ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
Σχολή Πολιτικών Μηχανικών
Τομέας Μεταφορών και Συγκοινωνιακής Υποδομής

Αναγνώριση του Μέσου Μεταφοράς από Δεδομένα Έξυπνου Τηλεφώνου με τη χρήση Αλγορίθμων Μηχανικής Μάθησης

Εκπόνηση Διπλωματικής εργασίας:

Ευθυμίου Αλέξανδρος

Επιβλέπουσα Καθηγήτρια: Ελένη Βλαχογιάννη, Επίκουρη Καθηγήτρια Σχολής Πολιτικών Μηχανικών ΕΜΠ

Αθήνα, 2017

Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Πολιτικών Μηχανικών

Τομέας Μεταφορών και Συγκοινωνιακής Υποδομής

Αναγνώριση του Μέσου Μεταφοράς από Δεδομένα Έξυπνου Τηλεφώνου με τη χρήση Αλγορίθμων Μηχανικής Μάθησης

Συγγραφέας διπλωματικής εργασίας: Ευθυμίου Αλέξανδρος

Επιβλέπουσα Καθηγήτρια: Ελένη Βλαχογιάννη

Αθήνα, 2017

National Technical University of Athens

School of Civil Engineering

Department of Transportation Planning and Engineering

Transportation Mode Detection on Data Collected from Smartphones with the use of Machine Learning Algorithms

Thesis author: Efthymiou Alexandros

Supervising Professor: Eleni I. Vlahogianni

Athens, 2017

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω θερμά την επιβλέπουσα καθηγήτριά μου, κυρία Ελένη Βλαχογιάννη, Επίκουρη Καθηγήτρια στον Τομέα Μεταφορών και Συγκοινωνιακής Υποδομής της Σχολής Πολιτικών Μηχανικών για τη δυνατότητα που μου έδωσε να ασχοληθώ με το συγκεκριμένο κομμάτι των μεταφορών και για τη βοήθειά της στην κατανόησή της λογικής της μηχανικής μάθησης. Επίσης, την ευχαριστώ για τη στήριξή της σε όλο το διάστημα της συνεργασίας μας. Στη συνέχεια, θα ήθελα να ευχαριστήσω τον Διδάκτορα Εμμανουήλ Μπαρμπουνάκη για όλες τις συμβουλές και την καθοδήγησή του για την εκπόνηση αυτής της διπλωματικής εργασίας.

Τέλος, θέλω να ευχαριστήσω τους φίλους μου και κυρίως την οικογένειά μου για την εμπιστοσύνη της και τη στήριξή της σε όλα τα φοιτητικά μου χρόνια. Ιδιαίτερα, θέλω να ευχαριστήσω τον αδερφό μου, Δημήτρη, ο οποίος με συμβουλεύει και κατευθύνει σε μια σειρά θεμάτων μέχρι και σήμερα, την μητέρα μου, Σοφία και τον πατέρα μου, Λάκη, ο οποίος δεν είναι πια μαζί μας.

ΠΕΡΙΛΗΨΗ

Σκοπός της παρούσας διπλωματικής εργασίας είναι η πρόβλεψη του μέσου μεταφοράς που χρησιμοποιούν οι χρήστες σε κάθε τους ταξίδι με τη χρήση δεδομένων από τους αισθητήρες έξυπνων κινητών. Για το σκοπό αυτό αναπτύσσεται μια μεθοδολογία που αποτελείται από δύο βήματα: Αρχικά, οι χρονοσειρές από το επιταχυνσιόμετρο, το γυροσκόπιο και τον αισθητήρα προσανατολισμού του κινητού ανά ταξίδι αναλύονται σε κυλιόμενα χρονικά παράθυρα και υπολογίζονται τα βασικά στατιστικά μεγέθη των χρονοσειρών αυτών σε κάθε παράθυρο. Στη συνέχεια, αναπτύσσονται πρότυπα μηχανικής μάθησης που χρησιμοποιούν ως δεδομένα τα υπολογισμένα στατιστικά μεγέθη με στόχο την πρόβλεψη του μέσου μετακίνησης του χρήστη ο οποίος χρησιμοποίησε είτε τα Μέσα Μαζικής Μεταφοράς είτε κάποιο άλλο μέσο. Η μεθοδολογία αυτή εφαρμόζεται σε ταξίδια διαφορετικών χρηστών και αναλύονται τα αποτελέσματα. Τέλος, παρατίθεται σχολιασμός για τους περιορισμούς της παρούσας έρευνας και τις προεκτάσεις για περαιτέρω έρευνα.

Λέξεις – Κλειδιά: μηχανική μάθηση, τυχαίο δάσος, έξυπνα κινητά τηλέφωνα, αισθητήρες, εντοπισμός μέσου μεταφοράς, αυτοκίνητο, Μέσα Μαζικής Μεταφοράς

ABSTRACT

The purpose of this diploma thesis is the prediction of transportation mode on each trip completed by users, using data from mobile sensors. For this purpose, a two-step methodology is developed: First, time series from accelerometer, gyroscope and orientation sensor are analysed on sliding time windows and, for each time window, the basic statistical measures of these time series are calculated. Then, machine learning models are developed to predict whether the user travels with public transport or other means of transport by using the calculated statistical measures. This methodology is applied on trips completed by different users and the results are analysed. Finally, discussion is made about the limitations of this research and further research is proposed.

Keywords: machine learning, random forest, smartphones, sensors, transportation mode detection, car, public transport

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΥΡΕΤΗΡΙΟ ΠΙΝΑΚΩΝ

ΕΥΡΕΤΗΡΙΟ ΔΙΑΓΡΑΜΜΑΤΩΝ

ΕΥΡΕΤΗΡΙΟ ΕΙΚΟΝΩΝ

1. ΕΙΣΑΓΩΓΗ

1.1. Η ΑΝΘΡΩΠΙΝΗ ΣΥΜΠΕΡΙΦΟΡΑ ΚΑΤΑ ΤΗ ΔΙΑΡΚΕΙΑ ΤΟΥ ΤΑΞΙΔΙΟΥ

Οι αυξημένες δυνατότητες των σύγχρονων κινητών τηλεφώνων σε συνδυασμό με τον εύκολο προγραμματισμό τους, το μεγάλο ποσοστό διείσδυσής τους στην αγορά και την αποτελεσματική προσέγγισή τους από τρίτες εταιρείες λογισμικού με σκοπό την ανάπτυξη εφαρμογών, έχουν συμβάλλει στην ωρίμανσή τους και τη μετατροπή τους σε ένα αποτελεσματικό εργαλείο παρακολούθησης της ανθρώπινης συμπεριφοράς (Mitchell, 2009). Η παρούσα διπλωματική εργασία επικεντρώνεται σε ένα κομμάτι της ανθρώπινης συμπεριφοράς, αυτής που αναπτύσσεται κατά τη μεταφορά του ατόμου από ένα σημείο σε ένα άλλο και γίνεται προσπάθεια αυτόματου εντοπισμού του μέσου μεταφοράς που χρησιμοποίησε.

Ο όρος «ανθρώπινη συμπεριφορά κατά τη διάρκεια του ταξιδιού» περιλαμβάνει όλες τις φυσικές κινήσεις του χρήστη όταν βρίσκεται στο αυτοκίνητο, ή σε κάποιο άλλο μέσο μεταφοράς, όπως αυτές καταγράφονται από τους αισθητήρες του τηλεφώνου. Για παράδειγμα, η χρήση του κινητού από τον χρήστη όταν εκείνος μιλάει ή στέλνει μηνύματα είναι μοναδική, όπως η ταχύτητα, η επιτάχυνση και το φρενάρισμα των οχημάτων διαφέρει σε Ι.Χ. και τρένο.

Η ικανότητα καταγραφής της συμπεριφοράς των ατόμων κατά τη διάρκεια του ταξιδιού τους από τα smartphones, μέσω των αισθητήρων τους, έχει θετικό αντίκτυπο σε πολλούς ερευνητικούς τομείς. Για παράδειγμα, η παρακολούθηση της ανθρώπινης κινητικότητας μπορεί να ωφεληθεί άμεσα από τη δυνατότητα παρακολούθησης της συμπεριφοράς των ατόμων κατά τη μεταφορά τους από ένα σημείο σε ένα άλλο (Lazer, et al., 2009) & (Song, et al., 2010). Αυτό με τη σειρά του θα επέτρεπε τη βελτίωση του πολεοδομικού σχεδιασμού (Zheng, et al., 2011), την παρακολούθηση και αντιμετώπιση της εξάπλωσης των ασθενειών καθώς και άλλων πιθανών κινδύνων, καθώς και την παροχή πληροφοριών έκτακτης ανάγκης

σχετικά με την ταχύτερη διαδρομή για να βοηθηθούν οι χαμένοι ή οι τραυματίες (Soper, 2012). Οι αλγόριθμοι εντοπισμού της τοποθεσίας θα μπορούσαν να βελτιωθούν με την κατασκευή πιο εξειδικευμένων μοντέλων κίνησης, εφόσον παρέχονται πληροφορίες του τρέχοντος τρόπου μεταφοράς του ατόμου (Nurmi, et al., 2010).

1.2. ΑΥΤΟΜΑΤΟΣ ΕΝΤΟΠΙΣΜΟΣ ΤΟΥ ΜΕΣΟΥ ΜΕΤΑΦΟΡΑΣ

1.2.1. ΣΗΜΑΣΙΑ

Η αξιόπιστη διάκριση μεταξύ των διαφορετικών τρόπων μεταφοράς με μηχανοκίνητο μέσο θα παρέχει λεπτομερέστερες πληροφορίες σχετικά με την ανθρώπινη συμπεριφορά κατά τη διάρκεια του ταξιδιού, για παράδειγμα θα επιτρέπει την αυτόματη εκτίμηση της κατανάλωσης διοξειδίου του άνθρακα ανά άτομο ή θα βοηθήσει στην καλύτερη κατανόηση των συνθηκών μετακίνησης των αστικών πολιτών. Επίσης, θα μπορεί να γίνει διαχωρισμός του τρένου από το αυτοκίνητο ώστε εταιρείες όπως η Google να μπορεί να υπολογίσει με βάση τους χρήστες τις, την κίνηση που θα έχει ο δρόμος χωρίς να περιλαμβάνει στα δεδομένα της και τους επιβάτες του τρένου. Επιπλέον, με τη γνώση του μεταφορικού μέσου που χρησιμοποιεί το κάθε άτομο, δίνεται η δυνατότητα στη δημιουργία ιδανικού προφίλ κάθε χρήστη, με σκοπό τον προορισμό σε αυτόν συγκεκριμένες διαδρομές ταξιδιών σε πραγματικό χρόνο ή την προβολή σε αυτόν στοχευμένες και εξιδανικευμένες διαφημίσεις. Για παράδειγμα, αν η Μαρία οδηγεί αυτοκίνητο μπορεί να σταλούν στο κινητό της κουπόνια βενζίνης ή κάποια έκπτωση για τη συντήρηση (service) του αυτοκινήτου της. Τέλος, ο αυτόματος διαχωρισμός των μέσων μεταφοράς είναι αναγκαίος για τις ασφαλιστικές εταιρείες, ώστε να μπορούν να διαχωρίσουν τους χρήστες του αυτοκινήτου από όλους τους υπόλοιπους, να παρακολουθήσουν την οδηγική του συμπεριφορά και να τιμολογήσουν με βάση αυτή.

1.2.2. ΕΞΥΠΝΑ ΚΙΝΗΤΑ ΚΑΙ ΑΙΣΘΗΤΗΡΕΣ

Με την εξέλιξη της τεχνολογίας, τα έξυπνα κινητά τηλέφωνα (smartphones) παίζουν πλέον καθοριστικό ρόλο στις μεθοδολογίες που χρησιμοποιούνται. Οι αισθητήρες με τους οποίους

είναι εξοπλισμένο το κινητό το έχουν μετατρέψει σε ένα απαραίτητο εργαλείο στον επιστημονικό κλάδο. Συλλέγονται ακατέργαστα δεδομένα ανάλογα με τη συχνότητα που έχει επιλέξει ο ερευνητής με τα οποία, μετά από κατάλληλη επεξεργασία, μπορεί να γίνει πρόβλεψη του τρόπου μετακίνησης των χρηστών χωρίς να χρειαστεί κάποια επιβεβαίωση από τον χρήστη. Ο ενσωματωμένος δέκτης Παγκόσμιου Συστήματος Τοποθεσίας (Global Positioning System – GPS), το επιταχυνσιόμετρο και το γυροσκόπιο, είναι μερικοί από τους αισθητήρες Που διαθέτουν τα σύγχρονα έξυπνα κινητά τηλέφωνα.

Ενώ, λοιπόν, η ιδέα της χρησιμοποίησης των smartphones για την παρακολούθηση της συμπεριφοράς της μεταφοράς των ατόμων δεν είναι καινούργια, σχεδόν όλες οι προηγούμενες έρευνες επικεντρώθηκαν κατά κύριο λόγο στην χρήση του ενσωματωμένου δέκτη GPS του τηλεφώνου. Παρόλο που οι μέθοδοι με βάση το GPS μπορούν να γίνουν αποδοτικές όταν υπάρχει κάλυψη του σήματος GPS, υφίστανται κάποιους σημαντικούς περιορισμούς. Πρώτον, οι ενσωματωμένοι δέκτες GPS είναι γνωστό ότι υποφέρουν από υψηλή κατανάλωση ενέργειας, επομένως αυτές οι προσεγγίσεις καταστρέφουν γρήγορα τη μπαταρία της συσκευής. Δεύτερον, η ανάγκη του δέκτη GPS να βρίσκεται σε μια περιοχή χωρίς εμπόδια ώστε να μπορέσουν οι δορυφόροι να ¨δουν¨ τη συσκευή, δημιουργεί πρόβλημα σε πολλές κοινές περιπτώσεις της αστικής μεταφοράς για παράδειγμα όταν ο χρήστης κινείται υπόγεια, μέσα σε έναν σταθμό ή όταν βρίσκεται ανάμεσα σε ουρανοξύστες. Τρίτον, οι τωρινές λύσεις που βασίζονται στο GPS παρέχουν μέτρια ακρίβεια όταν απαιτείται λεπτομερής διάκριση των τρόπων μεταφοράς, δηλαδή μεταξύ των μηχανοκίνητων μέσων. Το τελευταίο συμβαίνει, διότι το GPS μετράει την ταχύτητα οπότε ακόμα και αν ορίσουμε ένα εύρος ταχυτήτων για το κάθε μεταφορικό μέσο χωριστά, δε θα μπορέσει το μοντέλο να προβλέψει με απόλυτη ακρίβεια τον τρόπο μεταφοράς, βασισμένο μόνο σε αυτή την πληροφορία.

1.3. ΣΚΟΠΟΣ ΤΗΣ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Η παρούσα διπλωματική προτείνει μια μεθοδολογία για την αναγνώριση του μέσου με το οποίο μετακινείται ένας χρήστης οδικού δικτύου μέσα από δεδομένα που προέρχονται αποκλειστικά από αισθητήρες κινητών τηλεφώνων. Γίνεται η θεώρηση δύο διαφορετικών κατηγοριών μηχανοκίνητου μέσου (Ι.Χ. και ΜΜΜ) με τη συλλογή δεδομένων από επιταχυνσιόμετρο, γυροσκόπιο, αισθητήρα προσανατολισμού και την εφαρμογή προχωρημένων μοντέλων μηχανικής μάθησης.

Στη συγκεκριμένη διπλωματική εργασία και ενώ συλλέχθηκαν δεδομένα GPS, χρησιμοποιήθηκαν μόνο για την επαλήθευση των ταξιδιών, ώστε να ξεκινούν οι αισθητήρες την καταγραφή αυτόματα τη στιγμή που θα αντιληφθούν κίνηση του χρήστη με αυξημένη ταχύτητα. Για τον εντοπισμό του μεταφορικού μέσου όμως, που είναι και ο στόχος της εργασίας, δε χρησιμοποιήθηκαν καθόλου δεδομένα του GPS, καθώς σκοπός ήταν να γίνει μια πιο αποτελεσματική προσέγγιση του συγκεκριμένου θέματος χωρίς να χρειαστεί να θυσιαστεί η μπαταρία του τηλεφώνου.

1.4. ΔΟΜΗ ΤΗΣ ΕΡΓΑΣΙΑΣ

Στη συνέχεια δίνεται μια συνοπτική περιγραφή των κεφαλαίων που θα αναλυθούν σε αυτή τη διπλωματική εργασία.

ΚΕΦΑΛΑΙΟ 2
Γίνεται μια λεπτομερής βιβλιογραφική ανασκόπηση άλλων ερευνών που έχουν γίνει πάνω στο συγκεκριμένο θέμα. Περιλαμβάνει την εξέλιξη των μέσων και τεχνολογιών που έχουν χρησιμοποιηθεί, παρουσιάζοντας ταυτόχρονα τη σταδιακή βελτίωση των μοντέλων πρόβλεψης. Επίσης, γίνεται μια προσπάθεια κατηγοριοποίησης των ερευνών που έχουν διεξαχθεί στο συγκεκριμένο επιστημονικό κλάδο.

ΚΕΦΑΛΑΙΟ 3
Σε αυτό το κεφάλαιο αναλύονται οι αισθητήρες και τα δεδομένα που χρησιμοποιήθηκαν καθώς και ο τρόπος συλλογής τους. Περιγράφεται η σημασία τους τόσο γραφικά όσο και πρακτικά.

ΚΕΦΑΛΑΙΟ 4
Παρουσιάζεται το πρόβλημα και το μοντέλο που εφαρμόστηκε στην πρόβλεψη. Γίνεται αναφορά του λογισμικού που χρησιμοποιήθηκε και επισημαίνεται η μέθοδος αξιολόγησης του μοντέλου.

ΚΕΦΑΛΑΙΟ 5
Εκπαίδευση και δοκιμή του μοντέλου. Παρουσιάζονται τα αποτελέσματα και αξιολογείται η μέθοδος που χρησιμοποιήθηκε.

ΚΕΦΑΛΑΙΟ 6
Τέλος, αναλύονται τα συμπεράσματα που προέκυψαν και γίνονται προτάσεις για περαιτέρω έρευνα.

Μετά το πέρας και του 6^{ου} κεφαλαίου παρατίθεται η βιβλιογραφία που χρησιμοποιήθηκε στην παρούσα διπλωματική εργασία.

2. ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΑΝΑΣΚΟΠΗΣΗ

Το παρόν κεφάλαιο αφορά τη βιβλιογραφική ανασκόπηση του θέματος με σκοπό τη σύγκριση παρόμοιων μεθόδων. Αρχικά, γίνεται μια ιστορική αναδρομή των μεθοδολογιών που έχουν χρησιμοποιηθεί από άλλους ερευνητές, ενώ στη συνέχεια γίνεται διαχωρισμός των ερευνών με βάση τη χρησιμοποίηση ή μη του GPS. Έπειτα, αναφέρονται μελέτες που έγιναν χρησιμοποιώντας μηχανική μάθηση στα μοντέλα πρόβλεψης και τέλος επισημαίνονται συγκεκριμένα στοιχεία που διαφέρουν από έρευνα σε έρευνα.

2.1. ΤΕΧΝΟΛΟΓΙΕΣ

Η αναγνώριση του μέσου μεταφοράς αποτελεί ερευνητική επέκταση της αυτόματης μέτρησης του αριθμού των οχημάτων που περνούν από ένα σημείο του δρόμου, η οποία αφορά μόνο τα μηχανοκίνητα μέσα. Ενδεικτικά αναφέρονται στη συνέχεια οι υπάρχουσες τεχνολογίες για αυτή τη μέτρηση, οι οποίες και χωρίζονται σε τρεις βασικές κατηγορίες:

1. Αισθητήρες επί του οδοστρώματος (όπως οι μετρητές με πεπιεσμένο αέρα) ή εντός αυτού (για παράδειγμα οι μαγνητικοί βρόχοι)

2. Αισθητήρες έξω από το οδόστρωμα (ανιχνευτές υπερήχων, υπερύθρων)

3. Τεχνολογίες εντοπισμού οχήματος (για παράδειγμα το ίδιο το όχημα μπορεί να λειτουργήσει ως κινούμενος αισθητήρας με έναν ειδικό εξοπλισμό).

Στην τελευταία κατηγορία ανήκει πλέον και το GPS. Ωστόσο, οι παραδοσιακοί αισθητήρες υποφέρουν από βασικά μειονεκτήματα όπως είναι το υψηλό κόστος εγκατάστασης και συντήρησής τους, ενώ δυσλειτουργούν σε συγκεκριμένες καταστάσεις όπως όταν υπάρχουν δυσμενείς συνθήκες ή κορεσμένη κυκλοφορία. Τέλος, δε μπορούν να συλλέξουν δεδομένα από ένα ολόκληρο ταξίδι, παρά μόνο από συγκεκριμένες τοποθεσίες (Φραντζεσκάκης, et al., 2009).

Ο εντοπισμός του τρόπου μεταφοράς, λοιπόν, είναι μια νέα πρόκληση για τους ερευνητές. Οι μέθοδοι που εφαρμόζονται για τη συλλογή των δεδομένων ως προς τον συγκεκριμένο σκοπό έχουν υποστεί αλλαγές με την πάροδο του χρόνου, ξεκινώντας από τις συμβατικές προσωπικές και τις γραπτές συνεντεύξεις τη δεκαετία του 1950. Το υψηλό κόστος καθώς και διάφορα ζητήματα ασφαλείας, αποδείχθηκαν σημαντικά προβλήματα σε αυτή την προσέγγιση. Για να ξεπεραστούν αυτά τα μειονεκτήματα, πραγματοποιήθηκαν νέες έρευνες με τη βοήθεια του υπολογιστή τη δεκαετία του 1980. Αυτές οι έρευνες περιλάμβαναν τηλεφωνική συνέντευξη με τη βοήθεια του υπολογιστή (computer-assisted telephone interview, CATI) καθώς και αυτοματοποιημένη συνέντευξη πάλι με τη βοήθεια του υπολογιστή (computer-assisted self-interview, CASI) (Stopher, 2009) & (Jean, et al., 2001).

Οι έρευνες αυτές αποδείχθηκε πως ήταν βελτιωμένες σε σχέση με τις προσωπικές συνεντεύξεις (Hato, 2006), ωστόσο παρέμεναν σημαντικά μειονεκτήματα όσον αφορά στη συλλογή προσωπικών δεδομένων από τα ταξίδια. Για παράδειγμα, υπήρχαν ανακρίβειες στην καταχώρηση των χρόνων έναρξης και λήξης της διαδρομής και τα σύντομα ταξίδια δεν καταγραφόντουσαν από τους χρήστες (McGowen, et al., 2007) & (Hato, 2010). Η πηγή όλων αυτών των προβλημάτων ήταν η μεγάλη επιβάρυνση των ερωτηθέντων να απαντήσουν σε τεράστιο αριθμό ερωτήσεων με βάση τις αναμνήσεις τους. Για την αντιμετώπιση αυτού του προβλήματος, η τεχνολογία GPS χρησιμοποιήθηκε στα τέλη της δεκαετίας του 1990, παρέχοντας το σημείο εκκίνησης για μια νέα γενιά μεθόδων έρευνας των ταξιδιών (Wagner, 1997).

Αρχικά, πραγματοποιήθηκαν έρευνες με τη χρήση του GPS ως συμπληρωματικές, για την αξιολόγηση της ακρίβειας των παραδοσιακών μεθόδων. Στη συνέχεια όμως, έγινε πλήρης αντικατάσταση (Zito, et al., 1995), (Wagner, 1997) & (Sermons, et al., 1996). Στην αρχή, συσκευές GPS τοποθετήθηκαν σε οχήματα, συνεπώς παρακολουθήθηκε μόνο η συμπεριφορά των ατόμων που χρησιμοποιούσαν οχήματα. Στις αρχές τις δεκαετίας του 2000, η ραγδαία εξέλιξη της τεχνολογίας άνοιξε το δρόμο για την ανάπτυξη των φορετών συσκευών καταγραφής δεδομένων GPS (wearable GPS data loggers) (Gong, et al., 2014).

Με την εισαγωγή ελαφριών, φορητών και εύχρηστων συστημάτων καταγραφής δεδομένων GPS, όλα τα μέσα μεταφοράς μπορούν να παρακολουθούνται, ενώ τοποθεσίες και χρόνοι ταξιδιών καταγράφονται πλέον με ευκολία. Ωστόσο, οι συσκευές GPS αυτές, παρόλα τα πλεονεκτήματά τους έναντι των παραδοσιακών μεθόδων έρευνας, έχουν και σημαντικά μειονεκτήματα όπως το υψηλό κόστος τους. Επίσης, οι χρήστες ξεχνούσαν πολλές φορές να

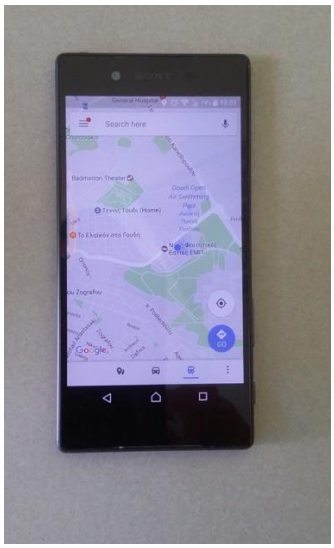
πάρουν μαζί τους τη συσκευή GPS όταν ταξίδευαν και υπήρχε μεγάλη αναξιοπιστία του σήματος GPS σε ορισμένες τοποθεσίες (Zhao, et al., 2015).

Πριν από την εμφάνιση των smartphones, είχε διερευνηθεί η πιθανότητα χρήσης κινητών τηλεφώνων για τη συλλογή δεδομένων χρησιμοποιώντας την τεχνολογία GSM (Wermuth, et al., 2003). Αντί της εφαρμογής του GPS, έχει γίνει προσπάθεια εντοπισμού της τοποθεσίας από πύργους κινητής τηλεφωνίας (Krygsman, et al., 2008). Μέσα σε λίγο καιρό, διερευνήθηκαν περισσότερες τεχνολογικές λύσεις, όπως το Bluetooth, WiFi, RFID και οι έξυπνες κάρτες (Stopher, 2009). Τα τηλεφωνικά συστήματα PHS (Personal handy-phone systems) έγιναν πολύ δημοφιλή στην Ιαπωνία για την καταγραφή γεωγραφικών τοποθεσιών. Αυτά τα συστήματα εντοπίζουν τη συσκευή με τη βοήθεια συγκεκριμένων σταθμών (Asakura, et al., 1999) & (Asakura, et al., 2004). Πάνω από 20 έρευνες έχουν διεξαχθεί στην Ιαπωνία χρησιμοποιώντας PHS από το 2003 ((Sugino, et al., 2005), (Itsubo, et al., 2006) & (Yatsumoto, et al., 2006)).

Τα τελευταία χρόνια, η εκρηκτική εξάπλωση των έξυπνων τηλεφώνων (smartphones) προσέφερε στην κοινότητα των μεταφορών ένα νέο δυναμικό και, πλέον, διεξάγεται συνεχής έρευνα με τη χρησιμοποίηση των smartphones για τη συλλογή δεδομένων ταξιδιού. Αυτό το ενδιαφέρον κυρίως οφείλεται στο γεγονός ότι οι αισθητήρες GPS ενσωματώνονται στα σύγχρονα κινητά τηλέφωνα, καθιστώντας δυνατή την αντικατάσταση των μηχανημάτων GPS που χρησιμοποιούνταν προηγουμένως. Επίσης, τα smartphones έχουν το πλεονέκτημα ότι είναι ένας μόνιμος σύντροφος ταξιδιού, οπότε μπορούν να παρακολουθούν τα μοτίβα ταξιδιών για παρατεταμένες χρονικές περιόδους. Από την άλλη, οι συσκευές με αποκλειστική λειτουργία την καταγραφή δεδομένων GPS θεωρούνται πλέον επιβαρυντικές ως προς τη μεταφορά τους και καταργούνται. (Shafique, et al., 2016). Στις Εικόνες 2.1 και 2.2 φαίνεται η συσκευή GPS και η μείωση που υπέστη το μέγεθός της με την εξέλιξη της τεχνολογίας, ενώ στην Εικόνα 2.3 φαίνεται ένα σύγχρονο κινητό τηλέφωνο το οποίο έχει ενσωματωμένο το δέκτη GPS στο εσωτερικό του.

Εικόνα 2.1. Αρχική μορφή GPS δίπλα σε κινητό τηλέφωνο (Stopher, et al., 2008)

Εικόνα 2.2. Εξελιγμένη μορφή GPS δίπλα σε κινητό τηλέφωνο (Stopher, et al., 2008)



Εικόνα 2.3 Σημερινό «έξυπνο» κινητό τηλέφωνο με ενσωματωμένο δέκτη GPS

Επίσης, αξίζει να αναφερθεί πως μια έρευνα με την ονομασία FMS (Future Mobility Survey) από τον Zhao (Zhao, et al., 2015), σύγκρινε τα αποτελέσματα των παραδοσιακών ερευνών με αυτές που γίνονται πλέον μέσω των smartphones. Είναι μέρος ενός ερευνητικού προγράμματος που ξεκίνησε από τη συνεργασία της Σιγκαπούρης και του Ινστιτούτου Τεχνολογίας της

Μασαχουσέτης (MIT). Η έρευνα απέδειξε ότι οι συμμετέχοντες τείνουν να υπερεκτιμούν το χρόνο ταξιδιού στις παραδοσιακές έρευνες.

Τα smartphones εξελίσσονται γρήγορα, αποκτούν περισσότερες υπολογιστικές ικανότητες, έχουν εύκολη πρόσβαση στο διαδίκτυο και πλέον μπορούν να λειτουργούν ως αισθητήρες. Είναι συνήθως εξοπλισμένα με επιταχυνσιόμετρο, γυροσκόπιο, μαγνητόμετρο, αισθητήρα βαρύτητας, βαρόμετρο, αισθητήρα φωτός, πυξίδα και άλλους. Οι παραπάνω αισθητήρες, επιτρέπουν την πλούσια εφαρμογή εξόρυξης δεδομένων, όπως την αναγνώριση της δραστηριότητας των χρηστών, συμπεριλαμβανομένων των ταξιδιωτικών δραστηριοτήτων από μια απλή μετακίνηση (για παράδειγμα περπάτημα, τζόκινγκ, περπάτημα στον κάτω όροφο) μέχρι πολύπλοκες δραστηριότητες όπως η οδήγηση και το πότε βλέπει ο χρήστης ταινία. Επομένως, το smartphone μπορεί να θεωρηθεί ως μία από τις καλύτερες πηγές συλλογής δεδομένων όσων αφορά την αναγνώριση δραστηριότητας του χρήστη (Su, et al., 2015).

2.2. ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΒΙΒΛΙΟΓΡΑΦΙΑΣ ΜΕ ΒΑΣΗ ΤΟΥΣ ΑΙΣΘΗΤΗΡΕΣ

Οι έρευνες πάνω στον εντοπισμό του μέσου μεταφοράς μπορούν να διαχωριστούν σε αυτές που χρησιμοποιούν GPS και σε αυτές που σκόπιμα το παραλείπουν.

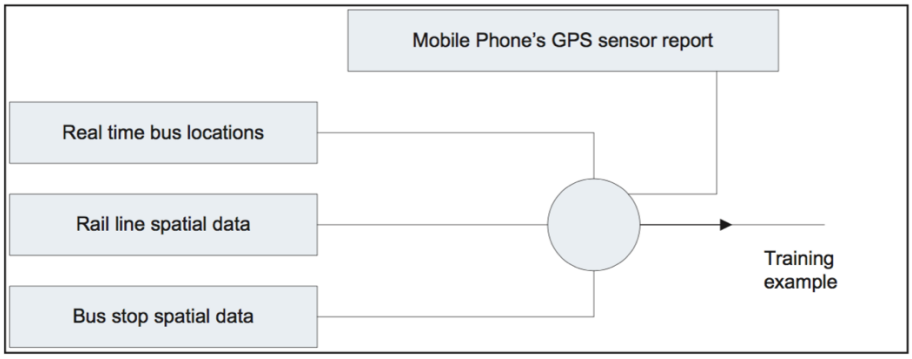
2.2.1. ΜΕ ΤΗ ΧΡΗΣΗ ΤΟΥ GPS

Η τεράστια δημοτικότητα των smartphones, καθώς και η αυξανόμενη διείσδυσή τους στην καθημερινότητα, έχει προσελκύσει πολύ ερευνητική προσοχή στο ρόλο τους ως εργαλεία προσδιορισμού του τρόπου μεταφοράς. Για τον συγκεκριμένο επιστημονικό κλάδο μάλιστα, οι περισσότερες υπάρχον έρευνες κάνουν αποκλειστική χρήση του GPS, όπως φαίνεται στη συνέχεια (Zong, et al., 2015) & (Huss, et al., 2014).

Η έρευνα των Tsui και Shalaby (Tsui, et al., 2006) έγινε με δεδομένα GPS που συλλέχθηκαν από το Τορόντο. Επιταχύνσεις, μέσες και μέγιστες τιμές, υπολογίστηκαν από δεδομένα GPS και μαζί με πληροφορίες των διαδρομών των δημοσίων συγκοινωνιών, χρησιμοποιήθηκαν για να προβλέψουν τον τρόπο μεταφοράς, καταφέρνοντας να φτάσουν σε ακρίβεια μεγαλύτερη

του 90%. Μια άλλη έρευνα που είχε παρόμοια χαρακτηριστικά, χρησιμοποίησε δεδομένα από μόνο έναν συμμετέχοντα ο οποίος έκανε 60 ταξίδια που καταγράφηκαν στο TTTS (Toronto Transportation Tomorrow Survey) φορώντας μια συσκευή GPS (Chung, et al., 2005). Μετά από τη συλλογή των δεδομένων GPS και συνδυάζοντάς τα με διαθέσιμες πληροφορίες GIS, ο τρόπος μεταφοράς προβλέφθηκε με ακρίβεια μεγαλύτερη του 92%.

Η ακρίβεια της αναγνώρισης του τρόπου μεταφοράς αυξάνεται όταν δεδομένα από GPS συνδέονται με μια πλατφόρμα GIS (Oliver, et al., 2010). Σε μια έρευνα, πέρα από τη ταχύτητα, τη μέση επιτάχυνση και τη μέση αλλαγή της κατεύθυνσης από τον ¨πραγματικό βορρά¨ χρησιμοποιήθηκε η τοποθεσία των λεωφορείων και οι στάσεις τους, καθώς και η απόσταση του χρήστη από αυτά σε πραγματικό χρόνο, οι τοποθεσίες των γραμμών των τρένων και η απόσταση του χρήστη από αυτές πάλι σε πραγματικό χρόνο. Η ακρίβεια της πρόβλεψης αυξήθηκε κατά 20% όταν πέρα από το GPS, χρησιμοποιήθηκαν και οι παράμετροι αυτοί μέσω της πλατφόρμας GIS (Stenneth, et al., 2011). Ένα διάγραμμα των δεδομένων που χρησιμοποιήθηκαν φαίνεται στην Εικόνα 2.4.



