1η Εργασία (Επιβλεπόμενη μάθηση – Ταξινόμηση)

Α' Μέρος [10 μονάδες]

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

Έχετε πάει πολύ καλά στην αρχική συνέντευξη και στη συνέχεια ακολουθεί η πρακτική αξιολόγηση. Σας παραδίδουν ένα αρχείο excel 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 Α έως και Η)
2. Τρεις δυϊκοί δείκτες δραστηριοτήτων (στήλες Ι, J, Κ)
3. Η κατάσταση της εταιρείας (1 όλα καλά, 2 έχει κηρύξει χρεωκοπία)
4. Το έτος στο οποίο αφορούν τα ως άνω μεγέθη.

Η δοκιμασία σας είναι η ακόλουθη:

I. Να δημιουργήσετε το καλύτερο δυνατό μοντέλο ταξινόμησης που θα εντοπίζει τις εταιρείες εκείνες που θα χρεωκοπήσουν, παίρνοντας ως είσοδο τις τιμές στις στήλες Α έως Κ. Φυσικά, όλα τα μοντέλα που θα αναπτύξετε πρέπει να είναι υλοποιημένα σε Python και να παραδοθούν για αξιολόγηση από τους υπευθύνους.

Προσοχή στους ακόλουθους δύο περιορισμούς:

1. Το μοντέλο πρέπει να βρίσκει με ποσοστό επιτυχίας **τουλάχιστον 62%** τις εταιρείες που θα πτωχεύσουν.
2. Το μοντέλο πρέπει να βρίσκει με ποσοστό επιτυχίας **τουλάχιστον 70%** τις εταιρείες που **δεν** θα πτωχεύσουν.

Μετά από σκέψη, αποφασίζετε να συντάξετε κώδικα στον οποίο:

1. Διαβάζετε τα δεδομένα από το excel, κάνετε κανονικοποίηση, και τα χωρίζετε σε training/test sets (αγνοήστε τη στήλη Μ με το έτος).
2. Εκπαιδεύετε και αξιολογείτε τα ακόλουθα μοντέλα επιβλεπόμενης μάθησης πάνω στα set που δημιουργήσατε, δηλαδή τα:
    - Linear Discriminant Analysis
    - Logistic Regression
    - Decision Trees

- k-Nearest Neighbors
- Naïve Bayes
- Support Vector Machines
- Neural Networks

3. Περνάτε τα αποτελέσματα των πειραμάτων σε ένα αρχείο excel όπου κάθε γραμμή έχει τις ακόλουθες τιμές:

```
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
```

Με βάση τα αποτελέσματα πάνω στα test data, υπάρχει κάποιο μοντέλο που να ικανοποιεί τους περιορισμούς απόδοσης;

II. Επαναλάβετε το παραπάνω πείραμα ως εξής:

Αφού φτιάξετε τα training και test sets, ελέγξτε το training set. 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.

Οδηγίες:

- Οι εργασίες είναι **ατομικές**.
- Οι εργασίες θα πρέπει να αναρτώνται στο eClass σε ένα αρχείο zip (όχι rar) εντός της προβλεπόμενης προθεσμίας.
- Κάθε εργασία πρέπει να συνοδεύεται από:
  - (υποχρεωτικά) Το MainScript.py: αυτό είναι το βασικό αρχείο στο οποίο έχετε γράψει τον κώδικά σας. Το αρχείο πρέπει να περιέχει στις πρώτες γραμμές, μέσα σε σχόλια και με λατινικούς χαρακτήρες, τα ακόλουθα:
    - Ονοματεπώνυμο, Τμήμα, πανεπιστημιακό email και αριθμό μητρώου
  - (εφόσον ζητείται) Έναν υποφάκελο με το όνομα InputData: σε αυτό τον φάκελο θα βρίσκονται τα αρχεία τα οποία διαβάζει το πρόγραμμά σας, π.χ. txt, csv, xlsx, jpg, bmp κ.λπ.
  - (εφόσον ζητείται) Έναν υποφάκελο με το όνομα OutputData: σε αυτόν θα υπάρχει το αρχείο results.txt ή results.xlsx στο οποίο θα καταγράφονται τα αποτελέσματα της εκτέλεσης του MainScript.py καθώς και οποιοδήποτε άλλο file παράγει το MainScript.
  - (υποχρεωτικά) Μια **αναφορά** σε Word ή σε pdf με τα ακόλουθα στοιχεία:

- Εξώφυλλο: 1 σελίδα, περιλαμβάνει τα στοιχεία του φοιτητή, όνομα μαθήματος, ημερομηνία, Τμήμα και λοιπές σχετικές πληροφορίες.
- 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 6 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 από μια τουλάχιστον παράγραφο με **σχολιασμό**.

- Ο κώδικας πρέπει να συνοδεύεται απαραιτήτως από κατάλληλα σχόλια.
- Οι εικόνες να ***μην*** έχουν προκύψει από print screen. Αν το πρόγραμμα δημιουργεί μια εικόνα αποθηκεύστε την κανονικά (jpg ή png) και εισάγετέ την στο Word.
- Οι γραφικές παραστάσεις να περιλαμβάνουν ονόματα στους άξονες και λεζάντα, στην οποία θα παραθέτετε μια σύντομη περιγραφή για το τι είναι αυτό που βλέπουμε.
- Αν κάτι δεν διευκρινίζεται, έχετε το δικαίωμα να κάνετε όποια υλοποίηση σας βολεύει. Φροντίστε οι υλοποιήσεις σας να τρέχουν στους υπολογιστές του εργαστηρίου ή στον φορητό σας υπολογιστή και να μπορείτε να εξηγήσετε τι ακριβώς κάνουν όταν εξεταστείτε.

Καταληκτική Ημερομηνία Παράδοσης: **Δευτέρα 31 Νοεμβρίου 2022**