Αναφορά Project

Στοιχεία Μελών Ομάδας:

Παπαδόπουλος Λάμπρος: ΑΜ: 1054433 Έτος Σπουδών: 4ο

Email: up104433@upnet.gr

Κωστορρίζος Δημήτριος: ΑΜ: 1054419 Έτος Σπουδών: 4ο

Email: up104419@upnet.gr

Καταγραφή του περιβάλλοντος

Για το project χρησιμοποιήθηκαν βασικές βιβλιοθήκες της Python version 3.7, καθώς και οι βιβλιοθήκες scikit-learn και NTLP. Για την εγκατάσταση του Project, δεν απαιτείται τίποτα περισσότερο από να οριστεί ως source φάκελος του project, ο φάκελος Source.

Διαδικασία Υλοποίησης

Το project είναι χωρισμένο σε 2 main function. Η SVM main περιλαμβάνει τον κώδικα για το 1ο Ερώτημα της άσκησης, ενώ η NLP main τον κώδικα για το 2ο Ερώτημα.

WineQualityMetrics: Η κλάση για τις μετρήσεις του κρασιού και η συνάρτηση που μετατρέπει το instance σε λίστα.

SVMHelperMethods: Το αρχείο που περιέχει της βοηθητικές συναρτήσεις. Πιο συγκεκριμένα, στο συγκεκριμένο αρχείο περιέχονται οι συναρτήσεις για import του csv αρχείου, η συνάρτηση για την μετατροπή των στιγμιότυπων στην μορφή λίστας από λίστας καθώς και οι συναρτήσεις που εφαρμόζουν την επεξεργασία στο dataset για το ερώτημα Β.

SVM main: Χρησιμοποιείται η συνάρτηση csv importer, ώστε να γίνει import το περιεχόμενο του αρχείου. Μετά το περιεχόμενο του αρχείου, χωρίζεται τυχαία σε training sample και test sample και μετασχηματίζεται στην απαραίτητη μορφή, δηλαδή στην μορφή: λίστα με λίστες, όπου οι εμφωλευμένες λίστες περιέχουν τα samples. Έπειτα, γίνεται πρόβλεψη των τιμών της ποιότητας τους κρασιού. Με βάση τις τιμές της πρόβλεψης και τις αρχικές τιμές για την ποιότητα του κρασιού,

υπολογίζονται οι τιμές για τις μετρικές F1 Score, Recall και Precision. Έπειτα, επαναλαμβάνεται η αντίστοιχη διαδικασία για τις παρακάτω περιπτώσεις:

- Αφαίρεση της στήλης του pH από το test και το train set.
- Συμπλήρωση της στήλης του pH του 30% του train set, με τον μέσο όρο του υπόλοιπου 60% του train set.
- Συμπλήρωση της στήλης του pH του 30% του train set, με τις τιμές από την εφαρμογή Logistic Regression στο υπόλοιπο 60% του train set.
- Συμπλήρωση της στήλης του pH του 30% του train set, με την μέση τιμή, των τιμών από την εφαρμογή K-Means Clustering στο υπόλοιπο 60% του train set.

NLP main: Χρησιμοποιείται η συνάρτηση csv importer, ώστε να γίνει import το περιεχόμενο του αρχείου. Έπειτα, οι τίτλοι χωρίζονται σε λέξεις. Στις λέξεις αυτές εφαρμόζεται word stemming. Από τις λέξεις αυτές, αφαιρούνται οι stopword λέξεις. Για την λίστα που περιέχει τις λίστες με τις εναπομείναντες λέξεις, υπολογίζονται οι τιμές Tf-Idf για την κάθε λέξη. Το μητρώο των αποτελεσμάτων μετασχηματίζεται σε λίστα με λίστες, όπου οι εμφωλευμένες λίστες περιέχουν τα samples. Το σύνολο των δειγμάτων χωρίζεται σε training και test set. Τα set αυτά χρησιμοποιούνται για να εκπαιδευτεί το Multi-Layer Perceptron. Με βάση τις τιμές της πρόβλεψης και τις αρχικές τιμές για την για το είδος του τίτλου, υπολογίζονται οι τιμές για τις μετρικές F1 Score, Recall και Precision.

NLPHelperMethods: Περιέχει την συνάρτηση για το import του dataset από το csv αρχείο.

Σχολιασμός τελικών αποτελεσμάτων

Σημείωση: Οι παράμετροι των SVM, LogisticRegresssion, K-means και Multi-Layer Perceptron object, έχουν ρυθμιστεί ώστε να συγκλίνουν στο επιθυμητό αποτέλεσμα σε λογικά πλαίσια χρόνου. Οι μετρήσεις F1 Score, Recall και Precision ανήκουν στα επιθυμητά όρια, ωχτόσο θα μπορούσαν να βελτιωθούν ακόμη περισσότερο, προσαρμόζοντας τις παραμέτρους των παραπάνω αντικειμένων ώστε να ανταποκρίνονται καλύτερα στο εκάστοτε dataset.

Μετρήσεις για το Support Vector Machine Classifier

Χωρίς επεξεργασία των δεδομένων εισόδου

- **F1 Score: 0.71**
- **Precision: 0.67**
- **Recall Score: 0.77**

Διαγραφή της στήλης του pH

- **F1 Score: 0.71**
- **Precision: 0.67**
- **Recall Score: 0.77**

Μέση τιμή για την στήλη του pH

- **F1 Score: 0.71**
- **Precision: 0.67**
- **Recall Score: 0.77**

Εφαρμογή Logistic Regression Classification στην στήλη του pH

- **F1 Score: 0.71**
- **Precision: 0.67**
- **Recall Score: 0.77**

Εφαρμογή K-Means Clustering στην στήλη του pH

- **F1 Score: 0.71**
- **Precision: 0.67**
- **Recall Score: 0.77**

Μετρήσεις για το Multi-Layer Perceptron Classification

- **F1 Score: 0.81**
- **Precision: 0.85**
- **Recall Score: 0.96**

Με βάση τις παραπάνω μετρήσεις, παρατηρούμε ότι το SVM classification, εμφανίζει τις ίδιες μετρήσεις ανεξαρτήτως της επεξεργασία που εφαρμόζουμε στα δεδομένα εισόδου. Πιο συγκεκριμένα, στις περιπτώσεις όπου εσκεμμένα προσπαθούμε να προβλέψουμε το 33% τους dataset, καταφέραμε να προβλέψουμε τα αρχικά “χαμένα” δεδομένα, με κατάλληλη ακρίβεια ώστε να μην δημιουργήσουν σφάλματα στην ολική πρόβλεψη.

Με βάση τις μετρήσεις για το Multi-Layer Perceptron, παρατηρούμε ότι το νευρωνικό δίκτυο πετυχαίνει αρκετά μεγάλη ακρίβεια στην πρόβλεψη των δεδομένων.