



Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Πολιτικών Μηχανικών
Τομέας Μεταφορών και Συγκοινωνιακής Υποδομής

**Ανάλυση Συμπεριφοράς Οδηγού με Χρήση
Δεδομένων από Διασυνδεδεμένα Οχήματα**
Διπλωματική Εργασία



Εγγλέζος Ανδρέας Γεώργιος

Επιβλέπων: Γιώργος Γιαννής, Καθηγητής ΕΜΠ

Ευχαριστίες

Με την παρούσα Διπλωματική Εργασία ολοκληρώνεται ο κύκλος σπουδών μου στη σχολή Πολιτικών Μηχανικών του Ε.Μ.Π.

Θα ήθελα να ευχαριστήσω θερμά τον κ. Γιώργο Γιαννή, Καθηγητή της Σχολής Πολιτικών Μηχανικών του Ε.Μ.Π., για την ανάθεση της παρούσας Διπλωματικής Εργασίας, την υποστήριξη και την πολύτιμη καθοδήγησή του σε όλα τα στάδια εκπόνησής της, καθώς και την εξαιρετική συνεργασία μας.

Επιπλέον οφείλω να ευχαριστήσω εξίσου την κα. Αρμίρα Κονταξή, Διδάκτωρ του Ε.Μ.Π., για τις συμβουλές και τις παραγωγικές υποδείξεις της πάνω σε σημαντικά θέματα της Διπλωματικής Εργασίας, καθώς και το εξαιρετικό κλίμα συνεργασίας που διαμόρφωσε.

Επίσης, ιδιαίτερες ευχαριστίες ανήκουν και στην εταιρεία τηλεματικής OSeven, η οποία παραχώρησε τα δεδομένα που χρησιμοποιήθηκαν στην παρούσα Διπλωματική Εργασία.

Τέλος, θέλω να ευχαριστήσω την οικογένεια και τους φίλους μου για την πλήρη στήριξη καθ' όλη την διάρκεια των ακαδημαϊκών μου χρόνων.

Αθήνα, Νοέμβριος 2025

Εγγλέζος Ανδρέας Γεώργιος

Ανάλυση Συμπεριφοράς Οδηγού με Χρήση Δεδομένων από Διασυνδεδεμένα Οχήματα

Εγγλέζος Ανδρέας Γεώργιος

Επιβλέπων: Γιώργος Γιαννής, Καθηγητής ΕΜΠ

ΣΥΝΟΨΗ

Στόχος της παρούσας Διπλωματικής Εργασίας είναι η διερεύνηση της οδηγικής συμπεριφοράς μέσω της ανάλυσης δεδομένων που συλλέχθηκαν από διασυνδεδεμένα οχήματα. Για την υλοποίηση της έρευνας εξετάστηκαν μεταβλητές που σχετίζονται με την ταχύτητα, τη θερμοκρασία κινητήρα, το σύστημα αντιμπλοκαρίσματος τροχών και άλλα χαρακτηριστικά της οδήγησης, που συλλέχθηκαν εντός διαστήματος τριών μηνών, προκειμένου να εντοπιστούν διαφορετικά προφίλ διαδρομών ως προς την συμπεριφορά του οδηγού. Εφαρμόστηκε η μέθοδος ταξινόμησης K-Means για τη διάκριση προτύπων οδηγικής συμπεριφοράς και διαπιστώθηκε ότι η ταξινόμηση των διαδρομών σε τρεις ομάδες παρέχει ικανοποιητική ανάλυση στα διαγράμματα και στην αξιολόγηση τους. Έπειτα εφαρμόστηκε ο αλγόριθμος Random Forest με εξαρτημένη μεταβλητή το σύστημα αντιμπλοκαρίσματος των τροχών, με σκοπό τον υπολογισμό της σημαντικότητας των ανεξάρτητων μεταβλητών. Από την εφαρμογή του αλγορίθμου προέκυψε ότι η μεταβλητή με την μεγαλύτερη τιμή σημαντικότητας είναι η θερμοκρασία λαδιού του κινητήρα. Τέλος, εφαρμόστηκε η Διωνυμική Λογιστική Παλινδρόμηση για τη διερεύνηση του βαθμού επιρροής των ανεξάρτητων μεταβλητών στην πιθανότητα ενεργοποίησης του συστήματος αντιμπλοκαρίσματος των τροχών, από την οποία προέκυψε ότι η μεταβλητή με το μεγαλύτερο βαθμό επιρροής είναι η ενεργοποίηση του κινητήρα.

Λέξεις κλειδιά: Οδηγική συμπεριφορά, επεξεργασία δεδομένων ,διασυνδεδεμένα οχήματα , K-Means, Random Forest, Διωνυμική Λογιστική Παλινδρόμηση

Driving Behavior Analysis

Using Connected Vehicle Data

Englezos Andreas Georgios

Supervisor: George Yannis, Professor, NTUA

ABSTRACT

The aim of this Thesis is to investigate driving behavior through the analysis of data collected from connected vehicles. For the purposes of this research, variables related to speed, engine temperature, the anti-lock braking system (ABS), and other driving characteristics were examined. These data were collected over a three-month period to identify different route profiles in terms of driver behavior. The K-Means clustering method was applied to distinguish patterns of driving behavior, and it was found that the classification of trips into three groups provided a satisfactory analytical results. Following, the Random Forest algorithm was applied, using the anti-lock braking system as the dependent variable, with the aim of determining the importance of the independent variables. From the application of the algorithm, it was found that the variable with the highest importance value was the engine oil temperature. Finally, a Binary Logistic Regression was performed to examine the extent to which the independent variables influence the probability of ABS activation, which showed that the variable with the greatest influence was engine activation.

Keywords: Driving Behavior, Data Analysis, Connected Vehicles, K-Means, Random Forest, Binary Logistic Regression

Περίληψη

Οι οδικές μεταφορές αποτελούν κύριο πυλώνα στην ανάπτυξη και στην εξέλιξη του κοινωνικού συνόλου. Η οδηγική συμπεριφορά είναι το κέντρο του ενδιαφέροντος αφού είναι άμεσα συνδεδεμένη με την οδική ασφάλεια. Η παρούσα Διπλωματική Εργασία επικεντρώνεται στη **διερεύνηση της οδηγικής συμπεριφοράς μέσω της ανάλυσης δεδομένων που συλλέχθηκαν από διασυνδεδεμένα οχήματα**. Πιο συγκεκριμένα, θα εξεταστούν μεταβλητές που σχετίζονται με χαρακτηριστικά της οδήγησης, προκειμένου να εντοπιστούν διαφορετικά προφίλ διαδρομών ως προς την συμπεριφορά του οδηγού.

Αρχικά πραγματοποιήθηκε η **βιβλιογραφική ανασκόπηση** (Κεφάλαιο 2), η οποία εστίασε σε έρευνες που αξιοποιούν δεδομένα από διασυνδεδεμένα οχήματα, συμβάλλοντας έμπρακτα στην περεταίρω ανάλυση της οδηγικής συμπεριφοράς και στην γενικότερη βελτίωση της οδικής ασφάλειας.

Στο κεφάλαιο 3, παρουσιάζονται **βασικές έννοιες στατιστικής ανάλυσης**. Πραγματοποιείται η επεξήγηση των κύριων μεθόδων ταξινόμησης και αλγορίθμων μηχανικής μάθησης που χρησιμοποιήθηκαν στην παρούσα Διπλωματική Εργασία. Έπειτα αναλύονται βασικές μαθηματικές κατανομές καθώς και κριτήρια αξιολόγησης και αποδοχής ενός μοντέλου.

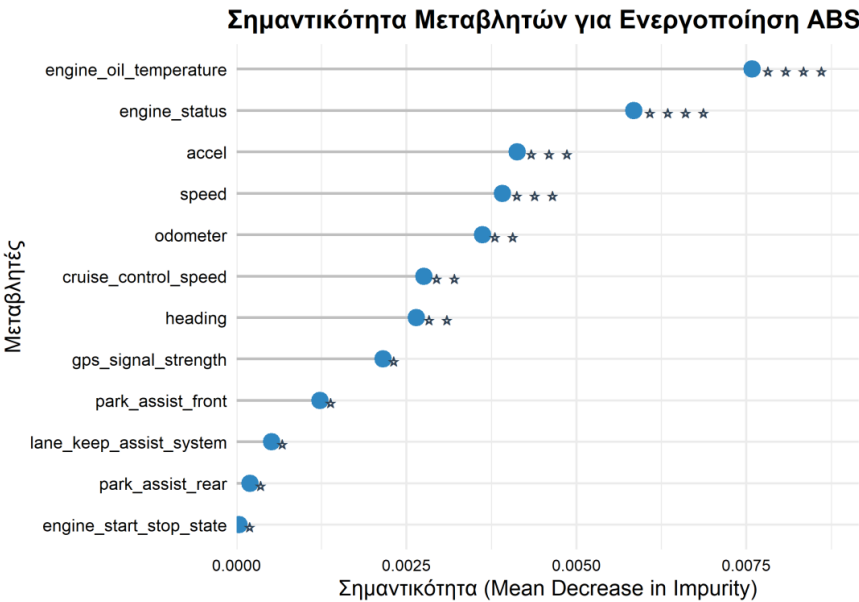
Στη συνέχεια, στο κεφάλαιο 4 επεξηγείται η διαδικασία της συλλογής και επεξεργασίας των δεδομένων για την υποστήριξη της μελέτης. Τα δεδομένα παραχωρήθηκαν από την εταιρεία OSeven και προέρχονται από διασυνδεδεμένα οχήματα. Η ανάλυση των χαρακτηριστικών εκτελείται με την χρήση του λογισμικού R, με δημιουργία πινάκων και απαραίτητων επεξηγηματικών σχολίων. Πραγματοποιείται μια **προκαταρκτική ανάλυση** των δεδομένων και εξάγονται χρήσιμα συμπεράσματα αναφορικά με το προφίλ οδήγησης των οχημάτων.

Τέλος, στο **κεφάλαιο της ανάλυσης των δεδομένων** εφαρμόστηκαν κατάλληλοι στατιστικοί μέθοδοι και μαθηματικά μοντέλα. Αρχικά, έγινε η ταξινόμηση των διαδρομών για κάθε όχημα βάσει κύριων χαρακτηριστικών όπως: η μέση και μέγιστη ταχύτητα, διακύμανση της ταχύτητας, σύστημα αντιμπλοκαρίσματος τροχών και απότομα φρεναρίσματα. Έπειτα, για την αξιολόγηση των μεταβλητών εφαρμόστηκε ο αλγόριθμος Random Forest θέτοντας ως εξαρτημένη μεταβλητή την ενεργοποίηση του συστήματος αντιμπλοκαρίσματος τροχών (ABS). Τέλος, εφαρμόστηκε η Διωνυμική Λογιστική Παλινδρόμηση για την διερεύνηση του βαθμού επιρροής των ανεξάρτητων μεταβλητών στην πιθανότητα ενεργοποίησης του συστήματος αντιμπλοκαρίσματος των τροχών

Παρακάτω παρατίθενται βασικοί πίνακες και διαγράμματα των αποτελεσμάτων από τις μεθόδους που χρησιμοποιήθηκαν:

Πίνακας Ι: Πίνακας μέσων τιμών των βασικών χαρακτηριστικών ανά συστάδα (Μέθοδος K-Means)

cluster	max_speed	sd_speed	avg_engine_temp	abs_per_100km	brakes_per_100km
1	104,94	18,28	70,97	18215	7
2	59,99	9,11	41,64	56627	36
3	155,03	35,11	86,34	7720	4



Διάγραμμα Ι: Διάγραμμα σημαντικότητας μεταβλητών (Μέθοδος Random Forest)

Πίνακας ΙΙ: Αποτελέσματα Βασικών Μετρικών Δεικτών (Μέθοδος Random Forest)

Random Forest	
Accuracy	0,981
Sensitivity	0,985
Specificity	0,654

Πίνακας ΙΙΙ:Πίνακας αποτελεσμάτων των μεταβλητών του μοντέλου (Διωνυμική Λογιστική Παλινδρόμηση)

term	estimate	p.value	Odds_Ratio	Σημαντικότητα
(Intercept)	2,526	<0,001	12,499	<0,001
engine_oil_temperature	0,001	0,204	1,001	-
engine_status (on)	1,300	<0,001	3,671	<0,001
accel	-0,049	0,199	0,952	-
speed	0,056	<0,001	1,058	<0,001
odometer	0,000	0,078	1,000	-
cruise_control_speed	0,002	0,007	1,002	0,007
heading_sin	-0,076	<0,001	0,927	<0,001
heading_cos	-0,101	<0,001	0,904	<0,001

Κατά την εκπόνηση της παρούσας Διπλωματικής Εργασίας προέκυψαν χρήσιμα συμπεράσματα που συνδέονται άμεσα με τους αρχικούς στόχους της. Τα κύρια συμπεράσματα συνοψίζονται ως εξής:

- Μέσω της μεθόδου Elbow και Silhouette διαπιστώθηκε ότι η ταξινόμηση των διαδρομών σε 3 ομάδες για την εφαρμογή του αλγόριθμου **K-Means** παρέχει ικανοποιητική ανάλυση στα διαγράμματα. Πιο συγκεκριμένα, εξετάζοντας τα διαγράμματα προέκυψε ότι μεγαλύτερη συχνότητα παρουσιάζει η συστάδα με ενδιάμεσες τιμές ταχυτήτων που υποδηλώνουν ένα τυπικό προφίλ οδήγησης. Επίσης η ομάδα που χαρακτηρίζεται από συντηρητική οδήγηση εμφανίζει τις υψηλότερες τιμές ενεργοποίησης του συστήματος αντιμπλοκαρίσματος των τροχών.
- Για την ανάλυση της σημαντικότητας των μεταβλητών ορίστηκε ως εξαρτημένη μεταβλητή το σύστημα αντιμπλοκαρίσματος των τροχών και μέσω του αλγόριθμου **Τυχαία Δάση** παρατηρήθηκε ότι οι μεταβλητές που αφορούν τη **θερμοκρασία του λαδιού του κινητήρα**, την **κατάσταση κινητήρα**, την **επιτάχυνση** και την **ταχύτητα** παρουσιάζουν, κατά σειρά, την **μεγαλύτερη συμβολή** στην πρόβλεψη της. Επίσης η ορθότητα (Accuracy) του μοντέλου κυμαίνεται σε υψηλά επίπεδα 98% συνεπώς είναι ικανοποιητικό και έχει ευαισθησία 98,5%, δηλαδή εντοπίζει εύκολα πότε το σύστημα αντιμπλοκαρίσματος των τροχών ενεργοποιείται. Τέλος η ειδικότητα παρουσίασε μέτριες προς υψηλές τιμές (65,4 %).
- Η εφαρμογή της **Διωνυμικής Λογιστικής Παλινδρόμησης** έδειξε ότι η ενεργοποίηση του συστήματος αντιμπλοκαρίσματος των τροχών δεν εξαρτάται μόνο από τις υπάρχουσες μεταβλητές αλλά και από άλλους μη μετρήσιμους παράγοντες (π.χ. οι καιρικές συνθήκες την περίοδο που συλλέχθηκαν τα δεδομένα ή μικρά συμβάντα κατά την διάρκεια των διαδρομών).
- Η ενεργοποίηση του κινητήρα αυξάνει την πιθανότητα ενεργοποίησης του συστήματος αντιμπλοκαρίσματος των τροχών. Ο ενεργός κινητήρας θέτει τα συστήματα ασφαλείας του οχήματος σε πλήρη ηλεκτρική λειτουργία. Συνεπώς υπάρχει αυξημένη πιθανότητα ενεργοποίησης του συστήματος αντιμπλοκαρίσματος των τροχών.
- Η αύξηση της ταχύτητας συμβάλλει στην πιθανότητα ενεργοποίησης του συστήματος αντιμπλοκαρίσματος των τροχών. Πιο συγκεκριμένα η αύξηση της ταχύτητας προκαλεί άνοδο στη κινητική ενέργεια του οχήματος, με αποτέλεσμα να υπάρχει μεγαλύτερη διάρκεια και δύναμη στην πέδηση, προκαλώντας μπλοκάρισμα στους τροχούς. Επίσης το σύστημα αυτό

ενεργοποιείται ακόμα και σε μικρής διάρκειας πέδησης όταν το όχημα κινείται με εξαιρετικά υψηλές ταχύτητες.

- Μικρή επίδραση έχει στο μοντέλο η μεταβλητή της ταχύτητας αυτόματου ελέγχου. Πιο συγκεκριμένα αύξηση της συγκεκριμένης ταχύτητας προκαλεί αύξηση της πιθανότητα ενεργοποίησης του συστήματος αντιμπλοκαρίσματος των τροχών. Η ταχύτητα αυτόματου ελέγχου δημιουργεί μια σταθερή και ελεγχόμενη οδήγηση που δεν δημιουργεί απότομες μεταβολές στην πέδηση που θα μπλοκάρουν τους τροχούς.
- Η μορφολογία του εδάφους και ο τύπος των οδών επηρεάζει στην πιθανότητα ενεργοποίησης του συστήματος. Πιο συγκεκριμένα αν οι διαδρομές πραγματοποιούνται σε οδούς με ανηφορική κλίση ή με μεγάλο ευθύγραμμο τμήμα μειώνεται η πιθανότητα ενεργοποίησης του συστήματος αντιμπλοκαρίσματος των τροχών.
- Συνολικά, διαπιστώθηκε ότι εξήχθησαν χρήσιμα συμπεράσματα αναφορικά με την οδηγική συμπεριφορά. Οι μέθοδοι ταξινόμησης και οι στατιστικές αναλύσεις βοήθησαν στην περαιτέρω μελέτη και κατανόηση του προφίλ οδήγησης μέσω της ανάλυσης δεδομένων από συνδεδεμένο όχημα.

Περιεχόμενα

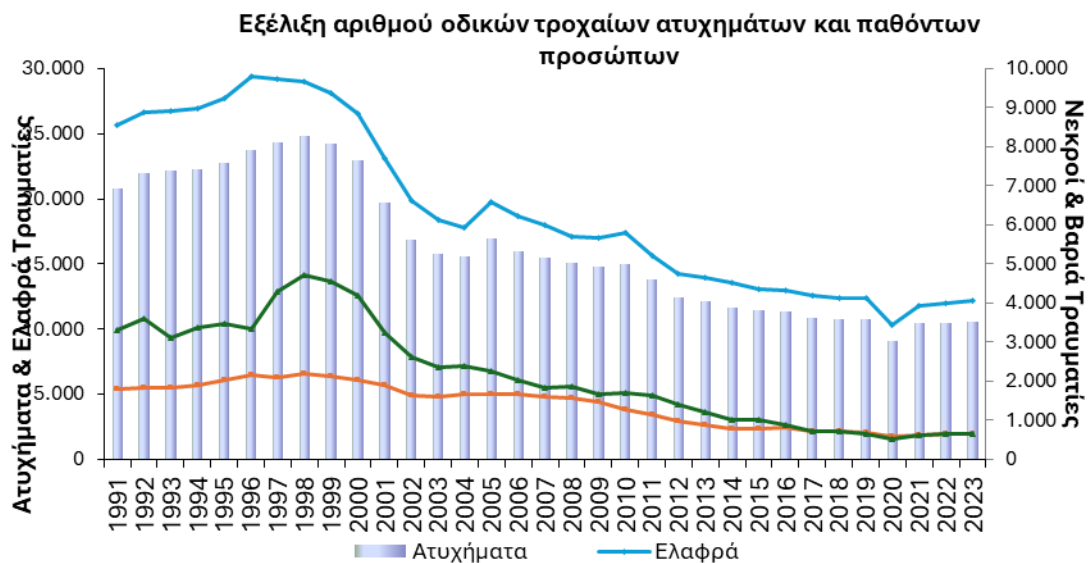
.....	1
1 Εισαγωγή.....	1
1.1 Γενική Ανασκόπηση	1
1.2 Στόχος Διπλωματικής Εργασίας	3
1.4 Δομή Διπλωματικής Εργασίας.....	5
2 Βιβλιογραφική Ανασκόπηση	7
2.1 Εισαγωγή	7
2.2 Διασυνδεδεμένα Οχήματα	7
2.3 Συνεργατικά Ευφυή Συστήματα Μεταφορών C-ITS.....	8
2.4 Συναφείς Έρευνες.....	9
2.4.1 Μεθοδολογία και Ταξινόμηση Ερευνών	9
2.4.2 Μέθοδοι Ανάλυσης	10
2.5 Σύνοψη Βιβλιογραφίας.....	14
3. Θεωρητικό Υπόβαθρο.....	15
3.1 Εισαγωγή	15
3.2 Βασικές Έννοιες Στατιστικής Ανάλυσης.....	15
3.3 Ερμηνεία Μαθηματικών Μοντέλων-Κατανομών.....	17
3.3.1 Κατανομή Poisson.....	17
3.3.2 Λογιστική Παλινδρόμηση (Logistic Regression).....	18
3.3.3 Διωνυμική Λογιστική Παλινδρόμηση.....	18
3.4 Κριτήρια Αποδοχής Μοντέλου	19
3.4.1 Ερμηνεία των Προσέμων και των Τιμών των Συντελεστών.....	19
3.4.2 Στατιστική Σημαντικότητα και Ελαστικότητα	19
3.4.3 Συντελεστής Προσαρμογής R^2 – Αξιολόγηση Ποιότητας	19
3.5 Αλγόριθμος K-Means.....	20
3.6 Αλγόριθμος Τυχαία Δάση (Random Forest)	21
4. Συλλογή και Επεξεργασία Στοιχείων.....	22
4.1 Εισαγωγή	22
4.2 OSeven και Συλλογή Δεδομένων	22
4.3 Επεξεργασία Δεδομένων	25
4.3.1 Εισαγωγή Βάσης Δεδομένων στο Λογισμικό Στατιστικής Ανάλυσης	26
4.3.2 Προκαταρκτική Ανάλυση.....	28
5. Εφαρμογή Μεθοδολογίας - Αποτελέσματα	32
5.1 Εφαρμογή Αλγόριθμου K-Means	32

5.1.1 Ποιότητα μοντέλου	34
5.1.2 Σχολιασμός Αποτελεσμάτων K-Means	35
5.2 Εφαρμογή Αλγόριθμου Random Forest	36
5.2.1 Εκπαίδευση του Αλγόριθμου Random Forest	36
5.2.2 Ποιότητα μοντέλου Random Forest	37
5.2.3 Σχολιασμός αποτελεσμάτων Random Forest	38
5.3 Εφαρμογή Διωνυμικής Λογιστικής Παλινδρόμησης	38
5.3.1 Στατιστικός Έλεγχος Μοντέλου	39
5.4 Αποτελέσματα	40
6. Συμπεράσματα.....	42
6.1 Σύνοψη Αποτελεσμάτων	42
6.2 Σύνοψη Συμπερασμάτων	44
6.3 Προτάσεις για Βελτίωση της Οδικής Ασφάλειας.....	45
6.4 Προτάσεις για Περαιτέρω Έρευνα	46
7. Βιβλιογραφικές Αναφορές.....	47

1 ΕΙΣΑΓΩΓΗ

1.1 ΓΕΝΙΚΗ ΑΝΑΣΚΟΠΗΣΗ

Στην σύγχρονη εποχή οι οδικές μεταφορές αποτελούν βασικό πυλώνα στην ανάπτυξη και στην εξέλιξη του κοινωνικού συνόλου. Η οδηγική συμπεριφορά είναι το κέντρο του ενδιαφέροντος αφού είναι άμεσα συνδεδεμένη με την οδική ασφάλεια. Πάρα το γεγονός ότι τα οδικά ατυχήματα στην Ευρώπη το 2023 παρουσίασαν μείωση 1% σε σχέση με το 2022, ο αριθμός των 20.000 νεκρών και πάνω από 100.000 με σοβαρούς τραυματισμούς αποτελεί ζήτημα ιδιαίτερης σημασίας. Η Ελλάδα το διάστημα 2007-2023 παρουσίασε πτωτική τάση στα θανατηφόρα ατυχήματα κατά 57,7%, στους σοβαρούς τραυματισμούς κατά 29,3% και συνολικά στα ατυχήματα κατά 31,9%. Παρόλο που παρουσιάζει μια αισιόδοξη εικόνα η καμπτική πορεία των ατυχημάτων, οι ηλικιακές τάξεις των παθόντων και η οικονομική επιβάρυνση αποτυπώνει πλήρως τη σοβαρότητα του ζητήματος. Επιπλέον για το έτος 2023 το ποσοστό των θανόντων ηλικίας 18-44 ανέρχεται σε 39,8% ενώ, λαμβάνοντας υπόψη και την ηλικιακή ομάδα 45-64 ο αριθμός εκτινάσσεται στο 66,9%. Αξίζει να σημειωθεί ότι για το έτος 2017 το κόστος των οδικών ατυχημάτων εκτιμήθηκε στα 2,7 δισεκατομμύρια (~ 1,5% του ΑΕΠ).



