Μηχανική Μάθηση

1η Εργασία – Classification problems

Το πρόβλημα που θα αντιμετωπίσετε αφορά στον εντοπισμό εταιρειών που θα κηρύξουν χρεωκοπία. Τα δεδομένα εισόδου που θα χρησιμοποιήσετε παρέχονται έτοιμα από κάποιον οργανισμό.

Σκοπός της εργασίας είναι να παραδώσετε μια αναφορά η οποία παραθέτει αναλυτικά συγκριτικά αποτελέσματα σχετικά με την ικανότητα διαφορετικών τεχνικών ταξινόμησης να ανταπεξέλθουν στο συγκεκριμένο πρόβλημα.

Προσοχή: θεωρείστε ότι το συγκεκριμένο πρόβλημα σας έχει δοθεί στα πλαίσια αξιολόγησης για μια θέση στην οποία είστε υποψήφιος(ια). Η αναφορά πρέπει να είναι προσεκτικά δομημένη και τα αποτελέσματα που παρουσιάζετε να είναι ερμηνεύσιμα και κατανοητά.

Τα δεδομένα που θα χρησιμοποιήσετε βρίσκονται στο αρχείο: "Dataset2Use_Assignment1.xlsx". Κάθε γραμμή αφορά ξεχωριστή εταιρία. Συνοπτικά, έχετε διαθέσιμο ένα μεγάλο πλήθος διαφορετικών τιμών:

1. Οι δείκτες απόδοσης των εταιρειών (στήλες A έως και H)
2. Τρεις δυαδικοί δείκτες δραστηριοτήτων (στήλες I, J, K)
3. Η κατάσταση της εταιρείας (1 όλα καλά, 2 έχει κηρύξει χρεωκοπία)
4. Το έτος στο οποίο αφορούν τα ως άνω μεγέθη.

Στην άσκηση θα παραδώσετε: α) κώδικα και β) αναφορά. Ακολουθεί περιγραφή για το τι πρέπει να υλοποιήσετε.

Ο κώδικας που θα παραδώσετε, πρέπει να υλοποιεί τα ακόλουθα:

1. Να διαβάζει τα δεδομένα από το αρχείο excel.

   Παρατηρήσεις: α) η βιβλιοθήκη Pandas θα σας φανεί χρήσιμη. β) θα χρειαστεί να ανεβάσετε το excel στο drive σας και να το προσπελάσετε από εκεί.

2. Να τυπώνει, σε γραφήματα τα ακόλουθα στοιχεία:
   a. Figure 1: Αριθμό υγιών και χρεωκοπημένων επιχειρήσεων, για κάθε έτος.
   b. Figure 2: Την min, max, average τιμή για κάθε δείκτη. Προσοχή: εδώ θα έχετε 2 subfigures. Ένα θα έχει τις τιμές υγειών εταιρειών και ένα διαφορετικό τις τιμές για τις πτωχευμένες.
3. Να ελέγχει για τυχόν ελλείπεις εγγραφές [π.χ. NaNs] και να ειδοποιεί τον χρήστη με σχετικό μήνυμα.
4. Να κανονικοποιεί τα δεδομένα στο διάστημα $[0,1]$ με χρήση τεχνικής τύπου map minmax.

5. Να κάνει χρήση του Stratified kfold ώστε να δημιουργεί 4 folds.
6. Να τυπώνει στην οθόνη, για κάθε fold, πόσες χρεωκοπημένες και πόσες υγιείς εταιρείες υπάρχουν στο a) train και στο b) test set.
7. Αν η κατανομή είναι πάνω από 3 υγιείς επιχειρήσεις για κάθε χρεωκοπημένη, διαλέξτε με τυχαίο τρόπο όσες υγιείς εταιρείες χρειαστεί, ώστε η αναλογία στο training set να είναι 3 υγιείς / 1 χρεωκοπημένη.

Παρατήρηση: Οι εταιρείες που βγήκαν από το train set θα μεταφερθούν στο test set. Τυπώστε εκ νέου τις κατανομές στα train/test sets.

