ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ «Ανταγωνιστικότητα Επιχειρηματικότητα και Καινοτομία»

ΑΞΟΝΑΣ ΠΡΟΤΕΡΑΙΟΤΗΤΑΣ 03 «Ανάπτυξη επιχειρηματικότητας με Τομεακές προτεραιότητες»

ΔΡΑΣΗ «ΕΡΕΥΝΩ – ΔΗΜΙΟΥΡΓΩ – ΚΑΙΝΟΤΟΜΩ»

LinkGeoML: Αυτοματοποιημένη και ακριβής διασύνδεση γεωχωρικών δεδομένων με τη χρήση μεθόδων μηχανικής μάθησης

ΚΩΔΙΚΟΣ ΟΠΣ «5030745»

ΤΙΤΛΟΣ ΠΑΡΑΔΟΤΕΟΥ

Π2.1: «Αλγόριθμοι μηχανικής μάθησης για διασύνδεση»

Πακέτο Εργασίας	**ΠΕ2: Ανάπτυξη μεθόδων μηχανικής μάθησης για διασύνδεση**
Υπεύθυνος Φορέας	**Ερατοσθένης ΑΕ**
Είδος Παραδοτέου	**Λογισμικό**
Ενδεικτικός Μήνας Παράδοσης	**Μ12**
Ημερομηνία Παράδοσης	**8/7/2019 (Μ12)**

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ ΥΠΟΥΡΓΕΙΟ ΟΙΚΟΝΟΜΙΑΣ & ΑΝΑΠΤΥΞΗΣ ΕΙΔΙΚΗ ΓΡΑΜΜΑΤΕΙΑ ΕΠΑ & ΤΣ ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ ΕΠΑνΕΚ
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

ΙΣΤΟΡΙΚΟ ΑΛΛΑΓΩΝ

Έκδοση	Ημερομηνία	Εργασίες	Συγγραφείς
0.1	08/05/2019	Δομή και πίνακας περιεχομένων του παραδοτέου	Γιώργος Γιαννόπουλος (ΑΘ.), Δημήτριος Σκούτας (ΑΘ.) Νώντας Τσάκωνας (ΕΡ.)
0.2	16/05/2019	Προσθήκη υλικού στις βιβλιοθήκες μηχανικής μάθησης	Βασίλης Καφφές (ΑΘ.), Νίκος Κωσταγιόλας(ΑΘ.)
0.3	23/05/2019	Προσθήκη υλικού στην αρχιτεκτονική	Νώντας Τσάκωνας (ΕΡ.), Γιώργος Γιαννόπουλος (ΑΘ.) .), Μιχάλης Αεράκης (GE.)
0.4	31/05/2019	Προσθήκη υλικού στις βιβλιοθήκες μηχανικής μάθησης	Νίκος Κωσταγιόλας(ΑΘ.), Γιώργος Ευταξίας (ΕΡ.), Μιχάλης Αεράκης (GE.)
0.5	13/06/2019	Προσθήκη υλικού στην αξιολόγηση των μεθόδων	Βασίλης Καφφές (ΑΘ.), Μανώλης Αλεξανδράκης(ΕΡ.),
0.6	24/06/2019	Διάφορες προσθήκες και βελτιώσεις	Νώντας Τσάκωνας (ΕΡ.), Γιώργος Ευταξίας (ΕΡ.), Γιώργος Γιαννόπουλος (ΑΘ.), Μιχάλης Αεράκης (GE.)
0.7	26/06/2019	Εσωτερικά επιθεωρημένη έκδοση	Δημήτριος Σκούτας (ΑΘ.), Γιώργος Γιαννόπουλος (ΑΘ.
1.0	05/07/2019	Τελική έκδοση	Δημήτριος Σκούτας (ΑΘ.), Γιώργος Γιαννόπουλος (ΑΘ.), Νώντας Τσάκωνας (ΕΡ.)

ΠΙΝΑΚΑΣ ΣΥΝΤΜΗΣΕΩΝ

ΣΕ	Σημείο Ενδιαφέροντος
MLP	Multi-Layer Perceptron
SVM	Support Vector Machines

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΠΕΡΙΛΗΨΗ

Το παραδοτέο περιγράφει την πρώτη έκδοση των βιβλιοθηκών κώδικα που υλοποιούν μοντέλα μηχανικής μάθησης, οι οποίες αναπτύχθηκαν στο έργο για τη διασύνδεση, επισημείωση και ολοκλήρωση γεωχωρικών δεδομένων. Τα υλοποιημένα μοντέλα μηχανικής συνίστανται στα χαρακτηριστικά εκπαίδευσης που ορίστηκαν στο Παραδοτέο 1.2: «Προδιαγραφή εξειδικευμένων χαρακτηριστικών και κανόνων εκπαίδευσης», καθώς και σε ένα σύνολο από αλγόριθμους μηχανικής μάθησης με τους οποίους συνδυάζονται. Τα χαρακτηριστικά εκπαίδευσης που υλοποιούνται στο τρέχον παραδοτέο οργανώνονται ανά ομάδες, ανάλογα με το επιμέρους πρόβλημα διασύνδεσης/σενάριο χρήσης στο οποίο εφαρμόζονται, ακολουθώντας την ομαδοποίηση των προηγούμενων παραδοτέων (Π1.1, Π1.2). Κατόπιν, συνδυάζονται με ένα σύνολο από διαφορετικούς αλγόριθμους μηχανικής μάθησης σε πλήρεις μεθόδους, οι οποίες αξιολογούνται μέσω διαφορετικών παραμετροποιήσεων, όσον αφορά την αποτελεσματικότητά τους (ακρίβεια διασύνδεσης/επισημείωσης/ ολοκλήρωσης) στο εκάστοτε σενάριο. Ο αποτελεσματικότερος συνδυασμός χαρακτηριστικών εκπαίδευσης, αλγορίθμου μηχανικής μάθησης και παραμετροποίησης, δίνει, για κάθε σενάριο χρήσης, το τελικό μοντέλο μηχανικής μάθησης για εκτέλεση στο αντίστοιχο πρόβλημα.

Η παραπάνω διαδικασία υλοποιείται μέσω μίας αναλυτικής ακολουθίας διεργασιών μηχανικής μάθησης (machine learning pipeline) η οποία είναι κοινή για όλα τα σενάρια χρήσης και κατ' επέκταση για τις βιβλιοθήκες κώδικα που υλοποιούν τα αντίστοιχα μοντέλα μηχανικής μάθησης. Η συγκεκριμένη ακολουθία διεργασιών προδιαγράφει όλα τα ουσιώδη βήματα εκτέλεσης διαδικασιών μηχανικής μάθησης, όπως επιλογή χαρακτηριστικών (feature selection), εμφωλευμένη συγκριτική αποτίμηση αλγορίθμων (nested cross validation) επιλογή βέλτιστων υπερ-παραμέτρων με αναζήτηση πλέγματος (grid search), επιλογή βέλτιστου συνδυασμού αλγορίθμων και υπερ-παραμέτρων, εκπαίδευση μοντέλου σε ιστορικά δεδομένα και εκτέλεση μοντέλου σε νέα δεδομένα. Η παραπάνω ακολουθία διεργασιών συνιστά το πρότυπο σχεδιασμού και υλοποίησης που υιοθετείται σε όλες τις βιβλιοθήκες μοντέλων μηχανικής μάθησης που υλοποιούνται κατά τη διάρκεια του έργου.

Το παραδοτέο δομείται ως εξής:

Στην Ενότητα 1 περιγράφονται οι στόχοι του παραδοτέου, όσον αφορά την ανάπτυξη της πρώτης έκδοσης των μοντέλων μηχανικής μάθησης για διασύνδεση, επισημείωση και ολοκλήρωση γεωχωρικών δεδομένων.

Στην Ενότητα 2 Επιπλέον, περιγράφονται οι βασικές αρχές που καθοδήγησαν την ανάπτυξη των βιβλιοθηκών, καθώς και η γενική αρχιτεκτονική όλων των υλοποιημένων βιβλιοθηκών μηχανικής μάθησης, η οποία συνίσταται σε μία ακολουθία διεργασιών που καλύπτουν όλο το φάσμα μίας διαδικασίας εκπαίδευσης και εφαρμογής μοντέλων μηχανικής μάθησης.

Στην Ενότητα 3, περιγράφεται το λογισμικό που υλοποιήθηκε στη μορφή βιβλιοθηκών μηχανικής μάθησης, οργανωμένο ανά σενάριο χρήσης/επιλυόμενο πρόβλημα μηχανικής μάθησης για διασύνδεση, κατηγοριοποίηση, γεωκωδικοποίηση και ολοκλήρωση γεωχωρικών δεδομένων. Οι περιγραφές περιλαμβάνουν πληροφορίες υλοποίησης,

τεκμηρίωσης, άδειας χρήσης και πρόσβασης του υλοποιημένου κώδικα, καθώς και οδηγούς εγκατάστασης και εκτέλεσης των βιβλιοθηκών.

Στην Ενότητα 4 παρουσιάζονται τα αποτελέσματα της πρώτης πειραματικής αξιολόγησης των υλοποιημένων μεθόδων, σε γεωχωρικά δεδομένα χρησιμοποιούμενα σε προβλήματα πραγματικού κόσμου.

Στην Ενότητα 5 συνοψίζεται το παραδοτέο και παρουσιάζονται μελλοντικές κατευθύνσεις για την περαιτέρω επέκταση και βελτίωση των βιβλιοθηκών μηχανικής μάθησης που υλοποιήθηκαν.

1. Εισαγωγή

Τα γεωχωρικά[1] δεδομένα είναι σύνθετες οντότητες που παρουσιάζουν μεγάλη ετερογένεια ως προς τις ιδιότητες που τα χαρακτηρίζουν, αλλά και τα σενάρια χρήσης τους. Επιπλέον, λόγω της εξαιρετικά μεγάλης σημασίας τους σε διάφορα εμπορικά σενάρια χρήσης (κτηματογράφηση, γεωχωρική ανάλυση, πλοήγηση, κ.α.), είναι συχνό το φαινόμενο της παραγωγής διαφορετικών συνόλων γεωχωρικών δεδομένων που αναφέρονται στις ίδιες γεωχωρικές περιοχές, ακολουθώντας διαφορετικές διαδικασίες και πρότυπα συλλογής, επεξεργασίας, καθαρισμού, ολοκλήρωσης και επισημείωσης. Ως αποτέλεσμα, σε πολλά σενάρια χρήσης πραγματικού κόσμου, μία εταιρία (πάροχος ή καταναλωτής) γεωχωρικών δεδομένων πρέπει να επιλύσει προβλήματα διασύνδεσης, επισημείωσης και γενικότερα ολοκλήρωσης γεωχωρικών δεδομένων από διαφορετικές πηγές.

Στο πλαίσιο του έργου, αντιμετωπίζουμε τα παραπάνω προβλήματα πραγματοποιώντας εμπορική έρευνα και ανάπτυξη αλγορίθμων μηχανικής μάθησης για την επίλυσή τους. Τελικός στόχος του έργου είναι η ανάπτυξη μίας ευρείας βιβλιοθήκης από χαρακτηριστικά εκπαίδευσης και αλγόριθμους μηχανικής μάθησης για διασύνδεση, επισημείωση και ολοκλήρωση γεωχωρικών δεδομένων. Για το σκοπό αυτό, ξεκινήσαμε ορίζοντας τέσσερα σενάρια χρήσης τα οποία αντιπροσωπεύουν υπαρκτά προβλήματα διασύνδεσης των επιχειρηματικών εταίρων του έργου (Π1.1). Πάνω σε αυτά τα σενάρια ορίσαμε αντίστοιχα προβλήματα μηχανικής μάθησης και αντίστοιχα χαρακτηριστικά εκπαίδευσης, τα οποία αφορούν μεν τα τέσσερα ορισμένα σενάρια χρήσης, δύνανται να χρησιμοποιηθούν δε και σε τυχόν παρεμφερή προβλήματα μηχανικής μάθησης για διασύνδεση γεωχωρικών δεδομένων (Π1.2). Τρίτο βήμα της εργασίας μας στο έργο αποτέλεσε η υλοποίηση βιβλιοθηκών κώδικα μηχανικής μάθησης για καθένα από τα παραπάνω σενάρια χρήσης/προβλήματα.

Σε πρώτη φάση, υλοποιήσαμε και πειραματιστήκαμε με ένα αρχικό σύνολο χαρακτηριστικών εκπαίδευσης και αλγορίθμων μηχανικής μάθησης, όπως προδιαγράφεται κυρίως στην περιγραφή του Π2.1 στο Τεχνικό Παράρτημα του έργου, χτίζοντας τη βάση των διαδικασιών, αλγορίθμων και μοντέλων που θα επεκταθούν και βελτιωθούν στη συνέχεια του έργου. Επιπλέον όμως, ορίσαμε μία πλήρη ακολουθία διεργασιών εκπαίδευσης και εκτέλεσης μοντέλων μηχανικής μάθησης που επιτρέπει την αποτελεσματική και αποδοτική αξιολόγηση και εκπαίδευση των υλοποιημένων μοντέλων, με στόχο τη βέλτιστη εκτέλεσή τους για την επίλυση των θεωρηθέντων προβλημάτων. Τα παραπάνω υλοποιήθηκαν ως τέσσερις βιβλιοθήκες ανοικτού κώδικα, οι οποίες επιλύουν προβλήματα διασύνδεσης, κατηγοριοποίησης, γεωκωδικοποίησης και ολοκλήρωσης χωρο-κειμενικών δεδομένων, και παρουσιάζονται αναλυτικά στην Ενότητα 3.

[1] Στην πλειονότητα των περιπτώσεων που αφορούν το έργο, αλλά και γενικότερα εφαρμογές διαχείρισης και ολοκλήρωσης γεωχωρικών δεδομένων, τα δεδομένα, πέρα από τις γεωχωρικές ιδιότητές τους, συνοδεύονται και από κάποιας μορφής κειμενική πληροφορία (όνομα, όνομα ιδιοκτήτη, περιγραφή, κτλ.). Για αυτό, οι όροι «γεωχωρικό» και «χωρο-κειμενικό» χρησιμοποιούνται εναλλάξ στο παρόν παραδοτέο, έχοντας την ίδια σημασία.

2. Αρχιτεκτονική βιβλιοθηκών

Οι βασικές αρχές και προδιαγραφές που καθοδήγησαν το σχεδιασμό και την υλοποίηση των βιβλιοθηκών (διεργασιών, χαρακτηριστικών εκπαίδευσης, αλγορίθμων) μηχανικής μάθησης συνοψίζονται στα παρακάτω σημεία:

- **Κάλυψη όσο το δυνατόν περισσότερων εμπορικών προβλημάτων διασύνδεσης γεωχωρικών δεδομένων**. Μπορεί η διασύνδεση γεωχωρικών οντοτήτων να είναι ένα ουσιώδες πρόβλημα που αφορά μία ευρεία γκάμα και σενάρια χρήσης γεωχωρικών δεδομένων, αλλά δεν είναι το μοναδικό. Η προδιαγραφή σεναρίων χρήσης που πραγματοποιήθηκε στην ΕΕ1 του έργου ανέδειξε επιπλέον υπαρκτά προβλήματα, τα οποία είτε είναι συμπληρωματικά/παρεμφερή με τη διασύνδεση γεωχωρικών δεδομένων, αφορώντας για παράδειγμα στην επισημείωση/κατηγοριοποίηση ή γενικότερα ολοκλήρωση γεωχωρικών δεδομένων, είτε χρησιμοποιούν εσωτερικά τη διασύνδεση γεωχωρικών δεδομένων για βελτίωση των αποτελεσμάτων (για παράδειγμα τη διασύνδεση/συγχώνευση πηγών γεωκωδικοποίησης).
- **Εκτεταμένη βιβλιοθήκη χαρακτηριστικών εκπαίδευσης για χρήση σε διάφορα είδη δεδομένων και σενάρια**. Ένας από τους βασικούς στόχους μας είναι να υλοποιήσουμε μία ευρεία γκάμα χαρακτηριστικών εκπαίδευσης, τα οποία θα δύνανται να επαναχρησιμοποιηθούν σε διαφορετικούς αλγόριθμους μηχανικής μάθησης, για την επίλυση παραλλαγών ή ακόμα και αρκετά διαφορετικών προβλημάτων από αυτά που εξετάζουμε, στην περιοχή των χωρο-κειμενικών δεδομένων. Στο παραδοτέο Π1.2 «Προδιαγραφή εξειδικευμένων χαρακτηριστικών και κανόνων εκπαίδευσης» προδιαγράφεται το παραπάνω εκτεταμένο και ετερογενές σύνολο χαρακτηριστικών. Στο πλαίσιο του παρόντος παραδοτέου, τα χαρακτηριστικά αυτά υλοποιούνται με αρθρωμένο και επεκτάσιμο τρόπο, έτσι ώστε να καθίστανται αυτόνομα, επεκτάσιμα και πλήρως αξιοποιήσιμα σε τρίτους αλγόριθμους και εφαρμογές.
- **Πλήρης παραμετροποίηση των διαδικασιών εκπαίδευσης και εκτέλεσης των υλοποιημένων μεθόδων**. Οι βιβλιοθήκες μηχανικής μάθησης υλοποιήθηκαν με τέτοιο τρόπο ώστε να είναι όσο το δυνατόν πιο παραμετροποιήσιμες όσον αφορά τις υπερ-παραμέτρους και τους αλγόριθμους μηχανικής μάθησης που εκπαιδεύουν, αξιολογούν συγκριτικά και εφαρμόζουν στα τελικά δεδομένα. Παράλληλα, μέσω εκτενών, δομημένων και τεκμηριωμένων αρχείων παραμετροποίησης (configuration files) προσφέρονται στους τελικούς χρήστες δύο διαφορετικές επιλογές παραμετροποιημένης αξιολόγησης και επιλογής βέλτιστων μοντέλων: (α) Άμεση επιλογή των κατάλληλων τιμών υπερ-παραμέτρων και αλγορίθμων για εξειδικευμένους χρήστες (β) Αυτοματοποιημένη επιλογή των βέλτιστων συνδυασμών υπερ-παραμέτρων και αλγορίθμων, μέσω αναζήτησης πλέγματος (grid search) για ανειδίκευτους σε έννοιες μηχανικής μάθησης χρήστες.
- **Ευέλικτη και αρθρωματοποιημένη υλοποίηση των βιβλιοθηκών σε διακριτά αρθώματα (components)**. Αυτό επιτυγχάνεται σχεδιάζοντας τις βιβλιοθήκες με την μορφή ενός συνόλου αρθρωμάτων, το καθένα από τα οποία είναι υπεύθυνο για μία συγκεκριμένη βασική διεργασία ή μέρος αυτής. Με τη σειρά του, κάθε άρθρωμα αποτελείται από επιμέρους υποσυστήματα (modules) που αναλαμβάνουν

επιμέρους λειτουργικότητες και δύνανται να αξιοποιούνται από διαφορετικά αρθρώματα. Με την παραπάνω λογική, τα υποσυστήματα των βιβλιοθηκών δύνανται να κληθούν ανεξάρτητα το ένα από το άλλο, ανάλογα με τη συγκεκριμένη χρήση που απαιτείται, ενώ είναι εύκολα επεκτάσιμα και προσαρμόσιμα.

Προκειμένου να ικανοποιήσουμε τις παραπάνω αρχές, σχεδιάσαμε μία κοινή, γενική αρχιτεκτονική, η οποία ακολουθείται από όλες τις υλοποιημένες βιβλιοθήκες μηχανικής μάθησης και η οποία ενσωματώνει όλα τα βασικά βήματα μίας πλήρους ροής μηχανικής μάθησης: αξιολόγηση και επιλογή του βέλτιστου αλγόριθμου μηχανικής μάθησης, αξιολόγηση και επιλογή του βέλτιστου μοντέλου μηχανικής μάθησης (αλγόριθμος και υπερ-παράμετροι), εκπαίδευση του επιλεγμένου μοντέλου στα δεδομένα εκπαίδευσης, εκτέλεση του εκπαιδευμένου μοντέλου σε νέα δεδομένα. Οι τέσσερις αυτές διακριτές φάσεις, υλοποιούνται σε τέσσερα διακριτά αρθρώματα, όπως παρουσιάζεται στην αρχιτεκτονική της Εικόνα 1.

Το Άρθρωμα 1 εκτελεί το πρώτο βήμα της διαδικασίας, δηλαδή την εξαντλητική σύγκριση διαφορετικών αλγορίθμων μηχανικής μάθησης και υπερ-παραμετροποιήσεών τους, ώστε να βρεθεί ο βέλτιστος αλγόριθμος, δηλαδή ο αλγόριθμος που επιτυγχάνει τη μέγιστη, κατά μέσο όρο, ακρίβεια στα δεδομένα εκπαίδευσης (αρχική είσοδος). Έξοδος αυτού του αρθρώματος-βήματος, είναι ο επιλεγμένος αλγόριθμος, ο οποίος δίνεται ως είσοδος στο Άρθρωμα 2. Στο συγκεκριμένο άρθρωμα εκτελείται εξαντλητική σύγκριση διαφορετικών υπερ-παραμετροποιήσεων του επιλεγμένου αλγορίθμου, ούτως ώστε να επιλεγεί η βέλτιστη, η οποία μαζί με τον αλγόριθμο συγκροτεί το βέλτιστο μοντέλο. Αυτό δίνεται ως είσοδος στο Άρθρωμα 3, το οποίο εκπαιδεύει το επιλεγμένο μοντέλο στο σύνολο των δεδομένων εκπαίδευσης που έχει δοθεί ως είσοδος στη ροή μηχανικής μάθησης. Έξοδος του αρθρώματος αποτελεί το εκπαιδευμένο μοντέλο, το οποίο δύναται πλέον να χρησιμοποιηθεί σε νέα δεδομένα, στο Άρθρωμα 4, προς επίλυση του εκάστοτε προβλήματος.

