



ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Πρόγραμμα Μεταπτυχιακών Σπουδών (ΠΜΣ) στις
«Ψηφιακές Εφαρμογές και την Καινοτομία»

CAPSTONE PROJECT

*«Μελέτη διεθνούς τάσης στην εξόρυξη γνώσης από εκπαιδευτικά δεδομένα.
Συλλογή εκπαιδευτικών δεδομένων (ερωτηματολόγια/καταγραφή μαθησιακής
συμπεριφοράς/καταγραφή μαθησιακής απόδοσης). Εξαγωγή γνώσης με χρήση
εργαλείων μηχανικής μάθησης»*

ΠΑΝΤΑΖΗ ΣΤΑΜΑΤΙΑ
ΠΚΜ201910

Τριμελή Επιτροπή :

Επιβλέπων καθηγήτρια : Κάτια Λήδα Κερμανίδου

Μέλος : Μαραγκουδάκης Εμμανουήλ

Μέλος : Έξαρχος Θεμιστοκλής

ΚΕΡΚΥΡΑ
ΦΕΒΡΟΥΑΡΙΟΣ 2021

Πίνακας περιεχομένων

Περίληψη.....	3
Abstract.....	4
Κεφάλαιο 1.....	5
1.1 Εισαγωγή στην εξόρυξη δεδομένων.....	5
1.2 Ανακάλυψη Γνώσης από Βάσεις Δεδομένων.....	5
1.3 Στάδια ανακάλυψης γνώσης.....	6
1.4 Βασικές Έννοιες Μηχανικής Μάθησης.....	6
Κεφάλαιο 2.....	7
2.1 Περιγραφή WEKA.....	8
2.2 Πλατφόρμα εφαρμογής WEKA.....	8
2.3 Προεπεξεργασία δεδομένων.....	9
2.4 Επιλογή Χαρακτηριστικών.....	10
2.5 Κατηγοριοποίηση Δεδομένων.....	11
2.6 Μέθοδοι εξόρυξης δεδομένων.....	11
2.6.1 Δένδρα Απόφασης (Decision tree classifiers).....	12
2.6.2 Bayesian (Bayesian Classifiers).....	12
2.6.3 Ταξινομητές του πλησιέστερου γείτονα (Nearest Neighbour Classifiers) ...	12
2.6.4 Τεχνητά Νευρωνικά δίκτυα (Artificial Neural Networks).....	13
2.7 Τύποι δεδομένων του weka.....	13
2.8 Το περιβάλλον Explorer.....	13
Κεφάλαιο 3.....	17
3.1 Εφαρμογή στο Weka.....	17
3.2 Σύνολο δεδομένων –Περιγραφή.....	20
3.3 Προεπεξεργασία του dataset.....	22
3.3.1 Προεπεξεργασία - Attribute selection.....	31
3.4 Εξόρυξη δεδομένων – γνώσης.....	33
3.4.1 Ο αλγόριθμος j48 - Δέντρα απόφασης.....	33
3.4.2 Αλγόριθμος j48 pruned.....	34
3.4.3 Naïve bayes.....	35
3.5 Μέθοδος K πλησιέστερων γειτόνων (K- NN).....	36
3.5.1 Αλγόριθμος IBK (k-means) K=1.....	36
3.5.2 Αλγόριθμος IBK (k-means) K=3.....	37
3.5.3 Αλγόριθμος IBK , (k-means) K=5.....	38
3.5.4 J48 unpruned tree (attribute selection).....	39
3.5.5 J48 pruned tree (attribute selection).....	40
3.5.6 Naïve bayes(attribute selection).....	41
3.5.7 Αλγόριθμος IBK , (k-means) K=1.....	42
3.5.8 Αλγόριθμος IBK , (k-means) K=3.....	42
3.5.9 Αλγόριθμος IBK , (k-means) K=5.....	43
3.6 Σύγκριση Αλγορίθμων.....	44
Κεφάλαιο 4.....	46
4.1 ΣΥΜΠΕΡΑΣΜΑΤΑ.....	46
4.2Μελλοντική Έρευνα.....	46
Βιβλιογραφία.....	48

Περίληψη

Οι τεχνικές εξόρυξης γνώσης έχουν αναπτυχθεί ραγδαία τα τελευταία χρόνια. Οι αλγόριθμοι εξόρυξης γνώσης βρίσκουν ένα αρκετά ευρύ φάσμα εφαρμογών, είναι ο τομέας της Ιατρικής, των παιδαγωγικών θεμάτων και της διάγνωσης μίας πάθησης – νόσου. Το πρώτο βήμα μιας διαδικασίας data mining είναι η προεπεξεργασία των δεδομένων και πιθανά η μείωση των διαστάσεων της βάσης δεδομένων. Αυτό μπορεί να γίνει με τον καθαρισμό των δεδομένων, την εφαρμογή στατιστικών μεθόδων, αλλά και άλλων τεχνικών που παρέχουν λογισμικά όπως το Weka.

Στην εργασία αυτή – capstone project θα γίνει μια θεωρητική μελέτη στην εξόρυξη δεδομένων, θα αναλυθούν οι βασικές έννοιες μηχανικής μάθησης καθώς και η περιγραφή του εργαλείου μηχανικής μάθησης Weka. Επίσης γίνεται και μια παρουσίαση της κατηγορίας των αλγορίθμων κατηγοριοποίησης(classification) διότι θα τους χρησιμοποιήσουμε στην φάση της υλοποίησης.

Στη συνέχεια θα επικεντρωθούμε στην εφαρμογή του Weka και θα διερευνήσουμε την μέθοδο κατηγοριοποίησης και συγκεκριμένα θα ασχοληθούμε με την επίδοση μαθητών στο μάθημα μαθηματικών, δηλαδή στην εξόρυξη γνώσης σε εκπαιδευτικά δεδομένα.

Συγκεκριμένα θα υλοποιήσουμε αλγόριθμους κατηγοριοποίησης όπως η χρήση των δέντρων απόφασης τον J48, τους αλγόριθμους στατιστικής κατηγοριοποίησης τον Naïve Bayes και την μέθοδο των Κ πλησιέστερων γειτόνων (K-NN). Τέλος θα παρουσιάσουμε τα αποτελέσματα αυτών των αλγορίθμων καθώς και συμπεράσματα και θα αξιολογήσουμε αν αποτελεί μία αξιόπιστη μέθοδο.

Abstract

Knowledge mining techniques have developed rapidly in recent years. Knowledge mining algorithms find a fairly wide range of applications, is the field of Medicine, pedagogical issues and the diagnosis of a condition – disease. The first step in a data mining process is to preprocess the data and possibly reduce the dimensions of the database. This can be done by cleaning up data, applying statistical methods, and other techniques that provide software such as Weka.

In this work – capstone project will be a theoretical study in data mining , the basic concepts of machine learning as well as the description of the machine learning tool Weka will be analyzed. There is also a presentation of the category of categorization algorithms because we will use them in the implementation phase. We will then focus on the implementation of Weka and explore the categorization method and in particular we will deal with the performance of students in the mathematics course, i.e. the extraction of knowledge into educational data.

Specifically, we will implement categorization algorithms such as the use of decision trees in J48, the statistical categorization algorithms of Naïve Bayes and the K Nearest Neighbors (K-NN) method. Finally, we will present the results of these algorithms as well as conclusions and assess whether it is a reliable method.

ΚΕΦΑΛΑΙΟ 1

1.1 ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

Η εξέλιξη της τεχνολογίας έδωσε τη δυνατότητα εξάπλωσης του Internet. Με το πέρασμα του χρόνου, η πρόσβαση στο Internet έγινε προσιτή σε ολόένα και περισσότερους ανθρώπους. Αυτό οδήγησε στο να αναπτυχθούν περισσότεροι ιστότοποι και να χρησιμοποιηθούν βάσεις δεδομένων για την αποθήκευση των δεδομένων. Με τη δημιουργία εμπορικών και κοινωνικών ιστοσελίδων υπήρξαν τα πρώτα άλματα στις απαιτήσεις και ανάγκες για αποθήκευση και διαχείριση μεγάλου όγκου δεδομένων. Σήμερα, το πλήθος των διαθέσιμων δεδομένων είναι τεράστιο και αυξάνεται εκθετικά κάθε μέρα. Η μείωση στο κόστος συλλογής και της δυσκολίας στη συλλογή και αποθήκευση των δεδομένων συνετέλεσε σημαντικά στην ανάπτυξη του πεδίου αυτού. Ο τεράστιος όγκος δεδομένων, που συσσωρεύεται στις βάσεις δεδομένων και στις αποθήκες δεδομένων (data warehouses), δεν μπορεί να αξιοποιηθεί όπως είναι. Πρέπει αρχικά να γίνουν κάποιες ενέργειες για να δομηθούν κατάλληλα τα δεδομένα, ώστε στη συνέχεια να μπορούν να αξιοποιηθούν. [1]

Η Επιστήμη των Δεδομένων (DataScience), είναι ένας καινούριος όρος, ο οποίος ήρθε να αντικαταστήσει προγενέστερους όρους, όπως Ανακάλυψη Γνώσης από Βάσεις Δεδομένων (Knowledge Discovery in Data-base) ή Εξόρυξη Δεδομένων (Data Mining). Οι τρεις αυτοί όροι χρησιμοποιούνται με σκοπό να αναλύσει έναν μεγάλο όγκο δεδομένων που αφορούν ένα συγκεκριμένο πρόβλημα, συνήθως εμπορικού ή επιστημονικού ενδιαφέροντος.[1]...

1.2 Ανακάλυψη Γνώσης από Βάσεις Δεδομένων

Ανακάλυψη Γνώσης από Βάσεις Δεδομένων (ΑΓΒΔ) αποτελείται από συγκεκριμένα στάδια. Πρόκειται για την αποκάλυψη ή παραγωγή λειτουργικής γνώσης μέσα από την ανάλυση των δεδομένων. Αναφέρεται σε ολόκληρη τη διαδικασία, από τη συλλογή δεδομένων μέχρι την αξιοποίηση των αποτελεσμάτων σε πιο πρακτικό επίπεδο. Τα βασικά στάδια της είναι:

1. Συλλογή Δεδομένων (Data Collection)
2. Προεπεξεργασία Δεδομένων (Preprocessing)
3. Μετασχηματισμός Δεδομένων (Transformation)
4. Εξόρυξη Δεδομένων (Data Mining)
5. Διερμηνεία και Αξιολόγηση (Interpretation/Evaluation)

Ο όρος Ανακάλυψη Γνώσης από Βάσεις Δεδομένων συχνά ταυτίζεται με τον όρο Εξόρυξη Δεδομένων, που στην πραγματικότητα αποτελεί ένα μόνο επιμέρους βήμα της. Ο βασικός στόχος της Εξόρυξης Δεδομένων (ΕΔ) είναι η εξαγωγή μη τετριμμένης, προηγούμενα άγνωστης και πιθανά χρήσιμης πληροφορίας ή προτύπων από το σύνολο των δεδομένων. Ο όρος χρησιμοποιείται μάλλον καταχρηστικά αφού

στην ουσία δεν γίνεται καμία εξαγωγή δεδομένων. Αντίθετα, χρησιμοποιούνται τα προεπεξεργασμένα και (ενδεχομένως) μετασχηματισμένα δεδομένα για την εξαγωγή χρήσιμης πληροφορίας, η οποία αξιοποιείται για την απόκτηση γνώσης σε σχέση με κάποιο πρόβλημα. [1]

1.3 Στάδια ανακάλυψης γνώσης

1. Συλλογή Δεδομένων

Το πρώτο βήμα της ανακάλυψης γνώσης από βάσεις δεδομένων είναι η συλλογή και η αποθήκευση των δεδομένων.

2. Προεπεξεργασία Δεδομένων

Το δεύτερο και πιο σημαντικό στάδιο της ΑΓΒΔ είναι η προεπεξεργασία του συνόλου δεδομένων, η οποία γίνεται με στόχο τον καθαρισμό τους, δηλαδή την τακτοποίηση εσφαλμένων, προβληματικών ή ελλιπόντων δεδομένων.

3. Μετασχηματισμός Δεδομένων

Πρόκειται για τη μετατροπή των δεδομένων κάτω από ένα κοινό πλαίσιο, για επεξεργασία.

4. Εξόρυξη Δεδομένων

Σε αυτό το στάδιο της ΑΓΒΔ εφαρμόζεται κάποιος αλγόριθμος για την παραγωγή ενός μοντέλου. Έχοντας καθαρίσει και μετασχηματίσει τα δεδομένα, είναι έτοιμα να χρησιμοποιηθούν από κάποιον αλγόριθμο, ώστε να δημιουργηθεί κάποιο μοντέλο, συνήθως κατηγοριοποίησης ή πρόβλεψης.

5. Διερμηνεία και Αξιολόγηση

Στο τελευταίο στάδιο της ΑΓΒΔ γίνεται η διερμηνεία και η αξιολόγηση των αποτελεσμάτων.[1]

1.4 Βασικές Έννοιες Μηχανικής Μάθησης

Σε αυτό το κεφάλαιο θα ασχοληθούμε με τις βασικές έννοιες που συναντούμε στο data mining και την προεπεξεργασία δεδομένων και την κατηγοριοποίηση.

Μηχανική Μάθηση ονομάζεται η ικανότητα ενός υπολογιστικού συστήματος να δημιουργεί μοντέλα ή πρότυπα από ένα σύνολο δεδομένων.

Ο τομέας της Μηχανικής Μάθησης αναπτύσσει τρεις τρόπους μάθησης, ανάλογους με τους τρόπους με τους οποίους μαθαίνει ο άνθρωπος: επιβλεπόμενη μάθηση, μη επιβλεπόμενη μάθηση και ενισχυτική μάθηση. [4]

- Επιβλεπόμενη Μάθηση (Supervised Learning) είναι η διαδικασία όπου ο αλγόριθμος κατασκευάζει μια συνάρτηση που απεικονίζει δεδομένες εισόδους (σύνολο εκπαίδευσης) σε γνωστές επιθυμητές εξόδους, με στόχο τη γενίκευση της συνάρτησης αυτής και για εισόδους με άγνωστη έξοδο.
- Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning), όπου ο αλγόριθμος κατασκευάζει ένα μοντέλο για κάποιο σύνολο εισόδων υπό μορφή παρατηρήσεων χωρίς να γνωρίζει τις επιθυμητές εξόδους.[3]

Τα κύρια είδη μάθησης κατηγοριοποιούνται ως εξής:

- ✓ Ταξινόμηση (classification) : ταξινόμηση των υποδειγμάτων σε μια προκαθορισμένη τάξη
- ✓ Συσχέτιση (association): ανακάλυψη συσχετίσεων διαφόρων χαρακτηριστικών του συνόλου των δεδομένων
- ✓ Συσταδοποίηση ή Ομαδοποίηση δεδομένων (clustering): εύρεση ομάδων αντικείμενων με υψηλό βαθμό ομοιότητας και εκχώρηση υποδειγμάτων στις ομάδες αυτές πρόβλεψη μίας αριθμητικής ποσότητας.
- ✓ Αριθμητική πρόβλεψη διαδικασία είναι όμοια με την ταξινόμηση μόνο που η τάξη σε αυτή την περίπτωση είναι αριθμητική

ΚΕΦΑΛΑΙΟ 2

2.1 Περιγραφή WEKA

Το WEKA είναι ένα περιβάλλον ανάπτυξης εφαρμογών μηχανικής μάθησης και εξόρυξης γνώσης . Αναπτύχθηκε στο πανεπιστήμιο του Waikato της Ν. Ζηλανδίας και διατίθεται ως ελεύθερο λογισμικό. Το όνομα του Weka αποτελεί ακρωνύμιο της φράσης Waikato Environment for Knowledge Analysis, κυριολεκτικά, όμως, Weka είναι ένα πτηνό το οποίο ζει μόνο στη Νέα Ζηλανδία.

Το WEKA ανήκει στην κατηγορία του λεγόμενου "ελεύθερου λογισμικού" (freeware) και διατίθεται δημοσίως σύμφωνα με τους όρους της άδειας GNU General Public License, η οποία επιτρέπει στους χρήστες να χρησιμοποιούν, αλλά και να τροποποιούν ελεύθερα το λογισμικό. Επίσης ,αποτελεί ένα από τα δημοφιλέστερα εργαλεία μηχανικής μάθησης και χρησιμοποιείται στην έρευνα, στην εκπαίδευση και σε εφαρμογές.[2]

Τα χαρακτηριστικά και οι δυνατότητες που προσφέρει το weka

- είναι γραμμένο στη γλώσσα προγραμματισμού Java, ώστε να μπορεί να χρησιμοποιηθεί με όσο το δυνατόν περισσότερα λειτουργικά συστήματα
- Περιέχει ποικιλία μεθόδων για κατηγοριοποίηση, παλινδρόμηση, ανάλυση συστάδων, και κανόνες συσχέτισης. Επίσης, παρέχει δυνατότητες για προεπεξεργασία των δεδομένων, καθώς και εργαλεία οπτικοποίησης
- Είναι λογισμικό ανοικτού κώδικα. Αυτό σημαίνει ότι ο πηγαίος κώδικας είναι δημοσίως διαθέσιμος.

2.2 Πλατφόρμα εφαρμογής WEKA

Με την εκκίνηση του WEKA εμφανίζεται το παράθυρο της Εικόνας 1.1. Από το σημείο αυτό ο χρήστης μπορεί να εκκινήσει τις κύριες εφαρμογές του WEKA:

- Ο **Explorer** είναι η πιο δημοφιλής διεπαφή το οποίο και θα χρησιμοποιήσουμε στην ανάλυσή μας.
- Ο **Experimenter** είναι ένα περιβάλλον για διεξαγωγή πειραμάτων, όπου αξιολογούνται μέθοδοι κατηγοριοποίησης και παλινδρόμησης.
- **To Knowledge Flow** είναι ένα περιβάλλον που επιτρέπει τη διεξαγωγή των ιδίων εργασιών με τον Explorer, διαθέτει όμως διαφορετική διεπαφή (interface). Στο περιβάλλον αυτό χρησιμοποιούνται components, τα οποία συνδέονται μεταξύ τους με γραφικό τρόπο, ο οποίος ορίζει τη ροή εργασίας. [2]



Εικόνα 1- Εκκίνηση του weka

Ας δούμε την στιγμή που φορτώνουμε τα δεδομένα στο πρόγραμμα weka. Η μορφή που υποστηρίζει το weka είναι σε ARFF, όμως τα δεδομένα μπορούν να φορτώσουν και σε κάποια άλλη κατάληξη που είναι σε csv. Από την στιγμή που τα δεδομένα είναι έτοιμα για επεξεργασία, βλέπουμε την επιγραφή attributes(χαρακτηριστικά).

Πλαίσιο επιλεγμένα χαρακτηριστικά	Περιγραφή
Όνομα	Είναι το όνομα κάθε χαρακτηριστικού
Τύπος	Nominal(κατ'όνομα) Numeric(αριθμητικά)
Missing	Είναι ο αριθμός των περιπτώσεων που τα δεδομένα των χαρακτηριστικών είναι απροσδιόριστα
District(διακριτή)	Είναι ο αριθμός διαφορετικών τιμών που τα δεδομένα διαθέτουν για αυτό το χαρακτηριστικό.
Unique	Είναι ο αριθμός των περιπτώσεων των δεδομένων, έχοντας αξία για αυτό το χαρακτηριστικό, που κανένα άλλο δεν έχει.

Πίνακας 1 : Χαρακτηριστικά Weka

2.3 Προεπεξεργασία δεδομένων

Η προεπεξεργασία των δεδομένων αποτελεί ένα από τα βασικότερα σημεία στην εξόρυξη δεδομένων. Ένα πολύ συνηθισμένο πρόβλημα είναι η ύπαρξη χαμένων τιμών, η έλλειψη δηλαδή τιμών σε ορισμένα πεδία καταχωρημένων εγγραφών. Οι χαμένες τιμές μπορεί να οφείλονται σε πλήθος διαφορετικών λόγων. Ένα άλλο πρόβλημα των δεδομένων είναι ο λεγόμενος «θόρυβος». Τα δεδομένα μπορεί να περιέχουν λανθασμένες τιμές. Ο λόγος ύπαρξης λαθών είναι τα σφάλματα των χειριστών τη στιγμή της καταχώρησης των δεδομένων ή της χρήσης του λογισμικού, αλλά και τεχνικά προβλήματα, όπως η κακή λειτουργία συσκευών που καταγράφουν δεδομένα, καθώς και πιθανά προβλήματα στη μετάδοση των δεδομένων μέσω ενός δικτύου.

Για τον λόγο αυτό, απαιτείται η αντιμετώπιση των προβλημάτων σε χρόνο προηγούμενο από την καθαυτή ανάλυση. Η διαδικασία αντιμετώπισης των χαμένων τιμών, του θορύβου, των ασυνεπειών και άλλων προβλημάτων των δεδομένων ονομάζεται «καθαρισμός δεδομένων» (data cleansing) και αποτελεί μέρος των εργασιών της προεπεξεργασίας τους.

