ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΑΛΓΟΡΙΘΜΟΙ ΜΑΘΗΣΗ

ΥΛΟΠΟΙΗΤΙΚΗ ΑΣΚΗΣΗ

Επιβλέποντες καθηγητές: Χρήστος Μακρής, Βασίλης Μεγαλοοικονόμου

Παναγιώτης Καββαδίας

ΑΜ:1054350

Περιγραφή

Στα πλαίσια του project κληθήκαμε να υλοποιήσουμε κατηγοριοποιητές. Για το πρώτο ερώτημα έπρεπε να υλοποιηθεί αλγόριθμος της κατηγορίας SVM που δοθέντων των στοιχείων της χημικής σύστασης ενός κρασιού θα μπορεί να εκτιμήσει την ποιότητα του.

Για το δεύτερο ερώτημα έπρεπε να κατασκευαστεί και να εκπαιδευτεί ένα νευρωνικό δίκτυο το οποίο θα μπορεί να εκτιμήσει από τον τίτλο αν μια είδηση δημοσιεύθηκε ή οχι στο χιουμοριστικό website theonion.com.

Η υλοποίηση έγινε σε γλώσσα Python έκδοσης 3.7.7 στην πλατφόρμα Anaconda. Το IDE που χρησιμοποιήθηκε ήταν το Spyder στην έκδοση 4.1.3.

Χρησιμοποιήθηκαν οι βιβλιοθήκες nltk και scikit-learn καθώς και οι ενσωματωμένες βασικές βιβλιοθήκες της python.

Για την εκτέλεση των υλοποιήσεων αρκεί να αποσυμπιεστεί το .zip αρχείο που περιλαμβάνεται η αναφορά και να εκτελεστούν τα Wine.py και Onion.py για το πρώτο και το δεύτερο ερώτημα αντίστοιχα. Τα αρχεία πρέπει να βρίσκονται στον ίδιο φάκελο με τα .csv αρχεία. Για ευκολία στην εκτέλεση συμπεριλαμβάνονται και αυτά στο .zip της αναφοράς.

Πρώτο ερώτημα

Για το πρώτο ερώτημα υλοποιήθηκε το πρώτο ζητούμενο και τα πρώτα δυο ερωτήματα του δεύτερου ζητουμένου. Αρχικά διαβάζεται το csv αρχείο και αρχικοποιείται το Support Vector Machine. Γίνεται η εκπαίδευση του και εμφανίζεται η κατηγοριοποίηση με τα αρχικά δεδομένα. Ύστερα αφαιρούμε το 33% των τιμών από τη στήλη pH. Πρώτα εμφανίζουμε την κατηγοριοποίηση με τις τιμές αυτές κενές. Ύστερα οι κενές αυτές τιμές παίρνουν την τιμή του μέσου όρου της στήλης και ξανακάνουμε κατηγοριοποίηση.

```
Classification with unedited data:

              precision    recall  f1-score   support

           3       0.00      0.00      0.00         3
           4       0.00      0.00      0.00        15
           5       0.54      0.44      0.48       155
           6       0.47      0.75      0.58       171
           7       0.00      0.00      0.00        53
           8       0.00      0.00      0.00         3

    accuracy                           0.49       400
   macro avg       0.17      0.20      0.18       400
weighted avg       0.41      0.49      0.43       400

Classification with deleted pH column

              precision    recall  f1-score   support

           3       0.00      0.00      0.00         3
           4       0.00      0.00      0.00        15
           5       0.54      0.44      0.48       155
           6       0.47      0.75      0.58       171
           7       0.00      0.00      0.00        53
           8       0.00      0.00      0.00         3

    accuracy                           0.49       400
   macro avg       0.17      0.20      0.18       400
weighted avg       0.41      0.49      0.43       400
```

```
Classification with average pH to the column

              precision    recall  f1-score   support

           3       0.00      0.00      0.00         3
           4       0.00      0.00      0.00        15
           5       0.40      0.98      0.57       155
           6       0.72      0.08      0.14       171
           7       0.00      0.00      0.00        53
           8       0.00      0.00      0.00         3

    accuracy                           0.41       400
   macro avg       0.19      0.18      0.12       400
weighted avg       0.46      0.41      0.28       400
```

Από τα αποτελέσματα παρατηρώ πως οι διαφορές της κατηγοριοποίησης όταν έχει επεξεργαστεί η στήλη pH είναι. Πιθανώς αν το dataset ήταν μεγαλύτερο οι διαφορές να ήταν πιο ορατές.

Δεύτερο ερώτημα

Στο δεύτερο ερώτημα επέλεξα να υλοποιήσω ένα multilayer perceptron για το classification χρησιμοποιώντας την βιβλιοθήκη MLPClassifier του scikit-learn.

Χρησιμοποιήθηκε επίσης η βιβλιοθήκη tfidf vectorizer του scikit-learn για να δημιουργηθεί το tf-idf διάνυσμα των λέξεων.

Αρχικά υπάρχει συνάρτηση για την εισαγωγή δεδομένων από το csv αρχείο παρόμοια με αυτή που χρησιμοποιήθηκε στο προηγούμενο ερώτημα. Αφού εισαχθούν οι τίτλοι εφαρμόζεται stemming μέσω του porter stemmer της βιβλιοθήκης nltk ώστε να διατηρηθεί μόνο το θέμα των λέξεων. Ύστερα αφαιρούνται οι stopwords και στις εναπομείνασες λέξεις ανατίθεται ως βάρος η τιμή tf-idf. Δημιουργείται ένα perceptron (με χρήση της βιβλιοθήκης MLPClassifier) το οποίο εκπαιδεύεται με τα δεδομένα που έχουμε. Στο τέλος γίνεται η πρόβλεψη και μετράμε f1,precision, recall και support μέσω της classification_report της βιβλιοθήκης scikit learn .

	precision	recall	f1-score	support
0	0.83	0.86	0.85	1883
1	0.75	0.70	0.72	1117
accuracy			0.80	3000
macro avg	0.79	0.78	0.78	3000
weighted avg	0.80	0.80	0.80	3000

Να σημειωθεί ότι λόγω περιορισμών μνήμης του υπολογιστή μου οι υπολογισμοί με το αρχικό dataset onion-or-not.csv δεν μπορούσαν να εκτελεστούν κι έτσι αφαίρεσα το 50% των τιμών του δημιουργώντας το onion-or-not2.csv(περιλαμβάνεται στο .zip με τα παραδοτέα) ώστε να είναι επιτυχής η εκτέλεση.