Κάθε ένα από τα αρθρώματα δύναται να κληθεί με ανεξάρτητη κλήση API, ανάλογα με την εκάστοτε ανάγκη εκτέλεσης του χρήστη. Για παράδειγμα, ο χρήστης δύναται να τρέξει μόνο μία φορά το βήμα-‘Αρθρωμα 1, και στη συνέχεια να τρέχει μόνο τα υπόλοιπα βήματα όταν προκύπτει ένα νέο σύνολο δεδομένων εκπαίδευσης.

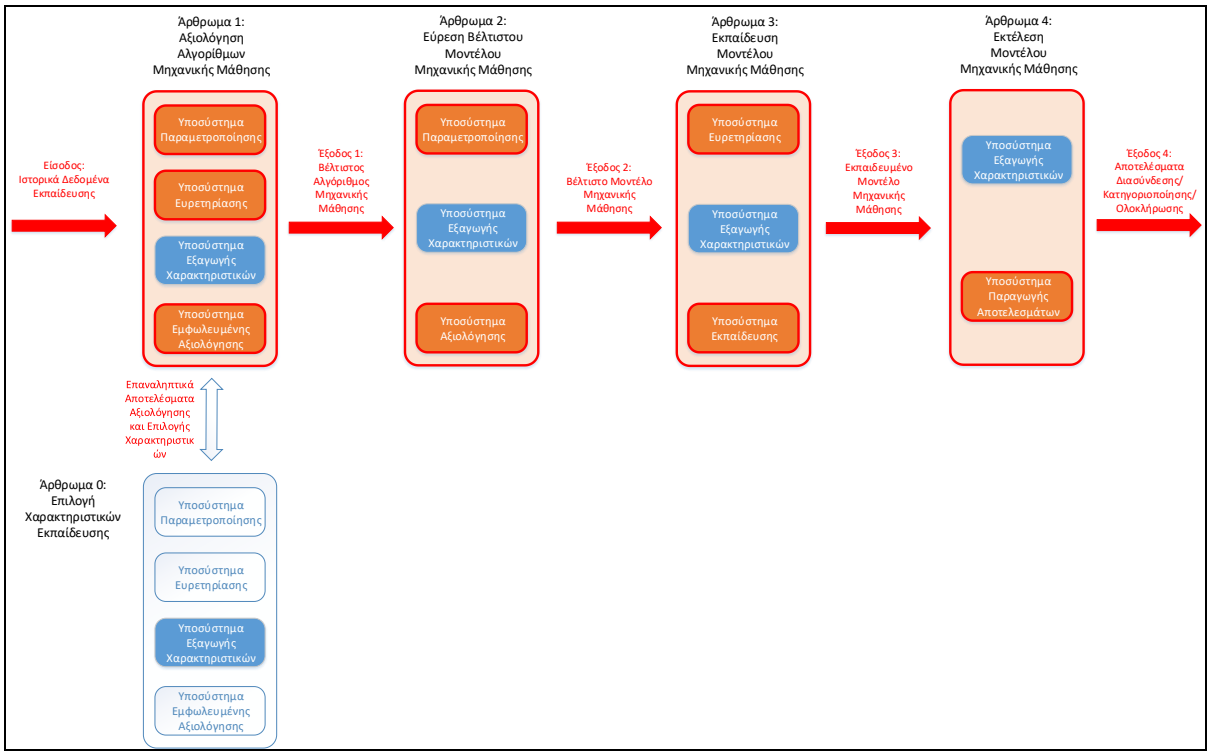
Κάθε άρθρωμα αποτελείται από επιμέρους υποσυστήματα (modules) τα οποία υλοποιούν επιμέρους λειτουργικότητα. Κάθε υλοποιημένη βιβλιοθήκη περιέχει μία ομάδα από υποσυστήματα που υλοποιούν τη συγκεκριμένη λειτουργικότητα του προβλήματος-ροής μηχανικής μάθησης που υλοποιεί. Παρόλα αυτά, όλες οι βιβλιοθήκες μοιράζονται ένα σύνολο από κοινά υποσυστήματα, τα οποία υλοποιούν βασικές για όλους τους αλγορίθμους λειτουργικότητες. Εν συντομία, τα υποσυστήματα αυτά είναι:

- *Υποσύστημα Παραμετροποίησης*: Το υποσύστημα αυτό αποτελεί το κύριο μέσο μέσω του οποίου ο χρήστης καθορίζει τη λειτουργικότητα της βιβλιοθήκης αφού παρέχει τη δυνατότητα επιλογής των τιμών μιας εκτενούς συλλογής (υπερ)παραμέτρων οι οποίες καθορίζουν σημαντικά βήματά της, από τη διαδικασία εξαγωγής χαρακτηριστικών ως και το επίπεδο κατηγοριών στο οποίο θα ανήκουν οι κατηγορίες κατάταξης στα πειράματα.
- *Υποσύστημα Εξαγωγής Χαρακτηριστικών*: Το υποσύστημα αυτό υποστηρίζει τη διαδικασία εξαγωγής κειμενικών και γεωχωρικών χαρακτηριστικών εκπαίδευσης, τα

οποία αποτυπώνουν ουσιαστική πληροφορία των χωρο-κειμενικών δεδομένων εισόδου που χρησιμοποιείται στην εκπαίδευση των αλγορίθμων μηχανικής μάθησης.

- *Υποσύστημα Ευρετηρίασης Δεδομένων*: Το υποσύστημα αυτό υποστηρίζει τη χρήση ευρετηρίων για την καλύτερη οργάνωση των γεωχωρικών και κειμενικών ιδιοτήτων των δεδομένων που χρησιμοποιεί η εκάστοτε βιβλιοθήκη για τις διάφορες λειτουργίες της, με σκοπό την επιτάχυνση της εκτέλεσής της. Αυτό επιτυγχάνεται μέσω προσεγγίσεων ευρετηρίασης που χρησιμοποιούν δενδρικά ευρετήρια (R-Tree, KD-Tree) και ανεστραμμένα ευρετήρια αντίστοιχα.
- *Υποσύστημα Εμφωλευμένης Αξιολόγησης*: Το υποσύστημα αυτό αφορά το πρώτο στάδιο των πειραμάτων και περιλαμβάνει την εκπαίδευση μιας συλλογής από αλγορίθμους κατάταξης χρησιμοποιώντας εμφωλευμένη συγκριτική αξιολόγηση (nested cross-validation), ώστε να βρεθεί εκείνος με την καλύτερη απόδοση, μέσω συνεχών διαφοροποιήσεων των υπερ-παραμέτρων τους. Από το βήμα αυτό προκύπτει ο βέλτιστος αλγόριθμος ο οποίος πρόκειται να χρησιμοποιηθεί στα επόμενα στάδια.
- *Υποσύστημα Αξιολόγησης*: Το υποσύστημα αυτό αφορά το δεύτερο στάδιο των πειραμάτων και περιλαμβάνει αναζήτηση του χώρου των υπερ-παραμέτρων του βέλτιστου αλγορίθμου που έχει προκύψει από την εκτέλεση του πρώτου σταδίου. Αυτό επιτυγχάνεται με τη χρήση συγκριτικής αξιολόγησης (cross-validation), από την οποία προκύπτει ο καλύτερος συνδυασμός υπερ-παραμέτρων.
- *Υποσύστημα Εκπαίδευσης*: Το υποσύστημα αυτό αφορά το τρίτο στάδιο των πειραμάτων και περιλαμβάνει την εκπαίδευση του καλύτερου αλγόριθμου κατάταξης ρυθμισμένου με τις καλύτερες υπερ-παραμέτρους (δηλαδή του βέλτιστου μοντέλου), στοιχεία που προέκυψαν από το πρώτο και το δεύτερο στάδιο των πειραμάτων και την αποθήκευσή του σε αρχείο με την κατάλληλη μορφή, για χρήση στην επίλυση του εκάστοτε προβλήματος.
- *Υποσύστημα Παραγωγής Αποτελεσμάτων*: Το υποσύστημα αυτό αφορά το τέταρτο στάδιο των πειραμάτων και περιλαμβάνει την παροχή αποτελεσμάτων διασύνδεσης, κατηγοριοποίησης, γεωκωδικοποίησης ή ολοκλήρωσης. Αναγκαία για τη λειτουργικότητα του υποσυστήματος αυτού είναι η ύπαρξη αρχείου στο οποίο βρίσκεται αποθηκευμένο ένα ήδη εκπαιδευμένο μοντέλο μηχανικής μάθησης, παραγμένο από το Υποσύστημα Εκπαίδευσης.

Στην Εικόνα 1 τονίζεται με διαφορετικό χρωματισμό το Υποσύστημα Εξαγωγής Χαρακτηριστικών, το οποίο, παρόλο που εμφανίζεται σε διάφορα αρθρώματα, αποτελεί ένα αρκετά ανεξάρτητο υποσύστημα, το οποίο θα μετασχηματιστεί σε ξεχωριστό άρθρωμα στις επόμενες φάσεις του έργου. Επιπλέον, με διαφορετικό χρωματισμό εμφανίζεται το Άρθρωμα Επιλογής Χαρακτηριστικών Εκπαίδευσης, το οποίο υλοποιείται στο πλαίσιο της Υποενότητας Εργασίας 3.1 και θα παρουσιαστεί στο Παραδοτέο 3.1. Το συγκεκριμένο άρθρωμα θα υποστηρίζει αυτοματοποιημένη επιλογή βέλτιστων χαρακτηριστικών, διαδικασία η οποία αυτή τη στιγμή επιτελείται χειροκίνητα μέσω ενός αρχείου παραμετροποίησης (configuration file).



Εικόνα 1: Αρχιτεκτονική Υλοποιημένων Βιβλιοθηκών Μηχανικής Μάθησης

3. Βιβλιοθήκες μηχανικής μάθησης

Ακολούθως, παρουσιάζονται οι τέσσερις βιβλιοθήκες μηχανικής μάθησης που υλοποιήθηκαν, στο πλαίσιο της υποενότητας εργασίας ΥΕ 2.1 του έργου, οι οποίες αντιστοιχούν στα τέσσερα βασικά σενάρια χρήσης, όπως παρουσιάζονται στο παραδοτέο Π1.1 «Προδιαγραφή περιπτώσεων χρήσης, βασικών δεικτών απόδοσης και συνόλου αναφοράς αξιολόγησης».

Σημειώνουμε ότι, λόγω της κοινής αρχιτεκτονικής που ακολουθήθηκε, υπάρχει μία αναπόφευκτη επανάληψη συγκεκριμένων κομματιών των τεσσάρων βιβλιοθηκών, για παράδειγμα κάποιων υποσυστημάτων ή των αλγορίθμων μηχανικής μάθησης που αξιολογούνται. Προκειμένου να αποφευχθεί μερικώς η παραπάνω επικάλυψη, παρακάτω περιγράφουμε συνοπτικά το σύνολο των αλγορίθμων μηχανικής μάθησης που ενσωματώθηκαν και αξιολογήθηκαν στις τέσσερις βιβλιοθήκες που υλοποιήθηκαν:

- *K-Nearest Neighbors (k-NN)*: Ο k-NN είναι ένας μη-παραμετρικός αλγόριθμος κατάταξης σύμφωνα με τον οποίο κάθε παράδειγμα στο σύνολο δεδομένων ταξινομείται στην κλάση στην οποία ανήκει η πλειοψηφία των k κοντινότερων γειτόνων του. Με τον όρο «k κοντινότεροι γείτονες» εννοούμε τα παραδείγματα εκείνα τα οποία απέχουν τη μικρότερη απόσταση, όπως αυτή ορίζεται στον αλγόριθμο, από το εξεταζόμενο παράδειγμα.
- *Support Vector Machines (SVM)*: Ο SVM είναι ένας αλγόριθμος κατάταξης σύμφωνα με τον οποίο τα παραδείγματα που βρίσκονται στο σύνολο δεδομένων αναπαρίστανται ως σημεία στο χώρο με τέτοιον τρόπο, ώστε να μεγιστοποιείται η απόσταση μεταξύ των παραδειγμάτων εκείνων που ανήκουν σε διαφορετικές κατηγορίες, μέσω μίας βέλτιστης διαχωριστικής επιφάνειας που υπολογίζεται από την εκπαίδευση του αλγορίθμου. Νέα παραδείγματα στη συνέχεια αντιστοιχούνται στον ίδιο χώρο, και η συγκεκριμένη επιφάνεια καθορίζει την κατάταξή τους.
- *Decision Trees (DT)*: Τα DT ή δένδρα απόφασης είναι ένας αλγόριθμος κατάταξης ο οποίος χρησιμοποιεί μια γραφική απεικόνιση όμοια της μορφής δένδρου, συμπεριλαμβάνοντας όλες τις πιθανές αποφάσεις, όλους τους παράγοντες επιρροής και όλα τα πιθανά αποτελέσματα και αποσκοπώντας στη σωστή κατάταξη των παραδειγμάτων που βρίσκονται σε ένα σύνολο δεδομένων.
- *Random Forests (RF)*: Τα RF αποτελούν μια ειδική κατηγορία των συνδυαστικών μεθόδων κατάταξης η οποία χρησιμοποιεί επιμέρους δένδρα απόφασης. Η διαδικασία κατάταξης παραδειγμάτων πραγματοποιείται μέσω της διάσχισης των δένδρων του δάσους ξεκινώντας από τη ρίζα και καταλήγοντας σε ένα από τα φύλλα του δένδρου και στη συνέχεια συνδυάζοντας τις προβλέψεις των επιμέρους δένδρων απόφασης βάσει ενός πλειοψηφικού συστήματος ψηφοφορίας. Κάθε παράδειγμα ανατίθεται στην πλειοψηφούσα κατηγορία.
- *Adaboost*: Ο Adaboost αλγόριθμος κατάταξης είναι μια εκ των συνδυαστικών μεθόδων η οποία χρησιμοποιείται σε συνδυασμό με άλλα είδη αλγορίθμων μάθησης ώστε να βελτιώσει την απόδοσή τους. Ο τελικός αλγόριθμος κατάταξης προκύπτει μέσα από το συνδυασμό των επιμέρους αλγορίθμων μάθησης (weak learners) μέσω ενός αθροίσματος βαρύτητας.
- *Naive Bayes* (NB): Ο NB είναι ένας αλγόριθμος κατάταξης ο οποίος βασίζεται στον υπολογισμό της εκ των υστέρων πιθανότητας, όπως υπολογίζεται από τον κανόνα

του Bayes, μοντελοποιώντας την πιθανοτική σχέση μεταξύ του συνόλου χαρακτηριστικών και της κατηγορίας. Συγκεκριμένα, δοθέντων των τιμών των χαρακτηριστικών ενός νέου παραδείγματος, στόχος του NB είναι να υπολογίσει τις υπό συνθήκη πιθανότητες για όλες τις πιθανές κατηγορίες και να αναθέσει το κάθε παράδειγμα στην κατηγορία για την οποία η αναμενόμενη πιθανότητα σφάλματος ελαχιστοποιείται.

- *Multi-layer Perceptron* (MLP): Τα MLP είναι ένας αλγόριθμος κατάταξης ο οποίος αντιπροσωπεύει την απλούστερη εκδοχή των νευρωνικών δικτύων. Στόχος του αλγορίθμου είναι να καθορίσει τα βάρη των συνδέσεων μεταξύ των νευρώνων με στόχο να μειώσει έτσι το ποσοστό σφάλματος κατάταξης. Κάθε παράδειγμα κατατάσσεται εφαρμόζοντας τις τιμές των χαρακτηριστικών του στην είσοδο του νευρωνικού δικτύου, το οποίο στη συνέχεια καθορίζει το αποτέλεσμα κατάταξης.
- *Gaussian Process*: Ο Gaussian Process είναι ένας αλγόριθμος κατάταξης ο οποίος βασίζεται στη χρήση μιας Γκαουσιανής διαδικασίας η οποία, σε συνδυασμό με τεχνικές lazy learning και μιας μετρικής ομοιότητας μεταξύ σημείων, οδηγεί σε προβλέψεις σχετικά με την κατηγορία στην οποία ανήκει το κάθε παράδειγμα σε ένα σύνολο δεδομένων.
- *Extra Trees*: Ο Extra Trees είναι ένας αλγόριθμος κατάταξης ο οποίος μοιράζεται πολλά κοινά στοιχεία με τον Random Forests, με την κύρια διαφορά να βρίσκεται στο γεγονός ότι τα τελικά δένδρα απόφασης δεν επιλέγονται βάσει κάποιου είδους ψηφοφορίας, αλλά τυχαία.
- *eXtreme Gradient Boosting (XGBoost)*: Ο XGBoost είναι ένας αλγόριθμος κατάταξης ο οποίος, όπως και το Extra Trees, μοιράζεται πολλά κοινά στοιχεία με τον Random Forest, με τη διαφορά ότι τα τελικά δένδρα απόφασης κατασκευάζονται διαδοχικά, προσθέτοντας ένα δέντρο κατάταξης σε κάθε βήμα, με σκοπό την βελτίωση των προβλέψεων του προηγούμενου δέντρου κατάταξης. Το πλεονέκτημα αυτού του αλγορίθμου είναι ότι επιτυγχάνει χαμηλή μεροληψία ως προς τις σχέσεις μεταξύ των χαρακτηριστικών εκπαίδευσης και των στόχων εξόδου.

Ακολούθως, σε κάθε επιμέρους υποενότητα, γίνεται απλά απαρίθμηση των συγκεκριμένων αλγορίθμων μηχανικής μάθησης που ενσωματώθηκαν στην αντίστοιχη βιβλιοθήκη. Αντιθέτως, για να διατηρηθεί η πληρότητα της τεκμηρίωσης, για κάθε επιμέρους βιβλιοθήκη περιγράφουμε το σύνολο των βασικών υποσυστημάτων που την απαρτίζουν.

3.1. Βιβλιοθήκη Διασύνδεσης Τοπωνυμίων

3.1.1. Σύντομη περιγραφή

Η βιβλιοθήκη LGM-Interlinking υλοποιεί μία πλήρη ακολουθία διεργασιών εκπαίδευσης και εκτέλεσης μοντέλων μηχανικής μάθησης με στόχο την αποδοτική επίλυση του προβλήματος της διασύνδεσης τοπωνυμίων μέσω δυαδικής κατάταξης (binary classification). Οι διεργασίες αυτές περιλαμβάνουν την υλοποίηση ευρείας συλλογής από χαρακτηριστικά εκπαίδευσης σχετικά με την ομοιότητα των συμβολοσειρών σε ζεύγη υποψήφιων τοπωνυμίων και τεχνικές αναζήτησης πλέγματος (grid-search) και συγκριτικής αξιολόγησης (cross-validation) για την αξιολόγηση μιας σειράς διαφορετικών μοντέλων μηχανικής μάθησης για κατάταξη, για την κατασκευή του αποδοτικότερου μοντέλου για τα δεδομένα που εξετάζουμε. Η εκπαίδευση και αξιολόγηση των διαφόρων μοντέλων μηχανικής μάθησης γίνεται σε επισημειωμένα σύνολα δεδομένων που αφορούν ζεύγη υποψήφιων τοπωνυμίων ως προς το αν αντιπροσωπεύουν ίδιες οντότητες ή όχι.

Η βιβλιοθήκη LGM-Interlinking παρέχεται δωρεάν και ο πηγαίος κώδικας είναι διαθέσιμος στο GitHub (https://github.com/LinkGeoML/LGM-Interlinking). Μπορεί να διανεμηθεί και να τροποποιηθεί σύμφωνα με τους όρους της μη περιοριστικής MIT άδειας[2].

3.1.2. Αλγόριθμοι μηχανικής μάθησης

Η αναζήτηση του καλύτερου μοντέλου μηχανικής μάθησης γίνεται μεταξύ των ακόλουθων αλγορίθμων αιχμής:

- Support Vector Machines
- Decision Trees
- Random Forests
- Multi-layer Perceptron
- Extra Trees
- eXtreme Gradient Boosting

3.1.3. Πληροφορίες υλοποίησης και τεκμηρίωση

Η βιβλιοθήκη LGM-Interlinking είναι υλοποιημένη στη γλώσσα Python. Οι λειτουργίες μηχανικής μάθησης καλύπτονται κυρίως από τη βιβλιοθήκη scikit-learn[3], καθώς και την xgboost[4] για την αποδοτική υλοποίηση του αλγορίθμου eXtreme Gradient Boosting. Όσον αφορά τις μετρικές ομοιότητας, γίνεται χρήση των βιβλιοθηκών jellyfish[5], για τις μετρικές Jaro και Jaro-Winkler, και pyxdameraulevenshtein[6], για την μετρική Damerau-Levenshtein.

[2] https://opensource.org/licenses/MIT

[3] https://scikit-learn.org/stable/

[4] https://xgboost.readthedocs.io/en/latest/

[5] https://pypi.org/project/jellyfish/

[6] https://pypi.org/project/pyxDamerauLevenshtein/

Τέλος, η αποδοτική διαχείριση και ανάλυση στα σύνολα δεδομένων γίνεται με τις βιβλιοθήκες pandas[7], numpy[8] και scipy[9].

Στη διεύθυνση https://linkgeoml.github.io/LGM-Interlinking/ υπάρχει αναλυτική τεκμηρίωση των διαφόρων υποστηριζόμενων μεθόδων (API) της βιβλιοθήκης LGM-Interlinking.

3.1.4. Βασικά υποσυστήματα

Τα βασικά υποσυστήματα που απαρτίζουν τη βιβλιοθήκη LGM-Interlinking είναι τα εξής:

- Διεπαφή Γραμμής Εντολών: Το υποσύστημα αυτό λαμβάνει παραμέτρους εισόδου για σύνολα δεμένων από τοπωνύμια τα οποία θα χρησιμοποιηθούν από τα διάφορα στάδια εκτέλεσης της βιβλιοθήκης. Συγκεκριμένα, ο χρήστης μπορεί να καθορίσει τη διαδρομή των αρχείων που θα χρησιμοποιηθούν για την εκπαίδευση και τον έλεγχο των αλγορίθμων μηχανικής μάθησης, καθώς και το αλφάβητο χαρακτήρων που χρησιμοποιείται σε αυτά.
- Εξωτερική Ρύθμιση Παραμέτρων: Ο σκοπός αυτού του υποσυστήματος είναι να επιτρέπει στον χρήστη να καθορίζει τη λειτουργικότητα της βιβλιοθήκης. Έτσι, παρέχει τη δυνατότητα επιλογής των τιμών μιας εκτενούς συλλογής παραμέτρων, οι οποίες καθορίζουν σημαντικά βήματά της εκτέλεσης των διαδικασιών μηχανικής μάθησης, όπως την ομάδα των χαρακτηριστικών που θα επιλέξουμε, το εύρος τιμών και τον τύπο πλέγματος αναζήτησης που θα χρησιμοποιηθούν για την εύρεση των βέλτιστων υπερ-παραμέτρων.
- Μετρικές Ομοιότητας: Το υποσύστημα αυτό υλοποιεί τις διάφορες μετρικές ομοιότητας που χρησιμοποιούνται για την κατασκευή των υποστηριζόμενων ομάδων χαρακτηριστικών που σχετίζονται με τα τοπωνύμια.
- Εξαγωγή Χαρακτηριστικών: Το υποσύστημα αυτό υποστηρίζει τη διαδικασία εξαγωγής κειμενικών χαρακτηριστικών τα οποία περιγράφουν τη σχέση ζευγαριών από επισημειωμένα τοπωνύμια και θα χρησιμοποιηθούν ως είσοδος για τα μοντέλα κατάταξης στα επόμενα βήματα.
- Επιλογή Αλγορίθμου Κατάταξης: Το υποσύστημα αυτό αφορά το πρώτο στάδιο των πειραμάτων και περιλαμβάνει την εκπαίδευση μιας συλλογής από αλγορίθμους κατάταξης, χρησιμοποιώντας εμφωλευμένη συγκριτική αξιολόγηση (nested cross-validation), ώστε να βρεθεί εκείνος με την καλύτερη απόδοση, μέσω συνεχών διαφοροποιήσεων των υπερ-παραμέτρων τους. Από το βήμα αυτό προκύπτει ο βέλτιστος αλγόριθμος, ο οποίος πρόκειται να χρησιμοποιηθεί στα επόμενα στάδια.
- Βελτιστοποίηση Αλγορίθμου Κατάταξης: Το υποσύστημα αυτό αφορά το δεύτερο στάδιο των πειραμάτων και περιλαμβάνει αναζήτηση του χώρου των υπερ-παραμέτρων του βέλτιστου αλγορίθμου που έχει προκύψει από την εκτέλεση του

[7] https://pandas.pydata.org/

[8] https://www.numpy.org/

[9] https://www.scipy.org/

πρώτου σταδίου. Αυτό επιτυγχάνεται με τη χρήση συγκριτικής αξιολόγησης (cross-validation), από την οποία προκύπτει ο καλύτερος συνδυασμός υπερ-παραμέτρων.

- Κατασκευή Μοντέλου Κατάταξης: Το υποσύστημα αυτό αφορά το τρίτο στάδιο των πειραμάτων και περιλαμβάνει την εκπαίδευση του καλύτερου μοντέλου κατάταξης στο σύνολο των δεδομένων εκπαίδευσης. Το μοντέλο αυτό μπορεί να χρησιμοποιηθεί στη συνέχεια για την κατάταξη ζευγών υποψήφιων τοπωνυμίων από νέα σύνολα δεδομένων.
- Έλεγχος Μοντέλου Κατάταξης: Το υποσύστημα αυτό αφορά το τέταρτο στάδιο των πειραμάτων και περιλαμβάνει την κατάταξη ζεύγων υποψήφιων τοπωνυμίων (από νέο συνόλων δεδομένων, το οποίο παρέχει ο χρήστης κατά την εκτέλεση) ως προς το αν αντιπροσωπεύουν ίδιες οντότητες ή όχι.
- Κύρια Ακολουθία Διεργασιών: Το υποσύστημα αυτό υλοποιεί όλα τα στάδια που απαρτίζουν την πλήρη ακολουθία διεργασιών εκπαίδευσης και εκτέλεσης μοντέλων μηχανικής μάθησης. Εκτός από τα βασικά τέσσερα στάδια πειραμάτων που περιγράφηκαν παραπάνω, στη διαδικασία αυτή περιλαμβάνεται η φόρτωση των κατάλληλων συνόλων δεδομένων που απαιτούνται και η αποδοτική κατασκευή των χαρακτηριστικών εκπαίδευσης που έχουν επιλεγεί. Τα αποτελέσματα κάθε φάσης εμφανίζονται στο χρήστη μέσα από τη γραμμή εντολών.

3.1.5. Οδηγός χρήσης

Η εκτέλεση της βιβλιοθήκη LGM-Interlinking υποστηρίζεται μέσα από γραμμή εντολών, καθώς και αντίστοιχο API[10]. Ο χρήστης εισάγει το σύνολο δεδομένων από τοπωνύμια για το οποίο ενδιαφέρεται να εξετάσει την απόδοση διαφόρων αλγορίθμων μηχανικής μάθησης. Η βιβλιοθήκη επιστρέφει, για το συγκεκριμένο σύνολο δεδομένων, τον καλύτερο αλγόριθμο σε σχέση με την ακρίβεια των αποτελεσμάτων που επιτυγχάνει, τις βέλτιστες υπερ-παραμέτρους του, διάφορες μετρικές που ποσοτικοποιούν αυτό το αποτέλεσμα και αναλυτική πληροφόρηση της χρονικής διάρκειας κάθε βήματος της διαδικασίας. Ακολουθεί αναλυτική περιγραφή των βημάτων που απαιτούνται για την εκτέλεση της βιβλιοθήκης.

3.1.5.1. Εγκατάσταση

Προαπαιτούμενα/εξαρτήσεις

Για την απρόσκοπτη εκτέλεση της βιβλιοθήκης, απαιτείται η εγκατάσταση των παρακάτω βιβλιοθηκών:

- jellyfish - 0.6.1
- numpy - 1.14.3
- pandas - 0.23.0
- pyxDamerauLevenshtein - 1.4.1
- scikit-learn - 0.20.3
- scipy - 1.2.1
- xgboost - 0.82

[10] https://linkgeoml.github.io/LGM-Interlinking/

- alphabet-detector - 0.0.7
- docopt - 0.6.2
- text-unidecode - 1.2
- kitchen - 1.2.5
- pycountry_convert - 0.7.2

Οι παραπάνω βιβλιοθήκες περιέχονται στο αρχείο pip_requirements.txt και η εγκατάστασή τους γίνεται ως εξής:

```
$ pip install –r pip_requirements.txt
```

Οδηγίες εγκατάστασης

Αρχικά ελέγχουμε εάν είναι εγκατεστημένη η Python 2.7 στη γραμμή εντολών:

```
$ python
Python 2.7.15 (default, Mar 26 2019, 21:43:19)
[GCC 7.3.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
```

Το παραπάνω μήνυμα δείχνει ότι η python έχει εγκατασταθεί σωστά. Στην περίπτωση που αυτό δεν ισχύει, προχωράμε στην εγκατάστασή της προτεινόμενης έκδοσης με βάση το λειτουργικό σύστημα που χρησιμοποιούμε.

Προτείνεται, χωρίς να είναι απαραίτητο, να δημιουργήσουμε ένα εικονικό περιβάλλον που θα φιλοξενεί τις διάφορες βιβλιοθήκες και τις εκδόσεις τους που είναι απαραίτητες για τη λειτουργία της βιβλιοθήκης LGM-Interlinking, το οποίο θα είναι απομονωμένο από το υπόλοιπο λειτουργικό σύστημα. Αυτό γίνεται ως εξής:

```
$ virtualenv -p `which python2.7` <path/to/new/virtualenv/>
$ source <path/to/new/virtualenv/>/bin/activate
```

Έχοντας ενεργοποιήσει το εικονικό περιβάλλον, μπορούμε να εγκαταστήσουμε τα προαπαιτούμενα:

```
$ pip install –r pip_requirements.txt
```

Κατεβάζουμε την τελευταία έκδοση του πηγαίου κώδικα της LGM-Interlinking βιβλιοθήκης:

```
$ git clone https://github.com/LinkGeoML/LGM-Interlinking.git
$ cd LGM-Interlinking
$ python run.py --version
LGM-Interlinking 0.1.0
```

Εάν τυπωθεί το παραπάνω στην γραμμή εντολών, τότε έχουμε εγκαταστήσει σωστά την LGM-Interlinking βιβλιοθήκη. Τέλος, η εκτέλεση μίας πλήρους ακολουθίας διεργασιών εκπαίδευσης και εκτέλεσης μοντέλων μηχανικής μάθησης γίνεται ως εξής:

```
$ python run.py --dtrain <path/to/train-dataset> --dtest <path/to-test-
dataset>
```

3.1.5.2. Παραμετροποίηση

Στο αρχείο config.py, της βιβλιοθήκης LGM-Interlinking, υπάρχουν μια σειρά από πεδία, εμφωλευμένα στην κλάση MLConf, που δίνουν τη δυνατότητα παραμετροποίησης διαφόρων λειτουργιών της. Τα πεδία αυτά είναι τα ακόλουθα:

- *k_fold_parameter*: η εξωτερική διαμέριση των δεδομένων, στο πλαίσιο της διαδικασίας k-fold cross-validation, σε δύο υποσύνολα, όπου το ένα χρησιμοποιείται για εκπαίδευση και το άλλο για τον έλεγχο της απόδοσης του μοντέλου.
- *k_fold_inner_parameter*: η εσωτερική διαμέριση του υποσυνόλου δεδομένων που έχει προκύψει στο πλαίσιο της διαδικασίας k-fold cross-validation και χρησιμοποιείται για την εκπαίδευση του μοντέλου. Η δεύτερη αυτή διαμέριση επιτρέπει τη βέλτιστη επιλογή υπερ-παραμέτρων κατά την εκπαίδευση.
- *classification_method*: δηλώνεται η ομάδα από χαρακτηριστικά εκπαίδευσης που επιθυμούμε, όπως έχουν περιγραφεί στο Π1.2. Οι διαθέσιμες έγκυρες επιλογές είναι: *basic* - ομοιότητα των αρχικών συμβολοσειρών, *basic*_sorted - ομοιότητα των ταξινομημένων συμβολοσειρών και *lgm* - ομοιότητα των εξειδικευμένα προ-επεξεργασμένων συμβολοσειρών.
- *hyperparams_search_method*: παράμετρος για τον τρόπο αναζήτησης βέλτιστων υπερ-παραμέτρων. Οι διαθέσιμες επιλογές είναι: *grid* - , *randomized* -
- *max_iter*: ο αριθμός των διαφορετικών συνδυασμών υπερ-παραμέτρων που εξετάζονται. Η παράμετρος αυτή έχει ισχύ όταν στη *hyperparams_search_method* έχει ανατεθεί η τιμή *randomized*.
- *n_jobs*: ο αριθμός των διεργασιών που τρέχουν παράλληλα.
- *{SVM,DecisionTree,RandomForest,MLP}_hyperparameters*: λεξικό που περιέχει αντιστοιχίες μεταξύ των ονομάτων των υπερ-παραμέτρων και της συλλογής από τιμές που θέλουμε να διερευνήσουμε στα πλαίσια της *grid* αναζήτησης για την εκπαίδευση του μοντέλου SVM, Decision Tree, Random Forest και Multi-Layer Perceptron. Οι υπερ-παράμετροι που έχουν δηλωθεί κάτω από το πεδίο RandomForest_hyperparameters χρησιμοποιούνται και από το συγγενικό μοντέλο Extra Trees.
- *{SVM,DecisionTree,RandomForest,MLP}_hyperparameters_dist*: λεξικό που περιέχει αντιστοιχίες μεταξύ των ονομάτων των υπερ-παραμέτρων και της συλλογής από συνεχόμενες κατανομές τιμών που θέλουμε να διερευνήσουμε στα πλαίσια της *randomized* αναζήτησης για την εκπαίδευση του μοντέλου SVM, Decision Tree, Random Forest και Multi-Layer Perceptron. Οι υπερ-παράμετροι που έχουν δηλωθεί κάτω από το πεδίο RandomForest_hyperparameters χρησιμοποιούνται και από το συγγενικό μοντέλο Extra Trees.

3.1.5.3. Εκτέλεση

Για τη λειτουργία της βιβλιοθήκης απαιτείται η παροχή ενός Tab-Separated αρχείου το οποίο θα περιλαμβάνει μια συλλογή από πληροφορίες για ένα ή παραπάνω ζεύγη τοπωνυμίων. Συγκεκριμένα το αρχείο πρέπει να περιέχει, τουλάχιστον, τα εξής πεδία/στήλες:

- Το όνομα του πρώτου τοπωνυμίου.

- Το όνομα του δεύτερου τοπωνυμίου.
- Επισημείωση ως προς τη διασύνδεση των δύο τοπωνυμίων, δηλαδή αν αναφέρονται στην ίδια οντότητα ή όχι ({True, False}).

Ακολουθεί η περιγραφή του συνόλου των λειτουργιών που καλύπτονται από τη βιβλιοθήκη, οι οποίες συνίστανται σε τέσσερα ξεχωριστά στάδια. Να σημειώσουμε ότι όλες οι συναρτήσεις που περιγράφονται παρακάτω ανήκουν στην κλάση *ParamTuning*, στο αρχείο *src/param_tuning.py* στο GitHub.

Αξιολόγηση/επιλογή αλγορίθμου

Η συνάρτηση *getBestClassifier* είναι υπεύθυνη για την εύρεση του καλύτερου μοντέλου μηχανικής μάθησης για το σύνολο δεδομένων που εξετάζουμε. Τα δεδομένα εισόδου της συνάρτησης είναι:

1. *X:* Χαρακτηριστικά εκπαίδευσης σε μορφή πίνακα που χρησιμοποιούνται για την εκπαίδευση των μοντέλων και έχουν προκύψει από διαμέριση (fold).
2. *y:* Επισημειωμένα δεδομένα σε μορφή πίνακα που αντιστοιχούν στα χαρακτηριστικά εκπαίδευσης X και έχουν προκύψει από διαμέριση (fold).

Για την έξοδο της συνάρτησης έχουμε ένα λεξικό με τους εξής όρους:

- *accuracy*: το σκορ ομοιότητας που πέτυχε το καλύτερο μοντέλο
- *classifier*: το όνομα του καλύτερου μοντέλου

Βελτιστοποίηση αλγορίθμου

Το στάδιο αυτό υλοποιείται από τη συνάρτηση *fineTuneClassifier* και παίρνει ως είσοδο:

1. *X:* Χαρακτηριστικά εκπαίδευσης σε μορφή πίνακα που χρησιμοποιούνται για την εκπαίδευση των μοντέλων και έχουν προκύψει από διαμέριση (fold).
2. *y:* Επισημειωμένα δεδομένα σε μορφή πίνακα που αντιστοιχούν στα χαρακτηριστικά εκπαίδευσης X και έχουν προκύψει από διαμέριση (fold).
3. *best_clf*: το λεξικό όρων που έχει προκύψει σαν έξοδος της προηγούμενης συνάρτησης, *getBestClassifier*.

Η συνάρτηση *fineTuneClassifier* επιστρέφει τα εξής ορίσματα:

1. Ένα μοντέλο του αλγορίθμου *best_clf* με τις βέλτιστες υπερ-παραμέτρους που έχουν βρεθεί.
2. Τις βέλτιστες υπερ-παραμέτρους σε μορφή λεξικού.
3. Το accuracy σκορ που πετυχαίνει.

Εξαγωγή Μοντέλου Κατάταξης

Η *trainClassifier* εκπαιδεύει το μοντέλο που έχει προκύψει στο προηγούμενο στάδιο σε όλο το σύνολο δεδομένων που έχει επιλεγεί για λόγους εκπαίδευσης, δηλαδή δεν γίνεται χρήση διαμερίσεων (folds). Η είσοδος είναι:

1. *X_train:* Χαρακτηριστικά εκπαίδευσης σε μορφή πίνακα που χρησιμοποιούνται για την εκπαίδευση των μοντέλων, χωρίς διαμέριση (fold).
2. *Y_train:* Επισημειωμένα δεδομένα σε μορφή πίνακα που αντιστοιχούν στα χαρακτηριστικά εκπαίδευσης X, χωρίς διαμέριση (fold).
3. *model*: το μοντέλο με βέλτιστες υπερ-παραμέτρους από τη συνάρτηση *getBestClassifier*.

Στην έξοδο της συνάρτησης παίρνουμε:

- Ένα εκπαιδευμένο μοντέλο στο σύνολο των δεδομένων εκπαίδευσης (χωρίς folds).

Εξαγωγή προβλέψεων σε νέα σύνολα δεδομένων

Ο έλεγχος ενός εκπαιδευμένου μοντέλου σε νέα σύνολα δεδομένων γίνεται από τη συνάρτηση *testClassifier*. Η είσοδος παίρνει τα εξής ορίσματα:

1. *X_train:* Χαρακτηριστικά εκπαίδευσης σε μορφή πίνακα που χρησιμοποιούνται για την εκπαίδευση των μοντέλων, χωρίς διαμέριση (fold).
2. *Y_train:* Επισημειωμένα δεδομένα σε μορφή πίνακα που αντιστοιχούν στα χαρακτηριστικά εκπαίδευσης X, χωρίς διαμέριση (fold).
3. *model*: ένα εκπαιδευμένο μοντέλο που έχει προκύψει από τη συνάρτηση *trainClassifier*.

Στην έξοδο της *testClassifier* παίρνουμε τα εξής σκορ, που περιγράφουν την αξιοπιστία του μοντέλου:

1. accuracy
2. precision
3. recall
4. f1-score

Επιπλέον των παραπάνω βασικών συναρτήσεων, γίνεται χρήση δύο ακόμη συναρτήσεων που είναι απαραίτητες για την ομαλή εκτέλεση της διαδικασίας. ΟΙ συναρτήσεις αυτές είναι οι εξής:

- *load_data*: μεταφορτώνει τα κατάλληλα δεδομένα που απαιτούνται, το σύνολο από ζεύγη τοπωνυμίων και τους συχνότερους όρους που περιέχει. Τα ορίσματα που παίρνει η συνάρτηση αφορούν το *μονοπάτι του αρχείου* στο δίσκο που περιέχει τα δεδομένα, καθώς και το *αλφάβητο των χαρακτήρων*, δηλαδή αν περιορίζεται σε λατινικούς χαρακτήρες (*latin*) ή όχι (*global*).
- *build*: κατασκευάζει τα διάφορα χαρακτηριστικά γνωρίσματα από τα τοπωνύμια των δεδομένων που έχουν μεταφορτωθεί.

Οι παραπάνω συναρτήσεις βρίσκονται στο αρχείο *src/featuresConstruction.py* στη κλάση *Features.*

3.2. Βιβλιοθήκη Κατηγοριοποίησης Σημείων Ενδιαφέροντος

3.2.1. Σύντομη περιγραφή

Η βιβλιοθήκη LGM-Classification είναι μια βιβλιοθήκη python που υλοποιεί μια πλήρη ροή εργασιών μηχανικής μάθησης για την εκπαίδευση αλγορίθμων σε επισημειωμένα σύνολα δεδομένων που αφορούν Σημεία Ενδιαφέροντος (ΣΕ), με στόχο την παραγωγή μοντέλων για την ακριβή κατάταξή ΣΕ σε κατηγορίες. Η βιβλιοθήκη LGM-Classification υλοποιεί μια συλλογή από χαρακτηριστικά εκπαίδευσης σχετικά με ιδιότητες των ΣΕ και τις σχέσεις τους με τα γειτονικά τους ΣΕ. Επιπλέον, περιλαμβάνει τεχνικές grid-search και cross-validation, βασισμένες στο εργαλείο scikit-learn, με σκοπό την αξιολόγηση μιας σειράς διαφορετικών μοντέλων κατάταξης και παραμετροποιήσεών τους, ώστε να παράγεται το πιο ταιριαστό μοντέλο για τα δεδομένα που είναι κάθε φορά διαθέσιμα.

Η βιβλιοθήκη LGM-Classification παρέχεται δωρεάν και ο πηγαίος κώδικας είναι διαθέσιμος στο GitHub (https://github.com/LinkGeoML/LGM-Classification). Μπορεί να διανεμηθεί και να τροποποιηθεί σύμφωνα με τους όρους της μη περιοριστικής MIT άδειας[11].

