ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΗΠΕΙΡΟΥ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ Τ.Ε.

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

«Εξεύρεση συσχετίσεων με χρήση εξόρυξης δεδομένων σε δεδομένα πολιτικοκοινωνικών απόψεων και ιδεολογίας»

ΚΛΕΑΝΘΙΝΗ ΣΧΙΖΑ - Α.Μ.10674

Επιβλέπων καθηγητής

Νικόλαος Γιαννακέας

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ Τ.Ε.

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Εξεύρεση συσχετίσεων με χρήση εξόρυξης δεδομένων σε δεδομένα πολιτικοκοινωνικών απόψεων και ιδεολογίας

ΚΛΕΑΝΘΙΝΗ ΣΧΙΖΑ- Α.Μ.10674

Επιβλέπων Καθηγητής

Νικόλαος Γιαννακέας

- Άρτα 2016 -

ΕΥΧΑΡΙΣΤΙΕΣ

Επιθυμώ να εκφράσω όλο μου το σεβασμό και την ευγνωμοσύνη στον επιβλέποντα καθηγητή μου κ.Γιαννακέα Ν., ο οποίος με εμπιστεύθηκε και με καθοδήγησε καθ' όλη τη διάρκεια εκπόνησης της πτυχιακής μου εργασίας. Θα ήθελα επίσης να ευχαριστήσω όλους τους καθηγητές του τμήματος Μηχανικών Πληροφορικής Τ.Ε. , για τις πολύτιμες γνώσεις που μου παρείχαν κατά τη διάρκεια της φοίτησής μου.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου για την αμέριστη συμπαράσταση και ηθική στήριξη που μου πρόσφερε.

Περίληψη

Στην παρούσα πτυχιακή εργασία πραγματοποιήθηκε μια καινοτόμα έρευναγια την συσχέτιση των απόψεων των πολιτών σε κοινωνικοπολιτικά θέματα σε σχέση με την ιδεολογία τους, αξιοποιώντας τεχνικές μηχανικής μάθησης. Για την σκιαγράφηση των απόψεων-ιδεολογιών της κοινής γνώμης καταρτήθηκε ηλεκτρονικό ερωτηματολόγιο, το οποίο μοιράστηκεσε εκατοντάδες άτομα μέσω ηλεκτρονικού ταχυδρομείου και κοινωνικών δικτύων. Τα δεδομένα που συλλέχτηκαναπό το ερωτηματολόγιο (δείγμα 150 ατόμων), επεξεργάστηκαν με τεχνικές ταξινόμησης με σκοπό να αναδειχθεί η πιθανότητα πρόβλεψης της ιδεολογίας ενός ατόμου, δοθέντων των αντιλήψεων του σε θέματα της επικαιρότητας.Συγκεκριμένα, έγινε χρήση τριών αλγορίθμων: α)του δένδρου απόφασηςC4.5, των μηχανών διανυσμάτων υποστήριξης (SVM) καιτου ταξινομητήNaïveBayes. Στο παρόν σύγγραμμα αναλύονται οι βασικές έννοιες της εξόρυξης δεδομένων και του λογισμικούweka που χρησιμοποιήθηκε για την επεξεργασία των δεδομένων, ενώ παρουσιάζονται ποσοτικά τα αποτελέσματα του ερωτηματολογίου. Μέσα από την διερεύνηση των αποτελεσμάτων διαπιστώθηκε ότι με την χρήση του αλγορίθμουNaïveBayesεπετεύχθη η υψηλότερη ακρίβεια σε σχέση με τους άλλους δύο αλγορίθμους. Συμπερασματικά, διαπιστώθηκε ότι η ακρίβεια των αποτελεσμάτων είναι χαμηλή,αναδεικνύονταςτο γεγονός ότι η συσχέτιση της ιδεολογίας ενός ατόμου με την αντίληψη του στα καθημερινά κοινωνικοπολιτικά προβλήματα δεν είναι πολύ ισχυρή.

Abstract

The current work presents an innovative investigation of the correlation between human sociopolitical aspects and their ideology, using machine learning techniques. In order to collect human opinions, regarding several sociopolitical issues and ideologies, an informative questionnaire is developed in google form. The questionnaire has sent to hundreds of peoples via email and social media, so that 150 responses are collected. Classification techniques are employed, in order to extract the prediction of human ideology, given their positions in current sociopolitical issues. More specifically, Naïve Bayes classifier, Support Vector Machines and C4.5 decision tree have been used. The manuscript describes in details the methodology. According to the conclusions, the association between person ideology and his sociopolitical aspects seems to be weak, due to low accuracy results of classification.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 3

Μεθοδολογία εργασίας

ΚΕΦΑΛΑΙΟ 4

Πειράματα και αποτελέσματα

ΕΙΣΑΓΩΓΗ

Στην παρούσα πτυχιακή εργασία διερευνήθηκαν οι απόψεις και οι ιδεολογίες της κοινής γνώμης σε ζητήματα της επικαιρότητας. Συγκεκριμένα, μελετήθηκε η άποψη της κοινής γνώμης σε κοινωνικά και πολιτικά θέματα, με σκοπό την ανάλυση, την επεξεργασία και την παρουσίαση των αποτελεσμάτων μέσα από την χρήση διαφόρων μεθόδων και τεχνικών της εξόρυξης δεδομένων.

Προκειμένου να καταγραφεί η κοινή γνώμη στα κοινωνικοπολιτικά θέματα, σχεδιάστηκε ένα ηλεκτρονικό ερωτηματολόγιο το οποίο μοιράστηκε με τη βοήθεια των ηλεκτρονικών μέσων. Στη συνέχεια χρησιμοποιήθηκε το γραφικό περιβάλλον weka, ώστε να γίνει η επεξεργασία των δεδομένων μέσα από τη χρήση των τριών αλγορίθμων J48, SMO και NaïveBayes. Συγκεκριμένα, δημιουργήθηκαν δύο αρχεία .arff, όπου η μόνη διαφορά τους ήταν το χαρακτηριστικό @attributeideology, όπου την πρώτη φορά είχε ως όρισμα πέντε κλάσεις και τη δεύτερη φορά είχε ως όρισμα τρεις κλάσεις. Και στις δύο περιπτώσεις, υλοποιήθηκαν οι τρεις αλγόριθμοι με χρήση όλων των χαρακτηριστικών καθώς επίσης και με επιλογή ορισμένων χαρακτηριστικών. Μέσα από αυτή τη διαδικασία εντοπίστηκε ότι ο βέλτιστος αλγόριθμος με το μεγαλύτερο ποσοστό των σωστά ταξινομημένων στοιχείων είναι ο NaïveBayes. Επίσης, διαπιστώθηκε ότι η ακρίβεια των δεδομένων είναι πολύ χαμηλή και ότι το καλύτερο χαρακτηριστικό από τα στοιχεία είναι η ηλικία.

Πιο αναλυτικά, στο πρώτο κεφάλαιο αναλύεται ο ορισμός και η σημασία της εξόρυξης δεδομένων καθώς επίσης και ο ορισμός της μηχανικής μάθησης. Στο δεύτερο κεφάλαιο γίνεται αναφορά στους τρεις αλγορίθμους που χρησιμοποιήθηκαν για την μελέτη των δεδομένων: J48, SMO και NaïveBayes, ενώ παράλληλα παρουσιάζεται και ερμηνεύεται ο πίνακας σύγχυσης. Στη συνέχεια στο τρίτο κεφάλαιο, παρουσιάζεται αναλυτικά η μεθοδολογία της εργασίας. Αρχικά γίνεται μία ανάλυση στις ερωτήσεις του ερωτηματολογίου, στη συνέχεια περιγράφεται το περιβάλλον weka που χρησιμοποιήθηκε, ενώ τέλος αναλύεται η δομή εγγραφής ενός αρχείου weka. Στο τέταρτο κεφάλαιο παρουσιάζονται ποσοτικά τα αποτελέσματα του ερωτηματολογίου , ενώ επιτυγχάνεται μία πλήρη ανάλυση του περιβάλλοντος weka μέσα από ένα στιγμιότυπο. Στο ίδιο κεφάλαιο αναλύονται και περιγράφονται όλα τα αποτελέσματα του ερωτηματολογίου από τους τρεις αλγορίθμους με τη χρήση

διαφορετικών ορισμάτων, καθώς επίσης και κάποια συμπεράσματα που σχετίζονται με τους αλγορίθμους.

Κεφάλαιο 1

Μηχανική μάθηση και εξόρυξη δεδομένων

1.1 Ορισμός Εξόρυξη Δεδομένων

Η εξόρυξη δεδομένων (DataMining) ή αλλιώς εξόρυξη γνώσης (KnowledgeDiscovery) σε βάσεις δεδομένων, είναι μια νέα δυναμική τεχνολογία που βοηθάει τις επιχειρήσεις να εστιάσουν στην σημαντική πληροφορία που βρίσκεται μέσα στις αποθήκες δεδομένων τους (datawarehouses).Οι τεχνικές της είναι σε θέση να αναζητήσουν και να βρουν βάσεις δεδομένων για την αναζήτηση κρυμμένων προτύπων (patterns) (Παρασύρη, 2014).

Είναι μία διαδικασία μη τετριμμένης εξαγωγής άγνωστων όπως συσχετίσεις, κανόνες γνώσης, πρότυπα, κανονικότητες και εύρεσης πληροφοριών από ένα μεγάλο όγκο δεδομένων (Λαδάς,2014). Ένας ευρύς ορισμός για την εξόρυξη γνώσης θα μπορούσε να είναι η διαδικασία ημι-αυτόματης ανάλυσης μεγάλων βάσεων δεδομένων με στόχο την εύρεση χρήσιμης πληροφορίας- «γνώσης» (Ταράτσα, 2011).

Οι βασικές τεχνικές εξόρυξης δεδομένων περιλαμβάνουν αλγορίθμους για κατηγοριοποίηση, συσταδοποίηση, παλινδρόμηση, ανακάλυψη ακολουθιακών προτύπων, εύρεση συσχετίσεων και πρόβλεψη συμβάντων (Λαδάς,2014).

Η εξόρυξη δεδομένων αποτελείται από πέντε βασικά βήματα υλοποίησης :

- Επιλογή δεδομένων
- Προεπεξεργασία δεδομένων
- Μετασχηματισμός δεδομένων
- Εξόρυξη δεδομένων / Εξόρυξη γνώσης
- Ερμηνεία της εξαγόμενης γνώσης

Εικόνα 1: Βήματα υλοποίησης εξόρυξης δεδομένων

Πιο αναλυτικά, αρχικά συλλέγονται δεδομένα από διάφορες βάσεις δεδομένων, αρχεία και μη ηλεκτρονικές πηγές (στάδιο επιλογής δεδομένων) και στη συνέχεια γίνεται εκτίμηση ελλιπών δεδομένων ή ακόμα και διόρθωση λανθασμένων δεδομένων (στάδιο προεπεξεργασίας δεδομένων). Ακολουθεί το στάδιο του μετασχηματισμού δεδομένων, όπου όλα τα δεδομένα μετατρέπονται σε ένα κοινό σχήμα ώστε να υπάρξει μία περαιτέρω επεξεργασία και ορισμένα από τα δεδομένα είτε κωδικοποιούνται, είτε μετασχηματίζονται. Στη συνέχεια με βάση το είδος της εξόρυξης που πρόκειται να εκτελεσθεί, εφαρμόζονται διάφοροι αλγόριθμοι στα τροποποιημένα δεδομένα, προκειμένου να προκύψουν τα αποτελέσματα (στάδιο εξόρυξης δεδομένων ή στάδιο εξόρυξης γνώσης από τα δεδομένα). Το τελευταίο βήμα υλοποίησης της εξόρυξης δεδομένων είναι η ερμηνεία της εξαγόμενης γνώσης, όπου στο στάδιο αυτό παρουσιάζονται και αξιολογούνται όλα τα αποτελέσματα μέσα από τη χρήση διαφόρων στρατηγικών οπτικοποίησης (Κεχαγιά, 2006).

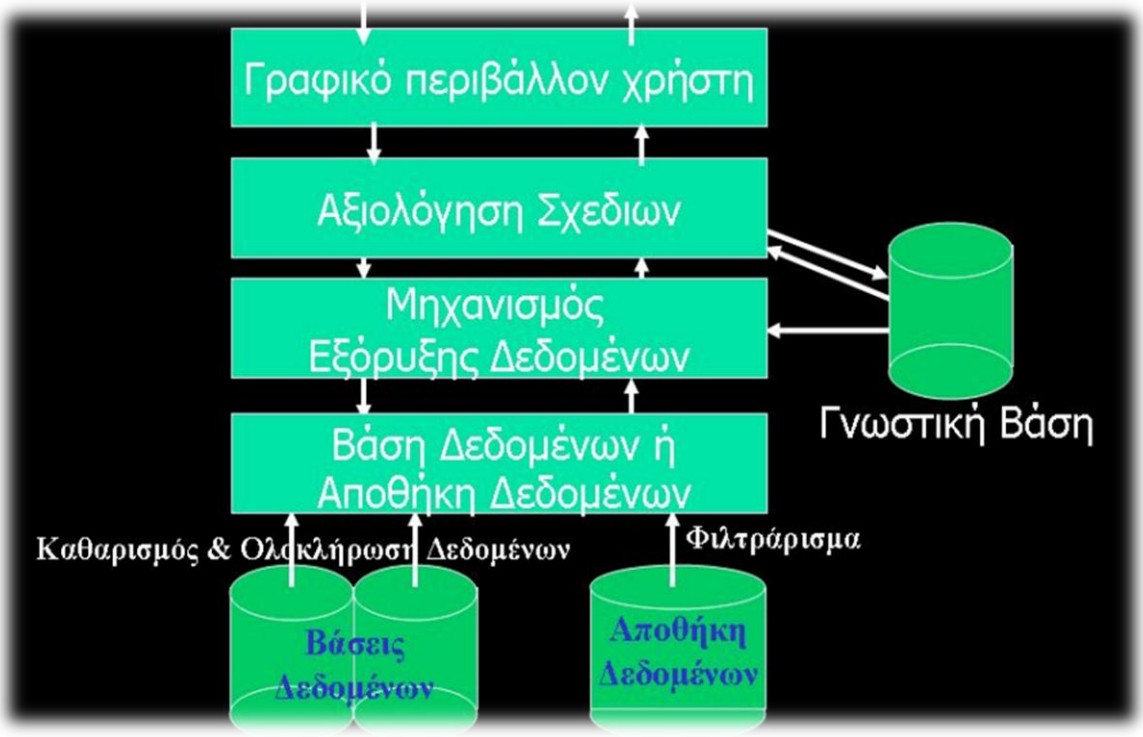
1.1.1 Πρότυπα (Patterns)

Τα πρότυπα, είναι ιδιαίτερα σημαντικά για το πεδίο της εξόρυξης δεδομένων. Οι αλγόριθμοι αναζητούν πρότυπα σε σύνολα από δεδομένα, τα οποία είναι έγκυρα και

χρήσιμα ώστε να αποκαλούνται γνώση. Σύμφωνα με τον Frawley, ένα πρότυπο σε ένα σύνολο δεδομένων είναι μια δήλωση που περιγράφει συσχετίσεις σε ένα υποσύνολο του συνόλου δεδομένων με κάποια βεβαιότητα, έτσι ώστε η δήλωση να είναι με κάποιο τρόπο περισσότερο απλή. Ο ίδιος τύπος προτύπου, μπορεί να χρησιμοποιηθεί σε διαφορετικούς αλγορίθμους εξόρυξης δεδομένων που αντιμετωπίζουν διαφορετικά προβλήματα (Ντάλλα, 2009).

1.2 Η σημασία της Εξόρυξης Δεδομένων

Η εξόρυξη δεδομένων είναι η εξαγωγή κρυμμένων πληροφοριών πρόβλεψης μέσα από μεγάλες βάσεις δεδομένων, όπου μέσα από την ανάκτηση της γνώσης αυτής προκύπτουν σημαντικές πληροφορίες. Η εξόρυξη γνώσης αποτελεί μια ισχυρή νέα τεχνολογία με πολλαπλές χρήσεις και συνδυάζει διάφορα επιστημονικά πεδία όπως η στατιστική, η τεχνολογία βάσεων δεδομένων, η τεχνητή νοημοσύνη, τα νευρωνικά δίκτυα και η μηχανική μάθηση.

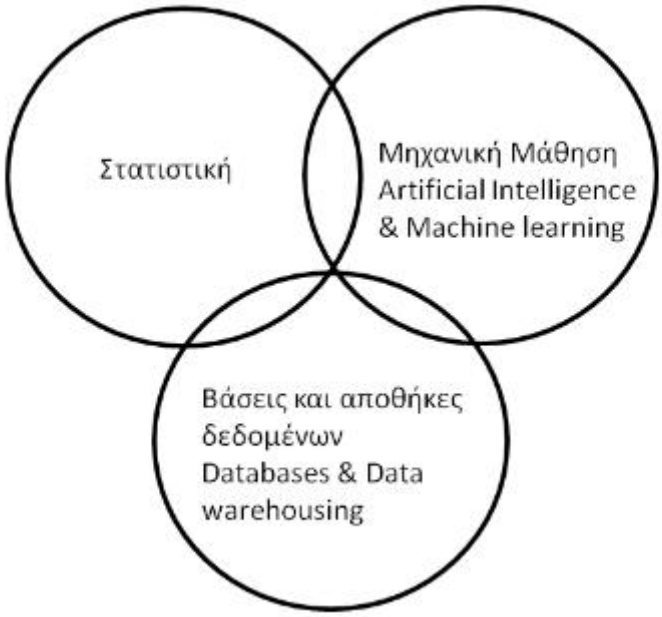


Εικόνα 2: Αρχιτεκτονική ενός τυπικού συστήματος Εξόρυξης Δεδομένων

Η γνώση που αποκτιέται μέσα από τα δεδομένα, μπορεί να χρησιμοποιηθεί στη διαχείριση της πληροφορίας, στη λήψη αποφάσεων και στην επεξεργασία ερωτημάτων για περαιτέρω απόκτηση πληροφοριών (Λαδάς, 2014). Σύμφωνα με τον Fayyad (όπ.αναφ. στο Λαδάς, 2014) η εξόρυξη δεδομένων έχει σαν βασικούς στόχους την εφαρμογή τεχνικών πρόβλεψης και συμπεριφοράς τάσεων (prediction), την αναγνώριση, την περιγραφή (description) σε μεγάλες βάσεις δεδομένων καθώς επίσης και την ταξινόμηση και βελτιστοποίηση των πόρων της. Αποτελεί σημαντικό στοιχείο πρόβλεψης για όλες τις επιχειρήσεις, αφού μέσα από αυτό τα στελέχη των επιχειρήσεων λαμβάνουν αποφάσεις και σχεδιάζουν στρατηγικές μάρκετινγκ (Λαδάς, 2014).

1.3 Ορισμός Μηχανικής Μάθησης

Ο τομέας της εξόρυξης γνώσης σχετίζεται με πολλούς τομείς όπως είναι η στατιστική, η τεχνητή νοημοσύνη, η μηχανική μάθησης, οι βάσεις δεδομένων, οι μηχανές αναζήτησης, τα συστήματα υποστήριξης αποφάσεων καθώς επίσης και το ταίριασμα προτύπων.



Εικόνα 3: Τα εργαλεία της Εξόρυξης Δεδομένων

Ο όρος μάθηση περιλαμβάνει διάφορες φράσεις όπως δεξιότητα, κατανόηση, απόκτηση γνώσης, τροποποίηση συμπεριφοράς βάσει εμπειρίας κ.ά. Έχουν δοθεί αρκετοί ορισμοί της μάθησης, όπως: «Μάθηση είναι η δημιουργία ή η αλλαγή της αναπαράστασης των εμπειριών» σύμφωνα με τον Ryszard (όπ.αναφ. στοΠαπαδόπουλου, 2011) ή «Μάθηση είναι να κάνουμε χρήσιμες αλλαγές στο μυαλό μας» σύμφωνα με τον Minsky(όπ.αναφ. στο Παπαδόπουλου, 2011). Σύμφωνα με τον Φλουρή, παρότι έχει διεξαχθεί πληθώρα σχετικών μελετών, η μάθηση παραμένει μια διαδικασία η οποία δεν έχει ερμηνευτεί και κατανοηθεί πλήρως. Συμπερασματικά, η μάθηση είναι μία διαδικασία κατά την οποία ο υποβαλλόμενος στην διαδικασία αποκτά γνώσεις, δεξιότητες, συμπεριφορές και αξίες μέσα από την παράθεση εκπαιδευτικού υλικού και με την εφαρμογή γνωστικών διαδικασιών.

Η μηχανική μάθηση είναι η διαδικασία με την οποία τα υπολογιστικά συστήματα βελτιώνουν την απόδοσή τους σε σχέση με το χρόνο και συχνά αντιμετωπίζουν προβλήματα της τεχνητής νοημοσύνης όπως διάγνωση, αναγνώριση, πρόβλεψη, σχεδιασμός κ.α. Οι αλγόριθμοι μηχανικής μάθησης μπορούν να κατηγοριοποιηθούν βάσει της μεθόδου μάθησης σε δύο κατηγορίες: τη μάθηση με επίβλεψη και τη μάθηση χωρίς επίβλεψη.

1.3.1 Μηχανική μάθηση με επίβλεψη

Στην επιβλεπόμενη μάθηση ο στόχος του αλγορίθμου είναι η δημιουργία συνάρτησης, η οποία εκφράζει το μοντέλο που μαθαίνει. Ο αλγόριθμος αυτός προσπαθεί να εκπαιδεύσει το μοντέλο κάνοντας χρήση δεδομένων εκπαίδευσης, τα οποία είναι συνήθως ζεύγη τιμών εισόδου και επιθυμητής εξόδου. Η συνάρτηση που προκύπτει από τον αλγόριθμο πιθανό να κάνει μια καλή προσέγγιση σε άγνωστα δεδομένα, εφόσον έχει προηγηθεί μία εκπαίδευση σε γνωστά δεδομένα. Η έξοδος που προκύπτει πιθανόν να είναι μια συνεχής ή διακριτή τιμή που προβλέπει μια τάξη. Στην πρώτη περίπτωση έχουμε παλινδρόμηση και στην δεύτερη περίπτωση ταξινόμηση. Οι κυριότερες τεχνικές επιβλεπόμενης μάθησης είναι τα δέντρα απόφασης, η μάθηση κανόνων, η μάθηση κατά περίπτωση, η μάθηση κατά Bayes, η μάθηση κατά νευρωνικών δικτύων, οι μηχανές διανυσμάτων υποστήριξης και η μάθηση εννοιών (Παπαδόπουλου, 2011).

1.3.2 Μηχανική μάθηση χωρίς επίβλεψη

Στην μη επιβλεπόμενη μάθηση στόχος είναι η αυτοματοποιημένη παραγωγή νέας γνώσης. Οι τεχνικές μάθησης που ανήκουν σε αυτή τη κατηγορία προσπαθούν να βρουν σχέσεις και ομάδες στα δεδομένα που τους δίνονται. Οι κυριότερες τεχνικές μη επιβλεπόμενης μάθησης είναι η εξαγωγή κανόνων συσχέτισης, ο αλγόριθμος Apriori, ο αλγόριθμος FP-Growth καθώς επίσης και διάφοροι αλγόριθμοι ομαδοποίησης όπως K-Means, DBSCAN, EM (Παπαδόπουλου, 2011).

