**Σχολή Οικονομίας & Τεχνολογίας**

**Τμήμα Ψηφιακών Συστημάτων**

[**Εξόρυξη Γνώσης και Μηχανική Μάθησης**](https://eclass.uop.gr/courses/2652/)

**Περιγραφή Εργασίας:**

Συσταδοποίηση και Κατηγοριοποίηση με το WEKA. Στόχος της εργασίας είναι (α) η ομαδοποίηση των βάσει των χαρακτηριστικών και (β) η κατασκευή ενός μοντέλου κατηγοριοποίησης



**Μαυρίκος Παναγιώτης-Γεώργιος**

**Α.Μ 2026201900049**

**Ζήτημα Α**

1. **Να οριστεί το πλήθος των συστάδων βάσει του SSE τεκμηριώνοντας την επιλογή αυτή με την παρουσίαση κατάλληλου διαγράμματος (πχ. στο MS Office Excel/LibreOffice Calc).**

Αρχικά, ανοίγουμε το file με τα δεδομένα «dmdata\_c» με την επιλογή «open file» (1).

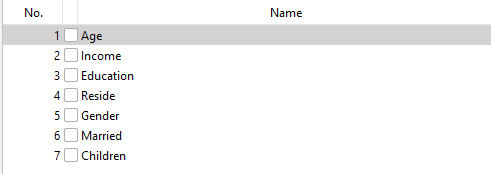
Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα



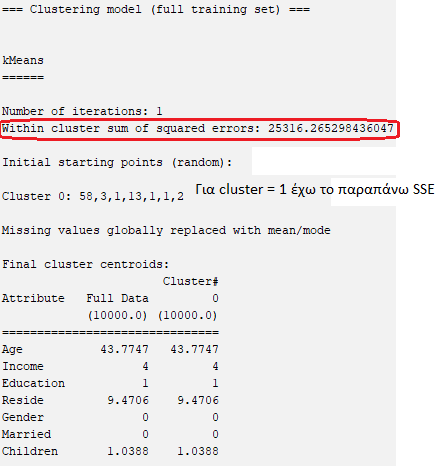
Εικόνα :Επιλογή Attributes και βασικές λειτουργίες

Έπειτα αφαιρούμε τα attributes -ID , responded και region επιλέγοντας το κατάλληλο κουτάκι και πατώντας remove (2) . Στην καρτέλα cluster (3) επιλέγουμε τον αλγόριθμο SimpleKmeans ,από εκεί έχουμε ακόμα την δυνατότητα να αλλάξουμε τον αριθμό των clusters.



Εικόνα :Κατaληλα Attributes

Για την επιλογή των κατάλληλων συστάδων χρησιμοποιούμε την μέθοδο του «αγκώνα». Για την μέθοδο αυτή αυξάνουμε κάθε φορά τον αριθμό τον clusters (+1) και παρατηρούμε το αντίστοιχο SSE.To σφάλμα άθροισμα τετραγώνων (SSE) είναι το άθροισμα της διαφοράς των τετραγώνων μεταξύ κάθε παρατήρησης και του μέσου όρου της ομάδας της.



Εικόνα :SSE για έναν cluster

Ακολουθώντας την παραπάνω λογική κατασκευάζουμε την παρακάτω γραφική παράσταση οπού στον κάθετο άξονα έχει τις τιμές sse και στο οριζόντιο τις τιμές των clusters.

