά και να παράγουν αποδοτικά αποτελέσματα.

Στο δεύτερο κεφάλαιο μιλήσαμε για το WEKA, το οποίο είναι ένα λογισμικό για μηχανική μάθηση και εξόρυξη δεδομένων. Αναπτύχθηκε στο Πανεπιστήμιο Waikato της Νέας Ζηλανδίας και διατίθεται ως ελεύθερο λογισμικό. Εκτός από το Weka υπάρχουν και άλλα ελεύθερα λογισμικά επιχειρηματικής Ευφυΐας και εξόρυξης δεδομένων. Μερικά από τα πιο γνωστά φαίνονται παρακάτω:

- IBM Watson Analytics.
- Microsoft Power BI.
- SAP Lumira Cloud.
- Pentaho Community Edition.
- Jaspersoft.
- Jedox Base Business Intelligence.
- SpagoBI.

- KNIME.
- Tableau Public.

Στο τρίτο κεφάλαιο κάναμε μία ανάλυση στο γραφικό περιβάλλον του Weka και στον τρόπο λειτουργίας του καθώς επίσης παρουσιάσαμε και ορισμένα από τα χαρακτηριστικά του. Οι αλγόριθμοι ταξινόμησης που είναι διαθέσιμοι για εκμάθηση είναι τα δέντρα αποφάσεων, οι μηχανές διανυσμάτων υποστήριξης, οι ταξινομητές που βασίζονται σε στιγμιότυπα, η παλινδρόμηση και τα δίκτυα Bayes.

Το WEKA χρησιμοποιεί την καρτέλα Cluster για να προβλέψει τις ομοιότητες στο σύνολο δεδομένων.

Το WEKA υποστηρίζει τη δισδιάστατη αναπαράσταση δεδομένων, τις τρισδιάστατες απεικονίσεις με περιστροφή και την 1Δ αναπαράσταση ενός μόνο χαρακτηριστικού.

Το τέταρτο κεφάλαιο αφορά τα βήματα εγκατάστασης ένα προς ένα αναλυτικά. Και το πέμπτο και τελευταίο κεφάλαιο αφορά τους αλγόριθμους μηχανικής μάθησης όπως:

K- means : Η χρήση του K-Means Clustering ως αλγόριθμος μάθησης χωρίς επίβλεψη στη μηχανική μάθηση ή στην επιστήμη δεδομένων μπορεί να λύσει προβλήματα σχετικά με την ομαδοποίηση. Αυτός ο αλγόριθμος επικεντρώνεται στη συσχέτιση κάθε συστάδας με ένα κέντρο, στοχεύοντας τελικά στην ελαχιστοποίηση των συνολικών αποστάσεων μεταξύ των σημείων δεδομένων και των αντίστοιχων συστάδων τους.

SVM : Η Μηχανική Εκμάθηση χρησιμοποιεί κυρίως Support Vector Machine (SVM), έναν από τους δημοφιλέστερους αλγόριθμους για εποπτευόμενη μάθηση. Ο αλγόριθμος Support Vector Machine στοχεύει στη δημιουργία ενός υπερεπίπεδου, το οποίο αποτελεί το καλύτερο όριο απόφασης που μπορεί να διαιρέσει τον n-διάστατο χώρο σε κλάσεις για εύκολη ταξινόμηση μελλοντικών σημείων δεδομένων και δέντρα αποφάσεων.

Decision Tree: Είναι μια τεχνική που χρησιμοποιείται για την επίλυση προβλημάτων ταξινόμησης, είναι ένας ταξινομητής δομημένου δέντρου που μπορεί επίσης να χειριστεί προβλήματα παλινδρόμησης. Αποτελεί μια εποπτευόμενη μέθοδος εκμάθησης όπου τα χαρακτηριστικά ενός συνόλου δεδομένων αντιπροσωπεύονται από εσωτερικούς κόμβους, οι κανόνες απόφασης αντιπροσωπεύονται από κλάδους και το αποτέλεσμα αντιπροσωπεύεται από κάθε κόμβο φύλλου.

Στο τελευταίο κεφάλαιο βρήκαμε ένα αρχείο .arff και τρέξαμε στο weka τους : K- means, SMO, J48 και πήραμε τα αντίστοιχα αποτελέσματα.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- Weka Tutorial – How To Download, Install And Use Weka Tool(2023). https://www.softwaretestinghelp.com/weka-tutorial/?fbclid=IwAR3z6GaQJZQw5m0hjJRnuBKMB_NQfKV7g-PRWK2HDKLhNf5RUtXfI1oMNhE

- Ελένη Γολέμη.,(2010).Κρυπτογραφία & Εξόρυξη Δεδομένων. http://nemertes.lis.upatras.gr/jspui/bitstream/10889/4791/1/ergasia-golemie.pdf

- Kantardzic, Mehmed (2003). Data Mining: Concepts, Models, Methods, and AlgorithmsΑπαιτείται δωρεάν εγγραφή. John Wiley & Sons. ISBN 0-471-22852-4. OCLC 50055336.

- Simmi Bagga., Dr. G.N. Singh., (2012).Applications of Data Mining. http://www.ijsett.com/images/P5.pdf

- Usama Fayyad· Gregory Piatetsky-Shapiro· Padhraic Smyth (1996). «From Data Mining to Knowledge Discovery in Databases» .

- Γούλου Ζωή.,(2010). Εφαρμογή μεθόδων εξόρυξης δεδομένων στη διαχείριση πελατειακών σχέσεων. http://dspace.lib.uom.gr/bitstream/2159/14808/6/GoulouZoiMsc2012.pdf

- http://www.ibm.com/developerworks/views/opensource/libraryview.jsp?search_by=data+mining+weka.

- Meta-guide.com. (2015). 100 Best Weka Tutorial Videos | Meta-Guide.com. http://meta-guide.com/videography/100-best-weka-tutorial-videos/.

- Slideshare.net. (2012). Data mining techniques using weka. http://www.slideshare.net/rathorenitin87/data-mining-techniques-using-weka.

- Technologyforge.net. (2015). Technology Forge - WEKA Tutorials. http://www.technologyforge.net/WekaTutorials/.

- http://ikee.lib.auth.gr/record/329260/files/GRI-2021-30208.pdf?fbclid=IwAR1NsehHzxkq9kchS16jIZhz4EKq1a7hRdR6pranqVU4QNRjqjbZxito-bg

- https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

- https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm?fbclid=IwAR2YLkKfadgdlKY0RrW1TZFcComY_U_9uFxHRiZTbYffo_seOgetVxaLIlI

- https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

- https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm?fbclid=IwAR20uDOsXQ_wlHhKFZaVOFgICE_TUfGBU2LJIEmR74vl8z-ai0Hd5jy2BIo