Κεφάλαιο 2

Περιγραφή των αλγορίθμων

2.1 Επιλογή χαρακτηριστικών

Η επιλογή χαρακτηριστικών ή αλλιώς μεταβλητή επιλογή, είναι η διαδικασία της επιλογής ενός υποσυνόλου των σχετικών χαρακτηριστικών (μεταβλητές, προγνωστικοί παράγοντες κ.ά.) για χρήση σε ένα μοντέλο κατασκευής. Η κεντρική ιδέα κατά τη χρήση μιας τεχνικής επιλογής χαρακτηριστικού, είναι ότι τα δεδομένα περιέχουν πολλά χαρακτηριστικά που είναι περιττά και με αυτό τον τρόπο μπορούν να αφαιρεθούν χωρίς να υποστεί μεγάλη απώλεια πληροφοριών. Η τεχνική επιλογής χαρακτηριστικών πραγματοποιείται με την εξαγωγή των χαρακτηριστικών. Η εξαγωγή χαρακτηριστικών δημιουργεί νέες λειτουργίες από τα αυθεντικά χαρακτηριστικά, ενώ η επιλογή χαρακτηριστικών επιστρέφει ένα υποσύνολο των χαρακτηριστικών (Βικιπαίδεια, 2015).

2.2 Αλγόριθμος C4.5

Ο αλγόριθμος C4.5 αποτελεί υλοποίηση ανοικτού κώδικα σε Java, βασίζεται στα δέντρα απόφασης (Μακρή, 2013) και αποτελεί επέκταση του αλγορίθμου ID3. Ο C4.5 δημιουργεί δέντρα απόφασης μέσα από ένα σύνολο εκπαίδευσης δεδομένων , χρησιμοποιώντας την έννοια της εντροπίας πληροφοριών.

Ο πρώτος που συνειδητοποίησε τη σχέση ανάμεσα σε εντροπία και πληροφορία ήταν ο Shannon. Κατά τον Shannon, η εντροπία της πληροφορίας ή απλώς εντροπία, είναι ένα μέτρο της ποσότητας πληροφορίας που περιέχεται σ’ ένα μήνυμα. Όπως δηλαδή κατά την κλασσική έννοια η εντροπία αποτελεί ένα μέτρο πλήθους των πιθανών

μικροκαταστάσεων ενός συστήματος, η εντροπία κατά τη θεωρία της πληροφορίας αποτελεί ένα μέτρο του πλήθους των πιθανών "μεταφράσεων" που περιέχει ένα μήνυμα (Τσελεντής, 2008).

Ο αλγόριθμος C4.5 κατασκευάστηκε από τον RossQuinlan και αναφέρεται συχνά ως στατιστικό ταξινομητής, επειδή χρησιμοποιείται συχνά για ταξινόμηση. Στον αλγόριθμο C4.5 τα δεδομένα εκπαίδευσης είναι ένα σύνολο S= s1,s2,.. από ήδη ταξινομημένα δείγματα. Κάθε δείγμα S1=x1,x2.. είναι ένα διάνυσμα όπου τα x1,x2.. αντιπροσωπεύουν τις ιδιότητες ή τα χαρακτηριστικά γνωρίσματα του δείγματος. Στα δεδομένα εκπαίδευσης του αλγορίθμου C4.5 ένα διάνυσμα C=c1,c2.. αντιπροσωπεύει την κατηγορία στην οποία ανήκει κάθε δείγμα. Είναι σημαντικό να αναφερθεί, ότι κάθε χαρακτηριστικό των δεδομένων μπορεί να χρησιμοποιηθεί για να λάβει μια απόφαση, η οποία χωρίζει τα δεδομένα σε μικρότερα υποσύνολα. Ο αλγόριθμος C4.5, εξετάζει το κέρδος των πληροφοριών (informationgain) που προκύπτει από την επιλογή ενός χαρακτηριστικού για το διαχωρισμό των δεδομένων. Το χαρακτηριστικό με το υψηλότερο κέρδος πληροφοριών, είναι αυτό που χρησιμοποιείται για να ληφθεί μια απόφαση (Τζετζούμης, 2012).

2.3 Αλγόριθμος SMO

Ο αλγόριθμος SVM (Support Vector Machines – Μηχανές Διανυσμάτων Υποστήριξης) εφευρέθηκε από τον JohnPlatt το 1998 και χρησιμοποιείται ευρέως σε μηχανές εκπαίδευσης διανυσμάτων. Είναι ένας αλγόριθμος για την αποτελεσματική επίλυση του προβλήματος της βελτιστοποίησης που προκύπτει κατά τη διάρκεια της εκπαίδευσης των μηχανών με διανύσματα υποστήριξης. Ο αλγόριθμος αυτός, διαχωρίζει ένα πρόβλημα σε επιμέρους προβλήματα, τα οποία έπειτα επιλύνονται αναλυτικά (Μακρή, 2013).

2.4 Αλγόριθμος NaïveBayes

Ο αλγόριθμος NaïveBayes, είναι ένας ταξινομητής που βασίζεται στο θεώρημα Bayes. Ο ταξινομητής αυτός θεωρεί, ότι όλα τα χαρακτηριστικά συμβάλλουν στην πιθανότητα μιας συγκεκριμένης απόφασης. Λαμβάνοντας υπόψη τη φύση του

μοντέλου πιθανοτήτων, ο αλγόριθμος μπορεί να εκπαιδευτεί αποτελεσματικά σε επιτηρούμενο περιβάλλον μάθησης και να εργάζεται σε σύνθετες καταστάσεις του πραγματικού κόσμου. Επειδή οι μεταβλητές θεωρούνται ανεξάρτητες, μόνο οι διακυμάνσεις των μεταβλητών για κάθε κλάση πρέπει να προσδιορίζονται και όχι ολόκληρη η μήτρα συνδιακύμανσης (Μακρή, 2013).

2.5 Πίνακας Σύγχυσης

Ο πίνακας σύγχυσης, δείχνει τον αριθμό των σωστών και λανθασμένων προβλέψεων που γίνονται για το μοντέλο ταξινόμησης που υλοποιείται σε σχέση με τα πραγματικά αποτελέσματα (τιμή στόχο) στα δεδομένα. Περιέχει επομένως πληροφορίες, σχετικά με την πραγματική και την προβλεπόμενη ταξινόμηση.

Οι βέλτιστες λύσεις του μοντέλου έχουν μηδενικές λύσεις περιμετρικά από την κύρια διαγώνιο στον πίνακα σύγχυσης, ενώ στην κύρια διαγώνιο του πίνακα εμφανίζονται τα ορθά στοιχεία ταξινόμησης, είτε είναι αληθώς θετικά (TruePositive-TP) είτε είναι αληθώς αρνητικά (TrueNegative-TN). Οι περιπτώσεις ψευδώς αρνητικά (FalseNegative-FN) και ψευδώς θετικά (FalsePositive-FP), αντιπροσωπεύουν τις εσφαλμένες ταξινομήσεις για τον υπολογισμό του συνολικού σφάλματος. Γενικότερα, η απόδοση μίας διαδικασίας ταξινόμησης μπορεί να περιγραφεί με ένα ‘‘πίνακα σύγχυσης’’ (confusion matrix), όπως αυτόν που απεικονίζεται πιο κάτω για την περίπτωση ενός προβλήματος ταξινόμησης με δύο κατηγορίες. Πιο αναλυτικά:

	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

Εικόνα 4 :Πίνακας σύγχυσης

- TN είναι ο αριθμός των αρνητικών παραδειγμάτων που έχουν ταξινομηθεί σωστά (True Negatives)
- FP είναι ο αριθμός των αρνητικών παραδειγμάτων που έχουν από ταξινομηθεί λάθος ως θετικά (False Positives)
- FN είναι ο αριθμός των θετικών παραδειγμάτων που έχουν από λάθος ταξινομηθεί ως αρνητικά (False Negatives)
- TP είναι ο αριθμός των θετικών παραδειγμάτων που έχουν ταξινομηθεί σωστά ως θετικά (True Positives).

Με βάση τους παραπάνω ορισμούς υπολογίζουμε τις παρακάτω ποσότητες οι οποίες είναι σημαντικές για τα αποτελέσματα του μοντέλου :

TPrate (αληθώς θετικά):

$$TP / (TP + FN)$$

FPrate (ψευδώς θετικά):

$$FP / (FP + TN)$$

Precision (ακρίβεια) :

$$TP / (TP+ FP)$$

Recall (ανάκληση):

$$TP / (TP+ FN) = TP \text{ rate}$$

F-measure: 2 x Recall x Precission / (Recall + Precission):

$$2 TP/ (2 TP+ FP+ FN)$$

TPR (Sensitivity, ευαισθησία: το αληθινό θετικό ποσοστό):

$$TP/P = TP / (TP+FN)$$

SPC (Ειδικότητα / εξειδίκευση: το αληθινό αρνητικό ποσοστό):

$$SPC=TN / N \quad \text{ή} \quad SPC = TN / (TN+FN)$$

Κεφάλαιο 3

Μεθοδολογία εργασίας

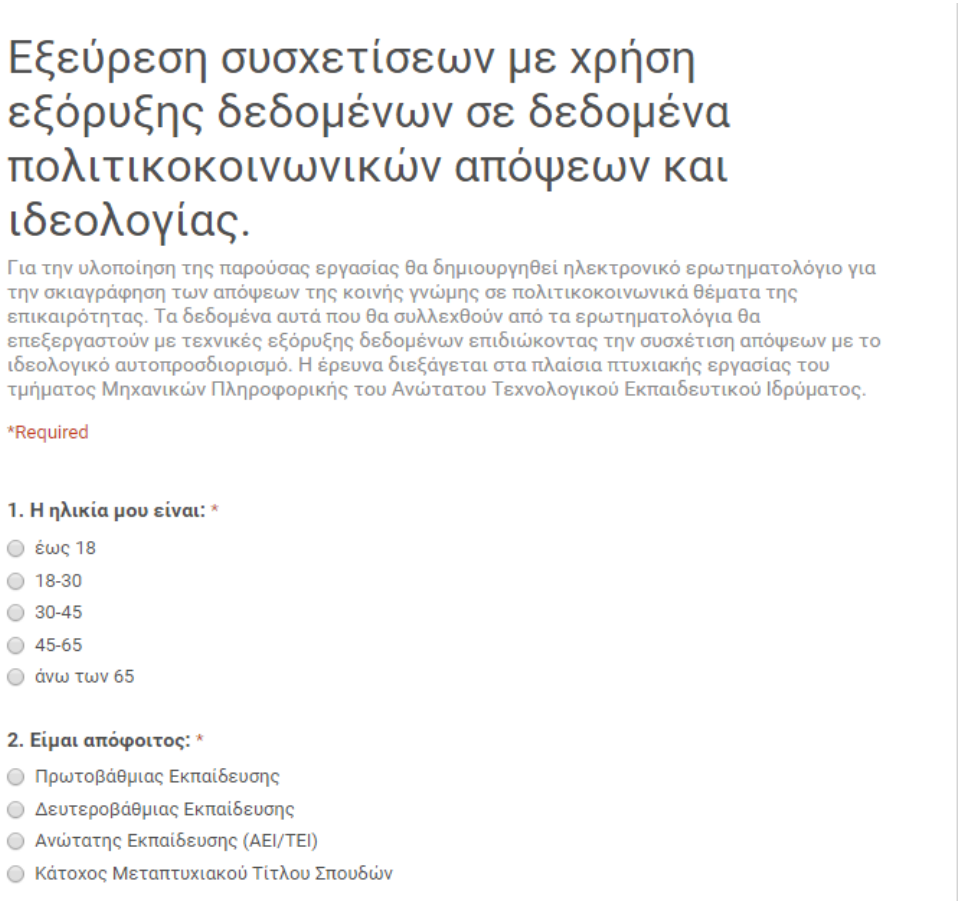
3.1 Παρουσίαση ερωτηματολογίου

Προκειμένου να υλοποιηθεί η παρούσα εργασία, δημιουργήθηκε ένα ηλεκτρονικό ερωτηματολόγιο για την σκιαγράφηση των απόψεων της κοινής γνώμης σε πολιτικοκοινωνικά θέματα της επικαιρότητας. Τα δεδομένα που συλλέχτηκαν από το ερωτηματολόγιο επεξεργάστηκαν με τεχνικές εξόρυξης δεδομένων επιδιώκοντας την συσχέτιση απόψεων με το ιδεολογικό αυτοπροσδιορισμό.

Το ερωτηματολόγιο αυτό αποτελείται από δεκαέξι ερωτήσεις στο σύνολο, εκ των οποίων οι πρώτες τρεις ερωτήσεις είναι δημογραφικές, ενώ ακολουθούν ερωτήσεις για πολιτικοκοινωνικά θέματα. Πιο αναλυτικά, οι τρεις πρώτες ερωτήσεις αναφέρονται στην ηλικία του ερωτηθέντος, το μορφωτικό του επίπεδο, καθώς επίσης και το πλήθος των κατοίκων στη περιοχή στην οποία κατοικεί. Στη συνέχεια, τίθεται το ερώτημα εάν ο δημόσιος τομέας αποτελείται από μεγάλο πλήθος υπαλλήλων και εάν οι ιδιωτικοποιήσεις φορέων μπορούν να αποτελέσουν μοχλό ανάπτυξης μέσω των επενδύσεων που θα επιφέρουν σε διάφορους τομείς. Ακολουθεί η ερώτηση για το αν η επίτευξη της συμφωνίας με τους δανειστές, θα βοηθήσει στην ανάπτυξη της οικονομίας και εάν το κράτος πρέπει να διατηρεί κέντρα διαμονής παράνομων μεταναστών. Έπειτα, ακολουθεί η ερώτηση σχετικά με την ιδιωτικοποίηση των πανεπιστημίων και για το αν θα πρέπει να διατηρηθεί η έννοια του άσυλου στα πανεπιστήμια. Επιπλέον, τίθεται το ερώτημα για το αν είναι σύμφωνοι με το κατ' οίκον περιορισμό ατόμων που έχουν διαπράξει παράνομες πράξεις αλλά έχουν προβλήματα υγείας τα οποία συνιστούν υψηλό βαθμό αναπηρίας.

Ακολουθούν πολιτικές ερωτήσεις όπως: εάν συμφωνούν με την πιθανότητα ενός δημοψηφίσματος, εάν η Ελλάδα πρέπει να παραμείνει στην Ευρωπαϊκή Ένωση, εάν η Ελλάδα ανήκει σε κάποια σφαίρα επιρροής και εάν θεωρούν ότι είναι πιθανό να προκύψει κάποιο επεισόδιο με την Τουρκία. Επίσης, τίθεται το ερώτημα εάν είναι ορθό να διεκδικηθούν οι γερμανικές αποζημιώσεις. Το ερωτηματολόγιο φτάνει στο

τέλος του, θέτοντας το ερώτημα σε ποια ιδεολογία κατατάσσονται (δεξιός, κεντροδεξιός, κεντρώος, κεντροαριστερός , αριστερός).



Εξεύρεση συσχετίσεων με χρήση εξόρυξης δεδομένων σε δεδομένα πολιτικοκοινωνικών απόψεων και ιδεολογίας.

Για την υλοποίηση της παρούσας εργασίας θα δημιουργηθεί ηλεκτρονικό ερωτηματολόγιο για την σκιαγράφηση των απόψεων της κοινής γνώμης σε πολιτικοκοινωνικά θέματα της επικαιρότητας. Τα δεδομένα αυτά που θα συλλεχθούν από τα ερωτηματολόγια θα επεξεργαστούν με τεχνικές εξόρυξης δεδομένων επιδιώκοντας την συσχέτιση απόψεων με το ιδεολογικό αυτοπροσδιορισμό. Η έρευνα διεξάγεται στα πλαίσια πτυχιακής εργασίας του τμήματος Μηχανικών Πληροφορικής του Ανώτατου Τεχνολογικού Εκπαιδευτικού Ιδρύματος.

*Required

1. Η ηλικία μου είναι: *

- έως 18
- 18-30
- 30-45
- 45-65
- άνω των 65

2. Είμαι απόφοιτος: *

- Πρωτοβάθμιας Εκπαίδευσης
- Δευτεροβάθμιας Εκπαίδευσης
- Ανώτατης Εκπαίδευσης (ΑΕΙ/ΤΕΙ)
- Κάτοχος Μεταπτυχιακού Τίτλου Σπουδών

3. Είμαι κάτοικος κοινότητας/Κωμόπολης/Πόλης *

- κάτω των 1000 κατοίκων
- μεταξύ 1000 και 10.000 κατοίκων
- Άνω των 10.000 κατοίκων

4. Ο δημόσιο τομέας της Ελλάδας είναι μεγαλύτερος από όσο θα έπρεπε *

- Συμφωνώ απόλυτα
- Συμφωνώ
- Ουδέτερος
- Διαφωνώ
- Διαφωνώ απόλυτα

5. Οι ιδιωτικοποιήσης φορέων και περιούσιας του δημοσίου μπορούν να αποτελέσουν μοχλό ανάπτυξης μέσω των επενδύσεων που θα επιφέρουν (Λιμάνια, Αεροδρόμια, Δημόσιες Επιχειρήσης). *

- Συμφωνώ απόλυτα
- Συμφωνώ
- Ουδέτερος
- Διαφωνώ
- Διαφωνώ απόλυτα

6. Η επίτευξη μιας συμφωνίας με τους δανειστές θα βοηθήσει την ανάπτυξη της οικονομίας *

- Συμφωνώ απόλυτα
- Συμφωνώ
- Ουδέτερος
- Διαφωνώ
- Διαφωνώ απόλυτα

7. Το κράτος πρέπει να διατηρεί κέντρα διαμονής παρανόμων μεταναστών *

- Συμφωνώ απόλυτα
- Συμφωνώ
- Ουδέτερος
- Διαφωνώ
- Διαφωνώ απόλυτα

8. Συμφωνείτε με την δημιουργία ιδιωτικών πανεπιστημίων *

- Συμφωνώ απόλυτα
- Συμφωνώ
- Ουδέτερος
- Διαφωνώ
- Διαφωνώ απόλυτα

9. Η αστυνομία δεν πρέπει να επεμβαίνει εντός του Πανεπιστημιακού ασύλου σε καμία περίπτωση *

- Συμφωνώ απόλυτα
- Συμφωνώ
- Ουδέτερος
- Διαφωνώ
- Διαφωνώ απόλυτα

10. Συμφωνείτε με τον κατ'οίκον περιορισμό ατόμων που έχουν διαπράξει παράνομες πράξεις αλλά έχουν προβλήματα υγείας τα οποία συνιστούν υψηλό βαθμό αναπηρίας *

- Συμφωνώ απόλυτα
- Συμφωνώ
- Ουδέτερος
- Διαφωνώ
- Διαφωνώ απόλυτα

11. Συμφωνείτε με την πιθανότητα ενός δημοψηφίσματος *

- Συμφωνώ απόλυτα
- Συμφωνώ
- Ουδέτερος
- Διαφωνώ
- Διαφωνώ απόλυτα

12. Η Ελλάδα πρέπει να παραμείνει στην Ευρωπαϊκή Ένωση *

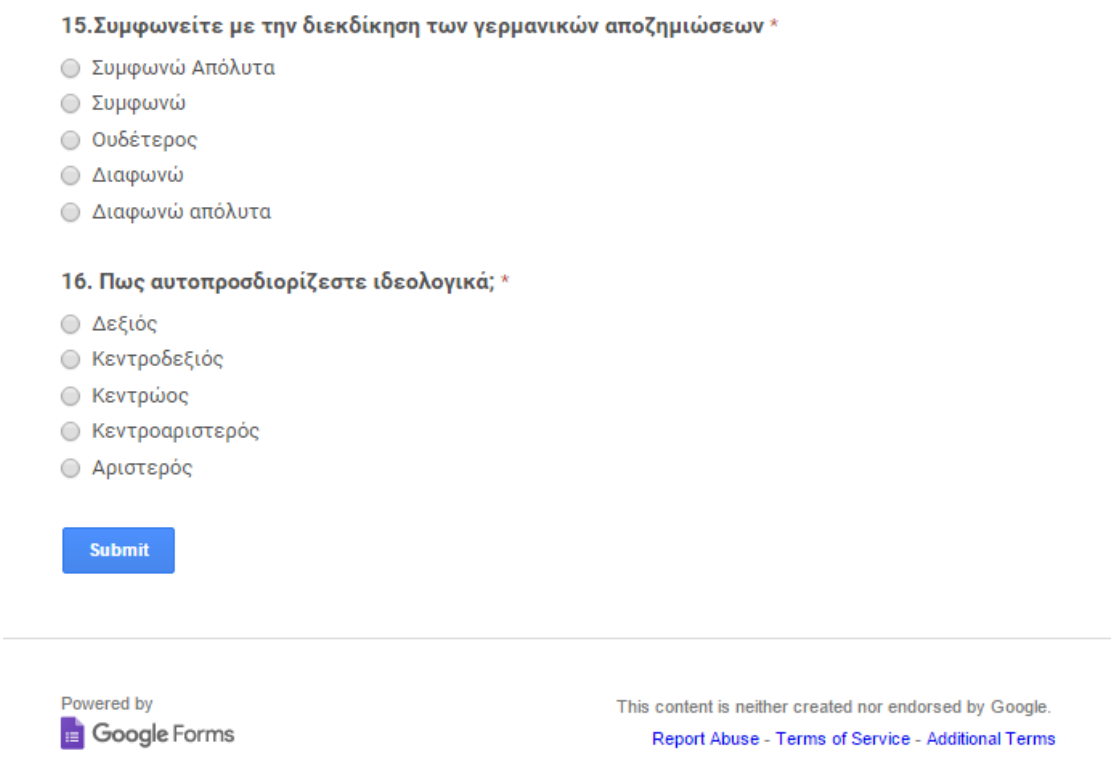
- Συμφωνώ απόλυτα
- Συμφωνώ
- Ουδέτερος
- Διαφωνώ
- Διαφωνώ απόλυτα

13. Είναι αναπόφευκτο η Ελλάδα να ανήκει σε κάποια από τις σφαίρες επιρροής (Δύση, Ρωσία, Κίνα) *

- Συμφωνώ απόλυτα
- Συμφωνώ
- Ουδέτερος
- Διαφωνώ
- Διαφωνώ Απολύτα

14. Θεωρείτε πιθανό ένα θερμό επεισόδιο με την Τουρκία που μπορεί να οδηγήσει σε σύρραξη *

- Συμφωνώ απόλυτα
- Συμφωνώ
- Ουδέτερος
- Διαφωνώ
- Διαφωνώ απόλυτα



Εικόνα 5: Παρουσίαση ερωτηματολογίου

3.2 Υλοποίηση ερωτηματολογίου

Η υλοποίηση του ερωτηματολογίου πραγματοποιήθηκε με τη βοήθεια του Google Drive, όπου είναι μία υπηρεσία αποθήκευσης και συγχρονισμού αρχείων που παρέχεται από την Google. Επιτρέπει την χρήση αποθήκευσης και διαμοιρασμού αρχείων, καθώς επίσης και την συνεργατική επεξεργασία από τον χρήστη. Το Google Drive προσφέρει στους χρήστες του έναν αρχικό online χώρο αποθήκευσης 15GB που μπορεί να χρησιμοποιηθεί από τις τρεις υπηρεσίες: το Google Drive, το Gmail και τις φωτογραφίες του Google + (Βικιπαίδεια, 2015).

3.3 Τρόποι προώθησης ερωτηματολογίου

Η προώθηση του ερωτηματολογίου που σχεδιάστηκε, πραγματοποιήθηκε με τη βοήθεια των μέσων κοινωνικής δικτύωσης όπως : facebook, twitter, skype καθώς επίσης και μέσω e-mail.