Όπως φαίνεται από τα παραπάνω γράφημα οι καλύτερες συστάδες είναι 2 .(κ=2)

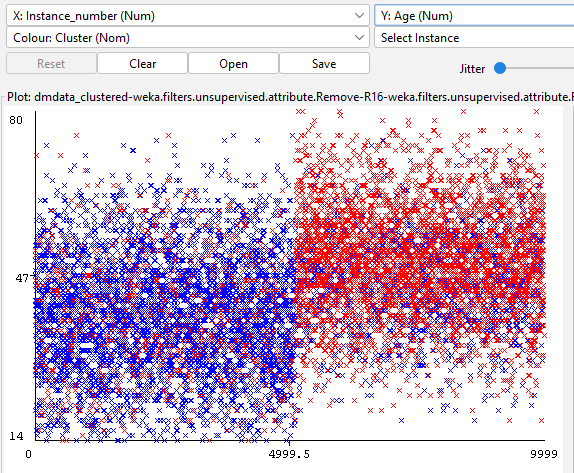
1. **Να περιγραφούν αναλυτικά οι συστάδες που προκύπτουν σκιαγραφώντας τα αντίστοιχα προφιλ των ομάδων πελατών παραθετοντας και τα αντίστοιχα στατιστικά στοιχεία.**

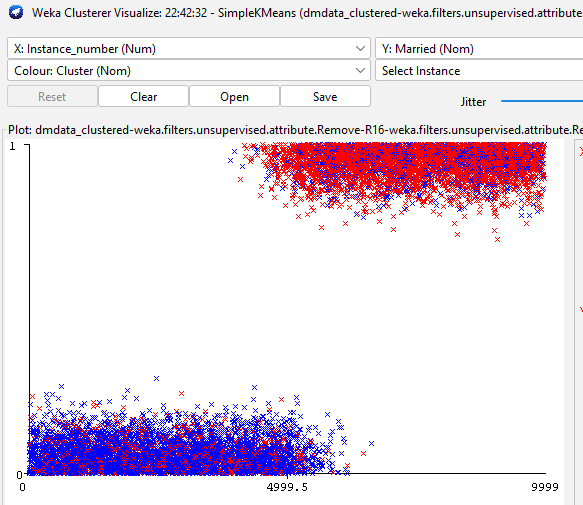
Για Κ=2 έχω το παρακάτω αποτέλεσμα :

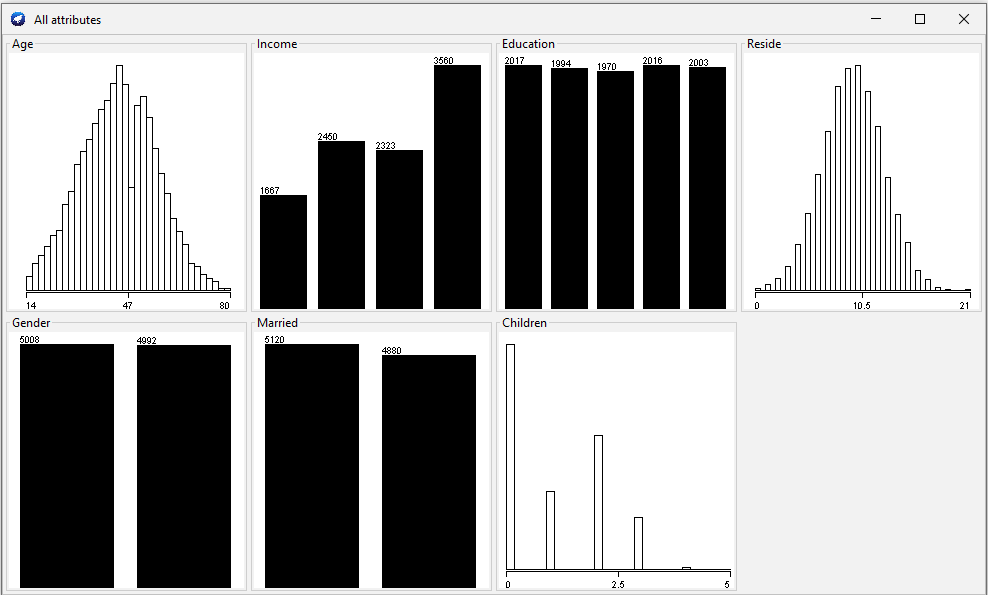
Εικόνα που περιέχει πίνακας

Περιγραφή που δημιουργήθηκε αυτόματα

Από την παραπάνω φωτογραφία παρατηρούμε τον αριθμό τον τιμών που καταλογίζονται σε κάθε Cluster ανάλογα τα Attribute.Στην παρακάτω εικόνα φαίνεται η κατανομή τον στοιχείων του Attribute age και married σε clusters.Ο cluster 0 απεικονίζεται με μπλε χρώμα ενώ ο cluster 1 με κόκκινο χρώμα .Παρατίθενται τα παρακάτω παραδείγματα:



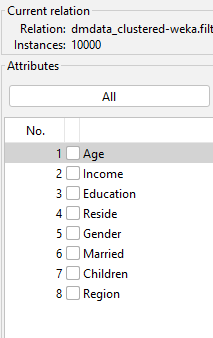




1. **Να αναφέρετε τον τρόπο με τον οποίο μετρώνται οι αποστάσεις κατά την εφαρμογή του αλγορίθμου.**

Ο αλγόριθμος k-means επιλεγεί τυχαία ένα αριθμό (στην συγκεκριμένη περίπτωση k=2) από τα κέντρα. Αρχικά για κάθε στιγμιότυπο υπολογίζει σε ποιο κέντρο είναι ποιο κοντά χρησιμοποιώντας την ευκλείδεια απόσταση ή την απόσταση Μανχάταν(στην συγκεκριμένη περίπτωση ο αλγόριθμος μας χρησιμοποίει για τον υπολογισμό της απόστασης την ευκλείδεια απόσταση).Έπειτα τα κέντρα μετακινούνται(προς τον μέσο ορό ) βρίσκοντας το μέσο ορό. Ο k-means επαναλαμβάνει την παραπάνω διαδικασία έως ότου δεν μετακινούνται τα κέντρα περεταίρω.

1. **Να συσχετίσετε διαγραμματικά τις συστάδες που θα προκύψουν με την περιοχή (region) που διαμένει ο κάθε πελάτης.**

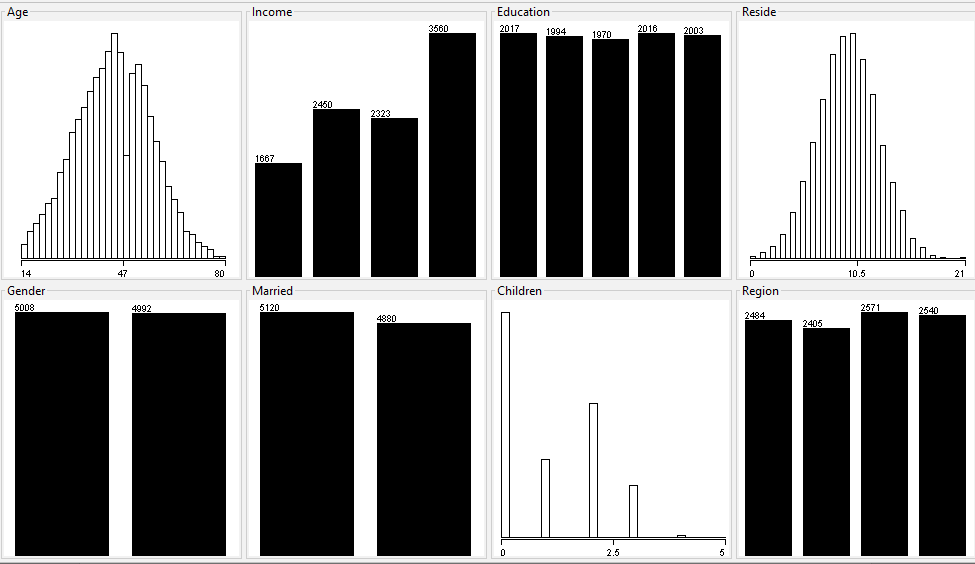


Τα Attributes που θα χρησιμοποιήσουμε

Οπότε εφαρμόζουμε το αλγόριθμο.

