Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

Μεταπτυχιακή Διατριβή
στα Πληροφοριακά και Επικοινωνιακά Συστήματα

Παραγωγή Κοινωνικών Συστάσεων Σε Συστήματα Υποστήριξης Απόφασης Ψήφου

Νικόλας Χατζηνικολάου

Επιβλέπων Καθηγητής
Ιωάννης Κατάκης

Ιανουάριος 2017

Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

Παραγωγή Κοινωνικών Συστάσεων Σε Συστήματα Υποστήριξης Απόφασης Ψήφου

Νικόλας Χατζηνικολάου

Επιβλέπων Καθηγητής
Ιωάννης Κατάκης

Η παρούσα μεταπτυχιακή διατριβή υποβλήθηκε
προς μερική εκπλήρωση των απαιτήσεων για απόκτηση

μεταπτυχιακού τίτλου σπουδών
στα Πληροφοριακά Συστήματα

από τη Σχολή Θετικών και Εφαρμοσμένων Επιστημών
του Ανοικτού Πανεπιστημίου Κύπρου

Ιανουάριος 2017

Περίληψη

Είναι ευρέως αποδεκτό ότι σε περίοδο προεκλογικής εκστρατείας, είτε για προεδρικές εκλογές είτε για βουλευτικές εκλογές, γίνονται αρκετές δημοσκοπήσεις όπου ο πολίτης καλείται να απαντήσει σε διάφορες ερωτήσεις πολιτικού περιεχομένου. Από αυτές τις δημοσκοπήσεις βγαίνουν αρκετά χρήσιμα συμπεράσματα που αφορούν του πολίτες αλλά και τα πολιτικά μας κόμματα.

Στην παρούσα διατριβή θα εξετάσουμε και θα αναλύσουμε τα συστήματα «Ηλεκτρονικοί Σύμβουλοι Ψήφου» γνωστά ως «Voting Advice Applications». Οι Ηλεκτρονικοί Σύμβουλοι Ψήφου δημιουργούν ένα προφίλ απόψεων του χρήστη αφού αυτός απαντήσει σε μία σειρά από ερωτήσεις. Οι απαντήσεις που καλείται να δώσει ο χρήστης είναι συνήθως του τύπου: «Συμφωνώ πλήρως», «Συμφωνώ», «Ούτε συμφωνώ, ούτε διαφωνώ», «Διαφωνώ», «Διαφωνώ πλήρως», «Χωρίς άποψη». Χρησιμοποιώντας αυτή την πληροφορία το σύστημα υπολογίζει την ταύτιση απόψεων του χρήστη με τους εκπρόσωπους των πολιτικών κομμάτων για τους οποίους υπάρχουν επίσης τα αντίστοιχα προφίλ. Τα εν λόγω εργαλεία επεξεργάζονται τις πληροφορίες αυτές και εξάγουν κάποια αποτελέσματα όπως για παράδειγμα την ταύτιση των απόψεων του πολίτη με τους συγκεκριμένους υποψήφιους. Με αυτό τον τρόπο ο πολίτης θα έχει την δυνατότητα να συμβουλευτεί αυτά τα αποτελέσματα ούτως ώστε να καθοδηγηθεί για τον υποψήφιο που θα προτιμήσει στις εκλογές. Τέτοια συστήματα ήταν το www.choose4greece.org και το www.choose4cyprus.org.

Στο πλαίσιο της εργασίας αυτής θα εξετάσουμε διάφορους αλγόριθμους ταξινόμησης όπως είναι οι Naïve Bayes, J48, IBK, Multi Layer Perceptron και SMO, θα δημιουργήσουμε δέντρα απόφασης, θα μειώσουμε τις διαστάσεις στα δέντρα απόφασης για καλύτερη απεικόνιση των δεδομένων μας και θα εκπαιδεύσουμε τους ταξινομητές μας. Στόχος όλων αυτών είναι να εκπαιδευτούν οι αλγόριθμοι μας με το λιγότερο κόστος και η βελτίωση της αποτελεσματικότητας των υπαρχουσών τεχνικών.

Summary

During a campaign it's widely acceptable, either for presidential elections or parliamentary elections, to carry out several polls in which citizens are called upon answering random questions with political content. Useful conclusions that matter both citizens and all political parties are extracted from this kind of polls.

In this research proposition we will study and analyze the electronic voting advise systems, also known as Voting Advice Applications. These systems create a perspective profile of the user after answering a series of questions. The options given to the user are usually: "Totally agree", "Neither agree nor disagree", "Agree", "Disagree", "Totally disagree" and "No decision". Then the system will use the data from polls in order to sum up the convergence of views from each participant against those of the representatives of the political parties were respectively profiles do exist. All the information is processed by these tools and extracts results such as convergence of views of citizens with specific candidates. In this way ability is given to citizens in order to make up their mind and determine in which candidate the vote will be given during the election process. Such systems were www.choose4greece.org and www.choose4syprus.org.

In this dissertation we will examine various sorting algorithms such as Naïve Bayes, J48, IBK, Multi Layer Perceptron and SMO. It will be used decision tree analysis, a dimensional reduction on decision trees for a better visualization of the results and we will train our classifiers. The target of the above is to train our classifiers with the minimum cost and to improve the efficiency of the existing techniques.

Ευχαριστίες

Η διπλωματική αυτή εργασία κλείνει τον μεταπτυχιακό κύκλο σπουδών μου, για τον οποίο εργάστηκα σκληρά και συστηματικά.

Αρχικά θα ήθελα να ευχαριστήσω τους καθηγητές του μεταπτυχιακού προγράμματος και ιδιαίτερα τον επιβλέποντα καθηγητή μου κ. Ιωάννη Κατάκη για την αμέριστη βοήθεια, υπομονή και συνεργασία του, χωρίς την οποία η διπλωματική εργασία δε θα είχε το ίδιο καλό αποτέλεσμα.

Τέλος θα ήθελα να ευχαριστήσω τους γονείς μου για την ηθική αλλά και υλική υποστήριξη που μου παρείχαν καθ' όλη την διάρκεια των σπουδών μου προπτυχιακά και μεταπτυχιακά.

Περιεχόμενα

Κεφάλαιο 1
Εισαγωγή

Σε αυτή την διατριβή θα μελετήσουμε και θα αναλύσουμε τα συστήματα «Ηλεκτρονικοί Σύμβουλοι Ψήφου» γνωστά ως «Voting Advice Applications».

Τα συστήματα αυτά εφαρμόζονται πριν από κάθε εκλογική αναμέτρηση, Προεδρικές ή Βουλευτικές εκλογές. Τα τελευταία χρόνια ενσωματώθηκαν σε κάποιες διαδικτυακές εφαρμογές όπου ο πολίτης έχει τη δυνατότητα να απαντήσει σε αυτές τις δημοσκοπήσεις διαδικτυακά. Δυο χαρακτηριστικά παραδείγματα είναι το www.choose4greece.org και το www.choose4cyprus.org. Ο εκάστοτε ψηφοφόρος καλείται να απαντήσει σε διάφορες ερωτήσεις που αντιπροσωπεύουν διάφορα θέματα όπως η Οικονομική Κρίση, ο Δημόσιος Τομέας, το Κυπριακό, η Ασφάλεια και Μετανάστευση και Κοινωνικά θέματα . Οι απαντήσεις στην κάθε ερώτηση είναι του τύπου Συμφωνώ πλήρως», «Συμφωνώ», «Ούτε συμφωνώ, ούτε διαφωνώ», «Διαφωνώ», «Διαφωνώ πλήρως», «Χωρίς άποψη».

Τα αποτελέσματα από αυτές τις σελίδες αναλύονται και προκύπτουν κάποια στατιστικά στοιχεία που αφορούν τους υποψήφιους των κομμάτων. Εκτός από τα συνηθισμένα στατιστικά στοιχεία υπάρχει και η δυνατότητα να εμφανίζεται σαν αποτέλεσμα και το ποσοστό ταύτισης των απόψεων του πολίτη με τους συγκεκριμένους υποψήφιους για τους οποίους υπάρχουν

επίσης τα αντίστοιχα προφίλ. Με αυτό τον τρόπο ο πολίτης θα έχει τη δυνατότητα να συμβουλευτεί αυτά τα αποτελέσματα ούτως ώστε να καθοδηγηθεί για τον υποψήφιο που θα προτιμήσει στις εκλογές.

Τα προσδοκώμενα θετικά αποτελέσματα από αυτή την έρευνα είναι να παραχθεί μία μέθοδος η οποία θα παράγει συστάσεις στο χρήστη με βάση την πρόθεση ψήφου χρηστών με παρόμοιες απόψεις και όχι με βάση την ομοιότητα των προφίλ από τους εκπροσώπους των πολιτικών κομμάτων. Αυτές οι συστάσεις ονομάζονται «κοινωνικές συστάσεις» (social recommendations). Στην ερευνητική αυτή πρόταση θα αξιολογηθούν διάφοροι αλγόριθμοι ταξινόμησης με στόχο τη βελτίωση της αποτελεσματικότητας των υπαρχουσών τεχνικών. Έμφαση θα δοθεί στην online ταξινόμηση, θεωρώντας τις απαντήσεις των ψηφοφόρων ως μία συνεχή ροή δεδομένων.

Επιπλέον μέσα από αυτή την έρευνα εκπαιδεύουμε τους αλγόριθμους ταξινόμησης σε online περιβάλλον. Σε αυτό θα βοηθήσει ο μεγάλος όγκος πληροφορίας από τις διάφορες απαντήσεις των χρηστών και μέσα από αυτές τις πληροφορίες παράγουμε μία μέθοδο η οποία προβλέπει την πρόθεση ψήφου χρηστών με παρόμοιες απόψεις. Ειδικότερα, μελετούμε τις ιδιότητες των ταξινομητών που αφορούν την ταξινόμηση ροών δεδομένων. Έμφαση δίνουμε σε ερωτήματα που σχετίζονται με τον όγκο των δεδομένων εκπαίδευσης που απαιτεί ο κάθε ταξινομητής, αλλά και την ανάγκη για την τακτική επανεκπαίδευσή του. Σημαντικό αντίκτυπο στα αποτελέσματα θα έχει ο τρόπος με τον οποίο εισέρχονται οι χρήστες σε ένα Ηλεκτρονικό Σύμβουλο Ψήφου.

Οι Ηλεκτρονικοί Σύμβουλοι Ψήφου προβλέπεται να αποκτήσουν σημαντικό ρόλο στην εκλογική διαδικασία. Κατά συνέπεια, είναι αναγκαίο να μελετηθούν σε βάθος όλες οι παράμετροι των συστημάτων αυτών όπως οι κοινωνικές συστάσεις. Μέσα από αυτή την έρευνα επιδιώκεται να δημιουργηθεί μία μέθοδος η οποία θα προβλέπει την πρόθεση ψήφου του χρήστη. Η έρευνα θα επικεντρωθεί στις ήδη υπάρχουσες πληροφορίες από τους διάφορους χρήστες, ακολούθως θα γίνει ανάλυση και επεξεργασία των δεδομένων αυτών για να επιτευχθούν τα επιθυμητά αποτελέσματα.

1.1 Διάρθρωση Διατριβής

Το υπόλοιπο της διατριβής είναι διαρθρωμένο ως εξής:

Το Κεφάλαιο 2 περιγράφει τις βασικές έννοιες γύρω από τις περιοχές της μηχανικής μάθησης και τις κατηγορίες των αλγορίθμων. Γίνεται μια αναφορά στο τι είναι ταξινόμηση δεδομένων, σε ποιες φάσεις χωρίζεται (training-testing), πως κατηγοριοποιούνται τα δεδομένα μας και ποιοι αλγόριθμοι μας βοηθάνε για να κάνουμε αυτή την ταξινόμηση (Naïve Bayes, Multi Layer Perceptron, J48, SMO, IBK)

Το Κεφάλαιο 3 περιγράφει τι είναι οι Ηλεκτρονικοί Σύμβουλο Ψήφου(VAA), πως λειτουργεί ο μηχανισμός των VAA, τι αποτελέσματα εξάγονται μέσα από αυτό τον μηχανισμό και σε ποιες χώρες εφαρμόστηκε αυτό το σύστημα.

Το Κεφάλαιο 4 αναλύει τη μορφή των δεδομένων μας, από πού βρήκαμε τα δεδομένα μας, τη μορφή των ερωτήσεων που καλείται να απαντήσει ο εκάστοτε χρήστης, καθώς επίσης και τη μορφή των απαντήσεων που έχει στη διάθεση του ο χρήστης.

Το Κεφάλαιο 5 αναλύει εκτενέστερα την πειραματική αξιολόγηση. Καθορίζονται οι στόχοι της πειραματικής αξιολόγησης, τα λογισμικά προγράμματα που χρησιμοποιήθηκαν για την ανάλυση των δεδομένων, καταγράφεται ο τρόπος που έχουν επεξεργαστεί τα δεδομένα και ποιες παράμετροι καθορίστηκαν σε αυτά. Επίσης γίνεται αναφορά στο τι είναι δέντρο απόφασης, ποιο είναι το βέλτιστο δέντρο απόφασης και ποιες παραμέτρους θέτουμε σε αυτό. Επιπλέον αξιολογείται η ορθότητα των αλγορίθμων από δεδομένα άλλων χωρών. Ακολούθως εξάγονται κάποια χρήσιμα συμπεράσματα που αφορούν τη συμπεριφορά των αλγορίθμων.

Τέλος, στο Κεφάλαιο 6 γίνεται μια ανασκόπηση των σημαντικότερων σημείων που διακρίναμε σε κάθε κεφάλαιο και προτείνονται κάποιες μελλοντικές μελέτες που θα μπορούσαν να γίνουν ως επέκταση αυτής της διατριβής.

Κεφάλαιο 2
Μηχανική Μάθηση και Ταξινόμηση

2.1 Εισαγωγή στη Μηχανική Μάθηση

Η «Μηχανική Μάθηση» (machine learning) είναι μια περιοχή της τεχνητής νοημοσύνης που έχει ως στόχο την αναγνώριση και τη μελέτη προτύπων. Αφορά αλγορίθμους και μεθόδους που επιτρέπουν σε υπολογιστικά συστήματα και προγράμματα να «μαθαίνουν», να αναπτύσσονται αλλά και να προσαρμόζονται σε δεδομένα που ήδη επεξεργάζονται όπως επίσης και σε νέα δεδομένα που θα προστεθούν σε μελλοντικό στάδιο. Ο Arthur Samuel είχε δώσει σαν ορισμό ότι «η Μηχανική Μάθηση είναι ένα πεδίο μελέτης όπου δίνει σε υπολογιστές τη δυνατότητα εκμάθησης δίχως τη χρήση εις βάθους προγραμματισμού» [11]. Μέσω της ανατροφοδότησης που λαμβάνει η Μηχανική Μάθηση, καθίσταται εφικτή η κατασκευή προγραμμάτων τα οποία είναι προσαρμοσμένα (adaptable) σε υπολογιστικά συστήματα που θα λειτουργούν αναλύοντας ένα σύνολο δεδομένων [18][19][30].

Για να κατανοήσουμε καλύτερα την ιδέα της μηχανικής μάθησης ούτως ώστε να αναπτύξουμε ένα σύστημα, πρέπει να ακολουθήσουμε ένα μαθησιακό μοντέλο όπου με βάση αυτό θα γίνει η υλοποίηση. Το μοντέλο αυτό χρειάζεται να είναι αρκετά εμπλουτισμένο σε πληροφορίες για τη σωστή και γενική αντίληψη έναντι των πτυχών που έχει το πρόβλημα αλλά συνάμα να είναι απλό στην κατανόησή του. Το μοντέλο πρέπει να απαντά σε μερικά ερωτήματα του τύπου:

- Από πού και πώς θα λαμβάνει τα δεδομένα;
- Πώς τα δεδομένα θα παρουσιάζονται, όλα μαζί ή λίγα κάθε φορά;
- Τι είναι αυτό που θα διδαχθεί;
- Ποιος ο στόχος εκμάθησης μέσω αυτού του μοντέλου;

2.2 Κατηγορίες αλγορίθμων

Οι αλγόριθμοι μηχανικής μάθησης κατηγοριοποιούνται ανάλογα με το επιθυμητό αποτέλεσμα του αλγορίθμου. Οι πιο δημοφιλείς κατηγορίες είναι η «μάθηση με επίβλεψη» (supervised learning), η «μάθηση χωρίς επίβλεψη» (unsupervised learning) και η «ενισχυτική μάθηση» (reinforcement learning).

2.2.1 Μάθηση με επίβλεψη – Supervised Learning

Στη «μάθηση με επίβλεψη» ή supervised learning το σύστημα καλείται να «μάθει» μια έννοια ή συνάρτηση από ένα σύνολο δεδομένων , η οποία αποτελεί περιγραφή ενός μοντέλου. Ονομάζεται έτσι επειδή θεωρείται ότι υπάρχει κάποιος «επιβλέπων» ο οποίος παρέχει τη σωστή τιμή εξόδου της συνάρτησης για τα δεδομένα που εξετάζονται. Ο αλγόριθμος κατασκευάζει μια συνάρτηση που απεικονίζει τις τιμές δεδομένων σαν εισόδους (οι οποίες τιμές είναι γνωστές) και τις αντιστοιχεί με επιθυμητές εξόδους (σύνολο εκπαίδευσης). Κάθε τιμή εισόδου αντιστοιχεί μόνο σε μια τιμή εξόδου[12].

2.2.2 Μάθηση χωρίς επίβλεψη (Unsupervised Learning)

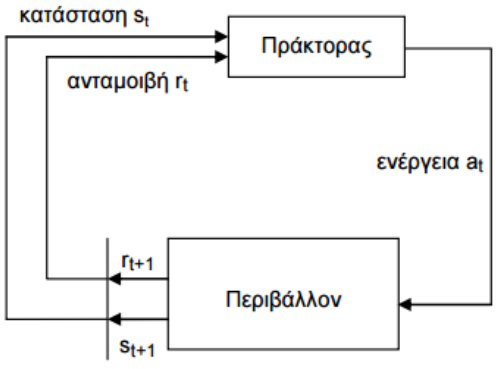
Στη «μάθηση χωρίς επίβλεψη» ή unsupervised learning το σύστημα πρέπει μόνο του να ανακαλύψει συσχετίσεις ή ομάδες σε ένα σύνολο δεδομένων, δημιουργώντας πρότυπα , χωρίς να είναι γνωστό αν υπάρχουν, πόσα και ποια είναι. Ο αλγόριθμος κατασκευάζει ένα μοντέλο για

κάποιο σύνολο εισόδων χωρίς να γνωρίζει τις επιθυμητές εξόδους για το σύνολο εκπαίδευσης [13].

2.2.3 Ενισχυτική Μάθηση (Reinforcement Learning)

Στην «ενισχυτική μάθηση» ή reinforcement learning το σύστημα μάθησης προσπαθεί να μάθει μέσω άμεσης αλληλεπίδρασης με το περιβάλλον. Ο αλγόριθμος μαθαίνει μια στρατηγική ενεργειών για μια δεδομένη παρατήρηση. Η ενισχυτική μάθηση μπορεί να εφαρμοστεί για να ελέγχει τις κινήσεις που κάνει ένα ρομπότ, για τη βελτιστοποίηση διάφορων εργασιών σε ένα εργοστάσιο, καθώς επίσης και την εκμάθηση επιτραπέζιων παιχνιδιών όπως το σκάκι. Στην ενισχυτική μάθηση το σύστημα δεν καθοδηγείται από κάποιον εξωτερικό επιβλέποντα για τις ενέργειες στις οποίες πρέπει να ακολουθήσει, αλλά πρέπει να ανακαλύψει μόνο του ποιες ενέργειες είναι αυτές που θα του αποφέρουν το μεγαλύτερο κέρδος.