3.4 Παρουσίαση WEKA

Το WEKA (WaikatoEnvironmentforKnowledgeAnalysis) είναι μια συλλογή απόαλγορίθμους μηχανικής μάθησης. Παρέχει δυνατότητα για:

- Προεπεξεργασία των δεδομένων (τα εργαλεία για την προεπεξεργασία στο περιβάλλον του WEKA ονομάζονται filters).
- Δημιουργία «μοντέλων» από τα δεδομένα με κάποια διαδικασία εκπαίδευσης
- Χρησιμοποίηση στατιστικών μεγεθών για την αξιολόγηση των διαφόρων αλγορίθμων μάθησης.
- Απεικόνιση τόσο των αρχικών δεδομένων όσο και των αποτελεσμάτων μετά τη διαδικασία της εκπαίδευσης.

Και όλα αυτά σε ένα γραφικό περιβάλλον (υπάρχει και η δυνατότητα χρήσης από τη γραμμή εντολών) το οποίο ονομάζεται «Explorer».

Εικόνα 6: Μενού επιλογών weka

3.4.1 Δομή αρχείων του WEKA

Τα βασικά αρχεία τα οποία δέχεται σαν είσοδο το WEKA έχουν την κατάληξη .arff (Attribute -Relation File Format) και πρόκειται για ένα αρχείο κειμένου χαρακτήρων ASCII (ASCII text file) το οποίο περιγράφει/περιέχει μια σειρά από παραδείγματα (instances) τα οποία «περιγράφονται» από χαρακτηριστικά (attributes).

Οι γραμμές που ξεκινάνε με % είναι σχόλια και δεν λαμβάνονται υπόψη όταν φορτώνεται το αρχείο. Μετά από τα σχόλια ακολουθεί η δήλωση του ονόματος, που περιγράφει το αρχείο η οποία δεν μπορεί να παραλειφθεί. Η γραμμή αυτής της δήλωσης ξεκινάει με: @relation +όνομα αρχείου. Μετά από αυτή τη γραμμή ακολουθεί η δήλωση όλων των χαρακτηριστικών που περιγράφουν το συγκεκριμένο σύνολο παραδειγμάτων. Η δήλωση αυτή γίνεται ως εξής: @attribute + <attribute-name> + <datatype>. Όπου <attribute-name> είναι το όνομα του χαρακτηριστικού, το οποίο πρέπει να ξεκινάει με γράμμα. Σε περίπτωση που ένα χαρακτηριστικό περιγράφεται με δύο ή περισσότερες λέξεις που χωρίζονται με κενό, τότε θα πρέπει όλες αυτές να περικλείονται σε εισαγωγικά (“ ”). Ενώ το όρισμα <datatype> καθορίζει τον τύπο του χαρακτηριστικού. Το Weka υποστηρίζει τέσσερις διαφορετικούς τύπους:

- Αριθμητικά δεδομένα (numeric)
- Δεδομένα που ορίζουν κατηγορία (ονομαστικά) (<nominal-specification>)
- Αλφαριθμητικά (string)
- Ημερομηνίες με συγκεκριμένο format (date[<date-format>])

Οι λέξεις κλειδιά numeric, string, date μπορούν να γραφούν είτε με κεφαλαία είτε με πεζά.

Αριθμητικών χαρακτηριστικών - Numeric attributes : Τα αριθμητικά χαρακτηριστικά μπορεί να είναι είτε πραγματικοί είτε ακέραιοι αριθμοί, στο παράδειγμα μας τέτοια είναι τα χαρακτηριστικά της θερμοκρασίας (temperature) και της υγρασίας (humidity).

«Ονομαστικά» χαρακτηριστικά - Nominal attributes : Τα χαρακτηριστικά που παίρνουν «ονομαστικές» τιμές ορίζονται χρησιμοποιώντας αγκύλες εντός των οποίων γράφονται όλες οι δυνατές «τιμές»: {<nominal-name1>, <nominal-name2>, <nominal-name3>, ...}

@attribute outlook {sunny, overcast, rainy}

Χαρακτηριστικά αλφαριθμητικών - String attributes: Τα χαρακτηριστικά αλφαριθμητικών επιτρέπουν τη δημιουργία αυθαίρετων αλφαριθμητικών δομών κάτι το οποίο είναι στην περίπτωση που ενδιαφερόμαστε για. text-mining applications. Ο ορισμός ενός τέτοιου χαρακτηριστικού έχει την παρακάτω μορφή

@attribute LCC string

Ημερομηνίες - Date attributes: Ο καθορισμός χαρακτηριστικών που παίρνουν ως τιμή ημερομηνίες γίνεται με την παρακάτω μορφή:

@attribute <name> date, όπου <name>είναι το όνομα του χαρακτηριστικού και <date> είναι η ημερομηνία σύμφωνα με το παρακάτω format: "yyyy-MM-dd'T'HH:mm:ss" (ISO-8601).

Π.χ. 2005-11-23-T11:50:25

Μετά από τη δήλωση των χαρακτηριστικών ακολουθεί η δήλωση ότι θα ακολουθήσουν τα δεδομένα. Η γραμμή : @data, που δηλώνει ότι θα ακολουθήσουν τα δεδομένα.

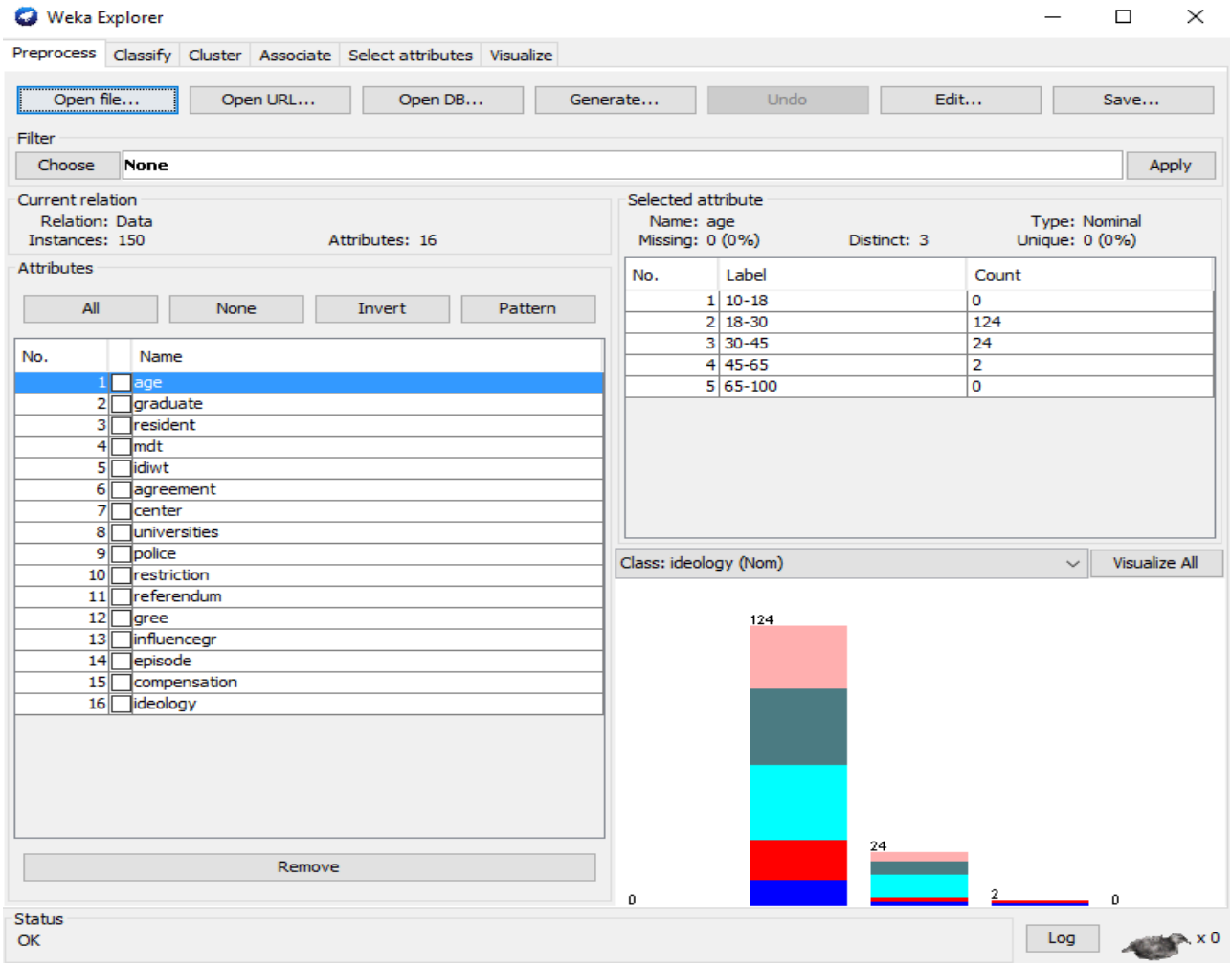
3.4.2 Το περιβάλλον WEKA

Ανοίγοντας το Weka, μέσα από την επιλογή Explorer επιτυγχάνεται η μεταφορά στο γραφικό περιβάλλον του προγράμματος ώστε να ανοιχθεί μέσα από την επιλογή του OpenFile το αρχείο που δημιουργήθηκε με βάση το ερωτηματολόγιο που σχεδιάστηκε. Το αρχείο .arff του ερωτηματολογίου απεικονίζεται παρακάτω:

```
@relation Data

@attribute age              {10-18,18-30,30-45,45-65,65-100}
@attribute graduate         {'Primary education','Secondary education','Higher education','Postgraduate degree'}
@attribute resident         {'Lower than 1000','Between 1000 and 10000','Up to 10000'}
@attribute mdt              {'Strongly agree',Agree,Neutral,Disagree,'Strongly disagree'}
@attribute idiwt            {'Strongly agree',Agree,Neutral,Disagree,'Strongly disagree'}
@attribute agreement        {'Strongly agree',Agree,Neutral,Disagree,'Strongly disagree'}
@attribute center           {'Strongly agree',Agree,Neutral,Disagree,'Strongly disagree'}
@attribute universities     {'Strongly agree',Agree,Neutral,Disagree,'Strongly disagree'}
@attribute police           {'Strongly agree',Agree,Neutral,Disagree,'Strongly disagree'}
@attribute restriction      {'Strongly agree',Agree,Neutral,Disagree,'Strongly disagree'}
@attribute referendum       {'Strongly agree',Agree,Neutral,Disagree,'Strongly disagree'}
@attribute gree             {'Strongly agree',Agree,Neutral,Disagree,'Strongly disagree'}
@attribute influencegr      {'Strongly agree',Agree,Neutral,Disagree,'Strongly disagree'}
@attribute episode          {'Strongly agree',Agree,Neutral,Disagree,'Strongly disagree'}
@attribute compensation     {'Strongly agree',Agree,Neutral,Disagree,'Strongly disagree'}
@attribute ideology         {Right,Center-right,Centrist,Center-left,Left}

@data

30-45,'Postgraduate degree','Up to 10000',Agree,'Strongly agree','Strongly agree',Agree,Agree,Neutral,Disagree,Neutral,'Strongly agree',
'Strongly agree',Disagree,Neutral,Centrist
18-30,'Secondary education','Between 1000 and 10000',Agree,'Strongly agree','Strongly agree',Disagree,Agree,Neutral,Neutral,Agree,Agree,Agree,
Agree,'Strongly agree',Left
18-30,'Higher education','Between 1000 and 10000',Agree,'Strongly agree',Neutral,Disagree,Agree,Neutral,Agree,Neutral,'Strongly agree',Agree,
Neutral,'Strongly agree',Left
18-30,'Secondary education','Up to 10000',Agree,Neutral,Agree,Agree,Disagree,'Strongly agree',Neutral,Neutral,Agree,Disagree,Agree,Agree,
Center-right
18-30,'Secondary education','Up to 10000',Neutral,'Strongly agree','Strongly agree',Agree,Disagree,Neutral,Agree,Neutral,Agree,Disagree,Agree,
'Strongly agree',Left
```

Εικόνα 7: Δεδομένα αρχείου .arff

Στη συνέχεια εμφανίζονται στο weka όλα τα δεδομένα του αρχείου, όπως απεικονίζονται παρακάτω:

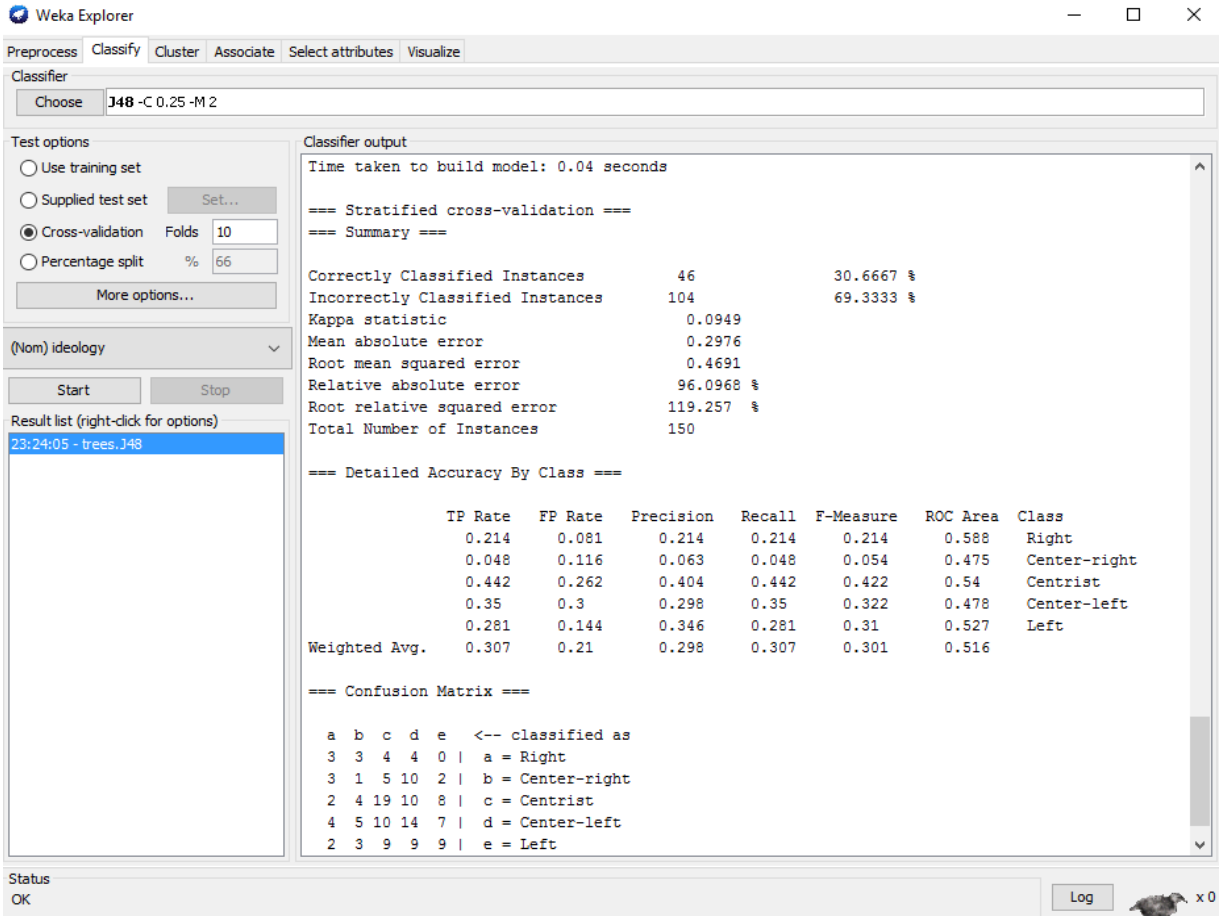


Εικόνα 8 : Παρουσίαση των δεδομένων του αρχείου .arff στο weka

Στο παράθυρο αυτό στο μέσο και αριστερά αναγράφεται ότι το συγκεκριμένο dataset ονομάζεται Data και περιλαμβάνει 150 παραδείγματα-απαντήσεις και κάθε παράδειγμα απαρτίζεται από 16 χαρακτηριστικά. Το πρώτο χαρακτηριστικό (age) είναι επιλεγμένο και στο κάτω δεξιό μέρος αναγράφεται ότι δεν έχει ελλιπείς καταχωρήσεις (missing: 0), πρόκειται για ένα χαρακτηριστικό το οποίο έχει ονομαστικές τιμές (10-18, 18- 30, 30-45, 45- 65, 65-100) οι οποίες εμφανίζονται 0, 124, 24, 2, 0 φορές αντίστοιχα.

Από την αριστερή πλευρά δίνονται: το όνομα του αρχείου, το σύνολο των απαντήσεων και τα χαρακτηριστικά. Από την δεξιά πλευρά εμφανίζονται όλα τα στοιχεία για κάθε χαρακτηριστικό που επιλέγεται. Πιο συγκεκριμένα, εμφανίζεται το πλήθος κάθε δυνατής απάντησης του κάθε χαρακτηριστικό. Ενώ τέλος, στο κάτω μέρος απεικονίζεται γραφικά το πλήθος των απαντήσεων του κάθε χαρακτηριστικού. Επιλέγοντας την καρτέλα Classify από το περιβάλλον, μας δίνεται η δυνατότητα να

επιλέξουμε ποιον αλγόριθμο επιθυμούμε να «τρέξουμε». Έστω ότι χρησιμοποιείται ο αλγόριθμος C4.5 και επιλέγοντας start προκύπτει:



Εικόνα 9: Παρουσίαση αποτελεσμάτων με τη χρήση του αλγορίθμου C4.5

Τέλος, στην καρτέλα Selectattributes, εμφανίζονται τα καλύτερα χαρακτηριστικά του ερωτηματολογίου που σχεδιάστηκε.

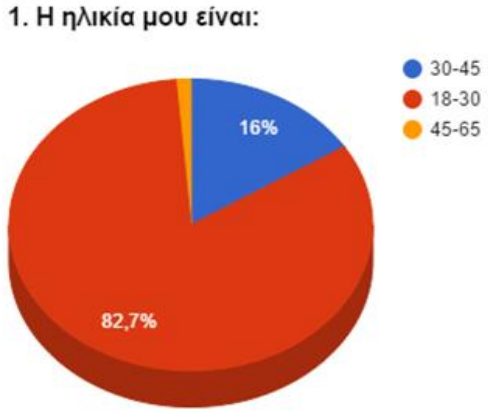
Κεφάλαιο 4

Πειράματα και αποτελέσματα

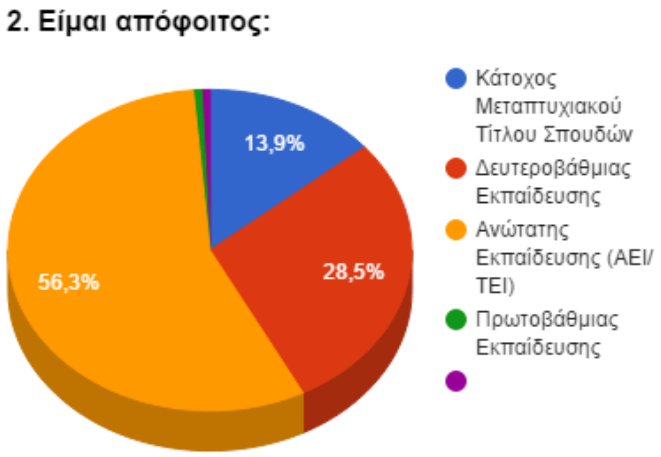
4.1 Παρουσίαση των αποτελεσμάτων του ερωτηματολογίου

Προκειμένου να ολοκληρωθεί η εργασία, πραγματοποιήθηκε μία έρευνα με τη χρήση ενός ερωτηματολογίου σχετικό με τα πολιτικοκοινωνικά θέματα της επικαιρότητας. Το ερωτηματολόγιο απαντήθηκε από εκατόν πενήντα άτομα, οι απαντήσεις των οποίων αναλύονται παρακάτω.

Αρχικά από το σύνολο των ερωτηθέντων, το μεγαλύτερο ποσοστό (82,7%) είχαν ηλικία 18-30, το 16% ήταν ηλικίας 30-45 και το υπόλοιπο ποσοστό αντιστοιχούσε σε ηλικίες 45-65 ετών.



Στη δεύτερη ερώτηση σχετικά με το μορφωτικό επίπεδο των ερωτηθέντων, το 56,3% είναι απόφοιτοι Ανώτατης Εκπαίδευσης, το 28,5% απόφοιτοι Δευτεροβάθμιας Εκπαίδευσης και το 13,9% είναι κάτοχοι Μεταπτυχιακών Σπουδών.



Στην τρίτη ερώτηση το 63,3% των ερωτηθέντων κατοικούν σε πόλη άνω των 10.000 κατοίκων, το 30% μεταξύ 1000 και 10.000 κατοίκων και μόλις το 6,7% σε περιοχή κάτω των 1000 κατοίκων.

Στην ερώτηση σχετικά με το αν ο δημόσιος τομέας της Ελλάδας είναι μεγαλύτερος απ' όσο θα έπρεπε, το μεγαλύτερο ποσοστό (46%) απάντησε συμφωνώ, το 24,7% συμφώνησε απόλυτα, το 19,3% ήταν ουδέτεροι και ακολουθούν με ελάχιστη διαφορά το «διαφωνώ» με το «διαφωνώ απόλυτα».

Στην πέμπτη ερώτηση, το 41,3% συμφωνεί ότι οι ιδιωτικοποιήσεις φορέων και περιουσίας του δημοσίου μπορούν να αποτελέσουν μοχλό ανάπτυξης μέσων των επενδύσεων, ενώ το 21,3% διαφωνεί με την άποψη αυτή.

Στην επόμενη ερώτηση σχετικά με το αν η επίτευξη μιας συμφωνίας με τους δανειστές θα βοηθήσει στην ανάπτυξη της οικονομίας, υπήρξε μία μικρή διαφορά σε ποσοστό μεταξύ αυτών που συμφωνούσαν (31,3%) και αυτών που διαφωνούσαν (28,7%).

Στην ερώτηση σχετικά με το αν το κράτος θα πρέπει να διατηρεί κέντρα διαμονής παρανόμων μεταναστών, το 29,3% συμφώνησαν στη γνώμη αυτή, το 24% είχαν ουδέτερη στάση και το 21,3% των ερωτηθέντων διαφώνησαν.

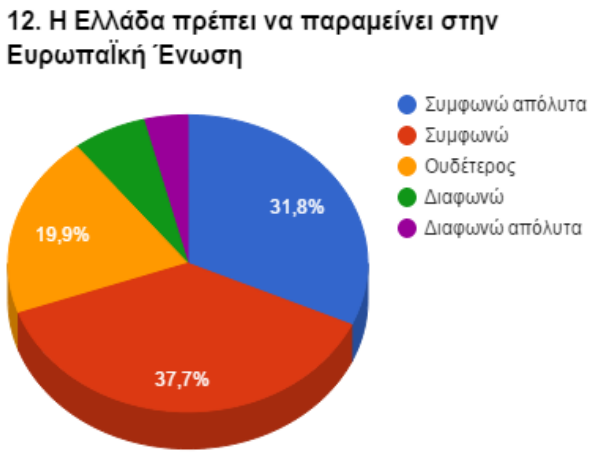
Στη συνέχεια, το 33,3% του συνόλου απάντησε ότι διαφωνεί με την δημιουργία ιδιωτικών πανεπιστημίων, ενώ μόλις το 10% του συνόλου συμφώνησε απόλυτα στη άποψη αυτή.