Διάγραμμα 1.1 Εξέλιξη αριθμού οδικών ατυχημάτων και παθόντων την περίοδο 1991-2023

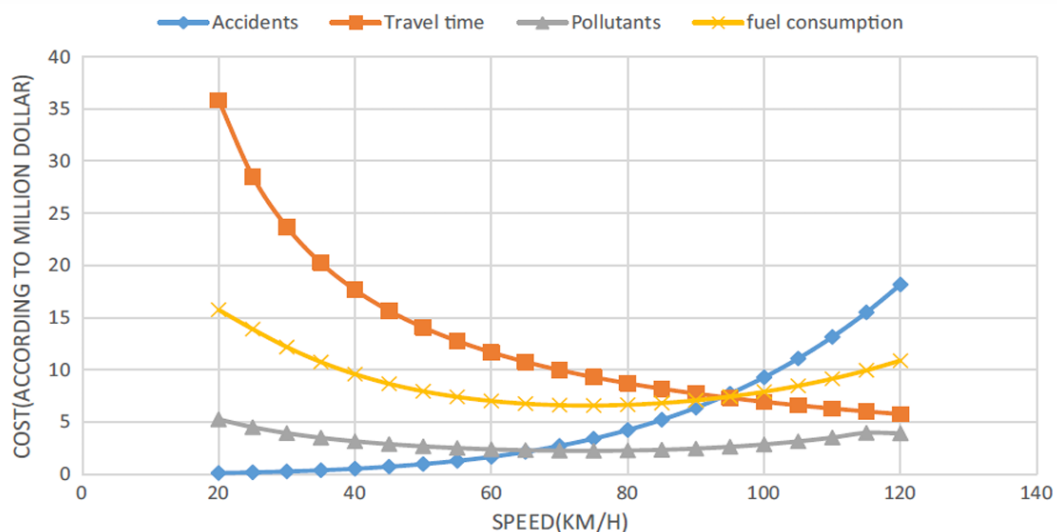
Πηγή: [ΕΛΣΤΑΤ \(Ελληνική Στατιστική Αρχή\)](#)

Οι κύριοι παράγοντες που επιδρούν στην εκδήλωση οδικών ατυχημάτων είναι: ο χρήστης της οδού, το όχημα και το οδικό περιβάλλον. Η συμπεριφορά του οδηγού αποτελεί την κύρια αιτία για ποσοστό 65-95%.

Στο πλαίσιο των φιλοδοξών στόχων της Ευρωπαϊκής Επιτροπής για τη μείωση των θανατηφόρων ατυχημάτων έως το 2030 κατά 50% μέχρι την πλήρη εξάλειψη τους το 2050, θεωρείται απαραίτητη η πλήρης κατανόηση των βασικών αιτίων των ατυχημάτων με κύριων αυτών να είναι:

- Υπερβολική ταχύτητα.
- Οδήγηση υπό την επήρεια αλκοόλ.
- Μη χρήση ζώνης.
- Μη χρήση κράνους.
- Απόσπασης προσοχής οδηγού.

Η ταχύτητα αποτελεί τον πυρήνα της ανάλυσης αφού επηρεάζει την πιθανότητα ατυχήματος αλλά και τον δείκτη σοβαρότητας τους. Χαρακτηριστικό είναι το γεγονός ότι η υπερβολική ταχύτητα ευθύνεται για το 30% των ατυχημάτων όπως επίσης ότι η μείωση της κατά 10 km/h μπορεί να οδηγήσει σε περιορισμό των θανατηφόρων συμβάντων κατά 50%.



Διάγραμμα 1.2 Η εκθετική επιρροή της ταχύτητας στο δείκτη οδικών ατυχημάτων

Πηγή: [NTUA Road Safety Driver Behavior](#)

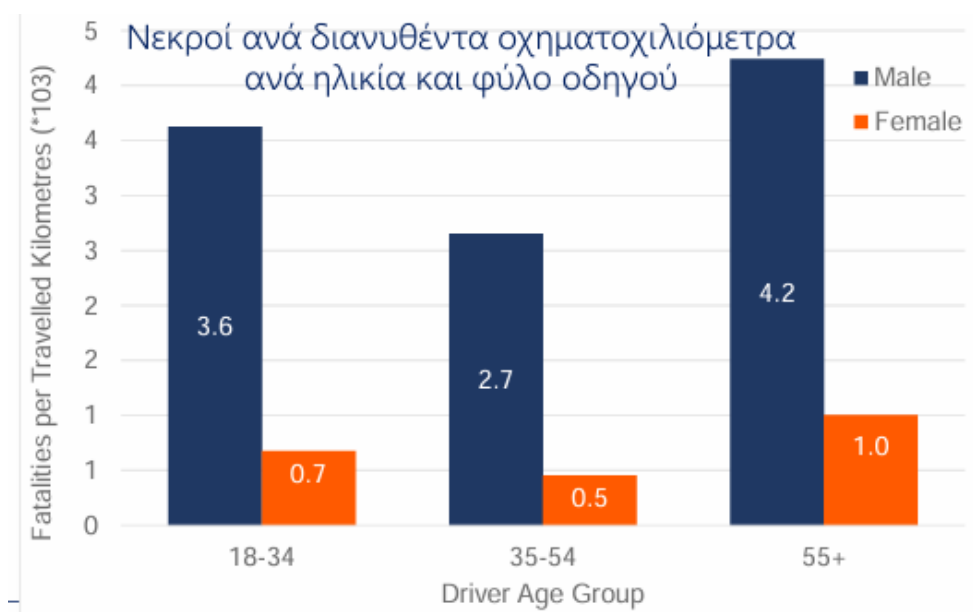
Σύμφωνα με το Βασικό Δείκτη Επίδοσης (Key Performance Indicator – (KPI)) το ποσοστό των οχημάτων που κινούνται στα προβλεπόμενα όρια ταχύτητας στο αστικό δίκτυο ανέρχεται στο 55,8% σε σχέση με τους αυτοκινητοδρόμους και τα υπεραστικά δίκτυα (76,9% και 84,0% αντίστοιχα).

Πίνακας 1.1 Βασικός Δείκτης επίδοσης (KPI)

Πηγή: [Baseline Project \(Μετρήσεις πεδίου 2022\)](#)

Τύπος Οδού	Ποσοστό οχημάτων εντός ορίων ταχύτητας			Μέση Ταχύτητα (χλμ/ώρα)	Τυπική Απόκλιση (χλμ/ώρα)	V85** (χλμ/ώρα)
	KPI	Κάτω όριο	Άνω όριο			
Αστικό δίκτυο	55,8%	55,0%	56,6%	48,4	9,3	57,0
Υπεραστικό δίκτυο*	84,0%	83,4%	84,6%	66,9	10,2	76,5
Αυτοκινητόδρομοι	76,9%	75,9%	77,8%	103,3	14,9	117,8

Αναλύοντας τους δείκτες επικινδυνότητας παρατηρείται ότι στα επιβατικά οχήματα ηλικίας άνω των 20 ετών, οι δείκτες είναι περίπου 4 φορές υψηλότεροι σε σύγκριση με τα νεότερα. Αναφορικά με τις ηλικιακές ομάδες, οι οδηγοί 35-54 ετών παρουσιάζουν χαμηλότερους δείκτες επικινδυνότητας σε σχέση με τους ηλικιωμένους και νέους. Τέλος οι άνδρες οδηγοί, με 4,6 φορές υψηλότερο δείκτη, παρουσιάζουν αποκλειστική υπευθυνότητα ή συνυπευθυνότητα στα ατυχήματα σε ποσοστό 63%.



Διάγραμμα 1.3: Νεκροί ανά διανυθέντα οχηματοχιλιόμετρα ανά ηλικία και φύλο οδηγού

Πηγή: [NTUA Road Safety Driver Behavior](#)

1.2 ΣΤΟΧΟΣ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Στόχος της παρούσας Διπλωματικής Εργασίας είναι η διερεύνηση της οδηγικής συμπεριφοράς μέσω της **ανάλυσης δεδομένων** που συλλέχθηκαν από **διασυνδεδεμένα οχήματα**. Πιο αναλυτικά, θα εξεταστούν μεταβλητές που σχετίζονται με την ταχύτητα, τη συχνότητα φρεναρισμάτων, τη θερμοκρασία κινητήρα, το σύστημα αντιμπλοκαρίσματος τροχών κλπ. και άλλα χαρακτηριστικά της οδήγησης, προκειμένου να εντοπιστούν διαφορετικά προφίλ διαδρομών ως προς

την συμπεριφορά του οδηγού, να εξεταστούν οι παράγοντες επιρροής της ενεργοποίησης συστημάτων ασφαλείας, καθώς και ο βαθμός επιρροής τους.

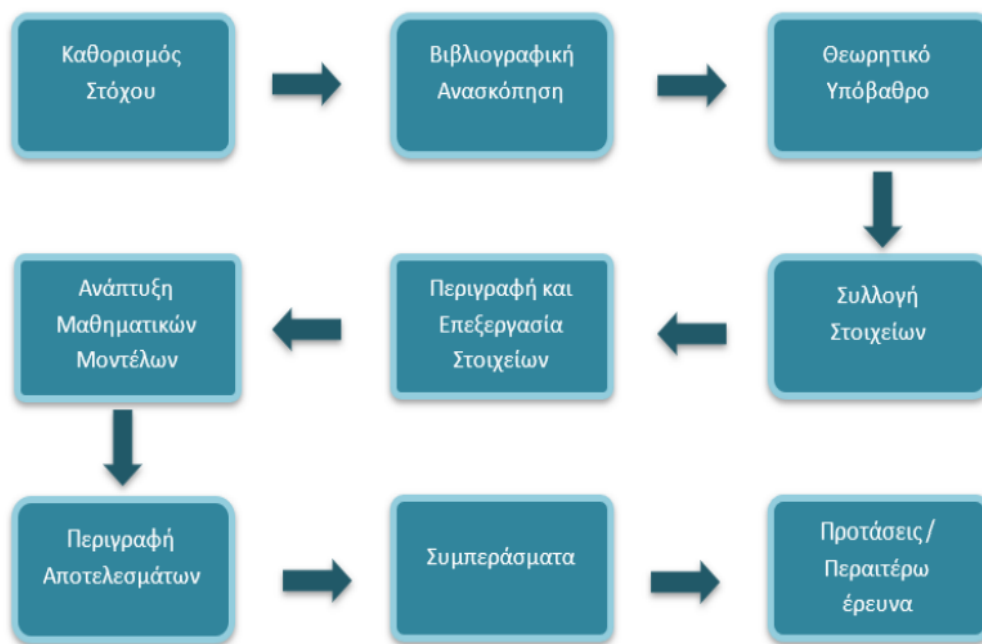
Στην συνέχεια, θα αναπτυχθούν και θα εφαρμοστούν κατάλληλα μοντέλα στατιστικής ανάλυσης και μέθοδοι μηχανικής μάθησης, τα οποία θα επιτρέψουν την ομαδοποίηση των διαδρομών βάσει των χαρακτηριστικών τους, καθώς και την ποσοτικοποίηση της επίδρασης επιμέρους παραγόντων στην πιθανότητα εμφάνισης κρίσιμων συμβάντων, όπως η ενεργοποίηση συστημάτων ασφαλείας (π.χ. ABS).

Τελικός στόχος της Διπλωματικής Εργασίας, είναι η εις βάθος μελέτη και κατανόηση των πρότυπων οδηγικής συμπεριφοράς με απώτερο σκοπό τη βελτίωση της οδικής ασφάλειας.

1.3 ΜΕΘΟΔΟΛΟΓΙΑ

Αρχικά καθορίστηκε το πεδίο που θα επικεντρωνόταν η παρούσα Διπλωματική Εργασία καθώς και ο πρωταρχικός στόχος της. Για την επίτευξη του στόχου διεξήχθη εκτεταμένη βιβλιογραφική ανασκόπηση στο τομέα της οδηγικής συμπεριφοράς και των συνδεμένων οχημάτων. Πραγματοποιήθηκε ανάλυση μελετών αναφορικά με μαθηματικές κατανομές και αλγορίθμων μηχανικής μάθησης. Στο πλαίσιο της ερευνάς κατέστη δυνατή η κατανόηση του αντικειμένου της εργασίας καθώς και η διαμόρφωση των κατάλληλων στατιστικών μεθόδων και αλγορίθμων που θα εφαρμοστούν.

Έπειτα της ολοκλήρωσης της συλλογής των δεδομένων από αισθητήρες των οχημάτων, καταχωρήθηκαν σε ηλεκτρονική βάση δεδομένων (αρχείο CSV). Ακολούθως πραγματοποιήθηκε η επιλογή της μεθόδου επεξεργασίας δεδομένων καθώς και η εισαγωγή σε ειδικό λογισμικό στατιστικής ανάλυσης (Rstudio). Έπειτα εφαρμόστηκε κατάλληλη ταξινόμηση κύριων χαρακτηριστικών μέσω του αλγόριθμου K-Means καθώς επίσης με την μέθοδο Τυχαία Δάση υπολογίστηκε η σημαντικότητα των μεταβλητών. Στην συνέχεια μέσω δοκιμών επιλέχθηκε και αναπτύχθηκε κατάλληλο μαθηματικό μοντέλο με το οποίο βρέθηκαν οι μεταβλητές οι οποίες επηρεάζουν σημαντικά την εξαρτημένη μεταβλητή.



Διάγραμμα 1.6 Διάγραμμα ροής της διπλωματικής εργασίας

1.4 ΔΟΜΗ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Το **πρώτο κεφάλαιο** αποτελεί μια εισαγωγή στην οδική ασφάλεια. Γίνεται αναφορά στα στατιστικά στοιχεία ατυχημάτων σε ευρωπαϊκό και ελληνικό επίπεδο. Έπειτα αναπτύσσεται ο κύριος στόχος της Διπλωματικής Εργασίας, η μεθοδολογία της, καθώς επίσης παρατίθεται σε συνοπτικό επίπεδο η δομή της (διάγραμμα ροής).

Στο **δεύτερο κεφάλαιο** γίνεται εκτενής παρουσίαση συμπερασμάτων που προέκυψαν από ανασκόπηση της υπάρχουσας βιβλιογραφίας σε έρευνες με παρεμφερές αντικείμενο.

Στο **τρίτο κεφάλαιο** αναλύεται το θεωρητικό υπόβαθρο που βασίζεται η παρούσα εργασία. Αρχικά γίνεται λεπτομερής αναφορά σε βασικές μαθηματικές και στατιστικές ορολογίες. Επίσης πραγματοποιείται επεξήγηση βασικών μαθηματικών κατανομών για την καλύτερη κατανόηση του μοντέλου που επιλέχθηκε.

Το **τέταρτο κεφάλαιο** πραγματεύεται την διαδικασία συλλογής και επεξεργασίας των δεδομένων. Παρατίθεται μια συνοπτική περιγραφή της εταιρείας OSeven, τους στόχους, τη δομή και το τρόπο λειτουργίας της. Στην συνέχεια, παρουσιάζεται το αρχείο της βάσης δεδομένων καθώς και τμήμα του κώδικα που εφαρμόστηκε για την δημιουργία συγκεντρωτικών γραφημάτων και αποτελεσμάτων στο πρόγραμμα της R. Τέλος πραγματοποιείται μια πρωταρχική ανάλυση των αποτελεσμάτων αναφορικά με κάθε συνδεδεμένο όχημα.

Στο **πέμπτο κεφάλαιο** περιγράφεται η εφαρμογή του αλγόριθμου K-Means, η διαδικασία ανάπτυξης των κατάλληλων μαθηματικών μοντέλων καθώς και τα ενδιάμεσα βήματα που χρησιμοποιήθηκαν για την εξαγωγή των τελικών

αποτελεσμάτων. Παρουσιάζονται τα δεδομένα εισόδου και εξόδου, με έμφαση στους στατιστικούς ελέγχους αξιοπιστίας των αποτελεσμάτων.

Στο **έκτο κεφάλαιο** παρουσιάζονται τα βασικά συμπεράσματα που προέκυψαν από την ανάλυση των δεδομένων και την ανάπτυξη των τελικών μοντέλων. Στο τέλος του κεφαλαίου διατυπώνονται προτάσεις που αποσκοπούν στην αξιοποίηση των αποτελεσμάτων της έρευνας και την βέλτιστη ανάλυση της συμπεριφοράς των οδηγών.

Στο **έβδομο κεφάλαιο** της Διπλωματικής Εργασίας παρατίθενται οι βιβλιογραφικές αναφορές, που αξιοποιήθηκαν για την εκπόνηση της παρούσας Διπλωματικής Εργασίας.

2 ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΑΝΑΣΚΟΠΗΣΗ

2.1 ΕΙΣΑΓΩΓΗ

Στο παρόν κεφάλαιο αναπτύσσονται οι κεντρικές έννοιες και τα θεωρητικά πλαίσια όπου επικεντρώνεται η παρούσα διπλωματική εργασία. Αρχικά πραγματοποιείται προσδιορισμός και επεξήγηση των όρων «Διασυνδεδεμένα Οχήματα» και «Συνεργατικά Ευφυή Συστήματα Μεταφορών» (C-ITS). Εν συνεχεία γίνεται ανασκόπηση ερευνών στον τομέα της οδικής ασφάλειας καθώς και παράθεση μελετών που έχουν ως κύριο αντικείμενο την οδηγική συμπεριφορά αξιοποιώντας δεδομένα που συλλέχθηκαν από διασυνδεδεμένα οχήματα. Παράλληλα, πραγματοποιείται παράθεση βασικών αλγορίθμων μηχανικής μάθησης και τεχνικές ταξινόμησης.

2.2 ΔΙΑΣΥΝΔΕΔΕΜΕΝΑ ΟΧΗΜΑΤΑ

Μία πρώτη προσέγγιση στον όρο «Διασυνδεδεμένο Όχημα» παρουσιάστηκε από την Uhlemann (2015), η οποία το προσδιορίζει ως ένα όχημα εξοπλισμένο με εφαρμογές, υπηρεσίες αλλά και τεχνολογίες οι οποίες συνδέονται με άλλα οχήματα ή την υποδομή. Τα συνδεδεμένα οχήματα έχουν ενσωματωμένη τεχνολογία ανταλλαγής δεδομένων, όπως επίσης συστήματα καμερών και καταγραφής κίνησης που τους παρέχει την δυνατότητα της συλλογής και διαμοιρασμού πληροφοριών σε μια βάση δεδομένων. Με αυτόν τον τρόπο οι εφαρμογές αυτές δύναται να συμβάλλουν στην βελτίωση της οδικής ασφάλειας και κυκλοφορίας όπως επίσης και στην πρόοδο της τηλεματικής και της αυτόνομης οδήγησης.



Εικόνα 2.1: Συνδεδεμένο Όχημα Πηγή: [FueLoyal](#)

2.3 ΣΥΝΕΡΓΑΤΙΚΑ ΕΥΦΥΗ ΣΥΣΤΗΜΑΤΑ ΜΕΤΑΦΟΡΩΝ C-ITS

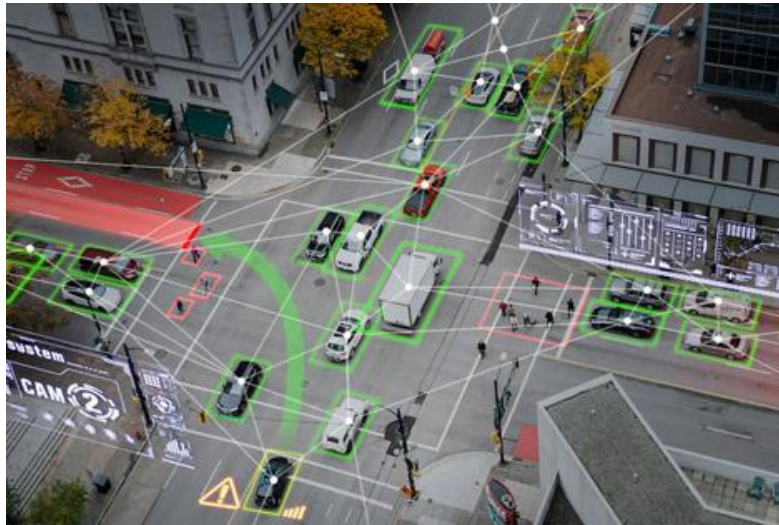
Πολλοί συγγραφείς έχουν δημοσιεύσει ερευνητικές μελέτες στο πεδίο των Συνεργατικών Ευφύων Συστημάτων Μεταφορών (Cooperative Intelligent Transport Systems C-ITS). Δεδομένου ότι τα C-ITS αποτελούν πεδίο μελέτης της παρούσας Διπλωματικής εργασίας κρίνεται αναγκαίο να παρουσιαστεί ο ορισμός τους. Σύμφωνα με το BSI (Bundesamt für Sicherheit in der Informationstechnik -Federal Office for Information Security) τα C-ITS επιτρέπουν την ανταλλαγή πληροφοριών σε πραγματικό χρόνο μεταξύ των οχημάτων, υποδομή και χρηστών με σκοπό την ενίσχυση της αυτοματοποίησης και διαχείριση της κυκλοφορίας. Μια επισκόπηση στην κεντρική έννοια και ιστορία των Συστημάτων Συνδεδεμένων και Αυτόνομων Οχημάτων παρουσιάστηκε από τον Shladover (2018). Αρχικά αναλύει την ποικιλομορφία των διασυνδεδεμένων οχημάτων, η οποία ενσωματώνεται σε:

- V2V (Vehicle to Vehicle)- Όχημα προς όχημα.
- V2I (Vehicle to Infrastructure)-Όχημα προς υποδομή.
- I2V (Infrastructure to Vehicle)-Υποδομή προς όχημα.
- V2P (Vehicle to Person)-Όχημα προς πεζό.
- V2X (Vehicle to Everything)-Όχημα προς τα πάντα.

