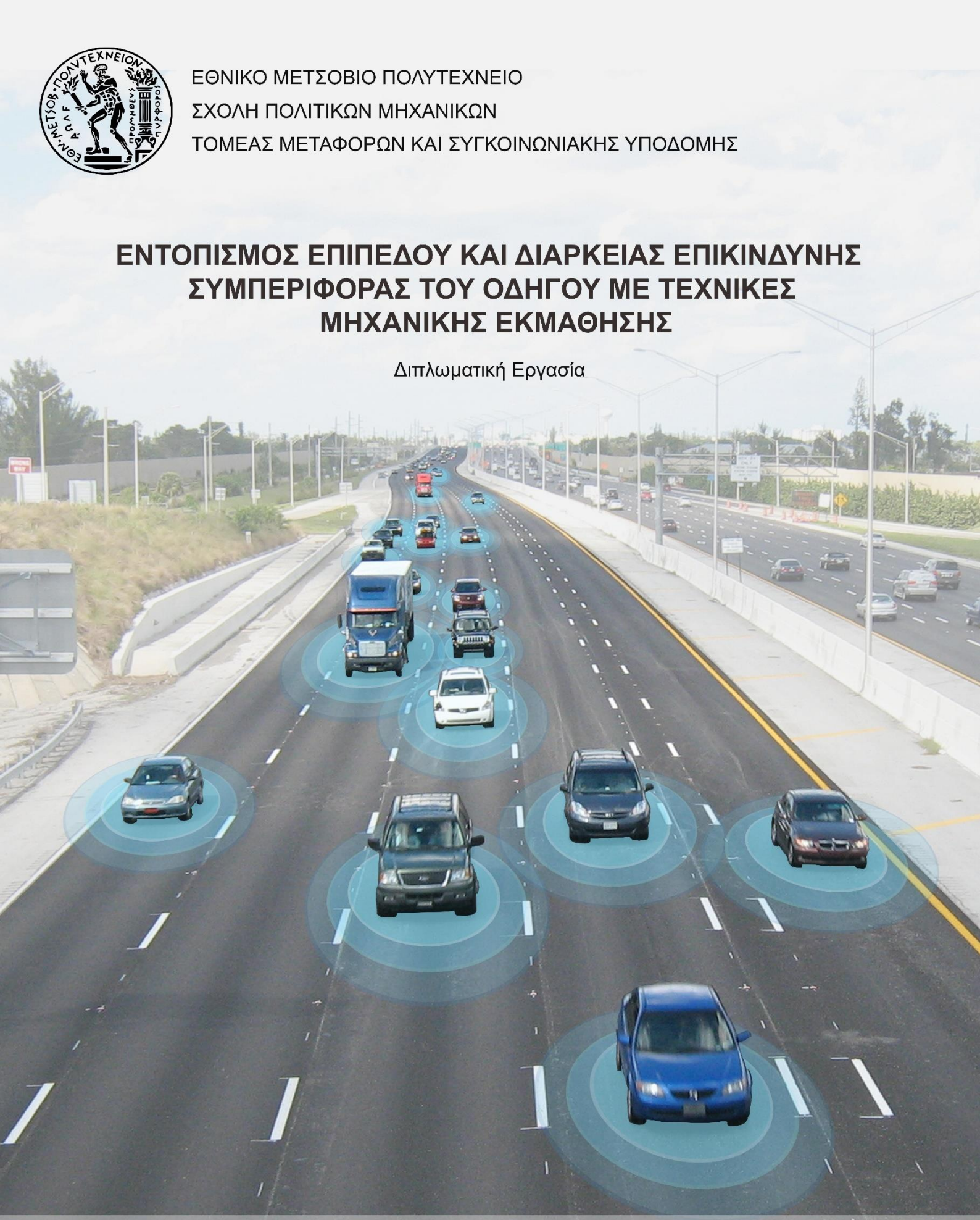
ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΠΟΛΙΤΙΚΩΝ ΜΗΧΑΝΙΚΩΝ
ΤΟΜΕΑΣ ΜΕΤΑΦΟΡΩΝ ΚΑΙ ΣΥΓΚΟΙΝΩΝΙΑΚΗΣ ΥΠΟΔΟΜΗΣ

ΕΝΤΟΠΙΣΜΟΣ ΕΠΙΠΕΔΟΥ ΚΑΙ ΔΙΑΡΚΕΙΑΣ ΕΠΙΚΙΝΔΥΝΗΣ ΣΥΜΠΕΡΙΦΟΡΑΣ ΤΟΥ ΟΔΗΓΟΥ ΜΕ ΤΕΧΝΙΚΕΣ ΜΗΧΑΝΙΚΗΣ ΕΚΜΑΘΗΣΗΣ

Διπλωματική Εργασία



Θεόδωρος Γαρεφαλάκης

Επιβλέπων: Γιώργος Γιαννής, Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2022

Ευχαριστίες

Με την ολοκλήρωση της παρούσας Διπλωματικής Εργασίας ολοκληρώνεται ο κύκλος των προπτυχιακών σπουδών μου στην Σχολή Πολιτικών Μηχανικών του Εθνικού Μετσόβιου Πολυτεχνείου.

Θα ήθελα πρωτίστως να ευχαριστήσω θερμά τον κ. Γ. Γιαννή, Καθηγητή της Σχολής Πολιτικών Μηχανικών ΕΜΠ, για την εμπιστοσύνη που μου έδειξε με την ανάθεση του θέματος, καθώς επίσης για την υποστήριξη και την καθοδήγηση σε όλα τα στάδια εκπόνησης της Διπλωματικής Εργασίας.

Επίσης θα ήθελα να ευχαριστήσω θερμά τον Δρ. Χρήστο Κατρακάζα, για την καθοριστική συνεισφορά του στη διεκπεραίωση της παρούσας εργασίας, μέσω των πολύτιμων συμβουλών και υποδείξεων του, καθώς και για το εξαιρετικό κλίμα συνεργασίας και επικοινωνίας που διαμόρφωσε.

Τέλος, θα ήθελα να ευχαριστήσω την Ανθή, την Λίλη, τους φίλους μου και κυρίως του γονείς μου που με στήριξαν καθ’ όλη τη διάρκεια των σπουδών μου.

Αθήνα, Μάρτιος 2022

Θεόδωρος Γαρεφαλάκης

Εντοπισμός επιπέδου και διάρκειας επικίνδυνης συμπεριφοράς του οδηγού με τεχνικές μηχανικής εκμάθησης

Θεόδωρος Γαρεφαλάκης

Επιβλέπων: Γιώργος Γιαννής, Καθηγητής Ε.Μ.Π.

Σύνοψη

Στόχος της παρούσας Διπλωματικής Εργασίας αποτελεί ο εντοπισμός του επιπέδου και της διάρκειας επικίνδυνης συμπεριφοράς του οδηγού με τεχνικές μηχανικής εκμάθησης. Για τον σκοπό αυτό συλλέχθηκαν χρήσιμα δεδομένα σχετιζόμενα με την συμπεριφορά του οδηγού μέσω προσομοιωτή οδήγησης. Με βάση την επεξεργασία και την ανάλυση των δεδομένων καθορίστηκαν τρία επίπεδα κινδύνου. Στο πρώτο μέρος των αναλύσεων αναπτύχθηκαν τέσσερεις αλγόριθμοι μηχανικής εκμάθησης με σκοπό την ταξινόμηση της συμπεριφοράς των οδηγών σε ένα από τα τρία επίπεδα ασφαλείας, με τον αλγόριθμο ‘Τυχαίων Δασών’ να σημειώνει την υψηλότερη επίδοση. Στο πλαίσιο διερεύνησης της επιρροής των παραγόντων οδήγησης στην αναγνώριση της επικίνδυνης οδήγησης, προέκυψαν ως σημαντικότερες η διανυθείσα απόσταση, η ταχύτητα και το όριο ταχύτητας. Στο δεύτερο μέρος των αναλύσεων εξετάστηκε η επίδραση των οδηγικών χαρακτηριστικών στη διάρκεια οδήγησης στα διαφορετικά στάδια με την ανάπτυξη τριών αλγορίθμων παλινδρόμησης για την πρόβλεψη της διάρκειας οδήγησης σε κάθε επίπεδο ασφαλείας. Η επίδραση των διαφορετικών μεταβλητών στη διαδικασία της πρόβλεψης καθορίστηκε με βάση τις επιδόσεις των μοντέλων και της στατιστικής σημαντικότητας τους. Από τα αποτελέσματα προέκυψε ως σημαντικότερη μεταβλητή η μέγιστη ταχύτητα η οποία επιδρά αρνητικά στην διάρκεια οδήγησης σε κάθε επίπεδο ασφαλείας.

Λέξεις κλειδιά: ανάλυση οδηγικής συμπεριφοράς, ταξινόμηση οδηγικής συμπεριφοράς, πρόβλεψη ατυχημάτων σε πραγματικό χρόνο, μηχανική μάθηση, μοντέλα ταξινόμησης, επιλογή χαρακτηριστικών, μη ισορροπημένο σύνολο δεδομένων, μέθοδοι επαναδειγματοληψίας, μοντέλα παλινδρόμησης, μηχανές διανυσμάτων υποστήριξης, τυχαία δάση, μοντέλο προσαρμοστικής ενδυνάμωσης, πολυεπίπεδο perceptron, παλινδρόμηση κορυφογραμμής, παλινδρόμηση lasso, παλινδρόμηση elastic net

Identification of driver's risky behavior level and duration with machine learning techniques

Theodoros Garefalakis

Supervisor: George Yannis, Professor NTUA

Abstract

The objective of this Thesis is the identification of driver's risky behavior level and duration with machine learning techniques. For this purpose, useful data related to driving behavior were collected through a driving simulator experiment. Based on the processing and analysis of the data, three levels of risk were defined. In the first part of the analysis, four machine learning algorithms were developed to classify driver behavior into one of three risk levels, with the 'Random Forests' algorithm scoring the highest performance. In the context of investigating the influence of driving factors to identify driving behavior, the distance traveled, speed and speed limit emerged as the most important. In the second part of the analysis, the effect of driving characteristics on driving duration at different stages was examined. To achieve the above goal, three regression algorithms were developed to predict driving duration at each safety level. The effect of different variables on the forecasting process was determined based on the performance of the models and their statistical significance. The results showed that the maximum speed was the most important variable, which negatively affects the driving duration at each safety level.

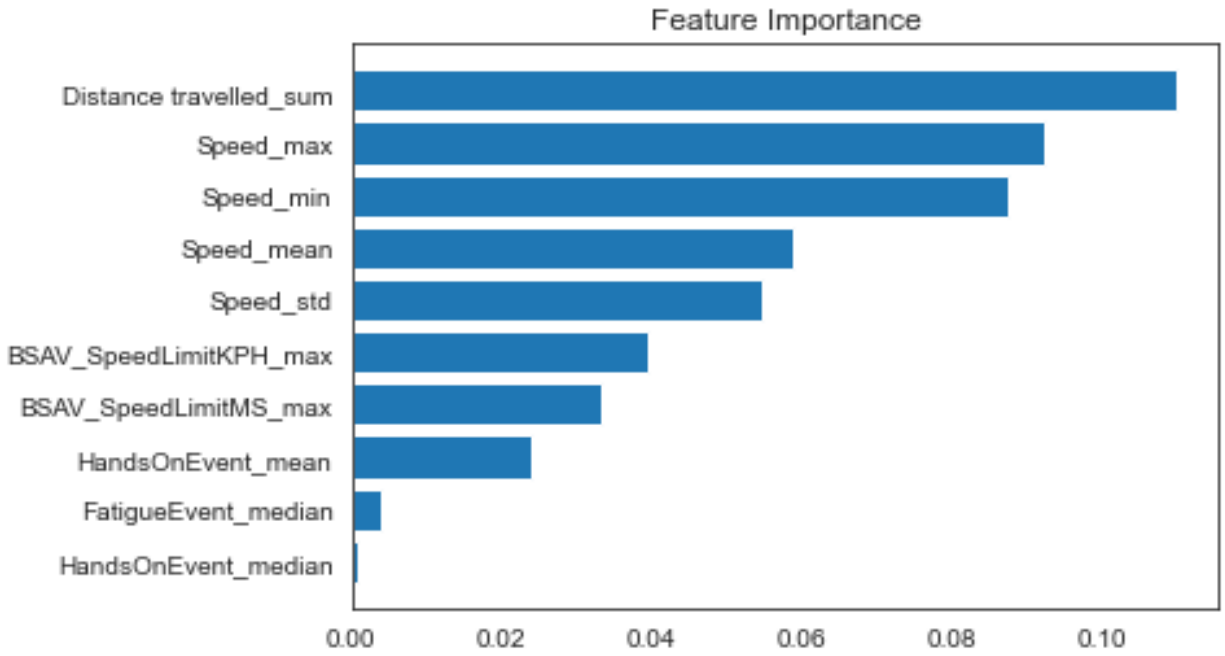
Key words: driving behavior analysis, driving behavior classification, real-time crash prediction, machine learning, classification models, feature selection, imbalanced dataset, resampling methods, regression models, support vector machines, random forests, adaboost, multilayer perceptron, ridge regression, lasso regression, elastic net regression

ΠΕΡΙΛΗΨΗ

Στόχος της παρούσας διπλωματικής εργασίας είναι **ο εντοπισμός του επιπέδου και της διάρκειας επικίνδυνης συμπεριφοράς του οδηγού (Ζώνης Ανοχής Ασφαλείας) με τεχνικές μηχανικής εκμάθησης**. Τα δεδομένα που αναλύθηκαν, συλλέχθηκαν από προσομοιωτή οδήγησης κατάλληλα διαμορφωμένο για το ερευνητικό έργο i-DREAMS. Για την ανάλυση της οδηγικής συμπεριφοράς ήταν αναγκαίο να οριστούν τα διαφορετικά επίπεδα της 'Ζώνης Ανοχής Ασφαλείας' βάσει ορισμένων τεχνικών. Τελικά ο καθορισμός των επιπέδων ασφαλείας πραγματοποιήθηκε με βάση την μεταβλητή Headway_min, καθώς η συγκεκριμένη τεχνική προσέφερε τη βέλτιστη, σύμφωνα με την βιβλιογραφία, κατανομή των δειγμάτων στα τρία επίπεδα:

- Επίπεδο 'Normal' (class: 0) : Headway_min > 2 δλ.
- Επίπεδο 'Dangerous' (class: 1) : Headway_min > 1.4 δλ. και Headway_min < 2 δλ.
- Επίπεδο 'Avoidable Accident' (class: 2) : Headway_min < 1.4 δλ.

Στο πρώτο μέρος των αναλύσεων αναπτύχθηκαν κατάλληλες τεχνικές προσδιορισμού της **σημαντικότητας των μεταβλητών στην πρόβλεψη του επιπέδου 'Ζώνης Ανοχής Ασφαλείας' που βρίσκεται ο οδηγός**. Επισημαίνεται ότι οι μεταβλητές Headway και TTC δεν λαμβάνονται υπόψη στο πρώτο μέρος των αναλύσεων καθώς θα αναπτύσσονταν προβλήματα μεροληψίας των μοντέλων ταξινόμησης. Η σημαντικότητα φαίνεται στο γράφημα που ακολουθεί.



Γράφημα 1: Σημαντικότητα μεταβλητών για την πρόβλεψη του επιπέδου 'Ζώνης Ανοχής Ασφαλείας'

Στη συνέχεια αξιοποιώντας τις σημαντικότερες μεταβλητές, αναπτύχθηκαν τέσσερεις αλγόριθμοι μηχανικής εκμάθησης με σκοπό **την ταξινόμηση της οδηγικής συμπεριφοράς σε ένα από τα τρία επίπεδα ασφαλείας**. Εφαρμόζοντας την 'Προσαρμοστική Συνθετική' (ADASYN) τεχνική επιλύθηκε το πρόβλημα άνισης κατανομής των δεδομένων εκπαίδευσης στις διαφορετικές κλάσεις. Η ονοματολογία και ο

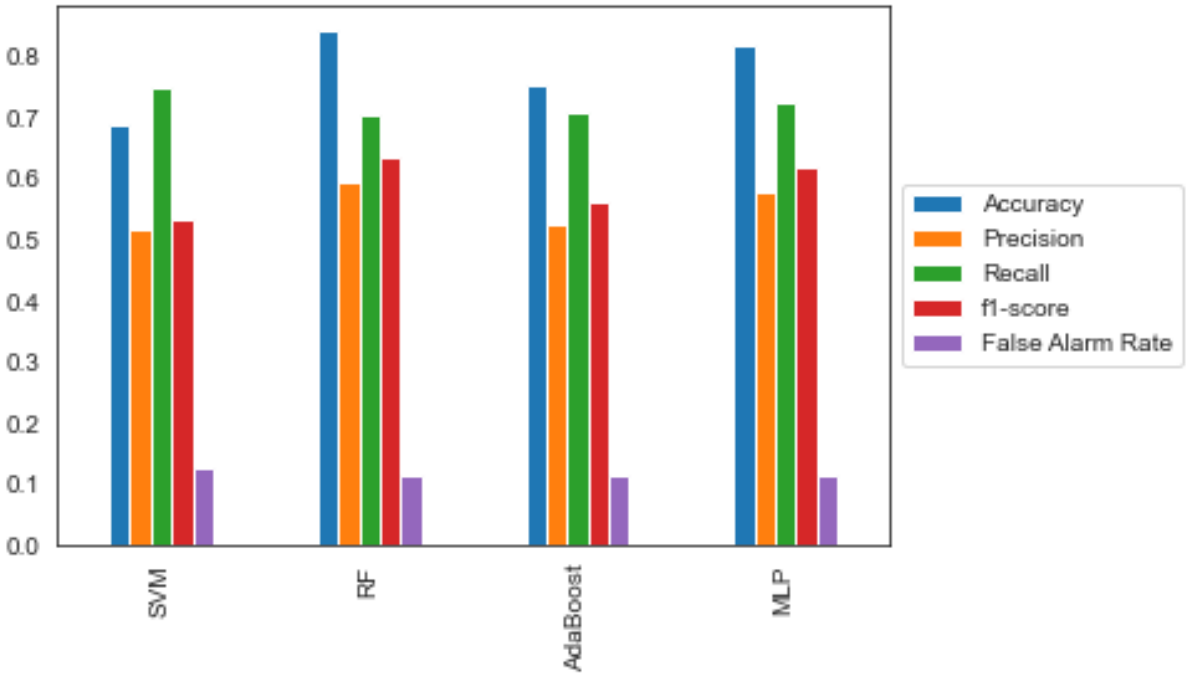
συμβολισμός των τεσσάρων αλγορίθμων παρατίθενται στον πίνακα ενώ οι επιδόσεις τους παρουσιάζονται στον πίνακα και στο γράφημα που ακολουθεί.

Πίνακας 1: Ονοματολογία και συμβολισμός μοντέλων ταξινόμησης

Όνομα μοντέλου (ελληνικά)	**Όνομα μοντέλου (αγγλικά)**	**Συμβολισμός μοντέλου**
Μηχανές Διανυσμάτων Υποστήριξης	Support Vector Machines	SVM
Ταξινομητής Τυχαίων Δασών	Random Forests Classifier	RF
Ταξινομητής AdaBoost	AdaBoost Classifier	AdaBoost
Ταξινομητής Πολυεπίπεδου Perceptron	Multilayer Perceptron Classifier	MLP

Πίνακας 2: Σύγκριση μετρικών αξιολόγησης των μοντέλων ταξινόμησης

	Ορθότητα	**Ακρίβεια**	**Ανάκληση**	**FPR**	**f1-score**
SVM	68,47 %	51,35 %	74,72 %	12,47 %	53,22 %
RF	84,00 %	59,41 %	70,27 %	11,47 %	63,42 %
AdaBoost	75,08 %	52,31 %	70,71 %	11,30 %	55,87 %
MLP	81,28 %	57,51 %	72,04 %	11,37 %	61,79 %



Γράφημα 2: Επίδοση των μοντέλων ταξινόμησης σύμφωνα με ορισμένες μετρικές αξιολόγησης

Στο δεύτερο μέρος των αναλύσεων **εξετάστηκε η διάρκεια που βρίσκεται κάθε οδηγός σε κάθε επίπεδο της 'Ζώνης Ανοχής Ασφαλείας'**. Για τον σκοπό αυτό αναπτύχθηκαν

τρία μοντέλα παλινδρόμησης μηχανικής εκμάθησης. Η επιλογή των ανεξάρτητων μεταβλητών πραγματοποιήθηκε με βάση την επίδοση των μοντέλων σε συνδυασμό με τη στατιστική σημαντικότητα και τη συσχέτιση μεταξύ των ανεξάρτητων μεταβλητών. Οι αλγόριθμοι που εφαρμόστηκαν λαμβάνουν υπόψη και αντιμετωπίζουν την πολύ-συγγραμμικότητα. Επίσης οι αλγόριθμοι πραγματοποιούν επιλογή χαρακτηριστικών μηδενίζοντας ή μειώνοντας τους συντελεστές των ανεξάρτητων μεταβλητών. Η τελική αξιολόγηση της επίδρασης των παραγόντων στην διάρκεια οδήγησης σε κάθε επίπεδο ασφαλείας προκύπτει με βάση τον συντελεστή κάθε ανεξάρτητης μεταβλητής στο μοντέλο της παλινδρόμησης. Στον πίνακα παρουσιάζεται η ονοματολογία και ο συμβολισμός των μοντέλων, ενώ στους πίνακες τα τελικά αποτελέσματα.

Πίνακας 3: Ονοματολογία και συμβολισμός μοντέλων παλινδρόμησης

Όνομα μοντέλου (ελληνικά)	**Όνομα μοντέλου (αγγλικά)**	**Συμβολισμός μοντέλου**
Παλινδρόμηση Κορυφογραμμής	Ridge Regression	RR
Παλινδρόμηση Lasso	Lasso Regression	LR
Παλινδρόμηση Elastic Net	Elastic Net Regression	ENR

Πίνακας 4: Σύνοψη μοντέλου παλινδρόμησης RR

Σύνοψη μοντέλου παλινδρόμησης RR				
	Συντελεστές	Τυπική απόκλιση	t value	p value
Σταθερός όρος	9966,716	472,905	21,076	0,000
Speed_max	-112,009	2,178	-51,441	0,000
Distance travelled_sum	0,001	0,001	8,896	0,000
R^2 = 0,8493		Adjusted R^2 = 0,8458		

Πίνακας 5: Σύνοψη μοντέλου παλινδρόμησης LR

Σύνοψη μοντέλου παλινδρόμησης LR				
	Συντελεστές	Τυπική απόκλιση	t value	p value
Σταθερός όρος	9967,358	472,905	21,077	0,000
Speed_max	-112,017	2,177	-51,445	0,000
Distance travelled_sum	0,001	0,001	8,896	0,000
R^2 = 0,8493		Adjusted R^2 = 0,8458		

Πίνακας 6: Σύνοψη μοντέλου παλινδρόμησης ENR

Σύνοψη μοντέλου παλινδρόμησης ENR				
	Συντελεστές	Τυπική απόκλιση	t value	p value
Σταθερός όρος	9697,044	472,981	20,459	0,000
Speed_max	-108,840	2,182	-49,873	0,000
Distance travelled_sum	0,001	0,001	8,955	0,000
R^2 = 0,8486		Adjusted R^2 = 0,8451		

Βάσει των αποτελεσμάτων που προέκυψαν κατά την εφαρμογή της μεθοδολογίας, προέκυψαν ορισμένα συμπεράσματα άμεσα σχετιζόμενα με τον στόχο της διπλωματικής εργασίας.

- Ο καθορισμός των επιπέδων ασφαλείας της 'Ζώνης Ανοχής Ασφαλείας' με βάση ορίων της μεταβλητής των ελάχιστων χρονο-αποστάσεων παρείχε **αποτελέσματα συναφή με τη διεθνή βιβλιογραφία** όσον αφορά στην κατανομή των δειγμάτων στις κλάσεις, σε σχέση με τις άλλες τεχνικές που εξετάστηκαν.

- Σύμφωνα με τα αποτελέσματα των αναλύσεων, η συνολική διανυθείσα απόσταση είναι η **σημαντικότερη μεταβλητή** για τον εντοπισμό της οδηγικής συμπεριφοράς. Ανάλογα με τη συνολική απόσταση που διανύει ο οδηγός μπορεί να παρατηρηθούν διαφορετικές οδηγικές συμπεριφορές. Για παράδειγμα οι οδηγοί που διανύουν μεγάλες αποστάσεις είναι πιθανό να εμφανίσουν σημάδια κούρασης και μειωμένης προσοχής, τα οποία οδηγούν σε επικίνδυνη οδηγική συμπεριφορά.

- Η ταχύτητα (μέγιστη, ελάχιστη, μέση τιμή, τυπική απόκλιση) είχαν εξίσου **σημαντική επιρροή** στη διαδικασία ταξινόμησης. Η ταχύτητα σχετίζεται άμεσα με την πιθανότητα εμφάνισης ατυχήματος καθώς επίσης και με τη σοβαρότητα αυτού. Όσο ο οδηγός αυξάνει την ταχύτητα οδήγησης, ελαχιστοποιείται ο χρόνος αντίδρασης του οδηγού.

- Τα όρια ταχύτητας τίθενται από τους αρμόδιους προκειμένου η οδήγηση να πραγματοποιείται με ασφάλεια. Η υπέρβαση του ορίου ταχύτητας σχετίζεται με την εμφάνιση ατυχημάτων. Με βάση τα αποτελέσματα, η μεταβλητή του ορίου ταχύτητας είναι **σημαντική για τον εντοπισμό της οδηγικής συμπεριφοράς**.

- Η κατάσταση του οδηγού και η αλληλεπίδραση του με το τιμόνι του οχήματος έχουν **μειωμένη επιρροή** στην αναγνώριση του επιπέδου ασφαλείας που βρίσκεται. Η σημαντικότητα των μεταβλητών Κούρασης και Απόσπασης Προσοχής είναι μικρότερη σε σχέση με τους υπόλοιπους οδηγικούς παράγοντες. Παρόλα αυτά, η κατάσταση του οδηγού και η αλληλεπίδραση του με το τιμόνι σχετίζεται με τους υπόλοιπους οδηγικούς παράγοντες (όπως η ταχύτητα ή η διανυθείσα απόσταση).

- Από τις διαφορετικές τεχνικές αντιμετώπισης του φαινομένου της άνισης κατανομής των δειγμάτων στις διαφορετικές κλάσεις, η 'Προσαρμοστική Συνθετική' (ADASYN) προσέφερε τα **βέλτιστα αποτελέσματα** για το σύνολο των ταξινομητών. Η τεχνική

ADASYN έχει το πλεονέκτημα να αντιμετωπίζει την μεροληψία ως προς την κυρίαρχη τάξη και να ωθεί τα όρια απόφασης της ταξινόμησης στα πιο δύσκολα παραδείγματα.

- Στην παρούσα εργασία αναπτύχθηκαν τέσσερεις αλγόριθμοι ταξινόμησης οι οποίοι σημείωσαν ικανοποιητικές επιδόσεις. Η μέθοδος 'Τυχαίων Δασών' (RF) και η μέθοδος 'Πολυεπίπεδου Perceptron' (MLP) σημείωσαν τις **υψηλότερες επιδόσεις** στην πλειοψηφία των μετρικών αξιολόγησης τους.

- Από το σύνολο των μεταβλητών που εξετάστηκαν, η μέγιστη ταχύτητα και η συνολική διανυθείσα απόσταση προσέφεραν **στατιστικά σημαντικά αποτελέσματα**. Με βάση τους συντελεστές παλινδρόμησης, η μέγιστη ταχύτητα έχει την κύρια, αρνητική επίδραση στην διάρκεια οδήγησης στα διαφορετικά επίπεδα ασφάλειας. Η ελαχιστοποίηση του συντελεστή της διανυθείσας απόστασης πραγματοποιείται στο πλαίσιο αντιμετώπισης της συγγραμμικότητας των μεταβλητών. Επομένως, η μέγιστη ταχύτητα είναι **ιδιαίτερα σημαντική** στην πρόβλεψη της διάρκειας οδήγησης σε κάθε επίπεδο.

- Τα τρία μοντέλα παλινδρόμησης (RR, LR, ENR) στο σύνολο τους έχουν **υψηλή προγνωστική ικανότητα**.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

1. ΕΙΣΑΓΩΓΗ

1.1 Γενική ανασκόπηση

Στη σημερινή πραγματικότητα οι οδικές μεταφορές έχουν σημαντική συνεισφορά στις καθημερινές και παραγωγικές δραστηριότητες των πολιτών και γενικότερα της κοινωνίας.

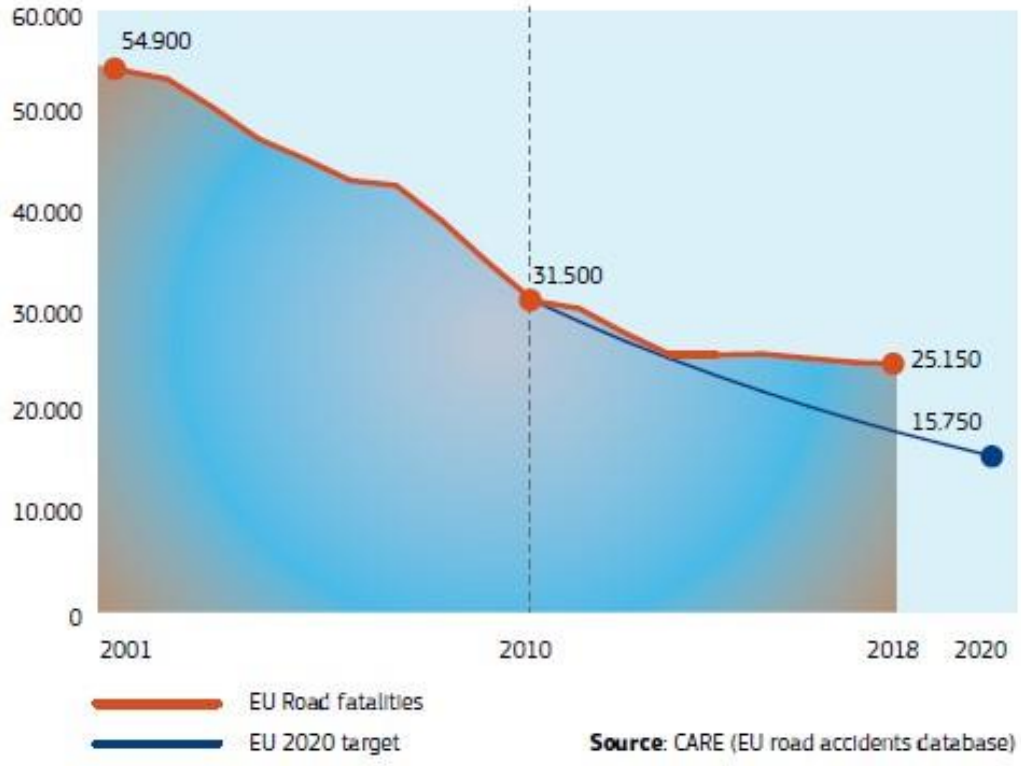
Ωστόσο, πέραν του σημαντικού ρόλου τους στην κοινωνία, οι οδικές μεταφορές αποτελούν σημαντική αιτία ατυχημάτων και απώλειας ανθρώπινων ζωών παγκοσμίως. Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας τα οδικά ατυχήματα αποτελούν την 8η αιτία θανάτου για τους ανθρώπους όλων των ηλικιών και την 1η αιτία θανάτου για νέους σε ηλικία 5 έως 29 ετών, καταγράφοντας περίπου 1.3 εκατ. απώλειες ζωών κάθε έτος (World Health Organization, 2018).

Τα τελευταία χρόνια η Ευρώπη έχει καταβάλει σημαντικές προσπάθειες μειώνοντας τους θανάτους από οδικά ατυχήματα κατά 43% μεταξύ 2001-2010 και 21% μεταξύ 2010 και 2018. Ειδικότερα η Ελλάδα την δεκαετία 2010-2020 κατάφερε να μειώσει κατά 51% τα οδικά ατυχήματα σημειώνοντας την υψηλότερη μείωση μεταξύ των κρατών μελών της ΕΕ (γράφημα 1.1).

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2010-2020
AT	66	62	63	54	51	56	50	47	46	47	*38*	-42.4%
BE	77	80	69	65	65	65	56	54	53	56	*44*	-42.9%
BG	105	89	82	83	91	98	99	96	87	90	*67*	-36.2%
CY	73	85	59	51	52	67	54	62	57	59	*54*	-26.0%
CZ	77	67	71	62	65	70	58	55	62	58	*48*	-37.7%
DE	45	50	45	41	42	43	39	39	40	37	*33*	-26.7%
DK	48	40	30	34	32	31	37	30	30	34	*27*	-43.8%
EE	59	76	66	61	59	51	54	36	51	39	*45*	-23.7%
EL	113	103	89	80	73	73	76	69	65	64	*54*	-52.3%
ES	52	44	41	36	36	36	39	39	39	37	*29*	-44.2%
FI	50	54	47	48	42	49	47	42	43	38	*40*	-20.0%
FR	62	61	56	50	51	52	54	51	49	50	*39*	-37.1%
HR	99	97	92	86	73	82	73	80	77	73	*58*	-41.4%
HU	74	64	61	60	63	65	62	64	65	62	*46*	-37.8%
IE	46	41	35	41	42	36	39	33	29	29	*30*	-34.8%
IT	66	64	63	57	56	56	54	56	55	53	*40*	-39.4%
LT	95	97	101	86	91	83	66	67	62	67	*63*	-33.7%
LU	64	64	65	84	64	64	56	42	60	36	*42*	-34.4%
LV	103	86	88	99	106	95	80	70	77	69	*74*	-28.2%
MT	36	51	22	40	24	26	51	41	38	32	*21*	-41.7%
NL	39	40	34	28	28	31	31	31	35	34	*31*	-20.5%
PL	103	110	94	88	84	77	80	75	75	77	*65*	-36.9%
PT	80	74	68	61	61	57	54	58	68	63	*52*	-35.0%
RO	117	100	102	93	91	95	97	99	96	96	*85*	-27.4%
SE	28	34	30	27	28	27	27	25	32	22	*18*	-35.7%
SI	67	69	63	61	52	58	63	50	44	49	*38*	-43.3%
SK	69	61	65	46	54	51	51	51	48	50	*45*	-34.8%
EU	67	65	60	55	54	55	53	53	52	51	*42*	-37.5%

Γράφημα 1.1: Αριθμός νεκρών ανά εκατομμύριο πληθυσμού στην ΕΕ
Πηγή: NTUA Road Safety Observatory (2022)

Στο γράφημα 1.2 που ακολουθεί διακρίνεται η σημαντική μείωση των θανατηφόρων οδικών ατυχημάτων στην ΕΕ καθώς και η απόκλιση από τον στόχο του έτους 2020.



Γράφημα 1.2: Εξέλιξη των θανατηφόρων ατυχημάτων στην ΕΕ και ο στόχος για το 2001-2020
Πηγή: CARE (EU road accidents database)

Παρόλα αυτά οι θάνατοι και σοβαροί τραυματισμοί παραμένουν σε υψηλά επίπεδα με σημαντικές επιπτώσεις.

Η Ευρωπαϊκή Ένωση έχει θέσει ως στόχο την μείωση των θανατηφόρων οδικών ατυχημάτων κατά 50% (**EU Road Safety Policy Framework 2021-2030**) (European Commission and Directorate-General for Mobility and Transport, 2020)όπως και ο Παγκόσμιος Οργανισμός Υγείας σε συνδυασμό με τα Ηνωμένα Έθνη (**Global Plan Decade of Action for Road Safety 2021-2030**) (World Health Organization, 2021). Για την επίτευξη του παραπάνω στόχου δίνεται ιδιαίτερη σημασία στην συνεισφορά των νέων τεχνολογιών του τομέα της αυτοκινητοβιομηχανίας και της αυτοματοποίησης στις μεταφορές, με στόχο την βελτίωση της οδικής ασφάλειας (Fagnant and Kockelman, 2015).

Τα οδικά ατυχήματα προκαλούνται από πολλούς διαφορετικούς παράγοντες όπως η κατάσταση του οδηγού, οι περιβαλλοντικές συνθήκες και οι κυκλοφοριακές συνθήκες (Aljanahi et al., 1999). Ωστόσο, ο ανθρώπινος παράγοντας αποτελεί κύρια αιτία οδικών ατυχημάτων. Η διαρκής ανάπτυξη στον τομέα των αυτόματων οχημάτων έχει ως στόχο τη βελτίωση της οδικής ασφάλειας, αφαιρώντας την πιθανότητα ανθρώπινου σφάλματος από την διαδικασία της οδήγησης (Katrakazas, 2017).

Η Ευρωπαϊκή Επιτροπή μέσω του προγράμματος-πλαισίου έρευνας στις μεταφορές Horizons 2020 χρηματοδοτεί το ερευνητικό έργο i-DREAMS (2022) (https://idreamsproject.eu/). Σκοπός του συγκεκριμένου έργου είναι η ανάπτυξη και αξιολόγηση μίας ‘Ζώνης Ανοχής Ασφαλείας’ (‘Safety Tolerance Zone’), που θα περιλαμβάνει διαφορετικά επίπεδα ασφάλειας. Με την αξιοποίηση ενός έξυπνου συστήματος παρακολούθησης των οδηγικών και περιβαλλοντικών χαρακτηριστικών θα

είναι δυνατή η αναγνώριση του επιπέδου που βρίσκεται κάθε οδηγός και η ανάπτυξη παρεμβάσεων προκειμένου αυτός να μην παρεκκλίνει από την ασφαλή οδήγηση. Οι παρεμβάσεις θα πραγματοποιούνται σε δύο φάσεις. Η πρώτη φάση σε πραγματικό χρόνο δηλαδή κατά την διαδικασία της οδήγησης και με στόχο ο οδηγός να προβεί άμεσα στις απαιτούμενες ενέργειες και η δεύτερη σε μετέπειτα χρόνο αποσκοπώντας στην βελτίωση της γνώσης και κατ' επέκταση της συμπεριφοράς του οδηγού.