Μεγάλο ποσοστό διαφωνίας παρατηρήθηκε στην ερώτηση σχετικά με το αν θα πρέπει να διατηρηθεί το άσυλο στα Πανεπιστήμια, όπου το 34,7% διαφώνησε με τη διατήρηση του άσυλου και το 21,3% συμφώνησε ότι η αστυνομία δεν θα πρέπει να επεμβαίνει εντός του πανεπιστημιακού ασύλου.

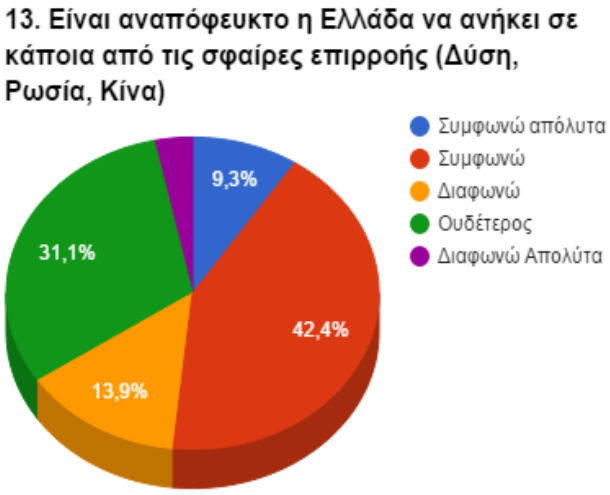
Στο ερώτημα για το αν είναι σύμφωνοι με το κατ' οίκον περιορισμό ατόμων που έχουν διαπράξει παράνομες πράξεις αλλά έχουν προβλήματα υγείας τα οποία συνιστούν υψηλό βαθμό αναπηρίας, το 38,7% συμφώνησε και το 19,3% διαφώνησε.

Ακολουθεί το ερώτημα για το αν είναι σύμφωνοι με την πιθανότητα ενός δημοψηφίσματος, όπου το 33,8% συμφώνησε, το 25,2% είχε ουδέτερη στάση, το 15,2% συμφώνησε απόλυτα, ενώ το 19,2% διαφώνησε.

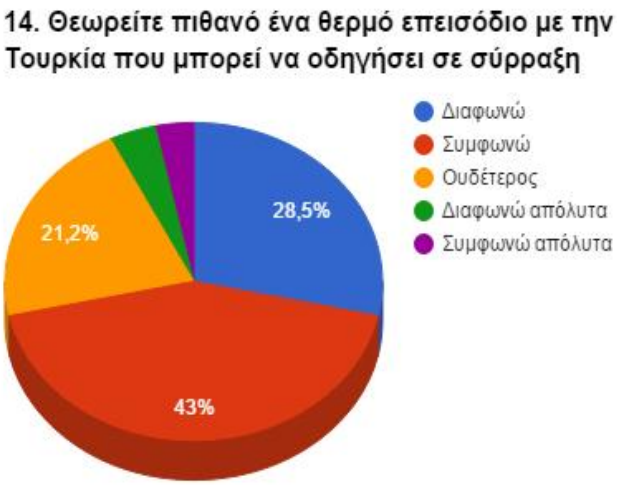
Στη συνέχεια, το 37,7% του συνόλου πιστεύει ότι η Ελλάδα πρέπει να παραμείνει στην Ευρωπαϊκή Ένωση, το 19,9% διατήρησε μία ουδέτερη στάση και το 31,8% του συνόλου συμφώνησε απόλυτα στην άποψη αυτή.



Στο ερώτημα εάν είναι αναπόφευκτο η Ελλάδα να ανήκει σε κάποια από τις σφαίρες επιρροής (Δύση, Ρωσία, Κίνα) , το 42,4% συμφώνησε με το ερώτημα ενώ το 31,1% κράτησαν ουδέτερη στάση.



Το 43% του συνόλου πιστεύει ότι είναι πιθανό να υπάρξει κάποιο θερμό επεισόδιο με την Τουρκία που θα οδηγήσει σε σύρραξη, ενώ το 28,5% διαφώνησε με αυτή την ιδεολογία.



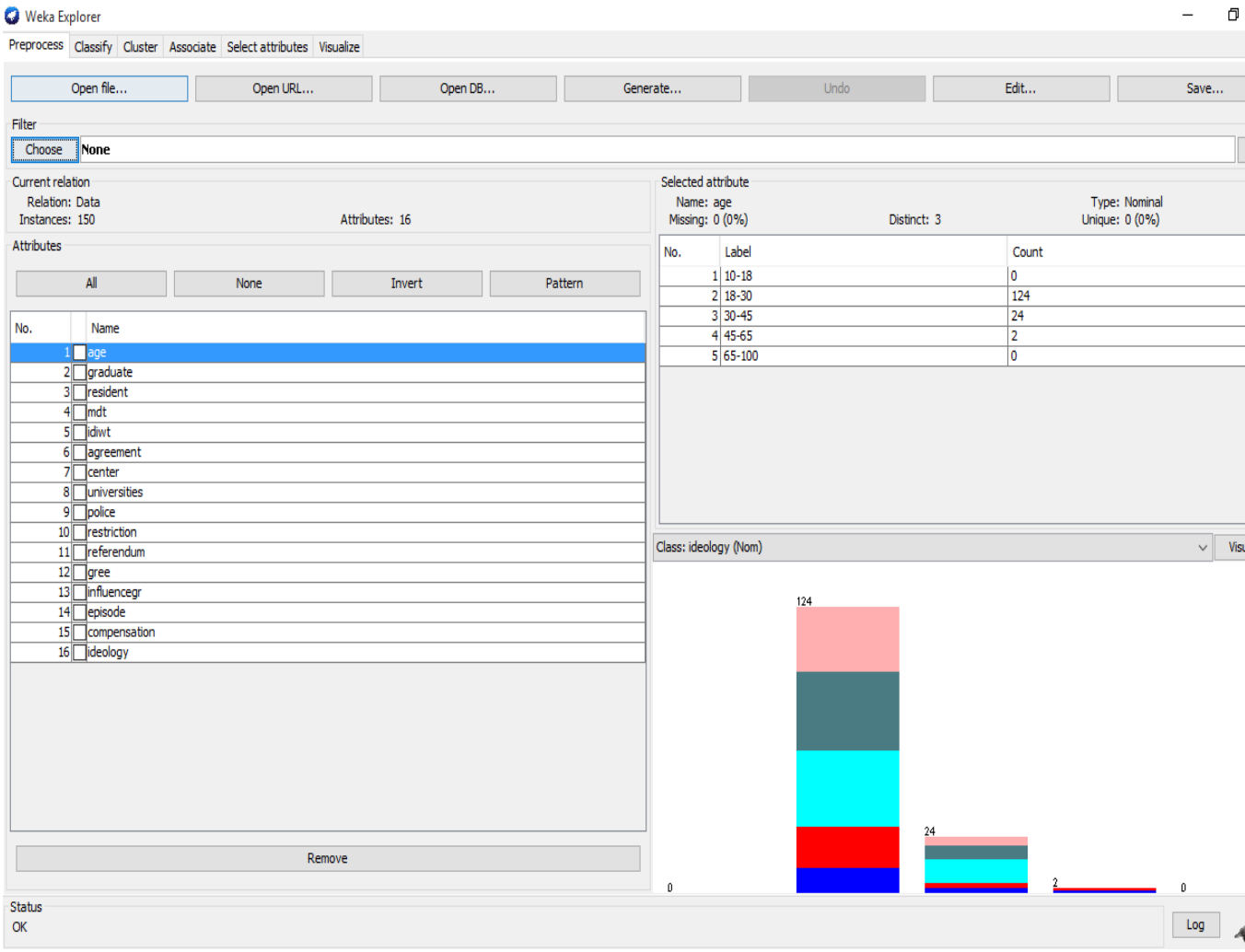
Στη συνέχεια, το 84,1% των ερωτηθέντων πιστεύει ότι θα πρέπει να διεκδικηθούν οι γερμανικές αποζημιώσεις, ενώ το ποσοστό διαφωνίας ήταν ελάχιστο.

Τέλος, στο ερώτημα για το πώς θα αυτοπροσδιοριζόταν ιδεολογικά το κάθε άτομο, το 28,5% απάντησε κεντρώος, το 27,2% κεντροαριστερός, το 21,2% αριστερός, το 13,9% κεντροδεξιός και το 9,3% του συνόλου δεξιός.

4.2 Περιγραφή δεδομένων με τη χρήση αλγορίθμων

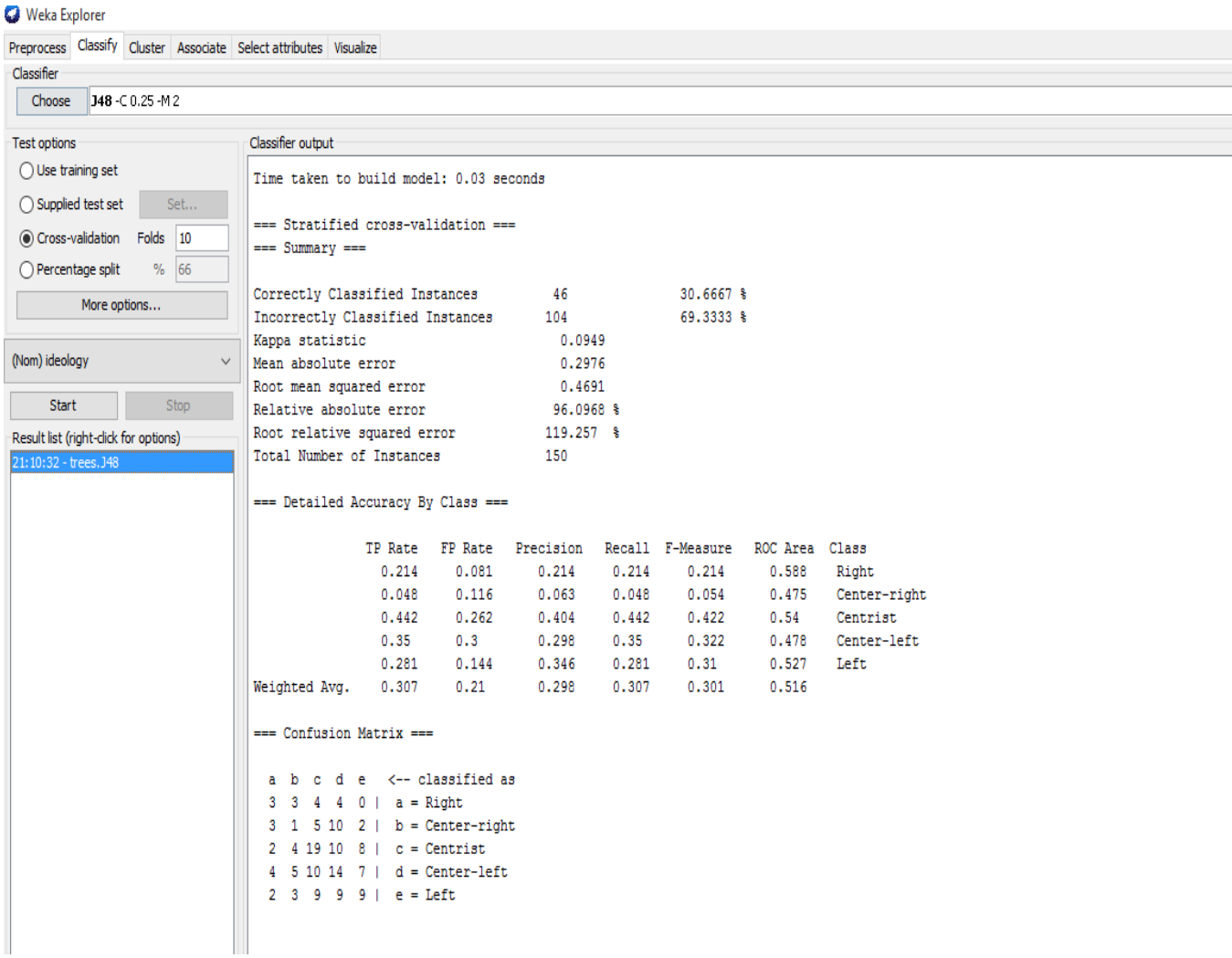
Στα πλαίσια υλοποίησης της εργασίας, δημιουργήθηκαν δύο αρχεία .arff με δεκαέξι χαρακτηριστικά το κάθε αρχείο. Η μόνη διαφορά μεταξύ των αρχείων αυτών ήταν το χαρακτηριστικό @attributeideology, όπου την πρώτη φορά είχε ως όρισμα πέντε κλάσεις και τη δεύτερη φορά είχε ως όρισμα τρεις κλάσεις. Προκειμένου να μελετήσουμε τα δεδομένα που συλλέχτηκαν χρησιμοποιήθηκαν και τα δύο αρχεία, ενώ σε κάθε αρχείο υλοποιήθηκαν οι αλγόριθμοι: C4.5,SVMκαι NaïveBayes.

Η διαδικασία που πραγματοποιήθηκε εξηγείτε παρακάτω αναλυτικά. Εφόσον πραγματοποιήθηκε το άνοιγμα του περιβάλλοντος weka, μέσα από την επιλογή openfile διαλέχτηκε το αρχείο answers.arff, το οποίο περιέχει τα αποτελέσματα του ερωτηματολογίου. Μόλις πραγματοποιηθεί το άνοιγμα του αρχείου, εμφανίζονται στα αριστερά τα δεκαέξι χαρακτηριστικά από το ερωτηματολόγιο (age, graduate….ideology). Στη δεξιά πλευρά, εμφανίζονται οι απαντήσεις από το ερωτηματολόγιο, για κάθε χαρακτηριστικό που επιλέγεται από την αριστερή στήλη. Δίπλα από τις απαντήσεις του ερωτηματολογίου, υπάρχει η στήλη count όπου δείχνει πόσες απαντήσεις επιλέχθηκαν. Παρακάτω δίνεται ένα στιγμιότυπο από το γραφικό περιβάλλον για την απεικόνιση αυτών που προαναφέρθηκαν.



Εικόνα 10: Φόρμα φόρτωσης και προ επεξεργασίας δεδομένων

Στη συνέχεια επιλέγεται από την καρτέλα classify, ο αλγόριθμος που επιθυμούμε κάθε φορά να εκτελεστεί. Αρχικά μελετήθηκε ο αλγόριθμος C4.5 και προέκυψαν τα αποτελέσματα που απεικονίζονται παρακάτω.



Εικόνα 11: Καρτέλα επιλογής κατηγοριοποίησης

Στη δεξιά πλευρά του γραφικού περιβάλλοντος εμφανίζονται όλα τα αποτελέσματα μετά από την εκτέλεση του αλγορίθμου. Πιο αναλυτικά αυτά που αναγράφονται έχουν την εξής ερμηνεία:

- **Timetakentobuildmodel**: ο χρόνος που χρειάστηκε για να υλοποιηθεί το μοντέλο σύμφωνα με τον αλγόριθμο που επιλέχθηκε ώστε να προκύψουν τα αποτελέσματα. Στη συγκεκριμένη περίπτωση αναγράφεται 0.03 sec.
- **Summary**: αποτελεί μία σύντομη περίληψη για το μοντέλο που σχεδιάστηκε δίνοντάς τα ακόλουθα αποτελέσματα:

CorrectlyClassifiedInstances: αναγράφεται ο αριθμός των απαντήσεων που έχουν ταξινομηθεί σωστά. Ο αριθμός αυτός δείχνει την ακρίβεια του δείγματος, η οποία αναγράφεται και σε ποσοστό.

Στη συγκεκριμένη περίπτωση:

CorrectlyClassifiedInstances: 46 30.66%

Από τις 150 απαντήσεις που δόθηκαν μόνο οι 46 από αυτές έχουν ταξινομηθεί σωστά και η ακρίβεια του δείγματος είναι 30,66%.

***IncorrectlyClassifiedInstances*:** αναγράφεται ο αριθμός των απαντήσεων που έχουν ταξινομηθεί εσφαλμένα. Ο αριθμός αυτός δείχνει την ακρίβεια του δείγματος, η οποία αναγράφεται και σε ποσοστό.

Στη συγκεκριμένη περίπτωση:

IncorrectlyClassifiedInstances: 104 69.33%

Από τις 150 απαντήσεις που δόθηκαν οι 104 από αυτές έχουν ταξινομηθεί εσφαλμένα και η ακρίβεια του δείγματος είναι 69,33%.

Kappastatistic: μετριέται κατά πόσο συμφωνούν η τιμή της πρόβλεψης με την πραγματική κατηγορία.

Στη συγκεκριμένη περίπτωση:

Kappastatistic 0.0949

Meanabsoluteerror: μετράει το μέσο μέγεθος των σφαλμάτων σε μια σειρά από προβλέψεις.

Στη συγκεκριμένη περίπτωση:

Meanabsoluteerror 0.2976

Rootmean squared error (RMSE): μετράει το μέσο μέγεθος του σφάλματος. Δηλαδή, τη σχέση μεταξύ των προβλέψεων σε σχέση με αυτές που παρατηρήθηκαν. Υπολογίζει την τετραγωνική ρίζα του μέσου όρου.

Στη συγκεκριμένη περίπτωση:

Rootmean squared error 0.4691

Relative absolute error: δίνεται μια σχετική ένδειξη για το πόσο καλή είναι μία μέτρηση σε σχέση με το μέγεθος που μετριέται (Σχετικό Σφάλμα).

Στη συγκεκριμένη περίπτωση:

Relative absolute error 96.0968 %

Root relative squared error: είναι η τετραγωνική ρίζα του σφάλματος και λαμβάνει το συνολικό τετραγωνικό σφάλμα. Συγκεκριμένα είναι ο μέσος όρος των πραγματικών τιμών.

Στη συγκεκριμένη περίπτωση:

Root relative squared error 119.257 %

Total Number of Instances: Οαριθμόςόλωντωνδειγμάτων.

Στη συγκεκριμένη περίπτωση:

Total Number of Instances 150

- **Detailed Accuracy By Class**

TPRate: ο αριθμός των παραδειγμάτων που προέβλεψε θετικά, και είναι πραγματικά θετικά. Δηλαδή, το ποσοστό των θετικών περιπτώσεων που έχουν αναγνωριστεί σωστά.

FPRate: ο αριθμός των παραδειγμάτων που προέβλεψε θετικά, και είναι πραγματικά αρνητικά. Δηλαδή, το ποσοστό των αρνητικών περιπτώσεων που είχαν ταξινομηθεί εσφαλμένα ως θετικά.

Precision: είναι η αναλογία των προβλεπόμενων θετικών κρουσμάτων που ήταν σωστή και υπολογίζεται από τη σχέση: TP Rate / θετικές προβλέψεις.

Recall: είναι η αναλογία των προβλεπόμενων πραγματικά θετικών κρουσμάτων που ήταν σωστή και υπολογίζεται από τη σχέση: TP Rate / πραγματικά θετικές προβλέψεις.

F-Measure: η σχέση μεταξύ της ακρίβειας και της ανάκλησης.

Όπου:

Ακρίβεια, είναι ο αριθμός των σωστών αποτελεσμάτων διαιρούμενο με τον αριθμό όλων των επιστρεφόμενων αποτελεσμάτων.

Ανάκληση, είναι ο αριθμός των σωστών αποτελεσμάτων διαιρούμενο με τον αριθμό των αποτελεσμάτων που θα έχουν επιστραφεί.

ROCArea: είναι ένας λειτουργικός δέκτης για να διακρίνει την απόδοση μίας δοκιμής. Χρησιμοποιείται για τη σύγκριση της απόδοσης των διαγνωστικών.

Class: οι κλάσεις ενός χαρακτηριστικού.

- **ConfusionMatrix:** είναι ο πίνακας σύγχυσης, ο οποίος δείχνει τον αριθμό των σωστών και λανθασμένων προβλέψεων που γίνονται για το μοντέλο ταξινόμησης που υλοποιείται σε σχέση με τα πραγματικά αποτελέσματα (τιμή στόχο) στα δεδομένα.

Στη συγκεκριμένη περίπτωση:

Η βέλτιστη λύση βρίσκεται στην $1^{η}$ γραμμή - $5^{η}$ στήλη, ενώ στην κύρια διαγώνιο απεικονίζονται τα ορθά στοιχεία της ταξινόμησης. Τα υπόλοιπα στοιχεία του πίνακα απεικονίζουν τις εσφαλμένες ταξινομήσεις.

```
=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
  3  3  4  4  0 |  a = Right
  3  1  5 10  2 |  b = Center-right
  2  4 19 10  8 |  c = Centrist
  4  5 10 14  7 |  d = Center-left
  2  3  9  9  9 |  e = Left
```

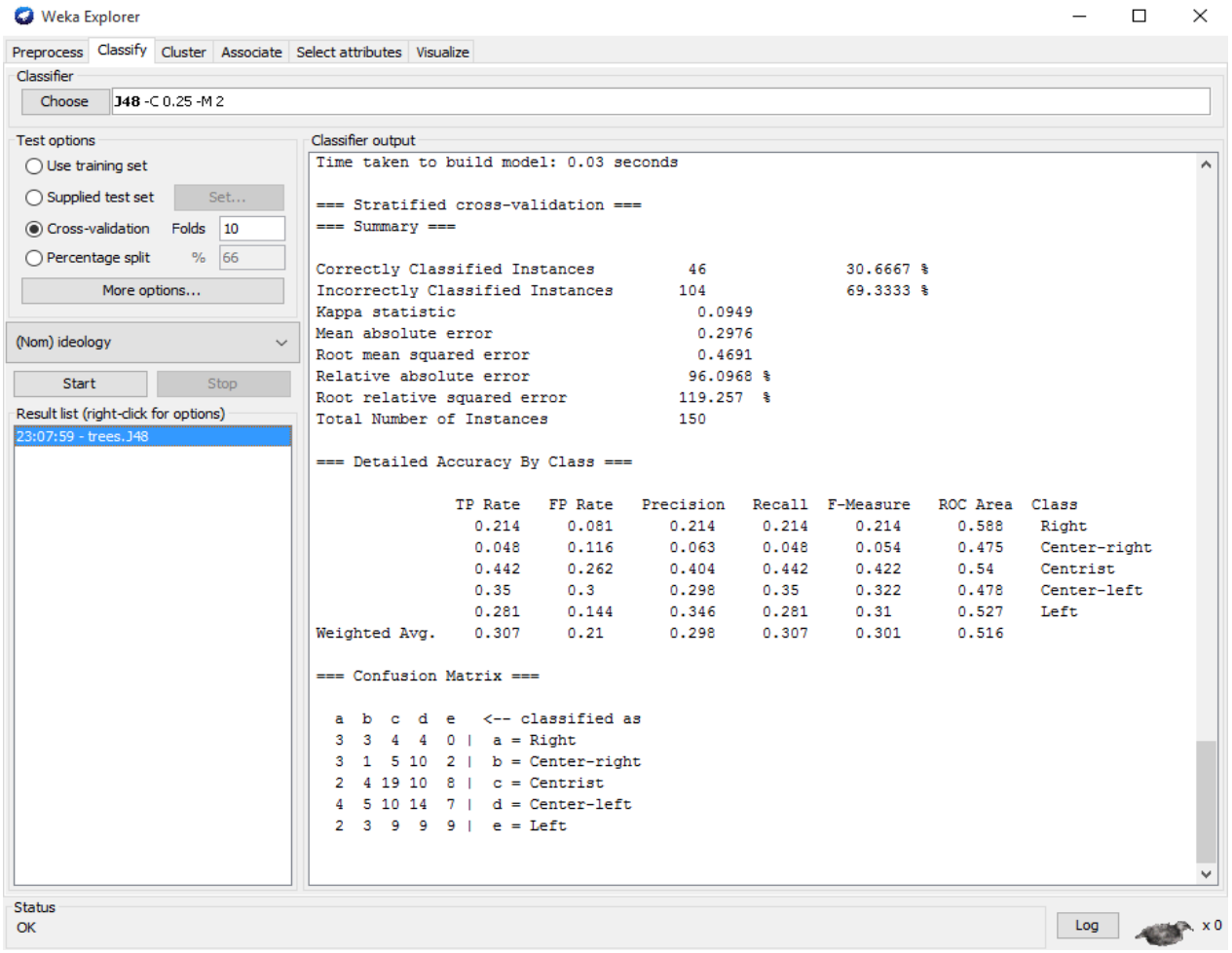
Εικόνα 12: Στιγμιότυπο του πίνακα σύγχυσης

4.3 Πρόβλημα πέντε κλάσεων

4.3.1 Πρόβλημα πέντε κλάσεων με χρήση όλων των χαρακτηριστικών

Το κύριο χαρακτηριστικό που μελετήθηκε στην εργασία ήταν το ideology, όπου την πρώτη φορά που πραγματοποιήθηκε η ανάλυσή του αποτελούνταν από πέντε κλάσεις. Συγκεκριμέναείχετηνμορφή: @attribute ideology {Right, Center-right, Centrist, Center-left, Left}. Με βάση αυτό το τύπο χαρακτηριστικού και τις κλάσεις που εμπεριέχονται σε αυτό, υλοποιήθηκαν οι αλγόριθμοι C4.5, SVM και NaïveBayes. Τα αποτελέσματα από τους τρεις αλγορίθμους αναλύονται παρακάτω.