Τα διασυνδεδεμένα οχήματα για την διασφάλιση της επικοινωνίας και ανταλλαγής δεδομένων στηρίζονται σε ένα ευρύ φάσμα ασυρμάτων τεχνολογιών. Στα πρώτα στάδια Vehicle to Infrastructure (V2I), το Υπουργείο μεταφορών των ΗΠΑ είχε εστιάσει αποκλειστικά στην τεχνολογία DSRC η οποία λειτουργεί στην συχνότητα 5,9 GHz. Η τεχνολογία αυτή είχε σχεδιαστεί λαμβάνοντας υπόψη το απαιτητικό περιβάλλον των κινητών εφαρμογών, παρέχοντας αξιοπιστία στην ασφάλεια των δεδομένων. Η υποστήριξη αυτών των εφαρμογών (Εξυπνων συστημάτων Μεταφορών-ITS) βασίζεται σε ένα ευρύ φάσμα ασυρμάτων τεχνολογιών με πλεονεκτήματα και περιορισμούς που διαφέρουν ανά περίπτωση. Πιο συγκεκριμένα:

- **5.9 GHz (DSRC):** Αποτελεί ειδική τεχνολογία τύπου Wi-Fi, με εφαρμογές στις οδικές μεταφορές, χρησιμοποιώντας αδειοδοτημένο και προστατευμένο φάσμα συχνοτήτων. Διαθέτει την δυνατότητα υποστήριξης μηνυμάτων τα οποία παρουσιάζουν αυξημένη απαίτηση χρονικής ακρίβειας και ασφάλεια.
- **Wi-Fi:** Μεγάλη καθυστέρηση σύνδεσης και είναι επιρρεπές σε απώλειες δεδομένων όταν είναι υπερφορτωμένο, με συνέπεια την αμφισβήτηση της αξιοπιστίας του για κρίσιμες πληροφορίες.
- **Κυψελοειδείς επικοινωνίες:** Η συγκεκριμένη κατηγορία περιλαμβάνει τεχνολογίες τύπου 4G LTE και WiMAX συμπεριλαμβανομένων και μελλοντικών εκδόσεων τους. Η εφαρμογή των κυψελοειδών επικοινωνιών δεν απαιτεί πρόσθετη επένδυση λόγω της ήδη υπάρχουσας υποδομής.
- Τα **Δορυφορικά συστήματα επικοινωνίας** είναι δυνατόν να χρησιμοποιηθούν, αντιμετωπίζοντας όμως ζητήματα όπως υψηλό κόστος, εύρος ζώνης και καθυστερήσεις.

- **Bluetooth:** Υποστηρίζεται μόνο από ορισμένες εφαρμογές και παρέχει υπηρεσίες μικρής εμβέλειας και περιορισμένου εύρους ζώνης.



Εικόνα 2.2: Συνεργατικά Ευφυή Συστήματα Μεταφορών Πηγή: [Swarco](#)

2.4 ΣΥΝΑΦΕΙΣ ΈΡΕΥΝΕΣ

2.4.1 ΜΕΘΟΔΟΛΟΓΙΑ ΚΑΙ ΤΑΞΙΝΟΜΗΣΗ ΕΡΕΥΝΩΝ

Η έρευνα που πραγματοποίησαν οι Zhou & Bridgeball (2020) αναδεικνύεται ως κομβικό σημείο του κεφαλαίου. Αρχικά επισημαίνουν το πρώιμο στάδιο της τεχνολογίας των συνδεδεμένων οχημάτων ενώ στην συνέχεια πραγματοποιούν μια συνολική ανασκόπηση στις υφιστάμενες μελέτες που χρησιμοποιούσαν δεδομένα από διασυνδεδεμένα οχήματα από το πρόγραμμα “Connected Vehicle-Safety Pilot Model Deployment” του Υπουργείου Μεταφορών των ΗΠΑ. Στο πλαίσιο της βιβλιογραφικής ανασκόπησης που πραγματοποίησαν αναφορικά με το SPMD (Safety Pilot Model Deployment) επισημαίνουν ότι οι μελέτες κατηγοριοποιούνται σε τρεις (3) άξονες:

➤ Στην αναγνώριση Προτύπων Οδήγησης

Τα δεδομένα (ταχύτητα, επιτάχυνση κλπ.) χρησιμοποιούνται για την ταξινόμηση προτύπων οδήγησης. Οι μεθοδολογίες βασίζονται στην ανταλλαγή των πληροφοριών μεταξύ οχημάτων και υποδομής και στην αναγνώριση επικίνδυνων χαρακτηριστικών όπως απότομες επιταχύνσεις, υψηλής διάρκειας φρεναρίσματα και απότομες αλλαγές λωρίδων.

➤ Ανάπτυξη Εναλλακτικών Δεικτών Οδικής Ασφάλειας

Εξαιτίας της περιορισμένης και χρονοβόρας διαδικασίας λήψης αξιόπιστων δεδομένων οι ερευνητές ανέπτυξαν εναλλακτικούς δείκτες οδικής ασφάλειας (SSs – Surrogate Safety Measures Development). Οι δείκτες αυτοί δεν βασίζονται σε πραγματικά οδικά ατυχήματα αλλά σε συμβάντα αυξημένου κινδύνου. Συνιστούν σύγχρονα εργαλεία αξιολόγησης της οδικής ασφάλειας (χωρίς την απαίτηση της εκδήλωσης οδικών ατυχημάτων) λαμβάνοντας τα δεδομένα από: διασυνδεδεμένα οχήματα, αισθητήρες στον δρόμο ή στα οχήματα και από εξοπλισμό πλευρικός της οδού που παρακολουθούν την κυκλοφορία.

➤ Βελτίωση Λειτουργίας Σηματοδότησεων και Διασταυρώσεων

Οι φωτεινοί σηματοδότες στους κόμβους είναι συχνά, σημεία με υψηλή κυκλοφοριακή συμφόρηση. Λόγω του υψηλού κόστους των «έξυπνων συστημάτων σηματοδότησης» μια πιο οικονομικά αποδεκτή λύση αποτελεί η ρύθμιση των σηματοδοτών λαμβάνοντας δεδομένα από συνδεδεμένα οχήματα. Πιο συγκεκριμένα οι Zheng και Liu για την κίνηση των οχημάτων και τον προγραμματισμό των φωτεινών σηματοδοτών πρότειναν μια μέθοδο εκτίμησης της κυκλοφοριακής ροής μέσα από δεδομένα που συλλέχθηκαν από συνδεδεμένα οχήματα.

2.4.2 ΜΕΘΟΔΟΙ ΑΝΑΛΥΣΗΣ

Αναφορικά με τις μεθόδους ανάλυσης, σημαντική είναι η έρευνα των Seraji et al. (2025), η οποία εξετάζει πως η μηχανική μάθηση, σε συνδυασμό με την επεξεργασία δεδομένων, αξιοποιείται για την βελτίωση της οδηγικής συμπεριφοράς. Αναλύονται οι κύριοι αλγόριθμοι ταξινόμησης καθώς και μοντέλα μηχανικής μάθησης με εφαρμογές στην οικολογική οδήγηση, στην ενεργειακή κατανάλωση και την ενίσχυση της οδικής ασφάλειας.

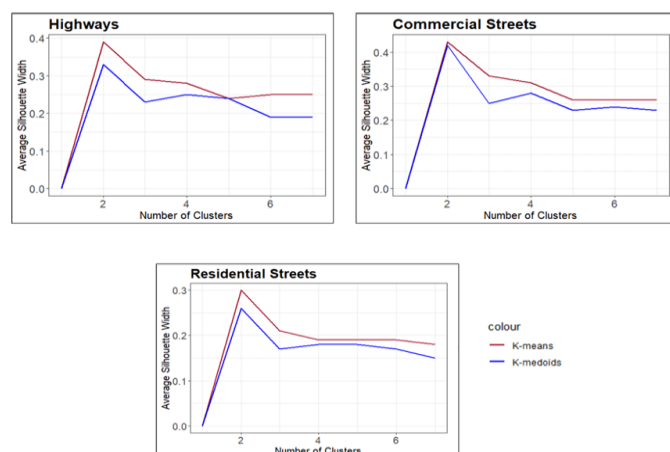
Σύμφωνα και με τους Ghosh & Dubey (2013) το **Fuzzy C-Means** αποτελεί μια μέθοδο ταξινόμησης δεδομένων που μπορεί να ανήκουν σε περισσότερες από μια κατηγορίες. Αποτελεί σημαντικό εργαλείο για την κατηγοριοποίηση οδηγών και οδηγικών συμπεριφορών βασιζόμενη σε δεδομένα που δεν απαιτείται να ανήκουν αποκλειστικά σε μια μόνο κατηγορία.

Παρόμοια ο αλγόριθμος **K-means**, αποτελεί μια από τις πιο γνωστές μεθόδους ταξινόμησης, στόχος του είναι να κατατάξει ένα σύνολο σε k ομάδες οι οποίες παρουσιάζουν υψηλό βαθμό ομοιότητας. Λειτουργεί επαναληπτικά και ακολουθεί την εξής διαδικασία: (1) επιλογή τυχαίου αριθμού k (centroids), (2) τα σημεία του συνόλου κατατάσσονται στην κατηγορία του πλησιέστερου κέντρου, σύμφωνα με κάποια μετρική απόσταση, (3) τα κέντρα αναπροσαρμόζονται υπολογίζοντας τον μέσο όρο των σημείων που ανήκουν, (4) επανάληψη των βημάτων (2) και (3) μέχρι να επιτευχθεί η σύγκλιση. Σε αντίθεση με το **K-means**, ο αλγόριθμος **DBSCAN**

(Density-Based Spatial Clustering of Applications with Noise) δεν απαιτεί προκαθορισμένο αριθμό ομάδων. Κατηγοριοποιεί τα δεδομένα σε τρεις τύπους σημείων: (1) core points (2) border points και (3) noise points τα οποία δεν ανήκουν σε καμία συστάδα.

Στο πλαίσιο της Μηχανικής Μάθησης, οι Seraji et al. (2025) υποστηρίζουν ότι δύο από τις πιο αποτελεσματικές μεθόδους είναι οι Support Vector Machines (SVM) και το Decision Trees (DTs). Ο αλγόριθμος SVM χρησιμοποιείται ευρέως για εργασίες ταξινόμησης και παλινδρόμησης. Βασική αρχή αποτελεί η αναζήτηση hyperplane που διαχωρίζει τις κατηγορίες δεδομένων με βέλτιστη ακρίβεια. Στόχος του αλγορίθμου είναι η μεγιστοποίηση της απόστασης μεταξύ του hyperplane και των πλησιέστερων δεδομένων κάθε κατηγορίας. Τα DTs αποτελούν επιβλεπόμενοι αλγόριθμοι μηχανικής μάθησης. Διαθέτουν ιεραρχική δομή και η λειτουργία τους μοιάζει με διάγραμμα ροής. Οι εσωτερικοί κομβοί αναπαριστούν ελέγχους χαρακτηριστικών, τα κλαδιά αντιστοιχούν σε τιμές των χαρακτηριστικών και τα φύλλα παρέχουν τις τελικές προβλέψεις.

Αξίζει να αναφερθεί η έρευνα των Mohammadnazar et al. (2021) η οποία επικεντρώνεται στην ταξινόμηση του στυλ οδήγησης ανάλογα με τον τύπο της οδού και το περιβάλλον, αξιοποιώντας αλγόριθμους μηχανικής μάθησης λαμβάνοντας δεδομένα από διασυνδεδεμένα οχήματα. Συλλέχθηκαν πληροφορίες από 2.836 οχήματα εξοπλισμένα με συστήματα V2V και V2I, τα οποία μετέδιδαν καταγραφές ταχυτήτων και επιτάχυνσης με συχνότητα 10 Hz. Σε πρώτο στάδιο της μελέτης πραγματοποιήθηκε κατάλληλη επεξεργασία των δεδομένων, περιορίστηκε η περιοχή μελέτης ενώ εντοπίστηκαν και αφαιρέθηκαν σφάλματα στα στοιχεία λόγω λανθασμένης κωδικοποίησης. Έπειτα κατηγοριοποιήθηκαν τα οδικά τμήματα βάσει των γεωμετρικών χαρακτηριστικών και ζωνών χρήσης γης σε αυτοκινητοδρόμους εμπορικούς και κατοικημένους δρόμους. Μέσω της εφαρμογής των μεθόδων K-Means και K-Medoids υπολογίστηκε το ASWC (Average Silhouette Width Criterion) το οποίο λειτουργεί ως κριτήριο αξιολόγησης της ποιότητάς τους λαμβάνοντας υπόψη την εσωτερική συνοχή των ομάδων.



Σχήμα 2.1: Αποδόσεις ομαδοποίησης K-means και K-medoids χρησιμοποιώντας τον δείκτη ASWC

Ο πίνακας 2.1 παρουσιάζει μια συνοπτική ανασκόπηση της κατανομής των προφίλ οδήγησης ανά τύπο οδού. Παρέχονται τα στατιστικά στοιχεία για την περιγραφή των διαφορετικών προτύπων συμπεριφοράς των οδηγών σε κάθε τύπο οδού. Τα αποτελέσματα της μελέτης εμφανίζονται στον πίνακα 2.2 και φανερώνουν αυξημένη συχνότητα επιθετικής οδήγησης σε εμπορικούς δρόμους σε ποσοστό 23,47%. Επιπλέον, οι μεταβολές της ταχύτητας και της επιτάχυνσης αναδεικνύονται ως οι κύριοι παράγοντες διαφοροποίησης των προτύπων οδήγησης σε διαφορετικούς τύπους οδού.

Πίνακας 2.1: Οι κανονικοποιημένες τιμές των κέντρων των ομάδων που αντιστοιχούν σε κάθε κατηγορία

Cluster	$Speed-V_f$	$Speed-D_{mean}$	$Speed-C_r$	$Speed-Q_{CV}$	$Accel-D_{mean}$	$Accel-D_{mean}$
Highways						
Cluster 1 (Aggressive)	1.572	1.705	1.971	1.940	1.096	0.654
Cluster 2 (Normal)	0.147	0.316	0.186	0.187	0.334	0.337
Cluster 3 (Calm)	-0.570	-0.775	-0.717	-0.710	-0.630	-0.514
Commercial streets						
Cluster 1 (Aggressive)	1.486	0.669	1.500	1.500	0.505	-0.245
Cluster 2 (Normal)	-0.182	0.712	-0.079	-0.073	0.761	0.818
Cluster 3 (Calm)	-0.569	-0.584	-0.617	-0.620	-0.533	-0.231
Residential streets						
Cluster 1 (Aggressive)	2.209	1.803	2.356	2.338	1.789	1.297
Cluster 2 (Normal)	0.399	0.607	0.494	0.493	0.332	0.151
Cluster 3 (Calm)	-0.558	-0.636	-0.639	-0.636	-0.458	-0.274

Πίνακας 2.2: Κατανομή των διαδρομών ανά ομάδα οδηγικής συμπεριφοράς

Type of road	Total num. of events	Total num. of drivers	Driving style					
			Aggressive		Normal		Calm	
			Number of events	Percent %	Number of events	Percent %	Number of events	Percent %
Highways	7537	1302	893	11.85	3326	44.13	3318	44.02
Commercial streets	3387	975	795	23.47	757	22.35	1835	54.18
Residential streets	1912	840	147	7.96	690	36.09	1075	56.22

Μία από τις μελέτες που έχει βασιστεί η παρούσα εργασία είναι αυτή των Koliou et al. (2025). Η έρευνα είναι σημαντική διότι αξιοποιεί την μηχανική μάθηση για την κατηγοριοποίηση οδικών τμημάτων της Αθήνας βάσει του κινδύνου ατυχήματος. Η μεθοδολογία οργανώνεται σε τρεις φάσεις: (1) συλλογή και προεπεξεργασία των δεδομένων, (2) μετατροπή των δεδομένων σε χρήσιμα χαρακτηριστικά με ενσωμάτωση στον κυκλοφοριακό φόρτο (3) εκπαίδευση και αξιολόγηση του μοντέλου ταξινόμησης.

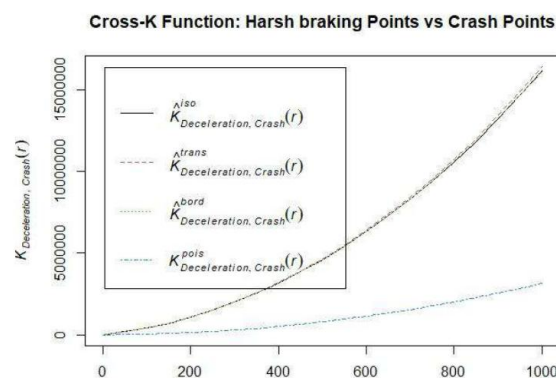
Για την προεπεξεργασία των δεδομένων, αξιοποιήθηκαν αναφορές ατυχημάτων που προέρχονται από τις αρμόδιες ελληνικές αρχές, οι οποίες καταγράφουν τις οδούς όπου σημειώθηκαν περιστατικά. Έπειτα δεδομένα τηλεματικής που προέκυψαν από 2.614 διαδρομές πραγματοποιημένες από 257 οδηγούς, αντιστοιχήθηκαν χωρικά με τα τμήματα οδικού δικτύου υψηλού κινδύνου. Το τελικό σύνολο δεδομένων περιλάμβανε 148 επιμέρους οδικά τμήματα, τα οποία ταξινομήθηκαν ως «Ασφαλή»

(≤ 2 ατυχήματα) ή ως «Μη ασφαλή» (> 2 ατυχήματα). Για την αξιολόγηση της απόδοσης δοκιμάστηκαν μοντέλα λογιστικής παλινδρόμησης και μηχανές υποστήριξης διανυσμάτων (SVM). Μέσω της ομαδοποίησης εντοπίστηκαν περιοχές με αυξημένο κίνδυνο ατυχημάτων, ενώ η χωρική ανάλυση αναγνώρισε συγκεκριμένους επικίνδυνους κόμβους του οδικού δικτύου. Τα ευρήματα της μελέτης επίσης αναφέρουν ότι χαρακτηριστικά όπως ταχύτητα, επιτάχυνση και συχνότητα απότομων φρεναρισμάτων αποτελούν βασικές ενδείξεις αυξημένης επικινδυνότητας σε ένα οδικό δίκτυο.

Ιδιαίτερο ενδιαφέρον παρουσιάζει η ερευνά των Ziakopoulos et al. (2025). Η μελέτη στοχεύει στην ανάλυση και πρόβλεψη της οδικής ασφάλειας αξιοποιώντας δεδομένα από ελαφρά φορτηγά στην περιοχή του Λονδίνου, καθώς και στην δυνατότητα αξιοποίησης υποκατάστατων δεικτών ασφαλείας (SSMs). Τα δεδομένα που συλλέχθηκαν περιλάμβαναν αναγνώριση οχήματος (vehicle ID), χρονική σήμανση, θέση γεωγραφικό πλάτος και μήκος, ταχύτητα και επιτάχυνση του οχήματος στους άξονες του εδάφους.

Μέσω της μεθόδου Cross Nearest Neighbor (cNN) υπολογίστηκε η απόσταση του συνόλου των κοντινότερων φρεναρισμάτων και των κοντινότερων ατυχημάτων. Χρησιμοποιήθηκε η συνάρτηση Cross-K Function (K12) για την ποσοτικοποίηση της χωρικής σχέσης μεταξύ των συνόλων των απότομων φρεναρισμάτων και των οδικών ατυχημάτων.

Η ανάλυση έδειξε ότι τα περισσότερα ζεύγη απότομου φρεναρίσματος και οδικών ατυχημάτων που περιλαμβάνουν ελαφρά φορτηγά απέχουν λιγότερο από 500 μέτρα υποδηλώνοντας έντονη χωρική συγκέντρωση. Η συνάρτηση K12 έδειξε ότι ο αριθμός των σημείων των οδικών ατυχημάτων εντός απόστασης r από τα σημεία απότομου φρεναρίσματος αυξάνονται αναλογικά με την απόσταση. Τέλος συμπεραίνεται ότι τα δεδομένα φρεναρίσματος μπορούν να αξιοποιηθούν ως δείκτες οδικής επικινδυνότητας (SSMs).



Σχήμα 2.2: Cross-K συναρτήσεις για διασυνδεδεμένα LCV οχήματα, απότομα φρεναρίσματα και συγκρούσεις

2.5 ΣΥΝΟΨΗ ΒΙΒΛΙΟΓΡΑΦΙΑΣ

Μια πρώτη προσέγγιση στον όρο «Συνδεδεμένο Όχημα» παρουσιάστηκε από την Uhlemann (2015), η οποία το προσδιορίζει ως ένα όχημα εξοπλισμένο με τεχνολογίες που επιτρέπουν την ανταλλαγή δεδομένων με άλλα οχήματα ή την υποδομή. Επίσης παρουσιάζεται η δομή λειτουργίας των Συνεργατικών Ευφυιών Συστημάτων Μεταφοράς και οι τεχνολογίες στις οποίες βασίζονται.

Παρόλο που η τεχνολογία των συνδεδεμένων οχημάτων βρίσκεται σε πρώιμο στάδιο, οι έρευνες αυτές συμβάλλουν σημαντικά στην εξαγωγή χρήσιμων συμπερασμάτων. Αρχικά οι Zhou & Bridgeball (2020) επισήμαναν ότι οι μελέτες κατηγοριοποιούνται σε τρεις άξονες: (α) στην ανάλυση και αξιολόγηση της οδηγικής συμπεριφοράς μέσω της επεξεργασίας δεδομένων (β) στη δημιουργία εναλλακτικών δεικτών οδικής ασφάλειας (SSMs) χωρίς την αναγκαιότητα λήψης δεδομένων από πραγματικά οδικά ατυχήματα και (γ) στη βελτίωση των φωτεινών σηματοδοτήσεων και διασταυρώσεων. Συνοψίζοντας, μέσω των ερευνών διαπιστώθηκε ότι η ανάλυση των δεδομένων από διασυνδεδεμένα οχήματα, συμβάλλει έμπρακτα στην ανάλυση της οδηγικής συμπεριφοράς και στην γενικότερη βελτίωση της οδικής ασφάλειας.