Το βασικό στοιχείο της ενισχυτικής μάθησης είναι η οντότητα που μαθαίνει και παίρνει αποφάσεις. Η οντότητα αυτή ονομάζεται «πράκτορας» (agent) ενώ οτιδήποτε άλλο εκτός του πράκτορα ονομάζεται «περιβάλλον». Ο πράκτορας και το περιβάλλον αλληλεπιδρούν συνεχώς, με τον πράκτορα να επιλέγει ενέργειες και το περιβάλλον να αποκρίνεται σε αυτές και να του παρουσιάζει καινούριες καταστάσεις. Το περιβάλλον δίνει στον πράκτορα «ανταμοιβές» (rewards) οι οποίες είναι ειδικές αριθμητικές τιμές. Τις τιμές αυτές ο πράκτορας προσπαθεί να τις μεγιστοποιήσει μακροπρόθεσμα.



Εικόνα 1 - Αλληλεπίδραση Πράκτορα με το Περιβάλλον[1]

Από το πιο πάνω σχήμα βλέπουμε την αλληλεπίδραση του πράκτορα με το περιβάλλον. Ο πράκτορας και το περιβάλλον αλληλεπιδρούν σε μια ακολουθία διακριτών χρονικών στιγμών t=0,1,2,3…

Σε μια χρονική στιγμή t, ο πράκτορας λαμβάνει μια αναπαράσταση της κατάστασης του περιβάλλοντος, $S_t \in S$ όπου S είναι το σύνολο των πιθανών καταστάσεων στις οποίες μπορεί να βρεθεί πράκτορας. Ακολούθως ο πράκτορας διαλέγει μια ενέργεια $\alpha_t \in A(S_t)$ όπου $A(S_t)$ είναι το σύνολο των ενεργειών που είναι διαθέσιμες τη δεδομένη κατάσταση S_t . Την επόμενη χρονική στιγμή σαν αποτέλεσμα της ενέργειας του ο πράκτορας λαμβάνει μια αριθμητική ανταμοιβή, $r_{t+1} \in R$ και μεταβαίνει σε μια καινούργια κατάσταση S_{t+1}. Σε κάθε χρονική στιγμή ο πράκτορας πραγματοποιεί μια απεικόνιση από τις πιθανές επιλογές κάθε δυνατής ενέργειας. Η απεικόνιση αυτή ονομάζεται «πολιτική του πράκτορα» και υποδηλώνεται ως π, όπου $\pi(S_t, \alpha_t)$ είναι η πιθανότητα να επιλεγεί η ενέργεια α_t στη κατάσταση S_t . Από τα πιο πάνω προκύπτει ότι η ενισχυτική μάθηση πραγματοποιείται σε πραγματικό χρόνο μέσω της αλληλεπίδρασης του πράκτορα με το περιβάλλον [6] [14].

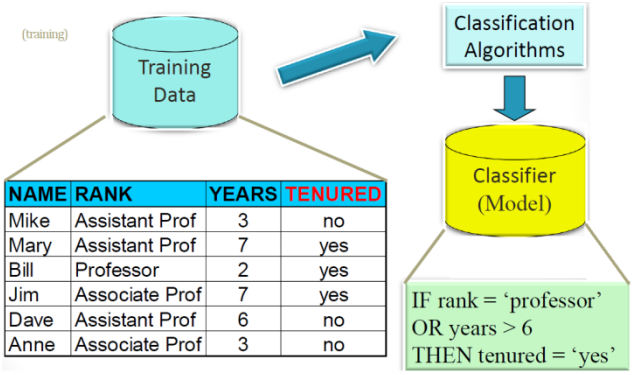
2.3 Ταξινόμηση Δεδομένων – Data Classification

Η «ταξινόμηση δεδομένων» (data classification) ή «διαχείριση δεδομένων» ως μέρος της διαχείρισης του κύκλου ζωής των πληροφοριών (ILM) μπορεί να οριστεί ως ένα εργαλείο το οποίο κατηγοριοποιεί τα δεδομένα. Με την κατηγοριοποίηση των δεδομένων αυτών μπορούμε εύκολα να βρούμε απαντήσεις σε διάφορα ερωτήματα όπως για παράδειγμα τι τύποι δεδομένων είναι διαθέσιμοι, πού βρίσκονται ορισμένα στοιχεία, ποια επίπεδα πρόσβασης υλοποιούνται, ποιο επίπεδο προστασίας υλοποιείται και εάν τηρούν τους κανόνες συμμόρφωσης [15][16].

Τα δεδομένα (data set) είναι ένα σύνολο από αντικείμενα δεδομένων (data objects). Άλλες ονομασίες για ένα αντικείμενο δεδομένων είναι η εγγραφή (record), σημείο (point), διάνυσμα (vector), πρότυπο (pattern), γεγονός (event), περίπτωση (case), δείγμα (sample), παρατήρηση (observation) ή οντότητα (entity). Ακολούθως κάθε αντικείμενο δεδομένων έχει ένα σύνολο από χαρακτηριστικά (attributes) που περιγράφουν τα βασικά χαρακτηριστικά ενός αντικειμένου. Ένα χαρακτηριστικό είναι μια ιδιότητα ή γνώρισμα ενός αντικειμένου το οποίο μπορεί να διαφέρει, είτε από ένα αντικείμενο σε ένα άλλο, είτε από μια χρονική στιγμή σε μια άλλη. Για κάθε αντικείμενο και κάθε χαρακτηριστικό υπάρχει και μια τιμή. Για κάθε αντικείμενο έχουμε και την τάξη (class) στην οποία ανήκει.

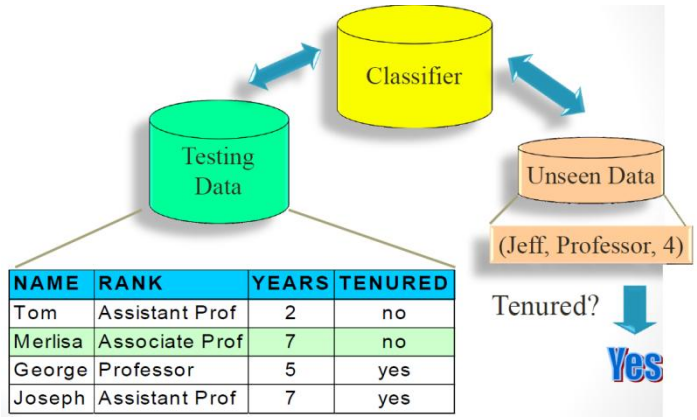
Σκοπός της ταξινόμησης είναι να κατηγοριοποιήσει τα εκάστοτε δεδομένα ανάλογα με τις ανάγκες και τις παραμέτρους που θα θέσει ο εκάστοτε χρήστης στους διάφορους ταξινομητές για την εμφάνιση συγκεκριμένων αποτελεσμάτων.

Υπάρχουν δύο φάσεις στην ταξινόμηση. Η πρώτη φάση είναι το training και η δεύτερη φάση είναι το testing. Στη φάση της εκπαίδευσης (training) ο ταξινομητής εκπαιδεύεται από τα υπάρχοντα δεδομένα. Έχοντας στη διάθεση του όλα τα δεδομένα εκπαίδευσης ο ταξινομητής μέσο ενός αλγορίθμου δημιουργεί κάποιους κανόνες ταξινόμησης. Με βάση αυτούς του κανόνες κατηγοριοποιεί τα δεδομένα σε διάφορες κλάσεις.



Εικόνα 2 – Εκπαίδευση Δεδομένων[2]

Στη φάση της αξιολόγησης (testing) ο ταξινομητής εφαρμόζει αυτά που έχει «μάθει» στην πρώτη φάση. Με βάση τους κανόνες ταξινόμησης που έχουν δημιουργηθεί και έχοντας στη διάθεση του κάποια καινούργια δεδομένα ή ακόμη και τα ίδια τα δεδομένα που χρησιμοποιήθηκαν στη φάση εκπαίδευσης προσπαθεί να προβλέψει σωστά σε ποια κλάση ανήκουν.

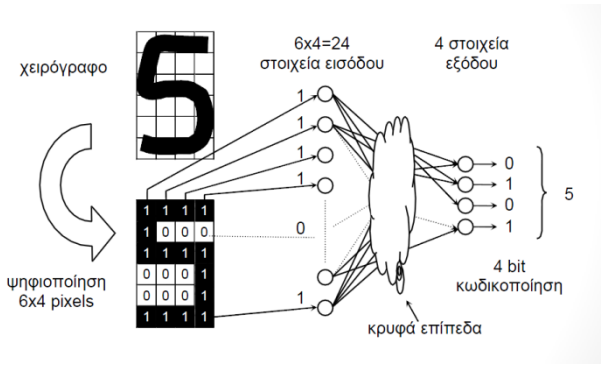


Εικόνα 3 - Αξιολόγηση Δεδομένων[2]

Οι διάφοροι μέθοδοι ταξινόμησης δεδομένων είναι τα δέντρα απόφασης ,τα νευρωνικά δίκτυα, ο ταξινομητής Bayes, η ταξινόμηση βάσει κανόνων, οι οκνηροί ταξινομητές και οι μηχανές διανυσμάτων υποστήριξης. Η κάθε μέθοδος χρησιμοποιεί και διαφορετικούς αλγόριθμους για

την ταξινόμηση των δεδομένων. Μερικοί από τους αλγόριθμους που θα αναλυθούν πιο κάτω είναι οι αλγόριθμοι Naive Bayes , Multi-Layer Perceptron, J48, SMO και ο IBk.

Αυτές οι διάφορες τεχνικές ταξινόμησης μπορούν να εφαρμοστούν σε διάφορους τομείς όπως για παράδειγμα η αναγνώριση χαρακτήρων σε ένα κείμενο. Για την αναγνώριση χαρακτήρων χρησιμοποιείται ο αλγόριθμος Multi-Layer Perceptron. Ο αλγόριθμος κάνει ψηφιοποίηση του εκάστοτε χαρακτήρα σε διάφορα pixels. Ακολούθως ελέγχει το κάθε pixel ξεχωριστά εάν έχει τιμή ή όχι. Μετά από διάφορα επίπεδα ελέγχου προκύπτει ως αποτέλεσμα η αναγνώριση του συγκεκριμένου χαρακτήρα.



Εικόνα 4 - Αναγνώριση Χαρακτήρων [1]

Διάφορα άλλα παραδείγματα εφαρμογής των διάφορων μεθόδων ταξινόμησης είναι:

i. Η ανίχνευση ανεπιθύμητων ηλεκτρονικών μηνυμάτων (spam). Βασιζόμενη στην επικεφαλίδα και το περιεχόμενο του μηνύματος κατηγοριοποιώντας τα μηνύματα σε κακόβουλα ή μη.

ii. Η κατηγοριοποίηση των κυττάρων του ανθρώπου σε καλοήθη ή κακοήθη βασιζόμενη στα αποτελέσματα εξετάσεων MRI (μαγνητική τομογραφία).

iii. Η κατηγοριοποίηση των γαλαξιών του πλανητικού μας συστήματος με βάση το σχήμα τους.

iv. Η αναγνώριση απάτης σε πιστωτικές κάρτες βρίσκοντας ποιες συναλλαγές δεν είναι από τον ιδιοκτήτη. Ο αλγόριθμος χρησιμοποιεί δεδομένα από προηγούμενες συναλλαγές της συγκεκριμένης κάρτας και πληροφορίες όπως τι αγοράζει , πότε το αγοράζει , από πού το αγοράζει και πόσο συχνά το αγοράζει. Εάν η επόμενη αγορά της πιστωτικής κάρτας δεν συνάδει ή δεν προσεγγίζει τα πιο πάνω δεδομένα τότε η συναλλαγή θεωρείται ύποπτη.

v. Να προβλέψει εάν ο επόμενος πελάτης που θα μπει στο εμπορικό κέντρο θα πληρώσει με πλαστικό χρήμα ή μετρητά. Πιθανά δεδομένα εισαγωγής στον αλγόριθμο είναι η μισθολογική κλίμακα του ατόμου, προηγούμενες του αγορές, μάρκα αυτοκινήτου και ο τόπος καταγωγής.

vi. Διαδικτυακή εφαρμογή μετάφρασης κειμένου. Οι δικτυακές εφαρμογές είναι ένα παράδειγμα μάθησης με επίβλεψη (supervised machine learning). Χρησιμοποιείται ένα σύνολο εκπαίδευσης από παραδείγματα χρησιμοποιώντας μεμονωμένες λέξεις ή ζεύγη λέξεων .

2.4 Αλγόριθμοι Ταξινόμησης

Πιο κάτω θα αναλύσουμε τις λειτουργίες και τα χαρακτηριστικά διάφορων αλγορίθμων ταξινόμησης (classifier). Οι αλγόριθμοι οι οποίοι θα αναλύσουμε εκτενέστερα είναι οι Naive Bayes , Multi-Layer Perceptron, J48, SMO και ο IBk.

2.4.1 Αλγόριθμος Naive Bayes

Η Μπεϋζιανή ταξινόμηση (Bayesian Classification) αντιπροσωπεύει μια μέθοδο επιβλεπόμενης μάθησης καθώς επίσης και μία στατιστική μέθοδο για την ταξινόμηση. Αυτή η ταξινόμηση ονομάστηκε έτσι από τον Thomas Bayes (1702-1761) ο οποίος πρότεινε το θεώρημα Bayes. Η Μπεϋζιανή ταξινόμηση παρέχει πρακτικούς αλγόριθμους εκμάθησης και την προγενέστερη γνώση καθώς επίσης και τα δεδομένα τα οποία θα παρατηρηθούν. Η Μπεϋζιανή ταξινόμηση παρέχει μια χρήσιμη προοπτική για την κατανόηση και την αξιολόγηση πολλών αλγορίθμων εκμάθησης. Υπολογίζει τις ρητές πιθανότητες για την υπόθεση του θορύβου στην είσοδο των δεδομένων [31].

Ο αλγόριθμος Naïve Bayes είναι ένας απλός πιθανολογικός ταξινομητής (probabilistic classifier) που υπολογίζει ένα σύνολο πιθανοτήτων μετρώντας τη συχνότητα και τον συνδυασμό των τιμών σε ένα σύνολο δεδομένων [32]. Ο αλγόριθμος χρησιμοποιεί το θεώρημα Bayes και υπολογίζει όλα τα χαρακτηριστικά που είναι ανεξάρτητα, λαμβάνοντας υπόψη την τιμή της μεταβλητής. Αυτή η υπό όρους προϋπόθεση ανεξαρτησίας εφαρμόζεται σπάνια στον πραγματικό κόσμο. Ο χαρακτηρισμός ως «απλοϊκός» αλγόριθμος αποδίδεται στο γεγονός ότι έχει καλή απόδοση και μαθαίνει γρήγορα σε διάφορα προβλήματα επιβλεπόμενης μάθησης

(supervised learning). Ο Bayes εφαρμόζεται στη λήψη αποφάσεων και στην επαγωγική στατιστική που ασχολείται με την εξαγωγή πιθανών συμπερασμάτων [20].

2.4.1.1 Θεώρημα Bayes

Η πιθανότητα ένα αντικείμενο d με ένα διάνυσμα x = < x1,...,xn> να ανήκει στην υπόθεση h ,

$$P(h/D) = \frac{P(D/h)\ P(h)}{P(D)}$$

P (h): Προγενέστερη πιθανότητα της υπόθεσης h

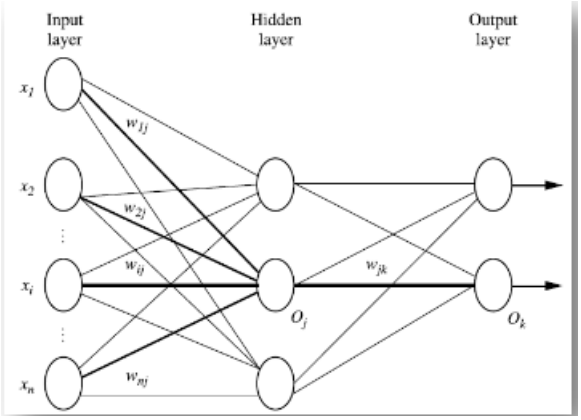
P (D): Προγενέστερη πιθανότητα των δεδομένων εκπαίδευσης xi

P (h / D): Πιθανότητα του h δεδομένου του x

P (D / h): Πιθανότητα του xi δεδομένου του h

2.4.2 Multi Layer Perceptron (MLP)

Ο Multi-Layer Perceptron (MLP) είναι ένα πολυστρωματικό νευρωνικό δίκτυο πρόσθιας τροφοδοσίας μοντέλο που χαρτογραφεί σύνολα δεδομένων εισόδου σε μια σειρά από κατάλληλες εξόδους. Ένα MLP αποτελείται από τρία ή περισσότερα στρώματα κόμβων σε ένα κατευθυνόμενο γράφημα τα οποία είναι συνδεδεμένα μεταξύ τους. Υπάρχει ένα στρώμα εισόδου, ένα στρώμα εξόδου και ενδιάμεσα υπάρχουν διάφορα κρυφά στρώματα για την επεξεργασία των χαρακτηριστικών. Εκτός από τους κόμβους εισόδου, κάθε κόμβος είναι ένας νευρώνας (ή στοιχείο επεξεργασίας) με μια μη γραμμική συνάρτηση ενεργοποίησης. Ο MLP χρησιμοποιεί τη μάθηση με επίβλεψη (supervised learning) που ονομάζεται ανάστροφη διάδοση (back propagation) για την εκπαίδευση του δικτύου. Ο MLP είναι μια τροποποίηση του γραμμικού ταξινομητή και μπορεί να διακρίνει τα δεδομένα που δεν είναι γραμμικά διαχωρίσιμα [21][33][34].



Εικόνα 5 – Multi Layer Feed Forward NN [2]

Στο στρώμα εισόδου εισάγονται τα χαρακτηριστικά και τροφοδοτούνται ταυτόχρονα στο σύστημα. Ακολούθως από το στρώμα εισόδου τροφοδοτείται η πληροφορία στα κρυφά επίπεδα για επεξεργασία. Τέλος από τα κρυφά στρώματα τροφοδοτείται η πληροφορία στα στρώματα εξόδου όπου εμφανίζεται η πρόβλεψη του δικτύου για το συγκεκριμένο χαρακτηριστικό που εισάγαμε στο στρώμα εισόδου.

Ο Multi-Layer Perceptron μπορεί να εφαρμοστεί σε διάφορους τομείς όπως είναι η αναγνώριση ομιλίας, αναγνώριση εικόνας και σε αυτόματες μηχανές μετάφρασης κειμένου.

2.4.3 Αλγόριθμος J48

Ο J48 είναι ένας ταξινομητής ο οποίος δημιουργεί ένα απλό C4.5 δέντρο απόφασης. Ο C4.5 είναι ένας αλγόριθμος που χρησιμοποιείται για να δημιουργηθεί το δέντρο απόφασης και αναπτύχθηκε από τον Ross Quinlan. Ο C.4.5 αποτελεί προέκταση του προηγούμενου αλγορίθμου ID3. Τα δέντρα απόφασης που δημιουργούνται χρησιμοποιούνται για την ταξινόμηση των δεδομένων και το δέντρο που δημιουργείται είναι δυαδικό. Τα δέντρα απόφασης είναι χρήσιμα για προβλήματα ταξινόμησης. Με την τεχνική αυτή το δέντρο κατασκευάζει και μοντελοποιεί τη διαδικασία ταξινόμησης. Από τα διάφορα δεδομένα που έχουμε τα εφαρμόζουμε στο δέντρο απόφασης που έχει δημιουργηθεί. Αναλόγως με την τιμή που έχει το κάθε δεδομένο ταξινομούνται κατάλληλα [3][4][22].

```
Algorithm [1] J48:
INPUT:
        D    //Training data
OUTPUT
        T    //Decision tree
DTBUILD (*D)
{
T=φ;
T= Create root node and label with splitting attribute;
T= Add arc to root node for each split predicate and
label;
For each arc do
        D= Database created by applying splitting
        predicate to D;
        If stopping point reached for this path, then
                T'= create leaf node and label with
                appropriate class;
        Else
                T'= DTBUILD(D);
        T= add T' to arc;
}
```

Εικόνα 6 - Ψευδοκώδικας J48 [3]

Καθώς δημιουργείται το δέντρο, ο J48 αγνοεί τις τιμές που λείπουν, δηλαδή εάν η τιμή για το συγκεκριμένο στοιχείο μπορεί να προβλεφθεί από πριν με βάσει το τι γνωρίζουμε για τις τιμές των χαρακτηριστικών που καταγράφονται. Η βασική ιδέα είναι να διαιρεθούν τα στοιχεία σε μια σειρά βασισμένη στις τιμές των χαρακτηριστικών για το συγκεκριμένο στοιχείο που βρήκαμε στα δεδομένα εκπαίδευσης. Ο J48 επιτρέπει την ταξινόμηση είτε μέσω των δέντρων απόφασης είτε των κανόνων απόφασης που παράγονται από αυτούς [5] [7].