3.2.2. Αλγόριθμοι μηχανικής μάθησης

Οι αλγόριθμοι που χρησιμοποιούνται για την αναζήτηση καλύτερου μοντέλου είναι οι ακόλουθοι:

- K-Nearest Neighbors
- Support Vector Machines
- Decision Trees
- Random Forests
- Adaboos
- Naive Bayes
- Multi-layer Perceptron
- Gaussian Process
- Extra Trees

3.2.3. Πληροφορίες υλοποίησης και τεκμηρίωση

Η βιβλιοθήκη αυτή έχει υλοποιηθεί με χρήση της γλώσσας python και οι λειτουργίες μηχανικής μάθησης που εφαρμόζει καλύπτονται από τη βιβλιοθήκη scikit-learn. Οι μέθοδοι επεξεργασίας γεωχωρικών δεδομένων που χρησιμοποιούνται καλύπτονται από μια συλλογή σχετικών βιβλιοθηκών της γλώσσας python (shapely[12], geopandas[13], osmnx[14]), ενώ

[11] https://opensource.org/licenses/MIT

[12] https://pypi.org/project/Shapely/

η επεξεργασία κειμενικών δεδομένων καλύπτεται από τα python εργαλεία nltk[15] και whoosh[16].

Στη διεύθυνση https://linkgeoml.github.io/LGM-Classification/ υπάρχει αναλυτική τεκμηρίωση των διαφόρων υποστηριζόμενων μεθόδων (API) της βιβλιοθήκης LGM-Classification.

3.2.4. Βασικά υποσυστήματα

- Υποσύστημα Διεπαφής Γραμμής Εντολών: Το υποσύστημα αυτό λαμβάνει παραμέτρους εισόδου από το χρήστη προκειμένου να καθοριστεί ο τρόπος εκτέλεσης καθενός από τα στάδια της βιβλιοθήκης. Μεταξύ άλλων, ο χρήστης μπορεί να καθορίσει την ονοματοδοσία των παραχθέντων αρχείων κατά τη διάρκεια της εκτέλεσης των σταδίων, να συγκεκριμενοποιήσει ποια αρχεία θα λειτουργήσουν ως είσοδοι καθενός από τα στάδια αλλά και να επιλέξει άλλες σημαντικές παραμέτρους όπως τον αριθμό προβλέψεων που θα παρέχει ένα εκπαιδευμένο μοντέλο.
- Υποσύστημα Εξωτερικής Ρύθμισης Παραμέτρων: Το υποσύστημα αυτό περιλαμβάνει το δεύτερο επίπεδο διεπαφής μεταξύ χρήστη και βιβλιοθήκης. Αποτελεί το κύριο μέσο μέσω του οποίου ο χρήστης καθορίζει τη λειτουργικότητα της βιβλιοθήκης, αφού παρέχει τη δυνατότητα επιλογής των τιμών μιας εκτενούς συλλογής παραμέτρων οι οποίες καθορίζουν σημαντικά βήματά της, από τη διαδικασία εξαγωγής χαρακτηριστικών ως και το επίπεδο κατηγοριών στο οποίο θα ανήκουν οι κατηγορίες κατάταξης στα πειράματα.
- Υποσύστημα Εξαγωγής Χαρακτηριστικών: Το υποσύστημα αυτό υποστηρίζει τη διαδικασία εξαγωγής κειμενικών και γεωχωρικών χαρακτηριστικών τα οποία αποτυπώνουν χρήσιμη πληροφορία για τα ΣΕ ,που χρησιμοποιείται από τα επόμενα βήματα για την εκπαίδευση και αξιολόγηση μοντέλων κατάταξης.
- Υποσύστημα Ευρετηρίασης Γεωχωρικών Δεδομένων: Το υποσύστημα αυτό υποστηρίζει τη χρήση ευρετηρίων για την καλύτερη οργάνωση των γεωχωρικών δεδομένων που χρησιμοποιεί η βιβλιοθήκη για τις διάφορες λειτουργίες της με σκοπό την επιτάχυνση των πειραμάτων και την καλύτερη οργάνωση της διαθέσιμης πληροφορίας. Αυτό επιτυγχάνεται μέσω προσεγγίσεων ευρετηρίασης που χρησιμοποιούν δενδρικά ευρετήρια R-Tree και KD-Tree.
- Υποσύστημα Ευρετηρίασης Κειμενικών Δεδομένων: Το υποσύστημα αυτό υποστηρίζει τη χρήση ευρετηρίων για την καλύτερη οργάνωση των κειμενικών δεδομένων που χρησιμοποιεί η βιβλιοθήκη για τις διάφορες λειτουργίες της με σκοπό την επιτάχυνση των πειραμάτων και την καλύτερη οργάνωση της διαθέσιμης πληροφορίας. Αυτό επιτυγχάνεται μέσω προσεγγίσεων ευρετηρίασης που

[13] http://geopandas.org/

[14] https://github.com/gboeing/osmnx

[15] https://www.nltk.org/

[16] https://whoosh.readthedocs.io/en/latest/intro.html

χρησιμοποιούν ανεστραμμένα ευρετήρια, υποστηριζόμενα μέσω του python εργαλείου whoosh.

- Υποσύστημα Επιλογής Αλγορίθμου Κατάταξης: Το υποσύστημα αυτό αφορά το πρώτο στάδιο των πειραμάτων και περιλαμβάνει την εκπαίδευση μιας συλλογής από αλγορίθμους κατάταξης χρησιμοποιώντας εμφωλευμένη συγκριτική αξιολόγηση (nested cross-validation), ώστε να βρεθεί εκείνος με την καλύτερη απόδοση, μέσω συνεχών διαφοροποιήσεων των υπερ-παραμέτρων τους. Από το βήμα αυτό προκύπτει ο βέλτιστος αλγόριθμος ο οποίος πρόκειται να χρησιμοποιηθεί στα επόμενα στάδια.
- Υποσύστημα Βελτιστοποίησης Αλγορίθμου Κατάταξης: Το υποσύστημα αυτό αφορά το δεύτερο στάδιο των πειραμάτων και περιλαμβάνει αναζήτηση του χώρου των υπερ-παραμέτρων του βέλτιστου αλγορίθμου που έχει προκύψει από την εκτέλεση του πρώτου σταδίου. Αυτό επιτυγχάνεται με τη χρήση cross-validation, από την οποία προκύπτει ο καλύτερος συνδυασμός υπερ-παραμέτρων για τον επιλεγμένο αλγόριθμο.
- Υποσύστημα Εξαγωγής Μοντέλου Κατάταξης: Το υποσύστημα αυτό αφορά το τρίτο στάδιο των πειραμάτων και περιλαμβάνει την εκπαίδευση του καλύτερου αλγορίθμου κατάταξης ρυθμισμένου με τις καλύτερες υπερ-παραμέτρους, στοιχεία που προέκυψαν από το πρώτο και το δεύτερο στάδιο των πειραμάτων, αντίστοιχα και την αποθήκευσή του σε αρχείο με την κατάλληλη μορφή. Το μοντέλο αυτό μπορεί να χρησιμοποιηθεί στη συνέχεια για την πρόβλεψη κατηγοριών Σημείων Ενδιαφέροντος που παρέχονται εκ νέου μέσω άλλων συνόλων δεδομένων.
- Υποσύστημα Παροχής Προβλέψεων για Νέο Σύνολο Δεδομένων: Το υποσύστημα αυτό αφορά το τέταρτο στάδιο των πειραμάτων και περιλαμβάνει την παροχή προβλέψεων κατηγοριών ταξινομημένων κατά σειρά πιθανοφάνειας για ΣΕ που αποτελούν μέρος νέων συνόλων δεδομένων τα οποία παρέχει ο χρήστης κατά την εκτέλεση. Αναγκαία για τη λειτουργικότητα του υποσυστήματος αυτού είναι η ύπαρξη αρχείου στο οποίο βρίσκεται αποθηκευμένο ένα ήδη εκπαιδευμένο μοντέλο αλγορίθμου κατάταξης.
- Υποσύστημα Αξιολόγησης Απόδοσης: Το υποσύστημα αυτό αναλαμβάνει την αξιολόγηση της απόδοσης των διαφορετικών συνδυασμών αλγορίθμων κατάταξης και αντίστοιχων υπερ-παραμέτρων. Κάτι τέτοιο επιτυγχάνεται με τη χρήση στρατηγικών cross-validation και κατάλληλων μετρικών ώστε κάθε φορά να προκύπτει μια όσο το δυνατόν αντικειμενικότερη αξιολόγηση των αποδόσεων παραμένοντας ανεξάρτητη από τη φύση του συνόλου δεδομένων εκπαίδευσης και τυχόν διαφοροποιήσεις σε αυτό ανά εκτέλεση.
- Υποσύστημα Εξαγωγής Αποτελεσμάτων: Το υποσύστημα αυτό αναλαμβάνει την εξαγωγή αποτελεσμάτων υπό τη μορφή αρχείων .csv, .pkl και .txt ώστε να εξασφαλίζεται τόσο η ομαλή και απρόσκοπτη λειτουργικότητα των διαφορετικών σταδίων των πειραμάτων της βιβλιοθήκης αλλά και να διατηρείται η αναγνωσιμότητα από τους χρήστες των αποτελεσμάτων της.

3.2.5. Οδηγός χρήσης

3.2.5.1. Εγκατάσταση

Προαπαιτούμενα/εξαρτήσεις

- python 3
- numpy
- pandas
- sklearn
- geopandas
- nltk
- matplotlib
- psycopg2
- osmnx
- shapely
- argparse
- whoosh

Οδηγίες εγκατάστασης

python 3

```
sudo add-apt-repository ppa:jonathonf/python-3.6
sudo apt-get update
sudo apt-get install python3.6
sudo update-alternatives --install /usr/bin/python3 python3
/usr/bin/python3.5 1
sudo update-alternatives --install /usr/bin/python3 python3
/usr/bin/python3.6 2
```

numpy

```
sudo apt-get install python-pip
sudo pip install numpy
```

pandas

```
pip install pandas
```

sklearn

```
pip install -U scikit-learn
```

geopandas

```
pip install geopandas
```

nltk

```
pip install -U nltk
```

matplotlib

```
python -mpip install matplotlib
```

psycopg2

```
sudo apt-get install python-psycopg2
```

osmnx

```
pip install osmnx
```

shapely

```
pip install Shapely
```

argparse

```
pip install argparse
```

whoosh

```
pip install Whoosh
```

3.2.5.2. Παραμετροποίηση

Για την ομαλή εκτέλεση των λειτουργιών της βιβλιοθήκης πρέπει στο φάκελο εκτέλεσης να περιλαμβάνεται ένα αρχείο διαμόρφωσης ονόματι config.py, στο οποίο τα απαραίτητα πεδία παραμετροποίησης πρέπει να βρίσκονται δηλωμένα εντός μιας initialConfig κλάσης. Τα πεδία αυτά πρέπει να είναι τα ακόλουθα:

- feature_list: μια λίστα από συμβολοσειρές οι οποίες αντιστοιχούν στους τίτλους των χαρακτηριστικών που θα χρησιμοποιούν κατά τη φάση εξαγωγής χαρακτηριστικών.
- included_features: μια λίστα από συμβολοσειρές οι οποίες αντιστοιχούν στους τίτλους των χαρακτηριστικών που θα χρησιμοποιούν κατά τη φάση εκπαίδευσης των αλγορίθμων.
- features_to_normalize: μια λίστα από συμβολοσειρές οι οποίες αντιστοιχούν στους τίτλους των χαρακτηριστικών των οποίων οι τιμές πρόκειται να κανονικοποιηθούν.
- root_path: το μονοπάτι στο δίσκο του φακέλου εκτέλεσης των πειραμάτων.
- experiment_folder: το όνομα του φακέλου στον οποίο θέλουμε να αποθηκεύονται τα αποτελέσματα των πειραμάτων. Αν επιθυμούμε να μην τον ονοματίσουμε, τότε αφήνουμε την τιμή ίση με None.
- k_fold_parameter: η παράμετρος που καθορίζει τον αριθμό των διαχωρισμών στο πλαίσιο της διαδικασίας k-fold cross-validation.
- classifiers: λίστα που περιέχει συμβολοσειρές που αντιπροσωπεύουν τα ονόματα των αλγορίθμων κατάταξης που θα χρησιμοποιηθούν κατά τη διάρκεια των πειραμάτων.
- kNN_hyperparameters: λεξικό που περιέχει αντιστοιχίες μεταξύ των ονομάτων των υπερ-παραμέτρων και της συλλογής από τιμές που θέλουμε να διερευνήσουμε κατά τη διάρκεια εκπαίδευσης του μοντέλου k-nearest neighbors.
- SVM_hyperparameters: λεξικό που περιέχει αντιστοιχίες μεταξύ των ονομάτων των υπερ-παραμέτρων και της συλλογής από τιμές που θέλουμε να διερευνήσουμε κατά τη διάρκεια εκπαίδευσης του μοντέλου SVM.
- DecisionTree_hyperparamaters: λεξικό που περιέχει αντιστοιχίες μεταξύ των ονομάτων των υπερ-παραμέτρων και της συλλογής από τιμές που θέλουμε να διερευνήσουμε κατά τη διάρκεια εκπαίδευσης του μοντέλου Decision Trees.
- RandomForest_hyperparameters: λεξικό που περιέχει αντιστοιχίες μεταξύ των ονομάτων των υπερ-παραμέτρων και της συλλογής από τιμές που θέλουμε να διερευνήσουμε κατά τη διάρκεια εκπαίδευσης των συγγενικών μοντέλων Random Forests / Extra Trees.
- MLP_hyperparameters: λεξικό που περιέχει αντιστοιχίες μεταξύ των ονομάτων των υπερ-παραμέτρων και της συλλογής από τιμές που θέλουμε να διερευνήσουμε κατά τη διάρκεια εκπαίδευσης του μοντέλου MLP.

- poi_id: συμβολοσειρά που αντιστοιχεί στο όνομα του πεδίου του αριθμού ταυτοποίησης των ΣΕ στο .csv αρχείο που παρέχεται στη βιβλιοθήκη αντιπροσωπεύοντας το σύνολο δεδομένων.
- name: συμβολοσειρά που αντιστοιχεί στο όνομα του πεδίου του ονόματος των ΣΕ στο .csv αρχείο που παρέχεται στη βιβλιοθήκη αντιπροσωπεύοντας το σύνολο δεδομένων.
- class_codes: συμβολοσειρά που αντιστοιχεί στα ονόματα των πεδίων του ονόματος κατηγορίας ΣΕ στο .csv αρχείο που παρέχεται στη βιβλιοθήκη αντιπροσωπεύοντας το σύνολο δεδομένων.
- x: συμβολοσειρά που αντιστοιχεί στο όνομα του πεδίου που αντιστοιχεί στην τετμημένη των ΣΕ στο .csv αρχείο που παρέχεται στη βιβλιοθήκη αντιπροσωπεύοντας το σύνολο δεδομένων.
- y: συμβολοσειρά που αντιστοιχεί στο όνομα του πεδίου που αντιστοιχεί στην τεταγμένη των ΣΕ στο .csv αρχείο που παρέχεται στη βιβλιοθήκη αντιπροσωπεύοντας το σύνολο δεδομένων.
- original_SRID: συμβολοσειρά που αντιστοιχεί στον κωδικό του συστήματος συντεταγμένων του συνόλου δεδομένων.
- threshold_distance_neighbor_pois: αριθμός σε μέτρα που αντιστοιχεί στο μέγεθος ακτίνας εντός της οποίας θεωρούνται γειτονικά ενός άλλου τα ΣΕ.
- num_poi_neighbors: αριθμός των γειτόνων ενός ΣΕ που θέλουμε να επεξεργαστούμε κατά τη φάση εξαγωγής χαρακτηριστικών.
- threshold_distance_neighbor_pois_roads: αριθμός σε μέτρα που αντιστοιχεί στο μέγεθος ακτίνας εντός της οποίας θεωρούνται γειτονικά ενός άλλου τα ΣΕ εφόσον αυτά βρίσκονται στον ίδιο δρόμο με αυτό.
- top_k_terms_percentage: αριθμός που αντιστοιχεί στο ποσοστό όρων που θέλουμε να συμπεριλάβουμε κατά την εξαγωγή κειμενικών χαρακτηριστικών βάσει συχνότητας στα πειράματά μας.
- top_k_character_ngrams_percentage: αριθμός που αντιστοιχεί στο ποσοστό ν-γραμμάτων χαρακτήρων που θέλουμε να συμπεριλάβουμε κατά την εξαγωγή κειμενικών χαρακτηριστικών βάσει συχνότητας στα πειράματά μας.
- character_n_gram_size: αριθμός που αντιστοιχεί στο μέγεθος των ν-γραμμάτων χαρακτήρων που θα χρησιμοποιήσουμε κατά την εξαγωγή κειμενικών χαρακτηριστικών.
- term_n_gram_size: αριθμός που αντιστοιχεί στο μέγεθος των ν-γραμμάτων όρων που θα χρησιμοποιήσουμε κατά την εξαγωγή κειμενικών χαρακτηριστικών.
- level: λίστα που περιέχει αριθμούς που αντιστοιχούν στα επίπεδα κατηγοριών για τα οποία θέλουμε να λάβουμε αποτελέσματα από την εκτέλεση των πειραμάτων.
- k_error: λίστα που περιέχει τους αριθμούς των πιο πιθανών κατηγοριών ανά πρόβλεψη παραδείγματος που θέλουμε να ληφθούν υπόψιν κατά την εξαγωγή αποτελεσμάτων.
- osmnx_bbox: τιμή αλήθειας η οποία δηλώνει την επιθυμία του χρήστη να φορτώσει γεωχωρικά δεδομένα βάσει χρήσης bounding box.
- osmnx_bbox_coordinates: συντεταγμένες των κορυφών του bounding box βάσει του οποίου θέλουμε να φορτώσουμε γεωχωρικά δεδομένα. Έχει νόημα μόνο όταν η παράμετρος osmnx_bbox είναι True.

- osmnx_placename: Συμβολοσειρά που αντιστοιχεί στο πλήρες όνομα της περιοχής από την οποία θέλουμε να φορτώσουμε γεωχωρικά δεδομένα. Έχει νόημα μόνο όταν η παράμετρος osmnx_bbox είναι False.

3.2.5.3. Εκτέλεση

Για τη λειτουργία της βιβλιοθήκης απαιτείται η παροχή, ως είσοδο δεδομένων εκπαίδευσης, ενός .csv αρχείου το οποίο θα περιλαμβάνει μια συλλογή από πληροφορίες για ένα ή παραπάνω ΣΕ. Συγκεκριμένα το .csv αρχείο πρέπει να περιέχει, για κάθε ΣΕ, πληροφορία που αντιστοιχεί στις ακόλουθες ιδιότητες του:

- τον αριθμό ταυτοποίησής του
- το όνομά του
- μια συλλογή από τα ονόματα των κατηγοριών στις οποίες ανήκει
- την τετμημένη του
- την τεταγμένη του

Το σύνολο των λειτουργιών που καλύπτονται από την βιβλιοθήκη μπορεί να χωριστεί σε 4 ξεχωριστά στάδια, η εκτέλεση των οποίων περιγράφεται ακολούθως:

Αξιολόγηση/επιλογή αλγορίθμου

Η εκτέλεση αυτού του βήματος γίνεται με την ακόλουθη εντολή:

```
python find_best_clf.py -pois_csv_name <όνομα του .csv αρχείου που περιέχει την πληροφορία για τα ΣΕ> -results_file_name <όνομα που θέλουμε να δώσουμε στο .csv που θέλουμε να περιέχει τα αποτελέσματα των μετρικών ανά fold> -hyperparameter_file_name <όνομα που θέλουμε να δώσουμε στο αρχείο το οποίο θέλουμε να περιέχει τις ιδανικές υπερ-παραμέτρους ανά fold>
```

Οι τελευταίες δύο παράμετροι γραμμής εντολών είναι προαιρετικές και, αν δεν δοθούν, οι τιμές τους παίρνουν τις προκαθορισμένες τιμές classification_report_<επίπεδο κατηγορίας>_<ώρα εκτέλεσης>.csv και hyperparameters_per_fold_<επίπεδο κατηγορίας>_<ώρα εκτέλεσης>.csv αντίστοιχα.

Η εκτέλεση αυτού του βήματος παράγει τρία αρχεία:

- Ένα αρχείο που παρουσιάζει τις μετρικές απόδοσης του κάθε αλγορίθμου κατάταξης ανά fold για το test set.
- Ένα αρχείο που περιέχει το όνομα του καλύτερου αλγορίθμου κατάταξης ως προς την αποτελεσματικότητα.
- Ένα αρχείο που περιέχει την καλύτερη συλλογή υπερ-παραμέτρων ανά fold για το συγκεκριμένο αλγόριθμο κατάταξης που επιλέχθηκε ως ο καλύτερος.

Βελτιστοποίηση αλγορίθμου

Η εκτέλεση αυτού του βήματος γίνεται με την ακόλουθη εντολή:

```
python finetune_best_clf.py -pois_csv_name <όνομα του .csv αρχείου που περιέχει την πληροφορία για τα ΣΕ> -best_clf_file_name <αρχείο που περιέχει το όνομα του αλγορίθμου κατάταξης που πρόκειται να χρησιμοποιηθεί> -best_hyperparameter_file_name <όνομα που θέλουμε να δώσουμε στο αρχείο το οποίο θέλουμε να περιέχει τις ιδανικές υπερ-παραμέτρους για τον αλγόριθμο>
```

Οι τελευταίες δύο παράμετροι γραμμής εντολών είναι προαιρετικές και, αν δεν δοθούν, οι τιμές τους παίρνουν τις προκαθορισμένες τιμές best_hyperparameters_<επίπεδο κατηγορίας>_<ώρα εκτέλεσης>.csv και το τελευταίο αρχείο που δημιουργήθηκε και αρχίζει με το πρόθεμα *best_clf_*.