Το WEKA προσφέρει ποικιλία μεθόδων για προεπεξεργασία δεδομένων. Μερικές φορές απαιτείται ο μετασχηματισμός των δεδομένων σε μορφή αποδοτική για τους αλγόριθμους που θέλουμε να εφαρμόσουμε στην ανάλυσή μας. Ο μετασχηματισμός των δεδομένων είναι απαραίτητος για την ανάδειξη νέων ιδιοτήτων του συνόλου των δεδομένων μας. [5]

Ορισμένες από τις εργασίες που μπορούν να εκτελεστούν είναι:

- Η προσθήκη νέων υπολογιζόμενων πεδίων.
- Η κανονικοποίηση μετατροπή αριθμητικών τιμών σε πιο κατάλληλες αριθμητικές τιμές.
- Η διακριτοποίηση μετασχηματισμό αριθμητικών τιμών σε ονομαστικές τιμές.
- Η μετατροπή αριθμητικών και ονομαστικών πεδίων σε δυαδικά.
- Η συγχώνευση δύο ονομαστικών πεδίων.
- Η μείωση των διαστάσεων με Ανάλυση Κυρίων Συνιστωσών.
- Η δημιουργία νέων συνόλων δεδομένων με εφαρμογή δειγματοληψίας[2]

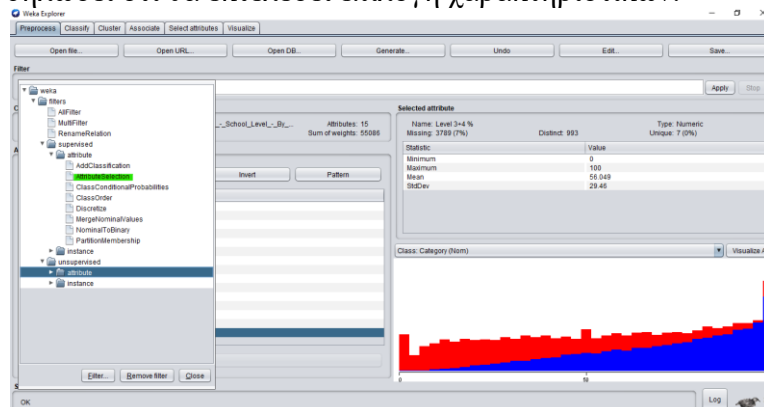
2.4 Επιλογή Χαρακτηριστικών

Μια σημαντική προεπεξεργασία δεδομένων είναι η επιλογή σημαντικών στηλών, στο "Select Attributes" του WEKA περιλαμβάνονται εργαλεία για την επιλογή σημαντικών χαρακτηριστικών.

Οι μέθοδοι επιλογής χαρακτηριστικών του WEKA αποτελούνται από δύο τμήματα:

- ✓ Μια μέθοδο αναζήτησης. Περιλαμβάνονται μέθοδοι πρόσθιας αναζήτησης, γενετικοί αλγόριθμοι κλπ.
- ✓ Μια μέθοδο αξιολόγησης. Διατίθεται μια σημαντική ποικιλία μεθόδων, η οποία περιλαμβάνει την CFS, στατιστική Chi-Square, ευαίσθητες στο κόστος μέθοδους, wrappers, κριτήριο Gain Ratio, χρήση κατηγοριοποιητή SVM, ανάλυση κυρίων συνιστωσών κλπ

Η επιλογή γίνεται από την αναδιπλούμενη λίστα που εμφανίζεται με το πάτημα του κουμπιού "Choose" στο πεδίο "Filter", ο χρήστης επιλέγει την εργασία weka/filters/supervised/attribute/AttributeSelection. Στην Εικόνα 1.2 ο χρήστης έχει δηλώσει ότι θα εκτελέσει επιλογή χαρακτηριστικών.



Εικόνα 2 – Attribute Selection

2.5 Κατηγοριοποίηση Δεδομένων

Η κατηγοριοποίηση (classification) εκπαιδεύει μία συνάρτηση, η οποία κατηγοριοποιεί κάποια δεδομένα σε μια από διάφορες κλάσεις που δημιουργούνται. Στην καρτέλα classify και συγκεκριμένα στην ενότητα Classifier μπορούμε να επιλέξουμε τη μέθοδο κατηγοριοποίησης. Στο σύνολο δεδομένων εκπαίδευσης και το σύνολο δοκιμής, οι επιλογές παρέχονται στην ενότητα test option και είναι οι εξής:

- Use training test: υπολογίζει τις επιδόσεις του μοντέλου, χρησιμοποιώντας το σύνολο εκπαίδευσης
- Supplied test set: χρησιμοποιεί για επικύρωση ένα διαφορετικό σύνολο δεδομένων
- Cross validation : είναι μια μέθοδο επικύρωσης 10 τμημάτων (10 fold cross validation) το σύνολο δεδομένων διαιρείται σε 10 υποσύνολα. Κάθε υποσύνολο περιέχει διαφορετικές παρατηρήσεις. Ένα από τα υποσύνολα χρησιμοποιείται ως σύνολο για εκπαίδευση και τα υπόλοιπα εννέα συνενώνονται και δημιουργούν το σύνολο εκπαίδευσης. Η διαδικασία επαναλαμβάνεται δέκα φορές, κάθε φορά χρησιμοποιώντας ένα διαφορετικό σύνολο ως σύνολο επικύρωσης και τα υπόλοιπα εννέα ως σύνολο εκπαίδευσης. Στο τέλος υπολογίζεται η μέση επίδοση του μοντέλου
- Percentage split: εφαρμόζει τη μέθοδο holdout και διασπά το σύνολο των παρατηρήσεων σε υποσύνολο εκπαίδευσης και υποσύνολο επικύρωσης, σύμφωνα με τα ποσοστά που ορίζει ο χρήστης.[3]

Για κάθε κατηγοριοποιητή παρουσιάζονται τα αποτελέσματα στο παράθυρο εξόδου (Classifier output), τα οποία περιλαμβάνουν τις ακόλουθες πληροφορίες:

- i) Πληροφορίες εκτέλεσης (Run information): Εμφανίζονται τα στοιχεία που αφορούν στον αλγόριθμο μάθησης και τις επιλογές του, στη μέθοδο ελέγχου του αλγορίθμου, καθώς και στο αρχείο εκπαίδευσης που χρησιμοποιήθηκε.
- ii) Μοντέλο κατηγοριοποιητή (Classifier model – full training set): Παρουσιάζει το μοντέλο κατηγοριοποιητή που δημιουργήθηκε από τα δεδομένα εκπαίδευσης.
- iii) Περίληψη (Summary): εμφανίζεται για μια λίστα στατιστικών στοιχείων για την ακρίβεια πρόβλεψης του αλγορίθμου, με βάση τη μέθοδο ελέγχου που επιλέχθηκε.
- iv) Η ακρίβεια του κατηγοριοποιητή για την πρόβλεψη της κάθε κατηγορίας.
- v) Πίνακας σύγχυσης (Confusion Matrix): αφορά έναν πίνακα που δείχνει πόσα παραδείγματα αποδόθηκαν σε κάθε κατηγορία.

2.6 Μέθοδοι εξόρυξης δεδομένων

Η κατηγοριοποίηση (classification) είναι πιο δημοφιλής τεχνική εξόρυξης γνώσης (data mining). Η τεχνική αυτή χρησιμοποιείται όταν θέλουμε να βρούμε ένα μοντέλο το οποίο να διαχωρίζει δεδομένα σε προκαθορισμένες κατηγορίες. Επομένως μπορούμε να πούμε ότι η κατηγοριοποίηση μας παρέχει ένα μοντέλο ώστε να

περιγράψουμε τα μελλοντικά μας δεδομένα. Οι μεθοδολογίες που έχουν χρησιμοποιηθεί για την κατηγοριοποίηση είναι :

- Δέντρα αποφάσεων (Decision Trees)
- Bayesian (Bayesian Classifiers)
- Ταξινομητές του πλησιέστερου γείτονα (Nearest Neighbour Classifiers)
- Τεχνητά Νευρωνικά δίκτυα (Artificial Neural Networks)

2.6.1 Δένδρα Απόφασης (Decision tree classifiers)

Τα Δένδρα Αποφάσεων είναι μια από τις βασικότερες και πιο δημοφιλείς μεθόδους κατηγοριοποίησης. Η κατασκευής τους είναι η διαδοχική διάσπαση του συνόλου των παρατηρήσεων σε υποσύνολα. Η διαδικασία των διαδοχικών διασπάσεων αναπαρίσταται με μια ανεστραμμένη δενδρική δομή. Στην κορυφή βρίσκεται ο κόμβος-ρίζα του δένδρου. Σε κατώτερα επίπεδα βρίσκονται επιπλέον κόμβοι, οι οποίοι συνδέονται με ακμές με άλλα στοιχεία του δένδρου. Στο κατώτερο επίπεδο κάθε κλάδου βρίσκονται τα φύλλα του δένδρου. Ο κόμβος - ρίζα έχει μόνο εξερχόμενες ακμές που τον συνδέουν με στοιχεία του κατώτερου επιπέδου. Τέλος, τα φύλλα έχουν μόνο εισερχόμενες ακμές, οι οποίες τα συνδέουν με τους κόμβους του ανώτερου επιπέδου.[6]

2.6.2 Bayesian (Bayesian Classifiers)

Τα Μπαϋεσιανά Δίκτυα (Bayesian Networks) είναι εργαλεία για αναπαράσταση σύνθετων σχέσεων μεταξύ μεταβλητών και για εξαγωγή συμπερασμάτων σε συνθήκες αβεβαιότητας. Ανήκουν στην κατηγορία των γραφικών πιθανοτικών μοντέλων, τα οποία αναπαριστούν σχέσεις με μορφή γράφων. Κάθε κόμβος του γράφου συμβολίζει μια στοχαστική μεταβλητή και κάθε βέλος συμβολίζει μια σχέση εξάρτησης ανάμεσα σε δύο μεταβλητές. Τα Μπαϋεσιανά Δίκτυα έχουν το θεωρητικό τους υπόβαθρο από τη στατιστική και πιο συγκεκριμένα από το θεώρημα του Bayes, που υπολογίζει την υπό συνθήκη πιθανότητα $P(H|X)$, δηλαδή την πιθανότητα να επαληθευτεί η υπόθεση H με δεδομένο ότι ισχύει το γεγονός X . [6]

2.6.3 Ταξινομητές του πλησιέστερου γείτονα (Nearest Neighbour Classifiers)

Ο αλγόριθμος ταξινόμησης με βάση τους k κοντινότερους γείτονες (k -Nearest Neighbor Algorithm – k -NN) είναι η πιο βασική μέθοδος μάθησης βασισμένη σε στιγμιότυπα. αρχικά ένα πρόβλημα κατηγοριοποίησης, όπου οι παρατηρήσεις αποτελούνται από δύο αριθμητικά πεδία και το γνώρισμα της κλάσης. Κάθε παρατήρηση μπορεί να θεωρηθεί ως ένα σημείο στον χώρο των δύο διαστάσεων. Μια παρατήρηση X απέχει από μια άλλη παρατήρηση Y , απόσταση $d(X,Y)$ μέσα στον διδιάστατο χώρο. Οι κατηγοριοποιητές k -NN διαθέτουν αξιολογικά πλεονεκτήματα: Είναι αποτελεσματικοί όταν υπάρχουν σύνθετες εξαρτήσεις μεταξύ των μεταβλητών.

Διαθέτουν απλό αλγόριθμο και σε πολλές περιπτώσεις επέτυχαν υψηλές επιδόσεις κατηγοριοποίησης.[7]

2.6.4 Τεχνητά Νευρωνικά δίκτυα (Artificial Neural Networks)

Τα Νευρωνικά Δίκτυα (Neural Networks) αποτελούν ένα από τα σημαντικότερα επιτεύγματα της Τεχνητής Νοημοσύνης. Εμπνευσμένα από το βιολογικό νευρικό σύστημα, και ειδικότερα από τον ανθρώπινο εγκέφαλο, διαθέτουν αξιοσημείωτα χαρακτηριστικά, όπως τη δυνατότητα τους να αναπαριστούν σύνθετες εξαρτήσεις ή την ικανότητα τους να προβλέπουν την κλάση άγνωστων παρατηρήσεων. Τα Νευρωνικά Δίκτυα είναι μια τεχνική ισχυρά καθοδηγούμενη από τα δεδομένα. Αυτό σημαίνει ότι δεν επιβάλλουν αυθαίρετες υποθέσεις και ότι τα μοντέλα τους πηγάζουν από την επεξεργασία των δεδομένων.[6]

2.7 Τύποι δεδομένων του weka

Οι τύποι δεδομένων που υποστηρίζει το πρόγραμμα Weka είναι:

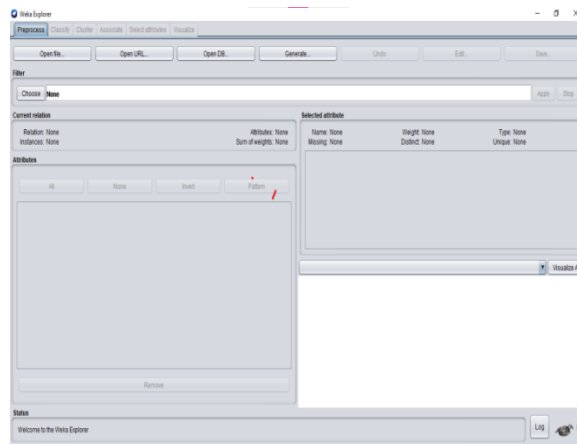
1. Αριθμητικός τύπος δεδομένων (numeric): Οι τιμές που παίρνει είναι αριθμητικές, ακέραιες ή πραγματικές.
2. Κατηγορικός τύπος δεδομένων (nominal): Πρόκειται για χαρακτηριστικά που παίρνουν κατηγορικές τιμές, οι οποίες παρατίθενται σαν λίστα στη δήλωση των ιδιοτήτων.
3. Αλφαριθμητικός τύπος δεδομένων (string): Πρόκειται για χαρακτηριστικά των οποίων οι τιμές μπορεί να είναι οποιοδήποτε τμήμα κειμένου.
4. Τύπος δεδομένων ημερομηνίας (date): Οι τιμές των χαρακτηριστικών αυτών είναι ημερομηνίες και με το προαιρετικό date-format καθορίζεται η μορφή με την οποία θα εμφανίζονται οι ημερομηνίες.

2.8 Το περιβάλλον Explorer

Ο Explorer είναι οργανωμένος σε έξι κατηγορίες λειτουργιών με τις αντίστοιχες ετικέτες:

❖ Preprocess

επιλέγουμε τα στοιχεία από μια βάση δεδομένων, ένα csv ή ένα arff αρχείο κ.λπ., και να προεπεξεργαστούμε αυτά τα δεδομένα χρησιμοποιώντας αλγόριθμους φιλτραρίσματος (Εικόνα 3).

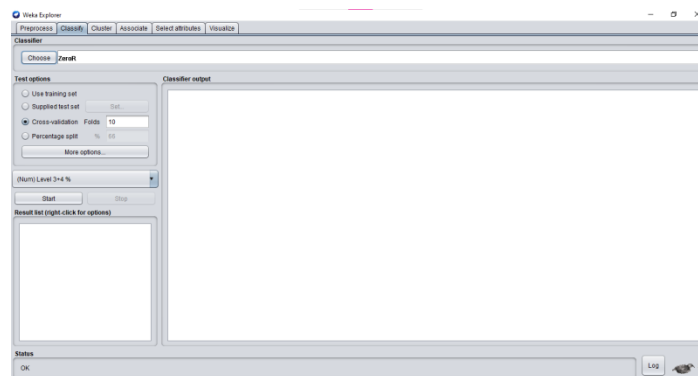


Εικόνα 3 – Preprocess

❖ Classify

Στο παράθυρο Classify μας δίνεται η δυνατότητα να εκπαιδεύσουμε έναν αλγόριθμο μηχανικής μάθησης (κατηγοριοποιητή) με βάση τα δεδομένα (Εικόνα 4). Επίσης μπορεί να εφαρμόσει τις τέσσερις παρακάτω επιλογές :

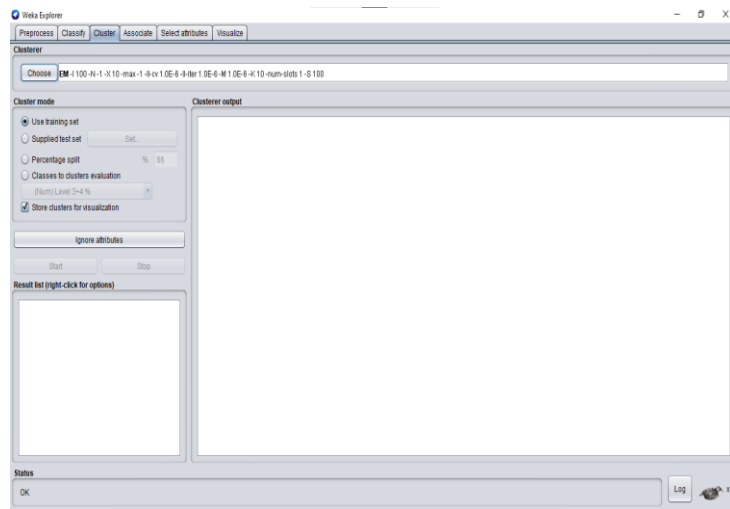
- I. **Use training set** : Εξετάζει τον ταξινομητή και εκτιμά την τάξη των περιπτώσεων στις οποίες εκπαιδεύτηκε.
- II. **Supplied test set** : Εξετάζει τον ταξινομητή και προβλέπει την τάξη ενός συνόλου που έχουν προταθεί από ένα αρχείο.
- III. **Cross-validation** : Χωρίζει το σύνολο εκπαίδευσης σε ανεξάρτητα υποσύνολα με αριθμό τεμαχισμού σε πακέτα των 10, αφήνοντας κάθε φορά ένα πακέτο ως σύνολο ελέγχου και τα υπόλοιπα ως σύνολα εκπαίδευσης.
- IV. **Percentage split** : Εξετάζει τον ταξινομητή σχετικά με το πόσο προβλέπει ένα ορισμένο ποσοστό των δεδομένων, τα οποία παρέχονται για δοκιμές.



Εικόνα 4 - Classify

❖ Cluster

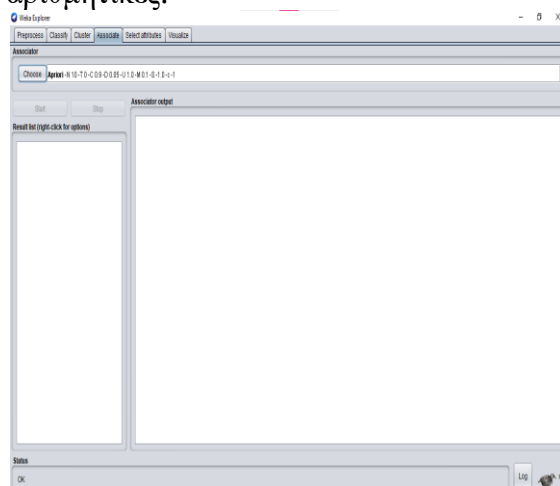
Αποτελεί κατηγορία όπου περιέχονται αλγόριθμοι για την εύρεση υποομάδων μέσα από το σύνολο δεδομένων.



Εικόνα 5 – Cluster

❖ Associate

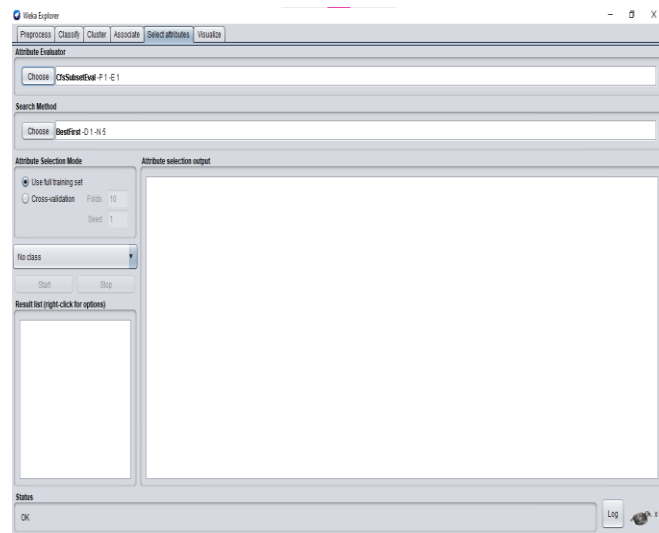
Στην κατηγορία αυτή περιέχονται αλγόριθμοι για την εύρεση κανόνων συσχέτισης στο σύνολο δεδομένων και οι μόνες τιμές που αναγνωρίζει είναι ονομαστικές και όχι αριθμητικές.