2.4.4 SMO (Sequential minimal optimization)

Ο SMO (Διαδοχική Ελάχιστη Βελτιστοποίηση) είναι ένας αλγόριθμος για την επίλυση δευτεροβάθμιων εξισώσεων που προκύπτουν κατά τη διάρκεια της εκπαίδευσης μηχανών υποστήριξης διανυσμάτων (Support Vector Machine-SVM). Εφευρέθηκε από τον John Platt το 1998 σε μια έρευνα της Microsoft. Ο SMO χρησιμοποιείται ευρέως για να εκπαιδεύσει μηχανές υποστήριξης διανυσμάτων (SVM). Ωστόσο είναι σημαντικό να σημειωθεί ότι η μέθοδος που περιγράφεται εδώ δεν είναι εφαρμόσιμη για να όλα τα σύνολα δεδομένων.

Ο SMO είναι ένας επαναληπτικός αλγόριθμος για την επίλυση προβλημάτων. Ο SMO σπάει το πρόβλημα σε μια σειρά από μικρότερα δυνατά υπό-προβλήματα τα οποία στη συνέχεια επιλύει αναλυτικά [23][24].

2.4.5 Αλγόριθμος iBK

Ο αλγόριθμος iBK ή k-NN (k Nearest Neighbors) είναι μια μη-παραμετρική μέθοδος που χρησιμοποιείται για την ταξινόμηση (classification) και την οπισθοδρόμηση (regression)[9]. Και στις δύο περιπτώσεις η είσοδος αποτελείται από τα k πλησιέστερα παραδείγματα εκπαίδευσης των χαρακτηριστικών διανυσμάτων. Η έξοδός του εξαρτάται από το εάν ο k-NN χρησιμοποιείται από την ταξινόμηση ή την παλινδρόμηση.

Στην ταξινόμηση k-NN, η έξοδος είναι μια ιδιότητα της κλάσης. Το αντικείμενο ταξινομείται με βάση τις k πλησιέστερες γειτονικές κλάσεις. Εάν το κ=1 τότε το αντικείμενο τοποθετείται στην πιο κοντινή κλάση[39].

Συνάρτηση απόστασης	
Ευκλείδεια απόσταση	$\sqrt{\sum_{i=1}^{k}(xi - yi)^2}$
Manhattan απόσταση	$\sum_{\iota=1}^{k} \lvert xi - yi \rvert$
Minkowski απόσταση	$[\sum_{i=1}^{kn}(\lvert xi - yi \rvert)^q\ ]^{1/q}$

Στην παλινδρόμηση k-NN, η έξοδος είναι η τιμή τις ιδιότητας του αντικειμένου. Η τιμή είναι ο μέσος όρος των τιμών των k πλησιέστερων γειτόνων της κλάσης. Ο k-NN είναι ένα είδος αργής μάθησης (lazy learning) όπου η συνάρτηση προσεγγίζεται μόνο τοπικά και όλοι οι υπολογισμοί αναβάλλονται μέχρι την ταξινόμηση. Ο αλγόριθμος k-NN είναι ο πιο απλός αλγόριθμος από όλους τους αλγόριθμους που χρησιμοποιούνται στην Μηχανική Μάθηση [26][27][28][35][36].

2.5 Αξιολόγηση Αλγορίθμων

Για την σωστή αξιολόγηση οποιουδήποτε αλγορίθμου θα πρέπει πρώτα να εκπαιδεύσουμε τον ταξινομητή με διάφορα δεδομένα. Όσο πιο πολλά δεδομένα υπάρχουν στην διαδικασία εκπαίδευσης τόσο καλύτερα θα εκπαιδευτεί ο ταξινομητής. Ακολούθως εφαρμόζουμε την διαδικασία αξιολόγησης δίνοντας τα ίδια δεδομένα ή ακόμα και διαφορετικά δεδομένα από τα δεδομένα εκπαίδευσης και βλέπουμε κατά πόσο ταξινομήθηκαν σωστά, καθώς επίσης και τα ποσοστά επιτυχίας του εκάστοτε αλγορίθμου. Οι μετρικές που παίζουν σημαντικό ρόλο στο

κατά πόσο ένας αλγόριθμος είναι καλός ή όχι είναι οι μετρικές Correctly Classified , Precision , Recall και F-Measure.

2.5.1 Ορθότητα (Accuracy)

Η μετρική Accuracy εμφανίζει το ποσοστό επιτυχίας ταξινόμησης των δεδομένων από τον εκάστοτε αλγόριθμο και καθορίζεται από τον ακόλουθο τύπο:

$$accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Όπου TP= True Positive ,
TN= True Negative,
FP= False Positive,
FN= False Negative

Τα True Positive είναι τα δεδομένα στα οποία έγινε ορθή πρόβλεψη, δηλαδή ταξινομήθηκαν σωστά.
Τα True Negative είναι τα δεδομένα τα οποία σωστά δεν ταξινομήθηκαν στις θετικές προβλέψεις.
Τα False Positive είναι τα δεδομένα τα οποία λανθασμένα ταξινομήθηκαν στις θετικές προβλέψεις.
Τα False Negative είναι τα δεδομένα τα οποία δηλώθηκαν ότι απέτυχαν να ταξινομηθούν σωστά ενώ στην πραγματικότητα έγινε επιτυχής ταξινόμηση τους [29][37][38].

2.5.2 Precision

Η μετρική Precision υποδηλώνει την ακρίβεια. Δηλαδή μας δείχνει πόσες πραγματικά σωστές θετικές προβλέψεις έκανε ο ταξινομητής από το σύνολο των δεδομένων που ταξινόμησε ως σωστές θετικές προβλέψεις. Το ποσοστό μέτρησης του Precision καθορίζεται από τον πιο κάτω τύπο:

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

2.5.3 Recall

Η μετρική Recall υποδηλώνει την ανάκληση. Η ανάκληση είναι το συνολικό ποσοστό των δεδομένων που έπρεπε να προβλεφτούν σωστά από όλο των σύνολο των σωστών προβλέψεων.

$$F = 2 * \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

2.5.4 F-Measure

Η μετρική F-Measure υποδηλώνει έναν αρμονικό μέσο όρο της ακρίβειας και της ανάκλησης και δίνεται από τον πιο κάτω τύπο:

$$F = 2 * \frac{\mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

Κεφάλαιο 3
Ηλεκτρονικοί Σύμβουλοι Ψήφου(Voting Advice Applications)

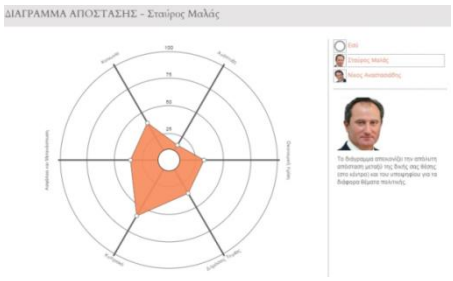
3.1 Τι είναι Ηλεκτρονικοί Σύμβουλοι Ψήφου (Voting Advice Applications)

Οι Ηλεκτρονικοί Σύμβουλοι Ψήφου (Voting Advice Applications) είναι μία διαδικτυακή εφαρμογή η οποία βοηθάει τους χρήστες να ανακαλύψουν τον βαθμό εγγύτητας των πολιτικών τους θέσεων με τις θέσεις των πολιτικών κομμάτων . Οι χρήστες καλούνται να απαντήσουν σε διάφορες ερωτήσεις. Ο αριθμός των ερωτήσεων καθώς επίσης και το περιεχόμενο των ερωτήσεων είναι διαφορετικά σε κάθε χώρα και καθορίζονται από τους εκάστοτε αναλυτές με βάση τι θέλουν να αναλύσουν, τι θέλουν να απαντηθεί, σε τι θέλουν να δώσουν έμφαση και ποια είναι τα σημαντικότερα ζητήματα στην συγκεκριμένη χώρα. Οι απαντήσεις είναι του τύπου «Συμφωνώ πλήρως», «Συμφωνώ», «Ούτε συμφωνώ, ούτε διαφωνώ», «Διαφωνώ», «Διαφωνώ πλήρως», «Χωρίς άποψη» και χωρίζονται σε έξι βαθμίδες. Μετά την συμπλήρωση του σχετικού ερωτηματολογίου οι Ηλεκτρονικοί Σύμβουλοι Ψήφου (VAA) δημιουργούν το προφίλ απόψεων

του χρήστη με βάση τις ερωτήσεις που έχει απαντήσει. Οι απαντήσεις συγκρίνονται με τις απαντήσεις που είχε δώσει το κάθε πολιτικό κόμμα ξεχωριστά. Ακολούθως παρουσιάζεται ο βαθμός συνάφειας του χρήστη με τους εκπροσώπους των πολιτικών κομμάτων. Τα αποτελέσματα εμφανίζονται σε διάφορα γραφήματα που αναπαριστούν πόσο κοντά βρίσκονται οι απόψεις τους με αυτές των κομμάτων. Πιο κάτω παρουσιάζονται μερικά παραδείγματα διαγραμμάτων από τα VAA. Τα συγκεκριμένα διαγράμματα πάρθηκαν από τις βουλευτικές εκλογές της Κύπρου το 2011.

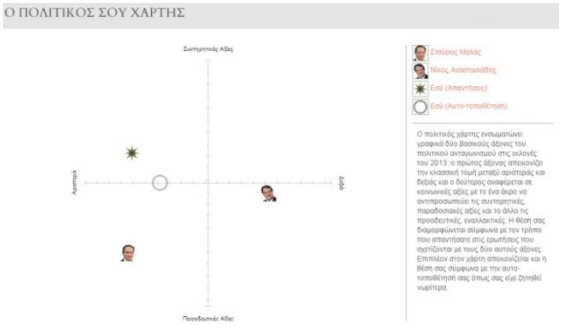
Εικόνα 7 - Ποσοστό Συνάφειας με Αρχηγούς Πολιτικών Κομμάτων

Στην Εικόνα 7 απεικονίζεται το ποσοστό συνάφειας των θέσεων του χρήστη με τις θέσεις των υποψηφίων.



Εικόνα 8 - Διάγραμμα απόστασης συγκεκριμένου υποψηφίου

Στην Εικόνα 8 απεικονίζεται η απόλυτη απόσταση μεταξύ της θέσης του ψηφοφόρου και του υποψηφίου για τα διάφορα θέματα πολιτικής όπως η Οικονομική Κρίση, ο Δημόσιος Τομέας, το Κυπριακό, η Ασφάλεια και Μετανάστευση και Κοινωνικά θέματα. Όσο μεγαλύτερος είναι ο αριθμός, τόσο μεγαλύτερη είναι η απόσταση ανάμεσα στο ψηφοφόρο και το κόμμα για το αντίστοιχο θέμα.



Εικόνα 9 - Πολιτικός Χάρτης Ψηφοφόρου

Σε αυτό το διάγραμμα απεικονίζεται ένας πολιτικός χάρτης. Ο πολιτικός χάρτης ενσωματώνει γραφικά δύο βασικούς άξονες του πολιτικού ανταγωνισμού στις εκλογές. Ο πρώτος άξονας απεικονίζει την κλασική τομή μεταξύ αριστεράς και δεξιάς. Ο δεύτερος άξονας αναφέρεται σε κοινωνικές αξίες με το ένα άκρο να αντιπροσωπεύει τις συντηρητικές, παραδοσιακές αξίες και το άλλο τις προοδευτικές, εναλλακτικές. Η θέση του χρήστη διαμορφώνεται σύμφωνα με τον τρόπο που απάντησε στις ερωτήσεις που σχετίζονται με τους δύο αυτούς άξονες. Επιπλέον στον χάρτη απεικονίζεται και η θέση του χρήστη σύμφωνα με την αυτό-τοποθέτηση του όπως του είχε ζητηθεί σε κάποια ερώτηση.

Τα δημοσιονομικά ελλείμματα πρέπει να καλυφθούν σε μεγάλο βαθμό με πρόσθετη φορολογία του πλούτου	Εσύ: Συμφωνώ πλήρως : Συμφωνώ πλήρως
Θα πρέπει να παραταθεί το χρονικό όριο του επιδόματος ανεργίας ακόμα και αν αυτό επιβαρύνει το έλλειμμα	Εσύ: Συμφωνώ : Συμφωνώ πλήρως
Οι κάτοχοι αξιογράφων θα πρέπει να αποζημιωθούν για την πλήρη αξία τους	Εσύ: Ούτε συμφωνώ, ούτε διαφωνώ : Ούτε συμφωνώ, ούτε διαφωνώ
Για την οικονομική κρίση στην Κύπρο κύρια ευθύνη έχουν οι τράπεζες παρά οι κυβερνήσεις	Εσύ: Συμφωνώ πλήρως : Συμφωνώ πλήρως
Για να αντιμετωπιστεί η κρίση είναι αναγκαίες οι περικοπές στις συντάξεις	Εσύ: Διαφωνώ πλήρως : Διαφωνώ πλήρως
Η Κύπρος θα πρέπει να εξετάσει την έξοδο από την ευρωζώνη αν της επιβάλλονται αυστηρά μέτρα λιτότητας	Εσύ: Συμφωνώ : Διαφωνώ πλήρως

Εικόνα 10 - Σύγκριση ερωτήσεων ψηφοφόρου χρήστη

Τέλος σε αυτό το διάγραμμα μπορούμε να δούμε και να συγκρίνουμε με ποιες ερωτήσεις ταυτίζεται ή διαφωνεί η άποψη του ψηφοφόρου με αυτές του εκάστοτε υποψηφίου.

Η ιδέα πίσω από τέτοιου είδους εφαρμογές είναι να επιτραπεί στους πολίτες να προσδιορίσουν καλύτερα τις δικές τους υποκειμενικές, πολιτικές προτιμήσεις. Το διαδικτυακό σύστημα τότε αντιπαραθέτει τις προτιμήσεις του πολίτη-χρήστη με τις θέσεις των κομμάτων (όπως τους έχουν αποδοθεί από ειδικούς αναλυτές). Με αυτόν τον τρόπο, επιτυγχάνεται ακόμα ένας στόχος που αφορά στο να κινητοποιηθούν οι πολίτες και να βοηθηθούν στην επιλογή της ψήφου τους μέσα

από την παρουσίαση μιας σειράς από πολιτικές επιλογές καθώς και της στάσης των κομμάτων για τα αντίστοιχα θέματα.

Τέλος τα VAA στοχεύουν στην πληρέστερη ενημέρωση των πολιτών για κρίσιμα θέματα της πολιτικής επικαιρότητας μέσα από την συνολική αλλά και λεπτομερή παρουσίαση και σύγκριση των θέσεων των κομμάτων για όλα αυτά τα θέματα. Έχουν υπάρξει μέχρι σήμερα διάφορες πρωτοβουλίες που βασίστηκαν σε τέτοιες εφαρμογές, οι οποίες είναι συχνά γνωστές και ως Ηλεκτρονικοί Σύμβουλοι Ψήφου (Voting Advice Application -VAAs), σε όλη την Ευρώπη και την Αμερική.

3.2 Μηχανισμός VAA

Για να μπορεί να λειτουργήσει σωστά ο μηχανισμός των VAA πρέπει πρώτα να έχουμε το προφίλ του εκάστοτε αρχηγού πολιτικού κόμματος ή βουλευτή και γενικότερα του εκάστοτε υποψήφιου. Πριν από κάθε εκλογική αναμέτρηση της εκάστοτε χώρας είτε είναι προεδρικές εκλογές, είτε είναι βουλευτικές, ο κάθε υποψήφιος θα πρέπει να απαντήσει ένα συγκεκριμένο ερωτηματολόγιο το οποίο θα καθοριστεί από τους διάφορους αναλυτές ούτως ώστε να κωδικοποιηθούν οι πολιτικές θέσεις του υποψηφίου στα διάφορα θέματα. Για την επικοινωνία με τους υποψηφίους η ερευνητική ομάδα που καθορίζει τις ερωτήσεις επικοινωνεί μαζί τους είτε τηλεφωνικά είτε ηλεκτρονικά για να συμπληρώσουν το σχετικό ερωτηματολόγιο. Ακολούθως το ερωτηματολόγιο του κάθε υποψηφίου καταχωρείται ξεχωριστά στη βάση δεδομένων της εφαρμογής για να μπορεί μετέπειτα να συγκριθούν οι ερωτήσεις τους με αυτές των χρηστών. Στη βάση δεδομένων καταχωρούνται οι υποψήφιοι με τους οποίους ήταν εφικτό να επικοινωνήσουν οι ερευνητές του ερωτηματολογίου.

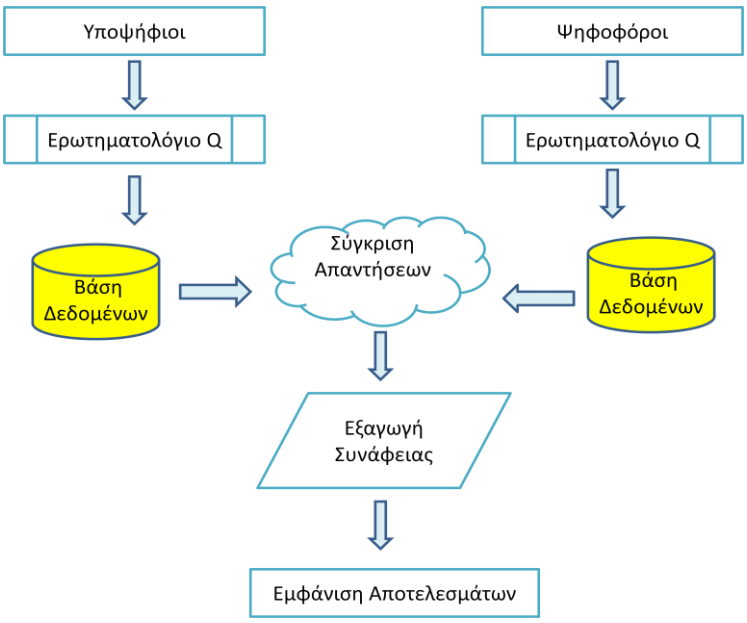
Να αναφέρουμε ότι η επιλογή των ερωτήσεων είναι αποκλειστικά αποτέλεσμα αυστηρά επιστημονικής έρευνας χωρίς καμία ανάμειξη των κομμάτων ή των υποψηφίων. Για την επιλογή των ερωτήσεων συμβάλλουν διάφοροι ερευνητές και ανεξάρτητοι ειδικοί ερευνητές. Το περιεχόμενο των ερωτήσεων διαφέρει από χώρα σε χώρα. Οι ερωτήσεις δημιουργούνται με βάση τις ιδιαιτερότητες κάθε χώρας, τα διάφορα θέματα πολιτικής επικαιρότητας καθώς επίσης και τα θέματα τα οποία θέλουν οι αναλυτές να δώσουν έμφαση. Πιο κάτω παρατίθενται μερικές ερωτήσεις οι οποίες υποβλήθηκαν στις βουλευτικές εκλογές τις Κύπρου το Μάιο του 2009.

1	Turkey should join the EU because this will help solve the Cyprus problem.
2	The position of Cyprus is in NATO.
3	Cyprus must apply for membership in the program "Partnership for Peace".
4	In the negotiations for the Cyprus' problem, the Government has made unacceptable concessions.
5	Cyprus should follow the example of other European countries and allow civil partnerships between homosexual couples.
6	The current process of resolving the Cyprus problem should be abandoned by replacing it with a five-party conference (Gr / Cypriots, T / Cypriots, Greece, Turkey and the EU).
7	The Cypriot economy has not benefited from EU membership.