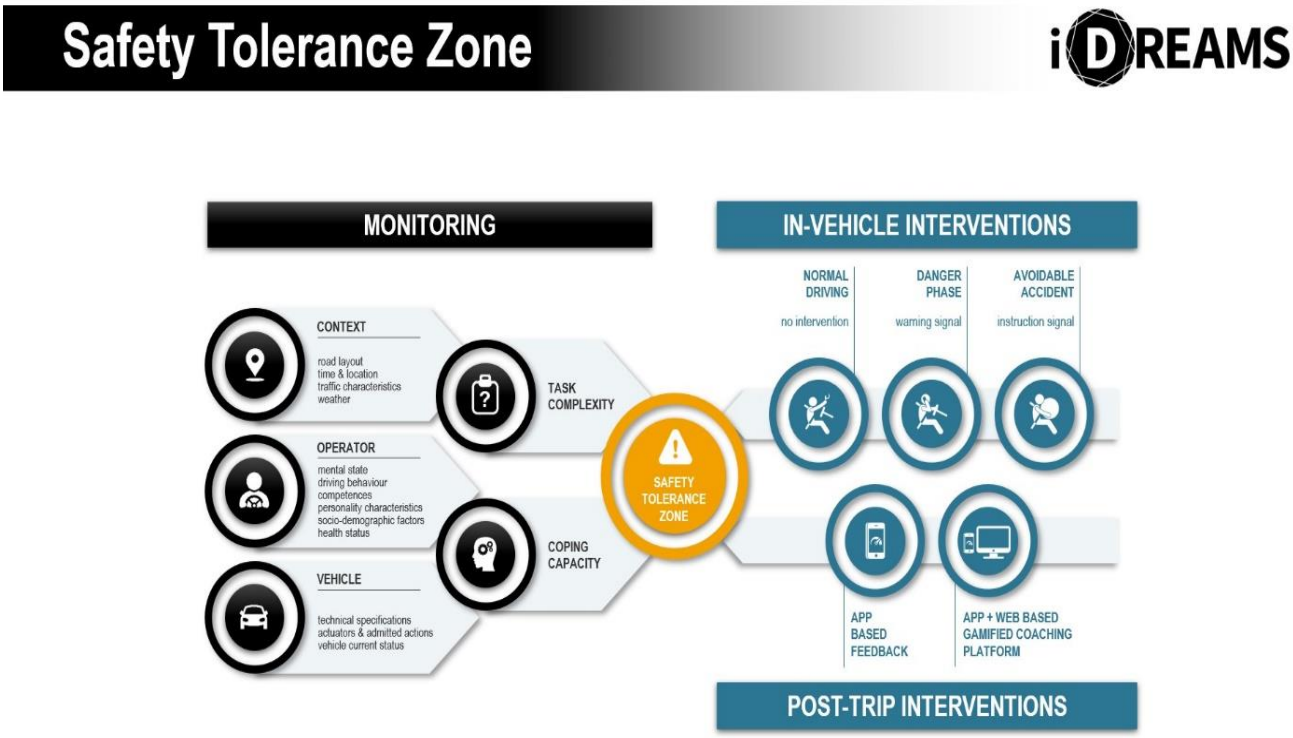
Η «Ζώνη Ανοχής Ασφάλειας» περιλαμβάνει τρία επίπεδα:

1) Κανονικό – Ασφαλές (Normal)

2) Επικίνδυνο (Dangerous)

3) Αποφυγής Ατυχήματος (Avoidable Accident)

Οι δοκιμές για την συλλογή σημαντικών δεδομένων πραγματοποιήθηκαν σε ένα περιβάλλον προσομοιωτή οδήγησης με την συμμετοχή 600 οδηγών σε 5 χώρες της ΕΕ.



Γράφημα 1.3: Μεθοδολογία ερευνητικού έργου i-DREAMS
Πηγή: i-DREAMS (2022)

Η ανάλυση της συμπεριφοράς του οδηγού με την αξιοποίηση αλγόριθμων μηχανικής εκμάθησης αποτελεί αντικείμενο έρευνας υψηλού ενδιαφέροντος τα τελευταία χρόνια (Peppes et al., 2021). Επιπρόσθετα, η χρήση έξυπνων συστημάτων παρακολούθησης της συμπεριφοράς του οδηγού με σκοπό τις παρεμβάσεις σε πραγματικό χρόνο, έχει αποδειχθεί ότι είναι ιδιαίτερα αποτελεσματική στην μείωση των ατυχημάτων (Michelaraki et al., 2021b). Η αναγκαιότητα ανάπτυξης ανάλογων συστημάτων με γνώμονα την βελτίωση της οδικής ασφάλειας καθιστά απαραίτητο τον εντοπισμό της επιρροής των διαφορετικών παραγόντων κίνδυνου κατά την οδήγηση.

Συνεπώς, ο εντοπισμός της επικίνδυνης συμπεριφοράς των οδηγών και των παραγόντων που επιδρούν σε αυτήν θα αποτελέσει κύριο αντικείμενο έρευνας στην παρούσα μελέτη.

1.2 Στόχος

Σύμφωνα με όσα προαναφέρθηκαν η παρούσα διπλωματική εργασία στοχεύει στον εντοπισμό του επιπέδου και της διάρκειας επικίνδυνης συμπεριφοράς του οδηγού (Ζώνης Ανοχής Ασφαλείας) μέσω:

1) Της ανάπτυξης μοντέλων ταξινόμησης, με σκοπό τον προσδιορισμό του επιπέδου της 'Ζώνης Ανοχής Ασφαλείας' που βρίσκεται κάθε οδηγός. Συγκεκριμένα θα αναπτυχθούν, εκπαιδευτούν και αξιολογηθούν αλγόριθμοι μηχανικής εκμάθησης οι οποίοι θα είναι σε θέση να ταξινομήσουν κάθε οδηγό σε ένα από τα τρία επίπεδα της 'Ζώνης Ανοχής Ασφαλείας'. Αυτό επιτυγχάνεται λαμβάνοντας ως δεδομένα εισόδου τα χαρακτηριστικά οδήγησης του κάθε οδηγού καθώς και του αντίστοιχου περιβάλλοντος οδήγησης. Η ταξινόμηση μέσω της μηχανικής μάθησης αποτελεί σημαντικό εργαλείο για την αναγνώριση της οδηγικής συμπεριφοράς και κατ' επέκταση τη βελτίωση της οδικής ασφάλειας (Meiring and Myburgh, 2015; Wu et al., 2016)

2) Της ανάπτυξης κατάλληλων μοντέλων παλινδρόμησης για την πρόβλεψη της διάρκειας που κάθε οδηγός βρίσκεται σε κάθε ένα από τα 3 επίπεδα της 'Ζώνης Ανοχής Ασφαλείας', προκειμένου να εξεταστούν και να αξιολογηθούν τα χαρακτηριστικά που επιδρούν στην διάρκεια οδήγησης κάτω από επικίνδυνες συνθήκες.

Ο εντοπισμός του επιπέδου της 'Ζώνης Ανοχής Ασφαλείας' θα προσφέρει σημαντικά συμπεράσματα για τους παράγοντες που επιδρούν στην αναγνώριση της επικίνδυνης οδήγησης. Επίσης, η ανάλυση της διάρκειας οδήγησης σε επικίνδυνες συνθήκες θα αποτελέσει εναλλακτικό δείκτη κινδύνου και η ανάλυση του θα ενισχύσει τα συμπεράσματα σχετικά με την επιρροή των παραγόντων οδήγησης.

Η συνεισφορά της παρούσας μελέτης είναι διττή αφού θα επιχειρήσει να προσφέρει επιπλέον γνώση και διευρύνει την υπάρχουσα στο τομέα ανάλυσης οδηγικής συμπεριφοράς και ανάπτυξης αυτόματων συστημάτων οδήγησης.

1.3 Μεθοδολογία

Στο παρόν υποκεφάλαιο περιγράφεται συνοπτικά η **μεθοδολογία** που ακολουθήθηκε για την επίτευξη του στόχου της διπλωματικής εργασίας.

Αρχικά, οριστικοποιήθηκε το θέμα της εργασίας και καθορίστηκε ο **στόχος** της μελέτης. Ακολούθως, πραγματοποιήθηκε η **βιβλιογραφική ανασκόπηση** κατά την οποία αναζητήθηκαν δημοσιεύσεις και παλαιότερες έρευνες, άμεσα συναφείς με το θέμα της διπλωματικής εργασίας, καθώς και με τις μεθόδους ανάλυσης που αξιοποιήθηκαν.

Στην συνέχεια πραγματοποιήθηκε η **συλλογή** και **επεξεργασία** των στοιχείων. Τα στοιχεία που συλλέχθηκαν παράχθηκαν από πείραμα σε προσομοιωτή οδήγησης στο

πλαίσιο του ερευνητικού έργου i-DREAMS και αφορούσαν στα χαρακτηριστικά οδήγησης 48 οδηγών καθώς και του αντίστοιχου περιβάλλοντος οδήγησης . Με την κατάλληλη επεξεργασία τα δεδομένα προετοιμάστηκαν για την ανάλυση τους.

Μετά την συλλογή και την επεξεργασία, ακολούθησε η **ανάπτυξη των κατάλληλων μοντέλων μηχανικής μάθησης,** ταξινόμησης και παλινδρόμησης. Η επεξεργασία, η ανάπτυξη των μοντέλων και οι αναλύσεις έγιναν με χρήση της γλώσσας προγραμματισμού Python αξιοποιώντας τη βιβλιοθήκη μηχανικής μάθησης scikit-learn και τη βιβλιοθήκη ανάλυσης δεδομένων pandas.

Τέλος, αξιολογήθηκαν τα αποτελέσματα με την εξαγωγή χρήσιμων **συμπερασμάτων και προτάσεων για περαιτέρω έρευνα**.

Παρακάτω παρουσιάζονται υπό την μορφή διαγράμματος ροής (γράφημα 1.4), τα διαδοχικά στάδια που ακολουθήθηκαν για την εκπόνηση της παρούσας διπλωματικής εργασίας.

Γράφημα 1.4: Διάγραμμα Ροής - Μεθοδολογία διπλωματικής εργασίας

1.4 Δομή διπλωματικής εργασίας

Στην παρούσα ενότητα παρουσιάζεται η **δομή της διπλωματικής εργασίας** μέσω της συνοπτικής περιγραφής του περιεχομένου κάθε κεφαλαίου.

Το **Κεφάλαιο 1** αποτελεί την **εισαγωγή** και την ανάδειξη του στόχου της διπλωματικής εργασίας. Αρχικά με την γενική ανασκόπηση παρουσιάζεται το πλαίσιο της διπλωματικής εργασίας που αφορά στην σοβαρή επιρροή των οδικών ατυχημάτων στην σύγχρονη κοινωνία. Παρατίθενται στατιστικά στοιχεία για την οδική ασφάλεια στην Ευρώπη και την Ελλάδα και γίνεται αναφορά στην συνεισφορά των σύγχρονων τεχνολογιών στην μείωση των θανατηφόρων οδικών ατυχημάτων με έμφαση στο ερευνητικό έργο i-DREAMS. Τέλος, περιγράφεται ο στόχος, η μεθοδολογία που ακολουθήθηκε για την επίτευξη του και η δομή της διπλωματικής εργασίας.

Το **Κεφάλαιο 2**, περιλαμβάνει την **βιβλιογραφική ανασκόπηση** στην οποία παρουσιάζονται συναφείς έρευνες τόσο με το αντικείμενο της διπλωματικής εργασίας όσο και με τις μεθοδολογίες που αξιοποιήθηκαν. Οι έρευνες προέρχονται από την Ελληνική και την Διεθνή Επιστημονική κοινότητα.

Στο **Κεφάλαιο 3**, γίνεται αναφορά στο **θεωρητικό υπόβαθρο** της έρευνας. Αρχικά αναλύονται οι τεχνικές επεξεργασίας των δεδομένων και δίνεται ιδιαίτερη έμφαση στην αναγκαιότητα αυτού του βήματος για την ανάπτυξη των μοντέλων. Στην συνέχεια παρουσιάζονται, οι διαφορετικοί αλγόριθμοι μηχανικής μάθησης που αναπτύχθηκαν για την ταξινόμηση και την παλινδρόμηση και περιγράφονται οι μετρικές αξιολόγησης των μοντέλων.

Στο **Κεφάλαιο 4**, περιγράφονται τα δεδομένα και η διαδικασία **συλλογής** τους από τον προσομοιωτή οδήγησης (i-DREAMS). Στη συνέχεια αναλύεται η διαδικασία και τα βήματα της **επεξεργασίας** των οδηγικών και περιβαλλοντικών χαρακτηριστικών προκειμένου να προετοιμαστούν για την περαιτέρω ανάλυση.

Το **Κεφάλαιο 5**, αποτελεί την κύρια ενότητα της διπλωματικής εργασίας καθώς περιλαμβάνει την αναλυτική παρουσίαση της **μεθοδολογίας ανάπτυξης των μοντέλων**. Η συγκεκριμένη υποενότητα χωρίζεται σε δύο τομείς, την ταξινόμηση και την παλινδρόμηση. Αρχικά επεξηγούνται τα βήματα που ακολουθήθηκαν για την εφαρμογή της μεθοδολογίας, αναλύεται η διαδικασία ανάπτυξης των μοντέλων μηχανικής μάθησης και περιγράφονται τα δεδομένα εισόδου και εξόδου. Τέλος παρουσιάζονται τα συνολικά αποτελέσματα της ανάλυσης συγκρίνοντας και περιγράφοντας τα διαφορετικά μοντέλα συνοδευόμενα από τις πολλαπλές μετρικές αξιολόγησης.

Το **Κεφάλαιο 6** περιλαμβάνει τα **συμπεράσματα** που προέκυψαν από τα τελικά αποτελέσματα του προηγούμενου κεφαλαίου. Στο τέλος παρουσιάζονται οι προτάσεις σχετικά για να συνδράμουν στην περαιτέρω έρευνα η οποία αφορά στην αξιοποίηση είτε διαφορετικών μεθόδων, είτε διαφορετικών δεδομένων.

Στο **Κεφάλαιο 7** παρατίθενται οι βιβλιογραφικές αναφορές, οι οποίες αξιοποιήθηκαν για την εκπόνηση της διπλωματικής εργασίας.

2. ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΑΝΑΣΚΟΠΗΣΗ

2.1 Εισαγωγή

Στο παρόν κεφάλαιο παρουσιάζονται **συναφείς έρευνες και μεθοδολογίες** σχετικά με το αντικείμενο της διπλωματικής εργασίας. Συγκεκριμένα αναζητήθηκαν στην διεθνή βιβλιογραφία δημοσιευμένες έρευνες που επικεντρώνονται στην ανάλυση, αναγνώριση της συμπεριφοράς του οδηγού καθώς και την πρόβλεψη συγκρούσεων σε πραγματικό χρόνο, αξιοποιώντας διάφορες τεχνικές μηχανικής εκμάθησης.

Μέσω της παρουσίασης των ερευνών θα προκύψει ο **στόχος** της παρούσας μελέτης καθώς και η κατάλληλη **μεθοδολογία** για την επίτευξη του. Ιδιαίτερη έμφαση θα δοθεί στα διαφορετικά μοντέλα μηχανικής εκμάθησης. Επισημαίνουμε ότι στη πλειονότητα των ερευνών που αναζητήθηκαν παρατηρείται το πρόβλημα της άνισης κατανομής των δεδομένων στις διαφορετικές τάξεις. Για τον λόγο αυτό θα παρουσιαστούν οι διαφορετικές τεχνικές επαναδειγματοληψίας δεδομένων που έχουν εφαρμοστεί σε προγενέστερες έρευνες.

2.2 Συναφείς έρευνες και μεθοδολογίες

Η επικίνδυνη συμπεριφορά κατά την οδήγηση επηρεάζεται από πολλούς και διαφορετικούς παράγοντες. Σύμφωνα με τους Wang et al. (2020) η σοβαρότητα της επικίνδυνης οδήγησης σχετίζεται με διάφορους κυκλοφοριακούς παράγοντες συμπεριλαμβανομένων των χαρακτηριστικών συμπεριφοράς του οδηγού, των χαρακτηριστικών του οχήματος και του περιβάλλοντος.

Τα τελευταία χρόνια **η ανάλυση της οδηγικής συμπεριφοράς** με την αξιοποίηση αλγορίθμων μηχανικής εκμάθησης βρίσκεται στο επίκεντρο της επιστημονικής κοινότητας ενώ παράλληλα η εξέλιξη των Ευφυών Μεταφορικών Συστημάτων (Intelligent Transportation Systems - ITS) έχει δημιουργήσει πρόσφορο έδαφος στον τομέα των αυτόνομων οχημάτων (Peppes et al., 2021).

Επίσης η αναγνώριση του επιπέδου ασφαλείας της οδηγικής συμπεριφοράς και ικανότητα πρόβλεψης ατυχήματος σε πραγματικό χρόνο μπορούν να συνεισφέρουν σημαντικά στην εξέλιξη των Αναπτυσσόμενων Συστημάτων Υποβοήθησης Οδηγού (Advanced Driver Assistance Systems – ADAS) (Yang et al., 2021).

Για την επίτευξη όλων των παραπάνω στόχων και την περαιτέρω εξέλιξη στον τομέα της αυτόματης οδήγησης αρκετές δημοσιοποιημένες έρευνες επικεντρώνονται στην κατανόηση της επίδρασης των διαφορετικών χαρακτηριστικών στην επικίνδυνη οδηγική συμπεριφορά. Η διερεύνηση της επιρροής δύναται να εξελίξει τα κατάλληλα μοντέλα αναγνώρισης επικίνδυνων συμπεριφορών κατά την οδήγηση και κατ’ επέκταση να βελτιώσει την αποδοτικότητα των συστημάτων υποβοήθησης του οδηγού.

2.2.1 Ανάλυση οδηγικής συμπεριφοράς

Οι **βασικές προσεγγίσεις** για την ανάλυση της επικίνδυνης οδηγικής συμπεριφοράς αποτελούνται από (α) μελέτες βασισμένες σε δεδομένα από έρευνες; (β) μακροσκοπική ανάλυση δεδομένων ατυχήματος; (γ) μελέτες βασισμένες σε πείραμα δεδομένα προσομοιωτή οδήγησης και (δ) νατουραλιστικές μελέτες οδήγησης (NDS). Σε προηγούμενες έρευνες έχουν διερευνηθεί ορισμένοι δείκτες ασφαλείας για την αξιολόγηση της πιθανότητας κινδύνου όπως είναι ο χρόνος πρόσκρουσης (TTC) (Shi et al., 2018).

Ωστόσο, είναι δύσκολος ο καθορισμός των ορίων για τους διαφορετικούς δείκτες καθιστώντας την διαδικασία περιγραφής των διαφορετικών επιπέδων κινδύνου ιδιαίτερα προβληματική (Shi et al., 2019). Για τον λόγο αυτό ορισμένες έρευνες καθορίζουν τα διαφορετικά επίπεδα κινδύνου αξιοποιώντας τεχνικές ομαδοποίησης.

Η απόσπαση της προσοχής του οδηγού αποτελεί βασικό παράγοντα επικίνδυνης οδηγικής συμπεριφοράς. Η μελέτη των Osman et al. (2019) επιχειρεί να αναπτύξει κατάλληλα μοντέλα αναγνώρισης του τύπου των **δευτερευόντων ενεργειών** στις οποίες εμπλέκεται ο οδηγός. Αξιοποιώντας ορισμένες παραμέτρους οδηγικής συμπεριφοράς όπως η ταχύτητα, η διαμήκης επιτάχυνση, η πλευρική επιτάχυνση, η θέση πεντάλ επιτάχυνσης και ο ρυθμός εκτροπής, οι δευτερεύουσες ενέργειες κατηγοριοποιούνται σε 4 ξεχωριστές κλάσεις. Στην συνέχεια αναπτύσσονται αλγόριθμοι μηχανικής εκμάθησης για την αναγνώριση της εμπλοκής του οδηγού σε δευτερεύουσα ενέργεια αλλά και την ταξινόμηση της δευτερεύουσας ενέργειας σε μία από τις 4 τάξεις. Σύμφωνα με τους ερευνητές, τα μοντέλα αναγνώρισης μπορούν να ενσωματωθούν σε συστήματα εντός του οχήματος για την αναγνώριση της επικίνδυνης οδηγικής συμπεριφοράς προκειμένου να ειδοποιείται ο οδηγός για πιθανή απόσπαση της προσοχής του.

Η μελέτη των Yi et al. (2019) επιδιώκει την ανάπτυξη και την επικύρωση **ενός συστήματος αναγνώρισης της κατάστασης του οδηγού**. Στόχος των ερευνητών είναι η δημιουργία ενός συστήματος αναγνώρισης το οποίο θα βασίζεται σε εύκολα προσβάσιμα κινηματικά δεδομένα μέσω κινητού τηλεφώνου. Προκειμένου τα συστήματα υποβοήθησης να είναι περισσότερο αποδεκτά, πιο φιλικά προς τον χρήστη και να έχουν υψηλότερη επίδοση, προτείνονται ορισμένοι αλγόριθμοι μηχανικής εκμάθησης εξατομικευμένοι για κάθε οδηγό και τα οδικά χαρακτηριστικά του. Ουσιαστικά προτείνεται ένα σύστημα αναγνώρισης βασιζόμενο στα ξεχωριστά οδηγικά χαρακτηριστικά του κάθε οδηγού. Σύμφωνα με τους μελετητές το πλαίσιο εργασίας που προτείνεται μπορεί να εφαρμοστεί σε μεγάλο πλήθος εφαρμογών που οι διαφορετικές συμπεριφορές ποικίλουν.

Η μέθοδος της επιλογής χαρακτηριστικών συναντάται συχνά στην ανάπτυξη μοντέλων πρόβλεψης της επικίνδυνης οδηγικής συμπεριφοράς. Η έρευνα των Shi et al. (2019) προτείνει ένα πλαίσιο εργασίας εξαγωγής και **επιλογής χαρακτηριστικών** για την αξιολόγηση της οδήγησης και την πρόβλεψη του επιπέδου επικινδυνότητας. Η ανάλυση βασίζεται σε δεδομένα οδηγικής συμπεριφοράς. Παράλληλα με την αξιοποίηση τεχνικών μη επιβλεπόμενης μάθησης, τα δεδομένα αυτά κατηγοριοποιούνται σε διαφορετικά επίπεδα επικινδυνότητας. Στη συνέχεια χρησιμοποιώντας κατάλληλο μοντέλο προσδιορίζονται τα βασικά χαρακτηριστικά σύμφωνα με την κατάταξη σημαντικότητας τους. Η πρόβλεψη του επιπέδου κινδύνου βασίζεται στα σημαντικότερα χαρακτηριστικά που επιλέχθηκαν.

Η έρευνα των Song et al. (2021) εστιάζει **στην σχέση των χαρακτηριστικών του οδηγού με τις επικίνδυνες ενέργειας κατά την οδήγηση**. Συγκεκριμένα αναλύεται η σχέση μεταξύ των δημογραφικών χαρακτηριστικών, της αναζήτησης περιπέτειας, της αντίληψης του κινδύνου και της επικίνδυνης οδηγικής συμπεριφοράς του οδηγού. Στο πλαίσιο της έρευνας με αξιοποίηση τεχνικών ομαδοποίησης ανάλογα το φύλλο του οδηγού καθορίζονται 3 ομάδες επικινδυνότητας, η ‘χαμηλή’, η ‘μεσαία΄ και η υψηλή’. Κατόπιν αναπτύσσεται και αξιολογείται κατάλληλος αλγόριθμος μηχανικής εκμάθησης για την ταξινόμηση της συμπεριφοράς του οδηγού σε ένα από τα 3 επίπεδα επικινδυνότητας. Από την αξιολόγηση των διαφορετικών χαρακτηριστικών του οδηγού προκύπτουν

σημαντικά συμπεράσματα για την μεταξύ τους σχέση και την σχέση τους με την επικίνδυνη οδήγηση.

Οι Ghandour et al. (2021) **μελετούν την οδηγική συμπεριφορά και τις διαφορετικές ψυχολογικές συνθήκες του οδηγού**. Όπως αναφέρουν έχουν προταθεί σε παλαιότερες έρευνες διάφορες μέθοδοι, οι οποίες ωστόσο πάσχουν από έλλειψη αποτελεσματικότητας σε πραγματικές συνθήκες. Στην έρευνα αναπτύσσονται και συγκρίνονται ορισμένα μοντέλα ταξινόμησης για την αναγνώριση της οδηγικής συμπεριφοράς και των συνθηκών απόσπασης προσοχής, τα οποία βασίζονται σε δεδομένα σχετιζόμενα με τις διαφορετικές συμπεριφορές (όπως είναι η επιθετική, η κουρασμένη και η φυσιολογική). Τα δεδομένα διαχωρίζονται σε δεδομένα 'ανίχνευσης λωρίδας' και σε δεδομένα 'κυκλοφοριακής κατάστασης'. Σύμφωνα με τα αποτελέσματα το δεύτερο σύνολο δεδομένων παρέχει σημαντικά και ποικίλα χαρακτηριστικά για τον προσδιορισμό της ψυχικής κατάστασης του οδηγού.

Στην έρευνα των Shangguan et al. (2021) προτείνεται μία μεθοδολογία **για την αξιολόγηση και την πρόβλεψη της κατάστασης κινδύνου** που βρίσκεται ο οδηγός σε πραγματικό χρόνο. Μέσω της ανάπτυξης αλγορίθμων ομαδοποίησης καθορίζονται 4 στάδια επικινδυνότητας. Επιπλέον για την πρόβλεψη της κατάστασης κινδύνου αναπτύσσονται ορισμένοι αλγόριθμοι ταξινόμησης μηχανικής εκμάθησης. Αναλύοντας την επιρροή των μεταβλητών προκύπτει ότι η διαφορά ταχύτητας, η απόσταση από το προπορευόμενο όχημα, η ταχύτητα και η επιτάχυνση είναι ιδιαίτερα σημαντικές για την πρόβλεψη της κατάστασης επικινδυνότητας του οδηγού.

Η έρευνα των Yang et al. (2021) που πραγματοποιήθηκε στο πλαίσιο του ερευνητικού έργου i-DREAMS, προτείνει ένα πλαίσιο εργασίας για **την ταξινόμηση και την αξιολόγηση διαφορετικών επιπέδων ασφαλείας της οδηγικής συμπεριφοράς**. Τα δεδομένα που αναλύονται έχουν συλλεχθεί από πείραμα προσομοιωτή οδήγησης και αφορούν διάφορα χαρακτηριστικά οδήγησης. Με την ανάπτυξη διαφορετικών τεχνικών ομαδοποίησης (clustering) προκύπτουν 4 επίπεδα ασφαλείας. Στην συνέχεια αναπτύσσονται και αξιολογούνται διαφορετικοί αλγόριθμοι ταξινόμησης προκειμένου να μπορεί να αναγνωριστεί το επίπεδο ασφαλείας που βρίσκεται κάθε οδηγός με βάση τα οδηγικά χαρακτηριστικά.

Στον πίνακα 2.1 παρουσιάζονται οι περιορισμοί, οι ελλείψεις και οι προτάσεις για περαιτέρω των παραπάνω μελετών.

Πίνακας 2.1: Ελλείψεις/Προτάσεις για μελλοντική διερεύνηση των ερευνών που παρουσιάστηκαν

Έρευνα	**Ελλείψεις**	**Προτάσεις για μελλοντική διερεύνηση**
Osman et al., 2019	Οι μελετητές δεν λαμβάνουν υπόψη την επίδραση του τύπου της οδού, των γεωμετρικών χαρακτηριστικών και των χαρακτηριστικών του οχήματος στην οδηγική συμπεριφορά	Προτείνεται η διερεύνηση της επίδρασης της κατηγορίας της οδού, των γεωμετρικών χαρακτηριστικών της καθώς και των χαρακτηριστικών του οχήματος στις μεταβλητές της οδηγικής συμπεριφοράς.
Yi et al., 2019	Βασική έλλειψη της μελέτης αφορά η απουσία σημαντικών πληροφοριών όπως η κυκλοφοριακή κατάσταση, οι καιρικές συνθήκες και ο συνεχής χρόνος οδήγησης.	Μελλοντικά θα μπορούσαν μέσω μεγαλύτερου αριθμού οδηγικών δεδομένων να ληφθούν οι μεταβλητές της κυκλοφοριακής κατάστασης, των καιρικών συνθηκών και του συνεχούς χρόνου οδήγησης.
Shi et al., 2019	Η έλλειψη περιπτώσεων συγκρούσεων στην διαβάθμιση της επικινδυνότητας καθιστά δύσκολη την διασύνδεση του υψηλού επιπέδου κινδύνου με την πραγματική εμφάνιση ατυχήματος.	Προτείνεται η σε βάθος εξαγωγή χαρακτηριστικών η οποία θα καλύπτει ένα μεγάλο εύρος οδηγικής συμπεριφοράς και επικίνδυνων συνθηκών, όπως η αλλαγή λωρίδας και η σύγκρουση μεταξύ μοτοσυκλέτας και οχήματος.
Song et al., 2021	Τα δεδομένα της έρευνας ενδέχεται να αντιμετωπίζουν το πρόβλημα της μεροληψίας καθώς προέρχονται από τα αποτελέσματα υποκειμενικών ερωτηματολογίων	Σε μελλοντική έρευνα χαρακτηριστικά της προσωπικότητας του οδηγού, όπως η επιθετική συμπεριφορά, θα μπορούσαν να ληφθούν υπόψη. Επιπρόσθετα Προτείνεται σε μελλοντική έρευνα ο συνδυασμός των αποτελεσμάτων των ερωτηματολογίων και αντικειμενικών δεδομένων οδήγησης προκειμένου τα αποτελέσματα να έχουν μεγαλύτερη ακρίβεια.

Ghandour et al., 2021	Η μελετητές δεν λαμβάνουν υπόψη ορισμένα πρόσθετα χαρακτηριστικά της οδού και της ψυχικής κατάστασης του οδηγού. Επίσης, η έρευνα περιορίζεται στην ανάλυση ενός τύπου οδού.	Με σκοπό την βελτίωση των αποτελεσμάτων της ταξινόμησης προτείνεται σε μελλοντική έρευνα να ληφθούν υπόψη πρόσθετοι παράγοντες όπως το όριο ταχύτητας της οδού και ο ψυχικός φόρτος εργασίας του οδηγού. Επίσης προτείνεται η ανάπτυξη ενός συστήματος ταξινόμησης βασισμένου στον συνδυασμό πολλαπλών μεθόδων.
Shangguan et al., 2021	Η έρευνα περιορίζεται μόνο στην επικίνδυνη επακολουθία των οχημάτων αγνοώντας επικίνδυνες συμπεριφορές κατά την αλλαγή λωρίδας. Επίσης σημαντικά χαρακτηριστικά του οχήματος και της οδού δεν λαμβάνονται υπόψη στην ανάλυση.	Κρίνεται αναγκαίο από τους μελετητές να συμπεριληφθούν επικίνδυνες ενέργειες αλλαγής λωρίδας σε μελλοντικές έρευνες. Επίσης μελλοντικά θα μπορούσαν να αναπτυχθούν ορισμένοι αλγόριθμοι βαθιάς εκμάθησης. Τέλος επιπρόσθετες μεταβλητές όπως τα χαρακτηριστικά του οχήματος και τα γεωμετρικά χαρακτηριστικά της οδού θα μπορούσαν να εξεταστούν για την βελτίωση και εξέλιξη του μοντέλου πρόβλεψης.
Yang et al., 2021	Τα δεδομένα που αναλύθηκαν δεν περιλάμβαναν την στάση, την αντίληψη και τα δημογραφικά χαρακτηριστικά των οδηγών.	Προτείνεται η εξέταση διαφορετικών χαρακτηριστικών του οδηγού. Επίσης, οι ερευνητές προτείνουν ότι μελλοντικές μελέτες θα μπορούσαν να εστιάσουν στις πιο σημαντικές μεταβλητές εξετάζοντας την σχέση τους με τα διαφορετικά επίπεδα ασφαλείας καθώς και την μεταξύ τους σχέση.

2.2.2 Αλγόριθμοι ταξινόμησης οδηγικής συμπεριφοράς

Οι επιστήμονες στον τομέα της οδικής ασφάλειας τείνουν να υιοθετούν όλο και περισσότερο **αλγόριθμους μηχανικής εκμάθησης** στον τομέα της οδικής ασφάλειας. Για την αναγνώριση της οδηγικής συμπεριφοράς και την ανάπτυξη μοντέλων πρόβλεψης συγκρούσεων σε πραγματικό χρόνο σχετικές μελέτες εστιάζουν στην ανάπτυξη διαφορετικών αλγορίθμων ταξινόμησης. Στον πίνακα 2.2 παρατίθενται οι αλγόριθμοι ταξινόμησης με την υψηλότερη επίδοση από τις έρευνες που παρουσιάστηκαν.

Πίνακας 2.2: Αποτελεσματικότεροι αλγόριθμοι ταξινόμησης ανά έρευνα ανάλυσης οδηγικής συμπεριφοράς

Έρευνα	**Σκοπός αλγορίθμων ταξινόμησης**	**Αλγόριθμοι ταξινόμησης με το υψηλότερο ποσοστό ορθών προβλέψεων**
Osman et al., 2019	Αναγνώριση ανάμιξης σε δευτερεύουσα ενέργεια	Δένδρα απόφασης (DT): ποσοστό ορθών προβλέψεων 78%
	Αναγνώριση τύπου δευτερεύουσας ενέργειας	Τυχαία δάση (RF): ποσοστό ορθών προβλέψεων 83%
Yi et al., 2019	Αναγνώριση της κατάστασης του οδηγού με βάση 3 κατηγορίες	Τυχαία δάση (RF): ποσοστό ορθών προβλέψεων 82%
Shi et al., 2019	Ταξινόμηση της οδηγικής συμπεριφοράς σε 4 επίπεδα	Ακραία Διαβαθμιζόμενη Ενδυνάμωση (XGBoost): ποσοστό ορθών προβλέψεων 89%
Song et al., 2021	Ταξινόμηση επικίνδυνης οδηγικής συμπεριφοράς σε 3 επίπεδα ασφαλείας	Τυχαία δάση (RF): ποσοστό ορθών προβλέψεων 90%
Ghandour et al., 2021	Αναγνώριση της κατάστασης του οδηγού με βάση 3 κατηγορίες	Διαβαθμιζόμενη ενδυνάμωση (GB): ποσοστό ορθών προβλέψεων 67%
Shangguan et al., 2021	Ταξινόμηση επικίνδυνης οδηγικής κατάστασης σε 4 επίπεδα	Πολυεπίπεδο Perceptron (MLP): ποσοστό ορθών προβλέψεων 85%
Yang et al., 2021	Ταξινόμηση οδηγικής συμπεριφοράς σε 4 επίπεδα ασφαλείας	Μηχανές διανυσμάτων υποστήριξης (SVM): ποσοστό ορθών προβλέψεων 95%

Όπως προκύπτει στην πλειονότητα των ερευνών που αναλύθηκαν ο αλγόριθμος ‘Τυχαίου δάσους’ σημειώνει την υψηλότερη επίδοση. Επίσης οι ‘Μηχανές διανυσμάτων υποστήριξης’ (Chandaka et al., 2009; Woo and Kulić, 2016) και το ‘Πολυεπίπεδο Perceptron’ (Assi, 2020; Pande and Abdel-Aty, 2006) που αποτελούν ευρέως διαδεδομένους αλγόριθμους στον τομέα της οδικής ασφάλειας, σημειώνουν υψηλές επιδόσεις.

2.2.3 Πρόβλημα ανισορροπίας δεδομένων σε κάθε τάξη

Στα προβλήματα του πραγματικού κόσμου συναντάται συχνά το **πρόβλημα της ανισορροπίας των δεδομένων όσον αφορά την κατανομή τους σε κάθε τάξη**. Συγκεκριμένα στις σχετικές μελέτες η επικίνδυνη οδηγική συμπεριφορά και η πιθανότητα ατυχήματος αποτελούν σπάνια φαινόμενα σε σχέση με την ασφαλή οδηγική συμπεριφορά και την μη πρόκληση ατυχήματος αντίστοιχα. Η κλάση με τα περισσότερα δεδομένα ονομάζεται κύρια κλάση ενώ εκείνη με τα λιγότερα ονομάζεται κλάση μειοψηφίας. Στα προβλήματα ανάλυσης συγκρούσεων σε πραγματικό χρόνο η αναλογία των συμβάντων ατυχήματος και μη κυμαίνεται από 1:5 (Roshandel et al., 2015) έως 1:20 (Xu et al., 2013).

Σύμφωνα με τους Elamrani Abou Elassad et al. (2020) υπάρχουν δύο βασικές προσεγγίσεις προκειμένου να αντιμετωπιστεί το πρόβλημα της ανισορροπίας: (i) Η προσθήκη συντελεστών βαρύτητας στα δεδομένα προκειμένου τα μοντέλα να δίνουν μεγαλύτερη έμφαση στην κλάση μειοψηφίας; (ii) Η ανάπτυξη κατάλληλων τεχνικών επαναδειγματοληψίας (resampling techniques) προκειμένου μειώνεται το μέγεθος της κύριας κλάσης ή να αυξάνεται εκείνο της κλάσης μειοψηφίας. Οι πιο γνωστές τεχνικές επαναδειγματοληψίας είναι η Τεχνική Συνθετικής Μειονοτικής Υπερδειγματοληψίας (Synthetic Minority Oversampling Technique, SMOTE) (Elamrani Abou Elassad et al., 2020; Guo et al., 2021; Morris and Yang, 2021; Shangguan et al., 2021) και η Προσαρμοστική Συνθετική Δειγματοληψία (Adaptive Synthetic Sampling, ADASYN) (Morris and Yang, 2021).

Επιπλέον αναζητώντας στην βιβλιογραφία προέκυψε ότι ο συνδυασμός της τεχνικής SMOTE με την τεχνική ENN (Edited Nearest Neighbor) είχε σημαντικά και ενδιαφέροντα αποτελέσματα στην αντιμετώπιση προβλημάτων πρόβλεψης και ταξινόμησης με ανομοιογενή δεδομένα (Katrakazas, 2017; Katrakazas et al., 2020, 2019).

Σε έρευνες διαφορετικού αντικειμένου εξετάζονται διάφορες τεχνικές όπως η Τυχαία Υπερδειγματοληψία (Random Oversampling), η SVM-SMOTE και η SMOTE-Tomek (Ghorbani & Ghousi, 2020). Όπως προτείνεται στην βιβλιογραφία είναι αναγκαίο να εξεταστούν οι διάφορες τεχνικές υποδειγματοληψίας και υπερδειγματοληψίας για την αντιμετώπιση της ανισορροπίας των δεδομένων καθώς η εφαρμογή τους θα διευρύνει περαιτέρω την έρευνα στα προβλήματα ανισορροπίας του τομέα της οδικής ασφάλειας.

