ΤΕΧΝΟΛΟΓΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΕΠΙΚΟΙΝΩΝΙΑΣ ΚΑΙ ΣΠΟΥΔΩΝ ΔΙΑΔΙΚΤΥΟΥ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

"ΣΥΓΚΡΙΣΗ ΜΕΘΟΔΩΝ ΠΡΟΒΛΕΨΗΣ ΨΗΦΟΥ ΣΤΙΣ ΗΛΕΚΤΡΟΝΙΚΕΣ ΠΛΑΤΦΟΡΜΕΣ ΣΥΜΒΟΥΛΩΝ ΨΗΦΟΥ"

Αρίστη Μακρή

Επιβλέπων καθηγητής: Νικόλας Τσαπατσούλης

Λεμεσός 2013

Πνευματικά δικαιώματα

Copyright © Όνομα Αρίστη Μακρή, 2013
Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Η έγκριση της πτυχιακής εργασίας από το Τμήμα Επικοινωνίας και Μέσων Ενημέρωσης του Τεχνολογικού Πανεπιστημίου Κύπρου δεν υποδηλώνει απαραιτήτως και αποδοχή των απόψεων του συγγραφέα εκ μέρους του Τμήματος.

Περίληψη

Η παρούσα πτυχιακή εργασία με τίτλο "Σύγκριση μεθόδων πρόβλεψης ψήφου στις Ηλεκτρονικές Πλατφόρμες Συμβούλων Ψήφου", εκπονήθηκε από την Αρίστη Μακρή, φοιτήτρια του 8ου εξαμήνου του Τμήματος ΕΣΔ του ΤΕΠΑΚ υπό την επίβλεψη του Επίκουρου Καθηγητή Νικόλα Τσαπατσούλη και ολοκληρώθηκε το Μάιο του 2013.

Η εργασία αυτή, σκοπό είχε την σύγκριση ορισμένων μεθόδων πρόβλεψης ψήφου για τον ηλεκτρονικό σύμβουλο ψήφου της Κύπρου. Οι μέθοδοι αυτοί είναι τύπου ταξινόμησης και ομαδοποίησης. Η μεθοδολογία που χρησιμοποιήθηκε για την εκτέλεση της εργασίας αυτής, ήταν κυρίως πειραματική και ποσοτική. Τα δεδομένα που χρησιμοποιήθηκαν στην μελέτη αυτή, έχουν συλλεχθεί από την έρευνα που έχει γίνει για τον ηλεκτρονικό σύμβουλο ψήφου της Κύπρου. Η έρευνα αυτή διήρκησε ένα μήνα περίπου και το δείγμα αφορούσε άτομα από όλη την Κύπρο, άντρες, γυναίκες διαφόρων ηλικιών. Το δείγμα της έρευνας αποτελείται από 477 άτομα. Στο σύνολο αυτό περιλαμβάνονται οι απαντήσεις 35 ερωτήσεων, η πρόθεση ψήφου για τον κάθε χρήστη και ορισμένα δημογραφικά χαρακτηριστικά. Επίσης, συμπεριλαμβάνονται και οι απαντήσεις των ερωτήσεων για τους τρείς κύριους υποψήφιους προέδρους της Κύπρου. Στην συνέχεια, χρησιμοποιώντας και δημιουργώντας κάποια διανύσματα με τα δεδομένα αυτά, γίνεται χρήση του κύριου εργαλείου εξόρυξης δεδομένων, weka, για να εξαχθούν τα αποτελέσματα. Για την εξαγωγή αποτελεσμάτων, τα δεδομένα χωρίστηκαν σε τρείς κατηγορίες. Η πρώτη περιλάμβανε τις ερωτήσεις και τα δημογραφικά χαρακτηριστικά, η δεύτερη τα δημογραφικά χαρακτηριστικά μόνο και η τρίτη μόνο τις ερωτήσεις. Από τα αποτελέσματα που προέκυψαν, συμπεραίνεται ότι το μεγαλύτερο ποσοστό ορθότητας το δίνουν οι αλγόριθμοι Bayes Net , Simple Cart και DTNB. Ο καλύτερος αλγόριθμος όμως φαίνεται να είναι ο DTNB με ποσοστό 65%. Εκτός από το ποσοστό ορθότητας των αλγορίθμων, σημαντικό ρόλο παίζει και ο μέσος όρος ακρίβειας των αλγορίθμων όπως αποδείχθηκε. Ο αλγόριθμος K-Star, έχει το υψηλότερο ποσοστό ακρίβειας από όλους τους αλγορίθμους. Το ποσοστό ανάκλησης του αλγορίθμου αυτού είναι 0.56 και ο μέσος όρος ακρίβειας είναι 0.59. Αυτό σημαίνει ότι ο αλγόριθμος K-Star, πρόβλεψε λιγότερους ψήφους, αλλά από αυτούς που πρόβλεψε δεν υπήρχαν πολλά περιττά ερωτηματολόγια στην κλάση, δηλαδή πρόβλεψε ορθά περισσότερους ψήφους. Τέλος, ένα άλλο σημαντικό συμπέρασμα, είναι ότι τα καλύτερα αποτελέσματα βγαίνουν από την κατηγορία όπου χρησιμοποιούνται μόνο οι ερωτήσεις των ερωτηματολογίων, αφού έχει υψηλά ποσοστά ορθότητας σε όλους τους αλγορίθμους.

Ευχαριστίες

Η παρούσα εργασία δεν θα μπορούσε να ολοκληρωθεί χωρίς την υποστήριξη του επιβλέποντος Επίκουρου Καθηγητή κ. Νικόλα Τσαπατσούλη και τη βοήθεια των συναδέλφων του κ. Ιωάννη Κατάκη και κ. Ζήνωνα Θεοδοσίου. Θα ήθελα να τους ευχαριστήσω για τις πολύτιμες συμβουλές τους καθ' όλη τη διάρκεια εκπόνησης της πτυχιακής μου.

Ιδιαίτερες ευχαριστίες στον αρραβωνιαστικό μου, την οικογένεια μου και στους φίλους μου που με στήριξαν στις ατελείωτες ώρες μελέτης και συγγραφής της εργασίας αυτής.

Πίνακας Περιεχομένων

Κατάλογος Πινάκων

Κατάλογος Εικόνων

Απόδοση όρων

Τεχνητή Νοημοσύνη	Artificial Intelligence
Εξόρυξη δεδομένων	Data Mining
Ταξινόμηση	Classification
Ομαδοποίηση	Clustering
Δένδρα Απόφασης	Decision trees
Πρόβλεψη	Prediction
Τμήματα	Segments
Διασταυρούμενη Επικύρωση	Cross-validation
Συνεργατικό Φιλτράρισμα	Collaborative Filtering
Ταξινομητής	Classifier
Ονομαστικός	Nominal
Αριθμητικές	Numeric
Αληθώς θετικό	True Positive (TP)
Ψευδώς θετικό	False Positive (FP)
Ακρίβεια	Precision
Ανάκληση	Recall
Κέντρα κλάσεων (βαρύκεντρο)	Centroids

Κεφάλαιο 1 : Εισαγωγή

"Τα τελευταία χρόνια με την ραγδαία ανάπτυξη του Παγκόσμιου Ιστού, ολοένα και περισσότεροι χρήστες κατακλύζονται με τεράστιους όγκους πληροφορίας. Καθημερινά, όλο και περισσότερη πληροφορία είναι εύκολα προσβάσιμη στον καθένα». Αυτό όμως είναι πρόβλημα επειδή οι χρήστες δεν μπορούν να ξεχωρίσουν ποια είναι η πληροφορία που πραγματικά τους ενδιαφέρει. Το παραπάνω φαινόμενο ανακαλείται «υπερφόρτωση πληροφοριών" (Φεσαλίδη, 2009).

Λόγω του φαινομένου της «υπερφόρτωσης πληροφοριών», έχουν αυξηθεί και οι αγορές μέσω του Διαδικτύου και αποτελούν ένα μεγάλο κομμάτι του μεριδίου αγοράς. Επομένως, οι άνθρωποι βασίζονται σε συστάσεις από άλλους ανθρώπους για να προβούν σε μια αγορά. Το παραπάνω φαινόμενο, μπορεί να χαρακτηριστεί και ως «πρόβλημα», αφού οι χρήστες δυσκολεύονται να επιλέξουν αυτό που πραγματικά τους ενδιαφέρει, επειδή υπάρχουν πολλές επιλογές. Το «πρόβλημα» αυτό λοιπόν, μπορεί να αντιμετωπιστεί με τα συστήματα συστάσεων. Τα συστήματα συστάσεων ευκολύνουν το χρήστη στις επιλογές τους. Ένα σύστημα συστάσεων θα ωφελήσει το χρήστη κάνοντας του προτάσεις αναφορικά με βιβλία, ταινίες, μουσική, ιστοσελίδες, άρθρα κτλ, οι οποίες εκτιμάται ότι θα ικανοποιήσουν τις απαιτήσεις του (Βοζαλής, 2007). Επίσης, τα συστήματα συστάσεων χρησιμοποιούνται αρκετά στο χώρο του ηλεκτρονικού επιχειρείν, αφού με τις προβλέψεις και συστάσεις αγορών που γίνονται στο χρήστη ωφελούνται και οι επιχειρήσεις. Ιστοχώροι όπως το Amazon και το eBay, χρησιμοποιούν τεχνικές παραγωγής συστάσεων για να διευκολύνουν τους χρήστες. Για παράδειγμα, το Amazon, καταγράφει όλες τις αγοραστικές συνήθειες των αγοραστών του, και όταν συνδεθούν στον ιστοχώρο, χρησιμοποιεί αυτές τις πληροφορίες για να τους προτείνει προϊόντα που μπορεί να τους αρέσουν. Ένα κατάστημα στο διαδίκτυο είναι σε καλύτερη θέση από ένα πραγματικό κατάστημα, για να δει και να καταγράψει όχι μόνο τις αγορές κάποιου πελάτη, αλλά και τι στοιχεία έχει σκεφτεί, κοιτάξει, και απορρίψει, άρα με βάση αυτό μπορεί να καταγράφει εύκολα το ιστορικό του κάθε χρήστη (Joseph, Riedl, 2012).

Μια τεχνική που χρησιμοποιούν τα συστήματα συστάσεων για να παράγουν συστάσεις είναι το «συνεργατικό φιλτράρισμα», που χρησιμεύει για τον υπολογισμό της ομοιότητας ανάμεσα στους χρήστες. Χρησιμοποιεί τη συμπεριφορά αυτών που μοιάζουν περισσότερο

με ένα συγκεκριμένο χρήστη ως μια λειτουργική βάση για τη δημιουργία προβλέψεων και προτάσεων για το συγκεκριμένο χρήστη.

Στα συστήματα συστάσεων οι χρήστες που είναι συνδεδεμένοι στο Διαδίκτυο λαμβάνουν συστάσεις από άλλους χρήστες με παρόμοια γούστα και προτιμήσεις. Η ομοιότητα των προτιμήσεων εκτιμάται από τις δράσεις των χρηστών, όπως για παράδειγμα αγορές προϊόντων, και τις αξιολογήσεις που κάνουν, τόσο οι χρήστες που ζητούν, όσο και οι χρήστες που παρέχουν συστάσεις. Οι συστάσεις αυτές, δημιουργούνται έμμεσα από την πραγματική συμπεριφορά των άλλων χρηστών. Κανένας όμως στην πραγματικότητα δεν ζητά να παρέχει συστάσεις, έτσι, η προστασία της ιδιωτικής ζωής και της ανωνυμίας των χρηστών διατηρείται (Tsapatsoulis & Georgiou, 2012).

Κεφάλαιο 2: Θεωρητική θεμελίωση

Στις επόμενες παραγράφους παρουσιάζουμε τη σημαντικότητα του προβλήματος που μελετάμε στην παρούσα εργασία, τα ερευνητικά ερωτήματα και την θεωρητική θεμελίωση μέσω ανασκόπησης της σχετικής βιβλιογραφίας.

2.1 Περιγραφή Προβλήματος-Αναγκαιότητα Μελέτης

Όσον αφορά τα συστήματα συστάσεων, είναι πολύ σημαντική η ύπαρξή τους, αφού όλοι μας χρησιμοποιούμε καθημερινά το Διαδίκτυο. Πολλές εφαρμογές των συστημάτων συστάσεων ευκολύνουν τους χρήστες για τις διάφορες επιλογές τους. Μερικές από τις εφαρμογές τους είναι στο Amazon, στο eBay και στην ιστοσελίδα reddit.com όπως αναφέρθηκε και πιο πάνω η λειτουργία τους. Μέσα από τις εφαρμογές αυτές, διαφαίνεται η σημαντικότητα των συστημάτων συστάσεων γενικότερα. Παρόλα αυτά, η συγκεκριμένη μελέτη, δίνει έμφαση στους μεθόδους πρόβλεψης ψήφου στις ηλεκτρονικές πλατφόρμες συμβούλων ψήφου. Οι σύμβουλοι αυτοί, παρουσίασαν μεγάλο ενδιαφέρον και εφαρμόστηκαν σε πολλές Ευρωπαϊκές χώρες. Είναι σημαντικό οι σύμβουλοι ψήφου να δίνουν όσο το δυνατόν πιο ακριβείς προτάσεις. Γι' αυτόν τον λόγο χρησιμοποιούνται συνεχώς όλο και περισσότεροι μέθοδοι πρόβλεψης. Επομένως, υπάρχει η ανάγκη αυτοί οι μέθοδοι πρόβλεψης να μελετηθούν εις βάθος, έτσι ώστε να αναλυθούν τα πλεονεκτήματα και τα μειονεκτήματα του καθενός και να προταθούν βελτιστοποιήσεις.

Η μελέτη αυτή συγκρίνει διάφορες μεθόδους πρόβλεψης ψήφων με βάση το σύμβουλο ψήφου της Κύπρου. Οι χρήστες του Choose4Cyprus έπρεπε να υποβάλουν τη γνώμη τους για 30 ερωτήσεις συν κάποιες συμπληρωματικές ερωτήσεις για δημογραφικές πληροφορίες, πρόθεση ψήφου και αυτό-τοποθέτηση σχετικά με τις κύριες πολιτικές απόψεις τους. Για κάθε ερώτηση ο χρήστης έπρεπε να επιλέξει μία από τις παρακάτω απαντήσεις: 1) Συμφωνώ Πλήρως, 2) Συμφωνώ, 3) Ούτε συμφωνώ ,ούτε διαφωνώ 4) Διαφωνώ 5) Διαφωνώ Πλήρως 6) Χωρίς άποψη. Για τις ίδιες δηλώσεις έδιναν τη γνώμη τους και τα πολιτικά κόμματα. Στη συνέχεια, τα συστήματα συστάσεων παρήγαγαν μια πρόταση στον κάθε χρήστη ανάλογα με τη ομοιότητα των απαντήσεων του με κάποια πολιτικά κόμματα (Katakis, Tsapatsoulis, Triga, Tziouvas, Mendez, n.d.). Δηλαδή όσες πιο πολλές κοινές απαντήσεις είχε ο χρήστης με το συγκεκριμένο πολιτικό κόμμα , τόσο πιο πολύ ταίριαζαν

οι απόψεις τους. Έτσι, η πρόβλεψη ψήφου, γινόταν με βάση αυτό το κριτήριο. Με αυτό τον τρόπο, συμβουλεύουν τους χρήστες τι να ψηφίσουν.

Αυτό είναι πολύ σημαντικό για ανθρώπους που δε γνωρίζουν τι θα ψηφίσουν. Απαντώντας σε κάποιες ερωτήσεις, το σύστημα προτείνει στους χρήστες να ψηφίσουν συγκεκριμένη πολιτική πλευρά ανάλογα με τις απαντήσεις που έχουν δώσει. Με αυτό τον τρόπο βοηθά τους χρήστες να παίρνουν γρήγορες αποφάσεις. Ένα άλλο παράδειγμα είναι, αν κάποιος πολίτης είναι σε δίλημμα και δεν έχει αποφασίσει τι θα ψηφίσει, ο σύμβουλος ψήφου του δίνει την εύκολη λύση. Η εφαρμογή αυτή βοηθά τον οποιοδήποτε πολίτη να κατανοήσει πλήρως τι πρεσβεύει το κάθε κόμμα και ποιες είναι ακριβώς οι πολιτικές του θέσεις. Πολλοί πολίτες νομίζουν ότι γνωρίζουν τι αντιπροσωπεύει το κόμμα το οποίο προτιμούν, όμως στην πραγματικότητα είτε δεν είναι αρκετά ενήμεροι για τις θέσεις του κόμματος αυτού, είτε είναι παραπληροφορημένοι, είτε απλά επιλέγουν το συγκεκριμένο κόμμα γιατί έτσι έμαθαν από τους γονείς τους. Γενικά, η εφαρμογή αυτή είναι αναγκαία για τους πολίτες όλων των κατηγοριών, από τον πιο ενημερωμένο και κατατοπισμένο πολιτικά μέχρι τον πιο αναποφάσιστο.