Πίνακας 1 - Ερωτήσεις από Ερωτηματολόγιο

Ακολούθως οι χρήστες μέσω της διαδικτυακής εφαρμογής θα απαντήσουν και αυτοί το ίδιο ερωτηματολόγιο. Στην συνέχεια μέσα από διάφορους αλγορίθμους το σύστημα θα συγκρίνει τις απαντήσεις του εκάστοτε χρήστη με αυτές των υποψηφίων. Σαν αποτέλεσμα αυτής της διαδικασίας θα έχουμε το βαθμό συνάφειας του χρήστη με τους διάφορους υποψηφίους.

Πιο κάτω παρουσιάζεται ένα διάγραμμα ροής των VAA



Εικόνα 11 - Διάγραμμα Ροής VAA

3.3 Διάδοση των VAA

Τέτοιες εφαρμογές όπως είναι οι Ηλεκτρονικοί Σύμβουλοι Ψήφου (Vote Advice Application), έχουν εφαρμοστοί σε διάφορες χώρες της Ευρώπης αλλά και στην Αμερική.

Πρωτοπόρος σε αυτή την διαδικασία ήταν η Ολλανδία. Το 1989 η εφαρμογή αυτή κυκλοφόρησε σε έντυπη μορφή και αργότερα ενσωματώθηκε σε δισκέτες υπολογιστών. Με την εξέλιξη της τεχνολογίας και του διαδικτύου οι Ηλεκτρονικοί Σύμβουλοι Ψήφου ενσωματώθηκαν και στο διαδίκτυο. Ακολούθως εφαρμόστηκαν και σε πολλές άλλες χώρες εκ των οποίων είναι η Ελβετία , Αγγλία , Ελλάδα και Κύπρος. Όσο αφορά την Ελλάδα η εφαρμογή αυτή εφαρμόστηκε στις βουλευτικές εκλογές του Σεπτεμβρίου 2015 και βρίσκεται στην ιστοσελίδα www.Choose4Greece.com. Το Choose4Greece ήταν ένα αποτέλεσμα μίας ομάδας ερευνητών από τα πανεπιστημιακά ιδρύματα α) e-Democracy Center, Center for Democracy Aarau, University of Zurich, β) Τμήμα Επικοινωνίας και Σπουδών Διαδικτύου, Τεχνολογικό Πανεπιστήμιο Κύπρου, γ)Τμήμα Ψυχολογίας, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, δ) Department of Public Administration, University of Twente, ε) Niffield Collage, University of Oxford και στ) Τμήμα Δημοσιογραφίας & ΜΜΕ, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης.

Αντίστοιχη διαδικτυακή εφαρμογή για τα δεδομένα τις Κύπρου βρίσκεται στην ιστοσελίδα www.choose4Cyprus.com.

Κεφάλαιο 4
Δεδομένα - Datasets

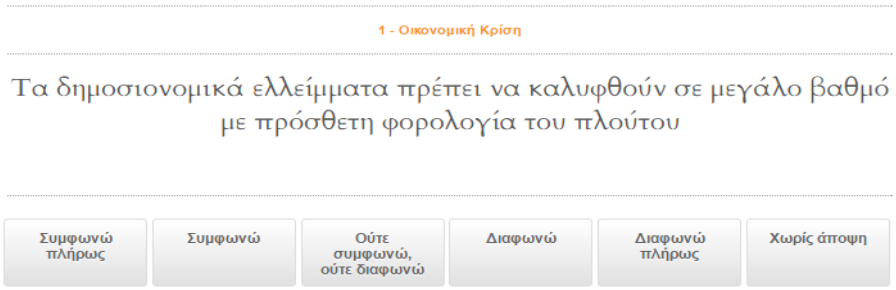
4.1 Εισαγωγή

Τα δεδομένα τα οποία θα αναλύσουμε πάρθηκαν από τις βουλευτικές εκλογές που έγιναν στην Κύπρο τον Μάιο του 2011, για τις Τουρκικές βουλευτικές εκλογές του 2011, για τις Ελληνικές βουλευτικές εκλογές του 2012 και για τις Βουλγαρικές βουλευτικές εκλογές του 2013.

Ενδεικτικά θα αναφερθούμε στις βουλευτικές εκλογές που έγιναν στην Κύπρο όπου συμμετείχαν συνολικά 11 κόμματα. Μερικά από τα κόμματα που συμμετείχαν είναι το ΑΚΕΛ, ΔΗΣΥ, ΔΗΚΟ, ΕΔΕΚ, ΕΥΡΩΚΟ, ΚΙΝΗΜΑ ΟΙΚΟΛΟΓΩΝ και το ΕΛΑΜ. Τα δεδομένα όλων των χωρών που θα χρησιμοποιηθούν βρίσκονται στο σύνδεσμο http://www.preferencematcher.org σε μορφή CSV. Μέσα σε αυτό το αρχείο αποθηκεύτηκαν κωδικοποιημένα σε αριθμούς όλες οι απαντήσεις των ερωτήσεων που έδωσαν οι χρήστες. Για να μπορέσουμε να αναλύσουμε και να αντιστοιχήσουμε τον κάθε αριθμό σε ποιο κόμμα ανήκει ή σε ποια απάντηση αντιστοιχεί ο κάθε αριθμός θα πρέπει να δούμε το αρχείο codebook. Στο αρχείο αυτό μπορούμε να δούμε τον αριθμό που αντιστοιχεί στην κάθε ερώτηση ή τον αριθμό που αντιστοιχεί στο κάθε κόμμα καθώς επίσης και όλες τις ερωτήσεις που κλήθηκε να απαντήσει ο κάθε ψηφοφόρος . Το αρχείο αυτό

μπορούμε να το βρούμε στον πιο πάνω σύνδεσμο και για κάθε εκλογική αναμέτρηση της εκάστοτε χώρας υπάρχει και ξεχωριστό codebook.

Για τα δεδομένα της Κύπρου οι ψηφοφόροι κλήθηκαν να απαντήσουν το διαδικτυακό ερωτηματολόγιο μέσω τις ιστοσελίδας www.choose4Cyprus.com. Συνολικά πήραν μέρος συνολικά 5470 χρήστες οι οποίοι κλήθηκαν να απαντήσουν σε 30 ερωτήσεις. Οι απαντήσεις των ερωτήσεων χωρίζονταν σε έξι βαθμίδες και είναι του τύπου «Συμφωνώ πλήρως», «Συμφωνώ», «Ούτε συμφωνώ, ούτε διαφωνώ», «Διαφωνώ», «Διαφωνώ πλήρως», «Χωρίς άποψη» , όπου το μηδέν(0) είναι το «Συμφωνώ πλήρως» μέχρι το πέντε(5) «Χωρίς άποψη» . Παράδειγμα των ερωτήσεων παρουσιάζεται πιο κάτω.



Εικόνα 12 – Ερώτηση από www.choose4cyprus.com

Επιπλέον ο χρήστης καλείται να απαντήσει και σε διάφορα άλλα δημογραφικά στοιχεία όπως είναι το έτος γέννησής του, το φύλο του, εάν είναι πρόσφυγας, το μορφωτικό του επίπεδο (απόφοιτος Δημοτικού, Γυμνασίου, Λυκείου, Τριτοβάθμια εκπαίδευση, Μεταπτυχιακές σπουδές), σε ποιο πολιτικό κόμμα νιώθει ότι είναι πιο κοντά, ποιο κόμμα προτίθεται να ψηφίσει (vote intention) και τον λόγο για τον οποίο ψηφίζει τον συγκεκριμένο υποψήφιο. Όλα αυτά τα δημογραφικά στοιχεία δεν τα χρησιμοποιούμε.

4.2 Ερωτηματολόγιο

Πιο κάτω παρατίθενται διάφορες ερωτήσεις στις οποίες καλείται να απαντήσει ο χρήστης καθώς και η ερώτηση για το τι προτίθεται να ψηφίσει σε αυτές τις εκλογές. Συγκεκριμένα βλέπουμε τις 30 ερωτήσεις που παρουσιάστηκαν στους ψηφοφόρους της Κύπρου για τις βουλευτικές εκλογές του 2009. Οι ερωτήσεις και ο αριθμός των ερωτήσεων διαφέρουν από χώρα σε χώρα.

1	In the negotiations for the Cyprus' problem, the Government has made unacceptable concessions.
2	A bi-zonal, bi-communal federation with one sovereignty and citizenship is an acceptable solution
3	The current process of resolving the Cyprus problem should be abandoned by replacing it with a five-party conference (Gr / Cypriots, T / Cypriots, Greece, Turkey and the EU).
4	Even if an agreed settlement is achieved, with the Turkish Cypriots we will have to live separately rather than together.
5	Turkey should join the EU because this will help solve the Cyprus problem.
6	The mobilization of Turkish-Cypriots for being released from Turkey is important in helping the prospect of a final settlement of the Cyprus problem.
7	Cyprus does not have to demand at the present moment the abolishing of the British bases
8	The position of Cyprus is in NATO.
9	Cyprus must apply for membership in the program "Partnership for Peace".
10	The Cypriot economy has not benefited from EU membership.
11	The corporate profits tax should be increased for two years from 10% to 11%.
12	The tax of property should be increased for the 5,000 biggest land owners.
13	The semi-governmental organizations in general should be privatized
14	The economy functions better as far as the state intervenes less and companies are provided with more liberty to operate.
15	The institution of the ATA is outdated and should be abolished.
16	The fiscal deficit should be covered largely by additional taxation of wealth
17	It is better to increase taxes than to cut spending on education and health.
18	Foreign investments, like the one by the state of Qatar, should be facilitated by the government.
19	Military spending should be increased even if this means increasing the public deficit.
20	The system of social welfare (social security and pensions) should be rather managed by the private sector.
21	The entry of foreign workers generally harms rather than benefits the Cypriot economy.
22	There should be found a way so that political asylum seekers and political refugees do not get any financial benefits.
23	The increasing presence of foreign immigrants strengthens the multicultural identity of Cyprus
24	The institutionalization of pupils unionism will help the more active involvement of

	youth in politics
25	Cyprus should follow the example of other European countries and allow civil partnerships between homosexual couples.
26	The views of the Church of Cyprus should be seriously taken into account regarding the formulation of the country's policy-making.
27	To increase the sense of security, civil and political liberties should be limited.
28	Criminality is due to the large number of immigrants who are in Cyprus.
29	The protection of the environment should not be an obstacle for economic development.
30	The owners of golf courts contribute vitally to the economy and this why they should not be further burdened with additional "green" taxes on excessive water consumption.

Πίνακας 2- Ερωτηματολόγιο Κύπρου

Επιπλέον ο ψηφοφόρος κλήθηκε να απαντήσει και στην ερώτηση για το ποιο κόμμα προτίθεται να ψηφίσει (vote intention) σε αυτές τις βουλευτικές εκλογές. Οι επιλογές οι οποίες είχε στην διάθεση του ήταν οι ακόλουθες:

1	ΑΚΕΛ
2	ΔΗΣΥ
3	ΔΗΚΟ
4	ΕΔΕΚ
5	ΕΥΡΟΚΟ
6	ΚΟΠ
7	ΕΛΑΜ
8	ΚΚΟ
9	ΖΥΓΟΣ
10	ΚΥΠΡΟΣ
11	ΠΑΣΟΚ
12	ΑΛΛΟ
13	ΚΑΝΕΝΑ
98	ΔΕΝ ΑΠΑΝΤΗΣΑΝ

Πίνακας 3 - Πρόθεση Ψήφου

Κεφάλαιο 5
Πειραματική Αξιολόγηση

5.1 Στόχος αξιολόγησης

Στόχος της συγκεκριμένης πειραματικής αξιολόγησης είναι να προβλέψουμε την πρόθεση ψήφου του ψηφοφόρου, να εκπαιδεύσουμε τους αλγόριθμους, να βρούμε τον ελάχιστο αριθμό δεδομένων ούτως ώστε να εκπαιδευτεί ικανοποιητικά ο αλγόριθμος και ποιος αλγόριθμος είναι ο καλύτερος σε κάθε χώρα. Για να το πετύχουμε αυτό θα πρέπει μέσω των αλγορίθμων να αναλύσουμε τις ερωτήσεις που απάντησε ο κάθε χρήστης ξεχωριστά.. Η πιο σημαντική ερώτηση που καλείται να απαντήσει ο χρήστης είναι η ερώτηση για την πρόθεση ψήφου. Μέσα από αυτή την ερώτηση μπορούμε στο τέλος της πειραματικής διαδικασίας να επαληθεύσουμε εάν έχει γίνει σωστή πρόβλεψη και το ποσοστό επιτυχίας του ταξινομητή.

Για την ανάλυση των δεδομένων χρησιμοποιήσαμε δύο λογισμικά προγράμματα. Το λογισμικό πρόγραμμα R Studio και το λογισμικό πρόγραμμα WEKA. Μέσα από την R Studio χρησιμοποιήσαμε μία συγκεκριμένη βιβλιοθήκη η οποία ονομάζεται Rattle, η οποία είναι μια γραφική διεπαφή με τον χρήστη για την εξόρυξη δεδομένων στην R. Την εγκατάσταση της R

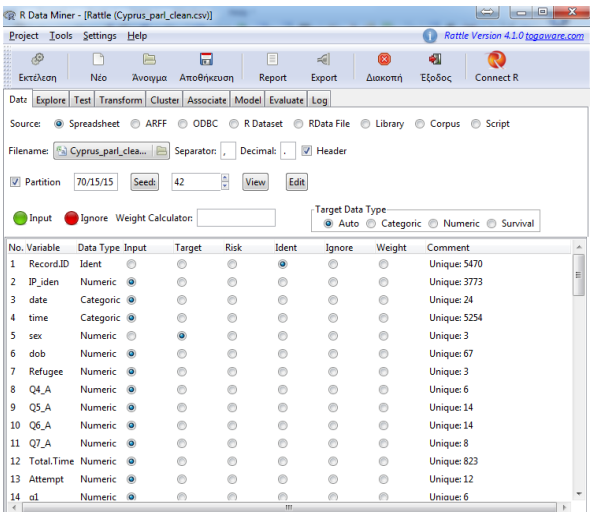
Studio την κάναμε από τον σύνδεσμο www.rstudio.com και του WEKA από τον σύνδεσμο http://www.cs.waikato.ac.nz/ml/weka/ .

5.1.1 Πρόθεση Ψήφου

Για την πρόβλεψη πρόθεσης ψήφου του κάθε χρήστη θα πρέπει αρχικά όλοι οι χρήστες να απαντήσουν ένα προκαθορισμένο ερωτηματολόγιο. Ακολούθως, από αυτό το ερωτηματολόγιο και με την βοήθεια κάποιων αλγορίθμων θα δημιουργήσουμε ένα δέντρο απόφασης. Από αυτό το δέντρο απόφασης θα έχουμε τη δυνατότητα να προβλέψουμε την πρόθεση ψήφου των επόμενων χρηστών που απαντούν το συγκεκριμένο ερωτηματολόγιο με βάση κάποιες συγκεκριμένες σημαντικές ερωτήσεις που βρίσκονται στο δέντρο απόφασης.

5.2 Επεξεργασία Δεδομένων

Στο αρχείο CSV υπάρχουν όλες οι ερωτήσεις που απάντησε ο χρήστης καθώς επίσης και διάφορες άλλες μεταβλητές. Μέσα από την Rattle θα πρέπει να καθοριστούν διάφοροι παράμετροι για τη σωστή επεξεργασία των δεδομένων. Σημαντικό είναι να καθορίσουμε τα δεδομένα ως κατηγορηματικά (categoric) και να δηλώσουμε και την ερώτηση στόχο (Target). Στη διεπαφή της Rattle αφού φορτώσουμε το αρχείο εμφανίζονται διάφορες μεταβλητές όπως είναι το Record.ID, date , time ,Q4_A, Q6_A , q1,t1 ,q2 ,t2, q3, t3 κτλ. Αυτό γίνεται για να γνωρίζουμε σε ποια ερώτηση αντιστοιχεί η κάθε μεταβλητή ούτως ώστε να κάνουμε τις απαραίτητες ρυθμίσεις στην rattle και να διαβάσουμε το codebook του συγκεκριμένου CSV αρχείου.



Εικόνα 13 – Interface of Rattle()

Το codebook είναι ένα αρχείο σε μορφή PDF το οποίο εμπεριέχει όλες τις ερωτήσεις που υποβλήθηκαν στον ψηφοφόρο καθώς επίσης και τις πολλαπλές απαντήσεις που είχε στην διάθεση του να επιλέξει. Το codebook το βρήκαμε στον σύνδεσμο: http://www.preferencematcher.org/.

Τα δεδομένα που χρησιμοποιήσαμε στην πειραματική αξιολόγηση είναι οι ερωτήσεις q1 έως και την ερώτηση q30 καθώς επίσης και η ερώτηση Q6_A. Η ερώτηση Q6_A είναι ερώτηση που καλείται να δηλώσει ο χρήστης πιο κόμμα προτίθεται να ψηφίσει. Είναι σημαντικό να δηλωθεί ως στόχος αυτή η ερώτηση γιατί μέσω αυτής μπορέσαμε να επαληθεύσουμε εάν ο ταξινομητής έχει προβλέψει σωστά την πρόθεση ψήφου του χρήστη.

Όπως έχουμε αναφέρει και στο δεύτερο κεφάλαιο υπάρχουν δύο φάσεις στην ταξινόμηση. Το Training (εκπαίδευση) και το Testing (επαλήθευση). Για αυτό και εμείς δηλώσαμε το ποσοστό των δεδομένων που θα εκπαιδεύσουμε στον ταξινομητή και το ποσοστό των δεδομένων που θα μπουν στην διαδικασία της επαλήθευσης.

Από το σύνολο των δεδομένων, δηλαδή από τους 5470 χρήστες που απάντησαν στα ερωτηματολόγια χρησιμοποιήσαμε το 66% των δεδομένων για να εκπαιδεύσουμε τον ταξινομητή και το υπόλοιπο 34% για να επαληθεύσουμε τα δεδομένα. Κατά την διάρκεια των απαντήσεων στις ερωτήσεις ο χρήστης κλήθηκε να απαντήσει και στην ερώτηση για το ποιο κόμμα προτίθεται να ψηφίσει. Έχοντας ως δεδομένο ποιο κόμμα προτίθεται να ψηφίσει μπορέσαμε εύκολα στο τέλος να επαληθεύσουμε εάν ο ταξινομητής προέβλεψε σωστά την πρόθεση ψήφου και το ποσοστό επιτυχίας του.

5.3 Δέντρα Απόφασης

Για να προβλέψουμε σωστά την πρόθεση ψήφου δημιουργήσαμε το «Δέντρο Απόφασης». Για την δημιουργία του τελικού Δέντρου Απόφασης (Decision Tree) χρησιμοποιήσαμε τον αλγόριθμο J48 και δώσαμε διάφορες τιμές στις παραμέτρους MinNumObj και confidenceFactor όπου MinNumObj είναι ο ελάχιστος αριθμός περιπτώσεων ανά φύλλο στο Δέντρο Απόφασης και confidenceFactor είναι ο συντελεστής πολυπλοκότητας του Δέντρου Απόφασης. Αρχικά δώσαμε μικρές τιμές για το MinNumObj και confidenceFactor. Για μικρές τιμές παρατηρούμε ότι δημιουργείται ένα πολύ μεγάλο δέντρο απόφασης το οποίο είναι δύσκολο να αναλυθεί και να διαβαστεί. Επιπλέον στο μεγάλο δέντρο απόφασης υπήρχε και μεγάλος συντελεστής

σφάλματος. Για να μειώσουμε τον συντελεστή σφάλματος δώσαμε σταδιακά μεγαλύτερες τιμές στο MinNumObj και στο confidenceFactor.