Εικόνα που περιέχει πίνακας

Περιγραφή που δημιουργήθηκε αυτόματα



**Ζήτημα Β**

Με βάση την 1η επισήμανση:

Κατά το άνοιγμα της σουίτας λογισμικού μηχανικής μάθησης WEKA επιλέγουμε Experimenter.

Εικόνα που περιέχει κείμενο, πουλί, υδρόβιο πουλί

Περιγραφή που δημιουργήθηκε αυτόματα

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, ουρανός

Περιγραφή που δημιουργήθηκε αυτόματαΗ επιλογή αυτή μας επιτρέπει να σχεδιάσουμε δικά μας πειράματα εκτέλεσης διαφορετικών αλγορίθμων σε διαφορετικά dataset και να συγκρίνουμε τα αποτελέσματα για το αν ένας αλγόριθμος είναι καλύτερος από τον άλλον και αν η διαφορά είναι στατιστικά σημαντική η όχι. Οπότε κλικ άρουμε στο κουμπί Experimenter και θα μας εμφανιστή το παρακάτω παράθυρο.

Αρχικά επιλέγουμε New (με κόκκινο χρώμα ) ώστε να δημιουργήσουμε νέο πείραμα και να μας ξεκλειδωθούν όλες οι επιλογές. Για να ορίσουμε το dataset που θα χρησιμοποιηθεί πατάμε Add new…(κάτω από κουτί Datasets δεξιά ) και επιλέγουμε dataset.Έπειτα (στο κουτί Algorithms αριστερά) επιλέγουμε τους αλγορίθμους που θέλουμε να χρησιμοποιήσουμε. Στην συγκεκριμένη περίπτωση οι αλγόριθμοι είναι οι εξης:

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα(α) δέντρο αποφάσεων (C4.5), (β) εγγύτεροι γείτονες (ibk), (γ) Naive Bayes

(στο κουτί με το μπλε χρώμα βάζουμε τον αριθμό των επαναλήψεων 1 )

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματαΑφού ολοκληρώσουμε τις παραπάνω ενέργειες πάμε στην δεύτερη καρτέλα Run και πατάμε START.

Τέλος πάμε στην καρτέλα Analuse

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματαΣτην καρτέλα Analyse επιλέγουμε Experiment ώστε να εμφανιστούν και οι υπόλοιπες επιλογές.

Ως comparison βάζουμε την επιλογή Percent\_correct , στο Rows και στο cols βάζουμε τις επιλογές dataset και scheme,scheme\_options,scheme\_version\_ID αντίστοιχα. Από το Actions (μπλε χρώμα) επιλέγω perform test.

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Mε βάση το ποσοστό επιτυχίας θα έχουμε τα παρακάτω αποτελέσματα:

(1) J48 75.71%

(2) Naïve Bayes 75%

(3) ibK 72.14%

Σημαντικά κριτήρια εκτός από την επιτυχία είναι η ευαισθησία (recall) και η ορθότητα (precision) της κλάσης που αφορά τα δάνεια που δεν εξυπηρετούνται.

Τώρα θα πρέπει να εφαρμόσουμε το F-measure. Για το κάνουμε αυτό θα πρέπει να αλλάξουμε το comparison σε F\_measure 

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Το f-measure υπολογίζεται με βάση την ακρίβεια και την ανάκληση. Ο υπολογισμός έχει ως εξής:

Ακρίβεια = t\_p / (t\_p + f\_p)

Ανάκληση = t\_p / (t\_p + f\_n)

F = 2 \* Ακρίβεια \* Ανάκληση / (Ακρίβεια + Ανάκληση)

Όπου t\_p είναι ο αριθμός των αληθινών θετικών, f\_p ο αριθμός των ψευδώς θετικών και f\_n ο αριθμός των ψευδώς αρνητικών.

Όπως παρατηρούμε από τα ποιο πάνω τεστ κατάταξη των αλγορίθμων είναι η εξής:

(1)J48

(2)NaïveBayes

(3) ibk

Όμως είναι πιθανόν κάποια γνωρίσματα του συνόλου δεδομένων να μην συνεισφέρουν στην ορθή πρόβλεψη και τελικά, μπορεί να βλάπτουν ένα ή περισσότερα από τα παραπάνω μέτρα αξιολόγησης. Για αυτό θα υπολογίσουμε το infogain.