Επομένως, η μελέτη αυτή, κρίνεται αναγκαία για το λόγο ότι οι σύμβουλοι ψήφου μπορούν να γίνουν ακόμη πιο ακριβείς και βοηθητικοί για τους αναποφάσιστους χρήστες που δεν ξέρουν τι να ψηφίσουν. Επίσης, θα ήταν χρήσιμη η σύγκριση των αλγορίθμων αυτών, αφού θα βοηθήσει και άλλους μελλοντικούς ερευνητές του γνωστικού αυτού αντικειμένου, να επιλέξουν τον καταλληλότερο και αποδοτικότερο αλγόριθμο για την έρευνά τους.

2.2 Ερευνητικό Ερώτημα

Ποια μέθοδος πρόβλεψης ψήφου είναι πιο αποτελεσματική με βάση τις πολιτικές απόψεις των χρηστών και τα δημογραφικά τους χαρακτηριστικά;

Σκοπός της παρούσας μελέτης είναι η σύγκριση των μεθόδων πρόβλεψης ψήφου στην ηλεκτρονική πλατφόρμα συμβούλου ψήφου της Κύπρου, λαμβάνοντας υπόψη τις πολιτικές απόψεις και τα δημογραφικά χαρακτηριστικά του κάθε χρήστη για την πρόταση που θα γίνεται. Για την διεκπεραίωση αυτής της μελέτης, έπρεπε αρχικά να γίνει μια έρευνα για την συλλογή των δεδομένων και ακολούθως να επεξεργαστούν τα δεδομένα αυτά στο πρόγραμμα εξόρυξης δεδομένων weka για να εξαχθούν τα κατάλληλα αποτελέσματα.

2.3 Ανασκόπηση Βιβλιογραφίας

Στο μέρος αυτό, της μελέτης, περιγράφονται εργασίες άλλων ερευνητών σχετικές με το θέμα της έρευνας.

Η μελέτη των Tsapatsoulis και Georgiou (2012), αναφέρεται στην επεκτασιμότητα των αλγορίθμων, στο ρόλο της μετρικής ομοιότητας και στη λίστα του σχήματος κατασκευής προτεινόμενων αντικειμένων. Επισημαίνεται ότι η καλύτερη μέθοδος για αντιμετώπιση της επεκτασιμότητας , είναι η ομαδοποίηση χρηστών. Ο όρος επεκτασιμότητα αναφέρεται στην ικανότητα να παρέχει επιτυχείς συστάσεις σε εύλογο χρονικό διάστημα όταν ασχολείται με πάρα πολύ μεγάλο αριθμό χρηστών. Επίσης, η μέθοδος του κοντινότερου χρήστη, είναι η πιο χρησιμοποιημένη για το συνεργατικό φιλτράρισμα. Βασίζεται στον K-NN αλγόριθμο και είναι συνήθως σε συνδυασμό με το Top-N σχέδιο σύστασης, προκειμένου να δημιουργηθεί η λίστα συστάσεων των στοιχείων. Στον αλγόριθμο K-NN, κάθε φορά που ένας χρήστης κάνει αιτήματα για συστάσεις, οι K πλησιέστεροι γείτονες προσδιορίζονται με βάση τα στοιχεία υψηλής βαθμολογίας και μια λίστα με συστάσεις παρέχεται στο χρήστη. Τέλος, ο K-NN, αποδείχθηκε ότι έχει την καλύτερη απόδοση πρόβλεψης.

Άλλες έρευνες που έχουν γίνει εφαρμόζουν τα συστήματα συστάσεων στις πλατφόρμες συμβούλων ψήφου, χωρίς όμως να λαμβάνουν υπόψη τα δημογραφικά χαρακτηριστικά των χρηστών στις συστάσεις που γίνονται. Οι περισσότερες έρευνες παράγουν συστάσεις με βάση τις συμπεριφορές των χρηστών και τις βαθμολογίες που δίνουν σε προϊόντα , όπως για παράδειγμα το Amazon.

Οι David και Balakrishnan (n.d.), στη μελέτη που έχουν εκπονήσει, χρησιμοποιούν παρόμοιους αλγόριθμους με την παρούσα έρευνα, όπως τον k-means και τον J48. Σκοπός της έρευνάς τους είναι να δείξει τη σημασία των δύο τεχνικών ταξινόμησης, δηλαδή των δέντρων αποφάσεων και της ομαδοποίησης, στην πρόβλεψη των μαθησιακών δυσκολιών των παιδιών σχολικής ηλικίας. Στην μελέτη τους έχουν χρησιμοποιήσει 125 πραγματικά σύνολα δεδομένων με 16 χαρακτηριστικά που τα περισσότερα παίρνουν δυαδικές τιμές για την ταξινόμηση. Τα αποτελέσματα της έρευνας τους είναι ότι ο αλγόριθμος J48, είναι κατάλληλος για τις ελλιπείς τιμές. Επίσης, ο J48 είναι καλός στην έννοια της απόδοσης και της πολυπλοκότητας. Τα δέντρα απόφασης δίνουν ποσοστό ορθότητας 77.6%. Για τον αλγόριθμο ομαδοποίησης, έχουν βρει την σημασία των χαρακτηριστικών στην ομαδοποίηση που γίνεται.

Στην έρευνά του ο Κωτσιαντής, 2005, βγάζει το συμπέρασμα ότι ο αλγόριθμος Bayes είναι αποδοτικότερος όταν ο αριθμός των μεταβλητών είναι μεγάλος. Όταν όμως το μέγεθος των δεδομένων εκπαίδευσης είναι μικρό και υπάρχει θόρυβος στα δεδομένα, οι αλγόριθμοι που βασίζονται στη βαθμολογία μπορούν συχνά να δώσουν τα ακριβέστερα αποτελέσματα.

Αλγόριθμος	**Ακρίβεια (%)**
Naïve Bayes	92.61
RBF networks	93.67
Trees-J48	92.97
Trees-CART	92.97
SVM-RBF kernel	98.06

Πίνακας 2.3.1: **Αποτελέσματα από την έρευνα των Aruna et al. (2011)**

Ακόμη, στην έρευνα τους οι Aruna, Rajagopalan & Nandakishore, 2011, μελετούν το κριτήριο της αποτελεσματικότητας των εργαλείων μηχανικής μάθησης για την ταξινόμηση του καρκίνου του μαστού. Συγκρίνουν τα εργαλεία εξόρυξης δεδομένων, όπως ο Naïve Bayes, μηχανές διανυσμάτων υποστήριξης Radial με βάση τα νευρωνικά δίκτυα, δέντρα απόφασης J48 και τον αλγόριθμο CART. Τα πειράματα τους πραγματοποιούνται σε WEKA. Ο στόχος αυτής της έρευνας είναι να μάθουν τον καλύτερο ταξινομητή όσον αφορά την ακρίβεια, την ευαισθησία και ειδικότητα στην ανίχνευση του καρκίνου του μαστού. Τα αποτελέσματα της έρευνας αυτής παρουσιάζονται στον πίνακα 2.3.1.

Τέλος, στην έρευνα του Κωτσόπουλου, 2012, γίνεται πάλι μια σύγκριση κάποιων μεθόδων εξόρυξης δεδομένων. Σκοπός της έρευνάς του είναι μέσα από ένα σύνολο δεδομένων τα όποια θα υποστούν την κατάλληλη επεξεργασία να αναδειχθούν γνωρίσματα και μοτίβα που διέπουν την ωοπαραγωγική διαδικασία των ορνίθων ώστε να βελτιστοποιήσουν την αποτελεσματικότητα των αντίστοιχων διαδικασιών. Χρησιμοποιεί την μέθοδο κατηγοριοποίησης για να βγάλει τα επιθυμητά αποτελέσματα. Τα αποτελέσματα απεικονίζονται στον πιο κάτω πίνακα.

Ποσοστά επιτυχούς πρόβλεψης της συνολικής ωοπαραγωγής
Το καλύτερο αποτέλεσμα ανά στήλη σημειώνεται με έντονη γραμματοσειρά.
CCI: Correctly classified instances, CC (%) = Percentage of correct classifications

Γνωστές ημέρες ωοπαραγωγής	Rules-DecisionTable		Rules-DTNB		Functions-Multilayer Perceptron	
	CCI	CC (%)	CCI	CC (%)	CCI	CC (%)
30 ημέρες	**8043**	**29.48**	7384	27.07	7938	29.10
45 ημέρες	8479	31.08	7878	28.88	**8481**	**31.09**
60 ημέρες	9244	33.89	8883	32.56	**9298**	**34.08**
75 ημέρες	10052	36.85	9578	35.11	**10082**	**36.96**
90 ημέρες	10909	39.99	10647	39.03	**11515**	**42.21**
105 ημέρες	12179	44.65	11558	42.37	**12890**	**47.25**
120 ημέρες	13916	51.01	12856	47.13	**14341**	**52.57**
135 ημέρες	15414	56.51	13982	51.26	**15942**	**58.44**
160 ημέρες	**19212**	**70.43**	17970	65.88	18926	69.38

Πίνακας 2.3 2: Αποτελέσματα από την έρευνα Κωτσόπουλου (2012)

Κεφάλαιο 3: Μεθοδολογία

Η μεθοδολογία που χρησιμοποιήθηκε στην παρούσα έρευνα για τη διεξαγωγή έγκυρων, αξιόπιστων αποτελεσμάτων και χρήσιμων συμπερασμάτων, είναι ο συνδυασμός ποσοτικής ανάλυσης και πειραματικής μεθόδου. Με την επιλογή αυτών των μεθόδων, έχει απαντηθεί το ερευνητικό ερώτημα που τέθηκε αρχικά. Επειδή έπρεπε να συγκρίνουμε τις μεθόδους πρόβλεψης ψήφου, δεν θα μπορούσε να χρησιμοποιηθεί οποιαδήποτε άλλη μέθοδος από τις προαναφερθείσες, εκτός από την πειραματική μέθοδο που έχει χρησιμοποιηθεί για να αναδείξει την πιο αποτελεσματική μέθοδο πρόβλεψης ψήφου.

Για την διεκπεραίωση της μελέτης αυτής, χρησιμοποιήθηκαν τα δεδομένα τα οποία έχουν συλλεχθεί από την έρευνα που έχει γίνει για το σύμβουλο ψήφου της Κύπρου. Η έρευνα αυτή διήρκησε ένα μήνα περίπου και το δείγμα αφορούσε άτομα από όλη την Κύπρο, άντρες, γυναίκες διαφόρων ηλικιών. Η έρευνα αυτή σκοπό είχε την επιλογή καλύτερων ερωτήσεων για την πλατφόρμα και αντιπροσώπευε και άτομα τα οποία δεν είχαν σύνδεση με το διαδίκτυο. Το δείγμα της έρευνας ήταν περίπου 850 άτομα. Μερικά όμως ερωτηματολόγια θεωρήθηκαν άκυρα για την έρευνα αυτή, αφού μερικά άτομα δεν ολοκλήρωναν το ερωτηματολόγιο ή δεν απαντούσαν σε όλες τις ερωτήσεις και έτσι το σύνολο των δειγμάτων αυτών δεν παρέμεινε 850, επομένως αυτού του είδους ερωτηματολόγια δεν λαμβάνονταν υπόψη. Στο σύνολο αυτό περιλαμβάνονται οι απαντήσεις των 35 ερωτήσεων, η πρόθεση ψήφου για τον κάθε χρήστη και ορισμένα δημογραφικά χαρακτηριστικά. Επίσης, στα ερωτηματολόγια οι υποψήφιοι πρόεδροι ήταν οκτώ αλλά στην έρευνα αυτή χρησιμοποιήθηκαν μόνο οι τρείς για καλύτερη διεξαγωγή αποτελεσμάτων. Το πραγματικό σύνολο των χρηστών είναι 477.

Αρχικά, τα δεδομένα που συλλέχθηκαν χωρίστηκαν σε τρείς ομάδες. Στην πρώτη ομάδα ήταν όλα τα δεδομένα μαζί, δηλαδή ερωτήσεις και δημογραφικά χαρακτηριστικά. Στην δεύτερη ομάδα ήταν μόνο τα δημογραφικά χαρακτηριστικά και στην τρίτη ομάδα μόνο οι ερωτήσεις που υπήρχαν στο ερωτηματολόγιο. Συνολικά οι ερωτήσεις ήταν 35 και τα δημογραφικά στοιχεία ήταν οκτώ (φύλο, έτος γέννησης, μορφωτικό επίπεδο, καταγωγή, αν ήταν πρόσφυγας, εισόδημα, θρησκεία και κριτήριο επιλογής ψήφου). Αυτά τα δημογραφικά χαρακτηριστικά θεωρήθηκαν ως τα καταλληλότερα γιατί με αυτά μπορούσαν να παραχθούν οι κατάλληλες προτάσεις για τους χρήστες. Δηλαδή, το φύλο και η ηλικία παίζουν σημαντικό ρόλο για τους χρήστες, γιατί θα ενημερώνονται ποιοι χρήστες της ίδιας

ηλικίας και φύλου απάντησαν με τον ίδιο τρόπο με αυτούς. Επίσης, το μορφωτικό επίπεδο είχε μεγάλη σημασία επειδή οι χρήστες είναι πιθανόν να ήθελαν να ξέρουν αν οι χρήστες που τους πρότεινε το σύστημα ως κοντινούς, είναι υψηλού επιπέδου μόρφωσης ή χαμηλού. Ο λόγος έγκειται στο γεγονός ότι είναι παρατηρημένο ότι πολίτες με ψηλότερο επίπεδο μόρφωσης τείνουν να είναι πιο ενημερωμένοι σε πολιτικά ζητήματα ή και να ασχολούνται ενεργά με την πολιτική.

Στη συνέχεια αφού δημιουργήθηκαν οι ομάδες των δεδομένων, έγινε πρόβλεψη για τον κάθε χρήστη για το τι σκόπευε να ψηφίσει στις εκλογές. Η πρόβλεψη, γινόταν εξαιτίας της ερώτησης που υπήρχε στο ερωτηματολόγιο «Τι σκοπεύετε να ψηφίσετε στις εκλογές τις 17 Φεβρουαρίου;». Ακολούθως, συγκρινόταν η πρόβλεψη με τη δήλωση του χρήστη και ανάλογα με το αν συμφωνούσε η πρόβλεψη με τη δήλωση, τότε θα ήταν επιτυχής η πρόβλεψη αλλιώς αν διαφωνούσαν θα ήταν αποτυχημένη. Σημειώνεται ότι για την σύγκριση της πρόβλεψης με τη δήλωση, χρησιμοποιήθηκαν ξεχωριστά κάθε φορά (α) μόνο τα διανύσματα με τις ερωτήσεις, (β) μόνο τα διανύσματα με τα δημογραφικά χαρακτηριστικά, και, (γ) συνδυάζοντας και τα δύο σύνολα διανυσμάτων.

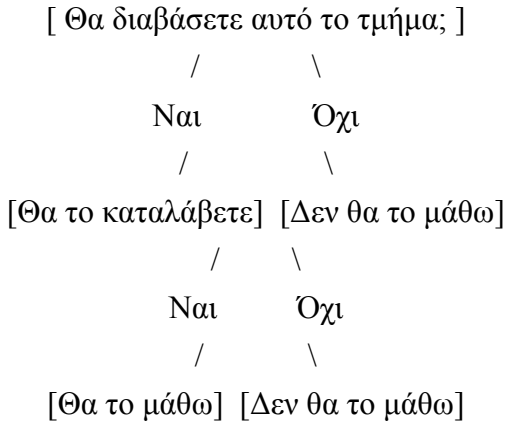
Στην παρούσα μελέτη, χρησιμοποιήθηκαν οι εξής μέθοδοι πρόβλεψης ψήφου:

K-Means , J48, Bayes, Bayes Net, Rule classifier, SMO classifier, DTNB Class, K-Star Class, Naive Bayes και JRip. Οι αλγόριθμοι αυτοί χωρίζονται σε τρεις κατηγορίες. Την Ταξινόμηση, την Ομαδοποίηση και τα Δέντρα Αποφάσεων. Οι μέθοδοι αυτοί περιγράφονται αναλυτικά πιο κάτω. Οι αλγόριθμοι JRip και Bayes δεν αναφέρονται πιο κάτω επειδή λειτουργούν με την ίδια λογική όπως οι κανόνες ταξινόμησης και Naïve Bayes αντίστοιχα.