Εικόνα 2.4 Τρόπος δημιουργίας μοντέλου εκπαίδευσης σε μια συγκεκριμένη έρευνα (Πηγή: (Stenneth, et al., 2011))

Η ακρίβεια βελτιώθηκε περαιτέρω, συνδυάζοντας δεδομένα από GPS και επιταχυνσιομέτρου για τον προσδιορισμό του τρόπου μεταφοράς (Feng, et al., 2013). Η συγκεκριμένη έρευνα προβλέπει το πότε ο χρήστης περπατάει, κάνει ποδήλατο, τρέχει, οδηγεί μοτοσικλέτα, βρίσκεται σε λεωφορείο, σε αμάξι, σε μετρό ή σε τραμ με σχετικά υψηλή ακρίβεια. Αρχικά, γίνεται δοκιμή πρόβλεψης των παραπάνω μόνο από το επιταχυνσιόμετρο, μετά μόνο από το GPS και τέλος γίνεται εφαρμογή και των δύο αισθητήρων μαζί όπου και προκύπτει μεγαλύτερη ακρίβεια από ότι χρησιμοποιώντας τον κάθε έναν χωριστά.

Σε μία άλλη έρευνα η οποία σκοπό είχε τη διάκριση του τρόπου μεταφοράς ανάμεσα σε περπάτημα, αυτοκίνητο και λεωφορείο, έγινε χρήση του GPS και του επιταχυνσιομέτρου (Zahra, et al., 2015). Σε αυτήν λοιπόν και ενώ υπήρχε μεγάλο ποσοστό ακρίβειας για τον

προσδιορισμό του Ι.Χ. της τάξης του 97%, τόσο στο λεωφορείο όσο και στο περπάτημα η ακρίβεια ήταν περίπου 10% χαμηλότερη. Αυτό συνέβη λόγω της ομοιότητας κάποιων χαρακτηριστικών (ταχύτητα, επιτάχυνση) που χρησιμοποιήθηκαν στην ανάπτυξη του μοντέλου πρόβλεψης και είναι κοινά ανάμεσα στα Ι.Χ. και τα λεωφορεία, καθώς και ανάμεσα στα ΙΧ και στην κίνηση με το περπάτημα, ειδικά τις ώρες αιχμής, όπου υπάρχει κίνηση και η ταχύτητα είναι χαμηλή.

Ένας από τους πρώτους που χρησιμοποίησε GPS και επιταχυνσιόμετρο από κινητό τηλέφωνο ήταν ο Reddy όπου ο αλγόριθμος που ανέπτυξε προέβλεπε το πότε ο χρήστης ήταν ακίνητος, περπατούσε, έτρεχε, έκανε ποδήλατο ή βρισκόταν σε μηχανοκίνητο όχημα (Reddy, et al., 2008) & (Reddy, et al., 2010). Στην πρώτη μελέτη του, τοποθέτησε το κινητό σε διαφορετικά σημεία του χρήστη όπως στο μπράτσο, στο χέρι, ακόμα και μέσα σε τσάντα, ώστε να συμπεράνει αν εμφανιστεί κάποια διαφορά στη συνολική ακρίβεια εντοπισμού του μέσου ανάλογα με την τοποθέτηση του τηλεφώνου. Παρότι βρέθηκαν διαφορές της τάξης του 2%, στην πραγματικότητα δεν είναι άνετο για τον χρήστη να πρέπει να τοποθετεί το κινητό του σε συγκεκριμένα σημεία. Επομένως, στη δεύτερη μελέτη του αφήνει ελεύθερο το χρήστη να τοποθετήσει το κινητό σε όποιο σημείο τον ικανοποιεί περισσότερο. Χρησιμοποίησε 4 αισθητήρες: επιταχυνσιόμετρο, GSM, GPS και WiFi και διαπίστωσε πως με το συνδυασμού GPS και επιταχυνσιομέτρου μόνο, η ακρίβεια έπεφτε πολύ λίγο σε σύγκριση με τη χρήση και των τεσσάρων αισθητήρων μαζί. Συγκεκριμένα, μειωνόταν κατά 0.6%, δηλαδή προέκυπτε συνολική ακρίβεια 91.3%. Επομένως, συμπέρανε πως WiFi και GSM προσέφεραν λίγες πληροφορίες σε σχέση με τους άλλους 2 αισθητήρες, με αποτέλεσμα να θεωρήσει αυτούς τους αισθητήρες μη σημαντικούς και να τους αφαιρέσει από το μοντέλο του.

Πλέον πολλές έρευνες (Zhao, et al., 2015), (Ferrer, et al., 2014) & (Shafique, et al., 2016), γίνονται μέσω εφαρμογών στο κινητό όπου συλλέγουν αυτόματα τα δεδομένα και ο χρήστης το μόνο που έχει να κάνει είναι να εξακριβώσει το μέσο με το οποίο ταξίδεψε, τους χρόνους εκκίνησης και διάρκειας ταξιδιού και γενικά να επιβεβαιώσει το ταξίδι του, μέσα από μια πλατφόρμα εξοικειωμένη στο χρήστη.

Μια τέτοια μελέτη που παρουσιάζει μεγάλο ενδιαφέρον, έχει γίνει χρησιμοποιώντας την εφαρμογή MovieSmarter (Geurs, et al., 2015). Σε αυτήν, τα ταξίδια που συλλέχθηκαν μέσα σε δύο εβδομάδες ήταν πάνω από 18.000, ενώ συμμετείχαν περίπου 600 άτομα. Η συγκεκριμένη εφαρμογή χρησιμοποιούσε το GPS μόνο όταν έβλεπε σημαντικές αλλαγές στην τοποθεσία του ατόμου, με αποτέλεσμα να αναγνωρίζει πως ο χρήστης βρίσκεται σε κίνηση. Μόνο τότε, λοιπόν, ξεκινούσε η συλλογή δεδομένων GPS με συχνότητα μια φορά ανά δύο δευτερόλεπτα

(0.5Hz). Αυτό επέτρεπε την ομαλή χρήση του τηλεφώνου κατά τη διάρκεια της ημέρας καθώς δε λειτουργούσε το GPS όλο το 24ωρο όπως συμβαίνει σε άλλες περιπτώσεις, με αποτέλεσμα να μην καταναλώνεται η μπαταρία του κινητού γρήγορα. Οι τρόποι μεταφοράς στους οποίους έγινε η πρόβλεψη ήταν το περπάτημα, το ποδήλατο, το αυτοκίνητο, το τρένο και σε μια κατηγορία μαζί ήταν το λεωφορείο, το τραμ και το μετρό.

Συνοψίζοντας, οι συσκευές GPS έχουν χρησιμοποιηθεί από πολλούς ερευνητές για την ανίχνευση του μέσου μετακίνησης, χρησιμοποιώντας είτε αλγορίθμους που βασίζονται σε κανόνες (Stopher, et al., 2008), (Bohte, et al., 2009), (Chen, et al., 2010) & (Gong, et al., 2012), είτε αλγορίθμους μηχανικής μάθησης (Bolbol, et al., 2012), (Maurer, et al., 2006), (Ellis, et al., 2014) & (Su, et al., 2015). Πρέπει να αναφερθεί πως έχει πραγματοποιηθεί μια πολύ ενδιαφέρουσα παρουσίαση όλων των μελετών που αφορούν την αναγνώριση του μέσου μεταφοράς, με βασικό αισθητήρα το GPS από τους Wu και Yang, οι οποίοι αναλύουν και παρουσιάζουν σε συγκεντρωτικούς πίνακες τους επιπλέον αισθητήρες που έχουν χρησιμοποιηθεί σε κάθε έρευνα μέχρι και τα μέσα του 2016, τα μοντέλα που έχουν χρησιμοποιηθεί για να γίνει ο εντοπισμός του μεταφορικού μέσου καθώς και τα ποσοστά ακριβείας που πέτυχαν τα μοντέλα αυτά (Wu, et al., 2016).

2.2.2. ΧΩΡΙΣ ΤΗ ΧΡΗΣΗ ΤΟΥ GPS

Ενώ μέχρι και σήμερα γίνεται πολύ συχνή χρήση του GPS για τον εντοπισμό του μέσου μεταφοράς, δεν παύουν να υπάρχουν αρκετά μειονεκτήματα στη χρήση του. Για αυτό το λόγο, στις περισσότερες νέες έρευνες απουσιάζει το GPS και τη θέση του ως βασικό αισθητήρα έχει πάρει το επιταχυνσιόμετρο. Αυτό γιατί το τελευταίο καταναλώνει πολύ λιγότερη ενέργεια από το GPS. Επίσης, χρειάζεται πολύ λίγο χρόνο για να ξεκινήσει και μπορεί να συλλέγει δεδομένα όλη την ώρα χωρίς να επηρεάζεται το κινητό ούτε στο ελάχιστο. Τέλος, το επιταχυνσιόμετρο είναι ανεξάρτητο από δορυφόρους και πύργους κινητής τηλεφωνίας (Wang, et al., 2010).

Σε μια τέτοια πρώτη προσέγγιση (Wang, et al., 2010), έγινε χρήση δεδομένων μόνο από το επιταχυνσιόμετρο τα οποία και συλλέχθηκαν με συχνότητα 35Hz. Για την πρόβλεψη, έγινε χρήση δένδρων απόφασης, της μεθόδου των K πλησιέστερων γειτόνων και των μηχανών διανυσμάτων υποστήριξης όπου και έγινε προσπάθεια εντοπισμού 6 τρόπων μετακίνησης: ποδήλατο, λεωφορείο, Ι.Χ., τρένο, ακινησία και περπάτημα με σχετικά μεγάλη ακρίβεια. Στη

συνέχεια, αναφέρονται άλλες έρευνες που παρουσιάζουν ενδιαφέρον και έγιναν πάλι με τη χρήση μόνο του επιταχυνσιομέτρου (Siirtola, et al., 2012) & (Shafique, et al., 2015).

Οι Su και Caceres χρησιμοποίησαν επιταχυνσιόμετρο, αισθητήρα βαρύτητας, βαρόμετρο, αισθητήρα φωτός και μαγνητόμετρο ώστε να συλλέξουν δεδομένα από κινητά. Η συχνότητα συλλογής τους ήταν στα 5Hz και χρησιμοποιήθηκαν κινούμενα παράθυρα όπου το κάθε ένα περιείχε 13 δευτερόλεπτα ταξιδιού. Η πρόβλεψη έγινε με μεθόδους μηχανικής εκπαίδευσης και οι κατηγορίες ήταν η κίνηση με λεωφορείο, με τρένο, με Ι.Χ., με ποδήλατο, περπάτημα, τζόκινγκ, ενώ η ακρίβεια ήταν κατά μέσο όρο στο 97%, με καλύτερη μέθοδο να αποδεικνύονται τα δίκτυα Bayes (Su, et al., 2015).

Επειδή η μη χρησιμοποίηση GPS δυσκολεύει τον εντοπισμό των χρόνων εκκίνησης και τερματισμού των ταξιδιών, δημιουργούνται κάποιοι κανόνες ώστε να γίνει αντιληπτή η κίνηση του ατόμου (Eftekhari, et al., 2016). Οι κανόνες αυτοί όμως, εκτός του ότι είναι χρονοβόροι στη δημιουργία τους, δεν είναι πάντα έγκυροι με αποτέλεσμα κάποια ταξίδια να χάνονται. Αυτό το πρόβλημα έρχονται να καλύψουν κάποιες μελέτες, οι οποίες έξυπνα χρησιμοποιούν τα δεδομένα GPS μόνο για την επιβεβαίωση των ταξιδιών. Στη συνέχεια, τα δεδομένα αυτά διαγράφονται ώστε να μη συμμετάσχουν στην εκπαίδευση του μοντέλου. Αποτέλεσμα αυτού, είναι η δημιουργία ενός μοντέλου πρόβλεψης που δε βασίζεται καθόλου σε δεδομένα GPS ενώ όλα τα ταξίδια καταγράφονται με την ακριβή ημερομηνία και διάρκεια πραγματοποίησής τους (Ferrer, et al., 2014) & (Shafique, et al., 2016).

2.3. ΠΡΟΤΥΠΑ ΠΡΟΒΛΕΨΗΣ ΜΕΣΟΥ ΜΕΤΑΚΙΝΗΣΗΣ

Παλαιότερα, η πρόβλεψη του μέσου μεταφοράς γινόταν με διάφορους κανόνες και αλγορίθμους που εφαρμόζονταν από τους ερευνητές. Για παράδειγμα, για κάθε εξαρτημένη μεταβλητή (μέσο μεταφοράς) , επιλέγονταν συγκεκριμένα διαστήματα και όρια τιμών των ανεξάρτητων μεταβλητών (δεδομένα αισθητήρων) και έτσι με βάση τα δεδομένα που συλλέγονταν από τα ταξίδια γινόταν η πρόβλεψη του μοντέλου. Πλέον, με την εξέλιξη της τεχνολογίας και της αυτοματοποίησης, η μηχανική μάθηση έχει κάνει την εμφάνισή της και πολλές έρευνες βασίζονται σε αυτή.

Βασικό κομμάτι των ερευνών που χρησιμοποιούν μηχανική μάθηση είναι ο καθορισμός των ποσοστών των δεδομένων, μετά την απαραίτητη επεξεργασία που έχουν αυτά υποστεί, που θα

λειτουργήσουν ως δεδομένα εκπαίδευσης, δοκιμής και επαλήθευσης, χωρίς πάντα η τελευταία κατηγορία να είναι απόλυτα αναγκαία. Σε μια έρευνα (Nham, et al., 2012), τα δεδομένα εκπαίδευσης και δοκιμής αποτελούσαν το 70% και 30% των συνολικών δεδομένων αντίστοιχα. Σε μια παρόμοια μελέτη, τα συνολικά δεδομένα είχαν χωριστεί σε δεδομένα εκπαίδευσης και δοκιμής με ποσοστό 90% και 10% αντίστοιχα (Nick, et al., 2010), ενώ πάλι σε άλλη μελέτη, τα δεδομένα εκπαίδευσης αφορούσαν το 50% των συνολικών δεδομένων ενώ τα δεδομένα δοκιμής αφορούσαν το υπόλοιπο 50% (Figo, et al., 2010). Σε κάποιες έρευνες (Ferrer, et al., 2014) γινόταν χρήση δεδομένων από GPS μόνο για την επικύρωση των ταξιδιών, ενώ ο εντοπισμός του τρόπου μεταφοράς γινόταν και πάλι από τα δεδομένα του επιταχυνσιομέτρου.

Διάφορες έρευνες έχουν συγκρίνει τη μέθοδο των τυχαίων δασών με άλλους αλγορίθμους για τον σκοπό της ανίχνευσης του τρόπου μετακίνησης και φτάνουν στο ίδιο συμπέρασμα, το οποίο είναι ότι η μέθοδος των τυχαίων δασών είναι ανώτερη των υπολοίπων ως προς τον επιδιωκόμενο σκοπό. Για παράδειγμα, μια μελέτη έκανε σύγκριση της μεθόδου των τυχαίων δασών με τους αλγορίθμους στατιστικής κατηγοριοποίησης όπως Naïve Bayes, τα δίκτυα Bayes, τη χρήση δένδρων απόφασης (decision trees) και τη χρήση πολυεπίπεδων Perceptrons (multilayer perceptrons) (Stenneth, et al., 2011). Μια άλλη έρευνα, συμπεριέλαβε τη χρήση νευρωνικών δικτύων και τις μηχανές διανυσμάτων υποστήριξης (Support Vector Machines SVM) μαζί με τη μέθοδο των τυχαίων δασών (Abdulazim, et al., 2013). Ακόμα, μια έρευνα χρησιμοποίησε τη μέθοδο των τυχαίων δασών, τη μέθοδο των Κ πλησιέστερων γειτόνων (KNN), τις μηχανές διανυσμάτων υποστήριξης, τους αλγορίθμους στατιστικής κατηγοριοποίησης Naïve Bayes και ταξινομητές με χρήση δένδρων απόφασης (decision trees) (Ellis, et al., 2014). Τέλος, έχει γίνει και σύγκριση των μηχανών διανυσμάτων υποστήριξης, της προσαρμοστικής ώθησης (adaptive boosting), τους ταξινομητές με χρήση δένδρων απόφασης με τη μέθοδο των τυχαίων δασών (random forests) (Shafique, et al., 2015). Όλες οι παραπάνω έρευνες, απέδειξαν πως η μέθοδος των τυχαίων δασών έχει τη μεγαλύτερη ακρίβεια ως προς τον προσδιορισμό του τρόπου ταξιδιού.

2.4. ΠΑΡΑΓΟΝΤΕΣ ΕΠΙΡΡΟΗΣ ΠΡΟΒΛΕΨΗΣ

2.4.1. ΤΥΠΟΣ ΜΕΣΟΥ

Παλαιότερα, οι έρευνες που γινόντουσαν με βάση τη συγκεκριμένη θεματολογία, δεν προέβλεπαν ακριβώς τον τρόπο μεταφοράς του χρήστη. Για παράδειγμα, δε γινόταν πρόβλεψη του τρόπου μεταφοράς του χρήστη για το αν κινούταν εκείνος με τα πόδια, ή με ποδήλατο ή με αμάξι ή ακόμα και με λεωφορείο, αλλά κατηγοριοποιούνταν όλα τα προηγούμενα σε μηχανοκίνητα ή μη μηχανοκίνητα μέσα (Oliver, et al., 2010) & (Zheng, et al., 2008). Αυτό συνέβαινε διότι δεν ήταν εύκολη η χρήση πολλών αισθητήρων μαζί όπως επιταχυνσιόμετρο και γυροσκόπιο μαζί με GPS, με αποτέλεσμα να είναι δύσκολος ο εντοπισμός του κάθε μέσου ξεχωριστά. Με την εξέλιξη της τεχνολογίας, οι προβλέψεις των μοντέλων μπορούσαν να είναι πλέον πιο συγκεκριμένες για τα μη μηχανοκίνητα μέσα. Για παράδειγμα το ποδήλατο, το περπάτημα, το τρέξιμο, το τζόκινγκ μπορούσαν να προβλεφθούν ξεχωριστά, ενώ τα μηχανοκίνητα παρέμεναν ακόμα όλα σε μια κατηγορία (Reddy, et al., 2008). Σήμερα, τα μοντέλα πρόβλεψης μπορούν να εντοπίσουν το μέσο που χρησιμοποιείται κάθε φορά είτε είναι ποδήλατο, είτε αμάξι, είτε λεωφορείο είτε τρένο, με μεγάλη ακρίβεια (Geurs, et al., 2015) & (Zong, et al., 2015).

2.4.2. ΣΥΧΝΟΤΗΤΑ ΔΕΔΟΜΕΝΩΝ

Σημαντικό ρόλο σε μια έρευνα, παίζει η συχνότητα με την οποία η συσκευή που χρησιμοποιείται συλλέγει τα δεδομένα. Για παράδειγμα, κάποιες μελέτες συλλέγουν δεδομένα από το επιταχυνσιόμετρο με συχνότητα 35Hz (Yang, et al., 2009), άλλες (Feng, et al., 2013) με συχνότητα 10Hz, ενώ άλλες (Eftekhari, et al., 2016) με συχνότητα 2Hz. Σε μια άλλη έρευνα, (Reddy, et al., 2008) δεδομένα από τον αισθητήρα επιταχυνσιομέτρου συλλέχθηκαν με συχνότητα 35Hz ενώ του GPS με συχνότητα 1Hz, ενώ αλλού συλλέχθηκαν δεδομένα από επιταχυνσιόμετρο (Ferrer, et al., 2014) με συχνότητα 1Hz ενώ δεδομένα από GPS (τα οποία δε χρησιμοποιήθηκαν στην πρόβλεψη) συλλέχθηκαν με συχνότητα 0.1Hz. Φαίνεται, λοιπόν, πως ο κάθε ερευνητής επιλέγει διαφορετικό τρόπο ώστε να εξάγει τα δεδομένα που θα χρησιμοποιήσει στο μοντέλο του. Όπως αναφέρει ο Hemminki (Hemminki, et al., 2013), η συχνότητα με την οποία συλλέγονται τα δεδομένα GPS δε χρειάζεται να ξεπερνάει το 1Hz,

αφού έτσι μαζεύονται όσες πληροφορίες χρειάζονται για την έρευνα ενώ ταυτόχρονα η μπαταρία της συσκευής διαρκεί περισσότερο. Από την άλλη, οι αισθητήρες επιταχυνσιομέτρου, μαγνητομέτρου, γυροσκοπίου και λοιποί, μπορούν να λειτουργούν όλη την ημέρα ξοδεύοντας σε πολύ μικρό ποσοστό τη μπαταρία της συσκευής. Οι συνηθισμένες συχνότητες που χρησιμοποιούνται για την εξαγωγή δεδομένων από το επιταχυνσιόμετρο είναι 1-10Hz (Reddy, et al., 2010), (Wang, et al., 2010) & (Montini, et al., 2015).

2.4.3. ΧΡΗΣΗ ΠΑΡΑΘΥΡΩΝ

Η χρήση παραθύρων χρησιμοποιήθηκε στην παρούσα διπλωματική εργασία επομένως περιγράφεται αναλυτικότερα στο Κεφάλαιο 3. Ενδεικτικά αναφέρεται, πως η χρήση των παραθύρων αυτών, έχει στόχο να μειώσει το θόρυβο των ακατέργαστων δεδομένων που συλλέγονται από τους αισθητήρες, έτσι ώστε με τα νέα και επεξεργασμένα πια δεδομένα, να μπορέσει να γίνει η πρόβλεψη του τρόπου μεταφοράς. Η διάρκεια των παραθύρων αυτών μπορεί να είναι από 1 δευτερόλεπτο (Reddy, et al., 2010) έως και 60 δευτερόλεπτα (Feng, et al., 2013). Τέλος, κάποια δευτερόλεπτα από δύο συνεχόμενα παράθυρα μπορεί να αλληλοκαλύπτονται, ανάλογα με τις προθέσεις και το σκοπό του εκάστοτε ερευνητή (Eftekhari, et al., 2016).

2.5. ΒΑΣΙΚΑ ΣΥΜΠΕΡΑΣΜΑΤΑ ΒΙΒΛΙΟΓΡΑΦΙΑΣ

Η χρήση του GPS είναι διαδεδομένη, κάτι που έγινε φανερό στο Κεφάλαιο 2, όμως η κατανάλωση της μπαταρίας που απαιτεί είναι τεράστια. Παρότι η χρήση του προσφέρει μεγάλη ακρίβεια στον εντοπισμό του μέσου μεταφοράς, οι σύγχρονες έρευνες το εγκαταλείπουν ψάχνοντας νέους τρόπους, με τη βοήθεια άλλων αισθητήρων, για τη δημιουργία των μοντέλων πρόβλεψης. Το επιταχυνσιόμετρο είναι ο πιο διαδεδομένος αισθητήρας από αυτούς και προσφέρει αρκετά καλά ποσοστά ακρίβειας στην πρόβλεψη, ωστόσο απαιτείται περεταίρω βελτίωση.

Επίσης, έχει γίνει φανερό πως ο διαχωρισμός του πεζού χρήστη με εκείνον που μετακινείται με μηχανοκίνητο όχημα, είναι εύκολος και τα μοντέλα πρόβλεψης λειτουργούν με πολύ μεγάλο ποσοστό επιτυχίας στη συγκεκριμένη περίπτωση. Ωστόσο, υπάρχει δυσκολία στην

ακριβή πρόβλεψη του μέσου μεταφοράς όταν αυτό κυμαίνεται μεταξύ των μηχανοκίνητων οχημάτων.

Τέλος, αξίζει να αναφερθεί πως η πλειοψηφία των ερευνών πλέον χρησιμοποιεί μηχανική μάθηση για τα μοντέλα πρόβλεψης. Έχει αποδειχθεί ότι αυτά προσφέρουν μεγαλύτερη ακρίβεια από οποιαδήποτε άλλη μέθοδο πρόβλεψης και συγκεκριμένα η χρήση των τυχαίων δασών αποδεικνύεται πως είναι η βέλτιστη από αυτές.

Η παρούσα εργασία έρχεται να καλύψει συγκεκριμένα κενά της υπάρχουσας βιβλιογραφίας, καθώς αποφεύγεται η χρήση του GPS στο μοντέλο πρόβλεψης και τα μέσα μεταφοράς που χρησιμοποιήθηκαν ήταν μηχανοκίνητα. Ταυτόχρονα, επιτυγχάνονται εξαιρετικά υψηλά ποσοστά επιτυχίας με τη χρήση των τυχαίων δασών.

3. ΣΥΛΛΟΓΗ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

Τα δεδομένα των ταξιδιών του κάθε ατόμου έχουν ζωτική σημασία για τη διαχείριση της σημερινής υποδομής των μεταφορών καθώς και για το σχεδιασμό μελλοντικών εγκαταστάσεων. Επίσης, παρέχουν τη βάση για νέες πολιτικές που εφαρμόζονται στο πλαίσιο της διαχείρισης της ζήτησης των μεταφορών.

Ένα αρχείο δεδομένων, περιέχει σειρές, στήλες και πίνακες, τα οποία είναι τακτοποιημένα με τρόπο όπου μπορεί να βρεθεί εύκολα όποια σχετική πληροφορία χρειάζεται. Τα δεδομένα αυτά ενημερώνονται, επεκτείνονται και διαγράφονται καθώς νέα δεδομένα προστίθενται. Στην παρούσα εργασία, τα δεδομένα που χρησιμοποιήθηκαν είναι επιταχύνσεις, γωνιακές ταχύτητες και αλλαγές της κατεύθυνσης ταξιδιού ως προς τον πραγματικό βορρά, με τη βοήθεια φυσικά των κινητών τηλεφώνων.

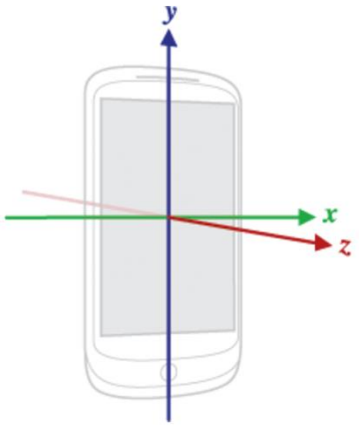
3.1. ΑΙΣΘΗΤΗΡΕΣ

Οι αισθητήρες ανήκουν στην κατηγορία των μικροηλεκτρομηχανικών συσκευών (Microelectromechanical systems, MEMS) και είναι τοποθετημένοι μέσα στο κινητό τηλέφωνο. Οι πιο γνωστοί από αυτούς είναι το επιταχυνσιόμετρο και το γυροσκόπιο και είναι αυτοί που χρησιμοποιήθηκαν για τη δημιουργία του μοντέλου στην παρούσα εργασία.

3.1.1. ΕΠΙΤΑΧΥΝΣΙΟΜΕΤΡΟ

Το επιταχυνσιόμετρο είναι μια ηλεκτρομηχανική συσκευή που μετράει την επιτάχυνση, συνήθως σε σχέση με την επιτάχυνση της βαρύτητας, η οποία αναπτύσσεται κατά την κίνηση της συσκευής. Μπορεί να ανιχνεύσει επιταχύνσεις ως προς τρεις άξονες (x , y, z) σε σχέση πάντα με την επιτάχυνση της βαρύτητας (Εικόνα 3.1). Για παράδειγμα, ένα επιταχυνσιόμετρο που βρίσκεται σε κατάσταση ηρεμίας στην επιφάνεια της Γης, θα καταγράψει επιτάχυνση προς τα πάνω $9.81 m/s^2$ λόγω της βαρύτητας της Γης. Αντίθετα, όταν βρίσκεται σε ελεύθερη πτώση

(που πέφτει προς το κέντρο της γης με ρυθμό περίπου 9.81m/s^2) θα μετρήσει επιτάχυνση ίση με μηδέν.



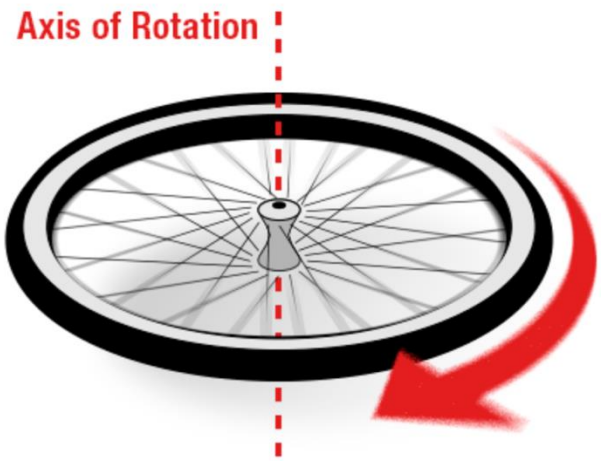
Εικόνα 3.1 Κατεύθυνση συντεταγμένων με την οποία καταγράφονται τα δεδομένα από το επιταχυνσιόμετρο (Πηγή:(Android, Google))

Τα επιταχυνσιόμετρα έχουν πολλαπλές εφαρμογές στη βιομηχανία και την επιστήμη. Τα υψηλής ευαισθησίας επιταχυνσιόμετρα χρησιμοποιούνται στην πλοήγηση αεροσκαφών και πυραύλων. Επίσης, χρησιμοποιούνται σε μηχανήματα και κατασκευές για τον εντοπισμό δονήσεων, σε τάμπλετ και σε ψηφιακές κάμερες ώστε η εικόνα που απεικονίζεται στην οθόνη να είναι πάντα προς τα πάνω, καθώς και σε drones για τη σταθεροποίησή τους στον αέρα (wikipedia).

Ο αισθητήρας αυτός, λοιπόν, εξάγει τα αποτελέσματά του σε ακατέργαστα δεδομένα με τη βοήθεια λογισμικού, όπου η πρώτη στήλη αφορά τη χρονική στιγμή όπου πάρθηκαν τα δεδομένα και οι υπόλοιπες τρεις αφορούν την επιτάχυνση που εντοπίστηκε εκείνη τη στιγμή στον κάθε άξονα ξεχωριστά, με συντελεστή την επιτάχυνση της βαρύτητας. Η συχνότητα εξαγωγής των δεδομένων διαφέρει, ανάλογα με το επιθυμητό αποτέλεσμα. Τα παραπάνω χαρακτηριστικά χρησιμοποιήθηκαν στην παρούσα εργασία στην ανάπτυξη του μοντέλου πρόβλεψης.

3.1.2. ΓΥΡΟΣΚΟΠΙΟ

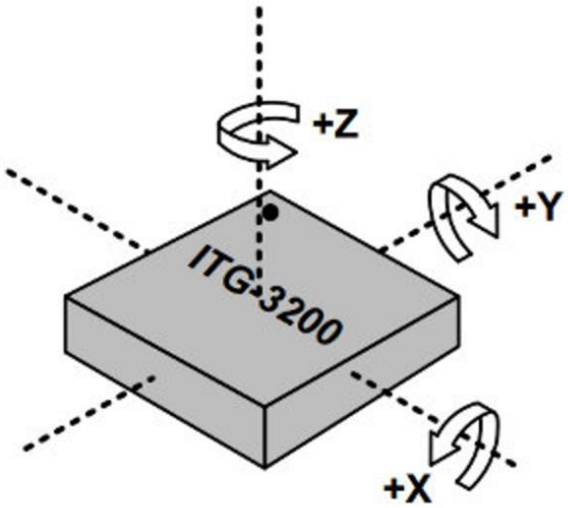
Όταν κάτι περιστρέφεται γύρω από έναν άξονα, η γωνιακή ταχύτητα κάνει την εμφάνισή της. Ένας περιστρεφόμενος τροχός, για παράδειγμα αυτός της Εικόνας 3.2, μπορεί να μετρήσει την ταχύτητα αυτή σε rad/s (sparkfun).



Εικόνα 3.2 Τροχός που περιστρέφεται ως προς τον άξονα z (Πηγή: (sparkfun))

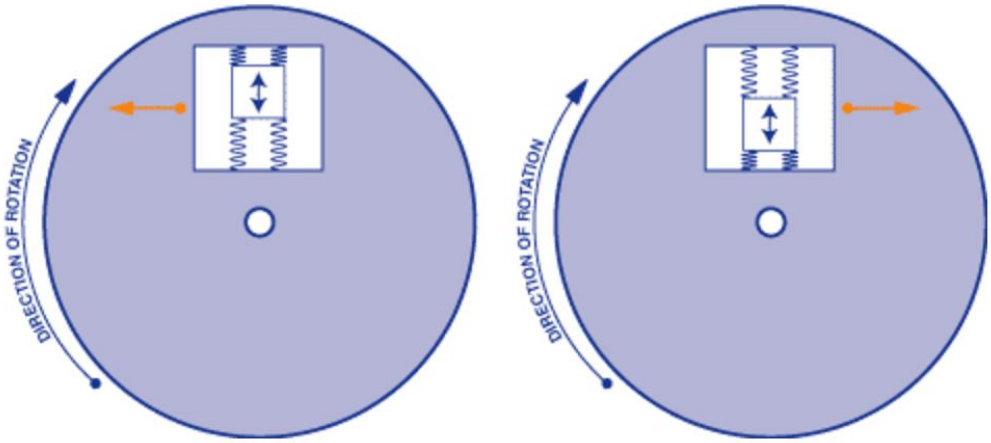
Αν συνδεθεί ένας αισθητήρας στον παραπάνω τροχό, μπορεί να μετρηθεί η γωνιακή ταχύτητα στον άξονα z του γυροσκοπίου, ενώ στους άλλους δύο άξονες δε θα υπάρξει καμία περιστροφή. Για παράδειγμα, αν ο τροχός κάνει μια ολόκληρη περιστροφή σε ένα δευτερόλεπτο τότε θα έχει γωνιακή ταχύτητα 2π rad/s. Η κατεύθυνση της περιστροφής είναι επίσης σημαντική, αν δηλαδή γίνεται σύμφωνα με τους δείκτες του ρολογιού ή αντίθετα.

Γυροσκόπια τα οποία εντοπίζουν τη γωνιακή ταχύτητα σε τρεις άξονες, όπως αυτό της Εικόνας 3.3, είναι πιο διαδεδομένα πλέον καθώς βρίσκονται και μέσα στα ίδια τα κινητά τηλέφωνα. Επίσης, χρησιμοποιούνται σε αντικείμενα τα οποία περιστρέφονται πολύ αργά, όπως είναι τα αεροσκάφη.



Εικόνα 3.3 Το γυροσκόπιο ITG3200 και οι θετικές κατευθύνσεις των αξόνων του (Πηγή: (sparkfun))

Τα αεροσκάφη περιστρέφονται λίγες μοίρες σε κάθε άξονα, οπότε ο εντοπισμός των μικρών αυτών αλλαγών στην κατεύθυνσή τους, που γίνεται με τη βοήθεια γυροσκοπίου και βοηθά στη σταθεροποίησή τους κατά τη διάρκεια της πτήσης. Όταν το γυροσκόπιο περιστρέφεται, μια μικρή μάζα μετακινείται καθώς η γωνιακή ταχύτητα αλλάζει. Η κίνηση αυτή μετατρέπεται σε χαμηλού ρεύματος ηλεκτρικά σήματα τα οποία μπορούν να ενισχυθούν και να διαβαστούν. Το εσωτερικό ενός γυροσκοπίου φαίνεται στην Εικόνα 3.4.