Εικόνα 6 – Associate

❖ Select Attributes

Αποτελεί κατηγορία στην οποία περιέχονται αλγόριθμοι για την επιλογή των πιο σχετικών χαρακτηριστικών μέσα από το σύνολο δεδομένων.



Εικόνα 7 - Select Attributes

Επισκόπηση Βιβλιογραφίας

Στο κεφάλαιο αυτό γίνονται αναφορές στην υπάρχουσα βιβλιογραφία από τις πηγές:
 [1] [2][3][4][5][6][7]

ΚΕΦΑΛΑΙΟ 3

3.1 Εφαρμογή στο Weka

Από το 2006 το Υπουργείο Παιδείας της Νέας Υόρκης επέκτεινε τα προγράμματα δοκιμών ELA και μαθηματικών σε βαθμούς 3-8. Οι κρατικές εξετάσεις χορηγήθηκαν στους βαθμούς 4 και 8 και οι δοκιμές σε όλη την πόλη χορηγήθηκαν στους βαθμούς 3, 5, 6, και 7.

Το 2006, NYSED αντιμετωπίζονται District 75 φοιτητές ως μια ξεχωριστή γεωγραφική περιοχή. Για την 2007-2011, οι μαθητές της Περιφέρειας 75 εκπροσωπούνται στις περιοχές και τους δήμους της πατρίδας τους. Τα λογιστικά φύλλα για την περιοχή και το δήμο δεν περιλαμβάνουν τους σπουδαστές περιοχής 75 το 2006. Ξεκινώντας το 2010, το NYSED άλλαξε τη βαθμολογία κλίμακας που απαιτείται για την κάλυψη κάθε ένα από τα επίπεδα επάρκειας, αυξάνοντας τον αριθμό των ερωτήσεων που οι μαθητές έπρεπε να απαντήσουν σωστά για να ανταποκριθούν.

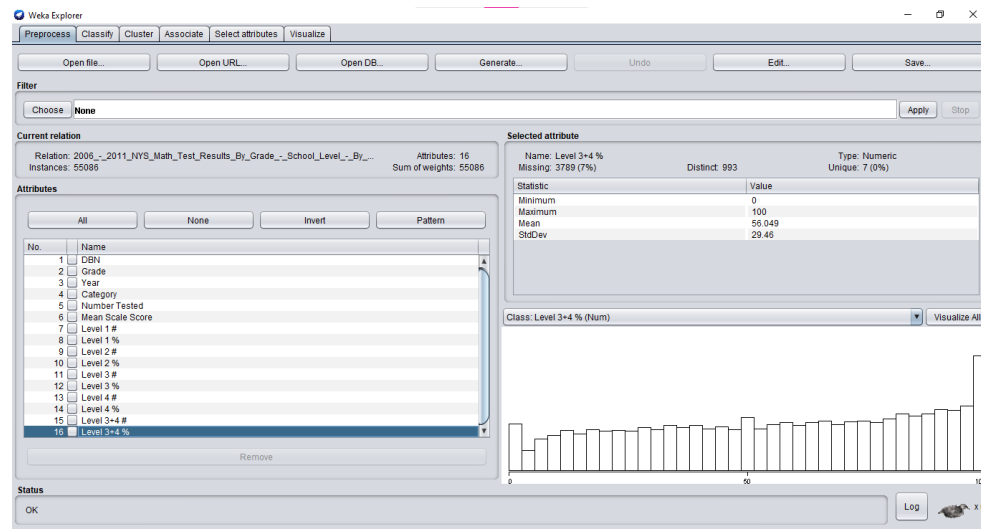
Στη συνέχεια παραθέτουμε την διαδικασία εισαγωγής του αρχείου που στο WEKA. Αφού φορτώσουμε το αρχείο “2006_-_2011_NYS_Math_Test_Results_By_Grade_-_School_Level_-_By_Disability_Status”, το οποίο αφορά τις δοκιμασίες μαθηματικών της πολιτείας της Νέας Υόρκης, βλέπουμε το αρχείο στο weka όπου τα δεδομένα εμφανίζονται με επιτυχία στο παρακάτω παράθυρο:



Εικόνα 8 - Αρχική οθόνη weka

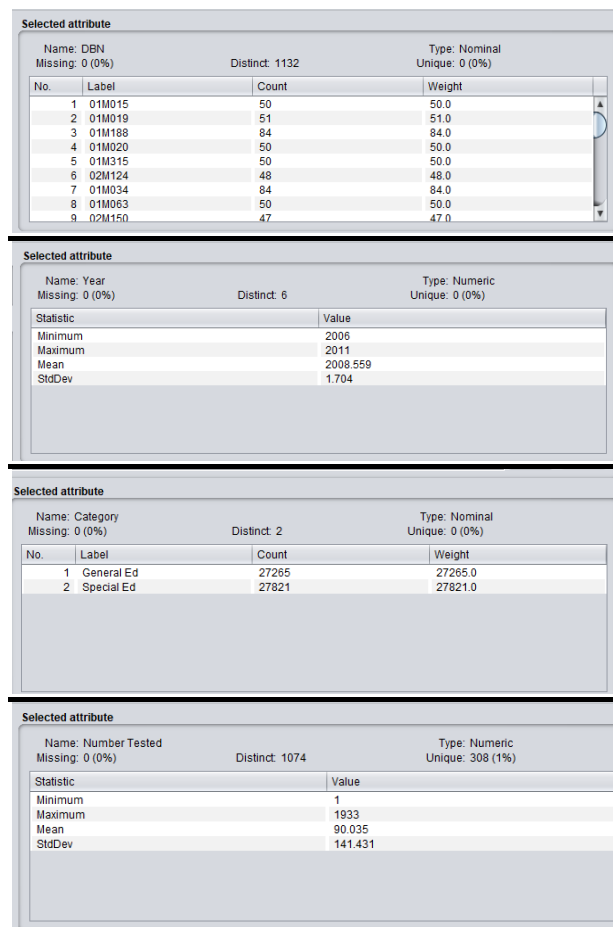
Στην εικόνα 9 παρατηρούμε ότι το δείγμα περιέχει τιμές που δεν έχουν καταγραφεί (missing values), κάτι που σημαίνει ότι τα δεδομένα είναι αξιόπιστα, και, συνεπώς, κατάλληλα για προβλέψεις.

Η παραπάνω εικόνα 8 αναπαριστά το panel που αντικρίζουμε αφού ολοκληρώσουμε την εισαγωγή του αρχείου στο WEKA. Αυτό σημαίνει πως η δομή του αρχείου που έχουμε εισάγει είναι η σωστή - σε διαφορετική περίπτωση το WEKA θα είχε εμφανίσει μήνυμα σφάλματος και δεν θα φόρτωνε το αρχείο.



Εικόνα 9- Άνοιγμα του συνόλου δεδομένων

Στη συνέχεια θα δούμε μια εικόνα του δείγματος ,στο οποίο δίνονται τα στατιστικά στοιχεία της μέσης τιμής (Mean), της τυπικής απόκλισης (StdDev), της ελάχιστης τιμής (Minimum) και της μέγιστης τιμής (Maximum) για κάθε αριθμητική μεταβλητή του συνόλου δεδομένων, όπως αυτά φαίνονται στο Weka.



Selected attribute		
Name: Mean Scale Score Missing: 14071 (26%)		Type: Numeric Unique: 12 (0%)
Distinct: 202		
Statistic	Value	
Minimum	539	
Maximum	772	
Mean	664.622	
StdDev	27.719	

Selected attribute		
Name: Level 1 # Missing: 3789 (7%)		Type: Numeric Unique: 52 (0%)
Distinct: 208		
Statistic	Value	
Minimum	0	
Maximum	428	
Mean	8.371	
StdDev	15.947	

Selected attribute		
Name: Level 1 % Missing: 3789 (7%)		Type: Numeric Unique: 79 (0%)
Distinct: 823		
Statistic	Value	
Minimum	0	
Maximum	100	
Mean	14.49	
StdDev	18.01	

Selected attribute		
Name: Level 2 # Missing: 3789 (7%)		Type: Numeric Unique: 95 (0%)
Distinct: 425		
Statistic	Value	
Minimum	0	
Maximum	726	
Mean	24.227	
StdDev	41.643	

Selected attribute		
Name: Level 2 % Missing: 3789 (7%)		Type: Numeric Unique: 26 (0%)
Distinct: 760		
Statistic	Value	
Minimum	0	
Maximum	100	
Mean	29.463	
StdDev	18.559	

Selected attribute		
Name: Level 3 # Missing: 3789 (7%)		Type: Numeric Unique: 153 (0%)
Distinct: 617		
Statistic	Value	
Minimum	0	
Maximum	1118	
Mean	43.207	
StdDev	70.798	

Selected attribute		
Name: Level 3 % Missing: 3789 (7%)		Type: Numeric Unique: 35 (0%)
Distinct: 895		
Statistic	Value	
Minimum	0	
Maximum	100	
Mean	40.994	
StdDev	20.347	

Selected attribute		
Name: Level 4 #	Distinct: 476	Type: Numeric
Missing: 3789 (7%)		Unique: 136 (0%)
Statistic	Value	
Minimum	0	
Maximum	994	
Mean	20.658	
StdDev	47.898	

Selected attribute		
Name: Level 4 %	Distinct: 933	Type: Numeric
Missing: 3789 (7%)		Unique: 66 (0%)
Statistic	Value	
Minimum	0	
Maximum	100	
Mean	15.055	
StdDev	17.881	

Selected attribute		
Name: Level 3+4 #	Distinct: 858	Type: Numeric
Missing: 3789 (7%)		Unique: 233 (0%)
Statistic	Value	
Minimum	0	
Maximum	1586	
Mean	63.865	
StdDev	110.174	

Selected attribute		
Name: Level 3+4 %	Distinct: 993	Type: Numeric
Missing: 3789 (7%)		Unique: 7 (0%)
Statistic	Value	
Minimum	0	
Maximum	100	
Mean	56.049	
StdDev	29.46	

Εικόνα 10 - Στατιστικά αριθμητικών μεταβλητών του συνόλου δεδομένων

3.2 Σύνολο δεδομένων -Περιγραφή

Το αρχείο των δεδομένων που θα αναλύσουμε καταγράφονται 16 χαρακτηριστικά που περιγράφουν την κάθε μια από τις 55.086 περιπτώσεις που θα μελετηθούν. Αρχικά, το αρχείο είναι σε μορφή .csv. Στη συνέχεια ακολουθεί η προεπεξεργασία των δεδομένων, η εφαρμογή διάφορων αλγορίθμων, η συγκέντρωση των αποτελεσμάτων τους και η αξιολόγησή τους. Τα δεδομένα που συγκεντρώθηκαν με στόχο την επεξεργασία και ανάλυσή τους αφορούν στην επίδοση των μαθητών στα μαθηματικά από το 2006 έως το 2011.

Το σύνολο δεδομένων που χρησιμοποιήθηκε είναι διαθέσιμο στην ιστοσελίδα <https://data.world/datasets/education>. Τα δεδομένα αντιπροσωπεύουν 55086 παραδείγματα. Από τα 16 χαρακτηριστικά του αρχικού σετ, επιλέχθηκαν τα 15 πιο σημαντικά γνωρίσματα για να μπορέσουμε να έχουμε καλύτερη εφαρμογή του μοντέλου που κατασκευάστηκε. Η μεταβλητή GRADE δεν την συμπεριλαμβάνουμε στο σετ. Η grade αναφέρεται στο επίπεδο βαθμού στο οποίο δεν είχαμε καθόλου στοιχεία και είναι τύπου string, καθώς τα στοιχεία δεν επηρεάζουν στο σύνολο δεδομένων. Τα χαρακτηριστικά των δεδομένων μας φαίνονται στον παρακάτω πίνακα.

Όνομα Μεταβλητής	Περιγραφή
DBN	Περιοχή, δήμο, αριθμός σχολείου ,
Year	Έτος που έγινε η δοκιμή
Number Tested	Συνολικός αριθμός μαθητών που δοκιμάστηκαν
Category	Γενική και Ειδική εκπαίδευση
Mean Scale Score	Μέσος όρος των συνολικών μαθητών που δοκιμάστηκαν
Num Level 1#	Αριθμός μαθητών που βαθμολογήθηκαν στο εύρος επιπέδου 1
Pct Level 1%	Ποσοστό μαθητών που βαθμολογήθηκαν στο εύρος επιπέδου 1
Num Level 2#	Αριθμός μαθητών που βαθμολογήθηκαν εύρος στο επίπεδο 2
Pct Level 2%	Ποσοστό μαθητών που βαθμολογήθηκαν στο εύρος επιπέδου 2
Num Level 3#	Αριθμός μαθητών που βαθμολογήθηκαν εύρος στο επίπεδο 3
Pct Level 3%	Ποσοστό μαθητών που βαθμολογήθηκαν στο εύρος επιπέδου 3
Num Level 4#	Αριθμός μαθητών που βαθμολογήθηκαν εύρος στο επίπεδο 4
Pct Level 4%	Ποσοστό μαθητών που βαθμολογήθηκαν στο εύρος επιπέδου 4
Num Level 3 and 4#	Αριθμός μαθητών που σημείωσαν συνδυασμούς στο επίπεδο 3 και 4
Pct Level 3+4%	Ποσοστό μαθητών που βαθμολογήθηκαν στο εύρος επιπέδου 3+4

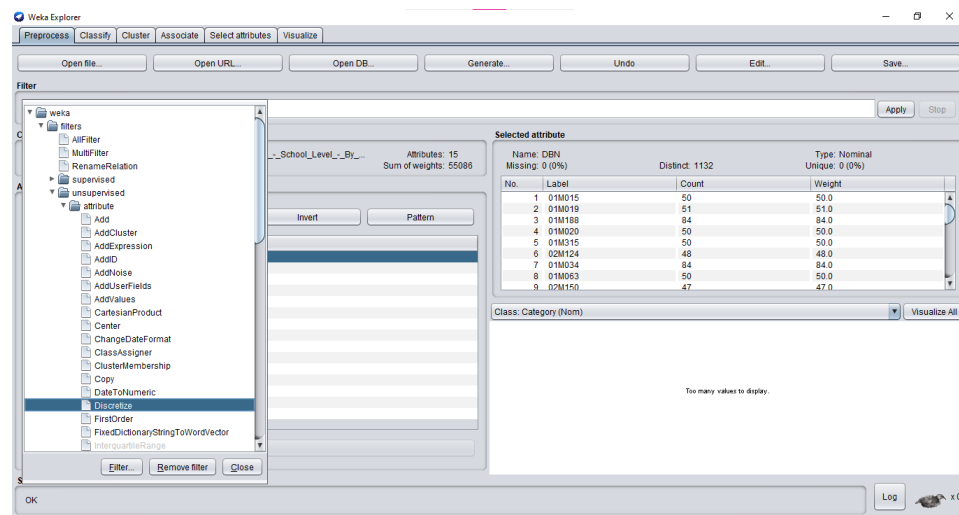
Πίνακας 2: Χαρακτηριστικά δεδομένων

- Num Level 1 = Not Meeting Learning Standards: Δεν πληρούν τα πρότυπα μάθησης
- Num Level 2 = Partially Meeting Learning Standards: Εν μέρει ικανοποίηση των μαθησιακών προτύπων
- Num Level 3 = Meeting Learning Standards: Πληρούν τα μαθησιακά πρότυπα
- Num Level 4 = Meeting Learning Standards with Distinction : πρότυπα μάθησης με διάκριση
- Num Level 3+4 = Level 3 and 4 combined: συνδυασμός επιπέδου 3+4

3.3 Προεπεξεργασία του dataset

Η διαδικασία που θα εφαρμόσουμε στο στάδιο της προεπεξεργασίας των δεδομένων στην παρούσα εργασία είναι η διακριτοποίηση των χαρακτηριστικών και η επιλογή κάποιων σημαντικών χαρακτηριστικών μέσω από το attribute selection που θα δούμε παρακάτω. Η διαδικασία της διακριτοποίησης είναι αναγκαία για το λόγο ότι κάποια χαρακτηριστικά είναι αριθμητικά.

Πρόκειται για τη διαδικασία μετατροπής αριθμητικών δεδομένων σε ονομαστικά δεδομένα. Η διακριτοποίηση των δεδομένων μπορεί να επιταχύνει τη διαδικασία εκπαίδευσης των μοντέλων και να βελτιώσει τις επιδόσεις τους. Για την μετατροπή των αριθμητικών χαρακτηριστικών σε δυαδικά απαιτείται διακριτοποίηση. Αυτό γίνεται με τη βοήθεια του weka.filters.unsupervised.attribute.Discretize στην εικόνα 11.

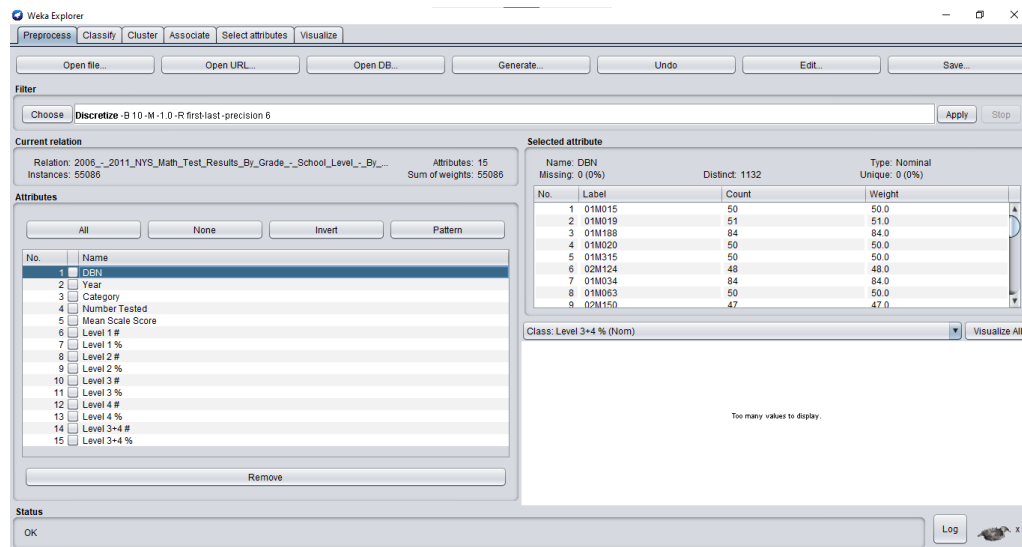


Εικόνα 11 - Επιλογή διακριτοποίησης

Τα χαρακτηριστικά παρουσιάζονται παρακάτω μετά την προεπεξεργασία:

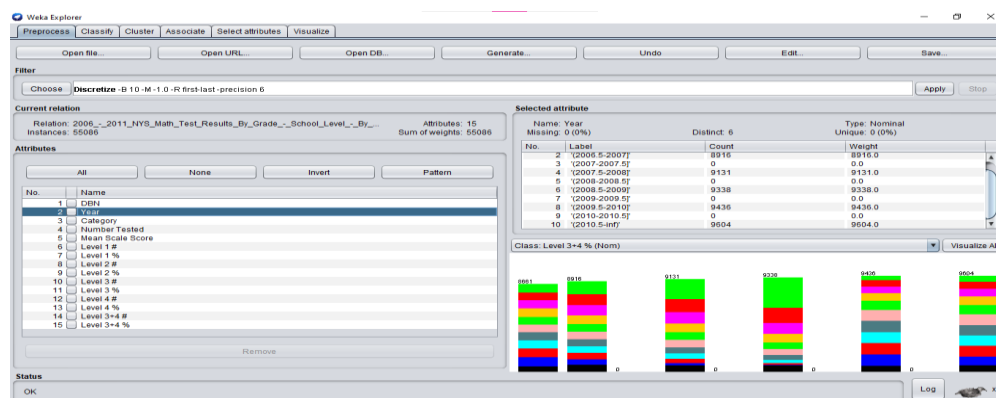
Παρατηρώντας την εικόνα 12 βλέπουμε πως η κλάση εξόδου ο αριθμός μαθητών και το ποσοστό που βαθμολογήθηκαν στο εύρος επιπέδου 3+4, level 3+4% είναι τύπου nominal μετά την διακριτοποίηση. Τα υπόλοιπα είναι χαρακτηριστικά εισόδου. Επιλέγοντας κάθε γνώρισμα μπορούμε να δούμε επιπλέον πληροφορίες εκτός από το όνομά του, όπως τον τύπο του και την κατανομή και το βάρος της κάθε τιμής που παίρνει κάθε ονομαστικό γνώρισμα.

