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 (ΑΝΑΠΛΗΡΩΤΗΣ ΚΑΘΗΓΗΤΗΣ)

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, Αναπληρωτή Καθηγητή του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς, καθώς και στην δρα. 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

Κατάλογος Διαγραμμάτων

Πίνακας Περιεχομένων

Εισαγωγή

Η κυκλοφοριακή ροή (Traffic Flow) 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** (Machine Learning - ML) και **βαθιάς μάθησης** (Deep Learning - DL). Το πρόβλημα προσεγγίζεται υπό ένα μοναδικό πρίσμα, μια μεθοδολογία που, από όσο γνωρίζουμε, δεν έχει χρησιμοποιηθεί στο παρελθόν για την επίλυση ενός τέτοιου προβλήματος.

Ο Κλάδος της Τεχνητής Νοημοσύνης

Όπως αναφέρθηκε προηγουμένως, το πρόβλημα που εστιάζει η παρούσα έρευνα συγκαταλέγεται στον τομέα της μηχανικής μάθησης. Η τελευταία ανήκει σε ένα ευρύτερο πεδίο που ονομάζεται τεχνητή νοημοσύνη (Artificial Intelligence - AI). Επομένως, καθίστανται αναγκαία η εξήγηση σχετικών με τον κλάδο αυτό όρων.

Η **τεχνητή νοημοσύνη** αντιπροσωπεύει μια σημαντική πτυχή της πληροφορικής που αποσκοπεί στη δημιουργία συστημάτων, τα οποία είναι σε θέση να εκτελούν εργασίες που απαιτούν ανθρώπινη νοημοσύνη και λογική σκέψη. Σε αυτήν την κατηγορία περιλαμβάνονται εφαρμογές όπως η αναγνώριση φωνής, η αναγνώριση εικόνας, η αυτόνομη οδήγηση και η αυτόματη μετάφραση. Η τεχνητή νοημοσύνη στηρίζεται στην υπολογιστική ισχύ για να αναλύσει και να επεξεργαστεί δεδομένα, προκειμένου να παράγει λογικά αποτελέσματα.

Συνεχίζοντας, η **μηχανική μάθηση** αποτελεί ένα σημαντικό υποσύνολο της τεχνητής νοημοσύνης. Στη μηχανική μάθηση, αλγόριθμοι και μοντέλα αναπτύσσονται, ώστε να μπορούν να «μαθαίνουν» από τα δεδομένα που τους παρέχονται. Αυτό σημαίνει ότι τα συστήματα αυτά μπορούν να αναγνωρίσουν μοτίβα και να κάνουν προβλέψεις με βάση την επεξεργασία των δεδομένων. Για παράδειγμα, ένα μοντέλο μηχανικής μάθησης μπορεί να εκπαιδευτεί για να αναγνωρίζει πρόσωπα σε φωτογραφίες ή να προβλέπει τις τιμές των μετοχών βάσει των προηγούμενων ιστορικών δεδομένων.

Τέλος, η **βαθιά μάθηση** αποτελεί ένα υποσύνολο της μηχανικής μάθησης που βασίζεται στη χρήση **νευρωνικών δικτύων** (Neural Networks – NN) με πολλά (συνήθως) επίπεδα. Αυτά τα βαθιά νευρωνικά δίκτυα προσομοιάζουν τον τρόπο με τον οποίο λειτουργεί ο ανθρώπινος εγκέφαλος, και έχουν την ικανότητα να εξάγουν υψηλού επιπέδου σχέσεις από τα δεδομένα. Αυτό επιτρέπει σε αυτά τα μοντέλα να αντιμετωπίσουν πιο πολύπλοκες εργασίες όπως η αναγνώριση αντικειμένων σε εικόνες, η ανάλυση φυσικής γλώσσας και η παραγωγή μουσικής.

Όλοι οι αλγόριθμοι που χρησιμοποιούν δεδομένα, για να παράγουν κάποια αποτελέσματα, ονομάζονται αλλιώς και μοντέλα (Models). Με άλλα λόγια, ένα μοντέλο στον χώρο της μηχανικής μάθησης αναφέρεται σε ένα συγκεκριμένο υπολογιστικό σύστημα ή αλγόριθμο που έχει εκπαιδευτεί για μια συγκεκριμένη εργασία. Επίσης, τα μοντέλα μηχανικής μάθησης μπορούν να είναι γραμμικά, όπως ο αλγόριθμος της γραμμικής παλινδρόμησης (Linear Regression), ή πιο πολύπλοκα, όπως ένα βαθύ νευρωνικό δίκτυο (Deep Neural Network – DNN). Κάθε μοντέλο, προκειμένου να λειτουργεί σωστά για κάθε σύνολο δεδομένων, περιέχει κάποιες

υπερπαραμέτρους (Hyperparameters), οι οποίες ρυθμίζονται από τον χρήστη κατά την κατασκευή του αλγορίθμου. Αυτές οι υπερπαράμετροι ρυθμίζουν ουσιαστικά το μοντέλο, βοηθώντας το να προσαρμοστεί καλύτερα στα δεδομένα που λαμβάνει ως είσοδο.

Στην παρούσα έρευνα έχουν χρησιμοποιηθεί τέσσερα μοντέλα μηχανικής και βαθιάς μάθησης που προσπαθούν να επιλύσουν το ίδιο πρόβλημα. Τα μοντέλα αυτά είναι αφενός τα **XGBoost** και **Random Forest** που βασίζονται σε αλγορίθμους δέντρων αποφάσεων (Decision Trees) και ανήκουν στην κατηγορία της μηχανικής μάθησης. Αφετέρου, χρησιμοποιούνται και δύο μοντέλα βαθιάς μάθησης, ένα αναδρομικό νευρωνικό δίκτυο με βραχυπρόθεσμη και μακροπρόθεσμη μνήμη (Long Short-Term Memory Recurrent Neural Network – LSTM RNN) και ένα μοντέλο Κωδικοποιητή - Αποκωδικοποιητή που βασίζεται σε νευρωνικά δίκτυα LSTM (Encoder – Decoder Model). Ο τρόπος λειτουργίας των μοντέλων αυτών αναλύεται περισσότερο στο παράρτημα Γ.

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

Στο τέταρτο κεφάλαιο γίνεται εκτενής περιγραφή της μεθοδολογίας που ακολουθούμε, προκειμένου να επιλύσουμε το συγκεκριμένο πρόβλημα.

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 (Path) του οδικού δικτύου εντός ενός συγκεκριμένου χρονικού διαστήματος. Για την καλύτερη παρακολούθηση του προβλήματος, ορίζονται οι επόμενες έννοιες:

- ο όρος «**ακμή**» (Edge) 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**» (Moving Object) 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 (AEM). Η ιδέα πίσω από αυτή την μεθοδολογία αναλύεται στο επόμενο υποκεφάλαιο. Έπειτα, ανάγεται το πρόβλημα σε χρονοσειρές (Timeseries) 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 (Trajectories) 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 χρησιμοποιώντας μοντέλα μηχανικής και βαθιάς μάθησης. 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 [3]

Η πρώτη μελέτη που παραθέτουμε εξετάζει τη σημασία της βραχυπρόθεσμης πρόβλεψης της ροής της κυκλοφορίας στις ευφυείς μεταφορές (Intelligent Transport – IT) και την εφαρμογή της στη διαχείριση της κυκλοφοριακής συμφόρησης, τη μείωση της ρύπανσης και την ενίσχυση της οδικής ασφάλειας. Επιπροσθέτως, επισημαίνονται και 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.

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 προηγούμενα δεδομένα 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 (Dataset) 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 [4]

Στην δεύτερη εργασία που παραθέτουμε, 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 (Intelligent Transport Systems - ITS) 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.

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 - BN) και την εισαγωγή των επιπέδων αγνόησης (Dropout Layers). 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 (Mean Absolute Percentage Error - 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 [5]

Σε αυτή την τρίτη έρευνα που συμπεριλαμβάνουμε, 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.

Η μεθοδολογία που χρησιμοποιείται στην παρούσα εργασία περιλαμβάνει την εφαρμογή ενός μοντέλου στοιβαγμένου κωδικοποιητή - αποκωδικοποιητή (Stacked Autoencoder - SAE), ο 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 [6]

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 (ΚΑ) 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.

Επιπρόσθετα, οι συγγραφείς προτείνουν ένα μοντέλο βαθιάς μάθησης που ονομάζεται «DeepCrowd». Το τελευταίο βασίζεται σε μία πυραμιδική αρχιτεκτονική αποτελούμενη από συνελικτικά νευρωνικά δίκτυα με βαχυπρόθεσμη και μακροπρόθεσμη μνήμη (Convolutional LSTM) και μηχανισμούς προσοχής υψηλής διάστασης (High-Dimensional Attention Mechanisms - HDAM). 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, οι Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE) και Mean Absolute Percentage Error (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.

2.5 Χρήση Γραφημάτων και Νευρωνικών Δικτύων [7]

Στην επόμενη 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.

Το προτεινόμενο μοντέλο που αναπτύσσεται στην παρούσα μελέτη ονομάζεται Graph Hierarchical Convolutional Recurrent Neural Network (GHCRNN). 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 (convolutional layers)**: 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 οργάνωσης (pooling units)**: οι μονάδες αυτές (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 (encoder – decoder)**: η συνολική αρχιτεκτονική του GHCRNN χρησιμοποιεί ένα δίκτυο (Sequence to Sequence - 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 Gated Recurrent Unit (GRU). Αυτό επιτρέπει στο μοντέλο να συλλαμβάνει τόσο μακροπρόθεσμες όσο και βραχυπρόθεσμες πληροφορίες.
6. **συνέλιξη γράφων και διαδικασία οργάνωσης (convolution of graphs and 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 [8]

Στην τελευταία εργασία που παραθέτουμε, 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.

Μάλιστα, η έρευνα αυτή συνδυάζει την τεχνική της αποσύνθεσης κυματιδίων (Wavelet Decomposition - WD) 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. Δηλαδή, χαμηλότερες τιμές των δύο παραπάνω μετρικών δηλώνουν και καλύτερη επίδοση του μοντέλου.

3. 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.

3.1 Η Γλώσσα Python

Η Python είναι μια γλώσσα προγραμματισμού υψηλού επιπέδου που δημιουργήθηκε από τον Guido van Rossum και πρωτοκυκλοφόρησε το 1991 [9]. 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.

3.1.1 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 [10] και η Numerical Python (NumPy) [11] για την ανάλυση και τον χειρισμό των δεδομένων, ενώ η βιβλιοθήκη Scikit-learn [12] 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 [13] και το Keras [14]. Αυτές οι βιβλιοθήκες επιτρέπουν τη δημιουργία μοντέλων πρόβλεψης, ταξινόμησης, συσταδοποίησης και πολλών άλλων.
- **οπτικοποίηση δεδομένων**: η Python παρέχει βιβλιοθήκες όπως η Matplotlib [15] και η Seaborn [16] για τη δημιουργία γραφημάτων που βοηθούν στην οπτικοποίηση (Visualization) 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.