Η εκτέλεση αυτού του βήματος παράγει ένα αρχείο το οποίο παρουσιάζει την καλύτερη συλλογή υπερ-παραμέτρων για τον αλγόριθμο τον οποίο επιλέξαμε να βελτιστοποιήσουμε.

Εξαγωγή Μοντέλου Κατάταξης

Η εκτέλεση αυτού του βήματος γίνεται με την ακόλουθη εντολή:

```
python export_best_model.py -pois_csv_name <όνομα του .csv αρχείου που περιέχει την πληροφορία για τα ΣΕ> -best_hyperparameter_file_name < όνομα που θέλουμε να δώσουμε στο αρχείο το οποίο θέλουμε να περιέχει τις ιδανικές υπερ-παραμέτρους για τον αλγόριθμο> -best_clf_file_name <αρχείο που περιέχει το όνομα του αλγορίθμου κατάταξης που πρόκειται να χρησιμοποιηθεί> -trained_model <όνομα αρχείου στο οποίο θα εξαχθεί το μοντέλο του αλγορίθμου>
```

Οι τελευταίες τρεις παράμετροι γραμμής εντολών είναι προαιρετικές και, αν δεν δοθούν, οι τιμές τους παίρνουν τις προκαθορισμένες τιμές του τελευταίου αρχείου που δημιουργήθηκε και έχει πρόθεμα *best_hyperparameters_*, του τελευταίου αρχείου που δημιουργήθηκε και έχει πρόθεμα *best_clf_* και trained_model_<επίπεδο>_<ώρα εκτέλεσης>.pkl αντίστοιχα.

Εξαγωγή προβλέψεων σε νέα σύνολα δεδομένων

Η εκτέλεση αυτού του βήματος γίνεται με την ακόλουθη εντολή:

```
python other_dataset_classification.py -pois_csv_name <όνομα του .csv αρχείου που περιέχει την πληροφορία για τα ΣΕ> -k <αριθμός των κατηγοριών που θέλουμε να αναθέσουμε σε κάθε Σημείο Ενδιαφέροντος με σειρά πιθανότητας> -results_file_name <όνομα του αρχείου στο οποίο θέλουμε να αποθηκευτούν οι προβλέψεις> -trained_model_file_name <όνομα του αρχείου που περιέχει το αποθηκευμένο μοντέλο το οποίο θέλουμε να αναλάβει τις προβλέψεις>
```

Οι τελευταίες τρεις παράμετροι γραμμής εντολών είναι προαιρετικές και, αν δεν δοθούν, οι τιμές τους παίρνουν τις προκαθορισμένες τιμές [5, 10], pred_categories_<επίπεδο κατηγορίας>_top<αριθμός που δόθηκε ως τιμή του k>categories.csv και το όνομα του τελευταίου αρχείου που δημιουργήθηκε με πρόθεμα trained_model_<επίπεδο>_<ώρα εκτέλεσης>.pkl αντίστοιχα.

3.3. Βιβλιοθήκη Γεωκωδικοποίησης Διευθύνσεων

3.3.1. Σύντομη περιγραφή

Η βιβλιοθήκη LGM-Geocoding είναι μια βιβλιοθήκη python η οποία υλοποιεί μια πλήρη ροή εργασιών μηχανικής μάθησης για την εκπαίδευση αλγορίθμων κατάταξης σε επισημειωμένα σύνολα δεδομένων που αφορούν αντιστοιχισμένα ζεύγη συντεταγμένων με την ιδανική πηγή γεωκωδικοποίησης, αποσκοπώντας στην παραγωγή μοντέλων για παροχή προβλέψεων σχετικά με την ιδανική πηγή γεωκωδικοποίησης για νέα ζεύγη συντεταγμένων. Κάθε στιγμιότυπο-παράδειγμα του προβλήματος αποτελείται από το σύνολο των ζευγών συντεταγμένων που προκύπτουν από όλες τις διαθέσιμες πηγές γεωκωδικοποίησης. Η βιβλιοθήκη LGM-Geocoding υλοποιεί μια συλλογή από χαρακτηριστικά εκπαίδευσης σχετικά με τις ιδιότητες των ζευγών συντεταγμένων που είναι διαθέσιμα ανά πηγή γεωκωδικοποίησης και τις σχέσεις τους με γειτονικά γεωχωρικά δεδομένα. Επιπλέον, περιλαμβάνει τεχνικές αναζήτησης πλέγματος (grid-search) και συγκριτικής αξιολόγησης (cross-validation), βασισμένες στο εργαλείο scikit-learn, με σκοπό την αξιολόγηση μιας σειράς διαφορετικών μοντέλων κατάταξης και παραμετροποιήσεών τους, ώστε να παράγεται το πιο ταιριαστό μοντέλο για τα δεδομένα που είναι κάθε φορά διαθέσιμα.

Η βιβλιοθήκη LGM-Geocoding παρέχεται δωρεάν και ο πηγαίος κώδικας είναι διαθέσιμος στο GitHub (https://github.com/LinkGeoML/LGM-Geocoding). Μπορεί να διανεμηθεί και να τροποποιηθεί σύμφωνα με τους όρους της μη περιοριστικής MIT άδειας[17].

3.3.2. Αλγόριθμοι μηχανικής μάθησης

Οι αλγόριθμοι που χρησιμοποιούνται για την αναζήτηση καλύτερου μοντέλου είναι οι ακόλουθοι:

- K-Nearest Neighbors
- Support Vector Machines
- Decision Trees
- Random Forests
- Adaboos
- Naive Bayes
- Multi-layer Perceptron
- Gaussian Process
- Extra Trees

3.3.3. Πληροφορίες υλοποίησης και τεκμηρίωση

Η βιβλιοθήκη αυτή έχει υλοποιηθεί με χρήση της γλώσσας python και οι αλγόριθμοι μηχανικής μάθησης που χρησιμοποιεί καλύπτονται από τη βιβλιοθήκη scikit-learn[18]. Οι

[17] https://opensource.org/licenses/MIT

μέθοδοι επεξεργασίας γεωχωρικών δεδομένων που χρησιμοποιούνται καλύπτονται από μια συλλογή σχετικών βιβλιοθηκών της γλώσσας python (shapely, geopandas, osmnx) ενώ η επεξεργασία κειμενικών δεδομένων καλύπτεται από τα εργαλεία nltk και whoosh.

Στη διεύθυνση https://linkgeoml.github.io/LGM-Geocoding/ υπάρχει αναλυτική τεκμηρίωση των διαφόρων υποστηριζόμενων μεθόδων (API) της βιβλιοθήκης LGM-Geocoding.

3.3.3.1. Βασικά υποσυστήματα

- Υποσύστημα Διεπαφής Γραμμής Εντολών: Το υποσύστημα αυτό λαμβάνει παραμέτρους εισόδου από τον χρήστη προκειμένου να καθοριστεί ο τρόπος εκτέλεσης καθενός από τα στάδια της βιβλιοθήκης. Μεταξύ άλλων ο χρήστης μπορεί να καθορίσει την ονοματοδοσία των παραχθέντων αρχείων κατά τη διάρκεια της εκτέλεσης των σταδίων, να συγκεκριμενοποιήσει ποια αρχεία θα λειτουργήσουν ως είσοδοι καθενός από τα στάδια αλλά και να επιλέξει άλλες σημαντικές παραμέτρους όπως τον αριθμό προβλέψεων που θα παρέχει ένα εκπαιδευμένο μοντέλο.
- Υποσύστημα Εξωτερικής Ρύθμισης Παραμέτρων: Το υποσύστημα αυτό περιλαμβάνει το δεύτερο επίπεδο διεπαφής μεταξύ χρήστη και βιβλιοθήκης. Αποτελεί το κύριο μέσο μέσω του οποίου ο χρήστης καθορίζει τη λειτουργικότητα της βιβλιοθήκης αφού παρέχει τη δυνατότητα επιλογής των τιμών μιας εκτενούς συλλογής παραμέτρων οι οποίες καθορίζουν σημαντικά βήματά της, όπως τη διαδικασία εξαγωγής χαρακτηριστικών.
- Υποσύστημα Γεωκωδικοποίησης: Το υποσύστημα αυτό υποστηρίζει την εξαγωγή συντεταγμένων δεδομένων διευθύνσεων με τη χρήση μιας συλλογής πηγών γεωκωδικοποίησης. Οι συντεταγμένες αυτές θα χρησιμοποιηθούν στη συνέχεια από το υποσύστημα εξαγωγής χαρακτηριστικών αφού προηγηθεί επισήμανση της κατάλληλης πηγής γεωκωδικοποίησης ανά ζεύγος συντεταγμένων.
- Υποσύστημα Εξαγωγής Χαρακτηριστικών: Το υποσύστημα αυτό υποστηρίζει τη διαδικασία εξαγωγής γεωχωρικών χαρακτηριστικών, τα οποία αποτυπώνουν χρήσιμη πληροφορία για τα ζεύγη συντεταγμένων και χρησιμοποιούνται από τα επόμενα βήματα για την εκπαίδευση και αξιολόγηση αλγορίθμων κατάταξης.
- Υποσύστημα Επιλογής Αλγορίθμου Κατάταξης: Το υποσύστημα αυτό αφορά το πρώτο στάδιο των πειραμάτων και περιλαμβάνει την εκπαίδευση μιας συλλογής από αλγορίθμους κατάταξης χρησιμοποιώντας εμφωλευμένη συγκριτική αξιολόγηση (nested cross-validation), ώστε να βρεθεί εκείνος με την καλύτερη απόδοση, μέσω συνεχών διαφοροποιήσεων των υπερ-παραμέτρων τους. Από το βήμα αυτό προκύπτει ο βέλτιστος αλγόριθμος ο οποίος πρόκειται να χρησιμοποιηθεί στα επόμενα στάδια.
- Υποσύστημα Βελτιστοποίησης Αλγορίθμου Κατάταξης: Το υποσύστημα αυτό αφορά το δεύτερο στάδιο των πειραμάτων και περιλαμβάνει αναζήτηση του χώρου των υπερ-παραμέτρων του βέλτιστου αλγορίθμου που έχει προκύψει από την εκτέλεση

[18] https://scikit-learn.org/stable/

του πρώτου σταδίου. Αυτό επιτυγχάνεται με τη χρήση της μεθόδου cross-validation, από την οποία προκύπτει ο καλύτερος συνδυασμός υπερ-παραμέτρων.

- Υποσύστημα Εξαγωγής Μοντέλου Κατάταξης: Το υποσύστημα αυτό αφορά το τρίτο στάδιο των πειραμάτων και περιλαμβάνει την εκπαίδευση του καλύτερου αλγορίθμου κατάταξης ρυθμισμένου με τις καλύτερες υπερ-παραμέτρους, στοιχεία που προέκυψαν από το πρώτο και το δεύτερο στάδιο των πειραμάτων, αντίστοιχα και την αποθήκευσή του σε αρχείο με την κατάλληλη μορφή. Το μοντέλο αυτό μπορεί να χρησιμοποιηθεί στη συνέχεια για την πρόβλεψη κατάλληλων πηγών γεωκωδικοποίησης για ζεύγη συντεταγμένων που παρέχονται εκ νέου μέσω άλλων συνόλων δεδομένων.
- Υποσύστημα Παροχής Προβλέψεων για Νέο Σύνολο Δεδομένων: Το υποσύστημα αυτό αφορά το τέταρτο στάδιο των πειραμάτων και περιλαμβάνει την παροχή προβλέψεων κατάλληλων πηγών γεωκωδικοποίησης ταξινομημένων κατά σειρά πιθανοφάνειας για ζεύγη συντεταγμένων που αποτελούν μέρος νέων συνόλων δεδομένων τα οποία παρέχει ο χρήστης κατά την εκτέλεση. Αναγκαία για τη λειτουργικότητα του υποσυστήματος αυτού είναι η ύπαρξη αρχείου στο οποίο βρίσκεται αποθηκευμένο ένα ήδη εκπαιδευμένο μοντέλο αλγορίθμου κατάταξης.
- Υποσύστημα Αξιολόγησης Απόδοσης: Το υποσύστημα αυτό αναλαμβάνει την αξιολόγηση της απόδοσης των διαφορετικών συνδυασμών αλγορίθμων κατάταξης και αντίστοιχων υπερ-παραμέτρων. Κάτι τέτοιο επιτυγχάνεται με τη χρήση στρατηγικών cross-validation και κατάλληλων μετρικών ώστε κάθε φορά να προκύπτει μια όσο το δυνατόν αντικειμενικότερη αξιολόγηση των αποδόσεων παραμένοντας ανεξάρτητη από τη φύση του συνόλου δεδομένων εκπαίδευσης και τυχόν διαφοροποιήσεις σε αυτό ανά εκτέλεση.
- Υποσύστημα Εξαγωγής Αποτελεσμάτων: Το υποσύστημα αυτό αναλαμβάνει την εξαγωγή αποτελεσμάτων υπό τη μορφή αρχείων .csv, .pkl και .txt ώστε να εξασφαλίζεται τόσο η ομαλή και απρόσκοπτη λειτουργικότητα των διαφορετικών σταδίων των πειραμάτων της βιβλιοθήκης αλλά και να διατηρείται η αναγνωσιμότητα από τους χρήστες των αποτελεσμάτων της.

3.3.4. Οδηγός χρήσης

3.3.4.1. Εγκατάσταση

Προαπαιτούμενα/εξαρτήσεις

- python 3
- numpy
- pandas
- sklearn
- geopandas
- matplotlib
- psycopg2
- osmnx
- shapely
- argparse

Οδηγίες εγκατάστασης

python 3

```
sudo add-apt-repository ppa:jonathonf/python-3.6
sudo apt-get update
sudo apt-get install python3.6
sudo update-alternatives --install /usr/bin/python3 python3
/usr/bin/python3.5 1
sudo update-alternatives --install /usr/bin/python3 python3
/usr/bin/python3.6 2
```

numpy

```
sudo apt-get install python-pip
sudo pip install numpy
```

pandas

```
pip install pandas
```

sklearn

```
pip install -U scikit-learn
```

geopandas

```
pip install geopandas
```

matplotlib

```
python -mpip install matplotlib
```

psycopg2

```
sudo apt-get install python-psycopg2
```

osmnx

```
pip install osmnx
```

shapely

```
pip install Shapely
```

argparse

```
pip install argparse
```

3.3.4.2. Παραμετροποίηση

Για την ομαλή εκτέλεση των λειτουργιών της βιβλιοθήκης πρέπει στο φάκελο εκτέλεσης να περιλαμβάνεται ένα αρχείο διαμόρφωσης ονόματι config.py, στο οποίο τα απαραίτητα πεδία παραμετροποίησης πρέπει να βρίσκονται δηλωμένα εντός μιας initialConfig κλάσης. Τα πεδία αυτά πρέπει να είναι τα ακόλουθα:

- feature_list: μια λίστα από συμβολοσειρές οι οποίες αντιστοιχούν στους τίτλους των χαρακτηριστικών που θα χρησιμοποιούν κατά τη φάση εξαγωγής χαρακτηριστικών.
- included_features: μια λίστα από συμβολοσειρές οι οποίες αντιστοιχούν στους τίτλους των χαρακτηριστικών που θα χρησιμοποιούν κατά τη φάση εκπαίδευσης των αλγορίθμων.
- root_path: το μονοπάτι στο δίσκο του φακέλου εκτέλεσης των πειραμάτων.

- experiment_folder: το όνομα του φακέλου στον οποίο θέλουμε να αποθηκεύονται τα αποτελέσματα των πειραμάτων. Αν επιθυμούμε να μην τον ονοματίσουμε, τότε αφήνουμε την τιμή ίση με None.
- k_fold_parameter: η παράμετρος που καθορίζει τον αριθμό των διαχωρισμών στο πλαίσιο της διαδικασίας k-fold cross-validation.
- classifiers: λίστα που περιέχει συμβολοσειρές που αντιπροσωπεύουν τα ονόματα των αλγορίθμων κατάταξης που θα χρησιμοποιηθούν κατά τη διάρκεια των πειραμάτων.
- kNN_hyperparameters: λεξικό που περιέχει αντιστοιχίες μεταξύ των ονομάτων των υπερ-παραμέτρων και της συλλογής από τιμές που θέλουμε να διερευνήσουμε κατά τη διάρκεια εκπαίδευσης του μοντέλου k-nearest neighbors.
- SVM_hyperparameters: λεξικό που περιέχει αντιστοιχίες μεταξύ των ονομάτων των υπερ-παραμέτρων και της συλλογής από τιμές που θέλουμε να διερευνήσουμε κατά τη διάρκεια εκπαίδευσης του μοντέλου SVM.
- DecisionTree_hyperparameters: λεξικό που περιέχει αντιστοιχίες μεταξύ των ονομάτων των υπερ-παραμέτρων και της συλλογής από τιμές που θέλουμε να διερευνήσουμε κατά τη διάρκεια εκπαίδευσης του μοντέλου Decision Trees.
- RandomForest_hyperparameters: λεξικό που περιέχει αντιστοιχίες μεταξύ των ονομάτων των υπερ-παραμέτρων και της συλλογής από τιμές που θέλουμε να διερευνήσουμε κατά τη διάρκεια εκπαίδευσης των συγγενικών μοντέλων Random Forests / Extra Trees.
- MLP_hyperparameters: λεξικό που περιέχει αντιστοιχίες μεταξύ των ονομάτων των υπερ-παραμέτρων και της συλλογής από τιμές που θέλουμε να διερευνήσουμε κατά τη διάρκεια εκπαίδευσης του μοντέλου MLP.

3.3.4.3. Εκτέλεση

Για τη λειτουργία της βιβλιοθήκης απαιτείται η παροχή, ως είσοδο δεδομένων εκπαίδευσης, ενός .csv αρχείου το οποίο θα περιλαμβάνει μια συλλογή από πληροφορίες για μία ή παραπάνω διευθύνσεις προς γεωκωδικοποίηση. Συγκεκριμένα το .csv αρχείο πρέπει να περιέχει, τουλάχιστον, για κάθε διεύθυνση προς γεωκωδικοποίηση, τις συντεταγμένες γεωκωδικοποίησης που επέστρεψε για αυτήν καθένας από γεωκωδικοποιητές (στο συγκεκριμένο σενάριο που εξετάσαμε, έχουμε συνολικά τρεις γεωκωδικοποιητές, στις κολώνες X2-Y2, X3-Y3 και X4-Y4 αντίστοιχα) και την ετικέτα που δηλώνει την προτιμότερη πηγή γεωκωδικοποίησης (κολώνα dset), η οποία λειτουργεί ως επισημείωση κατηγορίας για το πρόβλημα μηχανικής μάθησης κατάταξης που επιλύουμε.

Το σύνολο των λειτουργιών που καλύπτονται από την βιβλιοθήκη μπορεί να χωριστεί σε 4 ξεχωριστά στάδια η εκτέλεση των οποίων περιγράφεται ακολούθως:

Αξιολόγηση/επιλογή αλγορίθμου

Η εκτέλεση αυτού του βήματος γίνεται με την ακόλουθη εντολή:

```
python find_best_clf.py -geocoding_file_name <όνομα του .csv αρχείου που περιέχει την πληροφορία για τα σημεία γεωκωδικοποίησης> -results_file_name <όνομα που θέλουμε να δώσουμε στο .csv που θέλουμε να περιέχει τα αποτελέσματα των μετρικών ανά fold> -hyperparameter_file_name <όνομα που θέλουμε να δώσουμε στο αρχείο το οποίο θέλουμε να περιέχει τις ιδανικές υπερ-παραμέτρους ανά fold>
```

Οι τελευταίες δύο παράμετροι γραμμής εντολών είναι προαιρετικές και, αν δεν δοθούν, οι τιμές τους παίρνουν τις προκαθορισμένες τιμές classification_report_<επίπεδο κατηγορίας>_<ώρα εκτέλεσης>.csv και hyperparameters_per_fold_<επίπεδο κατηγορίας>_<ώρα εκτέλεσης>.csv αντίστοιχα.

Η εκτέλεση αυτού του βήματος παράγει τρία αρχεία:

- Ένα αρχείο που παρουσιάζει τις μετρικές απόδοσης του κάθε αλγορίθμου κατάταξης ανά fold για το test set.
- Ένα αρχείο που περιέχει το όνομα του καλύτερου αλγορίθμου κατάταξης ως προς την αποτελεσματικότητα.
- Ένα αρχείο που περιέχει την καλύτερη συλλογή υπερ-παραμέτρων ανά fold για τον συγκεκριμένο αλγόριθμο κατάταξης που επιλέχθηκε ως ο καλύτερος.

Βελτιστοποίηση αλγορίθμου

Η εκτέλεση αυτού του βήματος γίνεται με την ακόλουθη εντολή:

```
python finetune_best_clf.py -geocoding_file_name <όνομα του .csv αρχείου που περιέχει την πληροφορία για τα σημεία γεωκωδικοποίησης> -best_clf_file_name <αρχείο που περιέχει το όνομα του αλγορίθμου κατάταξης που πρόκειται να χρησιμοποιηθεί> -best_hyperparameter_file_name <όνομα που θέλουμε να δώσουμε στο αρχείο το οποίο θέλουμε να περιέχει τις ιδανικές υπερ-παραμέτρους για τον αλγόριθμο>
```

Οι τελευταίες δύο παράμετροι γραμμής εντολών είναι προαιρετικές και, αν δεν δοθούν, οι τιμές τους παίρνουν τις προκαθορισμένες τιμές best_hyperparameters_<επίπεδο κατηγορίας>_<ώρα εκτέλεσης>.csv και το τελευταίο αρχείο που δημιουργήθηκε και αρχίζει με το πρόθεμα *best_clf_*.