Στην έρευνα των Ghandour et al. 2021 ο αλγόριθμος ‘Διαβαθμιζόμενης Ενδυνάμωσης’ (Gradient Boosting) παρέχει σημαντικά και ενθαρρυντικά αποτελέσματα, παρά την ανισορροπία των δειγμάτων σε σχέση με την κατανομή τους στις διαφορετικές τάξεις. Επίσης αξίζει να σημειωθεί ότι τα τελευταία χρόνια οι ενισχυτικές συνδυαστικές μέθοδοι (boosting ensemble methods) χρησιμοποιούνται από τους ερευνητές με σκοπό να βελτιώσουν την επίδοση της ταξινόμησης ανομοιογενών δεδομένων. Ο αλγόριθμος ‘Προσαρμοστικής Ενδυνάμωσης’ (AdaBoost) αποτελεί γνωστή συνδυαστική μέθοδο με τους Ariannezhad et al. (2021) να αποδεικνύουν ότι αξιοποιώντας τον για την μείωση της διακύμανσης των ανομοιογενών δεδομένων βελτιώνεται σημαντικά η ακρίβεια των προβλέψεων στα μοντέλα πρόβλεψης ατυχήματος σε πραγματικό χρόνο.

2.3 Διάρκεια οδήγησης σε επικίνδυνες συνθήκες

Η δημοσίευση των Michelaraki et al. (2021) έχει εκπονηθεί στο πλαίσιο του ερευνητικού έργου i-DREAMS. Σκοπός της μελέτης είναι να προσφέρει οδηγίες για την χαρτογράφηση

των μεθόδων και των προσεγγίσεων που θα χρησιμοποιηθούν για την επίτευξη των στόχων του έργου i-DREAMS. Το επίπεδο της 'Ζώνης Ανοχής Ασφαλείας' είναι γνωστό για κάθε παράγοντα κινδύνου. Όπως αναφέρεται από τους ερευνητές, σε κάθε φάση στοχεύεται ένας συγκεκριμένος παράγοντας κινδύνου για την αναγνώριση του επιπέδου ασφαλείας, αγνοώντας άλλες σημαντικές μεταβλητές. Υπάρχει ανάγκη αξιοποίηση του συνόλου των παραγόντων στο ίδιο μοντέλο για την πρόβλεψη του κινδύνου οδήγησης. Κατ' επέκταση πρέπει να προβλεφθεί η διάρκεια οδήγησης σε κάθε ένα από τα τρία επίπεδα ασφαλείας ώστε να αξιολογηθεί η επίδραση όλων των μεταβλητών στην επικίνδυνη οδήγηση. Το παραπάνω χαρακτηρίζεται ως **πρόβλημα παλινδρόμησης σε πραγματικό χρόνο**. Παρεμφερής μεθοδολογία εφαρμόζεται στα προβλήματα βραχυπρόθεσμης πρόβλεψης κυκλοφορίας (short-time traffic prediction) τα οποία αποτελούν κεντρικό θέμα στην έρευνα των Ευφυών Μεταφορικών Συστημάτων (ITS) (Hinsbergen et al., 2007). Παλαιότερες έρευνες έχουν εστιάσει στην αντιμετώπιση των προβλημάτων βραχυπρόθεσμης πρόβλεψης κυκλοφορίας αναπτύσσοντας κατάλληλους αλγόριθμους παλινδρόμησης (Chen et al., 2019; Liu et al., 2018; Jiaqi Wang et al., 2021).

2.4 Σύνοψη

Η διπλωματική εργασία θα επιχειρήσει να αξιοποιήσει μεθοδολογίες παλαιότερων ερευνών καλύπτοντας παράλληλα τις ελλείψεις τους. Επίσης θα προσπαθήσει να προσφέρει επιπλέον γνώση στον τομέα των Ευφυών Μεταφορικών Συστημάτων (ITS).

Κατά συνέπεια με βάση την βιβλιογραφική ανασκόπηση καθορίζεται ο **στόχος** της διπλωματικής εργασίας που αφορά στον εντοπισμό του επιπέδου και της διάρκειας επικίνδυνης συμπεριφοράς του οδηγού. Παράλληλα αξιολογώντας τις προσεγγίσεις των παλαιότερων ερευνών επιλέγεται η **κατάλληλη μεθοδολογία** για την επίτευξη του επιλεγμένου στόχου.

Συγκεκριμένα θα μελετηθεί η επίδραση των διάφορων οδηγικών χαρακτηριστικών στην αναγνώριση των διαφορετικών επιπέδων της 'Ζώνης Ανοχής Ασφαλείας' και θα εξεταστεί η μεταξύ τους σχέση.

Ιδιαίτερη έμφαση θα δοθεί στις **τεχνικές επαναδειγματοληψίας** για την αντιμετώπιση του προβλήματος ανισορροπίας των δεδομένων σε κάθε τάξη. Στο πλαίσιο της διπλωματικής εργασίας θα αξιολογηθούν οι μέθοδοι που έχουν χρησιμοποιηθεί σε παλαιότερες έρευνες καθώς και εκείνες που έχουν προταθεί για μελλοντική έρευνα.

Επιπλέον, με βάση την αποτελεσματικότητα των μοντέλων ταξινόμησης των ερευνών που παρουσιάστηκαν, επιλέγονται οι εξής **αλγόριθμοι για την ταξινόμηση** των οδηγών σε ένα από τα τρία επίπεδα της 'Ζώνης Ανοχής Ασφαλείας':

1. Ο αλγόριθμος ΄Τυχαία Δάση' (Random Forests) που ανήκει στην οικογένεια των συνδυαστικών μεθόδων μάθησης (ensemble learning methods). Στην πλειονότητα των ερευνών που παρουσιάστηκαν ο αλγόριθμος του 'Τυχαίου Δάσους' σημείωσε την υψηλότερη επίδοση.

2. Ο αλγόριθμος 'Προσαρμοστικής Ενδυνάμωσης' (AdaBoost). Η οικογένεια των συνδυαστικών μεθόδων μάθησης (ensemble learning methods) αντιμετωπίζει επιτυχώς προβλήματα αναγνώρισης επικίνδυνης οδηγικής συμπεριφοράς. Επίσης όπως αναφέρεται στην βιβλιογραφία ο αλγόριθμος AdaBoost μπορεί να διαχειριστεί αποτελεσματικά δεδομένα με άνιση κατανομή στις διάφορες τάξεις.

3. Ο αλγόριθμος ΄Μηχανών Διανυσμάτων Υποστήριξης’ (Support Vector Machines) λόγω της συχνής εφαρμογής του στην βιβλιογραφία καθώς και της υψηλής επίδοσης του στην συναφή έρευνα των Yang et al. (2021).

4. Ο αλγόριθμος ‘Πολυεπίπεδου Perceptron’ (Multilayer Perceptron) ο οποίος είχε υψηλή επίδοση στην παρεμφερή, ως προς τα δεδομένα και τον στόχο, έρευνα των Shangguan et al. (2021).

Με βάση την αξιολόγηση των αλγορίθμων στην βιβλιογραφία θα εξεταστούν ορισμένες **μετρικές αξιολόγησης**.

Σύμφωνα με τις οδηγίες της έρευνας των Michelaraki et al. (2021), στο δεύτερο μέρος της παρούσας εργασίας θα **προβλεφθεί** η διάρκεια οδήγησης κάθε οδηγού σε επικίνδυνες συνθήκες και θα **αξιολογηθεί** η επίδραση των μεταβλητών κινδύνου σε αυτόν. Με βάση την βιβλιογραφική ανασκόπηση προκύπτουν οι εξής αλγόριθμοι παλινδρόμησης για την πρόβλεψη της διάρκειας οδήγησης σε κάθε ένα από τα τρία επίπεδα ασφαλείας:

1. Ο αλγόριθμος παλινδρόμησης κορυφογραμμής (Ridge Regression). Ο όρος ομαλοποίησης L_2 μειώνει την επίδραση των λιγότερο σημαντικών μεταβλητών στο μοντέλο πρόβλεψης.

2. Ο αλγόριθμος παλινδρόμησης Lasso (Lasso Regression) διότι με την εισαγωγή του όρου ομαλοποίησης L_1 πραγματοποιεί πρακτικά επιλογή των σημαντικότερων μεταβλητών.

3. Ο αλγόριθμος παλινδρόμησης Elastic Net (Elastic Net Regression) καθώς αποτελεί συνδυασμό των δύο παραπάνω αλγορίθμων.

Οι παραπάνω αλγόριθμοι έχουν την δυνατότητα να διαχειρίζονται δεδομένα με υψηλή συγγραμμικότητα (Voss, 2005) καθώς και να πραγματοποιούν επιλογή των σημαντικότερων μεταβλητών (Lee et al., 2021). Με βάση τα δεδομένα και τον σκοπό της παρούσας εργασίας κρίνονται κατάλληλοι για την επίτευξη του στόχου που έχει τεθεί.

Αξίζει να σημειωθεί ότι στο πλαίσιο της βιβλιογραφικής ανασκόπησης δεν εντοπίστηκε ερεύνα που να σχετίζεται με την πρόβλεψη και την ανάλυση της διάρκειας οδήγησης σε επικίνδυνες συνθήκες. Επομένως η διερεύνηση του με την εφαρμογή κατάλληλων αλγορίθμων παλινδρόμησης δύναται να συνεισφέρει στις μελλοντικές έρευνες.

3. ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

3.1 Εισαγωγή

Στην παρούσα ενότητα περιγράφεται το θεωρητικό υπόβαθρο που πραγματοποιήθηκε η επεξεργασία των δεδομένων καθώς και η ανάλυση τους. Αρχικά, αναλύονται οι τεχνικές επεξεργασίας και επαναδειγματοληψίας των δεδομένων καθώς υπάρχει ανισορροπία των δειγμάτων σε κάθε τάξη. Στην συνέχεια παρουσιάζονται τα μοντέλα μηχανικής εκμάθησης που αναπτύχθηκαν για την ταξινόμηση της οδηγικής συμπεριφοράς στα τρία επίπεδα της Ζώνης Ανοχής Ασφαλείας. Επιπλέον αναλύονται τα μοντέλα παλινδρόμησης που αξιοποιήθηκαν για την πρόβλεψη της διάρκειας που πραγματοποιείται η οδήγηση σε επικίνδυνες συνθήκες. Τέλος, δίνεται ιδιαίτερη έμφαση στην σημασία των μετρικών αξιολόγησης, των κριτηρίων αποδοχής των μοντέλων και των στατιστικών ελέγχων των αποτελεσμάτων.

3.2 Επιλογή χαρακτηριστικών (Feature Selection)

Η **επιλογή χαρακτηριστικών** αφορά στην διαδικασία βέλτιστης επιλογής δεδομένων από το σύνολο τους προκειμένου να μειωθεί ο αριθμός των μεταβλητών εισόδου. Αυτή η διαδικασία επιταχύνει την ταξινόμηση καθώς μειώνει την υπολογιστική πολυπλοκότητα και τα σφάλματα πρόβλεψης (Elamrani Abou Elassad et al., 2020). Για βέλτιστη επιλογή των χαρακτηριστικών της παρούσας εργασίας συνδυάστηκαν δύο μέθοδοι.

Η πρώτη αφορά στον προσδιορισμό της συσχέτισης (correlation) μεταξύ των ανεξάρτητων μεταβλητών. Ο δειγματικός συντελεστής γραμμικής συσχέτισης του Pearson για δύο μεταβλητές συμβολίζεται με r και ορίζεται από την μαθηματική εξίσωση 3.1.

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (3.1)$$

Όπου,

x_i, y_i οι τιμές των δύο μεταβλητών

$\bar{x}$, $\bar{y}$ ο μέσος όρος των τιμών

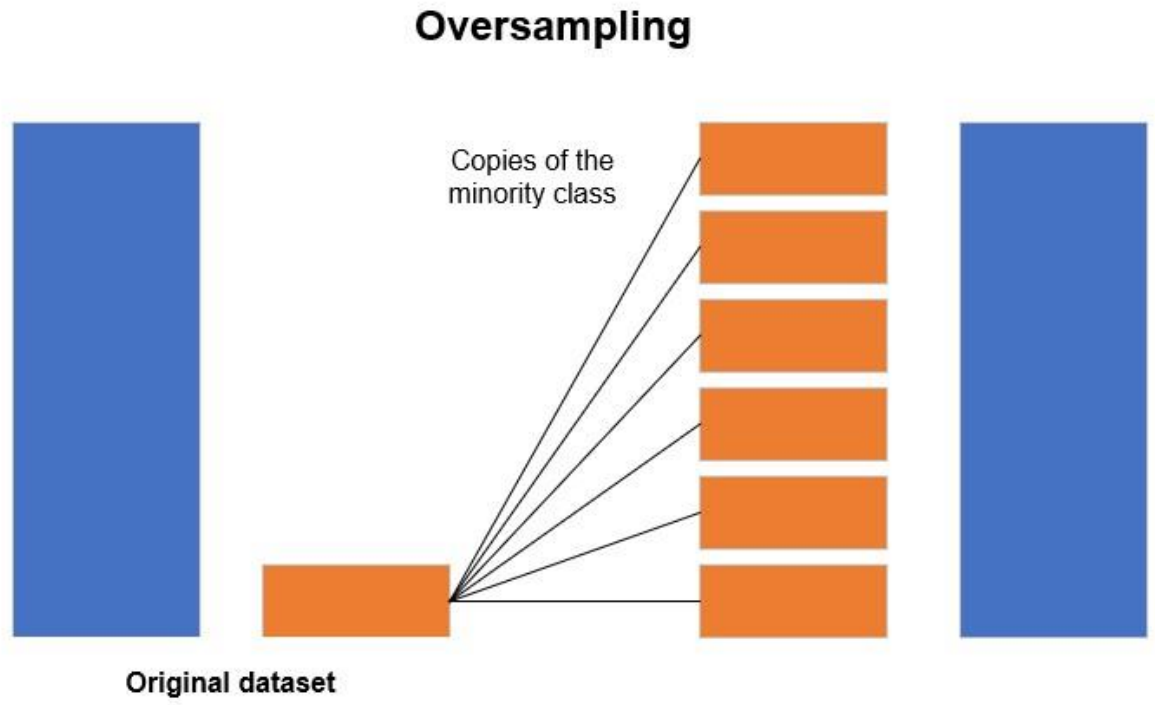
Οι τιμές κυμαίνονται μεταξύ -1 και 1, οπού r=0 σημαίνει μηδενική συσχέτιση, r=1 πλήρη θετική συσχέτιση και r=-1 πλήρη αρνητική συσχέτιση (Nettleton, 2014). Θετική συσχέτιση μεταξύ δύο μεταβλητών παρατηρείται όταν η τιμή της μίας αυξάνεται και ταυτόχρονα αυξάνεται η άλλη. Ενώ αρνητική συσχέτιση παρατηρείται όταν η μία αυξάνεται και η άλλη μειώνεται. Το βέλτιστο υποσύνολο αποτελείται από χαρακτηριστικά αρκετά συσχετισμένα με την προβλεπόμενη τάξη αλλά έχοντας ελάχιστη συσχέτιση μεταξύ τους (Hall, 2000). Επομένως στο πλαίσιο της μελέτης μεταξύ δύο χαρακτηριστικών με υψηλή συσχέτιση το ένα θα απορριφθεί για την περαιτέρω ανάλυση.

Η δεύτερη μέθοδος αφορά στην επιλογή χαρακτηριστικών βασιζόμενη στον εντοπισμό της σημαντικότητας χαρακτηριστικών (feature importance) του δείγματος. Η παραπάνω διαδικασία μπορεί να επιτευχθεί με την εκπαίδευση αλγορίθμων μηχανικής εκμάθησης όπως είναι ο ταξινομητής τυχαίου δάσους (random forest classifier) ώστε να προσδιοριστεί ο βαθμός επιρροής κάθε μεταβλητής στην τελική πρόβλεψη.

3.3 Μέθοδοι επαναδειγματοληψίας για προβλήματα ανισορροπίας ταξινόμησης

Οι περισσότεροι αλγόριθμοι μηχανικής εκμάθησης που χρησιμοποιούνται για ταξινόμηση, είναι βασισμένοι στην θεώρηση ότι όλες οι κλάσεις έχουν τον ίδιο αριθμό δεδομένων. Επίσης, τα περισσότερα προβλήματα του αληθινού κόσμου αφορούν **δεδομένα** με ανισορροπία όσον αφορά την κατανομή τους στις διαφορετικές κλάσεις (Kotsiantis et al., 2005) με αποτέλεσμα να αυξάνουν την μεροληψία (bias) του αλγορίθμου (Fernández et al., 2009) ως προς την κυρίαρχη κλάση (majority class) και να τον καθιστούν περισσότερο ευαίσθητο σε σφάλματα ταξινόμησης για την κλάση μειοψηφίας (minority class).

Δεδομένου ότι στην παρούσα μελέτη οι κλάσεις μειοψηφίας είναι τα δύο επίπεδα επικίνδυνης οδήγησης, ενώ η κυρίαρχη κλάση είναι το επίπεδο ασφαλούς οδήγησης καθίσταται σαφής η ανάγκη ανάπτυξης τεχνικών επαναδειγματοληψίας των δεδομένων εκπαίδευσης των αλγόριθμων. Οι επιπτώσεις στην ασφάλεια των οδηγών θα ήταν ιδιαίτερα σοβαρές εάν τα μοντέλα μηχανικής εκμάθησης ταξινομούσαν λανθασμένα επικίνδυνες συμπεριφορές ως ασφαλείς.



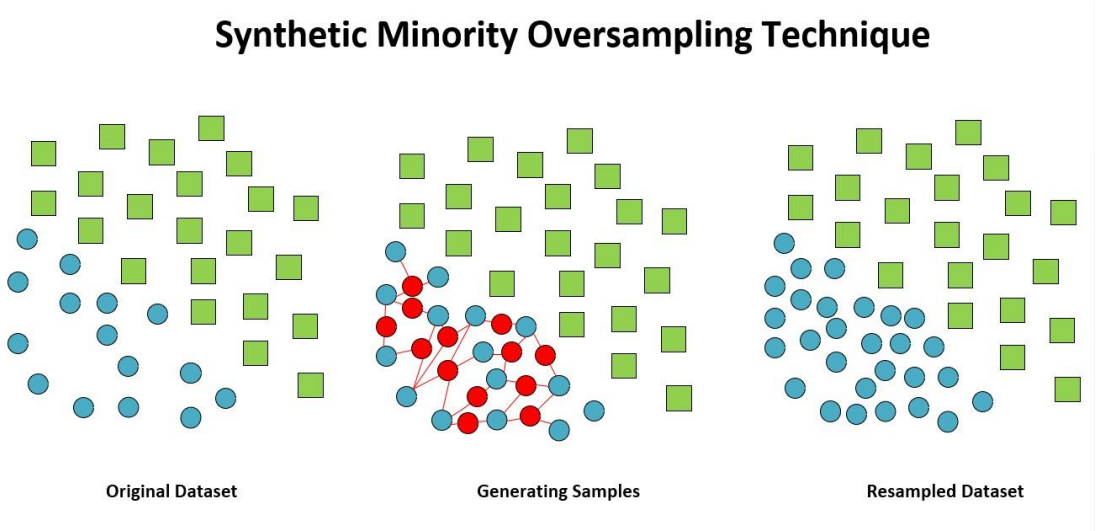
Γράφημα 3.1: Επαναδειγματοληψία δεδομένων που ανήκουν στην κλάση μειοψηφίας
Πηγή: Kaggle (2022)

3.3.1 Τεχνική Συνθετικής Μειονοτικής Υπερδειγματοληψίας (SMOTE)

Η **Τεχνική Συνθετικής Μειονοτικής Υπερδειγματοληψίας** (Synthetic Minority Over-sampling Technique, SMOTE) χρησιμοποιείται ευρέως στην επιστήμη της μηχανικής εκμάθησης και θεωρείται μία από τις ισχυρότερες τεχνικές επαναδειγματοληψίας για προβλήματα ανισόρροπης ταξινόμησης.

Ουσιαστικά η τεχνική δημιουργεί συνθετικά δεδομένα για την κλάση της μειοψηφίας με σκοπό να εξαλειφθεί η ανομοιογένεια των δειγμάτων στις κλάσεις. Αρχικά, εντοπίζονται οι k-κοντινότεροι γείτονες από κάθε δεδομένο της μειονοτικής κλάσης. Κατόπιν, επιλέγεται τυχαία ένας από τους k-γείτονες και υπολογίζεται η μεταξύ τους απόσταση. Τέλος, η

διαφορά τους πολλαπλασιάζεται με έναν τυχαίο αριθμό από το 0 έως το 1 και το νέο δεδομένο που δημιουργείται συνυπολογίζεται στην κλάση μειοψηφίας. (Chawla et al., 2002)



Γράφημα 3.2: SMOTE: Synthetic Minority Oversampling Technique
Πηγή: Medium (2022)

3.3.2 Προσαρμοστική Συνθετική Τεχνική (ADASYN)

Οι He et al., (2008) εισήγαγαν την **Προσαρμοστική Συνθετική Τεχνική** (Adaptive Synthetic, ADASYN) η οποία επικεντρώνεται στην παραγωγή δύσκολων ως προς την μάθηση δειγμάτων (Islam et al., 2021).

Η μέθοδος ADASYN χρησιμοποιεί μία σταθμισμένη κατανομή για διαφορετικά παραδείγματα της κλάσης μειοψηφίας ανάλογα με το επίπεδο δυσκολίας τους στη μάθηση. Με βάση αυτή την κατανομή παράγονται περισσότερα συνθετικά δεδομένα για τα παραδείγματα της κλάσης μειοψηφίας τα οποία είναι πιο δύσκολο να μαθευτούν σε σύγκριση με εκείνα που είναι πιο εύκολο.

Αποτέλεσμα της παραπάνω προσέγγισης είναι η βελτίωση της διαδικασίας μάθησης με δύο τρόπους:

1. Μειώνει την μεροληψία που προέκυψε από την ανισορροπία των κλάσεων.

2. Μετατοπίζει προσαρμοστικά τα όρια απόφασης της ταξινόμησης προς τα δύσκολα παραδείγματα.

Έρευνες κατά το παρελθόν έχουν δείξει ότι η χρήση της μεθόδου ADASYN έχει ικανοποιητικά αποτελέσματα για την αντιμετώπιση δεδομένων με ανισορροπία κατανομής στις διαφορετικές τάξεις (Islam et al., 2021; Morris and Yang, 2021; Y. Song et al., 2021).

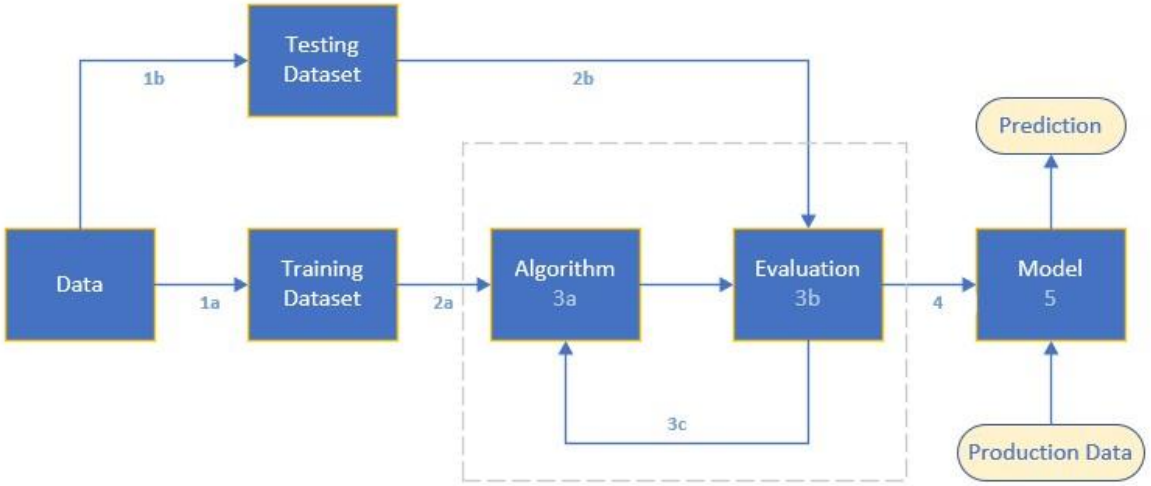
3.4 Αλγόριθμοι ταξινόμησης (Classification algorithms)

Η **ταξινόμηση** ανήκει στην επιβλεπόμενη μάθηση (supervised learning) καθώς η εκπαίδευση του αλγορίθμου γίνεται με βάση μεταβλητών γνωστής κλάσης (Junhua Wang et al., 2021).

Στην παρούσα έρευνα αναπτύχθηκαν 4 αλγόριθμοι μηχανικής εκμάθησης με σκοπό την ταξινόμηση δεδομένων πολλαπλών κλάσεων (multiclass classification).

Γενικά στην η διαδικασία ανάπτυξης αλγορίθμων μηχανικής εκμάθησης περιλαμβάνει ορισμένα σημαντικά βήματα. Αρχικά διαχωρίζονται τα δεδομένα σε δύο κατηγορίας, στα δεδομένα εκπαίδευσης (training dataset) και στα δεδομένα εξέτασης (testing dataset). Τα δεδομένα εκπαίδευσης εξασκούν τον αλγόριθμο και τα δεδομένα εξέτασης αξιοποιούνται για την αξιολόγηση του μοντέλου. Η αποτελεσματικότητα των μοντέλων κρίνεται βάση ορισμένων σημαντικών μετρικών αξιολόγησης.

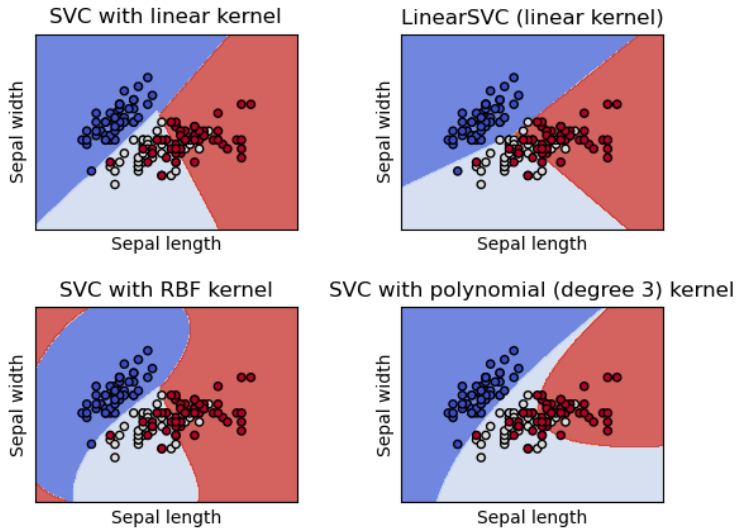
Στο γράφημα 3.2, περιγράφεται η διαδικασία ανάπτυξης, αξιολόγησης και λειτουργίας των μοντέλων μηχανικής εκμάθησης.



Γράφημα 3.3: Διαδικασία Μηχανικής Εκμάθησης
Πηγή: Towards Data Science (2022)

3.4.1 Μηχανές διανυσμάτων υποστήριξης (Support Vector Machines)

Οι **μηχανές διανυσμάτων υποστήριξης** (Support Vector Machines) αποτελούν μοντέλα επιβλεπόμενης μηχανικής εκμάθησης και χρησιμοποιούνται στην επίλυση προβλημάτων ταξινόμησης και παλινδρόμησης. Η μέθοδος στοχεύει στον εντοπισμό μίας εξίσωσης σε πολυδιάστατο χώρο η οποία θα μπορεί να διαχωρίσει τα δεδομένα εκπαίδευσης γνωστής κλάσης (Dukart, 2015). Ο διαχωρισμός πραγματοποιείται με την κατασκευή ενός υπερεπιπέδου μέγιστων περιθωρίων (maximum margin hyperplane) για την μείωση της απόστασης των λανθασμένα ταξινομημένων σημείων από τα όρια απόφασης (Yu & Abdel-Aty, 2013). Με την χρήση της μεθόδου των πυρήνων (kernel method) ο ταξινομητής διανυσμάτων υποστήριξης μπορεί να διαχειριστεί δεδομένα μη γραμμικά διαχωρίσιμα.



Γράφημα 3.4: Αλγόριθμος SVC με την χρήση διαφορετικών μεθόδων πυρήνων
Πηγή: Scikit-Learn (2022)

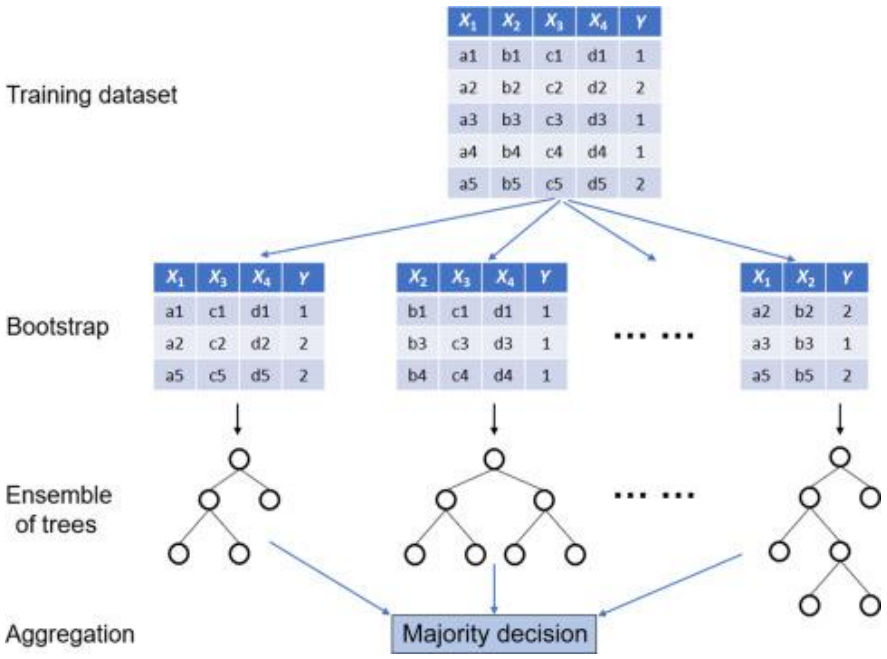
3.4.2 Τυχαία δάση (Random Forests)

Τα δένδρα απόφασης (decision trees) αποτελούν ευρέως διαδεδομένη τεχνική ταξινόμησης λόγω της απλότητας τους και της εύκολης κατανόησης. Έχουν δενδροειδή μορφή όμοια με τα διαγράμματα ροής και με βάση την αλληλουχία αποφάσεων κάθε κόμβος χωρίζεται σε δύο μέρη (Lugo Reyes, 2020). Η διαδικασία που ακολουθεί το δένδρο απόφασης είναι η εξής:

1. Αρχικοποίηση του κόμβου με το σύνολο των δεδομένων
2. Διάσπαση του κόμβου με βάση κάποιο κριτήριο διαχωρισμού σε κάποιο από τα γνωρίσματα.
3. Επανάληψη του βήματος 2 έως ότου ικανοποιηθεί το κριτήριο τερματισμού και τα δεδομένα έχουν ταξινομηθεί με βάση τα γνωρίσματα τους μέσω ενός συστήματος αποφάσεων

Ο δείκτης gini (gini index) και η εντροπία (entropy) αποτελούν τα κριτήρια υπολογισμού του κέρδους πληροφορίας. Οι αλγόριθμοι δένδρων απόφασης αξιοποιούν το κέρδος πληροφορίας για τον βέλτιστο αριθμό διαχωρισμών του κάθε κόμβου.

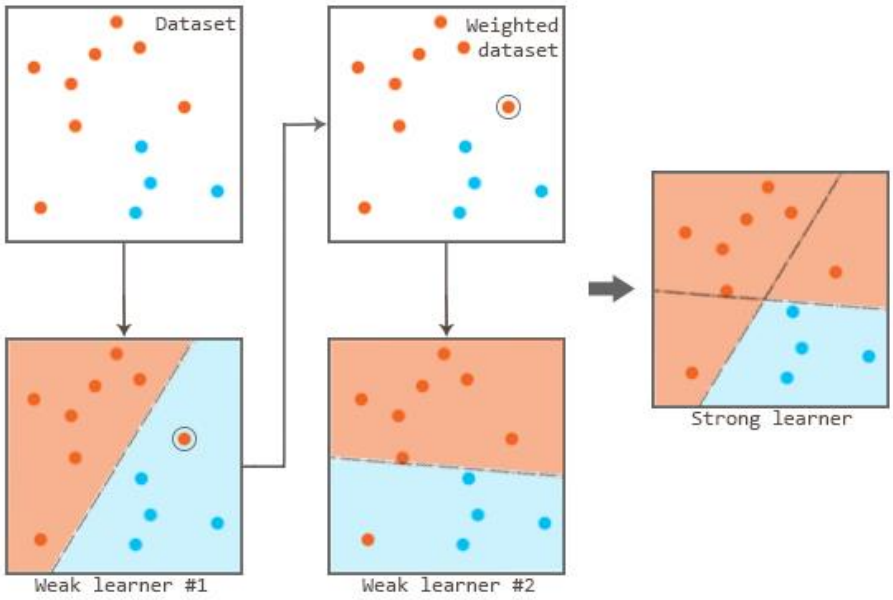
Το μοντέλο των **τυχαίων δασών** αποτελεί συνδυαστική μέθοδο που εκπαιδεύει παράλληλα πολλαπλά δένδρα απόφασης αξιοποιώντας την τεχνική bagging δηλαδή των συνδυασμό bootstrapping και aggregation (Misra and Li, 2020). Η τεχνική bootstrapping περιγράφεται ως η παράλληλη εκπαίδευση πολλαπλών δένδρων αποφάσεων χρησιμοποιώντας διαφορετικά υποσύνολα από το σύνολο των δεδομένων. Για την τελική απόφαση ο ταξινομητής συνδυάζει τις αποφάσεις των επιμέρους δένδρων απόφασης.



Γράφημα 3.5: Διαδικασία ταξινόμησης Τυχαίου Δάσους
Πηγή: Misra & Li (2020)

3.4.3 Προσαρμοστική ενδυνάμωση (AdaBoost)

Το μοντέλο **προσαρμοστικής ενδυνάμωσης** αποτελεί συνδυαστική μέθοδο ταξινόμησης. Με την υλοποίηση ενδυνάμωσης (boosting) το σύνολο των αδύναμων ταξινομητών συνδέεται σε σειρά προκειμένου ο κάθε ένας να προσπαθεί να βελτιώσει την ταξινόμηση των δειγμάτων που είχαν ταξινομηθεί εσφαλμένα από τον προηγούμενο (Misra and Li, 2020). Το βάρος των λανθασμένα ταξινομημένων δειγμάτων από το προηγούμενο δένδρο ενδυναμώνεται προκειμένου το επόμενο δένδρο να επικεντρωθεί στην σωστή ταξινόμηση τους.

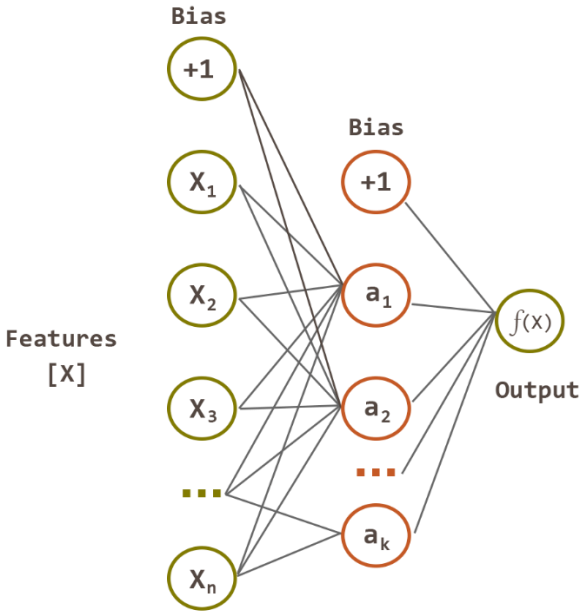


Γράφημα 3.6: Λειτουργία του αλγορίθμου AdaBoost
Πηγή: Misra & Li (2020)

3.4.4 Πολυεπίπεδο perceptron (Multilayer Perceptron)

Το μοντέλο **πολυεπίπεδου perceptron** ανήκει στην ευρύτερη κατηγορία των νευρικών δικτύων πρόσθιας τροφοδοσίας (feed forward neural network) (Abirami and Chitra, 2020). Αποτελείται από τρείς κατηγορίες επιπέδων (layers), το επίπεδο εισόδου (input), το κρυφό επίπεδο (hidden) και το επίπεδο εξόδου (output). Η διαδικασία της ταξινόμησης πραγματοποιείται από το επίπεδο εξόδου ενώ το κρυφό επίπεδο έχει καταλυτικό ρόλο στο σύνολο της διαδικασίας (Abirami and Chitra, 2020).

Η δομή του μοντέλου φαίνεται σχηματικά στο γράφημα 3.7. Το επίπεδο εισόδου αποτελείται από ένα πλήθος νευρώνων (x_1,x_2,..,x_n) που αντιπροσωπεύουν το πλήθος των χαρακτηριστικών εισόδου. Στην συνέχεια, το κρυφό επίπεδο μετασχηματίζει τις τιμές εισόδου με την χρήση σταθμισμένης γραμμικής άθροισης ($w_1x_1+w_2x_2+…+w_nx_n$) η οποία ακολουθείται από μία μη γραμμική εξίσωση ενεργοποίησης. Τέλος, το επίπεδο εξόδου λαμβάνει τις τιμές του κρυφού επιπέδου και τις μετασχηματίζει στις τελικές τιμές εξόδου.



Γράφημα 3.7: MLP με ένα κρυφό επίπεδο
Πηγή: Scikit-Learn (2022)

3.5 Αλγόριθμοι παλινδρόμησης (Regression algorithms)

Οι αλγόριθμοι μηχανικής εκμάθησης μπορούν να χωριστούν σε παραμετρικούς και μη παραμετρικούς. Τα παραμετρικά μοντέλα μηχανικής εκμάθησης έχουν την ικανότητα να απλοποιούν την διαδικασία της μάθησης. Η γραμμική παλινδρόμηση (linear regression) όπως και η λογιστική παλινδρόμηση (logistic regression) αποτελούν παραδείγματα παραμετρικών μοντέλων.

Η πολλαπλή γραμμική παλινδρόμηση (multiple linear regression) χρησιμοποιείται προκειμένου να εκτιμηθεί η στατιστική σημαντικότητα και η σχέση μεταξύ μίας εξαρτημένης μεταβλητής (y) και πολλαπλών ανεξάρτητων μεταβλητών (x_i) (Djuris et al., 2013). Η επίδραση της κάθε ανεξάρτητης μεταβλητής στην εξαρτημένη εκφράζεται μέσω των συντελεστών παλινδρόμησης (coefficient regression). Ο μαθηματικός τύπος της πολλαπλής γραμμικής παλινδρόμησης εκφράζεται από την μαθηματική εξίσωση 3.2.

$$y = b_0 + \sum_{i=1}^{n} b_i x_i + e \quad (3.2)$$

Όπου:
y: εξαρτημένη μεταβλητή
x_i: ανεξάρτητες μεταβλητές
b_i: συντελεστές παλινδρόμησης
e: σφάλμα παλινδρόμησης

Το άθροισμα τετραγώνων υπολοίπων (Residual Sum of Squares) εκφράζει την απόκλιση των δεδομένων από το βέλτιστο μοντέλο και ορίζεται από την εξίσωση 3.3.

