ΕΡΓΑΣΙΑ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ (2017-2018)

Η εργασία θα αξιοποιήσει επεξεργασμένα δεδομένα που έχουν συγκεντρωθεί από την εταιρία EMIS που παρέχει πληροφορίες για επιχειρήσεις σε ανερχόμενες αγορές. Τα δεδομένα αφορούν πάνω από 10000 επιχειρήσεις οι οποίες παρακολουθούνται από το 2000 και μετά.

Στο αρχείο εκπαίδευσης που θα σας δοθεί υπάρχουν οικονομικά στοιχεία για την πορεία 7000 περίπου επιχειρήσεων για την περίοδο 2000-2007. Συνολικά το αρχείο περιέχει 64 γνωρίσματα τα οποία αποτυπώνονται στον πίνακα:

X1 net profit / total assets X2 total liabilities / total assets X3 working capital / total assets X4 current assets / short-term liabilities X5 [(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] * 365 X6 retained earnings / total assets X7 EBIT / total assets X8 book value of equity / total liabilities X9 sales / total assets X10 equity / total assets X11 (gross profit + extraordinary items + financial expenses) / total assets X12 gross profit / short-term liabilities X13 (gross profit + depreciation) / sales X14 (gross profit + interest) / total assets X15 (total liabilities * 365) / (gross profit + depreciation) X16 (gross profit + depreciation) / total liabilities X17 total assets / total liabilities X18 gross profit / total assets X19 gross profit / sales X20 (inventory * 365) / sales X21 sales (n) / sales (n-1)	X22 profit on operating activities / total assets X23 net profit / sales X24 gross profit (in 3 years) / total assets X25 (equity - share capital) / total assets X26 (net profit + depreciation) / total liabilities X27 profit on operating activities / financial expenses X28 working capital / fixed assets X29 logarithm of total assets X30 (total liabilities - cash) / sales X31 (gross profit + interest) / sales X32 (current liabilities * 365) / cost of products sold X33 operating expenses / short-term liabilities X34 operating expenses / total liabilities X35 profit on sales / total assets X36 total sales / total assets X37 (current assets - inventories) / long-term liabilities X38 constant capital / total assets X39 profit on sales / sales X40 (current assets - inventory - receivables) / short-term liabilities X41 total liabilities / ((profit on operating activities + depreciation) * (12/365)) X42 profit on operating activities / sales	X43 rotation receivables + inventory turnover in days X44 (receivables * 365) / sales X45 net profit / inventory X46 (current assets - inventory) / short-term liabilities X47 (inventory * 365) / cost of products sold X48 EBITDA (profit on operating activities - depreciation) / total assets X49 EBITDA (profit on operating activities - depreciation) / sales X50 current assets / total liabilities X51 short-term liabilities / total assets X52 (short-term liabilities * 365) / cost of products sold) X53 equity / fixed assets X54 constant capital / fixed assets X55 working capital X56 (sales - cost of products sold) / sales X57 (current assets - inventory - short-term liabilities) / (sales - gross profit - depreciation) X58 total costs /total sales X59 long-term liabilities / equity X60 sales / inventory X61 sales / receivables X62 (short-term liabilities *365) / sales X63 sales / short-term liabilities X64 sales / fixed assets

Υπάρχει επίσης το γνώρισμα κλάσης σχετικά με το αν οι επιχειρήσεις αυτές πτώχευσαν στα επόμενα 5 έτη μετά την περίοδο παρακολούθησης, η οποία αποτυπώνεται με την τιμή 1 στην τελευταία στήλη κάθε γραμμής.

Το αρχείο περιέχει float τιμές στα 64 γνωρίσματα, ενώ έχει και ελλιπείς τιμές που συμβολίζονται με ?.

Δεδομένα

Τα δεδομένα εκπαίδευσης σας δίνονται στο αρχείο bankruptcy.zip που θα κατεβάσετε από το e-class.

Τα άγνωστα δεδομένα θα σας δοθούν μέσα στο Δεκέμβριο, οπότε και πρέπει να έχετε ολοκληρώσει το κομμάτι της εκπαίδευσης του μοντέλου σας. Αφού εκπαιδεύσετε και αξιολογήσετε το μοντέλο σας στα δεδομένα εκπαίδευσης θα σας δοθεί το σύνολο άγνωστων δεδομένων (θα αφορούν εταιρίες με τα ίδια γνωρίσματα και αντίστοιχη κατανομή σε πτωχευμένες και μη) και για το οποίο θα πρέπει να δώσετε την πρόβλεψή σας. Η αξιολόγηση θα ανακοινωθεί μαζί με τη βαθμολογία.

ΕΡΓΑΣΙΕΣ

Α) Προετοιμασία (10%)

Τα αρχεία που σας δίνονται είναι σε μορφή csv και arff οπότε μπορείτε να χρησιμοποιήσετε Scikit Learn ή Weka ή όποια άλλη πλατφόρμα επιθυμείτε και να κάνετε τις απαραίτητες μετατροπές σε τύπους δεδομένων (π.χ. numeric σε nominal), και να απορρίψετε γνωρίσματα που αποφασίζετε ότι δε χρειάζεστε.