Παρατηρήσαμε ότι όσο αυξάνουμε την τιμή του MinNumObj (ελάχιστος αριθμός περιπτώσεων ανά φύλλο) και του confidenceFactor (συντελεστής περιπλοκότητας του δέντρου) τόσο πιο πολύ μικραίνει το δέντρο απόφασης και ελαχιστοποιείται ο συντελεστής σφάλματος.

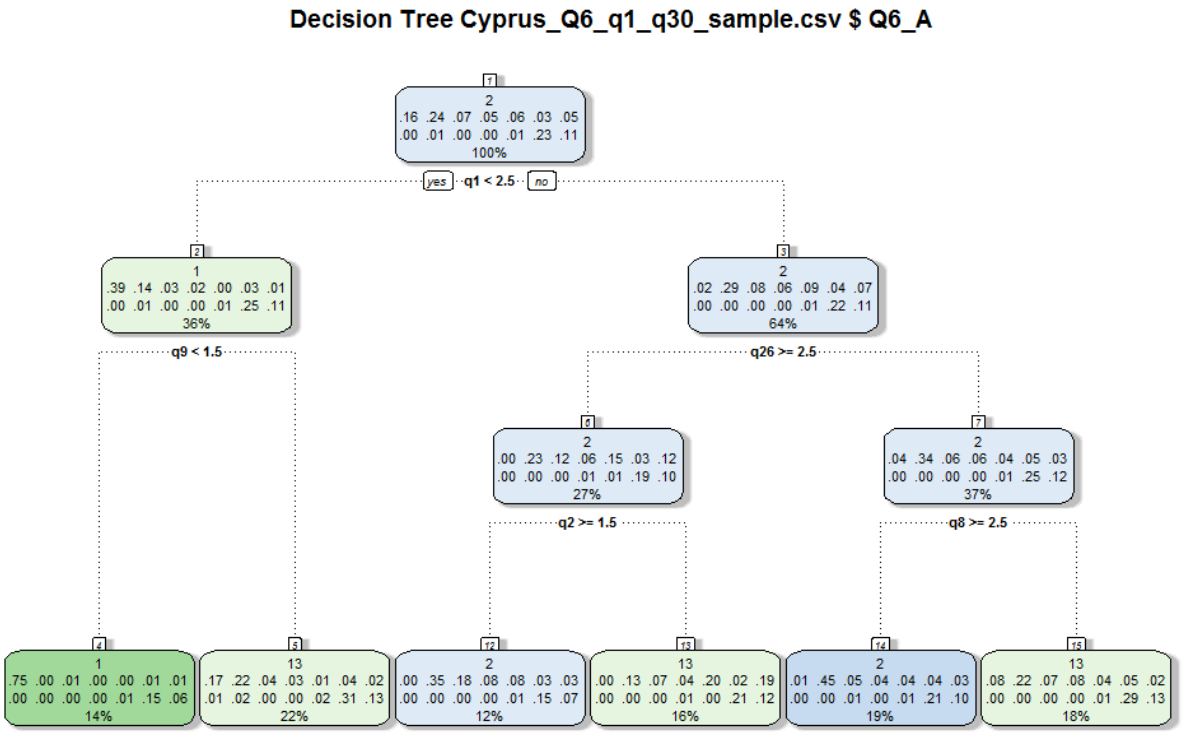
Πιο κάτω παρουσιάζονται αναλυτικά τα αποτελέσματα από την εκπαίδευση και την επαλήθευση των δεδομένων. Για την επιλογή του βέλτιστου δέντρου απόφασης θα πρέπει οι τιμές του F-Measure να είναι όσο το δυνατό πιο μεγάλες για να έχουμε καλύτερα ποσοστά επιτυχίας. Επίσης το μέγεθος του δέντρου θα πρέπει να μην είναι πολύ μεγάλο αλλά ούτε και πολύ μικρό για να μπορούμε πιο εύκολα να αναλύουμε και να προβλέπουμε την πρόθεση ψήφου μέσα από το δέντρο απόφασης. Δίνοντας διάφορες τιμές στον συντελεστή περιπλοκότητας του δέντρου απόφασης (Confidence Factor) και στον ελάχιστο αριθμό περιπτώσεων ανά φύλλο (MinNumObj) και προέκυψε ο πιο πάνω πίνακας:

	Αλγόριθμος J48	Correctly Classified	F-Measure	Χρόνος (Sec)	Μέγεθος Δέντρου
1st Case	**MinNumObj = 0.1 ConfidenceFactor =10**	**56.3986**	**0.513**	**0.16**	**Πολύ Μεγάλο**
2st Case	**MinNumObj = 0.01 ConfidenceFactor =40**	55.4926	0.459	**0.19**	**Μεγάλο**
3st Case	**MinNumObj = 0.001 ConfidenceFactor =80**	55.8324	0.485	**0.08**	**Ικανοποιητικό**
4st Case	**MinNumObj = 0.0001 ConfidenceFactor =90**	53.6806	0.416	**0.16**	**Ικανοποιητικό**
5st Case	**MinNumObj = 0.00001 ConfidenceFactor =190**	51.7554	0.406	**0.05**	**Μικρό**
6st Case	**MinNumObj = 0.00001 ConfidenceFactor =230**	51.7554	0.406	**0.09**	**Πολύ Μικρό**

Πίνακας 4 - Κλάδεμα Δέντρου Απόφασης

Στην 1η περίπτωση εμφανίζει πολύ καλύτερα αποτελέσματα στο F-Measure αλλά το μέγεθος του δέντρου είναι πολύ μεγάλο για να το αναλύσουμε και να το κατανοήσουμε. Ακολούθως δίναμε μεγαλύτερες τιμές στο MinNumObj Confidence Factor. Αυτή την επαναληπτική διαδικασία την κάναμε συνολικά 6 φορές. Από αυτή την πειραματική αξιολόγηση προέκυψε ότι η 3η περίπτωση είναι η πιο ιδανική γιατί το δέντρο απόφασης που προέκυψε δεν είναι ούτε πολύ μεγάλο αλλά ούτε πολύ μικρό και η τιμή στο F-Measure δεν έχει μεγάλη απόκλιση από την 1η περίπτωση όπου

είναι η πιο ιδανική. Δεδομένου των τιμών στην 3η πειραματική αξιολόγηση προκύπτει το πιο κάτω δέντρο απόφασης:



Εικόνα 14 – Τελικό Δέντρο Απόφασης

Από το τελικό δέντρο απόφασης κρατήσαμε τις δύο(2) πιο σημαντικές ερωτήσεις οι οποίες προβλέπουν πιο σωστά την πρόθεση ψήφου. Οι ερωτήσεις αυτές είναι οι δύο πρώτες ερωτήσεις από το τελικό δέντρο απόφασης δηλαδή η q1 και q9.

Από το πιο πάνω δέντρο απόφασης παρατηρούμε ο αρχικός κόμβος (root) αντιστοιχεί στην ερώτηση q1. Μέσα στον αρχικό κόμβο (root) υπάρχουν τα ποσοστά των απαντήσεων των χρηστών στην ερώτηση για το ποιο κόμμα προτίθενται να ψηφίσουν (vote intention). Τα 14 ποσοστά που εμφανίζονται μέσα στον κόμβο root αντιστοιχούν στο κάθε κόμμα ξεχωριστά . Η σειρά που εμφανίζονται τα κόμματα δεν είναι τυχαία αλλά παρουσιάζονται κατά σειρά τα κόμματα όπως έχουν δηλωθεί στην ερώτηση της πρόθεσης ψήφου. Το ποσοστό το οποίο αντιστοιχεί με κάθε κόμμα μπορούμε να το βρούμε στο codebook που αντιστοιχεί σε αυτό το ερωτηματολόγιο. Στη δική μας περίπτωση το 1ο ποσοστό αντιστοιχεί στο κόμμα ΑΚΕΛ, το 2ο στο κόμμα ΔΗΣΥ, το 3ο αντιστοιχεί στο κόμμα ΔΗΚΟ, το 4ο αντιστοιχεί στο κόμμα ΕΔΕΚ, το 5ο αντιστοιχεί στο κόμμα ΕΥΡΩΚΟ, το 6ο αντιστοιχεί στο κόμμα Κίνημα Οικολόγων Περιβαλλοντιστών (ΚΟΠ), το 7ο αντιστοιχεί στο κόμμα ΕΛΑΜ, το 8ο αντιστοιχεί στο κόμμα ΚΚΟ ,

το 9ο αντιστοιχεί στο κόμμα ΖΥΓΟΣ, το 10ο αντιστοιχεί στο κόμμα ΚΥ.ΠΡΟ.Σ, το 11ο αντιστοιχεί στο κόμμα ΠΑΣΟΚ , το 12ο αντιστοιχεί στην απάντηση ΑΛΛΟΣ, το 13ο αντιστοιχεί στο ΚΑΝΕΝΑΣ και το 14ο αντιστοιχεί σε αυτούς που δεν απάντησαν στην συγκεκριμένη ερώτηση. Στον αρχικό κόμβο φαίνεται ότι ποσοστό 16% προτίθεται να ψηφίσει ΑΚΕΛ , 24% προτίθεται να ψηφίσει ΔΗΣΥ κ.λπ.

Κάτω από τον αρχικό κόμβο βλέπουμε να αναγράφεται q1<2.5. Να υπενθυμίσουμε ότι οι απαντήσεις που καλείται να δώσει ο χρήστης είναι συνήθως του τύπου: «Διαφωνώ πλήρως», «Διαφωνώ», «Ούτε συμφωνώ, ούτε διαφωνώ», «Συμφωνώ», «Συμφωνώ πλήρως», «Χωρίς άποψη» όπου αντιστοιχούν από το μηδέν(0) μέχρι το έξι(6) αντίστοιχα. Στην περίπτωση που ο χρήστης στην ερώτηση q1 έχει απαντήσει από το «Διαφωνώ πλήρως», «Διαφωνώ», «Ούτε συμφωνώ, ούτε διαφωνώ» όπου αντιστοιχεί από το μηδέν(0) μέχρι το δύο(2) και σύμφωνα με την ισότητα του δέντρου απόφασης όπου q1<2.5 θα μεταβούμε στον κόμβο αριστερά του αρχικού κόμβου. Εάν ο χρήστης στον αρχικό κόμβο έχει απαντήσει «Συμφωνώ», «Συμφωνώ πλήρως», «Χωρίς άποψη» όπου οι επιλογές του αντιστοιχούν στους αριθμούς από 3 μέχρι 5 τότε θα μεταβούμε δεξιά του αρχικού κόμβου.

Το ίδιο συμβαίνει και στον κόμβο με αριθμό 2 ο οποίος βρίσκεται αριστερά του αρχικού. Η ανισότητα η οποία βρίσκεται στον κόμβο με αριθμό 2 είναι q9<1.5 . Αυτό μεταφράζεται στο ότι όσοι χρήστες απαντήσουν στην ερώτηση q9 με «Διαφωνώ πλήρως», «Διαφωνώ» θα πρέπει να μεταβούν στον κόμβο 4 όπου βρίσκεται αριστερά του κόμβου 2 , ενώ οι υπόλοιποι χρήστες που θα απαντήσουν από το «Ούτε συμφωνώ, ούτε διαφωνώ» μέχρι το «Χωρίς άποψη» τότε θα μεταβούν δεξιά του κόμβου 2.

Αυτή τη διαδικασία την κάνουμε για να προβλέψουμε όσο το δυνατό καλύτερα την πρόθεση ψήφου. Στην προκειμένη περίπτωση εάν η απάντηση του χρήστη στην ερώτηση q1 αντιστοιχεί αριθμητικά σε αριθμό μικρότερο του 2.5 και στην ερώτηση q9 αντιστοιχεί αριθμητικά σε αριθμό μικρότερο του 1.5 τότε υπάρχει πιθανότητα 75% κάποιος να ψηφίσει το κόμμα ΑΚΕΛ, 1% να ψηφίσει το κόμμα ΔΗΚΟ, 1% να ψηφίσει το κόμμα ΚΟΠ, 1% να ψηφίσει το κόμμα ΕΛΑΜ, 1% να ψηφίσει το κόμμα ΚΑΝΕΝΑ, 15% να ψηφίσει το κόμμα ΑΛΛΟ και 6% να μη δώσει καμία απάντηση.

5.3.1 Συμπεράσματα

Τα συμπεράσματα που εξαγάγαμε από την πιο πάνω πειραματική αξιολόγηση είναι ότι όσο αυξάνουμε το confidenceFactor (παράγοντας κλαδέματος δέντρου) και όσο αυξάνουμε το MinNumObj (ελάχιστος αριθμός περιπτώσεων ανά φύλλο), τόσο πιο πολύ μικραίνει το δέντρο βγάζοντας και μικρότερα ποσοστά σφάλματος.

Για την επιλογή των δύο πιο σημαντικών ερωτήσεων σχετικά με την πρόβλεψη πρόθεσης ψήφου του ψηφοφόρου καθορίζονται οι δύο πρώτες ερωτήσεις από το δέντρο απόφασης. Όπως βλέπουμε και στο τελικό δέντρο απόφασης οι δύο πρώτες ερωτήσεις του δέντρου απόφασης είναι η ερώτηση q1 και q9.

q1	In the negotiations for the Cyprus' problem, the Government has made unacceptable concessions.
q9	must apply for membership in the program "Partnership for Peace"

Πίνακας 5 - Σημαντικές Ερωτήσεις

5.4 Ομαδοποίηση χρηστών σε διανυσματικό χώρο

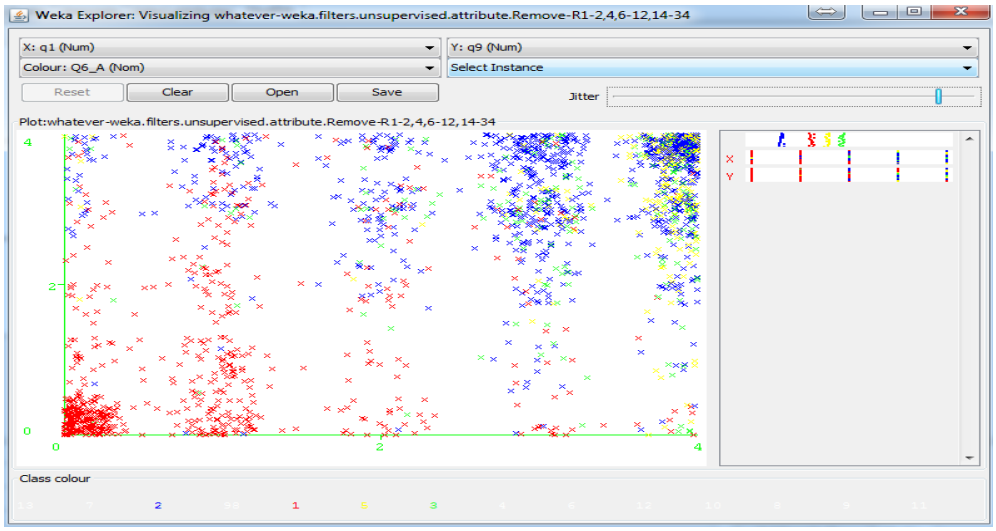
Για την ανάλυση των δεδομένων κρατήσαμε τα δεδομένα των χρηστών μόνα από τα τέσσερα (4) δημοφιλέστερα κόμματα. Τα τέσσερα (4) αυτά κόμματα αντιπροσωπεύουν την πλειοψηφία των ψηφοφόρων. Η επιλογή των τεσσάρων δημοφιλέστερων κομμάτων έγινε από τα αποτελέσματα της προηγούμενης πειραματικής αξιολόγησης. Με βάση αυτά τα αποτελέσματα επιλέξαμε τελικά τα τέσσερα δημοφιλέστερα κόμματα. Τα αποτελέσματα παρουσιάζονται στον πιο κάτω πίνακα.

Η συχνότητα προτίμησης στα κόμματα είναι η ακόλουθη:

%	Κωδικός Κόμματος	Όνομα
24	2	ΔΗΣΥ
23	13	ΚΑΝΕΝΑΣ
16	1	ΑΚΕΛ
11	98	Δεν Απάντησε
7	3	ΔΗΚΟ
6	5	ΕΥΡΩΚΟ
5	4	ΕΔΕΚ
5	7	ΕΛΑΜ
3	6	Κίνημα Οικολόγων

Πίνακας 6 - Ποσοστά Δημοφιλέστερων Κόμματων

Βάσει αυτών των δεδομένων τα τέσσερα δημοφιλέστερα κόμματα είναι το ΔΗΣΥ με 24%, το ΑΚΕΛ με 16% , το ΔΗΚΟ με 7% και το ΕΥΡΩΚΟ με 6%. Το κόμμα ΚΑΝΕΝΑΣ με ποσοστό 23% ήρθε δεύτερο στις προτιμήσεις των ψηφοφόρων αλλά δε θα το λάβουμε υπόψη. Το συνολικό ποσοστό και των τεσσάρων κομμάτων ανέρχεται στο 53%. Οι κωδικοί των 4 κομμάτων είναι οι εξής: 1,2,3,5. Πιο κάτω απεικονίζονται τα τέσσερα δημοφιλέστερα κόμματα και οι απαντήσεις των χρηστών στις ερωτήσεις q1 και q9 όπου είναι οι πιο σημαντικές.



Εικόνα 15 - Ποσοστά Δημοφιλέστερων Κόμματων

Στο πιο πάνω διάγραμμα με κόκκινο χρώμα απεικονίζεται το ΑΚΕΛ , με μπλε χρώμα το ΔΗΣΥ, με πράσινο χρώμα το ΔΗΚΟ και με κίτρινο χρώμα το ΕΥΡΩΚΟ.

Θυμίζουμε ότι, ο κάθε χρήστης απαντούσε στις 30 ερωτήσεις με τις ακόλουθες βαθμίδες «Διαφωνώ πλήρως», «Διαφωνώ», «Ούτε συμφωνώ, ούτε διαφωνώ», «Συμφωνώ», «Συμφωνώ πλήρως» , όπου το μηδέν(0) είναι το «Διαφωνώ Πλήρως» μέχρι το τέσσερα(4) «Συμφωνώ πλήρως».

Στην ερώτηση 1 «Κατά πόσο στις διαπραγματεύσεις για το κυπριακό η κυβέρνηση έχει κάνει απαράδεκτες παραχωρήσεις;» βλέπουμε μια ταύτιση απόψεων των ψηφοφόρων που είχαν πρόθεση να ψηφίσουν το ΑΚΕΛ.. Όπως βλέπουμε και στο διάγραμμα οι περισσότεροι χρήστες έχουν απαντήσει «Διαφωνώ Πλήρως», αρκετοί έχουν απαντήσει «Διαφωνώ» ενώ λίγοι είναι αυτοί που απάντησαν «Ούτε Συμφωνώ, ούτε Διαφωνώ» και «Συμφωνώ». Εάν αναλογιστούμε ότι στην Κυβέρνηση εκείνη την περίοδο ήταν το ΑΚΕΛ, τότε είναι λογικό οι ψηφοφόροι που είχαν πρόθεση να ψηφήσουν το ΑΚΕΛ να διαφωνούν με την ερώτηση κατά πόσο η κυβέρνηση έκανε απαράδεκτες υποχωρήσεις στις διαπραγματεύσεις για το κυπριακό.

Επιπλέον βλέπουμε πολλούς χρήστες που είχαν πρόθεση να ψηφίσουν το ΔΗΣΥ και απεικονίζονται με μπλε χρώμα να δηλώνουν «Συμφωνώ πλήρως» ότι η κυβέρνηση έχει κάνει απαράδεκτες υποχωρήσεις για το κυπριακό, ενώ αρκετοί είναι αυτοί που απάντησαν «Συμφωνώ». Υπάρχει και μία μικρή μερίδα που απάντησαν ότι «Ούτε συμφωνώ, ούτε διαφωνώ», «Διαφωνώ» ή «Διαφωνώ Πλήρως». Εάν αναλογιστούμε ότι εκείνη την περίοδο στην αντιπολίτευση βρισκόταν ο ΔΗΣΥ, τότε είναι λογικό να διαφωνούν με τους χειρισμούς της κυβέρνησης.