3.2 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.

3.2.1 Jupyter Notebook

Το Jupyter Notebook [17] 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 (Notebooks), 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.

3.2.2 Google Colab

Το Google Colab [18] 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.

3.3 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.

3.3.1 Γενικά Χαρακτηριστικά για την Βάση Δεδομένων

Η PostgreSQL, γνωστή και ως Postgres, αποτελεί ένα προηγμένο σχεσιακό Σύστημα Διαχείρισης Βάσεων Δεδομένων – ΣΔΒΔ (Database Management System – DBMS). 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.

3.3.2 Η Γλώσσα PL/pgSQL [19]

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.

3.4 Το Λογισμικό Docker

Το λογισμικό Docker [20] 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.

3.5 Το Λογισμικό Valhalla

Το λογισμικό Valhalla [21] 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.

3.5.1 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 [22] 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.

4. Μεθοδολογία

Στο κεφάλαιο αυτό γίνεται μία εκτενής ανάλυση της μεθοδολογίας που ακολουθήθηκε από την αρχή έως το τέλος, προκειμένου να οδηγηθούμε από τα απλά δεδομένα κυκλοφοριακής ροής που έχουμε στη διάθεσή μας, στις μελλοντικές προβλέψεις του κυκλοφοριακού φόρτου σε διάφορα μονοπάτια του οδικού δικτύου που έχουμε επιλέξει. Για την ευκολότερη κατανόηση της μεθοδολογίας, έχει γίνει τμηματοποίηση των σταδίων της σε υποκεφάλαια. Η εφαρμογή αυτής της μεθοδολογίας περιγράφεται στο πέμπτο κεφάλαιο.

4.1 Δεδομένα που Χρησιμοποιήθηκαν

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 κρίσιμης σημασίας 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.

4.1.1 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 [23]. 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 4.1: 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.

4.1.2 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 [24].

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.

4.2 Προεπεξεργασία Δεδομένων

Κατά το στάδιο της προεπεξεργασίας των δεδομένων, χρησιμοποιούμε τεχνικές που μπορούν να μετατρέψουν τα αρχικά δεδομένα κίνησης στο οδικό δίκτυο του San Francisco σε μία μορφή κατάλληλη για τις τεχνικές που θα αναφέρουμε σε λίγο.

Αρχικά, οι καταγραφές κίνησης για κάθε ταξί που κινείται μέσα στο οδικό δίκτυο της πόλης βρίσκονται σε ξεχωριστά αρχεία, τα οποία φέρουν τον τίτλο του εκάστοτε οχήματος. Τα αρχεία αυτά παραθέτουν δηλαδή την συνολική τροχιά που διένυσε κάθε ταξί. Όλες οι καταγραφές των δεδομένων κίνησης για κάθε όχημα έχουν συλλεχθεί κατά το διάστημα ολόκληρου του μήνα Μάιου, του έτους 2008. Για κάθε ταξί, η καταγραφή της θέσης του στο οδικό δίκτυο γίνεται ανά μερικά δευτερόλεπτα. Τα βήματα που πρέπει να γίνουν με τη σειρά κατά την προεπεξεργασία των δεδομένων είναι τα ακόλουθα:

1. δοθέντος αυτών των δεδομένων, απαραίτητο είναι να ενοποιηθούν όλα τα αρχεία των τροχιών κάθε ταξί σε ένα ενιαίο αρχείο - πίνακα.
2. στον νέο αυτό πίνακα, προσπαθούμε να διαχωρίσουμε τις ενιαίες τροχιές κάθε ταξί σε μικρότερες υποτροχιές (Sub Trajectory). Μετά τον διαχωρισμό των ενιαίων τροχιών, κάθε υποτροχιά θα περιέχει καταγραφές θέσεων που απέχουν όλες ανά δύο το πολύ **n** χρονικές μονάδες. Ο αριθμός n ορίζεται σε **δευτερόλεπτα**.
3. σε αυτό το βήμα έχει ήδη γίνει ο διαχωρισμός των τροχιών σε υποτροχιές. Ωστόσο, τα δεδομένα αυτά δεν έχουν αντιστοιχηθεί σωστά στο οδικό δίκτυο της πόλης, αφού κατά την συλλογή τους μπορεί να υπήρξε θόρυβος. Επομένως, μία διαδικασία αντιστοίχισης των τροχιών στο οδικό δίκτυο καθίσταται αναγκαία. Η διαδικασία αυτή γίνεται με το πρόγραμμα **Valhalla Meili**. Η έξοδος του προγράμματος αυτή θα περιέχει την πληροφορία των αντιστοιχιζόμενων τροχιών σε έναν νέο πίνακα.

4.3 Δημιουργία Χρονοσειρών

Ο τρόπος με τον οποίο προσεγγίζουμε το πρόβλημα στην πτυχιακή έρευνα είναι μέσω των δεδομένων χρονοσειράς. Ως χρονοσειρές ορίζουμε τα δεδομένα που έχουν διαταχθεί σε συνάρτηση με τον χρόνο. Με άλλα λόγια, αποτελούν δεδομένα που εξαρτώνται πλήρως από τον χρόνο.

Το επόμενο βήμα στην μεθοδολογία μας είναι να μετατρέψουμε τα δεδομένα που έχουμε στη διάθεσή μας σε χρονοσειρές. Υπενθυμίζεται ότι το πρόβλημα που επιλύεται είναι η πρόβλεψη της κυκλοφοριακής ροής σε ολόκληρα μονοπάτια εντός του οδικού δικτύου του San Francisco. Αρχικά, ορίζουμε **m** διαφορετικά μονοπάτια με τυχαία μήκη το κάθε ένα. Ο αριθμός m

προσδιορίζεται πάντα από το χρήστη και είναι θετικός αριθμός. Δεύτερο, κατακερματίζουμε το συνολικό χρονικό διάστημα του ενός μήνα (η συνολική διάρκεια καταγραφών για κάθε ταξί) σε μικρότερα ίσα χρονικά υποδιαστήματα. Επομένως, για κάθε μονοπάτι και για κάθε χρονικό υποδιάστημα, μετράμε την κυκλοφοριακή ροή μέσα σε αυτό χρησιμοποιώντας τον αλγόριθμο των ΑΕΜ. Το αποτέλεσμα θα είναι ένα καινούριο σύνολο δεδομένων, το οποίο για κάθε μονοπάτι θα αναφέρει το πλήθος των ταξί που το διένυσαν σε κάθε χρονικό υποδιάστημα, δημιουργώντας έτσι πληροφορία που εξαρτάται πλήρως από το μέγεθος του χρόνου.

Προκειμένου να καταστεί ευκολότερη η κατανόηση της μορφής ενός τέτοιου συνόλου, τα δεδομένα χρονοσειράς περιέχουν πληροφορίες όπως τα ονόματα των μονοπατιών, το μήκος τους και το πλήθος των ταξί που το διέτρεξαν σε κάθε ένα χρονικό διάστημα. Αυτές οι πληροφορίες αποθηκεύονται κατά μήκος των στηλών του. Από την άλλη, κατά μήκος των γραμμών είναι αποθηκευμένα τα χρονικά διαστήματα.

Μέχρι τώρα, το σύνολο δεδομένων περιέχει μόνο την πληροφορία της κυκλοφοριακής ροής μέσα σε ένα μονοπάτι για κάθε χρονικό διάστημα. Προκειμένου να καταστήσουμε ακριβέστερη την διαδικασία της πρόβλεψης, προσθέτουμε επιπλέον δεδομένα κατά μήκος των στηλών μέσα στα υπάρχοντα, συγχωνεύοντάς τα κατάλληλα, ώστε να μην υπάρχει αλλοίωση της πληροφορίας. Τα επιπλέον δεδομένα αναφέρονται σε δεδομένα **καιρού** και σε δεδομένα σχετικά με τον **χρόνο**, όπως η ώρα, η ημέρα, η εβδομάδα κ.α.

4.4 Η Τεχνική του Κυλιόμενου Παραθύρου

Σε προβλήματα, στα οποία προσπαθούμε να διαχειριστούμε χρονοσειρές προκειμένου να κάνουμε προβλέψεις, μία δημοφιλής τεχνική που εφαρμόζεται συχνά για την υποβοήθηση αυτής της διαδικασίας είναι το κυλιόμενο παράθυρο [25]. 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 κυλιόμενο παράθυρο δημιουργούνται διαδοχικά παράθυρα (δηλαδή τμήματα χρονοσειρών) 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 κυλιόμενο παράθυρο 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 (Training Set). Η ίδια ακριβώς τεχνική χρησιμοποιείται και στα δεδομένα ελέγχου (Test Set). Χρησιμοποιώντας διαφορετικά μήκη παραθύρου, δηλαδή το πόσα ιστορικά δεδομένα (Historical Data) 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.

Ως συνέχεια λοιπόν της μεθοδολογίας που εφαρμόζουμε, αφενός τμηματοποιούμε τις χρονοσειρές, εφαρμόζοντας σε κάθε μία από αυτές την παρούσα τεχνική, αφετέρου προσπαθούμε να προσδιορίσουμε το βέλτιστο μήκος παραθύρου (Window Length) για όλες τις χρονοσειρές. Για την εύρεση του βέλτιστου μήκους παραθύρου, χρησιμοποιούμε ένα μοντέλο μηχανικής μάθησης, το οποίο «μαθαίνει» τα δεδομένα που έχουν τμηματοποιηθεί σε κάθε περίπτωση με το συγκεκριμένο μήκος παραθύρου. Τελικά, επιλέγεται εκείνο το μήκος παραθύρου, για το οποίο η απόδοση του μοντέλου είναι βέλτιστη.

4.5 Χρήση Μοντέλων Μηχανικής Μάθησης

Το τελευταίο βήμα όλης της μεθοδολογίας αναφέρεται στην εκπαίδευση (Training) και την αξιολόγηση (Evaluation) διαφόρων μοντέλων μηχανικής και βαθιάς μάθησης. Στην έρευνά μας χρησιμοποιούμε, εκπαιδεύουμε και αξιολογούμε **τέσσερα** διαφορετικά μοντέλα μηχανικής μάθησης πάνω στα ίδια δεδομένα. Αυτά τα μοντέλα είναι ένα **XGBoost**, ένα **Random Forest**, ένα **νευρωνικό δίκτυο LSTM** και ένα **νευρωνικό δίκτυο Κωδικοποιητή - Αποκωδικοποιητή**. Περισσότερες λεπτομέρειες σχετικά με τον τρόπο λειτουργίας κάθε ενός από αυτούς τους αλγορίθμους βρίσκονται στο παράρτημα Γ του τόμου εργασίας. Κατά τον ορισμό των τεσσάρων αυτών μοντέλων, προσπαθούμε να βελτιστοποιούμε την συμπεριφορά τους στα συγκεκριμένα δεδομένα, εφαρμόζοντας διάφορες τεχνικές, οι οποίες προσδιορίζουν με τον καλύτερο δυνατό τρόπο τις ιδανικές υπερπαραμέτρους του εκάστοτε μοντέλου.

