



ΑΛΓΟΡΙΘΜΟΙ ΕΞΟΡΥΞΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΠΑΡΟΥΣΙΑΣΗ ΤΕΛΙΚΗΣ ΕΡΓΑΣΙΑΣ – Πανδή Αθηνά

- Εξερευνούμε οπτικά τα δεδομένα (εντολή visualize all)
- Αφαιρούμε από το train set τα attributes 36, 38 ,34 (remove)
- Αποθηκεύουμε το νέο train set με το όνομα train 11
- Κάνουμε δοκιμές τρέχοντας διάφορους αλγορίθμους στο train 11
- Οι αλγόριθμοι που επιλέγουμε πρέπει να είναι συμβατοί με numeric class attribute ώστε να μην χρειαστεί να μετατρέψουμε την μεταβλητή στόχο σε nominal
- Αν μετατρέψουμε την μεταβλητή στόχο χάνουμε πληροφορία

Μερικοί από τους αλγόριθμους που εφαρμόστηκαν στο train 11

ΑΛΓΟΡΙΘΜΟΣ	ΣΥΝΤΕΛΕΣΤΗΣ ΣΥΣΧΕΤΙΣΗΣ	SUCCESS RATE
RegressionByDiscretization->Bagging-> Bagging->PART	0.9251	0.955
RegressionByDiscretization-> AdaBoostM1->PART	0.9148	0.941
RegressionByDiscretization-> AdaBoostM1->Bagging->PART	0.912	0.955
RegressionByDiscretization->Bagging-> J48	0.9123	0.953
RandomSubSpace->RegressionByDiscretization-> J48Graft	0.9198	0.955

- 
- Επιλέγουμε για υποβολή στο test set τον αλγόριθμο

RandomSubSpace->RegressionByDiscretization-> J48Graft

- Έδωσε πολύ καλό συντελεστή συσχέτισης (0,9198)
- Έδωσε το μεγαλύτερο success rate (0.955) από τους 51 συνδυασμούς διαφορετικών αλγορίθμων που επιχειρήθηκαν στο quiz set
- Δεν προτιμήθηκαν οι άλλοι δύο αλγόριθμοι με success rate 0,955
 - Παρότι ο πρώτος είχε μεγαλύτερο συντελεστή συσχέτισης (0,9251)
 - Συνδύαζαν 3 meta αλγορίθμους
- **Σκοπός:** Αποφυγή overtraining
- Δεν προτιμήθηκε το φίλτρο attribute selection
 - Χαμηλά success rates
 - Ο αλγόριθμος 5 έδωσε 0,902 έναντι 0,955