Όσον αφορά του χρήστες που είχαν πρόθεση να ψηφίσουν το ΔΗΚΟ και απεικονίζονται με πράσινο χρώμα, στην ερώτηση 1 οι περισσότεροι έχουν δηλώσει ότι «Συμφωνώ Πλήρως», αρκετοί έχουν δηλώσει ότι «Συμφωνώ», ενώ λίγοι είναι αυτοί που δήλωσαν ότι «Ούτε συμφωνώ, ούτε διαφωνώ» , «Διαφωνώ» ή ότι «Διαφωνώ Πλήρως».

Τέλος οι χρήστες που είχαν πρόθεση να ψηφίσουν το ΕΥΡΩΚΟ και απεικονίζονται με κίτρινο χρώμα, στην ερώτηση 1 οι περισσότεροι έχουν δηλώσει «Συμφωνώ Πλήρως», πολύ λίγοι έχουν δηλώσει «Διαφωνώ», ενώ ελάχιστοι είναι αυτοί που δήλωσαν ότι «Ούτε συμφωνώ, ούτε διαφωνώ» ή ότι «Διαφωνώ Πλήρως».

Στην ερώτηση 9 όπου αναφέρει εάν η Κύπρος πρέπει να υποβάλει αίτηση για ένταξη στο πρόγραμμα «Συνεταιρισμός για την Ειρήνη» βλέπουμε πάλι μία ταύτιση απόψεων των χρηστών.

Από τους χρήστες που είχαν πρόθεση να ψηφίσουν το ΑΚΕΛ και απεικονίζονται με κόκκινο χρώμα, μια μεγάλη μερίδα είχε δηλώσει ότι «Διαφωνώ Πλήρως», αρκετοί χρήστες είχαν δηλώσει «Διαφωνώ» ή «Ούτε συμφωνώ, ούτε Διαφωνώ» , λίγοι χρήστες δήλωσαν «Συμφωνώ» , ενώ ελάχιστοι είναι αυτοί που δήλωσαν ότι «Συμφωνώ Πλήρως».

Επιπλέον βλέπουμε πολλούς χρήστες που είχαν πρόθεση να ψηφίσουν το ΔΗΣΥ και απεικονίζονται με μπλε χρώμα να δηλώνουν «Συμφωνώ πλήρως» ότι η Κύπρος πρέπει να υποβάλει αίτηση για ένταξη στο πρόγραμμα «Συνεταιρισμός για την Ειρήνη», αρκετοί έχουν δηλώσει «Συμφωνώ», λίγοι είναι αυτοί που δήλωσαν ότι «Ούτε συμφωνώ, ούτε διαφωνώ» , ενώ ελάχιστοι είναι αυτοί που δήλωσαν «Διαφωνώ» ή «Διαφωνώ Πλήρως».

Όσον αφορά του χρήστες που είχαν πρόθεση να ψηφίσουν το ΔΗΚΟ και απεικονίζονται με πράσινο χρώμα, στην ερώτηση 9 οι περισσότεροι έχουν δηλώσει ότι «Συμφωνώ», λίγοι είναι αυτοί που δήλωσαν ότι «Συμφωνώ Πλήρως», ελάχιστοι είναι αυτοί που δήλωσαν ότι «Ούτε

συμφωνώ , ούτε Διαφωνώ», ενώ πάρα πολύ λίγοι είναι αυτοί που δήλωσαν ότι «Διαφωνώ Πλήρως».

Τέλος οι χρήστες που είχαν πρόθεση να ψηφίσουν το ΕΥΡΩΚΟ και απεικονίζονται με κίτρινο χρώμα, στην ερώτηση 9 οι περισσότεροι έχουν δηλώσει ότι είτε «Συμφωνώ Πλήρως» είτε «Συμφωνώ», λίγοι έχουν δηλώσει «Ούτε συμφωνώ, ούτε διαφωνώ» , ενώ ελάχιστοι είναι αυτοί που δήλωσαν «Διαφωνώ» ή «Διαφωνώ Πλήρως».

5.4.1 Συμπεράσματα

Η κατανομή των χρηστών στην πιο πάνω εικόνα μας άφησε ικανοποιημένους. Στις ερωτήσεις 1 και 9 βλέπουμε μια συγκεντρωτική κατανομή των ψηφοφόρων του ΑΚΕΛ στις συντεταγμένες (0,0) όπου αντιστοιχούν στις απαντήσεις («Διαφωνώ Πλήρως», «Διαφωνώ Πλήρως») αντίστοιχα. Σωστά εμφανίζονται στις δύο διαστάσεις γιατί εάν αναλογιστούμε ότι στην Κυβέρνηση εκείνη την περίοδο ήταν το ΑΚΕΛ τότε είναι λογικό οι ψηφοφόροι που είχαν πρόθεση να ψηφήσουν το ΑΚΕΛ να διαφωνούν με τις ερωτήσεις «κατά πόσο η κυβέρνηση έκανε απαράδεκτες υποχωρήσεις στις διαπραγματεύσεις για το κυπριακό» και «στην ένταξη για τον συνεταιρισμό για την ειρήνη». Με αυτές τις απόψεις διαφωνεί και σαν κόμμα το ΑΚΕΛ.

Επίσης βλέπουμε ότι υπάρχει και μία συγκεντρωτική κατανομή των ψηφοφόρων του ΔΗΣΥ που απεικονίζονται με μπλε χρώμα στις συντεταγμένες (4,4) όπου αντιστοιχούν στις απαντήσεις («Συμφωνώ Πλήρως», «Συμφωνώ Πλήρως») αντίστοιχα. Και σε αυτή την περίπτωση σωστά εμφανίζονται εκεί τα αποτελέσματα γιατί εάν αναλογιστούμε ότι εκείνη την περίοδο στην αντιπολίτευση βρισκόταν ο ΔΗΣΥ, τότε είναι λογικό να διαφωνούν με τους χειρισμούς της κυβέρνησης. Επιπλέον όσον αφορά την ερώτηση 9 που αφορά την ένταξη στον συνεταιρισμό για την ειρήνη και εδώ είναι λογικό να διαφωνούν οι ψηφοφόροι του ΔΗΣΥ γιατί είναι πάγια πολιτική του ΔΗΣΥ να θέλει ένταξη στον συνεταιρισμό για την Ειρήνη.

5.5 Σύγκριση Αλγορίθμων ταξινόμησης

Σε αυτή την πειραματική αξιολόγηση αναλύσαμε την ορθότητα των αλγορίθμων Naïve Bayes , MLP , J48 , SMO και ο IBK .

Για να δούμε ποιος αλγόριθμος είναι πιο καλός και έχει τα πιο καλά αποτελέσματα με μικρότερο ποσοστό σφάλματος εφαρμόσαμε τον κάθε αλγόριθμο ξεχωριστά στα δεδομένα μας.

Υπενθυμίζουμε ότι τα δεδομένα μας πάρθηκαν από 5470 χρήστες οι οποίοι απάντησαν το ερωτηματολόγιο διαδικτυακά.

Τα αποτελέσματα από τους διάφορους αλγόριθμους τα βλέπουμε στον πιο κάτω πίνακα. Τα δεδομένα που χρησιμοποιήθηκαν είναι από τις δύο πιο σημαντικές ερωτήσεις καθώς επίσης και η ερώτηση για την πρόθεση ψήφου.

Αλγόριθμοι	Correctly Classified	Precision	Recall	F-Measure	Χρόνος (Sec)
Naive Bayes	**70.5972**	**0.573**	**0.706**	**0.626**	**0.03**
MLP	**70.4475**	**0.65**	**0.704**	**0.629**	**12.29**
J48	**69.615**	**0.563**	**0.696**	**0.614**	**0.04**
SMO	**70.4475**	**0.574**	**0.704**	**0.623**	**0.34**
IBk	**69.8751**	**0.564**	**0.699**	**0.617**	**0.02**

Πίνακας 7- Αποτελέσματα Αλγορίθμων

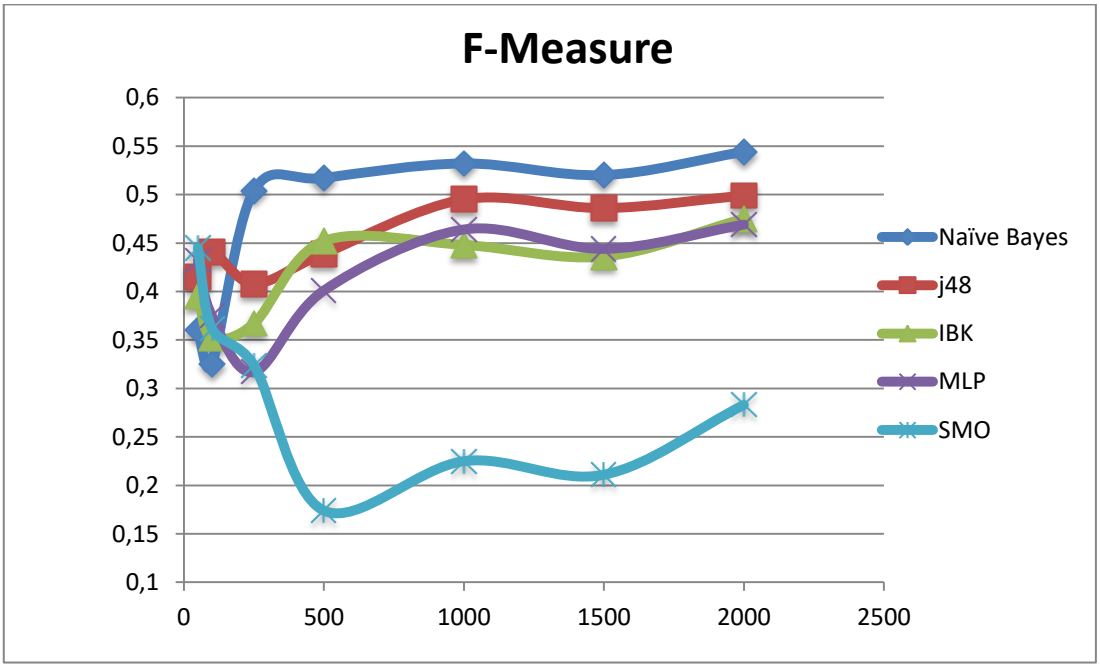
Βλέπουμε ότι ο αλγόριθμος Naive Bayes είναι ο πιο αξιόπιστος αλγόριθμος όσο αφορά τα ποσοστά επιτυχίας. Ο αλγόριθμος Naive Bayes συνδυάζει καλύτερα ποσοστά επιτυχίας και είναι και ο πιο γρήγορος . Ακολούθως έρχεται ο MLP , SMO , iBK και τέλος ο J48.

5.6 Δυναμική Αξιολόγηση Αλγορίθμων

Για την δυναμική πρόβλεψη ψήφου των χρηστών θέλουμε να δούμε ποιος ταξινομητής είναι ο καλύτερος στο να κάνει καλύτερες προβλέψεις με λίγα δεδομένα.

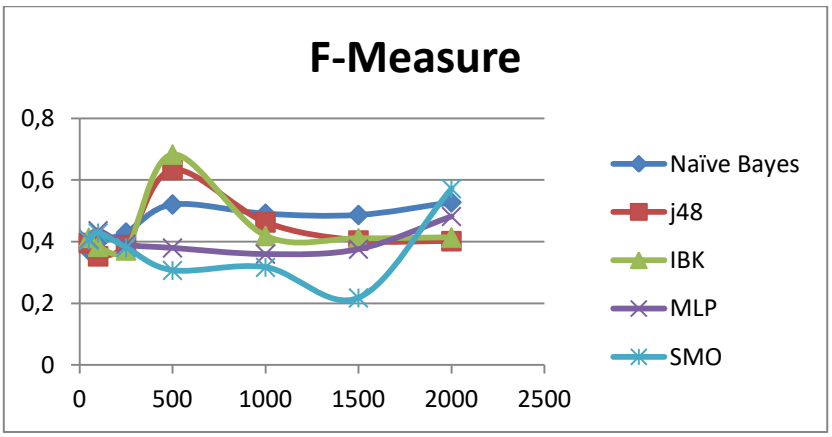
Για να το πετύχουμε αυτό αρχικά δώσαμε λίγα δεδομένα στους διάφορους ταξινομητές μας και είδαμε την ορθότητα τους. Στην συνέχεια δίναμε περισσότερα δεδομένα στους ταξινομητές μας και καταγράφαμε την ορθότητα του κάθε ταξινομητή μας ξεχωριστά. Η συχνότητα των δεδομένων που δίναμε κάθε φορά στους ταξινομητές μας για εκπαίδευση (training) ήταν τα πρώτα 50, 100, 250, 500, 1000, 1500, 2000 άτομα κατά σειρά προτεραιότητας όπως απάντησαν το ερωτηματολόγιο, ενώ τα δεδομένα που παίρναμε για τον έλεγχο (testing) της ορθότητας του αλγορίθμου ήταν ακριβώς τα επόμενα 500 άτομα κατά σειρά προτεραιότητας όπως απάντησαν το ερωτηματολόγιο. Με αυτό τον τρόπο καταγράψαμε την ορθότητα των αλγορίθμων δυναμικά. Όλα τα δεδομένα των χρηστών τα οποία χρησιμοποιήσαμε για την επιλογή του καλύτερου αλγόριθμου είναι 2598 χρήστες και είναι τα δεδομένα που πάρθηκαν από τις προεδρικές εκλογές 2009 . Στους χρήστες αυτούς δεν συμπεριλαμβάνονται οι χρήστες οι οποίοι απάντησαν στην ερώτηση της πρόθεσης ψήφου με το «Δεν ξέρω Δεν απαντώ» και το κόμμα «ΚΑΝΕΝΑ».

Με αυτή την διαδικασία επαληθεύσαμε ποιος αλγόριθμος κάνει καλύτερες προβλέψεις με μικρά δεδομένα, ποιος έχει καλύτερα αποτελέσματα για μεγάλα δείγματα και ποιος αλγόριθμος βελτιώνεται όσο μεγαλώνει το δείγμα μας.

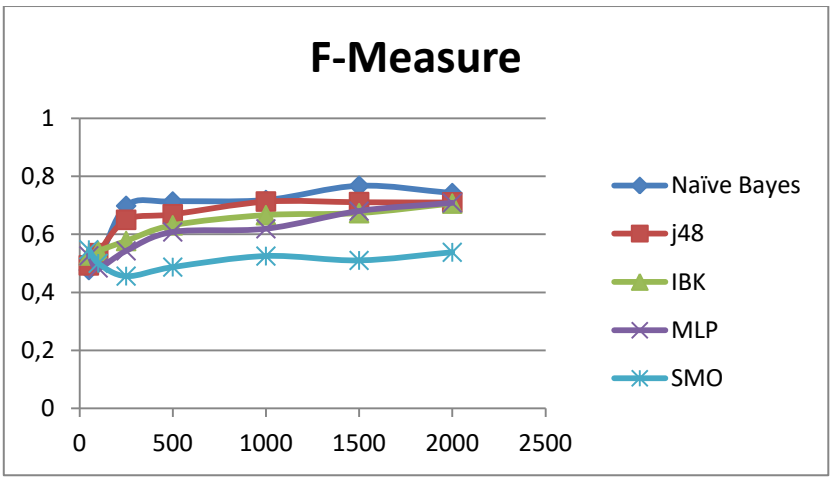


Γραφική Παράσταση 1 - Ορθότητα Αλγορίθμων, Κύπρος Βουλευτικές 2011

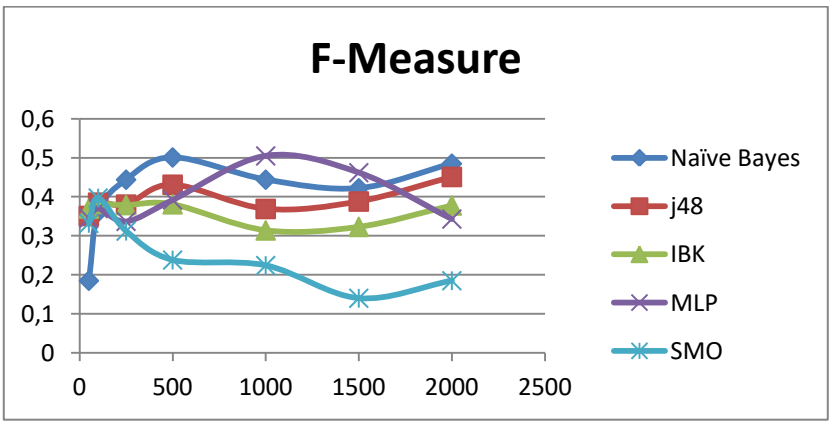
Πιο κάτω παρουσιάζονται τα αποτελέσματα των αλγορίθμων για τις χώρες Ελλάδα, Τουρκία και Βουλγαρία:



Γραφική Παράσταση 2 - Ορθότητα Αλγορίθμων, Ελλάδα Κοινοβουλευτικές Ιουνίου 2012



Γραφική Παράσταση 3 - Ορθότητα Αλγορίθμων, Τουρκία Κοινοβουλευτικές 2011



Γραφική Παράσταση 4 - Ορθότητα Αλγορίθμων, Βουλγαρία2013

Ένα γενικό συμπέρασμα το οποίο βγάζουμε αναλύοντας τις τέσσερις γραφικές παραστάσεις είναι ότι για μικρό δείγμα η ορθότητα των αλγορίθμων μας είναι μικρή ενώ όσο μεγαλώνει το δείγμα μας τα ποσοστά ορθότητας των αλγορίθμων αυξάνονται.

5.6.1 Ορθότητα Αλγορίθμων για Κύπρο

Για τα δεδομένα της Κύπρου ο αλγόριθμος Naïve Bayes για δείγμα εκπαίδευσης από 50 μέχρι 250 εμφανίζει μικρά ποσοστά ορθότητας δηλαδή ποσοστά από 0,36 μέχρι 0,504. Για δείγμα 500 μέχρι 2000 ατόμων το ποσοστό ορθότητας του αυξάνεται και κυμαίνεται από 0,517 μέχρι 0,544.

Ο αλγόριθμος J48 για δείγμα εκπαίδευσης από 50 μέχρι 250 άτομα το ποσοστό ορθότητας του είναι μικρό και κυμαίνεται από 0,415 μέχρι 0,408. Για δείγμα 500 μέχρι 2000 ατόμων το ποσοστό ορθότητας του αυξάνεται και κυμαίνεται από 0,439 μέχρι 0,499.

Ο αλγόριθμος IBK για δείγμα εκπαίδευσης από 50 μέχρι 250 άτομα το ποσοστό ορθότητας του είναι μικρό και κυμαίνεται από 0,395 μέχρι 0,367. Για δείγμα 500 μέχρι 2000 ατόμων το ποσοστό ορθότητας του αυξάνεται και κυμαίνεται από 0,452 μέχρι 0,475.

Ο αλγόριθμος Multilayer Perceptron (MLP) για δείγμα εκπαίδευσης από 50 μέχρι 250 άτομα το ποσοστό ορθότητας του κυμαίνεται από 0,432 μέχρι 0,317. Για δείγμα 500 μέχρι 2000 ατόμων το ποσοστό ορθότητας κυμαίνεται από 0,401 μέχρι 0,469.

Ο αλγόριθμος SMO για δείγμα εκπαίδευσης από 50 μέχρι 250 άτομα αρχικά το ποσοστό ορθότητας του είναι μεγάλο και σταδιακά μικραίνει δηλαδή από 0,445 κατεβαίνει στο 0,324. Για δείγμα 500 μέχρι 2000 ατόμων το ποσοστό ορθότητας αρχίζει να αυξάνεται σταδιακά και κυμαίνεται από 0,174 μέχρι 0,283. Όσο μεγαλώνει το δείγμα τόσο πιο μεγάλο είναι το ποσοστό ορθότητας του αλγορίθμου.