3.1 Ταξινόμηση

Η ταξινόμηση, γνωστή και ως δέντρα ταξινόμησης ή δέντρα απόφασης, είναι ένας αλγόριθμος εξόρυξης δεδομένων που δημιουργεί βήμα-προς-βήμα ένα οδηγό για το πώς να καθορίσει την παραγωγή ενός νέου στιγμιότυπου δεδομένων. Το δέντρο που δημιουργεί είναι ακριβώς αυτό: ένα δέντρο όπου κάθε κόμβος του δέντρου αντιπροσωπεύει ένα σημείο όπου μια απόφαση πρέπει να γίνεται με βάση την είσοδο, και να προχωρά στον επόμενο κόμβο και τον επόμενο μέχρι να φτάσει σε ένα φύλλο που λέει την προβλεπόμενη έξοδο. (Abernethy, 2010).

Ένα παράδειγμα ταξινόμησης είναι το πιο κάτω:



Αυτό το απλό δέντρο ταξινόμησης επιχειρεί να απαντήσει στο ερώτημα "Θα καταλάβετε τα δέντρα ταξινόμησης;" Σε κάθε κόμβο, απαντά στην ερώτηση και προχωρά στον ίδιο κλάδο, μέχρι να φτάσει σε ένα φύλλο που απαντά ναι ή όχι. Αυτό το μοντέλο μπορεί να χρησιμοποιηθεί για οποιοδήποτε άγνωστο παράδειγμα δεδομένων, και είναι σε θέση να προβλέψει αν αυτό το στιγμιότυπο με άγνωστα δεδομένα θα μάθει τα δέντρα ταξινόμησης, ζητώντας τους μόνο δύο απλές ερωτήσεις. Αυτό είναι φαινομενικά το μεγάλο πλεονέκτημα ενός δέντρου ταξινόμησης - δεν απαιτεί πολλές πληροφορίες σχετικά με τα δεδομένα να δημιουργήσει ένα δέντρο που θα μπορεί να είναι πολύ ακριβό και πολύ κατατοπιστικό.

3.2 Ομαδοποίηση

Επιτρέπει στον χρήστη να δημιουργήσει ένα συγκεκριμένο αριθμό ομάδων αναλόγως με τις επαγγελματικές του ανάγκες. Το πλεονέκτημα της ομαδοποίησης σε σχέση με την ταξινόμηση είναι ότι κάθε χαρακτηριστικό μέσα στο σύνολο δεδομένων θα χρησιμοποιηθεί για την ανάλυση των δεδομένων. Ένα κύριο μειονέκτημα του να χρησιμοποιεί κανείς την ομαδοποίηση, είναι ότι ο χρήστης πρέπει να ξέρει από την αρχή πόσες ομάδες θέλει να δημιουργήσει. Για έναν χρήστη χωρίς πραγματική γνώση των δεδομένων του, αυτό μπορεί να είναι δύσκολο. Με λίγα λόγια μπορεί να θεωρηθεί ως μια τμηματοποίηση των δεδομένων σε ομάδες, που μπορεί να είναι ή να μην είναι διακριτές μεταξύ τους. Αυτό επιτυγχάνεται με τον καθορισμό της ομοιότητας ανάμεσα στα δεδομένα. Τα δεδομένα που είναι σχετικότερα μεταξύ τους κατατάσσονται στις ίδιες ομάδες.

3.3 Δένδρα απόφασης (Decision trees)

Τα δέντρα αποφάσεων είναι εποπτευόμενοι αλγόριθμοι που χωρίζουν τα δεδομένα αναδρομικά, βασισμένοι στα χαρακτηριστικά των δεδομένων , μέχρι να ικανοποιηθεί μια τερματική συνθήκη. Ο ταξινομητής δένδρου απόφασης είναι μια από τις πιθανές προσεγγίσεις για πολυεπίπεδη λήψη αποφάσεων. Το πιο σημαντικό στοιχείο του ταξινομητή δένδρου απόφασης είναι η ικανότητα διάσπασης της διαδικασίας απόφασης σε μια συλλογή από πιο απλές αποφάσεις, με τέτοιο τρόπο, έτσι ώστε να παρέχουν μια λύση που είναι συνήθως πιο εύκολο να μεταφραστεί.

Η μεθοδολογία των δέντρων ταξινόμησης και οπισθοδρόμησης είναι πιθανώς η πιο γνωστή και διαδεδομένη μεθοδολογία. Χρησιμοποιεί διασταυρούμενη επικύρωση ή ένα μεγάλο ανεξάρτητο δείγμα δεδομένων για να επιλέξει το καλύτερο δέντρο από την ακολουθία δέντρων που θεωρούνται μέσα στη διαδικασία κλαδέματος. Ο βασικός αλγόριθμος αυτής της μεθοδολογίας είναι ένας άπληστος αλγόριθμος που επιλέγει το πιο διακριτό τοπικά χαρακτηριστικό σε κάθε βήμα της διαδικασίας. Αυτό είναι μη βέλτιστο, αλλά η πλήρης αναζήτηση ενός βέλτιστου συνόλου από ερωτήσεις θα ήταν υπολογιστικά πολύ ακριβό. Αυτή η προσέγγιση είναι εναλλακτική ως προς την παραδοσιακή μέθοδο για πρόβλεψη. Μέσα στην υλοποίηση αυτής της μεθοδολογίας, το σύνολο δεδομένων είναι διαχωρισμένο σε δύο υποσύνολα που διαφέρουν περισσότερο σε σχέση με το αποτέλεσμα. Αυτή η διαδικασία συνεχίζεται σε κάθε υποσύνολο μέχρι κάποιο ελάχιστο υποσύνολο να προσεγγιστεί.

3.4 Ο αλγόριθμος K-Means

"Ο αλγόριθμος k-means είναι αλγόριθμος ομαδοποίησης που χρησιμοποιεί μία απλή και εύκολη διαδικασία ανάθεσης των δειγμάτων ενός συνόλου δεδομένων σε έναν a priori καθορισμένο αριθμό ομάδων (έστω K). Η κύρια ιδέα είναι να καθοριστούν K κέντρα, ένα για κάθε ομάδα. Αυτά τα κέντρα πρέπει να τοποθετούνται με έξυπνο τρόπο, διότι για κάθε διαφορετική θέση προκύπτει διαφορετικό αποτέλεσμα. Έτσι, η καλύτερη επιλογή είναι να τοποθετηθούν όσο το δυνατόν μακρύτερα το ένα από το άλλο. Το επόμενο βήμα είναι να ληφθεί κάθε σημείο που ανήκει σε ένα δεδομένο σύνολο στοιχείων και να συνδεθεί με το κοντινότερο κέντρο. Όταν κανένα σημείο δεν εκκρεμεί πια, το πρώτο βήμα ολοκληρώνεται και έχει δημιουργηθεί ένα πρώιμο σύνολο ομάδων. Σε αυτό το σημείο πρέπει να υπολογίσουμε εκ νέου τα K κέντρα ως τα βαρύκεντρα των ομάδων του προηγούμενου βήματος. Εφόσον εντοπίσουμε αυτά τα νέα κέντρα, μια νέα σύνδεση πρέπει να γίνει μεταξύ

των στοιχείων δεδομένων και του κοντινότερου ως προς αυτά νέου κέντρου. Η διαδικασία (επαναϋπολογισμού των κέντρων και εκ νέου ανάθεσης των στοιχείων δεδομένων στις ομάδες που ορίζονται από τα κέντρα) εκτελείται επαναληπτικά, με αποτέλεσμα τα κέντρα να αλλάζουν τη θέση τους βαθμιαία. Η συνθήκη τερματισμού του αλγορίθμου ελέγχει πότε δεν γίνονται άλλες αλλαγές στις αναθέσεις των στοιχείων δεδομένων, ή ισοδύναμα, πότε τα κέντρα σταματούν να μετακινούνται"(Κωτσόπουλος, 2012).

3.5 Ο αλγόριθμος J48

Ο J48 αλγόριθμος αποτελεί υλοποίηση ανοικτού κώδικα σε Java του αλγόριθμου C4.5 μέσα στο εργαλείο εξόρυξης δεδομένων weka και βασίζεται στα δέντρα απόφασης. Ο C4.5 είναι ένα πρόγραμμα που δημιουργεί ένα δέντρο απόφασης βασισμένο σε ένα σύνολο από δεδομένα εισόδου με ετικέτες. Αυτός ο αλγόριθμος κατασκευάστηκε από τον Ross Quinlan. Τα δέντρα αποφάσεων που δημιουργούνται από τον C4.5, μπορούν να χρησιμοποιηθούν για ταξινόμηση. Για τον λόγο αυτό αναφέρεται συχνά σαν ένας στατιστικός ταξινομητής (Gholap).

3.6 Δίκτυο Bayes (Bayes Net)

Το δίκτυο Bayes είναι βασισμένο επίσης στα δέντρα απόφασης. Είναι ένας τρόπος να αναπαριστά την από κοινού κατανομή ενός συνόλου μεταβλητών με έναν τρόπο που είναι ιδιαίτερα χρήσιμος για την αναπαράσταση γνώσης. Για παράδειγμα η περιγραφή ενός σερβίτσιου για ένα γεύμα. Όλα τα αντικείμενα, δηλαδή το πιάτο, το σκεύος, η χαρτοπετσέτα, το μπολ, η κούπα και τα υλικά που μπορεί να είναι φτιαγμένα τα πιο πάνω, είναι μεταβλητές. Τα κύρια αντικείμενα υποδηλώνουν τα τμήματα της ρύθμισης του χώρου. Η μεταβλητή ρύθμιση έχει τέσσερις πιθανές τιμές (πρωινό, γεύμα, δείπνο, επιδόρπιο)που δηλώνουν τα αντίστοιχα είδη των γευμάτων. Ένα πιάτο μπορεί να είναι χάρτινο ή κεραμικό. Επίσης, υπάρχει πιθανότητα, τα αντικείμενα αυτά να έχουν σχέση μεταξύ τους. Για παράδειγμα, μια χαρτοπετσέτα αναμένεται σε κάθε ένα από τα πιθανά γεύματα (Rimey, 1992).

3.7 Ταξινόμηση με κανόνες (Rule classifier)

Δίνεται ένας πληθυσμός του οποίου τα μέλη μπορεί δυνητικά να διαχωρίζονται σε έναν αριθμό διαφορετικών συνόλων ή τάξεις, ένας κανόνας ταξινόμησης είναι μια διαδικασία στην οποία τα στοιχεία του συνόλου του πληθυσμού εκχωρούνται σε μια από τις κλάσεις.

Επιτυχημένη δοκιμή θεωρείται αυτή που έχει εκχωρήσει κάθε στοιχείο του πληθυσμού στην κατηγορία που πραγματικά ανήκει. Μια αποτυχημένη δοκιμή είναι αυτή που κατά την οποία θα εμφανιστούν μερικά λάθη και, στη συνέχεια θα πρέπει να εφαρμοστεί στατιστική ανάλυση για να αναλύσει την ταξινόμηση. Ένα ιδιαίτερο είδος του κανόνα ταξινόμησης είναι οι δυαδικές ταξινομήσεις.

3.8 Ταξινόμηση με διανύσματα υποστήριξης: SMO classifier

Διαδοχική ελάχιστη βελτιστοποίηση (SMO) είναι ένας αλγόριθμος για την αποτελεσματική επίλυση του προβλήματος βελτιστοποίησης που προκύπτει κατά τη διάρκεια της εκπαίδευσης των μηχανών με διανύσματα υποστήριξης. Ο SMO σπάει το πρόβλημα σε μια σειρά από μικρότερα επιμέρους προβλήματα, τα οποία στη συνέχεια λύνονται αναλυτικά. Εφευρέθηκε από τον John Platt το 1998. Χρησιμοποιείται ευρέως σε μηχανές εκπαίδευσης διανυσμάτων. Η δημοσίευση του αλγορίθμου αυτού το 1998, έχει δημιουργήσει μεγάλο ενθουσιασμό στην κοινότητα των μηχανών διανυσμάτων, αφού προηγουμένως, οι διαθέσιμες μέθοδοι για την εκπαίδευση ήταν πολύ πιο πολύπλοκη και δαπανηρή.

3.9 Υβριδικός ταξινομητής DTNB Class

Κλάση για κτίσιμο και χρήση ενός πίνακα απόφασης /υβριδικού naive bayes ταξινομητή. Σε κάθε σημείο της αναζήτησης, ο αλγόριθμος αξιολογεί την αξία διαχωρισμού των χαρακτηριστικών μέσα σε δύο διαχωρισμένα υποσύνολα: ένα για τον πίνακα απόφασης και το άλλο για τον Naive Bayes. Χρησιμοποιείται αναζήτηση εμπρόσθιας επιλογής, όπου σε κάθε βήμα, τα επιλεγμένα χαρακτηριστικά μοντελοποιούνται από τον naive Bayes και από το υπόλοιπο του πίνακα απόφασης, και όλα τα χαρακτηριστικά μοντελοποιούνται από τον πίνακα απόφασης αρχικά. Σε κάθε βήμα, ο αλγόριθμος λαμβάνει υπόψη την εξ' ολοκλήρου αφαίρεση ενός χαρακτηριστικού από το μοντέλο (Hall & Eibe, 2008).

3.10 Ταξινόμηση με βάση τα υποδείγματα: K-Star Class

K-Star είναι ένας ταξινομητής βασισμένος σε στιγμιότυπα, κι έτσι η κλάση ενός δοκιμαστικού στιγμιότυπου βασίζεται στην κλάση παρόμοιων στιγμιότυπων εκπαίδευσης, όπως καθορίζεται με κάποια συνάρτηση ομοιότητας. Διαφέρει από άλλους ταξινομητές βασισμένους σε στιγμιότυπα στο ότι χρησιμοποιεί μια συνάρτηση απόστασης βασισμένη στην εντροπία (Cleary & Trigg, 1995).

3.11 Πιθανοτική ταξινόμηση: Naive Bayes

Είναι ένας πιθανοτικός ταξινομητής που βασίζεται στο θεώρημα Bayes. Επίσης εξετάζει μια ισχυρή παραδοχή ανεξαρτησίας. Ο ταξινομητής αυτός θεωρεί ότι όλα τα χαρακτηριστικά ανεξάρτητα συμβάλλουν στην πιθανότητα μιας συγκεκριμένης απόφασης. Λαμβάνοντας υπόψη τη φύση του υποκείμενου μοντέλου πιθανοτήτων, ο αφελής ταξινομητής Bayes μπορεί να εκπαιδευτεί πολύ αποτελεσματικά σε επιτηρούμενο περιβάλλον μάθησης και να εργάζεται πολύ καλύτερα σε πολλές σύνθετες καταστάσεις του πραγματικού κόσμου. Επειδή οι μεταβλητές θεωρούνται ανεξάρτητες, μόνο οι διακυμάνσεις των μεταβλητών για κάθε κλάση πρέπει να προσδιορίζονται και όχι ολόκληρη η μήτρα συνδιακύμανσης. (Aruna, Rajagopalan & Nandakishore, 2011)
Οι παραπάνω αλγόριθμοι ανήκουν στην μέθοδο ταξινόμησης εκτός από τον αλγόριθμο k-means, ο οποίος ανήκει στην μέθοδο ομαδοποίησης. Οι δυο αυτοί όροι περιγράφονται πιο κάτω.

Όλες αυτές οι μέθοδοι και αλγόριθμοι που θα χρησιμοποιηθούν ανήκουν στην μελέτη της τεχνητής νοημοσύνης. Σύμφωνα με τους Luger και Stubblefield (όπ. αναφ. στο Ε.Κεραυνού, 2000), Τεχνητή Νοημοσύνη είναι η μελέτη των μηχανισμών που διέπουν ευφυή συμπεριφορά, μέσω της κατασκευής και αξιολόγησης συστημάτων τα οποία παριστάνουν αυτούς τους μηχανισμούς. Με την τεχνητή νοημοσύνη προσπαθούν οι επιστήμονες της πληροφορικής να εξελίξουν τις μηχανές σε τέτοιο σημείο ώστε να σκέφτονται και να κρίνουν σε τόσο καλό βαθμό όσο και ο άνθρωπος. Με λίγα λόγια, κύριος στόχος τους είναι να κάνουν τις μηχανές να έχουν παρόμοια σκέψη με τον άνθρωπο, να λύνουν δύσκολα προβλήματα, να σκέφτονται έξυπνα και όχι μηχανικά. Ένα αντικείμενο με το οποίο ασχολούνται οι επιστήμονες της τεχνητής νοημοσύνης είναι η εξόρυξη δεδομένων.

Εξόρυξη δεδομένων είναι η διαδικασία κατά την οποία ο υπολογιστής εκπαιδεύεται με τη χρήση μιας ή περισσότερων τεχνικών για να αναλύει αυτόματα και να εξάγει γνώσεις από δεδομένα που περιέχονται σε μια βάση δεδομένων (Roiger & Geatz, 2008). Σκοπός μιας συνεδρίας εξόρυξης πληροφορίας είναι να εντοπίσει μοτίβα στα δεδομένα. Για παράδειγμα στην παρούσα μελέτη εντοπίζει μοτίβα τα οποία περιέχουν τις απαντήσεις των ατόμων που απάντησαν τα ερωτηματολόγια για τον σύμβουλο ψήφου της Κύπρου. Κύριο εργαλείο για

την εξόρυξη δεδομένων θεωρείται το πρόγραμμα weka, που θα είναι χρήσιμο για τις προβλέψει ψήφων.

Είναι ένα λογισμικό ανοικτού κώδικα για εξόρυξη δεδομένων, γραμμένο σε γλώσσα προγραμματισμού java, το οποίο περιέχει υλοποιημένες μεθόδους για προεπεξεργασία δεδομένων, ταξινόμηση, ομαδοποίηση και εύρεση κανόνων συσχέτισης (Τσιράκης, 2008). Το weka χρησιμοποιεί αρχεία τύπου arff . Τα αρχεία αυτά αποτελούνται από μια κεφαλίδα, η οποία περιγράφει τους τύπους των χαρακτηριστικών και από την ενότητα των δεδομένων η οποία περιέχει τα δεδομένα χωρισμένα με κόμμα.

Γενικά οι υπολογιστές μπορούν να μάθουν γεγονότα, έννοιες, διαδικασίες και αρχές. Το γεγονός είναι μια απλή δήλωση αλήθειας. Η έννοια είναι μια ομάδα αντικειμένων, συμβόλων ή συμβάντων τα οποία ομαδοποιούνται λόγω του ότι έχουν συγκεκριμένα κοινά χαρακτηριστικά. Διαδικασία είναι μια σειρά από ενέργειες για την επίτευξη ενός σκοπού. Χρησιμοποιούμε διαδικασίες καθημερινά στη ζωή μας για την επίλυση δύσκολων προβλημάτων. Οι αρχές αντιπροσωπεύουν το υψηλότερο επίπεδο εκμάθησης. Είναι γενικές αλήθειες ή νόμοι οι οποίοι είναι βασικοί για άλλες αλήθειες.

Οι υπολογιστές είναι καλύτεροι στην εκμάθηση εννοιών. Οι έννοιες είναι το αποτέλεσμα μιας ολοκληρωμένης διαδικασίας εξόρυξης πληροφοριών. Υπεύθυνο για την εκμάθηση των εννοιών είναι το εργαλείο εξόρυξης πληροφορίας το οποίο υπαγορεύει την μορφή των δεδομένων. Στις συνήθεις δομές εννοιών εντάσσονται δένδρα, κανόνες, δίκτυα, και μαθηματικές εξισώσεις. Οι κανόνες και οι δενδροειδείς μορφές γίνονται πιο εύκολα κατανοητοί από τους ανθρώπους, ενώ αντίθετα τα δίκτυα και οι μαθηματικές εξισώσεις δεν γίνονται εύκολα κατανοητοί.

Οι κύριες στρατηγικές εξόρυξης πληροφορίας είναι η κατηγοριοποίηση, η εκτίμηση και η πρόβλεψη. Η κατηγοριοποίηση είναι η πιο κατανοητή στρατηγική εξόρυξης πληροφορίας. Οι εργασίες κατηγοριοποίησης έχουν τρία κοινά χαρακτηριστικά:

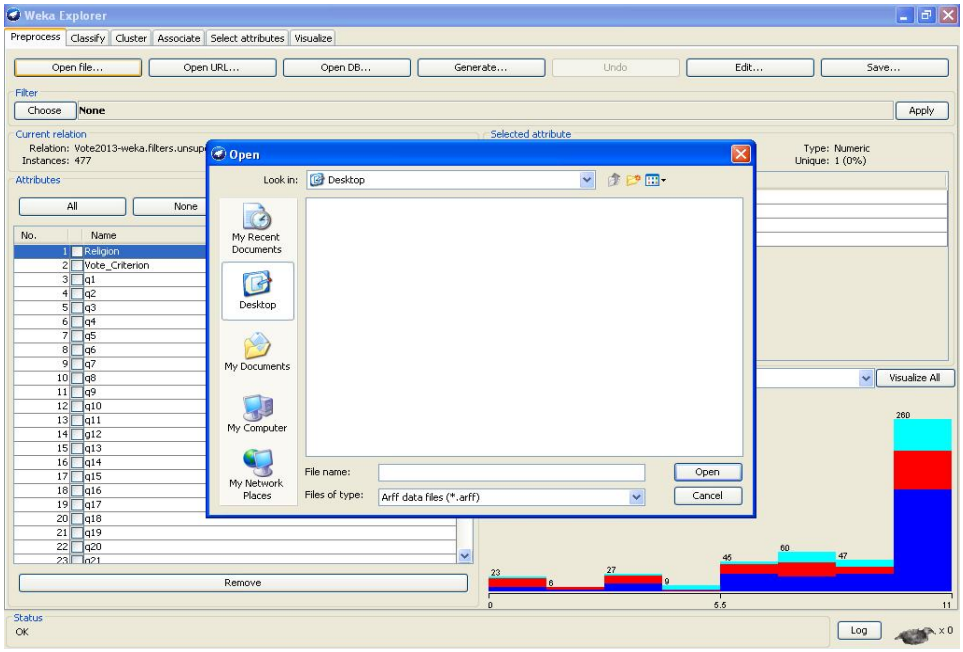
- η εκμάθηση είναι καθοδηγούμενη
- η εξαρτημένη μεταβλητή είναι μεταβλητή κατηγορίας
- κατασκευάζει μοντέλα που είναι ικανά να αντιστοιχίζουν νέα στιγμιότυπα σε μια κλάση.

Στην συγκεκριμένη περίπτωση, η εξαρτημένη μεταβλητή είναι η πρόθεση ψήφου των ερωτώμενων και με βάση τις απαντήσεις τους, προβλέπεται η ψήφος τους. Οι απαντήσεις των ερωτώμενων αντιστοιχίζονται με αριθμούς με βάση την κλίμακα Likert αφού και οι απαντήσεις είναι βασισμένες σε αυτή την κλίμακα. Η κάθε απάντηση αντιστοιχεί με ένα αριθμό από το 1 μέχρι και το 6. Για παράδειγμα, την πρώτη πολιτική δήλωση την υποστηρίζει ο υποψήφιος πρόεδρος Νίκος Αναστασιάδης. Όσα άτομα έδωσαν την πρώτη απάντηση, σημαίνει συμφωνούν με την δήλωση του κύριου Αναστασιάδη κ.ο.κ. Όσον αφορά την πρόβλεψη, δεν είναι καθόλου εύκολο να την διακρίνουμε από την κατηγοριοποίηση ή την εκτίμηση. Η πρόβλεψη καθορίζει τα μελλοντικά αποτελέσματα. Δηλαδή ανάλογα με τις απαντήσεις που έχουν δώσει οι ερωτώμενοι, προβλέπει το σύστημα την μελλοντική τους ψήφο μετά από εκπαίδευση που γίνεται στο σύστημα. Η εκπαίδευση γίνεται με βάση τα μοτίβα που δημιουργούνται, τα οποία περιέχουν αριθμούς ανάλογα με τις απαντήσεις που έδωσαν οι ερωτώμενοι.

Τα αποτελέσματα που εξάγει το weka μπορεί να είναι αριθμητικά ή να αντιστοιχούν σε κατηγορίες. Στο παράδειγμα αυτό, είναι αριθμητικά. Δηλαδή ο υποψήφιος πρόεδρος Νίκος Αναστασιάδης αντιστοιχεί με τον αριθμό 1, ο υποψήφιος πρόεδρος Σταύρος Μαλάς αντιστοιχεί με τον αριθμό 2 και ο υποψήφιος πρόεδρος Γιώργος Λιλλήκας αντιστοιχεί με τον αριθμό 3. Οι τρείς αυτές μεταβλητές έχουν οριστεί ως ονομαστικές, όπως ονομάζονται στο πρόγραμμα weka. Είναι μια καθορισμένη λίστα η οποία περιέχει τις τιμές για τους τρείς υποψήφιους. Τα υπόλοιπα χαρακτηριστικά, δημογραφικά και ερωτήσεις ορίζονται ως αριθμητικά αφού το καθένα αντιστοιχεί και με ένα ακέραιο αριθμό.

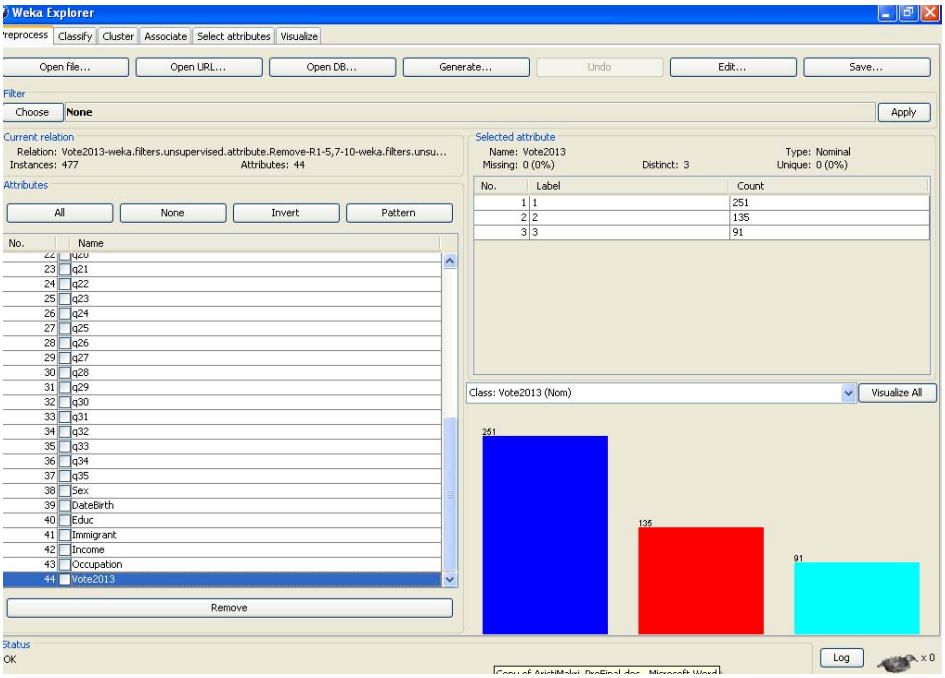
Τέλος, θα γίνει καταγραφή των σωστών και λανθασμένων προβλέψεων και ανάλογα με το πόσες σωστές ή λάθος προβλέψεις υπάρχουν θα εξαχθεί ένα ποσοστό επιτυχίας ή αποτυχίας του αλγορίθμου. Για να επιτευχθεί η διαδικασία αξιολόγησης των αλγορίθμων θα πρέπει να χωριστούν τα δεδομένα σε δεδομένα εκπαίδευσης και δεδομένα αξιολόγησης. Στη συνέχεια θα γίνει εκπαίδευση στο πρόγραμμα για να προβλέπει την ψήφο και μετά θα γίνεται αξιολόγηση του κάθε αλγορίθμου.

3.12 Περιβάλλον Weka



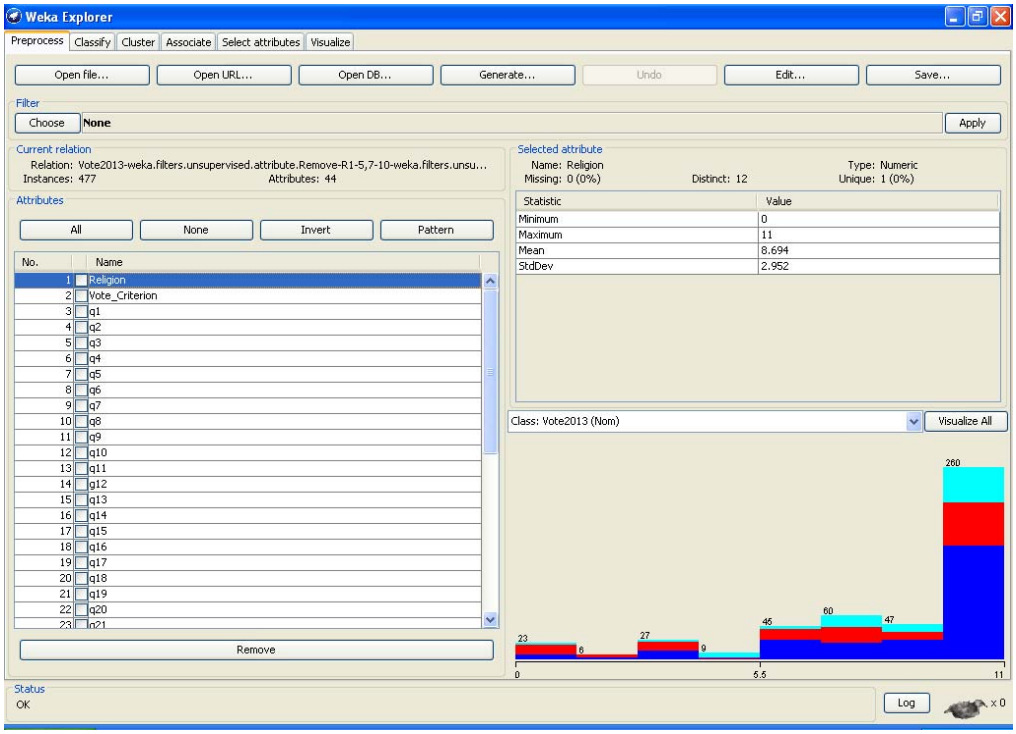
Εικόνα 3.12.1:

Στην μελέτη αυτή, χρησιμοποιήθηκαν τρία στάδια από το περιβάλλον weka. Το preprocess, το classify και το cluster. Αρχικά ξεκινήσαμε από το στάδιο preprocess και εισάγαμε το arff αρχείο μας από το μενού open file όπως δείχνει η Εικόνα 3.12.1.



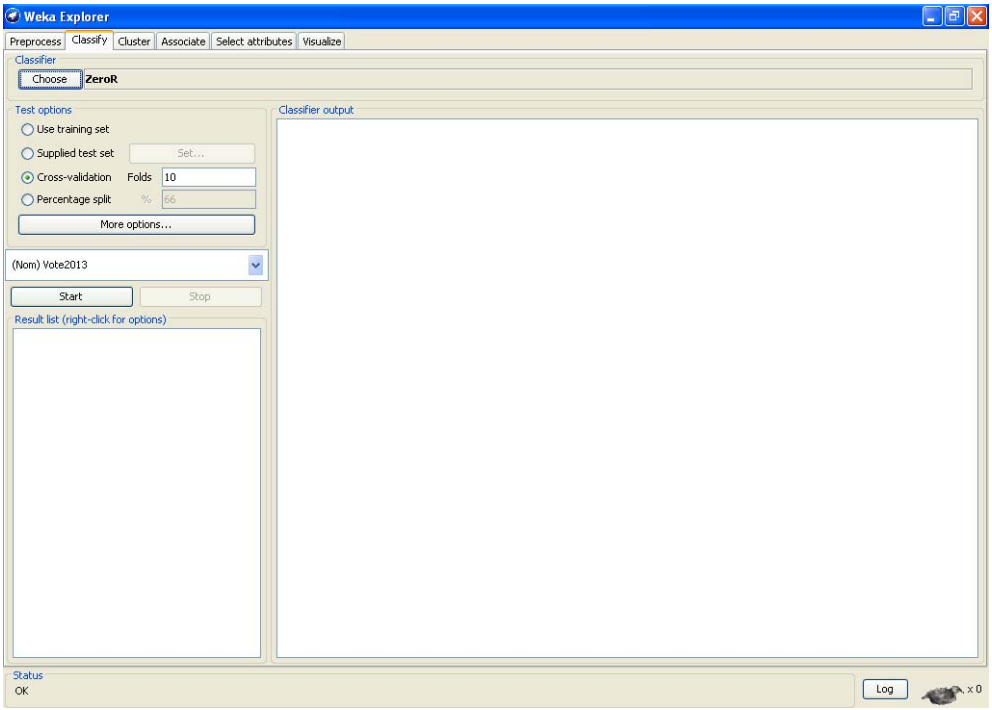
Εικόνα 3.12.2:

Στην συνέχεια εμφανίστηκαν τα δεδομένα μας στο περιβάλλον weka, όπως μας δείχνει η Εικόνα 3.12.2. Στα αριστερά φαίνονται όλα τα χαρακτηριστικά και στα δεξιά η πρόθεση ψήφου των χρηστών για το 2013 και για τους τρεις υποψήφιους. Έχοντας επιλεγμένο το τελευταίο στοιχείο από τα δεδομένα μας που είναι η κλάση(Vote2013), εμφανίζονται με μπλε χρώμα οι ψηφοφόροι του Νίκου Αναστασιάδη, με κόκκινο οι ψηφοφόροι του Σταύρου Μαλά και τιρκουάζ οι ψηφοφόροι του Γιώργου Λιλλήκα.



