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

Π3.1: «Μηχανισμοί παραμετροποίησης και επιλογής συνόλων χαρακτηριστικών εκπαίδευσης»

Πακέτο Εργασίας	**ΠΕ3: Προηγμένες μέθοδοι μάθησης διασύνδεσης**
Υπεύθυνος Φορέας	**Ε.Κ. «Αθηνά» / ΙΠΣΥ**
Είδος Παραδοτέου	**Λογισμικό**
Ενδεικτικός Μήνας Παράδοσης	**Μ21**
Ημερομηνία Παράδοσης	**8/4/2020 (Μ21)**

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	02/12/2019	Δομή και πίνακας περιεχομένων του παραδοτέου	Γιώργος Γιαννόπουλος (ΑΘ.), Βασιλική Χήρα (ΑΘ.)
0.2	06/12/2019	Προσθήκη εισαγωγικού υλικού για μεθόδους επιλογής χαρακτηριστικών	Βασιλική Χήρα (ΑΘ.), Κωνσταντίνος Αλέξης (ΑΘ.)
0.3	17/12/2019	Προσθήκη υλικού στην αξιολόγηση των μεθόδων	Βασιλική Χήρα (ΑΘ.), Βασίλης Καφφές (ΑΘ.), Γιώργος Γιαννόπουλος (ΑΘ.)
0.4	23/12/2019	Διάφορες προσθήκες και βελτιώσεις	Βασιλική Χήρα (ΑΘ.),
0.5	24/01/2020	Διάφορες προσθήκες και βελτιώσεις	Βασίλης Καφφές (ΑΘ.), Κωνσταντίνος Αλέξης (ΑΘ.)
0.6	28/02/2020	Διάφορες προσθήκες και βελτιώσεις	Γιώργος Γιαννόπουλος (ΑΘ.), Κωνσταντίνος Αλέξης (ΑΘ.)
0.7	31/03/2020	Εσωτερικά επιθεωρημένη έκδοση	Δημήτριος Σκούτας (ΑΘ.), Γιώργος Γιαννόπουλος (ΑΘ.), Βασίλης Καφφές (ΑΘ.), Κωνσταντίνος Αλέξης (ΑΘ.), Νώντας Τσάκωνας (ΕΡ.), Ηλίας Βάρκας (Ge.)
1.0	08/04/2020	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 επεκτεταμένη έκδοση των βιβλιοθηκών κώδικα που υλοποιούν μοντέλα μηχανικής μάθησης, οι οποίες αναπτύχθηκαν στο έργο για τη διασύνδεση, επισημείωση και ολοκλήρωση γεωχωρικών δεδομένων. Έχοντας ως βάση το Π2.1 «Αλγόριθμοι μηχανικής μάθησης για διασύνδεση», το παρόν παραδοτέο επικεντρώνει στους *μηχανισμούς επιλογής χαρακτηριστικών* (feature selection) που ενσωματώθηκαν στις παραπάνω βιβλιοθήκες μηχανικής μάθησης. Συγκεκριμένα, το παρόν παραδοτέο περιγράφει την υλοποίηση μηχανισμών επιλογής χαρακτηριστικών για τις βιβλιοθήκες *διασύνδεσης τοπωνυμίων*, *κατηγοριοποίησης Σημείων Ενδιαφέροντος (ΣΕ)* και *γεωκωδικοποίησης*[1]. Επιπλέον πραγματοποιείται συγκριτική αξιολόγηση της ακρίβειας των τριών βιβλιοθηκών, με χρήση και χωρίς χρήση των μηχανισμών επιλογής χαρακτηριστικών, και αναλύονται τα αποτελέσματα για την κάθε επιμέρους βιβλιοθήκη.

Το παρόν παραδοτέο περιγράφει ως επί το πλείστον τις εργασίες που πραγματοποιήθηκαν στο πλαίσιο της Υποενότητας Εργασίας 3.1. Συνοπτικά, εξετάσαμε, υλοποιήσαμε και ενσωματώσαμε μία σειρά από μεθόδους επιλογής χαρακτηριστικών, οι οποίες αντιπροσωπεύουν και τις τρεις ευρύτερες κατηγορίες της περιοχής: (α) Μέθοδοι περιτυλίγματος (wrapper methods), (β) Μέθοδοι διήθησης (filter methods) και (γ) ενσωματωμένες μέθοδοι (embedded methods). Με την προσθήκη των παραπάνω μηχανισμών, οι οποίοι είναι πλήρως παραμετροποιήσιμοι, ολοκληρώνεται η γενική, κοινή αρχιτεκτονική των υλοποιημένων βιβλιοθηκών, όπως παρουσιάζεται στο Κεφάλαιο 0 του παρόντος κειμένου και η οποία είχε αρχικά παρουσιαστεί στο Π2.1. Οι μηχανισμοί επιλογής χαρακτηριστικών εφαρμόζονται με σκοπό την επιλογή των πιο αντιπροσωπευτικών/χρήσιμων χαρακτηριστικών από το σύνολο των χαρακτηριστικών που υλοποιεί κάθε βιβλιοθήκη (τα σύνολα αυτά των χαρακτηριστικών τεκμηριώνονται αναλυτικά στο Π1.2 «Προδιαγραφή εξειδικευμένων χαρακτηριστικών και κανόνων εκπαίδευσης»). Ο στόχος των παραπάνω μεθόδων είναι η βελτίωση της απόδοσης, αλλά και της αποτελεσματικότητας των υλοποιημένων βιβλιοθηκών μηχανικής μάθησης, μέσω του φιλτραρίσματος των λιγότερο σημαντικών/επιβλαβών χαρακτηριστικών.

Το παραδοτέο δομείται ως εξής:

Στο Κεφάλαιο 1 πραγματοποιείται μία εισαγωγή σε μεθόδους επιλογής χαρακτηριστικών εκπαίδευσης και αναλύεται η σημασία τους σε σχέση με τις μεθόδους μηχανικής μάθησης για ολοκλήρωση χωρο-κειμενικών δεδομένων που αναπτύσσονται στο έργο.

Στο Κεφάλαιο 2 παρουσιάζονται οι μηχανισμοί επιλογής χαρακτηριστικών που υλοποιήθηκαν, καθώς και η ανανεωμένη γενική αρχιτεκτονική των βιβλιοθηκών μηχανικής

[1] Σημειώνουμε ότι για την τέταρτη βιβλιοθήκη που περιγράφεται στο Π2.1 δεν ενσωματώθηκαν μηχανισμοί επιλογής χαρακτηριστικών. Αυτό οφείλεται στο γεγονός ότι, μετά από βελτιώσεις και επεκτάσεις που πραγματοποιήθηκαν στη συγκεκριμένη βιβλιοθήκη, επιτυγχάνεται πλέον βέλτιστη ακρίβεια ολοκλήρωσης (~99%) και, ως εκ τούτου, η ενσωμάτωση μηχανισμών επιλογής χαρακτηριστικών θα ήταν περιττή. Οι συγκεκριμένες επεκτάσεις θα περιγραφούν αναλυτικά στο Π2.2 «Βελτιστοποιημένοι αλγόριθμοι μηχανικής μάθησης για διασύνδεση» το οποίο είναι προς παράδοση το Μ24 του έργου.

μάθησης, η οποία συνίσταται σε μία ακολουθία διεργασιών που καλύπτουν όλο το φάσμα μίας διαδικασίας εκπαίδευσης και εφαρμογής μοντέλων μηχανικής μάθησης, συμπεριλαμβανομένης πλέον και της διαδικασίας επιλογής χαρακτηριστικών.

Στο Κεφάλαιο 3, επικαιροποιείται η περιγραφή του λογισμικού που υλοποιήθηκε στη μορφή βιβλιοθηκών μηχανικής μάθησης, οργανωμένο ανά σενάριο χρήσης/επιλυόμενο πρόβλημα μηχανικής μάθησης για διασύνδεση, κατηγοριοποίηση και γεωκωδικοποίηση χωρο-κειμενικών δεδομένων. Οι περιγραφές περιλαμβάνουν πληροφορίες υλοποίησης, τεκμηρίωσης, άδειας χρήσης και πρόσβασης του υλοποιημένου κώδικα, καθώς και οδηγούς εγκατάστασης και εκτέλεσης των βιβλιοθηκών. Εξαιρείται η περιγραφή της βιβλιοθήκης ολοκλήρωσης γεωχωρικών δεδομένων, η οποία δεν κρίθηκε σκόπιμο να εμπλουτισθεί με μεθόδους επιλογής χαρακτηριστικών, και η ανανεωμένη έκδοση της οποίας θα παρουσιασθεί στο Π2.2.

Στο Κεφάλαιο 4 παρουσιάζονται τα αποτελέσματα της συγκριτικής πειραματικής αξιολόγησης των υλοποιημένων μεθόδων, με και χωρίς τη χρήση μεθόδων επιλογής χαρακτηριστικών, και αναλύονται οι διαφορές στην αποτελεσματικότητά τους.

Στο Κεφάλαιο 5 συνοψίζεται το παραδοτέο και γίνεται μία σύντομη παράθεση των ευρημάτων μας σχετικά με τη χρησιμότητα των υλοποιημένων μεθόδων επιλογής χαρακτηριστικών.

Σημειώνουμε ότι, για λόγους πληρότητας της τεκμηρίωσης και ευκολότερης κατανόησης του κειμένου από τον αναγνώστη, το παρόν παραδοτέο επαναχρησιμοποιεί περιγραφές από το Π2.1, ειδικά στο Κεφάλαιο 3, όπου γίνεται αναλυτική τεκμηρίωση των υλοποιημένων βιβλιοθηκών. Για αυτό το λόγο, όπου κρίνεται σκόπιμο, και ως επί το πλείστον στο Κεφάλαιο 3, χρησιμοποιείται έντονη και πλάγια γραφή (bold, italics) για να καταδείξει επεκτάσεις των βιβλιοθηκών που αφορούν τις μεθόδους επιλογής χαρακτηριστικών, κυρίως σε σχέση με τα προϋπάρχοντα υποσυστήματα και μηχανισμούς που παρουσιάστηκαν στο Π2.1.

1. Εισαγωγή

Στο παρόν κεφάλαιο, πρώτα γίνεται μία εισαγωγική παρουσίαση των βασικών εννοιών και τεχνικών επιλογής χαρακτηριστικών εκπαίδευσης και, στη συνέχεια, συζητείται η αναγκαιότητα υιοθέτησης και αξιολόγησης τέτοιων μεθόδων στα σενάρια μηχανικής μάθησης που εξετάζονται στο έργο.

1.1. Εισαγωγή στην επιλογή χαρακτηριστικών

Η επιλογή χαρακτηριστικών (feature selection), γνωστή και ως επιλογή μεταβλητών, είναι η διαδικασία με την οποία επιλέγεται ένα υποσύνολο σχετικών χαρακτηριστικών για την κατασκευή ενός μοντέλου. Οι τεχνικές επιλογής χαρακτηριστικών χρησιμοποιούνται για την αναγνώριση και αφαίρεση αλληλεπικαλυπτόμενων, μη σχετικών, περιττών ή μη χρήσιμων χαρακτηριστικών που, είτε δεν βελτιώνουν την αξιοπιστία ενός μοντέλου, είτε, μάλιστα, την μειώνουν. Ο μειωμένος αριθμός χαρακτηριστικών είναι επιθυμητός καθώς μειώνει τη πολυπλοκότητα του μοντέλου, μειώνει τις απαιτήσεις σε υπολογιστικούς πόρους, πόσο μάλλον σε μοντέλα που έχουν μεγάλο πλήθος χαρακτηριστικών εκπαίδευσης, και συμβάλλει στη καλύτερη κατανόηση της διαδικασίας παραγωγής αποτελεσμάτων από τα αντίστοιχα μοντέλα μηχανικής μάθησης. Περαιτέρω οφέλη που δύνανται να προσκομισθούν από αυτή την τεχνική είναι η διευκόλυνση της οπτικοποίησης και ανάλυσης των δεδομένων, η αποφυγή υπερ-εκπαίδευσης (overfitting), και άρα η καλύτερη γενίκευση (generalization) των μοντέλων μάθησης, η μείωση των απαιτήσεων αποθήκευσης, η μείωση των χρόνων εκπαίδευσης του μοντέλου και η αποτροπή της κατάρας της διαστασιμότητας (curse of dimensionality). Η επιλογή μεθόδου για την επιλογή χαρακτηριστικών ποικίλει ανάλογα με την εστίαση σε έναν ή περισσότερους από τους παραπάνω στόχους.