5.6.2 Ορθότητα Αλγορίθμων για Ελλάδα

Για τα δεδομένα της Ελλάδας ο αλγόριθμος Naïve Bayes για δείγμα εκπαίδευσης από 50 μέχρι 250 εμφανίζει μικρά ποσοστά ορθότητας δηλαδή ποσοστά από 0,372 μέχρι 0,43. Για δείγμα 500 μέχρι 2000 ατόμων το ποσοστό ορθότητας αυξάνεται κυμαίνεται από 0,521 μέχρι 0,528 .

Ο αλγόριθμος J48 για δείγμα εκπαίδευσης από 50 μέχρι 250 άτομα το ποσοστό ορθότητας του είναι μικρό και κυμαίνεται από 0,396 μέχρι 0,389. Για δείγμα 500 μέχρι 2000 ατόμων το ποσοστό ορθότητας κυμαίνεται από 0,633 μέχρι 0,402.

Ο αλγόριθμος IBK για δείγμα εκπαίδευσης από 50 μέχρι 250 άτομα το ποσοστό ορθότητας του είναι μικρό -κυμαίνεται από 0,413 μέχρι 0,372. Για δείγμα 500 μέχρι 2000 ατόμων το ποσοστό ορθότητας αυξάνεται και κυμαίνεται από 0,683 μέχρι 0,414.

Ο αλγόριθμος Multilayer Perceptron (MLP) για δείγμα εκπαίδευσης από 50 μέχρι 250 άτομα το ποσοστό ορθότητας του κυμαίνεται από 0,408 μέχρι 0,392. Για δείγμα 500 μέχρι 2000 ατόμων το ποσοστό ορθότητας κυμαίνεται από 0,38 μέχρι 0,482.

Ο αλγόριθμος SMO για δείγμα εκπαίδευσης από 50 μέχρι 250 άτομα αρχικά το ποσοστό ορθότητάς του είναι μεγάλο και σταδιακά μικραίνει δηλαδή από 0,405 μέχρι 0,381. Για δείγμα 500 μέχρι 2000 ατόμων το ποσοστό ορθότητας αρχίζει να αυξάνεται σταδιακά και κυμαίνεται από 0,307 μέχρι 0,572. Όσο μεγαλώνει το δείγμα τόσο πιο μεγάλο είναι το ποσοστό ορθότητας του αλγορίθμου.

5.6.3 Ορθότητα Αλγορίθμων για Τουρκία

Για τα δεδομένα της Τουρκίας ο αλγόριθμος Naïve Bayes για δείγμα εκπαίδευσης από 50 μέχρι 250 εμφανίζει μικρά ποσοστά ορθότητας δηλαδή ποσοστά από 0,477 μέχρι 0698. Για δείγμα 500 μέχρι 2000 ατόμων το ποσοστό ορθότητας αυξάνεται και κυμαίνεται από 0,713 μέχρι 0,742 .

Ο αλγόριθμος J48 για δείγμα εκπαίδευσης από 50 μέχρι 250 άτομα το ποσοστό ορθότητας του είναι μικρό και κυμαίνεται από 0,493 μέχρι 0,65. Για δείγμα 500 μέχρι 2000 ατόμων το ποσοστό ορθότητας κυμαίνεται από 0,669 μέχρι 0,71.

Ο αλγόριθμος IBK για δείγμα εκπαίδευσης από 50 μέχρι 250 άτομα το ποσοστό ορθότητας του είναι μικρό κυμαίνεται από 0,527 μέχρι 0577. Για δείγμα 500 μέχρι 2000 ατόμων το ποσοστό ορθότητας αυξάνεται και κυμαίνεται από 0,631 μέχρι 0,706.

Ο αλγόριθμος Multilayer Perceptron (MLP) για δείγμα εκπαίδευσης από 50 μέχρι 250 άτομα το ποσοστό ορθότητας του κυμαίνεται από 0,526 μέχρι 0,543. Για δείγμα 500 μέχρι 2000 ατόμων το ποσοστό ορθότητας κυμαίνεται από 0,608 μέχρι 0,708.

Ο αλγόριθμος SMO για δείγμα εκπαίδευσης από 50 μέχρι 250 άτομα αρχικά το ποσοστό ορθότητας του είναι μεγάλο και σταδιακά μικραίνει δηλαδή από 0,547 μέχρι 0,456. Για δείγμα

500 μέχρι 2000 ατόμων το ποσοστό ορθότητας αρχίζει να αυξάνεται σταδιακά και κυμαίνετε από 0,487 μέχρι 0,538. Όσο μεγαλώνει το δείγμα τόσο πιο μεγάλο είναι το ποσοστό ορθότητας του αλγορίθμου.

5.6.4 Ορθότητα Αλγορίθμων για Βουλγαρία

Για τα δεδομένα της Βουλγαρίας ο αλγόριθμος Naïve Bayes για δείγμα εκπαίδευσης από 50 μέχρι 250 εμφανίζει μικρά ποσοστά ορθότητας δηλαδή ποσοστά από 0,185 μέχρι 0444. Για δείγμα 500 μέχρι 2000 ατόμων το ποσοστό ορθότητας αυξάνεται κυμαίνεται από 0,501 μέχρι 0,485 .

Ο αλγόριθμος J48 για δείγμα εκπαίδευσης από 50 μέχρι 250 άτομα το ποσοστό ορθότητας του είναι μικρό και κυμαίνεται από 0,351 μέχρι 0,38. Για δείγμα 500 μέχρι 2000 ατόμων το ποσοστό ορθότητας κυμαίνεται από 0,431 μέχρι 0,451.

Ο αλγόριθμος IBK για δείγμα εκπαίδευσης από 50 μέχρι 250 άτομα το ποσοστό ορθότητας του είναι μικρό κυμαίνεται από 0,37 μέχρι 0,379. Για δείγμα 500 μέχρι 2000 ατόμων το ποσοστό ορθότητας αυξάνεται και κυμαίνεται από 0,381 μέχρι 0,378.

Ο αλγόριθμος Multilayer Perceptron (MLP) για δείγμα εκπαίδευσης από 50 μέχρι 250 άτομα το ποσοστό ορθότητας του κυμαίνεται από 0,345 μέχρι 0,337 . Για δείγμα 500 μέχρι 2000 ατόμων το ποσοστό ορθότητας αυξάνεται και κυμαίνεται από 0,392 μέχρι 0,343.

Ο αλγόριθμος SMO για δείγμα εκπαίδευσης από 50 μέχρι 250 άτομα αρχικά το ποσοστό ορθότητας του είναι μεγάλο και σταδιακά μικραίνει δηλαδή από 0,332 μέχρι 0,312. Για δείγμα 500 μέχρι 2000 ατόμων το ποσοστό ορθότητας κυμαίνεται από 0,328 μέχρι 0,185.

5.6.5 Κοινωνικές Συστάσεις (Social Recommendations) - Συμπεράσματα

Στόχος της διατριβής είναι να διαπιστωθεί μέχρι πιο σημείο εκπαιδεύονται ικανοποιητικά οι ταξινομητές μας. Ένα γενικό συμπέρασμα που εξάγεται αναλύοντας τις τέσσερις γραφικές παραστάσεις είναι ότι στα 1000 δεδομένα εκπαίδευσης το F-Measure των αλγορίθμων μας σχεδόν σταθεροποιείται, δηλαδή από ένα σημείο και μετά δεν φαίνεται να μεταβάλλεται ουσιαστικά, παρουσιάζει μόνο πολύ μικρές μεταβολές. Από αυτό συνεπάγεται ότι στα 1000

δεδομένα εκπαίδευσης ο αλγόριθμός μας εκπαιδεύεται ικανοποιητικά και μπορούμε να εξάγουμε ασφαλή συμπεράσματα..

Όπως αναφέρθηκε ήδη αναλύοντας τις τέσσερις γραφικές μας παραστάσεις καταλήγουμε ότι για μικρό δείγμα η ορθότητα του αλγορίθμου μας είναι μικρή ενώ όσο μεγαλώνει το δείγμα μας τα ποσοστά ορθότητας των αλγορίθμων μας αυξάνονται. Συγκρίνοντας τα δεδομένα (datasets) της κάθε χώρας παρατηρήσαμε ότι το F-Measure του κάθε αλγορίθμου διαφέρει σε κάθε χώρα.

Συγκρίνοντας γενικά τα αποτελέσματα του F-Measure της Κύπρου, Ελλάδας, Βουλγαρίας και Τουρκίας στο κρίσιμο σημείο των 1000 δειγμάτων παρατηρήσαμε ότι ο αλγόριθμος Naïve Bayes εμφανίζει τις πιο ψηλές τιμές στο F-Measure με εξαίρεση την Βουλγαρία όπου ο καλύτερος αλγόριθμος είναι ο MLP και ακολούθως ο αλγόριθμος Naïve Bayes με ελάχιστη τιμή διαφοράς.

Γενικά οι τιμές του F-Measure κυμαίνονται σε πολύ κοντινές μεταξύ τους τιμές και σε ανοδική πορεία με εξαίρεση το SMO όπου οι τιμές κυμαίνονται σε πολύ χαμηλά επίπεδα σε σύγκριση με τους άλλους τέσσερις και παρουσιάζει καθοδική πορεία και ελάχιστες αυξομειώσεις όσο εκπαιδεύεται. Ο αλγόριθμος που μας δίνει γενικά τις καλύτερες τιμές είναι στο F-Measure είναι ο Naïve Bayes και τις χαμηλότερες τιμές ο SMO.

5.6.5.1 Λεπτομερής Ανάλυση

Πιο κάτω αναλύεται λεπτομερώς η συμπεριφορά των αλγορίθμων σε κάθε χώρα.

5.6.5.1.1 Δεδομένα Κύπρου

Για τα δεδομένα της Κύπρου ο αλγόριθμος Naïve Bayes για λίγα δεδομένα εκπαίδευσης παρουσιάζει μικρό F-Measure. Ακολούθως όσο μεγαλώνει το δείγμα εκπαίδευσης, τόσο μεγαλώνει το F-Measure του. Από 50 μέχρι 500 δεδομένα εκπαίδευσης το F-Measure του παρουσιάζει μία αυξητική πορεία. Από 500 μέχρι 2000 δεδομένα εκπαίδευσης παρουσιάζει μια ελαφρώς αυξητική πορεία. Κάλλιστα μπορούμε να πούμε ότι από τα 500 μέχρι τα 2000 δεδομένα εκπαίδευσης το F-Measure του αλγορίθμου μας παραμένει σταθερό.

Ο αλγόριθμος J48 για 50 δεδομένα εκπαίδευσης παρουσιάζει μικρό F-Measure. Ακολούθως παρουσιάζει μια αυξητική πορεία μέχρι τα 1000 δεδομένα. Στα 1000 δεδομένα το F-Measure

είναι το δεύτερο πιο μεγάλο F-Measure από όλους τους αλγόριθμους. Από τα 1000 μέχρι τα 2000 δεδομένα εκπαίδευσης ο αλγόριθμός μας παρουσιάζει μια σταθερή πορεία και η τιμή του F-Measure δεν αλλάζει δραματικά.

Ο αλγόριθμος IBK για 50 δεδομένα εκπαίδευσης παρουσιάζει υψηλό ποσοστό F-Measure. Στη συνέχεια για 100 και 250 δεδομένα εκπαίδευσης παρουσιάζει πολύ χαμηλό F-Measure. Ακολούθως παρουσιάζει μία αυξητική πορεία μέχρι τα 500 δεδομένα εκπαίδευσης. Τέλος από τα 500 μέχρι τα 2000 δεδομένα οι τιμές του παραμένουν σχεδόν σταθερές.

Ο αλγόριθμος MLP από 50 μέχρι 250 δεδομένα εκπαίδευσης παρουσιάζει μία καθοδική πορεία. Στην συνέχεια από τα 250 μέχρι τα 1000 παρουσιάζει ανοδική πορεία. Από τα 1000 μέχρι τα 1500 παρουσιάζει ελαφρώς καθοδική πορεία και από τα 1500 μέχρι τα 2000 παρουσιάζει μια ελαφρώς ανοδική πορεία.

Ο αλγόριθμος SMO για 50 δεδομένα εκπαίδευσης εμφανίζει την πιο ψηλή τιμή στο F-Measure από όλους τους αλγόριθμους. Στην συνέχεια παρουσιάζει μία καθοδική πορεία μέχρι τα 500 άτομα. Η τιμή του F-Measure στα 500 δεδομένα είναι η πιο χαμηλή από όλους τους αλγόριθμους. Από τα 500 μέχρι τα 1000 παρουσιάζει μία μικρή άνοδο. Από τα 1000 μέχρι τα 1500 παρουσιάζει μια μικρή κάθοδο και από τα 1500 μέχρι τα 2000 πάλι μια μικρή άνοδο. Γενικά μπορούμε να πούμε ότι οι τιμές του αλγορίθμου μας είναι οι πολύ χαμηλές από όλους του υπόλοιπους αλγόριθμους.

5.6.5.1.2 Δεδομένα Βουλγαρίας

Για τα δεδομένα της Βουλγαρίας ο Naïve Bayes παρουσιάζει πολύ μικρό F-Measure για λίγα δεδομένα. Για 50 δεδομένα εκπαίδευσης το F-Measure του είναι το πιο χαμηλό από όλους τους αλγόριθμους. Ακολούθως το F-Measure του παρουσιάζει μία αυξητική πορεία μέχρι τα 500 δεδομένα. Το F-Measure του σε αυτό το σημείο είναι το πιο ψηλό από όλους τους άλλους αλγόριθμους. Από τα 500 μέχρι τα 1500 παρουσιάζει μία καθοδική πορεία και ακολούθως από τα 1500 μέχρι τα 2000 δεδομένα παρουσιάζει αυξητική τάση. Σε αυτό το σημείο μπορούμε να αναφέρουμε ότι από τα 1000 μέχρι τα 2000 δεδομένα το F-Measure του δεν αλλάζει δραματικά.

Ο αλγόριθμος J48 για λίγα δεδομένα εκπαίδευσης παρουσιάζει χαμηλό F-Measure. Στη συνέχεια παρουσιάζει μια ανοδική πορεία μέχρι τα 500 δεδομένα εκπαίδευσης. Από τα 500 μέχρι τα 1000

παρουσιάζει καθοδική πορεία και από τα 1000 μέχρι τα 2000 δεδομένα αρχίζει να παρουσιάζει ανοδική πορεία.

Ο αλγόριθμος IBK για λίγα δεδομένα παρουσιάζει ψηλό F-Measure. Στη συνέχεια μέχρι τα 500 δεδομένα εκπαίδευσης το F-Measure σχεδόν παραμένει σταθερό. Από 500 μέχρι 1000 παρουσιάζει καθοδική πορεία. Από 1000 μέχρι 2000 παρουσιάζει ανοδική πορεία. Στα 2000 δεδομένα η τιμή του F-Measure είναι σχεδόν η ίδια με τα αρχικά δεδομένα εκπαίδευσης.

Ο αλγόριθμος MLP για 50 δεδομένα εκπαίδευσης παρουσιάζει μικρό F-Measure. Στην συνέχεια παρουσιάζει μία αυξητική πορεία μέχρι τα 1000 δεδομένα εκπαίδευσης. Η τιμή του F-Measure στα 100 είναι η πιο ψηλή από όλους τους αλγόριθμους. Ακολούθως παρουσιάζει μία καθοδική πορεία μέχρι τα 2000 δεδομένα. Στα 2000 δεδομένα η τιμή του F-Measure του είναι η δεύτερη χειρότερη σε επίδοση.

Ο αλγόριθμος SMO για λίγα δεδομένα δηλαδή 50 δεδομένα εκπαίδευσης παρουσιάζει το ίδιο χαμηλό F-Measure με τους υπόλοιπους τέσσερις αλγόριθμους. Στην συνέχεια όσο εκπαιδεύεται, τόσο πιο χαμηλή τιμή δίνει το F-Measure του. Στα 2000 δεδομένα εκπαίδευσης έχει την πιο χαμηλή τιμή F-Measure από όλους τους αλγόριθμους.

Παρατηρώντας τη συμπεριφορά των 5 αλγόριθμων στην Βουλγαρία εξάγεται το συμπέρασμα ότι οι τέσσερις αλγόριθμοι (Naïve Bayes, J48, IBK, MLP) κυμαίνονται σε πολύ κοντινές μεταξύ τους τιμές και σε ανοδική πορεία με μικρές αυξομειώσεις. Οι αλγόριθμος SMO κυμαίνεται σε χαμηλότερα επίπεδα σε σύγκριση με τους άλλους τέσσερις με καθοδική πορεία. Ο αλγόριθμος που μας δίνει τις καλύτερα αποτελέσματα για 500 δεδομένα εκπαίδευσης είναι ο Naïve Bayes και για 1000 δεδομένα ο MLP. Όμως γενικά ο καλύτερος αλγόριθμος είναι ο Naïve Bayes γιατί από τα 500 μέχρι τα 2000 δεδομένα δεν παρατηρείται μεγάλη απόκλιση στην τιμή του F-Measure σε σύγκριση με τον MLP.

5.6.5.1.3 Δεδομένα Ελλάδας

Για τα δεδομένα της Ελλάδας το F-Measure του Naïve Bayes είναι μικρό για λίγα δεδομένα εκπαίδευσης. Στη συνέχεια ακολουθεί μια ανοδική πορεία μέχρι τα 500 δεδομένα εκπαίδευσης. Ακολούθως από τα 500 μέχρι τα 1000 δεδομένα εκπαίδευσης παρουσιάζει ελαφρώς καθοδική

πορεία. Από τα 1000 μέχρι τα 2000 το F-Measure του σχεδόν σταθεροποιείται με ελαφρώς ανοδική πορεία.

Ο αλγόριθμος J48 παρουσιάζει και αυτός χαμηλό F-Measure για λίγα δεδομένα. Ακολούθως αρχίζει μια αυξητική πορεία μέχρι τα 500 δεδομένα εκπαίδευσης. Η τιμή του F-Measure του είναι πιο μεγάλη από εκείνο του Naïve Bayes. Από τα 500 μέχρι τα 1000 δεδομένα το F-Measure του μειώνεται αρκετά, ενώ από τα 1000 μέχρι τα 2000 δεδομένα το F-Measure του μειώνεται ελάχιστα.

Ο αλγόριθμος IBK παρουσιάζει και αυτός χαμηλό F-Measure για λίγα δεδομένα εκπαίδευσης. Ακολούθως αρχίζει μια αυξητική πορεία μέχρι τα 500 δεδομένα εκπαίδευσης. Το F-Measure του στο σημείο αυτό είναι το πιο ψηλό από όλους τους αλγόριθμους. Στη συνέχεια παρουσιάζει μία καθοδική πορεία μέχρι τα 1000 δεδομένα εκπαίδευσης. Από τα 1000 μέχρι τα 2000 δεδομένα το F-Measure σχεδόν σταθεροποιείται.

Ο αλγόριθμος MLP παρουσιάζει σχεδόν μία σταθερή πορεία και για λίγα δεδομένα αλλά και για πολλά δεδομένα εκπαίδευσης. Από 50 μέχρι τα 1500 το F-Measure δεν παρουσιάζει δραματική αλλαγή. Από τα 1500 μέχρι τα 2000 το F-Measure του παρουσιάζει μια ανοδική πορεία.

Ο αλγόριθμος SMO για λίγα δεδομένα παρουσιάζει υψηλό F-Measure. Ακολούθως παρουσιάζει μια καθοδική πορεία μέχρι τα 1500. Από τα 1500 μέχρι τα 2000 δεδομένα εκπαίδευσης παρουσιάζει μια πολύ μεγάλη αυξητική πορεία. Η τιμή του F-Measure του στα 2000 είναι η πιο μεγάλη από όλους τους άλλους αλγόριθμους.