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{P} b_i x_{ij} \right)^2 \quad (3.3)$$

Η ελαχιστοποίηση του τετραγώνου των παρατηρήσεων επιδιώκεται στο πλαίσιοβελτιστοποίησης του μοντέλου παλινδρόμησης.

3.5.1 Παλινδρόμηση κορυφογραμμής (Ridge Regression)

Η πολύ-συγγραμμικότητα (multicollinearity) μεταξύ των δεδομένων αποτελεί συχνό φαινόμενο και οδηγεί σε ανακριβείς εκτιμήσεις των συντελεστών παλινδρόμησης μειώνοντας την ικανότητα πρόβλεψης του μοντέλου.

Για τον λόγο αυτό έχει εισαχθεί η τεχνική **παλινδρόμησης κορυφογραμμής** για την ανάλυση πολλαπλής γραμμικής παλινδρόμησης δεδομένων που πάσχουν από υψηλή συσχέτιση. Κύρια ιδέα του μοντέλου είναι η κανονικοποίηση του ελάχιστου αθροίσματος των τετραγώνων (least squares) με την χρήση μίας παραμέτρου κανονικοποίησης λ (regularization parameter) (Theodoridis, 2020). Ρόλος της παραμέτρου λ είναι να μειώσει το πλάτος των συντελεστών παλινδρόμησης (β) προς το 0, μειώνοντας έτσι την μεταβλητότητα των εκτιμήσεων. Οι εκτιμώμενες μεταβλητές (β) της παλινδρόμησης κορυφογραμμής ελαχιστοποιούν την συνάρτηση (James et al., 2013):

$$\sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{P} b_i x_{ij} \right)^2 + \lambda \sum_{j=1}^{P} b_j^2 = RSS + \lambda \sum_{j=1}^{P} b_j^2 \quad (3.4)$$

Η τεχνική κανονικοποίηση της παλινδρόμησης κορυφογραμμής ονομάζεται L_2 κανονικοποίηση (L_2 regularization).

3.5.2 Παλινδρόμηση Lasso (Lasso Regression)

Η **παλινδρόμηση Lasso** (Least Absolute Shrinkage and Selection Operator) αποτελεί παραλλαγή της παλινδρόμησης κορυφογραμμής αφού και εκείνη ομαλοποιεί την συνάρτηση κόστους με την χρήση μίας παραμέτρου κανονικοποίησης, αλλά ταυτόχρονα πραγματοποιεί επιλογή των σημαντικότερων επεξηγηματικών μεταβλητών του δείγματος αγνοώντας όσες έχουν ελάχιστη επίδραση στην πρόβλεψη. Με την χρήση της τεχνικής κανονικοποίησης L_1 (L_1 regularization) οι συντελεστές των λιγότερο σημαντικών μεταβλητών τείνουν προς το μηδέν πραγματοποιώντας πρακτικά επιλογή των περισσότερο επεξηγηματικών μεταβλητών (Ng, 2004). Οι εκτιμώμενες μεταβλητές (β) της παλινδρόμησης Lasso ελαχιστοποιούν την συνάρτηση (James et al., 2013):

$$\sum_{i=1}^{n}\left(y_i - b_0 - \sum_{j=1}^{P} b_i x_{ij}\right)^2 + \lambda \sum_{j=1}^{P} b_j^2 = RSS + \lambda_1 \sum_{j=1}^{P} |\beta j| + \lambda_2 \sum_{j=1}^{P} b_j^2 \qquad (3.5)$$

3.5.3 Παλινδρόμηση Elastic Net (Elastic Net Regression)

Η παλινδρόμηση Elastic Net (Zou and Hastie, 2005) είναι ο συνδυασμός των Ridge και Lasso παλινδρομήσεων. Αποτελεί έναν ιδιαίτερα αποδοτικό αλγόριθμο καθώς συνδυάζει τις ιδιότητες των άλλων δύο εισάγοντας δύο όρους κανονικοποίησης στην συνάρτηση κόστους. Ο ένας όρος είναι όμοιος με εκείνον του Ridge και ο άλλος όμοιος με εκείνον του Lasso. Οι εκτιμώμενες μεταβλητές (β) της παλινδρόμησης Lasso ελαχιστοποιούν την συνάρτηση:

$$\sum_{i=1}^{n}\left(y_i - b_0 - \sum_{j=1}^{P} b_i x_{ij}\right)^2 + \lambda \sum_{j=1}^{P} b_j^2 = RSS + \lambda \sum_{j=1}^{P} |\beta j| \qquad (3.6)$$

3.6 Μετρικές αξιολόγησης για ταξινόμηση (Evaluation metrics for classification)

3.6.1 Μήτρα σύγχυσης (Confusion matrix)

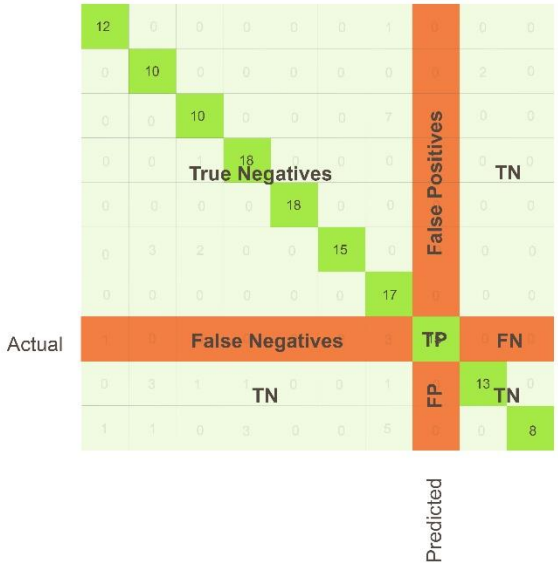
Η μήτρα σύγχυσης αποτελεί την σύνοψη της ταξινόμησης και δίνει σημαντικά συμπεράσματα για τον τύπο των σφάλματων του μοντέλου. Στην διαδικασία της επιβλεπόμενης μάθησης, σε συνέχεια από την εκπαίδευση του αλγορίθμου χρησιμοποιούνται τα δεδομένα εξέτασης για την αξιολόγηση του. Οι διαφορές μεταξύ των προβλεπόμενων και των πραγματικών κλάσεων αναπαρίστανται στην μήτρα σύγχυσης. Τα τέσσερα πιθανά αποτελέσματα είναι τα εξής:

- ‘Πραγματικά Θετικά’ (True Positive): Όσες περιπτώσεις ανήκουν στην κλάση i και ταξινομήθηκαν σε αυτήν.

- ‘Πραγματικά Αρνητικά’ (True Negative): Όσες περιπτώσεις δεν ανήκουν στην κλάση i και δεν ταξινομήθηκαν σε αυτήν.

- ‘Ψευδώς Θετικά’ (False Positive): Όσες περιπτώσεις δεν ανήκουν στην κλάση i αλλά ταξινομήθηκαν σε αυτήν. Τα ‘ψευδώς θετικά’ ορίζονται σαν τύπο σφάλματος 1.

- ΄Ψευδώς Αρνητικά’ (False Negative): Όσες περιπτώσεις ανήκουν στην κλάση i αλλά δεν ταξινομήθηκαν σε αυτήν. Τα ‘ψευδώς αρνητικά’ ορίζονται σαν τύπο σφάλματος 2.



Γράφημα 3.8: Confusion matrix for multiclass classification
Πηγή: Towards Data Science (2022)

3.6.2 Ορθότητα (Accuracy)

Η πιο γνωστή μετρική αξιολόγησης είναι η ορθότητα (accuracy) η οποία υπολογίζει το ποσοστό των στοιχείων που ταξινομήθηκαν σωστά από το σύνολο τους κατά την εξέταση του αλγορίθμου. Ουσιαστικά υποδεικνύει την πιθανότητα ότι η πρόβλεψη του μοντέλου είναι σωστή (Grandini et al., 2020).

$$\text{Ορθότητα (Accuracy)} = \frac{TP+TN}{TP+FP+FN+TN} \qquad (3.7)$$

Ωστόσο, ιδιαίτερα σε προβλήματα με ανομοιογενή δεδομένα δημιουργείται το λεγόμενο παράδοξο ορθότητας (‘Accuracy Paradox’), όπου η υπολογισμένη ορθότητα επηρεάζεται από την κυρίαρχη κλάση και δεν αντικατοπτρίζει την πραγματική κατάσταση. Για τον λόγο αυτό καταστρατηγούνται επιπλέον μετρικές αξιολόγησης.

3.6.3 Ακρίβεια (Precision)

Η ακρίβεια (precision) εκφράζει το ποσοστό των στοιχείων που πραγματικά ανήκουν στην κλάση i από το σύνολο των στοιχείων που αλγόριθμος ταξινόμησε στην κλάση i.

$$\text{Ακρίβεια (Precision)} = \frac{TP}{TP+FP} \qquad (3.8)$$

3.6.4 Ανάκληση (Recall)

Η ανάκληση (recall) περιγράφει το ποσοστό των στοιχείων που στην πραγματικότητα ανήκουν στην κλάση i και ο αλγόριθμος κατάφερε να τα ταξινομήσει ορθά στην κλάση i.

$$\text{Ανάκληση (Recall)} = \frac{TP}{TP+FN} \quad (3.9)$$

Στην παρούσα μελέτη οι επιπτώσεις της λανθασμένης ταξινόμησης μίας επικίνδυνης κλάσης σαν λιγότερο επικίνδυνη ή ασφαλή θα είχε σημαντικές επιπτώσεις. Για αυτό τον λόγο στην παρούσα μελέτη ο τύπος σφάλματος 2 είναι κρισιμότερος από τον τύπο σφάλματος 1 και καθιστά την μετρική ανάκλησης ιδιαίτερα σημαντική.

3.6.5 Ρυθμός λανθασμένων θετικών προβλέψεων (False positive rate)

Ο ρυθμός λανθασμένων θετικών προβλέψεων (False positive rate, FPR) ή ρυθμός λανθασμένου συναγερμού (False alarm rate, FAR) περιγράφεται από τον τύπο:

$$\text{FPR ή FAR} = \frac{FP}{FP+TN} \quad (3.11)$$

3.6.6 f1-score

Το f1-score αποτελεί τον αρμονικό μέσο όρο μεταξύ της ακρίβειας και της ανάκλησης.

$$\text{f1-score} = \frac{2x\ (Ακρίβεια)x\ (Ανάκληση)}{(Ακρίβεια)+(Ανάκληση)} \quad (3.11)$$

3.7 Κριτήρια αποδοχής μοντέλων παλινδρόμησης

3.7.1 Συντελεστής προσδιορισμού (Coefficient of determination)

Ο συντελεστής προσδιορισμού (coefficient of determination) R^2, υπολογίζει το ποσοστό διακύμανσης της εξαρτημένης μεταβλητής (Y) που ερμηνεύεται από τις ανεξάρτητες μεταβλητές (X).

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}-\bar{y})^2}{\sum_{i=1}^{n}(yi-\bar{y})^2} \quad (3.12)$$

Όπου:
n: αριθμός παρατηρήσεων
y_i: πραγματικές τιμές εξαρτημένης μεταβλητής Y
$\bar{y}$: μέση τιμή της μεταβλητής Y
$\hat{y}_i$: εκτιμημένες τιμής της μεταβλητής Y

Ο συντελεστής προσδιορισμού μετρά την ικανότητα των παραγόντων να ερμηνεύσουν ένα φαινόμενο και οι τιμές του κυμαίνονται από 0 έως 1. Το βέλτιστό μοντέλο, για το οποίο οι ανεξάρτητες μεταβλητές ερμηνεύουν 100% την διακύμανση της εξαρτημένης μεταβλητής, έχει συντελεστή R^2 ίσο με 1. Αντίθετα, όταν οι ανεξάρτητες μεταβλητές δεν

μπορούν να ερμηνεύσουν καθόλου την διακύμανση της εξαρτημένης μεταβλητής ο συντελεστής R^2 είναι ίσος με το 0.

3.7.2 Έλεγχος στατιστικής σημαντικότητας (Test of statistical significance)

Προκειμένου να αξιολογηθεί η επιρροή των μεταβλητών στην εμφάνιση ενός φαινομένου πρέπει πρώτα να εξεταστεί η στατιστική σημαντικότητα αυτών με τον έλεγχο στατιστικών υποθέσεων. Οι δύο στατιστικές υποθέσεις που μελετώνται είναι η μηδενική H_0 (null hypothesis) και η εναλλακτική H_1 (alternative hypothesis). Η μηδενική υπόθεση αφορά την συντηρητική υπόθεση του ερευνητικού προβλήματος ενώ η εναλλακτική την επιδιωκόμενη θεώρηση που προκύπτει από την απόρριψη της H_0.

Όταν σε ένα επίπεδο σημαντικότητας α απορρίπτεται η μηδενική υπόθεση (H_0) τότε το δείγμα χαρακτηρίζεται στατιστικά σημαντικό και υποδηλώνει ότι η επιρροή του στην εμφάνιση του φαινομένου δεν οφείλεται στην τυχαιότητα.

Η στατιστική σημαντικότητα ελέγχεται με την σύγκριση της p-value και του επιπέδου σημαντικότητας.

- Αν p-value < α: απορρίπτουμε την H_0
- Αν p-value > α: δεν απορρίπτουμε την H_0

Επίσης ο έλεγχος πραγματοποιείται μέσω σύγκρισης της τιμής t-value και της κατανομής t-student.

- Αν |t-value| > t-student: απορρίπτουμε την H_0
- Αν |t-value| < t-student: δεν απορρίπτουμε την H_0

Στην παρούσα μελέτη απαραίτητο βήμα αποτελεί ο έλεγχος στατιστικής σημαντικότητας των ανεξάρτητων μεταβλητών που λαμβάνονται υπόψιν για την πρόβλεψη της διάρκειας οδήγησης σε επικίνδυνες συνθήκες.

4. ΣΥΛΛΟΓΗ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑ ΣΤΟΙΧΕΙΩΝ

4.1 Εισαγωγή

Όπως αναφέρθηκε στις προηγούμενες ενότητες, στόχος του ερευνητικού έργου i-DREAMS είναι ο ορισμός, η ανάπτυξη, η δοκιμή και η επικύρωση μίας 'Ζώνης Ανοχής Ασφαλείας' ώστε να περιορίζεται η επικίνδυνη συμπεριφορά κατά την οδήγηση μέσω παρεμβάσεων σε πραγματικό ή μεταγενέστερο χρόνο. Η ανάλυση συγκεκριμένων δεδομένων της οδηγικής συμπεριφοράς και του οδικού περιβάλλοντος αποτελούν καίριο βήμα για την επίτευξη των παραπάνω στόχων. Με την ανάλυση δεδομένων της έρευνας επιδιώκεται:

1. Η αναγνώριση του επιπέδου της 'Ζώνης Ανοχής Ασφαλείας' που βρίσκεται ο οδηγός σε πραγματικό χρόνο με σκοπό την πρόκληση παρεμβάσεων.

2. Η αναγνώριση της σχέσης μεταξύ του κινδύνου και των άμεσα σχετιζόμενων με αυτόν παραγόντων. Επιδιώκεται η καλύτερη κατανόηση των παραγόντων της οδηγικής συμπεριφοράς και κατ' επέκταση την βελτίωση των παρεμβάσεων.

4.2 Πείραμα προσομοιωτή οδήγησης

4.2.1 Στόχος πειράματος

Στο πλαίσιο του ερευνητικού έργου i-DREAMS, 36 οδηγοί συμμετείχαν σε πείραμα προσομοιωτή οδήγησης το οποίο πραγματοποιήθηκε από 7/12/2020 έως 17/01/2021. Στόχος του πειράματος ήταν η συλλογή δεδομένων σχετιζόμενων με την οδηγική συμπεριφορά και το οδικό περιβάλλον προκειμένου να ακολουθήσει η ανάλυση τους για την επίτευξη των στόχων που έχουν τεθεί.

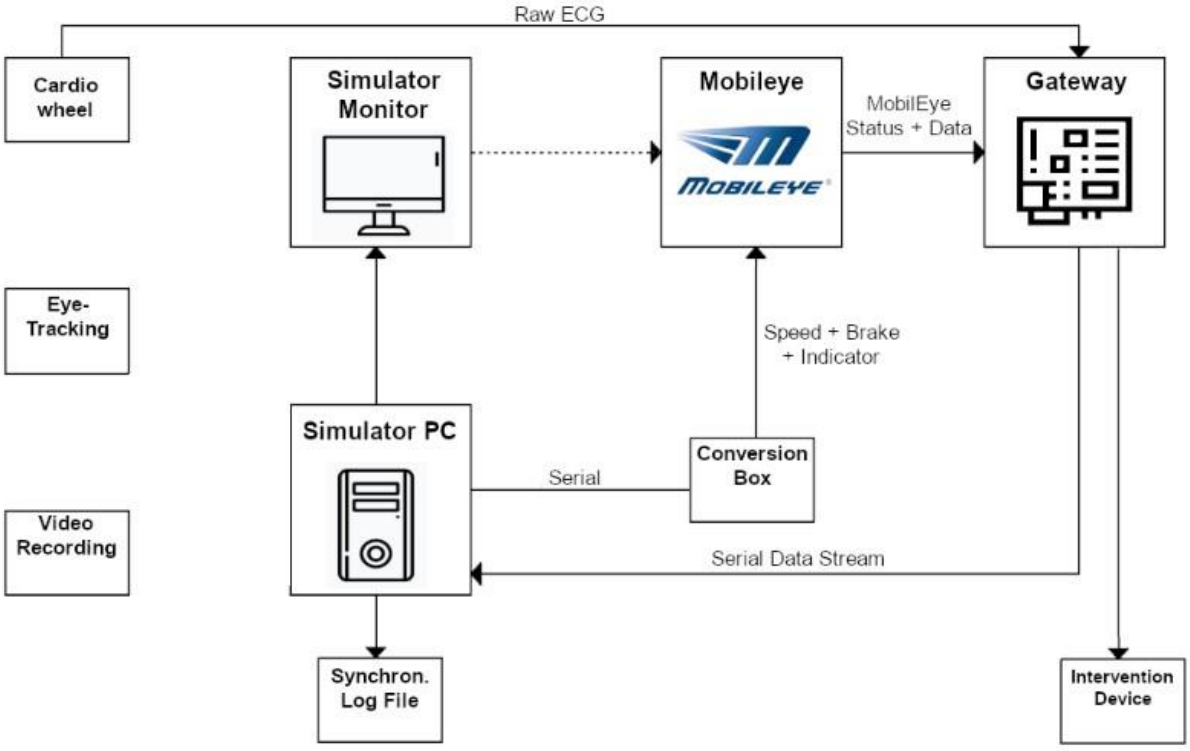
4.2.2 Προσομοιωτής οδήγησης

Ο προσομοιωτής οδήγησης, όπως φαίνεται στην εικόνα 4.1, σχεδιάστηκε και κατασκευάστηκε στο πλαίσιο του ερευνητικού έργου i-DREAMS. Ο προσομοιωτής βασίζεται στο μοντέλο Peugeot 206 από το οποίο χρησιμοποιούνται αρκετά αυθεντικά μέρη όπως το πλήρες ταμπλό, ο λειτουργικός πίνακας οργάνων και το κάθισμα οδήγησης, προκειμένου να αναπαραχθεί το πιλοτήριο του συγκεκριμένου οχήματος. Ο προσομοιωτής βασίζεται στο λογισμικό STISIM Drive 3 το οποίο αναπαρίσταται σε τρείς οθόνες 49 ιντσών με 4K ανάλυση, παρέχοντας με αυτό τον τρόπο ένα πεδίο ορατότητας 135º.

Εικόνα 4.1: Προσομοιωτής οδήγησης

4.2.3 Αρχιτεκτονική προσομοιωτή οδήγησης

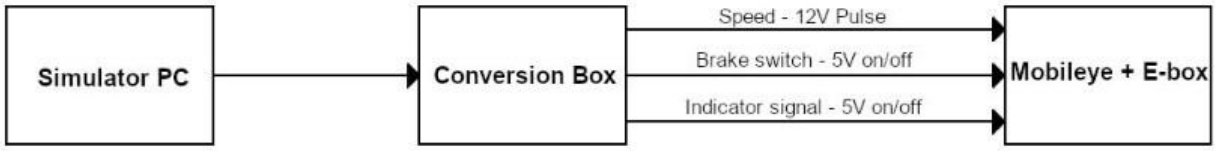
Η γενική περιγραφή της αρχιτεκτονικής του προσομοιωτή οδήγησης και ο τρόπος που εκείνος αλληλοεπιδρά με τον εξοπλισμό i-DREAMS φαίνεται στην εικόνα 4.2 Η κάμερα Mobileye, το ειδικό τιμόνι καρδιογραφήματος και το λογισμικό του προσομοιωτή χρησιμοποιήθηκαν ως αισθητήρες καταγραφής των δεδομένων σε πραγματικό χρόνο. Επιπρόσθετα, μπορεί να χρησιμοποιηθεί εξωτερικός εξοπλισμός όπως παρακολούθηση οφθαλμών και εγγραφή βίντεο ώστε να έχουμε περισσότερη πληροφόρηση για την οδηγική συμπεριφορά. Όπως στα πραγματικά οχήματα, η πύλη (gateway) του i-DREAMS είναι υπεύθυνη για την πρόκληση παρεμβάσεων σε πραγματικό χρόνο. Για τον προσομοιωτή οδήγησης τα δεδομένα δεν συλλέγονται από την πύλη ούτε αποθηκεύονται στο cloud. Αντ' αυτού η πύλη στέλνει όλα τα δεδομένα που συλλέγει και υπολογίζει πίσω στον προσομοιωτή οδήγησης μέσω μίας σειριακής διεπαφής. Τα δεδομένα αυτά συγχρονίζονται, συνδυάζονται με μεταβλητές προσομοίωσης και αποθηκεύονται τοπικά στον υπολογιστή του προσομοιωτή. Τα σειριακά δεδομένα έχουν κατεύθυνση από την πύλη προς τον προσομοιωτή οδήγησης που σημαίνει ότι δεν υπάρχει άμεση εισαγωγή των μεταβλητών προσομοίωσης στην πύλη. Επιλογή αυτής της διάταξης έγινε έτσι, ώστε τα δεδομένα που συλλέγονται από τους αισθητήρες καταγραφής στον προσομοιωτή να είναι κατά το δυνατόν παραπλήσια με του πραγματικού οχήματος.



Εικόνα 4.2: Περιγραφή αρχιτεκτονικής προσομοιωτή

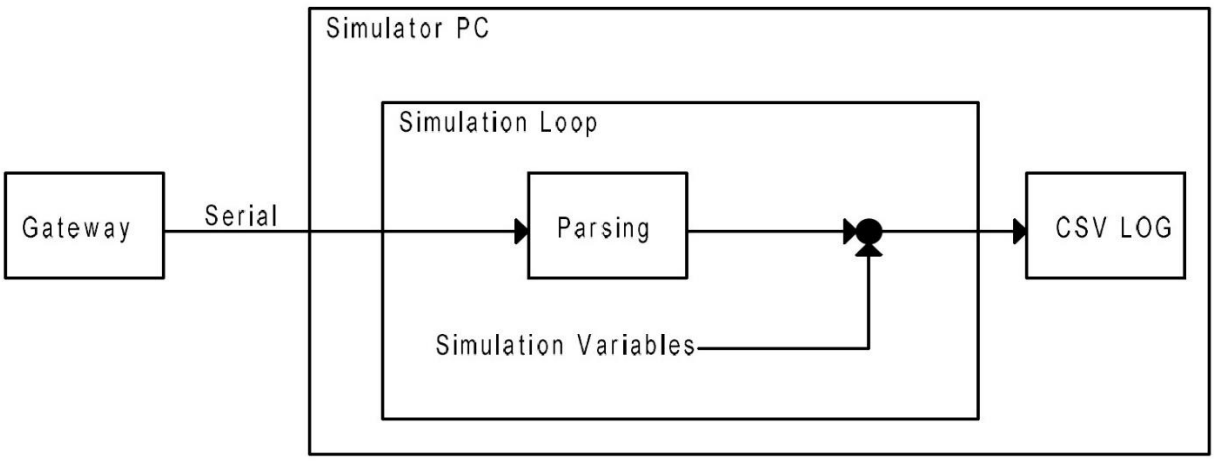
Σαν τα πραγματικά οχήματα, δεδομένα όπως η ταχύτητα, η θέση πέδησης και ο δείκτης χρήσης συλλέγονται από το Mobileye και είναι απαραίτητες προκειμένου να λειτουργεί σωστά. Το Mobileye χρησιμοποιεί τις τιμές αυτών των δεδομένων για τον υπολογισμό των δικών του παρεμβάσεων αλλά τις διαθέτει επίσης στην πύλη μέσω ειδικών μηνυμάτων. Αυτό προϋποθέτει οι μεταβλητές να μετατρέπονται σε συγκεκριμένο σήμα που είναι αποδεκτό από το Mobileye. Η μετατροπή πραγματοποιείται από έναν εξωτερικό ελεγκτή ο οποίος λαμβάνει τις μεταβλητές προσομοίωσης μέσω μίας σειριακής διεπαφής και τις

τροποποιεί σε φυσικά σήματα για την ταχύτητα, τον διακόπτη πέδησης και τον δείκτη αλλαγής πορείας (φλας). Η σχηματική αναπαράσταση φαίνεται στη εικόνα 4.3. Το σήμα ταχύτητας είναι η αναπαράσταση του σήματος VSS (Vehicle Speed Sensor) του οχήματος, το οποίο συνήθως παράγεται από έναν αισθητήρα Hall ή αντίστοιχου τύπου που μετατρέπει την περιστροφή σε παλμικό σήμα. Ο αισθητήρας μπορεί να εντοπιστεί στον εξερχόμενο άξονα του κιβώτιου ταχυτήτων ή να αποτελεί τμήμα του ABS (Anti-lock Braking System) για την μέτρηση της ταχύτητας περιστροφής του κάθε τροχού. Το σήμα είναι τετραγωνικό παλμικό σήμα και διαμορφωμένο σε συχνότητα 12V. Το σήμα πέδησης και το σήμα δείκτη αλλαγής κατεύθυνσης είναι ψηφιακά σήματα ενεργοποίησης/απενεργοποίησης (on/off). Το Mobileye δέχεται σήματα με μεγάλο εύρος τάσης από 5V έως 12V. Επίσης επειδή αντλεί ελάχιστο ρεύμα βολεύει και μπορεί να χρησιμοποιεί μία άμεση ψηφιακή έξοδο 5V από τον ίδιο ελεγκτή που χειρίζεται την μετατροπή του σήματος ταχύτητας.



Εικόνα 4.3: Μετατροπή σήματος από τον προσομοιωτή στο Mobileye

Τα δεδομένα του προσομοιωτή οδήγησης θα αποθηκευτούν τοπικά. Για να είναι χρήσιμα για ανάλυση είναι αναγκαίο τα εξωτερικά δεδομένα (από την πύλη) να συγχρονίζονται ταυτόχρονα με τα δεδομένα προσομοίωσης. Για αυτό τον σκοπό ο βρόγχος προσομοίωσης τροποποιήθηκε προκειμένου τα δεδομένα που συλλέγονται από την πύλη να συνδυάζονται με τα δεδομένα της προσομοίωσης σε κάθε χρονικό βήμα. Το αποτέλεσμα της παραπάνω διαδικασίας είναι συγχρονισμένα δεδομένα τα οποία καταγράφονται σε ένα αρχείο καταγραφής για κάθε βήμα με βάση ένα ειδικά διαμορφωμένο πρωτόκολλό (αποστέλλονται σε μορφή JSON). Στην εικόνα 4.4 αναπαρίσταται η διαδικασία.



Εικόνα 4.4: Διαδικασία συγχρονισμού εξωτερικών δεδομένων και δεδομένων προσομοίωσης

4.2.4 Σενάρια οδήγησης πειράματος

Εφαρμόστηκαν τρία διαφορετικά σενάρια οδήγησης σε οδό διπλής κατεύθυνσης. Σε κάθε ένα σενάριο, η οδός χωρίζεται σε τρία τμήματα με διαφορετικά χαρακτηριστικά. Παρακάτω παρουσιάζονται τα χαρακτηριστικά των οδών σε κάθε τμήμα για κάθε σενάριο.

Πίνακας 4.1: Διαφορετικά σενάρια που εφαρμόστηκαν κατά το πείραμα του προσομοιωτή οδήγησης.

Σενάριο	Οδικό Τμήμα	Αριθμός Λωρίδων	Όρια Ταχύτητας
A	0-6300 m	1x1	70 km/h
	6300-11300 m	2x2	90 km/h
	11300-16500 m	2x2	120 km/h
B	0-6100 m	2x2	90 km/h
	6100-12000 m	2x2	120 km/h
	12000-18200 m	1x1	70 km/h
C	0-6000 m	2x2	90 km/h
	6000-11000 m	2x2	120 km/h
	11000-17200 m	1x1	70 km/h

Κάθε οδηγός πραγματοποίησε τρείς ξεχωριστές διαδρομές:

- Διαδρομή 1: Χωρίς την πραγματοποίηση παρεμβάσεων
- Διαδρομή 2: Με την πραγματοποίηση παρεμβάσεων
- Διαδρομή 3: Με την πραγματοποίηση παρεμβάσεων σε μεταβαλλόμενες συνθήκες

4.2.5 Στοιχεία που συλλέχθηκαν από το πείραμα

Στον πίνακα 4.2 παρατίθεται η περιγραφή των δεδομένων που συλλέχθηκαν.

Πίνακας 4.2: Επεξήγηση μεταβλητών που συλλέχθηκαν από τον προσομοιωτή οδήγησης

Μεταβλητή	Περιγραφή	Μονάδες μέτρησης	Τύπος
TTC	Χρόνος πρόσκρουσης	δευτερόλεπτα	αριθμητική
Headway	Χρονική απόσταση από το προπορευόμενο όχημα	δευτερόλεπτα	αριθμητική
Speed	Ταχύτητα οχήματος	χιλιόμετρα ανά ώρα	αριθμητική
Distance_travelled	Απόσταση που διανύθηκε	μέτρα	αριθμητική
BSAV_SpeedLimitMS	Τρέχον όριο ταχύτητας	μέτρα ανά δευτερόλεπτα	αριθμητική

BSAV_SpeedLimitKPH	Τρέχον όριο ταχύτητας	χιλιόμετρα ανά ώρα	αριθμητική
HandsOnEvent	Ένδειξη ότι τα χέρια του οδηγού βρίσκονται στο τιμόνι	δύο / κανένα	διακριτή
FatigueEvent	KSS score	32 – 35 – 39	διακριτή

**Το KSS score αποτελεί μέτρο κούρασης των οδηγών και η τιμή έχει τρείς κατανομές KSS < 6 (1ο επίπεδο Ζώνης Ανοχής Ασφαλείας), 6<KSS<8 (2ο επίπεδο Ζώνης Ανοχής Ασφαλείας), KSS>8 (3ο επίπεδο Ζώνης Ανοχής Ασφαλείας)*

4.3 Επεξεργασία στοιχείων

Δεδομένου ότι κάθε οδηγός εκτέλεσε τρείς διαφορετικές διαδρομές (χωρίς παρεμβάσεις, με παρεμβάσεις, με παρεμβάσεις σε μεταβαλλόμενες συνθήκες) δημιουργήθηκαν τρία .csv αρχεία για κάθε οδηγό. Οι σχετικές πληροφορίες σχετικά με τον κωδικό του οδηγού, τον αριθμό της διαδρομής και το γράμμα του σεναρίου αναφέρονταν στα ονόματα των αρχείων.

Όλα τα αρχεία καταγραφής του προσομοιωτή τοποθετήθηκαν σε μία κοινή βάση δεδομένων. Αξιοποιώντας τα ονόματα των αρχείων, δημιουργήθηκαν τέσσερεις νέες στήλες με τον κωδικό του οδηγού, το γράμμα του σεναρίου, το νούμερο της διαδρομής και την ημερομηνία καταγραφής.

Προκειμένου να απλοποιηθεί η διαδικασία τα δεδομένα μορφοποιήθηκαν σε διαστήματα των 30 δευτερολέπτων. Συγκεκριμένα, για κάθε 30 δευτερόλεπτα υπολογίστηκαν τα περιγραφικά στατιστικά κάθε μεταβλητής όπως η μέση τιμή, η τυπική απόκλιση, η ελάχιστη τιμή, η μέγιστη τιμή και η διάμεσος. Στον πίνακα 4.3 παρατίθενται οι συγκεντρωμένες μεταβλητές των 30 δευτερολέπτων που προέκυψαν από το παραπάνω βήμα.

Πίνακας 4.3: Περιγραφή μεταβλητών μετά την επεξεργασία που αφορούν σε διαστήματα των 30 δλ.

Μεταβλητή	Περιγραφή
TTC_mean	Μέση τιμή της μεταβλητής TTC για το διάστημα των 30 δλ. (δλ.)
TTC_std	Τυπική απόκλιση της μεταβλητής TTC για το διάστημα των 30 δλ. (δλ.)
TTC_min	Ελάχιστη τιμή της μεταβλητής TTC για το διάστημα των 30 δλ. (δλ.)
TTC_max	Μέγιστη τιμή της μεταβλητής TTC για το διάστημα των 30 δλ. (δλ.)
Headway_mean	Μέση τιμή της μεταβλητής Headway για το διάστημα των 30 δλ. (δλ.)
Headway_std	Τυπική απόκλιση της μεταβλητής Headway για το διάστημα των 30 δλ. (δλ.)
Headway_median	Διάμεσος της μεταβλητής Headway για το διάστημα των 30 δλ. (δλ.)
Headway_min	Ελάχιστη τιμή της μεταβλητής Headway για το διάστημα των 30 δλ. (δλ.)

Headway_max	Μέγιστη τιμή της μεταβλητής Headway για το διάστημα των 30 δλ. (δλ.)
Speed_mean	Μέση τιμή της μεταβλητής Speed για το διάστημα των 30 δλ. (χλμ./ώρα)
Speed_std	Τυπική απόκλιση της μεταβλητής Speed για το διάστημα των 30 δλ. (χλμ./ώρα)
Speed_min	Ελάχιστη τιμή της μεταβλητής Speed για το διάστημα των 30 δλ. (χλμ./ώρα)
Speed_max	Μέγιστη τιμή της μεταβλητής Speed για το διάστημα των 30 δλ. (χλμ./ώρα)
Distance travelled_sum	Άθροισμα της μεταβλητής Distance travelled για το διάστημα των 30 δλ. (μ.)
BSAV_SpeedLimitMS_max	Μέγιστη τιμή της μεταβλητής BSAV_SpeedLimitMS για το διάστημα των 30 δλ. (μ./δλ.)
BSAV_SpeedLimitKPH_max	Μέγιστη τιμή της μεταβλητής BSAV_SpeedLimitKPH για το διάστημα των 30 δλ. (χλμ./ώρα)
HandsOnEvent_mean	Μέση τιμή της μεταβλητής HandsOnEvent για το διάστημα των 30 δλ.
HandsOnEvent_median	Διάμεσος της μεταβλητής HandsOnEvent για το διάστημα των 30 δλ.
FatigueEvent_median	Διάμεσος της μεταβλητής FatigueEvent για το διάστημα των 30 δλ.

4.4 Περιγραφική στατιστική δεδομένων

Αξιοποιώντας την βιβλιοθήκη ανάλυσης δεδομένων pandas στο προγραμματιστικό περιβάλλον python πραγματοποιήθηκε **περιγραφική στατιστική** των δεδομένων μετά την επεξεργασία τους. Στον πίνακα 4.4 παρατίθενται ορισμένα περιγραφικά στατιστικά στοιχεία των μεταβλητών που συλλέχθηκαν όπως η μέση τιμή, η τυπική απόκλιση, η ελάχιστη και η μέγιστη τιμή.