Στην ενότητα 2.4.2 παρατίθενται μελέτες οι οποίες αξιοποιούν κύριες μεθόδους ανάλυσης και μηχανικής μάθησης. Παρουσιάζονται βασικοί αλγόριθμοι όπως οι Fuzzy C-Means, K-Means, DBSCAN, SVM και Decision Trees. Γίνεται αναφορά στη μελέτη των Mohammadnazar et al. (2021) η οποία στοχεύει στην ταξινόμηση των προφίλ οδήγησης, μέσω αλγόριθμων μηχανικής μάθησης, αξιοποιώντας δεδομένα από διασυνδεδεμένα οχήματα. Σημαντικό έργο παρουσιάστηκε από Koliou et al. (2025) όπου χρησιμοποίησαν μεθόδους μηχανικής μάθησης για τον εντοπισμό οδών της Αθήνας με υψηλό κίνδυνο ατυχήματος και κατηγοριοποίησαν οδικά τμήματά της. Επίσης οι Ziakopoulos et al. (2025) πραγματοποίησαν μελέτη αξιοποιώντας δεδομένα από ελαφρά φορτηγά στο Λονδίνο και απέδειξαν ότι δεδομένα απότομου φρεναρίσματος μπορούν να αποτελέσουν δείκτες οδικής επικινδυνότητας (SSMs).

3. ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

3.1 ΕΙΣΑΓΩΓΗ

Το παρόν κεφάλαιο αφιερώνεται στην ανάπτυξη του θεωρητικού υπόβαθρου που αξιοποιήθηκε για την επίτευξη του στόχου της Διπλωματικής Εργασίας. Αρχικά, παρουσιάζονται συνοπτικά βασικές έννοιες στατιστικής ανάλυσης που χρησιμοποιούνται ευρέως σε εφαρμογές στον τομέα της οδικής ασφάλειας. Στη συνέχεια, ακολουθεί επεξήγηση των μαθηματικών μεθόδων που χρησιμοποιήθηκαν καθώς και κριτήρια αποδοχής ή απόρριψης ενός μοντέλου.

3.2 ΒΑΣΙΚΕΣ ΈΝΝΟΙΕΣ ΣΤΑΤΙΣΤΙΚΗΣ ΑΝΑΛΥΣΗΣ

Κάθε ανάλυση στηρίζεται σε βασικές έννοιες της στατιστικής. Ο πρώτος ορισμός είναι ο **πληθυσμός**. Στατιστικός πληθυσμός χαρακτηρίζεται ως το σύνολο όλων των στοιχείων (άτομα, μετρήσεις, γεγονότα) που έχουν κάποιο κοινό χαρακτηριστικό και αποτελούν αντικείμενο μελέτης. Σε πολλά πρακτικά προβλήματα λόγω του μεγάλου όγκου του υπό εξέτασης πληθυσμού, επιλέγεται ένα υποσύνολο του. Το υποσύνολο αυτό, ονομάζεται **δείγμα** και με κατάλληλες στατιστικές μεθόδους, προκύπτουν συμπεράσματα τα οποία στη συνέχεια εφαρμόζονται με ικανοποιητική ακρίβεια στον πληθυσμό. Η αξιοπιστία των συμπερασμάτων που προκύπτουν από την ανάλυση του δείγματος έχει πλήρης εξάρτηση με το αν το δείγμα είναι αντιπροσωπευτικό του πληθυσμού.

Εξίσου σημαντική στην στατιστική αποτελεί η έννοια της **μεταβλητής**, η οποία χρησιμοποιείται για να περιγράψει τα χαρακτηριστικά του πληθυσμού που διαφοροποιούνται από μέλος σε μέλος. Ανάλογα με την φύση τους οι μεταβλητές διακρίνονται σε ποσοτικές και ποιοτικές.

- **Ποιοτική μεταβλητή** είναι αυτή που εκφράζει χαρακτηριστικά του πληθυσμού που διαφοροποιούνται από μέλος σε μέλος.
- **Ποσοτική μεταβλητή** είναι εκείνη που μπορεί να μετρηθεί. Κατηγοριοποιούνται σε διακριτές και συνεχείς.

Στην επιστήμη της στατιστικής, δύο από τα πλέον θεμελιώδη μεγέθη είναι η **μέση τιμή** και η **διακύμανση**. Μέση τιμή $[\bar{X}]$ ορίζεται ως το πηλίκο του αθροίσματος των τιμών μιας μεταβλητής προς το πλήθος τους.

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} (x_1 + x_2 + \dots + x_n).$$

Η **διακύμανση** εκφράζει την διασπορά ενός συνόλου αριθμητικών δεδομένων από την μέση τους τιμή $[\bar{X}]$ και συμβολίζεται με S^2 ή $\text{Var}(X)$. Ο υπολογισμός της βασίζεται στα δεδομένα $x_1, \dots, x_n$ και υπολογίζεται με την παρακάτω σχέση:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})^2$$

Παρόμοια ερμηνεία με την διακύμανση έχει η **τυπική απόκλιση S** και ορίζεται ως η θετική τετραγωνική ρίζα της διακύμανσης.

$$S = +\sqrt{S^2}$$

Ένας επιπλέον διαχωρισμός των μεταβλητών, είναι σε **εξαρτημένες** και **ανεξάρτητες**. Εξαρτημένες είναι οι μεταβλητές των οποίων οι τιμές καθορίζονται από τις τιμές άλλων μεταβλητών, που ονομάζονται ανεξάρτητες. Αντίθετα οι τιμές των ανεξάρτητων μεταβλητών δεν εξαρτώνται από καμία άλλη μεταβλητή. Η διαδικασία με την οποία πραγματοποιείται η πρόβλεψη μίας μεταβλητής του δείγματος χρησιμοποιώντας τις υπόλοιπες ορίζεται ανάλυση παλινδρόμησης. Η εξαρτημένη μεταβλητή επιλέγεται από τον μελετητή, ενώ οι ανεξάρτητες δεν επιλέγονται αυθαίρετα, αλλά λαμβάνουν συγκεκριμένες τιμές. Η παλινδρόμηση χαρακτηρίζεται ως απλή όταν συμμετέχει μόνο μία ανεξάρτητη μεταβλητή ενώ αν χρησιμοποιούνται περισσότερες λέγεται πολλαπλή παλινδρόμηση. Επιπρόσθετα μια σημαντική έννοια είναι ο **συντελεστής συσχέτισης ρ**, είναι καθαρός αριθμός και εκφράζει τον βαθμό στον οποίο δύο μεταβλητές X και Y, μεταβάλλονται μεταξύ τους. Συμβολίζεται με **ρ(X,Y)**, λαμβάνει τιμές στο διάστημα [-1,1] και ορίζεται από τη σχέση:

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

Στην πράξη, ο συντελεστής συσχέτισης ρ, με πρακτικά δεδομένα $(x_1, y_1) (x_2, y_2), \dots, (x_n, y_n)$ εκτιμάται μέσω του συντελεστή συσχέτισης Pearson, συμβολίζεται με $r(X,Y)$ και εκφράζεται από τον τύπο:

$$r(X,Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

Αν η τιμή του $r(X,Y)$ λάβει θετικές τιμές δηλώνει ότι η μία μεταβλητή αυξάνεται όσο αυξάνεται η άλλη, ενώ σε αντίθετη περίπτωση (δηλαδή λάβει αρνητικές τιμές) δηλώνει ότι η μία μεταβλητή αυξάνεται όσο μειώνεται η άλλη. Αν $r(X,Y)=0$, δεν υπάρχει γραμμική συσχέτιση μεταξύ των X και Y. Τέλος, οι ανεξάρτητες μεταβλητές που εισάγονται στα μοντέλα παλινδρόμησης θα πρέπει να έχουν χαμηλό συντελεστή συσχέτισης $|\rho(x_i, x_j)| \leq 0,5-0,6$.

Υπενθύμιση: $COV(X,Y) = E\{(X-E(X))(Y-E(Y))\} = E(XY)-E(X)E(Y)$

3.3 ΕΡΜΗΝΕΙΑ ΜΑΘΗΜΑΤΙΚΩΝ ΜΟΝΤΕΛΩΝ-ΚΑΤΑΝΟΜΩΝ

3.3.1 ΚΑΤΑΝΟΜΗ POISSON

Η κατανομή Poisson, μία από τις πιο σημαντικές κατανομές της θεωρίας πιθανοτήτων, εφαρμόζεται σε περιπτώσεις που θέλουμε να απαριθμήσουμε τον αριθμό τυχαίων γεγονότων που συμβαίνουν σε ένα δεδομένο χρονικό διάστημα. Έστω X μία διακριτή τυχαία μεταβλητή (π.χ. εκφράζει αριθμό ατυχημάτων), με παράμετρο $\mu > 0$, ακολουθεί την κατανομή Poisson όταν έχει συνάρτηση μάζας πιθανότητας:

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, x = 0, 1, \dots, \lambda > 0$$

Από τα παραπάνω προκύπτει ότι τα οδικά ατυχήματα συχνά μοντελοποιούνται με κατανομή Poisson, καθώς αποτελούν διακριτά γεγονότα που μπορούν να εμφανιστούν τυχαία μέσα σε ένα σταθερό χρονικό διάστημα t . Επίσης η μέση τιμή και η διακύμανση της κατανομής Poisson ορίζονται από τις σχέσεις: $E(X) = \lambda$ και $Var(X) = \lambda$. Σε εφαρμογή της κατανομής Poisson για οδικά ατυχήματα θεωρούμε ότι $P(X)$ ορίζεται η πιθανότητα να εμφανιστούν X ατυχήματα στην εξεταζόμενη περιοχή στη χρονική περίοδο αναφοράς και λ ο αναμενόμενος αριθμός ατυχημάτων στην εξεταζόμενη περιοχή. Μέσω της παλινδρόμησης Poisson είναι εφικτό να υπολογιστεί η παράμετρος λ της κατανομής. Για τον υπολογισμό της χρησιμοποιείται μια σειρά μεταβλητών, πιο συγκεκριμένα: $\lambda_i = \text{EXP}(|\beta| |X_i|)$

Η γενική μαθηματική εξίσωση είναι: $\log(\lambda_i) = \beta_0 + \beta_1 \chi_1 + \beta_2 \chi_2 + \dots + \beta_p \chi_p$

όπου:

- λ_i = εξαρτημένη μεταβλητή.
- $\chi_1, \chi_2, \dots, \chi_p$ = ανεξάρτητες μεταβλητές.
- $\beta_1, \beta_2, \dots, \beta_p$ = συντελεστές των ανεξάρτητων μεταβλητών.
- β_0 = σταθερός συντελεστής.

3.3.2 ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ (LOGISTIC REGRESSION)

Είναι αναγκαίο αρχικά να διευκρινιστεί ότι η λογιστική παλινδρόμηση δεν αποτελεί αλγόριθμος παλινδρόμησης αλλά αλγόριθμος ταξινόμησης. Χρησιμοποιείται δηλαδή για την πρόβλεψη δυαδικών αποτελεσμάτων, όπως η ύπαρξη ή η απουσία ενός χαρακτηριστικού. Εφαρμόζεται κυρίως σε περιπτώσεις που η εξαρτημένη μεταβλητή είναι δυαδική, δηλαδή μπορεί να πάρει μόνο δύο τιμές (π.χ. 0 ή 1, ναι ή όχι), οι οποίες αντιστοιχούν σε δύο ενδεχόμενα. Είναι μια γενίκευση της απλής γραμμικής παλινδρόμησης. Στην λογιστική παλινδρόμηση η εκτίμηση των παραμέτρων γίνεται μέσω της εκτίμησης μέγιστης πιθανοφάνειας (Ε.Μ.Π), δηλαδή επιλέγονται οι τιμές των παραμέτρων, προκειμένου να οδηγήσουν στα παρατηρούμενα αποτελέσματα. Για να υπολογιστεί η πιθανότητα ενός γεγονότος Y_i αν γνωρίζουμε τις παραμέτρους $\beta_0, \beta_1, \dots, \beta_n$ υπάρχει το λογιστικό μοντέλο:

$$E(Y_i) = \frac{1}{1 + e^{-\beta_0 - x_i \beta_i}}$$

3.3.3 ΔΙΩΝΥΜΙΚΗ ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Η συγκεκριμένη παλινδρόμηση αποτελεί μια στατιστική μέθοδος που χρησιμοποιείται για την πρόβλεψη και τη μοντελοποίηση μιας εξαρτημένης μεταβλητής που λαμβάνει 2 πιθανές τιμές, είναι δηλαδή δυαδικής μορφής συνήθως 0 και 1. Ο βασικός στόχος είναι η πρόβλεψη της πιθανότητας της εξαρτημένης μεταβλητής Y να λάβει την τιμή 1, με βάση τις τιμές των ανεξάρτητων μεταβλητών X .

Παρακάτω παρουσιάζεται η συνάρτηση χρησιμότητας (U) πιθανότητας (P):

$$U_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + e_i$$

$$P_i = \frac{e^{U_i}}{1 + e^{U_i}}$$

U_i = Η χρησιμότητα που αντιστοιχεί στην επιλογή της υπόθεσης

P_i = Η πιθανότητα i να συμβεί μια συγκεκριμένη υπόθεση

3.4 ΚΡΙΤΗΡΙΑ ΑΠΟΔΟΧΗΣ ΜΟΝΤΕΛΟΥ

Η δημιουργία ενός αξιόπιστου μοντέλου για την πρόβλεψη της εξαρτημένης μεταβλητής δεν είναι απλή διαδικασία που βασίζεται σε αυθαίρετη επιλογή ανεξάρτητων μεταβλητών από το δείγμα. Είναι απαραίτητο να τηρούνται ορισμένα κριτήρια για την εγκυρότητα του μοντέλου. Βασικά κριτήρια αποτελούν τα πρόσημα και οι τιμές των συντελεστών β_i , η ελαστικότητα και στατιστική σημαντικότητα όπως επίσης η ποιότητα του μοντέλου και το σφάλμα εξίσωσης.

3.4.1 ΕΡΜΗΝΕΙΑ ΤΩΝ ΠΡΟΣΗΜΩΝ ΚΑΙ ΤΩΝ ΤΙΜΩΝ ΤΩΝ ΣΥΝΤΕΛΕΣΤΩΝ

Το πρόσημο του συντελεστή β_i υποδηλώνει τη συσχέτιση μεταξύ ανεξάρτητης και εξαρτημένης μεταβλητής. Θετικό πρόσημο σημαίνει αύξηση της εξαρτημένης μεταβλητής με την αύξηση της ανεξάρτητης, ενώ το αρνητικό δηλώνει μείωση της εξαρτημένης με αύξηση της ανεξάρτητης. Επίσης, θα πρέπει να έχει λογική ερμηνεία η τιμή του συντελεστή, διότι αύξηση της ανεξάρτητης μεταβλητής κατά μια μονάδα προκαλεί αύξηση της εξαρτημένης κατά β_i μονάδες.

3.4.2 ΣΤΑΤΙΣΤΙΚΗ ΣΗΜΑΝΤΙΚΟΤΗΤΑ ΚΑΙ ΕΛΑΣΤΙΚΟΤΗΤΑ

Ένας από τους βασικούς ελέγχους για την αξιολόγηση του μοντέλου είναι ο έλεγχος t-test. Μέσω του δείκτη t εκτιμάται η στατιστική σημαντικότητα των ανεξάρτητων μεταβλητών που συμβάλλουν στο τελικό μοντέλο. Ουσιαστικά εκτιμάται αν η επίδραση της αντίστοιχης μεταβλητής στο μοντέλο είναι σημαντική ή όχι (το s.e. εκφράζει τον συντελεστή σφάλματος).

$$t_{\text{stat}} = \frac{\beta_i}{\text{s.e.}}$$

Τέλος αναφορικά με την ελαστικότητα, πρόκειται για έναν δείκτη που αποτυπώνει την ευαισθησία μιας εξαρτημένης μεταβλητής στην μεταβολή των ανεξάρτητων μεταβλητών. Η ευαισθησία εκφράζεται ως η ποσοστιαία μεταβολή της εξαρτημένης που προκαλείται από μια μεταβολή 1% της ανεξάρτητης. $\epsilon_i = \beta_i (X_i / Y_i)$

3.4.3 ΣΥΝΤΕΛΕΣΤΗΣ ΠΡΟΣΑΡΜΟΓΗΣ R^2 – ΑΞΙΟΛΟΓΗΣΗ ΠΟΙΟΤΗΤΑΣ

Ο συντελεστής προσαρμογής R^2 εκφράζει τον δείκτη αξιολόγησης της ποιότητας του μοντέλου καθώς αποτελεί κριτήριο καλής προσαρμογής των δεδομένων και ορίζεται από τη σχέση:

$$R^2 = \frac{\sum_{t=1}^T (\hat{Y}_t - \bar{Y})^2}{\sum_{t=1}^T (Y_t - \bar{Y})^2}$$

Ο αριθμητής εκφράζει το **ερμηνευόμενο άθροισμα τετραγώνων**, είναι γνωστό ως ESS (Explained Sum of Squares) και ο παρονομαστής εκφράζει το συνολικό

άθροισμα τετραγώνων (Total Sum of Squares). Ο όρος $\hat{Y}_i$ συμβολίζει την προβλεπόμενη τιμή της εξαρτημένης μεταβλητής Y βάσει της αντίστοιχης ανεξάρτητης X . Ο συντελεστής R^2 λαμβάνει τιμές από 0 έως 1, όσο πλησιέστερη στη μονάδα είναι η τιμή τόσο πιο ισχυρή είναι γραμμική σχέση στις μεταβλητές X και Y . Δεν υπάρχει σαφής προσδιορισμός για το κατώτατο όριο αποδοχής ενός μοντέλου, προτιμάται κατά κανόνα το μοντέλο με την μεγαλύτερη τιμή. Όσον αφορά το σφάλμα εξίσωσης του μοντέλου, χρησιμοποιείται για να ληφθούν υπόψη επίδραση μεταβλητών που έχουν παραληφθεί ή για τυχόν αποκλίσεις που προκύπτουν από παραδοχές. Οφείλει να ικανοποιεί βασικές προϋποθέσεις για τη γραμμική παλινδρόμηση:

- Να ακολουθεί κανονική κατανομή.
- Να έχει σταθερή διασπορά $\text{Var}(\varepsilon_i)=c$.
- Να έχει μηδενική συσχέτιση $\rho(\varepsilon_i, \varepsilon_j)=0$.

3.5 ΑΛΓΟΡΙΘΜΟΣ K-MEANS

Ο αλγόριθμος K-Means όπως επισημαίνουν οι Ghosh & Dubey (2013) αποτελεί μια μέθοδο ταξινόμησης (clustering) που χρησιμοποιείται συχνά στην ανάλυση δεδομένων. Βασικός στόχος του αλγόριθμου είναι η ομαδοποίηση των δεδομένων σε K ομάδες με σκοπό τα σημεία κάθε συστάδας να παρουσιάζουν όσο το δυνατόν μεγαλύτερη ομοιότητα ενώ ταυτόχρονα οι ομάδες να είναι όσο το δυνατόν πιο απομακρυσμένες μεταξύ τους. Η μεθοδολογία είναι η εξής:

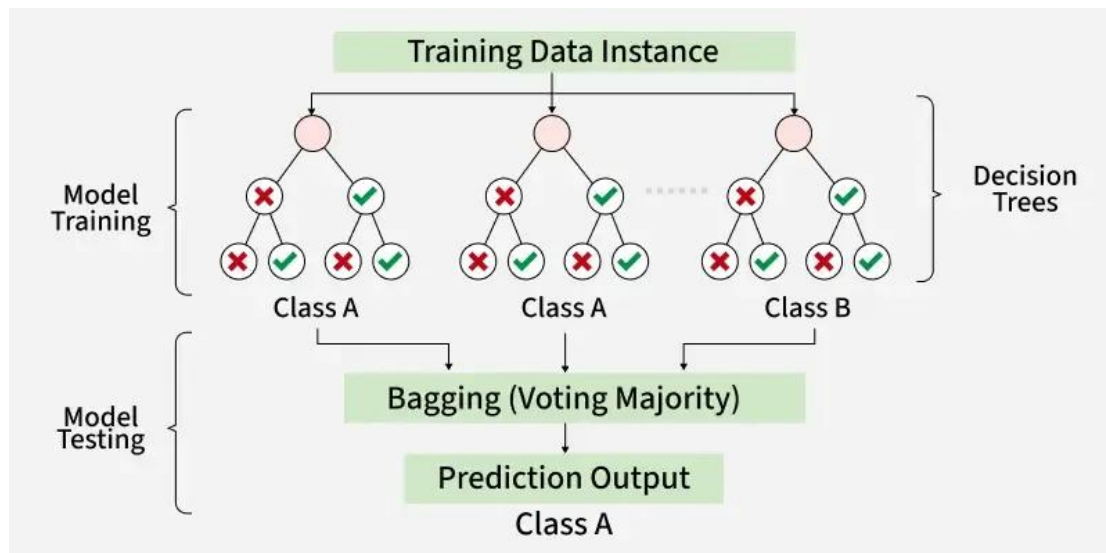
- 1) Καθορισμός του αριθμού των ομάδων (K).
- 2) Αρχική εκτίμηση των κεντροειδών των συστάδων.
- 3) Εξέταση και ανάθεση κάθε σημείου στην ομάδα με το πλησιέστερο κεντροειδές.
- 4) Με την ολοκλήρωση της ανάθεσης κάθε σημείου στην συστάδα του, υπολογίζεται εκ νέου τα K κεντροειδή.
- 5) Τα βήματα (3) και (4) επαναλαμβάνονται έως ότου κανένα σημείο δεν αλλάζει ομάδα.

Όπως διαπιστώθηκε η επιλογή του αριθμού των ομάδων αποτελεί κρίσιμο στάδιο για την σωστή εφαρμογή του αλγόριθμου K-Means. Οι Bholowalia & Kumar (2014) επισημαίνουν την μέθοδο του αγκώνα (Elbow Method). Είναι μια ευρέως χρησιμοποιούμενη τεχνική για τον υπολογισμό του κατάλληλου αριθμού K . Απεικονίζονται γραφικά οι τιμές του K ως προς το σφάλμα ομαδοποίησης, και το σημείο που η καμπύλη σχηματίζει κάμψη, υποδεικνύει τον βέλτιστο αριθμό συστάδων. Επισημαίνουν επίσης ότι για την αξιολόγηση των αριθμών και της

ποιότητας των ομάδων αξιοποιείται ο δείκτης Silhouette που συγκρίνει την ομοιότητα των σημείων της ομάδας μεταξύ τους.