Ένας αλγόριθμος επιλογής χαρακτηριστικών μπορεί να αποτελείται από ένα συνδυασμό τεχνικών εύρεσης υποσυνόλων χαρακτηριστικών που συνοδεύεται από ένα μέτρο αξιολόγησης που βαθμολογεί τα διαφορετικά υποσύνολα χαρακτηριστικών. Η επιλογή του μέτρου αξιολόγησης, καθώς και ο βαθμός ενσωμάτωσης της διαδικασίας επιλογής χαρακτηριστικών στη κατασκευή του εκάστοτε μοντέλου μάθησης διακρίνει τις τεχνικές σε τρεις επιμέρους κατηγορίες, όπως αναλύονται ακολούθως.

1.1.1. Μέθοδοι περιτυλίγματος (wrapper methods)

Οι συγκεκριμένες μέθοδοι χρησιμοποιούν ένα μοντέλο πρόβλεψης για να βαθμολογήσουν υποσύνολα χαρακτηριστικών. Κάθε νέο υποσύνολο χρησιμοποιείται για να εκπαιδεύσει το μοντέλο, το οποίο αξιολογείται σε ένα υποσύνολο των δεδομένων και εκ του οποίου προκύπτει μια βαθμολογία. Τα υποσύνολα δημιουργούνται ξεκινώντας: (α) είτε από ένα κενό σύνολο ή ένα σύνολο με ένα ελάχιστο αριθμό χαρακτηριστικών και στη συνέχεια προστίθενται χαρακτηριστικά ανάλογα με το αν το κάθε χαρακτηριστικό που προστίθεται βελτιώνει τη βαθμολογία του μοντέλου πρόβλεψης, (β) είτε αντίστροφα, ξεκινώντας δηλαδή με το σύνολο των χαρακτηριστικών εκπαίδευσης και αφαιρώντας χαρακτηριστικά. Βασικό πλεονέκτημα αυτών των μεθόδων είναι ότι λαμβάνουν υπόψη τον αλγόριθμο μηχανικής μάθησης και έτσι το σύνολο των χαρακτηριστικών που θα επιλεχθεί είναι

βελτιστοποιημένο ως προς αυτόν. Ωστόσο, σε αντιπαραβολή βρίσκεται το υπολογιστικό κόστος που απαιτείται για την εξέταση των υποσυνόλων, καθώς και ο κίνδυνος υπερ-εκπαίδευσης των μοντέλων σε περιπτώσεις μη επαρκών δεδομένων. Τεχνικές που ανήκουν σε αυτή τη κατηγορία είναι η *Backward Elimination* [WFH+17] και η *Forward Feature Selection* [AB96].

1.1.2. Μέθοδοι διήθησης (filter methods)

Οι συγκεκριμένες μέθοδοι επιλέγουν χαρακτηριστικά που έχουν επιτύχει τη μεγαλύτερη τιμή μίας συγκεκριμένης μετρικής/τεστ αξιολόγησης [SAT07]. Αυτό σημαίνει ότι αυτές οι μέθοδοι δε λαμβάνουν υπόψη κάποιο συγκεκριμένο μοντέλο μηχανικής μάθησης, αλλά βασίζονται περισσότερο σε στατιστικές που υποδηλώνουν το βαθμό συσχέτισης των χαρακτηριστικών μεταξύ τους αλλά και με τη μεταβλητή εξόδου (δηλαδή την πρόβλεψη του μοντέλου μηχανικής μάθησης). Προκύπτει, έτσι, μια ταξινόμηση χαρακτηριστικών που είναι πιο χρήσιμη στην αποκάλυψη της σχέσης μεταξύ των χαρακτηριστικών, που έχει μεν μικρότερη υπολογιστική επιβάρυνση σε σχέση με τις μεθόδους περιτυλίγματος, αλλά δεν παρέχει απαραίτητα το καλύτερο υποσύνολο χαρακτηριστικών. Αντιπρόσωποι αυτής της κατηγορίας είναι αλγόριθμοι που χρησιμοποιούν ως μετρικές την αμοιβαία πληροφορία (Mutual Information), το τεστ X^2 (Chi square) και το κατώφλι διακύμανσης (Variance Threshold).

1.1.3. Ενσωματωμένες μέθοδοι (embedded methods)

Οι μέθοδοι αυτές βρίσκονται ανάμεσα στις δύο παραπάνω κατηγορίες και πραγματοποιούν τη διαδικασία επιλογής των χαρακτηριστικών ως κομμάτι της διαδικασίας κατασκευής του μοντέλου και, ως εκ τούτου, εξειδικεύονται ανά αλγόριθμο μηχανικής μάθησης [CWE06]. Οι πιο συνήθεις μέθοδοι που ανήκουν σε αυτήν την κατηγορία είναι οι μέθοδοι κανονικοποίησης-περιορισμού (regularization). Αυτές επιβάλλουν έναν κανόνα "τιμωρίας" στα επιμέρους χαρακτηριστικά, δηλαδή περιορισμούς στη βελτιστοποίηση ενός αλγορίθμου πρόβλεψης, με στόχο να συμπιεστούν ή ακόμα και να μηδενιστούν οι βαρύτητες συγκεκριμένων χαρακτηριστικών, ώστε το επιστρεφόμενο μοντέλο μηχανικής μάθησης να είναι χαμηλότερης πολυπλοκότητας. Ο όρος περιορισμού (regularization term) που εισάγεται στη συνάρτηση κόστους-βελτιστοποίησης (optimization function) περιορίζει τα βάρη των χαρακτηριστικών που δεν προσφέρουν πληροφορία για τη βελτίωση του μοντέλου. Τέτοιες τεχνικές περιορισμού είναι οι Lasso[2], Ridge Regresssion[3] και Elastic Net[4].

[2] https://scikit-learn.org/stable/modules/linear_model.html#lasso

[3] https://scikit-learn.org/stable/modules/linear_model.html#ridge-regression-and-classification

[4] https://scikit-learn.org/stable/modules/linear_model.html#elastic-net

1.2. Σημασία επιλογής χαρακτηριστικών στην διασύνδεση χωρο-κειμενικών δεδομένων

Στο πλαίσιο του έργου, αντιμετωπίζουμε προβλήματα διασύνδεσης, επισημείωσης, εηεωκωδικοποίησης και γενικότερα ολοκλήρωσης χωρο-κειμενικών δεδομένων από διαφορετικές πηγές δεδομένων, με εφαρμογή σε διάφορα εμπορικά σενάρια χρήσης (κτηματογράφηση, γεωκωδικοποίηση, γεωχωρική ανάλυση). Για το σκοπό αυτό, στο πλαίσιο της ανάπτυξης πρωτότυπων μεθόδων μηχανικής μάθησης για την επίλυση των παραπάνω προβλημάτων, ορίσαμε και υλοποιήσαμε ένα εκτεταμένο σύνολο από χαρακτηριστικά εκπαίδευσης, βασισμένα στην εμπειρική γνώση πεδίου των εταίρων του έργου, τόσο γενικά στα γεωχωρικά δεδομένα, όσο και ειδικότερα στα επιμέρους προβλήματα. Τα ορισμένα χαρακτηριστικά εκπαίδευσης περιγράφονται αναλυτικά στο Π1.2.

Ένα υποσύνολο των παραπάνω χαρακτηριστικών βασίζονται και επεκτείνουν/βελτιώνουν προηγούμενες εργασίες στην περιοχή, όπως για παράδειγμα κάποια από τα χαρακτηριστικά εκπαίδευσης που χρησιμοποιούνται στη βιβλιοθήκη διασύνδεσης τοπωνυμίων (βλέπε Π1.2 – Κεφάλαιο 2.1.4), ενώ τα περισσότερα χαρακτηριστικά που ορίζονται στις υπόλοιπες βιβλιοθήκες (βλέπε Π1.2 – Κεφάλαια 2.2.4, 2.3.4, 2.4.4) αποτελούν πρωτότυπα χαρακτηριστικά που ενσωματώνονται πρώτη φορά σε αλγόριθμους μηχανικής μάθησης για την επίλυση εξειδικευμένων προβλημάτων ολοκλήρωσης.

Η πλειονότητα των οριζόμενων χαρακτηριστικών βασίζεται σε καλά θεμελιωμένη θεωρητική, εμπειρική και διαισθητική γνώση των προβλημάτων που επιλύονται και των αντίστοιχων δεδομένων. Παρόλα αυτά, είναι αναγκαία η εφαρμογή μίας μεθοδολογίας η οποία αξιολογεί με μεθοδικότερο και πιο εξαντλητικό τρόπο τη χρησιμότητα, τη μοναδικότητα και την αποτελεσματικότητα των ορισμένων αυτών χαρακτηριστικών, με στόχο τη βελτίωση της αποτελεσματικότητας των αλγορίθμων μηχανικής μάθησης στους οποίους χρησιμοποιούνται. Η μεθοδολογία αυτή έγκειται στην ενσωμάτωση μίας σειράς από μεθόδους επιλογής χαρακτηριστικών στη ροή μηχανικής μάθησης των επιμέρους βιβλιοθηκών, οι οποίες θα αξιολογούν και θα επιλέγουν τα πιο χρήσιμα χαρακτηριστικά εκπαίδευσης στα πρώτα βήματα κάθε ροής, με στόχο τη βελτίωση της τελικής ακρίβειας των μοντέλων διασύνδεσης, κατηγοριοποίησης, γεωκωδικοποίησης και ολοκλήρωσης δεδομένων.

2. Μηχανισμοί επιλογής χαρακτηριστικών και ανανεωμένη αρχιτεκτονική βιβλιοθηκών

Στο παρόν κεφάλαιο, πρώτα περιγράφονται οι μηχανισμοί επιλογής που υλοποιήθηκαν και οι οποίοι καλύπτουν τις τρεις βασικές μεθοδολογίες που παρουσιάστηκαν στο προηγούμενο κεφάλαιο και, στη συνέχεια, παρουσιάζεται η ανανεωμένη γενική αρχιτεκτονική των βιβλιοθηκών μηχανικής μάθησης, η οποία ενσωματώνει τους παραπάνω μηχανισμούς επιλογής χαρακτηριστικών.

2.1. Επιλογή Χαρακτηριστικών Εκπαίδευσης

Στη συνέχεια περιγράφονται αναλυτικά οι μέθοδοι επιλογής χαρακτηριστικών που ενσωματώθηκαν στις βιβλιοθήκες μηχανικής μάθησης.

2.1.1. Επαναλαμβανόμενη Κατάργηση Χαρακτηριστικού (Recursive Feature Elimination)

Η μέθοδος RFE είναι μια τεχνική που υιοθετεί τη λογική της προς τα πίσω εξάλειψης (backward elimination) και ανήκει στη κατηγορία των *μεθόδων περιτυλίγματος*, καθώς στην αρχή της επαναληπτικής διαδικασίας περιλαμβάνει το σύνολο των χαρακτηριστικών και σε κάθε επανάληψη αφαιρεί αυτά που είναι πιο "αδύναμα" χρησιμοποιώντας έναν εκτιμητή (estimator). Τα χαρακτηριστικά αφαιρούνται σταδιακά, έως ότου ο προκαθορισμένος (μέσω παραμετροποίησης) αριθμός χαρακτηριστικών συμπληρωθεί. Τα χαρακτηριστικά κατατάσσονται ανάλογα με το μοντέλο που εφαρμόζεται, είτε με βάση συντελεστές (coefficients), είτε με τη σημαντικότητα των χαρακτηριστικών (feature importance) και αφαιρούνται σε κάθε επανάληψη, καθώς ο αλγόριθμος προσπαθεί να εξαλείψει τη συγγραμικότητα (collinearity) και την αλληλεξάρτηση που μπορεί να υφίσταται στο μοντέλο. Στο πλαίσιο της συγκεκριμένης μεθόδου, δίνεται η δυνατότητα αυτόματης αναζήτησης του βέλτιστου αριθμού των χαρακτηριστικών, μέσω διαδικασίας συγκριτικής αξιολόγησης (cross-validation). Η υλοποίηση και ενσωμάτωση της συγκεκριμένης μεθόδου στις βιβλιοθήκες μηχανικής μάθησης βασίστηκε στην ανοικτού κώδικα βιβλιοθήκη scikit-learn[5].

2.1.2. Κατώφλι Διακύμανσης (Variance Threshold)

Είναι μια σχετικά απλή τεχνική για αφαίρεση μη σημαντικών χαρακτηριστικών. Βασίζεται στην ιδέα ότι όταν οι τιμές που λαμβάνει ένα χαρακτηριστικό έχουν σχετικά μικρή μεταβολή σε όλα στοιχεία ενός συνόλου δεδομένων (έχει δηλαδή μικρή διακύμανση) τότε

[5] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html

δε μπορεί να αξιοποιηθεί για την εύρεση μοτίβων που παρέχουν πληροφορία και άρα μπορεί να αφαιρεθεί από τα δεδομένα. Δεδομένου ενός ορίου (κατώφλι) διακύμανσης αφαιρούνται, λοιπόν, τα χαρακτηριστικά που έχουν διακύμανση μικρότερη αυτού. Η συγκεκριμένη τεχνική εμπίπτει στη κατηγορία των *μεθόδων διήθησης*. Η υλοποίηση και ενσωμάτωση της συγκεκριμένης μεθόδου στις βιβλιοθήκες μηχανικής μάθησης βασίστηκε στην ανοικτού κώδικα βιβλιοθήκη scikit-learn[6].

2.1.3. X2 (Chi-squared)

Το τεστ αυτό σημαντικότητας ελέγχει την ανεξαρτησία (independence) δύο μεταβλητών. Έχοντας τις τιμές που λαμβάνουν τα στοιχεία ενός συνόλου για δύο μεταβλητές, μπορούμε να υπολογίσουμε το παρατηρούμενο πλήθος και το αναμενόμενο πλήθος. Το X^2 μετράει πως διαφέρουν το αναμενόμενο πλήθος και το παρατηρούμενο. Σε σχέση με την επιλογή των χαρακτηριστικών, στόχος είναι να επιλέξουμε τα χαρακτηριστικά εκείνα που είναι σε μεγάλο βαθμό εξαρτημένα με την τιμή πρόβλεψης του αλγορίθμου (εν προκειμένω την κλάση/ετικέτα του κάθε στοιχείου). Υψηλότερη τιμή του X^2 σημαίνει ότι το χαρακτηριστικό και η τιμή πρόβλεψης είναι σε μεγάλο βαθμό εξαρτημένα, οπότε το συγκεκριμένο χαρακτηριστικό παρέχει πληροφορία και άρα πρέπει να επιλεχθεί. Η συγκεκριμένη τεχνική εμπίπτει στη κατηγορία *μεθόδων διήθησης*. Η υλοποίηση και ενσωμάτωση της συγκεκριμένης μεθόδου στις βιβλιοθήκες μηχανικής μάθησης βασίστηκε στην ανοικτού κώδικα βιβλιοθήκη scikit-learn[7], σε συνδυασμό με τη μέθοδο βαθμολόγησης/επιλογής των k καλύτερων χαρακτηριστικών SelectKBest[8].

2.1.4. Επιλογή μέσω του Μοντέλου (Select From Model)

Αυτή η τεχνική ανήκει στις *ενσωματωμένες μεθόδους* και εφαρμόζεται σε μοντέλα μηχανικής μάθησης που (α) είτε έχουν ενσωματωμένο κάποιον όρο κανονικοποίησης-περιορισμού όπως είναι οι νόρμες L1 ή L2 σε γραμμικούς αλγορίθμους μηχανικής μάθησης, (β) είτε υπολογίζουν τη σημαντικότητα των χαρακτηριστικών (feature importance) – όπως τα Δέντρα Αποφάσεων που χρησιμοποιούν το σκορ impurity σε κάθε κόμβο και, κατά την εκπαίδευση, μπορεί να υπολογιστεί κατά πόσο μειώνεται αυτό το σκορ από κάθε επιλογή χαρακτηριστικού και άρα να προκύψει έτσι μια κατάταξη των χαρακτηριστικών. Για αυτή τη μέθοδο πρέπει να οριστεί ένα κατώφλι το οποίο ορίζει πάνω από ποιο όριο ένα χαρακτηριστικό παραμένει στο βέλτιστο υποσύνολο χαρακτηριστικών. Η υλοποίηση και ενσωμάτωση της συγκεκριμένης μεθόδου στις βιβλιοθήκες μηχανικής μάθησης βασίστηκε στην ανοικτού κώδικα βιβλιοθήκη scikit-learn[9].

[6] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.VarianceThreshold.html

[7] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html

[8] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html

[9] https://scikit-learn.org/stable/modules/feature_selection.html#feature-selection-using-selectfrommodel

2.2. Ανανεωμένη γενική αρχιτεκτονική βιβλιοθηκών

Στο παραδοτέο Π2.1 παρουσιάστηκε η πρώτη έκδοση της γενικής αρχιτεκτονικής βιβλιοθηκών η οποία αποτελούταν από τέσσερα αρθρώματα:

- Άρθρωμα 1: Αξιολόγηση αλγορίθμων μηχανικής μάθησης και επιλογή βέλτιστου αλγορίθμου
- Άρθρωμα 2: Εύρεση βέλτιστου μοντέλου μηχανικής μάθησης (αλγόριθμος συνδυαστικά με τη βέλτιστη παραμετροποίησή του)
- Άρθρωμα 3: Εκπαίδευση μοντέλου μηχανικής μάθησης
- Άρθρωμα 4: Εκτέλεση μοντέλου μηχανικής μάθησης

Στο παρόν παραδοτέο ολοκληρώνεται η γενική αρχιτεκτονική των βιβλιοθηκών, προσθέτοντας το Άρθρωμα 0, το οποίο ενσωματώνει τους μηχανισμούς επιλογής χαρακτηριστικών εκπαίδευσης. Η ανανεωμένη γενική αρχιτεκτονική ενσωματώνει όλα τα βασικά βήματα μίας πλήρους ροής μηχανικής μάθησης: επιλογή χαρακτηριστικών, 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 πέντε αυτές διακριτές φάσεις υλοποιούνται στα πέντε διακριτά αρθρώματα, που περιγράφηκαν παραπάνω, όπως παρουσιάζεται στην αρχιτεκτονική της Εικόνα 1.

Συγκεκριμένα, το Άρθρωμα 0 εκτελεί το πρώτο βήμα της διαδικασίας, δηλαδή την επιλογή του βέλτιστου υποσυνόλου χαρακτηριστικών εκπαίδευσης. Βασιζόμενο στα επιλεγμένα χαρακτηριστικά από το προηγούμενο βήμα, το Άρθρωμα 1 εκτελεί 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 και, προαιρετικά, εκ νέου επιλογή χαρακτηριστικών με δεδομένο πλέον το βέλτιστο αλγόριθμο (με κλήση του Αρθρώματος 0), ούτως ώστε να επιλεγεί η βέλτιστη υπερπαραμετροποίηση, 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.

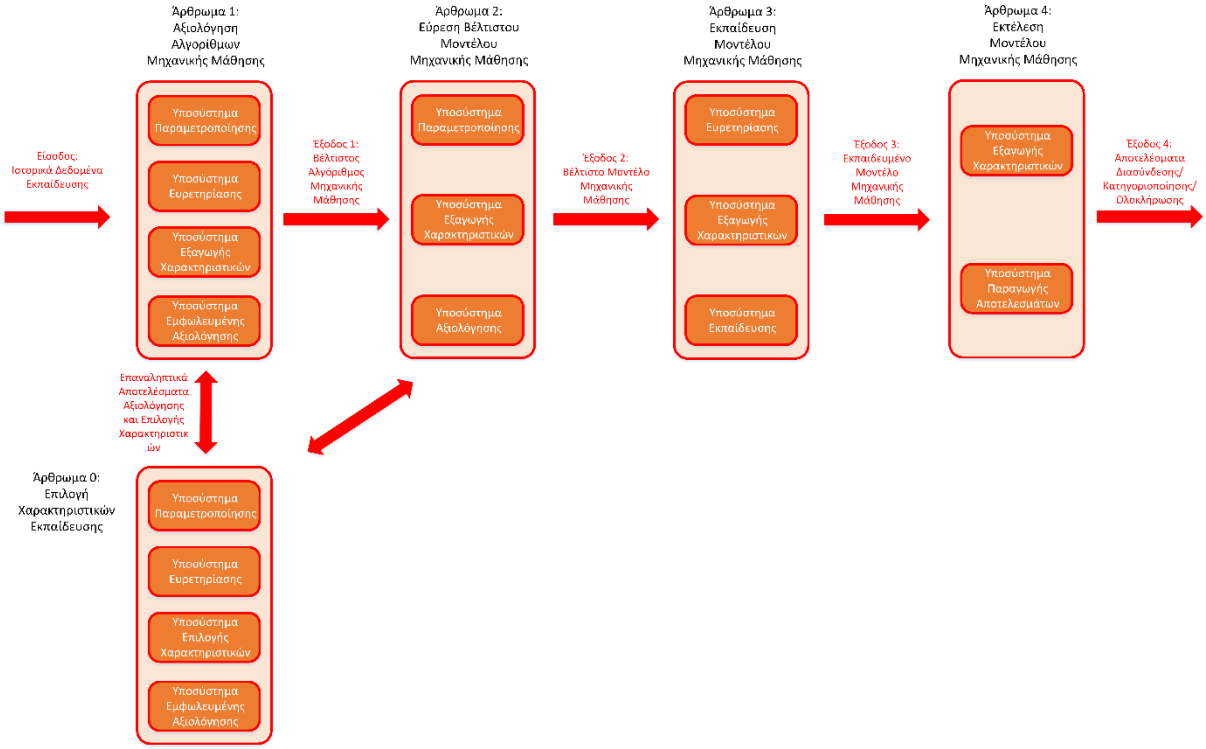
Κάθε ένα από τα αρθρώματα δύναται να κληθεί με ανεξάρτητη κλήση API, ανάλογα με την εκάστοτε ανάγκη εκτέλεσης του χρήστη. Για παράδειγμα, ο χρήστης δύναται να τρέξει μόνο μία φορά το βήμα-‘Άρθρωμα 1, και στη συνέχεια να τρέχει μόνο τα υπόλοιπα βήματα όταν προκύπτει ένα νέο σύνολο δεδομένων εκπαίδευσης.

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 Επιλογής Χαρακτηριστικών: Το υποσύστημα αυτό υποστηρίζει τη χρήση τεχνικών για την επιλογή του υποσυνόλου χαρακτηριστικών που επιτυγχάνουν τη μέγιστη ακρίβεια διασύνδεσης ενώ ταυτόχρονα μειώνουν τον όγκο των δεδομένων. Αυτό το σύστημα καλείται εμφωλευμένα στην αξιολόγηση των διαφορετικών αλγορίθμων και στην επιλογή των υπέρ-παραμέτρων τους.***
- *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 μέση 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.



Εικόνα 1: Ανανεωμένη Αρχιτεκτονική Υλοποιημένων Βιβλιοθηκών Μηχανικής Μάθησης

3. Βιβλιοθήκες μηχανικής μάθησης με επιλογή χαρακτηριστικών

Ακολούθως, παρουσιάζονται οι ανανεωμένες εκδόσεις των τριών από τις τέσσερις βιβλιοθήκες μηχανικής μάθησης που παρουσιάστηκαν στο Π2.1 του έργου: *βιβλιοθήκη διασύνδεσης τοπωνυμίων*, *βιβλιοθήκη κατηγοριοποίησης ΣΕ* και *βιβλιοθήκη γεωκωδικοποίησης*. Οι συγκεκριμένες βιβλιοθήκες επεκτάθηκαν με την προσθήκη παραμετροποιήσιμων μηχανισμών επιλογής χαρακτηριστικών εκπαίδευσης, και συγκεκριμένα με τις τέσσερις μεθόδους που περιγράφηκαν στο Κεφάλαιο 2.1. Σημειώνουμε ότι κρίθηκε περιττό η τέταρτη βιβλιοθήκη (ολοκλήρωσης γεωτεμαχίων) να επεκταθεί με μεθόδους επιλογής χαρακτηριστικών, αφού, μετά από βελτιστοποιήσεις που πραγματοποιήθηκαν[10], επιτυγχάνει σχεδόν βέλτιστα αποτελέσματα ακρίβειας.

Αρχικά, παρακάτω περιγράφουμε συνοπτικά το σύνολο των αλγορίθμων μηχανικής μάθησης που ενσωματώθηκαν και αξιολογήθηκαν στις τρεις βιβλιοθήκες που θα παρουσιαστούν στη συνέχεια:

- *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

[10] Οι συγκεκριμένες επεκτάσεις, καθώς και η αντίστοιχη πειραματική αξιολόγηση, τεκμηριώνονται στο Παραδοτέο 2.2.

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. Σημειώνουμε ότι οι επεκτάσεις που σχετίζονται με επιλογή χαρακτηριστικών σημειώνονται στη συνέχεια με έντονη και πλάγια γραφή.

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, επιλογή του βέλτιστου υποσυνόλου χαρακτηριστικών εκπαίδευσης 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/tree/feature_selection). Μπορεί να διανεμηθεί και να τροποποιηθεί σύμφωνα με τους όρους της μη περιοριστικής MIT άδειας[11].

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[12], καθώς και την xgboost[13] για την αποδοτική υλοποίηση του αλγορίθμου eXtreme Gradient Boosting. Όσον αφορά τις μετρικές ομοιότητας, γίνεται χρήση των βιβλιοθηκών jellyfish[14], για τις μετρικές

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

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

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

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

Jaro και Jaro-Winkler, και pyxdameraulevenshtein[15], για την μετρική Damerau-Levenshtein. Τέλος, η αποδοτική διαχείριση και ανάλυση στα σύνολα δεδομένων γίνεται με τις βιβλιοθήκες pandas[16], numpy[17] και scipy[18].

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 Χαρακτηριστικών: Το υποσύστημα αυτό υποστηρίζει τη διαδικασία επιλογής από το σύνολο των χαρακτηριστικών ένα υποσύνολο αυτών με την κατάλληλη στάθμιση τους, τα οποία επιτυγχάνουν το μέγιστο βαθμό διασύνδεσης ενώ ταυτόχρονα μειώνουν τον όγκο των δεδομένων προς επεξεργασία. Στα πλαίσια αυτού του υποσυστήματος χρησιμοποιούνται τεχνικές και αλγόριθμοι μηχανικής μάθησης τα οποία υπόκεινται σε διαδικασία συγκριτικής αξιολόγησης (cross-validation) για την επιλογή των βέλτιστων υπέρ-παραμέτρων τους και που***

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

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

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

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

δεδομένου ενός αλγορίθμου για την εσωτερική τους αξιολόγηση, εξάγουν το βέλτιστο υποσύνολο χαρακτηριστικών.

- 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 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 στάδια πειραμάτων που περιγράφηκαν παραπάνω, στη διαδικασία αυτή περιλαμβάνεται η φόρτωση των κατάλληλων συνόλων δεδομένων που απαιτούνται, η αποδοτική κατασκευή των χαρακτηριστικών εκπαίδευσης που έχουν επιλεγεί και η επιλογή του βέλτιστου υποσυνόλου χαρακτηριστικών εφόσον το έχει επιλέξει ο χρήστης. 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[19]. 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.

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

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
- alphabet-detector - 0.0.7
- docopt - 0.6.2
- text-unidecode - 1.2
- 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 3.6[20] στη γραμμή εντολών:

```
$ python
Python 3.6.9 (default, Nov  7 2019, 10:44:02)
[GCC 8.3.0] 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 python3` <path/to/new/virtualenv/>
$ source <path/to/new/virtualenv/>/bin/activate
```

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

[20] Έγινε μετάβαση από τη python 2 στη 3 διότι από 1η Ιανουαρίου του 2020 δεν υποστηρίζεται πια, επίσημα, με διορθώσεις σφαλμάτων, ενημερώσεις ή προσθήκες ασφαλείας (https://www.python.org/doc/sunset-python-2/).

```
$ 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*.
- ***feature_selection: λογική παράμετρος που υποδηλώνει κατά πόσο θα πραγματοποιηθεί επιλογή χαρακτηριστικών εκπαίδευσης ή όχι.***
- ***feature_selection_method: παράμετρος για την μέθοδο επιλογής χαρακτηριστικών που θα χρησιμοποιηθεί. Η παράμετρος αυτή έχει ισχύ όταν η feature_selection έχει τη τιμή True. Οι διαθέσιμες έγκυρες επιλογές είναι: VarianceThreshold, SelectKBest, RFE, SelectFromModel.***
- *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.
- ***{SelectKbest, VT, SFM}_hyperparameters: λεξικό που περιέχει αντιστοιχίες μεταξύ των ονομάτων των υπέρ-παραμέτρων και της συλλογής από ποσοστά που θέλουμε να διερευνήσουμε στα πλαίσια της συγκριτικής αξιολόγησης (grid search) για την εκπαίδευση των μοντέλων επιλογής χαρακτηριστικών.***

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.

Επιλογή αλγορίθμου και χαρακτηριστικών και βελτιστοποίηση αλγορίθμου

Το στάδιο αυτό υλοποιείται από τη συνάρτηση *fineTuneClassifier*, ενσωματώνοντας τα τρια πρώτα βήματα-αρθρώματα της ροής μηχανικής μάθησης. Παίρνει ως είσοδο:

1. *X:* Πίνακας με τα στοιχεία του συνόλου εκπαίδευσης αναπαριστώμενα από διανύσματα χαρακτηριστικών και έχουν προκύψει από διαμέριση (fold).
2. *y:* Διάνυσμα με τις κλάσεις/ετικέτες (classes/labels) που αντιστοιχούν στα παραπάνω στοιχεία του συνόλου εκπαίδευσης, τις οποίες προσπαθεί να προβλέψει ο αλγόριθμος μηχανικής μάθησης και έχουν προκύψει από διαμέριση (fold).

Η συνάρτηση *fineTuneClassifier* επιστρέφει τα εξής ορίσματα:

1. Ένα μοντέλο του επιλεγμένου αλγορίθμου *best_clf* με τις βέλτιστες υπερ-παραμέτρους που έχουν βρεθεί για το συγκεκριμένο αλγόριθμο.
2. Τον πίνακα με τα στοιχεία του συνόλου εκπαίδευσης αναπαριστώμενα από χαρακτηριστικά εκπαίδευσης X_train που έχουν υποστεί, ανάλογα με το αν το έχει

επιλέξει ο χρήστης, μετασχηματισμό, περιέχοντας πλέον μόνο το βέλτιστο υποσύνολο των χαρακτηριστικών κατόπιν της διαδικασίας επιλογής χαρακτηριστικών.

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:* Πίνακας με τα στοιχεία του συνόλου εκπαίδευσης αναπαριστώμενα από διανύσματα χαρακτηριστικών, χωρίς διαμέριση (fold).
2. *Y_train:* Διάνυσμα με τις κλάσεις/ετικέτες που αντιστοιχούν στα παραπάνω στοιχεία του συνόλου εκπαίδευσης, 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:* Πίνακας με τα στοιχεία του νέου συνόλου δεδομένων αναπαριστώμενα από διανύσματα χαρακτηριστικών, χωρίς διαμέριση (fold).
2. *(Προαιρετικά[21]) Y_train:* Διάνυσμα με τις κλάσεις/ετικέτες που αντιστοιχούν στα παραπάνω στοιχεία του συνόλου εκπαίδευσης, 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

[21] Το συγκεκριμένο στάδιο εκτελείται φυσικά και χωρίς της ύπαρξη ετικετών, μιας και αφορά την εκτέλεση του μοντέλου σε νέα δεδομένα, των οποίων τις ετικέτες/κλάσσεις/κατηγορίες θέλουμε να προβλέψουμε. Το διάνυσμα των ετικετών *Y_train* δίνεται ως είσοδος στην περίπτωση που θέλουμε να αξιολογήσουμε το μοντέλο σε ένα νέο σύνολο δεδομένων.

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/tree/feature_selection). Μπορεί να διανεμηθεί και να τροποποιηθεί σύμφωνα με τους όρους της μη περιοριστικής MIT άδειας**Error! Bookmark not defined.**.

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
- Adaboost
- 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**Error! Bookmark not defined.**. Οι μέθοδοι επεξεργασίας γεωχωρικών δεδομένων που χρησιμοποιούνται καλύπτονται από μια συλλογή σχετικών βιβλιοθηκών της γλώσσας python (shapely[22], geopandas[23]), ενώ η επεξεργασία κειμενικών δεδομένων καλύπτεται από το python εργαλείο whoosh[24] .

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

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. Ο χρήστης μπορεί να συγκεκριμενοποιήσει ποια αρχεία θα λειτουργήσουν ως είσοδοι καθενός από τα στάδια αλλά και να επιλέξει τον φάκελο του τρέχοντος πειράματος.
- 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 τον αριθμό των προβλέψεων που προβλέπει το μοντέλο για το κάθε ΣΕ.
- Υποσύστημα Εξαγωγής Χαρακτηριστικών: Το υποσύστημα αυτό αυτό αφορά το πρώτο στάδιο των πειραμάτων και 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 χρησιμοποιούν ανεστραμμένα ευρετήρια, υποστηριζόμενα μέσω του python εργαλείου whoosh.
- ***Υποσύστημα Επιλογής Χαρακτηριστικών: Το υποσύστημα αυτό υποστηρίζει τη διαδικασία επιλογής από το σύνολο των κειμενικών και γεωχωρικών χαρακτηριστικών, ένα υποσύνολο αυτών με την κατάλληλη στάθμιση τους, τα οποία επιτυγχάνουν τη μέγιστη ακρίβεια κατάταξης ΣΕ ενώ ταυτόχρονα μειώνουν τον όγκο των δεδομένων προς επεξεργασία. Στα πλαίσια αυτού του***

[23] http://geopandas.org/

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

υποσυστήματος χρησιμοποιούνται τεχνικές και αλγόριθμοι μηχανικής μάθησης τα οποία υπόκεινται σε διαδικασία συγκριτικής αξιολόγησης (cross-validation) για την επιλογή των βέλτιστων υπέρ-παραμέτρων τους και που δεδομένου ενός αλγορίθμου για την εσωτερική τους αξιολόγηση, εξάγουν το βέλτιστο υποσύνολο χαρακτηριστικών.

- 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 Επιλογής Βέλτιστου Μοντέλου: 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 Εκπαίδευσης 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
- shapely
- whoosh

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 3.6[25] στη γραμμή εντολών:

```
$ python
Python 3.6.9 (default, Nov 7 2019, 10:44:02)
[GCC 8.3.0] 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-CLassification, το οποίο θα είναι απομονωμένο από το υπόλοιπο λειτουργικό σύστημα. Αυτό γίνεται ως εξής:

```
$ virtualenv -p `which python3` <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
```

Τέλος, κατεβάζουμε την τελευταία έκδοση του πηγαίου κώδικα της LGM-Classification βιβλιοθήκης:

[25] Έγινε μετάβαση από τη python 2 στη 3 διότι από 1η Ιανουαρίου του 2020 δεν υποστηρίζεται πια, επίσημα, με διορθώσεις σφαλμάτων, ενημερώσεις ή προσθήκες ασφαλείας (https://www.python.org/doc/sunset-python-2/).

```
$ git clone https://github.com/LinkGeoML/LGM-Classification.git
$ cd LGM-Classification
```

3.2.5.2. Παραμετροποίηση

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

- poi_fpath: Το μονοπάτι στο δίσκο του csv αρχείου που περιέχει τα δεδομένα εκπαίδευσης.
- experiments_path: Το μονοπάτι στο δίσκο του φακέλου στον οποίο θέλουμε να αποθηκεύονται τα αποτελέσματα των πειραμάτων.
- supported_adjacency_features: Μια λίστα από συμβολοσειρές οι οποίες αντιστοιχούν στους τίτλους των χαρακτηριστικών γειτνίασης που υποστηρίζονται από τη βιβλιοθήκη.
- supported_textual_features: Μια λίστα από συμβολοσειρές οι οποίες αντιστοιχούν στους τίτλους των κειμενικών χαρακτηριστικών που υποστηρίζονται από τη βιβλιοθήκη.
- included_adjacency_features: Μια λίστα από συμβολοσειρές οι οποίες αντιστοιχούν στους τίτλους των χαρακτηριστικών γειτνίασης που θα χρησιμοποιούν κατά τη φάση εκπαίδευσης των αλγορίθμων.
- included_textual_features: Μια λίστα από συμβολοσειρές οι οποίες αντιστοιχούν στους τίτλους των κειμενικών χαρακτηριστικών που θα χρησιμοποιούν κατά τη φάση εκπαίδευσης των αλγορίθμων.
- normalized_features: Μια λίστα από συμβολοσειρές οι οποίες αντιστοιχούν στους τίτλους των χαρακτηριστικών των οποίων οι τιμές πρόκειται να κανονικοποιηθούν.
- classes_in_radius_thr: Μια λίστα παραμέτρων αποτελούμενη από αριθμούς σε μέτρα που αντιστοιχούν στην ακτίνα εντός της οποίας ΣΕ θα θεωρούνται γειτονικά του υπό εξέταση ΣΕ.
- classes_in_street_and_radius_thr: Μια λίστα παραμέτρων αποτελούμενη από αριθμούς σε μέτρα που αντιστοιχούν στην ακτίνα εντός της οποίας ΣΕ θα θεωρούνται γειτονικά του υπό εξέταση ΣΕ. Οι πιθανοί γείτονες εδώ αναφέρονται στα ΣΕ που βρίσκονται και στον ίδιο δρόμο από το υπό εξέταση ΣΕ.
- classes_in_neighbors_thr: Μια λίστα παραμέτρων αποτελούμενη από αριθμούς που αντιστοιχούν στο πλήθος των κοντινότερων γειτόνων που θα ληφθούν υπόψη για το υπό εξέταση ΣΕ.
- classes_in_street_radius_thr: Μια λίστα παραμέτρων αποτελούμενη από αριθμούς σε μέτρα που αντιστοιχούν στην ακτίνα από το δρόμο στον οποίο ανήκει το υπό εξέταση ΣΕ, εντός της οποίας ΣΕ θα θεωρούνται γειτονικά αυτού.
- top_k_terms_pct: Μια λίστα παραμέτρων αποτελούμενη από ποσοστά που αντιστοιχούν στην ποσόστοση των πιο συχνών όρων που θα ληφθεί υπόψη.
- top_k_trigrams_pct: Μια λίστα παραμέτρων αποτελούμενη από ποσοστά που αντιστοιχούν στην ποσόστοση των πιο συχνών τριγραμμάτων που θα ληφθεί υπόψη.

- top_k_fourgrams_pct: Μια λίστα παραμέτρων αποτελούμενη από ποσοστά που αντιστοιχούν στην ποσόστοση των πιο συχνών τετραγραμμάτων που θα ληφθεί υπόψη.
- n_folds: Η παράμετρος που καθορίζει τον αριθμό των διαχωρισμών στο πλαίσιο της διαδικασίας k-fold cross-validation.
- supported_classifiers: Μια λίστα από συμβολοσειρές οι οποίες αντιστοιχούν στα ονόματα των αλγορίθμων κατάταξης που υποστηρίζονται από τη βιβλιοθήκη.
- included_classifiers: Μια λίστα από συμβολοσειρές οι οποίες αντιστοιχούν στα ονόματα των αλγορίθμων κατάταξης που θα χρησιμοποιηθούν στο πείραμα.
- NaiveBayes_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 αλγορίθμου NaiveBayes.
- GaussianProcess_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 αλγορίθμου GaussianProcess.
- AdaBoost_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 αλγορίθμου AdaBoost.
- 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 αλγορίθμου kNN.
- LogisticRegression_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 αλγορίθμου LogisticRegression.
- SVM_hyperparameters: Λίστα από λεξικά που περιέχουν αντιστοιχίες μεταξύ των ονομάτων των υπερ-παραμέτρων και της συλλογής από τιμές που θέλουμε να διερευνήσουμε κατά τη διάρκεια εκπαίδευσης του αλγορίθμου SVM.
- 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 αλγορίθμου MLP.
- 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 αλγορίθμου 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 αλγορίθμων RandomForest / Extra Trees.
- top_k: Λίστα που περιέχει τους αριθμούς των πιο πιθανών κατηγοριών ανά πρόβλεψη παραδείγματος που θέλουμε να ληφθούν υπόψιν κατά την αξιολόγηση των αποτελεσμάτων.
- k_preds: Αριθμός που αντιστοιχεί στο πλήθος των πιο πιθανών κατηγοριών ανά πρόβλεψη παραδείγματος που θέλουμε να ληφθούν υπόψιν κατά την εξαγωγή αποτελεσμάτων.
- osm_crs: Αριθμός που αντιστοιχεί στον κωδικό του συστήματος συντεταγμένων που χρησιμοποιεί η πλατφόρμα OSM.

- id_col: Συμβολοσειρά που αναφέρεται στο όνομα της στήλης του αρχείου εισόδου, η οποία περιέχει το μοναδικό ID για το κάθε ΣΕ.
- name_col: Συμβολοσειρά που αναφέρεται στο όνομα της στήλης του αρχείου εισόδου, η οποία περιέχει το όνομα για το κάθε ΣΕ.
- label_col: Συμβολοσειρά που αναφέρεται στο όνομα της στήλης του αρχείου εισόδου, η οποία περιέχει την ετικέτα/κατηγορία για το κάθε ΣΕ.
- lon_col: Συμβολοσειρά που αναφέρεται στο όνομα της στήλης του αρχείου εισόδου, η οποία περιέχει την τετμημένη για το κάθε ΣΕ.
- lat_col: Συμβολοσειρά που αναφέρεται στο όνομα της στήλης του αρχείου εισόδου, η οποία περιέχει την τεταγμένη για το κάθε ΣΕ.
- poi_crs: Αριθμός που αντιστοιχεί στον κωδικό του συστήματος συντεταγμένων που χρησιμοποιεί το αρχείο εισόδου.
- ***SelectKbest_hyperparameters: Λεξικό που περιέχει τα ποσοστά του συνόλου των χαρακτηριστικών εκπαίδευσης που θέλουμε να διερευνήσουμε για να λάβουμε το βέλτιστο k αριθμό χαρακτηριστικών που θα περιέχει το υποσύνολο στο πέρας της διαδικασίας της επιλογής χαρακτηριστικών μέσω του αλγόριθμου SelectKbest.***
- ***VT_hyperparameters: Λεξικό που περιέχει τα ποσοστά που θέλουμε να διερευνήσουμε για να λάβουμε το βέλτιστο κατώφλι διακύμανσης(threshold) κάτω του οποίου τα χαρακτηριστικά δε θα περιλαμβάνονται στο υποσύνολο χαρακτηριστικών που θα προκύψει κατόπιν της διαδικασίας επιλογής τους μέσω του αλγορίθμου Variance Threshold.***
- ***SelectFromModel_hyperparameters: Λεξικό που περιέχει τα ποσοστά που θέλουμε να διευρευνήσουμε για να λάβουμε το βέλτιστο κατώφλι σημαντικότητας (importance) κάτω του οποίου τα χαρακτηριστικά δε θα περιλαμβάνονται στο υποσύνολο χαρακτηριστικών που θα προκύψει κατόπιν της διαδικασίας επιλογής τους μέσω του αλγορίθμου SelectFromModel.***

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
- το όνομα της κατηγορίας στην οποία ανήκει
- την τετμημένη του
- την τεταγμένη του

Το σύνολο των λειτουργιών που καλύπτονται από την βιβλιοθήκη μπορεί να χωριστεί σε 5 ξεχωριστά στάδια, η εκτέλεση των οποίων περιγράφεται ακολούθως:

Εξαγωγή χαρακτηριστικών εκπαίδευσης

Το βήμα αυτό αποτελεί το πρώτο βήμα κάθε πειράματος. Αρχικοποιεί ένα φάκελο, στον οποίο θα αποθηκεύονται τα αποτελέσματα όλων των βημάτων και αναλαμβάνει να δημιουργήσει τα χαρακτηριστικά εκπαίδευσης σε μορφή κατάλληλη για την απρόσκοπτη

λειτουργία της ροής μηχανικής μάθησης, όπως αυτή περιγράφεται στο αρχείο διαμόρφωσης config.py. Η εκτέλεση αυτού του βήματος γίνεται με την ακόλουθη εντολή:

```
$ python features_extraction.py
```

Αξιολόγηση/επιλογή αλγορίθμου

Η εκτέλεση αυτού του βήματος γίνεται με την ακόλουθη εντολή:

```
python algorithm_selection.py -experiment_path
<μονοπάτι/στον/φάκελο/του/τρέχοντος/πειράματος>
```

Η εκτέλεση αυτού του βήματος παράγει τέσσερα αρχεία:

- Ένα αρχείο στο οποίο παρουσιάζεται ο χώρος αναζήτησης που καλύπτεται. Συγκεκριμένα καταγράφεται το σύνολο των αλγορίθμων κατάταξης που περιλαμβάνονται στο πείραμα, μαζί με το χώρο των αντίστοιχων παραμέτρων που αξιολογείται για την εύρεση του καλύτερου μοντέλου.
- Τρία αρχεία στα οποία παρουσιάζονται (α) τα πλήρη αποτελέσματα του βήματος, (β) τα αποτελέσματα ομαδοποιημένα ανά fold κατά τη διαδικασία της μεθόδου cross-validation και (γ) τα αποτελέσματα ομαδοποιημένα ανά αλγόριθμο κατάταξης, έτσι ώστε να μπορεί να γίνει μια αξιολόγηση και σύγκριση μεταξύ των διάφορων αλγορίθμων.

Αξιολόγηση/επιλογή μοντέλου

Η εκτέλεση αυτού του βήματος γίνεται με την ακόλουθη εντολή:

```
python model_selection.py -classifier <όνομα του αλγορίθμου που πρόκειται
να βελτιστοποιηθεί> -experiment_path
<μονοπάτι/στον/φάκελο/του/τρέχοντος/πειράματος>
```

Η εκτέλεση αυτού του βήματος παράγει τα εξής τρία αρχεία:

- Ένα αρχείο στο οποίο παρουσιάζεται ο χώρος που εξετάζεται και περιλαμβάνει τον επιλεγμένο αλγόριθμο μαζί με το χώρο των αντίστοιχων υπερπαραμέτρων.
- Δύο αρχεία που παρουσιάζουν τα αποτελέσματα του βήματος (α) με εξαντλητικό τρόπο και (β) ομαδοποιημένα ανά fold και υπερπαραμετροποίηση.

Εκπαίδευση/εξαγωγή βέλτιστου μοντέλου

Η εκτέλεση αυτού του βήματος γίνεται με την ακόλουθη εντολή:

```
python model_training.py -experiment_path
<μονοπάτι/στον/φάκελο/του/τρέχοντος/πειράματος>
```

Το βήμα αυτό εκμεταλλεύεται την πληροφορία των προηγούμενων βημάτων ώστε να εκπαιδεύσει το πιο αποτελεσματικό μοντέλο (πιο ακριβής αλγόριθμος και καλύτερη παραμετροποίηση αυτού) πάνω στο σύνολο των παραδειγμάτων εκπαίδευσης. Στη συνέχεια το εκπαιδευμένο αυτό μοντέλο αποθηκεύεται στο δίσκο, σε μορφή κατάλληλη (αρχείο pickle) ώστε να μπορεί να χρησιμοποιηθεί από το επόμενο βήμα.

Εξαγωγή προβλέψεων σε νέα σύνολα δεδομένων

Η εκτέλεση αυτού του βήματος γίνεται με την ακόλουθη εντολή:

```
python model_deployment.py -experiment_path
<μονοπάτι/στον/φάκελο/του/τρέχοντος/πειράματος> -poi_fpath
<μονοπάτι/στο/αρχείο/που/περιέχει/τα/νέα/δεδομένα/προς/κατάταξη>
```

Το βήμα αυτό δημιουργεί ένα αρχείο csv, στο οποίο παρουσιάζονται οι προβλέψεις του μοντέλου σε ζεύγη της μορφής *κατηγορία – ποσοστό αξιοπιστίας* για κάθε ένα από τα δεδομένα προς κατάταξη.

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/tree/feature_selection). Μπορεί να διανεμηθεί και να τροποποιηθεί σύμφωνα με τους όρους της μη περιοριστικής MIT άδειας**Error! Bookmark not defined.**.

3.3.2. Αλγόριθμοι μηχανικής μάθησης

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

- Naive Bayes
- Nearest Neighbors
- Logistic Regression
- Support Vector Machines
- Multi-layer Perceptron
- Decision Tree
- Random Forest
- 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**Error! Bookmark not defined.**. Οι μέθοδοι επεξεργασίας γεωχωρικών δεδομένων που χρησιμοποιούνται καλύπτονται από μια συλλογή σχετικών βιβλιοθηκών της γλώσσας python (shapely**Error! Bookmark not defined.**, geopandas**Error! Bookmark not defined.**).

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 το χρήστη προκειμένου να καθοριστεί ο τρόπος εκτέλεσης καθενός από τα στάδια της βιβλιοθήκης. Ο χρήστης μπορεί να συγκεκριμενοποιήσει ποια αρχεία θα λειτουργήσουν ως είσοδοι καθενός από τα στάδια αλλά και να επιλέξει το φάκελο του τρέχοντος πειράματος.
- 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 και τη βελτιστοποίηση του αλγορίθμου κατάταξης.
- 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 αφορά το πρώτο στάδιο των πειραμάτων και 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 Χαρακτηριστικών: Το υποσύστημα αυτό υποστηρίζει τη διαδικασία επιλογής από το σύνολο των γεωχωρικών χαρακτηριστικών, ένα υποσύνολο αυτών με την κατάλληλη στάθμιση τους, τα οποία επιτυγχάνουν τη μέγιστη ακρίβεια κατάταξης των ζευγών ενώ ταυτόχρονα μειώνουν τον όγκο των δεδομένων προς επεξεργασία. Στα πλαίσια αυτού του υποσυστήματος χρησιμοποιούνται τεχνικές και αλγόριθμοι μηχανικής μάθησης τα οποία υπόκεινται σε διαδικασία συγκριτικής αξιολόγησης (cross-validation) για την επιλογή των βέλτιστων υπέρ-παραμέτρων τους και που δεδομένου ενός αλγορίθμου για την εσωτερική τους αξιολόγηση, εξάγουν το βέλτιστο υποσύνολο χαρακτηριστικών.***
- Υποσύστημα Επιλογής Αλγορίθμου Κατάταξης: Το υποσύστημα αυτό αφορά το δεύτερο 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 Επιλογής Βέλτιστου Μοντέλου: Το υποσύστημα αυτό αφορά το τρίτο στάδιο των πειραμάτων και περιλαμβάνει αναζήτηση του χώρου των υπερ-

παραμέτρων του βέλτιστου αλγορίθμου που έχει προκύψει από την εκτέλεση του δεύτερου σταδίου. Αυτό επιτυγχάνεται με τη χρήση της μεθόδου cross-validation, από την οποία προκύπτει ο καλύτερος συνδυασμός υπερ-παραμέτρων.

- Υποσύστημα Εκπαίδευσης Μοντέλου Κατάταξης: Το υποσύστημα αυτό αφορά το τέταρτο 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 πέμπτο 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 και .pkl και 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
- shapely

Οδηγίες εγκατάστασης

Αρχικά ελέγχουμε εάν είναι εγκατεστημένη η Python 3.6[26] στη γραμμή εντολών:

```
$ python
Python 3.6.9 (default, Nov  7 2019, 10:44:02)
[GCC 8.3.0] 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-Geocoding, το οποίο θα είναι απομονωμένο από το υπόλοιπο λειτουργικό σύστημα. Αυτό γίνεται ως εξής:

```
$ virtualenv -p `which python3` <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
```

Τέλος, κατεβάζουμε την τελευταία έκδοση του πηγαίου κώδικα της LGM-Geocoding βιβλιοθήκης:

```
$ git clone https://github.com/LinkGeoML/LGM-Geocoding.git
$ cd LGM-Geocoding
```

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 config κλάσης. Τα πεδία αυτά πρέπει να είναι τα ακόλουθα:

- n_folds: Η παράμετρος που καθορίζει τον αριθμό των διαχωρισμών στο πλαίσιο της διαδικασίας k-fold cross-validation.
- source_crs: Αριθμός που αντιστοιχεί στον κωδικό του συστήματος συντεταγμένων που χρησιμοποιεί το αρχείο δεδομένων που χρησιμοποιείται στην είσοδο του πειράματος.

[26] Έγινε μετάβαση από τη python 2 στη 3 διότι από 1η Ιανουαρίου του 2020 δεν υποστηρίζεται πια, επίσημα, με διορθώσεις σφαλμάτων, ενημερώσεις ή προσθήκες ασφαλείας (https://www.python.org/doc/sunset-python-2/).

- target_crs: Αριθμός που αντιστοιχεί στον κωδικό του συστήματος συντεταγμένων στο οποίο η βιβλιοθήκη μετατρέπει τις συντεταγμένες, έτσι ώστε οι αποστάσεις που υπολογίζονται να είναι σε μέτρα.
- clusters_pct: Ποσοστό των στοιχείων εκπαίδευσης, που επηρεάζει τον αριθμό των clusters που δημιουργούνται κατά τη διαδικασία εξαγωγής του δικτύου δρόμων από το Overpass API.
- osm_buffer: Απόσταση σε μέτρα που αναφέρεται στην απόσταση γύρω από ένα bounding box κατά τη διαδικασία εξαγωγής του δικτύου δρόμων από το Overpass API.
- osm_timeout: Χρόνος αναμονής ανά συγκεκριμένο αριθμό κλήσεων κατά τη διαδικασία εξαγωγής του δικτύου δρόμων από το Overpass API.
- distance_thr: Ανώτατη τιμή απόστασης (σε μέτρα). Τιμές κατά τη δημιουργία των χαρακτηριστικών εκπαίδευσης που είναι μεγαλύτερες από αυτή, θα μετατρέπονται σε αυτή.
- baseline_service: Συμβολοσειρά που αναφέρεται στην τρέχουσα πηγή γεωκωδικοποίησης των δεδομένων και χρησιμοποιείται για τον υπολογισμό της τρέχουσας ακρίβειας, χωρίς τη χρήση της βιβλιοθήκης.
- experiments_path: Μονοπάτι στο δίσκο του φακέλου στον οποίο θα αποθηκεύονται τα αποτελέσματα των πειραμάτων.
- services: Λίστα από συμβολοσειρές που αναφέρονται στα ονόματα των πηγών γεωκωδικοποίησης που θα χρησιμοποιηθουν στο πείραμα.
- supported_features: Λίστα από συμβολοσειρές οι οποίες αντιστοιχούν στους τίτλους των χαρακτηριστικών που υποστηρίζονται από τη βιβλιοθήκη.
- included_features: Λίστα από συμβολοσειρές οι οποίες αντιστοιχούν στους τίτλους των χαρακτηριστικών που θα χρησιμοποιηθούν στο πείραμα.
- normalized_features: Λίστα από συμβολοσειρές οι οποίες αντιστοιχούν στους τίτλους των χαρακτηριστικών των οποίων οι τιμές θα κανονικοποιηθούν.
- supported_classifiers: Λίστα από συμβολοσειρές οι οποίες αντιστοιχούν στα ονόματα των αλγορίθμων κατάταξης που υποστηρίζονται από τη βιβλιοθήκη.
- included_classifiers: Λίστα από συμβολοσειρές οι οποίες αντιστοιχούν στα ονόματα των αλγορίθμων κατάταξης που θα χρησιμοποιηθούν στο πείραμα.
- NB_hparams: 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 Naive Bayes.
- NN_hparams: 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 Nearest Neighbors.
- LR_hparams: 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 Logistic Regression.
- SVM_hparams: Λίστα από λεξικά που περιέχουν αντιστοιχίες μεταξύ των ονομάτων των υπερ-παραμέτρων και της συλλογής από τιμές που θέλουμε να διερευνήσουμε κατά τη διάρκεια εκπαίδευσης του μοντέλου Support Vector Machines.
- MLP_hparams: 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 Multi-Layer Perceptron.

- DT_hparams: 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 Tree.
- RF_hparams: 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 Random Forest.
- ET_hparams: 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 Extra Trees.
- ***SelectKbest_hyperparameters: Λεξικό που περιέχει τα ποσοστά του συνόλου των χαρακτηριστικών εκπαίδευσης που θέλουμε να διερευνήσουμε για να λάβουμε το βέλτιστο k αριθμό χαρακτηριστικών που θα περιέχει το υποσύνολο στο πέρας της διαδικασίας της επιλογής χαρακτηριστικών μέσω του αλγόριθμου SelectKbest.***
- ***VT_hyperparameters: Λεξικό που περιέχει τα ποσοστά που θέλουμε να διερευνήσουμε για να λάβουμε το βέλτιστο κατώφλι διακύμανσης(threshold) κάτω του οποίου τα χαρακτηριστικά δε θα περιλαμβάνονται στο υποσύνολο χαρακτηριστικών που θα προκύψει κατόπιν της διαδικασίας επιλογής τους μέσω του αλγορίθμου Variance Threshold.***
- ***SelectFromModel_hyperparameters: Λεξικό που περιέχει τα ποσοστά που θέλουμε να διευρευνήσουμε για να λάβουμε το βέλτιστο κατώφλι σημαντικότητας (importance) κάτω του οποίου τα χαρακτηριστικά δε θα περιλαμβάνονται στο υποσύνολο χαρακτηριστικών που θα προκύψει κατόπιν της διαδικασίας επιλογής τους μέσω του αλγορίθμου SelectFromModel.***

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 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) και την ετικέτα που δηλώνει την προτιμότερη πηγή γεωκωδικοποίησης (κολώνα label), 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 5 ξεχωριστά στάδια η εκτέλεση των οποίων περιγράφεται ακολούθως:

Εξαγωγή χαρακτηριστικών εκπαίδευσης

Το βήμα αυτό αποτελεί το πρώτο βήμα κάθε πειράματος. Αρχικοποιεί ένα φάκελο, στον οποίο θα αποθηκεύονται τα αποτελέσματα όλων των βημάτων και αναλαμβάνει να δημιουργήσει τα χαρακτηριστικά εκπαίδευσης που αντιστοιχούν στα στοιχεία του δοθέντος αρχείου, σε μορφή κατάλληλη για την απρόσκοπτη λειτουργία της ροής μηχανικής μάθησης, όπως αυτή περιγράφεται στο αρχείο διαμόρφωσης config.py. Η εκτέλεση αυτου του βήματος γίνεται με την ακόλουθη εντολή:

```
python features_extraction.py -fpath
<μονοπάτι/στο/αρχείο/που/περιέχει/τα/στοιχεία/εκπαίδευσης>
```

Αξιολόγηση/επιλογή αλγορίθμου

Η εκτέλεση αυτού του βήματος γίνεται με την ακόλουθη εντολή:

```
python algorithm_selection.py -experiment_path
<μονοπάτι/στο/φάκελο/του/τρέχοντος/πειράματος>
```

Η εκτέλεση αυτού του βήματος παράγει δύο αρχεία:

- Ένα αρχείο στο οποίο παρουσιάζονται τα πλήρη αποτελέσματα του βήματος.
- Ένα αρχείο στο οποίο παρουσιάζονται τα αποτελέσματα του βήματος ομαδοποιημένα με βάση τον αλγόριθμο, έτσι ώστε να είναι δυνατή η σύγκριση μεταξύ τους.

Αξιολόγηση/επιλογή μοντέλου

Η εκτέλεση αυτού του βήματος γίνεται με την ακόλουθη εντολή:

```
python model_selection.py -classifier <όνομα του αλγορίθμου που πρόκειται
να βελτιστοποιηθεί> -experiment_path
<μονοπάτι/στον/φάκελο/του/τρέχοντος/πειράματος>
```

Η εκτέλεση αυτού του βήματος παράγει τα εξής τρία αρχεία:

- Ένα αρχείο στο οποίο παρουσιάζεται ο χώρος που εξετάζεται και περιλαμβάνει τον επιλεγμένο αλγόριθμο μαζί με το χώρο των αντίστοιχων υπερπραμέτρων.
- Δύο αρχεία που παρουσιάζουν τα αποτελέσματα του βήματος (α) με εξαντλητικό τρόπο και (β) ομαδοποιημένα κατά υπερπαραμετροποίηση.

Εκπαίδευση/εξαγωγή βέλτιστου μοντέλου

Η εκτέλεση αυτού του βήματος γίνεται με την ακόλουθη εντολή:

```
python model_training.py -experiment_path
<μονοπάτι/στον/φάκελο/του/τρέχοντος/πειράματος>
```

Το βήμα αυτό εκμεταλλεύεται την πληροφορία των προηγούμενων βημάτων ώστε να εκπαιδεύσει το πιο αποτελεσματικό μοντέλο (πιο ακριβής αλγόριθμος και καλύτερη παραμετροποίηση αυτού) πάνω στο σύνολο των παραδειγμάτων εκπαίδευσης. Στη συνέχεια, το εκπαιδευμένο αυτό μοντέλο αποθηκεύεται στο δίσκο σε κατάλληλη μορφή (αρχείο pickle) ώστε να μπορεί να χρησιμοποιηθεί από το επόμενο βήμα.

Εξαγωγή προβλέψεων σε νέα σύνολα δεδομένων

Η εκτέλεση αυτού του βήματος γίνεται με την ακόλουθη εντολή:

```
python model_deployment.py -experiment_path
<μονοπάτι/στον/φάκελο/του/τρέχοντος/πειράματος> -fpath
<μονοπάτι/στο/αρχείο/που/περιέχει/τα/νέα/δεδομένα/προς/κατάταξη>
```

Το βήμα αυτό δημιουργεί ένα αρχείο csv, στο οποίο παρουσιάζονται οι προβλέψεις του μοντέλου σε ζεύγη της μορφής *κατηγορία – ποσοστό αξιοπιστίας* για κάθε ένα από τα δεδομένα προς κατάταξη.

4. Πειραματική αξιολόγηση

4.1. Βιβλιοθήκη Διασύνδεσης Τοπωνυμίων

4.1.1. Σύνολο αξιολόγησης

Το αρχικό σύνολο δεδομένων για την εκπαίδευση και αξιολόγηση των αλγορίθμων που αναπτύχθηκαν στο πλαίσιο της διασύνδεσης τοπωνυμίων έχει προκύψει από τη βάση τοπωνυμίων Geonames[27], 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.

Το παραπάνω σύνολο χρησιμοποιήθηκε για την εκτενή αξιολόγηση της βιβλιοθήκης διασύνδεσης τοπωνυμίων, στο Π2.1. Προκειμένου να συγκρίνουμε εκείνη, την αρχική έκδοση της βιβλιοθήκης, με την ανανεωμένη έκδοση που ενσωματώνει μεθόδους επιλογής χαρακτηριστικών, απομονώνουμε ένα υποσύνολο του αρχικού συνόλου, όπως περιγράφεται ακολούθως.

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

Από τα 100.000 ζεύγη τοπωνυμίων της κατηγορίας global, όπου 50.000 τέτοια ζεύγη είναι επισημειωμένα ως True, δηλαδή περιγράφουν το ίδιο τοπωνύμιο, και 50.000 ως False, δηλαδή αντιστοιχούν σε διαφορετικό τοπωνύμιο επιλέξαμε τυχαία ένα υποσύνολο 10.000 ζευγών της κατηγορίας True και αντίστοιχα 10.000 της κατηγορίας False. Επιπλέον, επιλέξαμε και πάλι τυχαία από τα 5 εκατομμύρια ζεύγη τοπωνυμίων, ένα υποσύνολο αξιολόγησης που αποτελείται από 10.000 ζεύγη επισημειωμένα ως True και αντίστοιχα 10.000 επισημειωμένα ως False.

	$\text{Train}_{\text{global}}$	Test
Συναφή (True)	10.000	10.000
Μη Συναφή (False)	10.000	10.000
Συνολικά	20.000	20.000

Πίνακας 1: Σύνολα δεδομένων συγκριτικής αξιολόγησης της Διασύνδεσης Τοπωνυμίων με και χωρίς επιλογή χαρακτηριστικών

4.1.2. Συνθήκες αξιολόγησης

Η αξιολόγηση των μεθόδων επιλογής χαρακτηριστικών πραγματοποιήθηκε συγκρίνοντας τις τιμές ακρίβειας που επιτυγχάνουν οι βιβλιοθήκες διασύνδεσης με και χωρίς την χρήση των μηχανισμών επιλογής χαρακτηριστικών. Για την αξιολόγηση της απόδοσης των μεθοδολογιών επιλογής χαρακτηριστικών χρησιμοποιήθηκαν οι ίδιες μετρικές από το χώρο της ανάκτησης πληροφορίας και της μηχανικής μάθησης οι οποίες αναφέρθηκαν και στη προηγούμενη πειραματική αξιολόγηση:

- *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 *ακρίβειας* και *ανάκλησης*.

4.1.3. Αποτελέσματα

Στο παρακάτω πίνακα παρουσιάζονται τα αποτελέσματα των μεθόδων επιλογής χαρακτηριστικών, σε σύγκριση με την εφαρμογή του μοντέλου χωρίς επιλογή χαρακτηριστικών (**None**). Με έντονη γραμματοσειρά παρουσιάζονται τα καλύτερα αποτελέσματα. Σημειώνουμε επίσης ότι η αξιολόγηση επικεντρώνεται στον καλύτερο από τους αλγόριθμους μηχανικής μάθησης που δοκιμάστηκαν (βλέπε Κεφάλαιο 3.1.2), δηλαδή τον Random Forest. Με έντονη γραμματοσειρά παρουσιάζονται τα καλύτερα αποτελέσματα.

Μέθοδος επιλογής Χαρακτηριστικών	Καλύτερος Αλγόριθμος	Ορθότητα	Ακρίβεια	Ανάκληση	Αρμονίκος Μέσος
None	Random Forest	**0.8689**	**0.8862**	**0.8466**	**0.8659**
Select From Model	Random Forest	0.8639	0.8819	0.8404	0.8606
RFE	Random Forest	0.8631	0.8848	0.8417	0.8627
Chi-Squared	Random Forest	0.8623	0.8803	0.8387	0.8590
Variance Threshold	Random Forest	0.86305	0.8827	0.8373	0.8594

Πίνακας 2: Συγκριτικά αποτελέσματα διασύνδεσης χωρίς και με επιλογή χαρακτηριστικών

Από τα αποτελέσματα διακρίνουμε ότι, στο συγκεκριμένο σενάριο, η επιλογή χαρακτηριστικών δεν προσφέρει κάποιο όφελος στην αποτελεσματικότητα του μοντέλου. Μάλιστα, μειώνει οριακά την αποτελεσματικότητά του (περίπου 0.5% στις διάφορες μετρικές).

Με βάση τα παραπάνω ευρήματα, καταλήγουμε στα ακόλουθα συμπεράσματα:

- Οι μέθοδοι επιλογής χαρακτηριστικών δεν ωφελούν το συγκεκριμένο πρόβλημα πιθανότατα λόγω του μικρού αριθμού (~40) χαρακτηριστικών εκπαίδευσης.
- Ενδεχομένως η φύση των συγκεκριμένων χαρακτηριστικών (ως επί το πλείστον χαρακτηριστικά που αντιπροσωπεύουν παραλλαγές της κειμενικής ομοιότητας ενός ζεύγους τοπωνυμίων) να τα καθιστά στην πλειονότητά τους απαραίτητα για την αποτελεσματική λειτουργία του αλγορίθμου μηχανικής μάθησης.
- Οι ενσωματωμένοι μηχανισμοί επιλογής χαρακτηριστικών που ήδη εμπεριέχονται στους επιλεγμένους αλγόριθμους μηχανικής μάθησης (π.χ. κανονικοποίηση στα μοντέλα SVM, μέγιστος αριθμός φύλλων και στοιχείων ανά φύλλο στα Δέντρα Αποφάσεων) επιτελούν αρκούντως αποτελεσματικά επιλογή χαρακτηριστικών ως μέρος του αλγόριθμου μηχανικής μάθησης, καθιστώντας περιττή (έως και ελαφρά επιζήμια) την επιπλέον, «εξωτερική» εφαρμογή επιλογής χαρακτηριστικών.

Ειδικότερα το τρίτο συμπέρασμα είναι και το σημαντικότερο στο σενάριο μας: οι αλγόριθμοι που υιοθετήθηκαν και προκρίθηκαν στην αξιολόγηση της διασύνδεσης τοπωνυμίων ήδη από το Π2.1 δύνανται να ρυθμίσουν εσωτερικά την επιλογή χαρακτηριστικών, επιτυγχάνοντας τη βέλτιστη ακρίβεια, στο συγκεκριμένο πρόβλημα.

4.2. Βιβλιοθήκη Κατηγοριοποίησης Σημείων Ενδιαφέροντος

4.2.1. Σύνολο αξιολόγησης

Το αρχικό σύνολο δεδομένων αξιολόγησης συνίσταται σε σύνολα δεδομένων ΣΕ που αποτελούν προϊόν της εταιρείας Geodata, τα οποία εμπορεύεται σε B2B επίπεδο. Ορισμένα στατιστικά ενδιαφέροντος παρουσιάζονται στον Πίνακας 3: Στατιστικές ιδιότητες του συνόλου δεδομένων Σημείων Ενδιαφέροντος της εταιρίας Geodata. 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 3: Στατιστικές ιδιότητες του συνόλου δεδομένων Σημείων Ενδιαφέροντος της εταιρίας Geodata

Το παραπάνω σύνολο χρησιμοποιήθηκε για την εκτενή αξιολόγηση της βιβλιοθήκης διασύνδεσης τοπωνυμίων, στο Π2.1. Προκειμένου να συγκρίνουμε εκείνη, την αρχική έκδοση της βιβλιοθήκης, με την ανανεωμένη έκδοση που ενσωματώνει μεθόδους επιλογής χαρακτηριστικών, επικεντρώνουμε στην αξιολόγηση των αλγορίθμων όσον αφορά στο δεύτερο επίπεδο κατηγοριών (71 κατηγορίες), μιας και στο συγκεκριμένο σενάριο αποδείχθηκε ότι οι αλγόριθμοι έχουν μεγαλύτερα περιθώρια βελτίωσης (βλέπε Π2.1, κεφάλαιο 4.2.3).

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, τέσσερα από τα πέντε μέρη χρησιμοποιούνται ως σύνολο εκπαίδευσης (με την εσωτερική χρήση υποσυνόλων εκπαίδευσης και επαλήθευσης – training/validation sets), 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 σε όλες τις διαμερίσεις (folds).

4.2.3. Αποτελέσματα

Ο παρακάτω πίνακας παρουσιάζει τα αποτελέσματα των μεθόδων επιλογής χαρακτηριστικών, σε σύγκριση με την εφαρμογή του μοντέλου χωρίς επιλογή χαρακτηριστικών (**None**). Η αξιολόγηση γίνεται για το δεύτερο επίπεδο κατηγοριοποίησης, το οποίο αποτελείται από 71 διακριτές κατηγορίες. Σημειώνουμε επίσης ότι η αξιολόγηση επικεντρώνεται στον καλύτερο από τους αλγόριθμους μηχανικής μάθησης που δοκιμάστηκαν (βλέπε Κεφάλαιο 3.2.2), δηλαδή τον Extra Trees. Οι μετρήσεις είναι σε κλίμακα %. Με έντονη γραμματοσειρά παρουσιάζονται τα καλύτερα αποτελέσματα.

Μέθοδος επιλογής Χαρακτηριστικών	Καλύτερος Αλγόριθμος	Ακρίβεια	Top-5 Ακρίβεια	Top-10 Ακρίβεια	F-Score
None	Extra Trees	0.6860	**0.8685**	0.92	0.6672
Select From Model	Extra Trees	0.6688	0.8436	0.9096	0.6688
RFE	Extra Trees	0.6737	0.8401	0.8995	0.6739
Chi-Squared	Extra Trees	**0.6882**	0.8557	**0.9245**	**0.6811**
Variance Threshold	Extra Trees	0.6786	0.8613	0.9177	0.6633

Πίνακας 4: Συγκριτικά αποτελέσματα κατηγοριοποίησης χωρίς και με επιλογή χαρακτηριστικών

Από τα αποτελέσματα διακρίνουμε ότι ο απλός αλγόριθμος επιλογής χαρακτηριστικών X^2 βελτιώνει ελαφρά την αποτελεσματικότητα των αλγορίθμων, κυρίως όσον αφορά την Top-10 ακρίβεια, καθώς και το F-score, ενώ μειώνει ελαφρά την Top-5 ακρίβεια. Αν και συνολικά οι μέθοδοι επιλογής χαρακτηριστικών δεν επιφέρουν, εκ πρώτης όψεως, κάποια αξιοσημείωτη βελτίωση των αποτελεσμάτων, μειώνουν τη διάσταση των χαρακτηριστικών εκπαίδευσης και ταυτόχρονα διατηρούν τα αποτελέσματα στα ίδια επίπεδα κάτι το οποίο σε σύνολα δεδομένων μεγάλων διαστάσεων (όπως το συγκεκριμένο πρόβλημα, στο οποίο ορίζεται ένας εξαιρετικά μεγάλος και αραιός χώρος χαρακτηριστικών) θα έχουν και βελτίωση στους υπολογιστικούς πόρους.

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 τρεις διαφορετικές πηγές γεωκωδικοποίησης: εσωτερική βάση Geodata, ArcGIS[28] και OpenStreetMap[29].

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, τέσσερα από τα πέντε μέρη χρησιμοποιούνται ως σύνολο εκπαίδευσης (με την εσωτερική χρήση υποσυνόλων εκπαίδευσης και επαλήθευσης – training/validation sets) 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 δύο 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 σε όλες τις διαμερίσεις (folds).

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

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

4.3.3. Αποτελέσματα

Ο παρακάτω πίνακας παρουσιάζει τα αποτελέσματα των μεθόδων επιλογής χαρακτηριστικών, σε σύγκριση με την εφαρμογή του μοντέλου χωρίς επιλογή χαρακτηριστικών (**None**). Σημειώνουμε επίσης ότι η αξιολόγηση επικεντρώνεται στον καλύτερο από τους αλγόριθμους μηχανικής μάθησης που δοκιμάστηκαν (βλέπε Κεφάλαιο 3.3.2), δηλαδή τον Random Forest. Οι μετρήσεις είναι σε κλίμακα %. Με έντονη γραμματοσειρά παρουσιάζονται τα καλύτερα αποτελέσματα.

Μέθοδος επιλογής Χαρακτηριστικών	Καλύτερος Αλγόριθμος	Ορθότητα	F1-macro	F1-micro	Αρμονίκος Μέσος
None	Random Forest	0.6314	0.4965	0.6314	0.6081
Select From Model	Random Forest	0.6402	0.5014	0.6402	0.6182
RFE	Random Forest	0.6364	0.5059	0.6352	0.6105
Chi-Squared	Random Forest	**0.6452**	**0.5477**	**0.6451**	**0.6240**
Variance Threshold	Random Forest	0.6402	0.5014	0.6402	0.6182

Πίνακας 5: Συγκριτικά αποτελέσματα γεωκωδικοποίησης χωρίς και με επιλογή χαρακτηριστικών

Από τα αποτελέσματα είναι εμφανές ότι στο σύνολο τους οι μέθοδοι επιλογής χαρακτηριστικών βελτιώνουν τα αποτελέσματα για κάθε μετρική. Αυτό που αξίζει να σημειωθεί είναι ότι η μέθοδος με τα καλύτερα αποτελέσματα είναι από τη κατηγορία των μεθόδων φιλτραρίσματος (ξανά η X^2), κάτι το οποίο υποδηλώνει, είτε τη στατιστική συνάφεια και αλληλοεπικάλυψη που μπορεί να έχουν τα χαρακτηριστικά μεταξύ τους, είτε ότι ακόμα και η αφαίρεση χαρακτηριστικών που έχουν πολύ χαμηλή διακύμανση μπορεί να βελτιώσει τα αποτελέσματα.

5. Σύνοψη

Στο παρόν Παραδοτέο 3.1 παρουσιάσαμε τις επεκτεταμένες εκδόσεις των μεθόδων μηχανικής μάθησης για διασύνδεση, κατηγοριοποίηση και γεωκωδικοποίηση χωρο-κειμενικών δεδομένων που υλοποιήσαμε στο πλαίσιο της ΕΕ3 του έργου, ενσωματώνοντας μηχανισμούς επιλογής χαρακτηριστικών. Παράλληλα, παρουσιάστηκε η ανανεωμένη γενική αρχιτεκτονική των βιβλιοθηκών, η οποία ενσωματώνει το επιπλέον άρθρωμα της επιλογής χαρακτηριστικών.

Η συγκριτική πειραματική αξιολόγηση των επεκτεταμένων μοντέλων σε σχέση με τα αρχικά κατέδειξε ετερογενή συμπεράσματα. Παρόλο που όλες οι βιβλιοθήκες επωφελούνται σε θέματα απόδοσης (οι αλγόριθμοι μηχανικής μάθησης εκπαιδεύονται ταχύτερα με μειωμένα σύνολα χαρακτηριστικών εκπαίδευσης), τα οφέλη στην αποτελεσματικότητα διαφέρουν:

- Στην περίπτωση *διασύνδεσης τοπωνυμίων*, η ακρίβεια των αλγορίθμων μειώνεται ελαφριά, πράγμα που καταδεικνύει ότι πιθανότατα οι ενσωματωμένοι μηχανισμοί των αλγορίθμων μηχανικής μάθησης επιτελούν την αναγκαία επιλογή χαρακτηριστικών και περαιτέρω βήματα επιλογής βλάπτουν το μοντέλο. Επιπλέον, τα αποτελέσματα της αξιολόγησης καταδεικνύουν ότι, ενδεχομένως, όλα τα υλοποιημένα χαρακτηριστικά εκπαίδευσης για το συγκεκριμένο μοντέλο μηχανικής μάθησης είναι αρκούντως χρήσιμα στο συγκεκριμένο σενάριο. Υπενθυμίζουμε ότι τα χαρακτηριστικά εκπαίδευσης για διασύνδεση τοπωνυμίων αποτελούνται στην πλειονότητά τους από παραλλαγές κειμενικής ομοιότητας μεταξύ των τοπωνυμίων. Αυτό σημαίνει ότι η πλειονότητα αυτών των παραλλαγών ομοιότητας μπορούν να ωφελήσουν τον αλγόριθμο μηχανικής μάθησης που αποφασίσει αν δύο τοπωνύμια είναι ίδια ή όχι.
- Στην περίπτωση *κατηγοριοποίησης ΣΕ*, η ακρίβεια αυξάνεται ελαφριά, αλλά υπάρχει το κέρδος της μείωσης ενός αρκετά μεγάλου χώρου χαρακτηριστικών. Στη συγκεκριμένη βιβλιοθήκη μάλιστα, τα αρχικά οριζόμενα χαρακτηριστικά δημιουργούν έναν αρκετά μεγάλο σε διαστάσεις και αρκετά αραιό χώρο. Για παράδειγμα, ένα μεγάλο πλήθος χαρακτηριστικών αφορά διακριτούς όρους (λέξεις) και διακριτές γειτονικές κατηγορίες για ένα ΣΕ. Δεδομένου αυτού, είναι σαφές το κέρδος από την εφαρμογή επιλογής χαρακτηριστικών στη μείωση της πολυπλοκότητας του μοντέλου, χωρίς πρακτικές απώλειες (συγκεκριμένα με ελαφρά αύξηση) στην αποτελεσματικότητα-ακρίβεια του μοντέλου κατηγοριοποίησης ΣΕ.
- Στην περίπτωση γεωκωδικοποίησης διευθύνσεων, τα κέρδη από την επιλογή χαρακτηριστικών είναι πιο εμφανή, βελτιώνοντας αρκετά την ακρίβεια των αλγορίθμων. Αυτό το αποτέλεσμα είναι επίσης συνεπές με το αρχικό σύνολο χαρακτηριστικών που ορίστηκαν για τη συγκεκριμένη βιβλιοθήκη μηχανικής μάθησης. Συγκεκριμένα, κάποιοι τύποι χαρακτηριστικών ορίστηκαν εν γνώσει μας έχοντας μερική επικάλυψη μεταξύ τους (για παράδειγμα «αποστάσεις σημείων» έναντι «αποστάσεων μεμονωμένων συντεταγμένων»). Η λογική της εισαγωγής μερικής επικάλυψης στο σύνολο των χαρακτηριστικών ήταν ακριβώς ότι οι αλγόριθμοι επιλογής χαρακτηριστικών θα ήταν σε θέση να απαλύνουν αυτό το

φαινόμενο, επιλέγοντας ταυτόχρονα μόνο τα βέλτιστα χαρακτηριστικά, μεταξύ των επικαλυπτόμενων χαρακτηριστικών. Η υπολογίσιμη αύξηση της ακρίβειας μέσω επιλογής χαρακτηριστικών στο συγκεκριμένο σενάριο επιβεβαιώνει τα παραπάνω.

Περαιτέρω επεκτάσεις και βελτιώσεις των παραπάνω βιβλιοθηκών, καθώς και η παρουσίαση της βελτιωμένης έκδοσης της τέταρτης βιβλιοθήκης (ολοκλήρωσης γεωτεμαχίων) θα καταγραφούν στο Παραδοτέο 2.2 «Βελτιστοποιημένοι αλγόριθμοι μηχανικής μάθησης για διασύνδεση» το οποίο είναι προς παράδοση το Μ24 του έργου.

6. Αναφορές

[AB96]	Aha D.W., Bankert R.L. A Comparative Evaluation of Sequential Feature Selection Algorithms. In: Fisher D., Lenz HJ. (eds) Learning from Data. Lecture Notes in Statistics, vol 112. Springer, New York, NY (1996)
[CWE06]	Lal T.N., Chapelle O., Weston J., Elisseeff A. Embedded Methods. In: Guyon I., Nikravesh M., Gunn S., Zadeh L.A. (eds) Feature Extraction. Studies in Fuzziness and Soft Computing, vol 207. (2006)
[DOR+14]	Nilesh Dalvi, Marian Olteanu, Manish Raghavan, and Philip Bohannon. 2014. Deduplicating a Places Database. In Proceedings of WWW '14.
[WFH+17]	Ian H. Witten, Eibe Frank, Mark A. Hall, Christopher J. Pal. Data Mining (Fourth Edition), Chapter 8 - Data transformations, Morgan Kaufmann, 2017, Pages 285-334, ISBN 9780128042915, https://doi.org/10.1016/B978-0-12-804291-5.00008-8.
[SAT07]	Sánchez-Maroño N., Alonso-Betanzos A., Tombilla-Sanromán M. Filter Methods for Feature Selection – A Comparative Study. In: Yin H., Tino P., Corchado E., Byrne W., Yao X. (eds) Intelligent Data Engineering and Automated Learning - IDEAL 2007. IDEAL 2007.