Πίνακας 4.4: Περιγραφική στατιστική αριθμητικών δεδομένων από τον προσομοιωτή οδήγησης

Μεταβλητή	Μέση τιμή	Τυπική απόκλιση	Ελάχιστη τιμή	Μέγιστη τιμή
TTC_mean	284.565	215.750	0.181	3868.964
TTC_std	525.368	3765.060	0.008	104788.018
TTC_min	3180.215	4535.467	0.017	11993.920
TTC_max	376013.640	3291508.167	24.269	93622860.000

Headway_mean	21.546	127.496	0.047	1880.767
Headway_std	43.952	237.035	0.000	1757.587
Headway_median	7.164	103.893	0.000	2704.396
Headway_min	37.284	99.836	0.000	4320.001
Headway_max	14693.979	81834.343	1.776	973035.900
Speed_mean	67.758	0.678	57.948	100.000
Speed_std	0.313	0.003	0.181	1.016
Speed_min	58.917	0.000	50.000	100.000
Speed_max	75.447	3.000	64.000	100.000
Distance travelled_sum	7006041.279	4176949.802	363.502	20023055.374
BSAV_SpeedLimitMS_max	26.648	5.821	20.968	34.857
BSAV_SpeedLimitKPH_max	95.943	20.956	75.495	125.500
HandsOnEvent_mean	0.024	0.016	0.000	0.050
HandsOnEvent_median	0.024	0.017	0.000	0.050
FatigueEvent_median	0.045	0.025	0.000	0.150

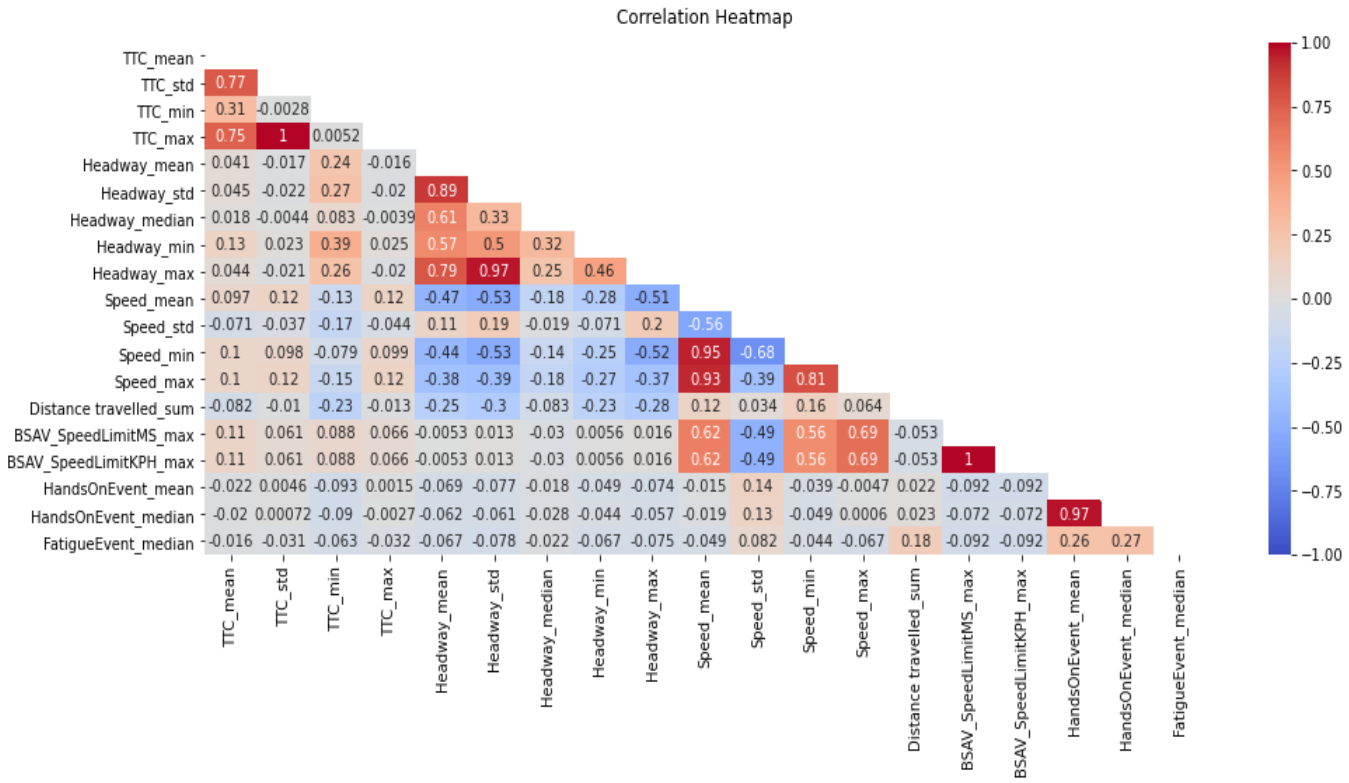
4.5 Συσχέτιση μεταβλητών

Για την ανάπτυξη των μοντέλων ταξινόμησης και παλινδρόμησης είναι απαραίτητο να διερευνηθεί η συσχέτιση μεταξύ των ανεξάρτητων μεταβλητών. Οι τιμές του συντελεστή συσχέτισης Pearson κυμαίνονται στο διάστημα [-1,1] και η σχέση των ανεξάρτητων μεταβλητών χαρακτηρίζεται ως εξής:

- Ελάχιστη συσχέτιση για $0.00 \leq |r| \leq 0.30$
- Μέτρια συσχέτιση για $0.31 \leq |r| \leq 0.70$
- Υψηλή συσχέτιση για $0.71 \leq |r| \leq 1.00$

Για τον λόγω αυτό, χρησιμοποιώντας την ίδια βιβλιοθήκη ανάλυσης στο προγραμματιστικό περιβάλλον python, αναπτύχθηκαν κατάλληλες τεχνικές υπολογισμού και απεικόνισης της συσχέτισης των μεταβλητών.

Στον παρακάτω τριγωνικό χάρτη θερμότητας παρουσιάζεται η συσχέτιση μεταξύ των διαφορετικών μεταβλητών. Η θετική συσχέτιση συμβολίζεται με θερμό χρώμα ενώ η αρνητική με ψυχρό.



Γράφημα 4.1: Τριγωνικός χάρτης συσχέτισης μεταβλητών

Από το γράφημα 4.1 προκύπτουν τα εξής συμπεράσματα:

- Μεταξύ των διαφορετικών περιγραφικών στατιστικών της ίδιας μεταβλητής παρατηρείται υψηλή συσχέτιση. Η παραπάνω υψηλή συσχέτιση είναι λογική δεδομένου ότι αφορά στη σχέση μεταξύ διαφορετικών εκφάνσεων του ίδιου στοιχείου.

- Μεταξύ των μεταβλητών Headway και Speed παρατηρείται σχετικά υψηλή αρνητική συσχέτιση. Η αύξηση της ταχύτητας οδηγεί στην μείωση του χρόνου επακολουθίας δύο οχημάτων.

- Η μεταβλητή της ταχύτητας (Speed) και των ορίων ταχύτητας (BSAV_SpeedLimit) παρουσιάζει σημαντική συσχέτιση. Η αύξηση του ορίου ταχύτητας προκαλεί αύξηση της ταχύτητας που αναπτύσσει ο οδηγός.

4.6 Σύνοψη

Συνοψίζοντας η συλλογή των δεδομένων πραγματοποιήθηκε μέσω του πειράματος προσομοιωτή οδήγησης και θα αντληθούν τα σχετιζόμενα με την οδηγική συμπεριφορά χαρακτηριστικά. Στην συνέχεια ακολούθησε η κατάλληλη επεξεργασία των δεδομένων και ο υπολογισμός των περιγραφικών στατιστικών στοιχείων για την καλύτερη κατανόηση τους. Τέλος εξετάστηκε η συσχέτιση μεταξύ των μεταβλητών το οποίο αποτελεί απαραίτητο βήμα για τις προκαταρτικές διαδικασίες των αναλύσεων που θα ακολουθήσουν.

5. ΕΦΑΡΜΟΓΗ ΜΕΘΟΔΟΛΟΓΙΑΣ - ΑΠΟΤΕΛΕΣΜΑΤΑ

5.1 Εισαγωγή

Το παρόν κεφάλαιο περιλαμβάνει την αναλυτική παρουσίαση της μεθοδολογίας που εφαρμόστηκε καθώς και τα αποτελέσματα που προέκυψαν στο πλαίσιο της μελέτης. Ο στόχος και η κατάλληλη μεθοδολογία για την επίτευξη του καθορίστηκε με βάση την βιβλιογραφική ανασκόπηση.

Για την διερεύνηση της επιρροής των διαφορετικών παραγόντων της οδηγικής συμπεριφοράς, σύμφωνα με τις μεθοδολογίες προγενέστερων ερευνών θα αναπτυχθούν κατάλληλοι **αλγόριθμοι μηχανικής εκμάθησης για την ταξινόμηση και την παλινδρόμηση.** Ειδικότερα θα αξιολογηθεί η σημαντικότητα των μεταβλητών στην ταξινόμηση καθώς και η ερμηνευτική τους ικανότητα στην παλινδρόμηση. Η ανάλυση θα χωριστεί σε δύο μέρη ώστε να εξεταστεί η επικίνδυνη οδήγηση βάσει δύο προσεγγίσεων.

Στο πρώτο μέρος των αναλύσεων θα αναπτυχθούν τα μοντέλα ταξινόμησης, με σκοπό την αναγνώριση του επιπέδου της ΄Ζώνης Ανοχής Ασφαλείας’ που βρίσκεται ο οδηγός για κάθε χρονικό πλαίσιο 30 δευτερολέπτων. Τα δεδομένα που συλλέχθηκαν από τον προσομοιωτή οδήγησης αποτελούν τις ενδογενείς μεταβλητές ενώ το επίπεδο της ‘Ζώνης Ανοχής Ασφαλείας΄ αποτελεί την εξωγενή μεταβλητή.

Στο δεύτερο μέρος των αναλύσεων θα αναπτυχθούν ορισμένα μοντέλα παλινδρόμησης με σκοπό την πρόβλεψη της διάρκειας οδήγησης σε κάθε ένα από τα τρία επίπεδα ασφαλείας για κάθε οδηγό. Η ανάλυση θα πραγματοποιηθεί υπολογίζοντας την διάρκεια που κάθε οδηγός βρίσκεται σε κάθε ένα από τα τρία επίπεδα της ‘Ζώνης Ανοχής Ασφαλείας’ και ομαδοποιώντας τα οδηγικά δεδομένα του κάθε οδηγού με βάση το επίπεδο ασφαλείας. Τα ομαδοποιημένα δεδομένα αποτελούν τις ανεξάρτητες μεταβλητές του μοντέλου, ενώ η διάρκεια οδήγησης στο κάθε επίπεδο την εξαρτημένη μεταβλητή.

Η αξιολόγηση της προγνωστικής ικανότητας των μοντέλων θα πραγματοποιηθεί αξιοποιώντας ορισμένες **μετρικές αξιολόγησης**.

Η ανάλυση θα πραγματοποιηθεί μέσω της προγραμματιστικής γλώσσας **Python** αξιοποιώντας τις εξής ειδικές βιβλιοθήκες και εργαλεία:

- Υπολογισμοί: NumPy
- Ανάλυση και χειρισμός δεδομένων: Pandas
- Χειρισμός ανομοιογένειας δεδομένων: Imbalanced Learn
- Γραφική απεικόνιση: Matplotlib, Seaborn
- Μηχανική εκμάθηση: Scikit-Learn

5.2 Εντοπισμός επιπέδου ‘Ζώνης Ανοχής Ασφαλείας’

Η αρχική προσέγγιση εστιάζει στον υπολογισμό της επιρροής του κάθε παράγοντα κινδύνου στην αναγνώριση της επικίνδυνης οδηγικής συμπεριφοράς του οδηγού. Η μεθοδολογία για την επίτευξη του παραπάνω στόχου περιλαμβάνει την ανάλυση των διαφορετικών παραγόντων κινδύνου με βάση την ανάπτυξη τεσσάρων αλγορίθμων ταξινόμησης. Η αξιολόγηση των κρίσιμων παραγόντων θα πραγματοποιηθεί με βάση την συνολική επίδοση των μοντέλων ταξινόμησης.

5.2.1 Καθορισμός επιπέδων ασφαλείας

Προτού αναπτυχθούν οι αλγόριθμοι ταξινόμησης και να διερευνηθεί η επιρροή των μεταβλητών στην επικίνδυνη οδήγηση, είναι απαραίτητο τα δεδομένα οδήγησης να **κατηγοριοποιηθούν** σε ένα από τα τρία επίπεδα της 'Ζώνης Ανοχής Ασφαλείας'. Με βάση την βιβλιογραφική ανασκόπηση εξετάστηκαν ορισμένες τεχνικές ομαδοποίησης οι οποίες έχουν χρησιμοποιηθεί σε παλαιότερες έρευνες. Συγκεκριμένα αναπτύχθηκαν το μοντέλο ομαδοποίησης κ-μέσων και το μοντέλο ιεραρχικής ομαδοποίησης χωρίς ωστόσο να προσφέρουν αποτελέσματα συνυφασμένα με την διεθνή βιβλιογραφία. Η κατανομή των δειγμάτων στις διαφορετικές τάξεις προέκυψε αντίθετη από την επιθυμητή καθιστώντας το επίπεδο ασφαλείας 'Avoidable Accident' ως την κυρίαρχη τάξη ενώ το επίπεδο ασφαλείας 'Normal' ως την τάξη μειοψηφίας.

Προκειμένου η ανάλυση της παρούσας μελέτης να εναρμονίζεται με την διεθνή βιβλιογραφία και τα δείγματα της επικίνδυνης οδήγησης να αποτελούν την κλάση μειοψηφίας, κρίθηκε απαραίτητο να εξεταστεί η τεχνική της κατηγοριοποίησης βάσει οριακών τιμών (threshold) συγκεκριμένων παραγόντων κινδύνου. Τα όρια που εξετάστηκαν αφορούσαν την ταχύτητα (Speed), τον χρόνο πρόσκρουσης (TTC) και την χρονική απόσταση από το προπορευόμενο όχημα (Headway). Στον πίνακα 5.1 παρατίθενται και συγκρίνονται αποτελέσματα από τις διαφορετικές μεθόδους.

Πίνακας 5.1: Σύγκριση αποτελεσμάτων διαφορετικών μεθόδων για τον καθορισμό του επιπέδου ασφαλείας

Μέθοδος	**Επίπεδο 'Ζώνης Ανοχής Ασφαλείας'**		
	Normal	Dangerous	Avoidable Accident
Ομαδοποίηση k-means	239	1483	1599
Ιεραρχική ομαδοποίηση	368	1204	1749
Όριο της μεταβλητής TTC_mean	3150	35	136
Όριο της μεταβλητής Speed_mean	3320	1	0
Όριο της μεταβλητής Headway_min	2820	338	163

Επομένως η μέθοδος του ορίου για την μεταβλητή Headway_min προσφέρει τα περισσότερο επιθυμητά αποτελέσματα. Για κάθε επίπεδο το εύρος τιμών της μεταβλητής Headway_min είναι:

- Επίπεδο 'Normal' (class: 0) : Headway_min > 2 δλ.
- Επίπεδο 'Dangerous' (class: 1) : Headway_min > 1.4 δλ. και Headway_min < 2 δλ.
- Επίπεδο 'Avoidable Accident' (class: 2) : Headway_min < 1.4 δλ.

Σύμφωνα με παλαιότερες έρευνες, χρονική απόσταση από το προπορευόμενο όχημα ίση με 1.1 έως 1.7 δευτερόλεπτα θεωρείται ως ανεκτό περιθώριο (Ohta, 1993). Ωστόσο, όταν

η τιμή της χρονικής απόστασης από το προπορευόμενο όχημα είναι μικρότερη από τα 2 δευτερόλεπτα αυξάνεται ο κίνδυνος και η δυσκολία στην οδήγηση (Lewis-Evans et al., 2010). Επιπλέον αρκετά εκπαιδευτικά προγράμματα δηλώνουν ότι τα 2 δευτερόλεπτα είναι η ελάχιστη χρονική απόσταση από το προπορευόμενο όχημα για την διατήρηση ασφαλούς επακολουθίας και την αποφυγή ατυχημάτων (Michael et al., 2000). Με βάση τα παραπάνω συμπεράσματα προέκυψε η διαβάθμιση των επιπέδων ασφαλείας για τα διαφορετικά όρια της μεταβλητής Headway_min.

Δεδομένης της διαδικασίας καθορισμού του επιπέδου ασφαλείας με βάση την μεταβλητή Headway_min, τα διάφορα περιγραφικά στατιστικά στοιχεία του παράγοντα Headway δεν θα αποτελέσουν μεταβλητές εισόδου στα μοντέλα. Ο συνυπολογισμός τους στην διαδικασία ταξινόμησης θα οδηγούσε σε μεροληψία του μοντέλου χωρίς εκείνο να προσφέρει χρήσιμα και σημαντικά αποτελέσματα. Επίσης τα διαφορετικά στοιχεία της μεταβλητής TTC θα εξαιρεθούν από τις μεταβλητές εισόδου. Η μαθηματική έκφραση της μεταβλητής Headway μπορεί να εκφραστεί σε συνάρτηση της μεταβλητής TTC για επακολουθία οχημάτων, γεγονός που τις καθιστά ιδιαίτερα συσχετισμένες.

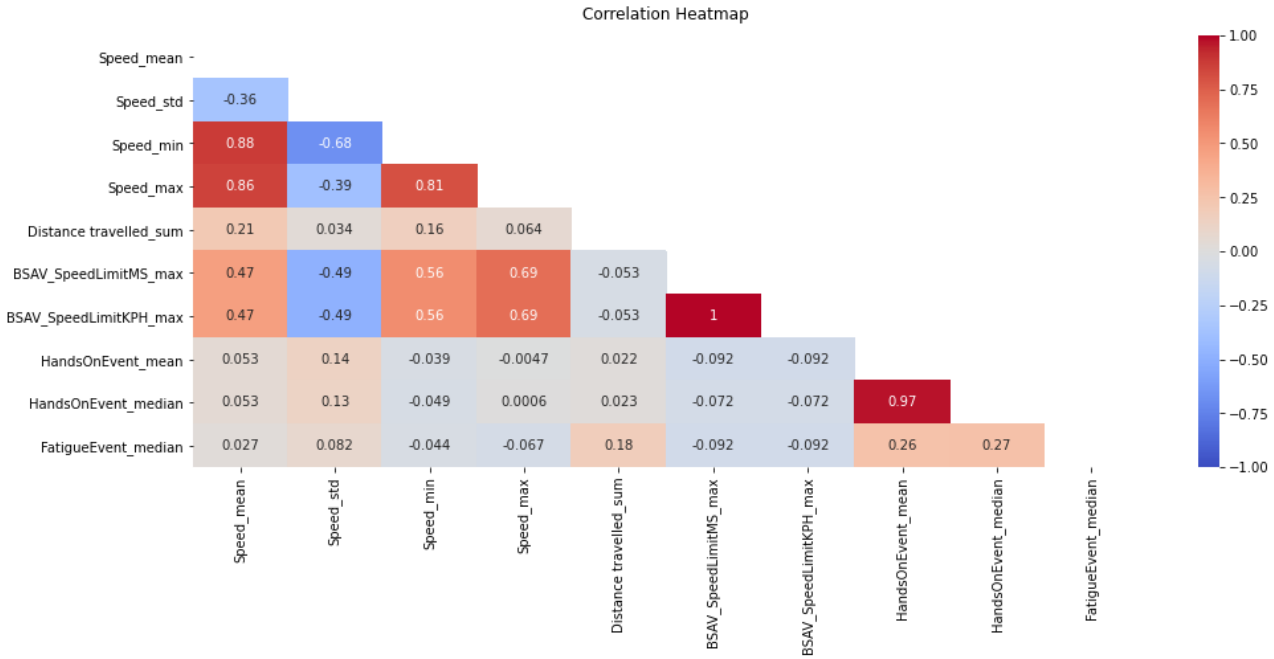
5.2.2 Επιλογή χαρακτηριστικών (Feature selection)

Η διαδικασία της **επιλογής χαρακτηριστικών** αποτελεί σημαντικό βήμα της μεθοδολογίας. Σκοπός της διαδικασίας είναι η ελαχιστοποίηση του υπολογιστικού κόστους και η βελτίωση της προγνωστικής απόδοσης του μοντέλου, μειώνοντας τον αριθμό των μεταβλητών εισόδου.

Η επιλογή των χαρακτηριστικών πραγματοποιείται με κριτήριο την **συσχέτιση** των μεταβλητών και την **επιρροή** κάθε μεταβλητής στην διαδικασία της ταξινόμησης. Η διαδικασία αυτή αποτελεί μία αρχική προσέγγιση για την μείωση των μεταβλητών εισόδου και την βελτίωση των μοντέλων.

Εξετάστηκαν διάφορα σύνολα συνδυάζοντας διαφορετικές μεταβλητές βάσει της συσχέτισης τους και της επιρροής τους στις προβλέψεις.

Στο γράφημα 5.1 απεικονίζεται η συσχέτιση των μεταβλητών που θα εξεταστούν η οποία προέκυψε με την χρήση των εργαλείων της βιβλιοθήκης Pandas.



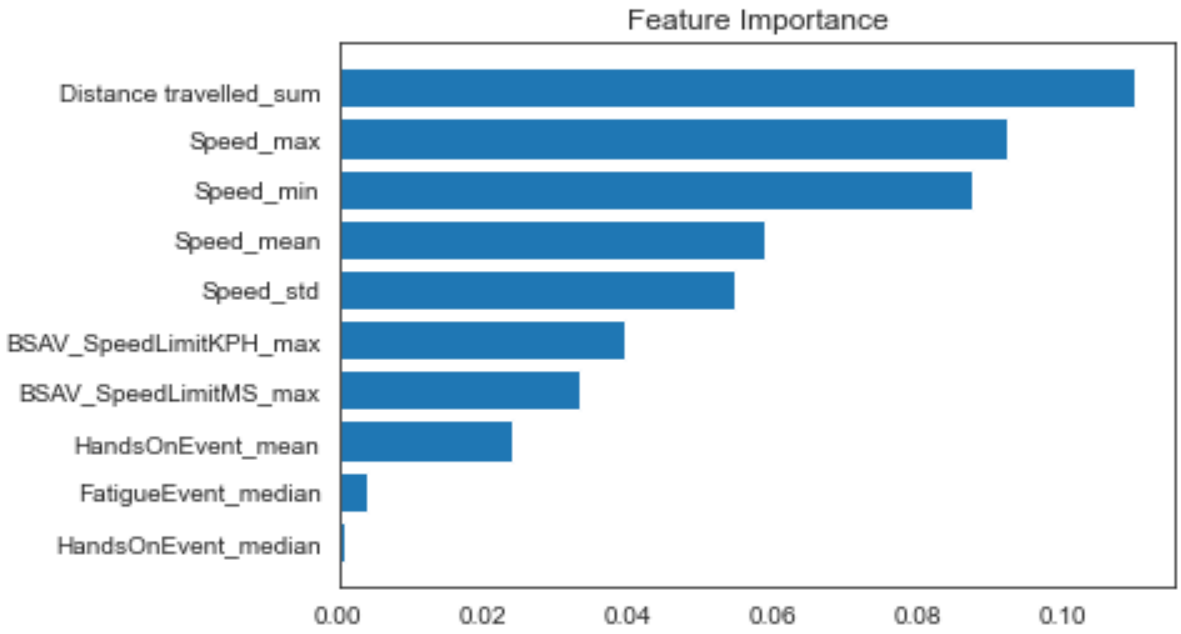
Γράφημα 5.1: Συσχέτιση μεταβλητών προς εξέταση

Από την συσχέτιση των μεταβλητών προκύπτουν οι εξής παρατηρήσεις:

- Υψηλή συσχέτιση μεταξύ περιγραφικών στοιχείων της ίδιας μεταβλητής.
- Μέτρια συσχέτιση μεταξύ στοιχείων της ταχύτητας (Speed) και των ορίων ταχύτητας (BSAV_SpeedLimit).
- Χαμηλή συσχέτιση μεταξύ HandsOnEvent, FatigueEvent, Distance travelled και όλων των υπολοίπων μεταβλητών.

Για τον εντοπισμό της σημαντικότητας των μεταβλητών στην ταξινόμηση χρησιμοποιήθηκε η τεχνική σημαντικότητας χαρακτηριστικών βάσει την μετάθεση χαρακτηριστικών (feature importance based on feature permutation). Αρχικά τα δεδομένα χωρίστηκαν σε δύο υποσύνολα. Το πρώτο υποσύνολο περιλάμβανε όλα τα δεδομένα που συλλέχθηκαν από τον προσομοιωτή οδήγησης και αποτελούσε τις μεταβλητές εισόδου της μεθόδου, ενώ το δεύτερο υποσύνολο αφορούσε την μεταβλητή εξόδου και αποτελούνταν από το επίπεδο της ‘Ζώνης Ανοχής Ασφαλείας’. Στην συνέχεια αξιοποιώντας τα ειδικά εργαλεία της βιβλιοθήκης scikit-learn, αναπτύχθηκε ο ταξινομητής ‘Τυχαίων Δασών’ (Random Forests classifier) και υπολογίστηκε η επιρροή κάθε μεταβλητής (feature importance) με βάση την μετάθεση χαρακτηριστικών (feature permutation) στην διαδικασία ταξινόμησης.

Στο γράφημα 5.2 καθώς και στον πίνακα 5.2 απεικονίζεται η επιρροή της κάθε μεταβλητής στην ταξινόμηση σε κλίμακα τιμών [0,1].



Γράφημα 5.2: Σημαντικότητα μεταβλητών σύμφωνα με την μέθοδο 'Τυχαίων Δασών'

Όπως προκύπτει από τον πίνακα 5.2, η διανυθείσα απόσταση, η ταχύτητα και τα όρια ταχύτητας έχουν την μεγαλύτερη επιρροή στην διαδικασία αναγνώρισης του επιπέδου της 'Ζώνης Ανοχής Ασφαλείας' που βρίσκεται ο οδηγός. Αντίθετα οι μεταβλητές HandsOnEvent και FatigueEvent σημειώνουν την χαμηλότερη επίδραση στην διαδικασία της ταξινόμησης.

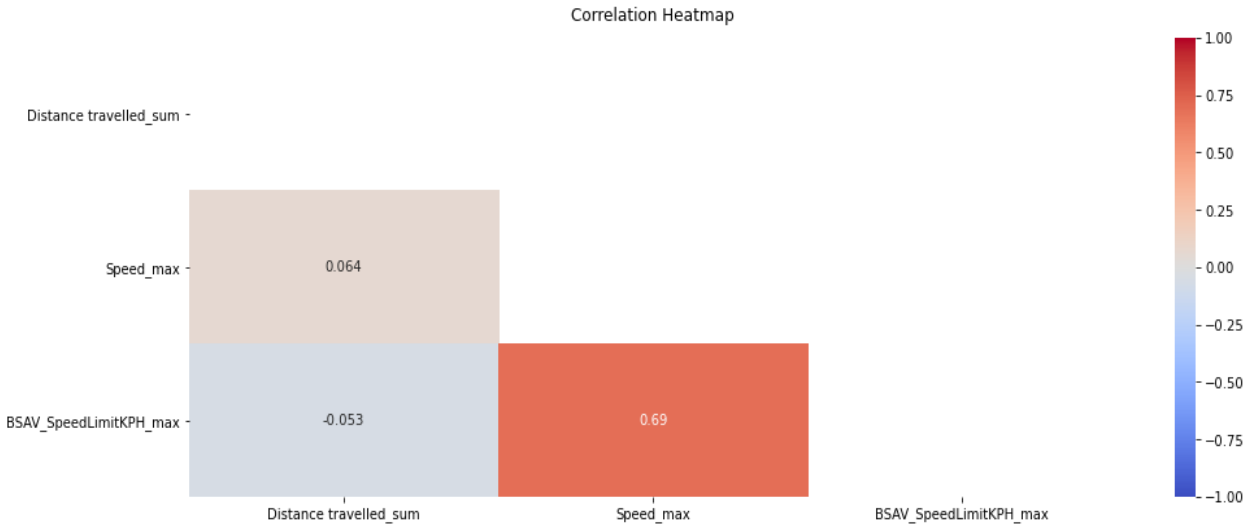
Πίνακας 5.2: Αριθμητικές τιμές της σημαντικότητας των μεταβλητών

Μεταβλητή	**Σημαντικότητα**
Distance travelled_sum	0.1102
Speed_max	0.0924
Speed_min	0.0874
Speed_mean	0.0587
Speed_std	0.0546
BSAV_SpeedLimitKPH_max	0.0397
BSAV_SpeedLimitMS_max	0.0333
HandsOnEvent_mean	0.0240
FatigueEvent_median	0.0040
HandsOnEvent_median	0.0007

Με βάση την συσχέτιση και την σημαντικότητα που προέκυψε γίνεται η επιλογή των μεταβλητών εισόδου στα μοντέλα ταξινόμησης.

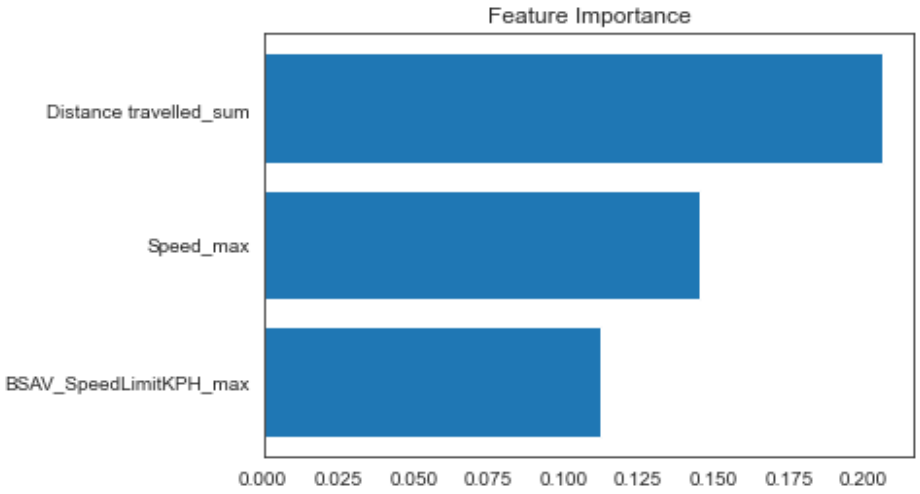
Το **τελικό σύνολο** μεταβλητών εισόδου μειώθηκε στις εξής **τρείς**:

1. Distance travelled_sum
2. Speed_max
3. BSAV_SpeedLimitKPH_max



Γράφημα 5.3: Συσχέτιση τελικών μεταβλητών εισόδου ταξινόμησης

Στον πίνακα 5.4, προκύπτει η σημαντικότητα των τελικών μεταβλητών εισόδου στην διαδικασία της ταξινόμησης.



Γράφημα 5.4: Σημαντικότητα τελικών μεταβλητών εισόδου ταξινόμησης

Όπως φαίνεται **η συνολική απόσταση που διανύθηκε** αποτελεί την σημαντικότερη μεταβλητή για την πρόβλεψη των διαφορετικών επιπέδων ασφαλείας, ενώ ακολουθεί η μέγιστη ταχύτητα και η μέγιστη τιμή του ορίου ταχύτητας.

5.2.3 Προετοιμασία δεδομένων

Όπως αναλύθηκε και στην διαδικασία επιλογής χαρακτηριστικών τα δεδομένα διαχωρίζονται: (1) στα δεδομένα εισόδου που αποτελούνται από τις 3 μεταβλητές που αναλύθηκαν προηγουμένως και (2) στις μεταβλητές εξόδου που είναι τα 3 επίπεδα της 'Ζώνης Ανοχής Ασφαλείας΄. Στην συνέχεια τα δύο σύνολα διαιρέθηκαν στα **δεδομένα εκπαίδευσης (training dataset)** και τα **δεδομένα εξέτασης (testing dataset)** με ποσοστό 90% και 10% αντίστοιχα. Σύμφωνα με την περιγραφή της λειτουργίας των μοντέλων μηχανικής εκμάθησης στο κεφάλαιο 3, τα δεδομένα εκπαίδευσης αξιοποιούνται για την εκπαίδευση του μοντέλου στην αναγνώριση του επιπέδου ασφαλείας δημιουργώντας μοτίβα αναγνώρισης βάσει συγκεκριμένων χαρακτηριστικών. Για την διαδικασία της αξιολόγησης το μοντέλο επεξεργάζεται τα εισαγόμενα δεδομένα εξέτασης και τα ταξινομεί σε ένα από τα τρία επίπεδα ώστε αυτά να συγκριθούν με τα πραγματικά επίπεδα ασφαλείας.

5.2.4 Αντιμετώπιση άνισης κατανομής δεδομένων στις κλάσεις

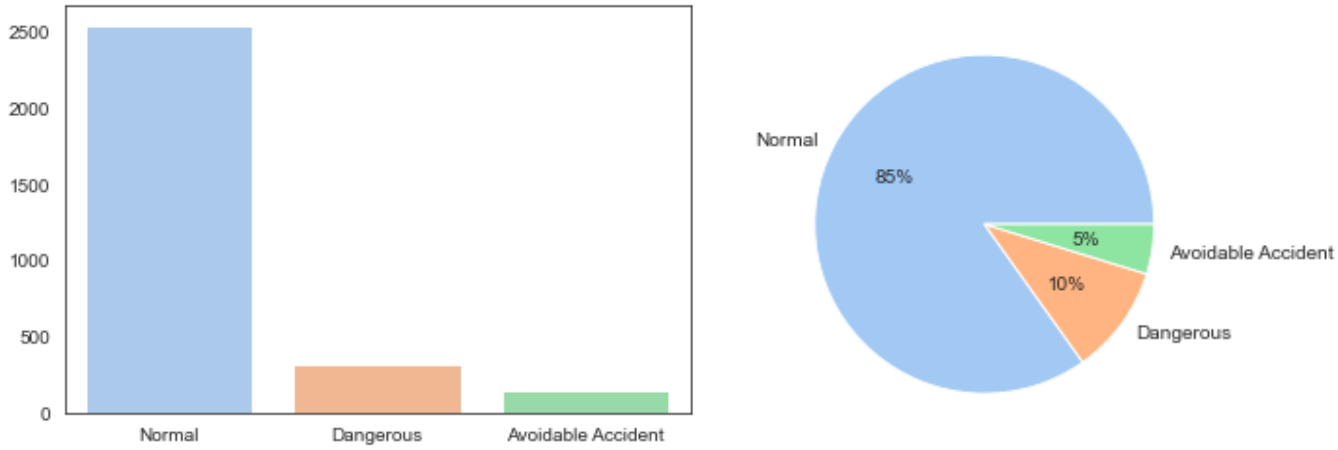
Σύμφωνα με την βιβλιογραφική ανασκόπηση στην πλειονότητα των ερευνών αντιμετωπίζεται το **πρόβλημα της ανισορροπίας** των δειγμάτων ως προς τις διαφορετικές κλάσεις, με τα δείγματα των επικίνδυνων συνθήκων να είναι σημαντικά μικρότερα σε σχέση με τα δείγματα των συνθηκών ασφαλούς οδήγησης. Επιπλέον, όπως αναφέρθηκε στο κεφάλαιο 3, τα μοντέλα ταξινόμησης θεωρούν την ίση κατανομή των δεδομένων στις κλάσεις καθιστώντας τα ιδιαίτερα ευαίσθητα σε σφάλματα ταξινόμησης για δεδομένα με άνιση κατανομή.

Μετά τον καθορισμό των επιπέδων ασφαλείας και την κατηγοριοποίηση των δεδομένων στα διαφορετικά επίπεδα προέκυψε η άνιση κατανομή που παρουσιάζεται στον πίνακα 5.3.

Πίνακας 5.3: Κατανομή δειγμάτων στα διαφορετικά επίπεδα ασφαλείας

Επίπεδο 'Ζώνης Ανοχής Ασφαλείας'	Αριθμός δειγμάτων	Ποσοστό δειγμάτων
Επίπεδο 0 (Normal)	2820	85 %
Επίπεδο 1 (Dangerous)	338	10 %
Επίπεδο 2 (Avoidable Accident)	163	5 %

Αντίστοιχή κατανομή έχουν και τα δεδομένα εκπαίδευσης που θα κληθούν να εκπαιδεύσουν τον αλγόριθμο ταξινόμησης. Η άνιση κατανομή των δεδομένων εκπαίδευσης (training dataset) με βάση τα οποία θα εκπαιδευτούν τα μοντέλα ταξινόμησης, απεικονίζεται στο γράφημα 5.5.

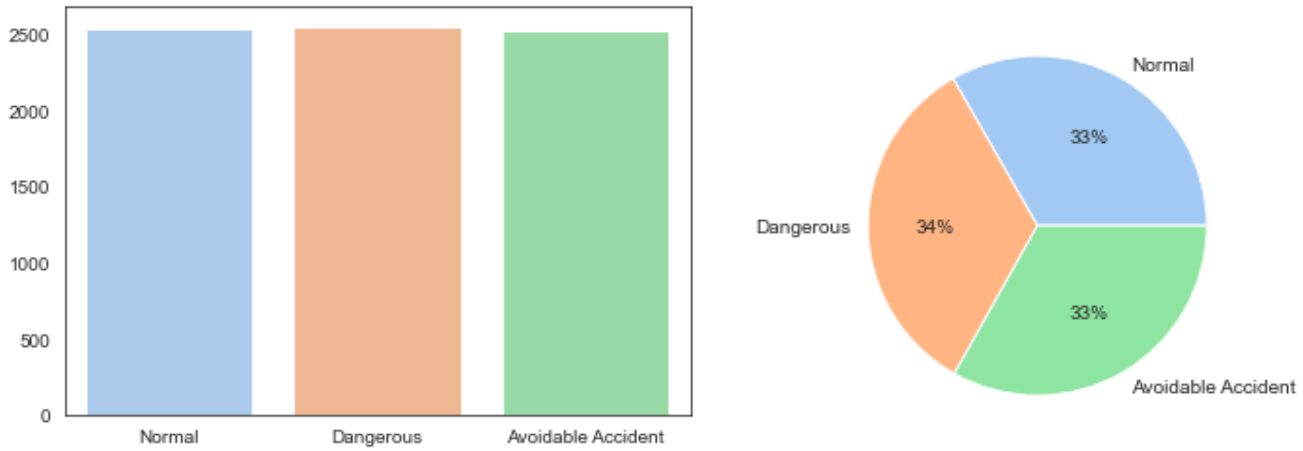


Γράφημα 5.5: Κατανομή δεδομένων εκπαίδευσης στα διαφορετικά επίπεδα ασφαλείας πριν την διαδικασία επαναδειγματοληψίας

Η εκπαίδευση των μοντέλων ταξινόμησης με τα ανομοιογενή δεδομένα εκπαίδευσης είναι πιθανό να αυξήσει τη μεροληψία των μοντέλων ως προς την κυρίαρχη τάξη ('Normal') και να οδηγήσει σε σφάλματα προβλέψεων σχετικά με τις τάξεις μειοψηφίας ('Dangerous', 'Avoidable Accident'). Οι κίνδυνοι ανάπτυξης ενός τέτοιου μοντέλου αναγνώρισης του επιπέδου ασφαλείας είναι ιδιαίτερα σημαντικοί αναφορικά με την οδική ασφάλεια.

Για την αντιμετώπιση της ανισορροπίας των δεδομένων στα επίπεδα ασφαλείας και κατ' επέκταση την εξασφάλιση της αμεροληψίας των μοντέλων, εξετάστηκαν ορισμένες **τεχνικές επαναδειγματοληψίας** σύμφωνα με την βιβλιογραφική ανασκόπηση. Ειδικότερα εξετάστηκε η τεχνική SMOTE, η SMOTE-ENN, η SMOTE-Tomek Links, η ADASYN και η τεχνική τυχαίας υπερδειγματοληψίας. Η εφαρμογή της κάθε τεχνικής επαναδειγματοληψίας πραγματοποιήθηκε παράλληλα με την ανάπτυξη των μοντέλων ταξινόμησης. Στο τέλος αξιολογήθηκε η επίδραση της κάθε μεθόδου επαναδειγματοληψίας στην πρόβλεψη της οδηγικής συμπεριφοράς.

Βέλτιστη μέθοδος κρίθηκε η **'Προσαρμοστική Συνθετική' (ADASYN)** καθώς προσέφερε της καλύτερες μετρικές αξιολόγησης για το σύνολο των ταξινομητών. Η τελική κατανομή των δεδομένων εκπαίδευσης που προέκυψε βάσει της επαναδειγματοληψίας παρουσιάζεται στο γράφημα 5.6.