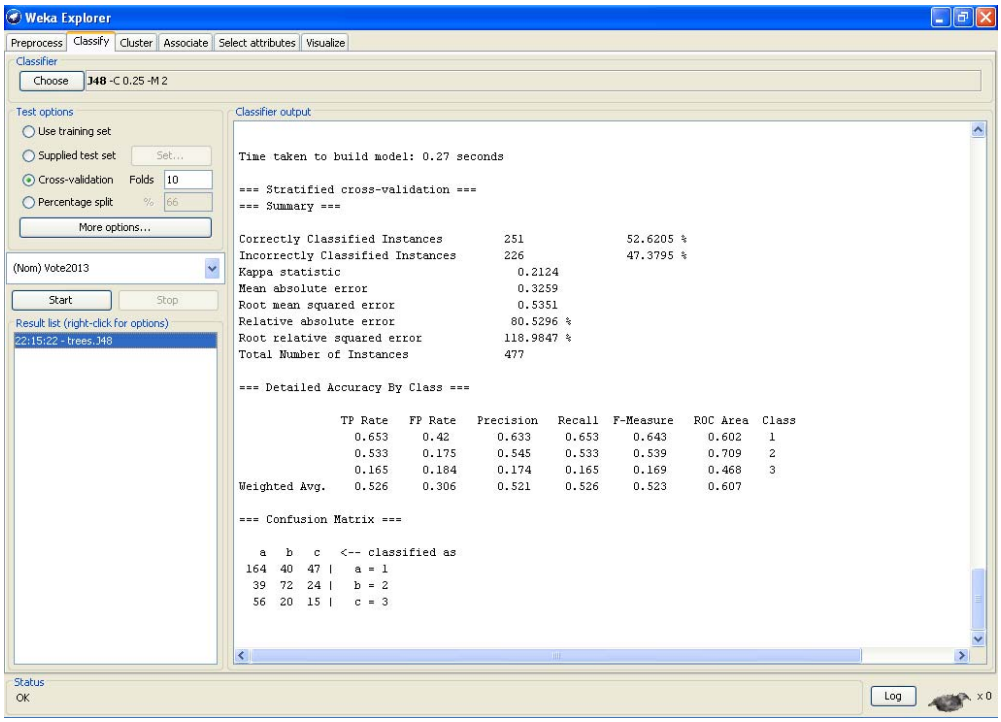
Εικόνα 3.12.3:

Η Εικόνα 3.12.3, μας δείχνει τα χαρακτηριστικά της κλάσης , όπως και την κύρια κλάση. Έχοντας επιλεγμένο ένα χαρακτηριστικό από την αριστερή στήλη, δεξιά εμφανίζονται οι συσχετίσεις του κάθε υποψήφιου και οι απαντήσεις των χρηστών. Δηλαδή με βάση την θρησκεία των χρηστών (όπως είναι επιλεγμένο πιο πάνω), η στήλη δεξιά δείχνει ότι για τους περισσότερους που ψήφισαν Νίκο Αναστασιάδη (μπλε χρώμα), παίζει σημαντικό ρόλο για άλλους η θρησκεία, ενώ για τους άλλους υποψήφιους το ποσοστό είναι μικρότερο.



Εικόνα 3.12.4:

Η Εικόνα 3.12.4 μας δείχνει το στάδιο classify. Μπορούσαμε να επιλέξουμε από το μενού classify την μέθοδο πρόβλεψης που θέλαμε και πατώντας το κουμπί start ξεκινούσε η επεξεργασία δεδομένων.



Εικόνα 3.12.5:

Η Εικόνα 3.12.5 απεικονίζει τα αποτελέσματα του αλγόριθμου J48. Με την ίδια λογική εργαστήκαμε και με το στάδιο cluster.

Κεφάλαιο 4: Αποτελέσματα

Τα αποτελέσματα της έρευνας αυτής έχουν προκύψει από την χρήση του προγράμματος Weka. Μέσα από τα αποτελέσματα, γίνεται αξιολόγηση της απόδοσης των διαφόρων αλγορίθμων του συνεργατικού φιλτραρίσματος στην πλατφόρμα σύμβουλου ψήφου της Κύπρου. Αρχικά, παρουσιάζονται τα αποτελέσματα τα οποία προέκυψαν από το συνδυασμό των δημογραφικών χαρακτηριστικών και των ερωτήσεων της πλατφόρμας, ακολούθως τα αποτελέσματα από τα δημογραφικά χαρακτηριστικά μόνο και τέλος, τα αποτελέσματα από τις ερωτήσεις μόνο. Για την κάθε περίπτωση ξεχωριστά, χρησιμοποιούνται οι ίδιοι αλγόριθμοι για αξιολόγηση της απόδοσής τους.

4.1 Ερμηνεία αποτελεσμάτων

Για την καλύτερη κατανόηση των αποτελεσμάτων χρειάζεται να δοθούν μερικοί ορισμοί:

Τα πραγματικά σύνολα των κλάσεων a,b,c είναι 251,135,91 αντίστοιχα.
Ερμηνεία πίνακα προβλέψεων

```
   a   b   c   <-- classified as
 184  36  31 |   a = 1
  36  83  16 |   b = 2
  49  26  16 |   c = 3
```

Τα πραγματικά σύνολα της κάθε κλάσης φαίνονται από τις γραμμές. Δηλαδή το πραγματικό σύνολο της κλάσης a είναι το άθροισμα των τριών αριθμών στην πρώτη γραμμή (184, 36, 31) . Στην κλάση a όμως πρόβλεψε ορθά μόνο 184. Τα 36 μπήκαν λανθασμένα στην κλάση b και τα 31 λανθασμένα στην κλάση c. Στις στήλες βλέπουμε τις προβλέψεις. Δηλαδή μια μέθοδος για την κλάση a πρόβλεψε συνολικά το άθροισμα των αριθμών της πρώτης στήλης. Τα 36 έπρεπε να μπουν στην κλάση b και τα 49 στην κλάση c. Στην διαγώνιο που φαίνεται, είναι οι ορθές προβλέψεις για την κάθε κλάση.

True Positive

Ποσοστό ορθών προβλέψεων της κλάσης X δια το πραγματικό σύνολο της κλάσης X, ισοδυναμεί με το recall. Δηλαδή πόσα ερωτηματολόγια μπήκαν σωστά σε κάθε κλάση, δια το πραγματικό σύνολο ερωτηματολογίων αυτής της κλάσης.

TP= Ποσοστό της κλάσης X/ πραγματικό σύνολο της κλάσης X.

Για παράδειγμα στον Bayes έχουμε το εξής αποτέλεσμα ορθών και λανθασμένων προβλέψεων :

```
   a   b   c   <-- classified as
 184  36  31 |   a = 1
  36  83  16 |   b = 2
  49  26  16 |   c = 3
```

Το TP της κλάσης a είναι:

(ποσοστό ορθών προβλέψεων για την κλάση a(184)/ πραγματικό σύνολο κλάσης a(184+36+31))

184/251=0.733

False Positive

Ποσοστό λανθασμένων χαρακτηριστικών που βρίσκονται στην κλάση X δια το πραγματικό σύνολο όλων των κλάσεων, εκτός της κλάσης X.

FP= Ποσοστό λανθασμένων χαρακτηριστικών που βρίσκονται στην κλάση X / πραγματικό σύνολο όλων των κλάσεων, εκτός της κλάσης X.

Για παράδειγμα στον αλγόριθμο Bayes έχουμε το εξής αποτέλεσμα ορθών και λανθασμένων προβλέψεων :

```
  a   b   c   <-- classified as
 184  36  31 |   a = 1
  36  83  16 |   b = 2
  49  26  16 |   c = 3
```

Το FP της κλάσης a είναι:

(Λανθασμένα χαρακτηριστικά στην κλάση a που έπρεπε να ήταν στην κλάση b(36)+ Λανθασμένα χαρακτηριστικά στην κλάση a που έπρεπε να ήταν στην κλάση c(49) / πραγματικό σύνολο κλάσης b(135)+πραγματικό σύνολο κλάσης c(91)).

(36+49)/(135+91)=0.376

Recall

Ποσοστό ορθών προβλέψεων της κλάσης X δια το πραγματικό σύνολο της κλάσης X.

Recall= Ποσοστό ορθών προβλέψεων της κλάσης X/ πραγματικό σύνολο της κλάσης X.

Για παράδειγμα στον Bayes έχουμε το εξής αποτέλεσμα ορθών και λανθασμένων προβλέψεων :

```
  a   b   c   <-- classified as
 184  36  31 |   a = 1
  36  83  16 |   b = 2
  49  26  16 |   c = 3
```

Το recall της κλάσης b είναι:

(ποσοστό ορθών προβλέψεων για την κλάση b(83)/ πραγματικό σύνολο κλάσης b(36+83+16))

83/135=0.615

Precision

Ποσοστό από τα παραδείγματα που έχει η κλάση Χ και είναι αληθές δια το σύνολο της κλάσης Χ.

Precision= Ποσοστό από τα παραδείγματα που έχει η κλάση Χ και είναι αληθές/ σύνολο της κλάσης Χ. Δηλαδή πόσα ερωτηματολόγια μπήκαν σωστά στην κλάση δια πόσα μπήκαν στην κλάση γενικά

Για παράδειγμα στον Bayes έχουμε το εξής αποτέλεσμα ορθών και λανθασμένων προβλέψεων :

```
  a   b   c   <-- classified as
 184  36  31 |   a = 1
  36  83  16 |   b = 2
  49  26  16 |   c = 3
```

Το precision της κλάσης a είναι:

(Ποσοστό από τα παραδείγματα που έχει η κλάση a και είναι αληθές(184) / σύνολο της κλάσης a(184+36+49))

184/269=0.684

F-measure

Είναι ένα μέτρο συνδυασμού του recall και του precision.

F-measure=2*Precision*Recall / (Precision + Recall).

Το F-measure της κλάσης a είναι:

2*0.684*0.733/(0.684+0.733)=1.0027/1.417=0.708

4.2 Δημογραφικά χαρακτηριστικά και ερωτήσεις πλατφόρμας

	Ορθότητα (Correctly)	**Αληθώς θετικό (TP)**	**Ψευδώς θετικό (FP)**	**Ακρίβεια (Precision)**	**Ανάκληση (Recall)**	**F-measure**
Bayes	59.33%	0.59	0.27	0.57	0.59	0.58
BayesNet	63.31%	0.63	0.3	0.59	0.63	0.59
DTNB	62.05%	0.62	0.29	0.58	0.62	0.59
J48	52.62%	0.53	0.31	0.52	0.53	0.52
K-Star	54.3%	0.54	0.27	0.56	0.54	0.55
Naïve Bayes Simple	57.02%	0.57	0.27	0.58	0.57	0.57
JRip	61%	0.61	0.34	0.56	0.61	0.56
Simple Cart	63.31%	0.63	0.28	0.52	0.63	0.57
K-Means	38.15%	-	-	-	-	-
SMO	62.68%	0.63	0.30	0.59	0.63	0.60

*Τα αποτελέσματα αυτά αποτελούν τον σταθμισμένο μέσο όρο των κλάσεων a,b και c.

4.3 Δημογραφικά Χαρακτηριστικά

	Ορθότητα (Correctly)	**Αληθώς θετικό (TP)**	**Ψευδώς θετικό (FP)**	**Ακρίβεια (Precision)**	**Ανάκληση (Recall)**	**F-measure**
Bayes	48.00%	0.48	0.45	0.40	0.48	0.42
BayesNet	52.62%	0.53	0.53	0.28	0.53	0.36
DTNB	51.36%	0.51	0.51	0.36	0.51	0.39
J48	50.73%	0.50	0.36	0.47	0.50	0.48
K-Star	48.43%	0.48	0.38	0.45	0.48	0.46
Naïve Bayes Simple	47%	0.47	0.43	0.42	0.47	0.43
JRip	51.36%	0.51	0.47	0.39	0.51	0.42
Simple Cart	50.31%	0.50	0.51	0.40	0.50	0.38
K-Means	38.15%	-	-	-	-	-
SMO	52.62%	0.53	0.53	0.28	0.53	0.36

*Τα αποτελέσματα αυτά αποτελούν τον σταθμισμένο μέσο όρο των κλάσεων a,b και c.

4.4 Ερωτήσεις Πλατφόρμας

	Ορθότητα (Correctly)	**Αληθώς θετικό (TP)**	**Ψευδώς θετικό (FP)**	**Ακρίβεια (Precision)**	**Ανάκληση (Recall)**	**F-measure**
Bayes	63.10%	0.63	0.25	0.61	0.63	0.62
BayesNet	63.31%	0.63	0.3	0.59	0.63	0.59
DTNB	65%	0.65	0.27	0.61	0.65	0.62
J48	53.9%	0.53	0.30	0.53	0.54	0.54
K-Star	56.39%	0.56	0.23	0.60	0.56	0.57
Naïve Bayes Simple	62.68%	0.62	0.25	0.61	0.63	0.62
JRip	59.33%	0.59	0.35	0.54	0.59	0.55
Simple Cart	63.31%	0.63	0.28	0.52	0.63	0.57
K-Means	52.62%	-	-	-	-	-
SMO	62.68%	0.63	0.30	0.60	0.63	0.60

*Τα αποτελέσματα αυτά αποτελούν τον σταθμισμένο μέσο όρο των κλάσεων a,b και c.

Κεφάλαιο 5: Συμπεράσματα

Το σημαντικότερο κριτήριο αξιολόγησης των αλγορίθμων είναι το ποσοστό ορθότητας ή αλλιώς ο μέσος όρος ανάκλησης. Αρχικά, μας ενδιαφέρει ο αλγόριθμος να κατορθώνει να προβλέπει σωστά την πρόθεση ψήφου από το κάθε ερωτηματολόγιο και έπειτα, άλλα κριτήρια, όπως για παράδειγμα ο μέσος όρος ακρίβειας του αλγορίθμου. Δηλαδή πόσα ερωτηματολόγια επέλεξε ορθά ο αλγόριθμος για να εισάγει στο σύνολο ορθότητας.

Όσον αφορά τις κατηγορίες που είναι χωρισμένα τα δεδομένα, οι αλγόριθμοι εξάγουν σχεδόν το ίδιο αποτέλεσμα. Γενικά όμως, το μεγαλύτερο ποσοστό ορθότητας το δίνουν οι αλγόριθμοι Bayes Net , Simple Cart και DTNB. Στην πρώτη κατηγορία που είναι όλα τα δεδομένα, ερωτήσεις και δημογραφικά χαρακτηριστικά, οι καλύτεροι αλγόριθμοι θεωρούνται οι Bayes Net και Simple Cart, αφού έχουν ποσοστό ορθότητας και οι δύο 63.3%. Στην δεύτερη κατηγορία όπου χρησιμοποιούνται μόνο τα δημογραφικά χαρακτηριστικά των ερωτώμενων, οι καλύτεροι σε ορθότητα αλγόριθμοι είναι οι Bayes Net και SMO, με ποσοστό 52.6% και οι δύο. Στην τελευταία κατηγορία όπου εξάγονται αποτελέσματα με βάση τις ερωτήσεις μόνο, ο καλύτερος αλγόριθμος με διαφορά 2% από τον δεύτερο είναι ο DTNB με ποσοστό 65%. Με αυτό συμπεραίνεται λοιπόν ότι από όλες τις κατηγορίες ο καλύτερος αλγόριθμος είναι ο DTNB. Στις άλλες δύο κατηγορίες, πάλι έχει ένα υψηλό ποσοστό ορθότητας.