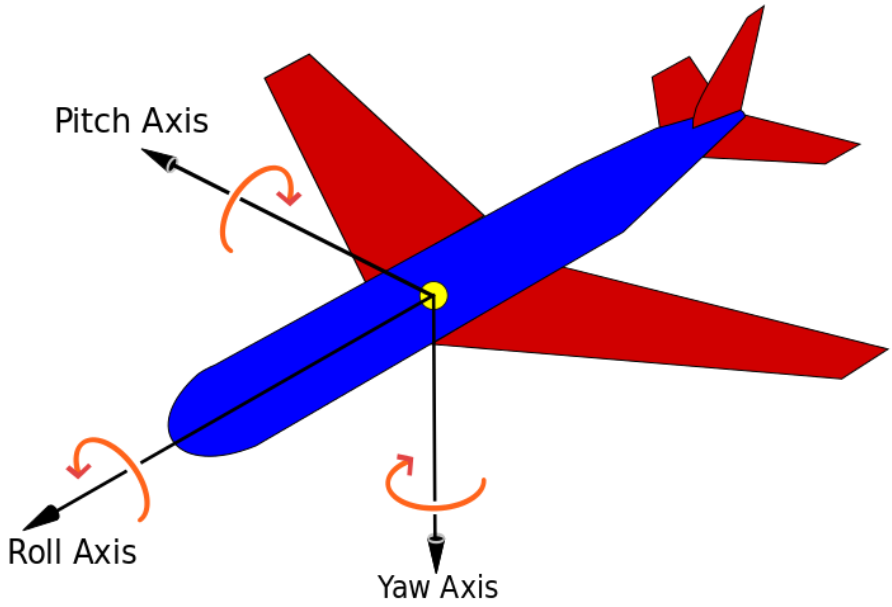


Εικόνα 3.4 Εσωτερικό αισθητήρα γυροσκοπίου (Πηγή: (sparkfun))

Στην παρούσα εργασία χρησιμοποιήθηκαν οι γωνιακές ταχύτητες που μετρήθηκαν από το γυροσκόπιο ως προς x, y, z για την ανάπτυξη του μοντέλου πρόβλεψης.

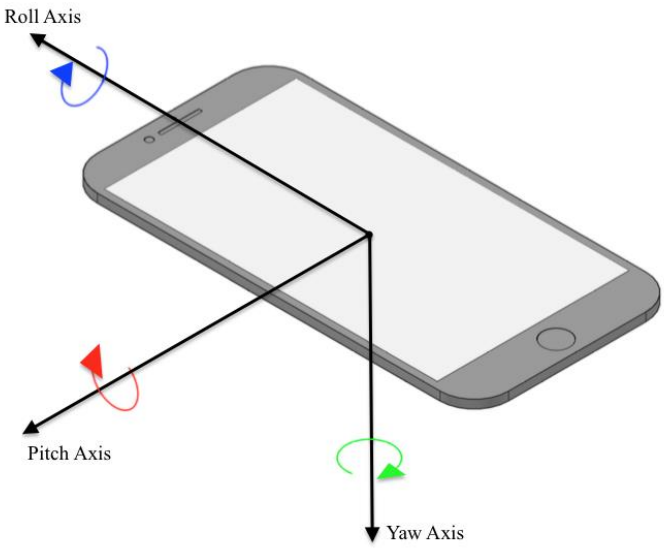
3.1.3. ΑΙΣΘΗΤΗΡΑΣ ΠΡΟΣΑΝΑΤΟΛΙΣΜΟΥ

Ο αισθητήρας αυτός, εμπλουτίζει τα δεδομένα που εξάγει το επιταχυνσιόμετο παρέχοντας πληροφορίες σχετικά με τη γωνιακή κίνηση. Συγκεκριμένα, μετράει τρεις διαφορετικές τιμές σχετικά με την τοποθέτηση του κινητού: yaw, δηλαδή την κατεύθυνση του τηλεφώνου σε μοίρες ως προς τον πραγματικό Βορρά, pitch, δηλαδή την περιστροφή του τηλεφώνου σε rad ως προς τον κατακόρυφο άξονα που ενώνει την κορυφή με τη βάση του και roll, δηλαδή την περιστροφή του τηλεφώνου σε rad ως προς τον οριζόντιο άξονα που ενώνει τις δύο πλευρές του. Τα συγκεκριμένα χαρακτηριστικά χρησιμοποιούνται στην κίνηση και τον προσανατολισμό των αεροπλάνων και γίνονται πιο κατανοητά με τη βοήθεια της Εικόνας 3.5.



Εικόνα 3.5 Η θετική κατεύθυνση των Yaw, Pitch και Roll, σύμφωνα με τον κανόνα του δεξιού χεριού (Πηγή: (wikipedia))

Η ίδια λογική όπου το αεροπλάνο κατά τη διάρκεια των περιστροφών του καταγράφονται τιμές για τις τρεις αυτές μεταβλητές στους παραπάνω άξονες, ισχύει και στα κινητά τηλέφωνα τα οποία με τη βοήθεια του αισθητήρα προσανατολισμού, ο οποίος βασίζεται σε λογισμικό, μπορούν να καταγράψουν τη θέση του κινητού οποιαδήποτε χρονική στιγμή. Στην Εικόνα 3.6 φαίνεται η θετική κατεύθυνση των αξόνων για αυτές τις μεταβλητές που ισχύουν στα κινητά τηλέφωνα.



Εικόνα 3.6 Θετική κατεύθυνση των Yaw, Pitch και Roll όπως καταγράφονται από το κινητό τηλέφωνο

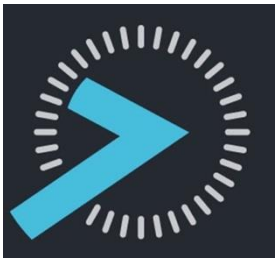
Στην παρούσα εργασία χρησιμοποιήθηκαν τα pitch και roll για την ανάπτυξη του μοντέλου πρόβλεψης, ενώ το yaw παραλήφθηκε.

3.2. ΤΡΟΠΟΣ ΣΥΛΛΟΓΗΣ ΔΕΔΟΜΕΝΩΝ

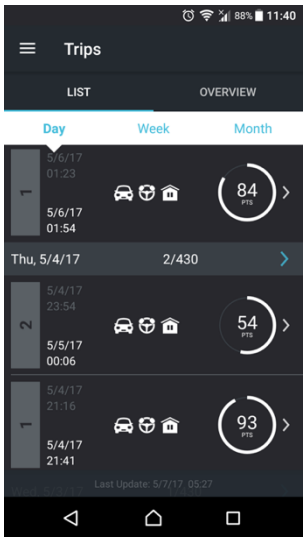
3.2.1. ΕΦΑΡΜΟΓΗ ΣΥΛΛΟΓΗΣ ΔΕΔΟΜΕΝΩΝ

Για τη συλλογή των δεδομένων σε κίνηση με μηχανοκίνητο όχημα, καθώς και για την μετέπειτα πρόβλεψη του μέσου μεταφοράς χρησιμοποιήθηκε η εφαρμογή Oseven Telematics (Εικόνα 37, Εικόνα 3.8). Όλα τα δεδομένα που απεικονίζονται σε πίνακες και διαγράμματα καθώς και τα μοντέλα που δημιουργήθηκαν για την πρόβλεψη του μέσου μεταφοράς, προήλθαν αποκλειστικά και μόνο από αυτή την εφαρμογή.

Στην παρούσα διπλωματική εργασία, τα δεδομένα που συλλέχθηκαν από το GPS αφαιρέθηκαν και δε χρησιμοποιήθηκαν ούτε στην εκπαίδευση του μοντέλου, αλλά ούτε και στην πρόβλεψη του μέσου μεταφοράς. Αυτό συνέβη καθώς στόχος της εργασίας είναι να γίνει μια πρόβλεψη του μέσου μεταφοράς του χρήστη η οποία θα είναι πιο αποτελεσματική ως προς τη μπαταρία του τηλεφώνου.



Εικόνα 3.7 Λογότυπο της εφαρμογής Oseven Telematics



Εικόνα 3.8 Περιήγηση στο μενού της Oseven Telematics

Η διεπαφή του χρήστη με τα δεδομένα που καταγράφει η εφαρμογή δεν είναι άμεση. Αυτό συμβαίνει, διότι σκοπός της εφαρμογής αυτής είναι να βαθμολογεί την οδήγηση του ατόμου, για αυτό και φαίνονται οι πόντοι που πήρε ο χρήστης δίπλα από κάθε ταξίδι του. Επίσης, τα δεδομένα που κατέγραφε η εφαρμογή από τους αισθητήρες, μετά το πέρας του ταξιδιού, ανέβαιναν στο διακομιστή της εταιρείας και δεν αποθηκεύονταν στο κινητό. Αποτέλεσμα αυτού, είναι η εξοικονόμηση αποθηκευτικού χώρου στο κινητό, καθώς η συγκεκριμένη εφαρμογή δεν καταγράφει μόνο δεδομένα επιταχυνσιομέτρου και γυροσκοπίου, αλλά και μαγνητόμετρου, GPS καθώς και άλλων αισθητήρων. Επομένως, τα αρχεία αυτά εξήχθησαν από τον διακομιστή της εταιρείας με τη χρήση των απαιτούμενων ψευδωνύμων και κωδικών ώστε να μπορέσει να γίνει στη συνέχεια η επεξεργασία τους.

3.2.2. ΣΥΧΝΟΤΗΤΑ ΣΥΛΛΟΓΗΣ ΔΕΔΟΜΕΝΩΝ

Η συλλογή των δεδομένων έγινε με συχνότητα 1Hz. Αυτό σημαίνει πως δεδομένα από τους αισθητήρες συλλέγονταν ανά 1 δευτερόλεπτο καθ'όλη τη διάρκεια των ταξιδιών. Όπως αναφέρεται και στο κεφάλαιο 2.4.2, η συγκεκριμένη συχνότητα είναι ιδανική, αφού έτσι δεν επηρεάζεται η μπαταρία του κινητού, ενώ τα δεδομένα που καταγράφονται είναι απόλυτα αντιπροσωπευτικά για το μοντέλο που πρόκειται να δημιουργηθεί. Επίσης, βοηθάει στον γρήγορο υπολογισμό των απαραίτητων πράξεων και επεξεργασίας δεδομένων καθώς δεν απαιτείται τεράστια υπολογιστική ισχύ, σε αντίθεση με τη συχνότητα των 10Hz στην οποία τα δεδομένα που συλλέγονται είναι δεκαπλάσια και η επεξεργασία τους χρονοβόρα. Έπειτα, με συχνότητα στα 0.1Hz σημαντικά δεδομένα θα χάνονταν. Τέλος, η συχνότητα που τέθηκε (1Hz), είναι εφικτή από όλα τα κινητά τηλέφωνα, με αποτέλεσμα να μπορούν να λειτουργήσουν εξίσου καλά ακόμα και παλαιότερα μοντέλα που έχουν τους αισθητήρες αυτούς.

3.2.3. ΑΠΑΙΤΗΣΕΙΣ ΧΡΗΣΤΗ

Οι απαιτήσεις από τον χρήστη για τη συλλογή των δεδομένων ήταν τρεις:

- Κατοχή κινητού τηλεφώνου με αισθητήρα επιταχυνσιομέτρου και γυροσκοπίου.
- Ενεργοποιημένο GPS στο κινητό, ώστε να ξεκινάει η καταγραφή των δεδομένων αυτόματα.

- Επιβεβαίωση του μέσου μεταφοράς μετά το πέρας του ταξιδιού, μέσα από την εφαρμογή.

Κατά την περίοδο συλλογής δεδομένων, ο χρήστης ήταν ελεύθερος να κινηθεί ή να ασχοληθεί με το κινητό του, όπως αυτός ήθελε κατά τη διάρκεια του ταξιδιού, παρόλο που μπορεί έτσι να δημιουργηθεί θόρυβος στα δεδομένα. Με αυτόν τον τρόπο, τα δεδομένα που συλλέχθηκαν αντλούνται από την καθημερινότητα των χρηστών κατά τη μεταφορά τους στο οδικό δίκτυο και δεν περιορίστηκαν από κανόνες για τη διεξαγωγή της έρευνας. Αυτό έγινε, καθώς το μοντέλο που εκπαιδεύτηκε αποσκοπεί σε όλους τους χρήστες μεταφορικών οχημάτων, ενώ είναι γνωστό πως η χρήση του κινητού και η τοποθέτησή του μέσα στο όχημα κατά τη διάρκεια του ταξιδιού διαφέρει από χρήστη σε χρήστη.

Από εκεί και πέρα, τα δεδομένα κάθε ταξιδιού αφορούσαν τις καταγραφές των αισθητήρων ολόκληρου του ταξιδιού αυτού, χωρίς να κόβονται οι στάσεις ή οι στιγμές όπου ο χρήστης κινούταν με σταθερή ταχύτητα. Η ταχύτητα ήταν γνωστή μέσω του GPS, όμως αυτή χρησιμοποιήθηκε μόνο για την επιβεβαίωση των ταξιδιών ενώ στη συνέχεια διαγράφηκε και δεν έπαιξε κανένα ρόλο στην εκπαίδευση αλλά ούτε και στην πρόβλεψη του μοντέλου.

3.2.4. ΜΕΣΑ ΜΕΤΑΦΟΡΑΣ

Τα μεταφορικά μέσα που χρησιμοποιήθηκαν για την εκπαίδευση και πρόβλεψη του μοντέλου ήταν λεωφορείο, τρόλεϊ, τραμ, τρένο και αυτοκίνητο. Λόγω της αρχιτεκτονικής της Oseven Telematics οι πιθανές επιλογές που είχε ο χρήστης για την επαλήθευση του μέσου μεταφοράς του ταξιδιού του ήταν δύο: Μέσα Μαζικής Μεταφοράς (ή αλλιώς ΜΜΜ) ή αυτοκίνητο. Συνεπώς, όλα τα ταξίδια που πραγματοποιήθηκαν με λεωφορείο, τρόλεϊ, τραμ ή τρένο συμπτύχθηκαν σε μία κατηγορία με την ονομασία ΜΜΜ ενώ η άλλη κατηγορία αποτελούταν αποκλειστικά από τα ταξίδια με αυτοκίνητο. Οι χρήστες που χρησιμοποιούσαν το αυτοκίνητο ήταν αποκλειστικά οδηγοί, ενώ εκείνοι που χρησιμοποιούσαν τα ΜΜΜ ήταν απλά επιβάτες. Η καταγραφή των δεδομένων ξεκινούσε αυτόματα, τη στιγμή που εντοπιζόταν μεγάλη αύξηση ταχύτητας μέσω του GPS. Οι χρήστες ανεξαρτήτου οχήματος, ήταν ελεύθεροι να χρησιμοποιήσουν το κινητό όπως αυτοί ήθελαν, καθώς δεν έπαιξε κανένα ρόλο, στο μοντέλο που δημιουργήθηκε, η τοποθέτηση του κινητού μέσα στο όχημα ή η χρήση του κατά τη διάρκεια του ταξιδιού.

Συνεπώς, τα δύο αυτά μεταφορικά μέσα αποτελούν τη μεταβλητή πρόβλεψης, η οποία έχει τις τιμές: 0 για τα ΜΜΜ και 1 για τα αυτοκίνητα. Ο λόγος που δεν προστέθηκαν στο μοντέλο και άλλες μεταβλητές πρόβλεψης, όπως ο πεζός χρήστης ή εκείνος που τρέχει είναι γιατί αυτές είναι πιο εύκολα αναγνωρίσιμες από τα μοντέλα που έχουν ήδη αναπτυχθεί από την υπάρχουσα βιβλιογραφία (Feng, και συν., 2013), (Stenneth, και συν., 2011) & (Reddy, και συν., 2008).

3.3. ΔΕΙΓΜΑ

Τα δεδομένα που χρησιμοποιήθηκαν για την πραγματοποίηση της παρούσας διπλωματικής εργασίας, αντλήθηκαν από την εφαρμογή Oseven Telematics. Συγκεκριμένα, χρησιμοποιήθηκαν όλα τα ταξίδια που έγιναν από τους χρήστες της εφαρμογής, συμπεριλαμβανομένου και του συγγραφέα του παρόν κειμένου, για το διάστημα μεταξύ Νοεμβρίου 2016 και Φεβρουαρίου 2017. Ο αριθμός των χρηστών, καθώς και ο αριθμός των ταξιδιών που πραγματοποιήθηκαν ανάλογα με το μέσο μεταφοράς που χρησιμοποιήθηκε, παρουσιάζονται στο Διάγραμμα 3.1, ενώ να αναφερθεί ξανά πως η κατηγορία ΜΜΜ περιλαμβάνει ταξίδια που πραγματοποιήθηκαν με λεωφορείο, τρόλεϊ, τραμ ή τρένο.

Διάγραμμα 3.1 Αριθμός χρηστών και ταξιδιών που χρησιμοποιήθηκαν για την ανάπτυξη του μοντέλου

Όπως αναφέρεται και στο Κεφάλαιο 3.2.4, οι χρήστες δεν είχαν κάποια συγκεκριμένη καθοδήγηση για την πραγματοποίηση της έρευνας διότι σκοπός ήταν να προβλεφθεί ο τρόπος μεταφοράς του χρήστη για τυχαία χρήση του κινητού κατά τη διάρκεια του ταξιδιού. Από τα

559 αυτά ταξίδια, τα 381 έγιναν με χρήση αυτοκινήτου ενώ τα 178 με χρήση ΜΜΜ. Επίσης, να αναφερθεί πως τα ταξίδια με το αυτοκίνητο αφορούν 387.978 δευτερόλεπτα ταξιδιού δηλαδή 107.5 ώρες ενώ των ΜΜΜ 141.763 δευτερόλεπτα, δηλαδή 39.5 ώρες. Συνεπώς, συνολικά έχουν καταγραφεί 529.741 δευτερόλεπτα, δηλαδή 147 ώρες ταξιδιού από 27 διαφορετικούς χρήστες και αποτελούν ένα ικανοποιητικό δείγμα για τη δημιουργία ενός αμερόληπτου μοντέλου πρόβλεψης. Στον Πίνακα 3.1 φαίνεται ο συνολικός αριθμός ταξιδιών που πραγματοποιήθηκαν για τις δύο κατηγορίες από κάθε έναν χρήστη.

Πίνακας 3.1 Κατηγοριοποίηση με βάση τους χρήστες και τον τρόπο μεταφοράς τους στο σύνολο των ταξιδιών τους

Κωδικός αριθμός του χρήστη	Συνολικός αριθμός ταξιδιών	Αριθμός ταξιδιών με ΜΜΜ	Αριθμός ταξιδιών με αυτοκίνητο
1	101	41	60
2	3	1	2
3	1	1	0
4	16	16	0
5	29	4	25
6	5	4	1
7	22	4	18
8	17	1	16
9	10	2	8
10	25	15	10
11	86	85	1
12	3	3	0
13	1	1	0
14	7	0	7
15	25	0	25
16	5	0	5
17	23	0	23
18	37	0	37
19	14	0	14
20	9	0	9
21	7	0	7
22	21	0	21
23	23	0	23
24	11	0	11
25	10	0	10
26	35	0	35
27	13	0	13

Από το Διάγραμμα 3.1 γίνεται αμέσως αντιληπτό πως τα δεδομένα που αφορούν στη μετακίνηση με τη χρήση αυτοκινήτου είναι πολύ περισσότερα από εκείνα με τη χρήση των ΜΜΜ και συγκεκριμένα 3 φορές περισσότερα. Επίσης, στον Πίνακα 3.1 φαίνεται πως πολλοί χρήστες δε χρησιμοποίησαν καθόλου λεωφορείο κατά τη διάρκεια των ημερών όπου συλλέχθηκαν τα δεδομένα. Για την αντιμετώπιση του φαινομένου αυτού, αναπτύχθηκε προσέγγιση που εξηγείται στο Κεφάλαιο 5.

3.4. ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

Τα κινητά τηλέφωνα τοποθετούνται συνήθως σε διαφορετικές θέσεις από τους χρήστες, για παράδειγμα, κάποιοι τα βάζουν στη μπροστινή τσέπη του παντελονιού τους, άλλοι στην πίσω, άλλοι μέσα στην τσάντα τους, ενώ άλλοι τα κρατούν στα χέρια τους όταν στέλνουν μηνύματα ή μιλάνε στο τηλέφωνο. Αυτές οι διαφορετικές τοποθεσίες του κινητού, καθιστούν δύσκολη τη χρησιμοποίηση του επιταχυνσιομέτρου αποκλειστικά ως προς τον κάθε άξονα στον οποίο μετρήθηκε η επιτάχυνση, αφού η επιτάχυνση αυτή μετριέται σε συνάρτηση με την επιτάχυνση της βαρύτητας και επομένως εξαρτάται σε ένα βαθμό από τη θέση του κινητού. Συνεπώς, διαφορετικοί προσανατολισμοί του κινητού επηρεάζουν την επιτάχυνση του κάθε άξονα διαφορετικά. Για να λυθεί αυτό το πρόβλημα, όπως έχει γίνει και με άλλες έρευνες (Eftekhari, et al., 2016), (Reddy, et al., 2010), (Wang, et al., 2010) & (Shafique, et al., 2016), αντί να χρησιμοποιηθεί το επιταχυνσιόμετρο ως προς τον κάθε άξονα ξεχωριστά, θα βρεθεί και θα χρησιμοποιηθεί η συνισταμένη των αξόνων του, όπως υπολογίζεται παρακάτω.

$$Συνισταμένη\ επιτάχυνσης = \boldsymbol{accTOT} = \boldsymbol{A} = \sqrt{\boldsymbol{A_x}^2 + \boldsymbol{A_y}^2 + \boldsymbol{A_z}^2} \qquad (3.1)$$

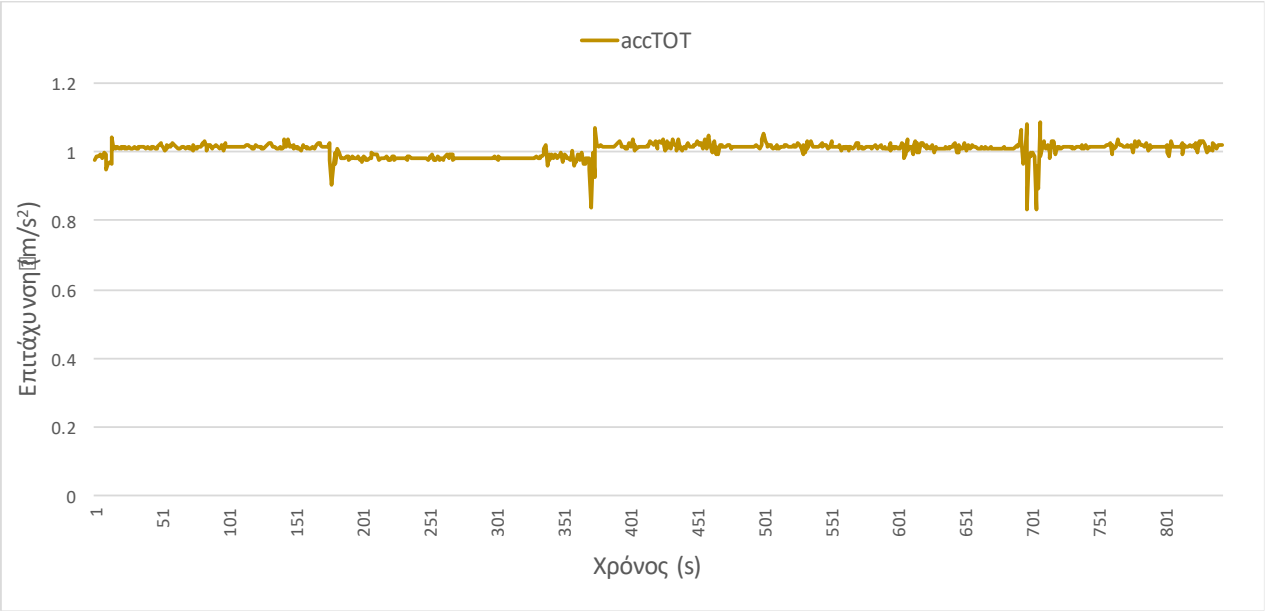
όπου:

$\boldsymbol{A_x}$, $\boldsymbol{A_y}$, $\boldsymbol{A_z}$: οι επιταχύνσεις που μετρήθηκαν από το επιταχυνσιόμετρο ως προς τους άξονες x, y και z αντίστοιχα.

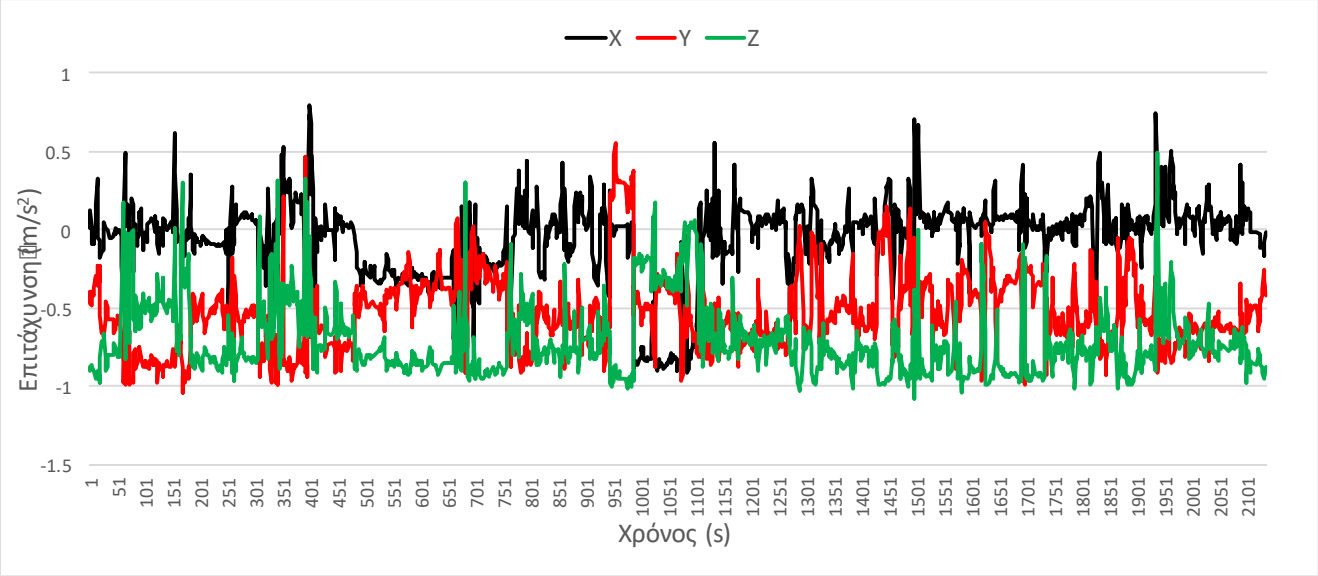
$\boldsymbol{A}$: η συνισταμένη επιτάχυνση των αξόνων.

Η διαφορά αυτή φαίνεται και γραφικά στα Διαγράμματα 3.2 και 3.3, όπου στο πρώτο απεικονίζεται γραφικά η επιτάχυνση του τηλεφώνου σε ένα ταξίδι με ΜΜΜ ως προς τους τρεις άξονες, ενώ στο δεύτερο απεικονίζεται το ίδιο ταξίδι χρησιμοποιώντας τη συνισταμένη τους. Το ίδιο φαίνεται και στα Διαγράμματα 3.4 και 3.5 τα οποία όμως αντιπροσωπεύουν ένα ταξίδι με τη χρήση αυτοκινήτου.

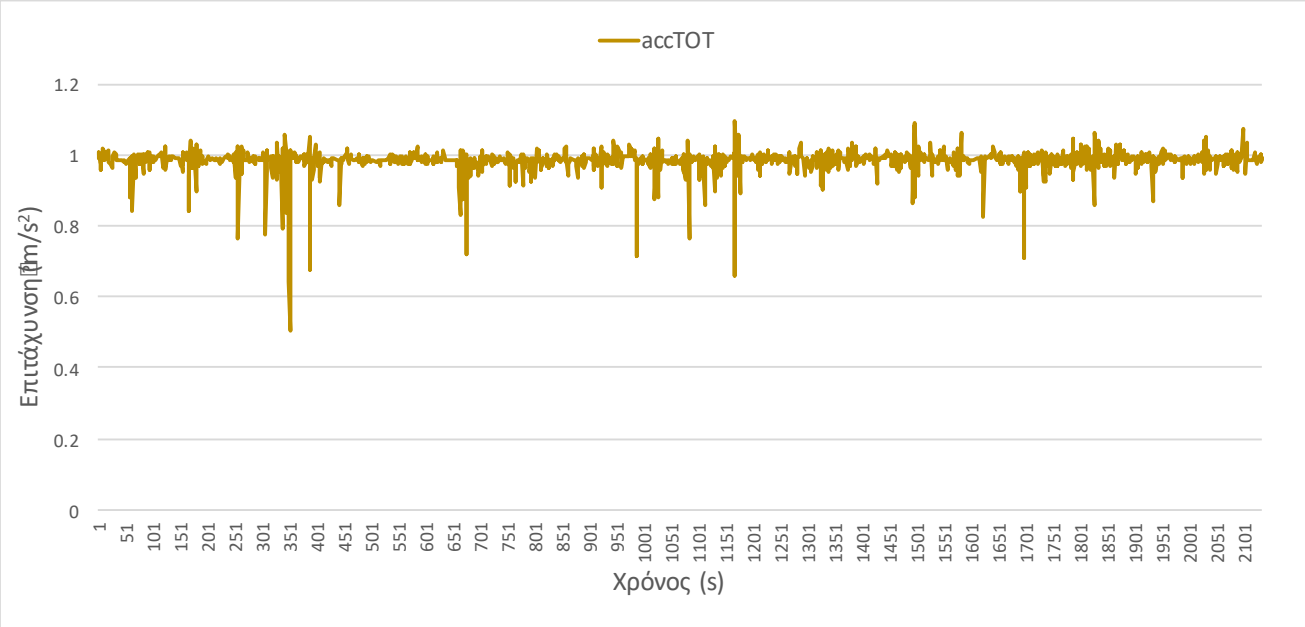
Διάγραμμα 3.2 Μετρήσεις επιταχυνσιομέτρου ενός ταξιδιού με τα ΜΜΜ με τη χρήση 3 αξόνων



Διάγραμμα 3.3 Μετρήσεις επιταχυνσιομέτρου ίδιου ταξιδιού με τα ΜΜΜ με τη χρήση της συνισταμένης



Διάγραμμα 3.4 Μετρήσεις επιταχυνσιομέτρου ενός ταξιδιού με αυτοκίνητο με τη χρήση 3 αξόνων



Διάγραμμα 3.5 Μετρήσεις επιταχυνσιομέτρου ίδιου ταξιδιού με αυτοκίνητο με τη χρήση της συνισταμένης

Το ίδιο έγινε και με το γυροσκόπιο, δηλαδή προτιμήθηκε και πάλι η χρήση μιας μεταβλητής στο μοντέλο πρόβλεψης, της συνισταμένης γωνιακής ταχύτητας του κινητού και όχι η γωνιακή ταχύτητά του ως προς κάθε άξονα χωριστά:

$$\text{Συνισταμένη γωνιακής ταχύτητας} = \mathbf{gyrTOT} = \boldsymbol{G} = \sqrt{\boldsymbol{G_x}^2 + \boldsymbol{G_y}^2 + \boldsymbol{G_z}^2} \qquad (3.2)$$

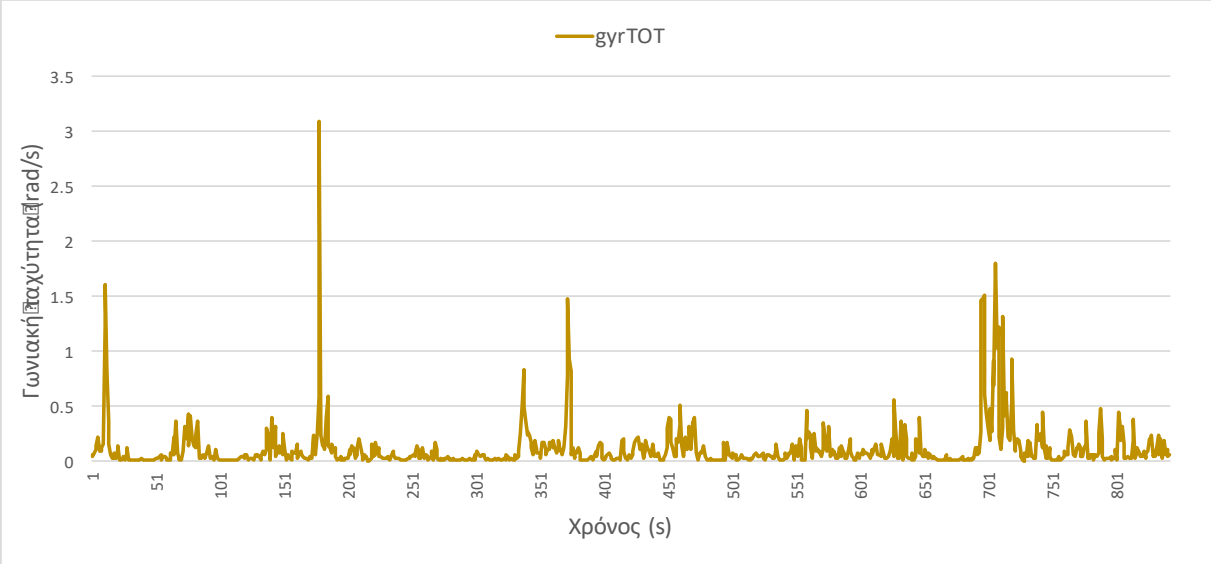
όπου:

$\boldsymbol{G_x}$, $\boldsymbol{G_y}$, $\boldsymbol{G_z}$: οι γωνιακές ταχύτητες που μετρήθηκαν από το γυροσκόπιο ως προς τους άξονες x, y και z αντίστοιχα.

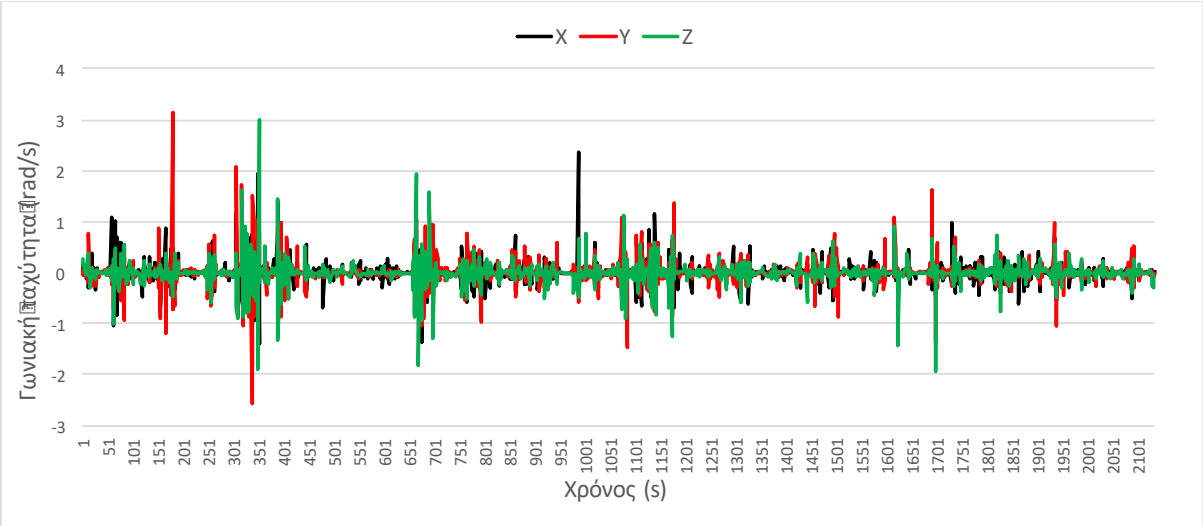
$\boldsymbol{G}$: η συνισταμένη γωνιακή ταχύτητα των αξόνων.