Γράφημα 5.6: Κατανομή δεδομένων εκπαίδευσης στα διαφορετικά επίπεδα ασφαλείας μετά την διαδικασία επαναδειγματοληψίας

Στην συνέχεια θα αναλυθούν τα μοντέλα ταξινόμησης που αναπτύχθηκαν καθώς και οι επιδόσεις που σημείωσαν, με βάση τις μεταβλητές εισόδου της υποενότητας 5.2.2 και την τεχνική επαναδειγματοληψίας **ADASYN**.

5.2.5 Ανάπτυξη μοντέλων ταξινόμησης

Όπως περιγράφεται στις προηγούμενες ενότητες, αναπτύχθηκαν ορισμένοι **αλγόριθμοι ταξινόμησης** με σκοπό την αναγνώριση του επιπέδου 'Ζώνης Ανοχής Ασφαλείας΄ που βρίσκεται σε κάθε χρονικό πλαίσιο των 30 δευτερολέπτων ο οδηγός. Η επιλογή των τεσσάρων μοντέλων έγινε με βάση την βιβλιογραφική ανασκόπηση. Στον πίνακα 5.4 επεξηγούνται η ονοματολογία και ο συμβολισμός των μοντέλων.

Πίνακας 5.4: Ονοματολογία και συμβολισμός μοντέλων ταξινόμησης

Όνομα μοντέλου (ελληνικά)	**Όνομα μοντέλου (αγγλικά)**	**Συμβολισμός μοντέλου**
Μηχανές Διανυσμάτων Υποστήριξης	Support Vector Machines	SVM
Ταξινομητής Τυχαίων Δασών	Random Forests Classifier	RF
Ταξινομητής AdaBoost	AdaBoost Classifier	AdaBoost
Ταξινομητής Πολυεπίπεδου Perceptron	Multilayer Perceptron Classifier	MLP

Λόγω της ιδιαιτερότητας του κάθε αλγορίθμου, σε ορισμένους από αυτούς κρίθηκε απαραίτητο να εφαρμοστούν ορισμένες πρόσθετες τεχνικές προ επεξεργασίας των δεδομένων πριν πραγματοποιηθεί η εκπαίδευση και αξιολόγηση τους. Η αναγκαιότητα του μετασχηματισμού των δεδομένων εντοπίζεται στο γεγονός ότι λόγω της διαφορετικής κλίμακας των τιμών τους δεν συνεισφέρουν ίσα στην εκπαίδευση του μοντέλου με αποτέλεσμα να κινδυνεύει το μοντέλο από μεροληψία. Για παράδειγμα ο αλγόριθμος SVM επιχειρεί να μεγιστοποίηση την απόσταση μεταξύ του διαχωριστικού υπερεπιπέδου και των διανυσμάτων υποστήριξης. Εάν μια μεταβλητή έχει πολύ υψηλές τιμές θα έχει μεγαλύτερη επιρροή σε σχέση με τις άλλες. Έτσι μετά τον μετασχηματισμό τα δεδομένα θα έχουν την ίδια επιρροή στη μέτρηση της απόστασης. Συγκεκριμένα για την βελτίωση των αλγορίθμων SVM και MLP, τα δεδομένα εισόδου μετασχηματίστηκαν με την μέθοδο Min-Max scaler. Η μέθοδος αυτή αποτελεί έναν τρόπο κανονικοποίησης των μεταβλητών εισόδου μετασχηματίζοντας τα σε ένα εύρος τιμών [0,1] με την τιμή 0 να αποτελεί την ελάχιστη τιμή ενώ την τιμή 1 την μέγιστη.

Στην περίπτωση των συνδυαστικών αλγορίθμων και των δένδρων αποφάσεων δεν είναι απαραίτητος ο μετασχηματισμός των μεταβλητών καθώς η απόσταση μεταξύ των τιμών των δεδομένων δεν θεωρείται σημαντική. Επομένως επειδή ο αλγόριθμος RF και ο αλγόριθμος AdaBoost είναι μοντέλα που βασίζονται σε δένδρα αποφάσεων, τα δεδομένα εισόδου τους δεν μετασχηματίζονται.

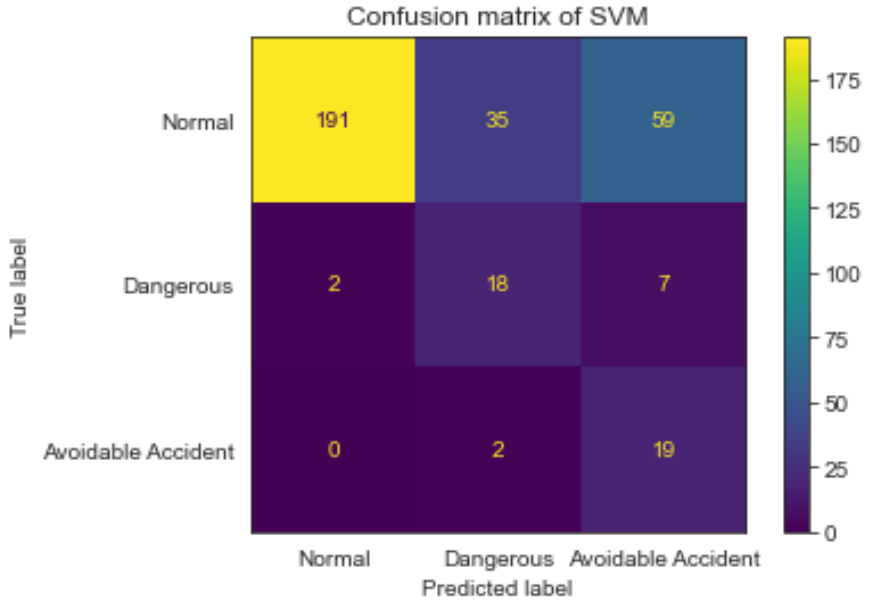
Στο σύνολο τον μοντέλων εφαρμόστηκε η τεχνική βελτιστοποίησης των παραμέτρων τους GridSearchCV μέσω της βιβλιοθήκης scikit-learn της python. Κατά αυτόν τον τρόπο βελτιώθηκε η επίδοση και των τεσσάρων αλγορίθμων.

Τα μοντέλα αναπτύχθηκαν αξιοποιώντας την βιβλιοθήκη scikit-learn της προγραμματιστικής γλώσσας python.

Παρακάτω παρατίθενται οι μήτρες σύγχυσης για την γραφική αναπαράσταση της επίδοσης κάθε αλγορίθμου. Επίσης παρουσιάζονται οι μετρικές αξιολόγησης που προέκυψαν μετά την εξέταση του κάθε μοντέλου.

1. Αλγόριθμος Μηχανών Διανυσμάτων Υποστήριξης (SVM)

Όπως φαίνεται στο γράφημα 5.7 ο αλγόριθμός SVM είχε υψηλότερα ποσοστά πρόβλεψης του επιπέδου ‘Avoidable Accident’ συγκριτικά με τα επίπεδα ‘Normal’ και ‘Dangerous’. Επομένως συνολικά θεωρείται ένα ικανοποιητικό μοντέλο που έχει υψηλή ικανότητα αναγνώρισης επικίνδυνων συμπεριφορών.



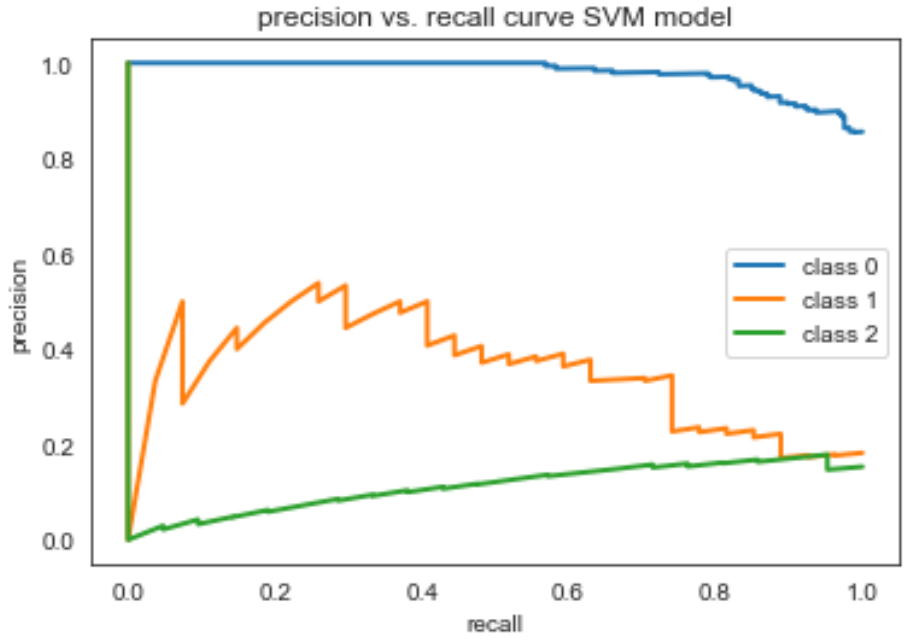
Γράφημα 5.7: Μήτρα σύγχυσης αλγόριθμου SVM

Με βάση τις μετρικές αξιολόγησης του μοντέλου της παρούσας έρευνας (πίνακας 5.5) παρατηρείται σημαντική διαφορά στο ποσοστό ορθών προβλέψεων σε σχέση με το αντίστοιχο ποσοστό του SVM στην έρευνα των (Yang et al., 2021) που είχε τιμή 95%.

Πίνακας 5.5: Σύνοψη μοντέλου SVM

Επίδοση μοντέλου ταξινόμησης SVM (ορθότητα: 68%)			
Επίπεδο 'Ζώνης Ανοχής Ασφαλείας'	Ακρίβεια	Ανάκληση	Σύνολο δεδομένων εξέτασης (333)
Normal	99%	67%	285
Dangerous	33%	67%	27
Avoidable Accident	22%	90%	21
Μέση τιμή	51%	75%	

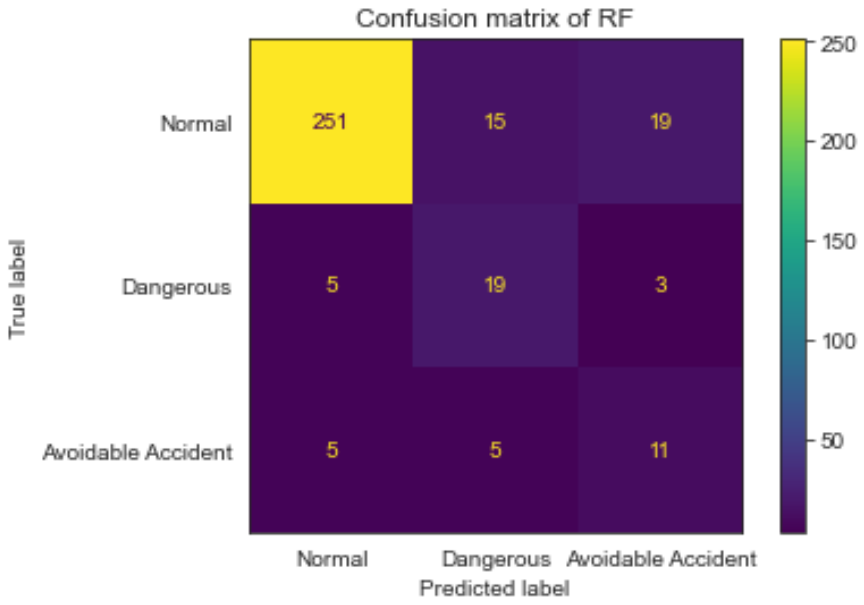
Η καμπύλη ακρίβειας – ανάκλησης μπορεί να χρησιμοποιηθεί όταν υπάρχει άνιση κατανομή των δεδομένων στις κλάσεις. Ουσιαστικά επεξηγεί την αντιστάθμιση μεταξύ προγνωστικής ισχύος (ακρίβεια) και πραγματικού θετικού ποσοστού (ανάκληση). Όπως αναπαρίσταται στο γράφημα 5.8, το επίπεδο 'Normal' (class 0) έχει υψηλό ποσοστό πραγματικών θετικών και υψηλή προγνωστική ικανότητα. Για τις άλλες δύο κλάσεις η σχέση μεταξύ των δύο μετρικών είναι σχετικά χαμηλή.



Γράφημα 5.8: Καμπύλη Ακρίβειας – Ανάκλησης του μοντέλου SVM

2. Αλγόριθμος Τυχαίων Δασών (RF)

Ο αλγόριθμος RF σημείωσε υψηλά ποσοστά αναγνώρισης και των τριών επιπέδων της ‘Ζώνης Ανοχής Ασφαλείας’. Το ποσοστό λανθασμένων προβλέψεων του επιπέδου ‘Normal’ είναι ιδιαίτερα χαμηλό με τιμή 12%. Ωστόσο σχετικά με το επίπεδο ασφαλείας ‘Avoidable Accident’ το αντίστοιχο ποσοστό έχει τιμή 48%.

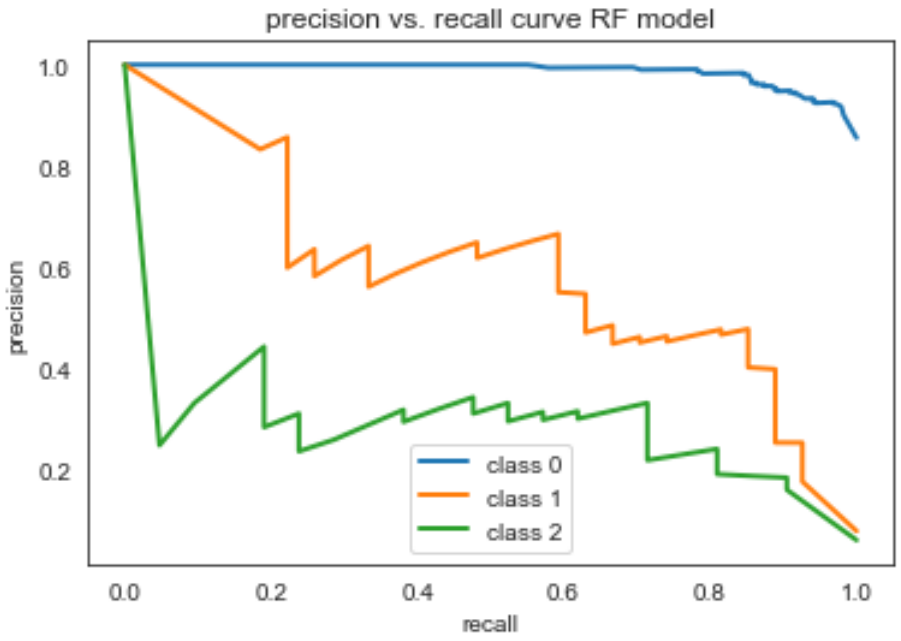


Γράφημα 5.9: Μήτρα σύγχυσης αλγόριθμου RF

Όπως αναλύθηκε και στο κεφάλαιο 2.2 ο αλγόριθμος RF είναι ιδιαίτερα διαδεδομένος στις έρευνες αναγνώρισης οδηγικής συμπεριφοράς. Συγκρίνοντας τις μετρικές αξιολόγησης του πίνακα 5.6 με τις αντίστοιχες των ερευνών που αναλύθηκαν, προκύπτει ότι η απόδοση του μοντέλου κινήθηκε σε όμοια ποσοστά. Με εξαίρεση την έρευνα των Song et al. (2021) που το ποσοστό ορθών προβλέψεων είχε τιμή 90%.

Πίνακας 5.6: Σύνοψη μοντέλου RF

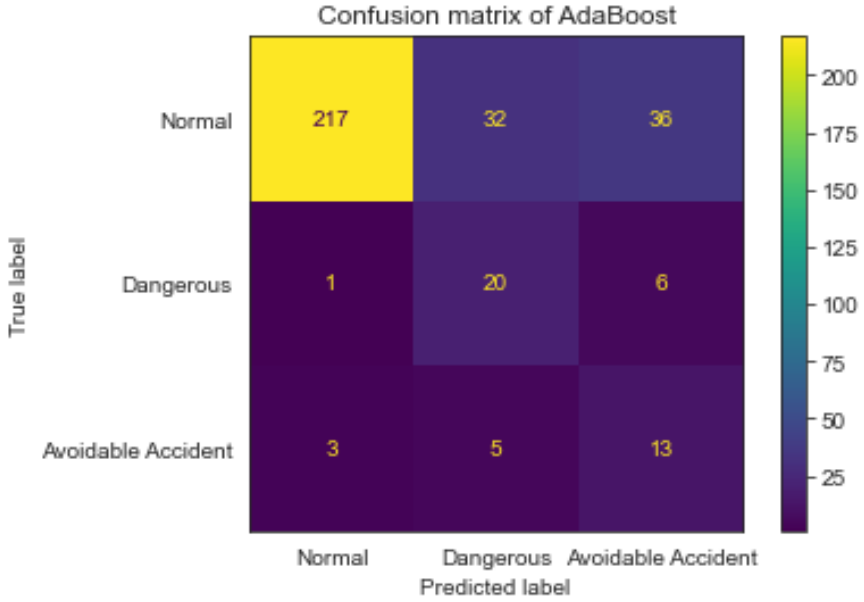
Επίδοση μοντέλου ταξινόμησης RF (ορθότητα: 84%)			
Επίπεδο ‘Ζώνης Ανοχής Ασφαλείας’	Ακρίβεια	Ανάκληση	Σύνολο δεδομένων εξέτασης (333)
Normal	96%	88%	285
Dangerous	49%	70%	27
Avoidable Accident	33%	52%	21
Μέση τιμή	59%	70%	



Γράφημα 5.10: Καμπύλη Ακρίβειας – Ανάκλησης του μοντέλου RF

3. Αλγόριθμος Προσαρμοστική Ενδυνάμωσης (AdaBoost)

Ομοίως με προηγουμένως ο αλγόριθμος AdaBoost σημείωσε ικανοποιητικά αποτελέσματα για το σύνολο των επιπέδων ασφαλείας. Τα ποσοστά λανθασμένων ταξινομήσεων για τα επίπεδα ‘Dangerous’ και ‘Avoidable Accident’ ήταν σχετικά χαμηλά με τιμές 26% και 38% αντίστοιχα.

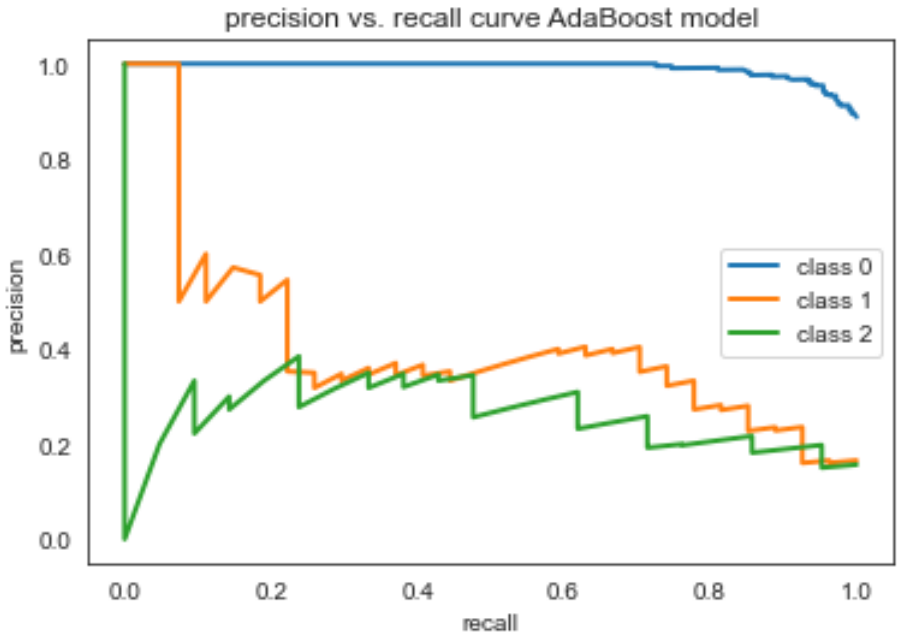


Γράφημα 5.11: Μήτρα σύγχυσης αλγόριθμου AdaBoost

Στην βιβλιογραφική ανασκόπηση δεν εντοπίστηκε η ανάπτυξη του αλγορίθμου AdaBoost. Παρόλα αυτά, όπως φαίνεται από τις μετρικές αξιολόγησης του πίνακα 5.7 το μοντέλο σημείωσε υψηλές επιδόσεις.

Πίνακας 5.7: Σύνοψη μοντέλου AdaBoost

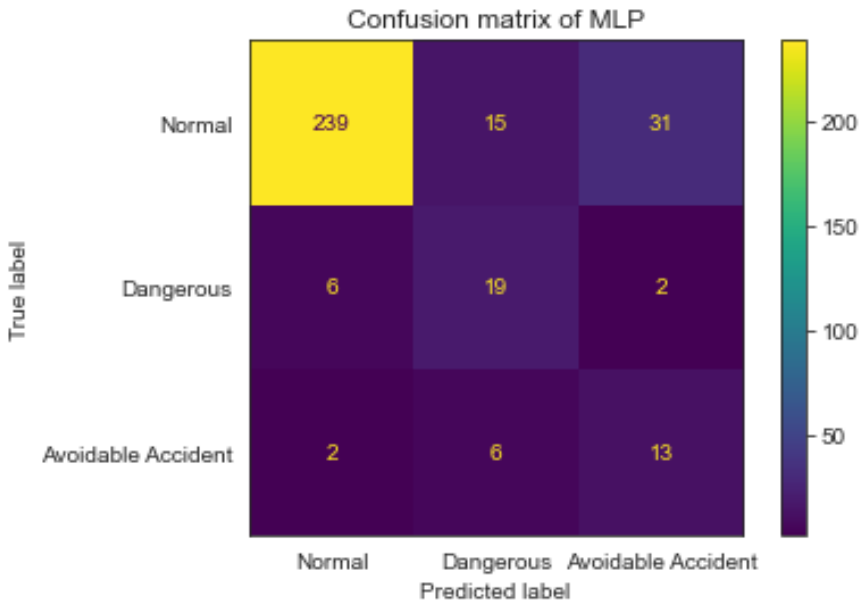
Επίδοση μοντέλου ταξινόμησης AdaBoost (ορθότητα: 75%)			
Επίπεδο 'Ζώνης Ανοχής Ασφαλείας'	Ακρίβεια	Ανάκληση	Σύνολο δεδομένων εξέτασης (333)
Normal	98%	76%	285
Dangerous	35%	74%	27
Avoidable Accident	24%	62%	21
Μέση τιμή	52%	71%	



Γράφημα 5.10: Καμπύλη Ακρίβειας – Ανάκλησης του μοντέλου AdaBoost

4. Αλγόριθμος Πολυεπίπεδου Perceptron (MLP)

Ο αλγόριθμος MLP κατάφερε να ταξινομήσει ορθά μεγάλο ποσοστό των δειγμάτων για κάθε επίπεδο της ‘Ζώνης Ανοχής Ασφαλείας’. Τα ποσοστά λανθασμένων προβλέψεων για τα δύο επίπεδα επικινδυνότητας έχουν τιμές όμοιες με εκείνες του μοντέλου AdaBoost. Επίσης για το επίπεδο ασφαλείας ‘Normal’ το ποσοστό λανθασμένων ταξινομήσεων έχει τιμή 16%

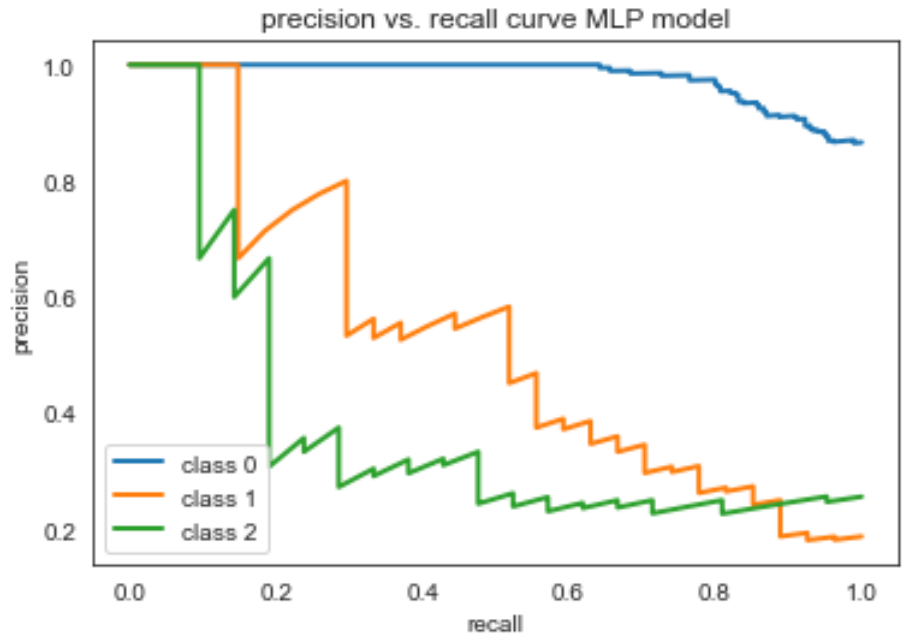


Γράφημα 5.11: Μήτρα σύγχυσης αλγόριθμου MLP

Σε σχέση με την έρευνα των Shangguan et al. (2021) που η ορθότητα του MLP ισούταν με 85%, η παρούσα μελέτη κατέγραψε μικρή διαφορά του ποσοστού ορθών προβλέψεων.

Πίνακας 5.8: Σύνοψη μοντέλου MLP

Επίδοση μοντέλου ταξινόμησης MLP (ορθότητα: 81%)			
Επίπεδο ‘Ζώνης Ανοχής Ασφαλείας’	Ακρίβεια	Ανάκληση	Σύνολο δεδομένων εξέτασης (333)
Normal	97%	84%	285
Dangerous	47%	70%	27
Avoidable Accident	28%	62%	21
Μέση τιμή	58%	72%	



Γράφημα 5.12: Καμπύλη Ακρίβειας – Ανάκλησης του μοντέλου MLP

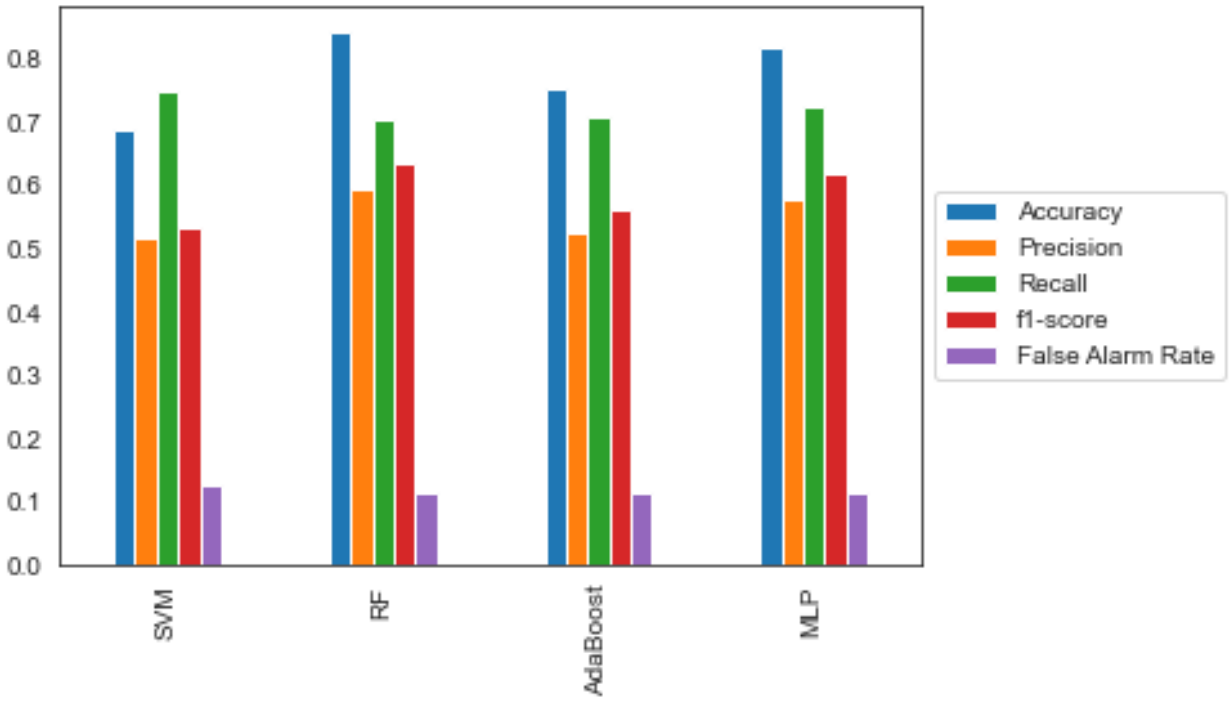
Στην υποενότητα 5.2.6 θα παρουσιαστούν συνολικά οι μετρικές αξιολόγησης σε έναν ενιαίο πίνακα με σκοπό την σύγκριση τους και την επιλογή των βέλτιστων μοντέλων.

5.2.6 Σύγκριση μετρικών αξιολόγησης των μοντέλων

Οι διαφορετικές τεχνικές επεξεργασίας των δεδομένων καθώς και η βελτιστοποίηση των παραμέτρων των αλγορίθμων είχαν ως στόχο την βελτίωση της προγνωστικής ικανότητας των μοντέλων. Στον πίνακα 5.9 και στο γράφημα 5.13 παρατίθενται ορισμένες σημαντικές **μετρικές αξιολόγησης** των τεσσάρων μοντέλων προς σύγκριση.

Πίνακας 5.9: Σύγκριση μετρικών αξιολόγησης των μοντέλων ταξινόμησης

	Ορθότητα	**Ακρίβεια**	**Ανάκληση**	**FPR**	**f1-score**
SVM	68,47 %	51,35 %	74,72 %	12,47 %	53,22 %
RF	84,00 %	59,41 %	70,27 %	11,47 %	63,42 %
AdaBoost	75,08 %	52,31 %	70,71 %	11,30 %	55,87 %
MLP	81,28 %	57,51 %	72,04 %	11,37 %	61,79 %



Γράφημα 5.13: Επίδοση των μοντέλων ταξινόμησης σύμφωνα με ορισμένες μετρικές αξιολόγησης

Με βάση τον πίνακα 5.10 στο σύνολο τους οι αλγόριθμοι σημειώνουν υψηλά ποσοστά ορθότητας (accuracy) και ανάκλησης (recall) συγκριτικά με την ακρίβεια (precision) και το f1-score. Όπως αναφέρθηκε στην ενότητα 3.6, η λανθασμένη ταξινόμηση δεδομένων επικίνδυνου επιπέδου σε λιγότερο επικίνδυνό θα είχε σοβαρές επιπτώσεις στην οδική ασφάλεια. Επομένως η ανάκληση αποτελεί αρκετά σημαντική μετρική αξιολόγησης. Ειδικά για το επίπεδο 'Avoidable Accident' η υψηλή ανάκληση σε συνδυασμό με χαμηλότερο ποσοστό ακρίβειας συνεπάγεται με υψηλή ικανότητα αναγνώρισης του πραγματικού επικίνδυνου επιπέδου αλλά υψηλότερο ποσοστό λανθασμένων ταξινομήσεων των επιπέδων 'Normal' και 'Dangerous' ως 'Avoidable Accident'. Στο πλαίσιο του συγκεκριμένου ζητήματος που εξετάζει η διπλωματική εργασία το παραπάνω σενάριο είναι ανεκτό. Σε περίπτωση αντίθετων αποτελεσμάτων θα υπήρχαν σοβαρά προβλήματα.

Με βάση την ορθότητα (accuracy), την ανάκληση (recall) και τον ρυθμό λανθασμένων θετικών προβλέψεων (false positive rate) των τεσσάρων μοντέλων τα καλύτερα αποτελέσματα προσφέρει ο αλγόριθμος 'Τυχαίων Δασών' (RF) και ο αλγόριθμος 'Πολυεπίπεδου Perceptron' (MLP).

5.3 Εντοπισμός διάρκειας οδήγησης σε επικίνδυνες συνθήκες

Με βάση τις οδηγίες του ερευνητικού έργου i-DREAMS τέθηκε σαν στόχος της διπλωματικής εργασίας η εξέταση της επίδρασης των παραγόντων κινδύνου στην **διάρκεια οδήγησης σε επικίνδυνες συνθήκες**. Για την επίτευξη του παραπάνω στόχου θα αναπτυχθούν τρία μοντέλα παλινδρόμησης και η αξιολόγηση των διαφορετικών παραγόντων και της επίδρασης τους στην διάρκεια οδήγησης θα πραγματοποιηθεί βάσει

τους συντελεστές (coefficients) των ανεξάρτητων μεταβλητών στο μοντέλο και της στατιστικής τους σημαντικότητας.

5.3.1 Υπολογισμός διάρκειας οδήγησης στα επίπεδα ασφαλείας

Όπως αναλύθηκε στην υποενότητα 5.2.1, προέκυψε το επίπεδο της 'Ζώνης Ανοχής Ασφαλείας' που βρίσκεται ο οδηγός σε κάθε χρονικό πλαίσιο 30 δευτερολέπτων. Επομένως αθροίζοντας τα χρονικά πλαίσια των 30 δευτερολέπτων προέκυψε **η συνολική διάρκεια οδήγησης** του κάθε οδηγού σε κάθε ένα από τα τρία επίπεδα ασφαλείας (πίνακας 5.10).

Πίνακας 5.10 Διάρκεια οδήγησης στα διαφορετικά επίπεδα της 'Ζώνης Ανοχής Ασφαλείας' για κάθε οδηγό

Οδηγός	**Επίπεδο 'Ζώνης Ανοχής Ασφαλείας'**	**Διάρκεια οδήγησης (δλ.)**
1	Normal	2310
	Dangerous	330
	Avoidable Accident	180
2	Normal	2310
	Dangerous	240
	Avoidable Accident	270
3	Normal	2460
	Dangerous	210
	Avoidable Accident	210
4	Normal	1860
	Dangerous	330
	Avoidable Accident	450
5	Normal	2490
	Dangerous	270
	Avoidable Accident	30
7	Normal	2610
	Dangerous	210
	Avoidable Accident	90
8	Normal	2280
	Dangerous	240
	Avoidable Accident	270
9	Normal	1560
	Dangerous	150
	Avoidable Accident	120
10	Normal	2400
	Dangerous	270

	Avoidable Accident	90
11	Normal	2460
	Dangerous	210
	Avoidable Accident	120
12	Normal	2490
	Dangerous	210
	Avoidable Accident	30
13	Normal	1650
	Dangerous	210
	Avoidable Accident	30
14	Normal	1860
	Dangerous	150
	Avoidable Accident	540
15	Normal	2670
	Dangerous	300
	Avoidable Accident	120
16	Normal	2190
	Dangerous	390
	Avoidable Accident	120
18	Normal	2610
	Dangerous	210
	Avoidable Accident	60
19	Normal	2340
	Dangerous	390
	Avoidable Accident	60
20	Normal	2190
	Dangerous	360
	Avoidable Accident	180
21	Normal	2070
	Dangerous	390
	Avoidable Accident	210
25	Normal	2370
	Dangerous	240
	Avoidable Accident	90
26	Normal	2430
	Dangerous	240

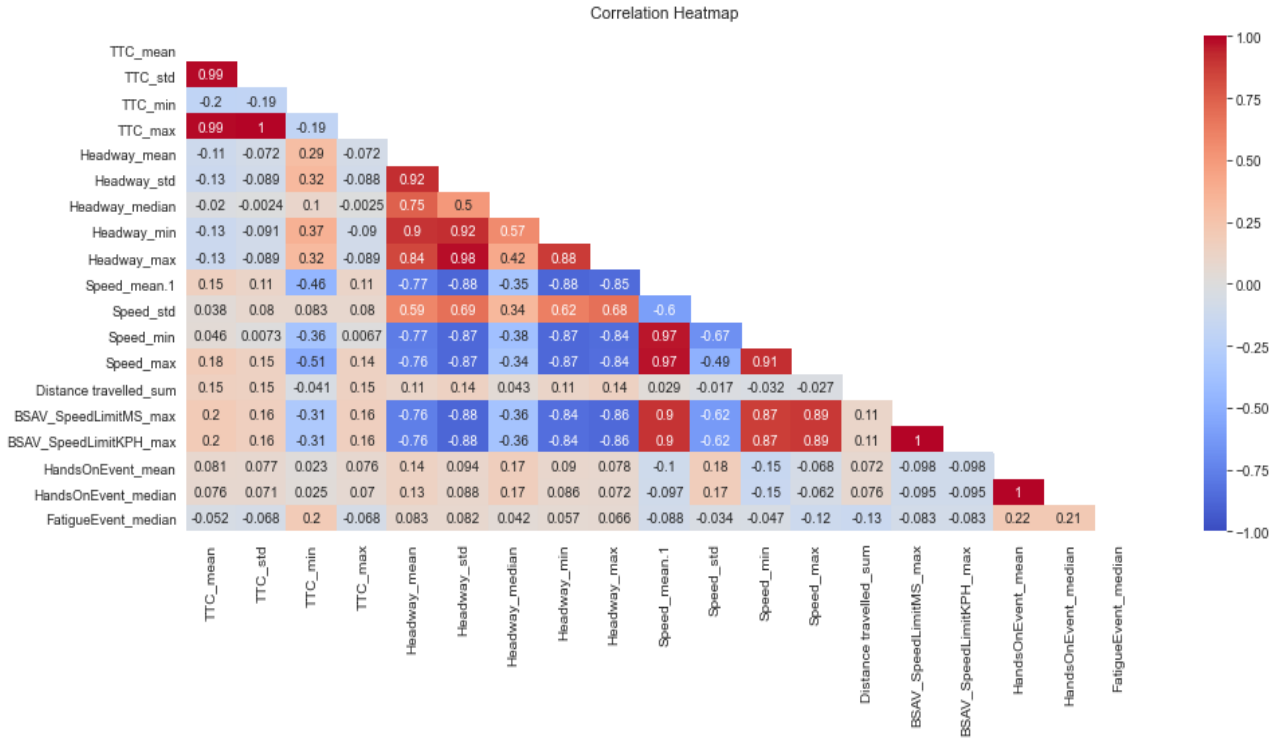
	Avoidable Accident	210
27	Normal	2490
	Dangerous	240
	Avoidable Accident	90
28	Normal	2670
	Dangerous	270
	Avoidable Accident	30
29	Normal	2760
	Dangerous	180
	Avoidable Accident	120
30	Normal	2430
	Dangerous	270
	Avoidable Accident	0
36	Normal	2070
	Dangerous	450
	Avoidable Accident	270
37	Normal	2910
	Dangerous	300
	Avoidable Accident	120
38	Normal	2790
	Dangerous	270
	Avoidable Accident	60
40	Normal	2250
	Dangerous	420
	Avoidable Accident	30
41	Normal	2700
	Dangerous	210
	Avoidable Accident	30
42	Normal	2460
	Dangerous	390
	Avoidable Accident	0
44	Normal	1860
	Dangerous	570
	Avoidable Accident	210
45	Normal	2400
	Dangerous	300

	Avoidable Accident	60
46	Normal	2220
	Dangerous	210
	Avoidable Accident	270
47	Normal	2400
	Dangerous	240
	Avoidable Accident	120
48	Normal	2580
	Dangerous	270
	Avoidable Accident	0

5.3.2 Επιλογή ανεξάρτητων μεταβλητών

Σε αντίθεση με την διαδικασία της ταξινόμησης, θα εξεταστούν όλες οι μεταβλητές που παρουσιάστηκαν του πίνακα 4.3. Στις αναλύσεις ωστόσο λόγω της σχέσης εξορισμού των μεταβλητών TTC και Headway, θα ληφθεί μία εκ των δύο. Στοχεύοντας στην συσχέτιση των διαφόρων μεταβλητών με την διάρκεια οδήγησης, υπολογίστηκε η **μέση τιμή** τους για κάθε οδηγό σε κάθε επίπεδο της 'Ζώνης Ανοχής Ασφαλείας'.