Εκτός από το ποσοστό ορθότητας των αλγορίθμων, σημαντικό ρόλο παίζει και ο μέσος όρος ακρίβειας των αλγορίθμων. Οι αλγόριθμοι J48, Naïve Bayes Simple και K-Star, δίνουν υψηλό ποσοστό ακρίβειας. Ο αλγόριθμος K-Star, έχει το υψηλότερο ποσοστό ακρίβειας από όλους τους αλγορίθμους. Το ποσοστό ανάκλησης του αλγορίθμου αυτού είναι 0.56 και ο μέσος όρος ακρίβειας είναι 0.59. Σε αυτό το παράδειγμα φαίνεται το ποσοστό ακρίβειας να είναι μεγαλύτερο από το ποσοστό ανάκλησης. Αυτό σημαίνει ότι ο αλγόριθμος K-Star, πρόβλεψε λιγότερους ψήφους, αλλά από αυτούς που πρόβλεψε δεν υπήρχαν πολλά περιττά ερωτηματολόγια στην κλάση, δηλαδή πρόβλεψε ορθά περισσότερους ψήφους. Οι J48 και Naïve Bayes Simple αλγόριθμοι έχουν ίδιο ποσοστό ορθότητας και ποσοστό ακρίβειας. Οι αλγόριθμοι αυτοί φαίνεται να είναι ισορροπημένοι αφού δεν κάνει πολλές λάθος προβλέψεις ούτε εισάγει λάθος ερωτηματολόγια στην κλάση. Συγκεκριμένα ο αλγόριθμος Naïve Bayes Simple, δεν παύει να είναι αρκετά

αποτελεσματικός, αφού δίνει μέσο όρο ακρίβειας 0.61 και ποσοστό ορθότητας 0.62. Άρα, και αυτός ο αλγόριθμος μπορεί να θεωρηθεί από τους καλύτερους στο θέμα ακρίβειας.

Οι υπόλοιποι αλγόριθμοι που δεν αναφέρονται σε αυτή την ενότητα, έχουν καλό μέσο όρο ακρίβειας αλλά όχι τόσο, όσο τους προαναφερθέντες αλγορίθμους. Έχουν υψηλό ποσοστό ορθότητας αλλά χαμηλότερο μέσο όρο ακρίβειας, που σημαίνει πώς εισάγει αρκετά λάθος ερωτηματολόγια μέσα στην κλάση.

Ένα άλλο σημαντικό συμπέρασμα, είναι ότι τα καλύτερα αποτελέσματα βγαίνουν από την κατηγορία όπου χρησιμοποιούνται μόνο οι ερωτήσεις των ερωτηματολογίων, αφού έχει υψηλά ποσοστά ορθότητας σε όλους τους αλγορίθμους. Συμπεραίνεται λοιπόν ότι κάνει πολλές σωστές προβλέψεις ψήφου.

Όσον αφορά τον αλγόριθμο K-means, στην πρώτη περίπτωση, όπως φαίνεται από τα αποτελέσματα χωρίζει τα δεδομένα σε τρείς ομάδες. Στην πρώτη ομάδα έχουν εισαχθεί 59 άτομα από τα 477. Στην δεύτερη ομάδα 216 και στην τρίτη 202. Τα άτομα αυτά έχουν επιλεχθεί με βάση κάποια κέντρα που καθορίζει το weka. Οι ομάδες αυτές αποτελούνται από ερωτηματολόγια που έχουν ομοιότητες μεταξύ τους, αλλά δεν έχουν απαραίτητα την ίδια πρόθεση ψήφου 2013. Το ποσοστό των λανθασμένων στιγμιότυπων ομαδοποίησης είναι 61.8449%. Από αυτό διαφαίνεται ότι δεν έχει κάνει και την καλύτερη ομαδοποίηση των στιγμιότυπων στις ομάδες τις οποίες έπρεπε να ανήκουν. Στην δεύτερη περίπτωση, χρησιμοποιεί τα δημογραφικά χαρακτηριστικά μόνο και έχει το ίδιο ποσοστό λανθασμένων στιγμιότυπων ομαδοποίησης, όπως και στο προηγούμενο παράδειγμα. Σε αυτή την περίπτωση διαφέρουν μόνο τα άτομα που έχουν εισαχθεί στις ομάδες. Στην πρώτη ομάδα έχουν εισαχθεί 234 άτομα. Στην δεύτερη ομάδα 149 και στην τρίτη 94. Στην τελευταία περίπτωση, στην πρώτη ομάδα έχουν εισαχθεί 92 άτομα. Στην δεύτερη ομάδα 125 και στην τρίτη 260 άτομα. Το ποσοστό λανθασμένων στιγμιότυπων είναι 47.3795. Στη συγκεκριμένη περίπτωση ο K-means έχει ομαδοποιήσει αρκετά καλά τα δεδομένα.

Συμπεραίνουμε λοιπόν ότι ο αλγόριθμος K-means είναι πιο αποδοτικός όταν τρέχει μόνο με τις ερωτήσεις που περιέχει η πλατφόρμα. Γενικά όμως, έχει χαμηλή απόδοση επειδή είναι αλγόριθμος ομαδοποίησης και έχει διαφορετική λειτουργία από τους αλγόριθμους ταξινόμησης. Σε αντίθεση με τους αλγόριθμους ταξινόμησης που η κύρια μέθοδος

ανάλυσης είναι η πρόθεση ψήφου 2013, ο k-means που είναι αλγόριθμος ομαδοποίησης, χρησιμοποιεί πιο αυθαίρετες μεθόδους ομαδοποίησης των ερωτηματολογίων.

Βιβλιογραφία

Abernethy, M. (2010). Data mining with WEKA, Part 2: Classification and clustering. In: http://www.ibm.com/developerworks/opensource/library/os-weka2/ . Retrieved on 11.05.2010.

Aruna,S., Rajagopalan,S.P. & Nandakishore,L.V. (2011). Knowledge based analysis of various statistical tools in detecting breast. In D.C. Wyld, et al. (CS & IT 02, pp. 37–45). CCSEA Retrieved from http://airccj.org/CSCP/vol1/csit1205.pdf

Βοζαλής, Γ.Ε. (2007). Ευφυείς πράκτορες και άλλες εφαρμογές της τεχνητής νοημοσύνης στο Διαδίκτυο. Μια μελέτη αλγορίθμων συνεργατικής διήθησης για Συστήματα Συστάσεων. Διδακτορική Εργασία. Πανεπιστήμιο Μακεδονίας,Θεσσαλονίκη.

Cleary,G.J. & Trigg,E.L (1995). K*: An Instance-based Learner Using an Entropic Distance Measure. In:*12th International Conference on Machine Learning*, pp.108-114.

David, J.M. & Balakrishnan,K. (n.d.). Significance of classification techniques in prediction of learning disabilitiew. In: http://arxiv.org/ftp/arxiv/papers/1011/1011.0628.pdf.

Hall,M.& Eibe,F. (2008). Combining Naive Bayes and Decision Tables. In:*Proceedings of the 21st Florida Artificial Intelligence Society Conference (FLAIRS)*,pp.318-319, AAAI press.

Joseph, A., Riedl,J. (2012). Deconstructing Recommender Systems. How Amazon and Netflix predict your preferences and prod you to purchase. Retrieved October 2012 from http://spectrum.ieee.org/computing/software/deconstructing-recommender-systems

Κεραυνού,Ε. (2000). *Τεχνητή Νοημοσύνη και Έμπειρα Συστήματα, Τόμος Α'*. Πάτρα: Ομάδα Εκτέλεσης Έργου ΕΑΠ/1997–2001.

Κωτσιαντής, Σ.Β. (2005). Ομάδες ταξινομητών για την αύξηση της ακρίβειας των μεθόδων της μηχανικής μάθησης και εξόρυξης γνώμης. Διδακτορική Διατριβή. Πανεπιστήμιο Πατρών.

Κωτσόπουλος, Β.Δ. (2012). Προηγμένες Μέθοδοι Ταξινόμησης για την Πρόβλεψη και την Ανίχνευση Μοτίβων σε Δεδομένα Ωοπαραγωγής. Πτυχιακή Εργασία. Θεσσαλονίκη: Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης.
Ανακτήθηκε από http://vivliothmmy.ee.auth.gr/1331/1/Diploma_thesis_Kotsopoulos.pdf

Katakis, I., Tsapatsoulis, N., Triga, V., Tziouvas ,C.& Mendez, F. (n.d.). Social Voting Advice Applications-Definitions, Challenges, Datasets and Evaluation. IEEE Transactions on Multimedia

Rimey,D.R. (1992). Where to look next using a bayes net: An overview. Retrieved from http://ray.rimey.org/publications/1992_IUW.pdf

Roiger,J.R. & Geatz,W.M. (2008). Εξόρυξη Πληροφορίας-Ένας εισαγωγικός οδηγός με παραδείγματα. Ευαγγελίδης,Γ., Σαμαράς,Ν. & Δέρβος,Δ.(Επιστημονική Επιμέλεια). Μαυρόπουλος, Γ. (μτφ.). Αθήνα: Εκδόσεις Κλειδάριθμος.

Segaram, T. (2007). Making Recommendations. In *Programming Collective Intelligence* (p.7-28). Unitted States of American: O'Reilly.

Τσιράκης, Ν. (2008). Weka.[Παρουσίαση μαθήματος «Εξόρυξη Δεδομένων και Αλγόριθμοι Μάθησης», 2ο Φροντιστήριο]. Ανακτήθηκε από: http://mmlabold.ceid.upatras.gr/courses/data_mining/docs/2008/2.WEKA.pdf

Tsapatsoulis, N. ,Georgiou, O. (2012). Investigating the scalability of algorithms, the role of similarity metric and the list of suggested items construction scheme in recommender systems. *International Journal on Artificial Intelligence Tools, 21(4),* World Scientific Publishing Company.

Φεσαλίδη, Γ. (2009). Σύστημα Παραγωγής Συστάσεων Ειδήσεων. Πτυχιακή Εργασία. Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης.

Παράρτημα 1: Κωδικοποίηση ερωτήσεων και δεδομένων

Part1_Relig

	Value	
Standard Attributes	Label	Axis of religion
Labeled Values	1.00	Not Important at all
	11.00	Very Important

Part2_VoteCriterion

	Value	
Standard Attributes	Label	What is the vote criterion for your vote choice in 2013
Labeled Values	1.00	Party Leader
	2.00	Party ID
	3.00	Ability to resolve the problems
	4.00	Positions of the candidate
	5.00	Honesty
	6.00	Personal relations with candidate (team)
	7.00	Strategic Voting

q1

	Value	
Standard Attributes	Label	Fiscal deficits should be largely covered by additional taxation of wealth
Labeled Values	1.00	Completely Agree
	2.00	Agree

	|3.00 |Neither Agree, Nor Disagree
	|4.00 |Disagree
	|5.00 |Completely Disagree]

q2

		Value
Standard Attributes	Label	You need to extend the time limit of unemployment benefits even if it burdens the deficit
Labeled Values	1.00	Completely Agree
	|2.00 |Agree
	|3.00 |Neither Agree, Nor Disagree
	|4.00 |Disagree
	|5.00 |Completely Disagree]

q3

		Value
Standard Attributes	Label	The quasi-governmental agencies should be privatized regardless of whether it is profitable
Labeled Values	1.00	Completely Agree
	|2.00 |Agree
	|3.00 |Neither Agree, Nor Disagree
	|4.00 |Disagree
	|5.00 |Completely Disagree]

q4

	Value	
Standard Attributes	Label	The institution of the ATA should be repealed
Labeled Values	1.00	Completely Agree
	2.00	Agree
	3.00	Neither Agree, Nor Disagree
	4.00	Disagree
	5.00	Completely Disagree]

q5

	Value	
Standard Attributes	Label	Labor rights of public employees should be equated with the rights of private sector employees
Labeled Values	1.00	Completely Agree
	2.00	Agree
	3.00	Neither Agree, Nor Disagree
	4.00	Disagree
	5.00	Completely Disagree]

q6

	Value	

Standard Attributes	Label	Allowances of political refugees should be cut
Labeled Values	1.00	Completely Agree
	2.00	Agree
	3.00	Neither Agree, Nor Disagree
	4.00	Disagree
	5.00	Completely Disagree]

q7

|---|---|---|
	Value
Standard Attributes	Label	The increase in unemployment is mainly due to the uncontrolled influx of foreign (EU and non) workers
Labeled Values	1.00	Completely Agree
	2.00	Agree
	3.00	Neither Agree, Nor Disagree
	4.00	Disagree
	5.00	Completely Disagree]

q8

|---|---|---|
	Value
Standard Attributes	Label	The securities holders should be compensated for the full value

|---|---|---|
Labeled Values	1.00	Completely Agree
	2.00	Agree
	3.00	Neither Agree, Nor Disagree
	4.00	Disagree
	5.00	Completely Disagree]

q9

|---|---|---|
	Value
Standard Attributes	Label	You should limit access to regulated T / K in free medical care unless they are residents of the free areas
Labeled Values	1.00	Completely Agree
	2.00	Agree
	3.00	Neither Agree, Nor Disagree
	4.00	Disagree
	5.00	Completely Disagree]

q10

|---|---|---|
| |Value |
|Standard Attributes|Label|A bi-zonal bi-communal federation on Cyprus will be sustainable|

|---|---|---|
Labeled Values	1.00	Completely Agree
	2.00	Agree
	3.00	Neither Agree, Nor Disagree
	4.00	Disagree
	5.00	Completely Disagree]

q11

|---|---|---|
	Value
Standard Attributes	Label	The new President of the Republic should be bound by previous agreements in Negotiations on Cyprus
Labeled Values	1.00	Completely Agree
	2.00	Agree
	3.00	Neither Agree, Nor Disagree
	4.00	Disagree
	5.00	Completely Disagree]

q12

|---|---|---|
| |Value |

|---|---|---|
Standard Attributes	Label	Cyprus should raise an abolition of the British bases before a comprehensive settlement of the Cyprus problem
Labeled Values	1.00	Completely Agree
	2.00	Agree
	3.00	Neither Agree, Nor Disagree
	4.00	Disagree
	5.00	Completely Disagree]

q13

|---|---|---|
	Value	
Standard Attributes	Label	The closing of the barricades should be used as leverage to solve the Cyprus problem
Labeled Values	1.00	Completely Agree
	2.00	Agree
	3.00	Neither Agree, Nor Disagree
	4.00	Disagree
	5.00	Completely Disagree]

q14

|---|---|
| |Value|

|---|---|---|
Standard Attributes	Label	Military service should be reduced to 18 months
Labeled Values	1.00	Completely Agree
	---	---
	2.00	Agree
	---	---
	3.00	Neither Agree, Nor Disagree
	---	---
	4.00	Disagree
	---	---
	5.00	Completely Disagree]

q15

|---|---|---|
	Value
Standard Attributes	Label	The crime will be tackled effectively if the limited number of non-EU migrants
Labeled Values	1.00	Completely Agree
	---	---
	2.00	Agree
	---	---
	3.00	Neither Agree, Nor Disagree
	---	---
	4.00	Disagree
	---	---
	5.00	Completely Disagree]

q16

|---|---|---|
	Value
Standard Attributes	Label	The role of the church should be focused on spiritual matters rather than matters of general policy of the State