Για να φανεί η διαφορά και εδώ γραφικά, παρουσιάζονται τα Διαγράμματα 3.6 – 3.9, όπου τα δύο πρώτα διαγράμματα αφορούν της καταγραφές του γυροσκοπίου σε ένα ταξίδι με MMM ενώ τα δύο τελευταία αφορούν τις καταγραφές του γυροσκοπίου για ένα ταξίδι με αυτοκίνητο.

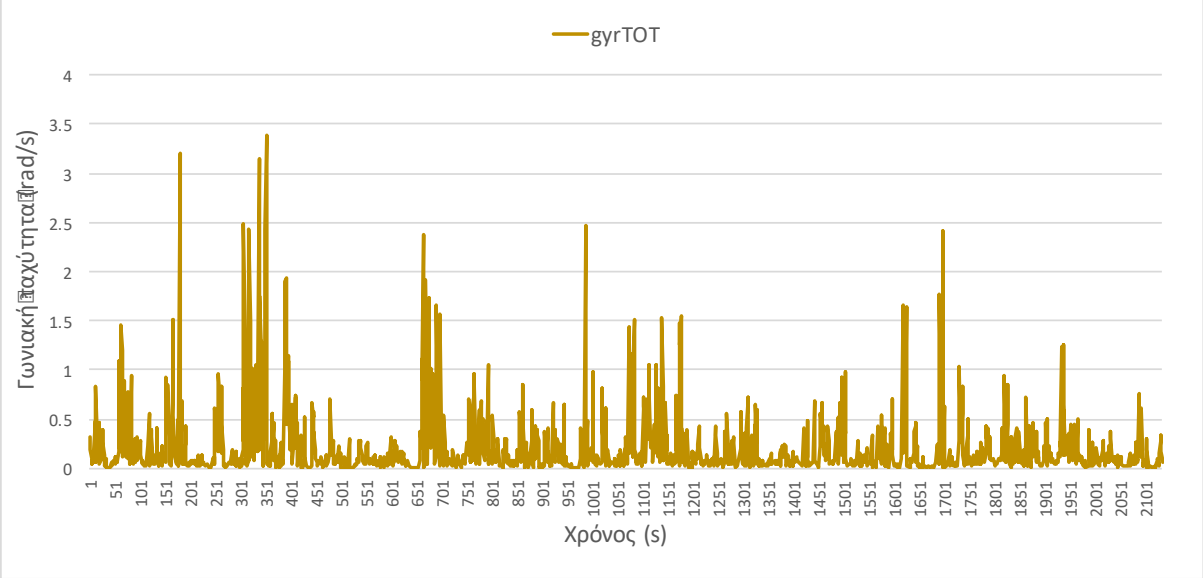
Διάγραμμα 3.6 Μετρήσεις γυροσκοπίου ενός ταξιδιού με τα MMM με τη χρήση 3 αξόνων



Διάγραμμα 3.7 Μετρήσεις γυροσκοπίου ίδιου ταξιδιού με MMM με τη χρήση της συνισταμένης



Διάγραμμα 3.8 Μετρήσεις επιταχυνσιομέτρου ενός ταξιδιού με αυτοκίνητο με τη χρήση 3 αξόνων



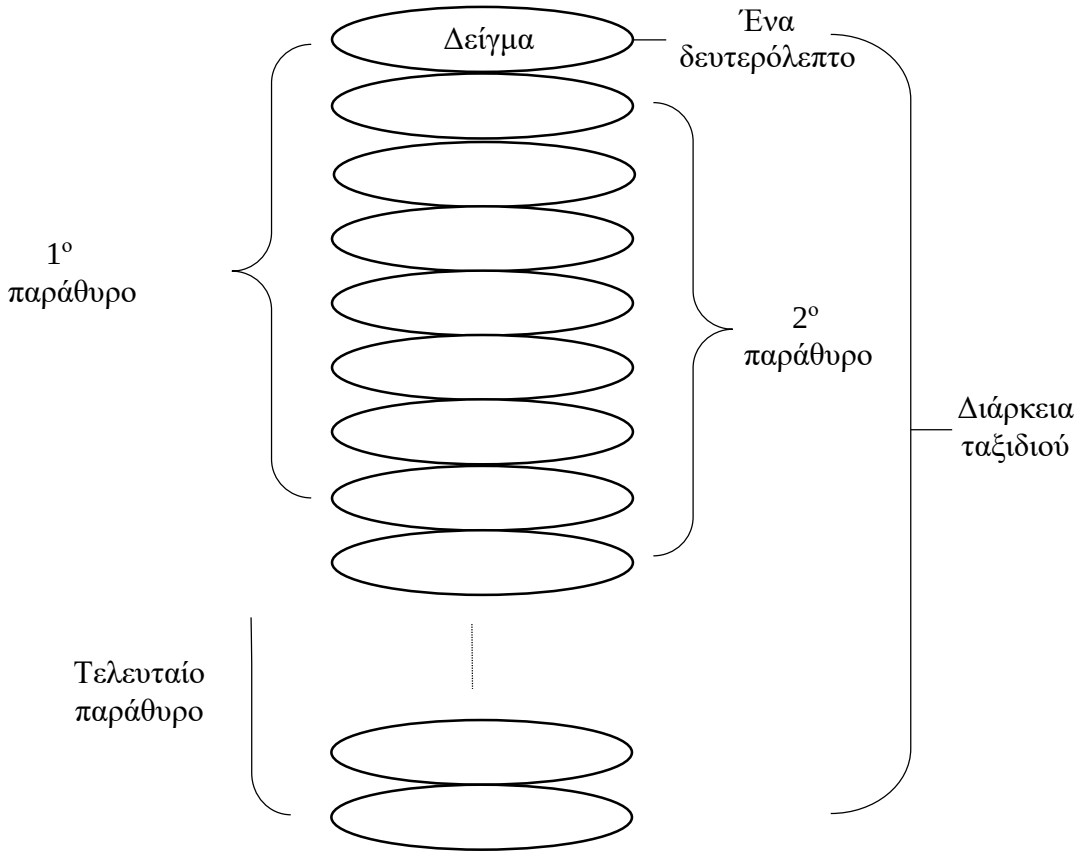
Διάγραμμα 3.9 Μετρήσεις επιταχυνσιομέτρου ίδιου ταξιδιού με αυτοκίνητο με τη χρήση της συνισταμένης

Εκτός από τα accTOT και gyrTOT, τα οποία πήραν το όνομά τους από της αγγλικές λέξεις *total* και τα αρχικά των αισθητήρων *accelerometer* και *gyroscope*, χρησιμοποιήθηκαν και οι τιμές Pitch και Roll, οι οποίες περιγράφονται στο Κεφάλαιο 3.1.3, όπως αυτές καταγράφτηκαν από την εφαρμογή. Αυτές οι τέσσερις μεταβλητές μετά από επεξεργασία, χρησιμοποιήθηκαν στην παρούσα εργασία για τη δημιουργία του μοντέλου πρόβλεψης.

Όλα τα ταξίδια που πραγματοποιήθηκαν ήταν συνολικά 559, όπως αναφέρθηκε προηγουμένως. Πριν όμως εισχωρήσουν τα ταξίδια αυτά στο μοντέλο εκπαίδευσης, έπρεπε

πρώτα να επεξεργαστούν εκτενώς. Συγκεκριμένα, το κάθε ένα ταξίδι αποτελούταν και από ένα αρχείο, επομένως έπρεπε να γίνει ένωση όλων των ταξιδιών σε ένα ώστε να μπορούν να επεξεργαστούν και να αναλυθούν πιο εύκολα από τον ερευνητή. Πρώτο βήμα λοιπόν ήταν η συγκέντρωση όλων των ταξιδιών που πραγματοποιήθηκαν με ΜΜΜ σε έναν φάκελο και όλων όσων πραγματοποιήθηκαν με αυτοκίνητο σε άλλον φάκελο. Μέσω αλγορίθμου, έγινε ένωση των αρχείων της κάθε κατηγορίας και υπολογίστηκαν αυτόματα οι συνισταμένες επιταχύνσεως ($1^η$ μεταβλητή) και γωνιακής ταχύτητας ($2^η$ μεταβλητή) από τις σχέσεις 3.1 και 3.2, ενώ στο νέο αρχείο μεταφέρθηκαν οι τιμές Pitch ($3^η$ μεταβλητή) και Roll ($4^η$ μεταβλητή) όπως αυτές καταγράφηκαν από τα κινητά. Επίσης, μεταφέρθηκαν τα ονόματα των αρχείων του κάθε ταξιδιού ($5^η$ μεταβλητή) ούτως ώστε να μπορεί να διαχωρίσει εύκολα ο ερευνητής κάποιο ταξίδι αν χρειαζόταν, ενώ έγινε και κατηγοριοποίηση των ταξιδιών όπου ο αριθμός 0 σηματοδοτούσε τη μεταφορά με ΜΜΜ ενώ ο αριθμός 1 τη μεταφορά με αυτοκίνητο ($6^η$ μεταβλητή). Επομένως το τελικό αρχείο περιείχε 6 μεταβλητές από τις οποίες οι 4 πρώτες δέχθηκαν περαιτέρω επεξεργασία η οποία αναλύεται στη συνέχεια.

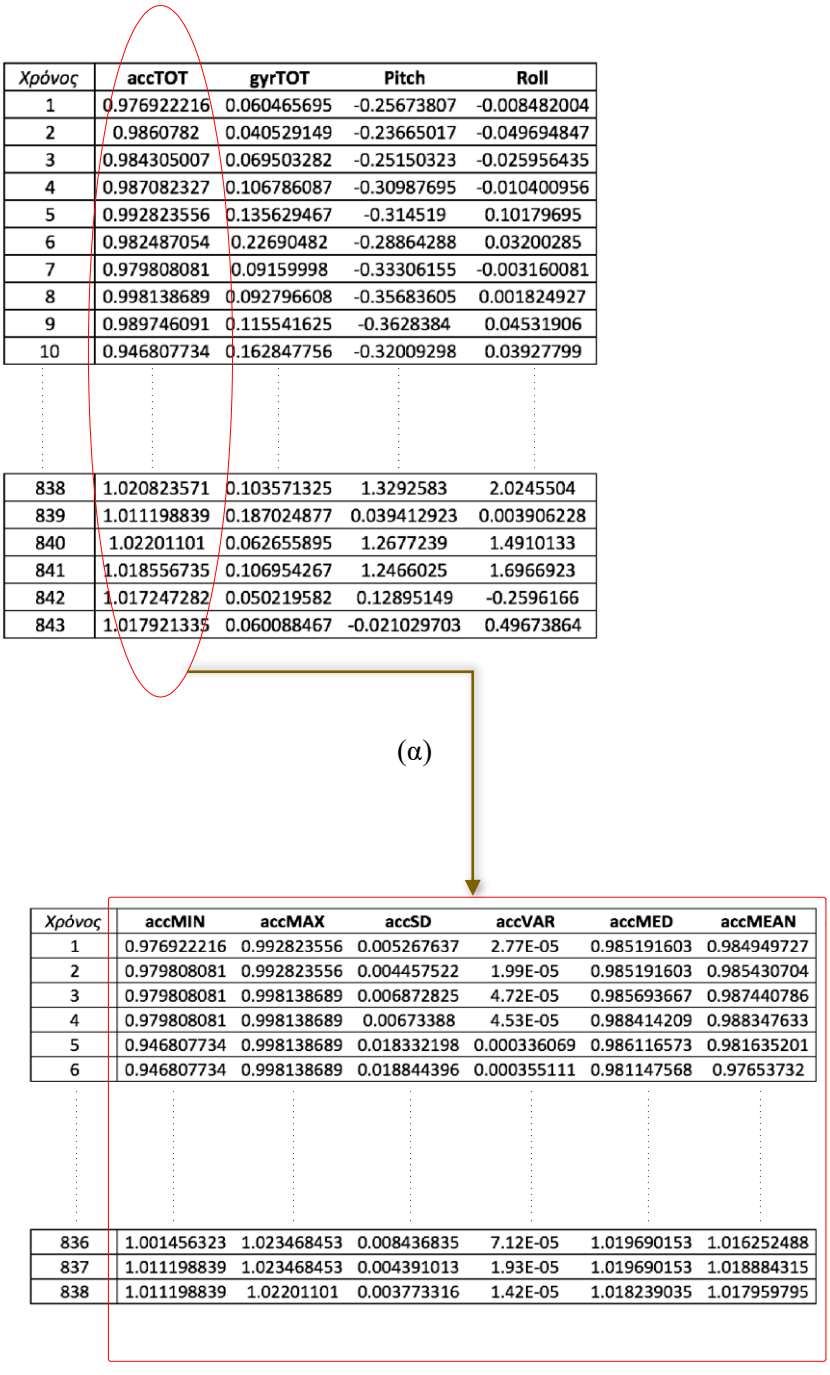
Η σημαντικότερη επεξεργασία που δέχθηκαν τα δεδομένα πριν εκπαιδευτούν, ήταν η ομαδοποίησή τους με τη δημιουργία των χρονικών παραθύρων. Συγκεκριμένα, μετά από δοκιμές που έγιναν, οι οποίες αναφέρονται αναλυτικότερα στο Κεφάλαιο 5, αποφασίστηκε να χρησιμοποιηθούν παράθυρα των 8 δευτερολέπτων, ενώ κάλυψη δύο συνεχόμενων παραθύρων γίνεται για 7 δευτερόλεπτα Αυτό σημαίνει πως το πρώτο παράθυρο ενός ταξιδιού περιείχε τις τιμές των 8 πρώτων δευτερολέπτων του, το δεύτερο παράθυρο περιείχε τιμές από το $2^ο$ μέχρι το $9^ο$ δευτερόλεπτο, το τρίτο από το $3^ο$ μέχρι το $10^ο$ δευτερόλεπτο και ότου καθεξής. Αυτή η τακτική παρουσιάζεται στο Διάγραμμα 3.10.



Διάγραμμα 3.10 Απεικόνιση της λειτουργίας του χρονικού παραθύρου

Το κάθε ένα δευτερόλεπτο περιέχει μία τιμή από κάθε μεταβλητή, δηλαδή μία τιμή του accTOT, μία του gyrTOT, μία του Pitch και μία του Roll, επομένως το χρονικό παράθυρο, το οποίο και αποτελείται από 8 δευτερόλεπτα, περιέχει 8 τιμές από κάθε μεταβλητή. Από τις τιμές αυτές, υπολογίζονται έξι νέα χαρακτηριστικά τα οποία αποτελούν τη μέγιστη και την ελάχιστη τιμή, την τυπική απόκλιση, τη διασπορά, τη διάμεσο και το μέσο όρο των τιμών του κάθε παραθύρου. Για παράδειγμα, σε κάθε παράθυρο, για τη μεταβλητή accTOT, υπολογίζονται τα accMIN, accMAX, accSD, accVAR, accMED και accMEAN, όπου accMIN αποτελεί την ελάχιστη τιμή του accTOT, accMAX τη μέγιστη τιμή του, accSD την τυπική απόκλιση, accVAR τη διασπορά, accMED τη διάμεσο και accMEAN το μέσο όρο για τις τιμές των 8 δευτερολέπτων. Οι νέες αυτές μεταβλητές φαίνονται γραφικά στο Διάγραμμα 3.11. Το ίδιο συμβαίνει και για τις άλλες τρεις μεταβλητές (gyrTOT, Pitch, Roll) ενώ αυτό εφαρμόζεται σε όλα τα ταξίδια. Συνεπώς, ο παλιός πίνακας που περιείχε τις 4 αρχικές μεταβλητές, μετατρέπεται σε έναν νέο πίνακα που περιέχει $4 * 6 = 24$ νέες μεταβλητές και με αυτές θα γίνουν οι δοκιμές εκπαίδευσης του μοντέλου. Η σημασία της κάθε μεταβλητής παρουσιάζεται στη συνέχεια στον Πίνακα 3.2.

Μετά τη δημιουργία των παραθύρων, τα τελευταία 7 δευτερόλεπτα κάθε ταξιδιού αφαιρούνταν καθώς δεν αρκούν για να γεμίσουν το παράθυρο των 8 δευτερολέπτων. Η λογική της δημιουργίας των χρονικών παραθύρων φαίνεται στο Διάγραμμα 3.11.



Χρόνος	accTOT	gyrTOT	Pitch	Roll
1	0.976922216	0.060465695	-0.25673807	-0.008482004
2	0.9860782	0.040529149	-0.23665017	-0.049694847
3	0.984305007	0.069503282	-0.25150323	-0.025956435
4	0.987082327	0.106786087	-0.30987695	-0.010400956
5	0.992823556	0.135629467	-0.314519	0.10179695
6	0.982487054	0.22690482	-0.28864288	0.03200285
7	0.979808081	0.09159998	-0.33306155	-0.003160081
8	0.998138689	0.092796608	-0.35683605	0.001824927
9	0.989746091	0.115541625	-0.3628384	0.04531906
10	0.946807734	0.162847756	-0.32009298	0.03927799
838	1.020823571	0.103571325	1.3292583	2.0245504
839	1.011198839	0.187024877	0.039412923	0.003906228
840	1.02201101	0.062655895	1.2677239	1.4910133
841	1.018556735	0.106954267	1.2466025	1.6966923
842	1.017247282	0.050219582	0.12895149	-0.2596166
843	1.017921335	0.060088467	-0.021029703	0.49673864

(α)

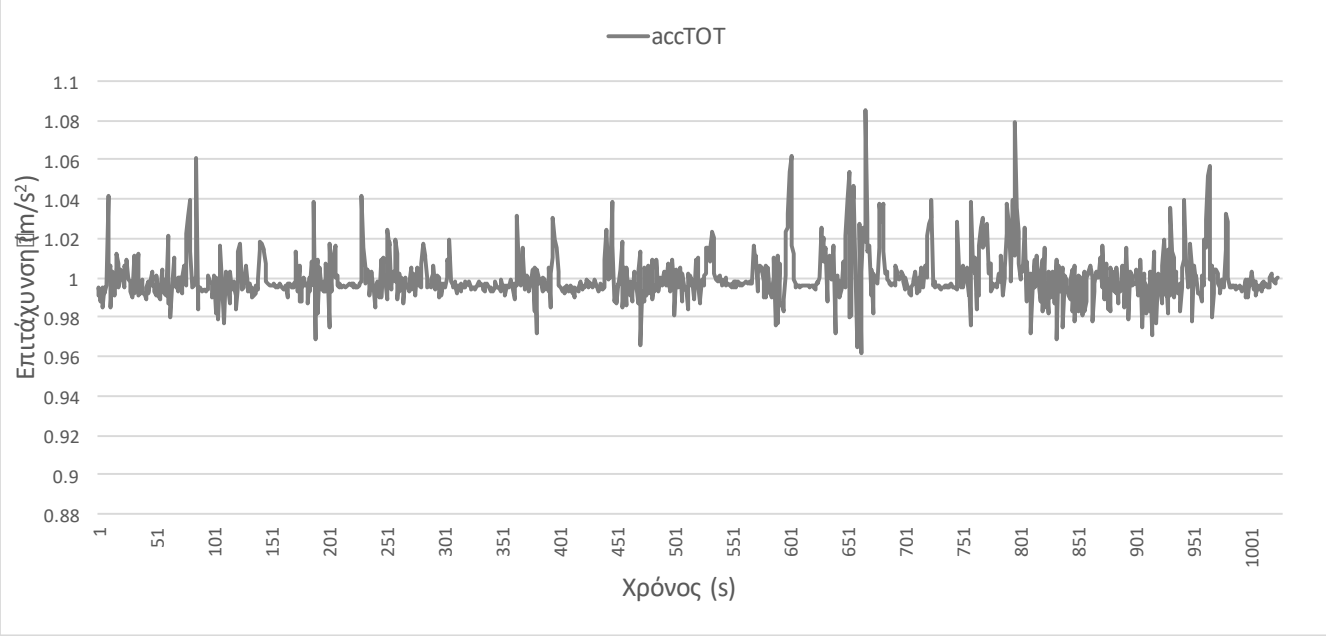
Χρόνος	accMIN	accMAX	accSD	accVAR	accMED	accMEAN
1	0.976922216	0.992823556	0.005267637	2.77E-05	0.985191603	0.984949727
2	0.979808081	0.992823556	0.004457522	1.99E-05	0.985191603	0.985430704
3	0.979808081	0.998138689	0.006872825	4.72E-05	0.985693667	0.987440786
4	0.979808081	0.998138689	0.00673388	4.53E-05	0.988414209	0.988347633
5	0.946807734	0.998138689	0.018332198	0.000336069	0.986116573	0.981635201
6	0.946807734	0.998138689	0.018844396	0.000355111	0.981147568	0.97653732
836	1.001456323	1.023468453	0.008436835	7.12E-05	1.019690153	1.016252488
837	1.011198839	1.023468453	0.004391013	1.93E-05	1.019690153	1.018884315
838	1.011198839	1.02201101	0.003773316	1.42E-05	1.018239035	1.017959795

(β)

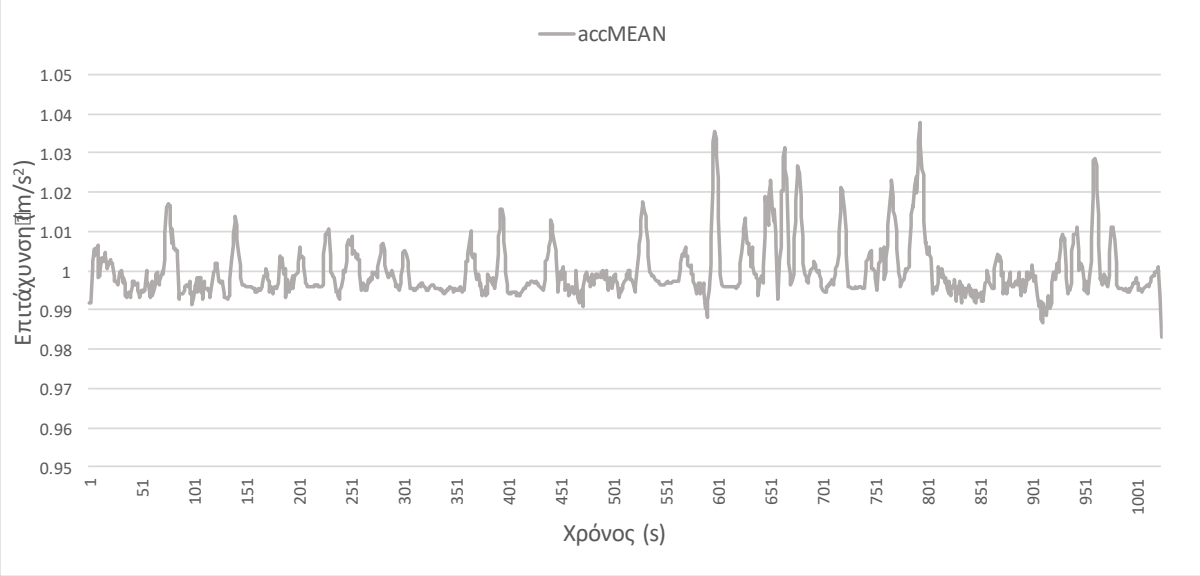
Διάγραμμα 3.11 *Στο **(α)** απεικονίζονται τα δεδομένα των αρχικών μεταβλητών accTOT, gyrTOT, Pitch και Roll ενός τυχαίου ταξιδιού ενώ στο **(β)** απεικονίζονται τα νέα δεδομένα που προκύπτουν με τη χρήση των χρονικών παραθύρων στις τιμές του επιταχυνσιομέτρου. Το ίδιο έγινε και για τις τιμές του gyrTOT, του Pitch και του Roll σε κάθε ταξίδι*

Σκοπός της χρήσης των χρονικών παραθύρων είναι η εξομάλυνση των δεδομένων, ώστε να απομακρυνθεί ο θόρυβος που δημιουργείται από τους αισθητήρες και την απότομη, καμία φορά, κίνηση του κινητού. Σε αντίθεση με άλλες έρευνες που οι χρονικές αποστάσεις δύο διαδοχικών παραθύρων ήταν μεγαλύτερες του ενός δευτερολέπτου (Eftekhari, et al., 2016), εδώ χρησιμοποιήθηκε το ένα δευτερόλεπτο. Ωστόσο, όπως παρουσιάζεται και στο Κεφάλαιο 5, έγιναν δοκιμές όπου η χρονική διαφορά δύο συνεχόμενων παραθύρων ήταν μεγαλύτερη του ενός δευτερολέπτου και τα αποτελέσματα έδειξαν μικρότερη ακρίβεια στην πρόβλεψη του μοντέλου.

Παρακάτω φαίνεται η διαφορά στη συνισταμένη επιτάχυνση χωρίς και με τη χρήση των χρονικών παραθύρων. Συγκεκριμένα, το Διάγραμμα 3.12 δείχνει τη συνισταμένη επιτάχυνση, ενώ το Διάγραμμα 3.13 δείχνει το μέσο όρο της συνισταμένης επιτάχυνσης με τη χρήση των χρονικών παραθύρων. Παρατηρείται πως το πρώτο διάγραμμα έχει μεγάλες αποκλίσεις στις τιμές του, ενώ δημιουργούνται συνέχεια απότομες γραμμές και κορυφές. Ωστόσο, το δεύτερο διάγραμμα είναι ομαλότερο, ευκολότερο να διαβαστεί από τον ερευνητή άρα και πιο αντιπροσωπευτικό για τη δημιουργία ενός ικανότερου μοντέλου πρόβλεψης.



Διάγραμμα 3.12 Γραφική απεικόνιση των τιμών της συνισταμένης επιτάχυνσης σε ένα τυχαίο ταξίδι με αυτοκίνητο



Διάγραμμα 3.13 Γραφική απεικόνιση των μέσων όρων των τιμών της συνισταμένης επιτάχυνσης που υπολογίστηκαν μέσα από κάθε ένα χρονικό παράθυρο για το ίδιο ταξίδι

Οι μεταβλητές που δημιουργήθηκαν από τη χρήση των χρονικών αυτών παραθύρων και αναφέρθηκαν προηγουμένως, παρουσιάζονται στον Πίνακα 3.2, ενώ αυτές αποτελούν και τις μεταβλητές εκπαίδευσης του μοντέλου.

Επομένως, το αρχείο με τα δεδομένα αποτελείται πλέον από 26 στήλες, όπου οι 24 πρώτες αφορούν τις μεταβλητές που υπολογίστηκαν από τα χρονικά παράθυρα, η 25^{η} μεταβλητή περιέχει τις ονομασίες, ή αλλιώς ετικέτες, των ταξιδιών και η 26^{η} μεταβλητή αποτελεί την κατηγορία του κάθε ταξιδιού δηλαδή το μέσο μεταφοράς που χρησιμοποιήθηκε. Συγκεκριμένα, η τελευταία αυτή μεταβλητή αποτελείται από επαναλαμβανόμενα 0 σε όλες τις γραμμές ενός ταξιδιού με χρήση των ΜΜΜ και επαναλαμβανόμενα 1 σε όλες τις γραμμές ενός ταξιδιού με χρήση αυτοκίνητο, ενώ αυτή είναι και η μεταβλητή πρόβλεψης. Συνοπτικά, οι 24 πρώτες μεταβλητές εκπαιδεύουν το μοντέλο, η 25^{η} υπάρχει για λόγους διευκόλυνσης του ερευνητή ενώ δε συμμετέχει ούτε στην εκπαίδευση ούτε στην πρόβλεψη του μοντέλου και η 26^{η} είναι εκείνη που το μοντέλο καλείται να προβλέψει για αυτό και ονομάζεται μεταβλητή πρόβλεψης.

Πίνακας 3.2 Ερμηνεία μεταβλητών μετά την εφαρμογή χρονικών παραθύρων

Μεταβλητές σε κάθε χρονικό παράθυρο	Μεγέθη	Ερμηνεία τιμών
accMIN	Ελάχιστο	ΤΙΜΕΣ ΕΠΙΤΑΧΥΝΣΕΩΣ
accMAX	Μέγιστο	
accSD	Τυπική απόκλιση	
accVAR	Διασπορά	
accMED	Διάμεσος	
accMEAN	Μέσος όρος	
pitMIN	Ελάχιστο	ΤΙΜΕΣ PITCH
pitMAX	Μέγιστο	
pitSD	Τυπική απόκλιση	
pitVAR	Διασπορά	
pitMED	Διάμεσος	
pitMEAN	Μέσος όρος	
rolMIN	Ελάχιστο	ΤΙΜΕΣ ROLL
rolMAX	Μέγιστο	
rolSD	Τυπική απόκλιση	
rolVAR	Διασπορά	
rolMED	Διάμεσος	
rolMEAN	Μέσος όρος	
gyrMIN	Ελάχιστο	ΤΙΜΕΣ ΓΩΝΙΑΚΗΣ ΤΑΧΥΤΗΤΑΣ
gyrMAX	Μέγιστο	
gyrSD	Τυπική απόκλιση	
gyrVAR	Διασπορά	
gyrMED	Διάμεσος	
gyrMEAN	Μέσος όρος	

Οι παραπάνω μεταβλητές χρησιμοποιήθηκαν στην εκπαίδευση και πρόβλεψη του μοντέλου και πάνω σε αυτές έγινε ο καθορισμός των παραμέτρων του μοντέλου. Τα αποτελέσματα πρόβλεψης ήταν υψηλά όταν χρησιμοποιήθηκαν όλες οι μεταβλητές μαζί, αλλά κρίθηκε αναγκαίο να γίνει ένας προσδιορισμός των πιο σημαντικών μεταβλητών. Για αυτόν τον λόγο, όπως φαίνεται στο Κεφάλαιο 5, κάποιες από τις παραπάνω μεταβλητές απομακρύνθηκαν από το μοντέλο καθώς η ακρίβεια των προβλέψεων παρέμενε σταθερή ακόμα και χωρίς τη χρησιμοποίησή τους.

Αν και το μοντέλο παρείχε υψηλή ακρίβεια στην πρόβλεψη των ταξιδιών των χρηστών, κρίθηκε αναγκαία η χρήση μια νέας μεταβλητής με στόχο την περαιτέρω βελτίωσή του η οποία αναλύεται στη συνέχεια, ενώ ο τελικός αριθμός των μεταβλητών που χρησιμοποιήθηκαν μαζί με τις αρχικές και τελικές τιμές της κάθε μεταβλητής φαίνεται στον Πίνακα 3.3.

Πίνακας 3.3 Στα (α), (β) και (γ) φαίνονται οι πρώτες και τελευταίες τιμές της κάθε μεταβλητής του μοντέλου

	accMIN	accMAX	accSD	accVAR	accMED	accMEAN	pitMIN	pitMAX	pitSD	pitVAR
1	0.96846	0.989214	0.007	4.41E-05	0.983623	0.9820924	-0.5042	-0.43375	0.026	0.00069
2	0.96846	0.995065	0.008	6.46E-05	0.984013	0.9838275	-0.5042	-0.43309	0.028	0.00079
3	0.96846	0.995065	0.009	7.85E-05	0.986202	0.9850872	-0.4961	-0.24826	0.075	0.00564
526008	0.97654	1.047025	0.023	0.00053	1.007889	1.0105126	-0.6153	-0.48681	0.055	0.003
526009	0.97654	1.047025	0.023	0.00053	1.010501	1.0111657	-0.6153	-0.32062	0.097	0.00938
526010	0.97654	1.047025	0.024	0.00056	1.009925	1.0098285	-0.6153	-0.32062	0.106	0.01116

(α)

	pitMED	pitMEAN	rolMIN	rolMAX	rolSD	rolVAR	rolMED	rolMEAN	gyrMIN	gyrMAX
1	-0.45115	-0.4609338	-0.1056	0.0062	0.041	0.0017	-0.06913	-0.059257	0.029549	0.16441
2	-0.44863	-0.4585659	-0.0954	0.04	0.049	0.0024	-0.04375	-0.041059	0.029549	0.24435
3	-0.44182	-0.4265764	-0.0954	0.1323	0.074	0.0054	-0.02485	-0.01312	0.033215	0.24435
526008	-0.5404	-0.5483275	-0.01852	0.0376	0.022	0.0005	-0.00441	0.0031197	0.025717	0.13085
526009	-0.5404	-0.5275531	-0.01852	0.0376	0.022	0.0005	-0.00441	0.003756	0.025717	0.13085
526010	-0.5404	-0.516521	-0.01439	0.0376	0.021	0.0005	-0.00441	0.00429	0.025717	0.13085

(β)

	gyrSD	gyrVAR	gyrMED	gyrMEAN	tripID	tripCHARACTERIZATION	driver
1	0.0573	0.00329	0.127729	0.1000938	mlinput--147--20160707120824--5C566D61-563F-4A03-AD7B-002BF4260615.csv	0	1
2	0.077	0.00593	0.137634	0.1157537	mlinput--147--20160707120824--5C566D61-563F-4A03-AD7B-002BF4260615.csv	0	1
3	0.0699	0.00488	0.137634	0.1234936	mlinput--147--20160707120824--5C566D61-563F-4A03-AD7B-002BF4260615.csv	0	1
526008	0.036	0.00129	0.040345	0.0532329	mlinput--94--20170502154658--606FFA57-451C-4C27-8BFC-1C38C9177588.csv	1	27
526009	0.0351	0.00123	0.04146	0.0570303	mlinput--94--20170502154658--606FFA57-451C-4C27-8BFC-1C38C9177588.csv	1	27
526010	0.0351	0.00123	0.052385	0.0616468	mlinput--94--20170502154658--606FFA57-451C-4C27-8BFC-1C38C9177588.csv	1	27

(γ)

Το όνομα του αρχείου του κάθε ταξιδιού έχει την μορφή που φαίνεται από τον Πίνακα 3.3 (γ) στην στήλη 'tripID'. Για παράδειγμα ο τίτλος 'mlinput--147--20160707120824--5C566D61-563F-4A03-AD7B-002BF4260615.csv' είναι ένα αρχείο .csv το οποίο περιέχει τα δεδομένα ενός ταξιδιού που καταγράφτηκε από την εφαρμογή Oseven Telematics. Οι πρώτοι αριθμοί του τίτλου δείχνουν ότι ο χρήστης που πραγματοποίησε το ταξίδι αυτό είναι εκείνος που έχει στην εφαρμογή τον κωδικό αριθμό 147. Έτσι, μπορούσε να γίνει διαχωρισμός των ταξιδιών ανά χρήστη. Για αυτό το λόγο, αλλά και για λόγους διευκόλυνσης της έρευνας, δόθηκε σε κάθε χρήστη νέα κωδική ονομασία από το 1 μέχρι το 27, για τους 27 χρήστες των οποίων τα δεδομένα αναλύθηκαν.

Η ονομασία αυτή των χρηστών φαίνεται και από την πρώτη στήλη του Πίνακα 3.1, ενώ στο τελικό μοντέλο πήρε την ονομασία 'driver' όπως φαίνεται στον Πίνακα 3.3. Η νέα αυτή μεταβλητή χρησιμοποιήθηκε στη συνέχεια στην εκπαίδευση και πρόβλεψη του μοντέλου. Στόχος της χρήσης της ήταν να αναγνωρίζει το μοντέλο ποιο ταξίδι ανήκει σε ποιον χρήστη, ώστε αν υπάρχει κάποιο συγκεκριμένο μοτίβο ή ιδιαιτερότητα του χρήστη να μπορεί να τα εντοπίσει και να κατηγοριοποιήσει το ταξίδι του ορθότερα. Τα μειονεκτήματα που ακολουθούν την εφαρμογή της συγκεκριμένης μεταβλητής αναλύονται στο Κεφάλαιο 6.