Η συσχέτιση μεταξύ των παραγόντων μετά την παραπάνω διαδικασία παρουσιάζεται στο γράφημα 5.14.



Γράφημα 5.14: Συσχέτιση ανεξάρτητων μεταβλητών

Από την συσχέτιση των μεταβλητών προκύπτουν οι εξής παρατηρήσεις:

- Ομοίως με την ανάλυση συσχέτισης στην διαδικασία της ταξινόμησης παρατηρείται υψηλή συσχέτιση μεταξύ περιγραφικών στοιχείων της ίδιας μεταβλητής.
- Υψηλή συσχέτιση μεταξύ στοιχείων της ταχύτητας (Speed) και των ορίων ταχύτητας (BSAV_SpeedLimit).
- Υψηλή συσχέτιση μεταξύ στοιχείων της ταχύτητας (Speed) και της χρονικής απόστασης από το προπορευόμενο όχημα (Headway).
- Χαμηλή συσχέτιση μεταξύ HandsOnEvent, FatigueEvent, Distance travelled και όλων των υπολοίπων μεταβλητών.

Αξιολογώντας την επίδοση των μοντέλων, την στατιστική σημαντικότητα και την συσχέτιση μεταξύ των μεταβλητών, επιλέχθηκαν οι εξής **δύο** παράγοντες ως ανεξάρτητες μεταβλητές του μοντέλου:

1. Speed_max
2. Distance travelled_sum

Η αξιολόγηση της επιρροής και της σημαντικότητας των μεταβλητών θα πραγματοποιηθεί με την ανάπτυξη των μοντέλων παλινδρόμησης και τον προσδιορισμό των συντελεστών των ανεξάρτητων μεταβλητών.

5.3.3 Προετοιμασία δεδομένων

Οι μεταβλητές Speed_max και Distance travelled_sum για κάθε οδηγό σε κάθε επίπεδο ασφαλείας, αποτελούν τις ανεξάρτητες μεταβλητές για τα μοντέλα παλινδρόμησης. Βασική επιδίωξη είναι η ανάπτυξη μοντέλων παλινδρόμησης με εξ' ολοκλήρου στατιστικά σημαντικές μεταβλητές. Επίσης η διάρκεια οδήγησης στα τρία επίπεδα που αναλύθηκε στην ενότητα 5.3.1 αποτελεί την εξαρτημένη μεταβλητή στην διαδικασία της παλινδρόμησης. Όμοια με την διαδικασία της ταξινόμησης, τα δύο σύνολα διαιρέθηκαν στα **δεδομένα εκπαίδευσης (training dataset)** και τα **δεδομένα εξέτασης (testing dataset)** με ποσοστό 85% και 15% αντίστοιχα.

5.3.4 Ανάπτυξη μοντέλων παλινδρόμησης

Βάσει της βιβλιογραφικής ανασκόπησης και για τον σκοπό της παρούσας μελέτης επιλέχθηκαν τρία **μοντέλα παλινδρόμησης** όπως παρουσιάστηκαν στην ενότητα 3.5. Η ονοματολογία και ο συμβολισμός των τριών αλγορίθμων παλινδρόμησης μηχανικής εκμάθησης παρατίθεται στον πίνακα 5.11.

Πίνακας 5.11: Ονοματολογία και συμβολισμός μοντέλων παλινδρόμησης

Όνομα μοντέλου (ελληνικά)	Όνομα μοντέλου (αγγλικά)	Συμβολισμός μοντέλου
Παλινδρόμηση Κορυφογραμμής	Ridge Regression	RR
Παλινδρόμηση Lasso	Lasso Regression	LR
Παλινδρόμηση Elastic Net	Elastic Net Regression	ENR

Αξιοποιώντας την τεχνική GridSearchCV της βιβλιοθήκης scikit-learn, εντοπίστηκαν οι βέλτιστες παράμετροι των μοντέλων με σκοπό την βελτίωση των επιδόσεων τους.

Οι τρείς αλγόριθμοι παλινδρόμησης αναπτύχθηκαν μέσω της βιβλιοθήκης scikit-learn της προγραμματιστικής γλώσσας python.

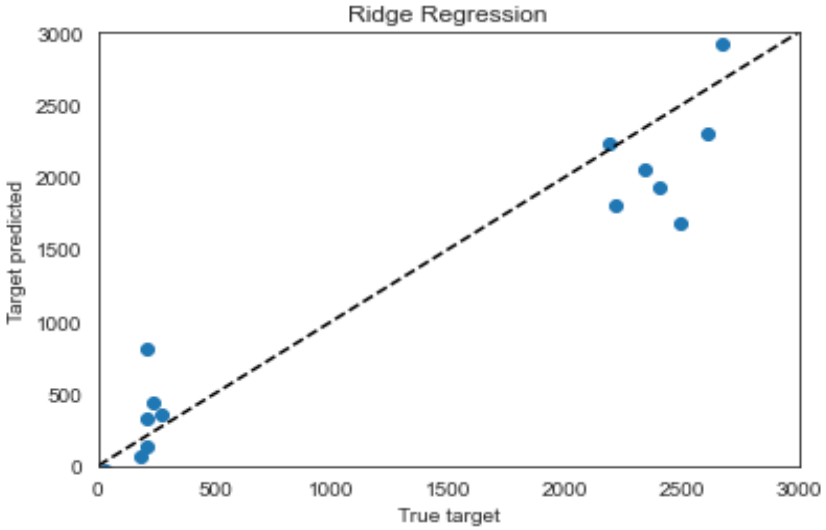
1) Αλγόριθμος Παλινδρόμησης Κορυφογραμμής (RR)

Όπως φαίνεται στον πίνακα 5.12, ο συντελεστής της μεταβλητής Distance travelled_sum μηδενίζεται στο πλαίσιο της διαδικασίας αντιμετώπισης της συγγραμμικότητας, που πραγματοποιεί ο αλγόριθμός παλινδρόμησης κορυφογραμμής.

Πίνακας 5.12: Σύνοψη μοντέλου παλινδρόμησης RR

Σύνοψη μοντέλου παλινδρόμησης RR				
	Συντελεστές	Τυπική απόκλιση	t value	p value
Σταθερός όρος	9966,716	472,905	21,076	0,000
Speed_max	-112,009	2,178	-51,441	0,000
Distance travelled_sum	0,001	0,000	8,896	0,000
R^2 = 0,8493		Adjusted R^2 = 0,8458		

Στο γράφημα 5.15, απεικονίζονται οι αποκλίσεις των προβλεπόμενων και των πραγματικών μεταβλητών. Παρατηρείται σχετική απόκλιση ιδιαίτερα για υψηλές τιμές.



Γράφημα 5.15: Σχέση προβλεπόμενων και πραγματικών τιμών μοντέλου RR

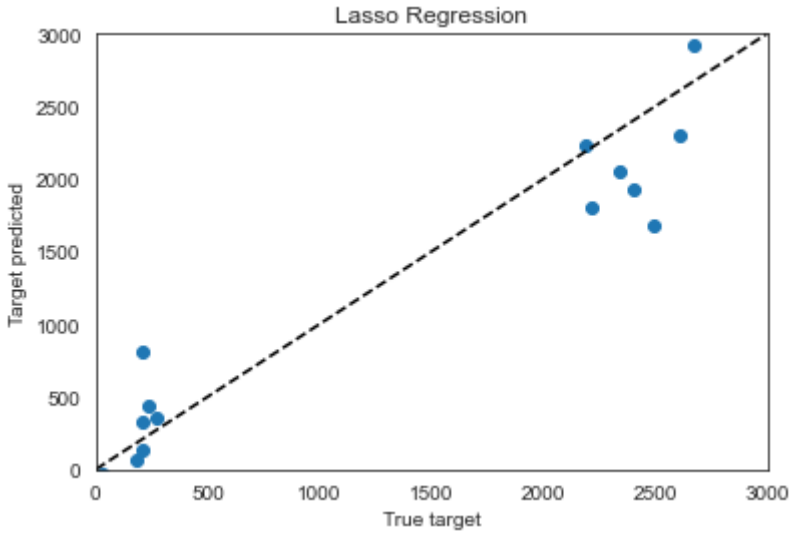
2) Αλγόριθμος Παλινδρόμησης Lasso

Αντίστοιχα με τον αλγόριθμό RR, ο αλγόριθμός παλινδρόμησης Lasso αντιμετωπίζει την συγγραμμικότητα των μεταβλητών.

Πίνακας 5.13: Σύνοψη μοντέλου παλινδρόμηση LR

Σύνοψη μοντέλου παλινδρόμησης LR				
	Συντελεστές	Τυπική απόκλιση	t value	p value
Σταθερός όρος	9967,358	472,905	21,077	0,000
Speed_max	-112,017	2,177	-51,445	0,000
Distance travelled_sum	0,001	0,000	8,896	0,000
R^2 = 0,8493		Adjusted R^2 = 0,8458		

Οι προβλεπόμενες και πραγματικές τιμές του γραφήματος 5.16, ομοίως με προηγουμένως, εμφανίζουν αποκλίσεις.



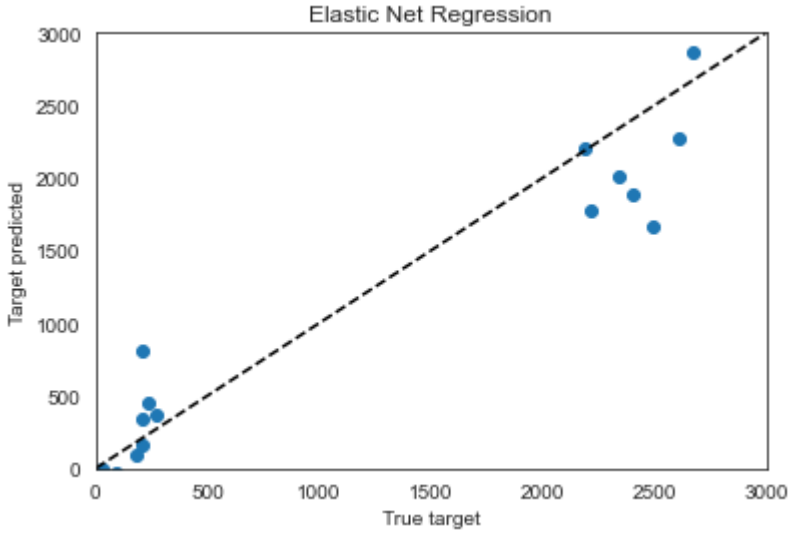
Γράφημα 5.16: Σχέση προβλεπόμενων και πραγματικών τιμών μοντέλου LR

3) Αλγόριθμος Παλινδρόμησης Elastic Net

Δεδομένου ότι ο αλγόριθμος Elastic Net αποτελεί συνδυασμό των μεθόδων RR και LL τα αποτελέσματα του πίνακα 5.14 και 5.17 είναι σχεδόν όμοια με προηγουμένως.

Πίνακας 5.14: Σύνοψη μοντέλου παλινδρόμηση ENR

Σύνοψη μοντέλου παλινδρόμησης ENR				
	Συντελεστές	Τυπική απόκλιση	t value	p value
Σταθερός όρος	9697,044	472,981	20,459	0,000
Speed_max	-108,840	2,182	-49,873	0,000
Distance travelled_sum	0,001	0,000	8,955	0,000
R^2 = 0,8486		Adjusted R^2 = 0,8451		



Γράφημα 5.17: Σχέση προβλεπόμενων και πραγματικών τιμών μοντέλου ENR

5.3.5 Αξιολόγηση μοντέλων παλινδρόμησης και αποτελεσμάτων

Με βάση τα αποτελέσματα των μοντέλων παλινδρόμησης παρατηρείται ότι στο σύνολο τους σημειώνουν υψηλές τιμές του συντελεστή προσδιορισμού R^2. Επομένως οι ανεξάρτητες μεταβλητές των μοντέλων έχουν υψηλή ικανότητα ερμηνείας της διακύμανσης της εξαρτημένης μεταβλητής.

Αξιολογώντας την σχέση μεταξύ των ανεξάρτητων μεταβλητών, την στατιστική σημαντικότητα τους και την επιρροή τους στις επιδόσεις και των τριών μοντέλων προέκυψε ως βέλτιστο το σύνολο μεταβλητών Speed_max και Distance travelled_sum. Με βάση τους συντελεστές της παλινδρόμησης προκύπτει ότι ο παράγοντας **Speed_max** έχει την υψηλότερη επίδραση στην διάρκεια οδήγησης σε επικίνδυνες συνθήκες.

5.4 Σύνοψη

Όπως αναλύθηκε στην ενότητα 5.2, εξετάστηκαν διάφοροι οδηγικοί παράγοντες που επηρεάζουν την επικίνδυνη οδήγηση. Η διερεύνηση της επιρροής των παραγόντων αυτών έγινε βάσει ενός συστήματος αναγνώρισης του επιπέδου της ‘Ζώνης Ανοχής Ασφαλείας’. Συγκεκριμένα, αναπτύχθηκαν τέσσερεις αλγόριθμοι ταξινόμησης και εξετάστηκε η επίδραση των διαφορετικών μεταβλητών στο σύνολο αυτών μέσω ορισμένων τεχνικών επιλογής και επεξεργασίας στοιχείων. Οι μεταβλητές που επιδρούσαν άμεσα στην αναγνώριση και κατ’ επέκταση στην ίδια την επικίνδυνη οδηγική συμπεριφορά ήταν **(1) η διανυθείσα απόσταση, (2) η μέγιστη ταχύτητα και (3) το μέγιστο όριο ταχύτητας** σε κάθε χρονικό διάστημα των 30 δευτερολέπτων.

Τα μοντέλα ταξινόμησης στο σύνολο τους είχαν ικανοποιητικά αποτελέσματα για την αναγνώριση του επιπέδου της ΄Ζώνης Ανοχής Ασφαλείας’ που βρίσκεται ο οδηγός σε κάθε χρονικό πλαίσιο των 30 δευτερολέπτων. Παρόλα αυτά συγκρίνοντας της μετρικές αξιολόγησης προέκυψε ότι το μοντέλο **‘Τυχαίων Δασών’ (Random Forests)** και το μοντέλο **‘Πολυεπίπεδου Perceptron’ (Multilayer Perceptron)** είχαν τα καλύτερα αποτελέσματα για το σύνολο των επιπέδων ασφάλειας.

Επίσης με βάση την δεύτερη προσέγγιση της παρούσας μελέτης προέκυψε ότι **η μέγιστη ταχύτητα** επιδρά αρνητικά στην διάρκεια οδήγησης σε επικίνδυνες συνθήκες. Συγκεκριμένα αναπτύχθηκαν τρία μοντέλα παλινδρόμησης και με βάση την στατιστική σημαντικότητα και την επιρροή στην επίδοση του μοντέλου επιλέχθηκε ένα σύνολο μεταβλητών. Η τελική αξιολόγηση της επίδρασης των μεταβλητών στην διαδικασία πρόβλεψης της διάρκειας οδήγησης σε επικίνδυνες συνθήκες πραγματοποιήθηκε βάσει του συντελεστή κάθε ανεξάρτητης μεταβλητής στο μοντέλο.

Αξίζει να αναλυθούν σε βάθος τα συμπεράσματα και τα αποτελέσματα που προέκυψαν ώστε να συμβάλλουν στην καλύτερη κατανόηση των παραγόντων που επιδρούν στην αναγνώριση της επικίνδυνης οδήγησης. Τέλος είναι αναγκαίο να επισημανθούν η συνεισφορά και οι σημαντικές ελλείψεις της έρευνα ώστε να αποτελέσουν την βάση για περαιτέρω διερεύνηση.

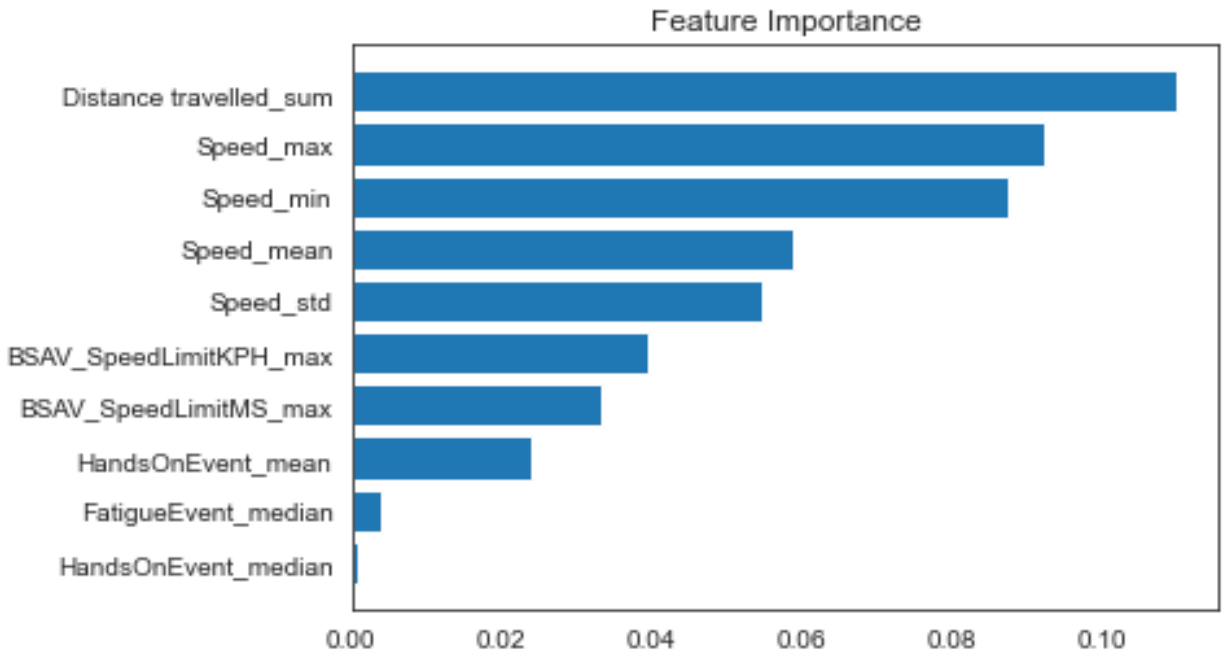
6. ΣΥΜΠΕΡΑΣΜΑΤΑ

6.1 Σύνοψη Αποτελεσμάτων

Στόχος της παρούσας διπλωματικής εργασίας είναι **ο εντοπισμός επιπέδου και διάρκειας επικίνδυνης συμπεριφοράς του οδηγού (Ζώνης Ανοχής Ασφαλείας) με τεχνικές μηχανικής εκμάθησης**. Τα δεδομένα που αναλύθηκαν, συλλέχθηκαν από προσομοιωτή οδήγησης κατάλληλα διαμορφωμένο για το ερευνητικό έργο i-DREAMS. Για την ανάλυση της οδηγικής συμπεριφοράς ήταν αναγκαίο να οριστούν τα διαφορετικά επίπεδα της 'Ζώνης Ανοχής Ασφαλείας' βάσει ορισμένων τεχνικών. Τελικά ο καθορισμός των επιπέδων ασφαλείας πραγματοποιήθηκε με βάση την μεταβλητή Headway_min, καθώς η συγκεκριμένη τεχνική προσέφερε τη βέλτιστη, σύμφωνα με την βιβλιογραφία, κατανομή των δειγμάτων στα τρία επίπεδα:

- Επίπεδο 'Normal' (class: 0) : Headway_min > 2 δλ.
- Επίπεδο 'Dangerous' (class: 1) : Headway_min > 1.4 δλ. και Headway_min < 2 δλ.
- Επίπεδο 'Avoidable Accident' (class: 2) : Headway_min < 1.4 δλ.

Στο πρώτο μέρος των αναλύσεων αναπτύχθηκαν κατάλληλες τεχνικές προσδιορισμού της **σημαντικότητας των μεταβλητών στην πρόβλεψη του επιπέδου 'Ζώνης Ανοχής Ασφαλείας' που βρίσκεται ο οδηγός**. Επισημαίνεται ότι οι μεταβλητές Headway και TTC δεν λαμβάνονται υπόψη στο πρώτο μέρος των αναλύσεων καθώς θα αναπτύσσονταν προβλήματα μεροληψίας των μοντέλων ταξινόμησης. Η σημαντικότητα φαίνεται στο γράφημα που ακολουθεί.



Γράφημα 1: Σημαντικότητα μεταβλητών για την πρόβλεψη του επιπέδου 'Ζώνης Ανοχής Ασφαλείας'

Στη συνέχεια αξιοποιώντας τις σημαντικότερες μεταβλητές, αναπτύχθηκαν τέσσερεις αλγόριθμοι μηχανικής εκμάθησης με σκοπό **την ταξινόμηση της οδηγικής συμπεριφοράς σε ένα από τα τρία επίπεδα ασφαλείας**. Εφαρμόζοντας την 'Προσαρμοστική Συνθετική' (ADASYN) τεχνική επιλύθηκε το πρόβλημα άνισης κατανομής

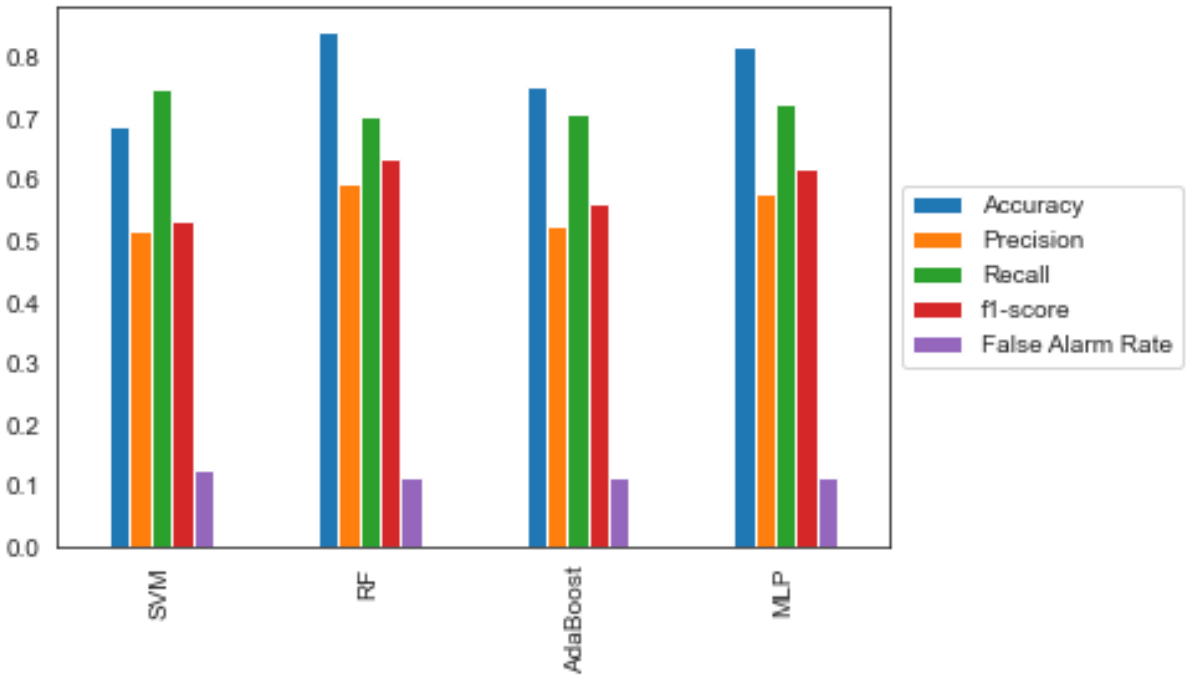
των δεδομένων εκπαίδευσης στις διαφορετικές κλάσεις. Η ονοματολογία και ο συμβολισμός των τεσσάρων αλγορίθμων παρατίθενται στον πίνακα ενώ οι επιδόσεις τους παρουσιάζονται στον πίνακα και στο γράφημα που ακολουθεί.

Πίνακας 1: Ονοματολογία και συμβολισμός μοντέλων ταξινόμησης

Όνομα μοντέλου (ελληνικά)	**Όνομα μοντέλου (αγγλικά)**	**Συμβολισμός μοντέλου**
Μηχανές Διανυσμάτων Υποστήριξης	Support Vector Machines	SVM
Ταξινομητής Τυχαίων Δασών	Random Forests Classifier	RF
Ταξινομητής AdaBoost	AdaBoost Classifier	AdaBoost
Ταξινομητής Πολυεπίπεδου Perceptron	Multilayer Perceptron Classifier	MLP

Πίνακας 2: Σύγκριση μετρικών αξιολόγησης των μοντέλων ταξινόμησης

	Ορθότητα	**Ακρίβεια**	**Ανάκληση**	**FPR**	**f1-score**
SVM	68,47 %	51,35 %	74,72 %	12,47 %	53,22 %
RF	84,00 %	59,41 %	70,27 %	11,47 %	63,42 %
AdaBoost	75,08 %	52,31 %	70,71 %	11,30 %	55,87 %
MLP	81,28 %	57,51 %	72,04 %	11,37 %	61,79 %



Γράφημα 2: Επίδοση των μοντέλων ταξινόμησης σύμφωνα με ορισμένες μετρικές αξιολόγησης

Στο δεύτερο μέρος των αναλύσεων **εξετάστηκε η διάρκεια που βρίσκεται κάθε οδηγός σε κάθε επίπεδο της 'Ζώνης Ανοχής Ασφαλείας'**. Για τον σκοπό αυτό αναπτύχθηκαν τρία μοντέλα παλινδρόμησης μηχανικής εκμάθησης. Η επιλογή των ανεξάρτητων μεταβλητών πραγματοποιήθηκε με βάση την επίδοση των μοντέλων σε συνδυασμό με τη στατιστική σημαντικότητα και τη συσχέτιση μεταξύ των ανεξάρτητων μεταβλητών. Οι αλγόριθμοι που εφαρμόστηκαν λαμβάνουν υπόψη και αντιμετωπίζουν την πολύ-συγγραμμικότητα. Επίσης οι αλγόριθμοι πραγματοποιούν επιλογή χαρακτηριστικών μηδενίζοντας ή μειώνοντας τους συντελεστές των ανεξάρτητων μεταβλητών. Η τελική αξιολόγηση της επίδρασης των παραγόντων στην διάρκεια οδήγησης σε κάθε επίπεδο ασφαλείας προκύπτει με βάση τον συντελεστή κάθε ανεξάρτητης μεταβλητής στο μοντέλο της παλινδρόμησης. Στον πίνακα παρουσιάζεται η ονοματολογία και ο συμβολισμός των μοντέλων, ενώ στους πίνακες τα τελικά αποτελέσματα.

Πίνακας 3: Ονοματολογία και συμβολισμός μοντέλων παλινδρόμησης

Όνομα μοντέλου (ελληνικά)	**Όνομα μοντέλου (αγγλικά)**	**Συμβολισμός μοντέλου**
Παλινδρόμηση Κορυφογραμμής	Ridge Regression	RR
Παλινδρόμηση Lasso	Lasso Regression	LR
Παλινδρόμηση Elastic Net	Elastic Net Regression	ENR

Πίνακας 4: Σύνοψη μοντέλου παλινδρόμησης RR

Σύνοψη μοντέλου παλινδρόμησης RR				
	Συντελεστές	Τυπική απόκλιση	t value	p value
Σταθερός όρος	9966,716	472,905	21,076	0,000
Speed_max	-112,009	2,178	-51,441	0,000
Distance travelled_sum	0,001	0,001	8,896	0,000
R^2 = 0,8493		Adjusted R^2 = 0,8458		

Πίνακας 5: Σύνοψη μοντέλου παλινδρόμησης LR

Σύνοψη μοντέλου παλινδρόμησης LR				
	Συντελεστές	Τυπική απόκλιση	t value	p value
Σταθερός όρος	9967,358	472,905	21,077	0,000
Speed_max	-112,017	2,177	-51,445	0,000
Distance travelled_sum	0,001	0,001	8,896	0,000
R^2 = 0,8493		Adjusted R^2 = 0,8458		

Πίνακας 6: Σύνοψη μοντέλου παλινδρόμησης ENR

Σύνοψη μοντέλου παλινδρόμησης ENR				
	Συντελεστές	Τυπική απόκλιση	t value	p value
Σταθερός όρος	9697,044	472,981	20,459	0,000
Speed_max	-108,840	2,182	-49,873	0,000
Distance travelled_sum	0,001	0,001	8,955	0,000
R^2 = 0,8486		Adjusted R^2 = 0,8451		

6.2 Σύνοψη Συμπερασμάτων

Βάσει των αποτελεσμάτων που προέκυψαν κατά την εφαρμογή της μεθοδολογίας, προέκυψαν ορισμένα συμπεράσματα άμεσα σχετιζόμενα με τον στόχο της διπλωματικής εργασίας.

- Ο καθορισμός των επιπέδων ασφαλείας της 'Ζώνης Ανοχής Ασφαλείας' με βάση ορίων της μεταβλητής headway_min παρείχε **αποτελέσματα συναφή με τη διεθνή βιβλιογραφία** όσον αφορά την κατανομή των δειγμάτων στις κλάσεις, σε σχέση με τις άλλες τεχνικές που εξετάστηκαν.

- Σύμφωνα με τα αποτελέσματα του γραφήματος 6.1, η συνολική διανυθείσα απόσταση είναι η **σημαντικότερη μεταβλητή** για την αναγνώριση της οδηγικής συμπεριφοράς. Ανάλογα με την συνολική απόσταση που διανύει ο οδηγός μπορεί να παρατηρηθούν διαφορετικές οδηγικές συμπεριφορές. Για παράδειγμα οι οδηγοί που διανύουν μεγάλες αποστάσεις είναι πιθανό να εμφανίσουν σημάδια κούρασης και μειωμένης προσοχής, τα οποία οδηγούν σε επικίνδυνη οδηγική συμπεριφορά.

- Η ταχύτητα (μέγιστη, ελάχιστη, μέση τιμή, τυπική απόκλιση) είχαν εξίσου **σημαντική επιρροή** στην διαδικασία ταξινόμησης. Η ταχύτητα σχετίζεται άμεσα με την πιθανότητα εμφάνισης ατυχήματος καθώς επίσης και με την σοβαρότητα αυτού. Όσο ο οδηγός αυξάνει την ταχύτητα οδήγησης, ελαχιστοποιείται ο χρόνος αντίδρασης του οδηγού.

- Τα όρια ταχύτητας τίθενται από τους αρμόδιους προκειμένου η οδήγηση να πραγματοποιείται με ασφάλεια. Η υπέρβαση του ορίου ταχύτητας σχετίζεται με την εμφάνιση ατυχημάτων. Με βάση το γράφημα 6.1, η μεταβλητή του ορίου ταχύτητας είναι **σημαντική για την αναγνώριση της οδηγικής συμπεριφοράς**.

- Η κατάσταση του οδηγού και η αλληλεπίδραση του με το τιμόνι του οχήματος έχουν **μειωμένη επιρροή** στην αναγνώριση του επιπέδου ασφαλείας που βρίσκεται. Η σημαντικότητα των μεταβλητών FatigueEvent και HandsOnEvent είναι μικρότερη σε σχέση με του υπόλοιπους οδηγικούς παράγοντες. Παρόλα η κατάσταση του οδηγού και η αλληλεπίδραση του με το τιμόνι σχετίζεται με τους υπόλοιπους οδηγικούς παράγοντες (όπως η ταχύτητα ή η διανυθείσα απόσταση).

- Από τις διαφορετικές τεχνικές αντιμετώπισης του φαινομένου της άνισης κατανομής των δειγμάτων στις διαφορετικές κλάσεις, η 'Προσαρμοστική Συνθετική' (ADASYN) προσέφερε τα **βέλτιστα αποτελέσματα** για το σύνολο των ταξινομητών. Όπως αναλύθηκε στο κεφάλαιο 3.3, η τεχνική ADASYN έχει το πλεονέκτημα να αντιμετωπίζει την μεροληψία ως προς την κυρίαρχη τάξη και να ωθεί τα όρια απόφασης της ταξινόμησης στα πιο δύσκολα παραδείγματα.

- Στην παρούσα εργασία αναπτύχθηκαν τέσσερεις αλγόριθμοι ταξινόμησης οι οποίοι σημείωσαν ικανοποιητικές επιδόσεις. Η μέθοδος 'Τυχαίων Δασών' (RF) και η μέθοδος 'Πολυεπίπεδου Perceptron' (MLP) σημείωσαν τις **υψηλότερες επιδόσεις** στην πλειοψηφία των μετρικών αξιολόγησης τους.

- Από το σύνολο των μεταβλητών που εξετάστηκαν, η μέγιστη ταχύτητα και η συνολική διανυθείσα απόσταση προσέφεραν **στατιστικά σημαντικά αποτελέσματα**. Με βάση τους συντελεστές παλινδρόμησης, η μέγιστη ταχύτητα έχει την κύρια, αρνητική επίδραση στην διάρκεια οδήγησης στα διαφορετικά επίπεδα ασφάλειας. Η ελαχιστοποίηση του συντελεστή της διανυθείσας απόστασης πραγματοποιείται στο πλαίσιο αντιμετώπισης της συγγραμμικότητας των μεταβλητών. Επομένως, η μέγιστη ταχύτητα είναι **ιδιαίτερα σημαντική** στην πρόβλεψη της διάρκειας οδήγησης σε κάθε επίπεδο.

- Τα τρία μοντέλα παλινδρόμησης (RR, LR, ENR) στο σύνολο τους έχουν **υψηλή προγνωστική ικανότητα**. Η ομοιότητα στην τιμή των συντελεστών παλινδρόμησης οφείλεται στο γεγονός ότι οι τρείς αλγόριθμοι έχουν παρόμοια λειτουργία, όπως παρουσιάστηκε στην ενότητα 3.5.

6.3 Προτάσεις για αξιοποίηση των αποτελεσμάτων

Με βάση τα αποτελέσματα και τα συμπεράσματα που εξήχθησαν κατά την εκπόνηση της μελέτης αυτής επιχειρείται η παράθεση μίας σειράς **προτάσεων αξιοποίησης των ευρημάτων**, οι οποίες ενδεχομένως θα μπορούσαν να συμβάλλουν στην καλύτερη κατανόηση της επιρροής των διαφόρων παραγόντων στην οδική ασφάλεια αλλά και στην εξέλιξη της έρευνας των Ευφυών Μεταφορικών Συστημάτων (ITS).

- **Αξιοποίηση των μοντέλων ταξινόμησης** για την αναγνώριση του επιπέδου ασφαλείας των οδηγών σε πραγματικές συνθήκες οδήγησης. Από τις επιδόσεις των τεσσάρων αλγορίθμων ταξινόμησης, προκύπτει ότι μπορούν να προσφέρουν ικανοποιητικά αποτελέσματα και έτσι θα μπορούσαν να αξιοποιηθούν για την περαιτέρω διερεύνηση της οδηγικής συμπεριφοράς.

- **Περαιτέρω διερεύνηση των κρισιμότερων παραγόντων** που επιδρούν στην αναγνώριση της επικίνδυνης οδηγικής συμπεριφοράς. Κατά αυτόν τον τρόπο θα ενισχυθεί η προσπάθεια της επιστημονικής κοινότητας και της αυτοκινητοβιομηχανίας για βελτίωση των προηγμένων συστημάτων υποστήριξης οδηγού.

- **Ανάπτυξη κατάλληλου συστήματος αναγνώρισης** του επιπέδου 'Ζώνης Ανοχής Ασφαλείας' που βρίσκεται ο οδηγός σε πραγματικό χρόνο εντός του οχήματος. Επομένως ο οδηγός κατά την διάρκεια της διαδρομής θα μπορεί να παρακολουθεί

την οδηγική του συμπεριφορά και να παρέμβει σε περίπτωση παρέκκλισης από την ασφαλή οδήγηση.

- Με βάση την πρόβλεψη της διάρκειας οδήγησης σε κάθε επίπεδο της 'Ζώνης Ανοχής Ασφαλείας' είναι δυνατόν **να συντονιστεί η συχνότητα και η ένταση των προειδοποιήσεων** κατά την διάρκεια της διαδρομής (Michelaraki et al., 2021).

- **Δημιουργία εφαρμογής σε έξυπνα κινητά τηλέφωνα** η οποία θα λαμβάνει τα δεδομένα οδήγησης και θα προβλέπει την διάρκεια που βρίσκεται ο οδηγός σε κάθε επίπεδο της 'Ζώνης Ανοχής Ασφαλείας'. Με αυτόν τον τρόπο μετά από την λήξη της διαδρομής ο οδηγός θα μπορεί να προβεί στις απαραίτητες παρεμβάσεις στην οδηγική του συμπεριφορά.

6.4 Προτάσεις για περαιτέρω έρευνα

Η υιοθέτηση σύγχρονων μεθόδων επεξεργασίας και ανάλυσης στον τομέα της οδικής ασφάλειας, αυξάνεται διαρκώς. Η ανάλυση της οδηγικής συμπεριφοράς με την αξιοποίηση μεθόδων μηχανικής εκμάθησης, αποτελεί αντικείμενο υψηλού ενδιαφέροντος για τους ερευνητές. Στις μελέτες που αναλύθηκαν προέκυψαν ορισμένα ζητήματα. Για την αντιμετώπιση αυτών οι ερευνητές πρότειναν την εξέταση επιπλέον παραγόντων και μεθόδων.

