Μηχανική Μάθηση

2η Εργασία – Classification problems

Έχοντας ολοκληρώσει τις σπουδές σας, δίνετε συνέντευξη για μια θέση με αντικείμενο "data analytics & financial risk estimations associate" σε ένα διεθνές χρηματοπιστωτικό ίδρυμα.

Στα πλαίσια της αξιολόγησης, σας παραδίδουν ένα αρχείο excel (Dataset2Use_Assignment2.xlsx) που περιέχει χρηματοπιστωτικούς δείκτες και μερικές ακόμα πληροφορίες για μια σειρά από ελληνικές εταιρείες.

Τα δεδομένα σας είναι:

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.

Το ζητούμενο είναι να δημιουργήσετε το καλύτερο δυνατό μοντέλο για το συγκεκριμένο πρόβλημα.

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 στην οθόνη ανά έτος τα ακόλουθα στοιχεία:
    a. Αριθμό υγιών και χρεωκοπημένων επιχειρήσεων
    b. Την min, max, average τιμή για κάθε δείκτη
3. Να κανονικοποιεί τα δεδομένα στο διάστημα $[0,1]$ και να τα χωρίζει σε train και test sets. Την αναλογία train/test data την καθορίζεται εσείς.
4. Να τυπώνει στην οθόνη πόσες χρεωκοπημένες και πόσες υγιείς εταιρείες υπάρχουν στο a) train και στο b) test set.
5. Να υλοποιεί (εκπαιδεύει) τέσσερα από τα ακόλουθα μοντέλα:
    a. Linear Discriminant Analysis
    b. Logistic Regression
    c. Decision Trees
    d. k-Nearest Neighbors
    e. Naïve Bayes
    f. Support Vector Machines
6. Θα τυπώνει confusion matrices τόσο στο train στο test set.

7. Θα υπολογίζει την επίδοση των εκπαιδευμένων μοντέλων τόσο στο train όσο και στο test set, με βάση τις ακόλουθες μετρικές: Accuracy, Precision, Recall, F1 score.
8. Θα περνάει τα αποτελέσματα σε ένα αρχείο csv. Κάθε γραμμή θα έχει τα ακόλουθα στοιχεία:
   Classifier Name | Training or test set| Number of training samples | Number of non-healthy companies in training sample | TP | TN | FP | FN | Precision | Recall | F1 score | Accuracy
   a. Classifier Name: το όνομα του ταξινομητή
   b. Training or test set: αναφορά αν τα αποτελέσματα της γραμμής αφορούν σε δεδομένα εκπαίδευσης ή αξιολόγησης
   c. Number of training samples: Αριθμός δεδομένων στο train set
   d. Number of non-healthy companies in training sample: Πόσες χρεωκοπημένες εταιρείες είχε το training set.
   e. TP: Αριθμός των true positive (χρεωκοπημένες εταιρεία που το μοντέλο τις έβγαλε ως χρεωκοπημένες).
   f. TN: Αριθμός των true negative (υγιείς εταιρείες που το μοντέλο τις έβγαλε ως υγιείς)
   g. FP: αριθμός των false positive
   h. FN: αριθμός των false negative
   i. Precision: η γνωστή μετρική
   j. Recall: η γνωστή μετρική
   k. F1 score: η γνωστή μετρική
   l. Accuracy: η γνωστή μετρική

Χρησιμοποιώντας το excel και το αρχείο csv από το ερώτημα 8, φτιάξτε γραφικές παραστάσεις που να δείχνουν πιο είναι το καλύτερο μοντέλο με βάση το F1 score. Πως δικαιολογείται η μεγάλη διαφορά στα αποτελέσματα μεταξύ train και test sets;

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».

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. Δεν θα δοθεί παράταση. **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)
    - Μέθοδοι που εφαρμόστηκαν: από 2 μέχρι 10 σελίδες, περιγράφετε τις μεθόδους που χρησιμοποιήσατε και παραθέτετε τα σχετικά αποτελέσματα. Φροντίστε να είναι ξεκάθαρο στο ποιο ερώτημα αναφέρεστε.
    - 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.

Καταληκτική Ημερομηνία Παράδοσης: **9 Ιανουαρίου 2023**. Δεν θα δοθεί παράταση.