Αφού γίνει η εκπαίδευση και η αξιολόγηση των μοντέλων, για την διαδικασία των προβλέψεων (Forecast) του μεγέθους της κυκλοφοριακής ροής σε κάθε μονοπάτι επιλέγουμε το μοντέλο με την καλύτερη επίδοση. Η επίδοση του μοντέλου μετριέται με την βοήθεια δύο τύπων: το **Μέσο Απόλυτο Σφάλμα** (Mean Absolute Error – Mae) και τη **Ρίζα του Μέσου Τετραγωνικού Σφάλματος** (Root Mean Square Error – RMSE).

- 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, κάνοντας τις προβλέψεις να απέχουν σημαντικά από τις πραγματικές τιμές. 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 (Evaluation Metrics) 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. Το μοντέλο με τις χαμηλότερες τιμές για κάθε μία από τις δύο αυτές μετρικές επιλέγεται ως το πλέον κατάλληλο για την διαδικασία των προβλέψεων. Εδώ υπονοείται ότι όσο χαμηλότερος είναι ο δείκτης κάθε μίας μετρικής, τόσο ακριβέστερες είναι οι προβλέψεις που παράγει το μοντέλο πάνω στα συγκεκριμένα δεδομένα.

Το τελευταίο βήμα της μεθοδολογίας αυτής είναι η **παραγωγή των προβλέψεων**. Αρχικά, οι προβλέψεις επιλέγουμε να γίνουν εντός ενός μικρού χρονικού διαστήματος στο μέλλον (Short-Term Predictions), καθώς η κυκλοφοριακή ροή είναι ένα μέγεθος απρόβλεπτο και μη γραμμικό (Non-Linear). Επομένως, μακροπρόθεσμες προβλέψεις (Long-Term Predictions) του κυκλοφοριακού φόρτου δεν δύνανται να έχουν νόημα. Οι προβλέψεις επιλέγουμε να γίνουν k ώρες στο μέλλον. Το k ορίζεται να είναι ένας θετικός αριθμός μικρότερος του δέκα. Επίσης, ως μέλλον ορίζουμε τα χρονικά διαστήματα, για τα οποία δεν έχουμε καμία πληροφορία για την κυκλοφοριακή ροή σε κάθε μονοπάτι, αλλά προσπαθούμε να την προβλέψουμε.

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 πέντε αρχεία, τα οποία φέρουν τους τίτλους «**Notebook1**», «**Notebook2**», «**Notebook3**», «**Notebook4**» και «**Notebook5**»:

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.

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

Επίσης, 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.

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

Τέλος, στο πέμπτο αρχείο του κώδικα, εφαρμόζεται ολόκληρη η μεθοδολογία που αναφέραμε πάνω σε ένα σύνολο δεδομένων που έχει δημιουργηθεί χωρίς την χρήση της αλγοριθμικής διαδικασίας των ΑΕΜ. Στο αρχείο αυτό καθίσταται ξεκάθαρη η αναγκαιότητα χρήσης των ΑΕΜ για την επίλυση του γενικού προβλήματος της πρόβλεψης της κυκλοφοριακής ροής σε ολόκληρα μονοπάτια του οδικού δικτύου.

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 (Rows). 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» (Trajectory Splitting). 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 AEM, καταλήγουμε σε ένα τελικό σύνολο δεδομένων που αποτελείται από χρονοσειρές. Μάλιστα, έχουμε υλοποιήσει την ιδέα των AEM σε δύο προγραμματιστικά περιβάλλοντα, επιλέγοντας στο τέλος την ταχύτερη λύση. Στο παρόν υποκεφάλαιο εξετάζονται όλα τα βήματα κατασκευής του τελικού συνόλου δεδομένων.

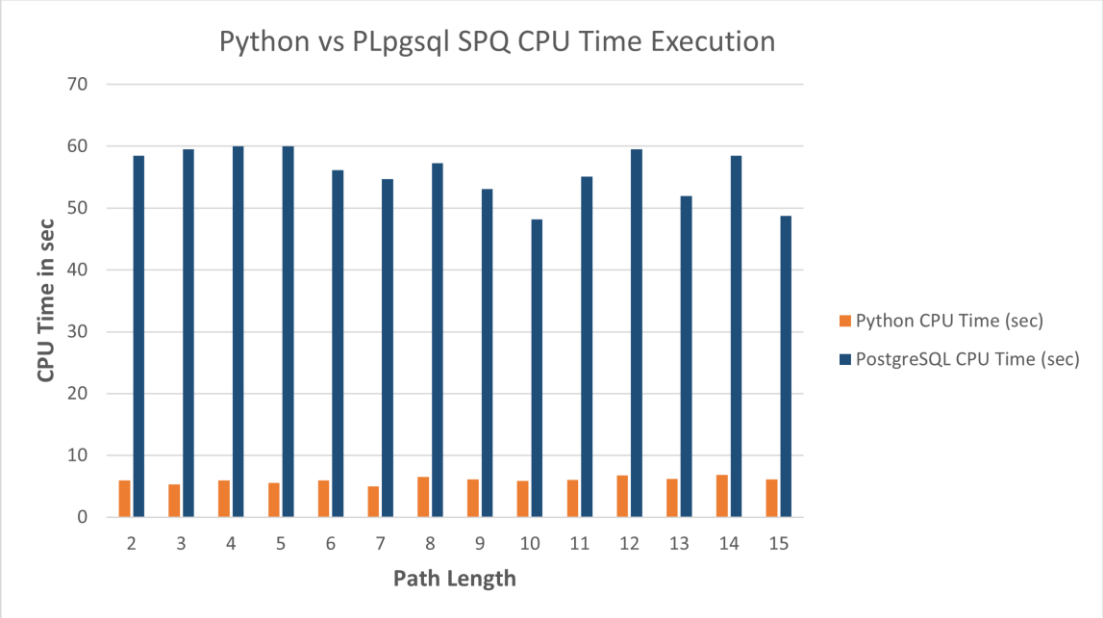
5.3.1 Υλοποίηση των Αυστηρών Ερωτημάτων Μονοπατιού

Ένα από τα πιο σημαντικά κομμάτια αυτής της πτυχιακής εργασίας είναι και η συγγραφή της συνάρτησης που υλοποιεί τα AEM. 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 γλώσσα **Python** σε μορφή συνάρτησης, ενώ η συνολική διάρκεια εκτέλεσης μίας μόνο 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 (Batches). 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 ΑΕΜ 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.

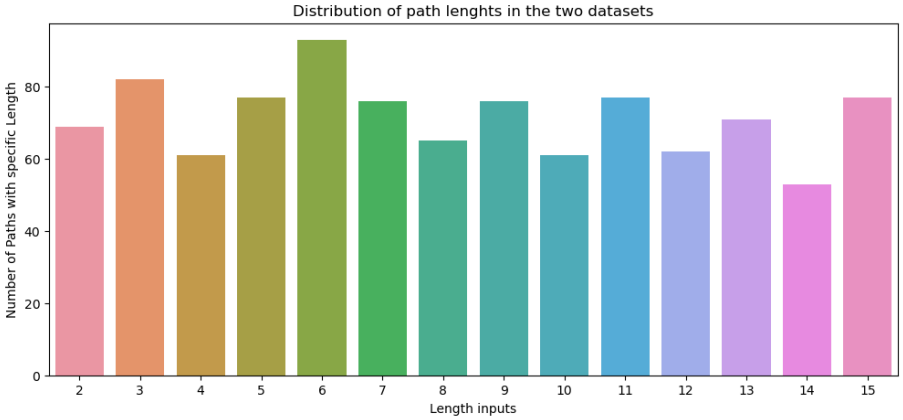
Συμπεραίνουμε ότι οι χρόνοι εκτέλεσης σε Python είναι ταχύτεροι, συγκριτικά με αυτούς σε PostgreSQL.

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 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 τέταρτο κεφάλαιο.

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

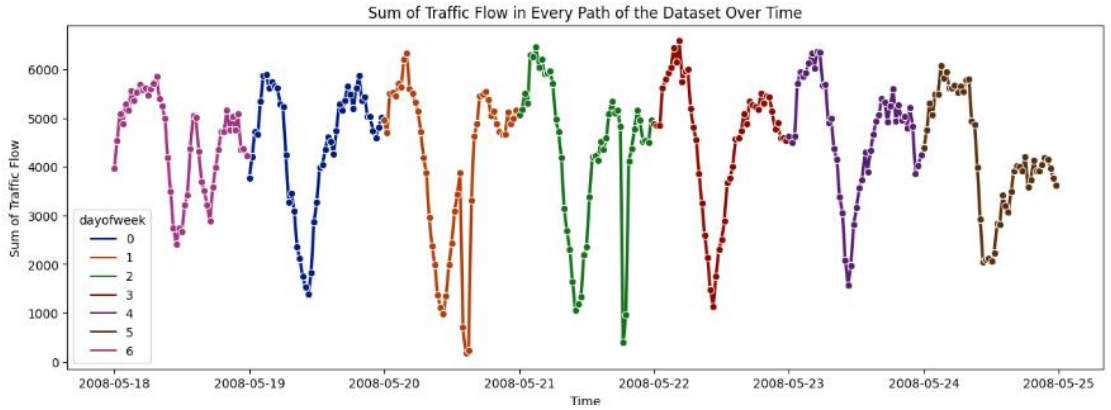
Πολλά χαρακτηριστικά που σχετίζονται με τον χρόνο εξάγονται από τις πληροφορίες χρονοσφραγίδων στο σύνολο δεδομένων. Τα χαρακτηριστικά αυτά περιλαμβάνουν την ώρα, την ημέρα της εβδομάδας, την ημέρα του μήνα και τα λεπτά. Εφαρμόζεται κυκλική κωδικοποίηση σε ορισμένα χαρακτηριστικά (ώρα, ημέρα της εβδομάδας, 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.

5.5 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.