- ❖ στο χαρακτηριστικό DBN αναγράφεται ότι δεν έχει ελλειπείς καταχωρήσεις (missing:0)



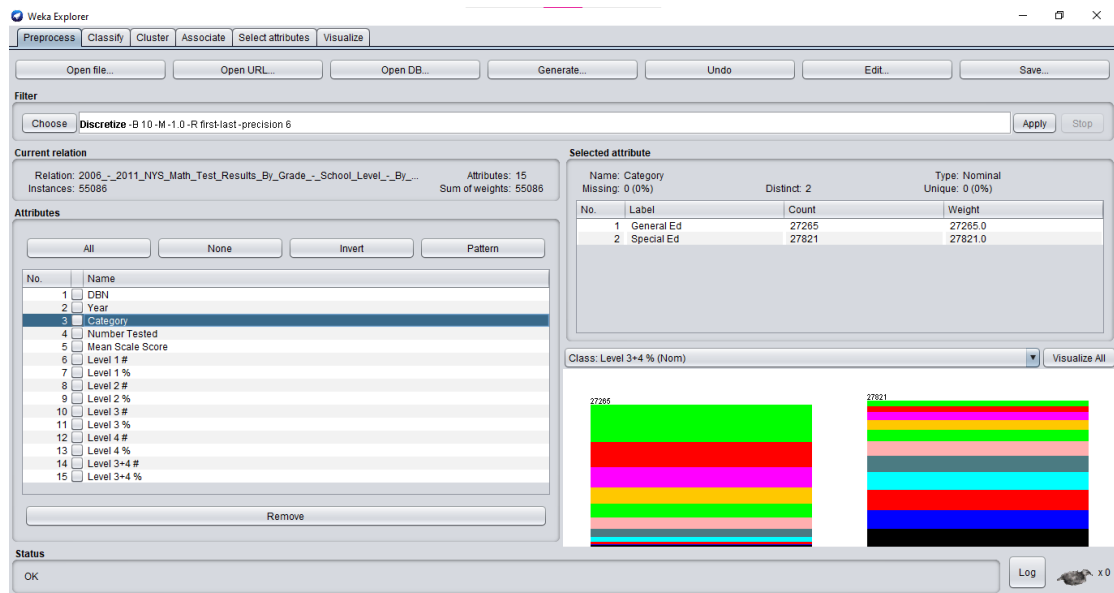
Εικόνα 12 - Χαρακτηριστικό DBN

- ❖ Στην εικόνα 13 το χαρακτηριστικό (year) είναι επιλεγμένο και στο κάτω δεξιό μέρος αναγράφεται ότι δεν έχει ελλειπείς καταχωρήσεις (missing:0), πρόκειται για ένα χαρακτηριστικό το οποίο έχει ονομαστικές τιμές με βάση το έτος (inf-2006.5, 2006.5-2007, 2007-2007.5, 2007.5-2008, 2008-2008.5, 2008.5-2009, 2009-2009.5, 2009.5-2010, 2010-2010.5, 2010.5-inf).



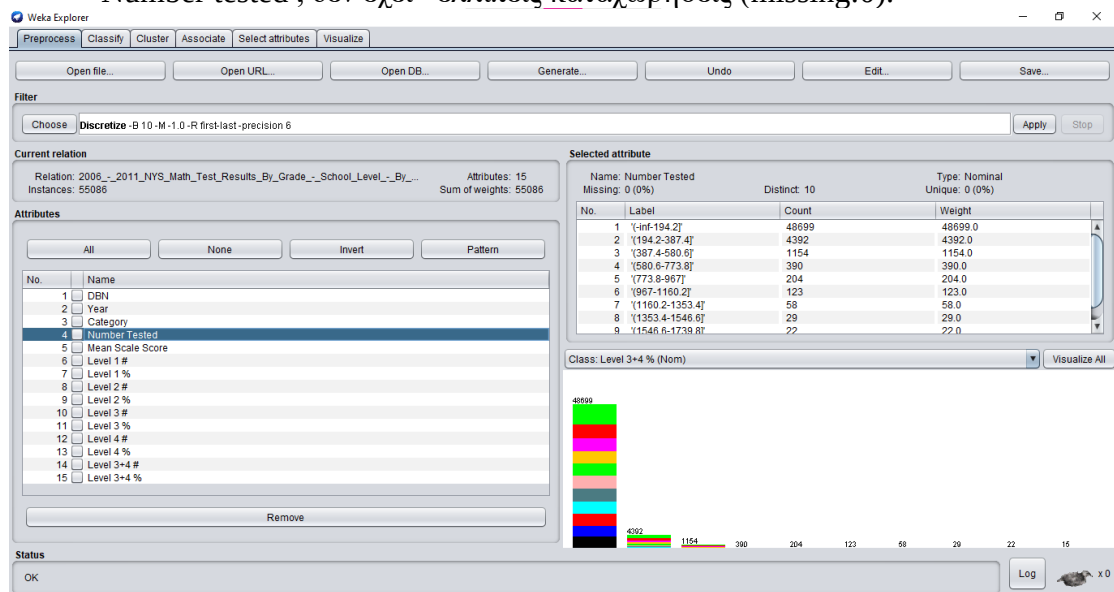
Εικόνα 13 - Χαρακτηριστικό year

- ❖ Στο χαρακτηριστικό «Category» έχουμε δύο κατηγορίες που εμφανίζονται, αφορά την γενική εκπαίδευση με 27265 μαθητές και η ειδική εκπαίδευση με 27821 μαθητές. Παρατηρούμε επίσης ότι το χαρακτηριστικό Category δεν έχει ελλειπείς καταχωρήσεις(missing:0).



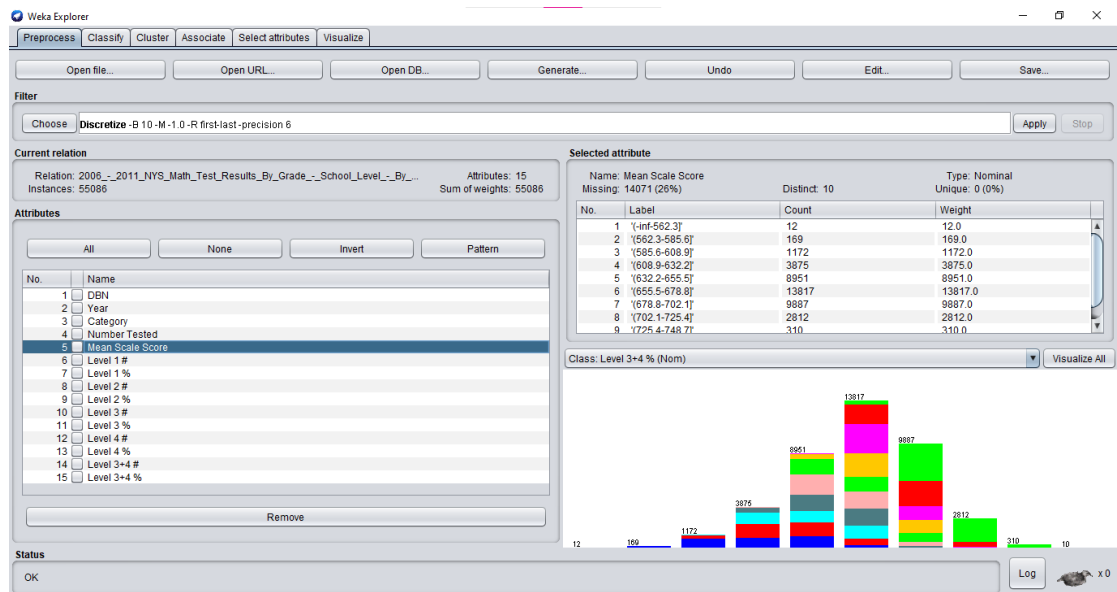
Εικόνα 14 - Χαρακτηριστικό Category

- ❖ Ο συνολικός αριθμός μαθητών που δοκιμάστηκαν στο χαρακτηριστικό Number tested, δεν έχει ελλειπίες καταχωρήσεις (missing:0).



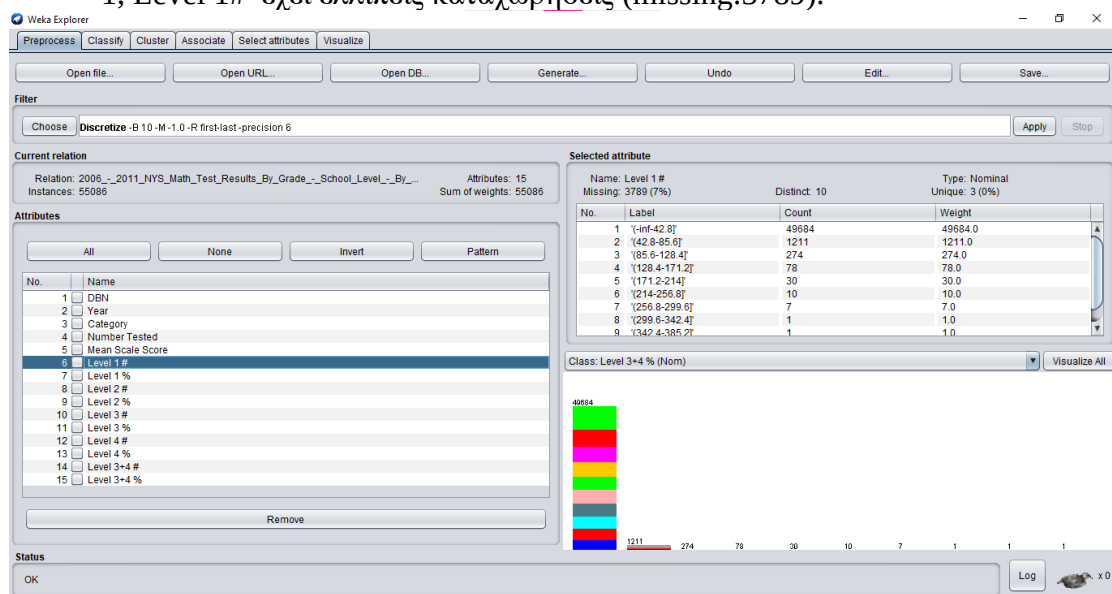
Εικόνα 15 - Χαρακτηριστικό Number tested

- ❖ Ο μέσος όρος των συνολικών μαθητών που δοκιμάστηκαν Mean scale score έχει ελλειπίες καταχωρήσεις (missing:14071).



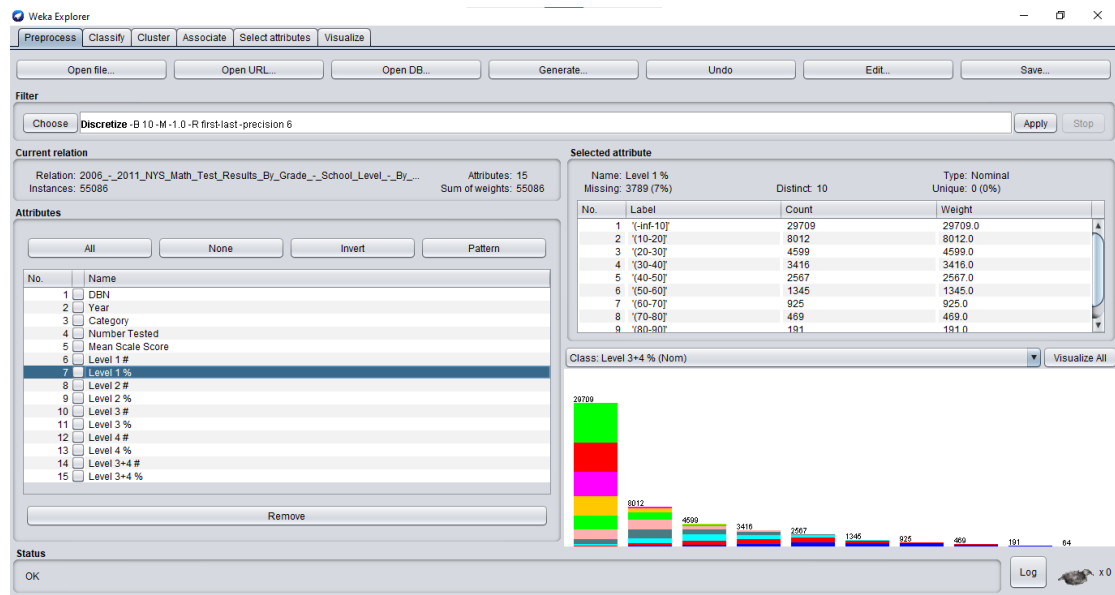
Εικόνα 16 - Χαρακτηριστικό mean scale score

- ❖ Ο αριθμός μαθητών και το ποσοστό που βαθμολογήθηκαν στο εύρος επιπέδου 1, Level 1# έχει ελλειπείς καταχωρήσεις (missing:3789).



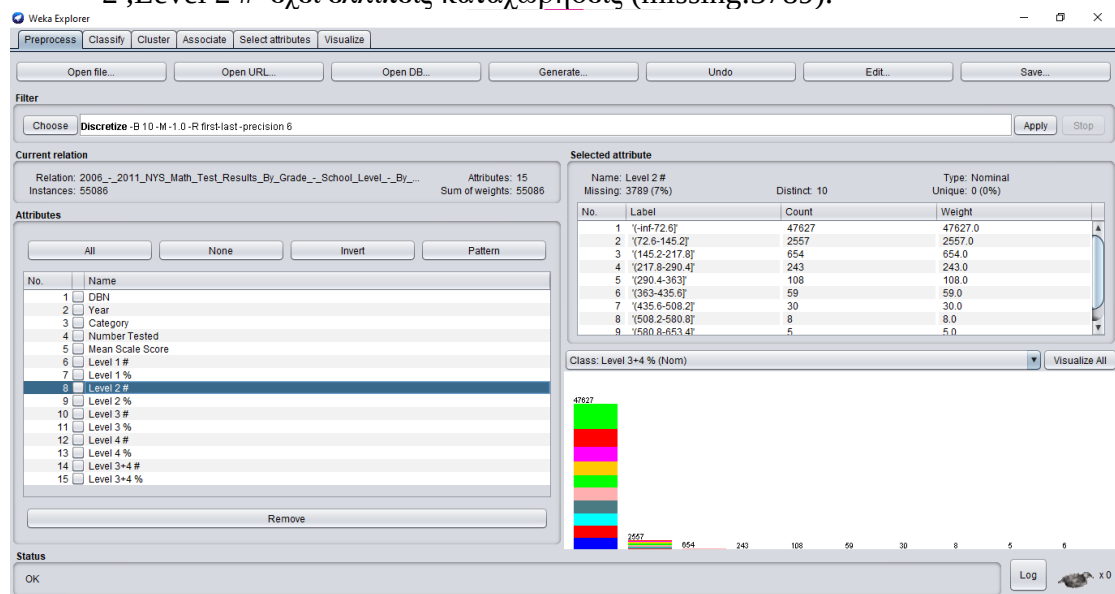
Εικόνα 17- Χαρακτηριστικό Level 1#

- ❖ Ο αριθμός μαθητών και το ποσοστό που βαθμολογήθηκαν στο εύρος επιπέδου Level % έχει ελλειπείς καταχωρήσεις (missing:3789).



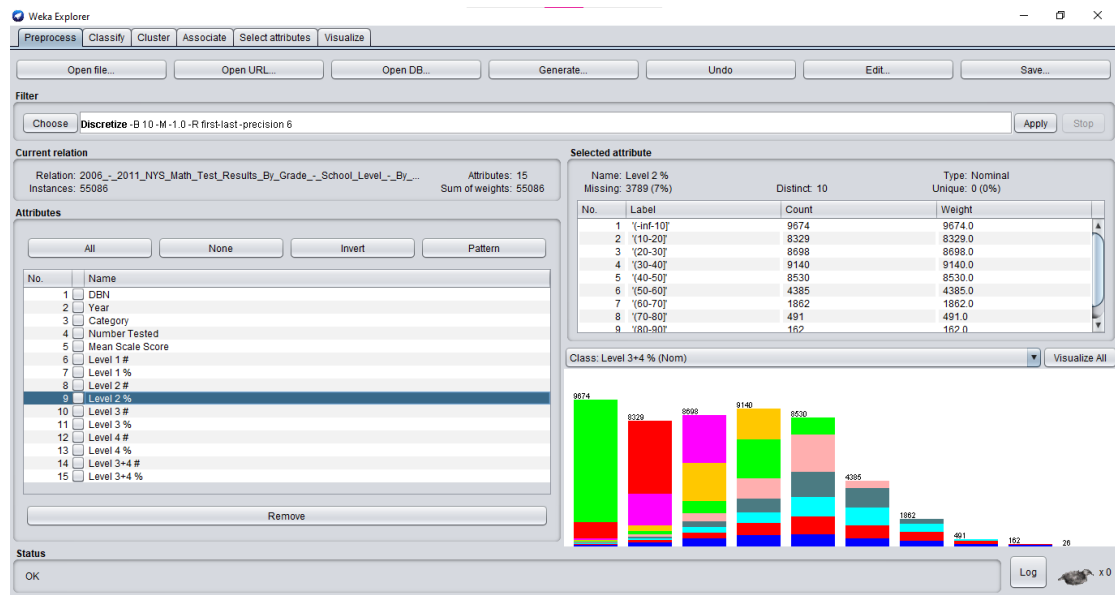
Εικόνα 18 - Χαρακτηριστικό Level 1 %

- ❖ Ο αριθμός μαθητών και το ποσοστό που βαθμολογήθηκαν στο εύρος επιπέδου 2 ,Level 2 # έχει ελλειπίες καταχωρήσεις (missing:3789).



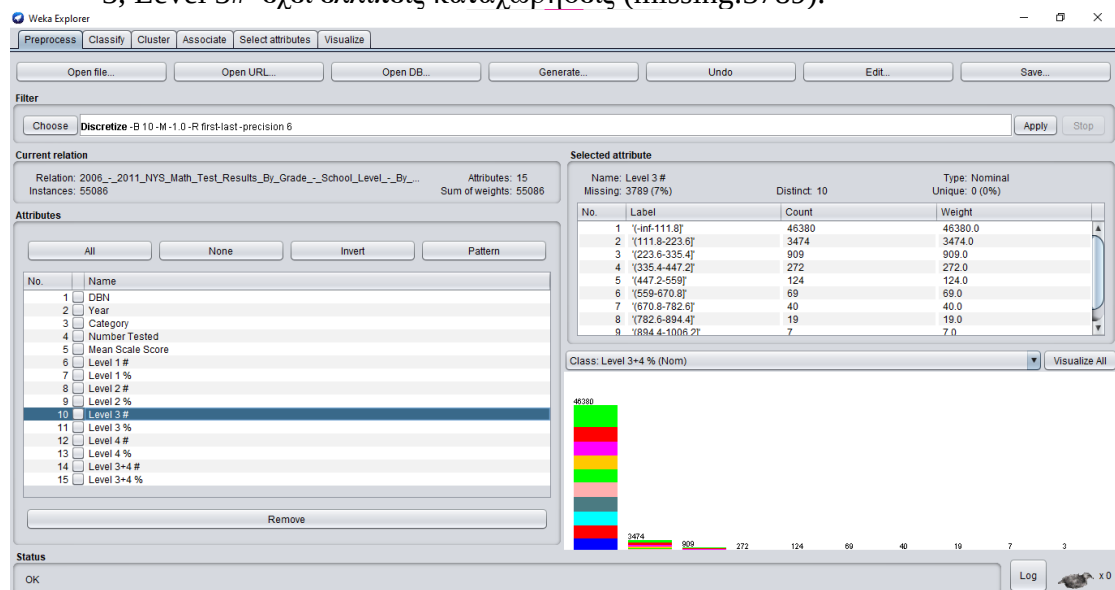
Εικόνα 19 - Χαρακτηριστικό Level 2#

- ❖ Ο αριθμός μαθητών και το ποσοστό που βαθμολογήθηκαν στο εύρος επιπέδου 2 ,level 2% έχει ελλειπίες καταχωρήσεις (missing:3789)



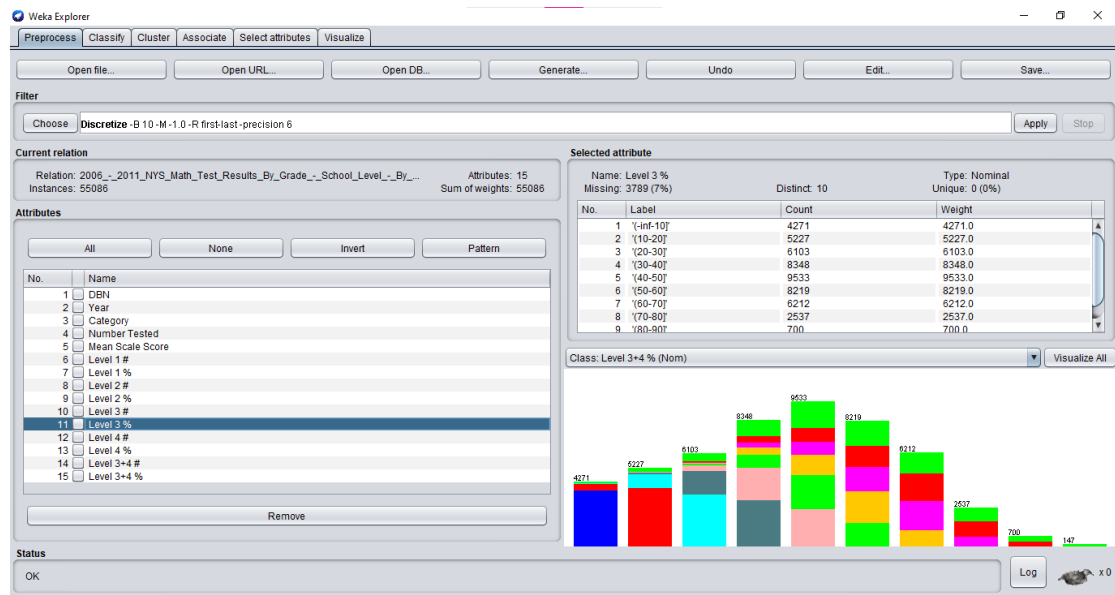
Εικόνα 20 - Χαρακτηριστικό Level 2%

- ❖ Ο αριθμός μαθητών και το ποσοστό που βαθμολογήθηκαν στο εύρος επιπέδου 3, Level 3# έχει ελλειπίες καταχωρήσεις (missing:3789).



