

ΑΛΓΟΡΙΘΜΟΙ ΕΞΟΡΥΞΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΤΕΛΙΚΗ ΕΡΓΑΣΙΑ

ΑΝΔΡΟΥΛΑΚΗΣ ΜΑΝΟΣ

ΣΤΑΔΙΑ ΠΟΥ ΑΚΟΛΟΥΘΗΣΑΜΕ

- Προεπισκόπηση-προεπεξεργασία δεδομένων
 - ✓ Αφαίρεση από το training set των μεταβλητών 34, 36, και 38
 - ✓ Αποθήκευση του καινούργιου αρχείου ως train1
 - ✓ Έλεγχος με Linear Regression αν καλώς πράξαμε
- Εφαρμογή διαφόρων αλγορίθμων στα training και quiz sets
 - ✓ Αρχικά αφαίρεση των ίδιων attributes στο quiz, και πρόσθεση ενός νέου attribute 'prediction' ώστε να είναι συμβατά τα δύο αρχεία
 - ✓ Αποθήκευση του καινούργιου αρχείου ως quizextended1

- ✓ Έλεγχος συμβατότητας των αλγορίθμων με το class attribute
- ✓ Εφαρμογή κάθε αλγορίθμου πρώτα στο train1 και έπειτα στο quizextended1
- ✓ Σύγκριση των αποτελεσμάτων → Correlation coefficients και σφάλματα στο train1 καθώς και success rates από το quizextended1

- Καταλήξαμε στη χρησιμοποίηση του αλγορίθμου
RandomSubSpace → RegressionByDescretization → J48graft

- Εν συνέχεια, εφαρμόσαμε τον αλγόριθμο στο test set
 - Να σημειωθεί, ότι αρχικά προστέθηκε και σε αυτό ένα καινούργιο attribute 'prediction' και αφαιρέθηκαν και οι μεταβλητές 34, 36, 38

- Πρόβλημα: το ενδεχόμενο ύπαρξης overtraining

- Ακολουθεί πίνακας με τους τέσσερις καλύτερους αλγόριθμους

Σύγκριση αλγορίθμων βάσει του success rate από το quizextended1

Αλγόριθμος	Success rate
EnsembleSelection→ Backward Elimination	0,948
Bagging→RegressionByDescretization →J48	0,953
RandomSubSpace→RegressionBy Descretization→J48	0,953
RandomSubSpace→RegressionBy Descretization→J48graft	0,955

Η παραπάνω επιλογή έγινε από ένα σύνολο 33 υποψήφιων αλγορίθμων