8. Να υλοποιεί (εκπαιδεύει) τα ακόλουθα μοντέλα:
    a. Linear Discriminant Analysis
    b. Logistic Regression
    c. Decision Trees
    d. Random Forests
    e. k-Nearest Neighbors
    f. Naïve Bayes
    g. Support Vector Machines
    h. Ένα επιπλέον μοντέλο της επιλογής σας (περιγράψτε ποιο)
9. Για κάθε εκπαιδευμένο μοντέλο:
    a. Να τυπώνει τα confusion matrices, ως figures, τόσο στο train όσο και στο test set.
    b. Θα υπολογίζει την επίδοση τόσο στο train όσο και στο test set, με βάση τις ακόλουθες μετρικές: Accuracy, Precision, Recall, F1 score, AUC ROC.
    c. Να τυπώνει τα αποτελέσματα στην οθόνη με ακρίβεια 2 δεκαδικών ψηφίων.
    d. Να καταχωρεί σε ένα dataframe (Pandas) σε μια νέα γραμμή τις ακόλουθες πληροφορίες:
        i. Classifier Name (str)
        ii. Training or test set (str)
        iii. Balanced or unbalanced train set (str)
        iv. Number of training samples (int)
        v. Number of non-healthy companies in training sample (int)
        vi. True positives TP (int)
        vii. True negatives TN (int)
        viii. False positives FP (int)
        ix. False negatives FN (int)
        x. ROC-AUC (double)

    Προσοχή: τα βήματα 6 μέχρι 9 θα γίνουν μέσα σε loop, μιας και έχετε πολλά folds.

10. Όταν ολοκληρωθεί η εκτέλεση τα αποτελέσματα θα αποθηκεύουν σε ένα αρχείο csv με όνομα balancedDataOutcomes.csv

Όταν τελειώσει η εκτέλεση και έχετε το αρχείο csv με όλα τα αποτελέσματα, προχωρήστε ως εξής:

1. Μετατρέψτε το csv σε αρχείο excel (.xls ή .xlsx).
2. Μεταβείτε στο https://en.wikipedia.org/wiki/Confusion_matrix. Χρησιμοποιείστε τους τύπους που δίνονται στο site και προσθέστε, στο αρχείο excel που έχετε, τις ακόλουθες εγγραφές (στήλες):
    a. Accuracy
    b. Precision
    c. Recall
    d. F1 score
    e. Μετρική 1
    f. Μετρική 2

Παρατήρηση: Οι μετρικές 1 και 2 είναι δική σας επιλογή.

3. Χρησιμοποιώντας το αρχείο excel, φτιάξτε γραφικές παραστάσεις που να δείχνουν πιο είναι το καλύτερο μοντέλο με βάση το F1 score.

Παρατήρηση: τα pivot tables (στο excel) θα σας φανούν πολύ χρήσιμα για την δημιουργία πινάκων και γραφικών παραστάσεων.

Το σύνολο των αποτελεσμάτων θα τα συγκεντρώσετε και θα τα υποβάλετε σε μορφή αναφοράς. Πληροφορίες για την δομή της αναφοράς θα βρείτε στην ενότητα «Οδηγίες».

Η αναφορά πρέπει να είναι δομημένη με τέτοιο τρόπο ώστε να απαντά, με σαφή τρόπο,

1. ποιο είναι το καλύτερο δυνατό μοντέλο ταξινόμησης.
2. Εάν πρέπει να πληρούνται οι ακόλουθοι δύο περιορισμοί στην απόδοση:
    a. Το μοντέλο πρέπει να βρίσκει με ποσοστό επιτυχίας τουλάχιστον **60%** τις εταιρείες που **θα πτωχεύσουν**.
    b. Το μοντέλο πρέπει να βρίσκει με ποσοστό επιτυχίας τουλάχιστον **70%** τις εταιρείες που **δεν θα πτωχεύσουν**.

    Υπάρχει κάποιο μοντέλο που να πληροί τις προϋποθέσεις;

Σημαντική παρατήρηση: εργασία που *δεν* συνοδεύεται από γραπτή αναφορά βαθμολογείται με μηδέν (0).