3.6 ΑΛΓΟΡΙΘΜΟΣ ΤΥΧΑΙΑ ΔΑΣΗ (RANDOM FOREST)

Ο Random Forest είναι μία μέθοδος μηχανικής μάθησης που βασίζεται στην δημιουργία πολλών δέντρων απόφασης, τα οποία συνεργάζονται για την παραγωγή ακριβών και σταθερών αποτελεσμάτων. Σύμφωνα με τους Salman et al. (2024) στόχος του αλγόριθμου είναι η βελτίωση της ακρίβειας της πρόβλεψης και η μείωση του κινδύνου υπερεκπαίδευσης (overfitting). Η μεθοδολογία βασίζεται στην τυχαία επιλογή δεδομένων και χαρακτηριστικών για την εκπαίδευση κάθε δέντρου, ώστε τα δέντρα να είναι διαφορετικά μεταξύ τους και να μην επηρεάζονται από κανένα άλλο. Στην συνέχεια ο αλγόριθμος υπολογίζει πόσα δέντρα πρόβλεψαν κάθε κατηγορία και αυτή με την μεγαλύτερη συχνότητα εμφάνισης αποτελεί την τελική απόφαση του αλγόριθμου.



Γράφημα 3.1: Διαδικασία Ταξινόμησης Τυχαίου Δάσους Πηγή: [GeeksForGeeks](https://www.geeksforgeeks.org/random-forest/)

4. ΣΥΛΛΟΓΗ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑ ΣΤΟΙΧΕΙΩΝ

4.1 ΕΙΣΑΓΩΓΗ

Η ολοκλήρωση της βιβλιογραφικής ανασκόπησης και του θεωρητικού υποβάθρου, συντέλεσε στην κατανόηση των βασικών θεωρητικών αρχών και την επιλογή των κατάλληλων στατιστικών μοντέλων. Στο κεφάλαιο που ακολουθεί, πραγματοποιείται επεξήγηση της συλλογής και επεξεργασίας των δεδομένων για την υποστήριξη της μελέτης. Αρχικά γίνεται αναφορά στη συλλογή των δεδομένων, τα οποία παραχωρήθηκαν από την OSeven και προέρχονται από διασυνδεδεμένα οχήματα. Η ανάλυση των δεδομένων πραγματοποιήθηκε με την χρήση του λογισμικού R, με δημιουργία πινάκων και απαραίτητων επεξηγηματικών σχολίων. Συνεπώς, για την επίτευξη του στόχου της Διπλωματικής Εργασίας που είναι η διερεύνηση της οδηγικής συμπεριφοράς μέσω της **ανάλυσης δεδομένων** που συλλέχθηκαν από **διασυνδεδεμένα οχήματα**, η διαδικασία που θα ακολουθηθεί διαδραματίζει καθοριστικό ρόλο.

4.2 OSEVEN ΚΑΙ ΣΥΛΛΟΓΗ ΔΕΔΟΜΕΝΩΝ

Η OSeven είναι μια τεχνολογικά προηγμένη εταιρεία τηλεματικής. Προσφέρει μια πρωτοποριακή πλατφόρμα πρόληψης κινδύνου βασισμένη σε τεχνολογίες τεχνητής νοημοσύνης πλήρως συμβατή με κάθε είδος συσκευών (Smartphones και αισθητήρες οχημάτων). Η συλλογή και επεξεργασία των δεδομένων, δίνει την δυνατότητα για λήψη χρήσιμων πληροφοριών αναφορικά με την ενίσχυση της οδικής ασφάλειας και της οικολογικής οδήγησης. Στόχος της εταιρείας είναι ο εντοπισμός και διόρθωση των αδύναμων σημείων των οδηγών, καθώς και η ενίσχυση της οικονομικής και κοινωνικής αποδοτικότητας του οδικού δικτύου.



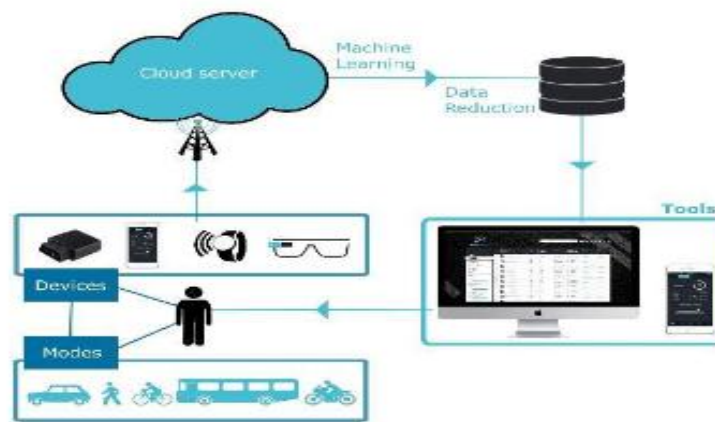
Εικόνα 4.1: Διάγραμμα ροής δεδομένων, Πηγή : [OSeven](#)

Η υποδομή της OSeven βασίζεται στην τεχνητή νοημοσύνη (AI). Το ερευνητικό έργο περιλαμβάνει τομείς όπως η ανάλυση της συμπεριφοράς, διερεύνηση οδικών ατυχημάτων και απόσπαση προσοχής κατά την οδήγηση. Επιπλέον εμπλέκονται πεδία όπως ο τομέας των συγκοινωνιών, μοντελοποίηση και επεξεργασία δεδομένων καθώς επίσης και Μηχανική Μάθηση. Έχει αναπτυχθεί ένα πλήρες σύνολο αλγορίθμων Μηχανικής Μάθησης, με εφαρμογές σε:

- Ανίχνευση επικίνδυνων ενεργειών και χρήση κινητού κατά την οδήγηση.
- Κατηγοριοποίηση μεταφορικού μέσου (αυτοκίνητο, μοτοσικλέτα, μέσα μαζικής μεταφοράς).

- Διάκριση όταν ο χρήστης είναι οδηγός ή επιβάτης.
- Χωροχρονική ανάλυση δεδομένων.
- Αξιολόγηση οικολογικής συμπεριφοράς οδηγού.
- Ανίχνευση σοβαρών συγκρούσεων.
- Ανίχνευση σοβαρότητας σύγκρουση.

Με την ολοκλήρωση της αποθήκευσης των δεδομένων, ακολουθεί η κεντρική επεξεργασία τους, με στόχο την μετατροπή τους σε ουσιώδεις πληροφορίες αναφορικά με την συμπεριφορά των οδηγών και την οδική ασφάλεια. Η OSeven χρησιμοποιεί προηγμένες τεχνικές κρυπτογράφησης για την διασφάλιση των δεδομένων σύμφωνα με την εθνική νομοθεσία και τις οδηγίες της Ευρωπαϊκής Ένωσης.



Εικόνα 4.2: Διάγραμμα διαχείρισης δεδομένων

Πηγή: [OSeven](#)



Εικόνα 4.3: Μοντέλο Αξιολόγησης

Πηγή: <https://www.nrso.ntua.gr/wp-content/uploads/OSeven-TelematicsWorkshop.pdf>

Στην παρούσα διπλωματική τα δεδομένα προέρχονται από αισθητήρες συνδεδεμένων οχημάτων. Δημιουργήθηκαν στο πλαίσιο ενός ερευνητικού προγράμματος (R&D) που ανέλαβε η OSeven με την ονομασία 07Connected. Στόχος του έργου ήταν η ανάπτυξη μίας νέας πλατφόρμας αντίστοιχης με την 07Platform η οποία θα μπορεί να αναλύει την οδηγική συμπεριφορά. Τα δεδομένα συλλέχθηκαν από δύο οχήματα που

ενοικιάστηκαν για περίοδο τριών μηνών και οδηγήθηκαν από τρεις διαφορετικούς υπάλληλους της εταιρείας. Γενικά, τα αυτοκίνητα παρείχαν πολλαπλά δεδομένα, επαρκή για την εκτέλεση αλγορίθμων εντοπισμού συμβάντων. Εκτός από την εκτέλεση των αλγορίθμων, τα δεδομένα μπορεί να αξιοποιηθούν και με άλλους τρόπους.

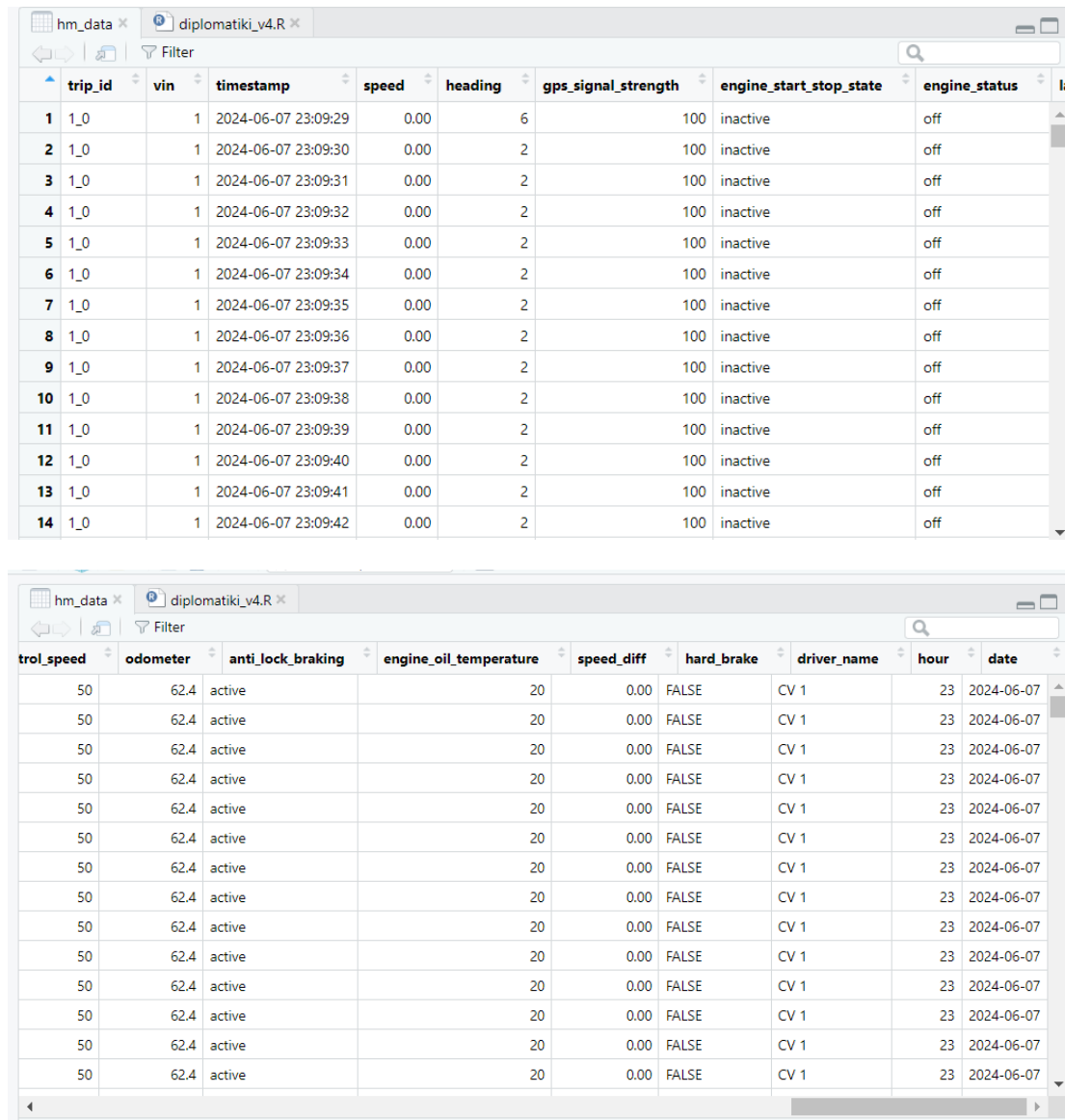
Τα δεδομένα που συλλέχθηκαν από τα διασυνδεδεμένα οχήματα είναι τα παρακάτω:

1. **Γεωγραφικό πλάτος, μήκος, κατεύθυνση.** Τα σήματα αυτά προέρχονται από το GPS και μεταδίδονται κάθε 6 δευτερόλεπτα.
2. **Ταχύτητα.** Το ίδιο όχημα αποτελεί την πηγή και όχι το GPS. Συχνότητα 1 Hz (1/δευτερόλεπτο).
3. **Οδόμετρο.** Έρχεται κάθε 1 λεπτό.
4. **Κατάσταση και λειτουργία κινητήρα.** Εμφανίζει πότε είναι σε λειτουργία ο κινητήρας.
5. **Δεδομένα καυσίμου.** Παρέχονται.
6. **Περιστατικά σύγκρουσης (εμπρός/πίσω/πλάγια) καθώς και είδος σύγκρουσης.** Πρόκειται για αισθητήρες που ανιχνεύουν συγκρούσεις σε κάθε πλευρά του αυτοκινήτου και εκπέμπουν κάθε λεπτό. Στα δικά μας δεδομένα όλες οι τιμές έχουν την ένδειξη repairs_not_needed.
7. **Σύστημα υποβοήθησης διατήρησης λωρίδας (Lane Keep Assist).** Για δεξιά/αριστερή λωρίδα και είτε ενεργή ή ανενεργή. Δεν ενεργοποιείται σε ταχύτητες μεγαλύτερες ή ίσες των 50 km/h.
8. **Υποστήριξη στάθμευσης με αισθητήρες εμπρός/πίσω.** Δείχνει την ενεργοποίηση park assist, ωστόσο είναι ενεργή πολλές φορές όταν η ταχύτητα είναι μεγαλύτερη των 20 km/h, οπότε δεν συνεπάγεται στάθμευση.
9. **Θερμοκρασία λαδιού κινητήρα.**
10. **Σύστημα μπλοκαρίσματος τροχών.** Είναι γνωστό ως ABS. Λαμβάνει τιμές «ενεργό» και «μη ενεργό» (active & inactive). Ενεργοποιείται πολλές φορές σε κάθε διαδρομή ακόμη και όταν το όχημα κινείται με σταθερή ταχύτητα. Αυτό σημαίνει ότι δεν ενεργοποιείται μόνο σε πολύ έντονα ή απότομα φρεναρίσματα αλλά και σε πιο ήπιες συνθήκες για την διατήρηση του ελέγχου του οχήματος.
11. **Cruise Control.** Σύστημα αυτόματου ελέγχου.

4.3 ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

Στην παρούσα υποενότητα παρουσιάζεται συνοπτικά η διαδικασία εισαγωγής των μεταβλητών στο λογισμικό στατιστικής ανάλυσης. Επίσης παρατίθενται ορισμένα διαγράμματα που δημιουργήθηκαν στη φάση της προκαταρκτικής ανάλυσης, με στόχο την καλύτερη κατανόηση των δεδομένων. Το τελικό στάδιο της επεξεργασίας των δεδομένων είναι η ανάπτυξη ενός μοντέλου αξιολόγησης της οδηγικής συμπεριφοράς.

Παρακάτω παρατίθεται ένα τμήμα της βάσης δεδομένων, με ονομασία `hm_data`, για καλύτερη κατανόηση του περιεχομένου και δομής του αρχείου.



trip_id	vin	timestamp	speed	heading	gps_signal_strength	engine_start_stop_state	engine_status
1	1_0	1 2024-06-07 23:09:29	0.00	6	100	inactive	off
2	1_0	1 2024-06-07 23:09:30	0.00	2	100	inactive	off
3	1_0	1 2024-06-07 23:09:31	0.00	2	100	inactive	off
4	1_0	1 2024-06-07 23:09:32	0.00	2	100	inactive	off
5	1_0	1 2024-06-07 23:09:33	0.00	2	100	inactive	off
6	1_0	1 2024-06-07 23:09:34	0.00	2	100	inactive	off
7	1_0	1 2024-06-07 23:09:35	0.00	2	100	inactive	off
8	1_0	1 2024-06-07 23:09:36	0.00	2	100	inactive	off
9	1_0	1 2024-06-07 23:09:37	0.00	2	100	inactive	off
10	1_0	1 2024-06-07 23:09:38	0.00	2	100	inactive	off
11	1_0	1 2024-06-07 23:09:39	0.00	2	100	inactive	off
12	1_0	1 2024-06-07 23:09:40	0.00	2	100	inactive	off
13	1_0	1 2024-06-07 23:09:41	0.00	2	100	inactive	off
14	1_0	1 2024-06-07 23:09:42	0.00	2	100	inactive	off

trol_speed	odometer	anti_lock_braking	engine_oil_temperature	speed_diff	hard_brake	driver_name	hour	date
50	62.4	active	20	0.00	FALSE	CV 1	23	2024-06-07
50	62.4	active	20	0.00	FALSE	CV 1	23	2024-06-07
50	62.4	active	20	0.00	FALSE	CV 1	23	2024-06-07
50	62.4	active	20	0.00	FALSE	CV 1	23	2024-06-07
50	62.4	active	20	0.00	FALSE	CV 1	23	2024-06-07
50	62.4	active	20	0.00	FALSE	CV 1	23	2024-06-07
50	62.4	active	20	0.00	FALSE	CV 1	23	2024-06-07
50	62.4	active	20	0.00	FALSE	CV 1	23	2024-06-07
50	62.4	active	20	0.00	FALSE	CV 1	23	2024-06-07
50	62.4	active	20	0.00	FALSE	CV 1	23	2024-06-07
50	62.4	active	20	0.00	FALSE	CV 1	23	2024-06-07
50	62.4	active	20	0.00	FALSE	CV 1	23	2024-06-07
50	62.4	active	20	0.00	FALSE	CV 1	23	2024-06-07
50	62.4	active	20	0.00	FALSE	CV 1	23	2024-06-07
50	62.4	active	20	0.00	FALSE	CV 1	23	2024-06-07

Εικόνα 4.4: Απόσπασμα Βάσης Δεδομένων στην R

4.3.1 ΕΙΣΑΓΩΓΗ ΒΑΣΗΣ ΔΕΔΟΜΕΝΩΝ ΣΤΟ ΛΟΓΙΣΜΙΚΟ ΣΤΑΤΙΣΤΙΚΗΣ ΑΝΑΛΥΣΗΣ

Για το αρχικό στάδιο της στατιστικής επεξεργασίας των δεδομένων, πραγματοποιήθηκε εισαγωγή του αρχείου CSV, στο πρόγραμμα της στατιστικής ανάλυσης R μέσω της εντολής `read_csv`. Ενδεικτικό τμήμα του κώδικα που χρησιμοποιήθηκε για την εμφάνιση των αποτελεσμάτων παρουσιάζεται στην παρακάτω εικόνα:



```
1 ##Diploma Thesis##
2
3 ##install useful packages
4 install.packages(c("ggplot2", "dplyr", "lubridate"))
5 install.packages("writexl")
6 library(readr)
7 ##read dataset
8 hm_data <- read_csv("C:/Users/User/OneDrive/Υπολογιστής/Diploma_Thesis/hm_data.csv")
9 min_speed <- min(hm_data$speed, na.rm = TRUE)
10 max_speed <- max(hm_data$speed, na.rm = TRUE)
11 mean_speed <- mean(hm_data$speed, na.rm = TRUE)
12 median_speed <- median(hm_data$speed, na.rm = TRUE)
13 print(median_speed)
14 print(min_speed)
15 print(max_speed)
16 print(mean_speed)
17 str(hm_data)
18
19 ##summary statistics
20
21 library(dplyr)
22
23 trip_summary <- hm_data %>%
24   group_by(trip_id, vin) %>%
25   summarise(
26     trip_duration = difftime(max(timestamp), min(timestamp), units = "mins"),
27     avg_speed      = mean(speed, na.rm = TRUE),
28     max_speed      = max(speed, na.rm = TRUE),
29     sd_speed       = sd(speed, na.rm = TRUE),
30     avg_engine_temp = mean(engine_oil_temperature, na.rm = TRUE),
31     abs_count      = sum(anti_lock_braking == "active", na.rm = TRUE),
32     .groups = "drop"
33   )
34
35 library(writexl)
36
37 # Save the summary to an Excel file
38 write_xlsx(trip_summary, "trip_summary.xlsx")
39
40
```

```

41 library(ggplot2)
42 library(dplyr)
43 library(lubridate)
44
45 # Histogram of speed by driver (vin)
46 ggplot(hm_data, aes(x = speed, fill = factor(vin))) +
47   geom_histogram(binwidth = 5, alpha = 0.6, position = "identity") +
48   labs(title = "Speed Distribution per Driver",
49        x = "Speed (km/h)", y = "Frequency", fill = "Driver (VIN)") +
50   theme_minimal()
51
52 ggplot(trip_summary, aes(x = factor(vin), y = sd_speed, fill = factor(vin))) +
53   geom_boxplot(alpha = 0.7) +
54   labs(title = "Trip-Level Speed Variability per Driver",
55        x = "Driver (VIN)", y = "Speed SD (km/h)") +
56   theme_minimal() +
57   theme(legend.position = "none")
58
59 # Example for one trip
60 trip_example <- hm_data %>% filter(trip_id == "1_0")
61
62 ggplot(trip_example, aes(x = timestamp, y = speed)) +
63   geom_line(color = "grey50") +
64   geom_point(data = subset(trip_example, anti_lock_braking == "active"),
65             aes(x = timestamp, y = speed), color = "red", size = 1.5) +
66   labs(title = "Speed Timeline with ABS Activation - Trip 1_0",
67        x = "Time", y = "Speed (km/h)") +
68   theme_minimal()
69
70 hm_data %>%
71   mutate(hour = hour(timestamp)) %>%
72   group_by(vin, hour) %>%
73   summarise(mean_speed = mean(speed, na.rm = TRUE)) %>%
74   ggplot(aes(x = hour, y = factor(vin), fill = mean_speed)) +
75   geom_tile(color = "white") +
76   scale_fill_viridis_c(option = "plasma") +
77   labs(title = "Heatmap of Average Speed by Hour and Driver",
78        x = "Hour of Day", y = "Driver (VIN)", fill = "Avg Speed (km/h)") +
79   theme_minimal()
80
81
82 ##Statistical analysis
83 #Μέση Ταχύτητα ανά συνδεδεμένο όχημα
84 library(dplyr)
85 library(ggplot2)
86
87 # Δημιουργούμε ένα πεδίο με πιο φιλικό όνομα για τα 2 συνδεδεμένα οχήματα
88 avg_speed_per_driver <- hm_data %>%
89   mutate(driver_name = ifelse(vin == unique(vin)[1], "Οδηγός 1", "Οδηγός 2")) %>%
90   group_by(driver_name) %>%
91   summarise(mean_speed = mean(speed, na.rm = TRUE))
92
93 # Γράφημα
94 ggplot(avg_speed_per_driver, aes(x = driver_name, y = mean_speed, fill = driver_name)) +
95   geom_bar(stat = "identity") +
96   labs(title = "Avg Speed per Connected Vehicle",
97        x = "Connected Vehicle", y = "Avg Speed (km/h)") +
98   theme_minimal() +
99   theme(legend.position = "none")
100
101
102 ##Μέση θερμοκρασία Κινητήρα σε Heatmap
103 library(dplyr)
104 library(lubridate)
105 library(ggplot2)
106
107 hm_data %>%
108   mutate(hour = hour(timestamp),
109          driver_name = ifelse(vin == unique(vin)[1], "CV 1", "CV 2")) %>%
110   group_by(driver_name, hour) %>%
111   summarise(mean_engine_temp = mean(engine_oil_temperature, na.rm = TRUE), .groups = "drop") %>%
112   ggplot(aes(x = hour, y = driver_name, fill = mean_engine_temp)) +
113   geom_tile(color = "white") +
114   scale_fill_viridis_c(option = "plasma") +
115   labs(title = "Engine Average Temperature Heatmap by Hour and Connected Vehicle",
116        x = "Ωρα της Ημέρας", y = "VIN", fill = "AVG Temperature (°C)") +
117   theme_minimal()

```

Εικόνα 4.5:Πεδίο Εισαγωγής Δεδομένων στην R

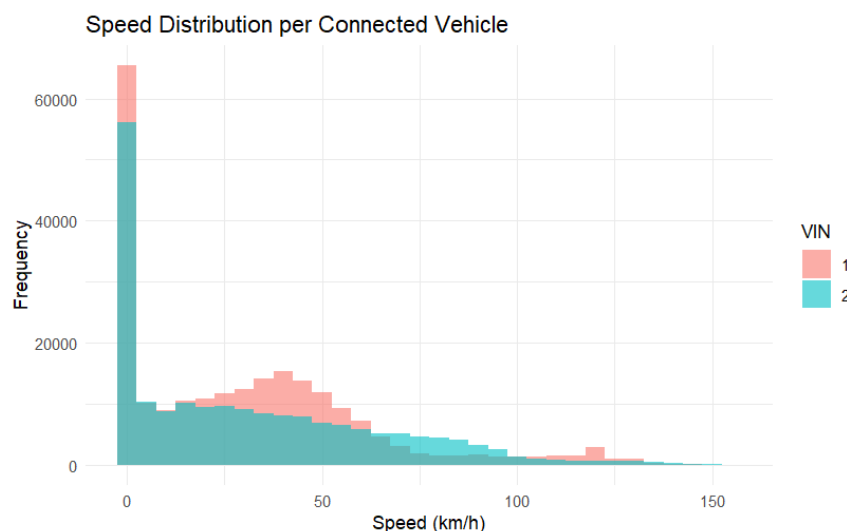
4.3.2 ΠΡΟΚΑΤΑΡΚΤΙΚΗ ΑΝΑΛΥΣΗ

Με την ολοκλήρωση των προαναφερθέντων βημάτων και την εκτέλεση του κώδικα, δημιουργήθηκαν χρήσιμα διαγράμματα που συμβάλλουν σε μια πρωταρχική κατανόηση των αποτελεσμάτων και στην εξαγωγή χρήσιμων συμπερασμάτων. Αρχικά παρατίθεται ο πίνακας με τα περιγραφικά στατιστικά των βασικών μεταβλητών που χρησιμοποιήθηκαν στην ανάλυση.