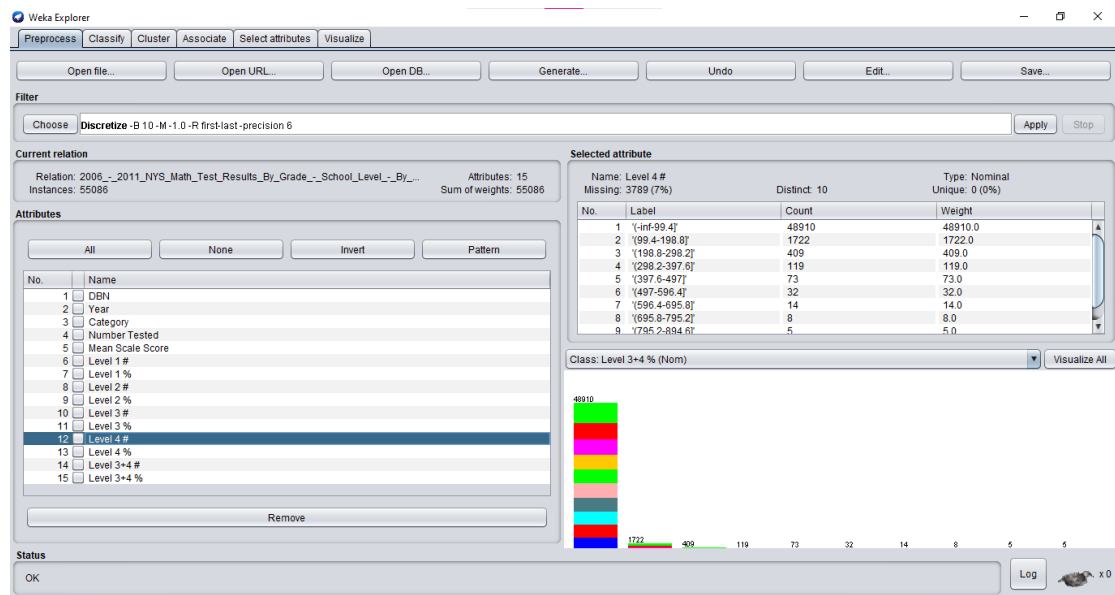
Εικόνα 21 - Χαρακτηριστικό Level 3#

- ❖ αριθμός μαθητών και το ποσοστό που βαθμολογήθηκαν στο εύρος επιπέδου 3, level 3% έχει ελλειπίες καταχωρήσεις (missing:3789).



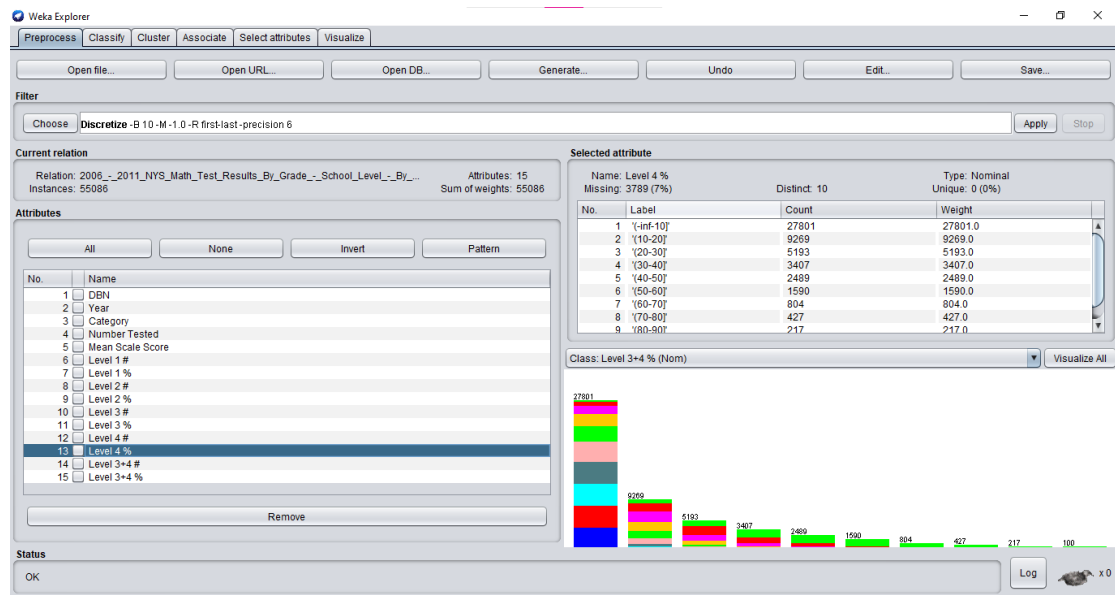
Εικόνα 22 - Χαρακτηριστικό Level 3%

- ❖ Ο αριθμός μαθητών και το ποσοστό που βαθμολογήθηκαν στο εύρος επιπέδου 4, Level 4# έχει ελλείψεις καταχωρήσεις (missing:3789)



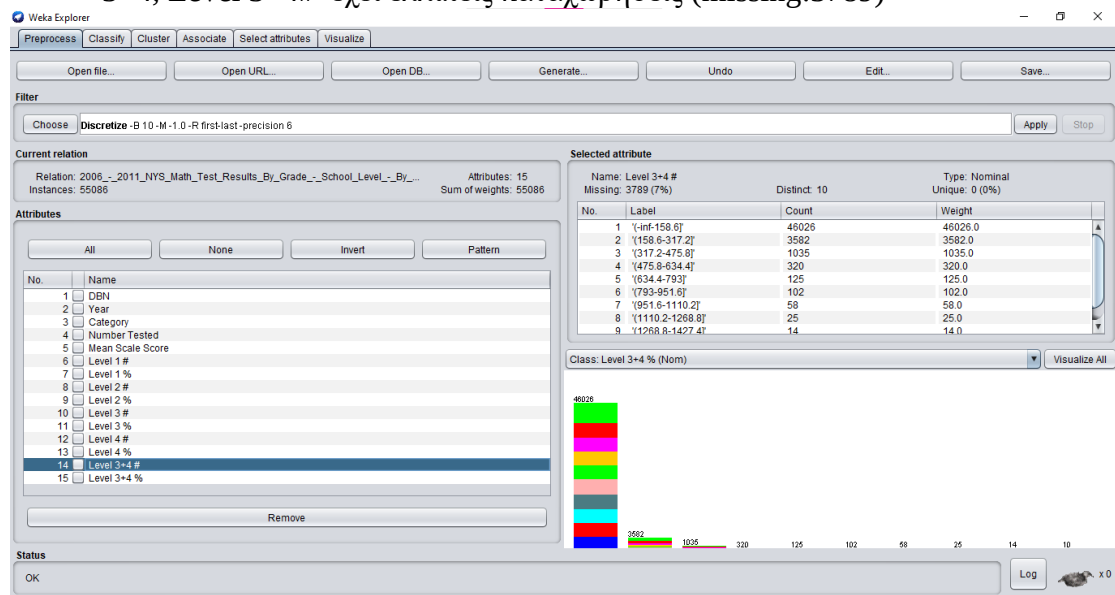
Εικόνα 23 - Χαρακτηριστικό Level 4#

- ❖ Ο αριθμός μαθητών και το ποσοστό που βαθμολογήθηκαν στο εύρος επιπέδου 4, level 4% έχει ελλείψεις καταχωρήσεις (missing:3789)



Εικόνα 24 - Χαρακτηριστικό Level 4%

- ❖ Ο αριθμός μαθητών και το ποσοστό που βαθμολογήθηκαν στο εύρος επιπέδου 3+4, Level 3+4# έχει ελλιπείς καταχωρήσεις (missing:3789)



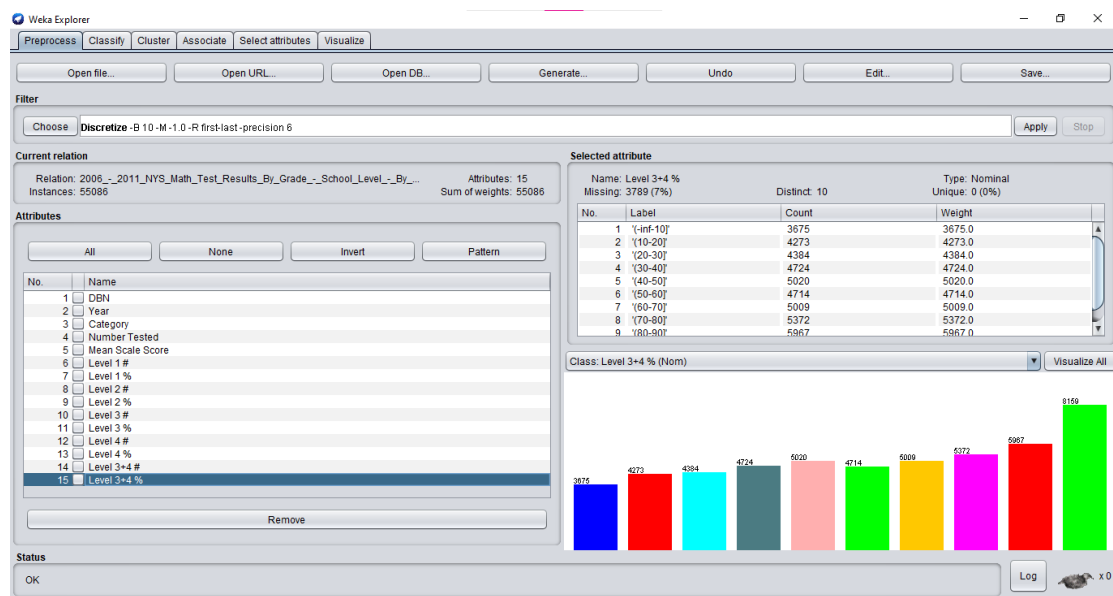
Εικόνα 25 - Χαρακτηριστικό Level 3+4#

- ❖ Ο αριθμός μαθητών και το ποσοστό που βαθμολογήθηκαν στο εύρος επιπέδου 3+4, level3+4% έχει ελλιπείς καταχωρήσεις (missing:3789). Στο παράθυρο αυτό, το χαρακτηριστικό εξόδου είναι επιλεγμένο και στο κάτω δεξιό μέρος αναγράφεται ότι έχει ελλιπείς καταχωρήσεις (missing:3789), πρόκειται για ένα χαρακτηριστικό το οποίο έχει ονομαστικές τιμές (inf-10, 10-20, 20-30,30-40, 40-50, 50-60,60-70,70-80,80-90,90-inf) .

Όπως φαίνεται στην εικόνα 26 η κάθε ομάδα των χαρακτηριστικών στην μεταβλητή εξόδου αναφέρεται με διαφορετικό χρώμα.

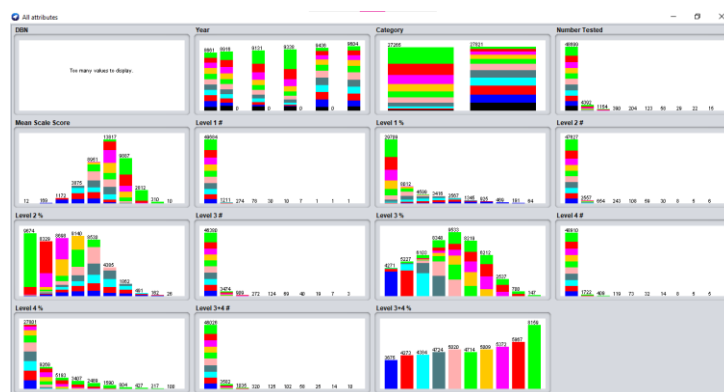
Έτσι έχουμε:

- μπλε χρώμα την πρώτη ομάδα inf-10
- κόκκινο χρώμα τη δεύτερη ομάδα 10- 20
- γαλάζιο τη Τρίτη ομάδα 20-30
- γκρι τη τέταρτη ομάδα 30-40
- ροζ τη πέμπτη ομάδα 40-50
- πράσινο την έκτη ομάδα 50-60
- φούξια την όγδοη ομάδα 70-80
- βαθύ κόκκινο την ένατη ομάδα 80-90
- ανοιχτό πράσινο την δέκατη ομάδα 90-inf



Εικόνα 26 - Συνδυασμός 3+4 μαθησιακών προτύπων Χαρακτηριστικό Level 3+4%

Τα παρακάτω ιστογράμματα εικόνα 27 μας βοηθούν στην κατανόηση των δεδομένων, δεδομένου ότι υποδεικνύουν την κατανομή των τιμών των ονομαστικών χαρακτηριστικών.



Εικόνα 27 - Πίνακας οπτικοποίησης στο σύνολο των χαρακτηριστικών

Σκοπός μας είναι να εξετάσουμε την δυνατότητα πρόβλεψης στην επίδοση των μαθηματικών για έτος 2006 -2011 και να μελετήσουμε τον κάθε αλγόριθμο που θα χρησιμοποιήσουμε. Για κάθε αλγόριθμο θα εξετάσουμε την απόδοσή του μοντέλου, όμως για να το επιτύχουμε αυτό θα χρησιμοποιήσουμε κάποιον αλγόριθμο αξιολόγησης χαρακτηριστικών.

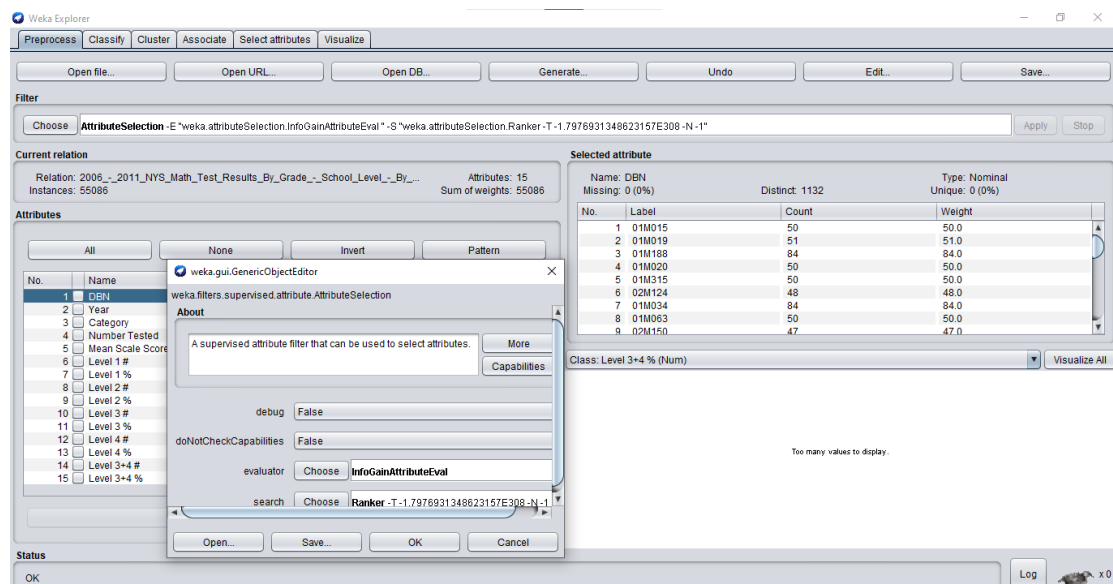
3.3.1 Προεπεξεργασία - Attribute selection

Η δεύτερη διαδικασία της προεπεξεργασίας είναι η επιλογή χαρακτηριστικών για να μας μειώσει τόσο τα δεδομένα όσο και την υπολογιστική πολυπλοκότητα. Η συλλογή που συλλέγεται είναι συνήθως μεγάλη, επομένως είναι επιθυμητό να επιλέξουμε ένα υποσύνολο δεδομένων δημιουργώντας διανύσματα χαρακτηριστικών που διαθέτουν υποσύνολο.

Η επιλογή της διαδικασίας αναγνώρισης και αφαίρεσης πολλών πιθανών περιπτώσεων και άσχετων πληροφοριών. Αυτό έχει ως αποτέλεσμα τη μείωση της διαστατικότητας των δεδομένων και με αυτόν τον τρόπο οι αλγόριθμοι μάθησης τρέχουν γρηγορότερα και πιο αποτελεσματικός τρόπος.

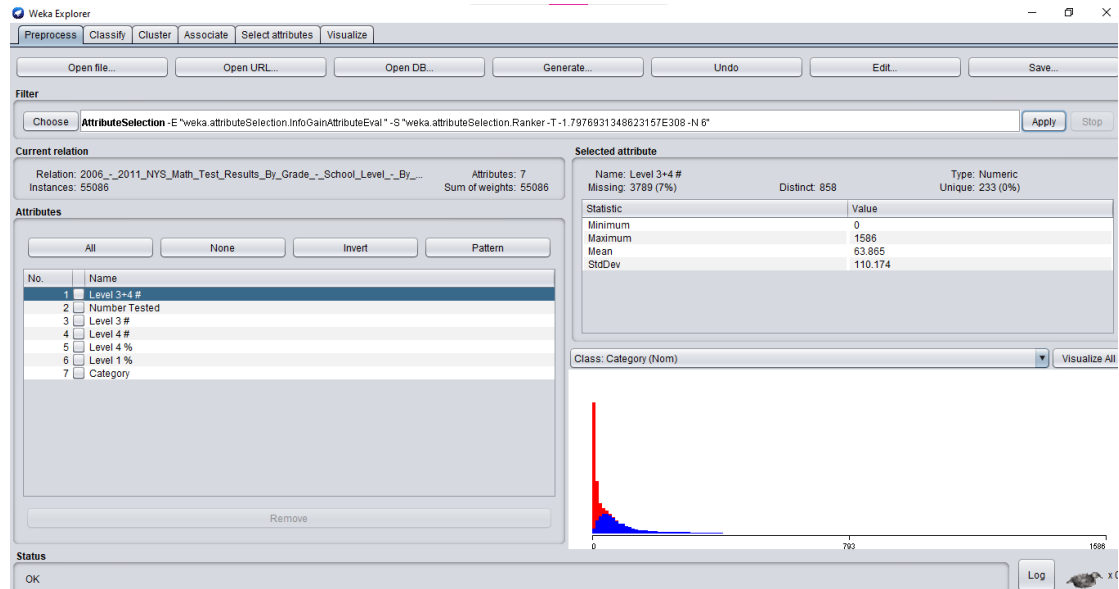
Η επιλογή του feature selection στο σύνολο δεδομένων μας, στο πεδίο “filter” - supervised και attribute selection, επιλέγουμε τα εξής evaluator - choose – infogainattributeEval. Το Information Gain είναι ένα σημαντικό μέτρο που χρησιμοποιείται για χαρακτηριστικά κατάταξης. Δεδομένου ότι η εντροπία είναι ένα κριτήριο σε ένα εκπαιδευτικό σύνολο, μπορούμε να ορίσουμε ένα μέτρο για επιπλέον πληροφορίες που παρέχονται από τα δεδομένα και αντιπροσωπεύει το πόσο η εντροπία μειώνεται.

Στη συνέχεια στο Search - Choose - Ranker, και μέσα στο ranker στο num to select= 7, για να μας κρατήσει το καλύτερο υποσύνολο με τα 7 καλύτερα χαρακτηριστικά(εικόνα 28).