4.3.1.1 C4.5 πέντε κλάσεων και χρήση όλων των χαρακτηριστικών



Εικόνα 13: Αποτελέσματα του αλγορίθμου C4.5 με πέντε κλάσεις και χρήση όλων των χαρακτηριστικών

Από την εκτέλεση του αλγορίθμου, προέκυψε ότι από τις 150 απαντήσεις που δόθηκαν, οι 46 από αυτές ταξινομήθηκαν σωστά ενώ οι 104 ταξινομήθηκαν εσφαλμένα. Η τιμή της πρόβλεψης συμφωνεί με την πραγματική κατηγορία κατά 0,0949 ενώ το μέσο μέγεθος σφαλμάτων από τις προβλέψεις που πραγματοποιήθηκαν ανέρχεται στο 0,2976. Αντίθετα, το μέσο μέγεθος του σφάλματος ήταν 0,4691. Το ποσοστό σχετικού σφάλματος ήταν 96%, ενώ το συνολικό τετραγωνικό σφάλμα ανέρχεται στο 119,25%. Στη συνέχεια ακολουθούν τα αποτελέσματα για κάθε κλάση του χαρακτηριστικού. Πιο αναλυτικά, για την κλάση right το ποσοστό των θετικών

περιπτώσεων που αναγνωρίστηκαν σωστά είναι 0,214 ενώ το ποσοστό των αρνητικών περιπτώσεων που ταξινομήθηκαν εσφαλμένα ως θετικά είναι 0,081. Από τα αποτελέσματα διαπιστώθηκε ότι η ακρίβεια, η ανάκληση και η σχέση μεταξύ τους είχαν την ίδια τιμή: 0,214, ενώ η απόδοση 0,588.

Για την κλάση center-right το ποσοστό των θετικών περιπτώσεων που αναγνωρίστηκαν σωστά είναι 0,048 ενώ το ποσοστό των αρνητικών περιπτώσεων που ταξινομήθηκαν εσφαλμένα ως θετικά είναι 0,116. Από τα αποτελέσματα διαπιστώθηκε ότι η ακρίβεια ήταν 0,063, η ανάκληση: 0,048 και η σχέση μεταξύ ακρίβειας και ανάκλησης: 0,054 . Ενώ η απόδοση της κλάσης ήταν 0,475. Επίσης, για την κλάση centrist το ποσοστό των θετικών περιπτώσεων που αναγνωρίστηκαν σωστά είναι 0,442 ενώ το ποσοστό των αρνητικών περιπτώσεων που ταξινομήθηκαν εσφαλμένα ως θετικά είναι 0,262. Από τα αποτελέσματα διαπιστώθηκε ότι η ακρίβεια ήταν 0,404, η ανάκληση: 0,442 και η σχέση μεταξύ ακρίβειας και ανάκλησης: 0,442 . Ενώ η απόδοση της κλάσης ήταν 0,54.

Χαμηλότερες τιμές διαπιστώθηκαν για την κλάση center-left όπου το ποσοστό των θετικών περιπτώσεων που αναγνωρίστηκαν σωστά είναι 0,35 ενώ το ποσοστό των αρνητικών περιπτώσεων που ταξινομήθηκαν εσφαλμένα ως θετικά είναι 0,3. Από τα αποτελέσματα διαπιστώθηκε ότι η ακρίβεια ήταν 0,298, η ανάκληση: 0,35 και η σχέση μεταξύ ακρίβειας και ανάκλησης: 0,322. Η απόδοση της κλάσης μετρήθηκε στο 0,527. Τέλος, για την κλάση left, διαπιστώθηκε ότι το ποσοστό των θετικών περιπτώσεων που αναγνωρίστηκαν σωστά είναι 0,281 ενώ το ποσοστό των αρνητικών περιπτώσεων που ταξινομήθηκαν εσφαλμένα ως θετικά είναι 0,144. Από τα αποτελέσματα διαπιστώθηκε ότι η ακρίβεια ήταν 0,346, η ανάκληση: 0,281 και η σχέση μεταξύ ακρίβειας και ανάκλησης: 0,31 ενώ η απόδοση της κλάσης είχε την τιμή: 0,527.

Ενώ τέλος, στον πίνακα σύγχυσης προκύπτουν τα ακόλουθα αποτελέσματα:

($1^{η}$ γραμμή,$1^{η}$ στήλη): Ο αλγόριθμος ταξινόμησε αληθώς σωστά τρεις ανθρώπους ως δεξιούς οι οποίοι δήλωσαν δεξιοί.

($1^{η}$ γραμμή, $2^{η}$ στήλη): Ο αλγόριθμος ταξινόμησε εσφαλμένα τρεις ανθρώπους ως κεντροδεξιούς ενώ δήλωσαν δεξιοί.

($1^{η}$ γραμμή, $3^{η}$ στήλη): Ο αλγόριθμος ταξινόμησε εσφαλμένα τέσσερις ανθρώπους κεντρώους ενώ δήλωσαν δεξιοί.

($1^{η}$ γραμμή, $4^{η}$ στήλη): Ο αλγόριθμος ταξινόμησε εσφαλμένα τέσσερις ανθρώπους ως κεντροαριστερούς, ενώ δήλωσαν δεξιοί.

($1^{η}$ γραμμή, $5^{η}$ στήλη): Ο αλγόριθμος εντόπισε τη βέλτιστη λύση.

($2^{η}$ γραμμή, $1^{η}$ στήλη): Ο αλγόριθμος ταξινόμησε εσφαλμένα τρεις ανθρώπους ως δεξιοί οι οποίοι δήλωσαν κεντροδεξιοί.

($2^{η}$ γραμμή, $2^{η}$ στήλη): Ο αλγόριθμος ταξινόμησε αληθώς σωστά έναν άνθρωπο ως κεντροδεξιό ,ο οποίος δήλωσε κεντροδεξιός.

($2^{η}$ γραμμή, $3^{η}$ στήλη): Ο αλγόριθμος ταξινόμησε εσφαλμένα πέντε ανθρώπους ως κεντρώους οι οποίοι δήλωσαν κεντροδεξιοί.

($2^{η}$ γραμμή, $4^{η}$ στήλη): Ο αλγόριθμος ταξινόμησε εσφαλμένα δέκα ανθρώπους ως κεντροαριστεροί, οι οποίοι δήλωσαν κεντροδεξιοί.

($2^{η}$ γραμμή, $5^{η}$ στήλη): Ο αλγόριθμος ταξινόμησε εσφαλμένα δύο ανθρώπους ως αριστερούς, οι οποίοι δήλωσαν κεντροδεξιοί.

($3^{η}$ γραμμή, $1^{η}$ στήλη): Ο αλγόριθμος ταξινόμησε εσφαλμένα δύο ανθρώπους ως δεξιούς, οι οποίοι δήλωσαν κεντρώοι.

($3^{η}$ γραμμή, $2^{η}$ στήλη): Ο αλγόριθμος ταξινόμησε εσφαλμένα τέσσερις ανθρώπους ως κεντροδεξιούς, οι οποίοι δήλωσαν κεντρώοι.

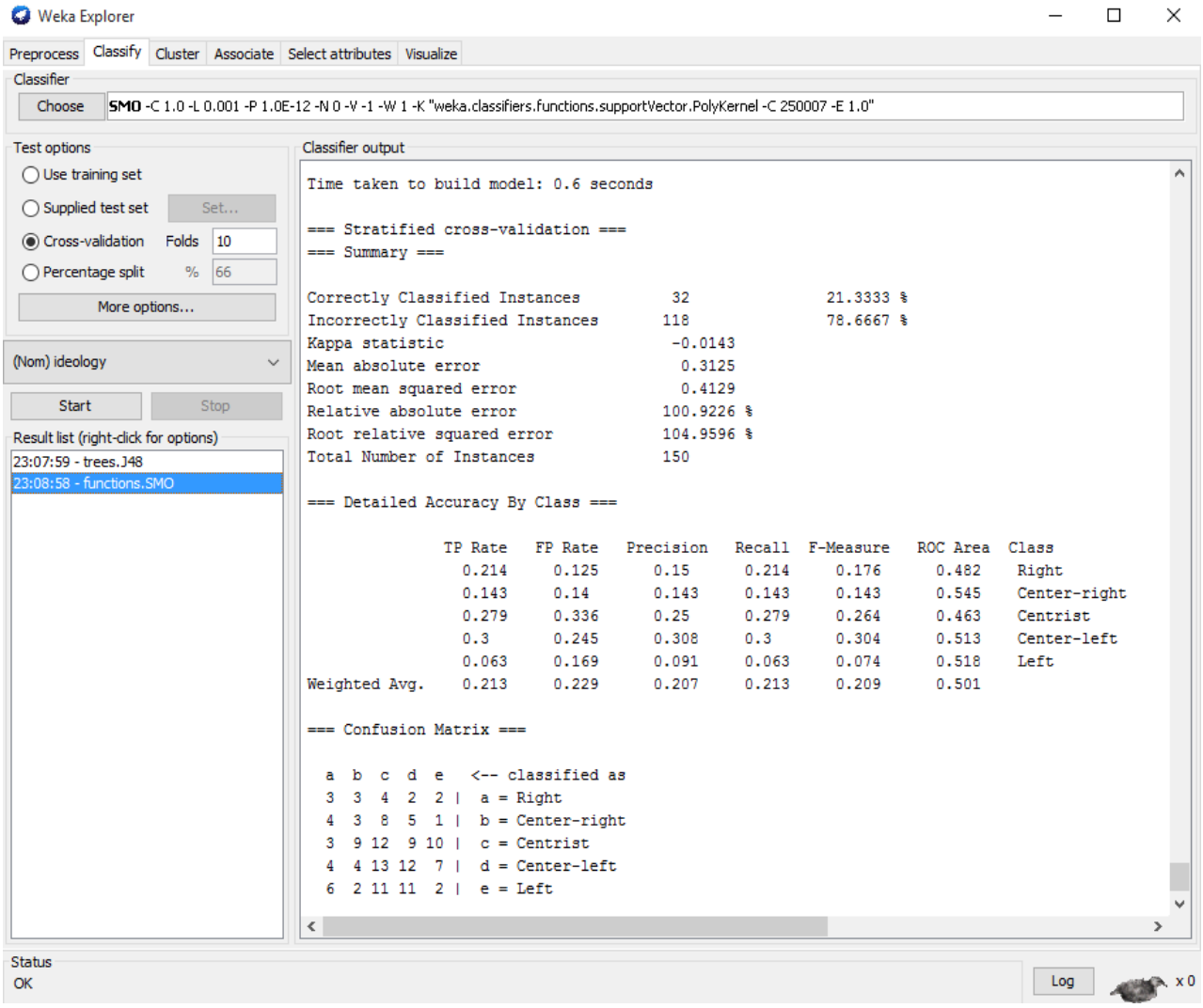
($3^{η}$ γραμμή, $3^{η}$ στήλη): Ο αλγόριθμος ταξινόμησε σωστά δεκαεννιά ανθρώπους ως κεντρώους, οι οποίοι δήλωσαν κεντρώοι.

($3^{η}$ γραμμή, $4^{η}$ στήλη): Ο αλγόριθμος ταξινόμησε εσφαλμένα δέκα ανθρώπους ως κεντροαριστερούς, οι οποίοι δήλωσαν κεντρώοι.

($3^{η}$ γραμμή, $5^{η}$ στήλη): Ο αλγόριθμος ταξινόμησε εσφαλμένα οχτώ ανθρώπους ως αριστερούς, οι οποίοι δήλωσαν κεντρώοι.

Η ίδια διαδικασία επεξήγησης ακολουθήθηκε σε ολόκληρο τον πίνακα σύγχυσης, όπου διαπιστώθηκε ότι η βέλτιστη λύση βρίσκεται στην $1^{η}$ γραμμή - $5^{η}$ στήλη, ενώ στην κύρια διαγώνιο απεικονίζονται τα ορθά στοιχεία της ταξινόμησης. Τα υπόλοιπα στοιχεία του πίνακα απεικονίζουν τις εσφαλμένες ταξινομήσεις.

4.3.1.2 SVM πέντε κλάσεων και χρήση όλων των χαρακτηριστικών

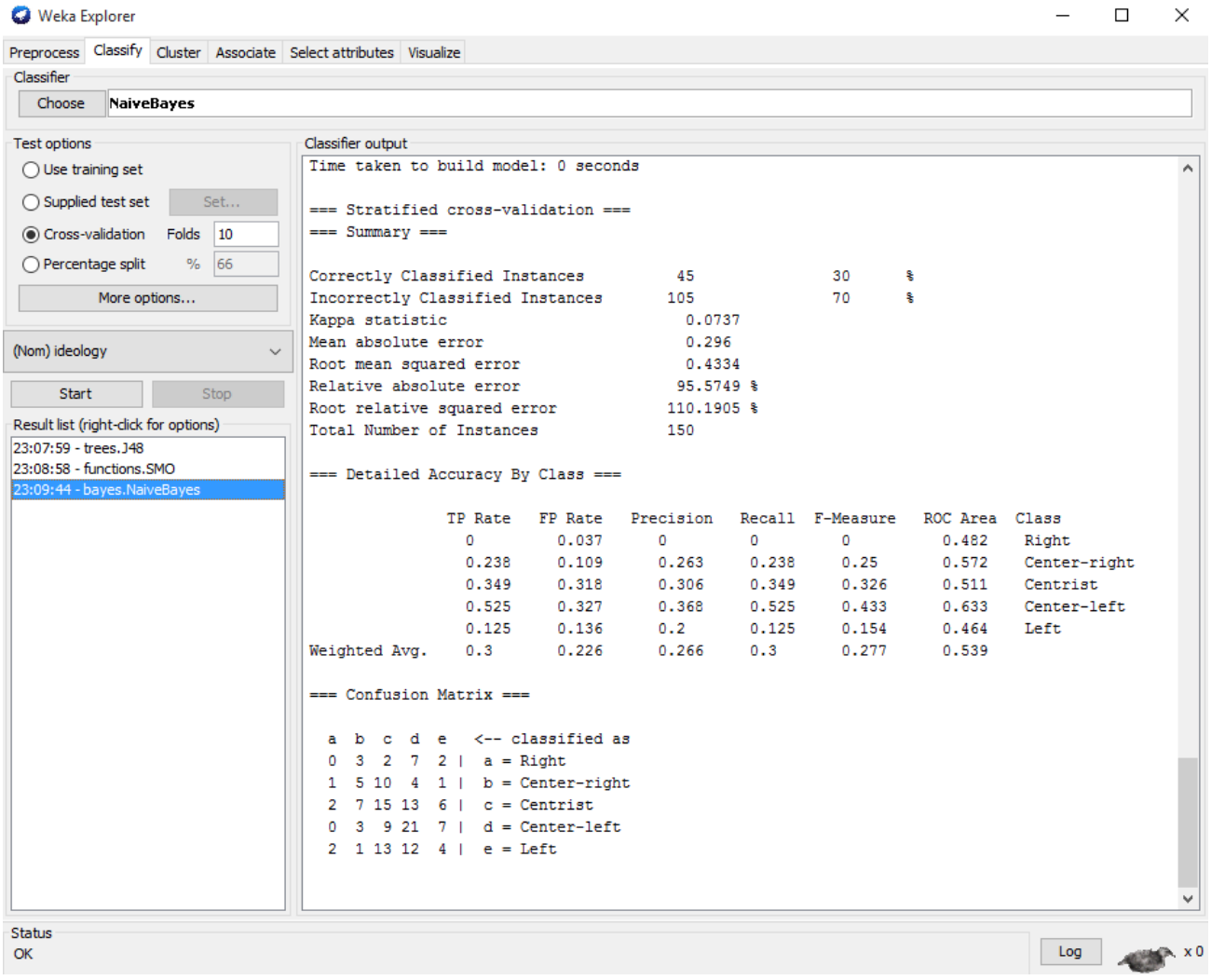


Εικόνα 14: Αποτελέσματα του αλγορίθμου SMO με πέντε κλάσεις και χρήση όλων των χαρακτηριστικών

Αντίθετα με τον προηγούμενο αλγόριθμο, στον SVM αλγόριθμο παρατηρήθηκε ότι ταξινομήθηκαν περισσότερα εσφαλμένα στοιχεία (118 από τα 150), μόνο τα 32

στοιχεία ταξινομήθηκαν σωστά, ενώ το ποσοστό σχετικού σφάλματος εντοπίστηκε στο 100%. Η κλάση με την υψηλότερη τιμή των σωστά αναγνωρισμένων θετικών περιπτώσεων την είχε η center-left, ενώ η κλάση με το υψηλότερο ποσοστό των αρνητικών περιπτώσεων που ταξινομήθηκαν εσφαλμένα ως θετικά ήταν η κλάση centrist. Μεγαλύτερη ακρίβεια διαπιστώθηκε στην κλάση center-left, ενώ μεγαλύτερη ανάκληση πραγματοποιήθηκε στην κλάση centrist. Όπως διαπιστώθηκε, η κλάση με την υψηλότερη απόδοση ήταν η center-right. Τέλος, από τον πίνακα σύγχυσης δεν εντοπίστηκε κάποια βέλτιστη τιμή, παρά μόνο τα ορθά στοιχεία της ταξινόμησης που βρίσκονται στην κύρια διαγώνιο.

4.3.1.3 NaïveBayes πέντε κλάσεων και χρήση όλων των χαρακτηριστικών

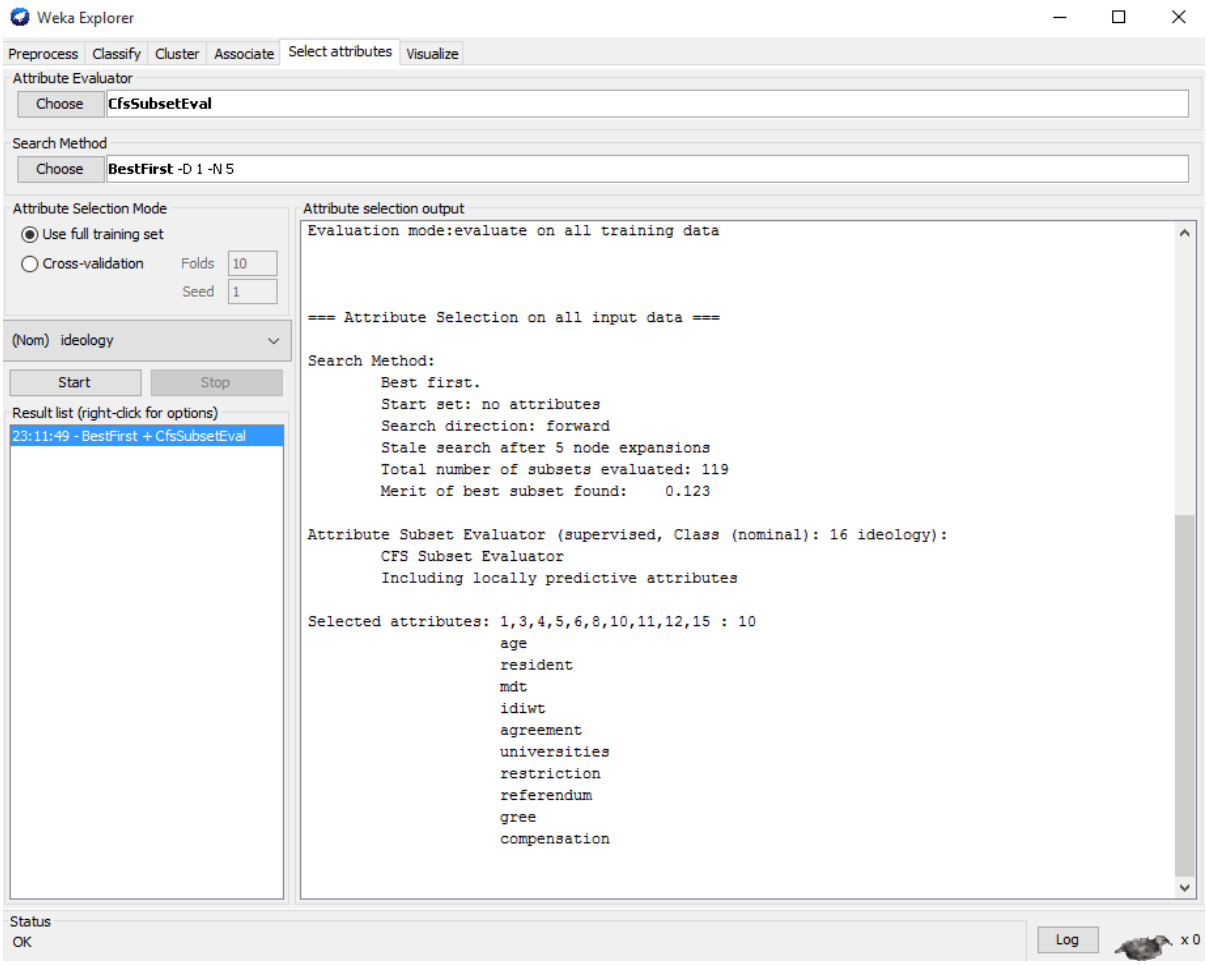


Εικόνα 15: Αποτελέσματα του αλγορίθμου NaïveBayes με πέντε κλάσεις και χρήση όλων των χαρακτηριστικών

Στον NaïveBayes αλγόριθμο παρατηρήθηκε ότι ταξινομήθηκαν περισσότερα εσφαλμένα στοιχεία (105 από τα 150) σε σχέση με τα σωστά ταξινομημένα στοιχεία (45), ενώ το ποσοστό σχετικού σφάλματος εντοπίστηκε στο 95%. Η κλάση με την υψηλότερη τιμή των σωστά αναγνωρισμένων θετικών περιπτώσεων και των αρνητικών περιπτώσεων που ταξινομήθηκαν εσφαλμένα ως θετικά ήταν η center-left. Μεγαλύτερη ακρίβεια, ανάκληση και απόδοση διαπιστώθηκε στην κλάση center-left. Τέλος, στον πίνακα σύγχυσης διαπιστώθηκαν δύο βέλτιστες τιμές στις θέσεις: $1^{η}$ γραμμή- $1^{η}$ στήλη και $4^{η}$ γραμμή-$1^{η}$ στήλη, ενώ στην κύρια διαγώνιο απεικονίζονται οι σωστά τα ορθά στοιχεία της ταξινόμησης.

4.3.2 Πρόβλημα πέντε κλάσεων με επιλογή χαρακτηριστικών

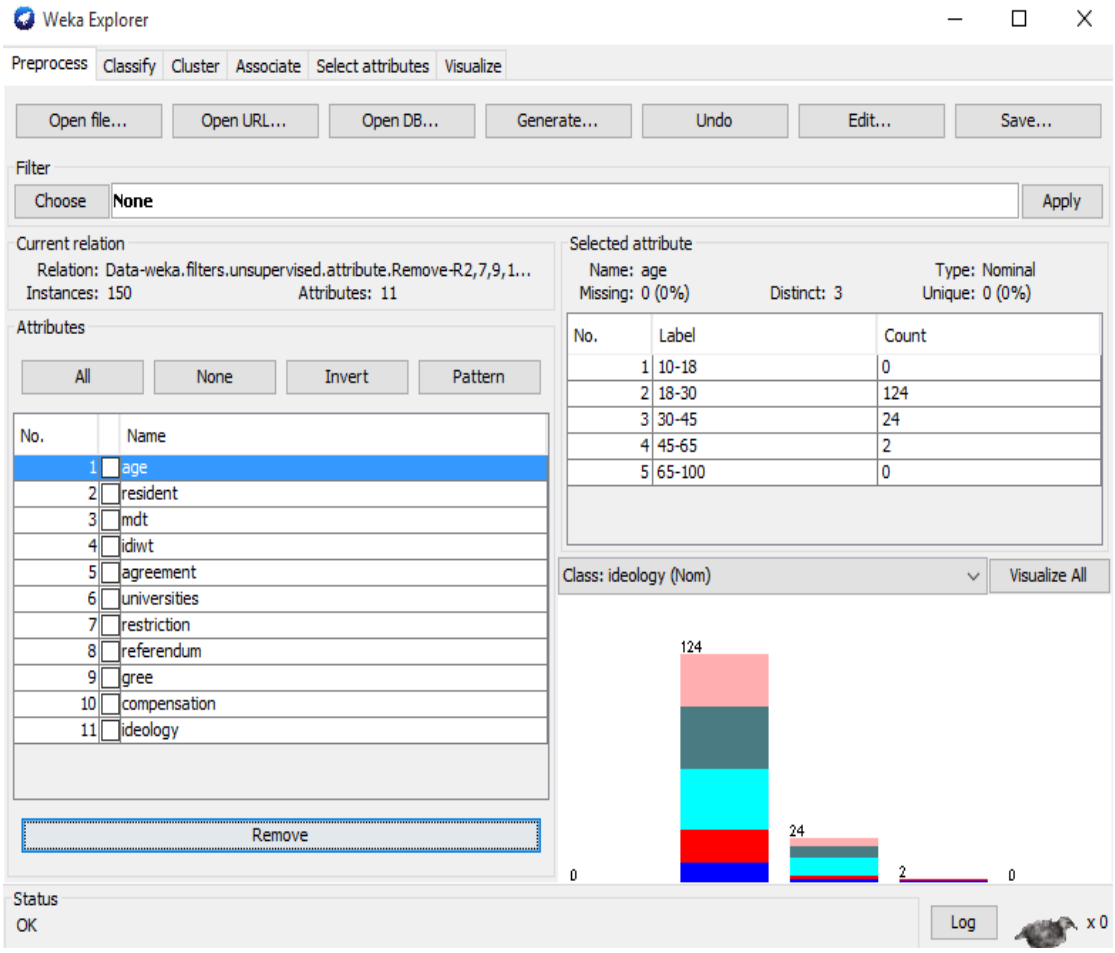
Επιλέγοντας το Selectattribute με 5 κλάσεις στο χαρακτηριστικό ideology προκύπτουν τα ακόλουθα καλύτερα χαρακτηριστικά:



Εικόνα 16: Επιλογή χαρακτηριστικών με πέντε κλάσεις

Σύμφωνα με τα αποτελέσματα δέκα από τα δεκαέξι χαρακτηριστικά είναι τα καλύτερα για το χαρακτηριστικό ideology με «όρισμα» πέντε κλάσεις.

Στη συνέχεια επιλέχθηκαν μόνο τα καλύτερα χαρακτηριστικά, με βάση το χαρακτηριστικό ideology με πέντε κλάσεις όπως φαίνεται παρακάτω:



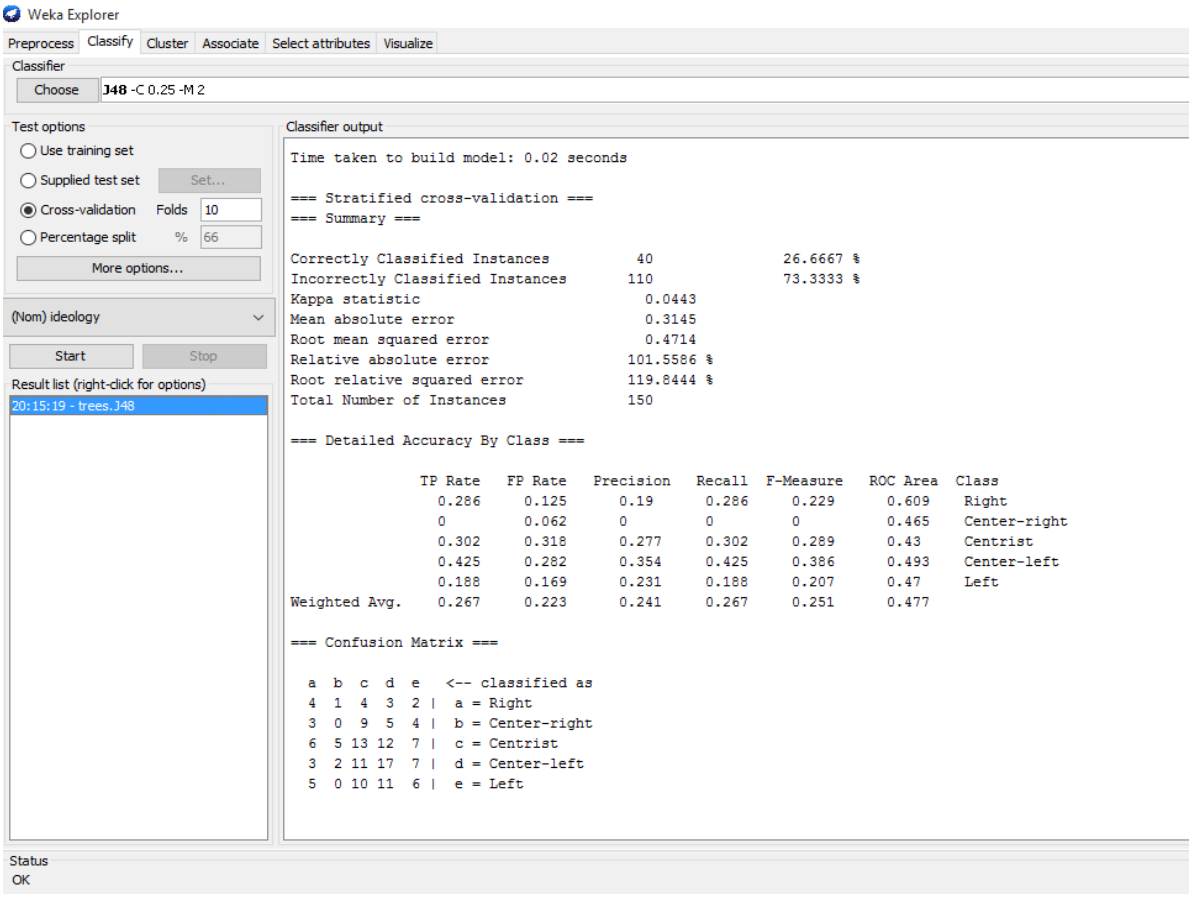
Εικόνα 17: Επιλογή των καλύτερων χαρακτηριστικών με πέντε κλάσεις

Με βάση τα δέκα αυτά καλύτερα χαρακτηριστικά επιλέχθηκαν οι τρεις αλγόριθμοι και προέκυψαν τα ακόλουθα αποτελέσματα.

4.3.2.1 C4.5 πέντε κλάσεων και επιλογή χαρακτηριστικών

Από την εκτέλεση του αλγορίθμου C4.5 παρατηρήθηκε ότι τα 110 από τα 150 στοιχεία ταξινομήθηκαν εσφαλμένα και μόνο τα 40 στοιχεία ταξινομήθηκαν σωστά. Το ποσοστό σχετικού σφάλματος εντοπίστηκε στο 100%. Η κλάση με την υψηλότερη τιμή των σωστά αναγνωρισμένων θετικών περιπτώσεων ήταν center-left και με το υψηλότερο ποσοστό των αρνητικών περιπτώσεων που ταξινομήθηκαν εσφαλμένα ως θετικά την είχε η κλάση centrist. Ενώ τέλος, μεγαλύτερη ακρίβεια, ανάκληση και απόδοση διαπιστώθηκε στην κλάση center-left. Αναφορικά με τον πίνακα σύγχυσης

εντοπίστηκαν δύο βέλτιστες τιμές στις θέσεις: $2^{η}$ γραμμή-$2^{η}$ στήλη και $5^{η}$ γραμμή-$2^{η}$ στήλη του πίνακα. Επίσης στην κύρια διαγώνιο του πίνακα απεικονίζονται τα ορθά στοιχεία της ταξινόμησης.

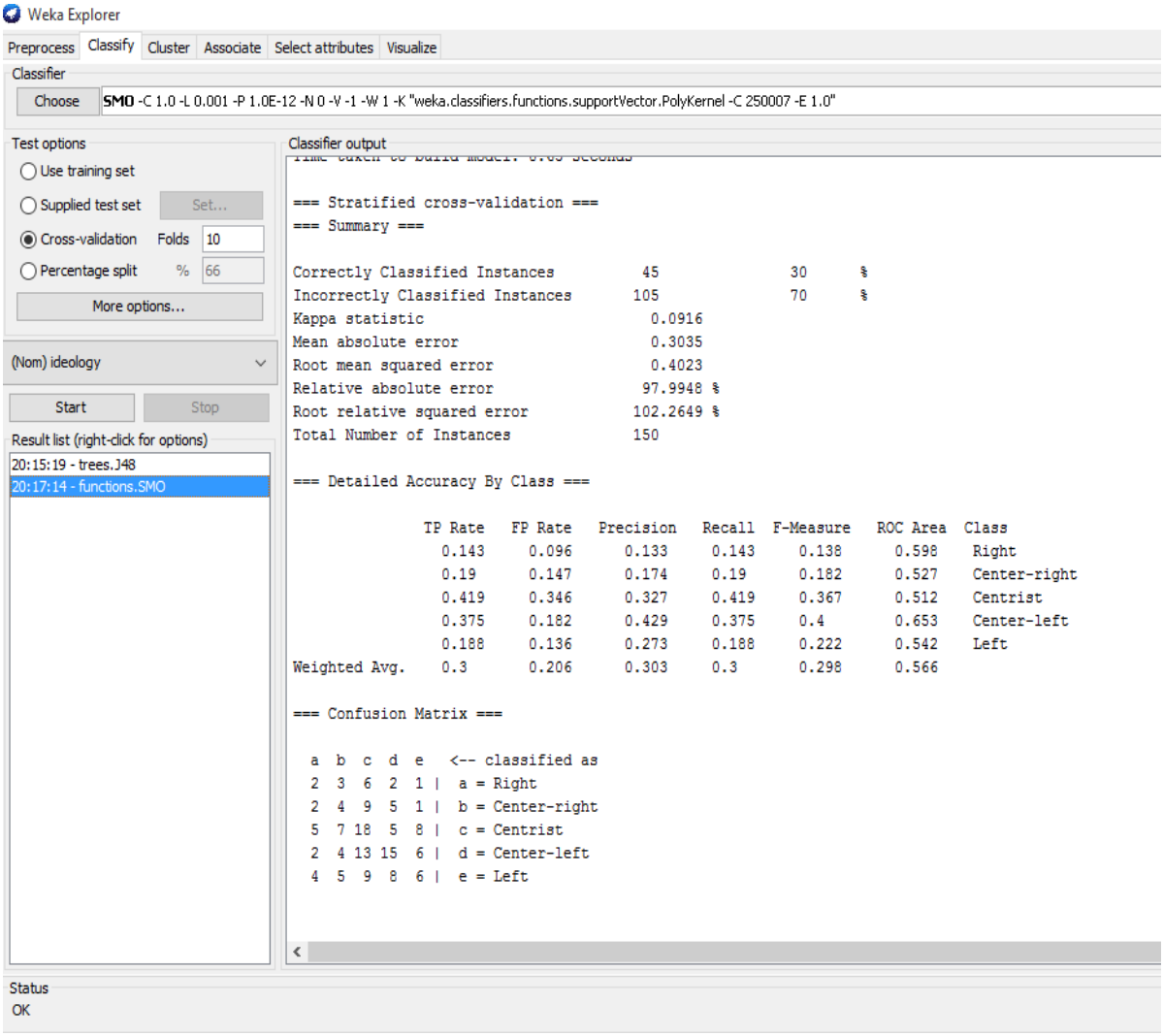


Εικόνα 18: Αποτελέσματα του αλγορίθμου C4.5 με πέντε κλάσεις και επιλογή των καλύτερων χαρακτηριστικών

4.3.2.2 SVM πέντε κλάσεων και επιλογή χαρακτηριστικών

Από την εκτέλεση του αλγορίθμου SVM παρατηρήθηκε ότι τα 105 στοιχεία του συνόλου ταξινομήθηκαν εσφαλμένα και μόνο τα 45 στοιχεία ταξινομήθηκαν σωστά, ενώ το ποσοστό σχετικού σφάλματος εντοπίστηκε στο 97%. Η κλάση με την υψηλότερη τιμή των σωστά αναγνωρισμένων θετικών περιπτώσεων και με το υψηλότερο ποσοστό των αρνητικών περιπτώσεων που ταξινομήθηκαν εσφαλμένα ως θετικά ανιχνεύθηκε στην centrist κλάση. Ενώ τέλος, μεγαλύτερη ακρίβεια εντοπίστηκε στην κλάση center-left, υψηλότερη ανάκληση στην centrist κλάση, ενώ μεγαλύτερη απόδοση διαπιστώθηκε στην κλάση center-left. Αναφορικά με τον πίνακα

σύγχυσης δεν εντοπίστηκαν βέλτιστες τιμές , ενώ στην κύρια διαγώνιο του πίνακα απεικονίζονται τα ορθά στοιχεία της ταξινόμησης.

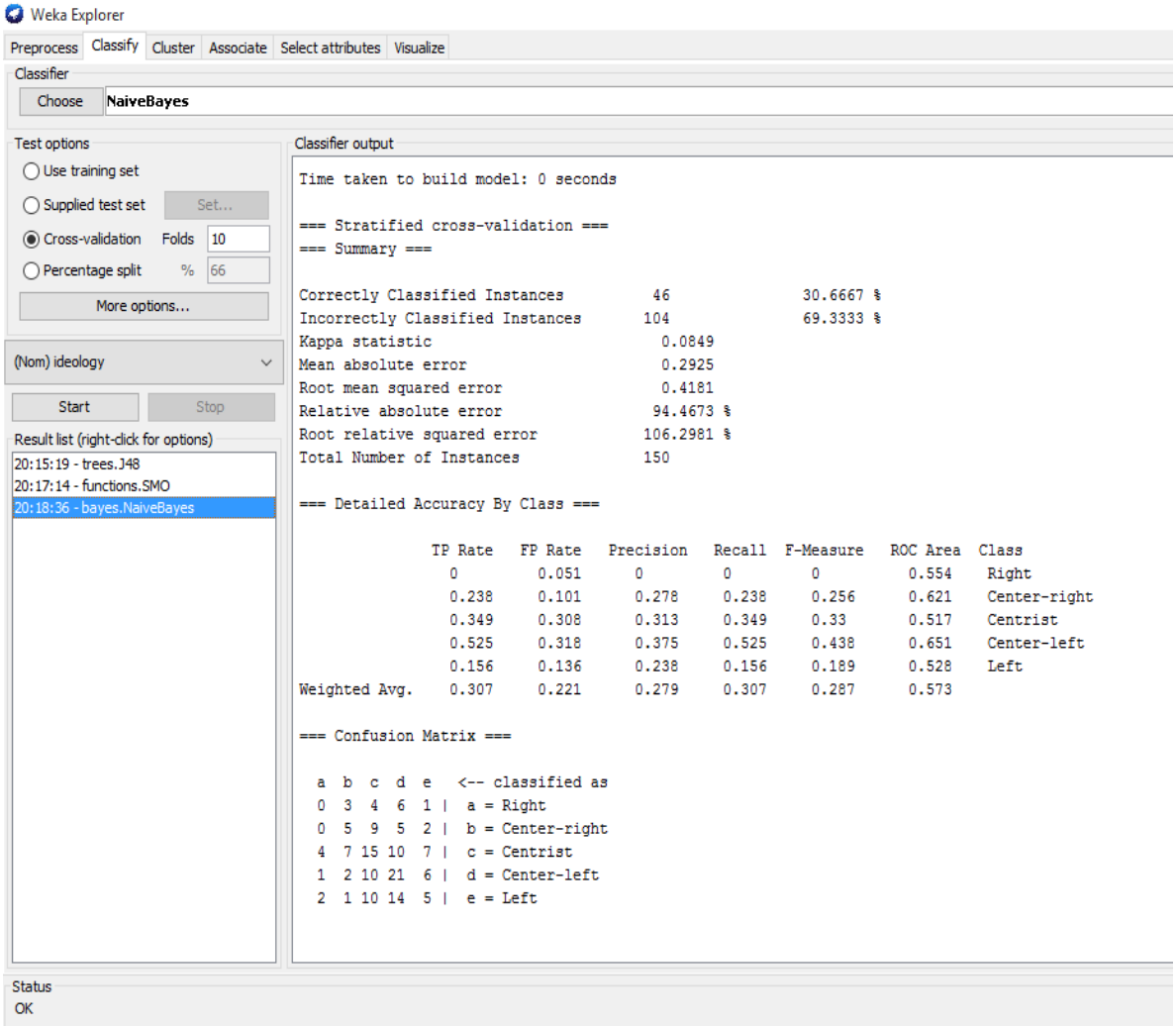


Εικόνα 18: Αποτελέσματα του αλγορίθμου SVM με πέντε κλάσεις και επιλογή των καλύτερων χαρακτηριστικών

4.3.2.3 NaïveBayes πέντε κλάσεων και επιλογή χαρακτηριστικών

Στον αλγόριθμο NaïveBayes παρατηρήθηκε ότι τα 104 στοιχεία του συνόλου ταξινομήθηκαν εσφαλμένα και μόνο τα 46 στοιχεία ταξινομήθηκαν σωστά, ενώ το ποσοστό σχετικού σφάλματος εντοπίστηκε στο 94%. Η κλάση με την υψηλότερη τιμή των σωστά αναγνωρισμένων θετικών περιπτώσεων και με το υψηλότερο ποσοστό των αρνητικών περιπτώσεων που ταξινομήθηκαν εσφαλμένα ως θετικά ανιχνεύθηκε στην center-left κλάση. Ενώ τέλος, υψηλότερη ακρίβεια, ανάκληση και απόδοση διαπιστώθηκε στην κλάση center-left. Αναφορικά με τον πίνακα σύγχυσης εντοπίστηκαν δύο βέλτιστες τιμές στις θέσεις: 1γραμμή-$1^{η}$ στήλη και $2^{η}$ γραμμή-$1^{η}$

στήλη , ενώ στην κύρια διαγώνιο του πίνακα απεικονίζονται τα ορθά στοιχεία της ταξινόμησης.



Εικόνα 19: Αποτελέσματα του αλγορίθμου NaïveBayes με πέντε κλάσεις και επιλογή των καλύτερων χαρακτηριστικών

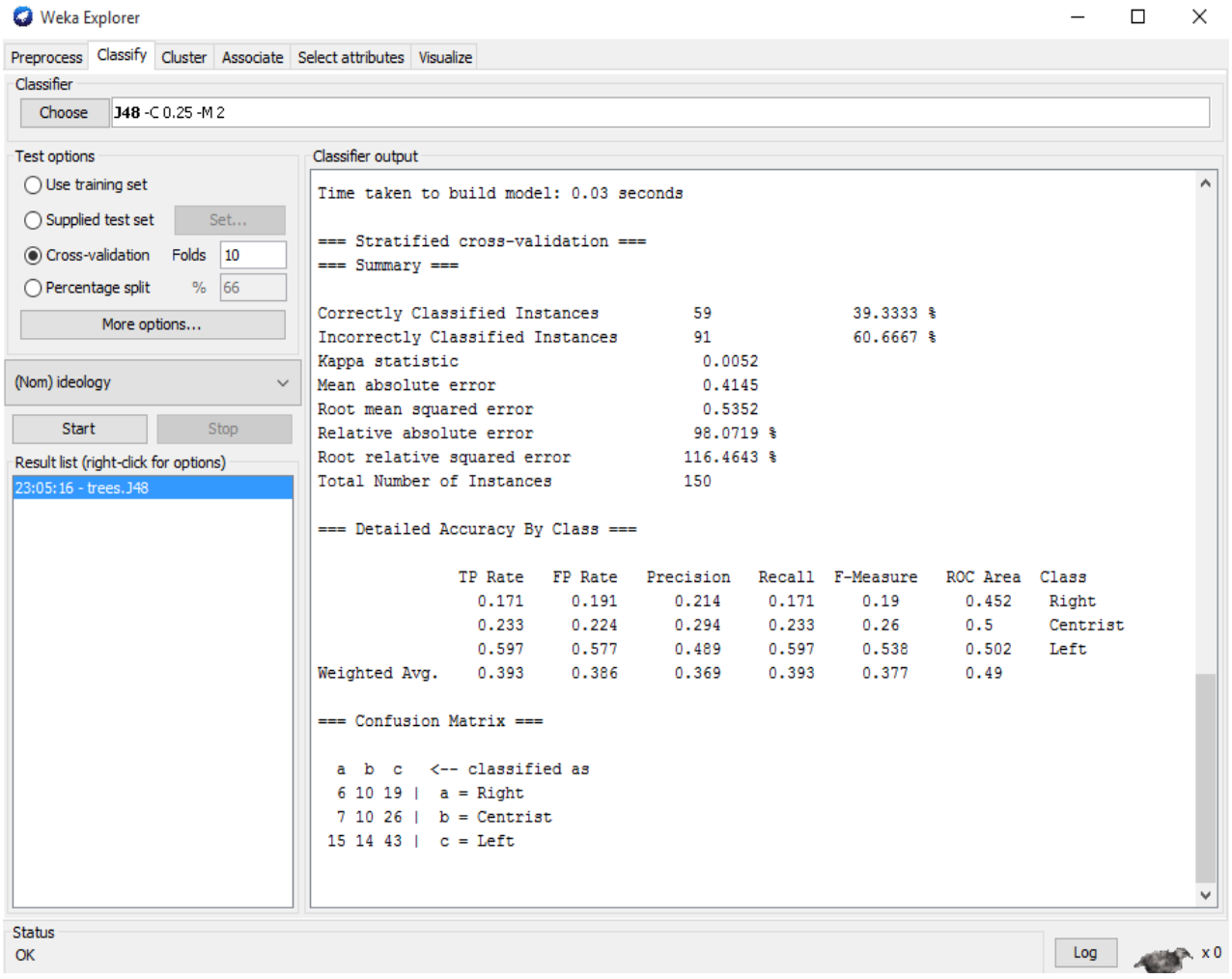
4.4 Πρόβλημα τριών κλάσεων

4.4.1 Πρόβλημα τριών κλάσεων με χρήση όλων των χαρακτηριστικών

Στη συνέχεια συγχωνεύτηκαν οι κλάσεις στο χαρακτηριστικό @attributeideology και απέκτησε τη μορφή: @attributeideology {Right, Centrist, Left}. Τα αποτελέσματα από τους τρεις αλγορίθμους που προέκυψαν απεικονίζονται παρακάτω.