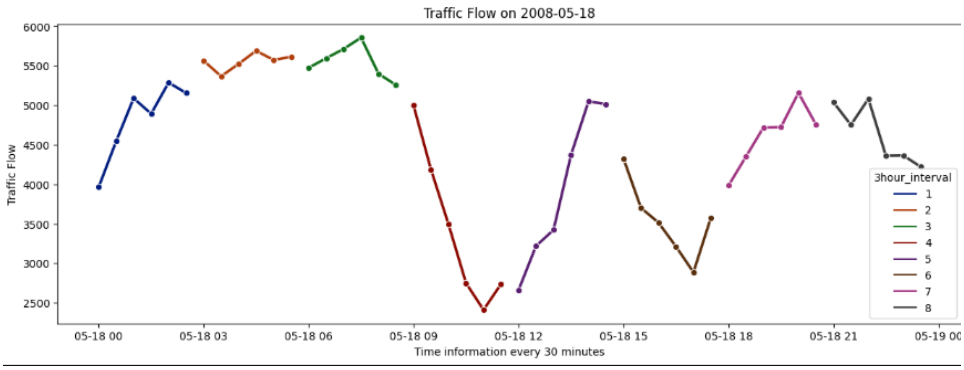


Διάγραμμα 5.3: Συνολική ροή κυκλοφορίας σε κάθε ημέρα.

Σε αυτό το διάγραμμα (Διάγραμμα 5.3), 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 5.4) 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 5.4 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.

5.6 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 περίπτωση.

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. Επιπρόσθετα, στο παράρτημα Γ αναφέρονται αναλυτικά οι τρόποι λειτουργίας κάθε ενός από τα τέσσερα μοντέλα που χρησιμοποιούμε.

5.6.1 Διαχωρισμός σε Σύνολα Εκπαίδευσης και Ελέγχου

Σε αυτό το βήμα, τα δεδομένα διαιρούνται σε δύο σύνολα: το σύνολο εκπαίδευσης και το σύνολο ελέγχου. 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 (Hyperparameters) 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 σύνολο επικύρωσης ταυτίζεται με το σύνολο ελέγχου 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

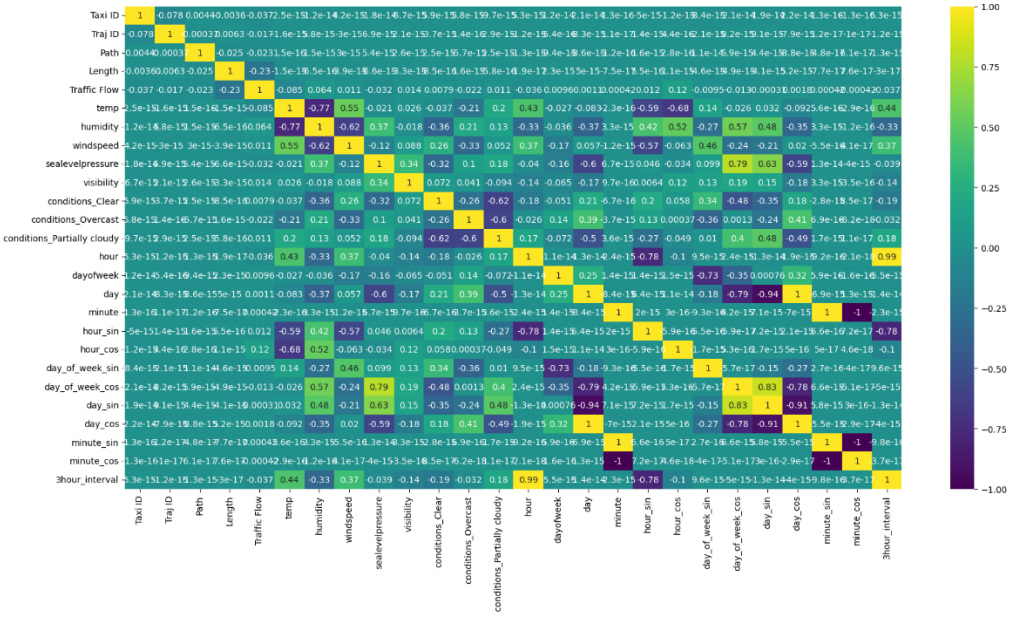
εκπαίδευσης να περιέχει όλα τα δεδομένα για κάθε μονοπάτι μέχρι και την 2008-05-23. Τα υπόλοιπα δεδομένα (τα πιο πρόσφατα) βρίσκονται στο σύνολο ελέγχου.

5.6.2 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 χαρακτηριστικά (Predictors or 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 5.5: 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.

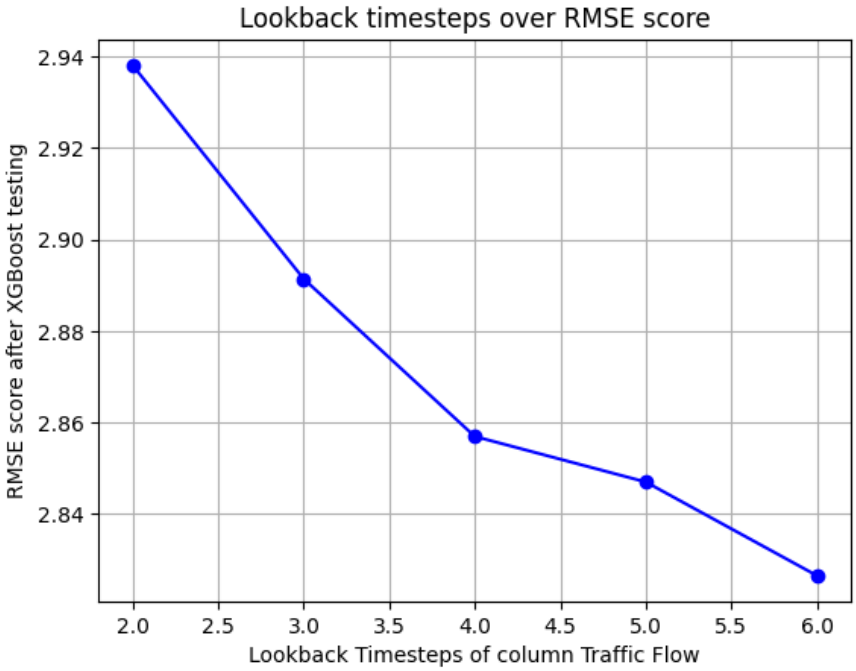
5.6.3 Εκπαίδευση των μοντέλων

Σε αυτήν την υποενότητα, αναφέρονται τα βασικά βήματα που ακολουθούμε, προκειμένου να εκπαιδεύσουμε τα τέσσερα μοντέλα μηχανικής μάθησης, ώστε να βρούμε εκείνον τον αλγόριθμο που μαθαίνει με τον καλύτερο τρόπο τις σχέσεις σε αυτά τα δεδομένα.

Αρχικά, όλα τα μοντέλα εκπαιδεύονται στο **ίδιο σύνολο δεδομένων**, χρησιμοποιώντας τα ίδια σύνολα εκπαίδευσης και ελέγχου, καθώς επίσης και τα ίδια features και labels. Το σύνολο label αναφέρεται στα χαρακτηριστικά προς πρόβλεψη. Στην προκειμένη περίπτωση, θεωρείται ως label η κυκλοφοριακή ροή.

Στην συνέχεια, καθίσταται απαραίτητες δύο ενέργειες: η πρώτη αφορά τον μετασχηματισμό των συνόλων εκπαίδευσης και ελέγχου σε πρόβλημα μηχανικής μάθησης, εφαρμόζοντας την τεχνική του συρόμενου παραθύρου. Πριν γίνει αυτό, είναι απαραίτητο να ευρεθεί το βέλτιστο μήκος αυτού του παραθύρου. Η διαδικασία αυτή γίνεται σε συνδυασμό με το μοντέλο XGBoost. Ο λόγος για τον οποίο επιλέγουμε αυτό το μοντέλο συγκεκριμένα, είναι επειδή ο αλγόριθμος αυτός είναι πολύ **ευέλικτος** και – ίσως – αποδειχτεί ο καλύτερος ως προς την απόδοση.

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

Παρατηρούμε, ότι όταν το μέγεθος παραθύρου περιλαμβάνει έξι χρονικά βήματα στο παρελθόν, η μετρική RMSE του μοντέλου XGBoost είναι βέλτιστη. Στο εξής, για κάθε μοντέλο που εκπαιδεύεται, θα θεωρούμε ότι λαμβάνει ως είσοδο έξι παρελθοντικές τιμές της κυκλοφοριακής ροής, προκειμένου να προβλέπει την αμέσως επόμενη τιμή του ίδιου μεγέθους.

Η δεύτερη ενέργεια που πρέπει να γίνει αφορά στην βελτιστοποίηση των μοντέλων. Κάθε μοντέλο που εξετάζουμε περιέχει ορισμένες υπερπαραμέτρους που ορίζονται από τον ίδιο τον χρήστη. Αυτό έχει ως αποτέλεσμα να ευρεθεί ένας τρόπος, ώστε τα μοντέλα αυτά να ορίζονται με τις βέλτιστες τιμές για κάθε μία υπερπαράμετρο που διαθέτουν. Ο τρόπος με τον οποίο ευρίσκουμε αυτές τις υπεραμέτρους είναι κοινός για τα μοντέλα XGBoost και Random Forest. Επίσης, με παρόμοιο, αλλά διαφορετικό τρόπο ευρίσκονται οι υπερπαράμετροι για το ζεύγος μοντέλων LSTM και Encoder-Decoder.

Στην περίπτωση των δύο πρώτων μοντέλων, χρησιμοποιείται η μέθοδος της **αναζήτησης πλέγματος πολλαπλής διεπικύρωσης** (Grid Search Cross Validation ή Grid Search CV). Η μέθοδος αυτή λειτουργεί εφαρμόζοντας τα παρακάτω βήματα:

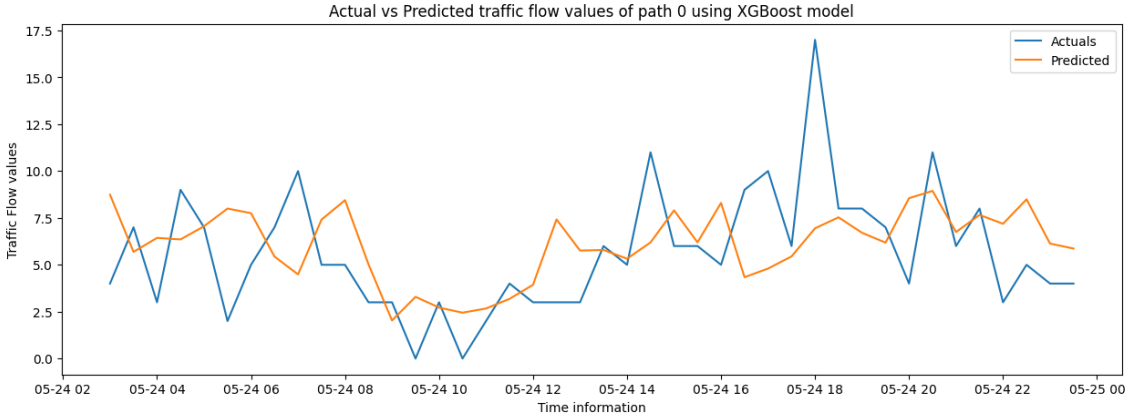
1. ορίζεται μια λίστα πιθανών τιμών για κάθε υπερπαράμετρο που χρειάζεται να βελτιστοποιηθεί.
2. η μέθοδος Grid Search CV δημιουργεί όλους τους δυνατούς συνδυασμούς των τιμών των υπερπαραμέτρων που έχουν καθοριστεί. Για παράδειγμα, εάν έχουν οριστεί δύο υπερπαράμετροι προς βελτιστοποίηση, καθεμία με 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 της πολλαπλής διεπικύρωσης (Cross Validation). Η μέθοδος αυτή θέλει το σύνολο εκπαίδευσης να διαχωρίζεται σε μικρότερα υποσύνολα, τα τμήματα (Folds). Κάθε φορά, ένα μόνο τμήμα 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.

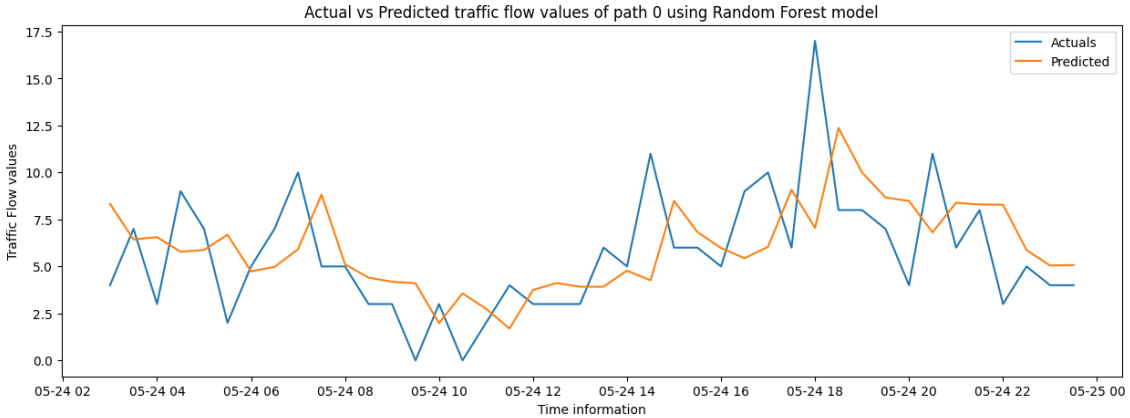
Συμπεραίνοντας, είναι κατανοητό ότι η παρούσα μεθοδολογία αποτελεί μια πολύ χρήσιμη τεχνική για την αυτοματοποίηση της διαδικασίας επιλογής υπερπαραμέτρων και τη βελτιστοποίηση της απόδοσης ενός μοντέλου.

Από την άλλη πλευρά, έγιναν βελτιστοποιήσεις και στα δύο μοντέλα νευρωνικών δικτύων. Για παράδειγμα, έχουν εισαχθεί **πυκνά επίπεδα** (Dense Layers) και **επίπεδα εγκατάλειψης** (Dropout Layers). Ο ρόλος των πρώτων είναι να βοηθούν το μοντέλο να παρατηρεί μη γραμμικές σχέσεις από τα δεδομένα, ενώ ο ρόλος των επιπέδων εγκατάλειψης είναι να «απενεργοποιούν» κατά την φάση της εκπαίδευσης έναν συγκεκριμένο αριθμό νευρώνων του δικτύου, προκειμένου να αποφευχθεί το φαινόμενο της υπερεκπαίδευσης.

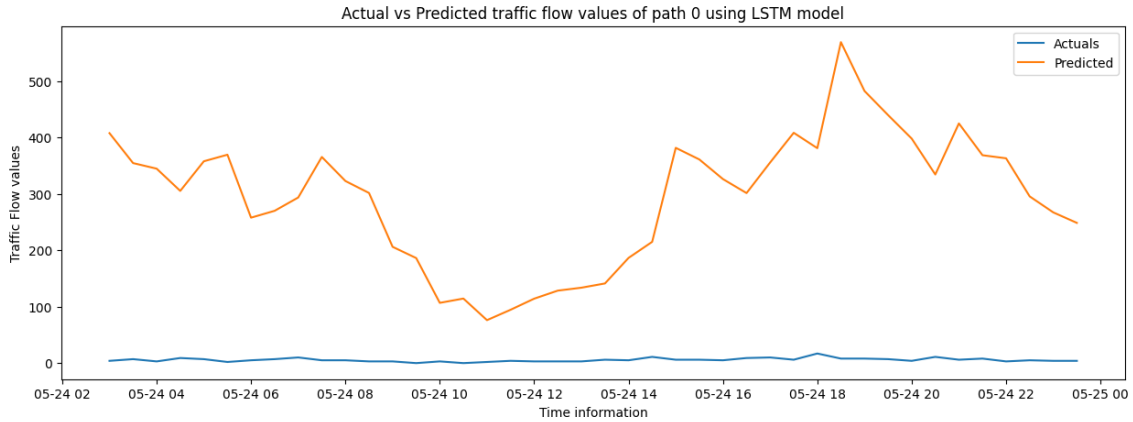
Μετά από όλη αυτή την διαδικασία που περιεγράφηκε, η οποία συνοψίζεται στα επόμενα βήματα: διαχωρισμός των δεδομένων σε σύνολα εκπαίδευσης και ελέγχου (βήμα 1), επιλογή των καλύτερων χαρακτηριστικών (σύμφωνα με την μήτρα συσχέτισης) που θα βοηθήσουν την εκπαίδευση του μοντέλου (βήμα 2), εφαρμογή του συρόμενου παραθύρου στα δεδομένα με το βέλτιστο μήκος (βήμα 3), χρήση τεχνικών βελτιστοποίησης των μοντέλων (βήμα 4), αναπαρίστανται στα επόμενα γραφήματα οι επιδόσεις των μοντέλων στα δεδομένα ελέγχου . Τα γραφήματα αυτά απεικονίζουν την πραγματική (μπλε χρώμα) και προβλεπόμενη τιμή (πορτοκαλί χρώμα) της κυκλοφοριακής ροής στο μονοπάτι με κωδικό μηδέν (0).



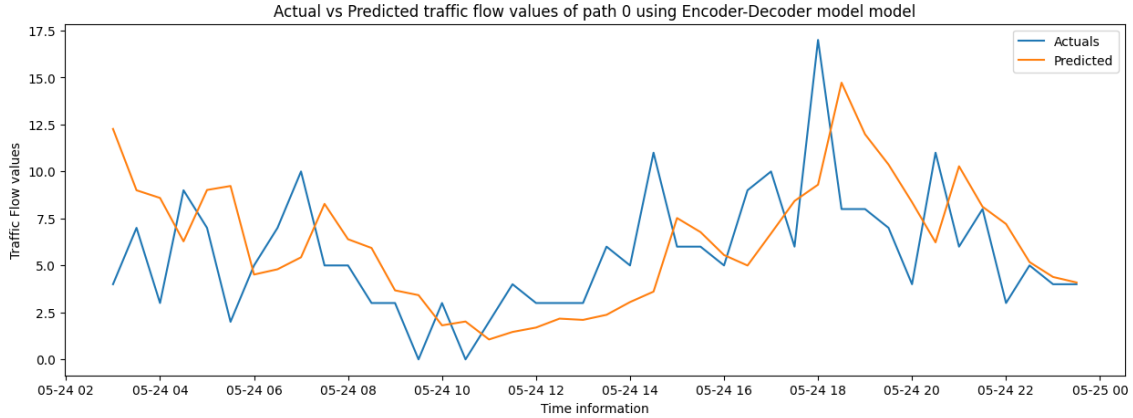
Διάγραμμα 5.7: Επίδοση του μοντέλου XGBoost στο σύνολο ελέγχου.



Διάγραμμα 5.8: Επίδοση του μοντέλου Random Forest στο σύνολο ελέγχου.



Διάγραμμα 5.9: Επίδοση του μοντέλου LSTM στο σύνολο ελέγχου.



Διάγραμμα 5.10: Επίδοση του μοντέλου Encoder-Decoder στο σύνολο ελέγχου.

Σε γενικές γραμμές, όλα τα μοντέλα – πλην του LSTM – έχουν καταφέρει να αναπαραστήσουν ικανοποιητικά τη συμπεριφορά της χρονοσειράς που αντιστοιχεί στο μονοπάτι με κωδικό μηδέν. Με τον όρο «συμπεριφορά», εννοούμε την εξέλιξη της χρονοσειράς, πότε δηλαδή παρουσιάζει καθόδους και ανόδους. Επίσης, εννοούμε και τις επαναλαμβανόμενες συμπεριφορές που παρατηρούνται ανά συγκεκριμένα χρονικά διαστήματα.

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	2.806625	1.726323
LSTM	3.073994	1.999086
Random Forest	2.874276	1.782795
Encoder Decoder	3.337230	2.116786

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

Σύμφωνα με τον παραπάνω πίνακα, εξάγουμε δύο συμπεράσματα. Πρώτο, το μοντέλο XGBoost είναι το καλύτερο μοντέλο: επιλύει δηλαδή το πρόβλημα με καλύτερες αποδόσεις σε σχέση με τα υπόλοιπα μοντέλα. Δεύτερο, 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.

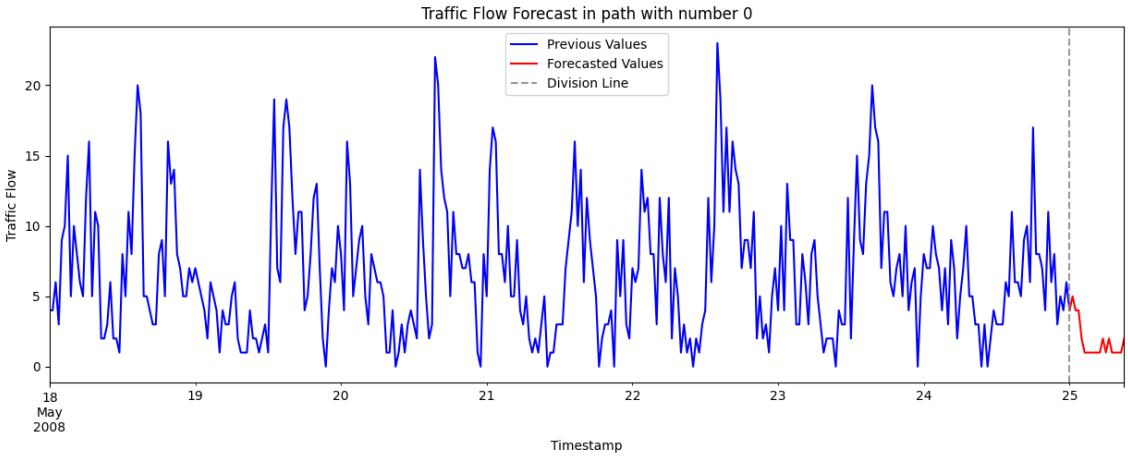
5.7 Αποτελέσματα

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:

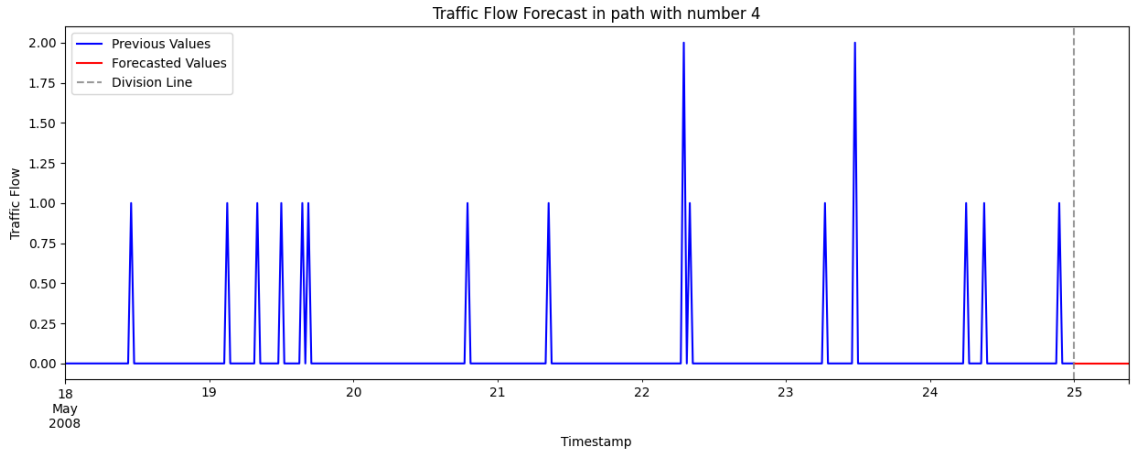
- έχουν δημιουργηθεί χρησιμοποιώντας τον αλγόριθμο XGBoost, λόγω του χαμηλότερου RMSE score που έχει, έναντι των υπολοίπων τριών μοντέλων.
- είναι βραχυπρόθεσμες. Αυτό συμβαίνει, διότι η κυκλοφοριακή ροή είναι ένα μέγεθος μη γραμμικό και πολυδιάστατο (Multidimensional): εξαρτάται, δηλαδή από πολλούς 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. κυλιόμενο παράθυρο) 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.

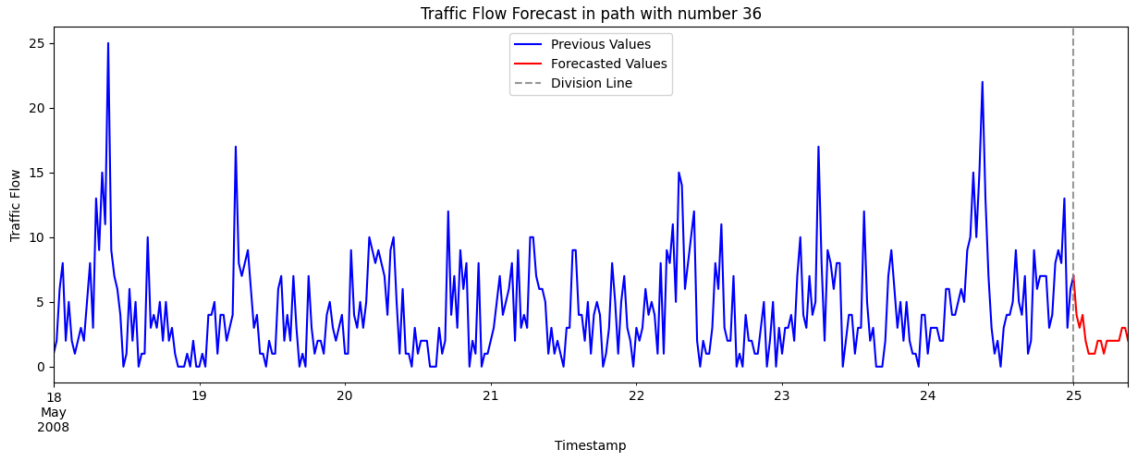
Στα παρακάτω διαγράμματα επιλέγουμε τυχαία μονοπάτια, στα οποία αναπαρίσταται οι γνωστές τιμές του μεγέθους της κυκλοφοριακής ροής (με μπλε χρώμα) και οι προβλεπόμενες τιμές του ίδιου μεγέθους εννέα ώρες στο μέλλον (με κόκκινο χρώμα).



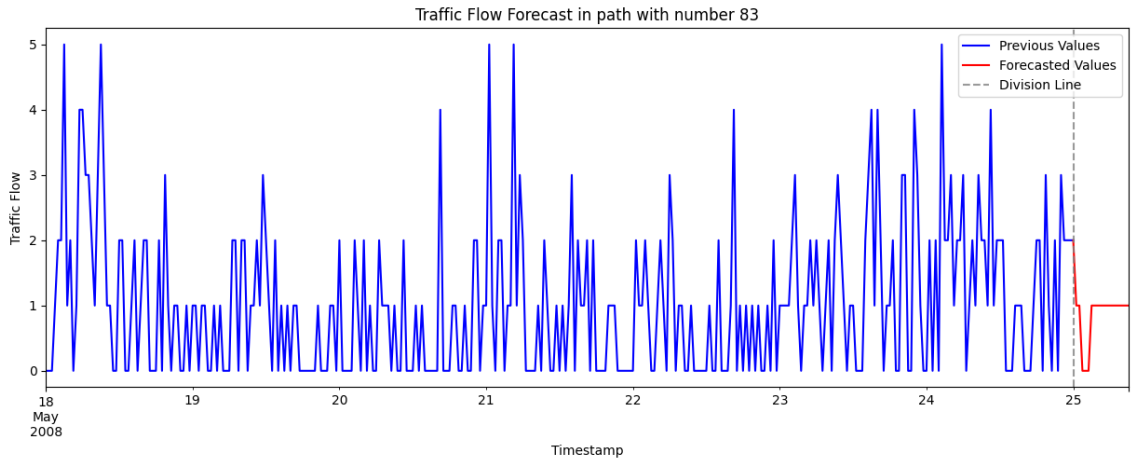
Διάγραμμα 5.11: Γνωστές και προβλεπόμενες τιμές του μεγέθους της κυκλοφοριακής ροής στο μονοπάτι με αριθμό 0.



Διάγραμμα 5.12: Γνωστές και προβλεπόμενες τιμές του μεγέθους της κυκλοφοριακής ροής στο μονοπάτι με αριθμό 4.



Διάγραμμα 5.13: Γνωστές και προβλεπόμενες τιμές του μεγέθους της κυκλοφοριακής ροής στο μονοπάτι με αριθμό 36.



Διάγραμμα 5.14: Γνωστές και προβλεπόμενες τιμές του μεγέθους της κυκλοφοριακής ροής στο μονοπάτι με αριθμό 83.

6. Συμπεράσματα και Προτάσεις για Βελτίωση

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.

Σε όλη το κείμενο τονίζουμε την αναγκαιότητα χρήσης της μεθοδολογίας των ΑΕΜ, προκειμένου να μετρήσουμε την κυκλοφοριακή ροή μέσα σε ολόκληρα μονοπάτι και όχι σε απλά τμήματα μεταξύ διασταυρώσεων. Αναφέρουμε ξανά ότι η επιλογή των ΑΕΜ στο πρόβλημα της πρόβλεψης 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. Παρακάτω, παρουσιάζονται δύο λόγοι, για τους οποίους η χρήση των ΑΕΜ είναι αναγκαία. Στην πρώτη περίπτωση δείχνουμε την διαφορά στην πληροφορία που έχουμε όταν χρησιμοποιούμε τα ΑΕΜ και όταν η χρήση τους απουσιάζει. Στην δεύτερη περίπτωση, εκπαιδεύουμε ξανά τον αλγόριθμο XGBoost στα δεδομένα που έχουν παρασκευαστεί χωρίς την χρήση των ΑΕΜ και συγκρίνουμε τις επιδόσεις του με την περίπτωση του κεφαλαίου πέντε.

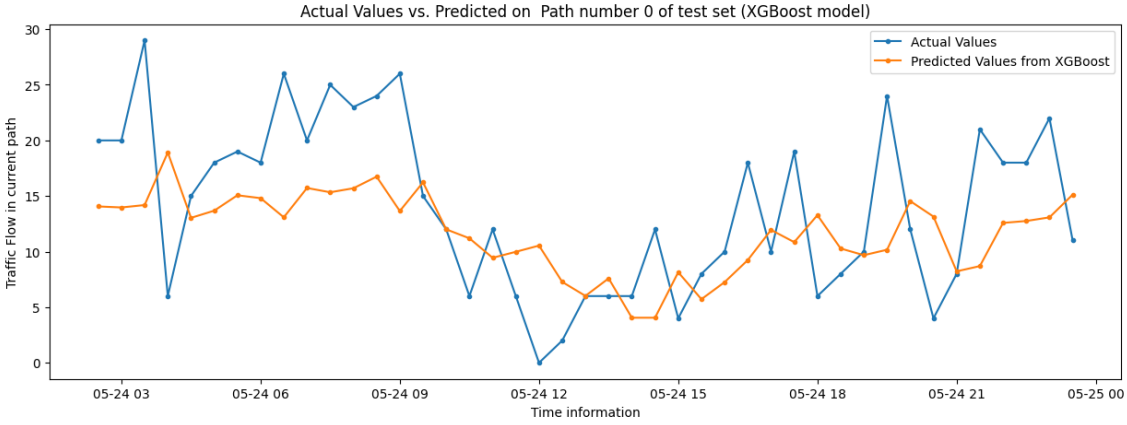
Για να ευρεθεί η διαφορά στα δεδομένα που συλλέγουμε όταν απουσιάζει η χρήση της μεθοδολογίας των ΑΕΜ, δημιουργούμε ένα καινούριο σύνολο δεδομένων χρονοσειρών. Η κύρια διαφορά τώρα είναι ότι σε κάθε μονοπάτι δεν μετράμε την κυκλοφοριακή ροή με την βοήθεια της συνάρτησης SPQ (παράρτημα Α). Σε αυτή την 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. Επομένως, δεν μας ενδιαφέρει εάν το κινούμενο όχημα παρέκκλινε της πορείας του ή εάν διέτρεξε τις ακμές του μονοπατιού με διαφορετική σειρά.

Το παρακάτω γράφημα δείχνει την διαφορά ανάμεσα στα δύο σύνολα δεδομένων. Με μπλε χρώμα απεικονίζεται η πληροφορία που υπάρχει στο πρώτο σύνολο δεδομένων εφαρμόζοντας την μέθοδο των ΑΕΜ, ενώ με πορτοκαλί χρώμα παρουσιάζεται η καταγραφή των δεδομένων κίνησης (δεύτερο σύνολο δεδομένων) όταν δεν γίνεται χρήση αυτής της μεθόδου. Επίσης, για κάθε χρονική στιγμή έχουμε συμπεριλάβει το άθροισμα όλων των ταξί που διένυσαν όλα τα μονοπάτια του συνόλου δεδομένων.

Διάγραμμα 6.1: 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 - Trend, περιοδικότητα - Seasonality) 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.

Επιπρόσθετα, έχουμε εκπαιδεύσει τον αλγόριθμο XGBoost στο σύνολο δεδομένων που έχει δημιουργηθεί εν απουσία της μεθόδου των AEM. Η διαδικασία εκπαίδευσης και αξιολόγησης είναι παρόμοια με αυτή που έχει εξηγηθεί στο κεφάλαιο τέσσερα. Μάλιστα, το μοντέλο αυτό έχει εκπαιδευτεί από την αρχή πάνω σε αυτά τα νέα δεδομένα ειδικά. Παρακάτω, φαίνονται οι πραγματικές τιμές της κυκλοφοριακής ροής (μπλε χρώμα) και οι αντίστοιχες προβλεπόμενες τιμές (πορτοκαλί χρώμα) για το μονοπάτι με κωδικό μηδέν.



Διάγραμμα 6.2: Απόδοση του αλγορίθμου XGBoost στο σύνολο ελέγχου. Το σύνολο αυτό περιέχει δεδομένα που έχουν παρασκευαστεί χωρίς την χρήση των AEM.

Η απόδοση του μοντέλου XGBoost σε όρους RMSE και MAE παρουσιάζεται στον επόμενο πίνακα:

Model	RMSE Score	MAE Score
XGBoost	16.972748	8.40884

Εικόνα 6.1: Απόδοση του μοντέλου XGBoost

Παρατηρούμε ότι όταν χρησιμοποιούμε δεδομένα που έχουν προέλθει ως εφαρμογή της μεθόδου των AEM, το μοντέλο XGBoost ανταποκρίνεται πολύ καλύτερα, δίνοντας ως έξοδο πιο

αξιόπιστες προβλέψεις. Αυτό συμβαίνει διότι τα AEM μετράνε την κίνηση μέσα στο μονοπάτι αυτή κάθε αυτή, απαλλάσσοντας το σύνολο δεδομένων από θόρυβο. Έτσι, ο αλγόριθμος μηχανικής μάθησης που εφαρμόζεται είναι φυσιολογικό να παρουσιάζει καλύτερη απόδοση.

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** (Spatial) και **χρονική διάταξη** (Temporal) 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. Με τις προαναφερθείσες προσεγγίσεις, οι προβλέψεις ενδέχεται να παρουσιάζουν αυξημένη ακρίβεια.

Πίνακας Ορολογιών

Σε αυτό το τμήμα του τόμου, παραθέτουμε την μετάφραση των ξενόγλωσσων όρων που υπάρχουν στο κείμενο στην ελληνική γλώσσα:

Ξενόγλωσσος όρος	Ελληνικός όρος
Artificial Intelligence	Τεχνητή Νοημοσύνη
Batch	Δέσμη
Big Data	Μεγάλος Όγκος Δεδομένων
Classification	Ομαδοποίηση / Συσταδοποίηση
Convolutional	Συνελικτικός
Container	Κουτί
Correlation Matrix	Μήτρα Συσχέτισης
Dataset	Σύνολο Δεδομένων
Date	Ημερομηνία
Decision Tree	Δέντρο Απόφασης
Decoder	Αποκωδικοποιητής
Deep Learning	Βαθιά Μάθηση
Deep Neural Network	Βαθύ Νευρωνικό Δίκτυο
Dense Layer	Πυκνό Στρώμα
Dropout Layer	Στρώμα Εγκατάλειψης
Edge	Ακμή
Encoder	Κωδικοποιητής

End Time	Χρόνος Τέλους
Ensemble Learning	Μάθηση συνόλου
Evaluation	Αξιολόγηση
Evaluation Metric	Μετρική Αξιολόγησης
Features	Χαρακτηριστικά
Forecast	Πρόβλεψη
Forget Gate	Πύλη Λησμόνησης
Gate	Πύλη
Gradient Boosting	Κάθοδος Βαθμίδας
Grid Search Cross Validation	Αναζήτηση πλέγματος με την μέθοδο της πολλαπλής διεπικύρωσης
Historical Data	Ιστορικά Δεδομένα
Hyperparameter	Υπερπαράμετρος
Index	Ευρετήριο
Input Gate	Πύλη Εισόδου
Intelligent Transport System	Ευφυή Συστήματα Μεταφορών
Label	Μέγεθος προς Πρόβλεψη
Latitude	Γεωγραφικό Πλάτος
Leaf Node	Φύλλο του Δέντρου
Length	Μήκος
Linear Regression	Γραμμική Παλινδρόμηση
Longitude	Γεωγραφικό Μήκος
Long-Term Prediction	Μακροπρόθεσμη Πρόβλεψη
Machine Learning	Μηχανική Μάθηση
Mean Absolute Error	Μέσο Απόλυτο Σφάλμα
Mean Absolute Percentage Error	Μέσο Απόλυτο Ποσοστιαίο Σφάλμα
Memory Cell	Κύτταρο Μνήμης
Model	Μοντέλο
Moving Object	Κινούμενο Αντικείμενο
Multidimensional	Πολυδιάστατος
Neural Network	Νευρωνικό Δίκτυο
Non-Linear	Μη Γραμμικός
Notebook	Σημειωματάριο
Occupied	Κατειλημμένος
Overfitting	Υπερεκπαίδευση
Output Gate	Πύλη Εξόδου
Path	Μονοπάτι
Pooling Unit	Μονάδα Ομαδοποίησης
Regression	Παλινδρόμηση
Root Mean Square Error	Ρίζα Μέσου Τετραγωνικού Σφάλματος
Root Node	Ριζικός Κόμβος
Row	Εγγραφή / Γραμμή
Seasonality	Περιοδικότητα

Server	Διακομιστής
Short-Term Prediction	Βραχυπρόθεσμη Πρόβλεψη
Sliding Window	Κυλιόμενο/Συρόμενο Παράθυρο
Spatial	Χωρικός
Start Time	Χρόνος Έναρξης
Strict Path Queries	Αυστηρά Ερωτήματα Μονοπατιού
Strong Learner	Ισχυρός μαθητής
Sub Trajectory	Υποτροχιά
Temporal	Χρονικός
Test Set	Σύνολο Ελέγχου
Time	Χρόνος
Timeseries	Χρονοσειρά
Traffic Flow	Κυκλοφοριακή Ροή
Training	Εκπαίδευση
Train Set	Σύνολο Εκπαίδευσης
Trajectory Splitting	Διαχωρισμός Τροχιάς
Trend	Τάση
Validation Set	Σύνολο Επικύρωσης
Visualization	Οπτικοποίηση
Weak Learner	Μη ισχυρός μαθητής
Window Length	Μέγεθος Παραθύρου

Πίνακας Συντμήσεων – Αρκτικόλεξων – Ακρωνύμιων

Σε αυτό το κεφάλαιο, δίνεται η πλήρης αντιστοιχία για κάθε ένα ακρωνύμιο που συναντάται στο κείμενο.

Αρκτικόλεξο	Πλήρης Σημασία
AEM	Ακριβές Ερώτημα Μονοπατιού
ΚΑ	Κινούμενο Αντικείμενο
ΣΔΒΔ	Σύστημα Διαχείρισης Βάσεων Δεδομένων
ARIMA	Autoregressive Integrated Moving Average
DBMS	Database Management System
Edge ID	Edge Identification Number
GPS	Global Positioning System
GRU	Gated Recurrent Unit
IP	Internet Protocol
ITS	Intelligent Transport Systems
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
NN	Neural Network
OSM	Open Street Map
OSM Way ID	Open Street Map Way Identification Number
PL/pgSQL	Procedural Language/PostgreSQL

RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
RW	Random Walk
SAE	Stacked Autoencoder
SARIMA	Seasonal AutoRegressive Integrated Moving Average
Seq2Seq	Sequence to Sequence
SPQ	Strict Path Query
SVM	Support Vector Machine
Taxi ID	Taxi Identification Number
Traj ID	Trajectory Identification Number
WD	Wavelet Decomposition

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.

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

Διαδικτυακές Αναφορές

[9] ‘Python History - javatpoint’. https://www.javatpoint.com/python-history.

[10] ‘pandas - Python Data Analysis Library’. https://pandas.pydata.org/.

[11] ‘NumPy’. https://numpy.org/.

[12] ‘scikit-learn: machine learning in Python — scikit-learn 1.3.0 documentation’. https://scikit-learn.org/stable/.

[13] ‘TensorFlow’. https://www.tensorflow.org/.

[14] ‘Keras: Deep Learning for humans’. https://keras.io/.

[15] ‘Matplotlib — Visualization with Python’. https://matplotlib.org/.

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

[17] ‘Project Jupyter | Home’. https://jupyter.org/.

[18] ‘Καλώς ορίσατε στο Colaboratory - Colaboratory’. https://colab.research.google.com/.

[19] ‘PostgreSQL: Documentation: 15: Chapter 43. PL/pgSQL — SQL Procedural Language’. https://www.postgresql.org/docs/current/plpgsql.html.

[20] ‘Docker: Accelerated Container Application Development’. https://www.docker.com/.

[21] ‘Valhalla Docs’. https://valhalla.github.io/valhalla/.

[22] ‘OpenStreetMap’. https://www.openstreetmap.org/#map=7/38.359/23.810.

[23] ‘Cabspotting | Stamen’. https://stamen.com/work/cabspotting/.

[24] ‘Weather Data Services | Visual Crossing’. https://www.visualcrossing.com/weather/weather-data-services.

[25] ‘Time Series Forecasting as Supervised Learning - MachineLearningMastery.com’. https://machinelearningmastery.com/time-series-forecasting-supervised-learning/.

[27] ‘Encoder Decoder What and Why ? - Simple Explanation’. https://inside-machinelearning.com/en/encoder-decoder-what-and-why-simple-explanation/.

[28] ‘Machine Learning Random Forest Algorithm - Javatpoint’. https://www.javatpoint.com/machine-learning-random-forest-algorithm.

Παραρτήματα

Τα παραρτήματα αποτελούν σημαντικό τμήμα αυτής της πτυχιακής εργασίας. Σε αυτά συμπεριλαμβάνονται πρόσθετες πληροφορίες και υλικό που είναι σχετικό με την έρευνα, αλλά δεν είναι αναγκαίο να ανακατεύονται με το κυρίως κείμενο. Συγκεκριμένα, στα παραρτήματα Α, Β και Γ αναφέρονται αναλύσεις για τους αλγορίθμους και τα μοντέλα που χρησιμοποιήθηκαν.