Ο Πίνακας 3.3 δείχνει το σύνολο των μεταβλητών που χρησιμοποιήθηκαν στην παρούσα διπλωματική εργασία, μετά την εφαρμογή των χρονικών παραθύρων των 8 δευτερολέπτων. Συγκεκριμένα δείχνει τις 3 πρώτες και τις 3 τελευταίες περιπτώσεις από τον συνολικό αριθμό των περιπτώσεων που καταγράφηκαν σε όλα τα ταξίδια της έρευνας. Η πρώτη στήλη δείχνει τον αριθμό της γραμμής στην οποία ανήκει η κάθε τιμή ενώ οι υπόλοιπες είναι οι μεταβλητές που χρησιμοποιήθηκαν στο μοντέλο. Να σημειωθεί πως ο τελικός αριθμός των γραμμών του πίνακα είναι ελάχιστα μικρότερος από τα συνολικά δευτερόλεπτα ταξιδιού που καταγράφηκαν, καθώς όπως αναφέρθηκε στο παρόν Κεφάλαιο, διαγράφονταν τα 7 τελευταία δευτερόλεπτα κάθε ταξιδιού καθώς δεν αρκούν για να συμπληρώσουν το χρονικό παράθυρο που έχει τεθεί.

Σε αυτό το σημείο, πρέπει να επισημανθεί πως η κατηγοριοποίηση των ταξιδιών σε 0 και 1 οδηγεί στο συμπέρασμα πως το μοντέλο που θα αναπτυχθεί στο Κεφάλαιο 5 θα είναι μοντέλο ταξινόμησης και όχι παλινδρόμησης και μάλιστα δυαδικό. Ο χαρακτηρισμός αυτός των ταξιδιών φαίνεται και στον Πίνακα 3.3 στη στήλη με το όνομα 'tripCHARACTERIZATION'.

4. ΜΕΘΟΔΟΛΟΓΙΚΗ ΠΡΟΣΕΓΓΙΣΗ

4.1. ΔΙΑΤΥΠΩΣΗ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ

Στην παρούσα διπλωματική εργασία γίνεται προσπάθεια αυτόματου εντοπισμού του μέσου μεταφοράς που χρησιμοποιεί ο χρήστης σε κάθε ένα ταξίδι του. Με την έννοια του ‘αυτόματου’, εννοείται πως μετά τη δημιουργία του τελικού μοντέλου, το μέσο μεταφοράς θα εντοπίζεται σε κάθε ταξίδι που πραγματοποιεί ο χρήστης χωρίς να απαιτείται καμία επιπλέον ενέργεια επιβεβαίωσης ή επεξεργασίας από εκείνον.

Αυτό μπορεί να επιτευχθεί, καθώς τα οχήματα διαφέρουν μεταξύ τους στον τρόπο με τον οποίο κινούνται στο οδικό δίκτυο. Για παράδειγμα, τα αυτοκίνητα στρίβουν απότομα, σε αντίθεση με τα ΜΜΜ που παίρνουν τις στροφές ομαλότερα. Από την άλλη τα ΜΜΜ κάνουν περισσότερες στάσεις από τα αυτοκίνητα. Οι μεταβολές αυτές στην κίνηση των οχημάτων καταγράφονται από τους αισθητήρες των κινητών τηλεφώνων διαφορετικά. Αποτέλεσμα, λοιπόν, είναι η δημιουργία μιας βάσης δεδομένων η οποία είναι ικανή, ύστερα από κατάλληλη επεξεργασία και εκπαίδευση με τη χρήση μοντέλων μηχανικής μάθησης, να προβλέψει το μέσο μεταφοράς του χρήστη.

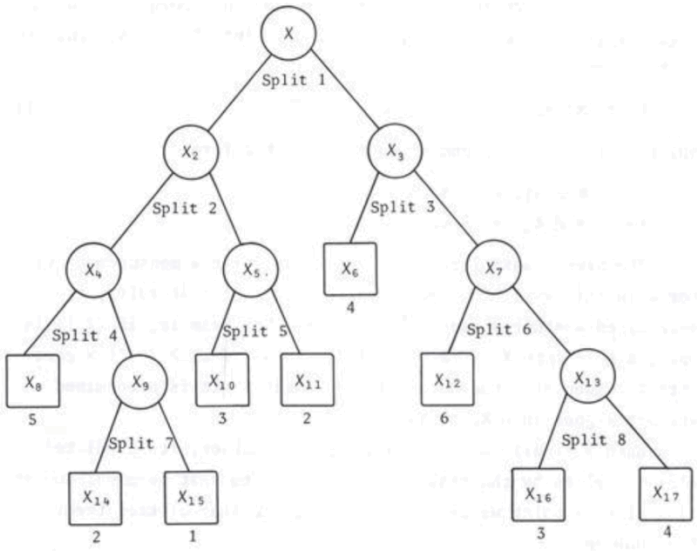
4.2. ΕΠΙΛΟΓΗ ΜΕΘΟΔΟΥ ΕΠΙΛΥΣΗΣ

Όπως έχει γίνει και αναφορά στο Κεφάλαιο 2.3, η πιο αποτελεσματική μέθοδος για τη δημιουργία μοντέλου πρόβλεψης του τρόπου μεταφοράς του χρήστη, είναι η χρήση μηχανικής μάθησης και συγκεκριμένα του μοντέλου του τυχαίου δάσους (random forest). Η νέα αυτή τεχνολογία έχει φέρει την επανάσταση στα μοντέλα πρόβλεψης καθώς τα αποτελέσματα που φέρει είναι αρκετά ακριβή και πλέον χρησιμοποιείται από πολλούς ερευνητές και εταιρείες. Η μέθοδος αυτή βασίζεται στο μοντέλο των δένδρων απόφασης, το οποίο και εξηγείται στη συνέχεια.

4.3. ΔΕΝΔΡΑ ΑΠΟΦΑΣΗΣ

Τα δένδρα απόφασης χωρίζονται σε δύο βασικές κατηγορίες. Τα δένδρα παλινδρόμησης (regression trees) και τα δένδρα ταξινόμησης (classification trees). Τα δένδρα αυτά είναι αρκετά όμοια μεταξύ τους στον τρόπο λειτουργίας τους, όμως έχουν μία βασική διαφορά. Τα δένδρα παλινδρόμησης χρησιμοποιούνται για να προβλέψουν ποσοτικές μεταβλητές ενώ τα δένδρα ταξινόμησης ποιοτικές. Τα τελευταία για παράδειγμα, χρησιμοποιούνται για να χωρίσουν ένα αρχείο δεδομένων σε τάξεις που ανήκουν στη μεταβλητή απόκρισης (response variable), ενώ συνήθως η μεταβλητή απόκρισης αυτή έχει δύο τάξεις: Ναι ή Όχι (1 ή 0 αντίστοιχα). Σε περίπτωση δυαδικού χωρισμού με 1 ή 0 (binary splits) χρησιμοποιείται μια μέθοδος με την ονομασία CART (Classification And Regression Tree) (Deshpande, 2011), ενώ όταν υπάρχουν πάνω από δύο κατηγορίες εφαρμόζεται ένας ελάχιστα διαφορετικός αλγόριθμος με το όνομα C4.5. Στη συνέχεια, παρουσιάζεται συνοπτικά ο τρόπος λειτουργίας ενός πολύ απλουστευμένου δένδρου απόφασης ταξινόμησης, χρησιμοποιώντας βιβλιογραφία από τον δημιουργό των τυχαίων δασών, Breiman (Breiman, et al., 1984). Έπειτα, αναλύονται βασικές έννοιες των δένδρων απόφασης και γίνεται σταδιακά η εισαγωγή στη μέθοδο των τυχαίων δασών με τη βοήθεια ενός αναλυτικού άρθρου του Manish (Team) αλλά και του Breiman (Breiman, et al.). Οποιαδήποτε άλλη πηγή έχει χρησιμοποιηθεί σε αυτό το κεφάλαιο θα αναφέρεται ξεχωριστά.

Οι δυαδικοί ταξινομητές που βασίζονται στο δένδρο απόφασης, είναι κατασκευασμένοι από επαναλαμβανόμενους χωρισμούς υποσυνόλων σε δύο νέα, ξεκινώντας από το σύνολο X. Αυτή η διαδικασία, για ένα υποθετικό εξατάξιο δένδρο φαίνεται στην Εικόνα 4.1. Εκεί, X_2 και X_3 έχουν διαχωριστεί ενώ $X = X_2 \cup X_3$. Ομοίως, X_4 και X_5 έχουν διαχωριστεί ενώ $X_4 \cup X_5 = X_2$ και $X_6 \cup X_7 = X_3$. Εκείνα τα υποσύνολα τα οποία δε μπορούν να χωριστούν άλλο, σε αυτή την περίπτωση τα X_6, X_8, X_{10}, X_{11}, X_{12}, X_{14}, X_{15}, X_{16} και X_{17}, ονομάζονται τερματικά υποσύνολα. Αυτά απεικονίζονται με ορθογώνιο, ενώ τα μη τερματικά υποσύνολα απεικονίζονται με κύκλο.



Εικόνα 4.1 Ένα τυχαίο δένδρο απόφασης (Πηγή: (Breiman, et al., 1984))

Τα τερματικά υποσύνολα σχηματίζουν ένα κομμάτι του συνόλου X και κάθε ένα από αυτά ορίζεται από μία τάξη. Μπορεί να είναι δύο ή περισσότερα τερματικά υποσύνολα με την ίδια τάξη. Ο ταξινομητής επομένως δημιουργείται συγκεντρώνοντας όλα τα τερματικά υποσύνολα που ανήκουν στην ίδια τάξη. Οπότε:

$A_1 = X_{15}$

$A_3 = X_{10} \cup X_{16}$

$A_5 = X_8$

$A_2 = X_{11} \cup X_{14}$

$A_4 = X_6 \cup X_{17}$

$A_6 = X_{12}$

Επίσης, οι χωρισμοί ενός συνόλου σε δύο υποσύνολα σχηματίζονται υπό συνθήκες του x = (x_1, x_2, …). Για παράδειγμα, ο χωρισμός 1 του X σε X_2 και X_3 θα μπορούσε να είναι της μορφής:

$$X_2 = \{x; x_4 \leq 7\}\ , \quad X_3 = \{x; x_4 > 7\} \qquad (4.1)$$

Ο χωρισμός 3 του X_3 σε X_6 και X_7 θα μπορούσε να είναι της μορφής

$$X_6 = \{x \in X_3; x_3 + x_5 \leq -2\} \qquad (4.2)$$

$$X_7 = \{x \in X_3; x_3 + x_5 > -2\} \qquad (4.3)$$

Ο ταξινομητής του δένδρου προβλέπει μια τάξη για τη μέτρηση του x με αυτό τον τρόπο: Από τον ορισμό του πρώτου χωρισμού, είναι καθορισμένο αν το x θα πάει στο X_2 ή στο X_3. Για παράδειγμα, αν χρησιμοποιηθεί η (4.1), το x θα πάει στο X_2 αν $x_4 \leq 7$ και στο X_3 αν $x_4 > 7$. Αν το x πάει στο X_3, τότε από τον ορισμό του χωρισμού 3, είναι καθορισμένο το αν το x θα πάει στο X_6 ή στο X_7.

Όταν το x τελικά μεταφέρεται σε κάποιο τερματικό υποσύνολο, η προβλεπόμενη τάξη του δίνεται από την τάξη με την οποία είναι ορισμένο το υποσύνολο αυτό.

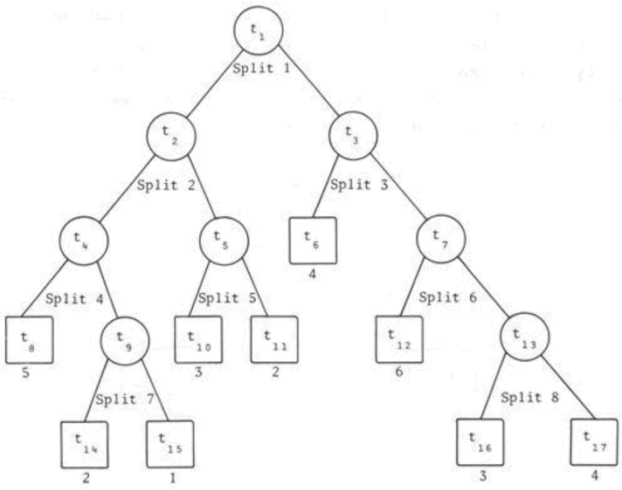
Τα παραπάνω έγιναν με σκοπό να γίνει κατανοητός ο τρόπος λειτουργίας του δένδρου απόφασης. Στη συνέχεια, θα χρησιμοποιούνται οι αντίστοιχοι όροι που εφαρμόζονται σε όλες τις βιβλιογραφίες:

- Ένα υποσύνολο του X = Ένας κόμβος t
- X = Η ρίζα του δένδρου t_1
- Τερματικά υποσύνολα = Τερματικοί κόμβοι ή αλλιώς φύλλα (leaves)
- Μη τερματικά υποσύνολα = Εσωτερικοί κόμβοι

Επομένως, η Εικόνα 4.1 μετατρέπεται στην Εικόνα 4.2. Επίσης, για λόγους διευκόλυνσης και κατανόησης, όταν από έναν κόμβο δημιουργούνται 2 ή περισσότεροι νέοι κόμβοι, ο πρώτος θα λέγεται κόμβος-πατέρας ενώ οι νέοι κόμβοι θα λέγονται κόμβοι-παιδιά.

Ολόκληρη η κατασκευή ενός δένδρου εξελίσσεται γύρω από τρεις ιδιαιτερότητες:

1. Την επιλογή του τρόπο χωρισμού των κόμβων.
2. Την απόφαση για το πότε πρέπει να κηρυχθεί ένας κόμβος τερματικός ή πρέπει να συνεχίσει να χωρίζεται.
3. Τον καθορισμό κάθε τερματικού κόμβου σε μία τάξη.



Εικόνα 4.2 Το ίδιο δένδρο απόφασης μετά από αλλαγή των όρων (Πηγή: (Breiman, et al., 1984))

Συνεπώς, τα κυρίως προβλήματα παρουσιάζονται στον τρόπο με τον οποίο θα χρησιμοποιηθούν τα δεδομένα ώστε να προσδιοριστεί ο χωρισμός των κόμβων, οι τερματικοί κόμβοι και η επιλογή των τάξεων τους. Τελικά, φαίνεται ότι το τελευταίο είναι αρκετά απλό. Η δυσκολία έγκειται στην εύρεση καλών χωρισμάτων και στον καθορισμό του σημείου που σηματοδοτεί τον τερματισμό της διάσπασης των κόμβων.

Η απόφαση με την οποία γίνεται ο διαχωρισμός ενός κόμβου επηρεάζει την ακρίβεια του δένδρου. Τα κριτήρια για την απόφαση αυτή διαφέρουν όταν υπάρχουν δένδρα ταξινομητές ή δένδρα παλινδρόμησης. Τα δένδρα αποφάσεων λοιπόν, χρησιμοποιούν πολλούς αλγορίθμους για την επιλογή του χωρισμού ενός κόμβου σε δύο ή περισσότερους κόμβους. Για τη διάσπαση αυτή, επιλέγονται κάθε φορά οι μεταβλητές που θα οδηγήσουν σε πιο ομοιογενείς νέους κόμβους. Στη συνέχεια, θα αναφερθούν τρεις αλγόριθμοι που χρησιμοποιούνται ευρέως στα δένδρα απόφασης και διαφέρουν στον τρόπο με τον οποίο παράγονται οι κόμβοι-παιδιά. Μόνο ο πρώτος αναλύεται, καθώς είναι και ο πιο διαδεδομένος, ενώ οι υπόλοιποι απλώς αναφέρονται.

Gini Index ή Gini Impurity (Μη καθαρός συντελεστής Τζίνι)

Ο συντελεστής Τζίνι μετράει την καθαρότητα του δείγματος. Για παράδειγμα, σε έναν καθαρό πληθυσμό, που έχει δηλαδή μία μόνο τάξη, ο συντελεστής Τζίνι είναι 0 αφού μια τυχαία επιλογή δείγματος από τον πληθυσμό αυτό έχει πιθανότητα 1 να ανήκει στην τάξη αυτή. Για τον υπολογισμό του χρησιμοποιείται ο παρακάτω τύπος:

$$Συντελεστής\ Τζίνι = 1 - \sum_j p_j^2 \qquad (4.4)$$

Όπου p_j η πιθανότητα να επιλεγεί ένα αντικείμενο, ενώ παίρνει τιμές από 0 μέχρι 1.

Ωστόσο, όταν χρησιμοποιείται για την επιλογή του βέλτιστου χωρισμού κόμβων, όπως συμβαίνει στη μέθοδο CART, πρέπει να βρεθεί ο σταθμισμένος μέσος του συντελεστή Τζίνι που υπολογίστηκε σε κάθε κόμβο-παιδί για όλους τους πιθανούς χωρισμούς. Για να γίνει πιο κατανοητός ο τρόπος εφαρμογής του συντελεστή αυτού, ακολουθεί το παρακάτω παράδειγμα (Lösch, 2017):

Έστω ότι υπάρχουν 3 τάξεις και 80 αντικείμενα από τα οποία τα 19 είναι στην τάξη 1, τα 21 στην τάξη 2 και τα υπόλοιπα 40 στην τάξη 3 (δηλαδή 19,21,40). Ο συντελεστής Τζίνι θα ήταν:

$$Κόστος_{πριν} = Τζίνι(19,21,40) = 1 - \left[\left(\frac{19}{80}\right)^2 + \left(\frac{21}{80}\right)^2 + \left(\frac{40}{80}\right)^2\right] = 0.6247 \qquad (4.5)$$

Για να αποφασιστεί που θα γίνει ο χωρισμός πρέπει να δοκιμαστούν όλοι οι πιθανοί χωρισμοί. Για παράδειγμα χωρίζοντας για 2.0623 έχει ως αποτέλεσμα σε έναν χωρισμό στα (16,9,0) και (3,12,40). Μετά από δοκιμές $x_1 < 2.0623$:

$$Κόστος_α = Τζίνι(16,9,0) = 0.4608 \qquad (4.6)$$

$$Κόστος_δ = Τζίνι(3,12,40) = 0.4205 \qquad (4.7)$$

Μετά βρίσκουμε τον σταθμισμένο μέσο:

$$Κόστος_{x_1<2.0623} = \frac{25}{80} * Κόστος_α + \frac{55}{80} * Κόστος_δ = 0.4331 \qquad (4.8)$$

Αυτό γίνεται για κάθε πιθανό χωρισμό. Έστω για $x_1 < 1$:

$$Κόστος_{x_1<1} = Κλάσμα_α * Τζίνι(8,4,0) + Κλάσμα_δ * Τζίνι(11,17,40)$$
$$= \frac{12}{80} * 0.4444 + \frac{68}{80} * 0.5653 = 0.5417 \qquad (4.9)$$

Τελικά, επιλέγεται ο χωρισμός με το μικρότερο κόστος. Στη συγκεκριμένη περίπτωση είναι ο χωρισμός για $x_1 < 2.0623$ με κόστος 0.4331

X^2 (Chi-Square)

Χρησιμοποιείται για να βρεθεί η στατιστική σημαντικότητα των διαφορών μεταξύ των κόμβων-παιδιών και του κόμβου-πατέρα. Υπολογίζεται από το άθροισμα των τετραγώνων των διαφορών μεταξύ της παρατηρούμενης και της προβλεπόμενης συχνότητας της μεταβλητής που αναζητούμε.

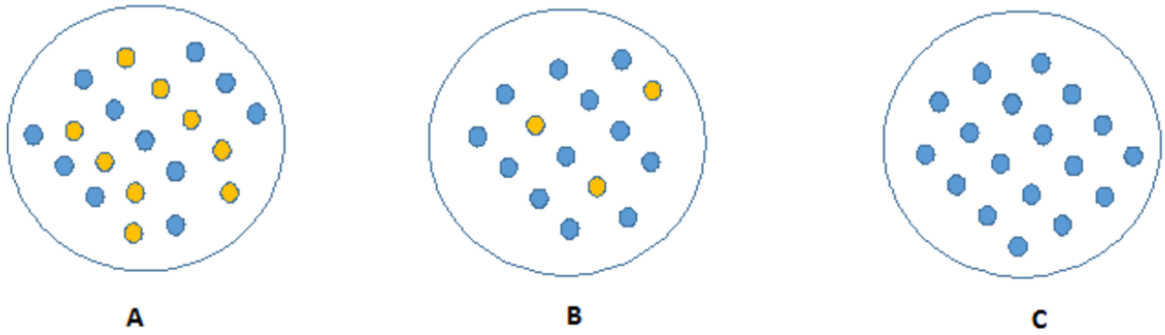
1. Μπορεί να δημιουργήσει δύο ή και περισσότερους χωρισμούς σε κάθε κόμβο.

2. Υψηλότερη τιμή του χ^2 σημαίνει και υψηλότερη στατιστική σημαντικότητα των διαφορών μεταξύ κόμβων-παιδιών και κόμβου-πατέρα.
3. Το χ^2 υπολογίζεται σε κάθε κόμβο από τον τύπο:

$$\chi^2 = \sqrt{\frac{(Πραγματικό - Προβλεπόμενο)^2}{Προβλεπόμενο}} \qquad (4.10)$$

Βάρος πληροφορίας (Information Gain)

Στη θεωρία της πληροφορίας και στη μηχανική μάθηση χρησιμοποιείται ο παραπάνω όρος για να εκτιμηθεί η ομοιογένεια ενός δείγματος και μετριέται με την εντροπία. Ένα παράδειγμα ομοιογένειας και μη φαίνεται στην Εικόνα 4.3.



Εικόνα 4.3 Παράδειγμα καθαρότητας δείγματος (Πηγή: (Team))

Αν κάποιος παρατηρήσει την Εικόνα 4.3 θα διαπιστώσει πως η τρίτη εικόνα χρειάζεται τη λιγότερη πληροφορία για να περιγραφεί. Η δεύτερη εικόνα χρειάζεται την αμέσως λιγότερη πληροφορία ενώ η πρώτη εικόνα χρειάζεται τη μέγιστη πληροφορία για να περιγραφεί. Με άλλα λόγια φαίνεται πως ο C είναι ένας καθαρός κόμβος, ο B λιγότερο καθαρός και ο A ακόμα λιγότερο. Συνεπώς, οι περισσότερο καθαροί κόμβοι χρειάζονται λιγότερη πληροφορία για να περιγραφούν. Για να προσδιορισθεί αυτού του είδους η «αταξία» στη θεωρία της πληροφορίας χρησιμοποιείται η εντροπία (entropy). Αν το δείγμα είναι ολόκληρο ομοιογενές τότε η εντροπία είναι μηδέν, ενώ αν το δείγμα είναι ισόποσα μοιρασμένο (50% - 50%) τότε η εντροπία είναι ένα. Υπολογίζεται από τον τύπο:

$$Entropy = \sum_j -p_j * \log_2 p_j \qquad (4.11)$$

Όπου p η πιθανότητα να επιλεγεί ένα αντικείμενο. Τελικά, προτιμάται ο χωρισμός που έχει τη μικρότερη εντροπία σε σχέση με τον κόμβο-πατέρα και άλλων πιθανών χωρισμών. Όσο πιο μικρή είναι, τόσο το καλύτερο.

Ωστόσο, κατά τη δημιουργία του δένδρου απόφασης, η κυριότερη πρόκληση που καλείται να αντιμετωπίσει είναι αυτό της υπερεκπαίδευσης (overfitting) του μοντέλου. Αυτό εμφανίζεται όταν ένα μοντέλο είναι υπερβολικά πολύπλοκο, ώστε να έχει πάρα πολλές παραμέτρους για σχετικά μικρό αριθμό παρατηρήσεων. Ένα μοντέλο το οποίο έχει υποστεί υπερεκπαίδευση έχει φτωχές επιδόσεις στην πρόβλεψη, αφού αντιδρά υπερβολικά σε μικρές μεταβολές στα δεδομένα εκπαίδευσης (Ove). Για την αποφυγή αυτού σε ένα μοντέλο δένδρου απόφασης:

- Τίθενται όρια στο μέγεθος του δένδρου.
- Πραγματοποιείται το λεγόμενο κλάδεμα δένδρου (tree pruning).

Όρια

Η τοποθέτηση ορίων στο μέγεθος του δένδρου μπορεί να προκύψει από τον καθορισμό ελάχιστου δείγματος που θα χρησιμοποιηθεί για τον χωρισμό ενός κόμβου ή για τη δημιουργία τερματικών κόμβων. Επίσης, μπορεί να τεθούν μέγιστα για παράδειγμα στο βάθος που θα έχει το δένδρο (max depth), στον αριθμό των τερματικών κόμβων ή ακόμα και στον αριθμό των χαρακτηριστικών που θα συμμετέχουν στο χωρισμό ενός κόμβου.

Κλάδεμα

Το κλάδεμα είναι μια τεχνική στη μηχανική μάθηση η οποία μειώνει το μέγεθος ενός δένδρου απόφασης αφαιρώντας κομμάτια δένδρου τα οποία ενισχύουν ελάχιστα την ακρίβεια του μοντέλου. Επίσης, ελαχιστοποιεί την πολυπλοκότητα του μοντέλου με αποτέλεσμα να αυξάνει την ακρίβεια πρόβλεψης ενώ ταυτόχρονα μειώνεται η πιθανότητα υπερεκπαίδευσής του αφού ένα μεγάλο δένδρο δε μπορεί να γενικεύσει νέα δείγματα. Από την άλλη, ένα μικρό δένδρο ίσως να μην καταγράψει σημαντικές πληροφορίες του δείγματος. Είναι επομένως δύσκολο για τον αλγόριθμο να εντοπίσει ακριβώς το σημείο στο οποίο πρέπει να σταματήσει να δημιουργεί νέους κόμβους, ώστε να μη μειωθεί η ακρίβεια του μοντέλου σε νέα δείγματα (Pru).

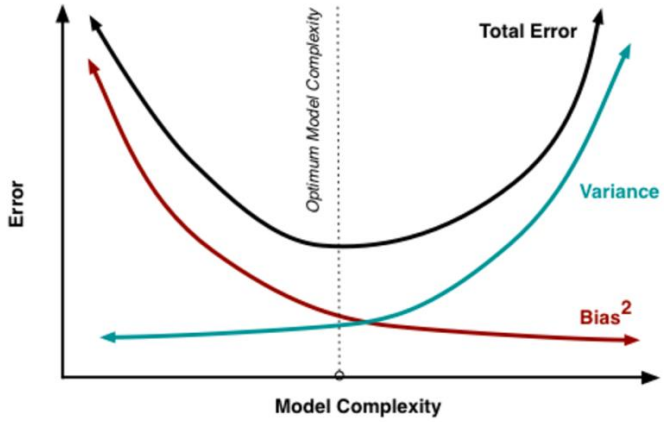
Ακόμα και με την τοποθέτηση των ορίων, η μέθοδος αυτή παραμένει μία «άπληστη» μέθοδος (greedy method) και αυτό γιατί ο αλγόριθμος νοιάζεται μόνο για τον τρέχον χωρισμό κόμβου και θα συνεχίσει να χωρίζει μέχρι να εντοπίσει κάποια συνθήκη λήξης του μοντέλου, ενώ δεν κοιτάει μελλοντικούς χωρισμούς που μπορεί να οδηγήσουν σε καλύτερα αποτελέσματα. Αν λοιπόν χρησιμοποιηθεί σωστά το «κλάδεμα» του δένδρου απόφασης, στην πραγματικότητα ο

αλγόριθμος θα μπορεί να κοιτάει μερικά βήματα μπροστά και να κάνει ορθότερα την επιλογή του. Για να γίνει αυτό, πρέπει αρχικά να δημιουργηθεί το δένδρο απόφασης με ένα μεγάλο βάθος και στη συνέχεια ξεκινώντας από κάτω προς τα πάνω, να αφαιρούνται τα φύλλα που δίνουν λανθασμένα αποτελέσματα συγκρίνοντας τα με τη ρίζα του δένδρου (κορυφή του). Αποτέλεσμα του κλαδέματος είναι η σημαντική μείωση του μεγέθους ενός δένδρου με αντίστοιχη μείωση της πιθανότητας υπερεκπαίδευσής του, χωρίς ταυτόχρονα να θυσιάζεται η ακρίβεια πρόβλεψης.

4.4. ΟΜΑΔΕΣ ΜΕΘΟΔΩΝ

Ενώ, λοιπόν, η χρήση των δένδρων απόφασης έχει διαδοθεί και τα ποσοστά ακριβείας που πετυχαίνουν τα μοντέλα αυτά είναι αυξημένα με τη βοήθεια διάφορων τεχνικών, υπάρχει ακόμα χώρος βελτίωσης. Αυτός ο χώρος έρχεται να καλυφθεί χρησιμοποιώντας ομάδες μεθόδων (ensemble methods) βασισμένων στο δένδρο απόφασης. Οι ομάδες μεθόδων περιλαμβάνουν πολλά μοντέλα πρόβλεψης μαζί και στόχο έχουν την αύξηση της ακρίβειας, της σταθερότητας και της αποτελεσματικότητας των μεθόδων που βασίζονται στο δένδρο απόφασης.

Όπως και κάθε άλλο μοντέλο, έτσι και το δένδρο απόφασης, υποφέρει από την πανούκλα της μεροληψίας (bias) και της διακύμανσης (variance). Με τον όρο μεροληψία εννοείται πόσο κατά μέσο όρο οι προβλεπόμενες τιμές διαφέρουν από τις πραγματικές. Με τον όρο διακύμανση εννοείται πόσο διαφορετικές θα είναι οι προβλέψεις στο ίδιο σημείο, αν ληφθούν διαφορετικά δείγματα από τον ίδιο πληθυσμό. Δημιουργώντας ένα μικρό δένδρο, το μοντέλο θα έχει μικρή διακύμανση και μεγάλη μεροληψία. Συνήθως, όταν αυξηθεί η πολυπλοκότητα του μοντέλου υπάρχει μείωση του σφάλματος στις προβλέψεις εξαιτίας της χαμηλής μεροληψίας του μοντέλου. Καθώς το μοντέλο γίνεται πιο πολύπλοκο, καταλήγει να υφίσταται υπερεκπαίδευση και ξεκινά να υποφέρει από υψηλή διακύμανση. Ένα μοντέλο για να είναι ακριβές, λοιπόν, πρέπει να διατηρεί μία ισορροπία ανάμεσα σε αυτούς τους δύο τύπους λαθών και αυτή η ισορροπία φαίνεται στο Διάγραμμα 4.1. Αυτό είναι γνωστό ως διαχείριση σφαλμάτων μεροληψίας – διακύμανσης (bias-variance trade-off). Η εφαρμογή των ομάδων μεθόδων που αναφέρθηκαν (ensemble methods) είναι ικανή για τη σωστή διαχείριση αυτών των σφαλμάτων.



Διάγραμμα 4.1 Απεικόνιση του βέλτιστου σημείου πολυπλοκότητας σε ένα μοντέλο (Πηγή: (Team))

Οι ομάδες μεθόδων χρησιμοποιούν κυρίως δύο διαφορετικές προσεγγίσεις του ίδιου προβλήματος, για αυτό και χωρίζονται στις μεθόδους που χρησιμοποιούν Bagging και στις μεθόδους που χρησιμοποιούν Boosting. Πριν αναλυθούν οι παραπάνω όροι, θα εξηγηθούν πρώτα δύο βασικές έννοιες. Η πρώτη είναι ο αδύναμος ταξινομητής (weak learner), ο οποίος είναι ένας ταξινομητής που οποιαδήποτε κατανομή και αν έχουν τα δεδομένα εκπαίδευσης, θα ανταποκρίνεται πάντα καλύτερα από την τύχη στην πρόβλεψη δεδομένων. Επομένως, θα υπάρχει ένα ποσοστό σφάλματος πάντα μικρότερο του 50% ενώ δεν υπάρχει πολύ μεγάλη ακρίβεια λόγω της αδυναμίας του ταξινομητή στην ανάλυση και μάθηση δεδομένων. Η άλλη έννοια αφορά στο διαχωρισμό των δειγμάτων σε δύο κατηγορίες: σε αυτά 1)με αντικατάσταση, στα οποία η κάθε περίπτωση δεδομένων μπορεί να επιλεγεί πάνω από μία φορά και σε αυτά 2)χωρίς αντικατάσταση όπου η κάθε περίπτωση δεδομένων μπορεί να επιλεγεί μόνο μία φορά.

Οι μέθοδοι που χρησιμοποιούν Boosting ακολουθούν τα εξής βήματα:

I. Χρησιμοποιείται ένα τυχαίο υποσύνολο α1 χωρίς αντικατάσταση του δείγματος εκπαίδευσης Α για την εκπαίδευση ενός αδύναμου ταξινομητή Β1
II. Χρησιμοποιείται ένα δεύτερο τυχαίο υποσύνολο α2 χωρίς αντικατάσταση του δείγματος (Seemakurthi, 2016)εκπαίδευσης Α και προστίθεται το 50% των δειγμάτων που ταξινομήθηκαν λάθος από τον προηγούμενο ταξινομητή για να εκπαιδεύσει έναν αδύναμο ταξινομητή Β2.
III. Εντοπίζει τα δείγματα α3 του δείγματος εκπαίδευσης Α όπου οι αδύναμοι ταξινομητές Β1 και Β2 διαφωνούν ώστε να εκπαιδεύσει με αυτά έναν τρίτο αδύναμο ταξινομητή Β3.
IV. Τέλος, συνδυάζει όλους τους αδύναμους ταξινομητές και επιλέγει πλειοψηφικά τον καλύτερο.

Η λέξη **Bagging** προέρχεται από τις λέξεις **B**ootstrap **Agg**regat**ing** όπου aggregating σημαίνει συγκέντρωση ενώ Bootstrap εννοείται η επιλογή ενός τυχαίου δείγματος με αντικατάσταση. Επομένως, τα βήματα που ακολουθούνται εδώ είναι:

I. Δημιουργούνται ν διαφορετικά bootstrap δείγματα εκπαίδευσης (τυχαία υποσύνολα με αντικατάσταση).
II. Ο αλγόριθμος εκπαιδεύεται σε κάθε ένα από αυτά τα δείγματα ξεχωριστά.
III. Τέλος, λαμβάνεται ο μέσος όρος των προβλέψεων των δειγμάτων.

Μία από τις βασικές διαφορές των Bagging και Boosting είναι ο τρόπος με τον οποίο χρησιμοποιούνται τα δείγματα εκπαίδευσης. Στο Bagging γίνονται με αντικατάσταση ενώ στο Boosting χωρίς. Στη θεωρία, το Bagging έχει στόχο να μειώσει τη διακύμανση ενώ το Boosting τη μεροληψία (Seemakurthi, 2016). Σε αυτό το σημείο λοιπόν, πρέπει να αναφερθεί πως υπάρχουν διάφορες εφαρμογές μοντέλων που χρησιμοποιούν κάποια από τις δύο αυτές μεθόδους. Στη συνέχεια θα αναλυθεί η μέθοδος του τυχαίου δάσους που εφαρμόστηκε στην παρούσα εργασία και ο πυρήνας της βασίζεται στη μέθοδο Bagging.

