Τεχνικές Εξόρυξης Δεδομένων
Εαρινό Εξάμηνο 2016-2017
Εργασία 2

ΚΟΡΟΜΠΙΛΗΣ ΗΛΙΑΣ 1115201200070
ΛΑΥΤΣΗΣ-ΣΚΡΕΤΑΣ ΠΑΡΙΣΗΣ 1115201000167

Στο αρχείο *graphs.py* υλοποιείται με την βοήθεια της βιβλιοθήκης *seaborn* η δημιουργία ενός histogram ή box plot για κάθε feature του αρχείου *train.tsv*. Τα παραγόμενα αρχεία είναι σε μορφή SVG. Παρατηρώντας τα histograms, εκτιμάμε ότι τα categorical features που θα παίξουν σημαντικό ρόλο για την κατηγοριοποίηση των πελατών, είναι αυτά που κάθε κατηγορίας του συγκεντρώνει διαφορετικό αριθμό τιμών, και μάλιστα αυτές οι τιμές προέρχονται κυρίως από μία κατηγορία πελατών (Good ή Bad). Όσο αφορά τα box plots, φαίνεται ότι θα βοηθήσουν numerical features, στα οποία οι κάθε κατηγορία πελατών συγκεντρώνει διαφορετικό εύρος τιμών.

Στο αρχείο *classifications.py* υλοποιείται με την βοήθεια της βιβλιοθήκης *sklearn* η αξιολόγηση των μεθόδων Classification: Linear SVC, Random Forest και Multinomia Naive Bayes χρησιμοποιώντας 10-fold Cross Validation. Στο παραγόμενο αρχείο *testSet_Predictions.csv* γίνεται εμφανές για κάθε μέθοδο ο μέσος όρος από τα 10 folds της μετρικής accuracy. Συμπεραίνεται ότι η πιο αποδοτική μέθοδος Classification είναι η Random Forest.

Στο αρχείο *features.py* υλοποιείται με την βοήθεια της βιβλιοθήκης *sklearn* ο υπολογισμός του information gain για κάθε feature, ενώ στη συνέχεια παρουσιάζεται στο παραγόμενο plot η μεταβολή του accuracy κατά τη διάρκεια της αφαίρεσης των features ένα προς ένα. Παρατηρείται ότι όσο πιο μεγάλο είναι το information gain του feature που αφαιρείται τόσο χειρότερο γίνεται το accuracy.

Στο αρχείο *predictions.py* υλοποιείται με χρήση της μεθόδου Random Forest, που αποδείχθηκε η αποδοτικότερη, η πρόβλεψη της κατηγορίας κάθε πελάτη από το αρχείο *test.tsv*. Κατά την προ-επεξεργασία των δεδομένων, πέρα από τις στήλες Label και Id, αφαιρέθηκαν ακόμη οι πέντε στήλες των οποίων τα features παρουσίαζαν το μικρότερο information gain. Το παραγόμενο αρχείο είναι στα πρότυπα των απαιτήσεων της εκφώνησης.