4.4.1.1 C4.5 τριών κλάσεων και χρήση όλων των χαρακτηριστικών

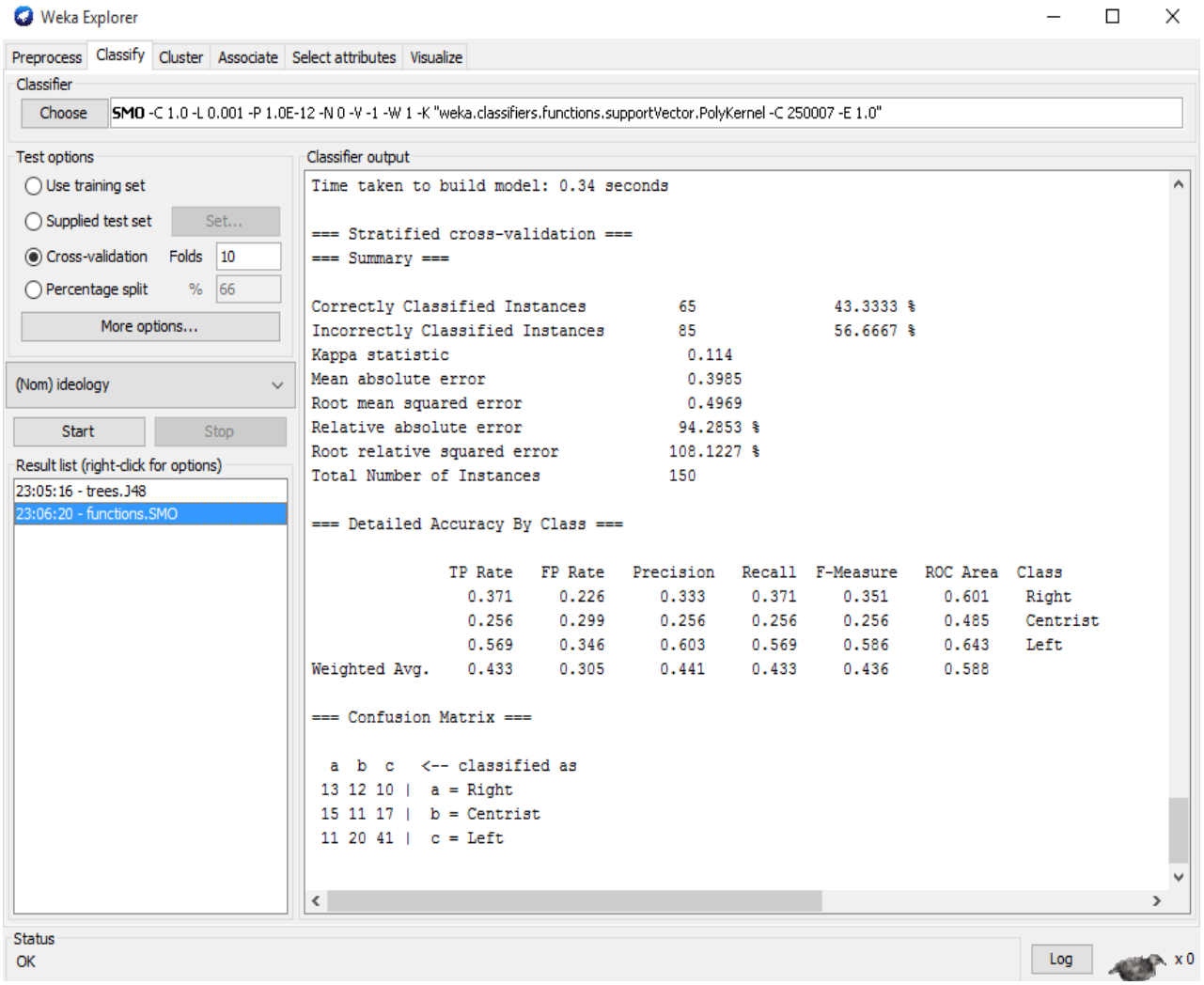
Από την εκτέλεση του αλγορίθμου C4.5 παρατηρήθηκε ότι τα 91 από τα 150 στοιχεία ταξινομήθηκαν εσφαλμένα και μόνο τα 59 στοιχεία ταξινομήθηκαν σωστά. Ενώ το ποσοστό σχετικού σφάλματος εντοπίστηκε στο 98%. Η κλάση με την υψηλότερη τιμή των σωστά αναγνωρισμένων θετικών περιπτώσεων και με το υψηλότερο ποσοστό των αρνητικών περιπτώσεων που ταξινομήθηκαν εσφαλμένα ως θετικά την είχε η κλάση left. Ενώ τέλος, μεγαλύτερη ακρίβεια, ανάκληση και απόδοση διαπιστώθηκε στην κλάση left. Αναφορικά με τον πίνακα σύγχυσης δεν εντοπίστηκαν βέλτιστες τιμές, παρά μόνο τα ορθά στοιχεία της ταξινόμησης που εντοπίζονται στην κύρια διαγώνιο του πίνακα.



Εικόνα 20: Αποτελέσματα του αλγορίθμου C4.5 με τρεις κλάσεις και χρήση όλων των χαρακτηριστικών

4.4.1.2 SMO τριών κλάσεων και χρήση όλων των χαρακτηριστικών

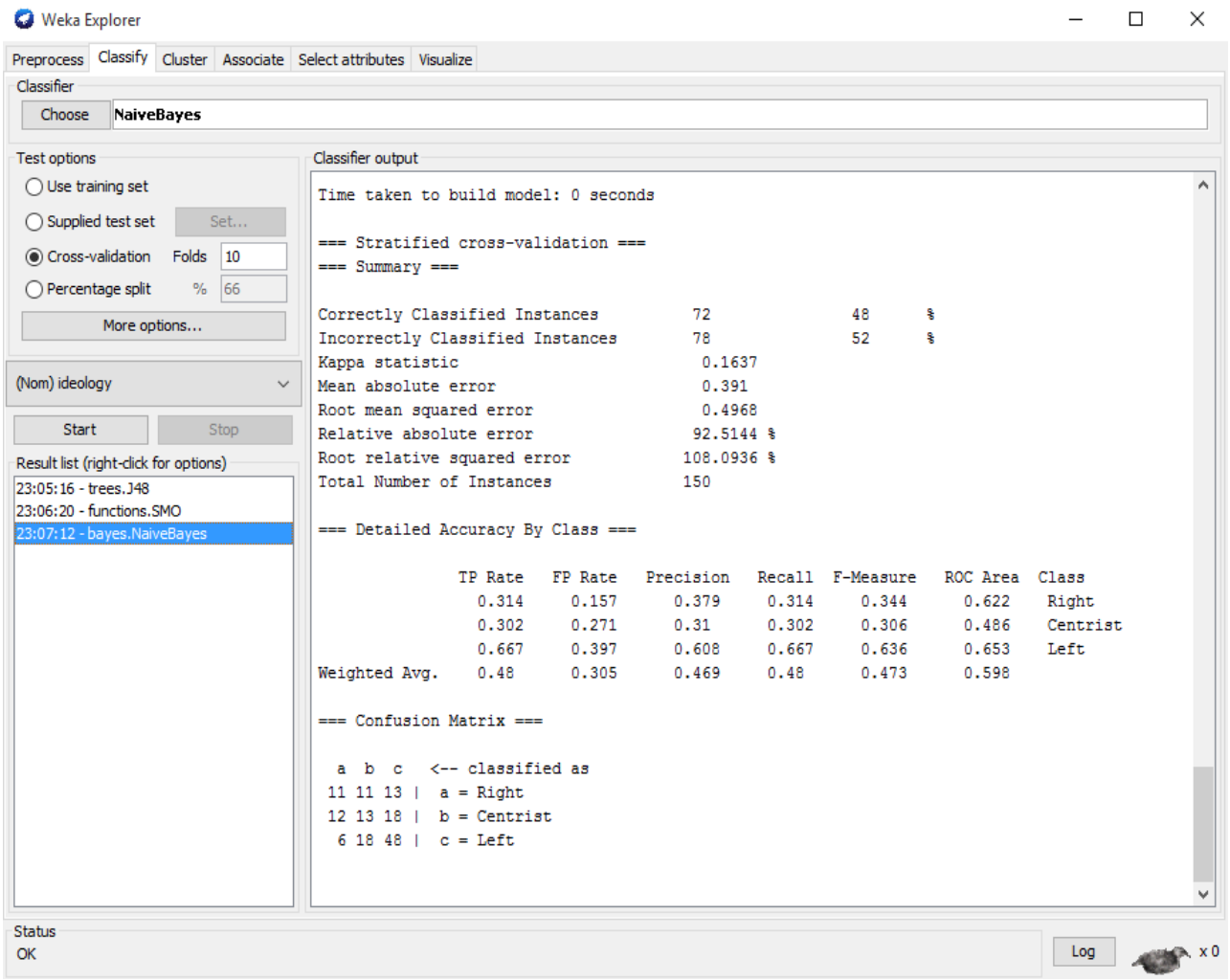
Στον SVM αλγόριθμο παρατηρήθηκε ότι τα 85 από τα 150 στοιχεία ταξινομήθηκαν εσφαλμένα, ενώ τα 65 στοιχεία ταξινομήθηκαν σωστά. Το ποσοστό σχετικού σφάλματος στον αλγόριθμο SVMεντοπίστηκε στο 94%. Η κλάση με την υψηλότερη τιμή των σωστά αναγνωρισμένων θετικών περιπτώσεων και με το υψηλότερο ποσοστό των αρνητικών περιπτώσεων που ταξινομήθηκαν εσφαλμένα ως θετικά την είχε η κλάση left. Ενώ τέλος, μεγαλύτερη ακρίβεια, ανάκληση και απόδοση διαπιστώθηκε στην κλάση left. Αναφορικά με τον πίνακα σύγχυσης δεν εντοπίστηκαν βέλτιστες τιμές, παρά μόνο τα ορθά στοιχεία της ταξινόμησης που εντοπίζονται στην κύρια διαγώνιο του πίνακα.



Εικόνα 21:Αποτελέσματα του αλγορίθμου SVM με τρεις κλάσεις και χρήση όλων των χαρακτηριστικών

4.4.1.3 NaïveBayes τριών κλάσεων και χρήση όλων των χαρακτηριστικών

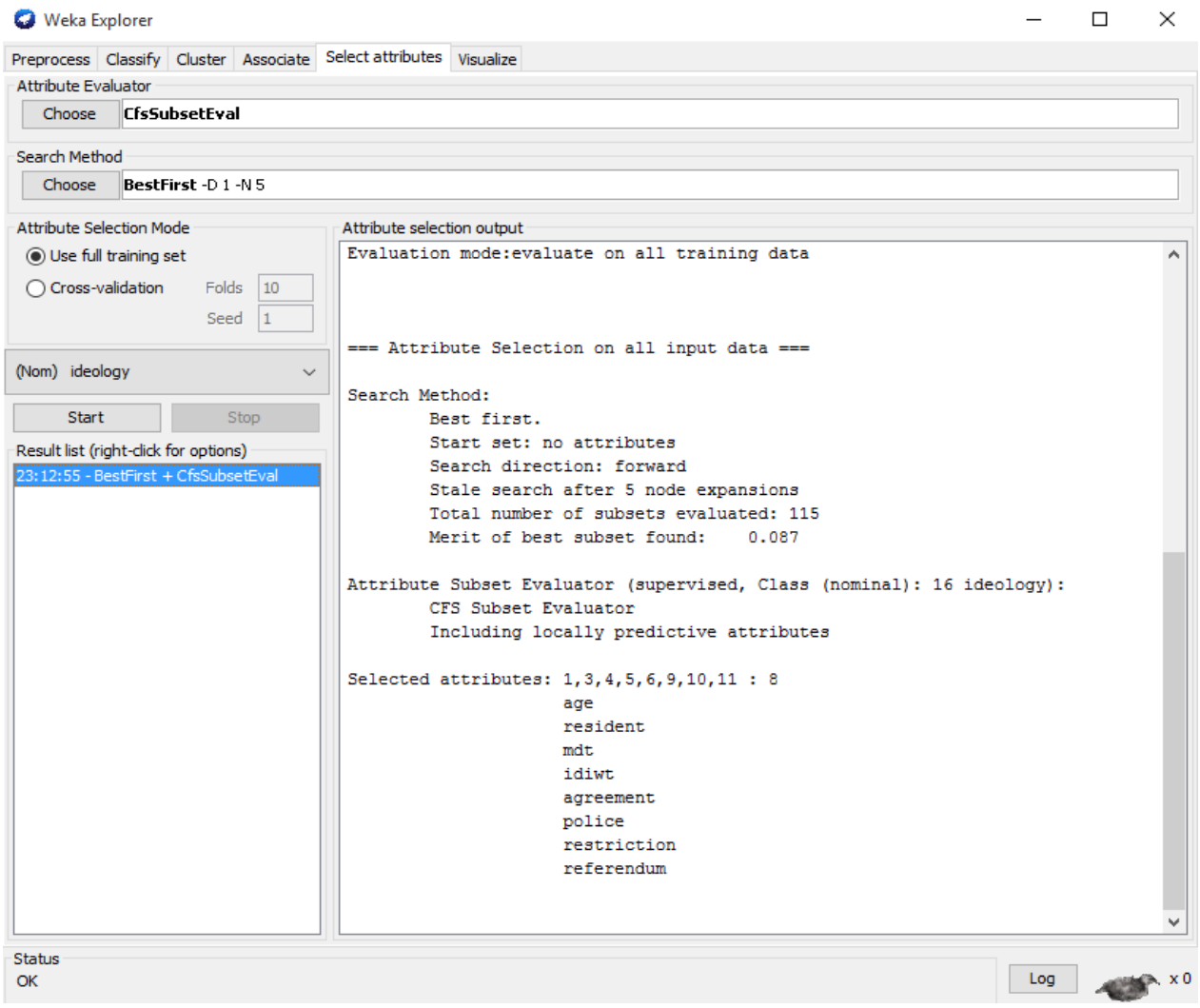
Στον NaïveBayes αλγόριθμο παρατηρήθηκε ότι ταξινομήθηκαν 78 εσφαλμένα στοιχεία και 72 σωστά. Το ποσοστό σχετικού σφάλματος εντοπίστηκε στο 92,5%. Η κλάση με την υψηλότερη τιμή των σωστά αναγνωρισμένων θετικών περιπτώσεων και των αρνητικών περιπτώσεων που ταξινομήθηκαν εσφαλμένα ως θετικά ήταν η κλάση left. Ομοίως και στην ακρίβεια, στην ανάκληση και στην απόδοση διαπιστώθηκε με υψηλότερα ποσοστά η κλάση left. Τέλος, στον πίνακα σύγχυσης δεν εντοπίστηκαν βέλτιστες τιμές, παρά μόνο τα ορθά στοιχεία της ταξινόμησης που εντοπίζονται στην κύρια διαγώνιο του πίνακα.



Εικόνα 22:Αποτελέσματα του αλγορίθμου NaïveBayes με τρεις κλάσεις και χρήση όλων των χαρακτηριστικών

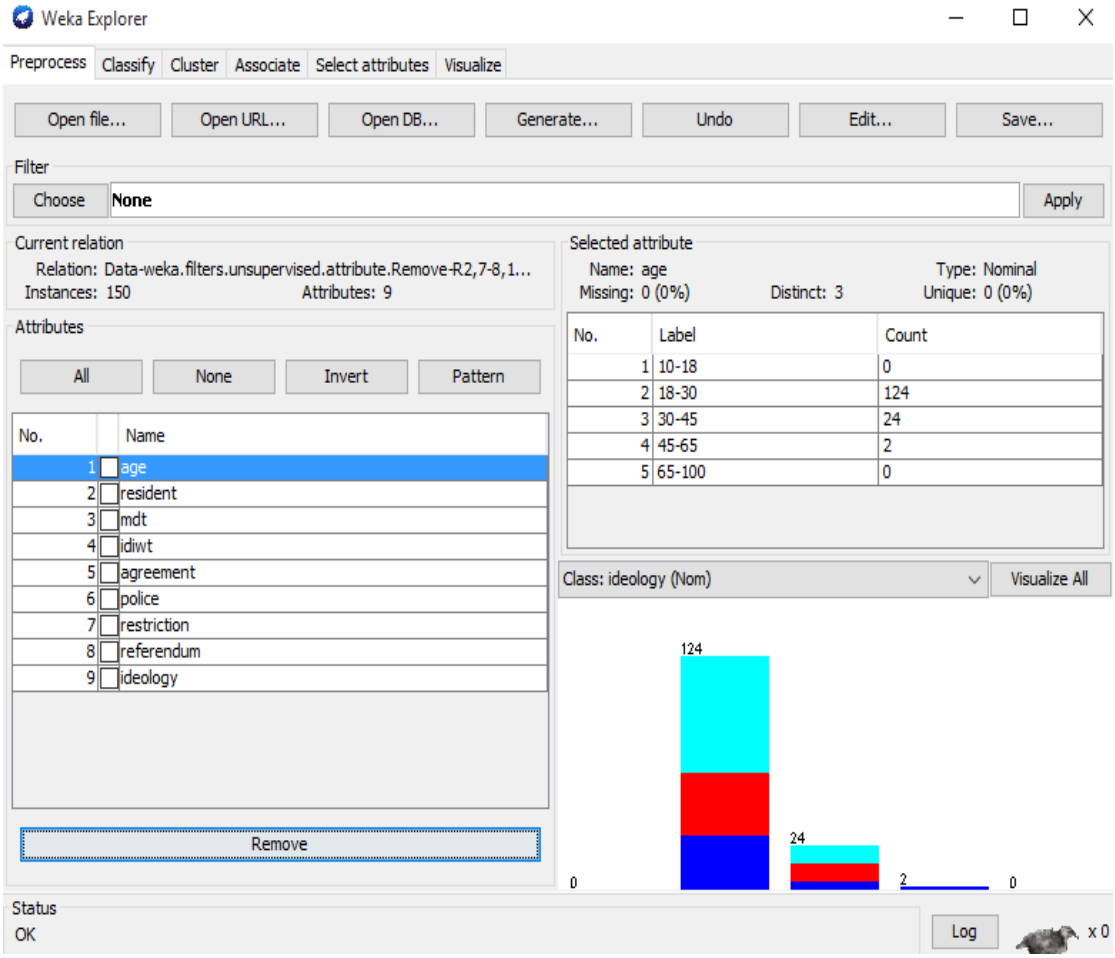
4.4.2 Πρόβλημα τριών κλάσεων με επιλογή χαρακτηριστικών

Επιλέγοντας το Selectattribute με τρεις κλάσεις στο χαρακτηριστικό ideology, προκύπτουν τα οχτώ καλύτερα χαρακτηριστικά όπως απεικονίζονται παρακάτω:



Εικόνα 23: Επιλογή χαρακτηριστικών με τρεις κλάσεις

Στη συνέχεια επιλέχθηκαν τα καλύτερα χαρακτηριστικά, με βάση το χαρακτηριστικό ideology με τρεις κλάσεις όπως φαίνεται παρακάτω.

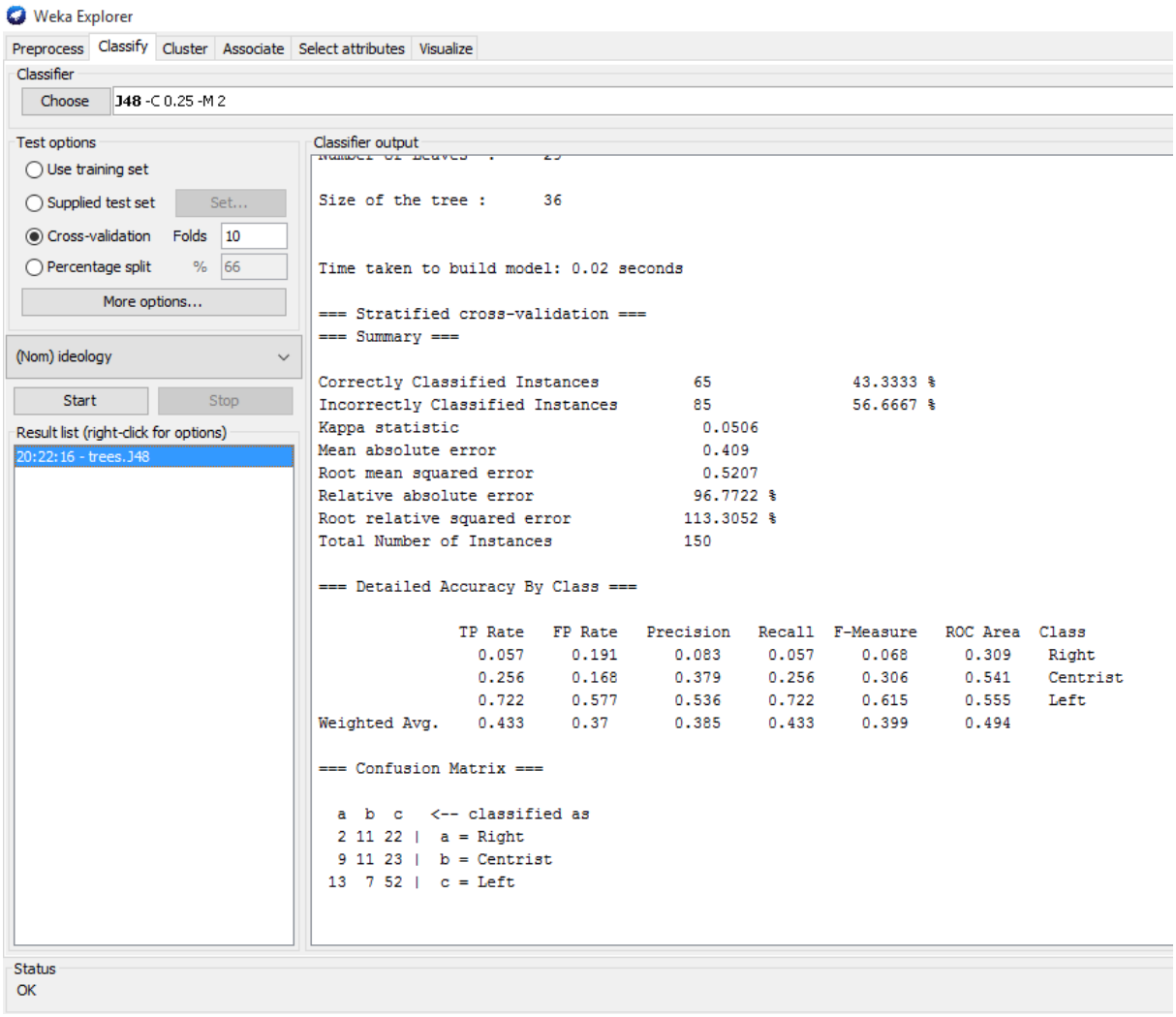


Εικόνα 24: Επιλογή των καλύτερων χαρακτηριστικών με τρεις κλάσεις

4.4.2.1C4.5 τριών κλάσεων και επιλογή χαρακτηριστικών

Από την εκτέλεση του αλγορίθμου C4.5 παρατηρήθηκε ότι τα 85 από τα 150 στοιχεία ταξινομήθηκαν εσφαλμένα και μόνο τα 65 στοιχεία ταξινομήθηκαν σωστά. Ενώ το ποσοστό σχετικού σφάλματος εντοπίστηκε στο 96,7%. Η κλάση με την υψηλότερη τιμή των σωστά αναγνωρισμένων θετικών περιπτώσεων και με το υψηλότερο ποσοστό των αρνητικών περιπτώσεων που ταξινομήθηκαν εσφαλμένα ως θετικά την είχε η κλάση left. Ενώ τέλος, μεγαλύτερη ακρίβεια, ανάκληση και απόδοση διαπιστώθηκε στην κλάση left. Αναφορικά με τον πίνακα σύγχυσης δεν εντοπίστηκαν

βέλτιστες τιμές, παρά μόνο τα ορθά στοιχεία της ταξινόμησης που εντοπίζονται στην κύρια διαγώνιο του πίνακα.