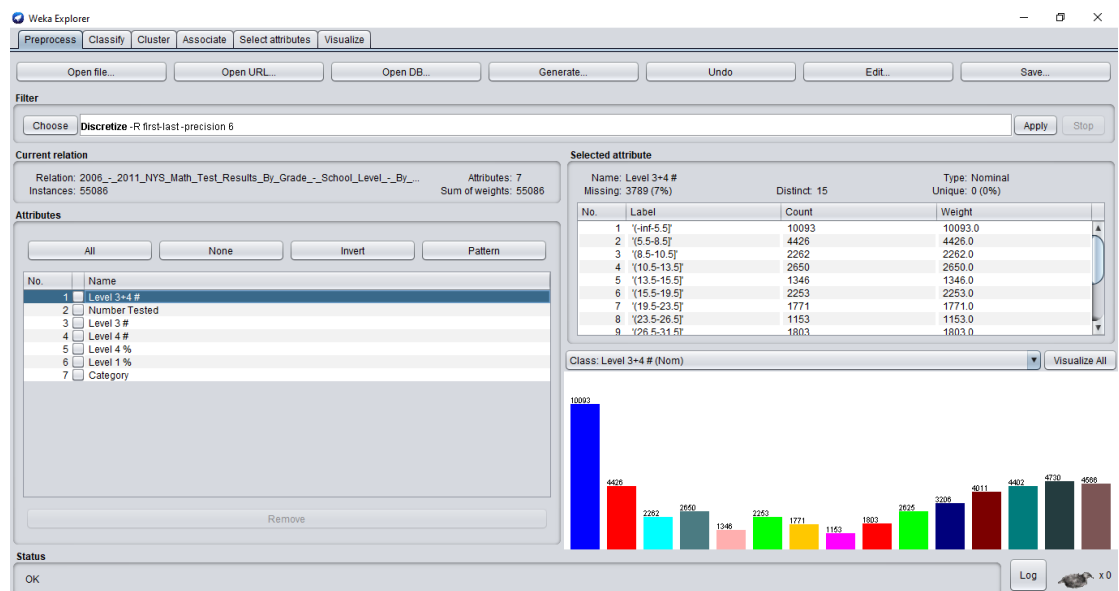
Εικόνα 28 - Attribute selection

Στην εικόνα 29 έχουμε τα σημαντικότερα χαρακτηριστικά των δεδομένων, τα οποία είναι level 3+4#,Number Tested,Level 3#,Level 4#,Level 4%,Level 1%,Category.Στην εξόρυξη γνώσης θα τρέξουμε το πείραμα με τα παραπάνω χαρακτηριστικά για να δούμε αν θα έχουμε καλύτερα αποτελέσματα των αλγορίθμων.



Εικόνα 29 – Μέθοδος προεπεξεργασίας Attribute selection

Στη συνέχεια θα γίνει μετατροπή των αριθμητικών χαρακτηριστικών σε ονομαστικά και θα εφαρμόσουμε την διακριτοποίηση. Αυτό γίνεται με τη βοήθεια του weka.filters.unsupervised.attribute.Discretize στην εικόνα 30.



Εικόνα 30- Διακριτοποίηση

3.4 Εξόρυξη δεδομένων – γνώσης

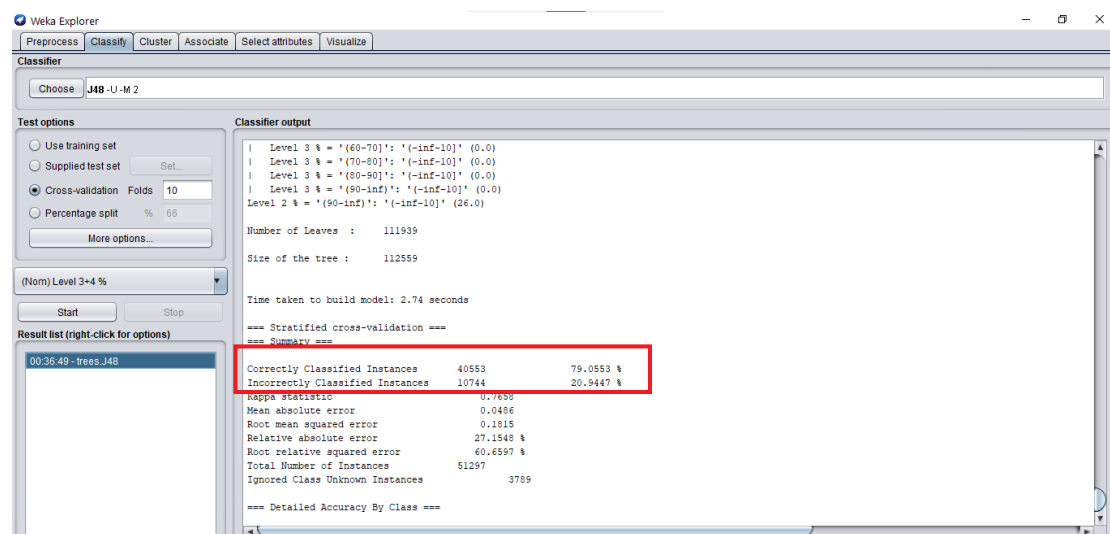
Στη συνέχεια θα κάνουμε σύγκριση μεταξύ διάφορων αλγορίθμων της κάθε κατηγορίας, για να δούμε ποιος αλγόριθμος δίνει τη μεγαλύτερη ακρίβεια. Αφού έχουμε κάνει το στάδιο της προ-επεξεργασία είμαστε στο σημείο να «τρέξουμε» διάφορους αλγορίθμους με σκοπό την κατασκευή δέντρου απόφασης. Οι αλγόριθμοι που προτείνονται συνήθως για την κατασκευή δέντρων απόφασης είναι ο αλγόριθμος J48. Έτσι θα πρέπει να αξιολογήσουμε την ακρίβεια του μοντέλου που δημιουργήθηκε με κάθε κατηγοριοποιητή.

3.4.1 Ο αλγόριθμος j48 - Δέντρα απόφασης

Ο αλγόριθμος που χρησιμοποιείται από το WEKA είναι γνωστός ως J48. Ο αλγόριθμος J48 αποτελεί έναν αλγόριθμο ο οποίος δημιουργεί ένα pruned tree. Βάση του οποίου γίνεται classify η κάθε εγγραφή του training set. Ο J48 αλγόριθμος δίνει αρκετές επιλογές σχετικά με δέντρα pruning. Επίσης το pruning δίνει λιγότερα και πιο εύκολα ερμηνεύσιμα αποτελέσματα.

Το πρώτο στοιχείο που διακρίνουμε στα αποτελέσματα είναι η ακρίβεια (Accuracy) της κατηγοριοποίησης, δηλαδή το ποσοστό των στιγμιότυπων που κατηγοριοποιήθηκαν σωστά και το ποσοστό των στιγμιότυπων που κατηγοριοποιήθηκαν λάθος. Ο αλγόριθμος J48 θα πραγματοποιηθεί μέσω της διαδικασίας Cross-validation. Στην διαδικασία αυτή μοιράζουμε τα δεδομένα σε 10 ίσα κομμάτια και στη συνέχεια εκπαιδεύουμε τα 9 κομμάτια και τα εφαρμόζουμε στο 1 που μένει.

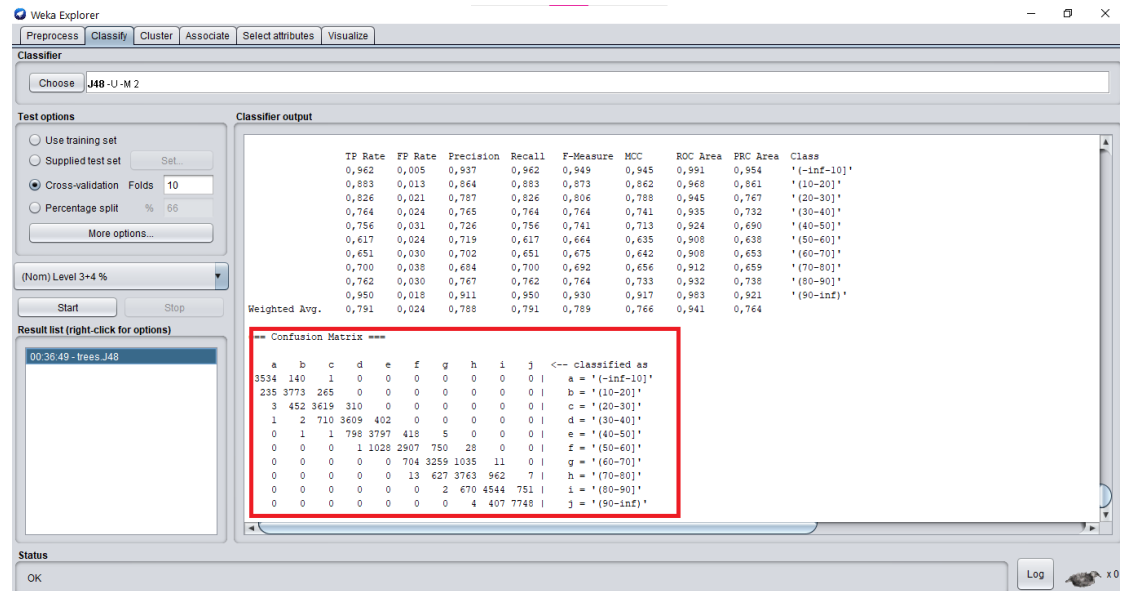
Αυτό που προκύπτει είναι πως από τα 55.086 στιγμιότυπα ταξινομήθηκαν σωστά τα 40553 και τα 10744 ταξινομήθηκαν λάθος. Δηλαδή, η ακρίβεια του classification στο μοντέλο μας είναι περίπου 79.055% του συνόλου των δεδομένων (Εικόνα 31).



Εικόνα 31 - Ταξινόμηση με τη χρήση του αλγορίθμου j48 unpruned

Η απόδοση του παραπάνω μοντέλου, μέσω της διαδικασίας cross-validation=10, είναι:

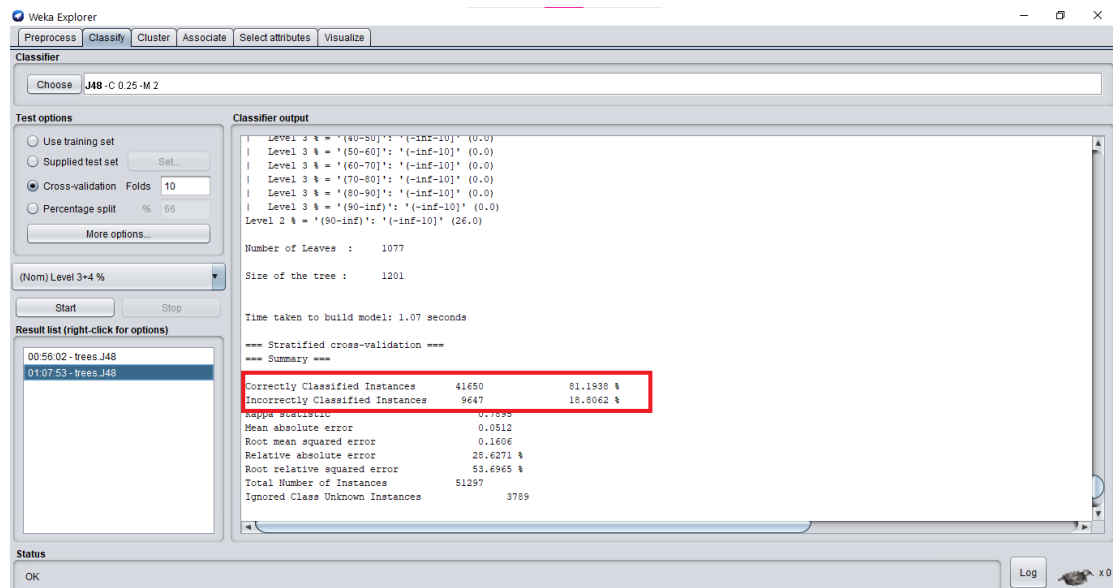
Correctly Classified Instances: 40553 79,05%
Incorrectly Classified Instances: 10744 20,94%



Εικόνα 32 - Ταξινόμηση με τη χρήση του αλγορίθμου j48 unpruned

3.4.2 Αλγόριθμος j48 pruned

Αυτό που προκύπτει από τη παρακάτω εικόνα είναι πως από τα 55.086 στιγμιότυπα ταξινομήθηκαν σωστά τα 41650 και τα 9647 ταξινομήθηκαν λάθος. Δηλαδή, η ακρίβεια του classification στο μοντέλο μας είναι περίπου 81,19% του συνόλου των δεδομένων (Εικόνα 33).

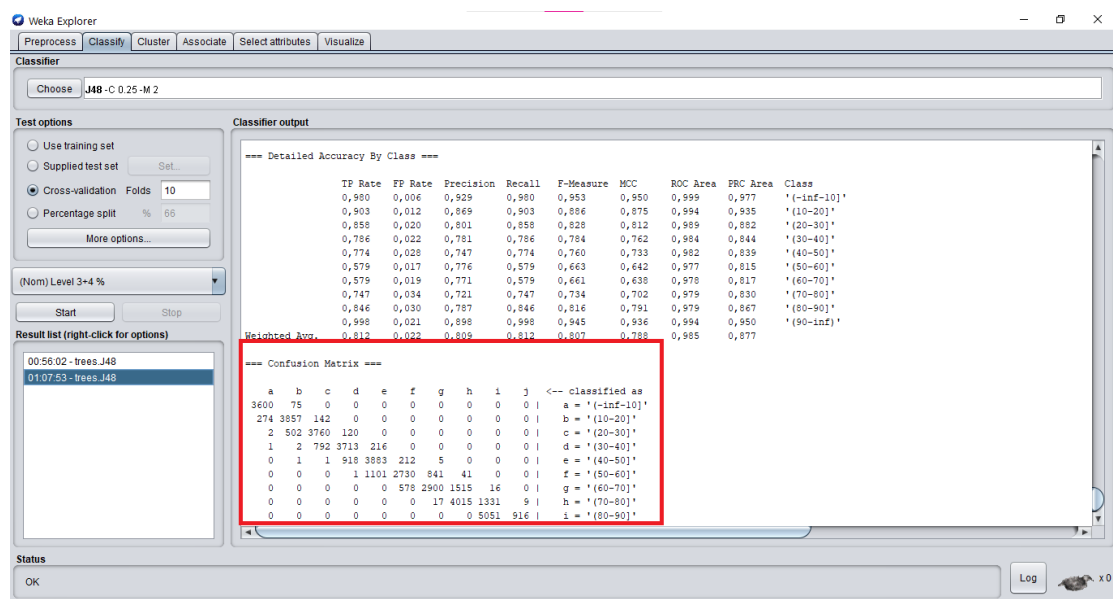


Εικόνα 33 – Ταξινόμηση με τη χρήση του αλγορίθμου j48 pruned

Η απόδοση του παραπάνω μοντέλου, μέσω της διαδικασίας cross-validation=10, είναι:

Correctly Classified Instances: 41650 81,19%

Incorrectly Classified Instances: 9647 18,80%



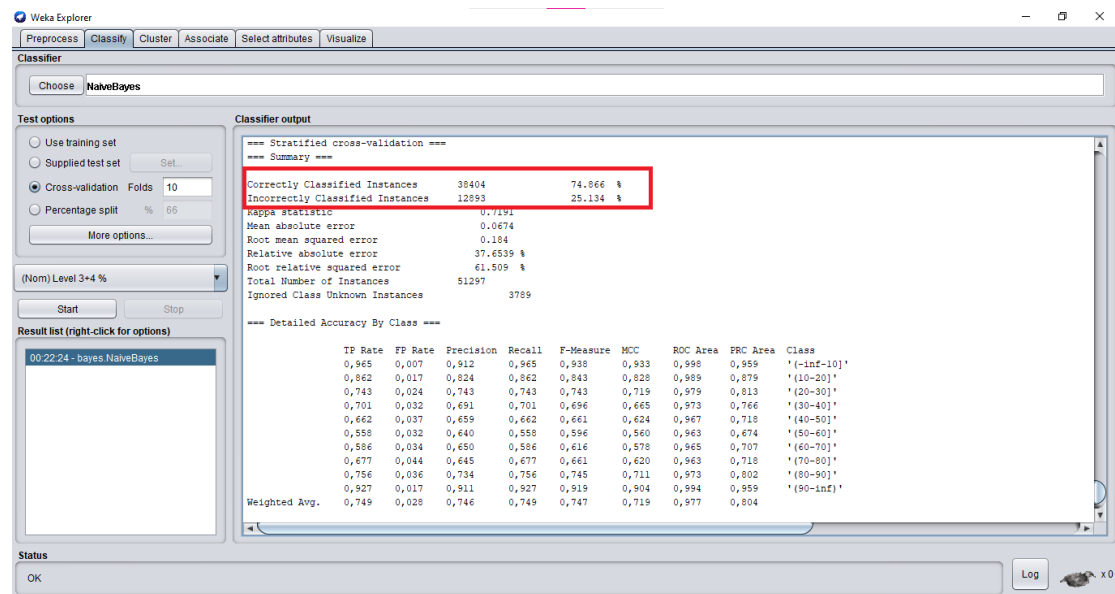
Εικόνα 34 – Ταξινόμηση με τη χρήση του αλγορίθμου j48 pruned

3.4.3 Naïve bayes

Στον NaïveBayes αλγόριθμο παρατηρήθηκε ότι ταξινομήθηκαν περισσότερα σωστά στοιχεία (38404 από τα 55086) σε σχέση με τα εσφαλμένα ταξινομημένα στοιχεία (12893), ενώ το ποσοστό σχετικού σφάλματος εντοπίστηκε στο 37.65%.

Η απόδοση του μοντέλου που προκύπτει με τον αλγόριθμο Naïve bayes , μέσω της διαδικασίας cross-validation, είναι:

Correctly Classified Instances 38404 74.86 %
Incorrectly Classified Instances 12893 25.13 %



Εικόνα 35 - Ταξινόμηση με τη χρήση του αλγορίθμου Naïve bayes

3.5 Μέθοδος K πλησιέστερων γειτόνων (K- NN)

Ο αλγόριθμος IBk είναι η υλοποίηση του Weka για τον αλγόριθμο των k κοντινότερων γειτόνων (kNN – k Nearest Neighbour). Ανήκει στην κατηγορία αλγορίθμων Μάθηση από παραδείγματα.

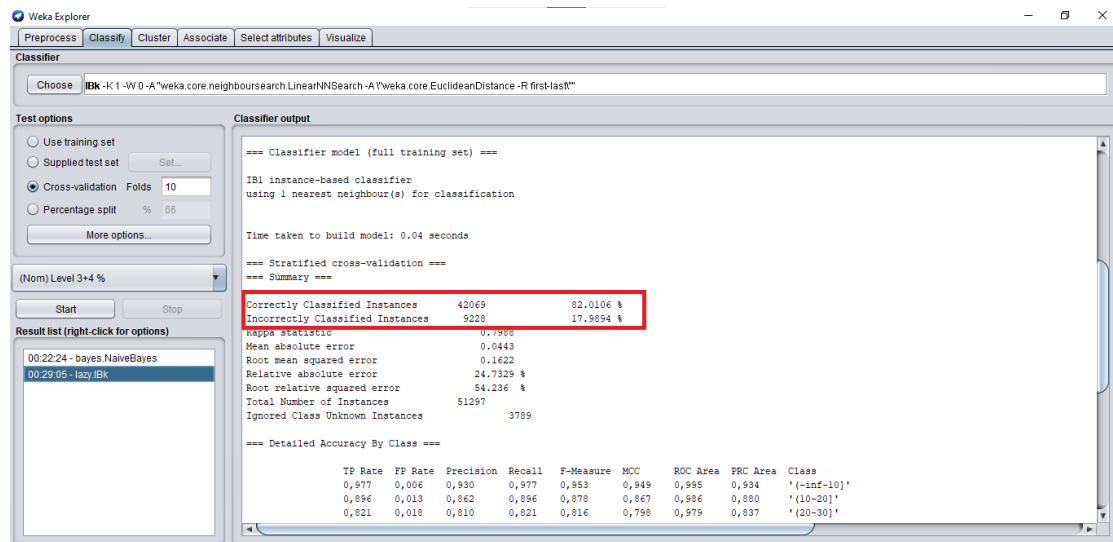
Οι παράμετροι που μπορούν να ρυθμιστούν για τον αλγόριθμο IBk είναι οι ακόλουθοι:

Η παράμετρος KNN καθορίζει τον αριθμό των κοντινότερων γειτόνων που χρησιμοποιεί ο αλγόριθμος για την πρόβλεψη (η default τιμή είναι 1). Η παράμετρος crossValidate επιλέγει τον αριθμό των γειτόνων που θα χρησιμοποιήσει ο αλγόριθμος με τη μέθοδο της διασταύρωσης hold-one-out.