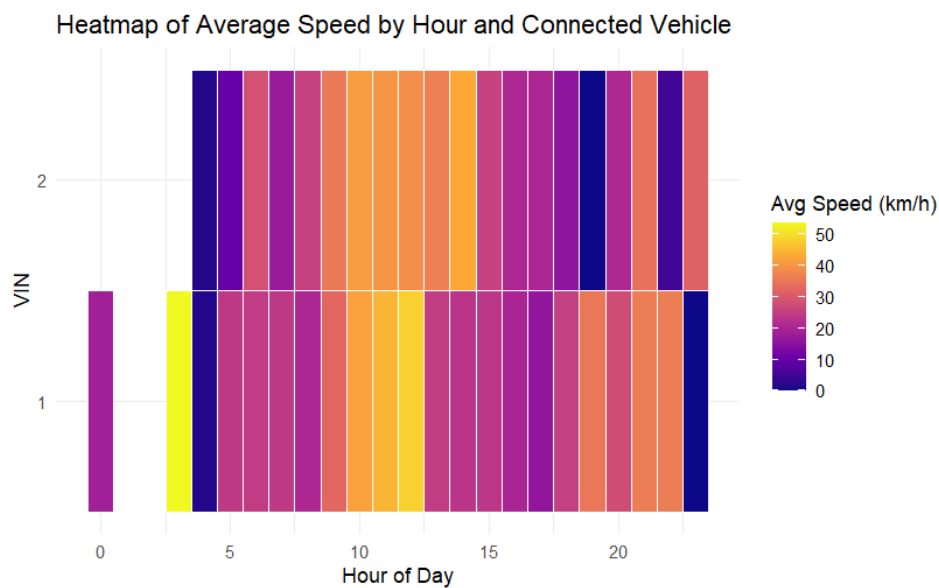
Πίνακας 4.1: Περιγραφικά Στατιστικά Ανά Συνδεδεμένο Όχημα

Vehicle	Variable	Speed (km/h)	Engine Temp. (°C)
CV1	Mean	30,2	76,8
CV1	SD	30,2	26,2
CV1	Min	0,0	20,0
CV1	Max	144,0	111,0
CV2	Mean	31,4	79,6
CV2	SD	32,0	25,6
CV2	Min	0,0	20,0
CV2	Max	155,0	117,0

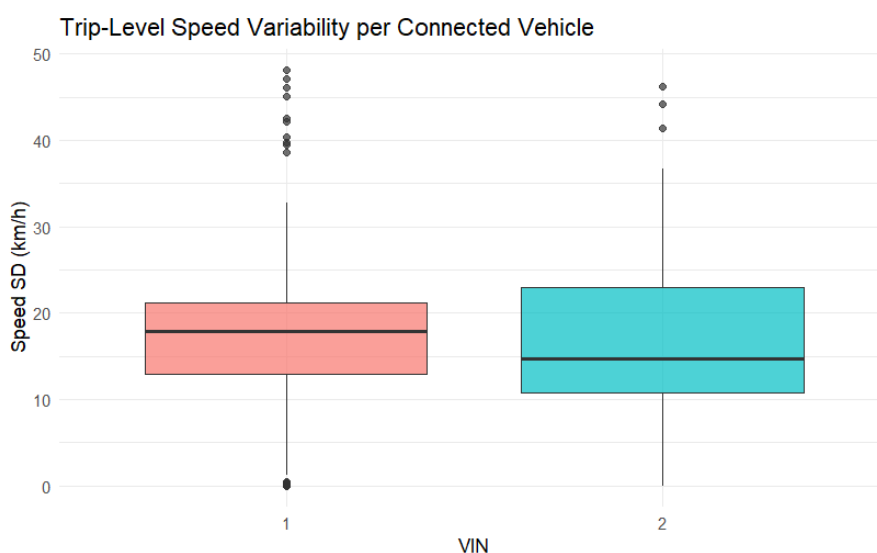
Παρακάτω παρουσιάζονται συνοπτικά τα σημαντικότερα διαγράμματα, σημειώνεται ότι ο όρος «VIN» αναφέρεται στον αριθμό αναγνώρισης των οχημάτων (Vehicle Identification Number).



Διάγραμμα 4.1: Διακύμανση της ταχύτητας ανά συνδεδεμένο όχημα



Διάγραμμα 4.2: Heatmap της μέσης ταχύτητας των συνδεδεμένων οχημάτων ανά ώρα κατά την διάρκεια της ημέρας

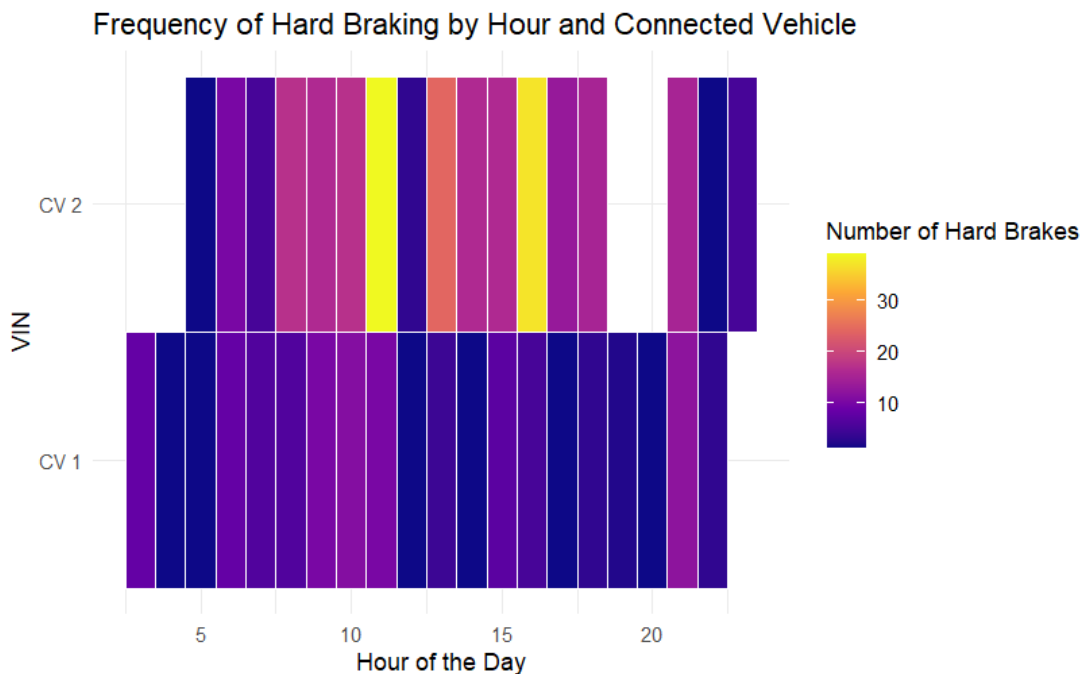


Διάγραμμα 4.3: Τυπική απόκλιση ταχύτητας ανά όχημα ανά διαδρομή

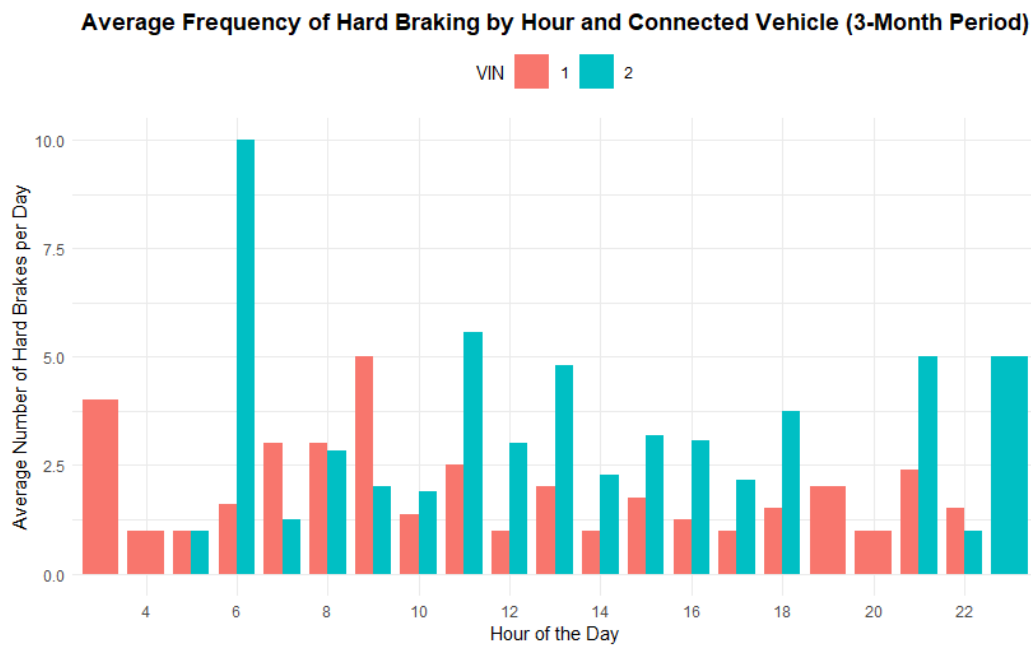
Εξετάζοντας τα ανωτέρω διαγράμματα διαπιστώνεται ότι:

- Το όχημα 1 κινείται συχνότερα με ταχύτητα μεταξύ 20-65 km/h, ενώ σε μερικές περιπτώσεις έχουν υπερβεί ταχύτητες μεγαλύτερες των 100 km/h.
- Το όχημα 2 κινείται σε μεγαλύτερο εύρος ταχυτήτων, από το όχημα 1, μεταξύ 20-100 km/h. Περιστασιακά έχει κινηθεί σε εξαιρετικά υψηλές ταχύτητες των 120-150 km/h.

- Από το διάγραμμα 4.2, παρατηρούνται χαμηλές μέσες τιμές κατά τις πρωινές (07:00-09:00) και απογευματινές (15:00-19:00) ώρες. Ώρες αιχμής δηλαδή που παρουσιάζουν χαμηλές τιμές ταχυτήτων λόγω της υψηλής κυκλοφοριακής συμφόρησης. Αντίθετα και τα δύο οχήματα καταγράφουν μεγαλύτερες μέσες τιμές κατά τις πρωινές προς μεσημεριανές (10:00-13:00) και νυχτερινές (21:00-24:00) ώρες, ενδεχομένως λόγω της μειωμένης κυκλοφορίας.
- Το πρώτο όχημα παρουσιάζει σχετικά υψηλή Speed SD (18 km/h), αυτό υποδεικνύει συχνές μεταβολές στην ταχύτητα (επιτάχυνση/επιβράδυνση) ή πιθανή οδήγηση σε συνθήκες κυκλοφοριακής συμφόρησης. Παρόμοια ισχύει και για το δεύτερο όχημα, που εμφανίζει διακυμάνσεις στην ταχύτητά του κατά περίπου 15 km/h. Η υψηλή Speed SD των αυτοκινήτων μπορεί να οφείλεται επίσης στην συχνότητα των απότομων φρεναρισμάτων. Πιο συγκεκριμένα σύμφωνα με το διάγραμμα 4.4 το όχημα 2 καταγράφει υψηλό πλήθος απότομων φρεναρισμάτων, εν αντιθέσει με το όχημα 1 που δεν παρουσιάζει απότομες αυξομειώσεις.
- Σημαντική παρατήρηση είναι ότι το όχημα 2 εμφανίζει τον μεγαλύτερο μέσο αριθμό απότομων φρεναρισμάτων στις 06:00 (διάγραμμα 4.5) ενώ σύμφωνα με το διάγραμμα 4.4, η μεγαλύτερη συχνότητα απότομων φρεναρισμάτων καταγράφεται τις μεσημβρινές ώρες και πιο συγκεκριμένα στις 11:00 και στις 16:00.

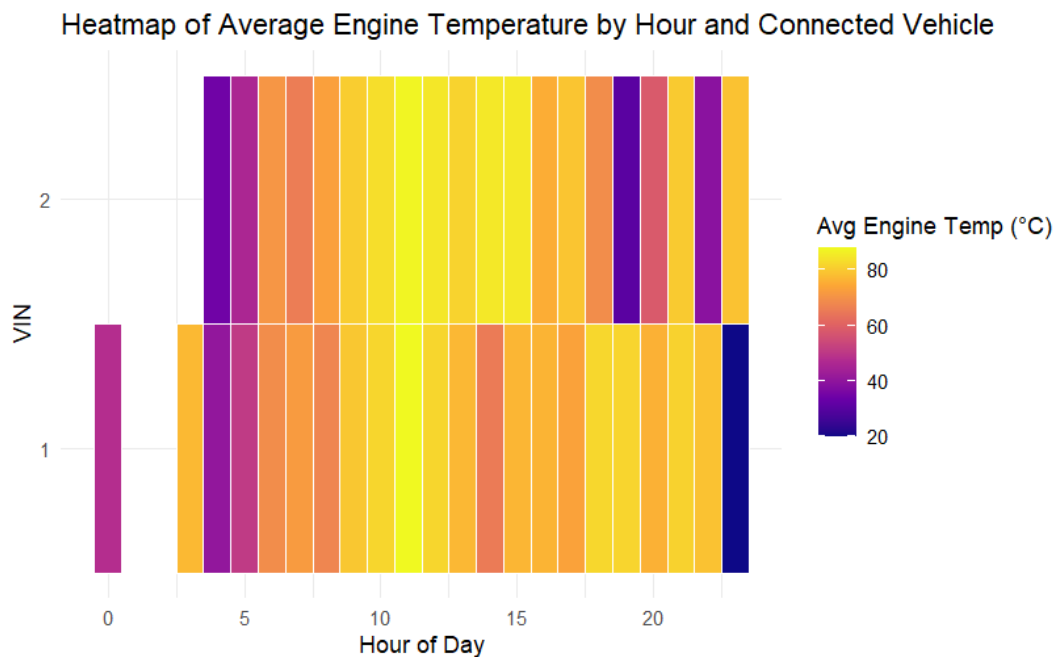


Διάγραμμα 4.4: Συχνότητα απότομων φρεναρισμάτων ανά ώρα ανά όχημα.



Διάγραμμα 4.5: Μέσος αριθμός απότομων φρεναρισμάτων ανά ώρα ανά όχημα.

Η ανάλυση των προηγούμενων διαγραμμάτων (μέση ταχύτητα, Speed SD, συχνότητα απότομων φρεναρισμάτων), σε συνδυασμό με τις εξωτερικές θερμοκρασίες στους θερινούς μήνες που συλλέχθηκαν τα δεδομένα, συμβάλουν στην αύξηση της θερμοκρασίας του κινητήρα. Το Heatmap της μέσης θερμοκρασίας του κινητήρα παρουσιάζει υψηλότερες τιμές όταν υπάρχει οδηγική δραστηριότητα κατά τη διάρκεια της ημέρας.



Διάγραμμα 4.6: Μέση θερμοκρασία κινητήρα ανά ώρα ανά όχημα.

5. ΕΦΑΡΜΟΓΗ ΜΕΘΟΔΟΛΟΓΙΑΣ - ΑΠΟΤΕΛΕΣΜΑΤΑ

Στο παρόν κεφάλαιο περιλαμβάνεται η αναλυτική παρουσίαση της μεθοδολογίας των μοντέλων που εφαρμόστηκαν και τα αποτελέσματά τους. Τα κεφάλαια της βιβλιογραφικής ανασκόπησης και του θεωρητικού υπόβαθρου συντέλεσαν σε ποιοτικότερη ανάλυση και επεξεργασία των δεδομένων. Επόμενο βήμα της Διπλωματικής Εργασίας αποτελεί η επιλογή της κατάλληλης μεθοδολογίας και του αντίστοιχου μοντέλου.

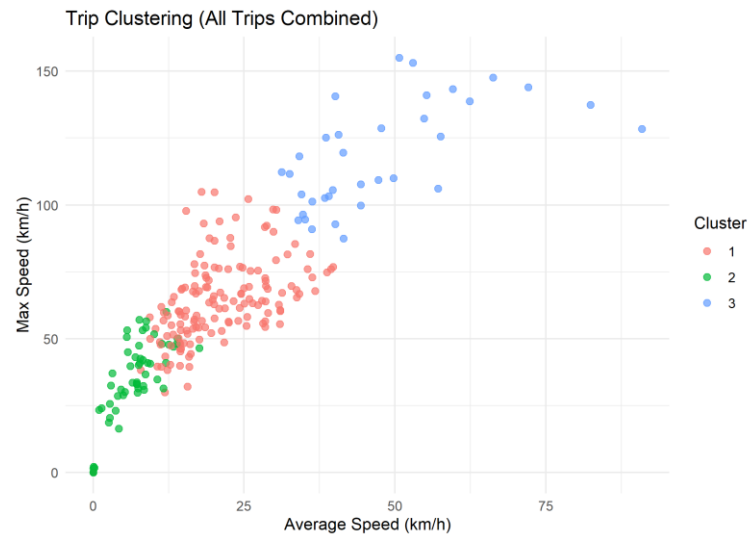
Για την ανάλυση των δεδομένων σε επίπεδο διαδρομών (trip level), θεωρήθηκε κατάλληλη μέθοδος αρχικής ανάλυσης η ομαδοποίηση των δεδομένων για εντοπισμό προφίλ ταξιδιών βάσει της οδηγικής συμπεριφοράς. Μετά από αρκετές δοκιμές, καταλληλότερο μοντέλο για ομαδοποίηση επιλέχθηκε ο αλγόριθμος **K-Means**. Στη συνέχεια, για περαιτέρω ανάλυση της οδηγικής συμπεριφοράς, επιλέχθηκε να διερευνηθεί ποιες μεταβλητές, συλλεχθέντες από τα διασυνδεδεμένα οχήματα, επηρεάζουν και σε ποιο βαθμό το σύστημα αντιμπλοκαρίσματος τροχών (ABS). Για το σκοπό αυτό επιλέχθηκε να χρησιμοποιηθεί η μέθοδος μηχανικής μάθησης Τυχαίων Δασών (Random Forest) για την αξιολόγηση της σημαντικότητας των μεταβλητών, ενώ στη συνέχεια χρησιμοποιήθηκε και η στατιστική μέθοδος της Διωνυμικής Λογιστικής Παλινδρόμησης με στόχο την περαιτέρω διερεύνηση και μοντελοποίηση της επιρροής των μεταβλητών.

Περιγράφονται αναλυτικά τα στάδια εφαρμογής των κατάλληλων μοντέλων δίνοντας έμφαση σε ζητήματα αξιοπιστίας των δεδομένων και στην αντιμετώπισή τους. Επιπλέον αναφορικά με την αποδοχή ή την απόρριψη μαθηματικών μοντέλων κρίσιμοι αποτελούν οι απαιτούμενοι στατιστικοί έλεγχοι.

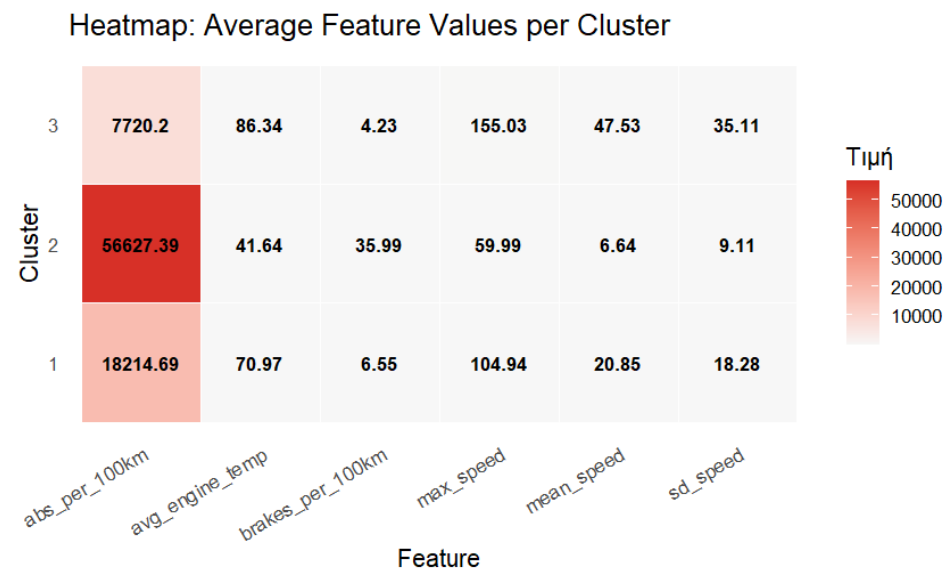
Τέλος, παρουσιάζεται η περιγραφή και ερμηνεία των αποτελεσμάτων των μεθοδολογιών, με βάση το συνολικό πεδίο που επικεντρώνεται η παρούσα Διπλωματική Εργασία.