Η εκτέλεση αυτού του βήματος παράγει ένα αρχείο το οποίο παρουσιάζει την καλύτερη συλλογή υπερ-παραμέτρων για τον αλγόριθμο τον οποίο επιλέξαμε να βελτιστοποιήσουμε.

Εξαγωγή Μοντέλου Κατάταξης

Η εκτέλεση αυτού του βήματος γίνεται με την ακόλουθη εντολή:

```
python export_best_model.py -geocoding_file_name <όνομα του .csv αρχείου που περιέχει την πληροφορία για τα σημεία γεωκωδικοποίησης> -best_hyperparameter_file_name < όνομα που θέλουμε να δώσουμε στο αρχείο το οποίο θέλουμε να περιέχει τις ιδανικές υπερ-παραμέτρους για τον αλγόριθμο> -best_clf_file_name <αρχείο που περιέχει το όνομα του αλγορίθμου κατάταξης που πρόκειται να χρησιμοποιηθεί> -trained_model <όνομα αρχείου στο οποίο θα εξαχθεί το μοντέλο του αλγορίθμου>
```

Οι τελευταίες τρεις παράμετροι γραμμής εντολών είναι προαιρετικές και, αν δεν δοθούν, οι τιμές τους παίρνουν τις προκαθορισμένες τιμές του τελευταίου αρχείου που δημιουργήθηκε και έχει πρόθεμα *best_hyperparameters_*, του τελευταίου αρχείου που δημιουργήθηκε και έχει πρόθεμα *best_clf_* και trained_model_<επίπεδο>_<ώρα εκτέλεσης>.pkl αντίστοιχα.

Εξαγωγή προβλέψεων σε νέα σύνολα δεδομένων

Η εκτέλεση αυτού του βήματος γίνεται με την ακόλουθη εντολή:

```
python other_dataset_classification.py -geocoding_file_name <όνομα του .csv αρχείου που περιέχει την πληροφορία για τα σημεία γεωκωδικοποίησης> -k <αριθμός των κατηγοριών που θέλουμε να αναθέσουμε σε κάθε Σημείο
```

```
Ενδιαφέροντος με σειρά πιθανότητας> -results_file_name <όνομα του αρχείου στο οποίο θέλουμε να αποθηκευτούν οι προβλέψεις> -trained_model_file_name <όνομα του αρχείου που περιέχει το αποθηκευμένο μοντέλο το οποίο θέλουμε να αναλάβει τις προβλέψεις>
```

Οι τελευταίες τρεις παράμετροι γραμμής εντολών είναι προαιρετικές και, αν δεν δοθούν, οι τιμές τους παίρνουν τις προκαθορισμένες τιμές [5, 10], pred_categories_<επίπεδο κατηγορίας>_top<αριθμός που δόθηκε ως τιμή του k>categories.csv και το όνομα του τελευταίου αρχείου που δημιουργήθηκε με πρόθεμα trained_model_<επίπεδο>_<ώρα εκτέλεσης>.pkl αντίστοιχα.

3.4. Βιβλιοθήκη Ολοκλήρωσης Γεωτεμαχίων

3.4.1. Σύντομη περιγραφή

Η βιβλιοθήκη LGM-PolygonClassification είναι μια βιβλιοθήκη python η οποία υλοποιεί μια πλήρη ροή εργασιών μηχανικής μάθησης για την εκπαίδευση αλγορίθμων κατάταξης σε επισημειωμένα σύνολα δεδομένων που αφορούν αντιστοιχισμένα ζεύγη πολυγώνων, κάθε ένα από τα οποία ανήκει σε μια διακριτή κατηγορία πολυγώνων. Η βιβλιοθήκη LGM-PolygonClassification υλοποιεί μια συλλογή από χαρακτηριστικά εκπαίδευσης σχετικά με τα ιδιαίτερα χαρακτηριστικά του κάθε πολυγώνου, αλλά και με τις γεωχωρικές σχέσεις ανάμεσα στα αντιστοιχισμένα πολύγωνα του κάθε ζευγαριού. Επιπλέον, περιλαμβάνει τεχνικές αναζήτησης πλέγματος (grid-search) και συγκριτικής αξιολόγησης (cross-validation), βασισμένες στο εργαλείο scikit-learn, με σκοπό την αξιολόγηση μιας σειράς διαφορετικών μοντέλων κατάταξης και παραμετροποιήσεών τους, ώστε να παράγεται το πιο ταιριαστό μοντέλο για τα δεδομένα που είναι κάθε φορά διαθέσιμα.

Η βιβλιοθήκη LGM-PolygonClassification παρέχεται δωρεάν και ο πηγαίος κώδικας είναι διαθέσιμος στο GitHub (https://github.com/LinkGeoML/LGM-PolygonClassification). Μπορεί να διανεμηθεί και να τροποποιηθεί σύμφωνα με τους όρους της μη περιοριστικής MIT άδειας[19].

3.4.2. Αλγόριθμοι μηχανικής μάθησης

Οι αλγόριθμοι που χρησιμοποιούνται για την αναζήτηση καλύτερου μοντέλου είναι οι ακόλουθοι:

- K-Nearest Neighbors
- Support Vector Machines
- Decision Trees
- Random Forests
- Adaboos
- Naive Bayes
- Multi-layer Perceptron
- Gaussian Process
- Extra Trees

3.4.3. Πληροφορίες υλοποίησης και τεκμηρίωση

Η βιβλιοθήκη έχει υλοποιηθεί με χρήση της γλώσσας python και οι λειτουργικότητες μηχανικής μάθησης που εφαρμόζει καλύπτονται από τη βιβλιοθήκη scikit-learn. Οι μέθοδοι επεξεργασίας γεωχωρικών δεδομένων που χρησιμοποιούνται καλύπτονται από μια συλλογή σχετικών βιβλιοθηκών της γλώσσας python (shapely, geopandas, osmnx) ενώ η επεξεργασία κειμενικών δεδομένων καλύπτεται από τα εργαλεία nltk και whoosh.

[19] https://opensource.org/licenses/MIT

Στη διεύθυνση https://linkgeoml.github.io/LGM-PolygonClassification/ υπάρχει αναλυτική τεκμηρίωση των διαφόρων υποστηριζόμενων μεθόδων (API) της βιβλιοθήκης LGM-PolygonClassification.

3.4.4. Βασικά υποσυστήματα

- Υποσύστημα Διεπαφής Γραμμής Εντολών: Το υποσύστημα αυτό λαμβάνει παραμέτρους εισόδου από τον χρήστη προκειμένου να καθοριστεί ο τρόπος εκτέλεσης καθενός από τα στάδια της βιβλιοθήκης. Μεταξύ άλλων ο χρήστης μπορεί να καθορίσει την ονοματοδοσία των παραχθέντων αρχείων κατά τη διάρκεια της εκτέλεσης των σταδίων, να συγκεκριμενοποιήσει ποια αρχεία θα λειτουργήσουν ως είσοδοι καθενός από τα στάδια αλλά και να επιλέξει άλλες σημαντικές παραμέτρους όπως τον αριθμό προβλέψεων που θα παρέχει ένα εκπαιδευμένο μοντέλο.
- Υποσύστημα Εξωτερικής Ρύθμισης Παραμέτρων: Το υποσύστημα αυτό περιλαμβάνει το δεύτερο επίπεδο διεπαφής μεταξύ χρήστη και βιβλιοθήκης. Αποτελεί το κύριο μέσο μέσω του οποίου ο χρήστης καθορίζει τη λειτουργικότητα της βιβλιοθήκης αφού παρέχει τη δυνατότητα επιλογής των τιμών μιας εκτενούς συλλογής παραμέτρων οι οποίες καθορίζουν σημαντικά βήματά της, από τη διαδικασία εξαγωγής χαρακτηριστικών ως και το επίπεδο κατηγοριών στο οποίο θα ανήκουν οι κατηγορίες κατάταξης στα πειράματα.
- Υποσύστημα Δημιουργίας Συνόλου Δεδομένων από Shapefiles: Το υποσύστημα αυτό υποστηρίζει τη δημιουργία του συνόλου δεδομένων το οποίο θα χρησιμοποιηθεί στη συνέχεια κατά τη διάρκεια εξαγωγής των απαραίτητων χαρακτηριστικών για τα πειράματα. Αυτό επιτυγχάνεται με την εύρεση αντιστοιχιών πολυγώνων ανάμεσα σε δύο ή περισσότερα shapefiles με τη χρήση κριτηρίων ποσοστού επικάλυψης ανάμεσα στα επιμέρους πολύγωνα.
- Υποσύστημα Εξαγωγής Χαρακτηριστικών: Το υποσύστημα αυτό υποστηρίζει τη διαδικασία εξαγωγής κειμενικών και γεωχωρικών χαρακτηριστικών τα οποία αποτυπώνουν χρήσιμη πληροφορία για τα διαθέσιμα πολύγωνα, που χρησιμοποιείται από τα επόμενα βήματα για την εκπαίδευση και αξιολόγηση μοντέλων κατάταξης.
- Υποσύστημα Επιλογής Αλγορίθμου Κατάταξης: Το υποσύστημα αυτό αφορά το πρώτο στάδιο των πειραμάτων και περιλαμβάνει την εκπαίδευση μιας συλλογής από αλγορίθμους κατάταξης χρησιμοποιώντας εμφωλευμένη συγκριτική αξιολόγηση (nested cross-validation), ώστε να βρεθεί εκείνος με την καλύτερη απόδοση, μέσω συνεχών διαφοροποιήσεων των υπερ-παραμέτρων τους. Από το βήμα αυτό προκύπτει ο βέλτιστος αλγόριθμος ο οποίος πρόκειται να χρησιμοποιηθεί στα επόμενα στάδια.
- Υποσύστημα Βελτιστοποίησης Αλγορίθμου Κατάταξης: Το υποσύστημα αυτό αφορά το δεύτερο στάδιο των πειραμάτων και περιλαμβάνει αναζήτηση του χώρου των υπερ-παραμέτρων του βέλτιστου αλγορίθμου που έχει προκύψει από την εκτέλεση του πρώτου σταδίου. Αυτό επιτυγχάνεται με τη χρήση της μεθόδου cross-validation, από την οποία προκύπτει ο καλύτερος συνδυασμός υπερ-παραμέτρων.
- Υποσύστημα Εξαγωγής Μοντέλου Κατάταξης: Το υποσύστημα αυτό αφορά το τρίτο στάδιο των πειραμάτων και περιλαμβάνει την εκπαίδευση του καλύτερου

αλγορίθμου κατάταξης ρυθμισμένου με τις καλύτερες υπερ-παραμέτρους, στοιχεία που προέκυψαν από το πρώτο και το δεύτερο στάδιο των πειραμάτων, αντίστοιχα και την αποθήκευσή του σε αρχείο με την κατάλληλη μορφή. Το μοντέλο αυτό μπορεί να χρησιμοποιηθεί στη συνέχεια για την πρόβλεψη κατηγοριών πολυγώνων που παρέχονται εκ νέου μέσω άλλων συνόλων δεδομένων.

- Υποσύστημα Παροχής Προβλέψεων για Νέο Σύνολο Δεδομένων: Το υποσύστημα αυτό αφορά το τέταρτο στάδιο των πειραμάτων και περιλαμβάνει την παροχή προβλέψεων κατηγοριών για νέα για πολύγωνα. Αναγκαία για τη λειτουργικότητα του υποσυστήματος αυτού είναι η ύπαρξη αρχείου στο οποίο βρίσκεται αποθηκευμένο ένα ήδη εκπαιδευμένο μοντέλο αλγορίθμου κατάταξης.
- Υποσύστημα Αξιολόγησης Απόδοσης: Το υποσύστημα αυτό αναλαμβάνει την αξιολόγηση της απόδοσης των διαφορετικών συνδυασμών αλγορίθμων κατάταξης και αντίστοιχων υπερ-παραμέτρων. Κάτι τέτοιο επιτυγχάνεται με τη χρήση στρατηγικών cross-validation και κατάλληλων μετρικών ώστε κάθε φορά να προκύπτει μια όσο το δυνατόν αντικειμενικότερη αξιολόγηση των αποδόσεων παραμένοντας ανεξάρτητη από τη φύση του συνόλου δεδομένων εκπαίδευσης και τυχόν διαφοροποιήσεις σε αυτό ανά εκτέλεση.
- Υποσύστημα Εξαγωγής Αποτελεσμάτων: Το υποσύστημα αυτό αναλαμβάνει την εξαγωγή αποτελεσμάτων υπό τη μορφή αρχείων .csv, .pkl και .txt ώστε να εξασφαλίζεται τόσο η ομαλή και απρόσκοπτη λειτουργικότητα των διαφορετικών σταδίων των πειραμάτων της βιβλιοθήκης αλλά και να διατηρείται η αναγνωσιμότητα από τους χρήστες των αποτελεσμάτων της.

3.4.5. Οδηγός χρήσης

3.4.5.1. Εγκατάσταση

Προαπαιτούμενα/εξαρτήσεις

- python 3
- numpy
- pandas
- sklearn
- geopandas
- matplotlib
- psycopg2
- osmnx
- shapely
- argparse

Οδηγίες εγκατάστασης

python 3

```
sudo add-apt-repository ppa:jonathonf/python-3.6
sudo apt-get update
sudo apt-get install python3.6
sudo update-alternatives --install /usr/bin/python3 python3
/usr/bin/python3.5 1
```

```
sudo update-alternatives --install /usr/bin/python3 python3
/usr/bin/python3.6 2
```

numpy

```
sudo apt-get install python-pip
sudo pip install numpy
```

pandas

```
pip install pandas
```

sklearn

```
pip install -U scikit-learn
```

geopandas

```
pip install geopandas
```

matplotlib

```
python -mpip install matplotlib
```

psycopg2

```
sudo apt-get install python-psycopg2
```

osmnx

```
pip install osmnx
```

shapely

```
pip install Shapely
```

argparse

```
pip install argparse
```

3.4.5.2. Παραμετροποίηση

Για την ομαλή εκτέλεση των λειτουργιών της βιβλιοθήκης, πρέπει στο φάκελο εκτέλεσης να περιλαμβάνεται ένα αρχείο διαμόρφωσης ονόματι config.py, στο οποίο τα απαραίτητα πεδία παραμετροποίησης πρέπει να βρίσκονται δηλωμένα εντός μιας initialConfig κλάσης. Τα πεδία αυτά πρέπει να είναι τα ακόλουθα:

- feature_list: μια λίστα από συμβολοσειρές οι οποίες αντιστοιχούν στους τίτλους των χαρακτηριστικών που θα χρησιμοποιούν κατά τη φάση εξαγωγής χαρακτηριστικών.
- included_features: μια λίστα από συμβολοσειρές οι οποίες αντιστοιχούν στους τίτλους των χαρακτηριστικών που θα χρησιμοποιούν κατά τη φάση εκπαίδευσης των αλγορίθμων.
- root_path: το μονοπάτι στο δίσκο του φακέλου εκτέλεσης των πειραμάτων.
- experiment_folder: το όνομα του φακέλου στον οποίο θέλουμε να αποθηκεύονται τα αποτελέσματα των πειραμάτων. Αν επιθυμούμε να μην τον ονοματίσουμε, τότε αφήνουμε την τιμή ίση με None.
- k_fold_parameter: η παράμετρος που καθορίζει τον αριθμό των διαχωρισμών στο πλαίσιο της διαδικασίας k-fold cross-validation.
- classifiers: λίστα που περιέχει συμβολοσειρές που αντιπροσωπεύουν τα ονόματα των αλγορίθμων κατάταξης που θα χρησιμοποιηθούν κατά τη διάρκεια των πειραμάτων.

- kNN_hyperparameters: λεξικό που περιέχει αντιστοιχίες μεταξύ των ονομάτων των υπερ-παραμέτρων και της συλλογής από τιμές που θέλουμε να διερευνήσουμε κατά τη διάρκεια εκπαίδευσης του μοντέλου k-nearest neighbors.
- SVM_hyperparameters: λεξικό που περιέχει αντιστοιχίες μεταξύ των ονομάτων των υπερ-παραμέτρων και της συλλογής από τιμές που θέλουμε να διερευνήσουμε κατά τη διάρκεια εκπαίδευσης του μοντέλου SVM.
- DecisionTree_hyperparameters: λεξικό που περιέχει αντιστοιχίες μεταξύ των ονομάτων των υπερ-παραμέτρων και της συλλογής από τιμές που θέλουμε να διερευνήσουμε κατά τη διάρκεια εκπαίδευσης του μοντέλου Decision Trees.
- RandomForest_hyperparameters: λεξικό που περιέχει αντιστοιχίες μεταξύ των ονομάτων των υπερ-παραμέτρων και της συλλογής από τιμές που θέλουμε να διερευνήσουμε κατά τη διάρκεια εκπαίδευσης των συγγενικών μοντέλων Random Forests / Extra Trees.
- MLP_hyperparameters: λεξικό που περιέχει αντιστοιχίες μεταξύ των ονομάτων των υπερ-παραμέτρων και της συλλογής από τιμές που θέλουμε να διερευνήσουμε κατά τη διάρκεια εκπαίδευσης του μοντέλου MLP.

3.4.5.3. Εκτέλεση

Για τη λειτουργία της βιβλιοθήκης απαιτείται η παροχή ενός .csv αρχείου το οποίο θα περιλαμβάνει μια συλλογή από πληροφορίες για ένα ή παραπάνω ζεύγη αντιστοιχισμένων πολυγώνων. Συγκεκριμένα το .csv αρχείο πρέπει να περιέχει, τουλάχιστον, για κάθε ζεύγος πολυγώνων πληροφορία που να αντιστοιχεί στις επιμέρους γεωμετρίες τους.

Το σύνολο των λειτουργιών που καλύπτονται από την βιβλιοθήκη μπορεί να χωριστεί σε 4 ξεχωριστά στάδια η εκτέλεση των οποίων περιγράφεται ακολούθως:

Αξιολόγηση/επιλογή αλγορίθμου

Η εκτέλεση αυτού του βήματος γίνεται με την ακόλουθη εντολή:

```
python find_best_clf.py -polygon_file_name <όνομα του .csv αρχείου που περιέχει την πληροφορία για τα πολύγωνα> -results_file_name <όνομα που θέλουμε να δώσουμε στο .csv που θέλουμε να περιέχει τα αποτελέσματα των μετρικών ανά fold> -hyperparameter_file_name <όνομα που θέλουμε να δώσουμε στο αρχείο το οποίο θέλουμε να περιέχει τις ιδανικές υπερ-παραμέτρους ανά fold>
```

Οι τελευταίες δύο παράμετροι γραμμής εντολών είναι προαιρετικές και, αν δεν δοθούν, οι τιμές τους παίρνουν τις προκαθορισμένες τιμές classification_report_<επίπεδο κατηγορίας>_<ώρα εκτέλεσης>.csv και hyperparameters_per_fold_<επίπεδο κατηγορίας>_<ώρα εκτέλεσης>.csv αντίστοιχα.

Η εκτέλεση αυτού του βήματος παράγει τρία αρχεία:

- Ένα αρχείο που παρουσιάζει τις μετρικές απόδοσης του κάθε αλγορίθμου κατάταξης ανά fold για το test set.
- Ένα αρχείο που περιέχει το όνομα του καλύτερου αλγορίθμου κατάταξης ως προς την αποτελεσματικότητα.
- Ένα αρχείο που περιέχει την καλύτερη συλλογή υπερ-παραμέτρων ανά fold για τον συγκεκριμένο αλγόριθμο κατάταξης που επιλέχθηκε ως ο καλύτερος.

Βελτιστοποίηση αλγορίθμου

Η εκτέλεση αυτού του βήματος γίνεται με την ακόλουθη εντολή:

```
python finetune_best_clf.py - polygon_file_name <όνομα του .csv αρχείου που περιέχει την πληροφορία για τα πολύγωνα> -best_clf_file_name <αρχείο που περιέχει το όνομα του αλγορίθμου κατάταξης που πρόκειται να χρησιμοποιηθεί> -best_hyperparameter_file_name <όνομα που θέλουμε να δώσουμε στο αρχείο το οποίο θέλουμε να περιέχει τις ιδανικές υπερ-παραμέτρους για τον αλγόριθμο>
```

Οι τελευταίες δύο παράμετροι γραμμής εντολών είναι προαιρετικές και, αν δεν δοθούν, οι τιμές τους παίρνουν τις προκαθορισμένες τιμές best_hyperparameters_<επίπεδο κατηγορίας>_<ώρα εκτέλεσης>.csv και το τελευταίο αρχείο που δημιουργήθηκε και αρχίζει με το πρόθεμα *best_clf_*.

Η εκτέλεση αυτού του βήματος παράγει ένα αρχείο το οποίο παρουσιάζει την καλύτερη συλλογή υπερ-παραμέτρων για τον αλγόριθμο τον οποίο επιλέξαμε να βελτιστοποιήσουμε.