Αρχικά στην καρτέλα preprocess βάζουμε το dataset και στην συνέχεια στην καρτέλα select attributes επιλέγουμε infogain και η gainratio και πατάμε start.

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Με βάση την 2η επισήμανση:

Εικόνα που περιέχει πίνακας

Περιγραφή που δημιουργήθηκε αυτόματαΕικόνα που περιέχει πίνακας

Περιγραφή που δημιουργήθηκε αυτόματα

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματαΑπό τα παραπάνω δεδομένα καταλήγουμε στο συμπέρασμα ότι το Attribute othdebt έχει εντροπία 0 οπότε μπορούμε να το διαγράψουμε. Αφού το αφαιρέσουμε θα πρέπει να ξανά συγκινούμε τους αλγορίθμους όπως παραπάνω.

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Με βάση το νέο dataset παρατηρούμε ότι η κατάταξη των αλγορίθμων είναι η εξής:

(1)naïve bayes

(2)j48

(3)ibk

Με βάση την 3η επισήμανση:

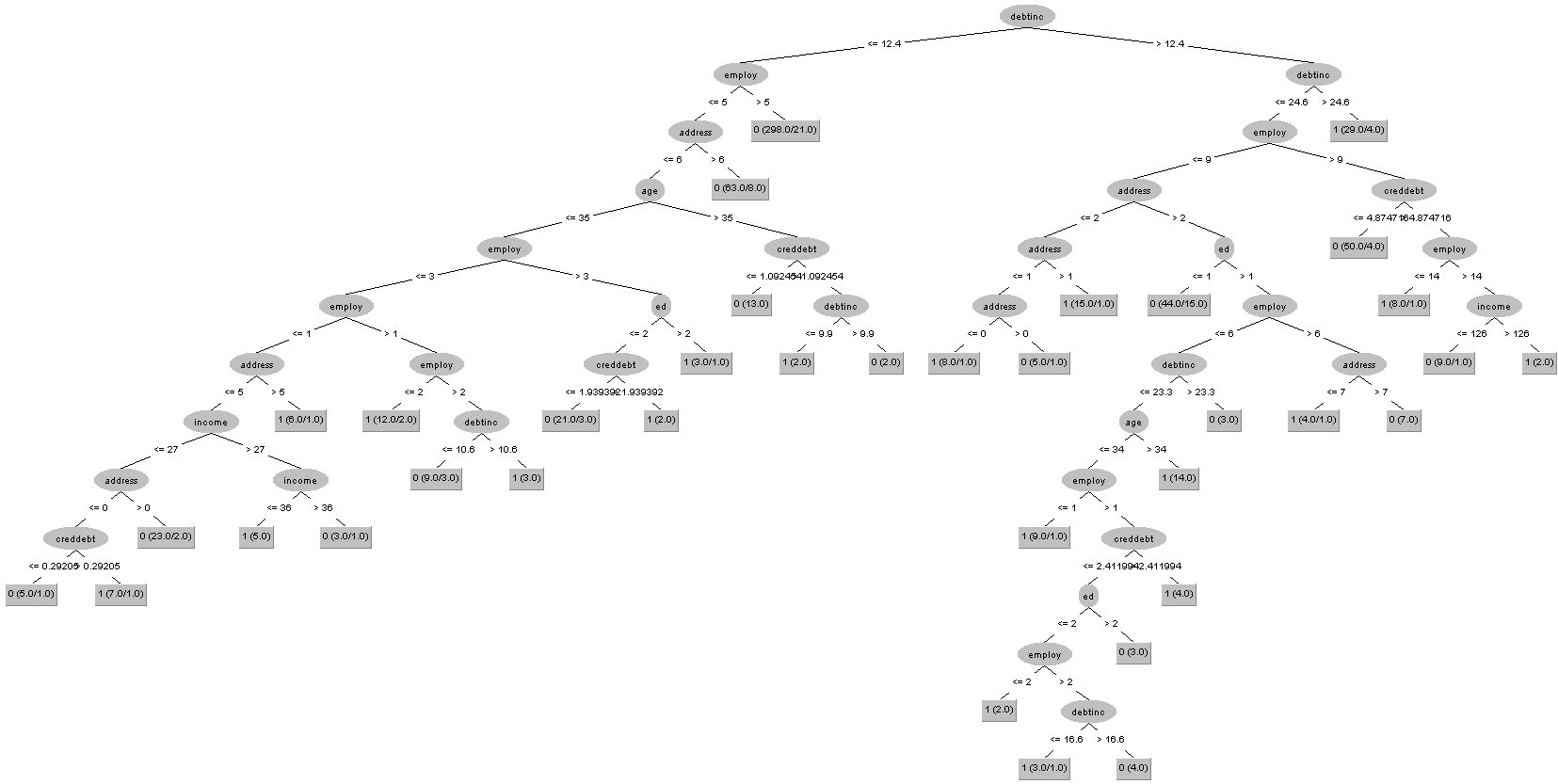
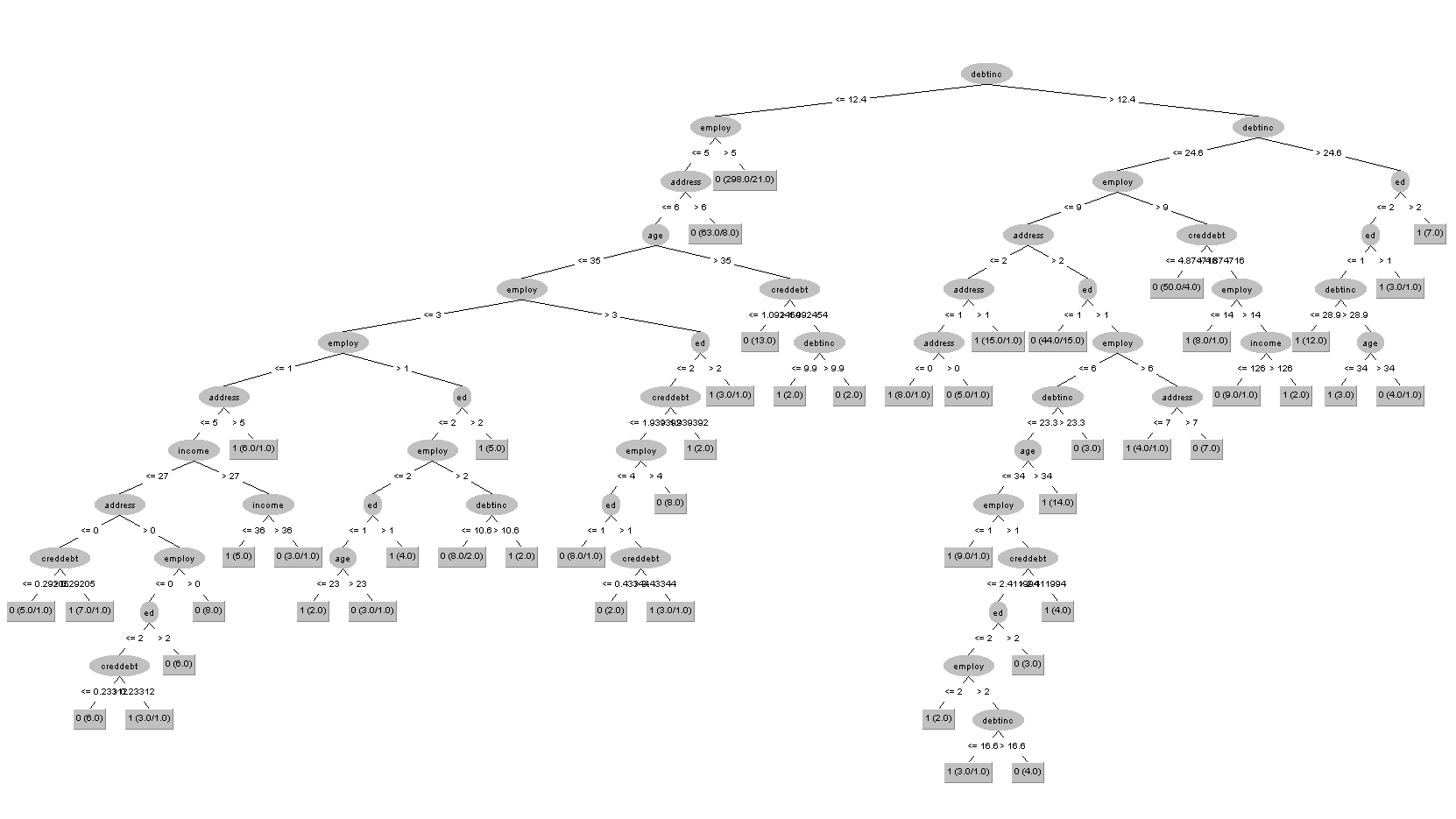
Το dataset που χρησιμοποιήθηκε για την κατασκευή των δέντρων δεν περιλαμβάνει το Attribute othdebt γιατί η εντροπία του είναι 0.