5.1 ΕΦΑΡΜΟΓΗ ΑΛΓΟΡΙΘΜΟΥ K-MEANS

Όπως προέκυψε από τις έρευνες της βιβλιογραφικής ανασκόπησης διαπιστώθηκε ότι ο αλγόριθμος **K-Means** υπερτερεί έναντί των εναλλακτικών μεθόδων λόγω της ικανότητας του να ομαδοποιεί τις διαδρομές με κοινά χαρακτηριστικά. Στην παρούσα Διπλωματική Εργασία πραγματοποιήθηκε ταξινόμησή των διαδρομών για κάθε όχημα βάσει κύριων χαρακτηριστικών όπως: η μέση και μέγιστη ταχύτητα, διακύμανση της ταχύτητας, σύστημα αντιμπλοκαρίσματος τροχών και απότομα φρεναρίσματα.



Διάγραμμα: 5.1: Ταξινόμηση των διαδρομών βάσει της μέσης και μέγιστης ταχύτητας.



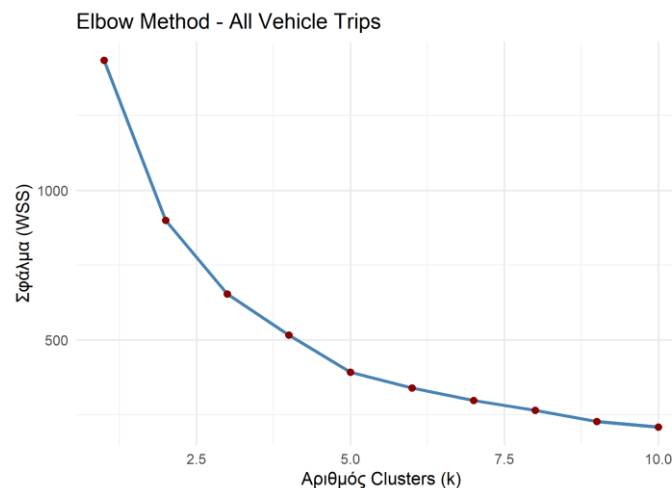
Διάγραμμα: 5.2: Heatmap Μέσων τιμών των χαρακτηριστικών ανά συστάδα.



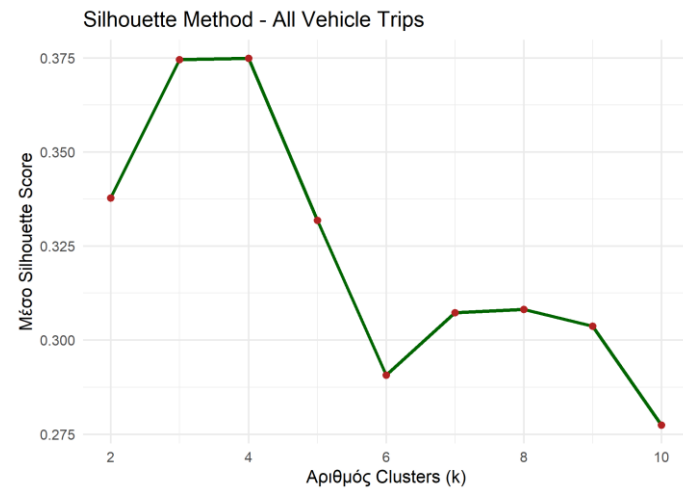
Διάγραμμα: 5.3: Κατανομή της Μέσης ταχύτητας ανά συστάδα.

5.1.1 ΠΟΙΟΤΗΤΑ ΜΟΝΤΕΛΟΥ

Μέσω της μεθόδου **Elbow** και **Silhouette** διαπιστώθηκε ότι η ταξινόμηση των διαδρομών σε τρεις ομάδες ($K=3$) παρέχει ικανοποιητική ανάλυση στα διαγράμματα και στην αξιολόγηση τους. Πιο συγκεκριμένα από την ανάλυση των διαγραμμάτων Elbow, διαπιστώθηκε ότι σαφές σημείο κάμψης συνολικά των διαδρομών και για τα δύο οχήματα είναι στο σημείο $K=3$. Επίσης οι τιμές των δεικτών Silhouette για $K=3$ έχουν την υψηλότερη τιμή υποδηλώνοντας ικανοποιητική συνοχή των ομάδων. Παρακάτω παρουσιάζονται τα γραφήματα των μεθόδων Elbow και Silhouette.



Διάγραμμα 5.4: Αποτέλεσμα μεθόδου Elbow για όλες τις διαδρομές των οχημάτων



Διάγραμμα 5.5: Αποτέλεσμα μεθόδου Silhouette για όλες τις διαδρομές των οχημάτων

5.1.2 ΣΧΟΛΙΑΣΜΟΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ K-MEANS

Από την ανάλυση και επεξεργασία των ανωτέρω διαγραμμάτων διαπιστώθηκαν κρίσιμα συμπεράσματα. Σημειώνεται πως το χρώμα των Cluster δεν εξαρτάται από την ταχύτητα των οχημάτων αλλά καθορίζεται από την συχνότητα εμφάνισης των διαδρομών. Πιο συγκεκριμένα το Cluster με κόκκινο χρώμα απαρτίζει την ομάδα με τις περισσότερες διαδρομές, ακολουθούν το πράσινο και τέλος το μπλε.

Αναφορικά με το διάγραμμα 5.3 το **Cluster 1** (κόκκινο χρώμα) εμφανίζεται με μεγαλύτερη συχνότητα έναντι των υπόλοιπων δυο. Δηλαδή η συστάδα αυτή, με ενδιάμεσες τιμές ταχυτήτων και μέτρια μεταβλητότητα, υποδηλώνει ένα **τυπικό προφίλ οδήγησης**. Στην συνέχεια το **Cluster 2** παρουσιάζει χαμηλές ταχύτητες με ελάχιστες διακυμάνσεις, αποτυπώνει δηλαδή ένα **προσεκτικό προφίλ οδήγησης**. Τέλος το **Cluster 3** χαρακτηρίζεται από υψηλές μέσες και μέγιστες ταχύτητες που αντιστοιχούν σε ένα **επιθετικό προφίλ οδήγησης**. Τα ανωτέρω επιβεβαιώνουν τα αποτελέσματα του διαγράμματος 5.5 (Μέσες τιμές των χαρακτηριστικών ανά συστάδα). Πιο συγκεκριμένα το **Cluster 2** που χαρακτηρίζεται από συντηρητική οδήγηση συγκεντρώνει τις **υψηλότερες** τιμές ενεργοποίησης του **συστήματος αντιμπλοκαρίσματος των τροχών**, ακολουθούν τα Cluster 1 και 3. Τέλος στο διάγραμμα 5.5 που αποτυπώνει την κατανομή της μέσης ταχύτητας των ομάδων, παρατηρείται σαφής διαχωρισμός μεταξύ των συστάδων και γενικότερα των προφίλ οδήγησης.

5.2 ΕΦΑΡΜΟΓΗ ΑΛΓΟΡΙΘΜΟΥ RANDOM FOREST

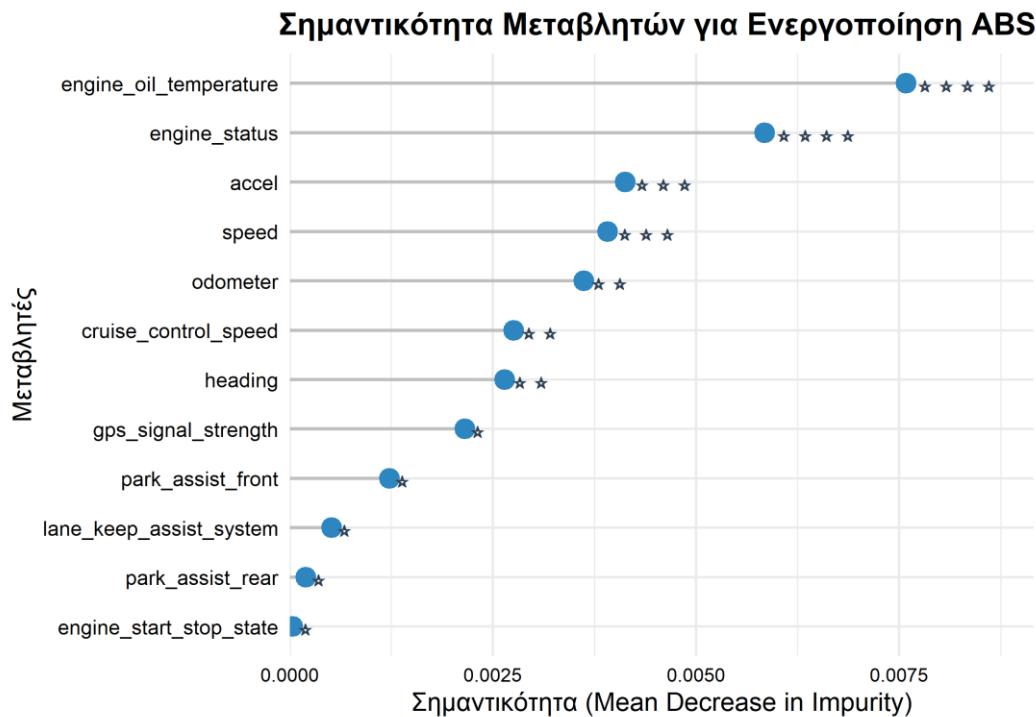
Ο αλγόριθμος Random Forest, μια μέθοδος μηχανικής μάθησης που στοχεύει, στην παρούσα έρευνα, στην αξιολόγηση της σημαντικότητας των μεταβλητών. Επιλέχθηκε η μεταβλητή που αφορά την ενεργοποίηση του **συστήματος αντιμπλοκαρίσματος τροχών (ABS)** ως εξαρτημένη μεταβλητή, καθώς αποτελεί αξιόπιστο δείκτη για την ανάλυση της οδηγικής συμπεριφοράς και συμβάλλει στην εξαγωγή ουσιαστικών συμπερασμάτων. Επισημαίνεται ότι πρόκειται για **δυναμική μεταβλητή** διότι λαμβάνει τιμές «ενεργό» και «μη ενεργό» (active & inactive).

5.2.1 ΕΚΠΑΙΔΕΥΣΗ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ RANDOM FOREST

Από το αρχείο `hm_data` αφαιρέθηκαν μεταβλητές όπως το οπίσθιο φως ομίχλης και κατάσταση σύγκρουσης, οι οποίες δεν σχετίζονται με το σύστημα μπλοκαρίσματος των τροχών. Για την εφαρμογή του αλγόριθμου Random Forest χρησιμοποιήθηκε η γλώσσα R και η βιβλιοθήκη `ranger`. Αρχικά μέσω της συνάρτησης `ranger()` πραγματοποιήθηκε η εκπαίδευση του μοντέλου, στην συνέχεια για την αξιολόγηση της σημαντικότητας κάθε μεταβλητής, εφαρμόστηκε η εντολή **`importance=permutation`** και τέλος ορίστηκε το πλήθος των δέντρων σε **300**. Παρακάτω παρατίθεται τμήμα του κώδικα του εκτελέστηκε και το διάγραμμα σημαντικότητας.

```
# Εκπαίδευση μοντέλου Ranger
set.seed(123)
rf_model <- ranger(
  anti_lock_braking ~ .,
  data = rf_data,
  importance = "permutation",
  num.trees = 300,
  classification = TRUE,
  num.threads = 2
)
```

Εικόνα 5.1: Απεικόνιση τμήματος του κώδικα εκπαίδευσης μοντέλου και σημαντικότητας μεταβλητών



Διάγραμμα: 5.6: Διάγραμμα σημαντικότητας μεταβλητών – Random Forest

5.2.2 ΠΟΙΟΤΗΤΑ ΜΟΝΤΕΛΟΥ RANDOM FOREST

Για την αξιολόγηση του μοντέλου Random Forest και την ορθότερη αξιολόγηση της σημαντικότητας των μεταβλητών, το σύνολο των δεδομένων χωρίστηκε σε σύνολο εκπαίδευσης 80% (training set) και ελέγχου 20% (test). Για τον υπολογισμό βασικών μετρικών αξιολόγησης του μοντέλου εφαρμόστηκε η εντολή **confusionMatrix()** της βιβλιοθήκης caret ενώ παράλληλα εξετάστηκε εκ νέου η σημαντικότητα των μεταβλητών με την εντολή **importance=permutation** της βιβλιοθήκης ranger. Στην συνέχεια λόγω της μεγάλης διαφοράς στη συχνότητα εμφάνισης των τιμών «ενεργό» και «μη ενεργό» της μεταβλητής, το μοντέλο εκπαιδεύτηκε με ενσωμάτωση Class Weights και ρύθμιση του ορίου απόφασης στο 0,9 για την βελτίωση της ικανότητας διάκρισης του, τα αποτελέσματα του κώδικα απεικονίζονται παρακάτω:

Πίνακας 5.1: Αποτελέσματα μετρικών δεικτών

Random Forest	
Accuracy	0,981
Sensitivity	0,985
Specificity	0,654

Η ορθότητα (Accuracy) κυμαίνεται σε υψηλά επίπεδα 98% συνεπώς είναι ικανοποιητικό. Επίσης το μοντέλο με ευαισθησία 98,5% εντοπίζει εύκολα τότε το σύστημα αντιμπλοκαρίσματος των τροχών ενεργοποιείται. Τέλος, μετά από μια σειρά δοκιμών του ορίου απόφασης η ειδικότητα παρουσίασε μέτριες προς υψηλές τιμές (65,4 %), δηλαδή αναγνωρίζει ένα ικανοποιητικό μέρος της μη ενεργοποίησης του συστήματος αντιμπλοκαρίσματος των τροχών.

5.2.3 ΣΧΟΛΙΑΣΜΟΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ RANDOM FOREST

Από την ανάλυση του διαγράμματος 5.6 διαπιστώνεται ότι οι μεταβλητές που αφορούν τη **θερμοκρασία του λαδιού του κινητήρα**, την **κατάσταση κινητήρα**, την **επιτάχυνση** και την **ταχύτητα** παρουσιάζουν, κατά σειρά, την μεγαλύτερη συμβολή στην πρόβλεψη της μεταβλητής. Στην συνέχεια ακολουθούν με μικρότερη αλλά σημαντική επίδραση στο μοντέλο ο χιλιομετρητής, η ταχύτητα αυτόματου ελέγχου και η κατεύθυνση. Τέλος οι μεταβλητές όπως η υποβοήθηση στάθμευση εμπρός και πίσω αλλά και το σύστημα διατήρησης λωρίδας παρουσίασαν μηδαμινή επιρροή στο μοντέλο.

5.3 ΕΦΑΡΜΟΓΗ ΔΙΩΝΥΜΙΚΗΣ ΛΟΓΙΣΤΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Όπως επισημάνθηκε παραπάνω η μεταβλητή του συστήματος αντιμπλοκαρίσματος τροχών είναι δυαδικής μορφής, καθώς λαμβάνει τιμές «ενεργό» και «μη ενεργό». Συνεπώς, επιλέχθηκε η μέθοδος της Διωνυμικής Λογιστικής Παλινδρόμησης. Η ανάλυση πραγματοποιήθηκε μέσω της γλώσσας R, χρησιμοποιώντας τη συνάρτηση `glm()` και ορίστηκε ως οικογένεια κατανομής τη `binomial`. Τονίζεται ότι πραγματοποιήθηκαν δοκιμές με πολλές μεταβλητές και επιλέχθηκε το κατάλληλο μοντέλο. Παρακάτω παρατίθεται τμήμα του κώδικα εφαρμόστηκε για την εξαγωγή των αποτελεσμάτων.

```
# Επιλογή μεταβλητών
selected_vars <- c(
  "engine_oil_temperature",
  "engine_status",
  "accel",
  "speed",
  "odometer",
  "cruise_control_speed",
  "heading_sin",
  "heading_cos"
)

# Δημιουργία dataset
model_df <- hm_data %>%
  select(anti_lock_braking, all_of(selected_vars)) %>%
  na.omit()

#
# Μοντέλο Διωνυμικής Λογιστικής Παλινδρόμησης
formula_log <- as.formula(
  paste("anti_lock_braking ~", paste(selected_vars, collapse = " + "))
)

model_log <- glm(formula_log, data = model_df, family = binomial)
```

Εικόνα 5.2 Απεικόνιση τμήματος του κώδικα για την εφαρμογή Διωνυμικής Λογιστικής Παλινδρόμησης

5.3.1 ΣΤΑΤΙΣΤΙΚΟΣ ΈΛΕΓΧΟΣ ΜΟΝΤΕΛΟΥ

Για την αποδοχή του μοντέλου και των αποτελεσμάτων είναι απαραίτητος ο στατιστικός έλεγχος του μοντέλου, ο οποίος θα πραγματοποιηθεί κατά την εκτέλεση του κώδικα στην εξαγωγή των μαθηματικών μοντέλων. Αρχικά παρουσιάζεται ο παρακάτω πίνακας που αποτυπώνονται οι σημαντικότερες μεταβλητές, οι αριθμητικές τιμές των συντελεστών των μεταβλητών (estimate), καθώς επίσης και οι p.value και Odds Ratio. Κατά την επιλογή των μεταβλητών, υιοθετήθηκε επίπεδο σημαντικότητας 95%, συνεπώς, οποιαδήποτε μεταβλητή με p.value μεγαλύτερη του 0,05 θεωρείται μη στατιστικά σημαντική για το μοντέλο. Το Odds Ratio προκύπτει από τη σχέση **Odds Ratio**= e^{estimate} και εκφράζει πόσες φορές αλλάζει το Odds για κάθε αύξηση μίας μονάδας της ανεξάρτητης μεταβλητής. Έπειτα για τον έλεγχο της πολυγγραμμικότητας αξιοποιήθηκε ο δείκτης VIF. Όπως προκύπτει από τον πίνακα 5.2, οι τιμές VIF των μεταβλητών κυμαίνονται μεταξύ του 1 έως 2, γεγονός που υποδηλώνει απουσία πολυγγραμμικότητας και θεωρούνται αποδεκτές για το τελικό μοντέλο.

Τέλος για την αξιολόγηση της εφαρμογής του μοντέλου της Διωνυμικής Λογιστικής Παλινδρόμησης, υπολογίστηκαν οι δείκτες Pseudo-R². Πιο συγκεκριμένα οι τιμές **McFadden R²=0,154** και **Cox & Snell=0,163** υποδηλώνουν ότι το μοντέλο παρουσιάζει ικανοποιητική προσαρμογή στα δεδομένα.

Πίνακας 5.2: Αποτελέσματα επιρροής των μεταβλητών στο μοντέλο

term	estimate	p.value	Odds_Ratio	Σημαντικότητα
(Intercept)	2,526	<0,001	12,499	<0,001
engine_oil_temperature	0,001	0,204	1,001	-
engine_status (on)	1,300	<0,001	3,671	<0,001
accel	-0,049	0,199	0,952	-
speed	0,056	<0,001	1,058	<0,001
odometer	0,000	0,078	1,000	-
cruise_control_speed	0,002	0,007	1,002	0,007
heading_sin	-0,076	<0,001	0,927	<0,001
heading_cos	-0,101	<0,001	0,904	<0,001

Πίνακας 5.3: Τιμές VIF των μεταβλητών

Variable	VIF
engine_oil_temperature	1,960
engine_status	2,033
accel	1,005
speed	1,183
odometer	1,006
cruise_control_speed	1,008
heading_sin	1,003
heading_cos	1,009

5.4 ΑΠΟΤΕΛΕΣΜΑΤΑ

Σε αυτή την ενότητα παρατίθεται η ερμηνεία των αποτελεσμάτων για την ανάπτυξη ενός μοντέλου πρόβλεψης ενεργοποίησης του συστήματος αντιμπλοκαρίσματος τροχών. Όπως αναφέρθηκε στην ενότητα 5.3.1, οι σημαντικές μεταβλητές που θα περιληφθούν στην εξίσωση χρησιμότητας είναι :

- Ο συντελεστής επίδρασης (estimate).
- Ενεργοποίηση κινητήρα.
- Ταχύτητα.
- Ταχύτητα αυτόματου ελέγχου
- Κατεύθυνση σε συνάρτηση γωνιακών μεταβλητών.

Συνεπώς η συνάρτηση χρησιμότητας που παρήγαγε αυτό το μοντέλο είναι:

$$U_{ABS}=2,526+1,3*engine_status+0,056*speed+0,002*cruise_control_speed-0,076*heading_sin-0,101*heading_cos$$

$$P(ABS=1)=1/(1+e^{-U_{ABS}})$$

Από την ανωτέρω συνάρτηση και με την βοήθεια των Odds Ratio, καθορίζονται τα εξής συμπεράσματα:

- Ο **σταθερός συντελεστής** της συνάρτησης παρουσιάζει **θετική τιμή** (2,526) γεγονός που δηλώνει ότι ακόμα και όλες οι μεταβλητές λάβουν την τιμή 0, υπάρχει η πιθανότητα ενεργοποίησης του συστήματος αντιμπλοκαρίσματος τροχών. Συνεπώς, ένα πρωταρχικό συμπέρασμα προκύπτει ότι η ενεργοποίηση του συστήματος δεν εξαρτάται μόνο από τις μεταβλητές αλλά και από άλλους μη μετρήσιμους παράγοντες (π.χ. οι καιρικές συνθήκες την περίοδο που συλλέχθηκαν τα δεδομένα ή μικρά συμβάντα κατά την διάρκεια των διαδρομών).