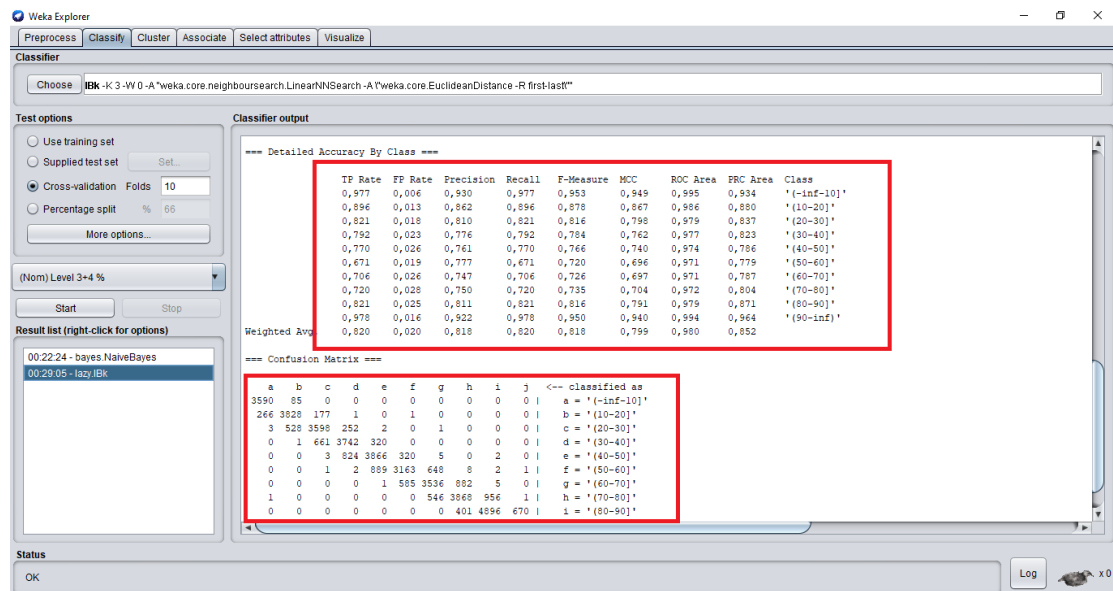
3.5.1 Αλγόριθμος IBK , (k-means) K=1

Η απόδοση του μοντέλου που προκύπτει εφαρμόζοντας τον αλγόριθμο IBK για KNN = 1, δηλαδή όταν για την ταξινόμηση κάθε νέας περίπτωσης χρησιμοποιείται 1 κοντινότερος γείτονας, μέσω της διαδικασίας cross-validation, είναι:

Correctly Classified Instances 42069 82,01 %
Incorrectly Classified Instances 9228 17,97%



Εικόνα 36 – Ταξινόμηση με τη χρήση του αλγορίθμου IBK

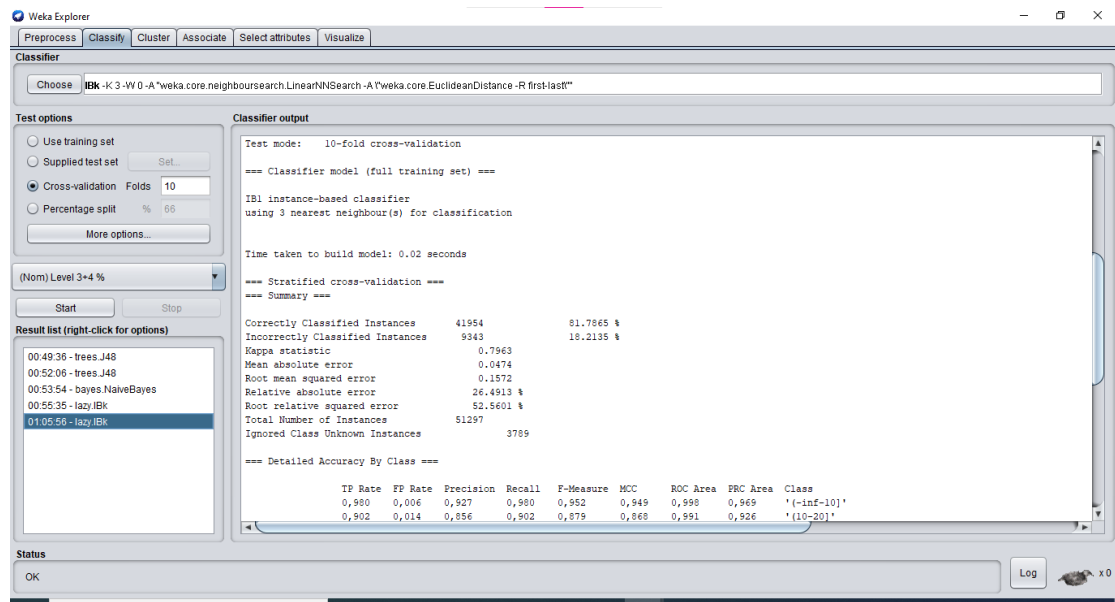


Εικόνα 37 – Confusion matrix

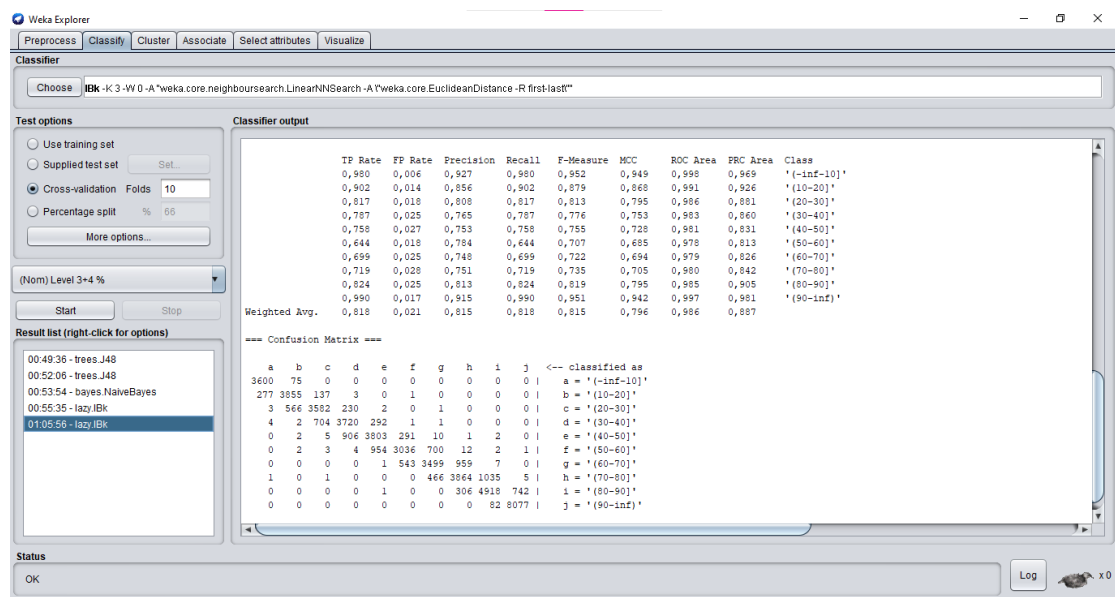
3.5.2 Αλγόριθμος IBK , (k-means) K=3

Η απόδοση του μοντέλου που προκύπτει εφαρμόζοντας τον αλγόριθμο IBK για KNN = 3, δηλαδή όταν για την ταξινόμηση κάθε νέας περίπτωσης χρησιμοποιούνται οι 3 κοντινότεροι γείτονες, μέσω της διαδικασίας cross-validation, είναι:

Correctly Classified Instances 41954 81,78 %
Incorrectly Classified Instances 9343 18,2 %

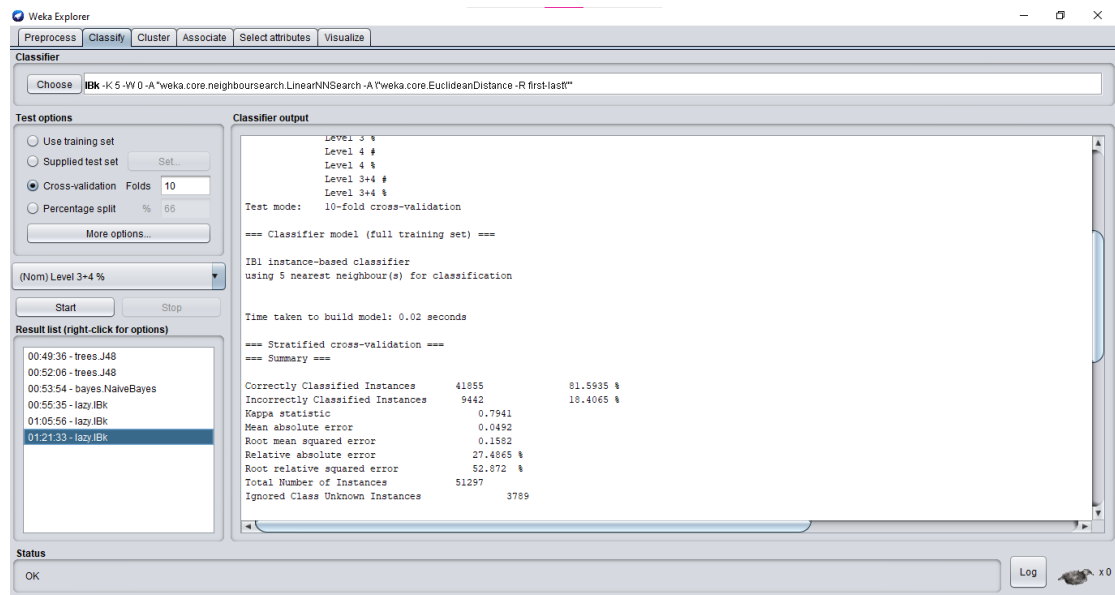


Εικόνα 38 - Ταξινόμηση με τη χρήση του αλγορίθμου IBK K=3

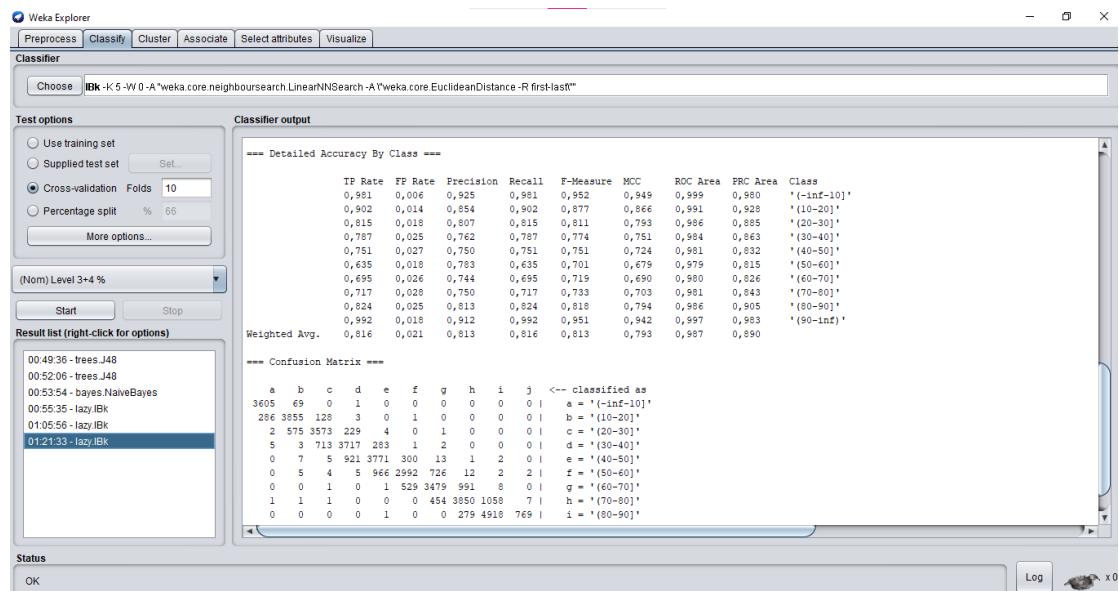


Εικόνα 39 – confusion matrix IBK K=3

3.5.3 Αλγόριθμος IBK , (k-means) K=5



Εικόνα 40 - Ταξινόμηση με τη χρήση του αλγορίθμου IBK K=5



Εικόνα 41 – confusion matrix IBK K=5

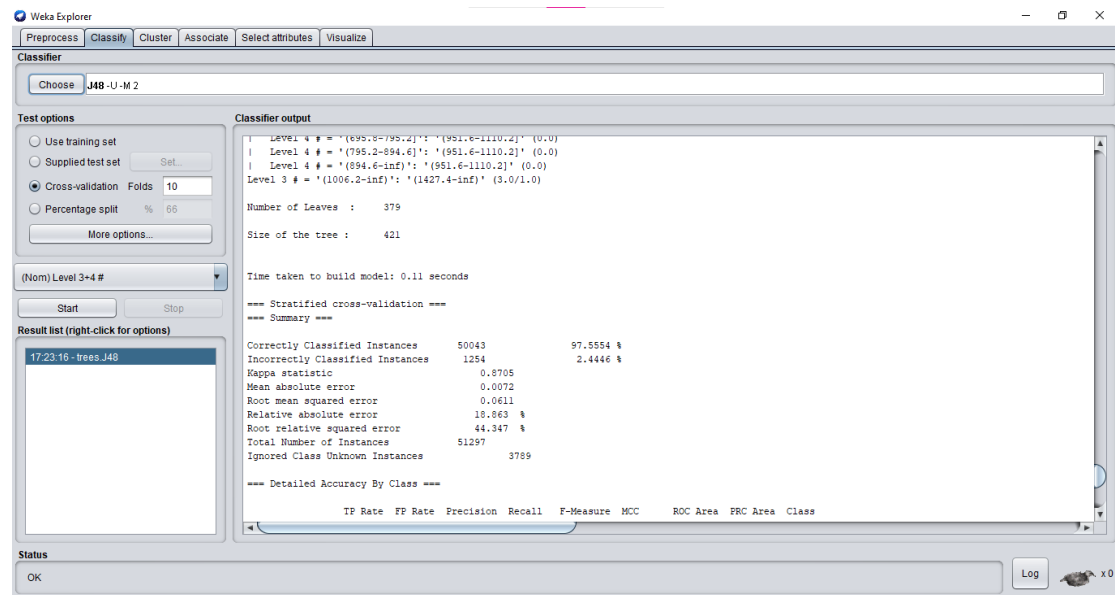
Η απόδοση του μοντέλου που προκύπτει εφαρμόζοντας τον αλγόριθμο IBK για KNN = 5, δηλαδή όταν για την ταξινόμηση κάθε νέας περίπτωσης χρησιμοποιούνται οι 5 κοντινότεροι γείτονες, μέσω της διαδικασίας cross-validation, είναι:

Correctly Classified Instances	41855	81.5935 %
Incorrectly Classified Instances	9442	18.4065 %

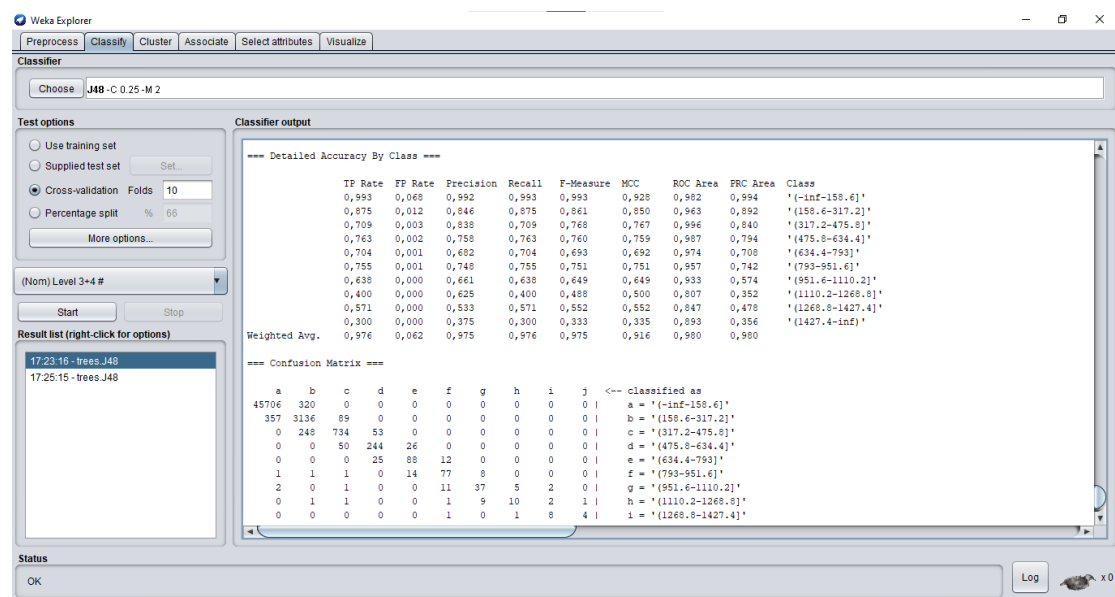
3.5.4 J48 unpruned tree (attribute selection)

Η απόδοση του μοντέλου που προκύπτει εφαρμόζοντας τον αλγόριθμο J48, μέσω της διαδικασίας cross-validation, είναι:

Correctly Classified Instances **50043** **97,55%**
Incorrectly Classified Instances **1254** **2,44 %**



Εικόνα 42- Αλγόριθμος j48

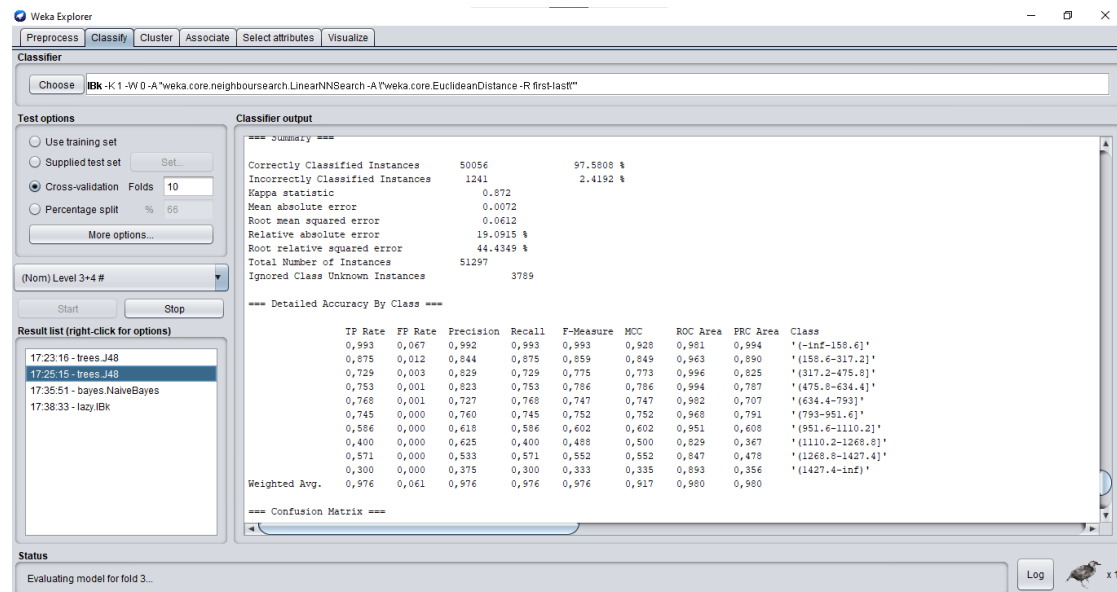


Εικόνα 43 – Πίνακας σύγχυσης

3.5.5 J48 pruned tree (attribute selection)

Η απόδοση του μοντέλου που προκύπτει εφαρμόζοντας τον αλγόριθμο J48, μέσω της διαδικασίας cross-validation, είναι:

Correctly Classified Instances **50056** **97,58%**
Incorrectly Classified Instances **1241** **2,41%**



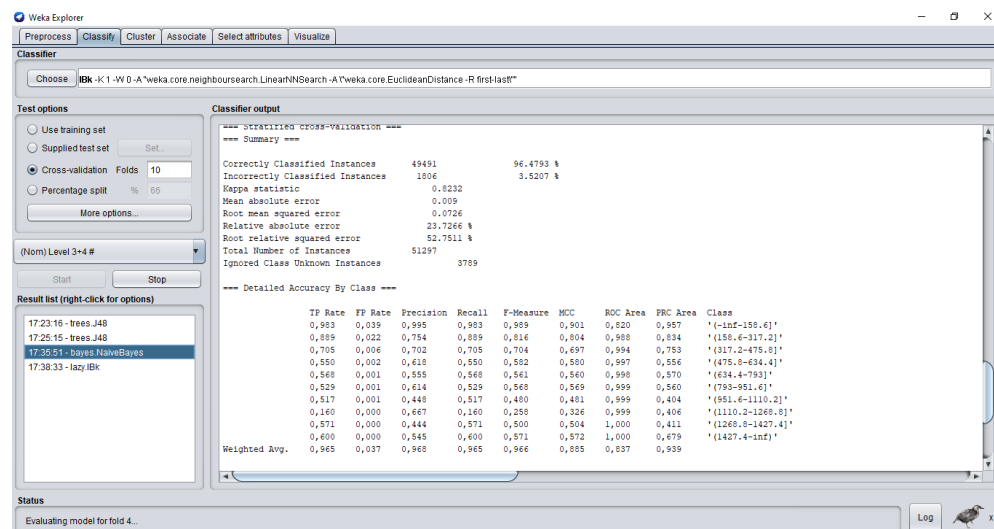
Εικόνα 44 –Αλγόριθμος J48 pruned tree

3.5.6 Naïve bayes(attribute selection)

Στον NaïveBayes αλγόριθμο παρατηρήθηκε ότι ταξινομήθηκαν περισσότερα σωστά στοιχεία (49491 από τα 55086) σε σχέση με τα εσφαλμένα ταξινομημένα στοιχεία (1806), ενώ το ποσοστό σχετικού σφάλματος εντοπίστηκε στο 23,72%.