Εξαγωγή Μοντέλου Κατάταξης

Η εκτέλεση αυτού του βήματος γίνεται με την ακόλουθη εντολή:

```
python export_best_model.py -polygon_file_name <όνομα του .csv αρχείου που περιέχει την πληροφορία για τα πολύγωνα> -best_hyperparameter_file_name <όνομα που θέλουμε να δώσουμε στο αρχείο το οποίο θέλουμε να περιέχει τις ιδανικές υπερ-παραμέτρους για τον αλγόριθμο> -best_clf_file_name <αρχείο που περιέχει το όνομα του αλγορίθμου κατάταξης που πρόκειται να χρησιμοποιηθεί> -trained_model <όνομα αρχείου στο οποίο θα εξαχθεί το μοντέλο του αλγορίθμου>
```

Οι τελευταίες τρεις παράμετροι γραμμής εντολών είναι προαιρετικές και, αν δεν δοθούν, οι τιμές τους παίρνουν τις προκαθορισμένες τιμές του τελευταίου αρχείου που δημιουργήθηκε και έχει πρόθεμα *best_hyperparameters_*, του τελευταίου αρχείου που δημιουργήθηκε και έχει πρόθεμα *best_clf_* και trained_model_<επίπεδο>_<ώρα εκτέλεσης>.pkl αντίστοιχα.

Εξαγωγή προβλέψεων σε νέα σύνολα δεδομένων

Η εκτέλεση αυτού του βήματος γίνεται με την ακόλουθη εντολή:

```
python other_dataset_classification.py -polygon_file_name <όνομα του .csv αρχείου που περιέχει την πληροφορία για τα πολύγωνα> -k <αριθμός των κατηγοριών που θέλουμε να αναθέσουμε σε κάθε Σημείο Ενδιαφέροντος με σειρά πιθανότητας> -results_file_name <όνομα του αρχείου στο οποίο θέλουμε να αποθηκευτούν οι προβλέψεις> -trained_model_file_name <όνομα του αρχείου που περιέχει το αποθηκευμένο μοντέλο το οποίο θέλουμε να αναλάβει τις προβλέψεις>
```

Οι τελευταίες τρεις παράμετροι γραμμής εντολών είναι προαιρετικές και, αν δεν δοθούν, οι τιμές τους παίρνουν τις προκαθορισμένες τιμές [5, 10], pred_categories_<επίπεδο κατηγορίας>_top<αριθμός που δόθηκε ως τιμή του k>categories.csv και το όνομα του τελευταίου αρχείου που δημιουργήθηκε με πρόθεμα trained_model_<επίπεδο>_<ώρα εκτέλεσης>.pkl αντίστοιχα.

4. Πειραματική αξιολόγηση

4.1. Βιβλιοθήκη Διασύνδεσης Τοπωνυμίων

4.1.1. Σύνολο αξιολόγησης

Το σύνολο δεδομένων για την εκπαίδευση και αξιολόγηση των αλγορίθμων που αναπτύχθηκαν στο πλαίσιο της διασύνδεσης τοπωνυμίων έχει προκύψει από τη βάση τοπωνυμίων Geonames[20], η οποία περιέχει πάνω από 11 εκατομμύρια εγγραφές με τοπωνύμια για περισσότερες από 250 χώρες από όλο τον πλανήτη. Συγκεκριμένα, κατασκευάσαμε δύο σύνολα εκπαίδευσης, κάθε ένα από τα οποία αποτελείται από 100.000 ζεύγη τοπωνυμίων, όπου 50.000 τέτοια ζεύγη είναι επισημειωμένα ως True, δηλαδή περιγράφουν το ίδιο τοπωνύμιο, και 50.000 ως False, δηλαδή αντιστοιχούν σε διαφορετικό τοπωνύμιο. Επιπλέον, ορίσαμε και ένα σύνολο αξιολόγησης με 5 εκατομμύρια ζεύγη τοπωνυμίων, όπου τα 2.5 εκατομμύρια είναι επισημειωμένα ως True και τα 2.5 εκατομμύρια ως False. Η διαδικασία που ακολουθήθηκε για την κατασκευή των παραπάνω συνόλων βασίστηκε στο γεγονός ότι η βάση Geonames περιέχει, μεταξύ άλλων πληροφοριών, το κύριο όνομα του κάθε τοπωνυμίου καθώς και μια σειρά από εναλλακτικά ονόματα τα οποία μπορεί να παρουσιάζουν από μικρές έως αρκετά μεγάλες διαφορές σε σχέση με το αρχικό κύριο όνομα και είναι η εξής:

- Για την κατασκευή ζευγαριών τοπωνυμίων που αντιστοιχούν στο ίδιο τοπωνύμιο, επιλέγουμε ένα κύριο όνομα και ένα εναλλακτικό όνομα από την ίδια εγγραφή στη βάση. Στην περίπτωση που υπάρχουν περισσότερα από ένα εναλλακτικά ονόματα, συνήθως, επιλέγουμε εκείνο που παρουσιάζει διαφορές με το κύριο όνομα, ώστε η διαδικασία αναγνώρισης ζευγών τοπωνυμίων που αντιστοιχούν στις ίδιες οντότητες να μην είναι τετριμμένη.
- Στην περίπτωση που τα ζεύγη τοπωνυμίων αντιστοιχούν σε διαφορετικά τοπωνύμια, οι όροι προέρχονται από ένα κύριο όνομα και ένα εναλλακτικό όνομα τα οποία, όμως, βρίσκονται σε διαφορετικές εγγραφές στη βάση. Όσον αφορά τα σύνολα εκπαίδευσης, το πρώτο σύνολο, $train_{latin}$, περιλαμβάνει ζεύγη τοπωνυμιών από χώρες τις Ευρώπης και της Βόρειας Αμερικής ενώ το δεύτερο, $train_{global}$, από όλον τον κόσμο, ακολουθώντας την κατανομή των χωρών που υπάρχει στο στο σύνολο αξιολόγησης, test.

	$Train_{latin}$	$Train_{global}$	Test
Συναφή (True)	50.000	50.000	2.500.000
Μη Συναφή (False)	50.000	50.000	2.500.000
Συνολικά	100.000	100.000	5.000.000

Πίνακας 1: Σύνολα δεδομένων αξιολόγησης της Διασύνδεσης Τοπωνυμίων

[20] http://download.geonames.org/export/dump/

4.1.2. Συνθήκες αξιολόγησης

Για την επίλυση του προβλήματος της διασύνδεσης τοπωνυμίων αξιολογήσαμε δύο διαφορετικές προσεγγίσεις: μία που βασίζεται απλά στη χρήση συναρτήσεων ομοιότητας για διασύνδεση τοπωνυμίων και μία που βασίζεται στην εκπαίδευση αλγορίθμων μηχανικής μάθησης για δυαδική κατάταξη. Στην πρώτη περίπτωση περιλαμβάνονται συναρτήσεις από απλές μετρικές ομοιότητας έως σύνθετες μετα-συναρτήσεις ομοιότητας (Πίνακας 2) οι οποίες, είτε χρησιμοποιούνται ευρέως στη βιβλιογραφία, είτε έχουν σχεδιαστεί με κύριο σκοπό να ενσωματώσουν χαρακτηριστικές ιδιομορφίες των τοπωνυμίων, όπως η LGM-Sim (Π1.2: «Προδιαγραφή εξειδικευμένων χαρακτηριστικών και κανόνων εκπαίδευσης» - Ενότητα 2.1.4.3). Στη δεύτερη περίπτωση, τα χαρακτηριστικά εκπαίδευσης που σχεδιάσαμε βασίζονται κυρίως στην εφαρμογή διαφόρων συναρτήσεων ομοιότητας συμβολοσειρών που χρησιμοποιήσαμε στην προηγούμενη περίπτωση. Συγκεκριμένα, σχεδιάσαμε τέσσερις ομάδες χαρακτηριστικών εκπαίδευσης, οι οποίες περιγράφονται αναλυτικά στο παραδοτέο Π1.2: «Προδιαγραφή εξειδικευμένων χαρακτηριστικών και κανόνων εκπαίδευσης» (Ενότητα 2.1.4).

Damerau-Levenshtein	Jaro
Jaro-Winkler	Jaro-Winkler Reversed
Sorted Jaro-Winkler	Cosine N-Grams
Jaccard N-Grams	Dice Bi-Grams
Jaccard Skip-grams	Monge-Elkan
Soft-Jaccard	Davis and De Salles
LGM Jaro-Winkler	LGM Jaro-Winkler Reversed

Πίνακας 2: Βασικές συναρτήσεις ομοιότητας που αξιολογήθηκαν

Για την αξιολόγηση της απόδοσης των μεθοδολογιών που σχεδιάσαμε χρησιμοποιήθηκαν ευρέως διαδεδομένες μετρικές από το χώρο της ανάκτησης πληροφορίας και της μηχανικής μάθησης και οι οποίες, συνοπτικά, είναι οι εξής:

- *Ορθότητα (Accuracy)*: είναι το ποσοστό των προβλέψεων/κατατάξεων ζευγαριών τοπωνυμίων (συναφή/μη συναφή – True/False) που είναι σωστές.
- *Ακρίβεια (Precision)*: είναι το ποσοστό των ανακτημένων κατατάξεων ζευγαριών τοπωνυμιών που είναι συναφή (True).
- *Ανάκληση (Recall)*: είναι το ποσοστό των συναφών (True) κατατάξεων ζευγαριών τοπωνυμίων που ανακτώνται.
- *Αρμονικός μέσος (F-score)*: Είναι ο αρμονικός μέσος όρος των μετρικών precision και recall.

Η διαδικασία της αξιολόγησης αποτελείται από δύο διακριτές φάσεις. Η πρώτη περιλαμβάνει την εύρεση των κατωφλιών (threshold) των μετρικών ομοιότητας για τα οποία επιτυγχάνεται το καλύτερο σκορ ορθότητας και την εκπαίδευση των τεχνικών μηχανικής μάθησης στα δύο σύνολα εκπαίδευσης, δηλαδή $train_{latin}$ και $train_{global}$. Έπειτα, υπολογίζεται η ορθότητα της κάθε μεθόδου και για τις διάφορες παραμέτρους που έχουν προκύψει από την πρώτη φάση, στο σύνολο ελέγχου (test) των 5 εκατομμυρίων ζευγαριών

τοπωνυμίων. Η παραπάνω διαδικασία, όπου η εκπαίδευση γίνεται σε σύνολα δεδομένων με διαφορετικά χαρακτηριστικά και ο έλεγχος σε κοινό σύνολο δεδομένων, επιτρέπει να ελέγξουμε όχι μόνο την αποτελεσματικότητα των διαφόρων μεθόδων ομοιότητας, αλλά και να αξιολογήσουμε την ευρωστία και τη δυνατότητα γενίκευσης των υπό μελέτη μεθόδων.

4.1.3. Αποτελέσματα

Στην παρουσίαση των αποτελεσμάτων παρουσιάζουμε τις μεθόδους που εξετάσαμε χωρισμένες σε δύο κατηγορίες: εκείνη που περιλαμβάνει τις βασικές μεθόδους (basic) που χρησιμοποιούνται ευρέως στη βιβλιογραφία και αποτελούν τη βάση σύγκρισης και εκείνη που περικλείει όλες τις μεθόδους που υλοποιήσαμε (LGM). Στους παρακάτω πίνακες παρουσιάζουμε, με έντονη γραμματοσειρά, τα καλύτερα αποτελέσματα που επιτυγχάνονται ξεχωριστά για κάθε κατηγορία μεθόδων, για κάθε μετρική αξιολόγησης και για κάθε μία από τις μεθοδολογίες που εφαρμόσαμε ξεχωριστά. Επίσης, παρουσιάζουμε τις καλύτερες 5 μεθόδους για κάθε μεθοδολογία ως προς την ορθότητα (accuracy) που πετυχαίνουν.

Μια πρώτη παρατήρηση είναι ότι οι μέθοδοι που έχουμε προτείνει (LGM) είναι συνολικά καλύτερες, και ως προς την ορθότητα αλλά και ως προς το F-score, από τις βασικές μεθόδους. Αυτό φαίνεται στους Πίνακας 3 καιΠίνακας 4, που παρουσιάζουν τα αποτελέσματα για την εκπαίδευση των μεθόδων στα σύνολα $train_{latin}$ και $train_{global}$ αντίστοιχα. Επίσης, παρατηρούμε ότι, αν και η διαφορά είναι μικρή μεταξύ των καλύτερων σκορ ορθότητας που πετυχαίνουν οι δυο κατηγορίες μεθόδων στο σύνολο δεδομένων $train_{latin}$, η διαφορά αυτή γίνεται αρκετά μεγάλη στην περίπτωση του συνόλου $train_{global}$, πάντα υπέρ των μεθόδων LGM. Αυτό είναι μια καλή ένδειξη ότι οι μετα-συναρτήσεις LGM-Sim και τα χαρακτηριστικά εκπαίδευσης που προκύπτουν από αυτές περιγράφουν καλύτερα και με πιο γενικό τρόπο γνωρίσματα που σχετίζονται με τοπωνύμια.

SIMILARITY METHODS	Accuracy		Precision		Recall		F-score	
	Basic	LGM	Basic	LGM	Basic	LGM	Basic	LGM
Damerau	0.721	0.751	0.763	**0.782**	**0.643**	0.698	**0.698**	**0.737**
Jaro	0.71	0.733	0.752	0.763	0.627	0.678	0.684	0.718
JW Reverse	0.727	0.739	0.782	0.768	0.629	0.684	0.697	0.723
Davis/De Salles	0.716	0.726	0.756	0.726	0.637	**0.726**	0.691	0.726
LGM JW Reverse	**0.729**	**0.74**	**0.806**	0.762	0.603	0.697	0.69	0.728
CLASSIFICATION METHODS								
Grad. Boost. Trees	**0.795**	**0.806**	0.814	0.819	**0.764**	**0.785**	**0.788**	**0.802**
SVM	0.74	0.754	0.794	0.804	0.647	0.671	0.713	0.732
Random Forests	0.787	0.801	0.803	0.817	0.76	0.776	0.781	0.796
Extr. Rand. Trees	0.781	0.793	0.803	0.813	0.745	0.761	0.773	0.786
Neural Network	0.73	0.742	0.781	**0.821**	0.639	0.618	0.703	0.705

Πίνακας 3: Αποτελέσματα των διαφόρων μεθόδων ομοιότητας στο σύνολο εκπαίδευσης $train_{latin}$

SIMILARITY METHODS	Accuracy		Precision		Recall		F-score	
	Basic	LGM	Basic	LGM	Basic	LGM	Basic	LGM
Damerau	0.652	0.789	0.79	0.821	0.414	0.738	0.543	0.778
Jaro	0.639	0.775	0.772	0.825	0.394	0.699	0.521	0.756
JW Reverse	0.651	0.8	0.779	0.832	0.421	0.751	0.547	0.79
Davis/De Salles	0.621	0.773	0.712	0.787	0.408	0.748	0.518	0.767
LGM JW Reverse	**0.655**	**0.802**	0.807	0.826	0.408	0.766	0.542	0.795
CLASSIFICATION METHODS								
Grad. Boost. Trees	**0.773**	**0.848**	0.759	0.864	0.799	**0.826**	**0.778**	**0.845**
SVM	0.723	0.826	0.69	0.865	0.808	0.772	0.744	0.816
Random Forests	0.769	0.846	0.764	**0.868**	0.779	0.817	0.771	0.841
Extr. Rand. Trees	0.766	0.843	**0.767**	0.871	0.763	0.805	0.765	0.836
Neural Network	0.721	0.821	0.682	0.844	**0.831**	0.789	0.749	0.815

Πίνακας 4: Αποτελέσματα των διαφόρων μεθόδων ομοιότητας στο σύνολο εκπαίδευσης $train_{global}$

Η αποτελεσματικότητα των μεθόδων στο σύνολο ελέγχου των 5 εκατομμυρίων ζευγαριών τοπωνυμίων παρουσιάζεται στους Πίνακας 5 καιΠίνακας 6. Συγκεκριμένα, όσον αφορά την αξιολόγηση των μεθόδων που παραμετροποιήθηκαν στο σύνολο $train_{latin}$ (Πίνακας 5), η καλύτερη μέθοδος LGM-Sim πετυχαίνει ποσοστό ορθότητας 79.3% σε σύγκριση με την καλύτερη βασική μέθοδο η οποία πετυχαίνει μόνο 65.1%. Τα αντίστοιχα ποσοστά ορθότητας που πετυχαίνουν τα μοντέλα μηχανικής μάθησης είναι 85.8% (LGM) και 78.6% (βασική). Παρόμοια αποτελέσματα βλέπουμε και ως προς το F1-score. Επίσης, παρατηρούμε ότι οι LGM-Sim μετρικές πετυχαίνουν εντυπωσιακά αποτελέσματα σε σύγκριση με τις βασικές μεθόδους όσον αφορά τα αποτελέσματα στην Ανάκληση (Recall), ενώ στην κατηγορία δυαδικής κατάταξης με τεχνικές μηχανικής μάθησης τα οφέλη εντοπίζονται στη Ακρίβεια (Precision). Ενδεικτική είναι και σε αυτή την περίπτωση η ευρωστία και γενίκευση που επιτυγχάνεται με τις μεθόδους LGM. Τέλος, παρόμοια συμπεράσματα προκύπτουν ως προς τα αποτελέσματα που επιτυγχάνουν οι μέθοδοι που παραμετροποιήθηκαν στο σύνολο $train_{global}$ και παρουσιάζονται στον Πίνακας 6. Συγκεκριμένα, στην κατηγορία των μετρικών ομοιότητας, η καλύτερη συνάρτηση LGM πετυχαίνει ποσοστό ορθότητας κατά 14.8% βελτιωμένο σε σχέση με τις βασικές μετρικές. Η αντίστοιχη αύξηση στη περίπτωση των μοντέλων μηχανικής μάθησης είναι 7%. Μια γενική παρατήρηση, εμφανής στα αποτελέσματα όλων των πινάκων, είναι ότι οι μετρικές ομοιότητας LGM-Sim πετυχαίνουν σκορ ορθότητας αρκετά πιο κοντά σε αυτό που επιτυγχάνεται με τις τεχνικές μηχανικής μάθησης, σε σύγκριση πάντα με τις βασικές συναρτήσεις ομοιότητας. Το γεγονός αυτό επιτρέπει τη χρήση των παραπάνω μετα-συναρτήσεων ομοιότητας στις περιπτώσεις που οι αρκετά πιο χρονοβόρες, όσον αφορά το στάδιο της προ-επεξεργασίας, τεχνικές μηχανικής μάθησης δεν αποτελούν επιλογή,

πετυχαίνοντας όμως ικανοποιητικά αποτελέσματα στο πρόβλημα της διασύνδεσης τοπωνυμίων.

SIMILARITY METHODS	Accuracy		Precision		Recall		F-score	
	Basic	LGM	Basic	LGM	Basic	LGM	Basic	LGM
Damerau	0.642	0.781	0.72	0.778	**0.465**	0.788	**0.565**	0.783
Jaro	0.634	0.763	0.721	0.76	0.437	0.769	0.544	0.764
JW Reverse	0.645	0.792	0.746	0.799	0.439	0.78	0.553	0.789
Davis/De Salles	0.618	0.762	0.715	0.739	0.394	**0.81**	0.508	0.773
LGM JW Reverse	**0.651**	**0.793**	**0.773**	0.792	0.426	0.795	0.55	**0.793**
CLASSIFICATION METHODS								
Grad. Boost. Trees	0.783	0.854	0.77	0.875	**0.809**	0.827	**0.789**	0.85
SVM	0.723	0.826	0.692	0.865	0.804	0.772	0.744	0.816
Random Forests	**0.786**	**0.858**	**0.781**	**0.881**	0.795	**0.827**	0.788	**0.853**
Extr. Rand. Trees	0.783	0.855	0.781	0.881	0.788	0.821	0.784	0.85
Neural Network	0.723	0.821	0.685	0.847	0.825	0.784	0.749	0.814

Πίνακας 5: Αποτελέσματα των διαφόρων μεθόδων ομοιότητας στο σύνολο αξιολόγησης με την παραμετροποίηση που έγινε στο $train_{latin}$

SIMILARITY METHODS	Accuracy		Precision		Recall		F-score	
	Basic	LGM	Basic	LGM	Basic	LGM	Basic	LGM
Damerau	0.647	0.786	0.79	0.825	0.401	0.726	0.532	0.772
Jaro	0.636	0.772	0.774	0.83	0.383	0.685	0.513	0.751
JW Reverse	0.648	0.797	0.781	0.835	**0.412**	0.74	**0.54**	0.785
Davis/De Salles	0.618	0.773	0.715	0.793	0.394	0.738	0.508	0.764
LGM JW Reverse	**0.651**	**0.799**	**0.806**	0.829	0.398	**0.755**	0.533	**0.79**
CLASSIFICATION METHODS								
Grad. Boost. Trees	0.783	**0.856**	0.77	0.879	**0.809**	**0.827**	**0.789**	**0.852**
SVM	0.723	0.826	0.692	0.866	0.804	0.771	0.744	0.816
Random Forests	**0.786**	0.855	**0.781**	**0.879**	0.795	0.823	0.788	0.85
Extr. Rand. Trees	0.783	0.85	0.781	0.878	0.788	0.813	0.784	0.844
Neural Network	0.723	0.822	0.685	0.842	0.825	0.794	0.749	0.817

Πίνακας 6: Αποτελέσματα των διαφόρων μεθόδων ομοιότητας στο σύνολο αξιολόγησης με την παραμετροποίηση που έγινε στο $train_{global}$

4.2. Βιβλιοθήκη Κατηγοριοποίησης Σημείων Ενδιαφέροντος

4.2.1. Σύνολο αξιολόγησης

Το σύνολο δεδομένων αποτελείται από σύνολα δεδομένων ΣΕ που αποτελούν προϊόν της εταιρείας Geodata, το οποίο εμπορεύεται σε B2B επίπεδο. Ορισμένα στατιστικά ενδιαφέροντος παρουσιάζονται στον Πίνακας 1. Συγκεκριμένα, το σύνολο δεδομένων οργανώνει τα ΣΕ σε δύο επίπεδα κατηγοριών αποτελούμενα από 13 και 71 κατηγορίες, αντίστοιχα. Η οργάνωση αυτή μας επιτρέπει την αξιολόγηση των μεθόδων μας σε δύο διαφορετικά επίπεδα όσον αφορά τον αριθμό των διαθέσιμων κατηγοριών. Επίσης, όπως είναι αντιληπτό από τα στατιστικά του πίνακα για το δεύτερο επίπεδο, 18 από τις 71 κατηγορίες παρουσιάζουν πολύ χαμηλή συχνότητα, πράγμα το οποίο εν δυνάμει επιδρά αρνητικά στην ακρίβεια της κατηγοριοποίησης. Παρόλα αυτά, το σύνολο δεδομένων χρησιμοποιείται όπως είναι, εφόσον ο κύριος στόχος μας είναι η μέτρηση της ακρίβειας της μεθόδου μας σε ρεαλιστικά σενάρια.