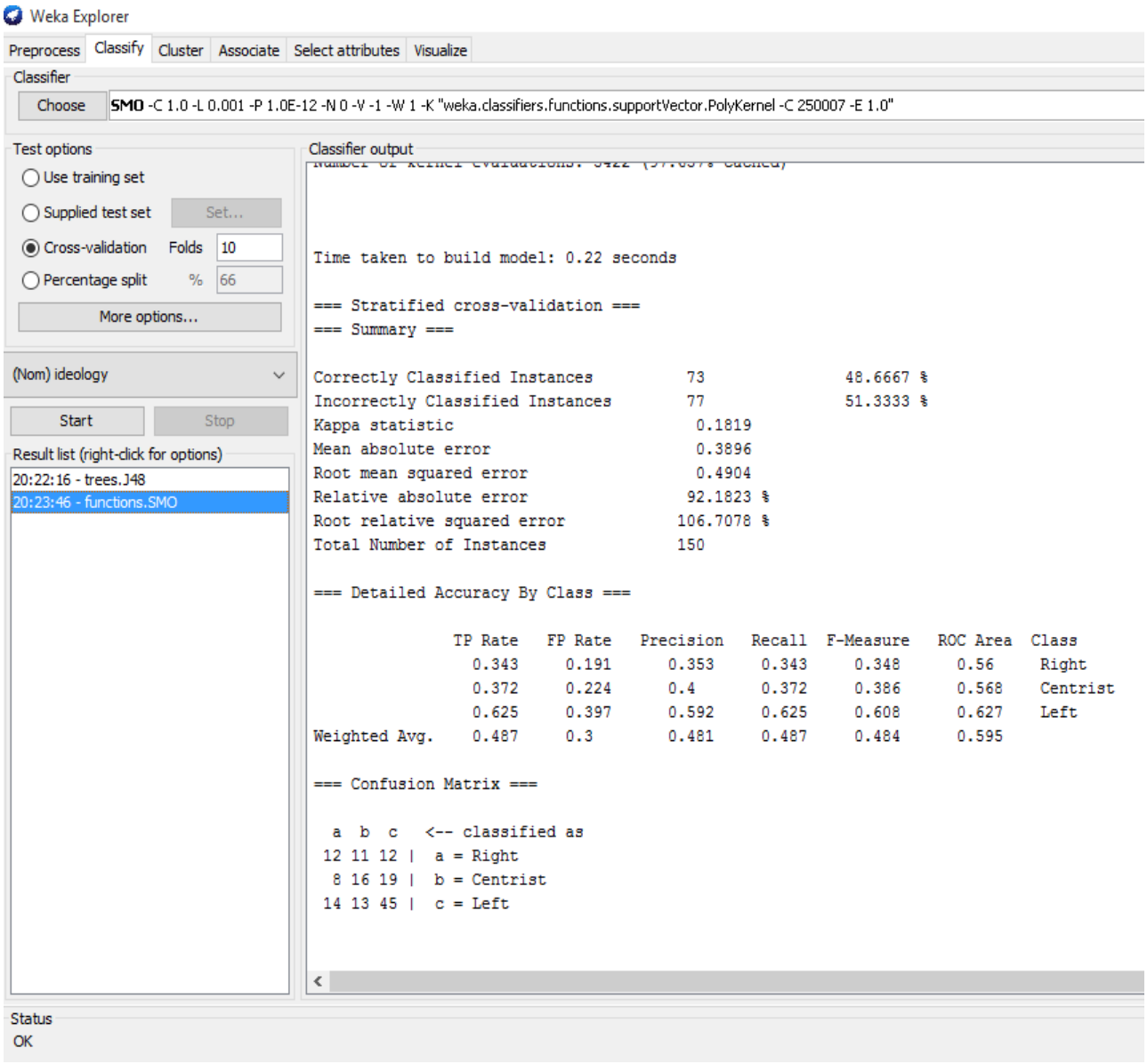


Εικόνα 25: Αποτελέσματα του αλγορίθμου C4.5 με τρεις κλάσεις και επιλογή των καλύτερων χαρακτηριστικών

4.4.2.2 SVM τριών κλάσεων και επιλογή χαρακτηριστικών

Από την εκτέλεση του αλγορίθμου SVM παρατηρήθηκε ότι τα 77 από τα 150 στοιχεία ταξινομήθηκαν εσφαλμένα, ενώ τα 73 στοιχεία ταξινομήθηκαν σωστά. Το το ποσοστό σχετικού σφάλματος εντοπίστηκε στο 92%. Η κλάση με την υψηλότερη τιμή των σωστά αναγνωρισμένων θετικών περιπτώσεων και με το υψηλότερο ποσοστό

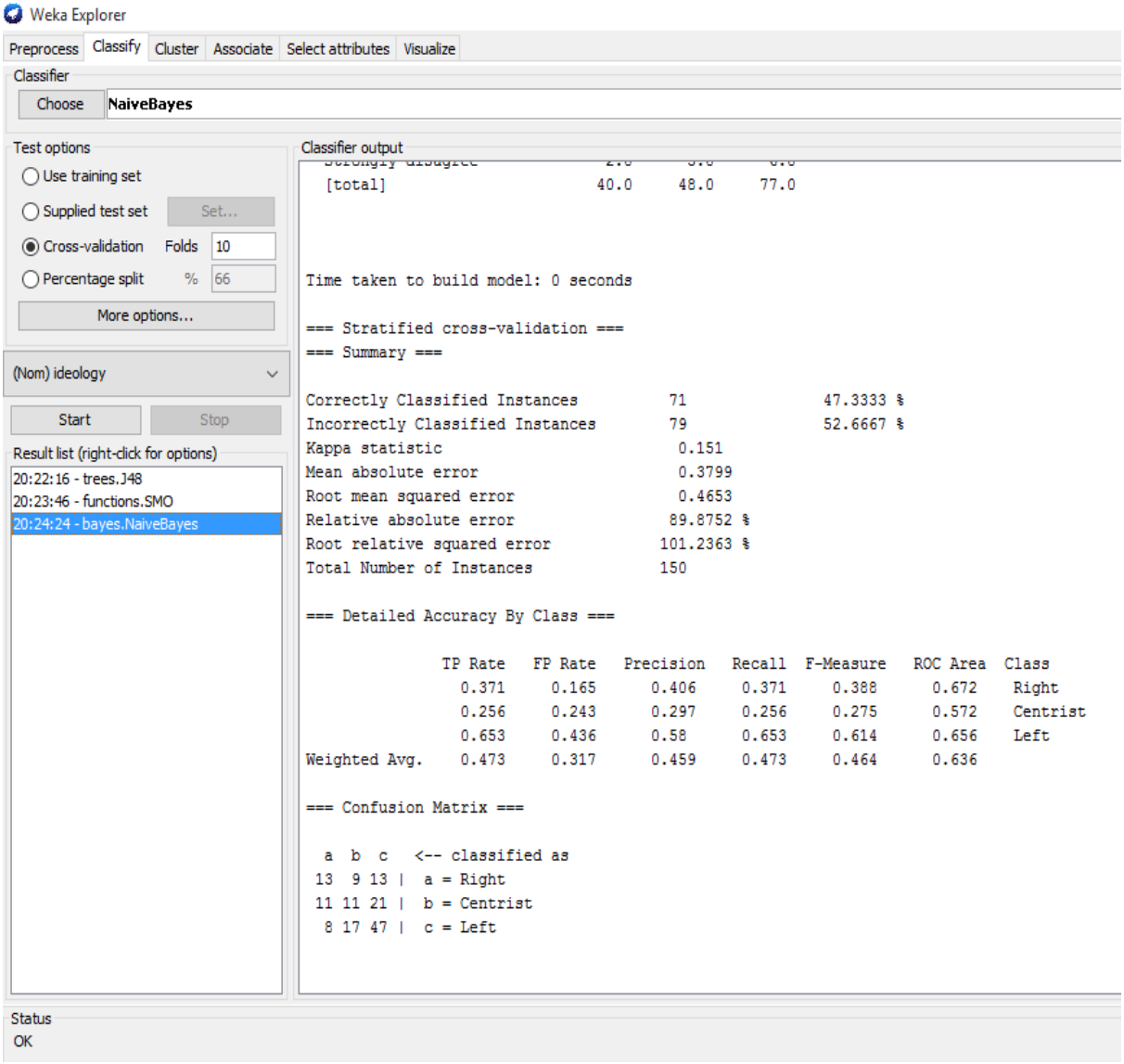
των αρνητικών περιπτώσεων που ταξινομήθηκαν εσφαλμένα ως θετικά την είχε η κλάση left. Ενώ τέλος, μεγαλύτερη ακρίβεια, ανάκληση και απόδοση διαπιστώθηκε στην κλάση left. Αναφορικά με τον πίνακα σύγχυσης δεν εντοπίστηκαν βέλτιστες τιμές, παρά μόνο τα ορθά στοιχεία της ταξινόμησης που εντοπίζονται στην κύρια διαγώνιο του πίνακα.



Εικόνα 26: Αποτελέσματα του αλγορίθμου SVM με τρεις κλάσεις και επιλογή των καλύτερων χαρακτηριστικών

4.4.2.3NaïveBayes τριών κλάσεων και επιλογή χαρακτηριστικών

Από την εκτέλεση του αλγορίθμου NaïveBayes παρατηρήθηκε ότι τα 79 από τα 150 στοιχεία ταξινομήθηκαν εσφαλμένα και 71 στοιχεία ταξινομήθηκαν σωστά, ενώ το ποσοστό σχετικού σφάλματος εντοπίστηκε στο 89%. Η κλάση με την υψηλότερη τιμή των σωστά αναγνωρισμένων θετικών περιπτώσεων και με το υψηλότερο ποσοστό των αρνητικών περιπτώσεων που ταξινομήθηκαν εσφαλμένα ως θετικά την είχε η κλάση left. Ενώ τέλος, μεγαλύτερη ακρίβεια, ανάκληση και απόδοση διαπιστώθηκε στην κλάση left. Αναφορικά με τον πίνακα σύγχυσης δεν εντοπίστηκαν βέλτιστες τιμές, παρά μόνο τα ορθά στοιχεία της ταξινόμησης που εντοπίζονται στην κύρια διαγώνιο του πίνακα.



Εικόνα 25: Αποτελέσματα του αλγορίθμου NaïveBayes με τρεις κλάσεις και επιλογή των καλύτερων χαρακτηριστικών

Κεφάλαιο 5

Συμπεράσματα

Στα πλαίσια αυτής της εργασίας διερευνήθηκε η δυνατότητα πρόβλεψης της ιδεολογίας ενός πολίτη, δεδομένων των θέσεων του σε συγκεκριμένα ζητήματα της επικαιρότητας. Συγκεκριμένα ο πολίτης ρωτήθηκε σε δεκαέξι ερωτήσεις που αφορούσαν κοινωνικά και πολιτικά θέματα. Μέσα από το ερωτηματολόγιο αυτό, διαπιστώθηκε ότι το 27% των ερωτηθέντων είναι κεντροαριστεροί όπου το 33,8% συμφωνεί στην πιθανότητα δημοψηφίσματος. Το 37,7% συμφωνεί στην παραμονή της Ελλάδας στην Ευρωπαϊκή Ένωση, ενώ πάνω από το 50% συμφώνησε στην διεκδίκηση των γερμανικών αποζημιώσεων. Επίσης, το 34,7% συμφωνεί ότι η αστυνομία πρέπει να επεμβαίνει εντός των Πανεπιστημίων ενώ διαφωνεί με τη δημιουργία ιδιωτικών πανεπιστημίων. Ουδέτερη στάση (24%) εκφράζουν οι πολίτες στη διατήρηση κέντρων διαμονής για παράνομους μετανάστες, ενώ το 46% του συνόλου συμφωνεί ότι ο δημόσιος τομέας είναι μεγαλύτερος απ' ότι θα έπρεπε.

Κατά την εκτέλεση των αλγορίθμων, παρατηρήθηκε η ακρίβειά τους με το πρόβλημα των πέντε και των τριών κλάσεων αντίστοιχα, όπου η υψηλή ακρίβεια πρόβλεψης σήμαινε ότι η ιδεολογία του κάθε πολίτη είναι εκείνη που καθορίζει την άποψή του. Αντίθετα η χαμηλή ακρίβεια πρόβλεψης, σήμαινε ότι ο κάθε πολίτης πλέον κρίνει το κάθε ζήτημα ξεχωριστά, με κριτήρια τα οποία είναι πέραν της ιδεολογίας του. Από την εκτέλεση των αλγορίθμων με πέντε κλάσεις και χρήση όλων των χαρακτηριστικών ο αλγόριθμος με την υψηλότερη ακρίβεια κατά μέσο όρο είναι ο αλγόριθμος C4.5, ενώ με την επιλογή των καλύτερων χαρακτηριστικών είναι ο SVM αλγόριθμος. Στην περίπτωση με ορίσματα τριών κλάσεων και χρήση όλων των χαρακτηριστικών, ο αλγόριθμος με την υψηλότερη ακρίβεια κατά μέσο όρο είναι ο NaïveBayes και με χαμηλότερη ακρίβεια ο C4.5 αλγόριθμος. Επίσης, κατά την επιλογή των καλύτερων χαρακτηριστικών με τρεις κλάσεις η υψηλότερη ακρίβεια κατά μέσο όρο εντοπίστηκε στον SVM αλγόριθμο, ενώ με χαμηλότερη ακρίβεια στον C4.5 αλγόριθμο.

Στη διάρκεια εκτέλεσης των τριών αλγορίθμων, παρατηρήθηκε ότι κανένας από τους αλγορίθμους δεν κατάφερε να ταξινομήσει σωστά όλα τα στοιχεία. Πιο αναλυτικά,

στη διάρκεια εκτέλεσης με ορίσματα πέντε κλάσεις και με χρήση όλων των χαρακτηριστικών ο αλγόριθμος C4.5 ταξινόμησε σωστά τα περισσότερα στοιχεία (30,6%), ενώ με επιλογή των καλύτερων χαρακτηριστικών ο αλγόριθμος Naïve Bayes ταξινόμησε σωστά το 30,6% των στοιχείων του. Αντίθετα, με όρισμα τρεις κλάσεις και χρήση όλων των χαρακτηριστικών ο Naïve Bayes ταξινόμησε σωστά το 48% των στοιχείων. Ενώ, με επιλογή των καλύτερων χαρακτηριστικών ο αλγόριθμος SVMταξινόμησε σωστά σχεδόν το 50% των στοιχείων του. Όπως απεικονίζεται και στον ακόλουθο πίνακα, οι καλύτεροι αλγόριθμοι των σωστά ταξινομημένων στοιχείων με όρισμα τριών κλάσεων είναι ο Naïve Bayes και ο SVM αλγόριθμος, ενώ οι καλύτεροι αλγόριθμοι των σωστά ταξινομημένων στοιχείων με όρισμα πέντε κλάσεων είναι ο Naïve Bayes και ο C4.5 αλγόριθμος.

	3 class			**5 class**		
	C4.5	**Naïve bayes**	**SVM**	**C4.5**	**Naïve bays**	**SVM**
Όλα τα χαρακτηριστικά	39,3%	48%	43,3%	30,6%	30%	21,3%
Με επιλογή χαρακτηριστικών	43,3%	47,3%	48,6%	26,6%	30,6%	30%

Εικόνα 26: Αποτελέσματα αλγορίθμων των σωστά ταξινομημένων στοιχείων

Τα αποτελέσματα που προκύπτουν από την εκτέλεση των αλγορίθμων δείχνουν ότι το ποσοστό πρόβλεψης σε σχέση με την τυχαία πρόβλεψη, έχει μία μικρή αύξηση η οποία δείχνει την συνεισφορά της ιδεολογίας του πολίτη. Βέβαια, επειδή η ακρίβεια που επιτεύχθηκε μέσα από την εκτέλεση των αλγορίθμων είναι πολύ χαμηλή, δεν απεικονίζεται απόλυτα η ιδεολογία των πολιτών. Σημαντικό στοιχείο της μελέτης αποτελεί και ο εντοπισμός του καλύτερου χαρακτηριστικού, όπου στη συγκεκριμένη περίπτωση είναι η ηλικία, με το μεγαλύτερο ποσοστό των ερωτηθέντων να είναι ηλικίας 18 έως 30. Μέσα από την ηλικία διακρίνεται και η ιδεολογία του κάθε πολίτη σε διάφορα ζητήματα. Οι νέοι άνθρωποι αντιμετωπίζουν τα πράγματα ελαφρότερα και κρίνουν τα ζητήματα μέσα από το ρομαντισμό. Αντίθετα, οι μεγαλύτερες ηλικίες

έχουν μεγαλύτερη πείρα αλλά και στερεότυπα που δύσκολα καταρρίπτονται. Επειδή το μεγαλύτερο ποσοστό των ερωτηθέντων ήταν νεότερης ηλικίας, οι απαντήσεις δεν ήταν απόλυτα ισορροπημένες.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Βικιπαίδεια (2015). *Επιλογή χαρακτηριστικών*. Διαθέσιμο στο διαδικτυακό τόπο: https://translate.google.gr/translate?hl=el&sl=en&u=https://en.wikipedia.org/wiki/Feature_selection&prev=search

Βικιπαίδεια (2015). *GoogleDrive*. Διαθέσιμο στο διαδικτυακό τόπο: https://el.wikipedia.org/wiki/Google_Drive .

Βικιπαίδεια, (2015). *Μάθηση*. Διαθέσιμο στο διαδικτυακό τόπο: https://el.wikipedia.org/wiki/%CE%9C%CE%AC%CE%B8%CE%B7%CF%83%CE%B7

Κεχαγιά, Ε. (2006). *Εφαρμογή σε Αλγορίθμους Συσταδοποίησης*, [Διπλωματική Εργασία]. Εθνικό Μετσόβιο Πολυτεχνείο, Σχολή Αγρονόμων και Τοπογράφων Μηχανικών. Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών «Γεωπληροφορική». Διαθέσιμο στο διαδικτυακό τόπο:

Λαδάς, Γ. (2014). *Βραχυπρόθεσμη πρόβλεψη ενεργειακής ζήτησης. Προσεγγίσεις βασισμένες στη Μηχανική Μάθηση*, [Διπλωματική Εργασία]. Διαθέσιμο στο διαδικτυακό τόπο: https://dspace.lib.ntua.gr/dspace2/bitstream/handle/123456789/40271/ladasp_forecasting.pdf?sequence=1

Μακρή, Α. (2013). *Σύγκριση Μεθόδων Πρόβλεψης Ψήφου στις ηλεκτρονικές πλατφόρμες συμβούλων ψήφων*, [Πτυχιακή Εργασία]. Τεχνολογικό Πανεπιστήμιο Κύπρου ,Τμήμα Επικοινωνίας και Σπουδών Διαδικτύου, Λεμεσός, 2013. Διαθέσιμο στο διαδικτυακό τόπο: http://ktisis.cut.ac.cy/bitstream/10488/3153/1/%CE%91%CF%81%CE%AF%CF%83%CF%84%CE%B7%20%CE%9C%CE%B1%CE%BA%CF%81%CE%AE.pdf

Ντάλλα, Μ. (2009). *Εφαρμογή Αλγορίθμων Επαγωγικού Λογικού Προγραμματισμού στη Σχεσιακή Εξόρυξη Δεδομένων* [Μεταπτυχιακή Εργασία]. Πανεπιστήμιο Πάτρας, Σχολή Θετικών Επιστημών, Τμήμα Μαθηματικών. Πάτρα, 2009. Διαθέσιμο στο διαδικτυακό τόπο: http://nemertes.lis.upatras.gr/jspui/bitstream/10889/2656/1/Nimertis_Dalla.pdf

Παπαδόπουλου, Δ. (2011). *Τεχνικές Εξόρυξης Δεδομένων σε Συστήματα Υποστήριξης Αποφάσεων Ακαδημαϊκών Επαγγελματικών Αποφάσεων*, [Διπλωματική Εργασία]. Πανεπιστήμιο Αιγαίου, Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων. Διαθέσιμο στο διαδικτυακό τόπο: http://www.icsd.aegean.gr/website_files/diplomatikes/undergraduate/415849017.pdf

Παρασύρη, Ε. (2014). *Εξόρυξη Γνώσης και Δεδομένων. Πλεονεκτήματα και Μειονεκτήματα σε μία επιχείρηση*, [Πτυχιακή Εργασία]. Τεχνολογικό Εκπαιδευτικό Ίδρυμα Κρήτης, Σχολή Διοίκησης και Οικονομίας, Τμήμα Λογιστικής. Ηράκλειο, 2014. Διαθέσιμο στο διαδικτυακό τόπο: http://nefeli.lib.teicrete.gr/browse/sdo/acfi/2014/ParasiriEleni/attached-document-1414665190-445690-12284/ParasiriEleni2014.pdf

Ταράτσα, Ν. (2011). *Εξόρυξη γνώσης σε κοινωνικά δίκτυα.* [Μεταπτυχιακή Διατριβή]. Πανεπιστήμιο Πειραιώς, Τμήμα Πληροφορικής. Διαθέσιμο στο διαδικτυακό τόπο:

Τζετζούμης, Ε. (2012). *Σύγκριση μεθόδων δημιουργίας έμπειρων συστημάτων με κανόνες για προβλήματα κατηγοριοποίησης από σύνολα δεδομένων*, [Διπλωματική Εργασία]. Πανεπιστήμιο Πατρών, Διατμηματικό ΠΜΣ «Μαθηματικά των υπολογιστών & των αποφάσεων», Τμήματα Μαθηματικών- Μηχανικών Η/Υ & Πληροφορικής. Πάτρα, 2012. Διαθέσιμο στο διαδικτυακό τόπο: http://nemertes.lis.upatras.gr/jspui/bitstream/10889/5777/1/%CE%94%CE%99%CE%A0%CE%9B%CE%A9%CE%9C%CE%91%CE%A4%CE%99%CE%9A%CE%97%20%CE%95%CE%A1%CE%93%CE%91%CE%A3%CE%99%CE%91_%CE%92%CE%91%CE%93%CE%93%CE%95%CE%9B%CE%97%CE%A3%20%CE%A4%CE%96%CE%95%CE%A4%CE%96%CE%9F%CE%A5%CE%9C%CE%97%CE%A3.pdf

Τσελεντής, Χ. (2008). *Εισαγωγή στη θεωρία της πληροφορίας και εντροπία.* Διαθέσιμο στο διαδικτυακό τόπο: http://christselentis.blogspot.gr/2008/08/blog-post_13.html