Καταληκτική Ημερομηνία Παράδοσης: **Παρασκευή 22 Δεκεμβρίου 2023**

Οδηγίες:

Α. Οι εργασίες είναι σε ομάδες μέχρι τέσσερα (4) άτομα. Κάθε άτομο μπορεί να υποβάλει εργασία σε μία μόνο ομάδα κάθε φορά.

Β. Οι εργασίες θα πρέπει να αναρτώνται στο eClass σε ένα αρχείο zip (όχι rar) εντός της προβλεπόμενης προθεσμίας. **Προσοχή**: Κάθε ομαδική εργασία θα υποβάλλεται μόνο από ένα μέλος της ομάδας (εσείς επιλέγετε ποιος/ποια).

Γ. Κάθε εργασία πρέπει να συνοδεύεται από:

- Ένα και μόνο ένα αρχείο .py θα περιέχει τις απαντήσεις στα ερωτήματα
- Μια **αναφορά** σε pdf με τα ακόλουθα στοιχεία:
    - Εξώφυλλο: 1 σελίδα, περιλαμβάνει τα στοιχεία των φοιτητών της ομάδος, όνομα μαθήματος, ημερομηνία, τμήμα και λοιπά σχετικά στοιχεία.
    - Συγκεντρωτικός πίνακας περιεχομένων, εικόνων, και λοιπών γραφημάτων που παραθέτετε στην αναφορά.
    - Ενότητα εισαγωγή: 1 σελίδα, περιγράφετε το πρόβλημα (*χωρίς* να αντιγράψετε αυτούσια την εκφώνηση της άσκησης)
    - Θεωρητικό υπόβαθρο: Μέθοδοι που εφαρμόστηκαν, από μέχρι 3 σελίδες, περιγράφετε τις μεθόδους που χρησιμοποιήσατε, εξηγώντας ποιες τιμές παραμέτρων (hyperparameters) επηρεάζουν την απόδοση και τι τιμές επιλέξατε εσείς για αυτές
    - Πειραματικά αποτελέσματα: Παραθέτετε τα σχετικά αποτελέσματα, μέχρι 5 σελίδες. Τα αποτελέσματα πρέπει να περιλαμβάνουν πίνακες, γραφικές παραστάσεις και εικόνες. Φροντίστε να υπάρχουν λεζάντες (captions) κάτω από κάθε γράφημα. Κάθε γραφική παράσταση πρέπει να συνοδεύεται από κείμενο που σχολιάζει τα αποτελέσματα.
    - Συμπεράσματα: 1 σελίδα, με βάση τα αποτελέσματα τι προτείνετε, ποιο μοντέλο αποδίδει καλύτερα, τι θα μπορούσε να γίνει για περαιτέρω βελτίωση στην απόδοση.
    - Η αναφορά θα περιέχει γραφικές παραστάσεις κάθε είδους και πίνακες αξιολόγησης των αποτελεσμάτων που πρέπει να συνοδεύονται (έκαστο) από μια τουλάχιστον παράγραφο με σχολιασμό.

Φροντίστε ώστε:

- Ο κώδικας να συνοδεύεται απαραιτήτως από κατάλληλα σχόλια.
- Να έχει γίνει συντακτικός και ορθογραφικός έλεγχος στην αναφορά που θα υποβάλετε.
- Οι προτάσεις να είναι κατανοητές και μικρές σε έκταση.
- Οι εικόνες να ***μην*** έχουν προκύψει από print screen. Αν το πρόγραμμα δημιουργεί μια εικόνα αποθηκεύστε την κανονικά (jpg ή png), πριν την χρησιμοποιήσετε.
- Οι γραφικές παραστάσεις να περιλαμβάνουν ονόματα στους άξονες και λεζάντα. Σκοπός είναι να γίνεται κατανοητό τι δείχνει, με μια ματιά.
- Αν κάτι δεν διευκρινίζεται, έχετε το δικαίωμα να κάνετε όποια υλοποίηση σας βολεύει. Φροντίστε να μπορείτε να εξηγήσετε τι ακριβώς κάνατε στον κώδικα.
- Οι βιβλιοθήκες που θα χρησιμοποιήσετε *πρέπει* να μπορούν να εγκατασταθούν μέσω του pip.
- Ο κώδικας *πρέπει* να τρέχει σε Google Colab.