Labeled Values	1.00	Completely Agree
	2.00	Agree
	3.00	Neither Agree, Nor Disagree
	4.00	Disagree
	5.00	Completely Disagree]

q17

	Value
Standard Attributes	Label	Possession of soft drugs (ie marijuana) for personal use should be decriminalized
Labeled Values	1.00	Completely Agree
	2.00	Agree
	3.00	Neither Agree, Nor Disagree
	4.00	Disagree
	5.00	Completely Disagree]

q18

	Value
Standard Attributes	Label	Should be institutionalized in a civil partnership same-sex couples
Labeled Values	1.00	Completely Agree
	2.00	Agree

		|3.00 |Neither Agree, Nor Disagree
		|4.00 |Disagree
		|5.00 |Completely Disagree]

q19

			Value
Standard Attributes	Label	Should be allowed to establish and operate casinos		
Labeled Values		|1.00 |Completely Agree
		|2.00 |Agree
		|3.00 |Neither Agree, Nor Disagree
		|4.00 |Disagree
		|5.00 |Completely Disagree]

q20

		|Value
Standard Attributes	Label	RECODE Multicultaralism positive	

q21

		|Value
Standard Attributes	Label	RECODE Banks responsability	

q22

	Value	
Standard Attributes	Label	The taxation of the residence is the first necessary steps to resolve the economic crisis
Labeled Values	1.00	Completely Agree
	2.00	Agree
	3.00	Neither Agree, Nor Disagree
	4.00	Disagree
	5.00	Completely Disagree]

q23

	Value	
Standard Attributes	Label	The next government should renegotiate the Memorandum on benefits cuts
Labeled Values	1.00	Completely Agree
	2.00	Agree
	3.00	Neither Agree, Nor Disagree
	4.00	Disagree
	5.00	Completely Disagree]

q24

	Value	

Standard Attributes	Label	Should pagopoiithoun promotions of civil servants for savings in the state budget
Labeled Values	1.00	Completely Agree
	2.00	Agree
	3.00	Neither Agree, Nor Disagree
	4.00	Disagree
	5.00	Completely Disagree]

q25

Value	
Standard Attributes	Label	To address the crisis necessary cuts to pensions
Labeled Values	1.00	Completely Agree
	2.00	Agree
	3.00	Neither Agree, Nor Disagree
	4.00	Disagree
	5.00	Completely Disagree]

q26

Value	
Standard Attributes	Label	Should be applied directly GHS merging private and public hospitals as this will lead to providing better care
Labeled Values	1.00	Completely Agree

| |-----|----------
----------|
| |2.00 |Agree |
| |-----|----------
----------|
| |3.00 |Neither Agree, Nor Disagree |
| |-----|----------
----------|
| |4.00 |Disagree |
| |-----|----------
----------|
| |5.00 |Completely Disagree] |
|---|-----|----------
---|

q27

|---|---|
	Value
Standard Attributes	Label	To tackle the crisis should be layoffs in the public sector
Labeled Values	1.00	Completely Agree
	-----	---
	2.00	Agree
	-----	---
	3.00	Neither Agree, Nor Disagree
	-----	---
	4.00	Disagree
	-----	---
	5.00	Completely Disagree]

q28

|---|---|
	Value
Standard Attributes	Label	The exploitation rights plot 12 is a safe solution for debt repayment
Labeled Values	1.00	Completely Agree
	-----	---
	2.00	Agree
	-----	---

		3.00	Neither Agree, Nor Disagree
		4.00	Disagree
		5.00	Completely Disagree]

q29

			Value
Standard Attributes	Label	The economy works best as it does not interfere the state and as much freedom do companies		
Labeled Values	1.00	Completely Agree
		2.00	Agree
		3.00	Neither Agree, Nor Disagree
		4.00	Disagree
		5.00	Completely Disagree]

q30

			Value
Standard Attributes	Label	The relations between Cyprus and Russia should strengthen		
Labeled Values	1.00	Completely Agree
		2.00	Agree
		3.00	Neither Agree, Nor Disagree
		4.00	Disagree

| | |-----|----------|
	|5.00 |Completely Disagree]	

q31

|---|---|
	|Value	
Standard Attributes	Label	The division of existing ministries should be changed with the creation of State Secretaries (eg environment, development, etc) even if it means higher government officials
Labeled Values |1.00 |Completely Agree	
	|-----|----------
	|2.00 |Agree	
	|-----|----------
	|3.00 |Neither Agree, Nor Disagree	
	|-----|----------
	|4.00 |Disagree	
	|-----|----------
	|5.00 |Completely Disagree]	

q32

|---|---|
	Value	
Standard Attributes	Label	Justice is justly attributed the responsibility for the tragedy of July 11, 2011

|---|---|---|
Labeled Values	1.00	Completely Agree
	2.00	Agree
	3.00	Neither Agree, Nor Disagree
	4.00	Disagree
	5.00	Completely Disagree]

q33

|---|---|---|
	Value
Standard Attributes	Label	The relations between Cyprus and Israel should be extended to areas other than the exploitation of hydrocarbons
Labeled Values	1.00	Completely Agree
	2.00	Agree
	3.00	Neither Agree, Nor Disagree
	4.00	Disagree
	5.00	Completely Disagree]

q34

|---|---|---|
| |Value |

Standard Attributes	Label	The introduction of tuition fees in public universities will improve the quality of education
Labeled Values	1.00	Completely Agree
	2.00	Agree
	3.00	Neither Agree, Nor Disagree
	4.00	Disagree
	5.00	Completely Disagree]

q35

	Value
Standard Attributes	Label	Cyprus should consider leaving the euro zone if the imposed strict austerity measures
Labeled Values	1.00	Completely Agree
	2.00	Agree
	3.00	Neither Agree, Nor Disagree
	4.00	Disagree
	5.00	Completely Disagree]

Part4_Sex

	Value
Standard Attributes	Label	Sex
Labeled Values	1.00	Female
	2.00	Male

|---|---|---|

Part4_DateBirth

	Value	
Standard Attributes	Label	Date of Birth

Part4_Educ

		Value	
Standard Attributes	Label	Level of Education	
Labeled Values	1.00	Primary School	
	2.00	Gymnasium (High School)	
	3.00	Lyceum	
	4.00	University/Public School	
	5.00	Post-Graduate Studies	
	6.00	"illiterate(no education)"	

Part4_Immigrant

		Value	
Standard Attributes	Label	Are you Immigrant?	
Labeled Values	1.00	Yes	
	2.00	No	

Part4_Income

	Value	
Standard Attributes	Label	Describe the current status of your income
Labeled Values	1.00	Comfortable
	2.00	Sufficient
	3.00	With Difficulty
	4.00	Very difficult

Part4_Occupation

	Value	
Standard Attributes	Label	Current Job/Occupation
Labeled Values	1.00	Employee (salary)
	2.00	Self-employed
	3.00	Vocational Training (unpaid)
	4.00	Unemployed (looking for a job)
	5.00	Unemployed (not actively searching)
	6.00	Chronic Illness (handicapo
	7.00	In pension
	8.00	Military Service
	9.00	Household
	10.00	Other

Part2_Vote2013

	Value	
Standard Attributes	Label	What will you vote i nthe Presidential Elections of 2013
Labeled Values	1.00	Nikos Anastasiadis
	2.00	Stavros Malas
	3.00	Giorgos Lilikas
	4.00	BLANK/INVALID
	5.00	I have not decided yet
	6.00	I will abstain

Παράρτημα 2: Αναλυτικά δεδομένα αποτελεσμάτων σε μορφή Weka

Δημογραφικά χαρακτηριστικά και ερωτήσεις πλατφόρμας

Αλγόριθμος Bayes

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	283	59.3291 %
Incorrectly Classified Instances	194	40.6709 %
Kappa statistic	0.313	
Mean absolute error	0.2942	
Root mean squared error	0.4357	
Relative absolute error	72.6895 %	
Root relative squared error	96.8818 %	
Total Number of Instances	477	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.733	0.376	0.684	0.733	0.708	0.744	1
	0.615	0.181	0.572	0.615	0.593	0.827	2
	0.176	0.122	0.254	0.176	0.208	0.601	3
Weighted Avg.	0.593	0.272	0.57	0.593	0.58	0.74	

=== Confusion Matrix ===

a	b	c	<-- classified as
184	36	31	a = 1
36	83	16	b = 2
49	26	16	c = 3

Αλγόριθμος BayesNet

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	302	63.3124 %
Incorrectly Classified Instances	175	36.6876 %
Kappa statistic	0.346	
Mean absolute error	0.2934	
Root mean squared error	0.4089	
Relative absolute error	72.5001 %	
Root relative squared error	90.94 %	
Total Number of Instances	477	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.829	0.469	0.662	0.829	0.736	0.763	1
	0.637	0.164	0.606	0.637	0.621	0.819	2
	0.088	0.034	0.381	0.088	0.143	0.643	3
Weighted Avg.	0.633	0.3	0.593	0.633	0.59	0.756	

=== Confusion Matrix ===

a	b	c	<-- classified as
208	34	9	a = 1
45	86	4	b = 2
61	22	8	c = 3

Αλγόριθμος DTNB

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	296	62.0545 %
Incorrectly Classified Instances	181	37.9455 %
Kappa statistic	0.3377	
Mean absolute error	0.3128	
Root mean squared error	0.4126	
Relative absolute error	77.2855 %	
Root relative squared error	91.7533 %	
Total Number of Instances	477	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.789	0.438	0.667	0.789	0.723	0.75	1
	0.652	0.158	0.62	0.652	0.635	0.804	2
	0.11	0.073	0.263	0.11	0.155	0.619	3
Weighted Avg.	0.621	0.289	0.576	0.621	0.59	0.74	

=== Confusion Matrix ===

a	b	c	<-- classified as
198	31	22	a = 1
41	88	6	b = 2
58	23	10	c = 3

Αλγόριθμος J48

Time taken to build model: 1.14 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	251	52.6205 %
Incorrectly Classified Instances	226	47.3795 %
Kappa statistic	0.2124	
Mean absolute error	0.3259	
Root mean squared error	0.5351	
Relative absolute error	80.5296 %	
Root relative squared error	118.9847 %	
Total Number of Instances	477	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.653	0.42	0.633	0.653	0.643	0.602	1
	0.533	0.175	0.545	0.533	0.539	0.709	2
	0.165	0.184	0.174	0.165	0.169	0.468	3
Weighted Avg.	0.526	0.306	0.521	0.526	0.523	0.607	

=== Confusion Matrix ===

a	b	c	<-- classified as
164	40	47	a = 1
39	72	24	b = 2
56	20	15	c = 3

Αλγόριθμος K-Star

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	259	54.2977 %
Incorrectly Classified Instances	218	45.7023 %
Kappa statistic	0.2629	
Mean absolute error	0.3067	
Root mean squared error	0.5125	
Relative absolute error	75.7805 %	
Root relative squared error	113.9744 %	
Total Number of Instances	477	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.629	0.345	0.669	0.629	0.649	0.711	1
	0.504	0.152	0.567	0.504	0.533	0.762	2
	0.363	0.228	0.273	0.363	0.311	0.602	3
Weighted Avg.	0.543	0.268	0.565	0.543	0.552	0.705	

=== Confusion Matrix ===

a	b	c	<-- classified as
158	34	59	a = 1
38	68	29	b = 2
40	18	33	c = 3

Αλγόριθμος Naïve Bayes Simple

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	272	57.0231 %
Incorrectly Classified Instances	205	42.9769 %
Kappa statistic	0.2955	
Mean absolute error	0.3033	
Root mean squared error	0.4497	
Relative absolute error	74.9426 %	
Root relative squared error	99.9982 %	
Total Number of Instances	477	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.661	0.358	0.672	0.661	0.667	0.727	1
	0.6	0.149	0.614	0.6	0.607	0.817	2
	0.275	0.189	0.255	0.275	0.265	0.596	3
Weighted Avg.	0.57	0.267	0.576	0.57	0.573	0.728	

=== Confusion Matrix ===

a	b	c	<-- classified as
166	31	54	a = 1
35	81	19	b = 2
46	20	25	c = 3

Αλγόριθμος JRip

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	291	61.0063 %
Incorrectly Classified Instances	186	38.9937 %
Kappa statistic	0.286	
Mean absolute error	0.3307	
Root mean squared error	0.4378	
Relative absolute error	81.7063 %	
Root relative squared error	97.3527 %	
Total Number of Instances	477	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.849	0.562	0.626	0.849	0.721	0.651	1
	0.548	0.146	0.597	0.548	0.571	0.71	2
	0.044	0.023	0.308	0.044	0.077	0.544	3
Weighted Avg.	0.61	0.342	0.557	0.61	0.556	0.647	

=== Confusion Matrix ===

a	b	c	<-- classified as
213	29	9	a = 1
61	74	0	b = 2
66	21	4	c = 3

Αλγόριθμος Simple Cart

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	302	63.3124 %
Incorrectly Classified Instances	175	36.6876 %
Kappa statistic	0.3536	
Mean absolute error	0.3464	
Root mean squared error	0.4176	
Relative absolute error	85.5832 %	
Root relative squared error	92.8684 %	
Total Number of Instances	477	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.801	0.407	0.686	0.801	0.739	0.659	1
	0.748	0.243	0.549	0.748	0.633	0.718	2
	0	0	0	0	0	0.478	3
Weighted Avg.	0.633	0.283	0.516	0.633	0.568	0.641	

=== Confusion Matrix ===

a	b	c	<-- classified as
201	50	0	a = 1
34	101	0	b = 2
58	33	0	c = 3

Αλγόριθμος K-Means

Number of iterations: 28
Within cluster sum of squared errors: 1261.8883298192109
Missing values globally replaced with mean/mode

Cluster centroids:

	Cluster#			
Attribute	Full Data	0	1	2
	(477)	(59)	(216)	(202)
Religion	8.6939	9.1525	8.5046	8.7624
Vote_Criterion	2.7736	2.7119	2.625	2.9505
q1	2.1342	2.1186	2.1991	2.0693
q2	2.3878	1.9322	2.412	2.495
q3	3.1195	2.3559	3.2685	3.1832
q4	3.3396	2.3729	3.412	3.5446
q5	2.2096	1.7119	2.0926	2.4802
q6	1.7547	1.9322	1.6759	1.7871
q7	1.8323	1.5932	1.8472	1.8861
q8	2.5849	1.322	2.9398	2.5743
q9	2.0503	1.6102	2.125	2.099
q10	2.7379	1.4407	2.9352	2.9059
q11	2.8616	1.7966	2.963	3.0644
g12	2.4822	1.6271	2.713	2.4851
q13	2.3187	1.5085	2.5509	2.3069
q14	2.1803	1.5932	2.3611	2.1584
q15	1.9266	1.3729	1.9861	2.0248
q16	2.1887	1.4746	2.1991	2.3861
q17	3.5996	2.2712	3.7546	3.8218
q18	3.3585	1.7119	3.7083	3.4653
q19	2.4172	2.4407	2.125	2.7228
q20	2.9518	1.9322	2.9907	3.2079
q21	2.9602	2.4576	2.9954	3.0693
q22	3.7463	2.9492	3.8565	3.8614
q23	2.2034	1.5932	2.3009	2.2772
q24	2.3187	2.1356	2.3287	2.3614
q25	3.7191	3.2034	3.7037	3.8861
q26	2	1.2373	2.0046	2.2178
q27	3.2075	2.661	3.1111	3.4703
q28	2.1845	1.322	2.3704	2.2376
q29	2.7715	1.5932	2.8056	3.0792
q30	1.9203	1.5932	1.8519	2.0891
q31	3.0901	1.5763	3.2917	3.3168
q32	3.6059	2.7797	3.7593	3.6832

q33	1.9182	1.3729	1.9074	2.0891
q34	3.434	2.1356	3.3843	3.8663
q35	3.0587	1.6441	3.4213	3.0842
Sex	1.4906	1.3051	2	1
DateBirth	1930.7652	1973.2712	1909.5787	1941.005
Educ	3.2201	3.1017	3.3194	3.1485
Immigrant	1.6415	1.678	1.588	1.6881
Income	2.4675	2.4407	2.4907	2.4505
Occupation	3.6184	3.0169	3.4352	3.9901

Time taken to build model (full training data) : 0.45 seconds

=== Model and evaluation on training set ===

Clustered Instances
0 59 (12%)
1 216 (45%)
2 202 (42%)

Class attribute: Vote2013
Classes to Clusters:

0	1	2	<-- assigned to cluster
28	114	109	| 1
21	56	58	| 2
10	46	35	| 3

Cluster 0 <-- 3
Cluster 1 <-- 1
Cluster 2 <-- 2

Incorrectly clustered instances : 295.0 61.8449 %

Αλγόριθμος SMO

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	299	62.6834 %
Incorrectly Classified Instances	178	37.3166 %

Kappa statistic	0.3412
Mean absolute error	0.3378
Root mean squared error	0.4342
Relative absolute error	83.4595 %
Root relative squared error	96.5546 %
Total Number of Instances	477

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.829	0.46	0.667	0.829	0.739	0.696	1
	0.563	0.132	0.628	0.563	0.594	0.761	2
	0.165	0.075	0.341	0.165	0.222	0.539	3
Weighted Avg.	0.627	0.294	0.594	0.627	0.599	0.684	

=== Confusion Matrix ===

a	b	c	<-- classified as
208	28	15	a = 1
45	76	14	b = 2
59	17	15	c = 3

Το ποσοστό ορθότητας είναι 62.7% και ο μέσος όρος ακρίβειας είναι 0.59.