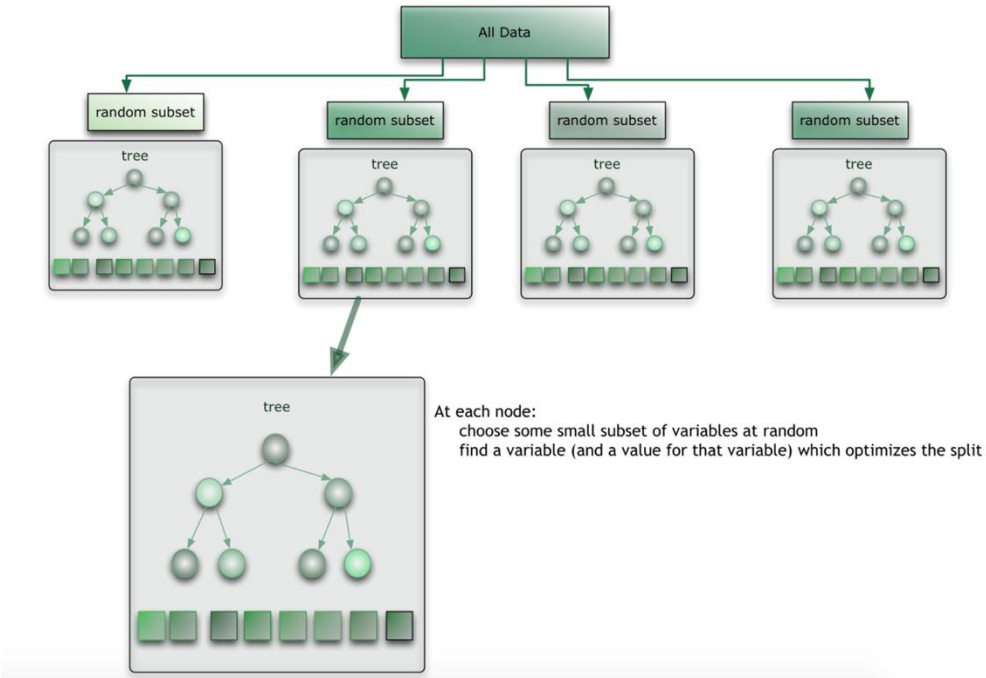
4.5. ΤΥΧΑΙΟ ΔΑΣΟΣ

Το τυχαίο δάσος (random forest) είναι μια ευέλικτη μέθοδος μηχανικής μάθησης που μπορεί να εφαρμοστεί τόσο σε προβλήματα παλινδρόμησης όσο και ταξινόμησης. Γενικά, αναλαμβάνει τη μείωση των διαστάσεων του μοντέλου, αντιμετωπίζει τις ελλιπείς τιμές και δίνει μεγάλη ώθηση στην ακρίβεια του μοντέλου υπεραναλύοντας τα δεδομένα. Είναι ένας τύπος από τις ομάδες μεθόδων που αναφέρθηκαν στο Κεφάλαιο 4.4, όπου αδύναμοι ταξινομητές (στην προκειμένη περίπτωση είναι τα δένδρα) συνδυάζονται για το σχηματισμό ενός δυνατού μοντέλου (δάσος).

Στα τυχαία δάση δημιουργούνται πολλά δένδρα, σε αντίθεση με το μοντέλο CART για παράδειγμα, που δημιουργεί ένα μόνο δένδρο και του οποίου η ανάλυση έγινε στο Κεφάλαιο 4.3. Για να ταξινομηθεί ένα νέο σύνολο που βασίζεται σε κάποια χαρακτηριστικά, κάθε δένδρο καταλήγει σε μία τάξη την οποία αυτό έχει «ψηφίσει». Το δάσος στην περίπτωση ενός προβλήματος ταξινόμησης, μετρώντας τις τάξεις όλων των δένδρων του δάσους, επιλέγει την τάξη που συναντάται περισσότερο (δηλαδή εκείνο με τις περισσότερες ψήφους). Στην περίπτωση ενός παλινδρομικού προβλήματος, επιλέγει το μέσο όρο των τελικών αποτελεσμάτων του κάθε δένδρου που δημιουργήθηκε στο δάσος.

Παράδειγμα της εφαρμογής του φαίνεται στην Εικόνα 4.4, ενώ πρακτικά λειτουργεί με τον εξής τρόπο:

1) Αν ο συνολικός αριθμός των σειρών σε ένα σύνολο δεδομένων εκπαίδευσης είναι N, τότε επιλέγονται N σειρές στην τύχη με αντικατάσταση από τα αρχικά δεδομένα. Αυτό το δείγμα θα αποτελεί τα δεδομένα εκπαίδευσης ώστε να αναπτυχθεί το δένδρο.
2) Αν υπάρχουν M μεταβλητές, ένας αριθμός μ<<M καθορίζεται έτσι ώστε σε κάθε κόμβο να επιλέγονται τυχαία μ μεταβλητές μέσα από τις M και με αυτές να πραγματοποιείται ο καλύτερος δυνατός χωρισμός του κόμβου-πατέρα σε κόμβους-παιδιά.
3) Κάθε δένδρο μεγαλώνει όσο του επιτρέπουν τα δεδομένα χωρίς άλλους περιορισμούς, ενώ δε γίνεται κλάδεμα.
4) Η ολική πρόβλεψη γίνεται αφού πρώτα συγκεντρωθεί η τελική πρόβλεψη του πρώτου μέχρι και του ν-οστού δένδρου που δημιουργήθηκαν στο δάσος. Συγκεκριμένα, για πρόβλημα ταξινόμησης χρησιμοποιείται η πλειοψηφία των προβλέψεων των δένδρων του δάσους ενώ για παλινδρόμησης χρησιμοποιείται ο μέσος όρος των προβλέψεων.



Εικόνα 4.4 Απεικόνιση της λειτουργίας του τυχαίου δάσους (Πηγή: (Benyamin, 2012))

Τα χαρακτηριστικά του τυχαίου δάσους είναι τα ακόλουθα:

- Λειτουργεί αποτελεσματικά σε μεγάλες βάσεις δεδομένων.
- Μπορεί να χειριστεί χιλιάδες μεταβλητές εισόδου χωρίς να διαγράφονται μεταβλητές.
- Δίνει εκτιμήσεις για το ποιες μεταβλητές είναι σημαντικές ώστε να γίνει αποτελεσματικότερα η ταξινόμηση.

- Δημιουργεί μια εσωτερική αμερόληπτη (unbiased) εκτίμηση του σφάλματος γενίκευσης καθώς το δάσος επεκτείνεται.
- Έχει μια αποτελεσματική μέθοδο εκτίμησης των δεδομένων που λείπουν και διατηρεί την ακρίβειά του ακόμα και όταν απουσιάζουν μεγάλες ποσότητες από αυτά.
- Το δάσος που δημιουργείται μπορεί να αποθηκευτεί και να χρησιμοποιηθεί και σε άλλη βάση δεδομένων που έχει ίδιες μεταβλητές.
- Υπολογίζονται διάφοροι όροι οι οποίοι δίνουν πληροφορίες για τη σχέση μεταξύ των μεταβλητών και της ταξινόμησης.
- Υπολογίζει την εγγύτητα (proximities) μεταξύ ζευγαριών περιπτώσεων η οποία μπορεί να χρησιμοποιηθεί για να βρεθούν τα ακρότατα (outliers) και δίνει μια διαφορετική εικόνα στα δεδομένα. (Οι δύο αυτοί όροι αναφέρονται λεπτομερέστερα στην επόμενη σελίδα.)
- Οι ικανότητές του, μπορούν να επεκταθούν και σε δεδομένα που δεν έχουν κατηγοριοποιηθεί (unlabelled data) οδηγώντας στη χρήση μη επιβλεπόμενης μάθησης (unsupervised learning).

Σε αυτό το σημείο αξίζει να σημειωθεί πως το τυχαίο δάσος δεν μπορεί να υποστεί υπερεκπαίδευση. Ο ερευνητής επιλέγει τον μέγιστο αριθμό δένδρων με τον οποίο θα τρέξει το μοντέλο, ενώ ταυτόχρονα η ανάλυση θα γίνει εξαιρετικά γρήγορα ακόμα και για μεγάλο όγκο δεδομένων. Η μόνη απαίτηση σε αυτή την περίπτωση είναι η επαρκής μνήμη αποθήκευσης, ώστε να μπορέσει η βάση δεδομένων να αποθηκευτεί στον σκληρό δίσκο. Αν υπολογιστεί και η εγγύτητα των δεδομένων, τότε η απαίτηση αποθηκευτικού χώρου αυξάνεται ραγδαία.

Τα τυχαία δάση για να αποκτήσουν όλες αυτές τις ιδιαιτερότητές τους, χρησιμοποιούν δύο συγκεκριμένες μεθόδους.

- Όταν τα δεδομένα εκπαίδευσης συγκεντρωθούν μετά από δειγματοληψία με αντικατάσταση για τη δημιουργία ενός δένδρου, περίπου το 1/3 του δείγματος μένει απ’ έξω. Αυτό το δείγμα (το οποίο φέρει την ονομασία «εκτός σακούλας» ή στα αγγλικά «out-of-bag data») χρησιμοποιείται για να εφαρμοστεί μια αμερόληπτη εκτίμηση του σφάλματος ταξινόμησης καθώς τα δένδρα αναπτύσσονται στο δάσος. Επίσης, με το δείγμα αυτό γίνονται και εκτιμήσεις για τη σημαντικότητα της κάθε μεταβλητής.
- Μετά τη δημιουργία κάθε δένδρου, όλα τα δεδομένα εφαρμόζονται στο δένδρο αυτό και η εγγύτητα υπολογίζεται για κάθε ζεύγος περιπτώσεων. Αν δύο περιπτώσεις καταλαμβάνουν τον ίδιο τερματικό κόμβο τότε η εγγύτητά τους αυξάνεται κατά ένα. Στο τέλος, αφού αναπτυχθούν όλα τα δένδρα, οι εγγύτητες κανονικοποιούνται,

διαιρώντας τες με τον αριθμό των δένδρων. Η εγγύτητα χρησιμοποιείται στην αντικατάσταση των δεδομένων που λείπουν και στην εύρεση των ακραίων τιμών (outliers).

Στη συνέχεια του παρών Κεφαλαίου, αναλύονται τα βασικότερα χαρακτηριστικά του τυχαίου δάσους.

Υπολογισμός των σφαλμάτων «εκτός σακούλας»

Στη μέθοδο του τυχαίου δάσους, δεν απαιτείται να γίνει χρήση ξεχωριστού δοκιμαστικού αρχείου δεδομένων (validation set) με σκοπό την αμερόληπτη εκτίμηση του δοκιμαστικού σφάλματος, όμως εφαρμόζεται κυρίως για προληπτικούς λόγους. Ωστόσο, υπάρχουν συγκεκριμένοι όροι οι οποίοι υπολογίζονται εσωτερικά, κατά τη διάρκεια ανάπτυξης του δάσους. Παρακάτω αναλύονται οι σημαντικότεροι από αυτούς τους όρους που χρησιμοποιούνται για την εκτίμηση σφαλμάτων στα μοντέλα παλινδρόμησης. Για τα μοντέλα ταξινόμησης χρησιμοποιούνται διαφορετικές μεταβλητές ο υπολογισμός των οποίων γίνεται στο Κεφάλαιο 4.7, καθώς είναι και αυτές που θα χρησιμοποιηθούν για την αξιολόγηση του μοντέλου που αναπτύσσεται στην παρούσα εργασία.

Ο πρώτος, ονομάζεται μέσος όρος απολύτου σφάλματος (Mean Absolute Error, MAE). Αυτός, μετράει το μέσο όρο του εύρους των σφαλμάτων, όπως καταγράφεται από τις προβλέψεις που εξάγονται από τα δεδομένα εκπαίδευσης που ανήκουν στην κατηγορία «εκτός σακούλας». Είναι ο μέσος όρος απολύτων διαφορών μεταξύ της πρόβλεψης και της πραγματικής τιμής και υπολογίζεται από τον παρακάτω τύπο:

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j| \qquad (4.12)$$

Ο δεύτερος, είναι η τετραγωνική ρίζα του μέσου των τετραγώνων του σφάλματος (Root Mean Squared Error, RMSE), δηλαδή του μέσου όρου των τετραγωνικών διαφορών μεταξύ της πρόβλεψης και της πραγματικής τιμής που επίσης μετράει το μέσο εύρος του σφάλματος.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2} \qquad (4.13)$$

Οι δύο αυτές τιμές εκφράζουν το μέσο σφάλμα πρόβλεψης του μοντέλου και παίρνουν τιμές από το 0 έως το ∞. Χαμηλότερες τιμές σημαίνει καλύτερο μοντέλο. Η κυριότερη διαφορά τους

είναι πως στον όρο RMSE τα σφάλματα τετραγωνίζονται πριν βρεθεί ο μέσος όρος τους επομένως δίνει μεγαλύτερη βαρύτητα στα μεγάλα σφάλματα. Οπότε, ο όρος αυτός είναι πιο χρήσιμος όταν τα μεγάλα σφάλματα είναι μη επιθυμητά. Για αυτό το λόγο η τιμή του RMSE θα είναι πάντα ίση ή μεγαλύτερη από την τιμή του MAE. Η περίπτωση της ισότητας ισχύει όταν όλα τα λάθη έχουν το ίδιο εύρος.

Σημαντικότητα μεταβλητής

Σε κάθε δένδρο του δάσους, μετριέται ο αριθμός των ψήφων, έστω P, για τη σωστή τάξη για τις περιπτώσεις «εκτός σακούλας». Στη συνέχεια, διαχωρίζονται τυχαία οι τιμές της μεταβλητής Μ από τις περιπτώσεις «εκτός σακούλας» και τοποθετούνται στο δένδρο. Έπειτα, αφαιρείται ο αριθμός των ψήφων για τη σωστή τάξη της μεταβλητής Μ μετά την τοποθέτηση των τιμών στο δένδρο από τον αριθμό P που υπολογίστηκε πριν. Ο μέσος όρος του αριθμού αυτού, λαμβάνοντας υπόψη όλα τα δένδρα του δάσους, είναι ο βαθμός σημαντικότητας της μεταβλητής Μ. Αν ο αριθμός των μεταβλητών είναι τεράστιος, τότε το δάσος μπορεί να αναπτυχθεί μια φορά με όλες τις μεταβλητές και στη συνέχεια να αναπτυχθεί ξανά μόνο με τις πιο σημαντικές μεταβλητές που καθορίστηκαν με τον τρόπο που αναφέρθηκε προηγουμένως. Στην παρούσα έρευνα, το ποσοστό σημαντικότητας της κάθε μεταβλητής παρουσιαζόταν σε έναν πίνακα που δημιουργούταν μετά την ανάπτυξη του δάσους. Αυτός έκρινε, κατά κύριο λόγο, τις μεταβλητές που θα συνεχίσουν να εκπαιδεύουν το μοντέλο και εκείνες που θα απομακρύνονταν από αυτό.

Αλληλεπίδραση

Ο ορισμός της αλληλεπίδρασης που χρησιμοποιείται είναι ότι οι μεταβλητές Μ και Κ αλληλεπιδρούν αν ο χωρισμός μιας μεταβλητής Μ σε ένα δένδρο πραγματοποιεί τον χωρισμό της Κ πιο συχνά ή πιο σπάνια. Αυτό βασίζεται στους δείκτες Τζίνι και συμβαίνει σε κάθε δένδρο στο δάσος.

Αντικατάσταση των τιμών που λείπουν στα δεδομένα εκπαίδευσης

Τα τυχαία δάση έχουν δύο τρόπους με τους οποίους αντικαθιστούν τις τιμές που λείπουν. Εδώ θα εξηγηθεί ο πρώτος, ο οποίος είναι και πιο γρήγορος στον υπολογισμό. Αν η Μ μεταβλητή δεν είναι κατηγορική, η μέθοδος υπολογίζει το διάμεσο όλων των τιμών αυτής της μεταβλητής στην τάξη Φ, μετά χρησιμοποιεί αυτή την τιμή για να αντικαταστήσει όλες τις τιμές που λείπουν από τη μεταβλητή Μ στην τάξη Φ. Αν η Μ μεταβλητή είναι κατηγορική, η αντικατάσταση θα γίνει από την πιο συχνή τιμή που συναντάται στην τάξη Φ.

Αντικατάσταση των τιμών που λείπουν στα δεδομένα δοκιμασίας

Όσον αφορά τα δεδομένα δοκιμασίας, υπάρχουν δύο διαφορετικές μέθοδοι αντικατάστασης που εξαρτώνται από το αν έχουν κατηγοριοποιηθεί τα δεδομένα αυτά σε τάξεις ή όχι. Αν έχουν, τότε οι αντικαταστάσεις που χρησιμοποιήθηκαν στα δεδομένα εκπαίδευσης χρησιμοποιούνται και εδώ. Αν δεν έχουν, τότε κάθε περίπτωση στα δεδομένα δοκιμασίας επαναλαμβάνεται ν φόρες (όπου ν ο αριθμός των τάξεων). Η πρώτη επανάληψη θεωρείται ότι βρίσκεται στην τάξη 1 και η τάξη αυτή χρησιμοποιείται για να συμπληρωθεί η τιμή που λείπει. Στη δεύτερη επανάληψη θεωρείται ότι βρίσκεται στην τάξη 2 και η τάξη αυτή χρησιμοποιείται για να συμπληρωθεί η τιμή που λείπει. Αυτό το αυξημένο αρχείο δεδομένων τρέχει στο δένδρο. Σε κάθε ομάδα επαναλήψεων, αυτή που λαμβάνει τις περισσότερους ψήφους καθορίζει και την τάξη της αρχικής περίπτωσης.

Περιπτώσεις δεδομένων με εσφαλμένη κατηγορία

Τα δεδομένα εκπαίδευσης συνήθως σχηματίζονται χρησιμοποιώντας την ανθρώπινη κρίση για να χωριστούν σε κατηγορίες. Σε κάποιους τομείς αυτό οδηγεί σε υψηλή συχνότητα εσφαλμένης κατηγοριοποίησης. Πολλές περιπτώσεις τέτοιων εσφαλμένων κατηγοριοποιήσεων μπορούν να ανιχνευθούν με τη χρήση των ακροτάτων (outlier measure).

Ακρότατα (outliers)

Ακρότατα γενικά θεωρούνται οι περιπτώσεις οι οποίες αφαιρούνται από το βασικό αρχείο δεδομένων. Αυτό συμβαίνει διότι τα ακρότατα είναι περιπτώσεις των οποίων η εγγύτητα σε σχέση με όλες τις υπόλοιπες περιπτώσεις στα δεδομένα είναι γενικά μικρή. Μια χρήσιμη αναθεώρηση για αυτά τα ακρότατα που δημιουργούνται είναι να καθοριστούν σε σχέση με την τάξη που ανήκουν. Συνεπώς, ένα ακρότατο στην τάξη Κ είναι μια περίπτωση της τάξης αυτής, της οποίας η εγγύτητα με όλες τις υπόλοιπες περιπτώσεις στην τάξη Κ είναι μικρή.

Ισορροπίες στο σφάλμα πρόβλεψης

Σε κάποια σύνολα δεδομένων, το σφάλμα πρόβλεψης μεταξύ των τάξεων είναι αρκετά μη ισορροπημένο. Μερικές τάξεις έχουν χαμηλό σφάλμα πρόβλεψης ενώ άλλες μεγάλο. Αυτό συμβαίνει συνήθως όταν ο αριθμός των περιπτώσεων που ανήκουν στη μία τάξη είναι πολύ μεγαλύτερος από τον αριθμό που ανήκουν στην άλλη. Σε αυτή την περίπτωση, τα τυχαία δάση προσπαθούν να ελαχιστοποιήσουν το ποσοστό σφάλματος, κρατώντας χαμηλό το ποσοστό αυτό στις μεγάλες τάξεις ενώ επιτρέπουν στις μικρότερες να έχουν ένα μεγαλύτερο ποσοστό σφάλματος. Η αλλαγή αυτή του ποσοστού σφάλματος μεταξύ των τάξεων γίνεται θέτοντας διαφορετικά βάρη σε κάθε τάξη. Όσο ψηλότερο βάρος δοθεί σε μία τάξη, το ποσοστό

σφάλματος μειώνεται. Ωστόσο, πρέπει να σημειωθεί πως στην προσπάθεια ισορροπίας του ποσοστού σφάλματος μεταξύ των τάξεων, το συνολικό ποσοστό σφάλματος αυξάνεται.

4.6. ΕΠΙΛΟΓΗ ΛΟΓΙΣΜΙΚΟΥ

Για την εκπόνηση της παρούσας διπλωματικής εργασίας επιλέχθηκε να χρησιμοποιηθεί η γλώσσα προγραμματισμού R και το φιλικό στον χρήστη πρόγραμμα R Studio. Η R είναι μια ανοιχτού κώδικα γλώσσα προγραμματισμού για στατιστικές αναλύσεις και ανάπτυξη προτύπων τεχνητής νοημοσύνης. Χρησιμοποιείται ευρέως από στατιστικολόγους και αναλυτές δεδομένων για την ανάπτυξη στατιστικών λογισμικών. Το βασικό πλεονέκτημά της είναι τα αναρίθμητα πακέτα που έχουν αναπτυχθεί από τους χρήστες της, με σκοπό τη διευκόλυνσή τους και την αυτοματοποίηση συγκεκριμένων διαδικασιών. Στην παρούσα εργασία χρησιμοποιήθηκε το πακέτο h2o το οποίο αναπτύχθηκε το 2011 και ανανεώνεται συνεχώς μέχρι και σήμερα, ενώ περιέχει πολλές επιλογές για τη δημιουργία μοντέλων μηχανικής μάθησης και επιταχύνει την ανάλυσή τους λόγω των αλγορίθμων που χρησιμοποιεί.

Η ανάλυση του τυχαίου δάσους μπορεί να γίνει πολύ δύσκολη αν αναλογιστεί κανείς τους περιορισμούς του χώρου του ηλεκτρονικού υπολογιστή. Η R αποθηκεύει τα προσωρινά της αρχεία στη μνήμη RAM του υπολογιστή ώστε να είναι εύκολα προσβάσιμα και να εκτελούνται οι εντολές πολύ πιο γρήγορα. Ωστόσο, αν τα δεδομένα απαιτούν μεγάλο χώρο και δεν υπάρχει διαθέσιμη μνήμη RAM, αρχίζει να γεμίζει ο αποθηκευτικός χώρος ενώ αν γεμίσει και αυτός τότε η διαδικασία σταματά. Στην παρούσα εργασία δεν προέκυψε πρόβλημα με τη μνήμη RAM, καθώς χρησιμοποιήθηκε νέος φορητός υπολογιστής με μνήμη RAM 8GB DDR3 στα 1867MHz.

4.7. ΜΕΘΟΔΟΣ ΑΞΙΟΛΟΓΗΣΗΣ

Για την αξιολόγηση του τυχαίου δάσους έγινε εφαρμογή του μοντέλου στα δεδομένα δοκιμής. Η συνολική ακρίβεια των προβλέψεων μετριέται και αποτελεί το βασικό μέτρο εκτίμησης της επιτυχίας του μοντέλου. Ωστόσο, χρησιμοποιούνται και κάποια άλλα στατιστικά μέτρα τα οποία οδηγούν στην επιτυχέστερη εξήγηση των αποτελεσμάτων του μοντέλου καθώς και στον καλύτερο προσδιορισμό των λαθών του. Επομένως, έγινε χρήση του παρακάτω πίνακα:

Πίνακας 4.1 Πραγματικές και προβλεπόμενες τιμές του μοντέλου

		ΠΡΑΓΜΑΤΙΚΟ		
		0	1	
ΠΡΟΒΛΕΨΗ	0	**Α**	**Β**	Σ_1
	1	**Γ**	**Δ**	Σ_2
		Σ_3	Σ_4	Σ

Ενώ τα στατιστικά μέτρα που χρησιμοποιήθηκαν είναι τα:

$$Sensitivity: \frac{A}{\Sigma_3} \qquad (4.14)$$

$$Specificity: \frac{\Delta}{\Sigma_4} \qquad (4.15)$$

$$Positive\ Predicted\ Value\ (PPV): \frac{A}{\Sigma_1} \qquad (4.16)$$

$$Negative\ Predicted\ Value\ (NPV): \frac{\Delta}{\Sigma_2} \qquad (4.17)$$

$$False\ Positive\ Rate\ (FPR): 1 - Specificity = \frac{B}{\Sigma_4} \qquad (4.18)$$

$$False\ Discovery\ Rate\ (FDR): 1 - PPV = \frac{B}{\Sigma_1} \qquad (4.19)$$

$$False\ Negative\ Rate\ (FNR): \frac{\Gamma}{\Sigma_3} \qquad (4.20)$$

$$Accuracy\ (ACC): \frac{A + \Delta}{\Sigma} \qquad (4.21)$$

$$F1: \frac{2 * A}{2 * A + B + \Gamma} \qquad (4.22)$$

$$MCC: \frac{A * \Delta - B * \Gamma}{\sqrt{\Sigma 1 * \Sigma 2 * \Sigma 3 * \Sigma 4}} \qquad (4.23)$$

Από τους παραπάνω τύπους φαίνεται πως στόχος είναι η μεγιστοποίηση των τιμών, εκτός των FPR, FDR, FNR τα οποία πρέπει να ελαχιστοποιηθούν. Ενδεικτικά, αναζητείται η μεγιστοποίηση της διαγωνίου Α-Δ και η ελαχιστοποίηση της διαγωνίου Β-Γ. Νοητικά, πρέπει το μοντέλο να προβλέψει τα πραγματικά μηδενικά σαν μηδέν και τους πραγματικούς άσους σαν άσους.

Εκτός από αυτά, σημαντικό ρόλο για την αξιολόγηση των μοντέλων που δημιουργήθηκαν κατά τη διάρκεια των δοκιμών, έπαιξε ο πίνακας που δείχνει τη σημαντικότητα της κάθε μεταβλητής, του οποίου η σημασία αναλύθηκε στο Κεφάλαιο 4.5. Αυτός ο πίνακας δεν καθορίζει το αν οι προβλέψεις του μοντέλου είναι σωστές ή όχι, αλλά ελέγχει τις μεταβλητές που βοήθησαν στη δημιουργία του δάσους και δίνει μεγαλύτερο συντελεστή σημαντικότητας σε εκείνες που βοήθησαν περισσότερο για τη δημιουργία του. Γενικά, όσο μεγαλύτερη τιμή έχει η μεταβλητή σε αυτόν τον πίνακα, τόσο πιο πολύτιμη είναι για την ανάπτυξη του δάσους.

Τελικά, οι τιμές που καθόρισαν την ακρίβεια του μοντέλου στα δεδομένα δοκιμής είναι οι TPR, SPC και ACC. Κύριο ρόλο έπαιξε το τελευταίο μέτρο όπου όσο μεγαλύτερο ήταν, τόσο μεγαλύτερη ήταν και η συνολική ακρίβεια των προβλέψεων του μοντέλου στα ταξίδια των δεδομένων δοκιμής. Πρέπει να επισημανθεί πως τα μέτρα που αναλύθηκαν παραπάνω και βγαίνουν από τον Πίνακα 4.1, αφορούν το σύνολο των ταξιδιών των δεδομένων δοκιμής. Αυτό σημαίνει πως περιλαμβάνεται η ακρίβεια ως προς το κάθε ένα ολοκληρωμένο ταξίδι και όχι για το κάθε δευτερόλεπτο ταξιδιού στο οποίο και έγινε πρόβλεψη ξεχωριστά. Για παράδειγμα, υπολογιζόταν ο αριθμός των δευτερολέπτων (έστω Α) ενός ταξιδιού που το μοντέλο προέβλεπε πως έγινε με αυτοκίνητο και ο αριθμός των δευτερολέπτων (έστω Β) στο ίδιο ταξίδι που το μοντέλο προέβλεπε πως έγινε με ΜΜΜ. Η πρόβλεψη γινόταν ώστε αν $A > B$ τότε το ταξίδι κατηγοριοποιούταν σαν ταξίδι με τρόπο μεταφοράς το αυτοκίνητο, ενώ αν $A \leq B$ τότε το ταξίδι κατηγοριοποιούταν σαν ταξίδι με τρόπο μεταφοράς τα ΜΜΜ.

5. ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΑΞΙΟΛΟΓΗΣΗ ΜΟΝΤΕΛΟΥ

Στο παρόν κεφάλαιο θα γίνει η παρουσίαση των αποτελεσμάτων της έρευνας, με τη χρήση δεδομένων που υπέστησαν την απαραίτητη επεξεργασία η οποία και αναλύθηκε στο Κεφάλαιο 3. Επίσης, θα γίνει η αξιολόγησή και ο σχολιασμός τους. Παρόλου που ο αριθμός των δοκιμών που έγιναν μέχρι να βρεθεί το βέλτιστο μοντέλο ήταν μεγάλος, εδώ θα αναλυθούν μόνο οι σημαντικότερες αλλαγές, οι οποίες βοήθησαν αισθητά στην αύξηση της ακρίβειας των αποτελεσμάτων. Με αυτόν τον τρόπο θα γίνει αντιληπτός ο τρόπος με τον οποίο βελτιώθηκε το μοντέλο, καθώς θα παρουσιάζονται τα βήματα που ακολουθήθηκαν.

Όπως αναλύθηκε και στο Κεφάλαιο 3, τα δεδομένα που χρησιμοποιήθηκαν προέρχονται από τους αισθητήρες επιταχυνσιομέτρου και γυροσκοπίου που βρίσκονται στο εσωτερικό του κινητού. Επίσης, χρησιμοποιήθηκε και ο αισθητήρας προσανατολισμού ο οποίος βασίζεται σε λογισμικό και εξάγει τα αποτελέσματά του με τη βοήθεια των άλλων δύο. Από τους αισθητήρες αυτούς, τα μεγέθη που εξήχθησαν είναι η επιτάχυνση του κινητού σε συνάρτηση με την επιτάχυνση της βαρύτητας, η γωνιακή του ταχύτητα όταν περιστρέφεται και η θέση του όπως αυτή καταγράφεται από τα Pitch και Roll. Στη συνέχεια, βρέθηκε η συνισταμένη της επιτάχυνσης και της γωνιακής ταχύτητας, διότι αυτά τα μεγέθη καταγράφονταν σε τρεις άξονες (x, y, z) και είναι μη αποτελεσματικό να συμπεριληφθούν αυτές οι τιμές καθώς εξαρτώνται σε πολύ μεγάλο βαθμό από τη θέση του κινητού. Η συνισταμένη, επειδή δεν επηρεάζεται από τη θέση του κινητού δίνει μια σωστή εικόνα για την επιτάχυνση και την γωνιακή ταχύτητα που αποκτά το κινητό κατά τη διάρκεια του ταξιδιού. Έπειτα, έγινε χρήση των χρονικών παραθύρων με σκοπό την εξομάλυνση των δεδομένων από πιθανές απότομες κινήσεις των κινητών. Με τα χρονικά αυτά παράθυρα δημιουργήθηκαν νέες μεταβλητές, οι οποίες αποτελούν τον κορμό του μοντέλου πρόβλεψης που αναπτύχθηκε. Συγκεκριμένα, ο μέσος όρος, η διάμεσος, η μέγιστη και η ελάχιστη τιμή, η τυπική απόκλιση και η διασπορά υπολογίστηκαν ανά οχτώ δευτερόλεπτα για κάθε ένα από τα τέσσερα μεγέθη που αναφέρθηκαν προηγουμένως και για όλα τα ταξίδια που συμπεριλήφθηκαν στη δημιουργία του μοντέλου. Αυτό το χρονικό διάστημα των οχτώ δευτερολέπτων λέγεται χρονικό παράθυρο, ενώ αλληλοκάλυψη μεταξύ δύο διαδοχικών παραθύρων εφαρμόστηκε για εφτά δευτερόλεπτα. Μετά τη δημιουργία των χρονικών παραθύρων, ξεκίνησαν οι δοκιμές του μοντέλου στις οποίες

αρχικά γινόντουσαν προσθήκες και αφαιρέσεις μεταβλητών, αλλαγές στα ποσοστά εκπαίδευσης και δοκιμής και αλλαγές των παραμέτρων του μοντέλου, με σκοπό φυσικά τη βελτίωση του μοντέλου και την επίτευξη υψηλότερης ακρίβειας στην πρόβλεψη. Οι σημαντικότερες αλλαγές που πραγματοποιήθηκαν στο μοντέλο πρόβλεψης μέχρι και την απόκτηση της τελικής του μορφής, αναλύονται στη συνέχεια του κεφαλαίου.

Μετά την επεξεργασία των δεδομένων, η διαδικασία εκπαίδευσης και πρόβλεψης του μοντέλου ήταν αρκετά απλή χάρη στο πακέτο h2o. Πιο συγκεκριμένα, με τη χρήση μίας συγκεκριμένης εντολής επιλέγονται τα δεδομένα εκπαίδευσης, οι μεταβλητές που θα συμμετέχουν στη δημιουργία του μοντέλου, η μεταβλητή στην οποία θα γίνει πρόβλεψη και τέλος οι παράμετροι του τυχαίου δάσους. Μετά τη δημιουργία του δάσους, αποθηκεύεται ο πίνακας σημαντικότητας των μεταβλητών και στη συνέχεια δοκιμάζεται το μοντέλο στα δεδομένα δοκιμής τα οποία φυσικά δε χρησιμοποιήθηκαν για την εκπαίδευσή του. Τέλος, δημιουργείται ο πίνακας με τις προβλέψεις του μοντέλου πάνω στα δεδομένα δοκιμής και υπολογίζονται οι όροι που αναφέρθηκαν στο Κεφάλαιο 4.7 ώστε να γίνει αξιολόγησή του.

Για τη δημιουργία του τυχαίου δάσους εφαρμόστηκαν συγκεκριμένοι παράμετροι οι οποίοι ήταν διαθέσιμοι για επεξεργασία από το πακέτο h2o που χρησιμοποιήθηκε. Η επιλογή τους εξαρτήθηκε από την ακρίβεια των προβλέψεων που είχε το μοντέλο μετά από πολλές δοκιμασίες διαφορετικών περιπτώσεων. Οι παράμετροι αυτοί φαίνονται στον Πίνακα 5.1 και προέκυψαν μετά από αρκετές δοκιμές με στόχο τη βελτιστοποίηση του μοντέλου.

Πίνακας 5.1 Παράμετροι μοντέλου

ΠΑΡΑΜΕΤΡΟΙ	
ΔΕΝΔΡΑ (ntrees)	250
ΜΕΓΙΣΤΟ ΒΑΘΟΣ (max_depth)	20
ΡΥΘΜΟΣ ΕΚΜΑΘΗΣΗΣ (sample_rate)	0.25
mtries	-1

Συγκεκριμένα, ο αριθμός των δένδρων που αναπτύχθηκαν στο δάσος τέθηκε στα 250. Μικρότερος αριθμός από αυτόν σηματοδοτούσε τη μείωση της αποτελεσματικότητας του μοντέλου ενώ μεγαλύτερος αριθμός δεν έδειξε καμία μεταβολή στις προβλέψεις. Επίσης, το μέγιστο βάθος του κάθε δένδρου οριοθετήθηκε στα 20. Αυτό εμποδίζει κατά κάποιον τρόπο την περαιτέρω ανάπτυξη του δένδρου και την πιθανή υπερεκπαίδευσή του μοντέλου.

Επιπλέον, μια άλλη σημαντική παράμετρος ανάπτυξης του μοντέλου είναι ο ρυθμός εκμάθησης. Αν ορισθεί η τιμή 1 τότε το κάθε δένδρο εκπαιδεύεται σε όλα τα δεδομένα εκπαίδευσης, ενώ αν τεθεί η τιμή 0.5 εκπαιδεύεται στα μισά. Όσο πιο μεγάλη είναι η τιμή του, τόσο πιο πολύ κίνδυνος υπάρχει για την υπερεκπαίδευση του μοντέλου. Η τελευταία τιμή που επιλέχθηκε για το μοντέλο είναι η mtries η οποία υπάρχει στο πακέτο h2o. Η παράμετρος αυτή επιλέγει τον αριθμό των μεταβλητών που θα είναι υποψήφιες για τον κάθε χωρισμό κόμβου στην ανάπτυξη των δένδρων. Με την τιμή $mtries = -1$ ο αριθμός αυτός είναι η τετραγωνική ρίζα του συνολικού αριθμού των μεταβλητών που χρησιμοποιούνται για την εκπαίδευση του μοντέλου.

Το πρώτο βήμα πριν ξεκινήσει η κανονική ροή των δοκιμών, είναι η πρόβλεψη του μοντέλου χωρίς τα δεδομένα να υποστούν κάποια επεξεργασία. Δηλαδή εκπαίδευση και πρόβλεψη του μοντέλου στα ακατέργαστα δεδομένα που συλλέχθηκαν από τους αισθητήρες. Οι μεταβλητές που χρησιμοποιήθηκαν για αυτή την εκπαίδευση είναι η επιτάχυνση και η γωνιακή ταχύτητα του κινητού ως προς τους 3 άξονες, το Pitch και Roll, δηλαδή συνολικά 8. Τα αποτελέσματα των προβλέψεων φαίνονται στον Πίνακα 5.2.

Πίνακας 5.2 Αποτελέσματα προβλέψεων σε ακατέργαστα δεδομένα

		Πραγματικό MMM	**Πραγματικό Αυτοκίνητο**	sum
Πρόβλεψη	**MMM**	18	8	26
Πρόβλεψη	**Αυτοκίνητο**	19	67	86
	sum	37	75	112
	TPR	0.486	**FDR**	0.308
	SPC	0.893	**FNR**	0.514
	PPV	0.692	**ACC**	0.759
	NPV	0.779	**F1**	0.571
	FPR	0.107	**MCC**	0.423

Στη συνέχεια, υπολογίστηκαν η συνισταμένη της επιτάχυνσης και της γωνιακής ταχύτητας, με αποτέλεσμα η επόμενη δοκιμή να γίνει με 4 μόνο μεταβλητές. Τις 2 συνισταμένες που αναφέρθηκαν και τις ακατέργαστες τιμές Pitch και Roll. Τα αποτελέσματα των προβλέψεων φαίνονται στον Πίνακα 5.3.

Πίνακας 5.3 Αποτελέσματα προβλέψεων με τη χρήση τεσσάρων μεταβλητών

		Πραγματικό MMM	**Πραγματικό Αυτοκίνητο**	sum
Πρόβλεψη	**MMM**	16	6	22
Πρόβλεψη	**Αυτοκίνητο**	21	69	90
	sum	37	75	112
	TPR	0.432	**FDR**	0.273
	SPC	0.92	**FNR**	0.568
	PPV	0.727	**ACC**	0.759
	NPV	0.767	**F1**	0.542
	FPR	0.08	**MCC**	0.417

Από τον παραπάνω πίνακα φαίνεται πως το μοντέλο με τη χρήση των συνισταμένων δε βελτιώθηκε καθώς η συνολική ακρίβεια παρέμεινε σταθερή. Για αυτό το λόγο τα δεδομένα έπρεπε να επεξεργαστούν περισσότερο ώστε νέες μεταβλητές να κάνουν την εμφάνισή τους με σκοπό τη βελτίωση του μοντέλου και έτσι επιλέχθηκε να γίνει χρήση των χρονικών παραθύρων. Ο καθορισμός του χρόνου διάρκειας κάθε χρονικού παραθύρου, αλλά και της χρονικής διαφοράς μεταξύ δύο διαδοχικών παραθύρων έγινε μετά από δοκιμές. Έτσι, η ακρίβεια των προβλέψεων ήταν βέλτιστη σε ότι είχε να κάνει με το μέγεθος των παραθύρων, επομένως διευκολύνθηκε η ανάλυση του μοντέλου, αφού οι μετέπειτα δοκιμές πραγματοποιήθηκαν με μεγαλύτερη ταχύτητα. Όπως αναφέρεται και στο Κεφάλαιο 2.4.3, η διάρκεια ενός παραθύρου μπορεί να είναι πολύ μικρή έως και πολύ μεγάλη. Εδώ λοιπόν, δοκιμάστηκαν πέντε διαφορετικές χρονικές διάρκειες παραθύρων, από έξι έως δέκα δευτερόλεπτα, ενώ η χρονική διαφορά δύο διαδοχικών παραθύρων τέθηκε στις δοκιμασίες από ένα μέχρι και τρία δευτερόλεπτα. Τα αποτελέσματα με την επικράτηση του βέλτιστου παραθύρου χρονικής διάρκειας οχτώ δευτερολέπτων και χρονικής διαφοράς δύο διαδοχικών παραθύρων στο ένα δευτερόλεπτο, φαίνεται στον Πίνακα 5.4, ενώ για λόγους σύγκρισης παρατίθενται και τα αποτελέσματα του δυσμενέστερου παραθύρου που εμφανίστηκε κατά τη διάρκεια των δοκιμών στον Πίνακα 5.5. Το δυσμενέστερο αυτό παράθυρο έχει χρονική διάρκεια έξι δευτερόλεπτα ενώ η χρονική διαφορά μεταξύ διαδοχικών παραθύρων είναι στα τρία δευτερόλεπτα.

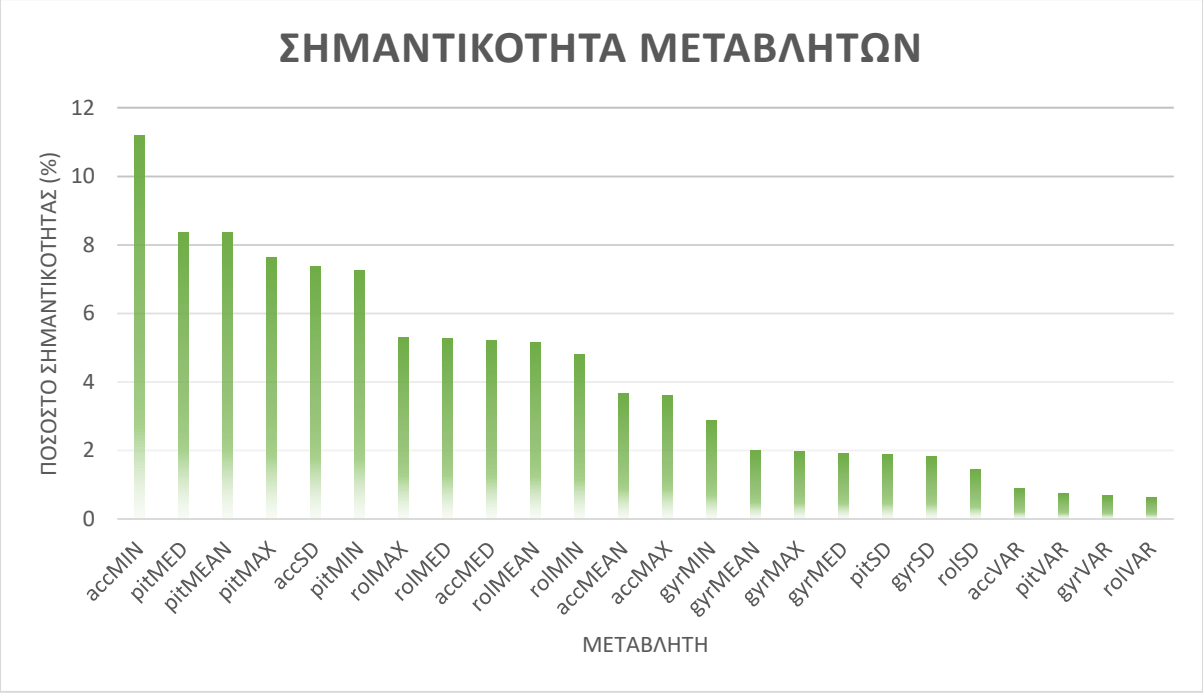
Πίνακας 5.4 Αποτελέσματα προβλέψεων με τη χρήση του βέλτιστου χρονικού παραθύρου

		Πραγματικό MMM	Πραγματικό Αυτοκίνητο	sum
Πρόβλεψη	**MMM**	22	8	30
Πρόβλεψη	**Αυτοκίνητο**	15	67	82
	sum	37	75	112
	TPR	0.595	**FDR**	0.267
	SPC	0.893	**FNR**	0.405
	PPV	0.733	**ACC**	0.795
	NPV	0.817	**F1**	0.657
	FPR	0.107	**MCC**	0.518

Πίνακας 5.5 Αποτελέσματα προβλέψεων με τη χρήση του δυσμενέστερου χρονικού παραθύρου

		Πραγματικό MMM	Πραγματικό Αυτοκίνητο	sum
Πρόβλεψη	**MMM**	21	9	30
Πρόβλεψη	**Αυτοκίνητο**	16	66	82
	sum	37	75	112
	TPR	0.568	**FDR**	0.3
	SPC	0.88	**FNR**	0.432
	PPV	0.7	**ACC**	0.777
	NPV	0.805	**F1**	0.628
	FPR	0.12	**MCC**	0.475

Στα δύο παραπάνω μοντέλα των πινάκων έγινε εκπαίδευση και πρόβλεψη με όλες τις διαθέσιμες μεταβλητές που υπολογίστηκαν από τα χρονικά παράθυρα. Οι μεταβλητές αυτές, όπως αναλύονται και στο Κεφάλαιο 3.4, είναι είκοσι τέσσερις, ενώ η σημαντικότητα της κάθε μία από αυτές για το μοντέλο του Πίνακα 5.4 φαίνεται στο Διάγραμμα 5.1.



Διάγραμμα 5.1 Σημαντικότητα μεταβλητών με τη χρήση του βέλτιστου χρονικού παραθύρου

Μετά την επιλογή του βέλτιστου χρονικού παράθυρου, χρησιμοποιήθηκε ακόμα ένας αισθητήρας ο οποίος δεν αναφέρεται καθόλου στην εργασία καθώς εγκαταλείφθηκε από πολύ νωρίς η χρήση του. Ο αισθητήρας αυτός είναι το μαγνητόμετρο, ο οποίος υπάρχει στα νέα κινητά τηλέφωνα, μαζί με το γυροσκόπιο και το επιταχυνσιόμετρο. Παρότι μέσω λογισμικού έχει και άλλες δυνατότητες, πρακτικά μετράει το μαγνητικό πεδίο της Γης στους τρεις άξονες και για μονάδα μέτρησης χρησιμοποιεί τα μικροTesla (μT). Για τις τιμές αυτές λοιπόν βρέθηκε η συνισταμένη τους και στη συνέχεια εφαρμόστηκαν τα χρονικά παράθυρα. Δυστυχώς, περίπου το 45% των ταξιδιών που χρησιμοποιήθηκαν για την εκπλήρωση της παρούσας εργασίας, δεν περιείχαν δεδομένα από το μαγνητόμετρο και αυτός ήταν ένας από τους λόγους που η χρήση του εγκαταλείφθηκε νωρίς. Ωστόσο, έγινε χρήση των τιμών που παρείχε το μαγνητόμετρο μαζί με όλες τις προηγούμενες μεταβλητές στο υπόλοιπο 55% και διαπιστώθηκε πως δεν επηρεάζει πολύ την ακρίβεια του μοντέλου. Ενδεικτικά, τα αποτελέσματα φαίνονται Πίνακα 5.6 όπου από τον συνολικό αριθμό ταξιδιών που καταγράφηκαν μετρήσεις μαγνητομέτρου, επιλέχθηκε το 75% αυτών για δεδομένα εκπαίδευσης και το υπόλοιπο 25% για δεδομένα δοκιμής. Ωστόσο, στα ίδια ακριβώς ταξίδια έτρεξε το μοντέλο και χωρίς τη χρήση του μαγνητομέτρου, τα αποτελέσματα του οποίου φαίνονται στον Πίνακα 5.7.

Πίνακας 5.6 Αποτελέσματα προβλέψεων με τη χρήση όλων των διαθέσιμων μεταβλητών (και του μαγνητόμετρου)

		Πραγματικό MMM	**Πραγματικό Αυτοκίνητο**	sum
Πρόβλεψη	**MMM**	31	4	35
Πρόβλεψη	**Αυτοκίνητο**	8	32	40
	sum	39	36	75
	TPR	0.795	**FDR**	0.114
	SPC	0.889	**FNR**	0.205
	PPV	0.886	**ACC**	0.84
	NPV	0.8	**F1**	0.838
	FPR	0.111	**MCC**	0.685
	MAE	0.042	**RMSE**	0.101

Πίνακας 5.7 Αποτελέσματα προβλέψεων με χρήση όλων των διαθέσιμων μεταβλητών (πλην του μαγνητόμετρου)

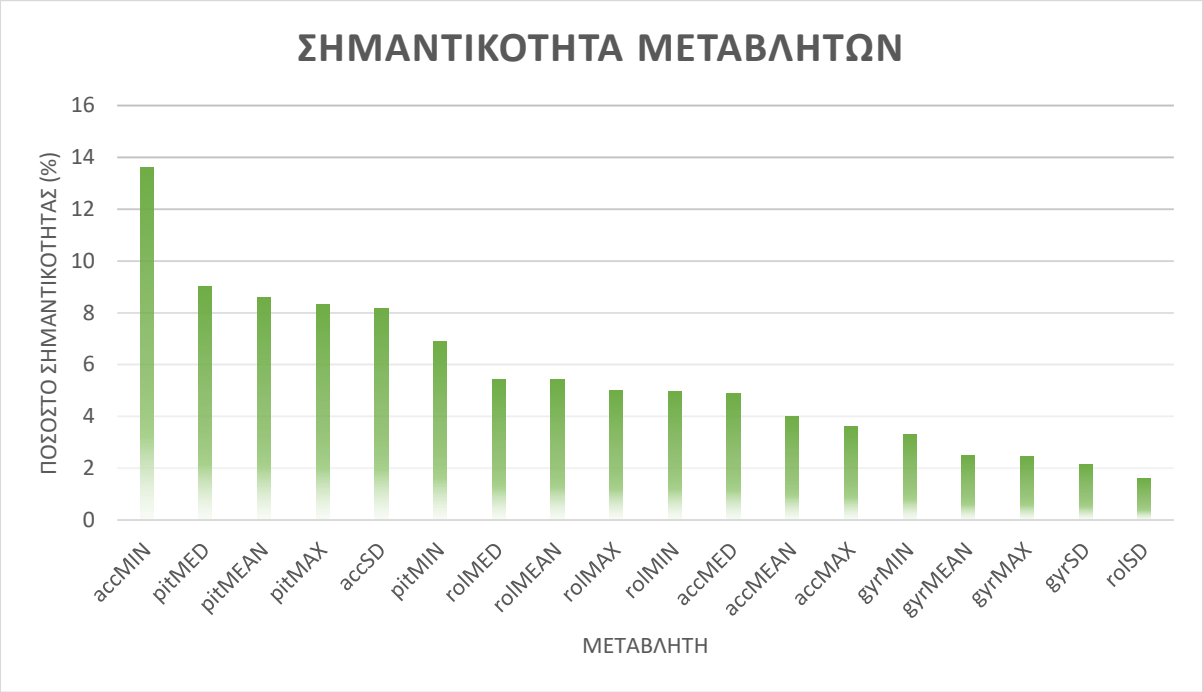
		Πραγματικό MMM	**Πραγματικό Αυτοκίνητο**	sum
Πρόβλεψη	**MMM**	31	3	34
Πρόβλεψη	**Αυτοκίνητο**	8	33	41
	sum	39	36	75
	TPR	0.795	**FDR**	0.088
	SPC	0.917	**FNR**	0.205
	PPV	0.912	**ACC**	0.853
	NPV	0.805	**F1**	0.849
	FPR	0.083	**MCC**	0.714
	MAE	0.055	**RMSE**	0.127

Παρατηρείται λοιπόν από τους παραπάνω πίνακες πως η χρήση του μαγνητόμετρου οδήγησε στη μείωση της συνολικής ακρίβειας του μοντέλου, αφού προέβλεψε ένα παραπάνω ταξίδι λάθος από ότι προέβλεψε το μοντέλο χωρίς τη χρήση του αισθητήρα αυτού. Επομένως, σε αυτό το σημείο κρίνεται απαραίτητη η απομάκρυνση όλων των μεταβλητών του μαγνητόμετρου, καθώς δεν αξίζει η μεγαλύτερη απαιτούμενη υπολογιστική ισχύ και η φόρτωση του μοντέλου με παραπάνω μεταβλητές εφόσον δεν υπάρχει αύξηση στην συνολική ακρίβεια του μοντέλου. Ωστόσο τα ταξίδια που κατέγραψαν δεδομένα από μαγνητόμετρο ήταν ελάχιστα και επομένως πρέπει να αναλυθεί περισσότερο η σημασία του συγκεκριμένου αισθητήρα στην ανάπτυξη παρόμοιων μοντέλων. Από το σημείο αυτό η χρησιμοποίηση του

μαγνητόμετρου στην παρούσα εργασία έλαβε τέλος και δε θα χρησιμοποιηθεί αλλά ούτε θα αναφερθεί ξανά. Τα υπόλοιπα αποτελέσματα που αναφέρονται στο παρόν κεφάλαιο δεν περιλαμβάνουν τιμές από μαγνητόμετρο. Επίσης να αναφερθεί πως, συγκρίνοντας τα αποτελέσματα των Πινάκων 5.4 και 5.7 παρατηρείται αύξηση στην ολική ακρίβεια πρόβλεψης του δεύτερου σε σχέση με τον πρώτο, ενώ όλες οι μεταβλητές και παράμετροι παραμένουν ίδιοι και στα δύο μοντέλα. Αυτό οφείλεται στο διαφορετικό δείγμα που χρησιμοποιήθηκε στα δύο μοντέλα, καθώς στην πρώτη περίπτωση χρησιμοποιήθηκαν όλα τα ταξίδια ενώ στη δεύτερη σχεδόν τα μισά λόγω της έλλειψης τιμών από το μαγνητόμετρο. Γίνεται φανερό λοιπόν πόσο πολύ μπορεί να επηρεάσει τα τελικά αποτελέσματα η ποσότητα και καθαρότητα του δείγματος.

Έπειτα, έγινε επιλογή των κατάλληλων μεταβλητών οι οποίες βοήθησαν στην αύξηση της αποτελεσματικότητας του μοντέλου και εν τέλει στην αύξηση της ακριβείας του. Πέρα από την ολοκληρωτική απομάκρυνση των τιμών του μαγνητόμετρου, κάποιες ακόμα μεταβλητές αναγκάστηκαν να διαγραφούν καθώς περισσότερο ¨βάραιναν¨ το μοντέλο παρά το βελτίωναν. Αρχικά, χρησιμοποιήθηκαν και οι 24 διαθέσιμες μεταβλητές ώστε να γίνει μια πρώτη εκτίμηση του μοντέλου και ανάλυση των αποτελεσμάτων του. Παράλληλα με τη δημιουργία του τυχαίου δάσους, παράγεται αυτόματα και ένας πίνακας που δείχνει τη σημαντικότητα των μεταβλητών στο μοντέλο αυτό, όπως αναφέρθηκε και στο Κεφάλαιο 4.7. Αρχικά, επιλέχθηκαν δύο μεταβλητές με τη μικρότερη σημαντικότητα και απομακρύνθηκαν από το μοντέλο. Στη συνέχεια, δημιουργήθηκε νέο τυχαίο δάσος και απομακρύνθηκαν πάλι 2 μεταβλητές με τη μικρότερη σημαντικότητα. Κάθε φορά που δημιουργούταν με αυτόν τον τρόπο ένα νέο τυχαίο δάσος υπήρχε βελτίωση στην ακρίβεια των προβλέψεων. Με αυτόν τον τρόπο το μοντέλο απέκτησε τη μέγιστη ακρίβειά του όταν έμειναν 18 μεταβλητές και αυτές καθόρισαν το μοντέλο που αναπτύχθηκε στην παρούσα εργασία. Ωστόσο, πρέπει να σημειωθεί πως η αλλαγή των μεταβλητών για την ανάπτυξη νέου τυχαίου δάσους δεν καθοριζόταν αποκλειστικά από τη σημαντικότητά τους, καθώς υπήρχαν περιπτώσεις στις οποίες ναι μεν απομακρύνονταν μεταβλητές με τη μικρότερη σημαντικότητα, αλλά το μοντέλο εξασθενούσε αντί να βελτιώνεται. Επομένως, επιπλέον προσθήκες και αφαιρέσεις μεταβλητών έγιναν ανάλογα με την κρίση του ερευνητή, που σκοπό είχαν τη βελτιστοποίηση του μοντέλου. Τελικά από τις 24 μεταβλητές έμειναν οι 18. Τα αποτελέσματα των προβλέψεων του μοντέλου τόσο με τη χρήση όλων των μεταβλητών όσο και με τη χρήση των 18 συμπίπτουν και είναι αυτά του Πίνακα 5.4. Μάλιστα, τα ταξίδια που δεν αναγνωρίστηκαν σωστά επίσης συμπίπτουν και για τις δύο περιπτώσεις. Ωστόσο περαιτέρω μείωση των μεταβλητών έριχνε την ποιότητα των προβλέψεων κάτι το οποίο δεν ήταν επιθυμητό. Από τις 6 μεταβλητές που αφαιρέθηκαν οι 4 αφορούν τη διασπορά που είχε υπολογιστεί στα χρονικά παράθυρα, ενώ αφαιρέθηκε και η

τυπική απόκλιση του Pitch και η διάμεσος της γωνιακής ταχύτητας. Οι τελικές μεταβλητές που παρέμειναν μέχρι και τη δημιουργία του τελικού μοντέλου καθώς και το ποσοστό σημαντικότητάς τους σε αυτό, φαίνεται στο Διάγραμμα 5.2.



Διάγραμμα 5.2 Σημαντικότητα μεταβλητών μετά την αφαίρεση των δυσμενέστερων

Στη συνέχεια, παρατηρήθηκε πως υπήρχε δυσκολία στη σωστή πρόβλεψη του τρόπου μεταφοράς για έναν από τους 27 χρήστες. Ο συγκεκριμένος χρήστης που στην έρευνα είχε τον κωδικό αριθμό 10, όπως φαίνεται και από τον Πίνακα 3.1 είχε πραγματοποιήσει συνολικά 25 ταξίδια, από τα οποία τα 10 ήταν με αυτοκίνητο και τα υπόλοιπα 15 με λεωφορείο. Σε κάθε εκπαίδευση και δοκιμή του μοντέλου που πραγματοποιήθηκε με τη συμμετοχή όλων των χρηστών, όσα ταξίδια του χρήστη αυτού και αν συμμετείχαν στην εκπαίδευση του μοντέλου, τα υπόλοιπα, στην πλειοψηφία τους, δεν εντοπίζονταν σωστά. Ήταν ο χρήστης για τον οποίο το μοντέλο έκανε τις περισσότερες λάθος προβλέψεις και επομένως έπεφτε η συνολική ακρίβεια του μοντέλου. Αυτό ίσως οφείλεται στο γεγονός ότι ο ίδιος δεν επικύρωνε σωστά τα ταξίδια του μετά την ολοκλήρωσή τους. Συνεπώς, θεωρήθηκε ορθά να γίνουν δοκιμές με την απομάκρυνση του συγκεκριμένου χρήστη από το μοντέλο και τελικά διαπιστώθηκε αύξηση στην απόδοσή του και στην περαιτέρω βελτιστοποίησή του. Επομένως, το τελικό μοντέλο που αναπτύχθηκε δεν περιλαμβάνει κανένα ταξίδι του συγκεκριμένου χρήστη. Παρακάτω, παρατίθεται ο πίνακας με τα αποτελέσματα πρόβλεψης χωρίς τη συμμετοχή του χρήστη αυτού στο μοντέλο και επομένως μπορεί να γίνει σύγκριση με τον Πίνακα 5.4 καθώς οι μεταβλητές,

οι παράμετροι και τα ποσοστά των δεδομένων εκπαίδευσης και δοκιμής που χρησιμοποιήθηκαν και στα δύο μοντέλα είναι ίδια. Ωστόσο διαφέρει το δείγμα καθώς τα αποτελέσματα του Πίνακα 5.8 δεν περιλαμβάνουν ταξίδια του χρήστη με το νούμερο 10, ενώ ο Πίνακας 5.4 περιλαμβάνει τα ταξίδια όλων των χρηστών.

Πίνακας 5.8 Αποτελέσματα προβλέψεων μετά την αφαίρεση του χρήστη με κωδικό αριθμό 10

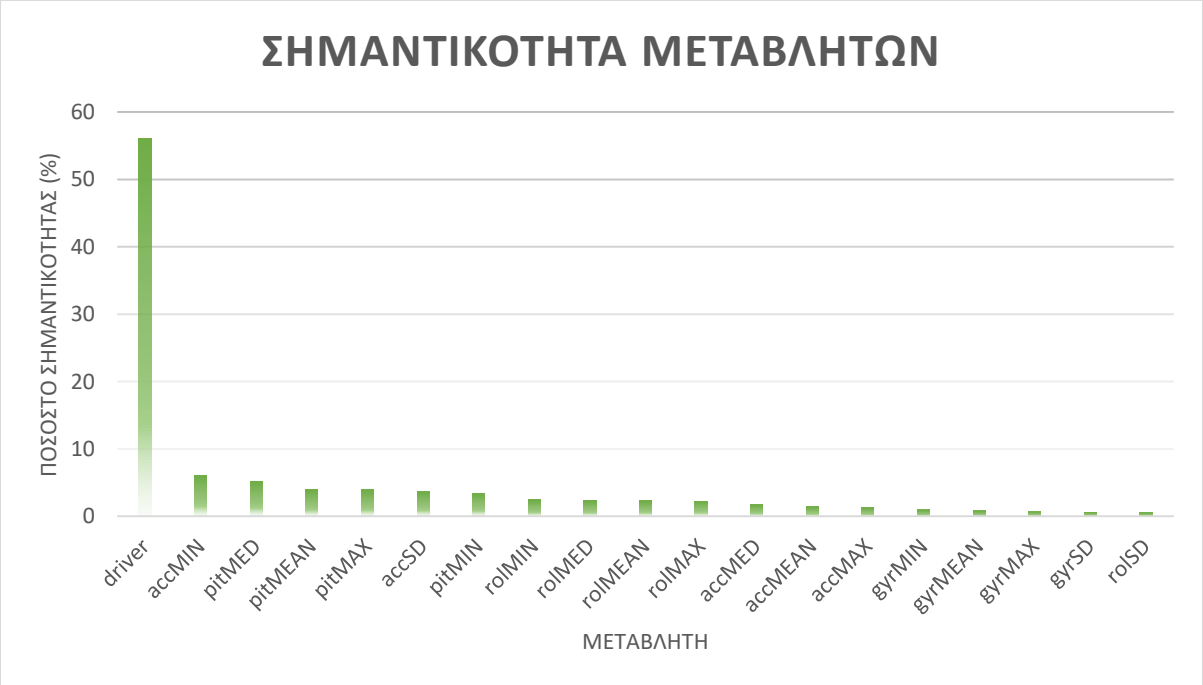
		Πραγματικό MMM	**Πραγματικό Αυτοκίνητο**	sum
Πρόβλεψη	**MMM**	29	5	34
Πρόβλεψη	**Αυτοκίνητο**	11	64	75
	sum	40	69	109
	TPR	0.725	**FDR**	0.147
	SPC	0.928	**FNR**	0.275
	PPV	0.853	**ACC**	0.853
	NPV	0.853	**F1**	0.784
	FPR	0.072	**MCC**	0.679

Πέρα από τις παραπάνω βελτιώσεις που εφαρμόστηκαν στο μοντέλο για την ενδυνάμωσή του, μια ακόμη σημαντική αλλαγή πραγματοποιήθηκε η οποία οδήγησε στην αύξηση κατά περίπου 8% της ακριβείας του. Μέχρι αυτό το σημείο, χρησιμοποιούνταν μόνο οι 18 μεταβλητές των οποίων η σημαντικότητα φαίνεται στο Διάγραμμα 5.2. Προστέθηκε, λοιπόν, ακόμη μια μεταβλητή η οποία περιλαμβάνει τους κωδικούς των χρηστών του Πίνακα 3.1. Η νέα αυτή μεταβλητή ονομάστηκε ‘driver’. Αφού ο αριθμός των χρηστών των οποίων τα ταξίδια χρησιμοποιήθηκαν τελικά στην παρούσα έρευνα, μετά την αφαίρεση του χρήστη με τον αριθμό 10, ήταν 26, το εύρος τιμών της μεταβλητής αυτής ήταν από το 1 έως το 26. Με αυτόν τον τρόπο, το μοντέλο ξεχωρίζει τους χρήστες μεταξύ τους και πλέον μαθαίνει και αναγνωρίζει τα πρότυπα και μοτίβα που ακολουθεί ο κάθε ένας από αυτούς, κάτι το οποίο παίζει σημαντικό ρόλο στην μετέπειτα πρόβλεψη του τρόπου μεταφοράς του. Στη συνέχεια, παρουσιάζεται ο Πίνακας 5.9 με τα αποτελέσματα του μοντέλου που εκπαιδεύτηκε και με τη χρήση της νέας αυτής μεταβλητής και μπορεί να γίνει σύγκρισή του με τον Πίνακα 5.8, αφού όλα τα υπόλοιπα χαρακτηριστικά του μοντέλου παρέμειναν σταθερά.

Πίνακας 5.9 Αποτελέσματα προβλέψεων με τη χρήση της νέας μεταβλητής "driver"

		Πραγματικό MMM	**Πραγματικό Αυτοκίνητο**	sum
Πρόβλεψη	**MMM**	37	1	38
Πρόβλεψη	**Αυτοκίνητο**	3	68	71
	sum	40	69	109
	TPR	0.925	**FDR**	0.026
	SPC	0.986	**FNR**	0.075
	PPV	0.974	**ACC**	0.963
	NPV	0.958	**F1**	0.949
	FPR	0.014	**MCC**	0.921

Παρατηρείται λοιπόν ότι η ακρίβεια των προβλέψεων έχει αυξηθεί αισθητά σε σχέση με τις πρώτες δοκιμές που έγιναν. Στο Διάγραμμα 5.3 φαίνονται οι μεταβλητές που χρησιμοποιήθηκαν για τη δημιουργία του τελικού μοντέλου σε συνδυασμό με τη σημαντικότητα της κάθε μεταβλητής. Αξιοσημείωτη είναι η απορρόφηση της σημαντικότητας στη νέα αυτή μεταβλητή.



Διάγραμμα 5.3 Σημαντικότητα μεταβλητών στη δημιουργία του τελικού μοντέλου

Πολύ σημαντική παράμετρος για τη δημιουργία του μοντέλου είναι η σωστή αναλογία μεταξύ των δεδομένων εκπαίδευσης και δοκιμής. Αξίζει να σημειωθεί πως στην έρευνα που πραγματοποιήθηκε για την παρούσα διπλωματική εργασία, δε χρησιμοποιήθηκαν δεδομένα επικύρωσης, παρά μόνο δεδομένα εκπαίδευσης και δοκιμής. Αυτό συνέβη καθώς ο αλγόριθμος των τυχαίων δασών δημιουργεί από μόνος του δεδομένα επικύρωσης για να ελέγξει την αποτελεσματικότητά του, εξάγοντας συγκεκριμένο αριθμό δεδομένων από αυτά της εκπαίδευσης. Ο τρόπος με τον οποίο επιτυγχάνεται αυτό έχει ήδη αναφερθεί στο Κεφάλαιο 4 σαν δεδομένα ¨εκτός σακούλας¨. Στη συνέχεια, απαιτούταν ο καθορισμός ενός ποσοστού μεταξύ των δεδομένων εκπαίδευσης και δοκιμής με στόχο τη βέλτιστη αποτελεσματικότητα του μοντέλου. Επειδή οι προβλέψεις μετριούνταν συνολικά για το κάθε ταξίδι, ο χωρισμός των δεδομένων σε εκπαίδευσης και δοκιμής πρέπει να γίνει με τον ίδιο τρόπο. Για αυτό το λόγο δημιουργήθηκε ένας αλγόριθμος που δίνει τη δυνατότητα στον ερευνητή να επιλέξει έναν αριθμό ταξιδιών που θα χρησιμοποιήσει για να εκπαιδεύσει το μοντέλο, ενώ τα υπόλοιπα ταξίδια θα λειτουργήσουν για τη δοκιμή του μοντέλου αυτού. Γενικά, όταν αυξάνονται τα δεδομένα εκπαίδευσης τότε το μοντέλο έχει μεγαλύτερη ακρίβεια λόγω του ότι έχει ¨μάθει¨ περισσότερα και επομένως μπορεί να προβλέψει καλύτερα.

Ωστόσο όταν χρησιμοποιείται ένα πολύ υψηλό ποσοστό για τα δεδομένα εκπαίδευσης, τότε το μοντέλο κινδυνεύει να υποστεί υπερεκπαίδευση και να χάσει την ικανότητα γενικοποίησής του. Αυτό σημαίνει πως η ακρίβειά του θα είναι μειωμένη όταν δοκιμαστεί σε νέο πληθυσμό. Παρόλα αυτά, δεν υπάρχει σωστός και λάθος χωρισμός των ποσοστών αυτών, καθώς τα αποτελέσματα μπορεί να διαφέρουν ανάλογα με το μοντέλο που δημιουργείται. Επομένως, η επιλογή των ποσοστών έγκειται στην κρίση του μελετητή. Στη συνέχεια, για να γίνει κατανοητό αυτό παρατίθενται δύο διαγράμματα που απεικονίζουν τις αλλαγές στη συνολική ακρίβεια πρόβλεψης σε σχέση με το ποσοστό των δεδομένων εκπαίδευσης και δοκιμής που επιλεγόταν κάθε φορά. Οι παράμετροι που χρησιμοποιήθηκαν είναι οι ίδιοι με τα δύο τελευταία μοντέλα που αναφέρθηκαν, δηλαδή η διαφορά τους είναι μόνο ότι το μοντέλο του Πίνακας 5.10 δεν περιλαμβάνει τη μεταβλητή “driver” ενώ του Πίνακα 5.11 την περιλαμβάνει. Σε αυτή την εργασία επιλέχθηκε να γίνει η εκπαίδευση του μοντέλου με το 80% των συνολικών δεδομένων και η δοκιμή του στο υπόλοιπο 20%, αφού σε αυτό το ποσοστό επιτυγχάνεται και η υψηλότερη ακρίβεια όπως φαίνεται από τους Πίνακες 5.10 και 5.11. Επίσης, αυτό το ποσοστό επιλέγεται επί το πλείστων στις έρευνες, χωρίς ωστόσο να σημαίνει ότι οποιαδήποτε απόκλιση από αυτό θα οδηγήσει σε λανθασμένη εκπαίδευση του μοντέλου αφού, όπως αναφέρθηκε, ο κάθε ερευνητής επιλέγει το ποσοστό που θα φέρει καλύτερα αποτελέσματα στο μοντέλο που δημιούργησε.