Παράρτημα Α: Η Συνάρτηση SPQ

Στο παράτημα αυτό του τόμου εργασίας περιγράφεται ο αλγόριθμος των AEM 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 AEM. 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: Η Συνάρτηση SPQ σε Γλώσσα PL/pgSQL

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.

Παράρτημα Γ: Τα Μοντέλα LSTM, Encoder – Decoder και Random Forest

Σε αυτό το τμήμα του τόμου εργασίας παρατίθενται περισσότερες πληροφορίες σχετικά με τα μοντέλα που εκμεταλλευτήκαμε για την διαδικασία της πρόβλεψης. Συγκεκριμένα, αναλύουμε τον τρόπο λειτουργίας των ακόλουθων μοντέλων: **XGBoost**, **Random Forest, Encoder–Decoder** και **LSTM**.

Ο Αλγόριθμος XGBoost

Ο αλγόριθμος XGBoost (Extreme Gradient Boosting) είναι ένας πανίσχυρος αλγόριθμος μηχανικής μάθησης που δημιουργήθηκε από τον Tianqi Chen [26]. 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 (Κάθοδος Βαθμίδας) εμπίπτει στην κατηγορία της μάθησης συνόλου (Ensemble Learning), η οποία συνδυάζει ασθενέστερα μοντέλα δέντρων (Weak 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. 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 την υπερεκπαίδευση.
3. υπερπαράμετρος **n_estimators**: καθορίζει τον τελικό αριθμό των δέντρων 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 υπερεκπαίδευση. 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.

Μοντέλο LSTM [3],

Το μοντέλο LSTM (Long Short-Term Memory) ανήκει στην κατηγορία των αναδρομικών νευρωνικών δικτύων (Recurrent Neural Networks - RNN). Επομένως, πρόκειται για έναν αλγόριθμο **βαθιάς μάθησης**. Συγκεκριμένα, αυτό το μοντέλο ειδικεύεται στην επεξεργασία και την πρόβλεψη δεδομένων, όπου ο χρόνος έχει σημασία.

Το νευρωνικό αυτό δίκτυο περιλαμβάνει τις ακόλουθες βελτιστοποιήσεις, για να προβαίνει σε προβλέψεις:

- **κύτταρα μνήμης (Memory Cells)**: το κύριο στοιχείο ενός LSTM είναι τα κύτταρα μνήμης, τα οποία είναι σε θέση να αποθηκεύουν πληροφορίες για μακροπρόθεσμα και βραχυπρόθεσμα χρονικά γεγονότα. Αυτά τα κύτταρα μπορούν να διατηρούν, να ενημερώνουν και να διαγράφουν πληροφορίες που επεξεργάζεται ο αλγόριθμος.
- **πύλες (Gates)**: οι πύλες στα LSTM είναι υπεύθυνες για τον έλεγχο της ροής των πληροφοριών μέσα στα κύτταρα μνήμης. Υπάρχουν τρεις κύριες πύλες στο μοντέλο LSTM:
    - **πύλη εισόδου (Input Gate)**: αποφασίζει ποιες πληροφορίες που έχουν δοθεί ως είσοδο θα εισαχθούν, τελικά, στα κύτταρα μνήμης.

- **πύλη εξόδου (Output Gate)**: αποφασίζει ποιες πληροφορίες από τα κύτταρα μνήμης θα δοθούν στην έξοδο του δικτύου.
- **πύλη λησμόνησης (Forget Gate)**: αποφασίζει ποιες πληροφορίες στα κύτταρα μνήμης θα διαγραφούν ή θα ξεχαστούν.

Κατά την φάση της εκπαίδευσης του αλγορίθμου, το LSTM μαθαίνει να προσαρμόζει διάφορες εσωτερικές παραμέτρους που διαθέτει, προκειμένου να επεξεργάζεται αποτελεσματικά τα χρονοσειριακά δεδομένα.

Άξιο παρατήρησης είναι ότι οι εφαρμογές των LSTM είναι πολλές την σήμερον ημέρα. Αυτές συμπεριλαμβάνουν προβλέψεις της καθημερινότητας, όπως: η πρόβλεψη των τιμών των μετοχών μίας εταιρείας, η φωνητική αναγνώριση, η μετάφραση κειμένου, η πρόβλεψη του καιρού κ.α.

Μοντέλο Encoder - Decoder [27]

Ένα μοντέλο Encoder-Decoder βασίζεται είτε στη χρήση LSTM (Long Short-Term Memory) μονάδων ή άλλων αντίστοιχων αρχιτεκτονικών όπως GRU (Gated Recurrent Unit) μονάδων για την επίλυση διαφόρων προβλημάτων, όπως η μετάφραση μηχανής, η σύνθεση κειμένου, η αναγνώριση προτύπων, κ.α. Στην πτυχιακή έρευνα συνδυάζουμε το μοντέλο αυτό με τα δίκτυα LSTM. Τα δομικά στοιχεία ενός μοντέλου κωδικοποιητή-αποκωδικοποιητή είναι τα επόμενα:

- **Τμήμα Encoder (Κωδικοποιητή):** Ο κωδικοποιητής αποδέχεται μία είσοδο στη μορφή ακολουθίας (π.χ., κείμενο σε φυσική γλώσσα). Κάθε τμήμα της εισόδου εισέρχεται στο τμήμα LSTM του κωδικοποιητή ένα προς ένα. Μετά την είσοδο, και αφού κάθε τμήμα διατρέξει την μονάδα LSTM, το μοντέλο κωδικοποιητή-αποκωδικοποιητή διατηρεί μια εσωτερική κατάσταση (Internal State) που περιέχει την πληροφορία από την είσοδο σε κωδικοποιημένη μορφή**.**
- **Τμήμα Decoder (Αποκωδικοποιητή)**: Ο αποκωδικοποιητής δέχεται την έξοδο του κωδικοποιητή ως είσοδο και ξεκινά τη διαδικασία παραγωγής της έξοδου του μοντέλου. Στον αποκωδικοποιητή γίνονται οι τελικές προβλέψεις του μοντέλου βασιζόμενες στην είσοδο που δίνεται κάθε φορά σε αυτόν.

Κατά την φάση της εκπαίδευσης, το μοντέλο χρησιμοποιεί ένα σύνολο ζευγαριών εισόδου-εξόδου για να προσαρμόσει τα βάρη του κωδικοποιητή και του αποκωδικοποιητή. Τόσο ο κωδικοποιητής, όσο και ο αποκωδικοποιητής είναι ένα μοντέλο LSTM ή GRU. Ουσιαστικά, το μοντέλο Encoder-Decoder στο σύνολό του συνδυάζει δίκτυα LSTM ή GRU και μία εσωτερική κατάσταση, για να προβαίνει σε προβλέψεις.

Μοντέλο Random Forest [28]

Ο αλγόριθμος Random Forest που χρησιμοποιείται στην πτυχιακή εργασία, είναι ένας αλγόριθμος μηχανικής μάθησης και ανήκει στην κατηγορία των μοντέλων συνόλου (Ensemble Model).

Το κύριο συστατικό αυτού του μοντέλου είναι το δέντρο απόφασης, δηλαδή μία δομή δεδομένων που έχει την μορφή δέντρου. Ένα δέντρο απόφασης αποτελείται από κόμβους, οι οποίοι έχουν τον ρόλο είτε της **ρίζας** (Root Node) του δέντρου, είτε τον ρόλο του **ενδιάμεσου κόμβου** (Internal Node), είτε το ρόλο του **φύλλου** (Leaf Node). Ο λόγος, για τον οποίο η δομή αυτή ονομάζεται δέντρο απόφασης είναι ότι κάθε φορά που προσπαθεί το μοντέλο να κάνει προβλέψεις, διατρέχει ένα συγκεκριμένο μονοπάτι από την ρίζα του δέντρου μέχρι να φτάσει σε ένα φύλλο του δέντρου. Το μονοπάτι που ακολουθείται δεν είναι το ίδιο κάθε φορά, αλλά εξαρτάται άμεσα από τις πληροφορίες που δίνονται ως είσοδος, βάσει των οποίων θα γίνουν και οι προβλέψεις. Στα δέντρα αποφάσεων αποθηκεύονται δεδομένα, τα οποία είναι σχετικά με αυτά που δίνονται ως είσοδος στον αλγόριθμό.

Κατά την εκπαίδευση του αλγορίθμου αυτού, στόχος είναι να δημιουργηθεί ένας αριθμός από διαφορετικά δέντρα απόφασης, η **δομή** των οποίων θα είναι **βέλτιστη** (π.χ το βάθος του δέντρου, ο αριθμός των φύλλων, τα δεδομένα που θα περιέχει κ.α.). Τα δέντρα που θα

δημιουργηθούν στο τέλος, θα περιγράφουν με τον καλύτερο δυνατό τρόπο το σύνολο δεδομένων εκπαίδευσης. Ο αριθμός των δέντρων που θα φτιαχτεί προσδιορίζεται από τον χρήστη.

Αφού γίνει η εκπαίδευση του μοντέλου, δηλαδή αφού κατασκευαστεί ένα σύνολο από δέντρα απόφασης που περιγράφουν με τον καλύτερο δυνατό τρόπο τα δεδομένα εκπαίδευσης, ο αλγόριθμος προβαίνει σε προβλέψεις με βάση τα δεδομένα ελέγχου. Κάθε φορά που χρειάζεται να γίνεται μία πρόβλεψη, ο αλγόριθμος διατρέχει το μονοπάτι όλων των δέντρων που έχουν δημιουργηθεί από την ρίζα έως ένα φύλλο του εκάστοτε δέντρου. Η τιμή που βρίσκεται στο φύλλο κάθε δέντρου είναι και η απόφαση – πρόβλεψη του συγκεκριμένου δέντρου. Η τελική πρόβλεψη του μοντέλου είναι, τελικά, ο μέσος όρος των τιμών των αποφάσεων όλων των δέντρων.

Γενικά, ο Random Forest χρησιμοποιείται σε πολλές εφαρμογές της καθημερινότητας, όπως η ταξινόμηση δεδομένων (π.χ. ανίχνευση spam email), η πρόβλεψη αριθμητικών τιμών (π.χ. πρόβλεψη τιμών ακινήτων), η ανίχνευση ανωμαλιών (π.χ. ανίχνευση απάτης σε χρεωστικές κάρτες). Μάλιστα, θεωρείται ένας ισχυρός και αξιόπιστος αλγόριθμος. Βέβαια, η ικανότητά του μοντέλου αυτού εξαρτάται δραματικά από τα δεδομένα που του δίνονται ως είσοδο.