Δημογραφικά Χαρακτηριστικά

Αλγόριθμος Bayes

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	229	48.0084 %
Incorrectly Classified Instances	248	51.9916 %
Kappa statistic	0.0353	
Mean absolute error	0.4023	
Root mean squared error	0.4765	
Relative absolute error	99.4012 %	
Root relative squared error	105.9718 %	
Total Number of Instances	477	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.753	0.726	0.535	0.753	0.626	0.547	1
	0.289	0.219	0.342	0.289	0.313	0.578	2
	0.011	0.023	0.1	0.011	0.02	0.47	3
Weighted Avg.	0.48	0.448	0.398	0.48	0.422	0.541	

=== Confusion Matrix ===

a	b	c	<-- classified as
189	56	6	a = 1
93	39	3	b = 2
71	19	1	c = 3

Αλγόριθμος BayesNet

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	251	52.6205 %
Incorrectly Classified Instances	226	47.3795 %
Kappa statistic	0	
Mean absolute error	0.4046	
Root mean squared error	0.4497	
Relative absolute error	99.9658 %	
Root relative squared error	99.9998 %	
Total Number of Instances	477	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	1	0.526	1	0.69	0.493	1
	0	0	0	0	0	0.488	2
	0	0	0	0	0	0.492	3
Weighted Avg.	0.526	0.526	0.277	0.526	0.363	0.491	

=== Confusion Matrix ===

a	b	c	<-- classified as
251	0	0	a = 1
135	0	0	b = 2
91	0	0	c = 3

Αλγόριθμος DTNB

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	245	51.3627 %
Incorrectly Classified Instances	232	48.6373 %
Kappa statistic	0.007	
Mean absolute error	0.4051	
Root mean squared error	0.4551	
Relative absolute error	100.0958 %	
Root relative squared error	101.2108 %	
Total Number of Instances	477	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.94	0.929	0.529	0.94	0.677	0.505	1
	0.067	0.061	0.3	0.067	0.109	0.521	2
	0	0.003	0	0	0	0.456	3
Weighted Avg.	0.514	0.507	0.363	0.514	0.387	0.5	

=== Confusion Matrix ===

a	b	c	<-- classified as
236	15	0	a = 1
125	9	1	b = 2
85	6	0	c = 3

Αλγόριθμος J48

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	242	50.7338 %
Incorrectly Classified Instances	235	49.2662 %
Kappa statistic	0.1369	
Mean absolute error	0.3705	
Root mean squared error	0.5109	
Relative absolute error	91.5473 %	
Root relative squared error	113.6127 %	
Total Number of Instances	477	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.737	0.535	0.605	0.737	0.664	0.577	1
	0.348	0.213	0.392	0.348	0.369	0.568	2
	0.11	0.106	0.196	0.11	0.141	0.477	3
Weighted Avg.	0.507	0.362	0.466	0.507	0.481	0.556	

=== Confusion Matrix ===

a	b	c	<-- classified as
185	47	19	a = 1
66	47	22	b = 2
55	26	10	c = 3

Αλγόριθμος K-Star

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	231	48.4277 %
Incorrectly Classified Instances	246	51.5723 %
Kappa statistic	0.0993	
Mean absolute error	0.372	
Root mean squared error	0.4877	
Relative absolute error	91.9154 %	
Root relative squared error	108.4565 %	
Total Number of Instances	477	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.705	0.571	0.578	0.705	0.636	0.599	1
	0.333	0.193	0.405	0.333	0.366	0.614	2
	0.099	0.132	0.15	0.099	0.119	0.467	3
Weighted Avg.	0.484	0.38	0.448	0.484	0.461	0.578	

=== Confusion Matrix ===

a	b	c	<-- classified as
177	45	29	a = 1
68	45	22	b = 2
61	21	9	c = 3

Αλγόριθμος Naïve Bayes Simple

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	224	46.9602 %
Incorrectly Classified Instances	253	53.0398 %
Kappa statistic	0.0429	
Mean absolute error	0.4084	
Root mean squared error	0.4828	
Relative absolute error	100.9186 %	
Root relative squared error	107.3613 %	
Total Number of Instances	477	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.725	0.681	0.542	0.725	0.62	0.521	1
	0.267	0.178	0.371	0.267	0.31	0.569	2
	0.066	0.098	0.136	0.066	0.089	0.453	3
Weighted Avg.	0.47	0.428	0.416	0.47	0.431	0.522	

=== Confusion Matrix ===

a	b	c	<-- classified as
182	46	23	a = 1
84	36	15	b = 2
70	15	6	c = 3

Αλγόριθμος JRip

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	245	51.3627 %
Incorrectly Classified Instances	232	48.6373 %
Kappa statistic	0.0492	
Mean absolute error	0.401	
Root mean squared error	0.4579	
Relative absolute error	99.0928 %	
Root relative squared error	101.825 %	
Total Number of Instances	477	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.865	0.823	0.538	0.865	0.664	0.505	1
	0.207	0.135	0.378	0.207	0.268	0.515	2
	0	0	0	0	0	0.516	3
Weighted Avg.	0.514	0.471	0.39	0.514	0.425	0.51	

=== Confusion Matrix ===

a	b	c	<-- classified as
217	34	0	a = 1
107	28	0	b = 2
79	12	0	c = 3

Αλγόριθμος Simple Cart

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	240	50.3145 %
Incorrectly Classified Instances	237	49.6855 %
Kappa statistic	-0.0085	
Mean absolute error	0.4086	
Root mean squared error	0.466	
Relative absolute error	100.9662 %	
Root relative squared error	103.6306 %	
Total Number of Instances	477	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.92	0.938	0.521	0.92	0.666	0.476	1
	0.052	0.056	0.269	0.052	0.087	0.49	2
	0.022	0.016	0.25	0.022	0.04	0.492	3
Weighted Avg.	0.503	0.512	0.398	0.503	0.383	0.483	

=== Confusion Matrix ===

a	b	c	<-- classified as
231	18	2	a = 1
124	7	4	b = 2
88	1	2	c = 3

Αλγόριθμος K-Means

Number of iterations: 7
Within cluster sum of squared errors: 152.23359140488768
Missing values globally replaced with mean/mode

Cluster centroids:

		Cluster#		
Attribute	Full Data	0	1	2
	(477)	(234)	(149)	(94)
Religion	8.6939	8.5085	8.2416	9.8723
Vote_Criterion	2.7736	2.6154	2.6779	3.3191
Sex	1.4906	2	1	1
DateBirth	1930.7652	1914.3248	1950.5302	1940.3617
Educ	3.2201	3.2821	3.6376	2.4043
Immigrant	1.6415	1.594	1.6846	1.6915
Income	2.4675	2.5214	2.1678	2.8085
Occupation	3.6184	3.3846	1.6174	7.3723

Time taken to build model (full training data) : 0.06 seconds

=== Model and evaluation on training set ===
Clustered Instances
0 234 (49%)
1 149 (31%)
2 94 (20%)

Class attribute: Vote2013
Classes to Clusters:

0	1	2	<-- assigned to cluster
121	82	48	| 1
62	44	29	| 2
51	23	17	| 3

Cluster 0 <-- 1
Cluster 1 <-- 2
Cluster 2 <-- 3

Incorrectly clustered instances : 295.0 61.8449 %

Αλγόριθμος SMO

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	251	52.6205 %
Incorrectly Classified Instances	226	47.3795 %
Kappa statistic	0	
Mean absolute error	0.3713	
Root mean squared error	0.4724	
Relative absolute error	91.7479 %	
Root relative squared error	105.049 %	
Total Number of Instances	477	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	1	0.526	1	0.69	0.5	1
	0	0	0	0	0	0.493	2
	0	0	0	0	0	0.492	3
Weighted Avg.	0.526	0.526	0.277	0.526	0.363	0.497	

=== Confusion Matrix ===

a	b	c	<-- classified as
251	0	0	a = 1
135	0	0	b = 2
91	0	0	c = 3

Ερωτήσεις Πλατφόρμας

Αλγόριθμος Bayes

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	301	63.1027 %
Incorrectly Classified Instances	176	36.8973 %
Kappa statistic	0.3743	
Mean absolute error	0.2826	
Root mean squared error	0.4154	
Relative absolute error	69.8182 %	
Root relative squared error	92.3644 %	
Total Number of Instances	477	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.773	0.358	0.705	0.773	0.738	0.771	1
	0.63	0.146	0.63	0.63	0.63	0.842	2
	0.242	0.117	0.328	0.242	0.278	0.646	3
Weighted Avg.	0.631	0.252	0.612	0.631	0.619	0.767	

=== Confusion Matrix ===

a	b	c	<-- classified as
194	26	31	a = 1
36	85	14	b = 2
45	24	22	c = 3

Αλγόριθμος BayesNet

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	302	63.3124 %
Incorrectly Classified Instances	175	36.6876 %
Kappa statistic	0.346	
Mean absolute error	0.2934	
Root mean squared error	0.4089	
Relative absolute error	72.5001 %	
Root relative squared error	90.94 %	
Total Number of Instances	477	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.829	0.469	0.662	0.829	0.736	0.763	1
	0.637	0.164	0.606	0.637	0.621	0.819	2
	0.088	0.034	0.381	0.088	0.143	0.643	3
Weighted Avg.	0.633	0.3	0.593	0.633	0.59	0.756	

=== Confusion Matrix ===

a	b	c	<-- classified as
208	34	9	a = 1
45	86	4	b = 2
61	22	8	c = 3

Αλγόριθμος DTNB

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	310	64.9895 %
Incorrectly Classified Instances	167	35.0105 %
Kappa statistic	0.3852	
Mean absolute error	0.3025	
Root mean squared error	0.4073	
Relative absolute error	74.7408 %	
Root relative squared error	90.5832 %	
Total Number of Instances	477	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.825	0.425	0.683	0.825	0.747	0.766	1
	0.667	0.14	0.652	0.667	0.659	0.815	2
	0.143	0.06	0.361	0.143	0.205	0.626	3
Weighted Avg.	0.65	0.275	0.613	0.65	0.619	0.753	

=== Confusion Matrix ===

a	b	c	<-- classified as
207	27	17	a = 1
39	90	6	b = 2
57	21	13	c = 3

Αλγόριθμος J48

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	257	53.8784 %
Incorrectly Classified Instances	220	46.1216 %
Kappa statistic	0.2325	
Mean absolute error	0.3188	
Root mean squared error	0.5243	
Relative absolute error	78.7653 %	
Root relative squared error	116.5929 %	
Total Number of Instances	477	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.661	0.42	0.636	0.661	0.648	0.614	1
	0.541	0.161	0.57	0.541	0.555	0.712	2
	0.198	0.181	0.205	0.198	0.201	0.508	3
Weighted Avg.	0.539	0.301	0.535	0.539	0.537	0.621	

=== Confusion Matrix ===

a	b	c	<-- classified as
166	33	52	a = 1
44	73	18	b = 2
51	22	18	c = 3

Αλγόριθμος K-Star

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	269	56.3941 %
Incorrectly Classified Instances	208	43.6059 %
Kappa statistic	0.3116	
Mean absolute error	0.3018	
Root mean squared error	0.4929	
Relative absolute error	74.5688 %	
Root relative squared error	109.6061 %	
Total Number of Instances	477	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.61	0.261	0.722	0.61	0.661	0.737	1
	0.563	0.181	0.551	0.563	0.557	0.767	2
	0.44	0.225	0.315	0.44	0.367	0.621	3
Weighted Avg.	0.564	0.232	0.596	0.564	0.575	0.723	

=== Confusion Matrix ===

a	b	c	<-- classified as
153	40	58	a = 1
30	76	29	b = 2
29	22	40	c = 3

Αλγόριθμος Naïve Bayes Simple

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	299	62.6834 %
Incorrectly Classified Instances	178	37.3166 %
Kappa statistic	0.3708	
Mean absolute error	0.2802	
Root mean squared error	0.4184	
Relative absolute error	69.2344 %	
Root relative squared error	93.0357 %	
Total Number of Instances	477	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.757	0.354	0.704	0.757	0.729	0.77	1
	0.644	0.143	0.64	0.644	0.642	0.841	2
	0.242	0.127	0.31	0.242	0.272	0.645	3
Weighted Avg.	0.627	0.251	0.61	0.627	0.617	0.766	

=== Confusion Matrix ===

a	b	c	<-- classified as
190	25	36	a = 1
35	87	13	b = 2
45	24	22	c = 3

Αλγόριθμος JRip

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	283	59.3291 %
Incorrectly Classified Instances	194	40.6709 %
Kappa statistic	0.2579	
Mean absolute error	0.3414	
Root mean squared error	0.4411	
Relative absolute error	84.3543 %	
Root relative squared error	98.0995 %	
Total Number of Instances	477	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.841	0.58	0.617	0.841	0.712	0.643	1
	0.496	0.117	0.626	0.496	0.554	0.702	2
	0.055	0.06	0.179	0.055	0.084	0.531	3
Weighted Avg.	0.593	0.349	0.536	0.593	0.547	0.638	

=== Confusion Matrix ===

a	b	c	<-- classified as
211	21	19	a = 1
64	67	4	b = 2
67	19	5	c = 3

Αλγόριθμος Simple Cart

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	302	63.3124 %
Incorrectly Classified Instances	175	36.6876 %
Kappa statistic	0.3536	
Mean absolute error	0.3464	
Root mean squared error	0.4176	
Relative absolute error	85.5832 %	
Root relative squared error	92.8684 %	
Total Number of Instances	477	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.801	0.407	0.686	0.801	0.739	0.659	1
	0.748	0.243	0.549	0.748	0.633	0.718	2
	0	0	0	0	0	0.478	3
Weighted Avg.	0.633	0.283	0.516	0.633	0.568	0.641	

=== Confusion Matrix ===

a	b	c	<-- classified as
201	50	0	a = 1
34	101	0	b = 2
58	33	0	c = 3

Αλγόριθμος K-Means

Number of iterations: 14
Within cluster sum of squared errors: 1025.8697717725752
Missing values globally replaced with mean/mode

Cluster centroids:

	Cluster#			
Attribute	Full Data	0	1	2
	(477)	(92)	(125)	(260)
q1	2.1342	1.6087	2.024	2.3731
q2	2.3878	2.1957	2.104	2.5923
q3	3.1195	3.7935	2.376	3.2385
q4	3.3396	3.8587	2.824	3.4038
q5	2.2096	2.2065	1.808	2.4038
q6	1.7547	2.1957	1.648	1.65
q7	1.8323	2.1413	1.544	1.8615
q8	2.5849	3.0435	1.664	2.8654
q9	2.0503	2.2609	1.528	2.2269
q10	2.7379	1.9239	2.056	3.3538
q11	2.8616	2.1522	2.464	3.3038
g12	2.4822	2.837	1.752	2.7077
q13	2.3187	3.0109	1.6	2.4192
q14	2.1803	1.8804	1.632	2.55
q15	1.9266	2.25	1.56	1.9885
q16	2.1887	1.7391	1.528	2.6654
q17	3.5996	3.6413	2.776	3.9808
q18	3.3585	3.1413	2.592	3.8038
q19	2.4172	2.5109	2.576	2.3077
q20	2.9518	2.8913	2.776	3.0577
q21	2.9602	1.9348	3.008	3.3
q22	3.7463	3.9565	3.384	3.8462
q23	2.2034	2.4239	1.656	2.3885
q24	2.3187	2.3804	1.92	2.4885
q25	3.7191	3.6739	3.552	3.8154
q26	2	1.8478	1.488	2.3
q27	3.2075	3.5326	2.888	3.2462
q28	2.1845	2.25	1.584	2.45
q29	2.7715	3.4348	1.832	2.9885
q30	1.9203	1.4891	1.872	2.0962
q31	3.0901	3.8261	1.968	3.3692
q32	3.6059	1.4022	3.776	4.3038
q33	1.9182	1.7065	1.736	2.0808
q34	3.434	3.587	2.696	3.7346
q35	3.0587	2.8043	2.208	3.5577

Time taken to build model (full training data) : 0.25 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 92 (19%)
1 125 (26%)
2 260 (55%)

Class attribute: Vote2013
Classes to Clusters:

0 1 2 <-- assigned to cluster
18 71 162 | 1
64 29 42 | 2
10 25 56 | 3

Cluster 0 <-- 2
Cluster 1 <-- 3
Cluster 2 <-- 1

Incorrectly clustered instances : 226.0 47.3795 %

Αλγόριθμος SMO

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	299	62.6834 %
Incorrectly Classified Instances	178	37.3166 %
Kappa statistic	0.3359	
Mean absolute error	0.3345	
Root mean squared error	0.4306	
Relative absolute error	82.6537 %	
Root relative squared error	95.756 %	
Total Number of Instances	477	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.833	0.487	0.655	0.833	0.733	0.684	1
	0.556	0.129	0.63	0.556	0.591	0.774	2
	0.165	0.062	0.385	0.165	0.231	0.572	3
Weighted Avg.	0.627	0.304	0.597	0.627	0.597	0.688	

=== Confusion Matrix ===

a	b	c	<-- classified as
209	27	15	a = 1
51	75	9	b = 2
59	17	15	c = 3