Στατιστική	**1ο επίπεδο κατηγοριών**	**2ο επίπεδο κατηγοριών**
Αριθμός Σημείων Ενδιαφέροντος	884	884
Αριθμός Διακριτών Κατηγοριών	13	71
Κατηγορίες με συχνότητα ≤ 5	0	18
Μέγιστη συχνότητα κατηγορίας	327	93

Πίνακας 7: Στατιστικές ιδιότητες του συνόλου δεδομένων Σημείων Ενδιαφέροντος της εταιρίας Geodata

4.2.2. Συνθήκες αξιολόγησης

Η αξιολόγηση ακολουθεί τα τυπικά πρότυπα αξιολόγησης με τη χρήση 5-fold cross-validation. Τα διανύσματα χαρακτηριστικών τροφοδοτούνται σε μια σειρά από αλγορίθμους μηχανικής μάθησης που αναλαμβάνουν την κατηγοριοποίηση. Το σύνολο δεδομένων χωρίζεται σε πέντε (5) ισοπληθή μέρη (folds) τα οποία διασχίζονται πέντε φορές. Κάθε φορά, τρία από τα πέντε μέρη χρησιμοποιούνται ως σύνολο εκπαίδευσης, ένα ως σύνολο επαλήθευσης και ένα ως σύνολο ελέγχου. Κατά αυτόν τον τρόπο, κάθε φορά που διασχίζουμε το σύνολο δεδομένων χρησιμοποιούνται και διαφορετικά μέρη του για την εκπαίδευση και την επαλήθευση του μοντέλου. Κάθε διάσχιση περιλαμβάνει, για κάθε επιμέρους αλγόριθμο κατηγοριοποίησης, αναζήτηση του χώρου των υπερ-παραμέτρων που τον συνοδεύουν, ώστε να εξασφαλιστεί ο βέλτιστος συνδυασμός τους. Τέλος, λαμβάνουμε τους μέσους όρους των τιμών των μετρικών ακρίβειας και F-score για τον κάθε αλγόριθμο κατηγοριοποίησης, διαλέγοντας τον καλύτερο συνδυασμό αλγορίθμου και υπερ-παραμέτρων για κάθε διάσχιση.

Χρησιμοποιούμε τρεις μετρικές αξιολόγησης απόδοσης:

- Ακρίβεια: Ελέγχεται μόνο η πρώτη κατηγορία που προβλέπει ο αλγόριθμος κατηγοριοποίησης. Η ακρίβεια, εν συνεχεία, ορίζεται ως η αναλογία των ορθών προβλέψεων με τον αριθμό των ΣΕ, για το σύνολο ελέγχου.
- Top-k ακρίβεια: Ελέγχονται οι πρώτες k κατηγορίες που προβλέπει ο αλγόριθμος κατηγοριοποίησης. Αν τουλάχιστον μία από αυτές αντιστοιχεί στην κατηγορία του ΣΕ υπό εξέταση, τότε η πρόβλεψη θεωρείται σωστή. Εν συνεχεία, η top-k ακρίβεια ορίζεται ως η αναλογία των ορθών προβλέψεων με τον αριθμό των ΣΕ, για το σύνολο ελέγχου. Εν προκειμένω γίνεται χρήση των top-5 και top-10 μετρικών.
- F-score: Είναι ο αρμονικός μέσος όρος των μετρικών precision και recall υπολογισμένων ανά τα ΣΕ που απαρτίζουν το σύνολο ελέγχου, ισοσκελισμένος κατάλληλα ώστε να συνυπολογίζεται και η πιθανή ανισορροπία κατηγοριών στο σύνολο δεδομένων.

Οι αναφερόμενες τιμές των μετρικών αντιστοιχούν στις μέσες τιμές τους και προκύπτουν λαμβάνοντας τους μέσους όρους, εν πρώτοις ως προς τα διαφορετικά fold ελέγχου ανά διαχωρισμό cross-validation και στη συνέχεια ως προς τον αριθμό των folds. Οι αλγόριθμοι συγκρίνονται κατόπιν με την απόδοση της αφελούς προσέγγισης η οποία αντιστοιχεί σε κάθε ΣΕ του συνόλου ελέγχου την πολυπληθέστερη κατηγορία του συνόλου εκπαίδευσης.

Επισημαίνουμε ότι στο συγκεκριμένο πρόβλημα δεν ήταν δυνατόν να οριστεί βασική μέθοδος σύγκρισης προερχόμενη από τις διαδικασίες της Geodata, αφού η λειτουργικότητα που προσφέρουν οι υλοποιημένες μέθοδοι, αυτή τη στιγμή, επιτελείται αποκλειστικά χειροκίνητα από της εταιρία.

4.2.3. Αποτελέσματα

Ο παρακάτω πίνακας παρουσιάζει τις μετρικές για το πρώτο επίπεδο κατηγοριοποίησης, το οποίο αποτελείται από 13 διακριτές κατηγορίες. Οι μετρικές της top-5 και top-10 ακρίβειας για το συγκεκριμένο επίπεδο παραλείπονται, αφού ο χαμηλός συνολικός αριθμός των διαθέσιμων κατηγοριών καθιστά ασήμαντη τη συνεισφορά τους, εφόσον θα οδηγούσε σε υπερβολικά υψηλές τιμές και άρα μη ρεαλιστικές για το πρόβλημά μας. Οι μετρήσεις είναι σε κλίμακα %.

Αλγόριθμος	**Ακρίβεια**	**F-score**
AdaBoost	23.4	22.7
Decision Tree	40.0	24.0
Gaussian Process	35.5	36.9
k-NN	49.4	42.7
MLP	66.9	67.0
Naive Bayes	51.8	51.9
Random Forest	**76.2**	**73.9**
SVM	61.8	59.5
Αφελής προσέγγιση	37.1	20.6

Πίνακας 8: Ακρίβεια μεθόδων για το πρώτο επίπεδο κατηγοριοποίησης

Ο παρακάτω πίνακας παρουσιάζει τις μετρικές για το δεύτερο επίπεδο κατηγοριοποίησης, το οποίο αποτελείται από 71 διακριτές κατηγορίες. Είναι άξιο αναφοράς το γεγονός ότι στο παρόν σενάριο όπου οι μετρικές top-5 και top-10 έχουν νόημα είναι και αρκετά υψηλές. Οι μετρήσεις είναι σε κλίμακα %.

Αλγόριθμος	**Ακρίβεια**	**Top-5 Ακρίβεια**	**Top-10 Ακρίβεια**	**F-score**
AdaBoost	38.6	52.0	64.5	29.6
Decision Tree	13.5	27.9	44.1	4.5
Gaussian Process	29.7	46.2	53.0	32.0
k-NN	34.4	46.5	52.7	32.8
MLP	54.6	70.0	78.4	53.9
Naive Bayes	48.1	54.5	59.9	46.7
Random Forest	**63.1**	**76.0**	**82.2**	**60.3**
SVM	45.3	62.5	72.0	44.3
Αφελής προσέγγιση	12.3	-	-	2.8

Πίνακας 9: Ακρίβεια μεθόδων για το δεύτερο επίπεδο κατηγοριοποίησης

Όπως φαίνεται συγκεντρωτικά στους παραπάνω δύο πίνακες, οι προτεινόμενες μέθοδοι ήδη βελτιώνουν αρχικά την αφελή μέθοδο, παράγοντας ποιοτικότερα, από άποψη ακρίβειας αποτελέσματα. Ασφαλώς, υπάρχουν περιθώρια για περαιτέρω βελτίωση, με περαιτέρω βελτίωση των χαρακτηριστικών εκπαίδευσης και των διαδικασιών μηχανικής μάθησης, τα οποία θα καταγραφούν σε επόμενα παραδοτέα των ΕΕ2 και ΕΕ3.

4.3. Βιβλιοθήκη Γεωκωδικοποίησης Διευθύνσεων

4.3.1. Σύνολο αξιολόγησης

Το σύνολο δεδομένων αποτελείται από πληροφορίες σχετικές με διευθύνσεις οι οποίες αποτελούν προϊόν εταιρείας των εταιριών Ερατοσθένης και Geodata, το οποίο εμπορεύεται σε B2B επίπεδο. Συγκεκριμένα, χρησιμοποιήθηκαν 976 διαφορετικές διευθύνσεις για τις οποίες ακολούθησε συλλογή ζευγών συντεταγμένων από 3 διαφορετικές πηγές γεωκωδικοποίησης: εσωτερική βάση Geodata, ArcGIS[21] και OpenStreetMap[22].

4.3.2. Συνθήκες αξιολόγησης

Η αξιολόγηση ακολουθεί τα τυπικά πρότυπα αξιολόγησης με τη χρήση 5-fold cross-validation. Τα διανύσματα χαρακτηριστικών τροφοδοτούνται σε μια σειρά από αλγορίθμους μηχανικής μάθησης που αναλαμβάνουν την κατηγοριοποίηση. Το σύνολο δεδομένων χωρίζεται σε πέντε (5) ισοπληθή μέρη (folds) τα οποία διασχίζονται πέντε φορές. Κάθε φορά, τρία από τα πέντε μέρη χρησιμοποιούνται ως σύνολο εκπαίδευσης, ένα ως σύνολο επαλήθευσης και ένα ως σύνολο ελέγχου. Κατά αυτόν τον τρόπο κάθε φορά που διασχίζουμε το σύνολο δεδομένων χρησιμοποιούνται και διαφορετικά μέρη του για την εκπαίδευση και την επαλήθευση του μοντέλου. Κάθε διάσχιση περιλαμβάνει, για κάθε επιμέρους αλγόριθμο κατηγοριοποίησης, αναζήτηση του χώρου των υπερ-παραμέτρων που τον συνοδεύουν ώστε να εξασφαλιστεί ο βέλτιστος συνδυασμός τους. Τέλος, λαμβάνουμε τους μέσους όρους των τιμών των μετρικών ακρίβειας και F-score για τον κάθε αλγόριθμο κατηγοριοποίησης, διαλέγοντας τον καλύτερο συνδυασμό αλγορίθμου και υπερ-παραμέτρων για κάθε διάσχιση.

Χρησιμοποιούμε τρεις μετρικές αξιολόγησης απόδοσης:

- Ακρίβεια: Ελέγχεται μόνο η πρώτη κατηγορία (συντεταγμένες γεωκωδικοποίησης) που προβλέπει ο αλγόριθμος κατηγοριοποίησης. Η ακρίβεια, εν συνεχεία, ορίζεται ως η αναλογία των ορθών προβλέψεων προς το συνολικό αριθμό των διευθύνσεων, για το σύνολο ελέγχου.
- F-score: Είναι ο αρμονικός μέσος όρος των μετρικών precision και recall υπολογισμένων ανά τις διευθύνσεις που απαρτίζουν το σύνολο ελέγχου, ισοσκελισμένος κατάλληλα ώστε να συνυπολογίζεται και η πιθανή ανισορροπία κατηγοριών στο σύνολο δεδομένων.

Οι αναφερόμενες τιμές των μετρικών αντιστοιχούν στις μέσες τιμές τους και προκύπτουν λαμβάνοντας τους μέσους όρους, εν πρώτοις ως προς τα διαφορετικά fold ελέγχου ανά διαχωρισμό cross-validation και στη συνέχεια ως προς τον αριθμό των folds. Στη συγκεκριμένη πειραματική διαδικασία, ως μέθοδος βάσης για σύγκριση θεωρείται η προϋπάρχουσα προσέγγιση των εταιριών Ερατοσθένης και Geodata, με βάση την οποία οι

[21] https://geocode.arcgis.com/arcgis/

[22] https://wiki.openstreetmap.org/wiki/Nominatim

συντεταγμένες γεωκωδικοποίησης προέρχονταν αποκλειστικά από το εσωτερικό σύστημα των εταιριών.

4.3.3. Αποτελέσματα

Ο παρακάτω πίνακας παρουσιάζει τις μετρικές για το πρόβλημα γεωκωδικοποίησης διευθύνσεων. Οι μετρήσεις είναι σε κλίμακα %.

Αλγόριθμος	Ακρίβεια	F-score
AdaBoost	56.0	55.0
Decision Tree	60.3	57.2
Gaussian Process	54.8	46.6
k-NN	52.3	51.1
MLP	56.3	49.1
Naive Bayes	16.4	16.2
Random Forest	**62.4**	**61.5**
SVM	56.3	46.5
Βασική (προϋπάρχουσα) προσέγγιση	44.0	-

Πίνακας 10: Ακρίβεια μεθόδων γεωκωδικοποίησης

Όπως φαίνεται στον πίνακα, οι προτεινόμενες μέθοδοι ήδη βελτιώνουν αρχικά την υφιστάμενη μέθοδο, παράγοντας ποιοτικότερα, από άποψη ακρίβειας αποτελέσματα. Συγκεκριμένα, η ακρίβεια πρόβλεψης της βέλτιστης πηγής γεωκωδικοποίησης και των αντίστοιχων συνεταγμένων, από 44% στην υφιστάμενη κατάσταση, ανεβαίνει στο 62.4% με τις προτεινόμενες μεθόδους. Ασφαλώς, υπάρχουν περιθώρια για περαιτέρω βελτίωση, με περαιτέρω βελτίωση των χαρακτηριστικών εκπαίδευσης και των διαδικασιών μηχανικής μάθησης, τα οποία θα καταγραφούν σε επόμενα παραδοτέα των ΕΕ2 και ΕΕ3.

4.4. Βιβλιοθήκη Ολοκλήρωσης Γεωτεμαχίων

4.4.1. Σύνολο αξιολόγησης

Το σύνολο δεδομένων αποτελείται από πληροφορίες γεωτεμαχίων-πολυγώνων οργανωμένες σε shapefiles που αποτελούν προϊόν της εταιρείας Ερατοσθένης και σχετίζεται με τις δραστηριότητές της στον τομέα της κτηματογράφησης. Τα πολύγωνα του συνόλου δεδομένων αντιστοιχούν σε δύο διακριτές κατηγορίες, τις Διανομές και τα PST (ενδιάμεσα αποτελέσματα κτηματογράφησης) και είναι οργανωμένα σε ζεύγη, ένα πολύγωνο από την κάθε κατηγορία, τα οποία γεωγραφικά επικαλύπτονται. Στη συγκεκριμένη πειραματική διαδικασία χρησιμοποιήθηκαν 1786 ζεύγη πολυγώνων.

4.4.2. Συνθήκες αξιολόγησης

Η αξιολόγηση ακολουθεί τα τυπικά πρότυπα αξιολόγησης με τη χρήση 5-fold cross-validation. Τα διανύσματα χαρακτηριστικών τροφοδοτούνται σε μια σειρά από αλγορίθμους μηχανικής μάθησης που αναλαμβάνουν την κατηγοριοποίηση. Το σύνολο δεδομένων χωρίζεται σε πέντε (5) ισοπληθή μέρη (folds) τα οποία διασχίζονται πέντε φορές. Κάθε φορά, τρία από τα πέντε μέρη χρησιμοποιούνται ως σύνολο εκπαίδευσης, ένα ως σύνολο επαλήθευσης και ένα ως σύνολο ελέγχου. Κατά αυτόν τον τρόπο κάθε φορά που διασχίζουμε το σύνολο δεδομένων χρησιμοποιούνται και διαφορετικά μέρη του για την εκπαίδευση και την επαλήθευση του μοντέλου. Κάθε διάσχιση περιλαμβάνει, για κάθε επιμέρους αλγόριθμο κατηγοριοποίησης, αναζήτηση του χώρου των υπερ-παραμέτρων που τον συνοδεύουν ώστε να εξασφαλιστεί ο βέλτιστος συνδυασμός τους. Τέλος, λαμβάνουμε τους μέσους όρους των τιμών των μετρικών ακρίβειας και f-score για τον κάθε αλγόριθμο κατηγοριοποίησης διαλέγοντας τον καλύτερο συνδυασμό αλγορίθμου και υπερ-παραμέτρων για κάθε διάσχιση.

Χρησιμοποιούμε τρεις μετρικές αξιολόγησης απόδοσης:

- Ακρίβεια: Ελέγχεται μόνο η πρώτη κατηγορία που προβλέπει ο αλγόριθμος κατηγοριοποίησης. Η ακρίβεια, εν συνεχεία, ορίζεται ως η αναλογία των ορθών προβλέψεων ως προς το συνολικό αριθμό ζευγών πολυγώνων, για το σύνολο ελέγχου.
- F-score: Είναι ο αρμονικός μέσος όρος των μετρικών precision και recall υπολογισμένων ανά τα ζεύγη πολυγώνων που απαρτίζουν το σύνολο ελέγχου, ισοσκελισμένος κατάλληλα ώστε να συνυπολογίζεται και η πιθανή ανισορροπία κατηγοριών στο σύνολο δεδομένων.

Οι αναφερόμενες τιμές των μετρικών αντιστοιχούν στις μέσες τιμές τους και προκύπτουν λαμβάνοντας τους μέσους όρους, εν πρώτοις ως προς τα διαφορετικά fold ελέγχου ανά διαχωρισμό cross-validation και στη συνέχεια ως προς τον αριθμό των folds.

Επισημαίνουμε ότι στο συγκεκριμένο πρόβλημα δεν ήταν δυνατόν να οριστεί βασική μέθοδος σύγκρισης, αφού η λειτουργικότητα που προσφέρουν οι υλοποιημένες μέθοδοι, αυτή τη στιγμή επιτελείται αποκλειστικά χειροκίνητα από την εταιρία Ερατοσθένης.

4.4.3. Αποτελέσματα

Ο παρακάτω πίνακας παρουσιάζει τις μετρικές για το πρόβλημα κατηγοριοποίησης των πολυγώνων. Οι μετρήσεις είναι σε κλίμακα %.

Αλγόριθμος	Ακρίβεια	F-score
AdaBoost	51.0	49.9
Decision Tree	51.3	49.5
Gaussian Process	**53.9**	49.8
k-NN	52.8	51.8
MLP	50.8	40.6
Naive Bayes	52.3	49.4
Random Forest	52.5	**52.2**
SVM	52.4	48.9

Πίνακας 11: Ακρίβεια μεθόδων ολοκλήρωσης γεωτεμαχίων

Παρατηρούμε ότι υπάρχουν μεγάλα περιθώρια για περαιτέρω βελτίωση, με περαιτέρω βελτίωση των χαρακτηριστικών εκπαίδευσης και των διαδικασιών μηχανικής μάθησης, τα οποία θα καταγραφούν σε επόμενα παραδοτέα των ΕΕ2 και ΕΕ3.

5. Σύνοψη

Στο Παραδοτέο 2.1 παρουσιάσαμε τις αρχικές εκδόσεις των μεθόδων μηχανικής μάθησης για διασύνδεση, κατηγοριοποίηση, γεωκωδικοποίηση και ολοκλήρωση χωρο-κειμενικών δεδομένων που υλοποιήσαμε στο πλαίσιο της ΕΕ2 του έργου. Η παρουσίαση των τεσσάρων αρχικών βιβλιοθηκών κώδικα ακολουθεί τον διαχωρισμό των τεσσάρων σεναρίων χρήσης που ορίστηκε στο Π1.1. Αρχικά, παρουσιάστηκε η κοινή γενική αρχιτεκτονική που μοιράζονται οι τέσσερις βιβλιοθήκες, οι οποία βασίζεται σε ουσιαστικές αρχές και απαιτήσεις υλοποίησης και λειτουργικότητας των εταίρων του έργου. Στη συνέχεια, παρουσιάστηκε αναλυτική τεκμηρίωση υλοποίησης και λειτουργικότητας των τεσσάρων βιβλιοθηκών, καθώς και αρχικά πειράματα αξιολόγησής τους στα σύνολα αξιολόγησης του έργου.

Επεκτάσεις και βελτιώσεις των παραπάνω βιβλιοθηκών, σε επίπεδο χαρακτηριστικών εκπαίδευσης, αλγορίθμων και διαδικασιών μηχανικής μάθησης, καθώς και πειράματα επαναξιολόγησής τους θα καταγράφονται στα επόμενα παραδοτέα των ΕΕ2 και ΕΕ3.

6. Αναφορές

[DOR+14]	Nilesh Dalvi, Marian Olteanu, Manish Raghavan, and Philip Bohannon. 2014. Deduplicating a Places Database. In Proceedings of WWW '14.