Παρατηρώντας την συμπεριφορά των 5 αλγόριθμων στην Ελλάδα εξάγεται το συμπέρασμα ότι και οι 5 αλγόριθμοι κυμαίνονται σε πολύ κοντινές μεταξύ τους τιμές. Ο αλγόριθμος που μας δίνει τις καλύτερες τιμές στο F-Measure είναι ο Naïve Bayes. Παρατηρείται επίσης ότι ο αλγόριθμος που μας δίνει τις χαμηλότερες τιμές είναι ο SMO. Όμως στα 2000 δεδομένα ο SMO μας δίνει την πιο ψηλή τιμή.

5.6.5.1.4 Δεδομένα Τουρκίας

Για τα δεδομένα της Τουρκίας ο Naïve Bayes εμφανίζει χαμηλό ποσοστό στο F-Measure. Ακολούθως παρουσιάζει μία ανοδική πορεία μέχρι τα 1000 δεδομένα εκπαίδευσης. Στο σημείο

αυτό έχει την πιο ψηλή τιμή από όλους τους αλγόριθμους. Στην συνέχεια από τα 1000 μέχρι τα 2000 το F-Measure του σχεδόν σταθεροποιείται με ελαφρώς ανοδική πορεία.

Ο αλγόριθμος J48 παρουσιάζει και αυτός χαμηλό F-Measure για λίγα δεδομένα, το οποίο έχει σχεδόν την ίδια τιμή με τους υπόλοιπους 4 αλγόριθμους. Ακολούθως αρχίζει μια ανοδική πορεία μέχρι τα 1000 δεδομένα εκπαίδευσης. Στην συνέχεια από τα 1000 μέχρι τα 2000 το F-Measure του σχεδόν σταθεροποιείται με ελαφρώς ανοδική πορεία.

Ο αλγόριθμος IBK παρουσιάζει και αυτός χαμηλό F-Measure για λίγα δεδομένα εκπαίδευσης. Ακολούθως αρχίζει μια ανοδική πορεία μέχρι τα 1000 δεδομένα εκπαίδευσης. Στην συνέχεια από τα 1000 μέχρι τα 2000 το F-Measure του σχεδόν σταθεροποιείται με ελαφρώς ανοδική πορεία.

Ο αλγόριθμος MLP παρουσιάζει και αυτός χαμηλό F-Measure για λίγα δεδομένα εκπαίδευσης. Ακολούθως αρχίζει μια ανοδική πορεία μέχρι τα 500 δεδομένα εκπαίδευσης. Από τα 500 μέχρι τα 1000 παρουσιάζει μία σταθερή πορεία και από τα 1000 μέχρι τα 2000 δεδομένα παρουσιάζει ανοδική πορεία.

Ο αλγόριθμος SMO για 50 δεδομένα εκπαίδευσης η τιμή του F-Measure είναι αρκετά ικανοποιητική σε σύγκριση με τις τιμές των άλλων αλγορίθμων. Ακολούθως παρουσιάζει μία καθοδική πορεία μέχρι τα 500 άτομα. Από τα 500 μέχρι τα 1000 παρουσιάζει ανοδική πορεία και από τα 1000 μέχρι τα 1500 παρουσιάζει ελαφρώς ανοδική πορεία.

Παρατηρώντας την συμπεριφορά των 5 αλγόριθμων στην πιο πάνω χώρα εξάγεται το συμπέρασμα ότι και οι 5 αλγόριθμοι κυμαίνονται σε πολύ κοντινές μεταξύ τους τιμές. Ο αλγόριθμος που μας δίνει τις καλύτερες τιμές στο F-Measure είναι ο Naïve Bayes. Παρατηρείται επίσης ότι ο αλγόριθμος που μας δίνει τις χαμηλότερες τιμές είναι ο SMO.

Κεφάλαιο 6
Επίλογος

Ολοκληρώνοντας την διατριβή θα επιχειρήσουμε να κάνουμε μια σύνοψη των σημαντικότερων σημείων που διακρίναμε σε κάθε κεφάλαιο.

Στο Κεφάλαιο 1 κάναμε μια εισαγωγή στα συστήματα Ηλεκτρονικοί Σύμβουλοι Ψήφου γνωστά ως Voting Advice Applications, για το πως λειτουργούν και που εφαρμόζονται.

Στο Κεφάλαιο 2 «Μηχανική Μάθηση και Ταξινόμηση» είδαμε τον ορισμό της μηχανικής μάθησης και ποιες κατηγορίες αλγορίθμων. Επιπλέον μελετήσαμε τι είναι ταξινόμηση δεδομένων, σε ποιες φάσεις χωρίζεται (Training – Testing), πως κατηγοριοποιούνται τα δεδομένα μας, ποιοι αλγόριθμοι είναι υπεύθυνοι για αυτή την ταξινόμηση (Naïve Bayes, Multi Layer Perceptron, J48, SMO και IBK) καθώς και την ορθότητα των αλγορίθμων μας.

Στο Κεφάλαιο 3 «Voting Advice Applications» αναλύσαμε εκτενέστερα τι είναι Ηλεκτρονικοί Σύμβουλοι Ψήφου, πού εφαρμόστηκαν, την δομή των ερωτήσεων που κλήθηκε να απαντήσει ο εκάστοτε ψηφοφόρος καθώς επίσης και τα διαγράμματα για τα ποσοστά συνάφειας με τα διάφορα πολιτικά κόμματα τα οποία προέκυψαν από τα VAA. Ακολούθως μελετήσαμε πώς

λειτουργεί ο μηχανισμός των VAA (Voting Advice Applications) και σε ποιες χώρες εφαρμόστηκε αυτό το σύστημα.

Στο κεφάλαιο 4 «Δεδομένα – Datasets» αναφερθήκαμε στην μορφή των δεδομένων που χρησιμοποιήσαμε (CSV αρχείο), από που βρήκαμε τα δεδομένα που αναλύσαμε (http://www.preferencematcher.org), πόσοι χρήστες συνολικά πήραν μέρος στο ερωτηματολόγιο, τον αριθμό των ερωτήσεων καθώς επίσης και τη μορφή των απαντήσεων («Συμφωνώ πλήρως», «Συμφωνώ», «Ούτε συμφωνώ, ούτε διαφωνώ», «Διαφωνώ», «Διαφωνώ πλήρως», «Χωρίς άποψη»). Τέλος είδαμε πως εφαρμόστηκαν τα VAA για τα δεδομένα της Κύπρου και ενδεικτικά κάποια παραδείγματα.

Στο κεφάλαιο 5 «Πειραματική Αξιολόγηση» παραμετροποιήσαμε κατάλληλα τα δεδομένα μας και τα επεξεργαστήκαμε με το κατάλληλο λογισμικό πρόγραμμα δημιουργώντας τα δέντρα απόφασης, αξιολογήσαμε δυναμικά τους 5 αλγόριθμους ταξινόμησης μεταξύ τους για να δούμε ποιος είναι ο καλύτερος όσο αφορά τα ποσοστά ορθότητας τους, εφαρμόσαμε αυτούς τους 5 αλγόριθμους και σε δεδομένα άλλων χωρών (Βουλγαρία, Ελλάδα, Τουρκία) μελετώντας επίσης την ορθότητα τους. Ακολούθως εξαγάγαμε κάποια συμπεράσματα που προκύπτουν από τη συμπεριφορά των αλγορίθμων μας καθώς επίσης και τον ελάχιστο αριθμό δεδομένων που χρειάζονται για να εκπαιδεύσουμε ένα ταξινομητή.

Συνοψίζοντας το γενικό συμπέρασμα που προκύπτει από αυτή τη διατριβή είναι ότι σε συστήματα ηλεκτρονικών συμβούλων ψήφων (VAA) ο καλύτερος αλγόριθμος για τη δυναμική επεξεργασία των δεδομένων είναι ο αλγόριθμος Naïve Bayes. Επιπλέον ο ελάχιστος αριθμός δεδομένων που χρειάζονται για να εκπαιδεύσουν κάποιο ταξινομητή είναι τα 1000 δεδομένα εκπαίδευσης.

6.1 Μελλοντική Εργασία

Έχοντας υπόψη τα πιο πάνω γενικά συμπεράσματα που εξαγάγαμε θα μπορούσαμε να επεκτείνουμε την πιο πάνω διατριβή ως ακολούθως.

Έχοντας ως δεδομένο το ελάχιστο σημείο εκπαίδευσης του αλγορίθμου μας θα μπορούμε να παράγουμε συστάσεις στο χρήστη με ποιο υποψήφιο συγκλίνουν οι απόψεις του με βάση τις απαντήσεις των άλλων χρηστών και όχι με αυτές των πολιτικών κομμάτων.

Επιπλέον να έχουμε την δυνατότητα να δίνουμε ανακατευθύνσεις στο χρήστη και πληροφορίες για το τι απάντησαν οι υπόλοιποι χρήστες μέχρι εκείνο το σημείο και σε ποιο υποψήφιο τείνει η μάζα των ψηφοφόρων για τις ανάλογες απαντήσεις.

Ακολούθως να τροποποιήσουμε τον κώδικα των αλγορίθμων μας ούτως ώστε να γίνει πιο βέλτιστος ως προς το ελάχιστο σημείο εκπαίδευσης και εμφανίζοντας καλύτερα ποσοστά ορθότητας. Αυτό θα είχε ως αποτέλεσμα τη βελτιστοποίηση του εκάστοτε αλγορίθμου εξοικονομώντας μας χρόνο και πόρους.

Βιβλιογραφία

[1]Ι. Βλαχάβας, Π. Κεφαλάς, Ν. Βασιλειάδης, Φ. Κόκκορας και Η. Σακελλαρίου. *Τεχνητή Νοημοσύνη*, Β' Έκδοση, 2006 σελ.508, σελ.522, σελ51

[2]Han J. and Kamber M*., Data Mining: Concepts and Techniques*, 2nd ed., San Francisco, Morgan Kauffmann Publishers,2001. Σελ. 287 , 328.

[3]Margaret H. Danham,S . Sridhar, " Data mining, Introductory and Advanced Topics", Person education , 1st ed., 2006.

[4]Aman Kumar Sharma, Suruchi Sahni, "A Comparative Study of Classification Algorithms for Spam Email Data Analysis", IJCSE, Vol. 3, No. 5, 2011, pp. 1890-1895

[5]William E. Spangler, Jerrold H. May and Luis G. Vargas Journal of Management Information Systems Vol. 16, No. 1 (Summer, 1999), pp. 37-62

[6]Williams, Ronald J. *(1987).* "A class of gradient-estimating algorithms for reinforcement learning in neural networks". *Proceedings of the IEEE First International Conference on Neural Networks.*

[7] William E. Spangler, Jerrold H. May and Luis G. Vargas, Journal of Management Information Systems Vol. 16, No. 1 (Summer, 1999), pp. 37-62

[8] *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines* , Platt, John (1998)

[9]*Nigsch F, Bender A, van Buuren B, Tissen J, Nigsch E, Mitchell JB (2006)"Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization" Journal of Chemical Information and Modeling. 46 (6): 2412–2422.*

[10] Altman, N. S. (1992)"An introduction to kernel and nearest-neighbor nonparametric regression" *The American Statistician*. 46 (3): 175–185

[11]Foster Provost (1998). «Glossary of terms». Machine Learning 30: 271–274.

[12]Wernick, Yang, Brankov, Yourganov and Strother, Machine Learning in Medical Imaging, *IEEE Signal Processing Magazine*, vol. 27, no. 4, July 2010, pp. 25-38

[13] Wernick, Yang, Brankov, Yourganov and Strother, Machine Learning in Medical Imaging, IEEE Signal Processing Magazine, vol. 27, no. 4, July 2010, pp. 25-38

[14]Sutton, Richard S. *(1984).* Temporal Credit Assignment in Reinforcement Learning *(PhD thesis). University of Massachusetts, Amherst, MA.*

[15]Francis, Bob. "SNIA nails down ILM definition." InfoWorld. 1 November 2004: 14

[16]Josh Judd and Dan Kruger (2005), Principles of SAN Design. Infinity Publishing

[17] Josh Judd and Dan Kruger (2005), Principles of SAN Design. Infinity Publishing

[18]Ι. Βλαχάβας, Π. Κεφαλάς, Ν. Βασιλειάδης, Φ. Κόκκορας και Η. Σακελλαρίου. *Τεχνητή Νοημοσύνη*, Β' Έκδοση, 2006 σελ.335-343

[19] Pierre Lison, Language Technology Group (LTG) Department of Informatics, An introduction to machine learning, HiOA, October 3 2012

[20]Norvig, Peter (2003) [1995]. *Artificial Intelligence: A Modern Approach* (2nd ed.). Prentice Hall. ISBN 978-0137903955.

[21] Rosenblatt, Frank. x. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan Books, Washington DC, 1961

[22]J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993

[23]John (1998), *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*

[24]Vapnik, V. (1995). "Support-vector networks". *Machine Learning*. 20 (3): 273–297

[25]Ben-Hur, Asa, Horn, David, Siegelmann, Hava, and Vapnik, Vladimir; "Support vector clustering" (2001) Journal of Machine Learning Research, 2: 125–137

[26]Altman, N. S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". *The American Statistician*. 46 (3): 175–185

[27]Campello, R. J. G. B. "Comparing Correlation Coefficients as Dissimilarity Measures for Cancer Classification in Gene Expression Data"

[28]D.L. Massart (1982). "Alternative k-nearest neighbour rules in supervised pattern recognition : Part 1. k-Nearest neighbour classification by using alternative voting rules". *Analytica Chimica Acta*. 136: 15–27

[29]Banerjee, A; Chitnis, UB; Jadhav, SL; Bhawalkar, JS; Chaudhury, S (2009). "Hypothesis testing, type I and type II errors". *Ind Psychiatry J*. 18: 127–31

[30] Phil Simon (March 18, 2013). Too Big to Ignore: The Business Case for Big Data. Wiley, σελ. 89. ISBN 978-1-118-63817-0.

[31]*Jeffreys, Harold (1973). Scientific Inference (3rd ed.). Cambridge University Press. p. 31. ISBN 978-0-521-18078-8.*

[32]Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). *The Elements of Statistical Learning* p. 348. [I]n data mining applications the interest is often more in the class probabilities themselves, rather than in performing a class assignment.

[33]Rumelhart, David E. Geoffrey E. Hinton, and R. J. Williams. "Learning Internal Representations by Error Propagation". David E. Rumelhart, James L. McClelland, and the PDP research group. (Editors), Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundations. MIT Press, 1986.

[34]Cybenko, G. 1989. Approximation by superpositions of a sigmoidal function *Mathematics of Control, Signals, and Systems*, 2(4), 303–314.

[35]Altman, N. S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". *The American Statistician*. 46 (3): 175–185

[36]Platt, John. Fast Training of Support Vector Machines using Sequential Minimal Optimization, in Advances in Kernel Methods – Support Vector Learning, B. Scholkopf, C. Burges, A. Smola, eds., MIT Press (1998).

[37] *"False Positive". WhatIs.com. Retrieved 26 August 2016*

[38]Banerjee, A; Chitnis, UB; Jadhav, SL; Bhawalkar, JS; Chaudhury, S (2009). "Hypothesis testing, type I and type II errors". *Ind Psychiatry* J. 18: 127–31

[39]*Tang B, He H (2015). "ENN: Extended Nearest Neighbor Method for Pattern Recognition [Research Frontier]". IEEE Computational Intelligence Magazine. 10 (3): 52–60*

Παράρτημα
Αποτελέσματα Αλγορίθμων

Πιο κάτω παρατίθενται τα αποτελέσματα του F-Measure από την εκπαίδευσης των αλγορίθμων μας για τις χώρες Κύπρου, Βουλγαρίας, Ελλάδας και Τουρκίας.

Α.1Πίνακας Αποτελεσμάτων Αλγορίθμου Naïve Bayes

Testing Data	F-Measure Κύπρου	F-Measure Βουλγαρίας	F-Measure Ελλάδας	F-Measure Τουρκίας
Training set 1 – 50 Testing set 51-550	0,425	0,185	0,372	0,477
Training set 1 – 100 Testing set 101-650	0,432	0,361	0,4	0,502
Training set 1 – 250 Testing set 251-750	0,444	0,444	0,43	0,698
Training set 1 – 500 Testing set 501-1000	0,517	0,501	0,521	0,713
Training set 1-1000 Testing set 1001-1500	0,532	0,444	0,491	0,718
Training set 1- 1500 Testing 1501-2000	0,52	0,422	0,487	0,767
Training set 1- 2000 Testing set 2001-2500	0,544	0,485	0,528	0,742

Πίνακας 8 - Αποτελέσματα Αλγορίθμου Naive Bayes

Α.2 Πίνακας Αποτελεσμάτων Αλγορίθμου J48

Testing Data	F-Measure Κύπρου	F-Measure Βουλγαρίας	F-Measure Ελλάδας	F-Measure Τουρκίας
Training set 1 – 50 Testing set 51-550	0,462	0,351	0,396	0,493
Training set 1 – 100 Testing set 101-650	0,401	0,385	0,353	0,535
Training set 1 – 250 Testing set 251-750	0,408	0,38	0,389	0,65
Training set 1 – 500 Testing set 501-1000	0,439	0,431	0,633	0,669
Training set 1-1000 Testing set 1001-1500	0,495	0,369	0,463	0,713
Training set 1- 1500 Testing 1501-2000	0,486	0,388	0,404	0,711
Training set 1- 2000 Testing set 2001-2500	0,499	0,451	0,402	0,71

Πίνακας 9 - Αποτελέσματα Αλγορίθμου J48

Α.3 Πίνακας Αποτελεσμάτων Αλγορίθμου IBK

Testing Data	F-Measure Κύπρου	F-Measure Βουλγαρίας	F-Measure Ελλάδας	F-Measure Τουρκίας
Training set 1 – 50 Testing set 51-550	0,52	0,37	0,413	0,527
Training set 1 – 100 Testing set 101-650	0,267	0,382	0,384	0,542
Training set 1 – 250 Testing set 251-750	0,367	0,379	0,372	0,577
Training set 1 – 500 Testing set 501-1000	0,452	0,381	0,683	0,631
Training set 1-1000 Testing set 1001-1500	0,448	0,314	0,419	0,666
Training set 1- 1500 Testing 1501-2000	0,436	0,323	0,411	0,673
Training set 1- 2000 Testing set 2001-2500	0,475	0,378	0,414	0,706

Πίνακας 10 - Αποτελέσματα Αλγορίθμου IBK

Α.4 Πίνακας Αποτελεσμάτων Αλγορίθμου MLP

Testing Data	F-Measure Κύπρου	F-Measure Βουλγαρίας	F-Measure Ελλάδας	F-Measure Τουρκίας
Training set 1 – 50 Testing set 51-550	0,44	0,345	0,408	0,526
Training set 1 – 100 Testing set 101-650	0,387	0,37	0,436	0,484
Training set 1 – 250 Testing set 251-750	0,317	0,337	0,392	0,543
Training set 1 – 500 Testing set 501-1000	0,401	0,392	0,38	0,608
Training set 1-1000 Testing set 1001-1500	0,464	0,505	0,36	0,619
Training set 1- 1500 Testing 1501-2000	0,444	0,462	0,376	0,68
Training set 1- 2000 Testing set 2001-2500	0,444	0,343	0,482	0,708

Πίνακας 11 - Αποτελέσματα Αλγορίθμου MLP

Α.5Πίνακας Αποτελεσμάτων Αλγορίθμου SMO

Testing Data	F-Measure Κύπρου	F-Measure Βουλγαρίας	F-Measure Ελλάδας	F-Measure Τουρκίας
Training set 1 – 50 Testing set 51-550	0,552	0,332	0,405	0,547
Training set 1 – 100 Testing set 101-650	0,314	0,397	0,429	0,5
Training set 1 – 250 Testing set 251-750	0,324	0,312	0,381	0,456
Training set 1 – 500 Testing set 501-1000	0,174	0,238	0,307	0,487
Training set 1-1000 Testing set 1001-1500	0,225	0,224	0,317	0,525
Training set 1- 1500 Testing 1501-2000	0,211	0,14	0,218	0,51
Training set 1- 2000 Testing set 2001-2500	0,283	0,185	0,572	0,51

Πίνακας 12 - Αποτελέσματα Αλγορίθμου SMO