- Η **ενεργοποίηση του κινητήρα** αυξάνει κατά **3,7 φορές περισσότερο** την πιθανότητα ενεργοποίησης του συστήματος αντιμπλοκαρίσματος των τροχών. Η αύξηση αυτή πιθανόν προκύπτει από το γεγονός ότι όταν ο κινητήρας είναι ενεργός προκαλεί υποβοήθηση στα φρένα και υπάρχει πλήρης ηλεκτρική λειτουργία των αισθητήρων και γενικά του συστήματος. Επίσης τα οχήματα τελευταίας τεχνολογίας, όπως είναι τα διασυνδεδεμένα οχήματα, με σκοπό την εξοικονόμηση ενέργειας και κατανάλωσης καυσίμου, απενεργοποιούν τον κινητήρα σε ειδικές περιπτώσεις όπως για παράδειγμα όταν το όχημα κινείται σε οδό με κατηφορική κλίση. Σε αυτήν την περίπτωση μπορεί το όχημα να παρουσιάζει υψηλές ταχύτητες αλλά το σύστημα αντιμπλοκαρίσματος των τροχών να είναι ανενεργό.
- Η **αύξηση της ταχύτητας κατά 1 km/h** προκαλεί **αύξηση κατά 5,8%** στην πιθανότητα ενεργοποίησης του συστήματος αντιμπλοκαρίσματος των τροχών. Η αιτία της αύξησης του ποσοστού εντοπίζεται στη φύση του ίδιου του συστήματος αντιμπλοκαρίσματος των τροχών. Πιο συγκεκριμένα η αύξηση της ταχύτητας προκαλεί άνοδο στη κινητική ενέργεια του οχήματος, με αποτέλεσμα να υπάρχει μεγαλύτερη διάρκεια και δύναμη στη πέδηση, προκαλώντας μπλοκάρισμα στους τροχούς. Επίσης, το σύστημα αυτό ενεργοποιείται ακόμα και σε μικρής διάρκειας πέδησης όταν το όχημα κινείται με εξαιρετικά υψηλές ταχύτητες.
- Μικρή επίδραση έχει στο μοντέλο η μεταβλητή της **ταχύτητας αυτόματου ελέγχου**. Πιο συγκεκριμένα **αύξηση 1 km/h** της συγκεκριμένης ταχύτητας προκαλεί **αύξηση κατά 0,2%** στην πιθανότητα ενεργοποίησης του συστήματος αντιμπλοκαρίσματος των τροχών. Η ταχύτητα αυτόματου ελέγχου δημιουργεί μια σταθερή και ελεγχόμενη οδήγηση που δεν δημιουργεί απότομες μεταβολές στην πέδηση που θα μπλοκάρουν τους τροχούς. Επίσης δεν δημιουργούνται απότομες επιβραδύνσεις που τείνουν να μπλοκάρουν τους τροχούς.
- Οι μεταβλητές **heading_cos** και **heading_sin** έχουν αρνητικούς συντελεστές. Γεγονός που φανερώνει ότι για κάθε μονάδα αύξησης των μεταβλητών προκαλείται μείωση της πιθανότητας ενεργοποίησης του συστήματος αντιμπλοκαρίσματος των τροχών. Πιο συγκεκριμένα **μειώνεται** η πιθανότητα του γεγονότος κατά **7,3%** και **9,6%** αντίστοιχα. Δεν υπάρχουν γεωγραφικές πληροφορίες αλλά η συσχέτιση αυτή πιθανόν οφείλεται στην μορφολογία της περιοχής ή τους τύπους των οδών που κινούνταν τα οχήματα. Για παράδειγμα υπάρχει η πιθανότητα η πλειονότητα των διαδρομών να πραγματοποιήθηκαν σε οδούς με ανηφορική κλίση ή με μεγάλο ευθύγραμμο τμήμα. Σε αυτές τις περιπτώσεις μειώνεται η πιθανότητα ενεργοποίησης του συστήματος αντιμπλοκαρίσματος των τροχών λόγω ότι μειώνεται η πέδηση στους εμπρός τροχούς.

6. ΣΥΜΠΕΡΑΣΜΑΤΑ

6.1 ΣΥΝΟΨΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

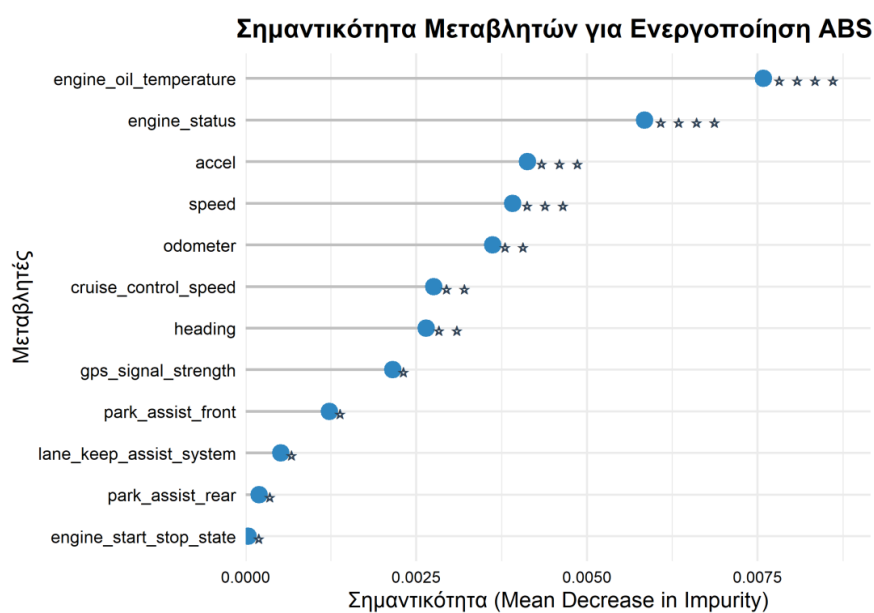
Αντικείμενο της παρούσας η Διπλωματικής Εργασίας αποτελεί η **διερεύνηση της οδηγικής συμπεριφοράς μέσω της ανάλυσης δεδομένων που συλλέχθηκαν από διασυνδεδεμένα οχήματα**. Ειδικότερα, εξετάστηκαν και αναλύθηκαν μεταβλητές που σχετίζονται με την ταχύτητα, τη συχνότητα φρεναρισμάτων, τη θερμοκρασία κινητήρα, το σύστημα αντιμπλοκαρίσματος τροχών κλπ. προκειμένου να εντοπιστούν διαφορετικά προφίλ διαδρομών ως προς την συμπεριφορά του οδηγού, να εξεταστούν οι παράγοντες επιρροής της ενεργοποίησης συστημάτων ασφαλείας, καθώς και ο βαθμός επιρροής τους.

Πιο συγκεκριμένα, τα δεδομένα προέρχονται από αισθητήρες συνδεδεμένων οχημάτων δημιουργήθηκαν στο πλαίσιο ενός ερευνητικού προγράμματος (R&D) που ανέλαβε η OSeven με την ονομασία 07Connected. Στόχος του έργου ήταν η ανάπτυξη μίας νέας πλατφόρμας η οποία θα μπορεί να αναλύει την οδηγική συμπεριφορά. Τα δεδομένα συλλέχθηκαν από δύο οχήματα που ενοικιάστηκαν για περίοδο τριών μηνών και οδηγήθηκαν από τρεις διαφορετικούς υπάλληλους της εταιρείας. Γενικά, τα αυτοκίνητα παρείχαν πολλαπλά δεδομένα, επαρκή για την εκτέλεση αλγορίθμων εντοπισμού συμβάντων.

Μετά από ενδελεχή μελέτη και μια σειρά δοκιμών αναπτύχθηκαν τα κατάλληλα μαθηματικά μοντέλα. Αρχικά, εφαρμόζοντας τον αλγόριθμο K-Means πραγματοποιήθηκε ταξινόμησή των διαδρομών για κάθε όχημα βάσει κύριων χαρακτηριστικών όπως η μέση και μέγιστη ταχύτητα. Στη συνέχεια, μέσω του αλγόριθμου Τυχαία Δάση και θέτοντας ως εξαρτημένη μεταβλητή το σύστημα αντιμπλοκαρίσματος των τροχών, ελέγχθηκε η σημαντικότητα των μεταβλητών. Τέλος, αξιοποιώντας την Διωνυμική Λογιστική Παλινδρόμηση και ύστερα από μια σειρά δοκιμών καθορίστηκε η τελική συνάρτηση χρησιμότητας του μοντέλου. Τα κύρια αποτελέσματα των μαθηματικών μοντέλων παρουσιάζονται παρακάτω.

Πίνακας: 6.1: Πίνακας μέσων τιμών των βασικών χαρακτηριστικών ανά συστάδα.

cluster	max_speed	sd_speed	avg_engine_temp	abs_per_100km	brakes_per_100km
1	104,94	18,28	70,97	18215	7
2	59,99	9,11	41,64	56627	36
3	155,03	35,11	86,34	7720	4



Διάγραμμα: 6.2: Διάγραμμα σημαντικότητας μεταβλητών – Random Forest

Πίνακας 6.1: Αποτελέσματα μετρικών δεικτών

Random Forest	
Accuracy	0,981
Sensitivity	0,985
Specificity	0,654

Πίνακας 6.2:Συνολικά αποτελέσματα του μοντέλου Διωνυμικής Λογιστικής Παλινδρόμησης

term	estimate	p.value	Odds_Ratio	Σημαντικότητα	VIF
(Intercept)	2,526	<0,001	12,499	<0,001	-
engine_oil_temperature	0,001	0,204	1,001	-	1,960
engine_status (on)	1,300	<0,001	3,671	<0,001	2,033
accel	-0,049	0,199	0,952	-	1,005
speed	0,056	<0,001	1,058	<0,001	1,183
odometer	0,000	0,078	1,000	-	1,006
cruise_control_speed	0,002	0,007	1,002	0,007	1,008
heading_sin	-0,076	<0,001	0,927	<0,001	1,003
heading_cos	-0,101	<0,001	0,904	<0,001	1,009
Pseudo-R ²	McFadden=0,154		Cox & Snell=0,163		

6.2 ΣΥΝΟΨΗ ΣΥΜΠΕΡΑΣΜΑΤΩΝ

Κατά την εκπόνηση της παρούσας Διπλωματικής Εργασίας προέκυψαν χρήσιμα συμπεράσματα που συνδέονται άμεσα με τους αρχικούς στόχους της. Τα κύρια συμπεράσματα συνοψίζονται ως εξής:

- Μέσω της μεθόδου Elbow και Silhouette διαπιστώθηκε ότι η ταξινόμηση των διαδρομών σε 3 ομάδες για την εφαρμογή του αλγόριθμου **K-Means** παρέχει ικανοποιητική ανάλυση στα διαγράμματα. Πιο συγκεκριμένα, εξετάζοντας τα διαγράμματα προέκυψε ότι μεγαλύτερη συχνότητα παρουσιάζει η συστάδα με ενδιάμεσες τιμές ταχυτήτων που υποδηλώνουν ένα τυπικό προφίλ οδήγησης. Επίσης η ομάδα που χαρακτηρίζεται από συντηρητική οδήγηση εμφανίζει υψηλότερες τιμές ενεργοποίησης του συστήματος αντιμπλοκαρίσματος των τροχών.
- Για την ανάλυση της σημαντικότητας των μεταβλητών ορίστηκε ως εξαρτημένη μεταβλητή το σύστημα αντιμπλοκαρίσματος των τροχών και μέσω του αλγόριθμου **Τυχαία Δάση** παρατηρήθηκε ότι οι μεταβλητές που αφορούν τη **θερμοκρασία του λαδιού του κινητήρα**, την **κατάσταση κινητήρα**, την **επιτάχυνση** και την **ταχύτητα** παρουσιάζουν, κατά σειρά, την **μεγαλύτερη συμβολή** στην πρόβλεψη της. Επίσης η ορθότητα (Accuracy) κυμαίνεται σε υψηλά επίπεδα 98% συνεπώς είναι ικανοποιητικό και έχει ευαισθησία 98,5%, δηλαδή εντοπίζει εύκολα πότε το σύστημα αντιμπλοκαρίσματος των τροχών ενεργοποιείται. Τέλος η ειδικότητα παρουσίασε μέτριες προς υψηλές τιμές (65,4 %).
- Η εφαρμογή της **Διωνυμικής Λογιστικής Παλινδρόμησης** έδειξε ότι η ενεργοποίηση του συστήματος αντιμπλοκαρίσματος των τροχών δεν εξαρτάται μόνο από τις υπάρχουσες μεταβλητές αλλά και από άλλους μη μετρήσιμους παράγοντες (π.χ. οι καιρικές συνθήκες την περίοδο που συλλέχθηκαν τα δεδομένα ή μικρά συμβάντα κατά την διάρκεια των διαδρομών).
- Η ενεργοποίηση του κινητήρα αυξάνει την πιθανότητα ενεργοποίησης του συστήματος αντιμπλοκαρίσματος των τροχών. Ο ενεργός κινητήρας θέτει τα συστήματα ασφαλείας του οχήματος σε πλήρη ηλεκτρική λειτουργία. Συνεπώς υπάρχει αυξημένη πιθανότητα ενεργοποίησης του συστήματος αντιμπλοκαρίσματος των τροχών.
- Η αύξηση της ταχύτητας συμβάλλει στην πιθανότητα ενεργοποίησης του συστήματος αντιμπλοκαρίσματος των τροχών. Πιο συγκεκριμένα η αύξηση της ταχύτητας προκαλεί άνοδο στη κινητική ενέργεια του οχήματος, με αποτέλεσμα να υπάρχει μεγαλύτερη διάρκεια και δύναμη στην πέδηση, προκαλώντας μπλοκάρισμα στους τροχούς. Επίσης το σύστημα αυτό

ενεργοποιείται ακόμα και σε μικρής διάρκειας πέδησης όταν το όχημα κινείται με εξαιρετικά υψηλές ταχύτητες.

- Μικρή επίδραση έχει στο μοντέλο η μεταβλητή της ταχύτητας αυτόματου ελέγχου. Πιο συγκεκριμένα αύξηση της συγκεκριμένης ταχύτητας προκαλεί αύξηση της πιθανότητα ενεργοποίησης του συστήματος αντιμπλοκαρίσματος των τροχών. Η ταχύτητα αυτόματου ελέγχου δημιουργεί μια σταθερή και ελεγχόμενη οδήγηση που δεν δημιουργεί απότομες μεταβολές στην πέδηση που θα μπλοκάρουν τους τροχούς.
- Η μορφολογία του εδάφους και ο τύπος των οδών επηρεάζει στην πιθανότητα ενεργοποίησης του συστήματος. Πιο συγκεκριμένα αν οι διαδρομές πραγματοποιούνται σε οδούς με ανηφορική κλίση ή με μεγάλο ευθύγραμμο τμήμα μειώνεται η πιθανότητα ενεργοποίησης του συστήματος αντιμπλοκαρίσματος των τροχών.
- Συνολικά, διαπιστώθηκε ότι εξήχθησαν χρήσιμα συμπεράσματα αναφορικά με την οδηγική συμπεριφορά. Οι μέθοδοι ταξινόμησης και οι στατιστικές αναλύσεις βοήθησαν στην περαιτέρω μελέτη και κατανόηση του προφίλ οδήγησης μέσω της ανάλυσης δεδομένων από συνδεδεμένο όχημα.

6.3 ΠΡΟΤΑΣΕΙΣ ΓΙΑ ΒΕΛΤΙΩΣΗ ΤΗΣ ΟΔΙΚΗΣ ΑΣΦΑΛΕΙΑΣ

Αξιοποιώντας τα αποτελέσματα και τα συνολικά συμπεράσματα της παρούσας Διπλωματικής Εργασίας, πραγματοποιείται η παράθεση μια σειράς προτάσεων που ενδεχομένως δύναται να συμβάλλουν στη βελτίωση της οδικής ασφάλειας. Πιο συγκεκριμένα:

- **Βελτίωση της υπάρχουσας οδικής υποδομής.** Με αυτόν τον τρόπο θα ενισχυθεί η ασφάλεια της κυκλοφορίας των οχημάτων. Επίσης σε παρόμοιες έρευνες θα υπάρχει βελτίωση στην εξαγωγή των δεδομένων καθώς τα δεδομένα που εξάγονται από τους αισθητήρες δεν θα επηρεάζονται από δονήσεις ή από λακκούβες της οδού.
- **Εναισθητοποίηση της πολιτείας και των εκπαιδευτικών φορέων.** Το άτομο από μικρή ηλικία θα πρέπει να αντιλαμβάνεται το ζήτημα της ασφάλειας και της ευθύνης του οδηγού. Είναι σημαντικό κατά τη διάρκεια των σχολικών ετών να εντάσσονται σε δραστηριότητες με θέμα την οδική ασφάλεια και τους κινδύνους που έχει ένας οδηγός όταν παραβιάζει τον Κώδικα Οδικής Κυκλοφορίας.
- **Αξιοποίηση ερευνών** που σχετίζονται με τον εντοπισμό φαινομένων επιθετικής συμπεριφοράς κατά τη διάρκεια της οδήγησης και να παρασχεθούν συμβουλές για την βελτίωση των οδηγών.

- Για την επίτευξη του περιορισμού των παραβιάσεων του Κώδικα Οδικής Κυκλοφορίας απαιτείται **συχνότερη και αυστηρότερη αστυνόμευση**. Με αυτόν το τρόπο θα επιτευχθεί ένα ασφαλές περιβάλλον όχι μόνο για τους οδηγούς αλλά και για τους υπόλοιπους χρήστες της οδού (π.χ. πεζούς ή χρήστες ηλεκτρικών πατινιών).
- Τέλος, οι έρευνες που αναλύουν την οδηγική συμπεριφορά με απώτερο σκοπό την οδική ασφάλεια θα πρέπει να πραγματοποιούνται με μεγαλύτερη συχνότητα.

6.4 ΠΡΟΤΑΣΕΙΣ ΓΙΑ ΠΕΡΑΙΤΕΡΩ ΈΡΕΥΝΑ

Για την περαιτέρω μελέτη του αντικειμένου της παρούσας Διπλωματικής Εργασίας ενδιαφέρον θα παρουσίαζε η εμβάθυνση στα παρακάτω ζητούμενα:

- **Μεγαλύτερο δείγμα οδηγών ή συνδεδεμένων οχημάτων.** Παρόλο που η παρούσα διπλωματική εργασία βασίστηκε στις καταγραφές των διαδρομών των οχημάτων, η αύξηση του δείγματος των οδηγών ή το συνδεδεμένων οχημάτων θα οδηγούσε σε πιο αξιόπιστα αποτελέσματα.
- Θα μπορούσαν να μελετηθούν ή να πραγματοποιηθούν παρόμοιες έρευνες που παρέχουν πληροφορίες για τις **συνθήκες κυκλοφορίας σε διαφορετικούς τύπους οδού**.
- Θα μπορούσε να πραγματοποιηθεί αντίστοιχη ερευνά αναλύοντας **διαφορετικές μεταβλητές** χρησιμοποιώντας διαφορετικούς αλγόριθμους μηχανικής μάθησης.
- Έρευνα με εναρμόνιση των μεταβλητών με **γεωγραφικές πληροφορίες** (π.χ. το όχημα επιταχύνει λόγω ότι βρίσκεται σε οδό με κατηφορική κλίση).

7. ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΑΝΑΦΟΡΕΣ

- ALEXIS AIVALIOTIS (2023) Improving driving behaviour through OSeven's telematics application [TelematicsWorkshop-4April2023-OSeven-v4.pptx](#)
- Eleonora Papadimitriou & Anastasios Dragomanovits (2024) Road safety thematic report – Main factors causing fatal crashes https://road-safety.transport.ec.europa.eu/document/download/a7428369-8eaf-4032-806e-ea08b46028c0_en?filename=ERSO-TR-MainCauses.pdf
- Bholowalia, P., & Kumar, A. (2014). EBK-means: A clustering technique based on elbow method and k-means in WSN. International Journal of Computer Applications, 105(9).
- BSI- [Cooperative intelligent transport systems](#)
- Ghosh, S., & Dubey, S. K. (2013). Comparative analysis of k-means and fuzzy c-means algorithms. International Journal of Advanced Computer Science and Applications, 4(4)
- Hahsler, M., Piekenbrock, M., & Doran, D. (2019). dbSCAN: Fast density-based clustering with R. Journal of Statistical Software, 91, 1-30
- Hong, Z., Chen, Y., & Wu, Y. (2020). A driver behavior assessment and recommendation system for connected vehicles to produce safer driving environments through a “follow the leader” approach. Accident Analysis & Prevention, 139, 105460.
- Koliou, P. Data-Driven Urban Road Safety Classification Integrating Telematics, Machine Learning, and Spatial Analysis.
- Mobini Seraji, M. H., Shaffiee Haghshenas, S., Shaffiee Haghshenas, S., Simic, V., Pamucar, D., Guido, G., & Astarita, V. (2025). A state-of-the-art review on machine learning techniques for driving behavior analysis: clustering and classification approaches. Complex & Intelligent Systems, 11(9), 1-28.
- Mohammadnazar, A., Arvin, R., & Khattak, A. J. (2021). Classifying travelers' driving style using basic safety messages generated by connected vehicles: Application of unsupervised machine learning. Transportation research part C: emerging technologies, 122, 102917
- Navada, A., Ansari, A. N., Patil, S., & Sonkamble, B. A. (2011, June). Overview of use of decision tree algorithms in machine learning. In 2011 IEEE control and system graduate research colloquium (pp. 37-42). IEEE.
- Noble, W. S. (2006). What is a support vector machine?. Nature biotechnology, 24(12), 1565-1567.
- OSeven oseven.io | [The No.1 customer-centric telematics platform](#)
- Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random forest algorithm overview. Babylonian Journal of Machine Learning, 2024, 69-79.ε

- Shladover, S. E. (2018). Connected and automated vehicle systems: Introduction and overview. *Journal of Intelligent Transportation Systems*, 22(3), 190-200.
- Uhlemann, E. (2015). Introducing connected vehicles [connected vehicles]. *IEEE vehicular technology magazine*, 10(1), 23-31.
- Zheng, J., and H. X. Liu. Estimating Traffic Volumes for Signalized Intersections using Connected Vehicle Data. *Transportation Research Part C: Emerging Technologies*, Vol. 79, 2017, pp. 347–362
- Zhou, Y., & Bridgelall, R. (2020). Review of usage of real-world connected vehicle data. *Transportation Research Record*, 2674(10), 939-950.
- Ziakopoulos, A., Karahlis, N., & Yannis, G (2025). Leveraging naturalistic Connected LCV data for spatial surrogate safety measure applications.
- Ziakopoulos, A., & Yannis, G. (2020). A review of spatial approaches in road safety. *Accident Analysis & Prevention*, 135, 105323.
- Ελληνική Στατιστική Αρχή (ΕΛΣΤΑΤ) (2023). Οδικά τροχαία ατυχήματα <https://www.statistics.gr/el/statistics/-/publication/SDT04/->
- Ειδικές Διακριτές Κατανομές. Οικονομικό Πανεπιστήμιο Αθηνών
- Ζωγράφος, Κ., & Τσαϊρίδης, Χ. (2024). Στατιστική και Στοιχεία Πιθανοτήτων.
- Μπαγατέλας, Α. (2020). Η λογιστική παλινδρόμηση για γεγονότα με χαμηλή συχνότητα εμφάνισης (Master's thesis, Πανεπιστήμιο Πειραιώς).
- Οδική Ασφάλεια, Συμπεριφορά Οδηγού και Ατυχήματα
- Οικονομού, S.(2020). Αλγόριθμοι μηχανικής μάθησης για την εξόρυξη δεδομένων.
- Παρμακσίζογλου, Η. Α. (2019). Χωρική ανάλυση συμπεριφοράς ασφάλειας οδηγών με δεδομένα από έξυπνα κινητά τηλέφωνα. (Diploma Thesis)
- Τζουτζούλης, Δ. Μ. (2020). Κρίσιμοι παράγοντες επιρροής της υπέρβασης των ορίων ταχύτητας με δεδομένα έξυπνων κινητών τηλεφώνων.(Diploma Thesis)
- Φραντζόλα, Ε. Κ. (2020). Χωρο-χρονική ανάλυση της ασφαλείας κυκλοφορίας με τη χρήση δεδομένων από τους αισθητήρες έξυπνων κινητών τηλεφώνων (Diploma Thesis)