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματαΕικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Για minNumobj =2 kai unpruned=false

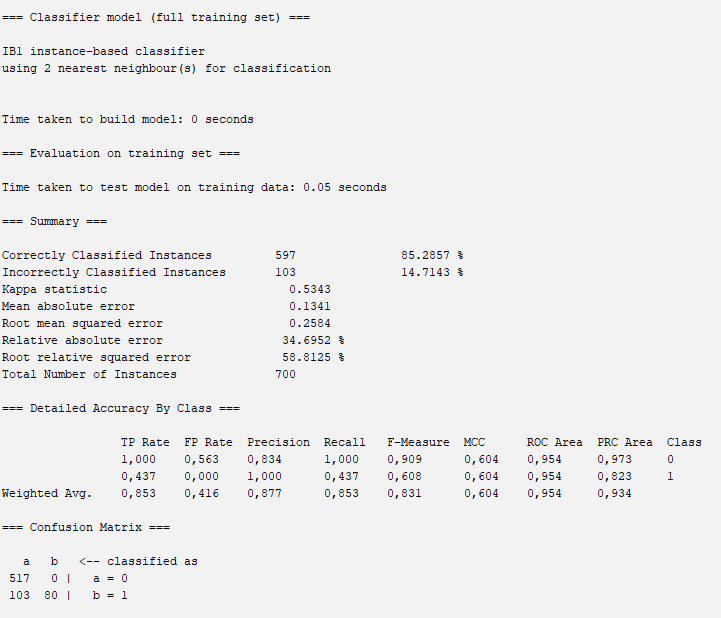
Για minNumobj =2 kai unpruned=true

Με βάση την 4η επισήμανση:

