ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

ΑΝΑΦΟΡΑ ΑΣΚΗΣΗΣ 1

ΜΕΛΕΤΗ ΜΕΘΟΔΩΝ ΤΑΞΙΝΟΜΗΣΗΣ ΔΕΔΟΜΕΝΩΝ

(DATA CLASSIFICATION ALGORITHMS)

Ομάδα:

Κώστας Ορφανούδης 3303

Σεραφείμ Θεμιστοκλέους 4555

Εαρινό Εξάμηνο 2021-2022

**Προεπεξεργασία Δεδομένων**

* Αρχικά φορτώνουμε τα αρχεία train και test και αντικαθιστούμε τις ετικέτες (labels) στο train\_data:

Ghoul -> 0

Goblin -> 1

Ghost -> 2

* Αποθηκεύουμε τις ετικέτες του train\_data στο tlabel.
* Αφαιρούμε από το train\_data την ετικέτα Type και το χαρακτηριστικό ID.
* Επίσης για το χαρακτηριστικό color αντικαθιστούμε το κάθε χρώμα με μια κανονικοποιημένη τιμή από το σύνολο {1/6, 2/6, 3/6, 4/6, 5/6 , 6/6}.
* Εφαρμόζοντας cross validation, διαχωρίζουμε το train\_data σε μερίδιο για προπόνηση (train) του μοντέλου και σε μερίδιο για έλεγχο (test) του μοντέλου δίνοντας το 25% του συνόλου δεδομένων στο μερίδιο ελέγχου.

**1η Μέθοδος:**

**k-NN Nearest Neighbor με Ευκλείδεια απόσταση**

Αρχικοποιούμε τον ταξινομητή με ευκλείδεια απόσταση για τις διάφορες τιμες του k = {1,3,5,10}. Στη συνέχεια προπονούμε τον ταξινομητή με τα χαρακτηριστικά του μεριδίου προπόνησης και τις αντίστοιχες ετικέτες του train\_data. Προβλέπουμε τις ετικέτες για τα χαρακτηριστικά στο μερίδιο ελέγχου του train\_data. Εκτυπώνουμε τα μέτρα αξιολόγησης Accuracy και F1 score του train\_data. Αφού προπονήσουμε το μοντέλο προβλέπουμε τις ετικέτες που αντιστοιχουν στο test\_data. Αποθηκεύουμε στη μορφή του sample\_submission τις ετικέτες που προέκυψαν μαζι με το αντίστοιχο ID. Τέλος ανεβάζουμε το csv αρχείο στη σελίδα του διαγωνισμού και παίρνουμε τα αποτελέσματα.

Παρακάτω παρατίθενται τα αποτελέσματα για τις τιμές του k:

**2η Μέθοδος:**

**Neural Networks με σιγμοειδή συνάρτηση ενεργοποίησης**

Αρχικοποιούμε τον ταξινομητή με ευκλείδεια απόσταση για τις διάφορες τιμες του k = {1,3,5,10}. Στη συνέχεια προπονούμε τον ταξινομητή με τα χαρακτηριστικά του μεριδίου προπόνησης και τις αντίστοιχες ετικέτες του train\_data. Προβλέπουμε τις ετικέτες για τα χαρακτηριστικά στο μερίδιο ελέγχου του train\_data. Εκτυπώνουμε τα μέτρα αξιολόγησης Accuracy και F1 score του train\_data. Αφού προπονήσουμε το μοντέλο προβλέπουμε τις ετικέτες που αντιστοιχουν στο test\_data. Αποθηκεύουμε στη μορφή του sample\_submission τις ετικέτες που προέκυψαν μαζι με το αντίστοιχο ID. Τέλος ανεβάζουμε το csv αρχείο στη σελίδα του διαγωνισμού και παίρνουμε τα αποτελέσματα.

Παρακάτω παρατίθενται τα αποτελέσματα για τις τιμές του k: