ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Πρόβλεψη Κυκλοφοριακού Φόρτου Σε Οδικά Δίκτυα

ΚΑΡΚΑΝΗΣ Ε. ΕΥΣΤΡΑΤΙΟΣ
Α.Μ.: Π19064
ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΠΕΛΕΚΗΣ ΝΙΚΟΛΑΟΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΠΕΙΡΑΙΑΣ, 2023

Ευχαριστίες

Κατά τη διάρκεια των φοιτητικών μου χρόνων στο Τμήμα Πληροφορικής του Πανεπιστημίου Πειραιώς, ανακάλυψα το βαθύ ενδιαφέρον μου με τον χώρο της μηχανικής μάθησης και της ανάλυσης των δεδομένων. Αυτή η ανακάλυψη ήταν αυτό που με ενέπνευσε να επιλέξω να αφιερώσω την πτυχιακή μου εργασία σε ένα θέμα που ανήκει σε αυτόν τον ευρύτερο τομέα. Επομένως, νιώθω ιδιαίτερα χαρούμενος και ευγνώμων που έχω αναλάβει να ασχοληθώ εκτενέστερα με ένα κλάδο που μου αρέσει πολύ. Ωστόσο, αυτή μου η προσπάθεια δεν θα ήταν δυνατόν να πραγματοποιηθεί χωρίς την συμβολή ορισμένων ατόμων.

Πρώτα και κύρια, θέλω να εκφράσω τις θερμές μου ευχαριστίες στον κο. Πελέκη Νικόλαο, Επίκουρο Καθηγητή του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς, καθώς και στην κα. Χονδροδήμα Εύα, μέλος του Data Science Lab του πανεπιστημίου Πειραιώς, για την ανεκτίμητη υποστήριξη, καθοδήγηση και εμπιστοσύνη που μου παρείχαν σε όλη τη διάρκεια εκπόνησης αυτής της μελέτης. Η συνεισφορά τους σε οργανωτικό και τεχνικό επίπεδο ήταν κρίσιμη για την πραγματοποίηση του έργου αυτού.

Δεν μπορώ να παραλείψω να ευχαριστήσω το εκπαιδευτικό προσωπικό του Τμήματος Πληροφορικής για την υψηλή ποιότητα εκπαίδευσης που έλαβα κατά τη διάρκεια των σπουδών μου. Η γνώση που αποκόμισα αποτέλεσε το θεμέλιο για την επιτυχημένη και ομαλότερη ολοκλήρωση αυτής της έρευνας.

Τέλος, θα ήθελα να εκφράσω την ευγνωμοσύνη μου προς όλους όσους με στήριξαν και με ενθάρρυναν σε όλη τη διάρκεια αυτής της πορείας, ιδίως προς την οικογένειά μου, που είναι πάντα στο πλευρό μου και με υποστηρίζει στις προσπάθειές μου.

Περίληψη

Η παρούσα πτυχιακή έρευνα επικεντρώνεται στον τομέα της μηχανικής μάθησης και συγκεκριμένα εξετάζει την πρόβλεψη της κυκλοφοριακής ροής των κίτρινων ταξί εντός του οδικού δικτύου της πόλης του San Francisco στην Καλιφόρνια. Εδώ ακολουθείται μία διαφορετική προσέγγιση για την διαδικασία των προβλέψεων, αφού η τελευταία γίνεται σε επίπεδο μονοπατιών. Ένα μονοπάτι είναι μία συνεχόμενη ακολουθία από τμήματα μεταξύ διασταυρώσεων εντός του ίδιου οδικού δικτύου. Η διαδικασία της πρόβλεψης αναπτύσσεται με την χρήση προηγμένων τεχνικών μηχανικής και βαθιάς μάθησης εξετάζοντας και αξιολογώντας τέσσερα μοντέλα: XGBoost, LSTM, Encoder-Decoder και Random Forest. Με βάση το χαμηλότερο RMSE score, το XGBoost επιλέγεται ως το ιδανικό μοντέλο για την εφαρμογή των βραχυπρόθεσμων προβλέψεων. Δεδομένης της πολυπλοκότητας και της πολυδιάστατης φύσης της ροής κυκλοφορίας, επιλέγονται να γίνουν μόνο βραχυπρόθεσμες προβλέψεις για το μέγεθος της κυκλοφοριακής ροής σε κάθε μονοπάτι.

Λέξεις Κλειδιά: πρόβλεψη κυκλοφοριακής ροής, μηχανική μάθηση, XGBoost, χρονοσειρές.

Abstract

This thesis focuses on the field of machine learning and specifically examines the prediction of yellow taxi traffic flow on the road network of the city of San Francisco, California. Here, a different approach is considered for the prediction process, as the latter is done at a path level. A path is a continuous sequence of segments between intersections within the same road network. The prediction process is developed using advanced machine and deep learning techniques by considering and evaluating four models: the XGBoost, LSTM, Encoder-Decoder and Random Forest. Based on the lowest RMSE score, XGBoost is selected as the ideal model for the short-term forecasting application. Given the complexity and multidimensional nature of traffic flow, only short-term predictions of the magnitude of traffic flow on each path are chosen.

Keywords: traffic flow forecasting, machine learning, XGBoost, time series.

Κατάλογος Εικόνων

Εικόνα 3.1: Το οδικό δίκτυο της πόλης του San Francisco, California. Επάνω σε αυτό το δίκτυο κινούνται τα ταξί, των οποίων την κίνηση μελετάμε……..…16

Εικόνα 5.1: Το τελικό σύνολο δεδομένων……..….26

Εικόνα 5.2: RMSE και MAE scores για κάθε ένα από τα μοντέλα που εκμεταλλευτήκαμε……..…37

Κατάλογος Διαγραμμάτων

Διάγραμμα 5.1: Στον οριζόντιο άξονα περιλαμβάνεται το μήκος των μονοπατιών που περιέχει κάθε δέσμη. Ο κατακόρυφος άξονας περιέχει τον χρόνο εκτέλεσης των είκοσι ερωτημάτων της δέσμης. Με μπλε χρώμα σημειώνεται η εκτέλεση στο περιβάλλον της PostgreSQL και με πορτοκαλί, στο περιβάλλον της Python……25

Διάγραμμα 5.2: Το μήκος του μονοπατιού κυμαίνεται από 2 έως 15 ακμές. Παρατηρούμε ότι στο σύνολο δεδομένων τα μήκη των μονοπατιών έχουν κατανεμηθεί με σχετικά ομοιόμορφο τρόπο.26

Διάγραμμα 5.3: Στον οριζόντιο άξονα απεικονίζεται ο χρόνος, ενώ ο κατακόρυφος άξονας μετράει το συνολικό άθροισμα της κυκλοφοριακής ροής σε όλα τα μονοπάτια……...27

Διάγραμμα 5.4: Συνολική ροή κυκλοφορίας σε κάθε ημέρα……28

Διάγραμμα 5.5: Η κυκλοφοριακή ροή κατά την ημέρα 2008-05-18 χωρισμένη σε διαστήματα τριών ωρών……..29

Διάγραμμα 5.6: Μήτρα συσχέτισης του συνόλου δεδομένων που χρησιμοποιείται στην έρευνα. Από αυτό το γράφημα προκύπτουν πολλές πληροφορίες για τις σχέσεις των χαρακτηριστικών. Για παράδειγμα, τα χαρακτηριστικά «hour» και «hour sin» φαίνεται να έχουν αρνητική γραμμική συσχέτιση (κοντά στο -1), ενώ τα «sea level pressure» και «day of week cos» έχουν θετική γραμμική συσχέτιση (κοντά στο 1). Τέλος, τα χαρακτηριστικά «Traffic Flow» και «Length» δεν έχουν γραμμική σχέση μεταξύ τους (τιμή κοντά στο 0)……31

Διάγραμμα 5.7: Απεικονίζεται η σχέση του RMSE (κατακόρυφος άξονας) με το μήκος του παραθύρου που εφαρμόζεται κάθε φορά στα ίδια δεδομένα (οριζόντιος άξονας)……34

Διάγραμμα 5.8: Επίδοση του μοντέλου XGBoost στο σύνολο εκπαίδευσης……35

Διάγραμμα 5.9:Επίδοση του μοντέλου XGBoost στο σύνολο ελέγχου……..35

Διάγραμμα 5.10: Γνωστές και προβλεπόμενες τιμές του μεγέθους της κυκλοφοριακής ροής στο μονοπάτι με αριθμό 0……38

Διάγραμμα 5.11: Γνωστές και προβλεπόμενες τιμές του μεγέθους της κυκλοφοριακής ροής στο μονοπάτι με αριθμό 4. ……..38

Διάγραμμα 5.12: Γνωστές και προβλεπόμενες τιμές του μεγέθους της κυκλοφοριακής ροής στο μονοπάτι με αριθμό 36……38

Διάγραμμα 5.13: Γνωστές και προβλεπόμενες τιμές του μεγέθους της κυκλοφοριακής ροής στο μονοπάτι με αριθμό 83……39

Πίνακας Περιεχομένων

Ευχαριστίες ... 1
Περίληψη ... 2
Abstract ... 2
Κατάλογος Εικόνων ... 3
Κατάλογος Διαγραμμάτων ... 4
Πίνακας Περιεχομένων ... 5
Εισαγωγή ... 7
Δομή Τόμου Εργασίας ... 7
1. Θεωρητικό Υπόβαθρο της Εφαρμογής ... 8
1.1 Παρουσίαση του Προβλήματος ... 8
1.2 Τα Αυστηρά Ερωτήματα Μονοπατιού ... 8
1.3 Εφαρμογή των ΑΕΜ στην Παρούσα Μελέτη ... 9
2. Σχετικές Εργασίες ... 10
2.1 Επίλυση του Προβλήματος με ένα Μοντέλο LSTM RNN ... 10
2.2 Πρόβλεψη με Απλά Νευρωνικά Δίκτυα ... 11
2.3 Μοντέλο Αποτελούμενο από Κωδικοποιητές και Αποκωδικοποιητές ... 12
2.4 Πρόβλεψη Κυκλοφοριακής Ροής με το Μοντέλο Deep Crowd ... 12
2.5 Χρήση Γραφημάτων και Νευρωνικών Δικτύων ... 13
2.6 Έρευνα Βασισμένη στον Αλγόριθμο XGBoost ... 14
3. Δεδομένα ... 15
3.1 Δεδομένα Τροχιών ... 15
3.2 Δεδομένα Καιρού ... 17
4. Τεχνολογίες ... 17
4.1 Η Γλώσσα Python ... 17
4.1.1 Python και Επιστήμη Δεδομένων ... 18
4.2 Jupyter Notebook και Google Colab ... 18
4.2.1 Jupyter Notebook ... 19
4.2.2 Google Colab ... 19
4.3 PostgreSQL και PLpgSQL [2] ... 19
4.3.1 Γενικά Χαρακτηριστικά για την Βάση Δεδομένων ... 19
4.3.2 Η Γλώσσα PL/pgSQL [21] ... 20
4.4 Το Λογισμικό Docker ... 21
4.5 Το Λογισμικό Valhalla ... 21
4.5.1 Η Υπηρεσία Valhalla Meili ... 21
5. Υλοποίηση της Εφαρμογής ... 22
5.1 Προεπεξεργασία των Δεδομένων ... 23
5.2 Αντιστοίχιση Τροχιών στο Οδικό Δίκτυο ... 23

5.3 Αναγωγή του Προβλήματος σε Χρονοσειρές 24
5.3.1 Υλοποίηση των Αυστηρών Ερωτημάτων Μονοπατιού 24
5.3.2 Το Τελικό Σύνολο Δεδομένων 25
5.4 Πώς Επηρεάζουν τα ΑΕΜ την Πληροφορία της Κυκλοφοριακής Ροής; 26
5.5 Προσθήκη Επιπλέον Πληροφορίας στο Τελικό Dataset 27
5.6 Οπτικοποίηση των δεδομένων 28
5.7 Χρήση Μοντέλων Μηχανικής και Βαθιάς Μάθησης 29
5.7.1 Διαχωρισμός σε Σύνολα Εκπαίδευσης και Ελέγχου 30
5.7.2 Διαχωρισμός των Χαρακτηριστικών σε Σύνολα Feature και Label 30
5.7.3 Εκπαίδευση του Αλγορίθμου XGBoost 31
5.8 Προβλέψεις 37
6. Συμπεράσματα και Προτάσεις για Βελτίωση 39
Πίνακας Ορολογιών 40
Πίνακας Συντμήσεων – Αρκτικόλεξων – Ακρωνύμιων 40
Βιβλιογραφικές Πηγές 40
Διαδικτυακές Αναφορές 41
Παράρτημα Α 42
Παράρτημα Β 42

Εισαγωγή

Η κυκλοφοριακή ροή αποτελεί ένα κρίσιμο και αναπόσπαστο μέρος της καθημερινότητας του ανθρώπου. Ανεξαρτήτως του αν βρίσκεται σε μια απομονωμένη κοινότητα ή στην καρδιά μιας πολυσύχναστης πόλης, η καθημερινή ρουτίνα εξαρτάται σε μεγάλο βαθμό από την απρόσκοπτη ροή των οχημάτων στον δρόμο. Προβλήματα όπως η κυκλοφοριακή συμφόρηση και ο χρόνος ταξιδιού καθίστανται ολοένα και πιο συνήθη, επιβάλλοντας την ανάγκη για καινοτόμες λύσεις.

Η τεχνολογική πρόοδος των τελευταίων ετών έχει ανοίξει τον δρόμο για την συλλογή, αποθήκευση και ανάλυση μεγάλου όγκου δεδομένων, γνωστών και ως «big data». Έτσι, η αυξημένη προσβασιμότητα σε δεδομένα κυκλοφοριακής ροής και η δυνατότητα της ταχείας επεξεργασίας τους, έχει επιτρέψει την ανάπτυξη μοναδικών μεθοδολογιών και προσεγγίσεων για την πρόβλεψη της κυκλοφοριακής ροής. Παράλληλα, η ανάγκη για βελτιωμένες πρακτικές μετακίνησης και για αντιμετώπιση των συνεχών προκλήσεων στον τομέα της κυκλοφοριακής ροής, έχει οδηγήσει πολλούς φοιτητές, ερευνητές και επιστήμονες σε εντατική έρευνα γύρω από αυτό το ζήτημα.

Στο πλαίσιο αυτό, η παρούσα πτυχιακή εργασία αποτελεί μία ακόμα έρευνα γύρω από αυτόν τον τομέα. Συγκεκριμένα, η μελέτη αυτή επιδιώκει την ανάπτυξη μιας προηγμένης προσέγγισης για την πρόβλεψη της κυκλοφοριακής ροής με την χρήση αλγορίθμων μηχανικής και βαθιάς μάθησης. Το πρόβλημα προσεγγίζεται υπό ένα μοναδικό πρίσμα, μια μεθοδολογία που, από όσο γνωρίζουμε, δεν έχει χρησιμοποιηθεί στο παρελθόν για την επίλυση ενός τέτοιου προβλήματος.

Η μηχανική μάθηση αποτελεί αναπόσπαστο κομμάτι της τεχνητής νοημοσύνης και αναζητά να αναπτύξει αλγόριθμους και μοντέλα που επιτρέπουν στα συστήματα να εκπαιδεύονται από δεδομένα και να βελτιώνουν την απόδοσή τους. Μια ειδική κατηγορία αυτής της προσέγγισης είναι η βαθιά μάθηση, η οποία βασίζεται σε τεχνητά νευρωνικά δίκτυα και αποσκοπεί στην ανάλυση και εξαγωγή υψηλού επιπέδου χαρακτηριστικών και μοτίβων από τα δεδομένα. Επικεντρωμένοι στην ενίσχυση της καθημερινής μας κινητικότητας μέσω της πρόβλεψης της ροής κυκλοφορίας, στην έρευνα αναλύουμε τέσσερα διαφορετικά μοντέλα, ένα δίκτυο Long Short - Term Memory (LSTM), ένα μοντέλο Encoder - Decoder, έναν αλγόριθμο Random Forest και ένα μοντέλο XGBoost. Με τη χρήση των ίδιων δεδομένων, αναπτύσσουμε αυτά τα μοντέλα και αξιολογούμε τις επιδόσεις τους στο ίδιο σύνολο ελέγχου.

Μέσω της παρούσας προσέγγισης, επιδιώκουμε να προσφέρουμε σημαντική συνεισφορά στο πεδίο της μηχανικής μάθησης και της κυκλοφοριακής ανάλυσης. Η μοναδική προσέγγιση που ακολουθείται, σε συνδυασμό με την σύγκριση διαφορετικών μοντέλων, διαφαίνουν τον δρόμο για βελτιωμένες λύσεις στον τομέα της κυκλοφοριακής ροής.

Δομή Τόμου Εργασίας

Η παρούσα εργασία αποτελείται από έξι κεφάλαια. Στο παρόν κεφάλαιο παρουσιάστηκε το πρόβλημα προς επίλυση, χωρίς να δίνεται έμφαση στις λεπτομέρειες. Επιπλέον, έγινε σαφής ο σκοπός της εργασίας.

Στο πρώτο κεφάλαιο του τόμου της εργασίας αναλύεται επακριβώς το πρόβλημα προς επίλυση από θεωρητική σκοπιά. Παράλληλα, δίνεται μία περιγραφή της μεθοδολογίας και των μηχανισμών που συμβάλλουν στην αντιμετώπιση του ζητήματος.

Στο δεύτερο κεφάλαιο γίνεται αναφορά σε παρόμοιες έρευνες που ασχολούνται με το θέμα της κυκλοφοριακής ροής. Για κάθε διαφορετική μελέτη που παρατίθεται αναφέρονται σε βάθος η προσέγγιση του προβλήματος και η μέθοδος που ακολουθείται για την επίλυσή του.

Στο τρίτο κεφάλαιο γίνεται εκτενής περιγραφή των συνόλων δεδομένων που υιοθετήθηκαν στην παρούσα έρευνα.

Στο τέταρτο κεφάλαιο αναφέρονται οι πλατφόρμες και τα πακέτα λογισμικού που χρησιμοποιήθηκαν για την υλοποίηση της παρούσας λύσης και την εμφάνιση των αποτελεσμάτων της.

Στο πέμπτο κεφάλαιο παρουσιάζεται με τεχνικές λεπτομέρειες η υλοποίηση της μεθοδολογίας που προτάθηκε, δίδονται παραδείγματα χρήσης της μέσω εκτέλεσης του κώδικα. Επίσης, αναλύεται η απόδοση των αλγορίθμων που την απαρτίζουν.

Στο τέλος αυτής της έρευνας παρουσιάζονται τα συμπεράσματα από την υλοποίηση της λύσης και ασκείται κριτική στα αδύναμα σημεία της, με παράλληλη παράθεση προτάσεων για βελτίωσή της.

1. Θεωρητικό Υπόβαθρο της Εφαρμογής

Υπάρχουν πολλά πρίσματα, υπό τα οποία μπορεί να εξετάσει κάποιος την κυκλοφοριακή ροή στους δρόμους. Για παράδειγμα, η τελευταία εξετάζεται σε μία ολόκληρη πόλη; σε ένα χωριό; σε μία περίοδο εορτής; ή κατά τη διάρκεια μίας καταιγίδας;

Στο παρόν κεφάλαιο παρατίθεται η μεθοδολογία που ακολουθήθηκε για την επίτευξη αυτού του στόχου. Συγκεκριμένα, ορίζεται ρητά το πρόβλημα προς επίλυση και η οπτική γωνία, υπό την οποία εξετάζεται. Δηλώνονται, επίσης, και οι ερμηνείες σημαντικών εννοιών, η κατανόηση των οποίων καθίσταται αναγκαία για την παρακολούθηση του κειμένου.

1.1 Παρουσίαση του Προβλήματος

Η παρούσα μελέτη επικεντρώνεται στο πρόβλημα της πρόβλεψης του κυκλοφοριακού φόρτου σε ένα οδικό δίκτυο μιας συγκεκριμένης περιοχής. Η έννοια του οδικού δικτύου ταυτίζεται με το σύστημα οδών, δρόμων και διασταυρώσεων της περιοχής. Με τον όρο κυκλοφοριακό φόρτο ή κυκλοφοριακή ροή εννοείται το πλήθος των κινούμενων αντικειμένων που διέρχονται από ένα καθορισμένο μονοπάτι του οδικού δικτύου εντός ενός συγκεκριμένου χρονικού διαστήματος. Για την καλύτερη παρακολούθηση του προβλήματος, ορίζονται οι επόμενες έννοιες:

- ο όρος «**ακμή**» αναφέρεται σε ένα τμήμα του οδικού δικτύου που βρίσκεται ανάμεσα σε δύο διασταυρώσεις.
- ο όρος «**μονοπάτι**» αναφέρεται σε μια ακολουθία από συνεχόμενες ακμές. Η συνεχόμενη φύση του μονοπατιού σημαίνει ότι η ακμή που ακολουθεί την προηγούμενη ακμή αποτελεί την αρχή του επόμενου τμήματος του μονοπατιού, διασφαλίζοντας τη συνεχόμενη σύνδεση των ακμών. Επιπλέον, δηλώνεται ρητά ότι ένα μονοπάτι πρέπει να περιλαμβάνει τουλάχιστον δύο ακμές.
- ο όρος «**κινούμενο αντικείμενο**» υποδηλώνει ένα κινητό στοιχείο που διασχίζει το οδικό δίκτυο.

Στο εξής, η εμβέλεια των παραπάνω ορισμών θα έχει ισχύ σε όλο το κείμενο.

Για την επίλυση του προβλήματος της πρόβλεψης της κυκλοφοριακής ροής σε ένα οδικό δίκτυο, ακολουθείται μία προσέγγιση που, εξ όσων γνωρίζουμε, είναι μοναδική. Με άλλα λόγια, δεν έχουμε καταφέρει να εντοπίσουμε έρευνες άλλων επιστημόνων που να ακολουθούν παρόμοια μεθοδολογία για την επίλυση του ίδιου προβλήματος.

Πιο συγκεκριμένα, γίνεται προσπάθεια να οριστεί ένας αριθμός από μοναδικά μονοπάτια μέσα στο οδικό δίκτυο, για τα οποία υπολογίζεται η κυκλοφοριακή ροή στο κάθε ένα. Για τη μέτρηση της ροής της κυκλοφορίας σε κάθε μονοπάτι, αξιοποιείται η μεθοδολογία των Αυστηρών Ερωτημάτων Μονοπατιού (ΑΕΜ). Η ιδέα πίσω από αυτή την μεθοδολογία αναλύεται στο επόμενο υποκεφάλαιο. Έπειτα, ανάγεται το πρόβλημα σε χρονοσειρές και γίνεται προσπάθεια υιοθέτησης διαφόρων αλγορίθμων μηχανικής μάθησης, για να διεξαχθούν οι προβλέψεις της κυκλοφοριακής ροής σε κάθε ένα μονοπάτι.

1.2 Τα Αυστηρά Ερωτήματα Μονοπατιού

Ο όρος «Αυστηρά Ερωτήματα Μονοπατιού» (ή στα αγγλικά ως «Strict Path Queries») αναφέρεται σε μια διαδικασία αναζήτησης που εκτελείται σε δεδομένα τροχιών κινούμενων αντικειμένων με στόχο την ανάκτηση όλων των τροχιών που διέρχονται αυστηρά από ένα προκαθορισμένο μονοπάτι, ακολουθώντας δηλαδή πιστά τις ακμές που αποτελείται το μονοπάτι μία προς μία και χωρίς να παρεκκλίνουν καθόλου από το μονοπάτι αυτό, μέσα σε ένα προεπιλεγμένο χρονικό διάστημα. [1], [2]

Στα πλαίσια της έρευνας, οι συγγραφείς, για να εξετάσουν την μεθοδολογία τους χρησιμοποιούν δεδομένα κίνησης κινούμενων αντικειμένων, τα οποία καταγράφονται μέσω ενός συστήματος GPS, παρέχοντας πληροφορίες σχετικά με τη θέση τους στον τρισδιάστατο χώρο (x, y, t). Το **x** αντιστοιχεί στο γεωγραφικό μήκος (longitude), το **y** αντιστοιχίζεται στο γεωγραφικό πλάτος (latitude) και το **t** αναπαριστά τον χρόνο (time). Η καταγραφή αυτών των δεδομένων κίνησης για κάθε κινούμενο όχημα ακολουθεί σταθερή περιοδικότητα. Κάθε αναφορά θέσης από το σύστημα GPS αναπαρίστανται από μια πλειάδα παραμέτρων της μορφής **loc=(moid, ts, pos)**, όπου:

- το στοιχείο «moid» (από το moving object id) αναπαριστά το αναγνωριστικό του κινούμενου αντικειμένου.
- το στοιχείο «ts» αντιστοιχεί σε συγκεκριμένη χρονική στιγμή.
- το στοιχείο «pos» (από το position) δηλώνει τη θέση του κινούμενου αντικειμένου κατά τη χρονική στιγμή «ts» με την χρήση χωρικών συντεταγμένων (latitude και longitude).

Μέσω μιας διαδικασίας αντιστοίχισης σημείων GPS σε ψηφιακούς χάρτες, τα αρχικά σημεία που παράγονται για κάθε κινούμενο όχημα αντιστοιχίζονται σε μια ακολουθία ακμών εντός του οδικού δικτύου. Με την εφαρμογή αυτής της διαδικασίας διαμορφώνεται μια τροχιά εντός του οδικού δικτύου για κάθε κινούμενο όχημα. Κάθε τροχιά που συσχετίζεται με ένα συγκεκριμένο κινούμενο όχημα, αποτελείται από πολλές εγγραφές της μορφής **locmm=(tid, eid, tsenter, tsleave)**. Στην εν λόγω αναπαράσταση το «tid» δηλώνει το αναγνωριστικό της τροχιάς, το «eid» αναφέρεται στο αναγνωριστικό της ακμής στο οδικό δίκτυο, ενώ τα «tsenter» και «tsleave» αναφέρονται στους χρόνους εισόδου και εξόδου του κινούμενου αντικειμένου από την ακμή με αναγνωριστικό «eid» αντίστοιχα. Επομένως, τα αρχικά δεδομένα που καταγράφονται από το σύστημα GPS υποβάλλονται σε μια διαδικασία μετατροπής σε εγγραφές locmm. Μια τροχιά t προκύπτει ως ένα σύνολο τέτοιων εγγραφών, δηλαδή η τροχιά t ορίζεται ως εξής: **t = [locmm1, locmm2, …, locmmn]**.

Όσον αφορά την επίδοση της μεθόδου, οι συγγραφείς υποστηρίζουν ότι τα αποτελέσματα που παρέχονται από τον αλγόριθμο SPQ μπορούν να θεωρηθούν ικανοποιητικά. Ο συγκεκριμένος αλγόριθμος διακρίνεται για την υψηλή του ακρίβεια και ταχύτητα εκτέλεσής του σε σύγκριση με άλλες προσεγγίσεις που έγιναν στα πλαίσια της ίδιας έρευνας.

Είναι σημαντικό να αναφερθεί ότι τα ΑΕΜ είναι απαραίτητα για πολλούς λόγους. Αφενός, παρέχουν την πληροφορία για το πόσες τροχιές διέσχισαν ένα συγκεκριμένο μονοπάτι από την αρχή του έως και το τέλος του, χωρίς να παρεκκλίνουν καθόλου από αυτό. Θέτοντάς το διαφορετικά, βοηθούν να προσδιοριστεί η ποσότητα της ροής των κινούμενων αντικειμένων εντός ενός ολοκλήρου μονοπατιού με μεγάλη ακρίβεια. Αφετέρου, υπάρχει δυνατότητα να οριστεί το χρονικό διάστημα που επιλέγεται να γίνει η αναζήτηση της κυκλοφοριακής ροής. Επομένως, με την χρήση αυτής της μεθοδολογίας μπορούν να εξαχθούν σημαντικά συμπεράσματα σχετικά με τη συμπεριφορά της κυκλοφορίας και της μετακίνησης κινούμενων οχημάτων, αναγνωρίζοντας ποια μονοπάτια διασχίζονται συχνότερα κατά τη διάρκεια διαφόρων χρονικών περιόδων, όπως οι ώρες αιχμής. Τέλος, μπορούν να ανακαλυφθούν μοτίβα συμφόρησης και να εντοπιστούν ανωμαλίες στην κυκλοφορία, επιτρέποντας την αντιμετώπιση πιθανών προβλημάτων.

1.3 Εφαρμογή των ΑΕΜ στην Παρούσα Μελέτη

Όπως δηλώθηκε παραπάνω, η έρευνα που διεξάγεται αναφέρεται στην πρόβλεψη της κυκλοφοριακής ροής κινούμενων αντικειμένων μέσα σε ολόκληρα μονοπάτια του οδικού δικτύου μίας περιοχής. Ο τρόπος με τον οποίο γίνεται αυτό περιγράφεται περιληπτικά στο παρόν κεφάλαιο, ενώ αναλυτικότερη εξήγηση δίνεται στο κεφάλαιο της υλοποίησης (πέμπτο κεφάλαιο).

Αρχικά, τα δεδομένα που υπάρχουν στην διάθεσή μας είναι ένα σύνολο από αναφορές θέσεων GPS διαφόρων κινούμενων αντικειμένων. Κάθε αναφορά θέσης αποτελείται από μία πλειάδα τεσσάρων στοιχείων (id,lat,lon,time), όπου το στοιχείο «id» αναφέρεται στο αναγνωριστικό του κινούμενου αντικειμένου, το στοιχείο «lat» αναφέρεται στο γεωγραφικό πλάτος, το στοιχείο «lon» παραπέμπει στο γεωγραφικό μήκος και το στοιχείο «time» στην χρονική στιγμή που έγινε η καταγραφή της θέσης του κινούμενου αντικειμένου.

Τα δεδομένα GPS που δίνονται αρχικά δεν έχουν αντιστοιχηθεί στο οδικό δίκτυο της περιοχής που διερευνείται. Για να γίνει αυτό, χρειάζεται να προηγηθεί μία διαδικασία αντιστοίχισης των GPS

δεδομένων σε ακμές του οδικού δικτύου. Αφού γίνει η διαδικασία της αντιστοίχισης, ορίζεται ένας αριθμός από τυχαία και μοναδικά μονοπάτια τυχαίου μήκους που πρόκειται να δημιουργηθούν. Ο τρόπος με τον οποίο παράγεται ένα μονοπάτι ακολουθεί αυστηρά τον ορισμό που δόθηκε παραπάνω, ενώ οι ακμές από τις οποίες απαρτίζεται κάθε ένα από αυτά προκύπτουν άμεσα από τα δεδομένα.

Στην συνέχεια, δημιουργούνται σταθερά χρονικά διαστήματα, για κάθε ένα από τα οποία μετριέται η κυκλοφοριακή ροή σε όλα τα μονοπάτια που έχουν δημιουργηθεί. Η μέτρηση της κυκλοφοριακής ροής των κινούμενων αντικειμένων πραγματοποιείται με την χρήση των ΑΕΜ. Με αυτόν τον τρόπο, τα αρχικά GPS δεδομένα μετατρέπονται πλέον σε δεδομένα χρονοσειρών, δηλαδή σε μια σειρά από μετρήσεις που καταγράφονται με χρονική σειρά και ανά σταθερά χρονικά διαστήματα μεταξύ τους. Οι χρονοσειρές αυτές αποτελούν τα ιστορικά δεδομένα κυκλοφοριακής ροής για κάθε μονοπάτι. Τέλος, εφαρμόζοντας αλγορίθμους μηχανικής και βαθιάς μάθησης, προβλέπεται η μελλοντική ροή της κυκλοφορίας σε όλα τα μονοπάτια χρησιμοποιώντας τα ιστορικά δεδομένα.

Παρατήρηση: όσα περισσότερα μοναδικά μονοπάτια δημιουργηθούν, τόσο αυξάνεται και η πιθανότητα να καλυφθεί ολόκληρο το προς μελέτη οδικό δίκτυο. Επομένως, η παρούσα έρευνα, αν και εστιάζει στην πρόβλεψη της κυκλοφοριακής ροής εντός ολόκληρων μονοπατιών, μπορεί να χρησιμοποιηθεί και για την πρόβλεψη της κυκλοφοριακής ροής σε ένα σύνολο μονοπατιών που απαρτίζουν ολόκληρο το οδικό δίκτυο.

2. Σχετικές Εργασίες

Σε αυτό το κεφάλαιο παρατίθενται έξι σχετικές εργασίες που έχουν υλοποιηθεί από άλλους ερευνητές. Στις έρευνες αυτές, το πρόβλημα προς επίλυση είναι το ίδιο με αυτό που παρουσιάστηκε προηγουμένως. Μάλιστα, ο τρόπος με τον οποίο ορίζεται η έννοια της κυκλοφοριακής ροής στις επόμενες έρευνες είναι ανάλογος με τον τρόπο που ορίζεται το μέγεθος αυτό στην παρούσα πτυχιακή εργασία.

Στατιστικά μιλώντας, οι περισσότεροι επιστήμονες υιοθετούν αναδρομικά νευρωνικά δίκτυα με μνήμη (LSTMs), απλά νευρωνικά δίκτυα και στατιστικά μοντέλα, όπως τα ARIMA και SARIMA, προκειμένου να επιλύσουν το ζήτημα αυτό. Για κάθε σχετική έρευνα που αναφέρεται, προσδιορίζονται το μοντέλο που χρησιμοποιήθηκε και τα δεδομένα τα οποία δόθηκαν ως είσοδο σε αυτό.

2.1 Επίλυση του Προβλήματος με ένα Μοντέλο LSTM RNN

Τίτλος: Predicting Short-term Traffic Flow by Long Short-Term Memory Recurrent Neural Network [3]

Εισαγωγή:

Η εργασία αυτή εξετάζει τη σημασία της βραχυπρόθεσμης πρόβλεψης της ροής της κυκλοφορίας στις ευφυείς μεταφορές και την εφαρμογή της στη διαχείριση της κυκλοφοριακής συμφόρησης, τη μείωση της ρύπανσης και την ενίσχυση της οδικής ασφάλειας. Επιπροσθέτως, επισημαίνονται οι προκλήσεις που σχετίζονται με την ακριβή πρόβλεψη της έντονα μη γραμμικής και στοχαστικής φύσης της ροής κυκλοφορίας. Αυτό οφείλεται σε ποικίλους παράγοντες, όπως οι μεταβαλλόμενες καιρικές συνθήκες και η μορφολογία του εδάφους, οι οποίοι έχουν αντίκτυπο στη ροή κυκλοφορίας. Στόχος είναι η πρόβλεψη της κυκλοφοριακής ροής σε ένα συγκεκριμένο χρονικό διάστημα με βάση ιστορικά δεδομένα που καταγράφονται σε διαστήματα των 15 λεπτών.

Στο πλαίσιο της παρούσας εργασίας, η κυκλοφοριακή ροή αναφέρεται στον όγκο των οχημάτων που διέρχονται από ένα συγκεκριμένο σταθμό παρατήρησης, τοποθετημένος σε έναν αυτοκινητόδρομο. Η ροή αυτή μετριέται σε διαστήματα των 15 λεπτών. Η εργασία έχει ως στόχο να προβλέψει την βραχυπρόθεσμη κυκλοφοριακή ροή με ακρίβεια, ώστε να παρέχει έγκαιρες και πολύτιμες πληροφορίες για ενδιαφερόμενους, συμπεριλαμβανομένων των ταξιδιωτών, των επιχειρήσεων και των κυβερνητικών υπηρεσιών.

Μεθοδολογία:

Στην έρευνα αυτή χρησιμοποιείται το μοντέλο Long Short-Term Memory Recurrent Neural Network (LSTM RNN) για την βραχυπρόθεσμη πρόβλεψη της ροής της κυκλοφορίας. Το LSTM RNN είναι ένας τύπος νευρωνικού δικτύου κατάλληλος για εργασίες πρόβλεψης χρονοσειρών. Για την εκτέλεση της πρόβλεψης, το μοντέλο LSTM RNN λαμβάνει ως είσοδο δεδομένα ιστορικής ροής κυκλοφορίας, τα οποία περιλαμβάνουν πληροφορίες σχετικές με προηγούμενους όγκους κίνησης (π.χ. τα τελευταία 30 λεπτά).

Έπειτα, το μοντέλο μαθαίνει τα μοτίβα και τις εξαρτήσεις που υπάρχουν στα ιστορικά δεδομένα, για να προβλέψει τη ροή κυκλοφορίας για το επόμενο διάστημα των 15 λεπτών. Η αρχιτεκτονική LSTM RNN περιλαμβάνει μπλοκ μνήμης που επιτρέπει στο δίκτυο να συλλαμβάνει και να αποθηκεύει πληροφορίες για μεγαλύτερες χρονικές περιόδους, αντιμετωπίζοντας την πρόκληση των χρονικών εξαρτήσεων στη ροή της κυκλοφορίας. Σε αντίθεση με τα παραδοσιακά μοντέλα νευρωνικών δικτύων, το LSTM RNN προσδιορίζει δυναμικά τις βέλτιστες χρονικές καθυστερήσεις και τις ενσωματώνει στη διαδικασία πρόβλεψης.

Όσον αφορά την διαδικασία εκπαίδευσης του μοντέλου, χρησιμοποιείται το σύνολο δεδομένων Caltrans Performance Measurement System (PeMS), το οποίο παρέχει στον ερευνητή έναν μεγάλο αριθμό ιστορικών δεδομένων κυκλοφοριακής ροής.

Η απόδοση του μοντέλου αξιολογείται και σε σύγκριση με άλλα μοντέλα, όπως το Random Walk (RW), τον αλγόριθμο Support Vector Machine (SVM) και τα Feed Forward Neural Networks (FFNN). Τα αποτελέσματα δείχνουν την υπεροχή του μοντέλου LSTM RNN, όσον αφορά την ακρίβεια πρόβλεψης και τις δυνατότητες γενίκευσης σε άλλα σύνολα δεδομένων (π.χ. τα δεδομένα ελέγχου). Το LSTM RNN υπερτερεί έναντι άλλων μοντέλων στην αποτύπωση των μη γραμμικών και στοχαστικών χαρακτηριστικών της ροής κυκλοφορίας.

2.2 Πρόβλεψη με Απλά Νευρωνικά Δίκτυα

Τίτλος: Supervised deep learning based for traffic flow prediction [4]

Εισαγωγή:

Στην παρούσα εργασία, ως ροή κυκλοφορίας ορίζεται η κίνηση των οχημάτων εντός ενός οδικού δικτύου ή ενός συγκεκριμένου τμήματος μίας οδού σε μία δεδομένη χρονική στιγμή. Οι συγγραφείς αναφέρουν ότι η ακριβής πρόβλεψη της κυκλοφοριακής ροής είναι ζωτικής σημασίας στα Ευφυή Συστήματα Μεταφορών (ΕΣΜ) για την αποτελεσματική μείωση της κυκλοφοριακής συμφόρησης. Οι επιστήμονες έχουν σαν βασικό στόχο την εκτίμηση του όγκου και της ταχύτητας των οχημάτων στο δρόμο για τη διευκόλυνση της αποτελεσματικής διαχείρισης της κυκλοφορίας.

Μεθοδολογία:

Το μοντέλο που προτείνεται στην παρούσα εργασία για την πρόβλεψη της κυκλοφοριακής ροής είναι ένα μοντέλο SDLTFP (Supervised Deep Learning Based Traffic Flow Prediction), το οποίο είναι ένας τύπος πλήρως συνδεδεμένου νευρωνικού δικτύου (Fully Conncted Deep Neural Network). Το μοντέλο SDLTFP λαμβάνει ιστορικά δεδομένα κυκλοφορίας ως είσοδο και προσπαθεί να προβλέψει την μελλοντική κυκλοφοριακή ροή, δηλαδή τον εκτιμώμενο όγκο των οχημάτων σε μια δεδομένη χρονική στιγμή στο μέλλον.

Στην εργασία, οι συγγραφείς εφαρμόζουν διάφορες τεχνικές βελτιστοποίησης, για να αναβαθμίσουν την απόδοση του μοντέλου. Αυτές οι τεχνικές περιλαμβάνουν την κανονικοποίηση δέσμης (Batch Normalization) και την εισαγωγή των επιπέδων «Dropout». Ο συνδυασμός αυτών των δύο μεθόδων συμβάλλει στην ευκολία γενίκευσης του μοντέλου και την αποφυγή της υπερεκπαίδευσης (overfitting).

Όσον αφορά τα αποτελέσματα, αναφέρεται ότι το μοντέλο SDLTFP επιτυγχάνει μέσο απόλυτο ποσοστιαίο σφάλμα (MAPE) 5% στα δεδομένα εκπαίδευσης. Για τα δεδομένα ελέγχου, ο προτεινόμενος αλγόριθμος καταφέρνει για την ίδια μετρική ένα σφάλμα μεταξύ 15% έως 20%. Τα αποτελέσματα αυτά δείχνουν ότι το μοντέλο αποδίδει καλά στην πρόβλεψη της ροής κυκλοφορίας, παρουσιάζοντας σχετικά χαμηλά σφάλματα πρόβλεψης.

2.3 Μοντέλο Αποτελούμενο από Κωδικοποιητές και Αποκωδικοποιητές

Τίτλος: Traffic Flow prediction with big data: A deep learning approach [5]

Εισαγωγή:

Στο παρόν έγγραφο, η ροή κυκλοφορίας αναφέρεται στην κίνηση των οχημάτων σε δίκτυα μεταφορών, όπως οι δρόμοι ή οι αυτοκινητόδρομοι. Συγκεκριμένα, η τελευταία ορίζεται ως την ποσότητα των οχημάτων που διέρχονται από μια συγκεκριμένη τοποθεσία σε διαφορετικά χρονικά διαστήματα. Η παρατηρούμενη ποσότητα ροής κυκλοφορίας, η οποία συμβολίζεται ως Xt_i, αντιπροσωπεύει τον όγκο κυκλοφορίας που μετρήθηκε στην i-οστή θέση παρατήρησης κατά τη διάρκεια του t-οστού χρονικού διαστήματος.

Η πρόβλεψη της κυκλοφοριακής ροής αποσκοπεί στην παροχή ακριβών και έγκαιρων πληροφοριών σχετικά με τις αναμενόμενες συνθήκες κυκλοφορίας, οι οποίες είναι ζωτικής σημασίας για τη διαχείριση των μεταφορών, για τα ευφυή συστήματα μεταφορών και για διάφορες εφαρμογές που στοχεύουν στον έλεγχο και τη βελτιστοποίηση της κυκλοφορίας.

Μεθοδολογία:

Η μεθοδολογία που χρησιμοποιείται στην παρούσα εργασία περιλαμβάνει την εφαρμογή ενός μοντέλου στοιβαγμένου κωδικοποιητή-αποκωδικοποιητή (Stacked Autoencoder), το οποίος αποτελεί αρχιτεκτονική βαθιάς μάθησης. Το μοντέλο SAE εκπαιδεύεται με τη χρήση του συνόλου δεδομένων PeMS της Caltrans που περιέχει πληροφορίες για την κίνηση και την κατάσταση των οδικών δικτύων σε διάφορες περιοχές της Καλιφόρνιας. Τα βήματα της μεθοδολογίας που ακολουθούνται από τους συντάκτες της έρευνας είναι τα ακόλουθα:

1. **συλλογή δεδομένων**: τα δεδομένα κυκλοφοριακής ροής συλλέγονται από τη βάση δεδομένων PeMS της Caltrans.
2. **προεπεξεργασία**: τα δεδομένα υποβάλλονται σε προεπεξεργασία για την απομάκρυνση τυχόν ακραίων τιμών ή ασυνεπειών σε αυτά. Στη συνέχεια, χωρίζονται σε σύνολα εκπαίδευσης και ελέγχου για την αξιολόγηση του μοντέλου.
3. **σχεδίαση του μοντέλου**: το μοντέλο SAE αποτελείται από πολλαπλά στρώματα autoencoder. Ο autoencoder θεωρείται τύπος νευρωνικού δικτύου βαθιάς μάθησης. Αποτελείται από έναν κωδικοποιητή και έναν αποκωδικοποιητή. Ο κωδικοποιητής συμπιέζει τα δεδομένα εισόδου σε μια αναπαράσταση χαμηλότερης διάστασης, ενώ ο αποκωδικοποιητής ανακατασκευάζει τα αρχικά δεδομένα από αυτή την αναπαράσταση. Το μοντέλο SAE χρησιμοποιεί τους autoencoders ως δομικά στοιχεία για τη δημιουργία ενός βαθιού δικτύου.
4. **εκπαίδευση και αξιολόγηση**: το μοντέλο SAE εκπαιδεύεται στα δεδομένα εκπαίδευσης. Έπειτα, η απόδοση του μοντέλου SAE αξιολογείται με τη χρήση διαφόρων μετρικών, όπως το μέσο απόλυτο σφάλμα (Mean Absolute Error - MAE) στα δεδομένα ελέγχου.

Η απόδοση του μοντέλου συγκρίνεται και με άλλα μοντέλα που χρησιμοποιούνται για τον ίδιο σκοπό, συμπεριλαμβανομένων των αλγορίθμων Support Vector Machine (SVM), Random Walk (RW) και Radial Basis Function Neural Network (RBF NN). Τελικά, η απόδοση του προτεινόμενου μοντέλου είναι ανώτερη από αυτές των υπολοίπων. Όμως, αξίζει να σημειωθεί, ότι το μοντέλο δυσκολεύεται να αποδώσει καλά όταν η κυκλοφοριακή ροή βρίσκεται σε χαμηλά επίπεδα, ενώ για υψηλότερους όγκους κυκλοφοριακής ροής το μοντέλο αποδίδει ικανοποιητικά.

2.4 Πρόβλεψη Κυκλοφοριακής Ροής με το Μοντέλο Deep Crowd

Τίτλος: DeepCrowd - A Deep Model for Large-Scale Citywide Crowd Density and Flow Prediction [6]

Εισαγωγή:

Το πρόβλημα που προσπαθεί να λύσει η παρούσα έρευνα είναι η πρόβλεψη της κυκλοφοριακής ροής. Συγκεκριμένα, οι επιστήμονες προσπαθούν να προβλέψουν αφενός τον αριθμό των κινούμενων αντικειμένων (KA) που θα βρίσκονται σε κάθε θέση μίας πόλης την επόμενη χρονική

στιγμή (το ονομάζουν πρόβλεψη πληθυσμιακής πυκνότητας), αφετέρου, προσπαθούν να εκτιμήσουν πόσα ΚΑ θα φύγουν από μία συγκεκριμένη θέση και θα επισκεφτούν μία άλλη στην επόμενη χρονική στιγμή (το ονομάζουν ροή εισόδου-εξόδου).

Προκειμένου να παρουσιάσουν μία καινοτόμα ιδέα, οι επιστήμονες προτείνουν μια λύση που περιλαμβάνει τη διαίρεση μιας μεγάλης αστικής περιοχής σε μικρά πλέγματα. Κάθε πλέγμα αναπαριστά μία θέση.

Μεθοδολογία:

Οι συγγραφείς προτείνουν ένα μοντέλο βαθιάς μάθησης που ονομάζεται «DeepCrowd». Το τελευταίο βασίζεται σε μία πυραμιδική αρχιτεκτονική αποτελούμενη από συνελικτικά νευρωνικά δίκτυα με μνήμη (Convolutional LSTM) και μηχανισμούς προσοχής υψηλής διάστασης. Τα Convolutional LSTM μπορούν να διαχειριστούν πολυδιάστατα δεδομένα, ενώ ο πυραμιδικός σχεδιασμός αξιοποιεί καλύτερα χαμηλής ή/και υψηλής ποιότητας χαρακτηριστικά.

Η αναπαράσταση της κυκλοφοριακής ροής σε όλη την πόλη γίνεται με την μορφή μίας τετραδιάστατης μήτρας. Οι τέσσερις διαστάσεις της είναι ο χρόνος, το ύψος, το πλάτος και το κανάλι. Το κανάλι λαμβάνει δύο τιμές, ανάλογα με το είδος της πρόβλεψης που γίνεται σε κάθε περίπτωση. Τα δύο είδη πρόβλεψης που μελετώνται στην έρευνα αυτή αναφέρθηκαν παραπάνω.

Αξιοπρόσεχτο είναι το γεγονός ότι για την διεκπεραίωση της μελέτης, οι επιστήμονες κατάφεραν να δημιουργήσουν ένα δικό τους σύνολο δεδομένων. Η εργασία χρησιμοποιεί δεδομένα μεγάλης κλίμακας αποτελούμενα από καταγραφές GPS σημείων σε πραγματικό χρόνο με την βοήθεια μίας εφαρμογής κινητού. Αυτό το σύνολο δεδομένων προσφέρει πλεονεκτήματα σε σχέση με τα ήδη υπάρχοντα, όχι μόνο επειδή καλύπτεται το εύρος μιας μεγάλης περιοχής, αλλά επίσης υπάρχουν καταγραφές από πολλούς χρήστες.

Για την αξιολόγηση των επιδόσεων του «DeepCrowd», οι συγγραφείς διεξάγουν ενδελεχή πειράματα και συγκρίνουν τα αποτελέσματά τους με άλλες σύγχρονες μεθόδους. Τέσσερις μετρικές (MSE, RMSE, MAE και MAPE) χρησιμοποιούνται για την αξιολόγηση. Τα αποτελέσματα των πειραμάτων αποδεικνύουν την αποτελεσματικότητα και την αποδοτικότητα του προτεινόμενου μοντέλου για την πρόβλεψη της κυκλοφοριακής ροής σε δύο μεγάλες πόλεις της Ιαπωνίας, το Τόκιο και την Οσάκα.

Συμπεραίνοντας, με την χρήση ενός ισχυρού μοντέλου όπως το προτεινόμενο, οι συγγραφείς είναι σε θέση να συλλάβουν και να αναλύσουν τις χωροχρονικές πτυχές της κινητικότητας στην πόλη και να ανακαλύψουν μοτίβα στον τομέα της κυκλοφορίας.

GitHub Link: https://github.com/deepkashiwa20/DeepCrowd

2.5 Χρήση Γραφημάτων και Νευρωνικών Δικτύων

Τίτλος: Graph Hierarchical Convolutional Recurrent Neural Network (GHCRNN) for Vehicle Condition Prediction [7]

Εισαγωγή:

Στην παρούσα εργασία, οι συγγραφείς ορίζουν τη κυκλοφοριακή ροή ως τον αριθμό των οχημάτων που διέρχονται από μια συγκεκριμένη τοποθεσία ή περιοχή κατά τη διάρκεια μιας χρονικής περιόδου. Το κύριο πρόβλημα προς επίλυση είναι η πρόβλεψη της κυκλοφοριακής ροής και της ταχύτητας των οχημάτων στο μέλλον. Το πρόβλημα αυτό ανάγεται σε χρονοσειρά. Στόχος είναι η πρόβλεψη των τιμών της με βάση διάφορα ιστορικά δεδομένα. Τέλος, λαμβάνονται υπόψη οι παράγοντες που επηρεάζουν τη κυκλοφοριακή ροή και την ταχύτητα, όπως είναι ο χρόνος, ο χώρος, οι καιρικές συνθήκες και οι δραστηριότητες (π.χ. εορτές), προκειμένου να αναπτυχθεί ένα μοντέλο που θα είναι όσο το δυνατόν ακριβέστερο.

Μεθοδολογία:

Το προτεινόμενο μοντέλο στην παρούσα μελέτη ονομάζεται GHCRNN (Graph Hierarchical Convolutional Recurrent Neural Network). Αυτό το μοντέλο έχει σχεδιαστεί για την πρόβλεψη της

κυκλοφοριακής ροής και της ταχύτητας των οχημάτων σε αστικές περιοχές. Επίσης, ενσωματώνει τόσο χωρικές όσο και ιεραρχικές πληροφορίες για τη βελτίωση της ακρίβειας πρόβλεψης. Ακολουθεί μια περιγραφή του τρόπου λειτουργίας του αλγορίθμου:

1. **ακολουθίες εισόδου και εξόδου**: το μοντέλο GHCRNN λαμβάνει ως είσοδο ιστορικά δεδομένα χρονοσειρών. Αυτό περιλαμβάνει πληροφορίες σχετικά με τη ροή και την ταχύτητα των οχημάτων κατά τη διάρκεια μιας χρονικής περιόδου. Το μοντέλο προβλέπει τη κυκλοφοριακή ροή και την ταχύτητα σε μία καθορισμένη χρονική στιγμή στο μέλλον και την παραδίδει ως έξοδο.
2. **συνελικτικά επίπεδα**: το νευρωνικό δίκτυο περιλαμβάνει συνελικτικές μονάδες (Conv0 και Conv1) για την εξαγωγή χωρικών πληροφοριών από το οδικό δίκτυο. Αυτά τα συνελικτικά στρώματα αναλύουν τις σχέσεις και τα μοτίβα μεταξύ διαφορετικών θέσεων στο δίκτυο.
3. **μονάδα pooling**: οι μονάδες pooling (Pooling0 και Pooling1) χρησιμοποιούνται για την αποτύπωση της ιεραρχικής δομής του οδικού δικτύου. Η λειτουργία αυτή βοηθά στην εξαγωγή ιεραρχικών πληροφοριών, στην εξάλειψη περιττών γνώσεων και στη μείωση της πολυπλοκότητας των δεδομένων.
4. **επίπεδο κωδικοποιητή - αποκωδικοποιητή**: η συνολική αρχιτεκτονική του GHCRNN χρησιμοποιεί ένα δίκτυο Seq2Seq που βασίζεται σε ένα μοντέλο κωδικοποιητή – αποκωδικοποιητή. Τα ιστορικά δεδομένα τροφοδοτούνται στον κωδικοποιητή, ο οποίος παράγει ενδιάμεσα κωδικοποιημένα αποτελέσματα (State0 και State1). Αυτά τα αποτελέσματα χρησιμεύουν ως είσοδος για τον αποκωδικοποιητή, ο οποίος ολοκληρώνει τη διαδικασία της πρόβλεψης.
5. **μονάδες GHCRNN**: τα επίπεδα κωδικοποιητών και αποκωδικοποιητών περιέχουν πολλαπλές μονάδες GHCRNN, επιτρέποντας στο μοντέλο να χειρίζεται τις ακολουθίες εισόδου και εξόδου. Κάθε μονάδα GHCRNN ενσωματώνει μονάδες GRU (Gated Recurrent Unit). Αυτό επιτρέπει στο μοντέλο να συλλαμβάνει τόσο μακροπρόθεσμες όσο και βραχυπρόθεσμες πληροφορίες.
6. **συνέλιξη γράφων και διαδικασία pooling**: σε κάθε μονάδα GHCRNN ενσωματώνονται λειτουργίες συνέλιξης και pooling για την εξαγωγή χωρικών και ιεραρχικών πληροφοριών. Αυτές οι λειτουργίες βρίσκουν τις σχέσεις μεταξύ των κόμβων του οδικού δικτύου και αποτυπώνουν τις ομοιότητες ή τις διαφορές στα χαρακτηριστικά και τις δομές τους.

Το σύνολο δεδομένων που χρησιμοποιείται στην παρούσα έρευνα είναι τα δεδομένα ταχύτητας του Los Angeles και αποτελείται από μετρήσεις των μέσων ταχυτήτων για διάφορους σταθμούς ανίχνευσης, οι οποίες συλλέγονται ανά χρονικά διαστήματα των πέντε λεπτών. Συνολικά επιλέχθηκαν 207 σταθμοί ανίχνευσης. Στο σύνολό τους, τα δεδομένα καλύπτουν ένα διάστημα τεσσάρων μηνών.

Στο πείραμα που έγινε, το γράφημα που χρησιμοποιείται αναπαρίσταται ως ένας πίνακας γειτνίασης που καθορίζει τη σχέση σύνδεσης μεταξύ των σταθμών ανίχνευσης στο οδικό δίκτυο. Δύο σταθμοί ανίχνευσης γειτνιάζουν αν υπάρχει άμεση ή έμμεση σύνδεση μεταξύ τους. Το βάρος που αποδίδεται στον πίνακα γειτνίασης καθορίζεται από τη σχέση απόστασης μεταξύ των δύο συνδεδεμένων σταθμών.

Επιπροσθέτως, οι συγγραφείς της έρευνας αναφέρουν ότι το μοντέλο μπορεί να εφαρμοστεί όχι μόνο σε δεδομένα κυκλοφορίας, αλλά και σε δεδομένα που διαθέτουν χρονικά και χωρικά χαρακτηριστικά. Το μοντέλο επιδεικνύει, επίσης, υψηλότερη ικανότητα και αποδοτικότητα κατά την επεξεργασία μεγάλων γράφων.

Συνοψίζοντας, το μοντέλο GHCRNN συνδυάζει την ισχύ των γραφημάτων, των συνελικτικών και αναδρομικών νευρωνικών δικτύων και των πράξεων pooling. Έτσι, παρέχει ακριβείς προβλέψεις της κυκλοφοριακής ροής και της ταχύτητας των οχημάτων, λαμβάνοντας υπόψη τις χρονικές, χωρικές και ιεραρχικές πτυχές του προβλήματος.

2.6 Έρευνα Βασισμένη στον Αλγόριθμο XGBoost

Τίτλος: Short-Term Traffic Flow Prediction Based on XGBoost [8]

Εισαγωγή:

Στην παρούσα εργασία, ως ροή κυκλοφορίας ορίζεται η κίνηση των οχημάτων εντός μιας συγκεκριμένης λωρίδας κυκλοφορίας στο οδικό δίκτυο σε μια δεδομένη χρονική στιγμή. Η κυκλοφοριακή ροή μετριέται με βάση παραμέτρους όπως ο αριθμός των διερχόμενων οχημάτων, ο μέσος όρος ταχύτητας και η πληρότητα (το ποσοστό του χρόνου κατά τον οποίο η λωρίδα είναι κατειλημμένη από οχήματα). Αυτές οι παράμετροι παρέχουν πληροφορίες σχετικά με τον όγκο και τα χαρακτηριστικά της κίνησης των οχημάτων και βοηθούν στην ανάλυση και την πρόβλεψη των μοτίβων κυκλοφορίας.

Μεθοδολογία:

Η έρευνα αυτή συνδυάζει την τεχνική της αποσύνθεσης κυματιδίων (wavelet decomposition) και τον αλγόριθμο XGBoost, θέτοντας ως στόχο τη βραχυπρόθεσμη πρόβλεψη της κυκλοφοριακής ροής. Η αποσύνθεση κυματιδίων χρησιμοποιείται για την εξαγωγή επιπρόσθετης πληροφορίας από το προς πρόβλεψη χαρακτηριστικό, ενώ η ανακατασκευή του συνδυάζει τις πληροφορίες χαμηλής και υψηλής συχνότητας που παρήχθησαν για τη δημιουργία του τελικού χαρακτηριστικού που θα χρησιμοποιηθεί στα δεδομένα εκπαίδευσης του αλγορίθμου.

Στη συνέχεια, αφού τα δεδομένα χωριστούν σε σύνολα εκπαίδευσης και ελέγχου, ο αλγόριθμος XGBoost μαθαίνει τα μοτίβα και τις σχέσεις στα δεδομένα εκπαίδευσης. Έπειτα, χρησιμοποιούνται δύο μετρικές για την αξιολόγηση του προτεινόμενου μοντέλου στα δεδομένα ελέγχου, το μέσο τετραγωνικό σφάλμα (RMSE) και το μέσο απόλυτο ποσοστιαίο σφάλμα (MAPE).

Οι προβλέψεις που παράγει το προτεινόμενο μοντέλο συγκρίνονται και με άλλες μεθόδους μηχανικής μάθησης, συμπεριλαμβανομένων των αλγορίθμων (Support Vector Machine) SVM και XGBoost χωρίς την μέθοδο της αποσύνθεσης κυματιδίων. Τα αποτελέσματα δείχνουν ότι η προτεινόμενη μέθοδος επιτυγχάνει χαμηλότερο RMSE και MAPE σε σύγκριση με τα υπόλοιπα μοντέλα.

3. Δεδομένα

Η κυκλοφοριακή ροή είναι ένα μέγεθος που δεν χαρακτηρίζεται από γραμμικότητα. Αυτό σημαίνει ότι ο αριθμός των οχημάτων που διέρχονται από ένα συγκεκριμένο σημείο ενός οδικού δικτύου κατά μια δεδομένη χρονική στιγμή δεν αυξάνεται ή μειώνεται σε σταθερό ρυθμό, αλλά μπορεί να διακυμαίνεται εξαιτίας διαφόρων παραγόντων, όπως ο χρόνος, οι καιρικές συνθήκες και τα τροχαία ατυχήματα. Επομένως, η χρήση κατάλληλων δεδομένων είναι κρίσιμη για την κατανόηση και την πρόβλεψη της κυκλοφοριακής ροής. Για τους σκοπούς της έρευνας, έχουμε εκμεταλλευτεί δύο σύνολα δεδομένων:

- τα **δεδομένα τροχιών** (trajectory data): αναφέρονται στις πληροφορίες που καταγράφουν τις θέσεις των κινούμενων αντικειμένων σε όλη τη διάρκεια του χρονικού διαστήματος που μας ενδιαφέρει.
- τα **δεδομένα καιρού** (weather data): αφορούν μετρήσεις και πληροφορίες σχετικά με τις καιρικές συνθήκες σε μια συγκεκριμένη περιοχή. Η καταγραφή και η ανάλυση των δεδομένων καιρού μπορεί να προσφέρει πληροφορίες σχετικές με τις συνθήκες που επηρεάζουν την κίνηση των οχημάτων, καθώς και την ασφάλεια και την απόδοσή τους στους δρόμους.

Και τα δύο προαναφερθέντα είδη δεδομένων συνδυάζονται συχνά με σκοπό την καλύτερη κατανόηση της κυκλοφοριακής ροής και της συμπεριφοράς των οχημάτων υπό διάφορες καιρικές συνθήκες.

3.1 Δεδομένα Τροχιών

Για να εξασφαλιστεί η υψηλή ποιότητα της έρευνας, έχουμε προσπαθήσει να αναζητήσουμε ένα σύνολο δεδομένων θέσεων GPS εντός μίας περιοχής, το οποίο ικανοποιεί ορισμένες αυστηρές προδιαγραφές. Συγκεκριμένα, κατά την επιλογή του ιδανικού συνόλου δεδομένων, πρέπει να συμπεριληφθούν υπόψιν τα ακόλουθα:

- πρώτο, περιέχει δεδομένα σημείων κινούμενων αντικειμένων GPS που καλύπτουν ένα σχετικά ευρύ γεωγραφικό χώρο (για παράδειγμα τον χώρο μίας ολόκληρης πόλης).

- δεύτερο, στο σύνολο δεδομένων είναι σημαντικό να υπάρχουν καταγραφές θέσεων GPS πολλών διαφορετικών κινούμενων αντικειμένων, ώστε η μελέτη να είναι όσο το δυνατόν ακριβέστερη και πληρέστερη.
- τρίτο, τα συνεχόμενα δεδομένα θέσεων GPS για κάθε κινούμενο όχημα πρέπει να καταγράφονται εντός ενός μικρού χρονικού διαστήματος το ένα με το άλλο.
- τέταρτο, ο χρόνος καταγραφής για κάθε σημείο GPS πρέπει να δίνεται ως πληροφορία στο σύνολο δεδομένων.

Ένα σύνολο δεδομένων που καλύπτει όλες αυτές τις προϋποθέσεις είναι γνωστό ως Cab Mobility Traces [9]. Αυτό το σύνολο δεδομένων αποτελεί προϊόν της συνεργασίας μεταξύ του Exploratorium (το μουσείο επιστήμης, τέχνης και ανθρώπινης αντίληψης του San Francisco) και του καλλιτέχνη Scott Snibbe.

Το συγκεκριμένο σύνολο δεδομένων προσφέρει μια λεπτομερή εικόνα των μοτίβων κίνησης των κίτρινων ταξί στην πόλη του San Francisco. Το σύστημα δρόμων της πόλης, πάνω στο οποίο έχει γίνει η καταγραφή των δεδομένων, φαίνεται στην ακόλουθη εικόνα:

Εικόνα 3.1: Το οδικό δίκτυο της πόλης του San Francisco, California. Επάνω σε αυτό το δίκτυο κινούνται τα ταξί, των οποίων την κίνηση μελετάμε.

Τα δεδομένα έχουν συγκεντρωθεί μέσω ενός καινοτόμου συστήματος παρακολούθησης GPS που έχει ενσωματωθεί σε κάθε κίτρινο ταξί της πόλης. Αυτό το σύστημα μεταδίδει δεδομένα πραγματικού χρόνου, περιλαμβάνοντας τον αριθμό ταυτοποίησης του ταξί, τις γεωγραφικές του συντεταγμένες (γεωγραφικό πλάτος και γεωγραφικό μήκος) εκείνη τη χρονική στιγμή, τον χρόνο που γίνεται η καταγραφή και την κατάσταση του ταξί, δηλαδή εάν εκτελεί δρομολόγιο ή όχι εκείνη τη χρονική στιγμή. Όλη αυτή η πληροφορία συγκεντρώνεται σε ένα κεντρικό διακομιστή.

Στο σύνολο δεδομένων Cab Mobility Traces καταγράφεται η κίνηση των ταξί κατά τον μήνα Μάιο του έτους 2008. Η πορεία κάθε ταξί ενσωματώνεται σε ένα ξεχωριστό αρχείο που φέρει ως όνομα το αναγνωριστικό του ταξί.

Το παρών σύνολο δεδομένων αντιπροσωπεύει μια επιστημονικά σημαντική πηγή, παρέχοντας σε ερευνητές και ακαδημαϊκούς μια μοναδική ευκαιρία να εξερευνήσουν τα μοτίβα

κίνησης στην αστική περιοχή του San Francisco και την αλληλεπίδραση των συστημάτων μεταφοράς με το αστικό περιβάλλον.

3.2 Δεδομένα Καιρού

Η συμπερίληψη δεδομένων καιρού είναι απαραίτητη κατά την πρόβλεψη της ροής κυκλοφορίας σε ένα οδικό δίκτυο, καθώς οι καιρικές συνθήκες έχουν έντονη επίδραση στη συμπεριφορά των οχημάτων και, κατά συνέπεια, στην κυκλοφοριακή κατάσταση των οδών. Ο καιρός μπορεί να επηρεάσει την ταχύτητα, την ορατότητα, την πρόσφυση των ελαστικών, τη συμπεριφορά των οδηγών και την κυκλοφοριακή ροή.

Επομένως, για να εξασφαλιστούν ακόμα καλύτερα αποτελέσματα, έχουν συμπεριληφθεί δεδομένα καιρού της πόλης του San Francisco κατά τον μήνα Μάιο του έτους 2008 [10]. Επιπλέον, τα δεδομένα καιρού που χρησιμοποιούνται έχουν καταγραφεί σε ωριαία βάση. Τα πιο σημαντικά στοιχεία που συμπεριλαμβάνονται σε αυτά είναι τα ακόλουθα:

- **θερμοκρασία:** η θερμοκρασία μπορεί να επηρεάσει την ταχύτητα των οχημάτων, καθώς και την ορατότητα.
- **υγρασία:** η υγρασία μπορεί να επηρεάσει την πρόσφυση των ελαστικών στον δρόμο και την ασφάλεια της οδήγησης.
- **ταχύτητα του ανέμου:** η ταχύτητα του ανέμου επηρεάζει την συμπεριφορά των οχημάτων και την σταθερότητα της κίνησης.
- **ατμοσφαιρική πίεση:** η ατμοσφαιρική πίεση μπορεί να προσφέρει πληροφορίες σχετικά με τις αλλαγές στις καιρικές συνθήκες και την πιθανή επίδρασή τους στην κυκλοφορία.
- **ορατότητα:** η ορατότητα είναι κρίσιμη για την ασφάλεια της οδήγησης και μπορεί να επηρεάσει την ταχύτητα και τον τρόπο κίνησης.

Αξίζει να σημειωθεί ότι η ανάλυση αυτών των δεδομένων παρέχει πολύτιμες πληροφορίες για τον τρόπο, με τον οποίο οι καιρικές συνθήκες επηρεάζουν την κυκλοφοριακή ροή. Η σύνδεση των δεδομένων καιρού με τα δεδομένα κίνησης μπορεί να αποκαλύψει ποιες συνθήκες οδήγησης έχουν τη μεγαλύτερη επίδραση στην κίνηση των οχημάτων, προσφέροντας ένα πληρέστερο και πιο περιεκτικό πλαίσιο για την κατανόηση των παραγόντων που επηρεάζουν την κυκλοφορία.

4. Τεχνολογίες

Το συγκεκριμένο κεφάλαιο έχει δημιουργηθεί για να παραθέσει περισσότερες πληροφορίες σχετικά με τις τεχνολογίες που υιοθετήθηκαν για την διεκπεραίωση της παρούσας έρευνας. Πιο συγκεκριμένα, αναλύονται πληροφορίες για την γλώσσα προγραμματισμού Python που χρησιμοποιήθηκε. Αναφέρονται, επίσης, πληροφορίες για τα περιβάλλοντα εκτέλεσης του κώδικα, δηλαδή τα Jupyter Notebook και Google Colab. Περιγράφεται, επιπλέον, ο τρόπος λειτουργίας της βάσης δεδομένων PostgreSQL και της γλώσσας PL/pgSQL που αξιοποιήθηκαν στην έρευνα μόνο για συγκριτικούς σκοπούς. Τέλος, αναλύεται ο τρόπος λειτουργίας των λογισμικών Docker και Valhalla, η σύνθεση των οποίων κατέστη εφικτή την διαδικασία αντιστοίχισης των τροχιών των κινούμενων αντικειμένων επάνω στο οδικό δίκτυο.

4.1 Η Γλώσσα Python

Η Python είναι μια γλώσσα προγραμματισμού υψηλού επιπέδου που δημιουργήθηκε από τον Guido van Rossum και πρωτοκυκλοφόρησε το 1991 [11]. Έχει κερδίσει ευρεία αναγνώριση χάρη στην απλή και ευανάγνωστη σύνταξή της, καθώς και στην ευελιξία και ισχυρή κοινότητα που την υποστηρίζει. Μερικά από τα βασικά χαρακτηριστικά της Python περιλαμβάνουν:

- **κατανοητή σύνταξη**: η Python χρησιμοποιεί απλή και ευανάγνωστη σύνταξη, που διευκολύνει την ανάπτυξη και τη συντήρηση του κώδικα.
- **διερμηνευμένη φύση**: η διερμηνευμένη φύση της Python αναφέρεται στο γεγονός ότι ο κώδικας δεν χρειάζεται να μεταγλωττιστεί προκαταβολικά σε γλώσσα μηχανής πριν από

την εκτέλεσή του. Αντί αυτού, ο κώδικας εκτελείται απευθείας από τον διερμηνέα της Python κατά τη διάρκεια της εκτέλεσης του προγράμματος.

- **δομημένη οργάνωση**: υποστηρίζεται η οργάνωση του κώδικα σε μονάδες, όπως οι συναρτήσεις και οι κλάσεις. Επομένως, η συγκεκριμένη γλώσσα συνδυάζει τα χαρακτηριστικά του διαδικαστικού και του αντικειμενοστραφούς προγραμματισμού.
- **δυναμική δήλωση τύπων**: δεν απαιτείται να δηλώνεται εξαρχής ο τύπος μιας μεταβλητής, καθώς ο τελευταίος αναγνωρίζεται αυτόματα κατά την εκτέλεση του προγράμματος.
- **υποστήριξη από πολυάριθμες βιβλιοθήκες**: η Python παρέχει μια πλούσια συλλογή από ενσωματωμένες βιβλιοθήκες για διάφορες εργασίες που υποστηρίζουν από ανάλυση δεδομένων μέχρι γραφικά και σχεδίαση ιστοσελίδων.

Η Python υποστηρίζει επίσης πολλούς τύπους δεδομένων, όπως ακέραιους αριθμούς, πραγματικούς αριθμούς, σύνθετους αριθμούς, συμβολοσειρές, λίστες, πλειάδες και λεξικά. Κάθε τύπος χρησιμοποιείται σε διαφορετικές περιστάσεις και συχνά συνδυάζονται, για να δημιουργήσουν πολύπλοκες δομές και λειτουργίες στα προγράμματά.

Συνολικά, η Python είναι μια απλή και ισχυρή γλώσσα προγραμματισμού με πολλές δυνατότητες. Για αυτό τον λόγο, αποτελεί μια εξαιρετική επιλογή τόσο για αρχάριους όσο και για προχωρημένους προγραμματιστές. Δεν είναι λοιπόν παράλογο το γεγονός ότι η συγκεκριμένη γλώσσα χρησιμοποιείται εκτεταμένα σε πολλούς τομείς, όπως είναι η ανάπτυξη λογισμικού, οι επιστημονικοί υπολογισμοί, η ανάλυση δεδομένων, η κατασκευή ιστοσελίδων, η τεχνητή νοημοσύνη και πολλοί άλλοι.

4.1.1 Python και Επιστήμη Δεδομένων

Στην παρούσα έρευνα, εκμεταλλευόμαστε τη γλώσσα προγραμματισμού Python για την ανάλυση και επίλυση ενός προβλήματος που ανήκει στο πεδίο της μηχανικής μάθησης, δηλαδή την πρόβλεψη της κυκλοφοριακής ροής σε ένα οδικό δίκτυο με αλγορίθμους που εκπαιδεύονται από δεδομένα. Στο πλαίσιο της επιστήμης των δεδομένων, η Python αποτελεί ένα ευρέως χρησιμοποιούμενο εργαλείο, υποστηρίζοντας τις δραστηριότητες ανάλυσης, εξόρυξης και ερμηνείας δεδομένων, προκειμένου να αντληθούν πληροφορίες και γνώση από αυτά. Οι πιο συνήθεις τρόποι χρήσης της Python στο πεδίο της επιστήμης των δεδομένων περιλαμβάνουν:

- **ανάλυση και εξόρυξη δεδομένων**: η Python παρέχει βιβλιοθήκες, όπως η Pandas [12] και η Numerical Python (NumPy) [13] για την ανάλυση και τον χειρισμό των δεδομένων, ενώ η βιβλιοθήκη Scikit-learn [14] προσφέρει μία ποικιλία από αλγορίθμους μηχανικής μάθησης προκειμένου να εξοριστεί γνώση από αυτά.
- **μηχανική μάθηση**: η Python αποτελεί συνήθως την πρώτη επιλογή για την ανάπτυξη μοντέλων μηχανικής μάθησης με τη χρήση βιβλιοθηκών όπως η Scikit-learn, το TensorFlow [15] και το Keras [16]. Αυτές οι βιβλιοθήκες επιτρέπουν τη δημιουργία μοντέλων πρόβλεψης, ταξινόμησης, συσταδοποίησης και πολλών άλλων.
- **οπτικοποίηση δεδομένων**: η Python παρέχει βιβλιοθήκες όπως η Matplotlib [17] και η Seaborn [18] για τη δημιουργία γραφημάτων που βοηθούν στην οπτικοποίηση των δεδομένων, ενισχύοντας την κατανόηση τους.

Επομένως, η ευελιξία της Python, μαζί με τις πλούσιες βιβλιοθήκες που προσφέρει, καθιστούν αυτήν τη γλώσσα ένα ισχυρό εργαλείο για την ανάλυση και την εξόρυξη δεδομένων. Μάλιστα, αυτό αποτέλεσε και το βασικό κίνητρο επιλογής της συγκεκριμένης γλώσσας στην έρευνά μας.

4.2 Jupyter Notebook και Google Colab

Όσον αφορά το προγραμματιστικό περιβάλλον που χρησιμοποιήθηκε, η παρούσα έρευνα που έχουμε διεξάγει έχει αναπτυχθεί κυρίως στα περιβάλλοντα του Jupyter Notebook και του Google Colab (ο όρος Colab προέρχεται από τη λέξη Collaboratory). Τόσο το Jupyter Notebook, όσο και το Google Colab αποτελούν δημοφιλή περιβάλλοντα προγραμματισμού, τα οποία

χρησιμοποιούνται ευρέως για την ανάπτυξη και εκτέλεση κώδικα Python, καθώς και για την οπτικοποίηση και ανάλυση δεδομένων. Επομένως, έχουν εξαιρετική απήχηση στον χώρο της επιστήμης των δεδομένων και της μηχανικής μάθησης.

4.2.1 Jupyter Notebook

Το Jupyter Notebook [19] αντιπροσωπεύει ένα περιβάλλον προγραμματισμού που επιτρέπει τη δημιουργία κώδικα Python και τον κοινό διαμοιρασμό εγγράφων που περιλαμβάνουν εκτελέσιμο κώδικα, κείμενο, εικόνες, γραφήματα και άλλα, με άλλους χρήστες.

Αναλυτικότερα, το Jupyter Notebook, όταν εγκατασταθεί τοπικά, λειτουργεί σε έναν τοπικό διακομιστή. Κατά την εκτέλεσή του, το γραφικό περιβάλλον της εφαρμογής εμφανίζεται μέσω ενός προγράμματος περιήγησης σε μια προκαθορισμένη διεύθυνση IP, συνήθως στην http://localhost:8888. Επιπρόσθετα, το κλείσιμο του προγράμματος περιήγησης δεν συνεπάγεται και την αποσύνδεση από τον διακομιστή που φιλοξενεί το Jupyter Notebook.

Όσον αφορά την σχεδίαση του Jupyter Notebook, αυτό είναι οργανωμένο σε σημειωματάρια, όπως βέβαια δηλώνει και το όνομα του προγράμματος. Κάθε σημειωματάριο είναι οργανωμένο σε εκτελέσιμα κελιά. Ένα κελί μπορεί να περιλαμβάνει διάφορα είδη πληροφοριών, όπως κώδικα Python ή άλλους τύπους πληροφορίας, για παράδειγμα κείμενο. Οι χρήστες έχουν τη δυνατότητα να εκτελούν κάθε ένα αυτά τα κελιά, ξεχωριστά το ένα με το άλλο, και να παρατηρούν άμεσα τα αποτελέσματα. Το Jupyter Notebook αποτέλεσε το βασικό εργαλείο συγγραφής κώδικα κατά την διάρκεια της έρευνας, λόγω της ευχρηστότητάς του.

4.2.2 Google Colab

Το Google Colab [20] αντιπροσωπεύει μια υπηρεσία που προσφέρεται από την Google και λειτουργεί παρόμοια με το Jupyter Notebook. Η μόνη διαφορά είναι ότι η εφαρμογή αυτή παρέχει ένα περιβάλλον προγραμματισμού που εκτελείται στο υπολογιστικό νέφος της Google προσφέροντας στον τελικό χρήστη δωρεάν ή προς πληρωμή πόρους, όπως επεξεργαστική ισχύ και μνήμη για την εκτέλεση κώδικα. Το Colab αποδεικνύεται ιδιαίτερα χρήσιμο για την εκπαίδευση μοντέλων μηχανικής μάθησης, αφού παρέχει πρόσβαση σε βιβλιοθήκες όπως το TensorFlow και το Keras. Τέλος, το Google Colab δίνει τη δυνατότητα της εύκολης κοινής χρήσης των σημειωματάριων με άλλους χρήστες.

Στην εν λόγω έρευνα, εκμεταλλευτήκαμε αρκετά τις κάρτες γραφικών που προσφέρει η υπηρεσία, ώστε να εκτελέσουμε τον τελικό κώδικα και να εκπαιδεύσουμε τα μοντέλα μηχανικής μάθησης που αναπτύχθηκαν σε αρκετά μικρότερο χρόνο.

4.3 PostgreSQL και PLpgSQL [2]

Το σύστημα διαχείρισης βάσεων δεδομένων της PostgreSQL είναι μία τεχνολογία, η οποία χρησιμοποιήθηκε στην έρευνά μόνο για συγκριτικούς σκοπούς. Περισσότερες πληροφορίες για την χρήση της δίνονται στο επόμενο κεφάλαιο. Λόγω του γεγονότος ότι η χρήση της είναι περιορισμένη, παρουσιάζονται τα πιο σημαντικά στοιχεία που αφορούν αυτήν την τεχνολογία.

4.3.1 Γενικά Χαρακτηριστικά για την Βάση Δεδομένων

Η PostgreSQL, γνωστή και ως Postgres, αποτελεί ένα προηγμένο σχεσιακό σύστημα διαχείρισης βάσεων δεδομένων (ΣΔΒΔ). Αυτό το ΣΔΒΔ είναι εύκολα επεκτάσιμο μέσω διάφορων προσθέτων, πολλά από τα οποία είναι ανοιχτού κώδικα και διαθέσιμα δωρεάν μέσω του διαδικτύου. Επομένως, πρόκειται για μία τεχνολογία με πολλές δυνατότητες.

Σε ένα σχεσιακό ΣΔΒΔ, τα δεδομένα αποθηκεύονται σε πίνακες με στήλες και γραμμές. Κάθε στήλη αντιπροσωπεύει έναν συγκεκριμένο τύπο δεδομένων, ενώ κάθε γραμμή αντιπροσωπεύει μια εγγραφή με συγκεκριμένες τιμές για κάθε στήλη. Το σχεσιακό μοντέλο επιτρέπει τη δημιουργία συσχετίσεων μεταξύ των πινάκων, δημιουργώντας ένα περίπλοκο δίκτυο συνδέσεων για την ανάκτηση σχετικών δεδομένων. Ένα σχεσιακό ΣΔΒΔ προσφέρει πολλά πλεονεκτήματα, όπως:

- **δομημένη οργάνωση**: τα δεδομένα οργανώνονται σε πίνακες ή σχέσεις, πράγμα που καθιστά εύκολη την οργάνωση και την ανάκτησή τους.
- **ευέλικτη ανάκτηση**: οι χρήστες μπορούν να κάνουν πολύπλοκα ερωτήματα με βάση διάφορα κριτήρια και συνθήκες.
- **κοινόχρηστη πρόσβαση**: πολλοί χρήστες μπορούν να έχουν πρόσβαση στα ίδια δεδομένα ή στην ίδια βάση δεδομένων ταυτόχρονα.
- **αποκλεισμός ανωμαλιών**: το ΣΔΒΔ παρέχει μηχανισμούς αποκλεισμού ανωμαλιών και ανάκαμψης μετά από κάποιο σφάλμα.

Ένα σημαντικό χαρακτηριστικό που υποστηρίζεται από την PostgreSQL είναι τα ευρετήρια (indexes). Τα ευρετήρια βελτιστοποιούν την απόδοση και την αναζήτηση δεδομένων στη βάση επιτρέποντας γρήγορη πρόσβαση σε συγκεκριμένες εγγραφές μίας σχέσης, μειώνοντας έτσι τον χρόνο αναζήτησης και ανάκτησης των δεδομένων. Η PostgreSQL υποστηρίζει διάφορους τύπους ευρετηρίων που εξυπηρετούν διάφορες ανάγκες. Τα πιο σημαντικά και ευρέως χρησιμοποιούμενα ευρετήρια στην postgres είναι τα ευρετήρια B+ δέντρων και τα ευρετήρια κατακερματισμού.

Τέλος, η PostgreSQL υποστηρίζει διεπαφές με πολλές γλώσσες προγραμματισμού, όπως είναι οι Java, Python, C, C++, PHP, C#, επιτρέποντας στις εφαρμογές να αλληλοεπιδρούν άμεσα με τη βάση δεδομένων.

4.3.2 Η Γλώσσα PL/pgSQL [21]

Η γλώσσα προγραμματισμού PL/pgSQL (από το Procedural Language/PostgreSQL) αποτελεί μια πολύτιμη επέκταση του διαχειριστικού συστήματος βάσεων δεδομένων της PostgreSQL. Πιο αναλυτικά, τα κυριότερα χαρακτηριστικά της PL/pgSQL συνοψίζονται παρακάτω:

- **δημιουργία συναρτήσεων και διαδικασιών**: με την PL/pgSQL, μπορούν να δημιουργηθούν συναρτήσεις που επιστρέφουν τιμές και διαδικασίες που εκτελούν ενέργειες χωρίς επιστροφή τιμών. Αυτό επιτρέπει την οργάνωση εντός της βάσης δεδομένων.
- **ενσωμάτωση με την PostgreSQL**: ένα από τα πλεονεκτήματα της PL/pgSQL είναι ότι είναι πλήρως ενσωματωμένη με το σύστημα διαχείρισης βάσεων δεδομένων της PostgreSQL. Αυτό σημαίνει ότι οι διαδικασίες και οι συναρτήσεις που γράφονται σε γλώσσα PL/pgSQL είναι άμεσα εκτελέσιμες από τον ίδιο τον εξυπηρετητή της βάσης. Με τη δυνατότητα εκτέλεσης πολύπλοκων λειτουργιών απευθείας στη βάση δεδομένων, εξοικονομούνται πόροι, βελτιώνοντας την απόδοση. Η PL/pgSQL αναδεικνύεται ως ένα απαραίτητο εργαλείο για τη δημιουργία πολυπλοκότερων και αποτελεσματικότερων διαχειριστικών λειτουργιών εντός του περιβάλλοντος της PostgreSQL.
- **εύκολη ανάπτυξη κώδικα**: η σύνταξη της PL/pgSQL είναι παρόμοια με αυτή της κλασικής SQL, επιτρέποντας στους προγραμματιστές και τους διαχειριστές βάσεων δεδομένων να εξοικειωθούν γρήγορα με τη νέα γλώσσα.
- **δήλωση μεταβλητών και σταθερών**: ένα από τα σημαντικότερα χαρακτηριστικά της PL/pgSQL είναι η δυνατότητα δήλωσης μεταβλητών για την αποθήκευση δεδομένων κατά τη διάρκεια της εκτέλεσης προγραμμάτων, καθώς και η δυνατότητα αποθήκευσης σταθερών τιμών, οι τιμές των οποίων παραμένουν σταθερές κατά τη διάρκεια της εκτέλεσης του κώδικα.
- **έλεγχος ροής**: η PL/pgSQL παρέχει δομές ακολουθίας, επιλογής και επανάληψης, οι οποίες επιτρέπουν τη διαχείριση της ροής εκτέλεσης του προγράμματος βάσει συνθηκών και κριτηρίων.
- **παραμετροποίηση**: η δυνατότητα χρήσης παραμέτρων και μεταβλητών στην PL/pgSQL παρέχει ευελιξία κατά την εκτέλεση των διαδικασιών, επιτρέποντας την προσαρμογή της συμπεριφοράς του προγράμματος βάσει δυναμικών παραμέτρων.
- **αναγνώριση σφαλμάτων**: σε περίπτωση εμφάνισης σφάλματος, η PL/pgSQL παρέχει σαφή μηνύματα σφάλματος προς τον προγραμματιστή, διευκολύνοντας τον εντοπισμό και τη διόρθωσή τους.

Συνολικά, η γλώσσα PL/pgSQL αντιπροσωπεύει μια ισχυρή εργαλειοθήκη που συμβάλλει στην επέκταση των δυνατοτήτων της PostgreSQL, καθιστώντας τη ένα περιβάλλον κατάλληλο όχι μόνο για την αποθήκευση δεδομένων, αλλά και για την εκτέλεση πολυπλοκότερων εργασιών και λογικών σεναρίων. Ουσιαστικά, η PL/pgSQL επιτρέπει στους χρήστες να εκμεταλλεύονται την ισχύ μίας γλώσσας προγραμματισμού, για να διαμορφώσουν τη βάση δεδομένων όπως αυτοί επιθυμούν.

4.4 Το Λογισμικό Docker

Το λογισμικό Docker [22] αποτελεί μια εξαιρετικά σημαντική τεχνολογία, η οποία συνδέεται στενά με τον τρόπο που αναπτύσσονται, μεταφέρονται και εκτελούνται εφαρμογές και λογισμικά. Το κύριο πλεονέκτημα του Docker είναι ότι επιτρέπει τη δημιουργία ενός «κουτιού εκτέλεσης» (container), το οποίο μπορεί να εκτελεστεί σε οποιοδήποτε σύστημα και λειτουργικό, εξαλείφοντας έτσι τις ανησυχίες περί συμβατότητας. Αναλυτικότερα, κάθε κουτί εκτέλεσης περιλαμβάνει:

- **εκτελέσιμο αρχείο:** το εκτελέσιμο αρχείο της εφαρμογής.
- **βιβλιοθήκες:** οι βιβλιοθήκες και οι εξαρτήσεις που απαιτούνται για τη λειτουργία της εφαρμογής. Αυτές χρησιμοποιούνται κατά τη διάρκεια της εκτέλεσης.
- **ρυθμίσεις περιβάλλοντος:** είναι τα πρότυπα που καθορίζουν πώς η εφαρμογή πρέπει να εκτελείται, όπως είναι οι μεταβλητές περιβάλλοντος και οι παράμετροι εκτέλεσης.
- **εκτελέσιμες εντολές:** οι εκτελέσιμες εντολές αναφέρονται στις ενέργειες που πρέπει να ληφθούν για να ξεκινήσει η εφαρμογή εντός του container.
- **δεδομένα εφαρμογής:** πρόκειται για οποιαδήποτε δεδομένα ή αρχεία που χρειάζεται η εφαρμογή για να λειτουργήσει σωστά.

Η αξία του Docker έγκειται στην ευκολία ανάπτυξης, στην ευελιξία και στην αξιοπιστία εκτέλεσης των εφαρμογών. Όσον αφορά την επιρροή του Docker στον παγκόσμιο χώρο, η τεχνολογία αυτή έχει υιοθετηθεί από πολυάριθμους χρήστες και επιχειρήσεις, διαμορφώνοντας έτσι νέα πρότυπα για την αποτελεσματική και αξιόπιστη διαχείριση των εφαρμογών.

4.5 Το Λογισμικό Valhalla

Το λογισμικό Valhalla [23] αποτελεί ένα πρόγραμμα ανοιχτού κώδικα που αναπτύχθηκε αρχικά από την εταιρεία Mapzen. Η συγκεκριμένη εταιρία ιδρύθηκε το 2013 και ειδικευόταν στις υπηρεσίες χαρτογράφησης γεωχωρικών δεδομένων. Η εφαρμογή έχει ως κύριο στόχο να εκτελεί υπηρεσίες χαρτογράφησης GPS δεδομένων, προσφέροντας ακριβείς δρομολογήσεις για αυτοκίνητα, ποδήλατα και άλλα μέσα μεταφοράς, καθώς επίσης και τη δυνατότητα αντιστοίχισης τροχιών GPS σε ψηφιακούς χάρτες.

Η διαδικασία ανάπτυξης του Valhalla εκκινείται το 2014 και ολοκληρώνεται το 2019, όταν η εταιρεία Mapzen αποσύρεται από τον επιχειρηματικό τομέα. Μετά το κλείσιμο της Mapzen, το Valhalla και τα σχετικά δεδομένα που το απαρτίζουν μεταφέρονται στον Οργανισμό Ελεύθερου Λογισμικού (Software Freedom Conservancy) για τη διαχείριση, ανάπτυξη και συντήρηση του λογισμικού Valhalla.

Στον πυρήνα του Valhalla είναι συγκεντρωμένοι αλγόριθμοι υπεύθυνοι για την επεξεργασία των γεωχωρικών δεδομένων, όπως η δρομολόγηση και η αντιστοίχιση τροχιών σε χάρτες, εξυπηρετώντας έτσι προγραμματιστές και χρήστες που ψάχνουν μία τέτοια υπηρεσία. Επιπρόσθετα, οι αλγόριθμοι είναι σχεδιασμένοι με τέτοιο τρόπο, ώστε να προσδίδουν ακρίβεια και αξιοπιστία στα αποτελέσματα που παράγουν.

Το λογισμικό Valhalla είναι συμβατό με διάφορα λειτουργικά συστήματα, όπως το Linux, το macOS και τα Windows. Επιπλέον, το τελευταίο είναι προσαρμόσιμο για εκτέλεση τόσο σε τοπικούς υπολογιστές, όσο και σε απομακρυσμένους διακομιστές (servers). Τέλος, λόγω του γεγονότος ότι το εν λόγω λογισμικό είναι ανοιχτού κώδικα, σημαίνει ότι διανέμεται και δωρεάν στους ενδιαφερόμενους χρήστες.

4.5.1 Η Υπηρεσία Valhalla Meili

Το Valhalla Meili αποτελεί ένα υποσύνολο του ευρύτερου project Valhalla και αναλαμβάνει να εκτελέσει μία συγκεκριμένη λειτουργία χαρτογράφησης. Η κύρια διαφορά μεταξύ του Valhalla Meili και του Valhalla είναι ότι το πρώτο είναι υπεύθυνο για την αντιστοίχιση τροχιών GPS σε ένα ψηφιακό χάρτη, ενώ το δεύτερο παρέχει στο χρήστη ένα ευρύτερο φάσμα λειτουργιών χαρτογράφησης, όπως η αναζήτηση της συντομότερης διαδρομής μεταξύ δύο σημείων.

Συγκεκριμένα, το Valhalla Meili αναλαμβάνει την λήψη της πορείας μιας παρατηρηθείσας τροχιάς αποτελούμενη από διαδοχικά GPS σημεία (ζεύγη από γεωγραφικά μήκη και γεωγραφικά πλάτη) και δίνει ως αποτέλεσμα μια πιθανή αντιστοιχισμένη διαδρομή στον ψηφιακό χάρτη.

Στην μελέτη μας, χρησιμοποιούμε ως ψηφιακό χάρτη τα δεδομένα από το Open Street Map (OSM) της περιοχής του San Francisco, California. Το OSM [24] είναι μια δωρεάν, δημόσια βάση γεωγραφικών δεδομένων που ενημερώνεται και συντηρείται συνεχώς από μια κοινότητα εθελοντών. Μέσα σε αυτή την βάση αποθηκεύονται πληροφορίες σχετικά με το οδικό δίκτυο μίας περιοχής (για παράδειγμα δρόμοι, διασταυρώσεις, εθνικές οδοί, πεζοδρόμια κ.α.) και την τοπολογία μίας περιοχής (παραδείγματος χάριν βουνά, λίμνες, ποτάμια, πεδιάδες, θάλασσες κ.α.). Το OpenStreetMap χρησιμοποιείται ευρέως για την εξαγωγή ηλεκτρονικών χαρτών σε μορφή αρχείων και την οπτικοποίηση γεωχωρικών δεδομένων, καθώς διατίθεται ελεύθερα υπό την άδεια της ανοιχτής βάσης δεδομένων.

Στην διερεύνηση που έχουμε διεξάγει, χρησιμοποιείται η υπηρεσία Valhalla Meili σε συνδυασμό με το λογισμικό Docker. Ουσιαστικά, έχουμε εγκαταστήσει το Valhalla Meili σε ένα κουτί (container) του Docker, μέσα στο οποίο το πρόγραμμα εκτελείται ανεξάρτητα από το λειτουργικό σύστημα και τους περιορισμούς του μηχανήματος που διαθέτουμε. Μετά την εγκατάσταση του Valhalla Meili, διαμορφώνεται ένας εξυπηρετητής που αναλαμβάνει να δέχεται αιτήματα αντιστοίχισης GPS σημείων και να επιστρέφει την αντιστοιχισμένη πληροφορία στον ψηφιακό χάρτη OSM που χρησιμοποιούμε.

Η απάντηση που λαμβάνουμε από τον εξυπηρετητή του Valhalla μετά από κάθε αίτημα αντιστοίχισης που στέλνουμε σε αυτόν είναι εμπλουτισμένη με χρήσιμες πληροφορίες. Οι πληροφορίες αυτές καθίστανται προσβάσιμες εξαιτίας της λειτουργίας trace attributes που διαθέτει το Valhalla Meili. Αυτό έχει ως αποτέλεσμα κάθε τροχιά να συνοδεύεται από επιπρόσθετα δεδομένα, όπως είναι η ταχύτητα του κινούμενου αντικειμένου, η κατάστασή του (π.χ. είναι σε κίνηση ή σε στάση), το είδος του κινούμενου οχήματος (π.χ. αυτοκίνητο, ποδήλατο, πεζός) κ.α.

Εξαιρετικά χρήσιμο χαρακτηριστικό της λειτουργίας αυτής αποτελεί επίσης και τη δυνατότητα να γνωρίζουμε το μοναδικό αναγνωριστικό της ακμής (Edge ID), στην οποία αντιστοιχίζεται το GPS σημείο που εξετάζουμε. Αυτά τα μοναδικά Edge IDs που παρέχει το Valhalla Meili δεν είναι τυχαία, αλλά προέρχονται απευθείας από τη βάση δεδομένων του Open Street Map. Αυτό σημαίνει ότι μπορούμε εύκολα να αναζητήσουμε την αντίστοιχη ακμή, απλά κάνοντας αναζήτηση στη βάση δεδομένων του OSM. Μία ακόμα λεπτομέρεια που αξίζει να αναφερθεί, είναι ότι κάθε Edge ID που υπάρχει στη βάση δεδομένων του OSM είναι μοναδικό.

Αντιστοιχίζοντας σημεία GPS με τα Edge IDs, ο αλγόριθμος μπορεί να ανακατασκευάσει ακριβώς τη διαδρομή που ακολούθησε το κινούμενο αντικείμενο, όπως ένα όχημα ή ένας χρήστης. Η διαδικασία αντιστοίχισης δεδομένων GPS που απαρτίζουν μία τροχιά στο οδικό δίκτυο περιλαμβάνει την εύρεση της ακολουθίας των Edge IDs με σωστή σειρά. Αυτή η ακολουθία των Edge IDs αντιστοιχίζεται όσο το δυνατόν καλύτερα στο οδικό δίκτυο δίνοντας ως αποτέλεσμα μια ακριβή ή αρκετά καλή αναπαράσταση της διαδρομής που διένυσε το κινούμενο αντικείμενο πάνω στον OSM χάρτη.

5. Υλοποίηση της Εφαρμογής

Στο κεφάλαιο αυτό παρέχεται όλη η απαραίτητη γνώση σχετικά με τον κώδικα που αναπτύχθηκε. Ταυτόχρονα, αναλύονται οι τεχνικές και οι βελτιστοποιήσεις που λήφθηκαν υπόψιν, ώστε να παραχθούν ακριβέστερα και ποιοτικότερα αποτελέσματα. Ο κώδικας έχει χωριστεί σε τέσσερα αρχεία, τα οποία φέρουν τους τίτλους «**Notebook1**», «**Notebook2**», «**Notebook3**» και «**Notebook4**»:

Στο πρώτο αρχείο γίνεται όλη η προεπεξεργασία των δεδομένων: η αντιστοίχιση των σημείων GPS στο οδικό δίκτυο του San Francisco και η αναγωγή του προβλήματος σε χρονοσειρές.

Στο δεύτερο αρχείο γίνεται κατανοητή η σημασία χρήση της μεθόδου των Αυστηρών Ερωτημάτων Μονοπατιού (ΑΕΜ) χρησιμοποιώντας γραφήματα.

Στο τρίτο αρχείο, εισάγονται τα δεδομένα καιρού, γίνονται οπτικοποιήσεις πάνω στα δεδομένα και συγκρίνονται τέσσερα μοντέλα μηχανικής μάθησης, όπου κάθε ένα από αυτά αποσκοπεί στην επίλυση του βασικού προβλήματος, δηλαδή της πρόβλεψης της κυκλοφοριακής ροής σε ολόκληρα μονοπάτια εντός του οδικού δικτύου του San Francisco.

Τέλος, στο τέταρτο αρχείο πραγματοποιούνται οι διαδικασίες των προβλέψεων του μεγέθους της κυκλοφοριακής ροής, κάνοντας χρήση του καλύτερου αλγορίθμου μηχανικής μάθησης που εκπαιδεύτηκε.

5.1 Προεπεξεργασία των Δεδομένων

Η αρχική φάση του κώδικα είναι αφιερωμένη στην συλλογή και την προετοιμασία του αρχικού συνόλου δεδομένων. Όσον αφορά την συλλογή των δεδομένων, λόγω του ότι αυτά βρίσκονται μοιρασμένα σε πολλά διαφορετικά αρχεία, προσπαθούμε να τα ενοποιήσουμε όλα μαζί σε ένα ενιαίο αρχείο. Το σύνολο δεδομένων που προκύπτει περιλαμβάνει όλες τις καταγραφές κίνησης των ταξί που κινούνται στην πόλη του San Francisco κατά τον μήνα Μάιο του έτους 2008 και είναι σχολαστικά δομημένο, ενσωματώνοντας χαρακτηριστικά όπως τα εξής:

- **Taxi ID:** το μοναδικό αναγνωριστικό του κάθε Ταξί.
- **Latitude:** το γεωγραφικό πλάτος της θέσης του ταξί όταν έγινε η καταγραφή.
- **Longitude:** το γεωγραφικό μήκος της θέσης του ταξί όταν έγινε η καταγραφή.
- **Occupied:** δηλώνει εάν το ταξί μετέφερε επιβάτη/επιβάτες ή όχι όταν έγινε η καταγραφή της θέσης του (είναι η μόνη πληροφορία που δεν θα χρειαστεί στην έρευνά).
- **Date Time:** η ημερομηνία και η ώρα που έγινε η καταγραφή. Είναι της μορφής χρονιά-μήνας-ημέρα ώρα:λεπτό:δευτερολεπτο.

Το σύνολο δεδομένων περιλαμβάνει περίπου έντεκα εκατομμύρια εγγραφές. Για να διευκολύνουμε την διαδικασία της έρευνάς, έχει περιοριστεί ο αριθμός των εγγραφών που χρησιμοποιούνται, απομονώνοντας τις εγγραφές που καταγράφηκαν κατά τη διάρκεια μιας εβδομάδας - συγκεκριμένα, από τις 18 έως τις και τις 25 Μαΐου. Μάλιστα, θεωρήθηκε φρόνιμο το γεγονός ότι η κυκλοφοριακή ροή είναι ένα μέγεθος που για να προβλεφθεί η συμπεριφορά του δεν χρειάζεται να ανατρέξουμε πολύ πίσω στον χρόνο.

Το συγκεκριμένο σύνολο δεδομένων είναι ιδανικό, καθώς περιέχει καταγραφές σημείων GPS ανά μικρά χρονικά διαστήματα. Ωστόσο, για κάθε διαφορετικό ταξί (δηλαδή για κάθε διαφορετικό «Taxi ID») υπάρχει μία συνεχόμενη ροή από καταγραφές θέσεων GPS για το χρονικό διάστημα ολόκληρου του μήνα Μάιου. Επομένως, για κάθε ταξί διατίθεται μία μονοκόμματη τροχιά. Ένα πρόβλημα που δημιουργείται είναι η διαδικασία διαχωρισμού της τροχιάς σε υποτροχιές, δηλαδή σε μικρότερες τροχιές της ίδιας μεγαλύτερης τροχιάς. Επιπλέον, θέλουμε σε κάθε υποτροχιά να περιλαμβάνονται σημεία GPS που απέχουν ανά δύο ένα συγκεκριμένο χρονικό διάστημα. Στην έρευνα ορίζεται ως μέγιστο χρονικό διάστημα τα ενενήντα δευτερόλεπτα. Με άλλα λόγια, κάθε τροχιά ενός ταξί διαμοιράζεται σε υποτροχιές που περιέχουν η κάθε μία διαδοχικά σημεία που απέχουν χρονικά ανά δύο έως και ενενήντα δευτερόλεπτα. Αυτή η διαδικασία ονομάζεται “Διαχωρισμός Τροχιάς”. Η προσέγγιση αυτή έρχεται να εισάγει μία καινούρια πληροφορία στο σύνολο δεδομένων, την στήλη «**Traj ID**» που δηλώνει το αναγνωριστικό της υποτροχιάς. Πλέον, κάθε μοναδικό ζεύγος (Taxi ID, Traj ID) ταυτοποιεί μοναδικά μία τροχιά. Μετά το πέρας αυτής της ενέργειας, το πλήθος των τροχιών που υπάρχει στο σύνολο δεδομένων αυξάνεται.

5.2 Αντιστοίχιση Τροχιών στο Οδικό Δίκτυο

Ένα επιπλέον πρόβλημα που προκύπτει με τα δεδομένα, είναι ότι οι τροχιές δεν έχουν αντιστοιχηθεί σε κάποιον ψηφιακό χάρτη. Αυτό οδηγεί στο ακόλουθο πρόβλημα: τα δεδομένα GPS ίσως να μην είναι ακριβή, αφού μπορεί να υπήρξε θόρυβος κατά την συλλογή τους. Επομένως, μία διαδικασία αντιστοίχισης των σημείων GPS στο οδικό δίκτυο που διερευνείται καθίσταται απαραίτητη.

Ο κώδικας εκκινεί τη διαδικασία αντιστοίχισης τροχιών στον χάρτη της περιοχής του San Francisco, εκμεταλλευόμενο το Valhalla Meili API. Αυτό καθιστά δυνατή την ευθυγράμμιση των πορειών GPS του κάθε ταξί με το υποκείμενο οδικό δίκτυο. Κάθε τροχιά που αναπαρίσταται από ζεύγη γεωγραφικού πλάτους και μήκους μέσα στο σύνολο δεδομένων, υποβάλλεται ως αίτημα στον εξυπηρετητή του Valhalla Meili, λαμβάνοντας ως έξοδο ένα νέο σύνολο δεδομένων που φέρει το όνομα **visited_segments** και περιέχει όλη την απαραίτητη πληροφορία, δηλαδή τις αντιστοιχιζόμενες τροχιές επάνω στο οδικό δίκτυο. Συγκεκριμένα, οι πληροφορίες εξάγονται από το trace attributes του Valhalla Meili και περιλαμβάνουν τις ακόλουθες στήλες:

- **αναγνωριστικό ταξί (Taxi ID):** το μοναδικό αναγνωριστικό που αντιστοιχεί σε κάθε ταξί, επιτρέποντας τη διάκριση των μονοκόμματων τροχιών.
- **αναγνωριστικό τροχιάς (Traj ID):** το αναγνωριστικό της υποτροχιάς της μονοκόμματης-κύριας τροχιάς που διένυσε το ταξί με αναγνωριστικό Taxi ID.
- **αναγνωριστικό διαδρομής OSM (OSM Way ID):** δηλώνει το αναγνωριστικό της αντιστοιχιζόμενης στο οδικό δίκτυο ακμής.
- **ώρα έναρξης (Start Time):** χρονοσφραγίδα που υποδηλώνει τη στιγμή που η τροχιά εισέρχεται στην ακμή με αναγνωριστικό OSM Way ID.
- **ώρα λήξης (End Time):** χρονοσφραγίδα που υποδηλώνει τη στιγμή που η τροχιά εξέρχεται από την ακμή με αναγνωριστικό OSM Way ID.

Σύμφωνα με τα παραπάνω, στο καινούριο σύνολο δεδομένων που δημιουργήθηκε, για κάθε ξεχωριστή τροχιά (δηλαδή ένα μοναδικό ζεύγος Taxi ID και Traj ID) είναι γνωστές οι διαδοχικές ακμές που αυτή διένυσε, όπως και την χρονική στιγμή που αυτή εισήλθε και εξήλθε από κάθε ακμή.

5.3 Αναγωγή του Προβλήματος σε Χρονοσειρές

Χρησιμοποιώντας τα δεδομένα που έδωσε ως έξοδο ο αλγόριθμος Valhalla Meili σε συνδυασμό με την μεθοδολογία των ΑΕΜ, καταλήγουμε σε ένα τελικό σύνολο δεδομένων που αποτελείται από χρονοσειρές. Μάλιστα, έχουμε υλοποιήσει την ιδέα των ΑΕΜ σε δύο προγραμματιστικά περιβάλλοντα, επιλέγοντας στο τέλος την ταχύτερη λύση. Στο παρόν υποκεφάλαιο εξετάζονται όλα τα βήματα κατασκευής του τελικού συνόλου δεδομένων.

5.3.1 Υλοποίηση των Αυστηρών Ερωτημάτων Μονοπατιού

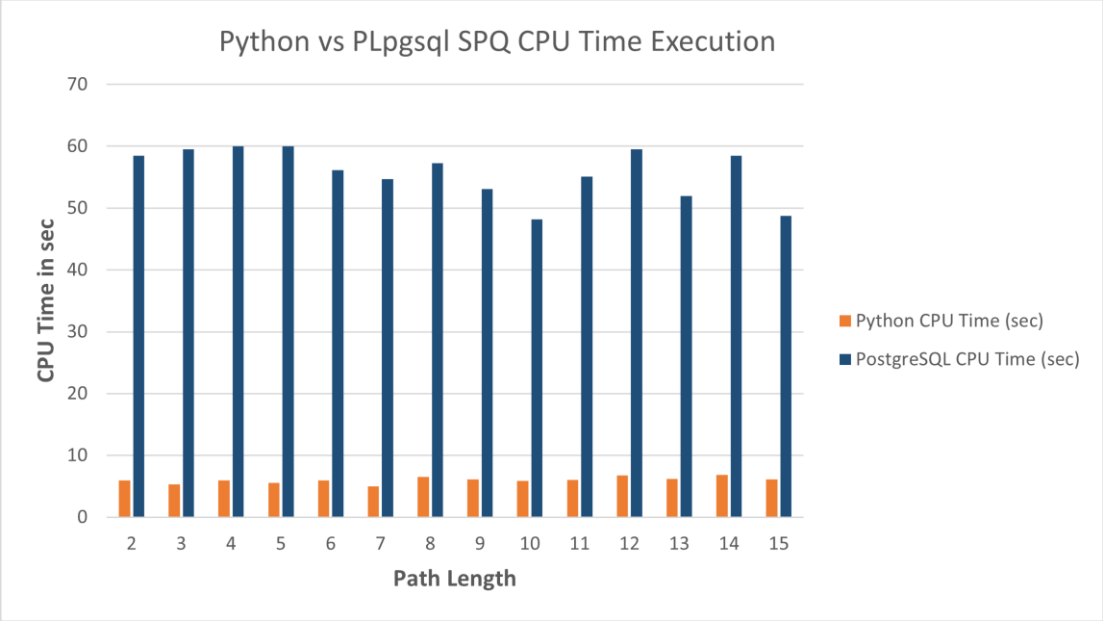
Ένα από τα πιο σημαντικά κομμάτια αυτής της πτυχιακής εργασίας είναι και η συγγραφή της συνάρτησης που υλοποιεί τα ΑΕΜ. Υπενθυμίζεται ότι ένα τέτοιο ερώτημα βρίσκει όλες τις τροχιές που ακολουθούν επακριβώς ένα συγκεκριμένο μονοπάτι οποιουδήποτε μήκους και εντός ενός συγκεκριμένου χρονικού διαστήματος. Η έναρξη αυτού του χρονικού διαστήματος υποδηλώνει τον χρόνο που η τροχιά εισέρχεται στο συγκεκριμένο μονοπάτι, ενώ η λήξη του χρονικού διαστήματος υποδηλώνει τον ανεκτό χρόνο που η εκάστοτε τροχιά θα πρέπει να έχει εξέλθει από αυτό.

Πρέπει να σημειωθεί ότι όλος ο κώδικας έχει γραφεί με **python** σε μορφή συνάρτησης, ενώ η συνολική διάρκεια εκτέλεσης μίας κλήσης της διαρκεί περίπου 250 – 400 ms κατά προσέγγιση. Οι συγκεκριμένοι χρόνοι είναι σχετικοί ως προς το μηχάνημα που εκτελείται ο κώδικας και προκύπτουν από ένα σύνολο δεδομένων με μέγεθος 3.300.000 εγγραφών. Κατά τη γνώμη μας, η συγκεκριμένη χρονική απόδοση είναι πάρα πολύ ικανοποιητική! Περισσότερες πληροφορίες σχετικά με τον κώδικα python δίνονται στο παράρτημα Α.

Στην παρούσα μελέτη έχουμε καταφέρει να υλοποιήσουμε την ίδια μεθοδολογία των ΑΕΜ και σε γλώσσα **PL/pgSQL** εκμεταλλευόμενοι τις δυνατότητες της βάσης περί ταχύτητας δημιουργώντας ευρετήρια B+ δέντρων σε στήλες και σε συνδυασμό στηλών του πίνακα «visited_segments» της βάσης. Περισσότερες αναφορές σχετικά με τα ευρετήρια και τον κώδικα παρατίθενται στο παράρτημα Β.

Αφού έχουμε στην διάθεσή μας δύο όμοιες συναρτήσεις υλοποιημένες σε διαφορετικές γλώσσες, μπορούμε να τις συγκρίνουμε ως προς την ταχύτητα. Για να είμαστε όσο τον δυνατόν ακριβέστεροι, στο πείραμα αυτό καλούμε κάθε συνάρτηση με τις ίδιες παραμέτρους. Πρώτον,

ορίζεται ένα σύνολο από δεκατέσσερις δέσμες. Κάθε δέσμη περιλαμβάνει είκοσι διαφορετικά μονοπάτια ίσου μήκους σε συγκεκριμένα χρονικά διαστήματα. Η πρώτη δέσμη περιλαμβάνει είκοσι μονοπάτια μήκους δύο, η δεύτερη δέσμη περιλαμβάνει είκοσι μονοπάτια μήκους τρία κ.ο.κ. Ουσιαστικά, κάθε δέσμη αποτελεί είκοσι κλήσεις της συνάρτησης των AEM που υλοποιήθηκε. Δεύτερο, σε κάθε προγραμματιστικό περιβάλλον (Python ή PostgreSQL) εκτελούμε μαζικά μία δέσμη και μετράμε τον συνολικό χρόνο εκτέλεσης των είκοσι κλήσεων της συνάρτησης SPQ σε αυτό. Το αποτέλεσμα αυτού του πειράματος συνοψίζεται στο ακόλουθο γράφημα:



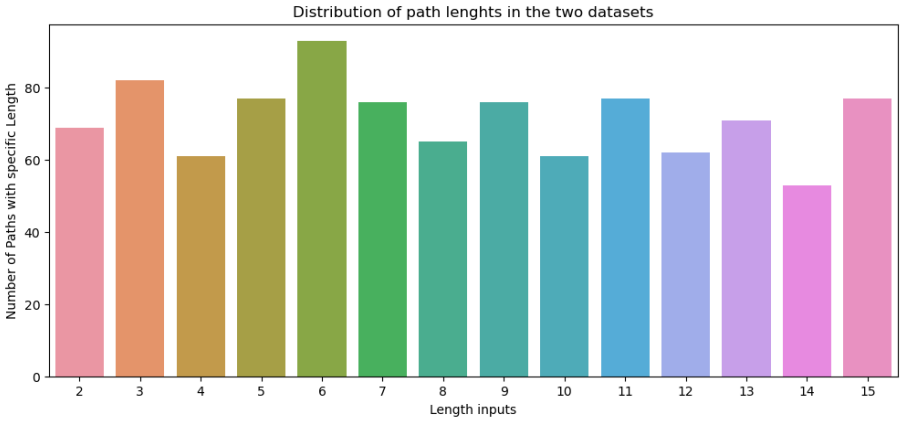
Διάγραμμα 5.1: Στον οριζόντιο άξονα περιλαμβάνεται το μήκος των μονοπατιών που περιέχει κάθε δέσμη. Ο κατακόρυφος άξονας περιέχει τον χρόνο εκτέλεσης των είκοσι ερωτημάτων της δέσμης. Με μπλε χρώμα σημειώνεται η εκτέλεση στο περιβάλλον της PostgreSQL και με πορτοκαλί, στο περιβάλλον της Python.

Συμπεραίνουμε ότι οι χρόνοι εκτέλεσης σε Python είναι ταχύτεροι, συγκριτικά με αυτούς σε PostgreSQL.

5.3.2 Το Τελικό Σύνολο Δεδομένων

Στην ανάλυση ασχολούμαστε με την κατασκευή ενός συνόλου δεδομένων που εστιάζει στην καταγραφή της ροής της κυκλοφορίας στο οδικό δίκτυο ανά μονοπάτι και ανά χρονικό διάστημα. Για αυτή την διαδικασία χρησιμοποιούνται όλα τα αντιστοιχισμένα στο οδικό δίκτυο δεδομένα που υπάρχουν στον πίνακα «visited_segments». Στην διάθεσή μας έχουμε δεδομένα μιας ολόκληρης εβδομάδας. Επομένως, χωρίζουμε αυτό το μεγάλο χρονικό διάστημα σε μικρότερα διαστήματα διάρκειας μισής ώρας έκαστο. Δηλαδή, σε κάθε ημέρα αντιστοιχούν (24 ώρες * 60 λεπτά)/30 λεπτά = 48 χρονικά διαστήματα μισής ώρας.

Προχωρώντας, πρέπει να οριστούν τα μονοπάτια, πάνω στα οποία θα γίνει η ανάλυση της κυκλοφοριακής ροής. Για αυτό τον λόγο, δημιουργούνται χίλια μοναδικά μονοπάτια διαφορετικού μήκους το κάθε ένα. Ο λόγος για τον οποίο επιλέγονται τόσα μονοπάτια είναι για να επιταχυνθεί η ταχύτητα εκτέλεσης του κώδικα. Κάθε μονοπάτι αντιπροσωπεύει μια ακολουθία συνεχόμενων οδικών ακμών που έχει διανύσει ένα ταξί. Οι ακμές αυτές δεν είναι τυχαίες, αλλά πηγάζουν άμεσα από τα δεδομένα, διασφαλίζοντας ότι τα μονοπάτια που δημιουργούνται ακολουθούν την ιδιότητα της συνέχειας, όπως ορίστηκε στο πρώτο κεφάλαιο. Το μήκος αυτών των μονοπατιών μπορεί να κυμαίνεται από δύο έως και δεκαπέντε ακμές. Στο παρακάτω γράφημα φαίνεται η κατανομή των μονοπατιών που δημιουργούνται ως προς το μήκος των ακμών τους.



Διάγραμμα 5.2: Το μήκος του μονοπατιού κυμαίνεται από 2 έως 15 ακμές. Παρατηρούμε ότι στο σύνολο δεδομένων τα μήκη των μονοπατιών έχουν κατανεμηθεί με σχετικά ομοιόμορφο τρόπο.

Πλέον, η κατασκευή του συνόλου δεδομένων χρονοσειρών είναι εύκολη υπόθεση, αφού τα δομικά συστατικά που αποτελείται είναι στην διάθεσή μας. Οι χρονοσειρές αποθηκεύονται σε έναν πίνακα με όνομα «time_series_SPQ». Κάθε εγγραφή σε αυτόν τον πίνακα περιλαμβάνει το μονοπάτι (δηλαδή την λίστα από τις ακμές που αποτελείται), την τροχιά (Taxi ID, Traj ID) που βρίσκεται αυτό το μονοπάτι, το μήκος του μονοπατιού ως προς τις ακμές του και το πλήθος των ταξί που διέσχισαν το συγκεκριμένο μονοπάτι ανά διάστημα μισής ώρας. Συγκεκριμένα, οι στήλες που αποτελείται ο νέος πίνακας είναι οι εξής: «Path», «Length», «Taxi ID», «Traj ID» και τα χρονικά διαστήματα. Κάθε χρονικό διάστημα είναι και μία διαφορετική στήλη στον πίνακα.

	Taxi ID	Traj ID	Path	Length	2008-05-18 00:00:00	2008-05-18 00:30:00	2008-05-18 01:00:00	2008-05-18 01:30:00	2008-05-18 02:00:00	2008-05-18 02:30:00	...
0	255	408	[38855344, 38855344]	2	1	2	4	4	1	14	...
1	111	199	[1112271467, 1112271467, 1112271468]	3	15	15	15	18	14	12	...
2	348	51	[1166095110, 1166095110, 1166095110, 397144264...	7	26	29	30	26	27	26	...
3	388	56	[225806030, 225806030]	2	10	20	27	29	39	29	...
4	151	268	[8921980, 48101169, 48191415, 839813773, 89155...	11	6	9	5	8	6	7	...

Εικόνα 5.1: Το τελικό σύνολο δεδομένων

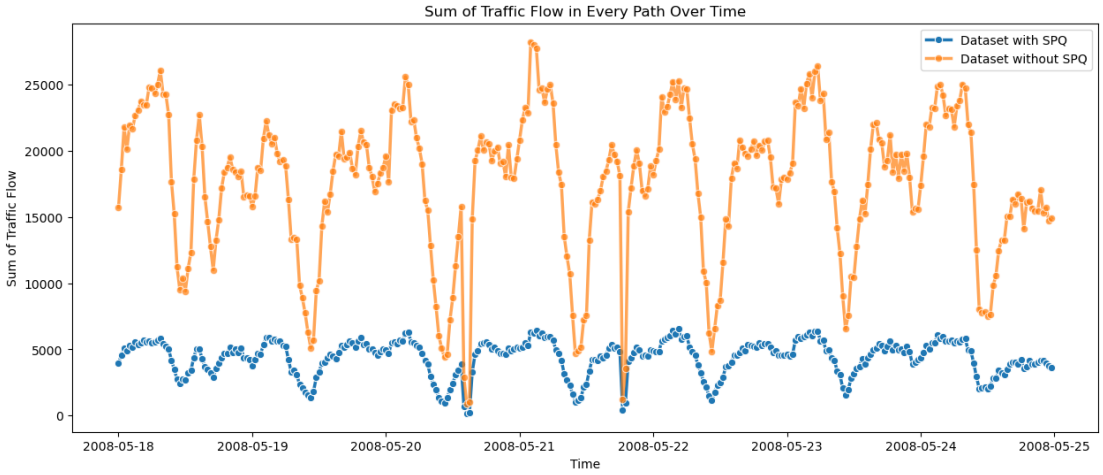
5.4 Πώς Επηρεάζουν τα ΑΕΜ την Πληροφορία της Κυκλοφοριακής Ροής;

Η επιλογή των ΑΕΜ στο πρόβλημα της πρόβλεψης της κυκλοφοριακής ροής εντός ολόκληρων μονοπατιών που βρίσκονται σε ένα οδικό δίκτυο, δηλαδή την πρόβλεψη του αριθμού των ταξί που θα διανύσουν ένα μονοπάτι την επόμενη χρονική στιγμή, ευθύνεται στην ακόλουθη γνώση: τα ΑΕΜ μας εγγυόνται ότι τα ταξί δεν θα παρεκκλίνουν της πορείας τους κατά την διάσχιση του μονοπατιού. Επίσης, η διάσχιση αυτή είναι σίγουρο ότι θα γίνεται με την σωστή σειρά, δηλαδή οι ακμές που αποτελείται το μονοπάτι θα διανύονται μία προς μία από την πρώτη έως και την τελευταία.

Για να ευρεθεί η διαφορά στην κυκλοφοριακή ροή που συλλέγουμε όταν απουσιάζει η χρήση της μεθοδολογίας των ΑΕΜ, δημιουργούμε ένα καινούριο σύνολο δεδομένων χρονοσειρών, με τον ίδιο τρόπο που περιεγράφηκε προηγουμένως. Η κύρια διαφορά τώρα είναι ότι σε κάθε

μονοπάτι δεν μετράμε την κυκλοφοριακή ροή με την βοήθεια της συνάρτησης SPQ. Σε αυτή την περίπτωση, η κυκλοφοριακή ροή σε κάθε μονοπάτι ορίζεται ως το πλήθος των τροχιών που έχουν διανύσει τουλάχιστον μία φορά όλες τις ακμές που απαρτίζουν το μονοπάτι εντός ενός συγκεκριμένου χρονικού διαστήματος.

Το παρακάτω γράφημα δείχνει την διαφορά ανάμεσα στα δύο σύνολα δεδομένων. Με μπλε χρώμα απεικονίζεται η πληροφορία που προκύπτει εφαρμόζοντας την μέθοδο των AEM, ενώ με πορτοκαλί χρώμα παρουσιάζεται η καταγραφή όταν απουσιάζει αυτή η μέθοδος. Επίσης, για κάθε χρονική στιγμή έχουμε συμπεριλάβει το άθροισμα των ταξί που διένυσαν όλα τα μονοπάτια του συνόλου δεδομένων.



Διάγραμμα 5.3: Στον οριζόντιο άξονα απεικονίζεται ο χρόνος, ενώ ο κατακόρυφος άξονας μετράει το συνολικό άθροισμα της κυκλοφοριακής ροής σε όλα τα μονοπάτια.

Αξιοσημείωτο είναι το γεγονός ότι η κυκλοφοριακή ροή παρουσιάζει παρόμοια συμπεριφορά (τάση, περιοδικότητα) και στα δύο σύνολα δεδομένων. Ωστόσο, η παρουσία περιορισμών που διακατέχει την μέθοδο των AEM οδηγεί σε εγκυρότερα δεδομένα. Προτιμάται, λοιπόν, η χρήση των AEM, αφού αναφερόμαστε σε κινητικότητα μέσα σε ολόκληρο μονοπάτι.

5.5 Προσθήκη Επιπλέον Πληροφορίας στο Τελικό Dataset

Ενσωμάτωση δεδομένων καιρού

Ένας παράγοντας που επηρεάζει συχνά την κυκλοφορία στους δρόμους είναι ο καιρός. Για αυτό τον λόγο, σε συνδυασμό με τα δεδομένα κίνησης προστίθενται και δεδομένα καιρού. Τα δεδομένα αυτά έχουν περιγραφεί αναλυτικά στο κεφάλαιο τρία.

Επομένως, έχουμε ενοποιήσει τα δύο σύνολα δεδομένων σε ένα, κάνοντας κατάλληλη επεξεργασία. Και τα δύο σύνολα δεδομένων ανταποκρίνονται στο ίδιο χρονικό πλαίσιο. Επιπλέον, τα δεδομένα κίνησης καταγράφονται ανά μισή ώρα, ενώ τα δεδομένα καιρού καταγράφονται ανά μία ώρα. Άρα, χρειάζεται να συνδεθούν με σωστό τρόπο οι δύο πίνακες: σε δύο εγγραφές δεδομένων κίνησης αντιστοιχίζεται μία εγγραφή δεδομένων καιρού. Με αυτόν τον τρόπο, η χρονική πληροφορία δεν χάνεται, αλλά παραμένει ακλόνητη. Το μόνο μειονέκτημα είναι ότι οι εγγραφές των δεδομένων καιρού αντιγράφονται συνολικά δύο φορές έκαστη.

Ενσωμάτωση χαρακτηριστικών που σχετίζονται με το χρόνο

Πολλά χαρακτηριστικά που σχετίζονται με τον χρόνο εξάγονται από τις πληροφορίες χρονοσφραγίδων στο σύνολο δεδομένων. Τα χαρακτηριστικά αυτά περιλαμβάνουν την ώρα, την ημέρα της εβδομάδας, την ημέρα του μήνα και τα λεπτά. Εφαρμόζεται κυκλική κωδικοποίηση σε ορισμένα χαρακτηριστικά (ώρα, ημέρα της εβδομάδας, ημέρα και λεπτό), η οποία αποτυπώνει την κυκλική φύση τους με την πάροδο του χρόνου. Επιπρόσθετα, το χαρακτηριστικό

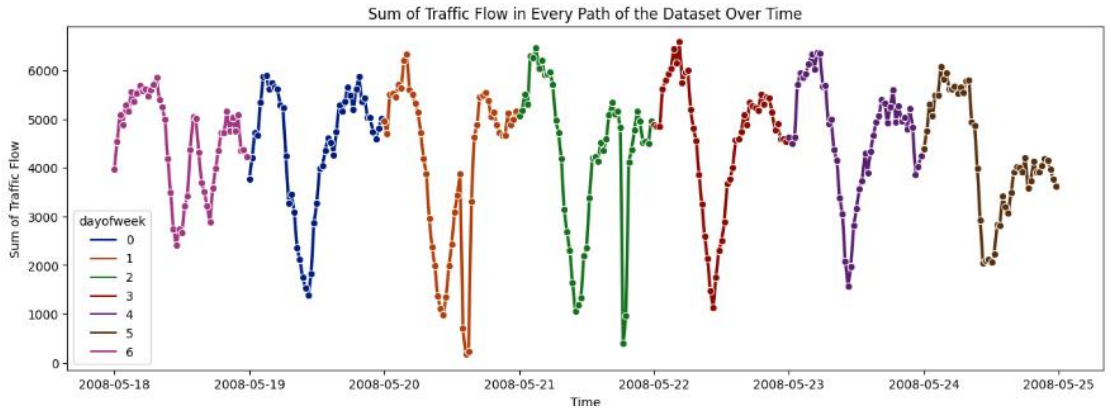
«3hour_interval» εισάγεται, για να υποδεικνύει σε ποιο 3-ωρο χρονικό διάστημα της ημέρας εντοπίζεται αυτή η καταγραφή. Το χαρακτηριστικό αυτό λαμβάνει τιμές από το ένα έως και το οκτώ (αφού 24 ώρες ανά ημέρα / 3 ώρες = 8 διαστήματα ανά ημέρα). Αυτή η πληροφορία μπορεί ενδεχομένως να καταγράψει τις διακυμάνσεις στη ροή της κυκλοφορίας κατά τη διάρκεια διαφορετικών τμημάτων της ημέρας.

Συμπέρασμα

Συνδυάζοντας δεδομένα καιρού και χαρακτηριστικά που σχετίζονται με τον χρόνο με τα υπάρχοντα δεδομένα ροής κυκλοφορίας, η προεπεξεργασία αποσκοπεί στη δημιουργία ενός ολοκληρωμένου συνόλου δεδομένων που ενσωματώνει τόσο εξωτερικούς περιβαλλοντικούς παράγοντες (καιρός) όσο και εγγενή χρονικά πρότυπα (χαρακτηριστικά που σχετίζονται με τον χρόνο). Αυτό το εμπλουτισμένο σύνολο δεδομένων μπορεί δυνητικά να ενισχύσει τις προγνωστικές δυνατότητες ενός μοντέλου πρόβλεψης ροής κυκλοφορίας, επιτρέποντάς του να εξετάσει ένα ευρύτερο φάσμα επιρροών στις μεταβολές της ροής κυκλοφορίας. Ο γενικός στόχος αυτών των προσθηκών είναι η βελτίωση της ακρίβειας και της αποτελεσματικότητας των προβλέψεων.

5.6 Οπτικοποίηση των δεδομένων

Μία από τις βασικές αρχές στην επιστήμη των δεδομένων αποτελεί η οπτικοποίηση των δεδομένων. Με αυτόν τον τρόπο, ο ερευνητής μπορεί εύκολα να κατανοήσει σημαντικές πτυχές στα δεδομένα που δεν μπορούν να παρατηρηθούν αλλιώς. Σε αυτό το υποκεφάλαιο προσπαθούμε να ανακαλύψουμε την συμπεριφορά της κυκλοφοριακής ροής στο χρονικό διάστημα της μίας εβδομάδας που εξετάζουμε χρησιμοποιώντας διαγράμματα.



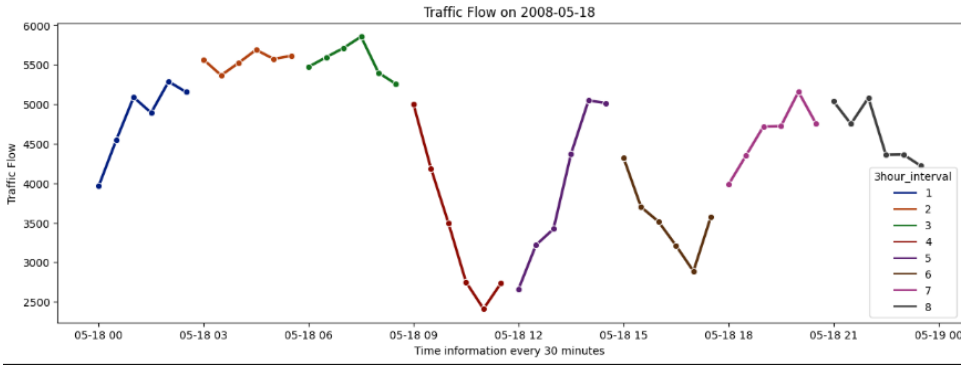
Διάγραμμα 5.4: Συνολική ροή κυκλοφορίας σε κάθε ημέρα.

Σε αυτό το διάγραμμα (Διάγραμμα 5.4), φαίνεται η συνολική ροή της κυκλοφορίας, δηλαδή το άθροισμα της κυκλοφοριακής ροής σε όλα τα μονοπάτια του συνόλου δεδομένων, ανά χρονικό διάστημα. Στον οριζόντιο άξονα έχει τοποθετηθεί ο χρόνος, ο οποίο κυμαίνεται από τις 18 έως και 24 του Μάιου (μία ολόκληρη εβδομάδα). Το χρώμα της γραμμής αντιπροσωπεύει την ημέρα της εβδομάδας.

Μέσα από αυτό το γράφημα μπορούμε να παρατηρήσουμε πως αλλάζει η ροή της κυκλοφορίας κατά τη διάρκεια της εβδομάδας. Συγκεκριμένα, σε κάθε ημέρα, κατά τις πρωινές και βραδινές ώρες η κυκλοφοριακή ροή είναι αυξημένη, ενώ τις μεσημεριανές και απογευματινές ώρες παρατηρείται μικρότερη κινητικότητα στα μονοπάτια. Επομένως, υπάρχει μία σταθερή περιοδικότητα στα δεδομένα.

Στην συνέχεια, γίνεται μία αναλυτικότερη απεικόνιση της κίνησης με βάση την ημέρα και το 3-ωρο χρονικό διάστημα. Για παράδειγμα, το επόμενο γράφημα (Διάγραμμα 5.5) χωρίζει το άθροισμα της κυκλοφοριακής ροής όλων των μονοπατιών κατά την ημέρα 2008-05-18 σε 3-ωρα χρονικά διαστήματα. Κάθε τρίωρο χρονικό διάστημα φαίνεται με διαφορετικό χρώμα. Το

υπόμνημα βοηθάει στην αποσαφήνιση τίνος τρίωρου αντιστοιχείται κάθε χρώμα. Η πληροφορία σε κάθε άξονα είναι η ίδια όπως και στο προηγούμενο γράφημα.



Διάγραμμα 5.5 Η κυκλοφοριακή ροή κατά την ημέρα 2008-05-18 χωρισμένη σε διαστήματα τριών ωρών.

Η παραπάνω απεικόνιση μας επιτρέπει να αναφέρουμε ότι κατά το τέταρτο τρίωρο της ημέρας παρατηρήθηκε η χαμηλότερη αθροιστική κυκλοφορία. Από την άλλη, κατά το τρίτο τρίωρο παρατηρήθηκε η υψηλότερη αθροιστική κυκλοφορία. Επιπλέον, μπορεί να εκφραστεί με σιγουριά ότι στα πρώτα τρία τρίωρα, δηλαδή τις πρώτες εννέα ώρες της ημέρας υπάρχει αυξημένη κινητικότητα, ενώ στο τέταρτο τρίωρο (δηλαδή για τις επόμενες τρεις ώρες) παρατηρείται χαμηλή κινητικότητα κ.ο.κ.

Με παρόμοια γραφήματα, εξετάζεται η κυκλοφοριακή ροή ανά τρίωρο για κάθε μία από τις υπόλοιπες ημέρες της εβδομάδας. Τα διαγράμματα αυτά είναι διαθέσιμα στο φάκελο **Images/ Info about time series dataset**. Γενικά, μέσω αυτών των διαγραμμάτων, γίνονται κατανοητά τα παρακάτω:

- συνολική κυκλοφορία κάθε ημέρας: καθίστανται ευδιάκριτες οι τάσεις και τα μοτίβα της κυκλοφορίας κατά τη διάρκεια της εβδομάδας. Παρατηρείται εάν υπάρχει κάποια συγκεκριμένη μέρα με υψηλότερη ή χαμηλότερη κυκλοφορία.
- κορυφές και κοιλάδες: μπορούμε να εντοπίζουμε τις ώρες κατά τις οποίες η κυκλοφορία είναι στο αποκορύφωμά της κατά τη διάρκεια μιας συγκεκριμένης ημέρας. Ταυτόχρονα, εντοπίζεται εύκολα πότε η κυκλοφοριακή ροή είναι χαμηλή.
- συγκρίσεις ημερών: δίνεται δυνατότητα να συγκρίνουμε την κυκλοφορία μεταξύ διαφορετικών ημερών της εβδομάδας και να παρατηρούμε αν υπάρχουν διαφορές στα μοτίβα κυκλοφορίας μεταξύ των ημερών.

5.7 Χρήση Μοντέλων Μηχανικής και Βαθιάς Μάθησης

Αφού έχει κατασκευαστεί το τελικό σύνολο δεδομένων και έχουν παραχθεί γραφήματα που εξηγούν αυτά τα δεδομένα, το επόμενο βήμα στην ανάλυση αυτή είναι να ορίσουμε αλγορίθμους μηχανικής και βαθιάς μάθησης, με στόχο την πρόβλεψη της κυκλοφοριακής ροής. Στην έρευνα έχουν χρησιμοποιηθεί τέσσερα μοντέλα (Random Forest, XGBoost, LSTM και Encoder – Decoder), για να επιλύσουν το ίδιο πρόβλημα στα ίδια δεδομένα. Ωστόσο, λόγω της διαφορετικής φύσης του κάθε αλγορίθμου, τα αποτελέσματα των προβλέψεων δεν είναι τα ίδια για κάθε μοντέλο.

Σε αυτό το υποκεφάλαιο δείχνουμε πως γίνεται η εκπαίδευση του καλύτερου μοντέλου που χρησιμοποιήθηκε, του XGBoost, και ποιες βελτιστοποιήσεις θεωρήσαμε υπόψιν. Στο τέλος του υποκεφαλαίου γίνεται μία συνοπτική αναφορά για την επίδοση των υπολειπόμενων τριών μοντέλων που εκμεταλλευτήκαμε.

5.7.1 Διαχωρισμός σε Σύνολα Εκπαίδευσης και Ελέγχου

Σε αυτό το βήμα, τα δεδομένα διαιρούνται σε δύο σύνολα: το σύνολο εκπαίδευσης (train set) και το σύνολο ελέγχου (test set). Το πρώτο σύνολο αποτελείται από δεδομένα που θα χρησιμοποιηθούν για την εκπαίδευση κάθε αλγορίθμου μηχανικής μάθησης, με σκοπό οι τελευταίοι να μάθουν διάφορα μοτίβα και σχέσεις σε αυτά. Με άλλα λόγια, κατά την φάση της εκπαίδευσης, το μοντέλο προσαρμόζει τις παραμέτρους του, για να μπορεί να προβλέπει σωστά τα αποτελέσματα σε νέα μη γνωστά δεδομένα. Αντιθέτως, τα δεδομένα ελέγχου αποτελούν δεδομένα, στα οποία το μοντέλο δεν έχει εκπαιδευτεί και εξετάζουν την ικανότητα γενίκευσής του. Για αυτό τον λόγο, η απόδοση του μοντέλου αξιολογείται με βάση το πόσο καλά προβλέπει τα μοτίβα που υπάρχουν στα δεδομένα ελέγχου. Τελευταίο αλλά εξίσου σημαντικό είναι και το σύνολο επικύρωσης (validation set) που χρησιμοποιείται εκτεταμένα στην έρευνά μας. Τα δεδομένα αυτά χρησιμοποιούνται για τη βελτιστοποίηση των υπερπαραμέτρων του μοντέλου κατά τη διάρκεια της εκπαίδευσης, εξασφαλίζοντας πως το μοντέλο γενικεύει καλά σε νέα δεδομένα, χωρίς να επηρεάζεται από την επίδοσή του στα δεδομένα εκπαίδευσης. Συχνά, το validation set ταυτίζεται με το test set στην μελέτη αυτή.

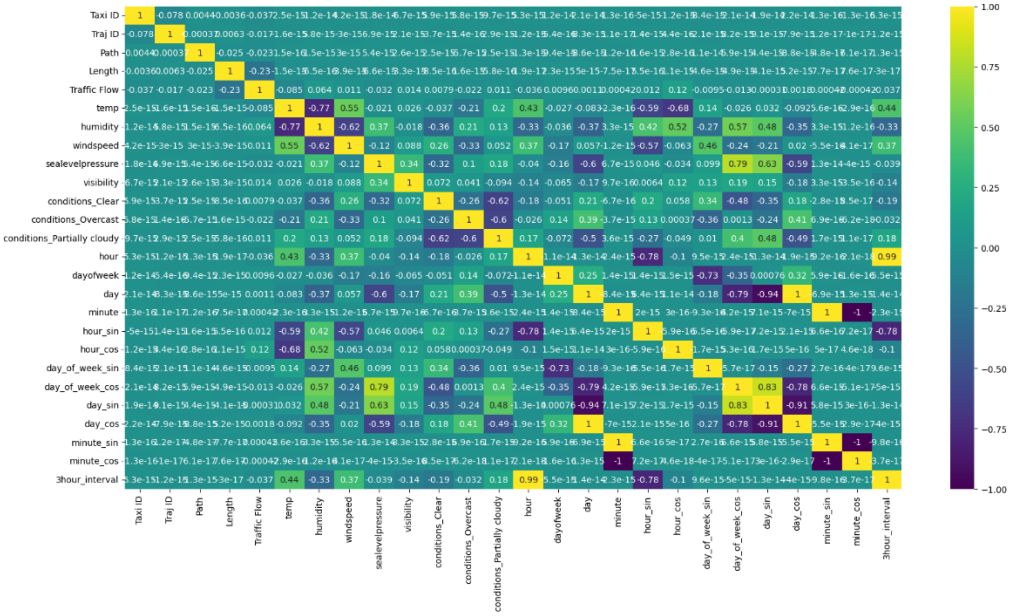
Το χρονικό διάστημα όλων των παρατηρήσεων που υπάρχουν στην διάθεσή μας κυμαίνεται μεταξύ των ημερομηνιών 2008-05-18 και 2008-05-24. Στην μελέτη αποφασίστηκε το σύνολο εκπαίδευσης (train set) να περιέχει όλα τα δεδομένα για κάθε μονοπάτι μέχρι και την 2008-05-23. Τα υπόλοιπα δεδομένα (τα πιο πρόσφατα) βρίσκονται στο σύνολο ελέγχου (test set).

5.7.2 Διαχωρισμός των Χαρακτηριστικών σε Σύνολα Feature και Label

Αφού γίνει ο διαχωρισμός σε σύνολα εκπαίδευσης και ελέγχου, προσπαθούμε να επιλέξουμε εκείνα τα χαρακτηριστικά που θα βοηθήσουν τους αλγορίθμους να προβλέψουν τις εκάστοτε τιμές της κυκλοφοριακής ροής, γνωστά και ως features. Ο πίνακας συσχέτισης (correlation matrix) είναι ένα εργαλείο που βοηθά στην επιλογή αυτών των χαρακτηριστικών κατά την ανάπτυξη ενός μοντέλου μηχανικής μάθησης. Ο ρόλος του είναι να παρουσιάσει τις συσχετίσεις μεταξύ των διαφόρων χαρακτηριστικών στα δεδομένα και να βοηθήσει τον προγραμματιστή στην κατασκευή ενός καλύτερου μοντέλου. Συγκεκριμένα, ο πίνακας συσχέτισης συμβάλλει στην:

- **βελτιστοποίηση του μοντέλου**: εάν το μοντέλο υποφέρει από υπερεκπαίδευση (overfitting), δηλαδή δεν μπορεί να γενικευτεί σε άλλα σύνολα δεδομένων, μια προσέγγιση που υιοθετείται είναι να μειωθεί ο αριθμός των χαρακτηριστικών που χρησιμοποιούνται κατά την εκπαίδευση. Ο πίνακας συσχέτισης μπορεί να βοηθήσει στην επιλογή των χαρακτηριστικών που πρέπει να διατηρηθούν, για να βελτιωθεί η γενίκευση του μοντέλου.
- **ανίχνευση κοινών χαρακτηριστικών**: κοινά χαρακτηριστικά που περιγράφουν τη ίδια πληροφορία μπορούν να εισαγάγουν θόρυβο στο μοντέλο μπερδεύοντάς το. Ο πίνακας συσχέτισης μπορεί να αναδείξει χαρακτηριστικά που έχουν υψηλή σχέση με άλλα χαρακτηριστικά, υποδηλώνοντας ότι μπορεί να είναι περιττά και ίσως να χρειαστεί να μην χρησιμοποιηθούν στην εκπαίδευση.

Κατά την εκπαίδευση των μοντέλων μηχανικής και βαθιάς μάθησης έχει χρησιμοποιηθεί ένας τέτοιος πίνακας συσχέτισης, όπως φαίνεται στην ακόλουθη εικόνα:



Διάγραμμα 5.6: Μήτρα συσχέτισης του συνόλου δεδομένων που χρησιμοποιείται στην έρευνα. Από αυτό το γράφημα προκύπτουν πολλές πληροφορίες για τις σχέσεις των χαρακτηριστικών. Για παράδειγμα, τα χαρακτηριστικά «hour» και «hour sin» φαίνεται να έχουν αρνητική γραμμική συσχέτιση (κοντά στο -1), ενώ τα «sea level pressure» και «day of week cos» έχουν θετική γραμμική συσχέτιση (κοντά στο 1). Τέλος, τα χαρακτηριστικά «Traffic Flow» και «Length» δεν έχουν γραμμική σχέση μεταξύ τους (τιμή κοντά στο 0).

Σε αυτό το γράφημα παρουσιάζεται ένας πίνακας διαστάσεων N*N, όπου N είναι το πλήθος των χαρακτηριστικών στο σύνολο δεδομένων. Οι χρωματικές αντιστοιχίες σε κάθε κελί χρησιμοποιούνται για να απεικονίσουν τις τιμές συσχέτισης μεταξύ των χαρακτηριστικών που αναλογούν σε αυτό το κελί.

Οι νόμιμες τιμές που λαμβάνει μία μήτρα συσχέτισης κυμαίνονται μεταξύ του μείον ένα και του ενός (από το -1 έως και το 1). Τα σκούρα χρώματα (μωβ, μπλε) αντιστοιχούν σε αρνητική τιμή συσχέτισης (κοντά στο -1) ή ακόμα και ανύπαρκτη συσχέτιση (κοντά στο 0).

Αν η τιμή της συσχέτισης είναι κοντά στο μείον ένα, τότε υπάρχει αρνητική γραμμική συσχέτιση μεταξύ των δύο μεταβλητών. Αυτό σημαίνει ότι όταν μία μεταβλητή αυξάνεται, η άλλη μειώνεται σύμφωνα με μια γραμμική σχέση.

Από την άλλη, όταν η τιμή της συσχέτισης είναι κοντά στο μηδέν, δεν υπάρχει γραμμική συσχέτιση ανάμεσα στις μεταβλητές. Αυτό σημαίνει ότι οι μεταβλητές δεν συσχετίζονται με τρόπο που να μπορεί να περιγραφεί με μια γραμμική σχέση. Ωστόσο, αυτό δεν σημαίνει απαραιτήτως ότι δεν υπάρχει καμία άλλη γραμμική συσχέτιση μεταξύ τους.

Τέλος, τα ανοιχτά χρώματα (κίτρινο, πράσινο) αντιστοιχούν σε υψηλή τιμή συσχέτισης (κοντά στο 1), δηλώνοντας ότι οι δύο μεταβλητές παρουσιάζουν θετική συσχέτιση. Κάθε αλλαγή στην τιμή της μίας μεταβλητής, επηρεάζει την άλλη κατά ανάλογο τρόπο. Με άλλα λόγια, υπάρχει μία γραμμική σχέση που διέπει τις δύο μεταβλητές.

5.7.3 Εκπαίδευση του Αλγορίθμου XGBoost

Σε αυτήν την υποενότητα, αναφέρονται τα βασικά βήματα που ακολουθήσαμε, προκειμένου να ορίσουμε ένα μοντέλο πρόβλεψης XGBoost, το οποίο θα είναι βέλτιστο για τα δεδομένα που έχουμε στη διάθεσή μας. Από όλα τα μοντέλα που έχουμε αξιοποιήσει, ο αλγόριθμος XGBoost φαίνεται να παράγει τις καλύτερες προβλέψεις. Αυτό υποθέτουμε ότι οφείλεται αφενός στην ευελιξία του αλγορίθμου και, αφετέρου, στην ικανότητά του να διαχειρίζεται μεγάλα σύνολα δεδομένων.

Η Τεχνική του Κυλιόμενου Παραθύρου

Το κυλιόμενο παράθυρο (ή sliding window) είναι μια τεχνική που χρησιμοποιείται στην ανάλυση χρονοσειρών εξασφαλίζοντας προβλέψεις με βάση προηγούμενες παρατηρήσεις (ιστορικά δεδομένα). Στην ουσία, χρησιμοποιείται ένα κινούμενο παράθυρο που κυλάει κατά μήκος της χρονοσειράς, επιτρέποντας τη δημιουργία πολλαπλών προβλέψεων. Σε ένα πρόβλημα ανάλυσης ή πρόβλεψης χρονοσειρών, η τεχνική του συρόμενου παραθύρου χρησιμοποιείται για διάφορους λόγους:

- οι παρατηρήσεις από το παρελθόν μπορούν να χρησιμοποιηθούν, για να προβλέψουν το μέλλον. Με το sliding window, δημιουργούνται διαδοχικά παράθυρα που περιλαμβάνουν τις προηγούμενες παρατηρήσεις και το μοντέλο εκπαιδεύεται σε αυτά τα παράθυρα για να παράγει τις μελλοντικές τιμές του μεγέθους προς πρόβλεψη.
- με την πρόοδο του χρόνου, το sliding window επιτρέπει τη συνεχή ενημέρωση του μοντέλου με νεότερες παρατηρήσεις, ενισχύοντας την ικανότητα πρόβλεψης με βάση τις τελευταίες πληροφορίες.

Η μεθοδολογία αυτή εγγυάται ότι ο αλγόριθμος εκπαιδεύεται σε ολόκληρη την χρονοσειρά (ή τις χρονοσειρές) που συμπεριλαμβάνεται (συμπεριλαμβάνονται) στα δεδομένα εκπαίδευσης. Η ίδια ακριβώς τεχνική χρησιμοποιείται και στα δεδομένα ελέγχου. Χρησιμοποιώντας διαφορετικά μήκη παραθύρου, δηλαδή το πόσα ιστορικά δεδομένα χρησιμοποιούνται κάθε φορά στο παράθυρο για την πρόβλεψη, μπορούμε να αξιολογήσουμε πως η εφαρμογή τους επηρεάζει την απόδοση του μοντέλου. Αυτό βοηθά να ευρεθεί το βέλτιστο μέγεθος παραθύρου για τη συγκεκριμένη χρονοσειρά.

Ο Αλγόριθμος XGBoost

Ο αλγόριθμος XGBoost (Extreme Gradient Boosting) είναι ένας πανίσχυρος αλγόριθμος μηχανικής μάθησης που δημιουργήθηκε από τον Tianqi Chen [25]. Βασίζεται σε δέντρα αποφάσεων και χρησιμοποιείται ευρέως για προβλήματα παλινδρόμησης (Regression) και ταξινόμησης (Classification).

Ο μηχανισμός αυτός έχει ως θεμέλιο λίθο την τεχνική Gradient Boosting. Η τελευταία πρόκειται για μια ευρέως χρησιμοποιούμενη μέθοδο μηχανικής μάθησης και απαντάται συχνά σε αλγορίθμους που υιοθετούν δέντρα, για να προβαίνουν σε προβλέψεις. Η τεχνική Gradient Boosting εμπίπτει στην κατηγορία της μάθησης συνόλου (ensemble learning), η οποία συνδυάζει ασθενέστερα μοντέλα δέντρων (week learners) για τη δημιουργία ενός ισχυρού (strong learner). Κάθε δέντρο κατασκευάζεται το ένα μετά το άλλο. Η ιδιαιτερότητα της μεθόδου είναι ότι το επόμενο δέντρο απόφασης που δημιουργείται προσπαθεί να μειώσει το σφάλμα του προηγούμενου. Η τελική πρόβλεψη που θα προκύψει, αποτελεί το άθροισμα όλων των προβλέψεων όλων των δέντρων.

Ο αλγόριθμος XGBoost είναι μια βελτιωμένη έκδοση αυτής της μεθόδου και έχει χρησιμοποιηθεί από πολλούς ερευνητές λόγω των εντυπωσιακών επιδόσεών του. Επιπρόσθετα, το μοντέλο συνδυάζεται με ένα πλήθος υπερπαραμέτρων που πρέπει να οριστούν από τον ερευνητή κατά την αρχικοποίησή του. Οι υπερπαράμετροι αυτοί χρησιμοποιούνται για την βελτιστοποίηση των προβλέψεων και την αποφυγή της υπερεκπαίδευσης (overfitting). Στην έρευνα που έχουμε κάνει, χρησιμοποιούνται πέντε υπερπαράμετροι, οι gamma, alpha, max_depth, n_estimators και learning_rate.

1. υπερπαράμετρος **gamma**: βοηθάει το μοντέλο να αποφεύγει το overfitting. Καθώς δημιουργούνται week learners, τα δεδομένα που αναπαρίστανται σε αυτά αποθηκεύονται σε κόμβους και κλαδιά. Το δέντρο μπορεί τελικά να φτάσει σε ένα μεγάλο βάθος (το βάθος ή τα επίπεδα που μπορεί να έχει το δέντρο επιλέγονται από τον προγραμματιστή). Όσο πιο βαθύ είναι το δέντρο, τόσο περισσότερο έχει εντρυφήσει το συγκεκριμένο δέντρο στο σύνολο εκπαίδευσης, οδηγώντας το σε υπερεκπαίδευση. Η υπερπαράμετρος αυτή χρησιμοποιείται για να «κλαδεύει» τα δέντρα απόφασης ώστε να μειωθεί το βάθος τους και να αποφευχθεί το overfitting.
2. υπερπαράμετρος **alpha**: προσθέτει έναν επιπλέον όρο στην συνάρτηση σφάλματος που χρησιμοποιείται κατά την εκπαίδευση. Ανάλογα με την τιμή του alpha, το μοντέλο γίνεται πιο αυστηρό ή ανεκτό σε σφάλματα, επηρεάζοντας ανάλογα και τις τιμές των παραμέτρων κατά την φάση της εκπαίδευσης. Τελικά, αυτός ο επιπλέον όρος ελέγχει την πολυπλοκότητα των δέντρων που δημιουργούνται και αποτρέπει το overfitting.

3. υπερπαράμετρος **n_estimators**: καθορίζει τον τελικό αριθμό των δέντρων (week learners) που θα δημιουργηθούν. Όσο περισσότερα δέντρα προστίθενται, τόσο πιο πολύπλοκο γίνεται το μοντέλο, ενέχοντας τον κίνδυνο υπερεκπαίδευσης. Η εύρεση της κατάλληλης τιμής της υπερπαραμέτρου αυτής είναι σημαντική για την ισορροπία μεταξύ απόδοσης του μοντέλου και χρόνου εκπαίδευσης.
4. υπερπαράμετρος **max_depth**: ορίζει το μέγιστο βάθος που μπορεί να έχει ένα δέντρο. Ένα βαθύ δέντρο μπορεί να προσδιορίσει σύνθετες σχέσεις στα δεδομένα, αλλά συνήθως οδηγεί σε overfitting. Η σωστή τιμή για το max_depth βοηθά στο να δημιουργηθούν δέντρα που γενικεύουν καλά τα δεδομένα.
5. υπερπαράμετρος **learning_rate**: ελέγχει το βήμα, με το οποίο το μοντέλο προσαρμόζεται στα δεδομένα εκπαίδευσης.

Εύρεση Βέλτιστων Υπερπαραμέτρων

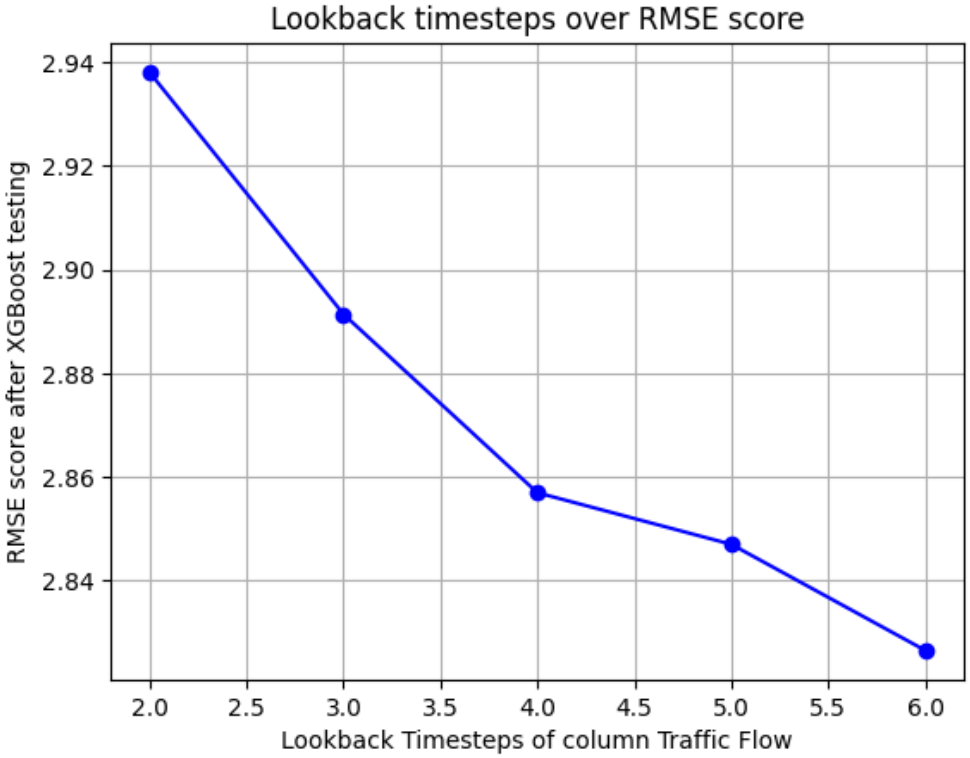
Ένα βασικό ερώτημα που προκύπτει είναι πώς θα ευρεθούν οι κατάλληλες τιμές για αυτές τις υπερπαραμέτρους του μοντέλου XGBoost. Την απάντηση σε αυτό το πρόβλημα δίνει η μέθοδος Grid Search Cross Validation (GridSearchCV). Πρόκειται για μια τεχνική που χρησιμοποιείται για την αυτόματη εύρεση των βέλτιστων τιμών των υπερπαραμέτρων ενός μοντέλου. Οι υπερπαράμετροι δεν μαθαίνονται από το μοντέλο κατά τη διάρκεια της εκπαίδευσης, αλλά πρέπει να προκαθοριστούν πριν από αυτή. Η μέθοδος GridSearchCV λειτουργεί εφαρμόζοντας τα παρακάτω βήματα:

1. ορίζεται μια λίστα πιθανών τιμών για κάθε υπερπαράμετρο που χρειάζεται να βελτιστοποιηθεί.
2. η μέθοδος GridSearchCV δημιουργεί όλους τους δυνατούς συνδυασμούς των τιμών των υπερπαραμέτρων που έχουν καθοριστεί. Για παράδειγμα, εάν έχουν οριστεί δύο υπερπαράμετροι προς βελτιστοποίηση, καθεμία με τρεις δυνατές τιμές, θα δημιουργηθούν συνολικά 3 * 3 = 9 διαφορετικοί συνδυασμοί υπερπαραμέτρων.
3. για κάθε συνδυασμό υπερπαραμέτρων, το μοντέλο εκπαιδεύεται με αυτές και αξιολογείται μέσω της τεχνικής Cross Validation. Η μέθοδος αυτή θέλει το σύνολο εκπαίδευσης να διαχωρίζεται σε μικρότερα υποσύνολα, τα «folds». Κάθε φορά, ένα μόνο fold αντιπροσωπεύει τα δεδομένα επικύρωσης (validation fold), ενώ τα υπόλοιπα χρησιμοποιούνται ως σύνολο εκπαίδευσης (training folds). Αυτή η διαδικασία επαναλαμβάνεται πολλές φορές, ώσπου κάθε fold να έχει αναλάβει τον ρόλο των δεδομένων επικύρωσης. Σε κάθε επανάληψη υπολογίζεται το σφάλμα των προβλέψεων. Στο τέλος, η μέση τιμή των σφαλμάτων κάθε επανάληψης χρησιμοποιείται, για να αξιολογηθεί η απόδοση του μοντέλου για τον συγκεκριμένο συνδυασμό υπερπαραμέτρων.
4. μετά την αξιολόγηση όλων των συνδυασμών, η GridSearchCV επιλέγει τον συνδυασμό υπερπαραμέτρων που παρήγαγε τα καλύτερα αποτελέσματα.

Συμπεραίνοντας, είναι κατανοητό ότι η παρούσα μεθοδολογία αποτελεί μια πολύ χρήσιμη τεχνική για την αυτοματοποίηση της διαδικασίας επιλογής υπερπαραμέτρων και τη βελτιστοποίηση της απόδοσης ενός μοντέλου.

Βελτιστοποίηση του Αλγορίθμου XGBoost

Για το σύνολο δεδομένων που υπάρχει στην κατοχή μας, έχουμε εκπαιδεύσει τον αλγόριθμο XGBoost χρησιμοποιώντας κάθε φορά διάφορα μήκη παραθύρου. Το παρακάτω γράφημα περιγράφει πως μειώνεται ή αυξάνεται το σφάλμα RMSE των προβλέψεων του συγκεκριμένου αλγορίθμου, καθώς αλλάζει το μήκος του παραθύρου.

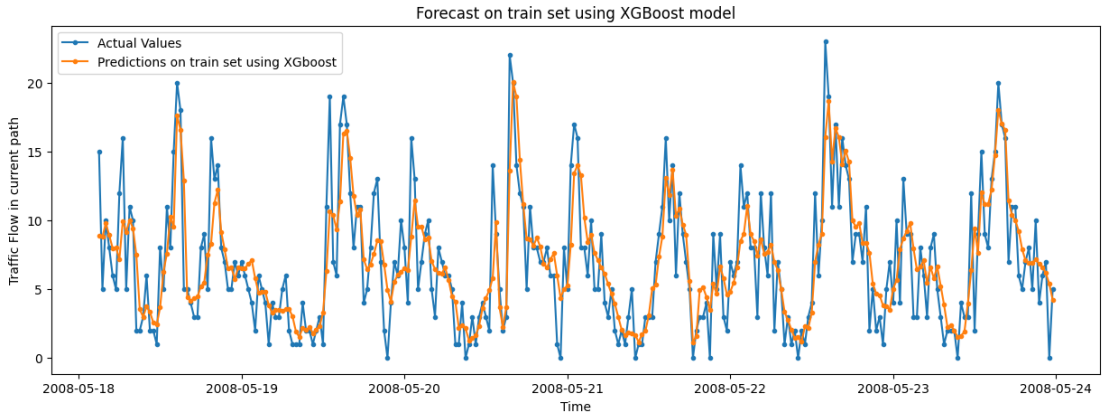


Διάγραμμα 5.7: Απεικονίζεται η σχέση του RMSE (κατακόρυφος άξονας) με το μήκος του παραθύρου που εφαρμόζεται κάθε φορά στα ίδια δεδομένα (οριζόντιος άξονας).

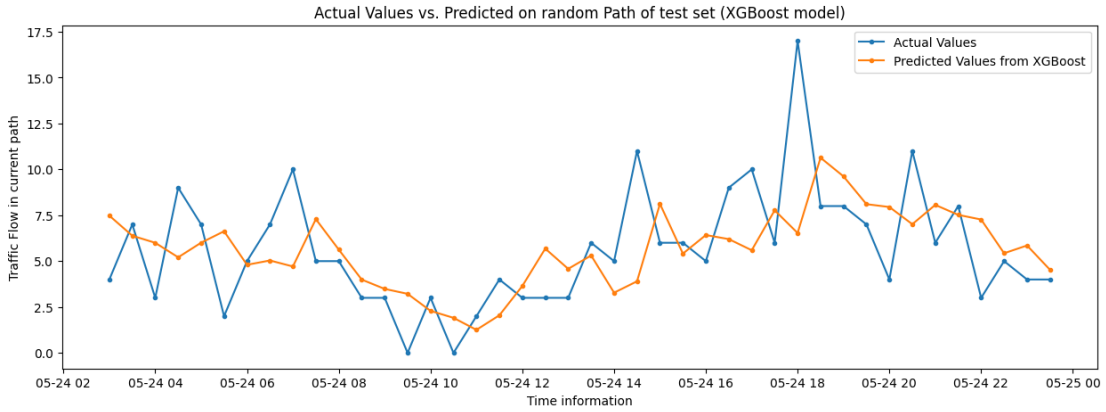
Ο λόγος που επιλέχθηκε ο αλγόριθμος XGBoost για αυτή την βελτιστοποίηση είναι διότι παρουσιάζει την καλύτερη επίδοση, σε σχέση με τα υπόλοιπα μοντέλα (LSTM, Random Forest και Encoder Decoder) που χρησιμοποιήθηκαν στην έρευνα, όπως αναφέρεται και σε μετέπειτα κεφάλαιο. Ταυτόχρονα, για την εύρεση των βέλτιστων υπερπαραμέτρων που θα ωφελήσουν το μοντέλο να παρουσιάσει υψηλή απόδοση στα συγκεκριμένα δεδομένα, έχει χρησιμοποιηθεί η μέθοδος Grid Search CV.

Αξιολόγηση του Αλγορίθμου XGBoost

Μετά από όλη αυτή την διαδικασία που περιεγράφηκε, η οποία συνοψίζεται στα επόμενα βήματα: διαχωρισμός των δεδομένων σε σύνολα εκπαίδευσης και ελέγχου (βήμα 1), επιλογή των καλύτερων χαρακτηριστικών (σύμφωνα με την μήτρα συσχέτισης) που θα βοηθήσουν την εκπαίδευση του μοντέλου (βήμα 2), εφαρμογή του συρόμενου παραθύρου στα δεδομένα με το βέλτιστο μήκος (βήμα 3), χρήση της τεχνικής GridSearchCV για βελτιστοποίηση των υπερπαραμέτρων (βήμα 4), αναπαρίστανται στα επόμενα δύο γραφήματα οι επιδόσεις του μοντέλου XGBoost:



Διάγραμμα 5.8: Επίδοση του μοντέλου XGBoost στο σύνολο εκπαίδευσης.



Διάγραμμα 5.9: Επίδοση του μοντέλου XGBoost στο σύνολο ελέγχου.

Τα παραπάνω δύο γραφήματα δείχνουν την επίδοση του μοντέλου XGBoost σε ένα τυχαία επιλεγμένο μονοπάτι. Με μπλε χρώμα προσδιορίζονται οι πραγματικές τιμές του μεγέθους της κυκλοφοριακής ροής στο μονοπάτι αυτό, ενώ με πορτοκαλί παρουσιάζονται οι προβλέψεις που εξήγε το μοντέλο για αυτό το μέγεθος.

Το πρώτο γράφημα, δείχνει πόσο καλά έχει αποκωδικοποιήσει ο αλγόριθμος τις σχέσεις μεταξύ των features και του προς πρόβλεψη μεγέθους (δηλαδή του Traffic Flow). Παρατηρούμε ότι ο αλγόριθμος έχει μάθει πολύ καλά τα δεδομένα και τα μοτίβα που υπάρχουν σε αυτά. Με άλλα λόγια, έχει αποσαφηνιστεί με ακρίβεια η τάση (δηλαδή οι αυξομειώσεις που αναπαρίστανται στο γράφημα) και η περιοδικότητα (δηλαδή τα μοτίβα που επαναλαμβάνονται ανά συγκεκριμένα χρονικά διαστήματα) της χρονοσειράς. Σε αυτή την κατάληξη έπαιξαν καταλυτικό ρόλο τα χρονικά δεδομένα και τα δεδομένα καιρού που εισήχθησαν, για να βοηθήσουν επιπλέον τον αλγόριθμο να ανακαλύψει περισσότερες σχέσεις στα δεδομένα.

Μάλιστα, ο αλγόριθμος XGBoost φαίνεται να έχει μικρή διακύμανση. Με τον όρο αυτό δηλώνεται η ικανότητα ενός μοντέλου μηχανικής μάθησης να γενικεύεται σε πολλά σύνολα δεδομένων. Στο δεύτερο γράφημα, παρατηρούμε τις σχέσεις που έχει διαγράψει το μοντέλο στα δεδομένα ελέγχου. Και σε αυτή την περίπτωση, η τάση και η περιοδικότητα της χρονοσειράς έχουν προσδιοριστεί αρκετά καλά, αλλά όχι τέλεια. Αυτό δεν είναι απαραίτητα ένα μειονέκτημα, καθώς οι τέλειες προβλέψεις συνήθως υπονοούν ότι ο αλγόριθμος έχει υπερεκπαιδευτεί και δεν μπορεί να γενικευτεί.

Σε γενικές γραμμές, ο αλγόριθμος αυτός έχει καταφέρει να προσδιορίσει τις τιμές της κυκλοφοριακής ροής, ενός μεγέθους που είναι δύσκολο να προβλεφθεί εξαιτίας της μη γραμμικής φύσης του και τους πολλούς αστάθμητους παράγοντες, από τους οποίους εξαρτάται, όπως είναι τα τροχαία ατυχήματα, ο καιρός και οι εορτές.

Τα μεγέθη που χρησιμοποιήθηκαν για να εξετάσουν την ικανότητα του αλγορίθμου XGBoost είναι τα Root Mean Square Error (RMSE) και το Mean Absolute Error (MAE).

- Το RMSE μετράει την τυπική απόκλιση των προβλέψεων από τις πραγματικές τιμές. Υψηλές τιμές του RMSE υποδεικνύουν μεγάλη διακύμανση μεταξύ των προβλέψεων και των πραγματικών τιμών. Ο τύπος για το RMSE score περιγράφεται στην εξίσωση 5.1:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - k)^2} \quad (5.1)$$

  Το n είναι ο αριθμός των παρατηρήσεων, y_i είναι η πραγματική τιμή και k είναι η πρόβλεψη του μοντέλου για την i-οστή παρατήρηση .

- Το MAE μετράει το μέσο απόλυτο σφάλμα μεταξύ των προβλέψεων και των πραγματικών τιμών. Ο τύπος για το MAE score δίνεται από τη σχέση 5.2:

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - k| \quad (5.2)$$

  Το n είναι ο αριθμός των παρατηρήσεων, y_i είναι η πραγματική τιμή και k είναι η πρόβλεψη του μοντέλου για την i-οστή παρατήρηση. Αυτό το μέτρο αγνοεί το πρόσημο του σφάλματος, δηλαδή αν η πρόβλεψη είναι πάνω ή κάτω από την πραγματική τιμή.

Συνολικά, και οι δύο μετρικές χρησιμοποιούνται σε μοντέλα παλινδρόμησης (Regression), για να εκτιμήσουν πόσο κοντά βρίσκονται οι προβλέψεις του μοντέλου στις πραγματικές τιμές. Όσο μικρότερες είναι αυτές οι τιμές, τόσο καλύτερες είναι και οι προβλέψεις που γίνονται.

Σύγκριση του XGBoost με Άλλα Μοντέλα Μηχανικής Μάθησης

Πλέον είναι γνωστό ότι ο αλγόριθμος XGBoost βασίζεται σε δέντρα αποφάσεων, για να κάνει τις όποιες προβλέψεις χρειάζεται. Ένας άλλος αλγόριθμος που έχει ενσωματωθεί στα πλαίσια της έρευνας αυτής και βασίζεται σε δέντρα αποφάσεων, είναι γνωστός με το όνομα Random Forest. Όπως και στην περίπτωση του XGBoost, ο τρόπος με τον οποίο έχει γίνει η εκπαίδευση και αξιολόγηση του μοντέλου αυτού είναι πανομοιότυπος. Επιπρόσθετα, έχει γίνει χρήση της μεθόδου Grid Search Cross Validation για την επιλογή των βέλτιστων υπερπαραμέτρων του.

Στην προσπάθειά να επιλύσουμε το πρόβλημα με όσο το δυνατόν καλύτερο τρόπο, έχουμε εκμεταλλευτεί και τις δυνατότητες των νευρωνικών δικτύων. Αυτοί οι αλγόριθμοι βαθιάς μάθησης είναι σε θέση να προσδιορίζουν μη γραμμικά μοτίβα στα δεδομένα που δέχονται ως είσοδο. Για αυτό τον λόγο, θεωρήθηκε χρήσιμη η δοκιμή ενός μοντέλου LSTM και η χρήση ενός Κωδικοποιητή – Αποκωδικοποιητή που ενσωματώνει επίπεδα LSTM. Όπως αναφέρεται και στο κεφάλαιο δύο, η παρουσία της ενσωματωμένης μνήμης στα μοντέλα LSTM αποδεικνύεται ιδιαίτερα ευεργετική για την πρόβλεψη χρονοσειρών. Αυτό καθιστά τα LSTM ως ιδανική επιλογή, καθώς επιτρέπουν την αποθήκευση πληροφοριών που σχετίζονται με το παρελθόν. Επομένως, μέσω της ανάκτησης και αξιοποίησης αυτών των πληροφοριών, τα μοντέλα LSTM αναδύονται ως πολύτιμο εργαλείο για την πρόβλεψη πολύπλοκων χρονοσειρών.

Όσον αφορά τις βελτιστοποιήσεις που έγιναν στα δύο αυτά μοντέλα, έχουν εισαχθεί επίπεδα Dense και Dropout. Ο ρόλος των πρώτων είναι να βοηθούν το μοντέλο να δημιουργεί μη γραμμικές σχέσεις στα δεδομένα, ενώ ο ρόλος των Dropout επιπέδων είναι να «απενεργοποιούν» κατά την φάση της εκπαίδευσης έναν συγκεκριμένο αριθμό νευρώνων του δικτύου, προκειμένου να αποφευχθεί το φαινόμενο της υπερεκπαίδευσης.

Για να έχει νόημα η σύγκριση με τα υπόλοιπα μοντέλα, έχουμε ακολουθήσει ακριβώς την ίδια μέθοδο εκπαίδευσης και αξιολόγησης που περιεγράφηκε για το μοντέλο XGBoost. Δηλαδή, τα δεδομένα που δόθηκαν ως είσοδο στα νευρωνικά δίκτυα είναι ίδια με αυτά που περιεγράφηκαν σε προηγούμενη ενότητα, υιοθετώντας το ίδιο μέγεθος συρόμενου παραθύρου.

Στον επόμενο πίνακα φαίνονται για κάθε ένα από τα τέσσερα μοντέλα που χρησιμοποιήθηκαν, οι επιδόσεις τους σε όρους σφαλμάτων RMSE και MAE. Υπενθυμίζεται ότι όσο μικρότερος είναι ο δείκτης για κάθε μία από αυτές τις δύο μετρικές, τόσο καλύτερες είναι και οι αποδόσεις του εκάστοτε μοντέλου.

