ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ & ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ

ΕΡΓΑΣΙΑ 1: ΕΠΙΒΛΕΠΟΜΕΝΗ ΜΑΘΗΣΗ - ΤΑΞΙΝΟΜΗΣΗ

ΚΑΓΙΑΣ ΑΝΤΩΝΙΟΣ - aid23003

aid23003@uom.edu.gr

Πίνακας περιεχομένων

Εισαγωγή

Το πρόβλημά μας έχει να κάνει με τον εντοπισμό εταιρειών που θα πτωχεύσουν μετά από επεξεργασία ενός αρχείου excel που περιλαμβάνει διάφορα στοιχεία εταιρειών. Ο εντοπισμός αυτός θα γίνει με τη χρήση μοντέλων επιβλεπόμενης μάθησης και συγκεκριμένα με τα επτά παρακάτω μοντέλα:

- Linear Discriminant Analysis
- Logistic Regression
- Decision Trees
- k-Nearest Neighbors
- Naïve Bayes
- Support Vector Machines
- Neural Networks

Οι περιορισμοί του προβλήματος είναι δύο. Πρώτον, το μοντέλο πρέπει να βρίσκει με ποσοστό επιτυχίας τουλάχιστον 62% τις εταιρείες που θα πτωχεύσουν και δεύτερον να βρίσκει με ποσοστό τουλάχιστον 70% τις εταιρείες που δεν θα πτωχεύσουν.

Τα αποτελέσματα των πειραμάτων θα περαστούν στη συνέχεια σε ένα αρχείο excel το οποίο αποτελεί και παραδοτέο. Παραδοτέο αποτελεί επίσης και το αρχείο .py που περιλαμβάνει όλο τον κώδικα ο οποίος συνοδεύεται με τα απαραίτητα σχόλια καθώς το Google Colaboratory αρχείο με κατάληξη .ipynb για καλύτερη οπτικοποίηση του κώδικα και των αποτελεσμάτων.

Μέθοδοι που εφαρμόστηκαν

Αφού φορτώσουμε το dataset και δούμε με την εντολή display τα δεδομένα μας, κάνουμε την απαραίτητη κανονικοποίηση και χωρίζουμε το dataset σε inputData και outputData. Το πρώτο **δεν** περιλαμβάνει τη στήλη με την πληροφορία για πτώχευση ή όχι της εταιρείας ενώ το δεύτερο την περιλαμβάνει. Στη συνέχεια, τα παραπάνω χωρίζονται σε train και test sets με αναλογία 80/20.

Ακολουθεί η εφαρμογή των μοντέλων επιβλεπόμενης μάθησης και η χρήση των εκπαιδευμένων ταξινομητών και μετά ο υπολογισμός των μετρικών accuracy, precision, recall και F1 score για train και test sets. Έπειτα, επειδή είναι απαραίτητο για το excel με τα αποτελέσματα που θα δημιουργήσουμε, θα υπολογίσουμε τους confusion matrices για κάθε μοντέλο που αναπτύχθηκε έτσι ώστε να αντλήσουμε τα true positives, false positives, true negatives και false negatives.

Συμπεράσματα

Σύμφωνα και με το excel που έχουμε δημιουργήσει, φαίνεται πως το μοντέλο Naive Bayes έχει τις χειρότερες τιμές στις μετρικές που υπολογίστηκαν ανάμεσα στα άλλα μοντέλα (ο χρωματισμός των τιμών στο excel βοηθάει στην οπτικοποίηση του παραπάνω ισχυρισμού). Από εκεί και πέρα, φυσιολογικό είναι να παρατηρούνται γενικά καλύτερες τιμές στις μετρικές για τα train sets στα οποία ξεχωρίζουν τα μοντέλα Decision Trees και Support Vector Machines ενώ το μοντέλο των k-Nearest Neighbors αποτελεί το “best of the rest”.

Όσον αφορά τα test sets, στην καλύτερη θέση και μάλιστα με ίδια σκορ βρίσκονται τα μοντέλα Logistic Regression και Support Vector Machines. Όπως και πριν, το επόμενο καλύτερο μοντέλο με διαφορά από τα υπόλοιπα αποτελεί το k-Nearest Neighbors. Ενδιαφέρον είναι το γεγονός πως παρά την πολύ καλή απόδοση στα train sets, τα Decision Trees αποδίδουν χειρότερα από οποιοδήποτε άλλο μοντέλο στα test sets.

Εξετάζοντας τους αρχικούς περιορισμούς του προβλήματος, βλέπουμε πως ενώ όλα τα μοντέλα πετυχαίνουν ποσοστά επιτυχίας στην αναγνώριση των εταιρειών που δε θα πτωχεύσουν άνω του 70% που ζητείται (και μάλιστα το μικρότερο ποσοστό ήταν 95%), αυτό δεν ισχύει για την αναγνώριση των εταιρειών που θα πτωχεύσουν καθώς το υψηλότερο ποσοστό που παρατηρείται είναι 21%, δηλαδή κατά πολύ χαμηλότερο του ορίου του 62% που θα θέλαμε. Αυτό μπορεί να οφείλεται στον πολύ μικρό αριθμό των εταιρειών που κήρυξαν πτώχευση ανάμεσα σε αυτές που εξετάζονται καθώς σε test set 2144 εταιρειών, αυτές που χρεωκόπησαν ήταν μόλις 57, δηλαδή το 2,65%.