Η παρούσα μελέτη επιχείρησε να καλύψει το κενό που προέκυψε από την βιβλιογραφική ανασκόπηση εξετάζοντας διαφορετικές τεχνικές μηχανικής εκμάθησης και αποσκοπώντας να αποτελέσει βάση για ακόμα περισσότερες συγκριτικές αναλύσεις. Παρ΄ όλα αυτά κατά την ανάπτυξη της μεθοδολογίας και την αξιολόγηση των αποτελεσμάτων εντοπίστηκαν ορισμένες ελλείψεις οι οποίες θα μπορούσαν να ληφθούν υπόψη σε μελλοντικές έρευνες. Παρακάτω παρουσιάζονται **προτάσεις για περαιτέρω έρευνα** οι οποίες θα μπορούσαν να συμβάλουν στην εμβάθυνση, στην πληρέστερη κατανόηση και κατ' επέκταση στην αντιμετώπιση των ζητημάτων που προέκυψαν.

- **Αξιοποίηση μεγαλύτερου όγκου δεδομένων** με σκοπό την βελτίωση της προγνωστικής ικανότητας των μοντέλων ταξινόμησης και παλινδρόμησης. Όσο αυξάνεται ο αριθμός των δεδομένων, παράλληλα μειώνεται η πιθανότητα σφάλματος του μοντέλου.

- **Εξέταση επιπλέον τεχνικών ομαδοποίησης** για τον καθορισμό του επιπέδου της 'Ζώνης Ανοχής Ασφαλείας'. Δεδομένου ότι οι τεχνικές ομαδοποίησης που αναπτύχθηκαν για τον καθορισμό του επιπέδου ασφαλείας δεν οδήγησαν σε ορθή κατανομή των τριών επιπέδων, ο καθορισμός πραγματοποιήθηκε με βάση την μεταβλητή Headway. Η μεταβλητή Headway καθώς και η μεταβλητή TTC εξαιρέθηκαν από την διαδικασία ταξινόμησης, ώστε να μην εμφανιστεί μεροληψία στα μοντέλα. Κατά αυτόν τον τρόπο δεν εξετάζεται η επιρροή των δύο μεταβλητών στην επικίνδυνη οδήγηση.

- **Ανάπτυξη εναλλακτικών τεχνικών εξέτασης σημαντικότητας χαρακτηριστικών (feature importance).** Η περαιτέρω διερεύνηση της σημαντικότητας των μεταβλητών

μπορεί να προσδιορίσει με μεγαλύτερη ακρίβεια την σχέση των μεταβλητών με την ικανότητα αναγνώρισης του επιπέδου ασφαλείας που βρίσκεται κάθε οδηγός.

- **Ανάπτυξη πρόσθετων αλγορίθμων παλινδρόμησης και τεχνικών εξέτασης** των παραγόντων που επιδρούν στην διάρκεια οδήγησης σε επικίνδυνες συνθήκες. Όπως αναφέρθηκε στην βιβλιογραφική ανασκόπηση, δεν εντοπίστηκαν έρευνες ανάλυσης της διάρκειας οδήγησης σε επικίνδυνες συνθήκες. Επομένως βάση της προσέγγισης και των αποτελεσμάτων που προέκυψαν στην παρούσα εργασία, κρίνεται αναγκαία η περαιτέρω διερεύνηση του συγκεκριμένου ζητήματος.

- **Εξέταση μοντέλων βαθιάς εκμάθησης (deep learning).** Η βαθιά εκμάθηση αποτελείται από μία σύνθετη δομή αλγορίθμων μηχανικής εκμάθησης που έχει διαμορφωθεί με βάση τον ανθρώπινο εγκέφαλο. Η βαθιά εκμάθηση αφαιρεί την χειροκίνητη αναγνώριση χαρακτηριστικών των δεδομένων. Αντ' αυτού βασίζεται σε οποιαδήποτε εκπαιδευτική διαδικασία προκειμένου να ανακαλύψει τα χρήσιμα μοτίβα στα παραδείγματα εισόδου. Με αυτόν τον τρόπο η διαδικασία επιταχύνεται και οδηγεί σε καλύτερα αποτελέσματα.

- **Ανάπτυξη κατάλληλων μοντέλων για ταξινόμηση ακολουθίας (sequence classification)** του επιπέδου 'Ζώνης Ανοχής Ασφαλείας' που θα βρίσκεται ο οδηγός στο επόμενο χρονικό πλαίσιο των 30 δευτερολέπτων.

- **Διερεύνηση της επιρροής πρόσθετων παραγόντων**. Με βάση την παρούσα μελέτη αλλά και τις έρευνες που αναζητήθηκαν κατά την βιβλιογραφική ανασκόπηση, οι παράγοντες που θα μπορούσαν μελλοντικά να εξεταστούν αφορούν τις καιρικές συνθήκες, τα στοιχεία της οδού, τα χαρακτηριστικά και τις αντιλήψεις (σχετικά με την επικινδυνότητα κατά την οδήγηση) των οδηγών.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Abirami, S., Chitra, P., 2020. Energy-efficient edge based real-time healthcare support system, in: Advances in Computers. Academic Press Inc., pp. 339–368. https://doi.org/10.1016/bs.adcom.2019.09.007
2. Aljanahi, A.A.M., Rhodes, A.H., Metcalfe, A.V., 1999. Speed, speed limits and road traffic accidents under free flow conditions. Accident Analysis & Prevention 31, 161–168. https://doi.org/10.1016/S0001-4575(98)00058-X
3. Ariannezhad, A., Karimpour, A., Qin, X., Wu, Y.-J., Salmani, Y., 2021. Handling Imbalanced Data for Real-Time Crash Prediction: Application of Boosting and Sampling Techniques. Journal of Transportation Engineering, Part A: Systems 147, 04020165. https://doi.org/10.1061/JTEPBS.0000499
4. Assi, K., 2020. Traffic Crash Severity Prediction—A Synergy by Hybrid Principal Component Analysis and Machine Learning Models. International Journal of Environmental Research and Public Health 17. https://doi.org/10.3390/ijerph17207598
5. Chandaka, S., Chatterjee, A., Munshi, S., 2009. Cross-correlation aided support vector machine classifier for classification of EEG signals. Expert Systems with Applications 36, 1329–1336. https://doi.org/10.1016/j.eswa.2007.11.017
6. Chawla, N. v., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. https://doi.org/10.1613/jair.953
7. Chen, X. (Michael), Zhang, S., Li, L., 2019. Multi-model ensemble for short-term traffic flow prediction under normal and abnormal conditions. IET Intelligent Transport Systems 13, 260–268. https://doi.org/10.1049/iet-its.2018.5155
8. Djuris, J., Ibric, S., Djuric, Z., 2013. 4 - Chemometric methods application in pharmaceutical products and processes analysis and control, in: Djuris, J. (Ed.), Computer-Aided Applications in Pharmaceutical Technology. Woodhead Publishing, pp. 57–90. https://doi.org/10.1533/9781908818324.57
9. Dukart, J., 2015. Basic Concepts of Image Classification Algorithms Applied to Study Neurodegenerative Diseases, in: Toga, A.W. (Ed.), Brain Mapping. Academic Press, Waltham, pp. 641–646. https://doi.org/10.1016/B978-0-12-397025-1.00072-5
10. Elamrani Abou Elassad, Z., Mousannif, H., al Moatassime, H., 2020. A real-time crash prediction fusion framework: An imbalance-aware strategy for collision avoidance systems. Transportation Research Part C: Emerging Technologies 118, 102708. https://doi.org/10.1016/j.trc.2020.102708
11. European Commission, Directorate-General for Mobility and Transport, 2020. Next steps towards ‘Vision Zero’: EU road safety policy framework 2021-2030. Publications Office. https://doi.org/doi/10.2832/261629
12. Fagnant, D.J., Kockelman, K., 2015. Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. Transportation Research Part A: Policy and Practice 77, 167–181. https://doi.org/https://doi.org/10.1016/j.tra.2015.04.003
13. Fernández, A., del Jesus, M.J., Herrera, F., 2009. Hierarchical fuzzy rule-based classification systems with genetic rule selection for imbalanced data-sets. International Journal of Approximate Reasoning 50, 561–577. https://doi.org/10.1016/j.ijar.2008.11.004

14. Ghandour, R., Potams, A.J., Boulkaibet, I., Neji, B., al Barakeh, Z., 2021. Driver Behavior Classification System Analysis Using Machine Learning Methods. Applied Sciences 11. https://doi.org/10.3390/app112210562

15. Ghorbani, R., Ghousi, R., 2020. Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques. IEEE Access 8, 67899–67911. https://doi.org/10.1109/ACCESS.2020.2986809

16. Grandini, M., Bagli, E., Visani, G., 2020. Metrics for Multi-Class Classification: an Overview. ArXiv abs/2008.05756.

17. Guo, M., Zhao, X., Yao, Y., Yan, P., Su, Y., Bi, C., Wu, D., 2021. A study of freeway crash risk prediction and interpretation based on risky driving behavior and traffic flow data. Accident Analysis & Prevention 160, 106328. https://doi.org/10.1016/j.aap.2021.106328

18. Hall, M.A., 2000. Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning, in: ICML.

19. He, H., Bai, Y., Garcia, E.A., Li, S., 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning, in: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). pp. 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

20. Hinsbergen, C., Lint, J.W.C., Sanders, F., 2007. Short Term Traffic Prediction Models. 14th World Congress on Intelligent Transport Systems, ITS 2007 7.

21. i-DREAMS [WWW Document], 2022. URL https://idreamsproject.eu/wp/ (accessed 2.10.22).

22. Imbalanced learn [WWW Document], 2022. URL https://imbalanced-learn.org/stable/ (accessed 2.15.22).

23. Islam, Z., Abdel-Aty, M., Cai, Q., Yuan, J., 2021. Crash data augmentation using variational autoencoder. Accident Analysis & Prevention 151, 105950. https://doi.org/10.1016/j.aap.2020.105950

24. James, Gareth., Witten, Daniela., Hastie, Trevor., Tibshirani, Robert., 2013. An Introduction to Statistical Learning with Applications in R, 1st ed. 2013. ed, Springer Texts in Statistics, 103. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4614-7138-7

25. Kaggle: Your Machine Learning and Data Science Community [WWW Document], 2022. URL https://www.kaggle.com/ (accessed 1.4.22).

26. Katrakazas, C., 2017. Developing an advanced collision risk model for autonomous vehicles.

27. Katrakazas, C., Antoniou, C., Yannis, G., 2020. Identification of driving simulator sessions of depressed drivers: A comparison between aggregated and time-series classification. Transportation Research Part F: Traffic Psychology and Behaviour 75, 16–25. https://doi.org/10.1016/j.trf.2020.09.015

28. Katrakazas, C., Antoniou, C., Yannis, G., 2019. Time series classification using imbalanced learning for real-time safety assessment, in: Proceedings of the Transportation Research Board (TRB) 98th Annual Meeting. Washington DC, United States.

29. Kotsiantis, S., Kanellopoulos, D., Pintelas, P., 2005. Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering 30, 25–36.

30. Lee, J.H., Shi, Z., Gao, Z., 2021. On LASSO for predictive regression. Journal of Econometrics. https://doi.org/10.1016/j.jeconom.2021.02.002

31. Lewis-Evans, B., de Waard, D., Brookhuis, K.A., 2010. That's close enough—A threshold effect of time headway on the experience of risk, task difficulty, effort, and comfort. Accident Analysis & Prevention 42, 1926–1933. https://doi.org/10.1016/j.aap.2010.05.014

32. Liu, W., Dou, Z., Wang, W., Liu, Y., Zou, H., Zhang, B., Hou, S., 2018. Short-Term Load Forecasting Based on Elastic Net Improved GMDH and Difference Degree Weighting Optimization. Applied Sciences 8. https://doi.org/10.3390/app8091603

33. Lugo Reyes, S.O., 2020. Chapter 21 - Artificial intelligence in precision health: Systems in practice, in: Barh, D. (Ed.), Artificial Intelligence in Precision Health. Academic Press, pp. 499–519. https://doi.org/10.1016/B978-0-12-817133-2.00021-5

34. Matplotlib: Visualization with Python [WWW Document], 2022. URL https://matplotlib.org/ (accessed 2.15.22).

35. Medium [WWW Document], 2022. URL https://medium.com/ (accessed 2.12.22).

36. Meiring, G.A.M., Myburgh, H.C., 2015. A Review of Intelligent Driving Style Analysis Systems and Related Artificial Intelligence Algorithms. Sensors 15, 30653–30682. https://doi.org/10.3390/s151229822

37. Michael, P.G., Leeming, F.C., Dwyer, W.O., 2000. Headway on urban streets: observational data and an intervention to decrease tailgating. Transportation Research Part F: Traffic Psychology and Behaviour 3, 55–64. https://doi.org/10.1016/S1369-8478(00)00015-2

38. Michelaraki, E., Katrakazas, C., Brijs, T., Yannis, G., 2021a. Modelling the Safety Tolerance Zone: Recommendations from the i-DREAMS project, in: 10th International Congress on Transportation Research. Rhodes Island, Greece.

39. Michelaraki, E., Katrakazas, C., Yannis, G., Konstantina Frantzola, E., Kalokathi, F., Kaiser, S., Brijs, K., Brijs, T., 2021b. A Review of Real-Time Safety Intervention Technologies, in: 7th Humanist Conference. Rhodes Island, Greece.

40. Misra, S., Li, H., 2020. Chapter 9 - Noninvasive fracture characterization based on the classification of sonic wave travel times, in: Misra, S., Li, H., He, J. (Eds.), Machine Learning for Subsurface Characterization. Gulf Professional Publishing, pp. 243–287. https://doi.org/10.1016/B978-0-12-817736-5.00009-0

41. Morris, C., Yang, J.J., 2021. Effectiveness of resampling methods in coping with imbalanced crash data: Crash type analysis and predictive modeling. Accident Analysis & Prevention 159, 106240. https://doi.org/10.1016/j.aap.2021.106240

42. Nettleton, D., 2014. Chapter 6 - Selection of Variables and Factor Derivation, in: Nettleton, D. (Ed.), Commercial Data Mining. Morgan Kaufmann, Boston, pp. 79–104. https://doi.org/10.1016/B978-0-12-416602-8.00006-6

43. Ng, A.Y., 2004. Feature Selection, L1 vs. L2 Regularization, and Rotational Invariance, in: Proceedings of the Twenty-First International Conference on Machine Learning, ICML '04. Association for Computing Machinery, New York, NY, USA, p. 78. https://doi.org/10.1145/1015330.1015435

44. NTUA Road Safety Observatory [WWW Document], 2022. URL https://www.nrso.ntua.gr/ (accessed 2.1.22).

45. NumPy library for the Python programming language [WWW Document], 2022. URL https://numpy.org/ (accessed 3.10.22).

46. Ohta, H., 1993. Individual differences in driving distance headway. Vision in vehicles 4, 91–100.

47. Osman, O.A., Hajij, M., Karbalaieali, S., Ishak, S., 2019. A hierarchical machine learning classification approach for secondary task identification from observed driving behavior data. Accident Analysis & Prevention 123, 274–281. https://doi.org/10.1016/j.aap.2018.12.005

48. Pandas - Python Data Analysis Library [WWW Document], 2022. URL https://pandas.pydata.org/ (accessed 2.10.22).

49. Pande, A., Abdel-Aty, M., 2006. Assessment of freeway traffic parameters leading to lane-change related collisions. Accident Analysis & Prevention 38, 936–948. https://doi.org/10.1016/j.aap.2006.03.004

50. Peppes, N., Alexakis, T., Adamopoulou, E., Demestichas, K., 2021. Driving Behaviour Analysis Using Machine and Deep Learning Methods for Continuous Streams of Vehicular Data. Sensors 21. https://doi.org/10.3390/s21144704

51. Roshandel, S., Zheng, Z., Washington, S., 2015. Impact of real-time traffic characteristics on freeway crash occurrence: Systematic review and meta-analysis. Accident Analysis & Prevention 79, 198–211. https://doi.org/10.1016/j.aap.2015.03.013

52. Scikit-learn: Machine Learning in Python [WWW Document], 2022. URL https://scikit-learn.org/ (accessed 2.5.22).

53. Seaborn: statistical data visualization [WWW Document], 2022. URL https://seaborn.pydata.org/ (accessed 2.1.22).

54. Shangguan, Q., Fu, T., Wang, J., Luo, T., Fang, S., 2021. An integrated methodology for real-time driving risk status prediction using naturalistic driving data. Accident Analysis & Prevention 156, 106122. https://doi.org/10.1016/j.aap.2021.106122

55. Shi, X., Wong, Y.D., Li, M.Z.F., Chai, C., 2018. Key risk indicators for accident assessment conditioned on pre-crash vehicle trajectory. Accident Analysis & Prevention 117, 346–356. https://doi.org/10.1016/j.aap.2018.05.007

56. Shi, X., Wong, Y.D., Li, M.Z.-F., Palanisamy, C., Chai, C., 2019. A feature learning approach based on XGBoost for driving assessment and risk prediction. Accident Analysis & Prevention 129, 170–179. https://doi.org/10.1016/j.aap.2019.05.005

57. Song, X., Yin, Y., Cao, H., Zhao, S., Li, M., Yi, B., 2021. The mediating effect of driver characteristics on risky driving behaviors moderated by gender, and the classification model of driver's driving risk. Accident Analysis & Prevention 153, 106038. https://doi.org/10.1016/j.aap.2021.106038

58. Song, Y., Kou, S., Wang, C., 2021. Modeling crash severity by considering risk indicators of driver and roadway: A Bayesian network approach. Journal of Safety Research 76, 64–72. https://doi.org/10.1016/j.jsr.2020.11.006

59. Theodoridis, S., 2020. Chapter 6 - The Least-Squares Family, in: Theodoridis, S. (Ed.), Machine Learning (Second Edition). Academic Press, pp. 253–299. https://doi.org/10.1016/B978-0-12-818803-3.00015-5

60. Towards Data Science [WWW Document], 2022. URL https://towardsdatascience.com/ (accessed 2.2.22).

61. Voss, D.S., 2005. Multicollinearity, in: Kempf-Leonard, K. (Ed.), Encyclopedia of Social Measurement. Elsevier, New York, pp. 759–770. https://doi.org/10.1016/B0-12-369398-5/00428-X

62. Wang, J., Huang, H., Li, Y., Zhou, H., Liu, J., Xu, Q., 2020. Driving risk assessment based on naturalistic driving study and driver attitude questionnaire analysis. Accident Analysis & Prevention 145, 105680. https://doi.org/10.1016/j.aap.2020.105680

63. Wang, Jiaqi, Ma, Y., Yang, X., Li, T., Wei, H., 2021. Short-Term Traffic Prediction considering Spatial-Temporal Characteristics of Freeway Flow. Journal of Advanced Transportation 2021. https://doi.org/10.1155/2021/5815280

64. Wang, Junhua, Song, H., Fu, T., Behan, M., Jie, L., He, Y., Shangguan, Q., 2021. Crash prediction for freeway work zones in real time: A comparison between Convolutional Neural Network and Binary Logistic Regression model. International Journal of Transportation Science and Technology. https://doi.org/10.1016/j.ijtst.2021.06.002

65. Woo, C., Kulić, D., 2016. Manoeuvre segmentation using smartphone sensors, in: 2016 IEEE Intelligent Vehicles Symposium (IV). pp. 572–577. https://doi.org/10.1109/IVS.2016.7535444

66. World Health Organization, 2021. Global Plan: Decade of Action for Road Safety 2021-2030 [WWW Document]. World Health Organization. URL https://www.who.int/teams/social-determinants-of-health/safety-and-mobility/decade-of-action-for-road-safety-2021-2030 (accessed 2.2.22).

67. World Health Organization, 2018. Global Status Report On Road Safety 2018 [WWW Document]. World Health Organization. URL https://www.who.int/publications/i/item/9789241565684 (accessed 2.2.22).

68. Wu, M., Zhang, S., Dong, Y., 2016. A Novel Model-Based Driving Behavior Recognition System Using Motion Sensors. Sensors 16. https://doi.org/10.3390/s16101746

69. Xu, C., Tarko, A.P., Wang, W., Liu, P., 2013. Predicting crash likelihood and severity on freeways with real-time loop detector data. Accident Analysis & Prevention 57, 30–39. https://doi.org/10.1016/j.aap.2013.03.035

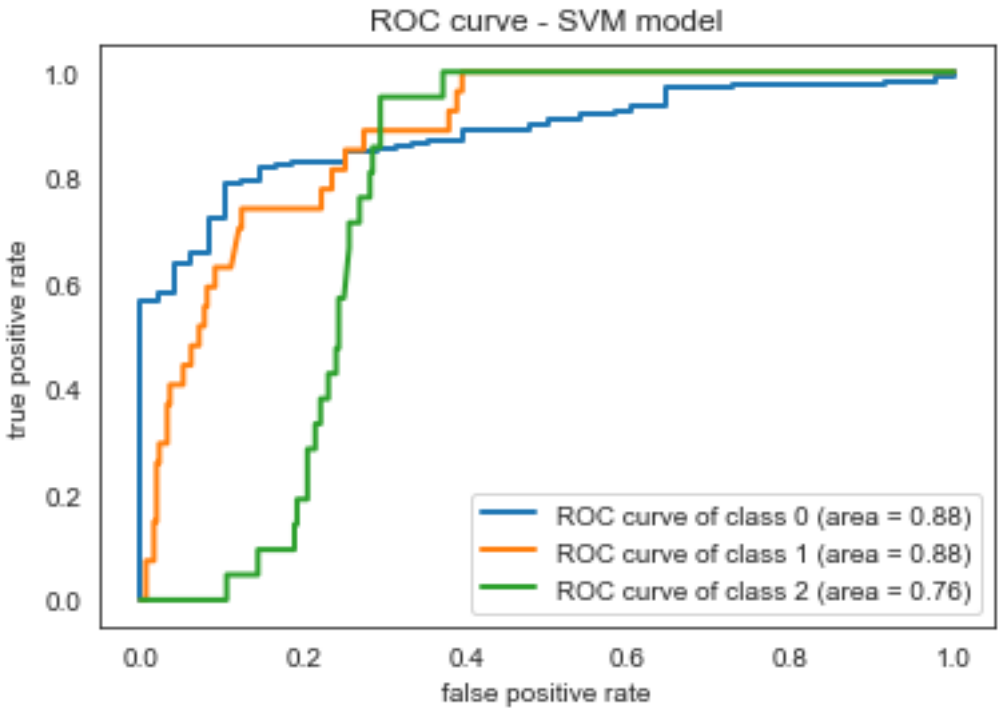
70. Yang, K., Haddad, C. al, Yannis, G., Antoniou, C., 2021. Driving Behavior Safety Levels: Classification and Evaluation, in: 2021 7th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS). pp. 1–6. https://doi.org/10.1109/MT-ITS49943.2021.9529309

71. Yi, D., Su, J., Liu, C., Quddus, M., Chen, W.-H., 2019. A machine learning based personalized system for driving state recognition. Transportation Research Part C: Emerging Technologies 105, 241–261. https://doi.org/10.1016/j.trc.2019.05.042

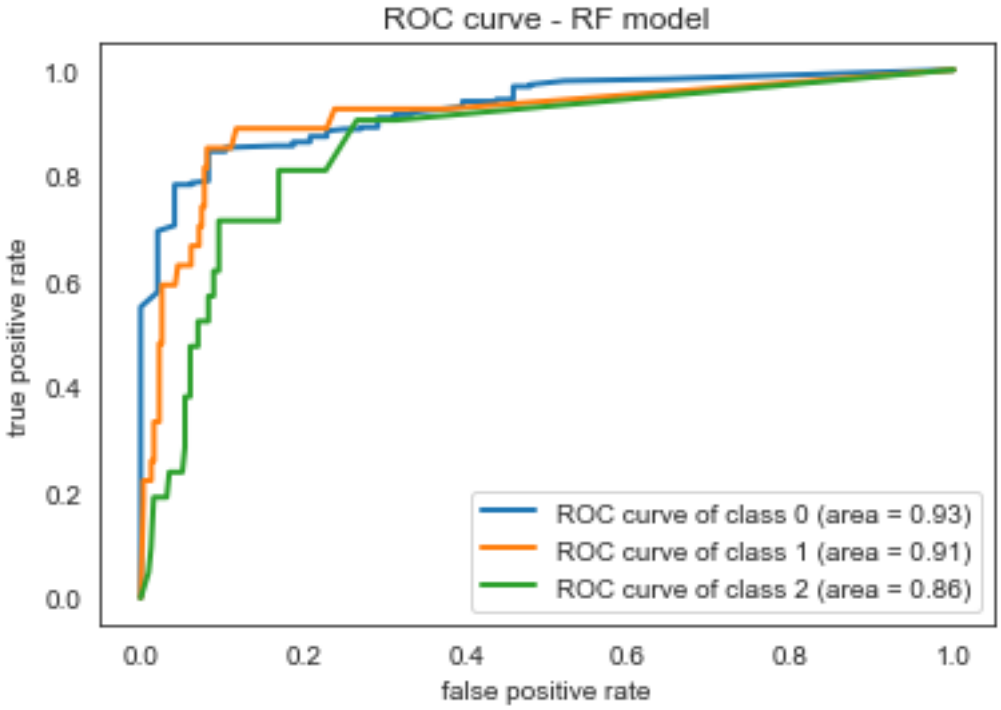
72. Yu, R., Abdel-Aty, M., 2013. Utilizing support vector machine in real-time crash risk evaluation. Accident Analysis & Prevention 51, 252–259. https://doi.org/10.1016/j.aap.2012.11.027

73. Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

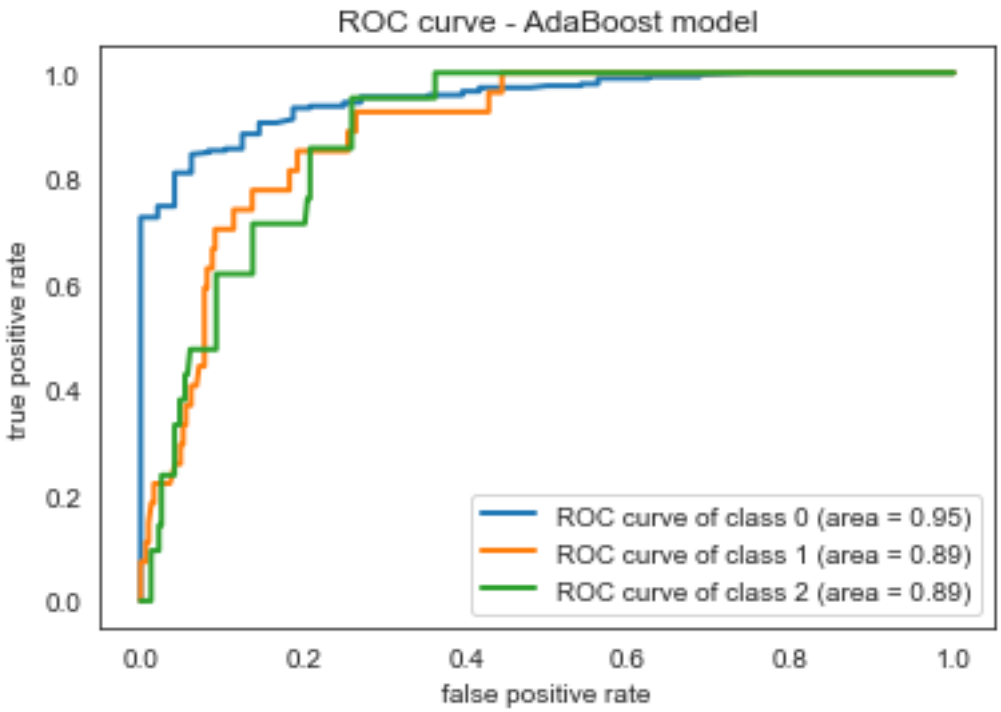
ΠΑΡΑΡΤΗΜΑΤΑ



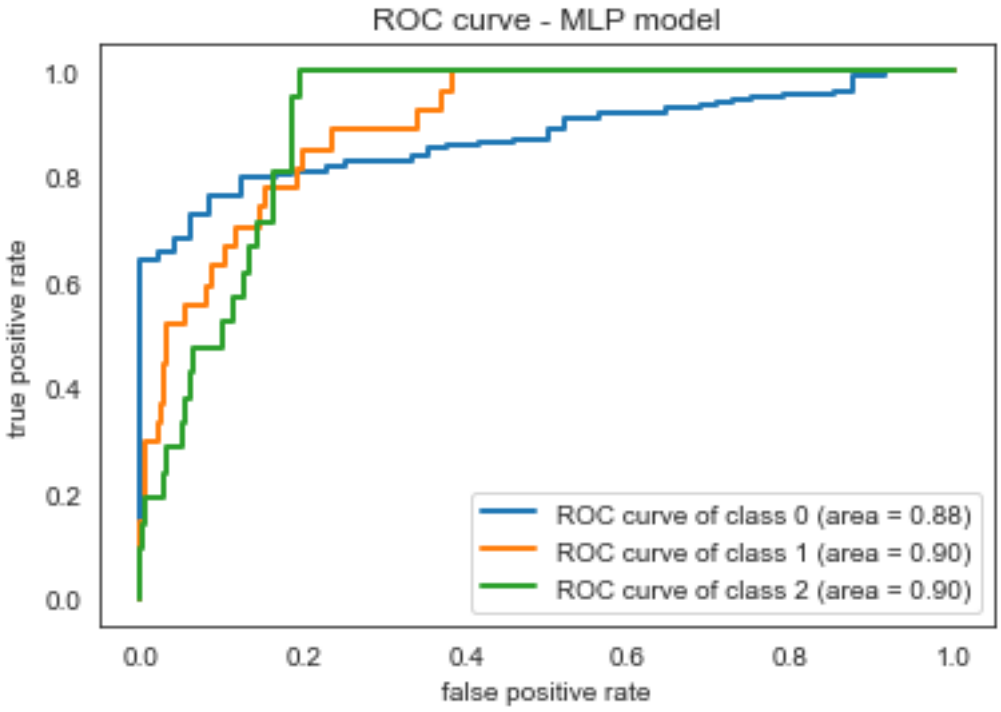
Γράφημα 5: Καμπύλη Διαχείρισης Λειτουργικών Χαρακτηριστικών (ROC) για το μοντέλο SVM



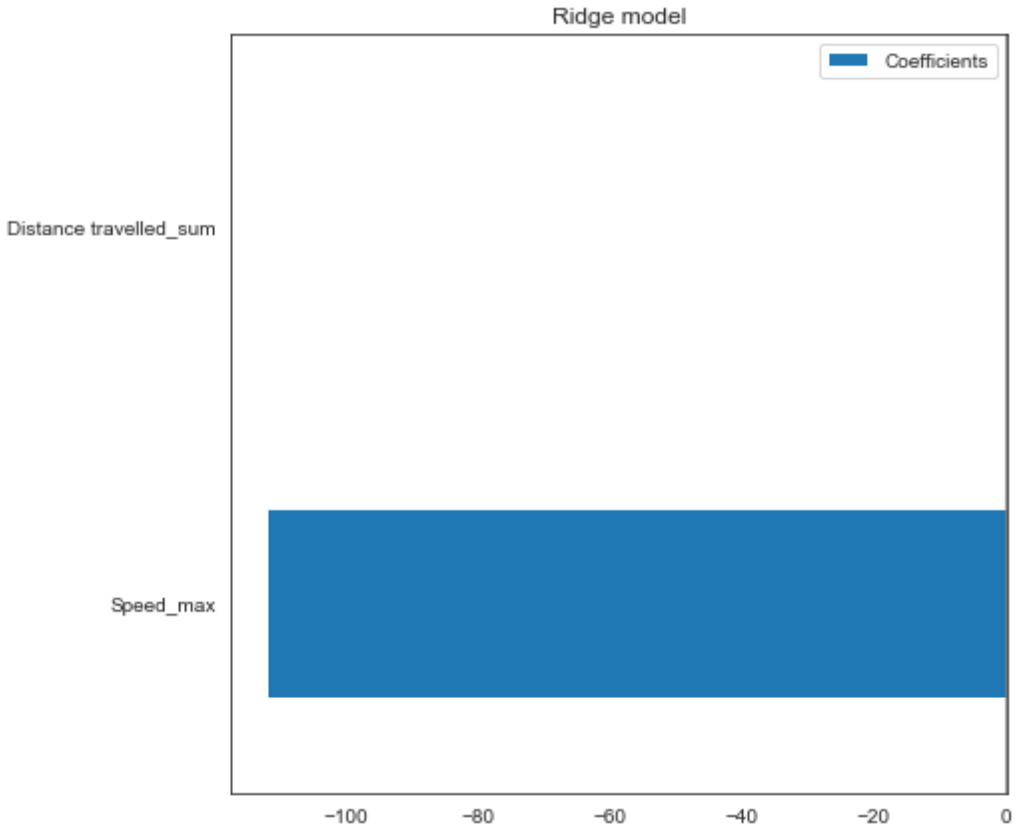
Γράφημα 6: Καμπύλη Διαχείρισης Λειτουργικών Χαρακτηριστικών (ROC) για το μοντέλο RF



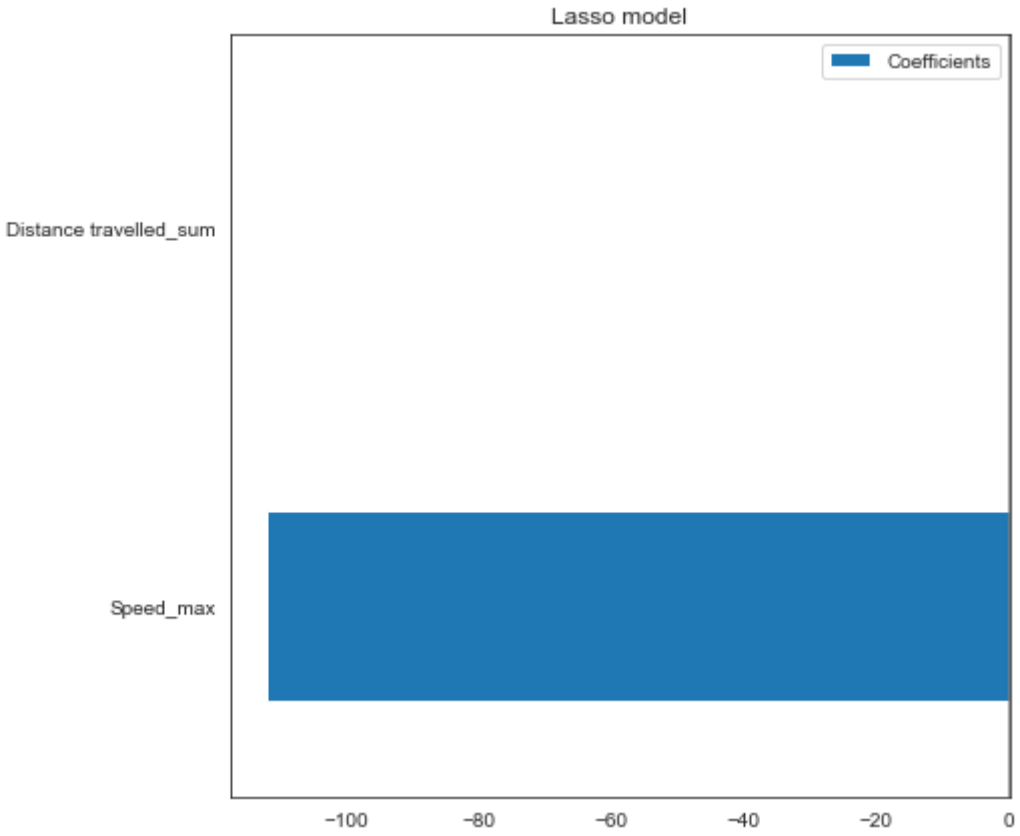
Γράφημα 7: Καμπύλη Διαχείρισης Λειτουργικών Χαρακτηριστικών (ROC) για το μοντέλο AdaBoost



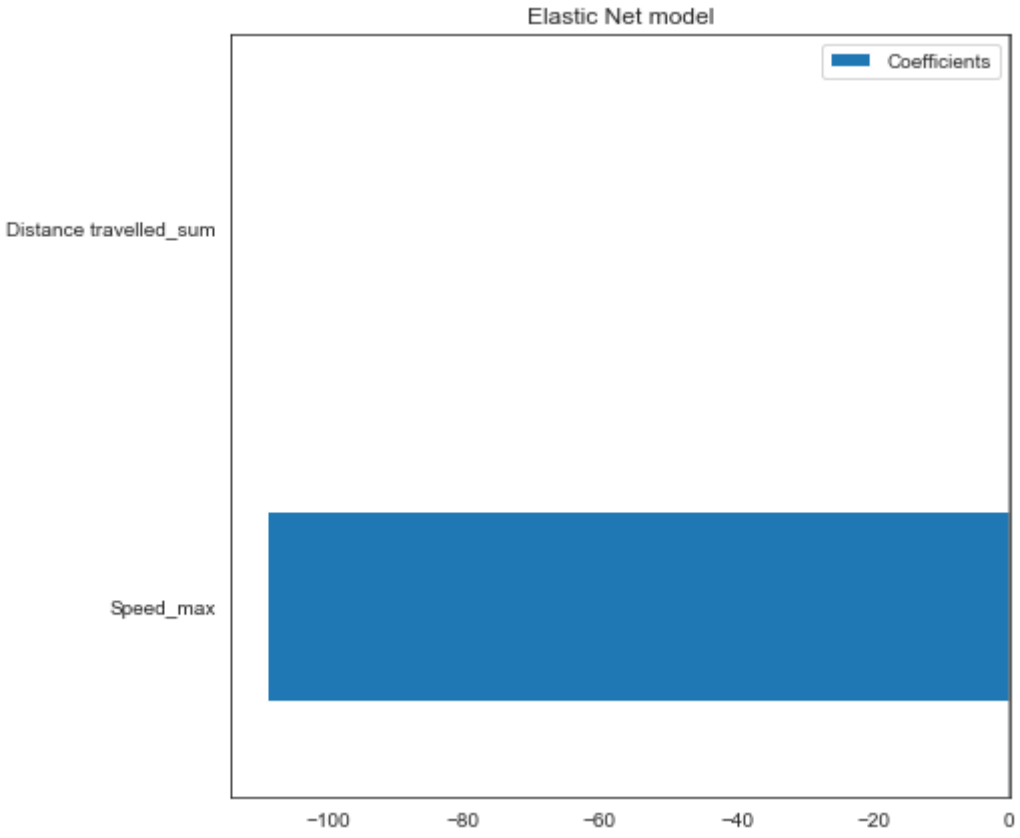
Γράφημα 8: Καμπύλη Διαχείρισης Λειτουργικών Χαρακτηριστικών (ROC) για το μοντέλο MLP



Γράφημα 9: Συντελεστές παλινδρόμησης μοντέλου RR



Γράφημα 10: Συντελεστές παλινδρόμησης μοντέλου LR



Γράφημα 11: Συντελεστές παλινδρόμησης μοντέλου ENR