Β) Κατηγοριοποίηση (60-70%)

Το μοντέλο που θα εκπαιδεύσετε θα πρέπει να μπορεί να κατηγοριοποιεί κάθε εταιρία σε ένα από τους 2 τύπους (1-χρεωκοπία ή 0-όχι χρεωκοπία). Θα πρέπει να δοκιμάσετε αλγορίθμους κατηγοριοποίησης της επιλογής σας με στόχο να έχετε όσο το δυνατό καλύτερες επιδόσεις στην κατηγοριοποίηση του ίδιου του συνόλου εκπαίδευσης.

Βεβαιωθείτε ότι έχετε αποφύγει να υπερεκπαιδεύσετε το μοντέλο σας.

Θα πρέπει να δοκιμάστε τον καλύτερό σας αλγόριθμο στα άγνωστα δεδομένα ελέγχου για τα οποία δεν έχετε ετικέτες. Θα πρέπει να παράγετε ένα αρχείο που να περιέχει την ετικέτα που προβλέψατε για κάθε εταιρία σε ξεχωριστή γραμμή. Οι απαντήσεις σας θα συγκριθούν με τις σωστές απαντήσεις και θα ανακοινωθούν τα αποτελέσματα για κάθε ομάδα. Οι δύο εργασίες που θα πετύχουν το υψηλότερο F-measure στα άγνωστα δεδομένα θα έχουν +10% στον τελικό βαθμό.

Γ) Αξιολόγηση Γνωρισμάτων - Παλινδρόμηση (30%)

Εφαρμόζοντας τεχνικές συσχέτισης και αξιολόγησης γνωρισμάτων καταλλήξτε σε υποσύνολα γνωρισμάτων που ενδέχεται να προβλέπουν καλύτερα την κλάση στόχο. Μπορείτε να κ αταλ λ ξετε σε μια κατάταξη των επιχειρήσεων σε σχέση με τον κίνδυνο να χρεωκοπήσουν στα επόμενα 5 χρόνια;

Σχετική βιβλιογραφία

- Zieba, M., Tomczak, S. K., & Tomczak, J. M. (2016). Ensemble Boosted Trees with Synthetic Features Generation in Application to Bankruptcy Prediction. Expert Systems with Applications.

ΠΑΡΑΔΟΣΗ
Η ημερομηνία παράδοσης είναι στις 15/1/2018 στο eclass.

1) Στην αναφορά που θα παραδώσετε θα πρέπει να αναλύσετε τη διαδικασία που ακολουθήσατε σε κάθε εργασία (προετοιμασία, εκπαίδευση ταξινομητή, αξιολόγηση ταξινομητή σε γνωστά και άγνωστα δείγματα) και αφορά: τους μετασχηματισμούς που κάνατε στο σύνολο δεδομένων, τον αλγόριθμο που χρησιμοποιήσατε, τις παραμέτρους που δοκιμάσατε και πως καταλήξατε σε αυτές, τις επιδόσεις που είχατε στα δεδομένα εκπαίδευσης αλλά και στα δεδομένα ελέγχου κλπ.
2) Θα πρέπει επίσης να δώσετε τις προβλέψεις του μοντέλου σας για τα άγνωστα δείγματα που θα σας δοθούν μέσα στο Δεκέμβριο. Αυτές θα είναι σε ένα αρχείο που θα έχει τόσες γραμμές όσα τα άγνωστα δείγματα που θα σας δοθούν και σε κάθε γραμμή θα έχει μόνο την τιμή που προβλέψατε για την κλάση.
3) Θα πρέπει να βρείτε, και να εξηγήσετε πως, ένα υποσύνολο 10 το πολύ γνωρισμάτων και να συγκρίνετε την απόδοσή του στην κατηγοριοποίηση με τον καλύτερό σας αλγόριθμο.
4) Θα πρέπει να δώσετε τις 50 εταιρίες που φαίνεται πιθανότερο να χρεωκοπήσουν στο άγνωστο dataset. Για το σκοπό αυτό χρησιμοποιήστε το rowid στο αρχείου που θα σας δοθεί (η αρίθμηση από το 1).

Οι εργασίες θα βαθμολογηθούν:
ι) για την προετοιμασία: η περιγραφή των ενεργειών (10%)
ιι) για την κατηγοριοποίηση: α) η περιγραφή των ενεργειών (30%), β) οι επιδόσεις που είχατε στα δεδομένα εκπαίδευσης (20%), γ) η επίδοση του αλγορίθμου σας στα άγνωστα δεδομένα (10% - 20%).
ιιι) για τη μεθοδολογία που ακολουθήσατε για την πρόβλεψη της κατάταξης και του υποσυνόλου των γνωρισμάτων (30%)

Η εργασία είναι για 3 (το πολύ) άτομα.