Model	RMSE Score	MAE Score
XGBoost	2.825479	1.747307
LSTM	3.042226	2.082228
Random Forest	2.873084	1.781851
Encoder Decoder	3.605352	2.607833

Εικόνα 5.2: RMSE και MAE scores για κάθε ένα από τα μοντέλα που εκμεταλλευτήκαμε.

Αν και τα νευρωνικά δίκτυα που εφαρμόστηκαν στην μελέτη αυτή μπορούν να διαχειρίζονται τις χρονοσειρές με αποδοτικό τρόπο, συμπεραίνουμε ότι έχουν χειρότερες επιδόσεις από τους δενδροειδής αλγορίθμους. Μία τέτοια παρατήρηση είναι χρήσιμη καθώς αποδεικνύεται ότι τα νευρωνικά δίκτυα με μνήμη δεν ανταποκρίνονται τέλεια σε κάθε σύνολο δεδομένων χρονοσειρών. Επομένως, η απόδοση ενός αλγορίθμου μηχανικής μάθησης εξαρτάται και από τα ίδια τα δεδομένα που του δίνονται ως είσοδο.

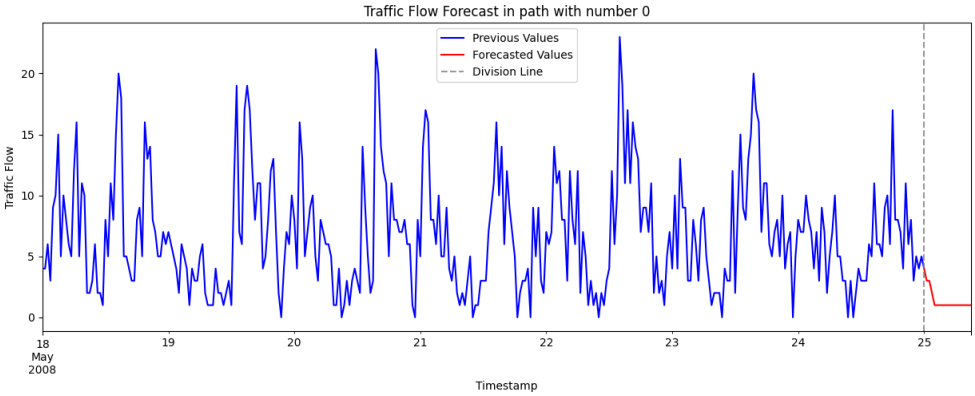
5.8 Προβλέψεις

Σε αυτό το κεφάλαιο παρουσιάζονται οι προβλέψεις που έχουν πραγματοποιηθεί για το μέγεθος της κυκλοφοριακής ροής σε κάθε ένα από τα 1000 μονοπάτια που έχουμε ορίσει στην έρευνά μας. Αξίζει να σημειωθεί ότι οι προβλέψεις παρουσιάζουν τα ακόλουθα χαρακτηριστικά:

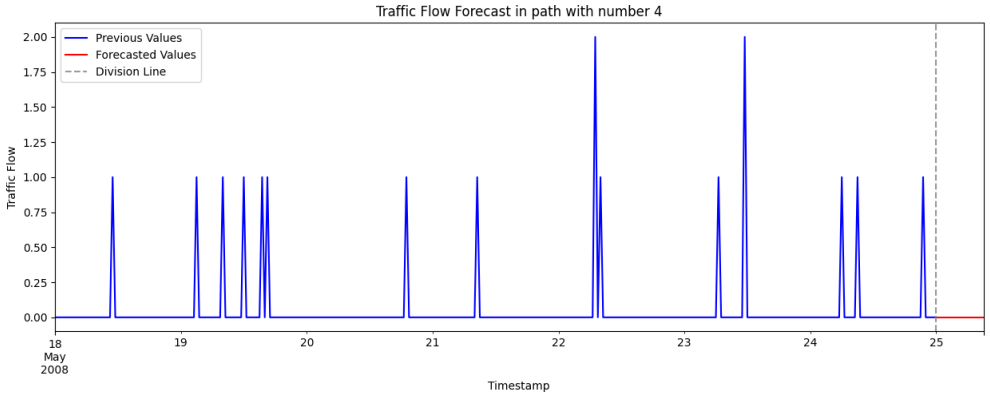
- έχουν δημιουργηθεί χρησιμοποιώντας τον αλγόριθμο XGBoost, λόγω του χαμηλότερου RMSE score που έχει, έναντι των υπολοίπων τριών μοντέλων.
- είναι βραχυπρόθεσμες. Αυτό συμβαίνει, διότι η κυκλοφοριακή ροή είναι ένα μέγεθος μη γραμμικό και πολυδιάστατο: εξαρτάται, δηλαδή από πολλούς παράγοντες, όπως τα τροχαία ατυχήματα, οι εορτές και ο καιρός. Ούτε αυτοί οι παράγοντες είναι γραμμικοί. Επομένως, η μακροπρόθεσμη πρόβλεψη ενός τέτοιου μεγέθους, όπως το κυκλοφοριακό φόρτο, φαντάζει μία διαδικασία δύσκολη, αν όχι μη έγκυρη.

Πέρα από αυτό, η διαδικασία των προβλέψεων έχει πραγματοποιηθεί χρησιμοποιώντας δεδομένα (π.χ. καιρού) και τεχνικές (π.χ. sliding window) παρόμοιες με αυτές που εξηγήθηκαν παραπάνω. Επίσης, ο χρονικός ορίζοντας που εξάγουμε τις προβλέψεις, ανέρχεται στις πρώτες εννέα ώρες μετά από το διάστημα της μίας εβδομάδας που μελετάμε. Με άλλα λόγια, το διάστημα στο οποίο γίνεται η εκπαίδευση και η αξιολόγηση των μοντέλων είναι από 2008-05-18 00:00:00 έως και 2008-05-24 23:30:00. Οι προβλέψεις γίνονται στο διάστημα από 2008-05-25 00:00:00 έως και 2008-05-25 09:00:00.

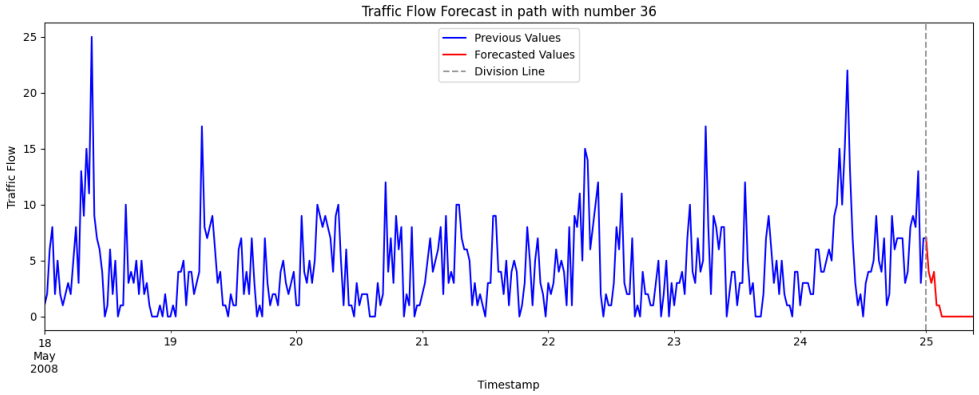
Στα παρακάτω διαγράμματα επιλέγουμε τυχαία μονοπάτια, στα οποία αναπαρίσταται οι γνωστές τιμές του μεγέθους της κυκλοφοριακής ροής (με μπλε χρώμα) και οι προβλεπόμενες τιμές του ίδιου μεγέθους εννέα ώρες στο μέλλον (με κόκκινο χρώμα).



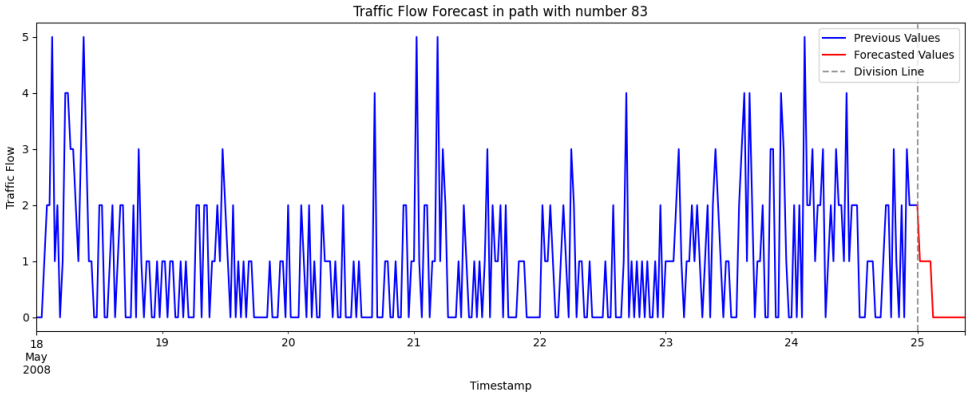
Διάγραμμα 5.10: Γνωστές και προβλεπόμενες τιμές του μεγέθους της κυκλοφοριακής ροής στο μονοπάτι με αριθμό 0.



Διάγραμμα 5.11: Γνωστές και προβλεπόμενες τιμές του μεγέθους της κυκλοφοριακής ροής στο μονοπάτι με αριθμό 4.



Διάγραμμα 5.12: Γνωστές και προβλεπόμενες τιμές του μεγέθους της κυκλοφοριακής ροής στο μονοπάτι με αριθμό 36.



Διάγραμμα 5.13: Γνωστές και προβλεπόμενες τιμές του μεγέθους της κυκλοφοριακής ροής στο μονοπάτι με αριθμό 83.

6. Συμπεράσματα και Προτάσεις για Βελτίωση

Συμπεράσματα

Η κυκλοφοριακή ροή στις οδούς είναι ένα μέγεθος που ασκεί άμεση και έμμεση επίδραση στις αποφάσεις και στις ενέργειες των ανθρώπων. Συγκεκριμένα, η ροή των οχημάτων επηρεάζει την ασφάλεια των οδηγών, τις χρονικές καθυστερήσεις και την ατμοσφαιρική ρύπανση. Έτσι, η πρόβλεψη της μελλοντικής κυκλοφοριακής ροής αναδύεται ως αναγκαία, διότι μπορεί να συμβάλει στην επίλυση αυτών των σημαντικών ζητημάτων.

Η παρούσα πτυχιακή εργασία αποκάλυψε ότι το προς πρόβλεψη μέγεθος είναι πολύπλοκο, αφού εξαρτάται από πολλούς παράγοντες και παρουσιάζει μη γραμμικές σχέσεις. Αυτή η πολυπλοκότητα αποτέλεσε μια σημαντική πρόκληση κατά την διαδικασία ανάλυσης και πρόβλεψης του εν λόγω μεγέθους. Παρ' όλα αυτά, καταφέραμε να αναπτύξουμε έναν αλγόριθμο μηχανικής μάθησης που ανταποκρίνεται ικανοποιητικά στις απαιτήσεις της έρευνας και μπορεί να γενικευτεί σε άγνωστα δεδομένα κίνησης της ίδιας δομής. Αυτό αποδεικνύεται τόσο στις επιδόσεις του μοντέλου στο σύνολο ελέγχου, όσο και στις προβλέψεις που αυτό έχει παράξει.

Προτάσεις για Βελτίωση

Στο πλαίσιο της παρούσας μελέτης, υπήρξε περιορισμός στις δυνατότητες που χρησιμοποιήθηκαν για την εκτέλεση των προβλέψεων. Αρχικά, για τον σκοπό της ανάλυσης, επικεντρωθήκαμε σε δεδομένα κίνησης ταξί εντός της πόλης του San Francisco. Παρόλο που τα εν λόγω δεδομένα εξυπηρέτησαν τον σκοπό της παρούσας πτυχιακής εργασίας, πρέπει να αποσαφηνιστεί ότι δεν αντιπροσωπεύουν εντελώς την πραγματική κυκλοφοριακή κίνηση στην εν λόγω πόλη. Με άλλα λόγια, δεν λαμβάνονται υπόψιν οι κινήσεις άλλων μεταφορικών μέσων, όπως τα λεωφορεία και τα αυτοκίνητα. Για τη βελτίωση αυτής της κατάστασης, προτείνεται η προσθήκη δεδομένων από διάφορους τύπους οχημάτων στο σύνολο δεδομένων που χρησιμοποιείται.

Παράλληλα, κατά την εξέλιξη της διαδικασίας πρόβλεψης, χρησιμοποιήθηκαν και δεδομένα καιρού. Ο καιρός αποτελεί έναν από τους βασικούς παράγοντες που επηρεάζουν την κυκλοφοριακή ροή. Επιπλέον δεδομένα, όπως οι εορτές και οι αργίες εντός της χρονικής περιόδου που εξετάζεται, αποδεικνύονται ιδιαιτέρως σημαντικά και άξια εκμετάλλευσης για τη διαδικασία της πρόβλεψης.

Τέλος, όσον αφορά τα μοντέλα μηχανικής μάθησης, η ενδεχόμενη ενσωμάτωση ενός μοντέλου που θα λαμβάνει υπόψιν τη χωρική και χρονική διάταξη του προβλήματος αποτελεί μία προοπτική βελτίωσης. Υπό αυτήν την προσέγγιση, θα λαμβάνονται υπόψιν οι συνδέσεις μεταξύ μονοπατιών εντός του οδικού δικτύου, καθιστώντας εφικτή την εκμετάλλευση της πληροφορίας σε γειτονικά μονοπάτια. Η χωρική αυτή διάταξη θα δίνεται ως πληροφορία στον μοντέλο μαζί με τον χρόνο. Παράδειγμα ενός τέτοιου μοντέλου αποτελεί το Graph Neural Network (GNN). Με τις προαναφερθείσες προσεγγίσεις, οι προβλέψεις ενδέχεται να παρουσιάζουν αυξημένη ακρίβεια.

Πίνακας Ορολογιών

Ξενόγλωσσος όρος	Ελληνικός όρος
Big Data	Μεγάλος Όγκος Δεδομένων
Latitude	Γεωγραφικό Πλάτος
Longitude	Γεωγραφικό Μήκος
Dropout Layer	Στρώμα Αγνόησης
Convolutional	Συνελικτικός
Pooling Unit	Μονάδα Ομαδοποίησης
Index	Ευρετήριο
Dense Layer	Πυκνό Στρώμα
Grid Search Cross Validation	Αναζήτηση πλέγματος με την μέθοδο της πολλαπλής διεπικύρωσης
Correlation Matrix	Μήτρα Συσχέτισης
Ensemble Learning	Μάθηση συνόλου
Strong Learner	Ισχυρός μαθητής
Weak Learner	Μη ισχυρός μαθητής
Decision Tree	Δέντρο Απόφασης
Classification	Ομαδοποίηση / Συσταδοποίηση
Regression	Παλινδρόμηση

Πίνακας Συντμήσεων – Αρκτικόλεξων – Ακρωνύμιων

Αρκτικόλεξο	Πλήρης Σημασία
GPS	Global Positioning System
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
Seq2Seq	Sequence to Sequence
OSM	Open Street Map
SPQ	Strict Path Query
AEM	Ακριβές Ερώτημα Μονοπατιού
SAE	Stacked Autoencoder
KA	Κινούμενο Αντικείμενο
LSTM	Long Short-Term Memory
ARIMA	Autoregressive integrated moving average
SARIMA	Seasonal AutoRegressive Integrated Moving Average
SVM	Support Vector Machine
RW	Random Walk
MAPE	Mean Absolute Percentage Error

Βιβλιογραφικές Πηγές

[1] B. Krogh, N. Pelekis, Y. Theodoridis, and K. Torp, ‘Path-based Queries on Trajectory Data’.

[2] ‘ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ ΣΧΟΛΗ ΑΓΡΟΝΟΜΩΝ ΚΑΙ ΤΟΠΟΓΡΑΦΩΝ ΜΗΧΑΝΙΚΩΝ’.

[3] X. Liu et al., 2015 IEEE International Conference on Smart City : SmartCity 2015 : proceedings : 19-21 December 2015, Chengdu, China.

[4] ‘Supervised_Deep_Learning_Based_for_Traffic_Flow_Prediction’.

[5] Y. Lv, Y. Duan, W. Kang, Z. Li, and F. Y. Wang, ‘Traffic Flow Prediction with Big Data: A Deep Learning Approach’, IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 2, pp. 865–873, Apr. 2015, doi: 10.1109/TITS.2014.2345663.

[6] R. Jiang et al., ‘DeepCrowd: A Deep Model for Large-Scale Citywide Crowd Density and Flow Prediction’, IEEE Trans Knowl Data Eng, vol. 35, no. 1, pp. 276–290, Jan. 2023, doi: 10.1109/TKDE.2021.3077056.

[7] M. Lu, K. Zhang, H. Liu, and N. Xiong3, ‘Graph Hierarchical Convolutional Recurrent Neural Network (GHCRNN) for Vehicle Condition Prediction’.

[8] X. Dong, T. Lei, S. Jin, and Z. Hou, ‘Short-term traffic flow prediction based on XGBoost’, Proceedings of 2018 IEEE 7th Data Driven Control and Learning Systems Conference, DDCLS 2018, pp. 854–859, Oct. 2018, doi: 10.1109/DDCLS.2018.8516114.

[25] T. Chen and C. Guestrin, ‘XGBoost: A Scalable Tree Boosting System’, Mar. 2016, doi: 10.1145/2939672.2939785.

Διαδικτυακές Αναφορές

[9] ‘Cabspotting | Stamen’. https://stamen.com/work/cabspotting/ (accessed Aug. 17, 2023).

[10] ‘Weather Data Services | Visual Crossing’. https://www.visualcrossing.com/weather/weather-data-services (accessed Aug. 17, 2023).

[11] ‘Python History - javatpoint’. https://www.javatpoint.com/python-history (accessed Aug. 17, 2023).

[12] ‘pandas - Python Data Analysis Library’. https://pandas.pydata.org/ (accessed Aug. 17, 2023).

[13] ‘NumPy’. https://numpy.org/ (accessed Aug. 17, 2023).

[14] ‘scikit-learn: machine learning in Python — scikit-learn 1.3.0 documentation’. https://scikit-learn.org/stable/ (accessed Aug. 17, 2023).

[15] ‘TensorFlow’. https://www.tensorflow.org/ (accessed Aug. 17, 2023).

[16] ‘Keras: Deep Learning for humans’. https://keras.io/ (accessed Aug. 17, 2023).

[17] ‘Matplotlib — Visualization with Python’. https://matplotlib.org/ (accessed Aug. 17, 2023).

[18] M. Waskom, ‘seaborn: statistical data visualization’, J Open Source Softw, vol. 6, no. 60, p. 3021, Apr. 2021, doi: 10.21105/JOSS.03021.

[19] ‘Project Jupyter | Home’. https://jupyter.org/ (accessed Aug. 17, 2023).

[20] ‘Καλώς ορίσατε στο Colaboratory - Colaboratory’. https://colab.research.google.com/ (accessed Aug. 17, 2023).

[21] ‘PostgreSQL: Documentation: 15: Chapter 43. PL/pgSQL — SQL Procedural Language’. https://www.postgresql.org/docs/current/plpgsql.html (accessed Aug. 17, 2023).

[22] ‘Docker: Accelerated Container Application Development’. https://www.docker.com/ (accessed Aug. 17, 2023).

[23] ‘Valhalla Docs’. https://valhalla.github.io/valhalla/ (accessed Aug. 17, 2023).

[24] ‘OpenStreetMap’. https://www.openstreetmap.org/#map=7/38.359/23.810 (accessed Aug. 17, 2023).

Παράρτημα Α

Στο παράτημα αυτό του τόμου εργασίας περιγράφεται ο αλγόριθμος των ΑΕΜ που έχει υλοποιηθεί σε γλώσσα Python.

Πιο συγκεκριμένα, η συνάρτηση προσπελάβνει κάθε φορά τον πίνακα «visited_segments». Οι παράμετροι που λαμβάνει σαν είσοδο η συνάρτηση SPQ είναι οι ακόλουθες:

- **path:** είναι η διαδρομή που πρέπει να ακολουθήσουν οι τροχιές ακριβώς (ακμή προς ακμή), χωρίς να παρεκκλίνουν από αυτή. Αυτό το μονοπάτι μπορεί να έχει οποιοδήποτε μήκος ακμών μεγαλύτερο ή ίσο των δύο.
- **time_enter:** ο χρόνος, κατά τον οποίο η τροχιά θα πρέπει να έχει εισέλθει στην πρώτη ακμή της διαδρομής που δίνεται ως είσοδος.
- **time_leave:** ο χρόνος κατά τον οποίο η τροχιά πρέπει να έχει εγκαταλείψει την τελευταία ακμή της διαδρομής που δίνεται ως είσοδος.

Ακολουθεί τώρα ο ψευδοκώδικας που περιγράφει την συνάρτηση που υλοποιεί τα ΑΕΜ. Το όνομα της συνάρτησης είναι SPQ:

ΣΥΝΑΡΤΗΣΗ SPQ

Παράμετροι Συνάρτησης: ένα μονοπάτι, ο χρόνος time_enter, ο χρόνος time_leave.

Κώδικας Συνάρτησης:

0. Υπολόγισε το μήκος του μονοπατιού και αποθήκευσέ το σε μία μεταβλητή path_length

1. Φτιάξε μία κενή λίστα trajectories

2. Βρες όλες τις εγγραφές του πίνακα visited_segments που έχουν Start Time >= time_enter και End Time <= time_leave, αποθήκευσέ τες σε μία μεταβλητή με όνομα examined_data.

3. Βρες όλα τα αναγνωριστικά των γραμμών που περιέχουν σαν OSM Way ID την πρώτη ακμή στο μονοπάτι π και αποθήκευσέ τα σε μία λίστα needed_indexes.

4. Για κάθε στοιχείο index στη λίστα needed_indexes επανέλαβε:

4.1 Βρες το Taxi ID του στοιχείου index και αποθήκευσέ το σε μία μεταβλητή taxi_id

4.2 Βρες το Traj ID του στοιχείου index και αποθήκευσέ το σε μία μεταβλητή traj_id

4.3 Όρισε την τιμής μίας νέας μεταβλητής inter = 1

4.4 Από i = 1 έως path_length επανέλαβε:

4.4.1 Έλεγξε εάν η γραμμή με αναγνωριστικό index+i περιέχει σαν Taxi ID == taxi_id ΚΑΙ Traj ID == traj_id ΚΑΙ OSM Way ID την επόμενη σε σειρά ακμή στο μονοπάτι.

4.4.2 Εάν ισχύει η παραπάνω συνθήκη αύξησε τον μετρητή inter κατά 1

4.5 Τέλος εσωτερικού βρόγχου

4.6 Εάν path_length == inter, πρόσθεσε το ζευγάρι (taxi_id,traj_id) στη λίστα trajectories

4.7 Τέλος εξωτερικού βρόγχου

5. Διέγραψε τα διπλότυπα ζευγάρια (εάν υπάρχουν) από τη λίστα trajectories και επέστρεψε το μήκος της.

ΤΕΛΟΣ ΣΥΝΑΡΤΗΣΗΣ SPQ

Παράρτημα Β

Σε αυτό το παράρτημα, εξηγείται η τεχνική που ακολουθήθηκε για την υλοποίηση της συνάρτησης των ΑΕΜ στο περιβάλλον της PostgreSQL.

Όσον αφορά τα **ευρετήρια**, δημιουργείται, αρχικά, ένα ευρετήριο πάνω στην στήλη «OSM_Way_ID». Αυτό το ευρετήριο βοηθά στην ταχεία ανάκτηση δεδομένων όταν αναφερόμαστε σε συγκεκριμένες ακμές του δρόμου. Στην συνέχεια, δηλώνεται ένα δεύτερο ευρετήριο πάνω στις στήλες «Time_Enter» και «Time_Leave». Αυτό το ευρετήριο επιτρέπει την αποτελεσματική αναζήτηση και το φιλτράρισμα εγγραφών με βάση τα χρονικά διαστήματα

εισόδου και εξόδου. Τέλος, παράγεται ένα τρίτο ευρετήριο πάνω στις στήλες «OSM_Way_ID», «Traj_ID» και «Taxi_ID». Αυτό το ευρετήριο βοηθά στην γρήγορη αναζήτηση και συγκριτική ανάλυση εγγραφών με βάση τα κριτήρια που περιλαμβάνουν τα αναγνωριστικά των ακμών και των τροχιών.

Έπειτα, υλοποιείται η συνάρτηση SPQ σε γλώσσα PL/pgSQL με τον ίδιο τρόπο, όπως ακριβώς υλοποιήθηκε και στην γλώσσα Python, χρησιμοποιώντας τις ίδιες παραμέτρους σε κάθε κλήση της. Ο κώδικας και τα σχόλια που τον συνοδεύουν παρατίθεται στο αρχείο «**SQP func in PLpgsql.txt**» στο GitHub Repository που συνοδεύει την εργασία αυτή.