|  |  |
| --- | --- |
| **new_eap_logo.jpg** | **ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ** |

**Εργασία:** Ομαδικό Προγραμματιστικό Project

**Ακαδημαϊκό Έτος :** 2021-2022

**Τίτλος Project:** *DT\_Titanic Ανάπτυξη μοντέλου κατηγοριοποίησης*

**ID Project:** 38

Η Μηχανική Μάθηση (Machine Learning) είναι ένα πεδίο της επιστήμης των υπολογιστών, που ασχολείται με τη μελέτη της αναγνώρισης προτύπων (μοτίβων) και την  υπολογιστική θεωρία μάθησης στην τεχνητή νοημοσύνη.

Στο πεδίο της ανάλυσης δεδομένων, η μηχανική μάθηση είναι μια μέθοδος που χρησιμοποιείται για την επινόηση πολύπλοκων μοντέλων και αλγορίθμων που οδηγούν στην πρόβλεψη. Τα αναλυτικά μοντέλα επιτρέπουν στους επιστήμονες και τους αναλυτές να παράγουν αξιόπιστες αποφάσεις και αποτελέσματα και να αναδείξουν αλληλοσυσχετίσεις μέσω της μάθησης (απόκτησης γνώσης) από ιστορικές σχέσεις και τάσεις στα δεδομένα.

Το σύνολο δεδομένων Titanic (Titanic.xlsx) περιλαμβάνει 2201 εγγραφές (στιγμιότυπα). Κάθε στιγμιότυπο περιλαμβάνει πληροφορίες για τους επιβάτες και το πλήρωμα του Τιτανικού όταν βυθίστηκε στις 15 Απριλίου του 1912. Κάθε στιγμιότυπο προσδιορίζεται από 4 χαρακτηριστικά:

* Το είδος της θέσης στο πλοίο (πρώτη, δεύτερη, τρίτη, πλήρωμα)
* Την ηλικία του ατόμου (ενήλικος ή παιδί)
* Το φύλο του ατόμου (άνδρας ή γυναίκα)
* Το τέταρτο χαρακτηριστικό είναι το χαρακτηριστικό εξόδου, και δηλώνει αν το άτομο που επιβίωσε διηγήθηκε την ιστορία του τραγικού συμβάντος στις αρχές.

Αναπτύξτε το προφίλ του διασωθέντος που διηγήθηκε την ιστορία χρησιμοποιώντας τεχνική κατηγοριοποίησης «δέντρο απόφασης (decision tree)».

**Ζητούμενα:**

Να γράψετε κώδικα στη γλώσσα προγραμματισμού python ο οποίος θα

(i) διαβάσει το αρχείο Titanic.xlsx.

(ii) επιλέξει ως training data set τυχαία το 80% των εγγραφών με το εξής κριτήριο: στο 80% των εγγραφών του training data set θα υπάρχει η ίδια αναλογία μεταξύ των δυο κλάσεων όπως αυτή εμφανίζεται στο σύνολο των δεδομένων. Το υπόλοιπο 20% των εγγραφών θα χρησιμοποιηθεί ως test data set.

(iii) χρησιμοποιήσει τον αλγόριθμο DecisionTreeClassifier για την ανάπτυξη μοντέλου πρόβλεψης όπου η δεσμευμένη μεταβλητή θα είναι η Survived και ανεξάρτητες οι υπόλοιπες μεταβλητές.

(iv) θα υπολογίζει τη μήτρα σύγχυσης (confusion matrix) για την απόδοση του προηγούμενου μοντέλου.

(v) θα υλοποιεί ένα απλό γραφικό περιβάλλον χρήστη (μπορείτε να κάνετε χρήση του Tkinter GUI: <https://docs.python.org/3/library/tkinter.html>.) όπου θα μπορούν να γίνονται το άνοιγμα του αρχείου, η εκτέλεση του αλγορίθμου και η εμφάνιση του δέντρου απόφασης με τη μήτρα σύγχυσης.

**HINT!** Μπορείτε να χρησιμοποιήσετε τις βιβλιοθήκες pandas, numpy, sklearn και όποιες άλλες κρίνετε απαραίτητο.

**Παραδοτέα:**

Παραδίδετε ένα αρχείο zip που περιέχει:

1. Τον κώδικά σας.
2. Οδηγίες εγκατάστασης και χρήσης της εφαρμογής σας.
3. Έκθεση (5-10 σελίδων pdf) που θα εξηγεί: τί κάνατε, πώς το κάνατε, πώς μοιράσατε τη δουλειά, ποια ήταν τα αποτελέσματα. Βάλτε screenshots από την εκτέλεση της εφαρμογής σας (όχι screenshots του κώδικα!!!)
4. Επιπλέον, κάθε φοιτητής ξεχωριστά θα πρέπει να γράψει μια ατομική έκθεση 1-2 σελίδων pdf που θα εξηγεί τί ακριβώς έκανε ο ίδιος, πόσες ώρες δούλεψε και κατάλογο με βιβλιογραφία (τί μελέτησε, από ποιες πηγές, ακόμα κι αν αυτές δεν είναι βιβλία, αλλά ιστοσελίδες ή video). Μην ξεχάσετε να βάλετε το όνομά σας στο αρχείο.
5. Αρχείο pdf με τις διαφάνειες της παρουσίασής σας.
6. Προαιρετικά, μπορείτε να φτιάξετε video με την παρουσίασή σας στο youtube και να αναρτήσετε τον κώδικά σας στο github. Σε αυτήν την περίπτωση βάλτε τα links στο κείμενο της έκθεσής σας.