Η απόδοση του μοντέλου που προκύπτει με τον αλγόριθμο **Naïve bayes** , μέσω της διαδικασίας cross-validation, είναι:

Correctly Classified Instances **49491** **96,47%**
Incorrectly Classified Instances **1806** **3,52%**

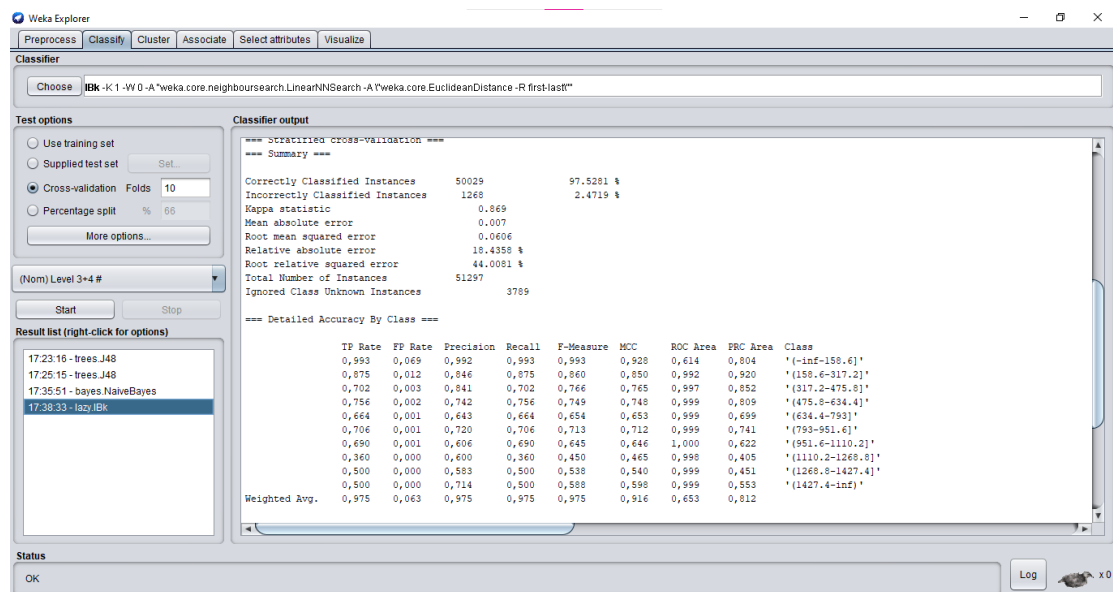


Εικόνα 45 - Naïve bayes

3.5.7 Αλγόριθμος IBK , (k-means) K=1(attribute selection)

Το συγκεκριμένο σύνολο κανόνων παρουσιάζει ικανοποιητική ισχύ, διότι μόλις το 97,52% των περιπτώσεων επαληθεύονται. Η απόδοση του αλγορίθμου είναι σε καλά επίπεδα και ο αλγόριθμος κρίνεται κατάλληλος για ταξινόμηση

Correctly Classified Instances 50029 97,52%
Incorrectly Classified Instances 1268 2,47%



3.5.8 Αλγόριθμος IBK , (k-means) K=3(attribute selection)

Η απόδοση του μοντέλου που προκύπτει εφαρμόζοντας τον αλγόριθμο IBK για KNN = 3, δηλαδή όταν για την ταξινόμηση κάθε νέας περίπτωσης χρησιμοποιούνται οι 3 κοντινότεροι γείτονες, μέσω της διαδικασίας cross-validation, είναι:

Correctly Classified Instances 49991 97,45%
Incorrectly Classified Instances 1306 2,54%

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose IBK -K 5 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A "weka.core.EuclideanDistance -R first-last"

Test options

Use training set
Supplied test set Set...
Cross-validation Folds 10
Percentage split % 66
More options...

(Nom) Level 3+4 #

Start Stop

Result list (right-click for options)

- 17:23:16 - trees.J48
- 17:25:15 - trees.J48
- 17:35:51 - bayes.NaiveBayes
- 17:38:33 - lazy.IBK
- 17:49:47 - lazy.IBK
- 18:05:34 - lazy.IBK

Classifier output

```

=== stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 49961      97.454 %
Incorrectly Classified Instances 1306      2.546 %
Kappa statistic 0.865
Mean absolute error 0.0072
Root mean squared error 0.061
Relative absolute error 15.9602 %
Root relative squared error 44.2826 %
Total Number of Instances 51297
Ignored Class Unknown Instances 3789

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	FRC Area	Class
	0.993	0.070	0.992	0.993	0.993	0.927	0.614	0.804	'(-inf-158.6)'
	0.876	0.012	0.845	0.876	0.860	0.850	0.992	0.921	'(158.6-317.2)'
	0.700	0.003	0.830	0.700	0.759	0.758	0.997	0.851	'(317.2-475.8)'
	0.709	0.002	0.723	0.709	0.716	0.714	0.999	0.798	'(475.8-634.4)'
	0.632	0.001	0.598	0.632	0.615	0.614	0.999	0.683	'(634.4-793)'
	0.657	0.001	0.670	0.657	0.663	0.663	0.999	0.724	'(793-951.6)'
	0.552	0.001	0.492	0.552	0.520	0.521	0.999	0.559	'(951.6-1110.2)'
	0.120	0.000	0.375	0.120	0.182	0.212	0.998	0.370	'(1110.2-1268.8)'
	0.643	0.000	0.692	0.643	0.667	0.667	0.999	0.475	'(1268.8-1427.4)'
	0.400	0.000	0.800	0.400	0.533	0.566	1.000	0.590	'(1427.4-inf)'
Weighted Avg.	0.975	0.064	0.974	0.975	0.974	0.915	0.653	0.812	

Status

Evaluating model for fold 1...

Log x1

3.5.9 Αλγόριθμος IBK , (k-means) K=5

Η απόδοση του μοντέλου που προκύπτει εφαρμόζοντας τον αλγόριθμο IBK για KNN = 5, δηλαδή όταν για την ταξινόμηση κάθε νέας περίπτωσης χρησιμοποιούνται οι 5 κοντινότεροι γείτονες, μέσω της διαδικασίας cross-validation, είναι:

Correctly Classified Instances 49963 97,39%
Incorrectly Classified Instances 1334 2,60%

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose IBK -K 5 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A "weka.core.EuclideanDistance -R first-last"

Test options

Use training set
Supplied test set Set...
Cross-validation Folds 10
Percentage split % 66
More options...

(Nom) Level 3+4 #

Start Stop

Result list (right-click for options)

- 17:23:16 - trees.J48
- 17:25:15 - trees.J48
- 17:35:51 - bayes.NaiveBayes
- 17:38:33 - lazy.IBK
- 17:49:47 - lazy.IBK
- 18:05:34 - lazy.IBK

Classifier output

```

=== stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 49963      97.3995 %
Incorrectly Classified Instances 1334      2.6005 %
Kappa statistic 0.8622
Mean absolute error 0.0074
Root mean squared error 0.0615
Relative absolute error 15.385 %
Root relative squared error 44.6285 %
Total Number of Instances 51297
Ignored Class Unknown Instances 3789

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	FRC Area	Class
	0.993	0.069	0.992	0.993	0.993	0.927	0.614	0.804	'(-inf-158.6)'
	0.877	0.012	0.842	0.877	0.859	0.849	0.992	0.921	'(158.6-317.2)'
	0.687	0.003	0.816	0.687	0.746	0.744	0.997	0.844	'(317.2-475.8)'
	0.706	0.002	0.704	0.706	0.705	0.703	0.998	0.770	'(475.8-634.4)'
	0.584	0.001	0.608	0.584	0.596	0.595	0.999	0.659	'(634.4-793)'
	0.608	0.001	0.633	0.608	0.620	0.619	0.999	0.709	'(793-951.6)'
	0.500	0.001	0.460	0.500	0.479	0.479	0.999	0.529	'(951.6-1110.2)'
	0.000	0.000	0.000	0.000	0.000	-0.000	0.998	0.286	'(1110.2-1268.8)'
	0.643	0.000	0.692	0.643	0.667	0.667	0.999	0.604	'(1268.8-1427.4)'
	0.400	0.000	1.000	0.400	0.571	0.632	1.000	0.763	'(1427.4-inf)'
Weighted Avg.	0.974	0.063	0.973	0.974	0.974	0.914	0.653	0.811	

Status

OK

Log x0

3.6 Σύγκριση Αλγορίθμων

Στην πίνακα 3 φαίνεται η συγκριτική απόδοση των αλγορίθμων που δοκιμάστηκαν. Από το πίνακα διαφαίνεται πως η ακρίβεια των μοντέλων που προέκυψαν κυμαίνεται από 74,86% έως και 82,01%. Η ακρίβεια τους, επομένως, μπορεί να θεωρηθεί ικανοποιητική. Το μοντέλο που παρουσιάζει τη μεγαλύτερη ακρίβεια είναι αυτό που προκύπτει από την εφαρμογή του αλγορίθμου IBK K=1, ενώ τη μικρότερη αυτό που προκύπτει από τον αλγόριθμο Naïve bayes.

Αλγόριθμοι	j48 (unpruned)	j48 (pruned)	Naïve bayes	IBk K=1	IBk K=3	IBk K=5
Correctly Classified Instances	79,05%	81,19%	74.86 %	82,01%	81,78%	81.59%
Incorrectly Classified Instances	20,94%	18,80%	25.13 %	17,97%	18,2%	25.13%
Precision	0,937	0,929	0,912	0,930	0,980	0,925
Recall	0,962	0,980	0,965	0,977	0,952	0,981
TP Rate	0,962	0,980	0,965	0,977	0,980	0,981
FP Rate	0,005	0,006	0,007	0,006	0,006	0,006

Πίνακας 3 : Συγκεντρωτικός Πίνακας Αλγορίθμων

Στον πίνακα 4 παρουσιάζονται τα αποτελέσματα που δίνουν οι αλγόριθμοι με το attribute selection στην φάση της προεπεξεργασίας. Από το πίνακα φαίνεται πως η ακρίβεια των μοντέλων που προέκυψαν κυμαίνεται από 96,47% έως και 97,58%. Η ακρίβεια τους, επομένως, μπορεί να θεωρηθεί αρκετά ικανοποιητική. Το μοντέλο που παρουσιάζει τη μεγαλύτερη ακρίβεια είναι αυτό που προκύπτει από την εφαρμογή του αλγορίθμου j48 (pruned), ενώ τη μικρότερη αυτό που προκύπτει από τον αλγόριθμο Naïve bayes.

Αλγόριθμοι	j48 (unpruned)	j48 (pruned)	Naïve bayes	IBk K=1	IBk K=3	IBk K=5
Correctly Classified Instances	97,55%	97.58%	96,47%	97,52%	97,45%	97,39%
Incorrectly Classified Instances	2,44%	2,41%	3,52%	2,47%	2,57%	2,60%
Precision	0,992	0,992	0,995%	0,992%	0,992	0,992
Recall	0,993	0,993	0,983	0,993%	0,993	,0993

Πίνακας 4 : Συγκεντρωτικός πίνακας αλγορίθμων με τα σημαντικότερα χαρακτηριστικά

Στη διάρκεια εκτέλεσης των αλγορίθμων, ο αλγόριθμος IBK $K=1$ ταξινομήσε σωστά τα περισσότερα στοιχεία (82,01%), ενώ με επιλογή των καλύτερων χαρακτηριστικών ο αλγόριθμος j48 (pruned) ταξινομήσε σωστά το 97,58% των στοιχείων του. Τα αποτελέσματα που προκύπτουν από την εκτέλεση των αλγορίθμων δείχνουν ότι το μοντέλο πρόβλεψης με τα 16 χαρακτηριστικά σε σχέση με την μοντέλο πρόβλεψης με τα 6 χαρακτηριστικά έχει μια μεγαλύτερη αύξηση η οποία δείχνει την επίδοση του μαθητή στο αν πληροί τα μαθησιακά πρότυπα και τα μαθησιακά πρότυπα με διάκριση.

Κεφάλαιο 4

4.1 ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην παρούσα εργασία capstone project έγινε μία προσπάθεια για την έγκυρη και έγκαιρη πρόβλεψη επίδοσης των μαθητών στα μαθηματικά σε ένα σύστημα γενικής και ειδικής εκπαίδευσης. Με τη χρήση του λογισμικού WEKA και την εφαρμογή γνωστών αλγορίθμων ταξινόμησης προέκυψαν οι κανόνες. Ακόμα εφαρμόστηκαν διάφοροι αλγόριθμοι εξόρυξης δεδομένων σε εκπαιδευτικά δεδομένα που περιείχαν στοιχεία των μαθητών και η επίδοσή τους στο μάθημα των μαθηματικών από 2006-2011.

Η ακρίβεια των μοντέλων που προέκυψαν κυμαίνεται από 74,05% έως και 82,01%. Επομένως, η ακρίβεια τους μπορεί να θεωρηθεί ικανοποιητική και κατά συνέπεια θα μπορούσαν τα μοντέλα αυτά να χρησιμοποιηθούν για μελλοντικές εγγραφές. Στη συνέχεια εφαρμόστηκαν τα κατάλληλα χαρακτηριστικά με την επιλογή του attribute selection και παρατηρήσαμε ότι η ακρίβεια των μοντέλων που προέκυψε ήταν από 96,47% έως και 97,58%, επομένως η ακρίβεια μπορεί να θεωρηθεί αρκετά ικανοποιητική.

Συμπερασματικά και στις δύο περιπτώσεις οι κατηγοριοποιητές έδωσαν διαφορετικά αποτελέσματα. Στην πρώτη περίπτωση που χρησιμοποιήσαμε όλα τα χαρακτηριστικά του data set παρατηρούμε ότι οι αλγόριθμοι δεν πήγαν και τόσο καλά. Στην δεύτερη περίπτωση που μειώσαμε τα χαρακτηριστικά με την μέθοδο attribute selection οι αλγόριθμοι ταξινομήθηκαν αρκετά ικανοποιητικά. Αυτό οφείλεται στο γεγονός ότι στο dataset συναντήσαμε ελλιπής καταχωρίσεις στα χαρακτηριστικά μας. Τα δεδομένα μας δεν ήταν «καθαρά».

Στο στάδιο της προεπεξεργασίας (κεφάλαιο 2.3) αναφέραμε ότι ένα πολύ συνηθισμένο πρόβλημα είναι η ύπαρξη χαμένων τιμών, η έλλειψη δηλαδή τιμών σε ορισμένα πεδία καταχωρημένων εγγράφων, αυτό σημαίνει ότι κάποια χαρακτηριστικά μας δεν καταγράφηκαν ή μπορεί να καταγράφηκαν λανθασμένα ή μπορεί να λείπουν κάποια ενδιαφέροντα χαρακτηριστικά που να μην θεωρήθηκαν σημαντικά. Ένα άλλο πρόβλημα θεωρείται και ο «θόρυβος» που μπορεί κάποιες τιμές να περιέχουν λάθη ή τιμές που διαφέρουν από την πλειοψηφία.

Στην δεύτερη περίπτωση που οι αλγόριθμοι έδωσαν καλύτερη απόδοση διότι χρησιμοποιήθηκαν λιγότερα χαρακτηριστικά με ελλιπής καταχωρίσεις είδαμε τους καλύτερους να φτάνουν σε ακρίβεια 97,58% και κατάφεραν να κατασκευάσουν μοντέλα επίδοσης μαθηματικών με μεγαλύτερη πιθανότητα να υλοποιηθεί και να πετύχει.

4.2 Μελλοντική Έρευνα

Στην παρούσα εργασία capstone project έγινε μία προσπάθεια για την έγκυρη και έγκαιρη πρόβλεψη επίδοσης των μαθητών στα μαθηματικά σε ένα σύστημα γενικής και ειδικής εκπαίδευσης. Θα μπορούσαν να γίνουν προσπάθειες για την ενημέρωση του αρχείου δεδομένων με περισσότερες εγγραφές μαθητών με βελτιωμένα χαρακτηριστικά ή ακόμα και με αντικατάσταση κάποιων χαρακτηριστικών με κάποια άλλα που να παρουσιάζουν περισσότερο ενδιαφέρον. Αυτό θα είχε σαν αποτέλεσμα

να έχουμε αποτελεσματικά συμπεράσματα και μοντέλα πρόβλεψης ώστε η εκπαίδευση να βελτιώσει την αποτελεσματικότητα προς όφελος των μαθητών. Στο στάδιο της προεπεξεργασίας των δεδομένων θα μπορούσε να γίνει η επιλογή των κατάλληλων χαρακτηριστικών χρησιμοποιώντας διάφορα φίλτρα, ανάλογα με τους αλγορίθμους που θα επιλεγούν για να χρησιμοποιηθούν. Τέλος η δημιουργία εκτενών πειραμάτων με εφαρμογή διαφόρων αλγορίθμων εξόρυξης δεδομένων και συνδυασμού αυτών, ώστε να προκύψουν τα διάφορα μοντέλα πρόβλεψης.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Βερύκιος, Β., Καγκλής, Β., Σταυρόπουλος, Η. 2015. Εισαγωγή στην Εξόρυξη Δεδομένων. [Κεφάλαιο Συγγράμματος]. Στο Βερύκιος, Β., Καγκλής, Β., Σταυρόπουλος, Η. 2015. Η επιστήμη των δεδομένων μέσα από τη γλώσσα R. [ηλεκτρ. βιβλ.] Αθήνα:Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. κεφ 1. Διαθέσιμο στο: <http://hdl.handle.net/11419/2966>

[2] Κύρκος, Ε. 2015. Οδηγός WEKA. [Κεφάλαιο Συγγράμματος]. Στο Κύρκος, Ε. 2015. Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων. [ηλεκτρ. βιβλ.] Αθήνα:Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. κεφ 13. Διαθέσιμο στο: <http://hdl.handle.net/11419/1239>

[3] Κύρκος, Ε. 2015. Εναλλακτικές Μέθοδοι και ειδικά θέματα Κατηγοριοποίησης. [Κεφάλαιο Συγγράμματος]. Στο Κύρκος, Ε. 2015. Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων. [ηλεκτρ. βιβλ.] Αθήνα:Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. κεφ 10. Διαθέσιμο στο: <http://hdl.handle.net/11419/1237>

[4] Γεωργούλη, Α. 2015. Μηχανική Μάθηση. [Κεφάλαιο Συγγράμματος]. Στο Γεωργούλη, Α. 2015. Τεχνητή νοημοσύνη. [ηλεκτρ. βιβλ.] Αθήνα:Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. κεφ 4. Διαθέσιμο στο: <http://hdl.handle.net/11419/3382>

[5] Κύρκος, Ε. 2015. Προεπεξεργασία Δεδομένων. [Κεφάλαιο Συγγράμματος]. Στο Κύρκος, Ε. 2015. Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων. [ηλεκτρ. βιβλ.] Αθήνα:Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. κεφ 7. Διαθέσιμο στο: <http://hdl.handle.net/11419/1234>

[6] Κύρκος, Ε. 2015. Κατηγοριοποίηση. [Κεφάλαιο Συγγράμματος]. Στο Κύρκος, Ε. 2015. Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων. [ηλεκτρ. βιβλ.] Αθήνα:Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. κεφ 9. Διαθέσιμο στο: <http://hdl.handle.net/11419/1236>

[7] Κύρκος, Ε. 2015. Εναλλακτικές Μέθοδοι και ειδικά θέματα Κατηγοριοποίησης. [Κεφάλαιο Συγγράμματος]. Στο Κύρκος, Ε. 2015. Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων. [ηλεκτρ. βιβλ.] Αθήνα:Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. κεφ 10. Διαθέσιμο στο: <http://hdl.handle.net/11419/1237>

Διαδικτυακοί τόποι

https://repository.kallipos.gr/pdfviewer/web/viewer.html?file=/bitstream/11419/1226/3/Epixeirimatiki_Efyia_kai_exorixi_Dedomenon_pdf.pdf

<http://www.cs.waikato.ac.nz/~ml/weka>

https://repository.kallipos.gr/bitstream/11419/3382/1/02_chapter_04.pdf

<https://knowledgesociety.usal.es/sites/default/files/MANUAL%20INGLE%CC%81S.pdf>

<https://waikato.github.io/weka-wiki/documentation/>

https://repository.kallipos.gr/bitstream/11419/1233/2/Kef._6.pdf

[https://repository.kallipos.gr/bitstream/11419/1234/2/Kef. 7.pdf](https://repository.kallipos.gr/bitstream/11419/1234/2/Kef._7.pdf)
<https://slideplayer.gr/slide/2312082/>
[https://repository.kallipos.gr/bitstream/11419/1236/2/Kef. 9.pdf](https://repository.kallipos.gr/bitstream/11419/1236/2/Kef._9.pdf)
[https://repository.kallipos.gr/bitstream/11419/1239/2/Kef. 13.pdf](https://repository.kallipos.gr/bitstream/11419/1239/2/Kef._13.pdf)