Πίνακας 5.10 Ακρίβεια πρόβλεψης ανάλογα με το ποσοστό δεδομένων εκπαίδευσης/δοκιμής χωρίς τη χρήση της μεταβλητής "driver"

ΔΕΔΟΜΕΝΑ		ΑΚΡΙΒΕΙΑ ΠΡΟΒΛΕΨΗΣ (%)
ΕΚΠΑΙΔΕΥΣΗΣ (%)	ΔΟΚΙΜΗΣ (%)	
30	70	84.72
35	65	85.59
40	60	86.88
45	55	86.35
50	50	86.52
55	45	85.42
60	40	85.45
65	35	84.41
70	30	86.88
75	25	86.47
80	20	84.91
85	15	88.75
90	10	86.79

Πίνακας 5.11 Ακρίβεια πρόβλεψης ανάλογα με το ποσοστό δεδομένων εκπαίδευσης/δοκιμής με τη χρήση της μεταβλητής "driver"

ΔΕΔΟΜΕΝΑ		ΑΚΡΙΒΕΙΑ ΠΡΟΒΛΕΨΗΣ (%)
ΕΚΠΑΙΔΕΥΣΗΣ (%)	ΔΟΚΙΜΗΣ (%)	
30	70	93.30
35	65	94.24
40	60	93.75
45	55	93.86
50	50	93.63
55	45	92.92
60	40	94.37
65	35	94.09
70	30	94.38
75	25	95.49
80	20	97.17
85	15	96.25
90	10	98.11

Πλέον, το μοντέλο έχει αποκτήσει μια εξαιρετική ακρίβεια στις προβλέψεις του και δεν χρειάζεται περαιτέρω εκπαίδευση. Η αποτελεσματικότητά του έχει αυξηθεί αισθητά σε σύγκριση με αυτήν που είχε στις πρώτες δοκιμές που έγιναν και αποτελεί μια επιτυχημένη προσπάθεια εντοπισμού του τρόπου μεταφοράς χωρίς τη χρήση του GPS. Οι στόχοι της παρούσας εργασίας επιτεύχθηκαν καθώς παρουσιάζεται μια αποτελεσματική λύση για την αναγνώριση του τρόπου μεταφοράς των χρηστών η οποία δεν καταναλώνει τη μπαταρία του κινητού και προσφέρει πολύ μεγάλη ακρίβεια στις προβλέψεις.

Όπως φαίνεται από τις προβλέψεις, η προσθήκη της μεταβλητής που απαριθμεί τους χρήστες, παρείχε πολύ μεγάλη βελτίωση στην ακρίβεια των αποτελεσμάτων. Αυτό όμως μπορεί να οδηγήσει και σε προβλήματα. Στην παρούσα εργασία, οι χρήστες που συμπεριλήφθηκαν τελικά στις δοκιμές ήταν μόνο 26, πράγμα που σημαίνει ότι δεν έχει γίνει εκτενής δοκιμή της συγκεκριμένης μεταβλητής. Το μοντέλο μπορεί να μην συμπεριφέρεται το ίδιο καλά όταν υπάρχουν 10.000 ή 100.000 χρήστες των οποίων τα ταξίδια πρέπει να κατηγοριοποιηθούν. Επιπλέον, η προσθήκη αυτής της μεταβλητής στο μοντέλο, σημαίνει πως το μοντέλο θα δυσκολευτεί στη σωστή πρόβλεψη του τρόπο μεταφοράς νέων χρηστών που δεν συμπεριλήφθηκαν στην εκπαίδευση. Τέλος, πρέπει να επισημανθεί πως παρά τα υψηλά ποσοστά ακριβείας που επιτεύχθηκαν για τον εντοπισμό του τρόπου μεταφοράς των χρηστών,

δεν αναιρείται το γεγονός πως απαιτείται περεταίρω δοκιμή σε περισσότερα ταξίδια και χρήστες κάτι το οποίο δεν ήταν εφικτό εδώ καθώς δεν υπήρχαν άλλα διαθέσιμα δεδομένα.

6. ΣΥΜΠΕΡΑΣΜΑΤΑ

6.1. ΓΕΝΙΚΑ

Στόχος της παρούσας διπλωματικής εργασίας είναι η πρόβλεψη του μέσου μεταφοράς ενός χρήστη οδικού δικτύου, ανάμεσα στη χρήση των Μέσων Μαζικής Μεταφοράς (ΜΜΜ) και τα αυτοκίνητα, με την εφαρμογή μηχανικής μάθησης και με τη βοήθεια δεδομένων που αντλήθηκαν αποκλειστικά από τους αισθητήρες κινητών τηλεφώνων. Οι κατηγορίες αυτές, ΜΜΜ και αυτοκίνητο, είναι οι βασικότερες για κίνηση σε οδικό δίκτυο για αυτό επιλέχθηκαν.

Η σταδιακή βελτίωση του μοντέλου παρουσιάζεται στη συνέχεια, ώστε να γίνουν αντιληπτά τα βήματα που ακολουθήθηκαν με αποτέλεσμα να μπορούν στη συνέχεια να υιοθετηθούν εύκολα από άλλους ερευνητές με σκοπό την περαιτέρω αύξηση της αποτελεσματικότητας του μοντέλου.

Αρχικά, έγινε επεξεργασία των δεδομένων με διάφορες τεχνικές ώστε αυτά να ομαλοποιηθούν για να γίνουν πιο ευανάγνωστα από τον ερευνητή και ταυτόχρονα για να δημιουργηθούν νέες μεταβλητές οι οποίες θα παίξουν σημαντικό ρόλο στη μετέπειτα εκπαίδευση του μοντέλου. Τα δεδομένα που αντλήθηκαν από τους αισθητήρες ήταν η επιτάχυνση που κατέγραφε το κινητό κατά τη διάρκεια του ταξιδιού, η γωνιακή του ταχύτητα και οι τιμές Pitch και Roll. Έπειτα, βρέθηκαν οι συνισταμένες επιταχύνσεως και γωνιακής ταχύτητας καθώς η χρήση τριών αξόνων μπορεί να μπερδέψει και να ξεγελάσει το μοντέλο, αφού οι άξονες εξαρτώνται αρκετά από τη θέση που έχει το κινητό του χρήστη τη δεδομένη χρονική στιγμή, κάτι που δεν ήταν επιθυμητό στην παρούσα έρευνα και δεν βοηθάει τον ερευνητή στη σωστή αξιολόγηση του μοντέλου.

Στη συνέχεια, η χρήση των χρονικών παραθύρων έδωσε τη δυνατότητα σε νέες μεταβλητές να κάνουν την εμφάνισή τους, όπως ο μέσος όρος, το ελάχιστο και μέγιστο τιμών, η διάμεσος, η τυπική απόκλιση και η διασπορά, ενώ όλα αυτά υπολογίστηκαν για κάθε παράθυρο που αναπτύχθηκε. Επίσης, τα χρονικά παράθυρα εξομάλυναν τα ακατέργαστα δεδομένα που

εξάγονταν από τους αισθητήρες του κινητού, καθώς απότομες κινήσεις αυτού επηρεάζουν αρνητικά τα δεδομένα αυτά και δεν ανταποκρίνονταν εξολοκλήρου στην πραγματικότητα.

Με αυτά τα νέα και επεξεργασμένα δεδομένα ξεκίνησε η εκπαίδευση του μοντέλου με τη χρήση την τυχαίων δασών. Από την αρχή έγινε αντιληπτό πως οι τιμές της διασποράς επιβάρυναν το μοντέλο και αφαιρέθηκαν, ενώ στη συνέχεια αφαιρέθηκαν κάποιες ακόμα μεταβλητές που μείωναν την αποτελεσματικότητα του μοντέλου. Τα δεδομένα εκπαίδευσης αποτελούσαν το 80% των συνολικών δεδομένων ενώ τα δεδομένα δοκιμής, πάνω στα οποία έγινε η πρόβλεψη, το υπόλοιπο 20%.

Έπειτα, δημιουργήθηκαν συγκεκριμένοι όροι οι οποίοι έπαιζαν το ρόλο του "κριτή", καθώς εκείνοι ήταν υπεύθυνοι για την αξιολόγηση του μοντέλου και οι τιμές τους βοήθησαν στην εξέλιξή του. Οι όροι αυτοί αναλύθηκαν στο Κεφάλαιο 4.7 και πρέπει να επισημανθεί πως η εφαρμογή τους έγινε τόσο στα δεδομένα εκπαίδευσης, κατά τη διάρκεια δηλαδή της εκπαίδευσης του μοντέλου, όσο και στα δεδομένα δοκιμής όπου οι τιμές παρουσιάζονται στους πίνακες του Κεφαλαίου 5 για όλες τις περιπτώσεις των μοντέλων που αναπτύχθηκαν.

Εκτός από αυτά, πρέπει να σημειωθεί πως είναι κρίσιμο για την έρευνα τα άτομα που συμμετέχουν στην έρευνα να είναι αξιόπιστα. Στην παρούσα εργασία αφαιρέθηκε 1 χρήστης από τους 27 αφού οι προβλέψεις του τρόπου μεταφοράς ήταν λανθασμένες και επιβάρυναν το μοντέλο. Αυτό πιθανώς οφείλεται στην λάθος επικύρωση που έκανε το συγκεκριμένο άτομο τη δεδομένη χρονική στιγμή. Παρόλα αυτά ο λόγος δεν είναι αποκλειστικά αυτός, καθώς ο αριθμός των ταξιδιών που ήταν διαθέσιμος για την παρούσα εργασία ήταν περιορισμένος και επομένως κρίνεται αναγκαία η καταγραφή επιπλέον ταξιδιών του συγκεκριμένου χρήστη και την μετέπειτα εφαρμογή τους στο μοντέλο που αναπτύχθηκε.

Τέλος, αξιοσημείωτη είναι η σημαντικότητα της συνεχόμενης δοκιμής και προσθήκης νέων μεταβλητών στο μοντέλο, αφού μπορεί κάποια από αυτές ή ο συνδυασμός αυτών να οδηγήσει σημαντικά στην αύξηση της αποτελεσματικότητάς του. Αυτό φαίνεται και σε αυτή την εργασία, αφού η προσθήκη μιας νέας μεταβλητής βελτίωσε σημαντικά την απόδοση του μοντέλου και συγκεκριμένα αύξησε τα ποσοστά ακριβείας του περίπου κατά 8%.

Στη συνέχεια, αναφέρονται τα βασικά συμπεράσματα της ανάλυσης που πραγματοποιήθηκε ενώ δίνονται προτάσεις για περαιτέρω έρευνα πάνω στο θέμα της παρούσας διπλωματικής εργασίας.

6.2. ΒΑΣΙΚΑ ΣΥΜΠΕΡΑΣΜΑΤΑ ΑΝΑΛΥΣΗΣ

Με την συνεχή εξέλιξη της τεχνολογίας, τα κινητά τηλέφωνα πλέον παίζουν πολύ σημαντικό ρόλο στην καθημερινότητα των ανθρώπων και δύσκολα τα αποχωρίζεται κανείς. Αυτό και σε συνδυασμό με τους αισθητήρες με τους οποίους είναι πλέον εξοπλισμένα, τα καθιστούν χρήσιμα στη συλλογή δεδομένων. Με αυτόν τον τρόπο, συλλέγονται περισσότερα δεδομένα και αποτελεσματικότερα, χωρίς δυσκολίες και επιβαρύνσεις με επιπλέον συσκευές από τον χρήστη.

Φαίνεται πως η χρήση των χρονικών παραθύρων είναι απαραίτητη, γιατί τα δεδομένα ομαλοποιούνται και μπορούν από αυτά να εξαχθούν νέες μεταβλητές που θα κρίνουν την αξιοπιστία του μοντέλου, χωρίς να χάνεται η χρονική εξέλιξη και η δομή τους στο χρόνο που αποτελεί ένα δομικό χαρακτηριστικό που τα διαφοροποιεί μεταξύ των μέσων μεταφοράς. Στην παρούσα έρευνα έγινε χρήση των παραθύρων αυτών και μετά από πολλές δοκιμές καθορίστηκε η διάρκειά τους καθώς και η χρονική διαφορά μεταξύ δύο διαδοχικών παραθύρων. Ωστόσο διαπιστώθηκε πως η μεταβολή των χρονικών παραθύρων δεν είχε πολύ μεγάλες μεταβολές στην ακρίβεια του μοντέλου, καθώς τα ποσοστά του παρέμεναν υψηλά. Παρόλα αυτά, σε αυτή την έρευνα χρησιμοποιήθηκαν παράθυρα 8 δευτερολέπτων με χρονική διαφορά ενός δευτερολέπτου καθώς αυτά παρήγαγαν τα καλύτερα αποτελέσματα.

Επίσης, πρέπει να αναφερθεί πως είναι πολύ σημαντικό τα άτομα που θα συμμετέχουν στη συλλογή των δεδομένων για την διεξαγωγή της έρευνας να είναι αξιόπιστα και συνεπή, καθώς οποιαδήποτε αδιαφορία ή μη ενασχόλησή τους με τις απαραίτητες ενέργειες επικύρωσης του ταξιδιού τους, μπορεί να οδηγήσει σε εσφαλμένα συμπεράσματα της έρευνας.

Στη συνέχεια, φάνηκε πως η διασπορά των τιμών στα χρονικά παράθυρα δε βοηθάει το μοντέλο στο σωστό εντοπισμό του μέσου μεταφοράς ενώ η τυπική απόκλιση βοηθάει ελάχιστα. Αντιθέτως, οι υπόλοιπες τιμές που υπολογίστηκαν από τα χρονικά παράθυρα ήταν καθοριστικές για την αποτελεσματικότητα και την ακρίβεια του μοντέλου. Οι τιμές της περιστροφής του τηλεφώνου ως προς τον κατακόρυφο άξονα που ενώνει την κορυφή με τη βάση του, καθώς και της περιστροφής στον άξονα κάθετα σε αυτόν (δηλαδή οι τιμές των Pitch και Roll αντίστοιχα) βοήθησαν περισσότερο το μοντέλο ενώ οι τιμές του επιταχυνσιομέτρου ήταν εξίσου σημαντικές. Το γυροσκόπιο επίσης βοήθησε στη βελτίωση του μοντέλου αλλά όχι στον ίδιο βαθμό με τα υπόλοιπα. Παρόλα αυτά συμμετέχει ενεργά στη δημιουργία του μοντέλου. Τελικά, προκύπτει πως η θέση του κινητού μέσα στο μέσο με το οποίο μεταφέρεται

ο χρήστης παίζει σημαντικό ρόλο στην ακριβή πρόβλεψη του μέσου αυτού και σε αυτό ευθύνονται οι τιμές τον Pitch και Roll.

Επιπλέον, η προσθήκη μιας μεταβλητής αύξησε την ακρίβεια των προβλέψεων κατά 8%. Η μεταβλητή αυτή είναι στην ουσία ο κωδικός των χρηστών, με τον οποίο μπορεί το μοντέλο να αναγνωρίζει ποιανού είναι το κάθε ταξίδι και να μαθαίνει τις ιδιαιτερότητες των κινήσεων και τα μοτίβα που ακολουθεί το κάθε άτομο κατά τη διάρκειά του ταξιδιού του. Η συγκεκριμένη μεταβλητή οφείλεται για τα πολύ υψηλά ποσοστά ακρίβειας στις προβλέψεις του τελικού μοντέλου. Ωστόσο, δεν έχει εφαρμοστεί σε μεγάλο αριθμό χρηστών και παραμένει άγνωστη η λειτουργία του όταν οι χρήστες αυξηθούν.

Τέλος, πρέπει να τονιστεί ότι η ακρίβεια των αποτελεσμάτων ήταν πολύ υψηλή από τις πρώτες κιόλας δοκιμές χάρη στο πακέτο h2o που χρησιμοποιήθηκε, ενώ πέρα από το τυχαίο δάσος το οποίο εφαρμόστηκε στην παρούσα εργασία, διαθέτει πολλούς ακόμα αλγορίθμους μοντέλων μηχανικής μάθησης. Επιπλέον, αξιοσημείωτη είναι η ευκολία χρήσης της γλώσσας προγραμματισμού R λόγω της μεγάλης κοινότητας που διαθέτει με αποτέλεσμα να υπάρχουν πολλές διευκρινήσεις ως προς τον τρόπο λειτουργίας της στο διαδίκτυο.

6.3. ΠΡΟΤΑΣΕΙΣ ΓΙΑ ΠΕΡΑΙΤΕΡΩ ΕΡΕΥΝΑ

Παρά τα υψηλά ποσοστά ακρίβειας των προβλέψεων που επιτυγχάνονται σε αυτή την έρευνα, υπάρχουν ακόμα περιθώρια βελτίωσης του μοντέλου. Αρχικά, απαιτείται να γίνει ανάλυση σε περισσότερα δεδομένα και παραπάνω χρήστες ώστε να εξεταστούν και άλλες περιπτώσεις ατόμων. Ο τρόπος με τον οποίο χρησιμοποιείται το κινητό διαφέρει από άτομο σε άτομο με αποτέλεσμα να υπάρχουν χιλιάδες επιπλέον πιθανές χρήσεις του κινητού κατά τη διάρκεια ενός ταξιδιού. Επίσης, υπήρχε μια δυσαναλογία στα δεδομένα καθώς αυτά που πραγματοποιήθηκαν με αυτοκίνητο ήταν 3πλάσια σε αριθμό από εκείνα που πραγματοποιήθηκαν με τη χρήση των ΜΜΜ. Επομένως, πρέπει να γίνει μελέτη με δεδομένα στα οποία τα θα έχουν ίσο αριθμό ταξιδιών με τη χρήση και των δύο κατηγοριών ώστε να εκπαιδευτεί το μοντέλο ορθότερα.

Επιπλέον, πρέπει να γίνει περαιτέρω ανάλυση και εκπαίδευση του μοντέλου και με δεδομένα από αισθητήρα μαγνητόμετρου, καθώς αν και στην παρούσα εργασία φαίνεται πως δεν προσέφεραν αύξηση στις προβλέψεις του μοντέλου, τα δεδομένα που συλλέχθηκαν από αυτόν τον αισθητήρα ήταν ελάχιστα και η εκπαίδευση και δοκιμή του μοντέλου έγινε σε πολύ μικρό αριθμό ταξιδιών.

Πιο συγκεκριμένοι τρόποι μεταφοράς μπορούν να αναλυθούν εφόσον έχουν συλλεχθεί δεδομένα από κάθε ένα μέσο ξεχωριστά, δηλαδή να μην υπάρχει μια κατηγορία που περιλαμβάνει όλα τα ταξίδια με την χρήση των Μέσων Μαζικής Μεταφοράς, αλλά κάθε κατηγορία για κάθε ένα μέσο ξεχωριστά όπως είναι το τρένο, ο ηλεκτρικός, το λεωφορείο. Έτσι, το μοντέλο που θα δημιουργηθεί θα είναι πιο εξειδικευμένο και νέες έρευνες θα μπορούν να πραγματοποιηθούν μελετώντας τις επιπτώσεις του κάθε μέσου στο οδικό δίκτυο.

Εκτός από τα παραπάνω, προτείνεται η χρήση διαφορετικών μοντέλων μηχανικής μάθησης, αφού στην παρούσα εργασία χρησιμοποιήθηκε και αναλύθηκε μόνο το μοντέλο των τυχαίων δασών. Παρόλο που η βιβλιογραφία έχει δείξει ότι αυτή η μέθοδος είναι η πιο αποτελεσματική για τον εντοπισμό του τρόπου μεταφοράς, δεν παύουν να υπάρχουν εξαιρέσεις σε κάποιες έρευνες και αυτό εξαρτάται από το είδος των δεδομένων που έχουν συλλεχθεί και την επεξεργασία που έχουν αυτά υποστεί.

Τέλος, η χρήση επιπλέον μεταβλητών που θα ομαδοποιούν τα δεδομένα είναι επιτακτική ανάγκη, καθώς η αποκλειστική χρήση της αρίθμησης των χρηστών που εφαρμόστηκε στο

μοντέλο της παρούσας διπλωματικής μπορεί να προκαλέσει πρόβλημα όταν νέοι χρήστες εισέρχονται στις δοκιμές, ή όταν ο αριθμός των χρηστών γίνει πολύ μεγάλος. Η προσθήκη μεταβλητής που θα καθορίζει το φύλο των χρηστών ή η χρησιμοποίηση μεταβλητών που θα κατηγοριοποιούν τους χρήστες ανάλογα με την ηλικία τους, μπορεί να βοηθήσουν το μοντέλο αισθητά χωρίς απόλυτα να το ιδανικεύουν.

7. ΒΙΒΛΙΟΓΡΑΦΙΑ

Φραντζεσκάκης, Ι. Μ., Γκόλιας, Ι. Κ., & Πιτσιάβα-Λατινοπούλου, Μ. Χ. (2009). *Κυκλοφοριακή Τεχνική.* Αθήνα: Παπασωτηρίου.

Abdulazim, T., Abdelgawad, H., Habib, K. N., & Abdulhai, B. (2013, December). Using Smartphones and Sensor Technologies to Automate Collection of Travel Data. *Transportation Research Record Journal of the Transportation Research Board*, 44-52.

Accelerometer. (n.d.). Ανάκτηση από wikipedia: https://en.wikipedia.org/wiki/Accelerometer

Android, Google. (n.d.). *SensorEvent*. Ανάκτηση από Developers Android: https://developer.android.com/reference/android/hardware/SensorEvent.html

Asakura, Y., & Hato, E. (2004, June-August). Tracking survey for individual travel behaviour using mobile communication instruments. *Transportation Research Part C: Emerging Technologies, 12*(3-4), 273-291.

Asakura, Y., Hato, E., Nishibe, Y., Daito, T., Tanabe, J., & Koshima, H. (1999). Monitoring travel behavior using PHS based location positioning service system. *Proceedings of 6th World Congress on Intelligent Transport Systems (ITS).* Toronto, ON, Canada.

Benyamin, D. (2012). *A Gentle Introduction to Random Forests, Ensembles, and Performance Metrics in a Commercial System*. Ανάκτηση από blog.citizennet.com: http://blog.citizennet.com/blog/2012/11/10/random-forests-ensembles-and-performance-metrics

Bohte, W., & Maat, K. (2009, June). Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. *Transportation Research Part C: Emerging Technologies, 17*(3), 285-297.

Bolbol, A., Cheng, T., Tsapakis, I., & Haworth, J. (2012, November). Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification. *Computers, Environment and Urban Systems, 36*(6), 526-537.

Breiman, Friedman, Olshen, & Stone. (1984). *CLASSIFICATION AND REGRESSION TREES.*

Breiman, L., & Cutler, A. (n.d.). *Random Forests*. Ανάκτηση από stat.berkeley.edu: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

Chen, C., Gong, H., Lawson, C., & Bialostozky, E. (2010, December). Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study. *Transportation Research Part A: Policy and Practice, 44*(10), 830-840.

Chung, E. H., & Shalaby, A. (2005). A Trip Reconstruction Tool for GPS-based Personal Travel Surveys. *Transportation Planning and Technology*, (σσ. 381-401).

Deshpande, B. (2011). *2 main differences between classification and regression trees*. Ανάκτηση από simafore: http://www.simafore.com/blog/bid/62482/2-main-differences-between-classification-and-regression-trees

Eftekhari, H. R., & Ghatee, M. (2016, August). An inference engine for smartphones to preprocess data and detect stationary and transportation modes. *Transportation Research Part C: Emerging Technologies, 69*, 313-327.

Ellis, K., Godbole, S., Marshall, S., Lanckriet, G., Staudenmayer, J., & Kerr, J. (2014). Identifying Active Travel Behaviors in Challenging Environments Using GPS, Accelerometers, and Machine Learning Algorithms . *Frontiers in Public Health , 2*, 2-36.

Feng, T., & Timmermans, H. J. (2013, December). Transportation mode recognition using GPS and accelerometer data. *Transportation Research Part C: Emerging Technologies, 37*, 118-130.

Ferrer, S., & Ruiz, T. (2014, December 19). Travel behavior characterization using raw accelerometer data collected from smartphones. *Procedia - Social and Behavioral Sciences, 160*, 140-149.

Figo, D., Diniz, P. C., Ferreira, D. R., & Cardoso, J. M. (2010, October). Preprocessing techniques for context recognition from accelerometer data. *Personal and Ubiquitous Computing, 14*(7), 645-662.

Froehlich, J., Dillahunt, T., Klasnja, P., Mankoff, J., Consolvo, S., Harrison, B., & Landay, J. A. (2009). UbiGreen: investigating a mobile tool for tracking and supporting green transportation habits. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, (σσ. 1043-1052). Boston, MA, USA.

Geurs, K. T., Thomas, T., Bijsma, M., & Douhou, S. (2015). Automatic trip and mode detection with MoveSmarter: first results from the Dutch Mobile Mobility Panel. *Transportation Research Procedia, 11*, 247-262.

Gong, H., Chen, C., Bialostozky, E., & Lawson, C. T. (2012, March). A GPS/GIS method for travel mode detection in New York City . *Computers, Environment and Urban Systems, 36*(2), 131-139.

Gong, L., Morikawa, T., Yamamoto, T., & Sato, H. (2014, July 14). Deriving Personal Trip Data from GPS Data: A Literature Review on the Existing Methodologies. *Procedia - Social and Behavioral Sciences, 138*, 557-565.

Hato, E. (2006). Development of MoALs (Mobile Activity Loggers supported by gps-phones) for travel behavior analysis. *Transportation Research Board 85th Annual Meeting.* Washington, DC United States: Transportation Research Board.

Hato, E. (2010, February). Development of behavioral context addressable loggers in the shell for travel-activity analysis. *Transportation Research Part C: Emerging Technologies, 18*(1), 55-67.

Hemminki, S., Nurmi, P., & Tarkoma, S. (2013). Accelerometer-Based Transportation Mode Detection on Smartphones. *Proceedings of the 11th ACM Conference on Embedded Networked.* Roma, Italy.

Huss, A., Beekhuizen, J., Kromhout, H., & Vermeulen, R. (2014). Using GPS-derived speed patterns for recognition of transport modes in adults. *International Journal of Health Geographics*.

Itsubo, S., & Hato, E. (2006). Effectiveness of household travel survey using GPS-equipped cell phones and Web diary: Comparative study with paper-based travel survey. *Proceedings of the Transportation Research Board 85th Annual Meeting.* Washington, DC, USA.

Jean, W., Randall, G., & William, B. (2001). Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data. *Transportation Research Record, 1768*(1), 125-134.

Krygsman, S., Nel, J., & Jong, T. (2008). The use of cellphone technology in activity and travel data collection in developing countries. *Proceedings of the 18th International Conference on Transport Survey Methods.*

Lösch, R. (2017). *How to compute impurity using Gini Index?* Ανάκτηση από ResearchGate: https://www.researchgate.net/post/How_to_compute_impurity_using_Gini_Index

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A. L., Brewer, D., . . . Alstyne, M. V. (2009, February 6). Computational Social Science. *Science, 323*(5915), 721-723.

Maurer, U., Smailagic, A., Siewiorek, D., & Deisher, M. (2006). Activity recognition and monitoring using multiple sensors on different body positions. *International Workshop on Wearable and Implantable Body Sensor Networks.* Cambridge, MA, USA: IEEE.

McGowen, P., & McNally, M. (2007). Evaluating the Potential To Predict Activity Types from GPS and GIS Data. *Transportation Research Board 86th Annual Meeting.* Washington, DC, USA: Transportation Research Board.

Mitchell, T. M. (2009). Mining Our Reality. *Science, 326*(5960), 1644-1645.

Montini, L., Prost, S., Schrammel, J., Rieser-Schüssler, N., & Axhausen, K. W. (2015). Comparison of travel diaries generated from smartphone data and dedicated GPS devices. *Transportation Research Procedia, 11*, 227-241.

Murakami, E., & Wagner, D. P. (1999, April-June). Can using global positioning system (GPS) improve trip reporting? *Transportation Research Part C: Emerging Technologies, 7*(2-3), 149-165.

Nham, B., Siangliulue, K., & Yeung, S. (2012, April). *Predicting Mode of Transport from Iphone Accelerometer Data.* Stanford University, Stanford, CA, USA.

Nick, T., Coersmeier, E., Geldmacher, J., & Goetze, J. (2010). Classifying means of transportation using mobile sensor data. *The 2010 International Joint Conference on Neural Networks.* Barcelona, Spain: IEEE.

Nurmi, P., Bhattacharya, S., & Kukkonen, J. (2010). A grid-based algorithm for on-device GSM positioning. *Proceedings of the 12th ACM international conference on Ubiquitous computing*, (σσ. 227-236). Copenhagen, Denmark.

Oliver, M., Badland, H., Mavoa, S., Duncan, M. J., & Duncan, S. (2010). Combining GPS, GIS, and Accelerometry: Methodological Issues in the Assessment of Location and Intensity of Travel Behaviors. *Journal of physical activity & health*.

Overfitting. (n.d.). Ανάκτηση από wikipedia: https://en.wikipedia.org/wiki/Overfitting

Pruning. (n.d.). Ανάκτηση από wikipedia: https://en.wikipedia.org/wiki/Pruning_(decision_trees)

Reddy, S., Burke, J., Estrin, D., M.Hansen, & M.Srivastava. (2008). Determining transportation mode on mobile phones. *12th IEEE International Symposium on Wearable Computers.* Pittsburgh, PA, USA.

Reddy, S., Mun, M., Burke, J., & Srivastava, M. B. (2010, February). Using Mobile Phones to Determine Transportation Modes. *ACM Transactions on Sensor Networks*.

Seemakurthi, P. (2016, June). *What's the difference between boosting and bagging?* Ανάκτηση από quora: https://www.quora.com/Whats-the-difference-between-boosting-and-bagging

Sermons, M., & Koppelman, F. S. (1996, April). Use of vehicle positioning data for arterial incident detection. *Transportation Research Part C: Emerging Technologies, 4*(2), 87-96.

Shafique, M. A., & Hato, E. (2015). Modelling of Accelerometer Data for Travel Mode Detection by Hierarchical Application of Binomial Logistic Regression. *Transportation Research Procedia, 10*, 236-244.

Shafique, M. A., & Hato, E. (2015). Use of acceleration data for transportation mode prediction. *Transportation*, 163-188.

Shafique, M., & E.Hato. (2016, May 18). Travel Mode Detection with Varying Smartphone Data Collection Frequencies. *Sensors 2016*.

Shen, L., & Stopher, P. R. (2014). Review of GPS Travel Survey and GPS Data-Processing Methods. *Transport Reviews, 34*(3), 316-334.

Siirtola, P., & Röning, J. (2012, June). Recognizing Human Activities User-independently on Smartphones Based on Accelerometer Data. *International Journal of Interactive Multimedia and Artificial Intelligence*, 38-45.

Song, C., Qu, Z., Blumm, N., & Barabasi, A. L. (2010, February 19). Limits of Predictability in Human Mobility. *Science, 327*(5968), 1018-1021.

Soper, D. (2012, 4). Is human mobility tracking a good idea? *Communications of the ACM, 55*(4), 35-37.

sparkfun. (n.d.). *sparkfun*. Ανάκτηση από sparkfun: https://learn.sparkfun.com/tutorials/gyroscope/how-a-gyro-works

Stenneth, L., Wolfson, O., Yu, P. S., & Xu, B. (2011). Transportation Mode Detection using Mobile Phones and GIS Information. *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, (σσ. 54-63). Chicago, Illinois.

Stopher, P. (2009). The Travel Survey Toolkit: Where to From Here. *Transport Survey Methods: Keeping Up with a Changing World*, 15-46.

Stopher, P., FitzGerald, C., & Zhang, J. (2008, June). Search for a global positioning system device to measure person travel. *Transportation Research Part C: Emerging Technologies, 16*(3), 350-369.

Stopher, P., FitzGerald, C., & Zhang, J. (2008, June). Search for a global positioning system device to measure person travel. *16*(3), 350-369.

Su, X., Caceres, H., Tong, H., & He, Q. (2015). Travel Mode Identification with Smartphones. *TRB 94th Annual Meeting for Presentation.* Washington D.C.

Sugino, K., Yano, S., Hato, E., & Asakura, Y. (2005). Empirical analysis of sightseeing behaviour using probe person survey data. *Proceedings of the Infrastructure Planning.* Miyazaki, Japan.

Team, A. V. (n.d.). *A Complete Tutorial on Tree Based Modeling from Scratch*. Ανάκτηση από analyticsvidhya: https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/

The Accelerometer, our everyday sensor. (n.d.). Ανάκτηση από IMALAB: http://www.imalab.net/news/accelerometer-our-everyday-sensor

Tsui, S., & Shalaby, A. S. (2006, January). Enhanced System for Link and Mode Identification for Personal Travel Surveys Based on Global Positioning Systems. *Transportation Research Record Journal of the Transportation Research Board*, 38-45.

Wagner, D. (1997). *Global Positioning Systems for Personal Travel Surveys: Lexington Area Travel Data Collection Test* . Columbus, OH, USA: Battelle Transportation Division.

Wang, S., Chen, C., & Ma, J. (2010). Accelerometer based transportation mode recognition on mobile phones. *Asia-Pacific Conference on Wearable Computing Systems.* Shenzhen, China, China.

Wermuth, M., Sommer, C., & Kreitz, M. (2003). Impact of New Technologies in Travel Surveys. *Transport Survey Quality and Innovation*, 455-481.

wikipedia. (n.d.). *Aircraft principal axes*. Ανάκτηση από wikipedia: https://en.wikipedia.org/wiki/Aircraft_principal_axes

Wu, L., Yang, B., & Jing, P. (2016). Travel Mode Detection Based on GPS Raw Data Collected by Smartphones: A Systematic Review of the Existing Methodologies. *Information 2016*.

Yang, J., & Nokia, C. R. (2009). Toward Physical Activity Diary: Motion Recognition Using Simple Acceleration Features with Mobile Phones. *IMCE '09 Proceedings of the 1st international workshop on Interactive multimedia for consumer electronics*, (σσ. 1-10). Beijing, China.

Yatsumoto, H., Kitazawa, T., Nakagawa, S., Okamoto, A., & Asakura, Y. (2006). Analysis of route choice behavior under flexible toll system of urban expressway based on probe person trip survey. *Proceedings of the 33rd Meeting of Infrastructure Planning.* Sendai, Japan.

Zahra, A. L., & Amir, G. (2015). Automated Transportation Mode Detection Using Smart Phone Applications via Machine Learning: Case Study Mega City of Tehran. *Transportation Research Board 94th Annual Meeting.* Washington, DC United States.

Zhao, F., A.Ghorpade, F.C.Pereira, C.Zegras, & M.Ben-Akiva. (2015). Stop Detection in Smartphone-based Travel Surveys. *Transportation Research Procedia, 11*, 218-226.

Zheng, Y., Liu, L., Wang, L., & Xie, X. (2008). Learning Transportation Mode from Raw GPS Data for Geographic Applications on the Web. *Proceedings of the 17th international conference on World Wide Web*, (σσ. 247-256). Beijing, China.

Zheng, Y., Liu, Y., Yuan, J., & Xie, X. (2011). Urban computing with taxicabs. *Proceedings of the 13th international conference on Ubiquitous computing*, (σσ. 89-98). Beijing, China.

Zito, R., D'Este, G., & Taylor, M. (1995, August). Global positioning systems in the time domain: How useful a tool for intelligent vehicle-highway systems? *Transportation Research Part C: Emerging Technologies*, *3*(4), 193-209.

Zong, F., Bai, Y., Wang, X., Yuan, Y., & He, Y. (2015). Identifying Travel Mode with GPS Data Using Support Vector Machines and Genetic Algorithm. *Information*, 212-227.