Για ΚΚΝ=2 και crossvalidate=true

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

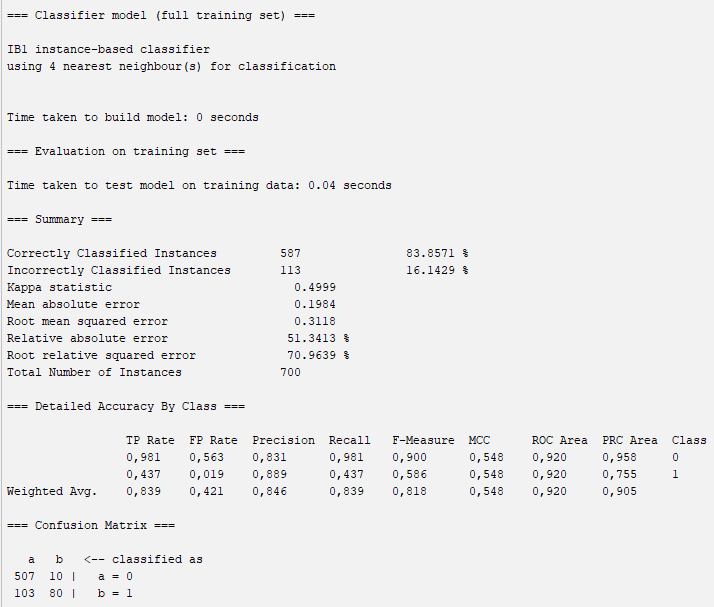


**85.2857% correct**

Για ΚΚΝ=4 και crossvalidate=true

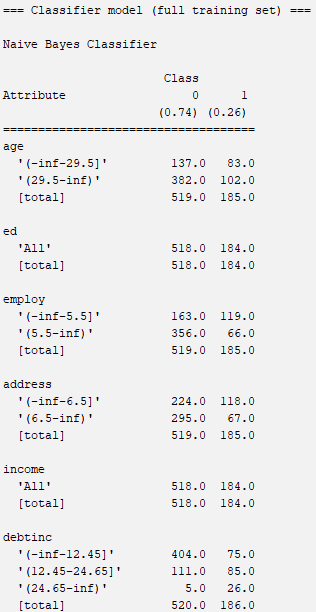
Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα



**83.8571% correct**

Με βάση την 5η επισήμανση:

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Εικόνα που περιέχει πίνακας

Περιγραφή που δημιουργήθηκε αυτόματα

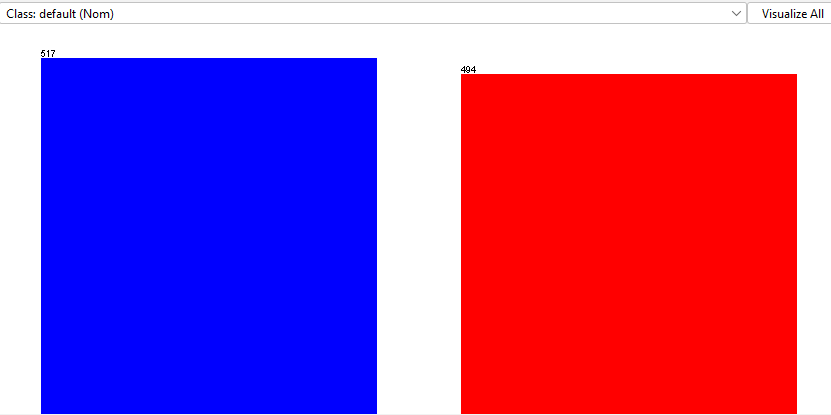
Με βάση την 6η επισήμανση:

Αρχικά εγκαθιστώ το SMOTE με βάση τις οδηγίες που αναφέρονται στο pdf.Έπειτα ανοίγω το dataset (χωρις Attribute othdebt ) και τους εφαρμόζω το φίλτρο ώστε να εξισορροπήσουν οι τιμές.

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα





Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματαΤώρα θα πρέπει να ξανασυγκρινουμε τους αλγορίθμους με το διπλά(χωρις Attribute othdebt, oversampling) τροποποιημένο dataset οπότε:

**Αρά ο καλύτερος αλγόριθμος είναι ο IBK!!!!!!!**

**Βιβλιογραφία**

* [**https://www.youtube.com/watch?v=WqEAD\_Z6h0E**](https://www.youtube.com/watch?v=WqEAD_Z6h0E)
* [**https://machinelearningmastery.com/compare-performance-machine-learning-algorithms-weka/**](https://machinelearningmastery.com/compare-performance-machine-learning-algorithms-weka/)
* [**https://stackoverflow.com/questions/21342449/what-is-f-measure-for-each-class-in-weka**](https://stackoverflow.com/questions/21342449/what-is-f-measure-for-each-class-in-weka)