ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΠΟΛΙΤΙΚΩΝ ΜΗΧΑΝΙΚΩΝ
ΤΟΜΕΑΣ ΜΕΤΑΦΟΡΩΝ ΚΑΙ ΣΥΓΚΟΙΝΩΝΙΑΚΗΣ ΥΠΟΔΟΜΗΣ

ΕΝΤΟΠΙΣΜΟΣ ΕΠΙΚΙΝΔΥΝΗΣ ΣΥΜΠΕΡΙΦΟΡΑΣ ΟΔΗΓΟΥ ΜΕ ΔΕΔΟΜΕΝΑ ΕΥΡΕΙΑΣ ΚΛΙΜΑΚΑΣ ΑΠΟ ΕΞΥΠΝΑ ΣΥΣΤΗΜΑΤΑ ΚΑΤΑΓΡΑΦΗΣ ΚΑΙ ΤΕΧΝΙΚΕΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

ΈΚΤΟΡΑΣ ΚΑΜΒΟΥΣΙΩΡΑΣ

ΕΠΙΒΛΕΠΩΝ: ΓΙΩΡΓΟΣ ΓΙΑΝΝΗΣ, ΚΑΘΗΓΗΤΗΣ Ε.Μ.Π.

ΑΘΗΝΑ, ΙΟΥΛΙΟΣ 2022

ΕΥΧΑΡΙΣΤΙΕΣ

Με την παρούσα Διπλωματική Εργασία ολοκληρώνεται ο κύκλος των προπτυχιακών σπουδών μου στη Σχολή Πολιτικών Μηχανικών του Εθνικού Μετσόβιου Πολυτεχνείου.

Θα ήθελα πρωτίστως να ευχαριστήσω θερμά τον κύριο Γιώργο Γιαννή, Καθηγητή του Εθνικού Μετσόβιου Πολυτεχνείου, για την ανάθεση και επίβλεψη της παρούσας Διπλωματικής Εργασίας και για την καθοδήγηση του καθ' όλη τη διάρκεια της εκπόνησης της.

Επίσης, θα ήθελα να ευχαριστήσω εξίσου θερμά τον Δρ. Χρήστο Κατρακάζα για την πολύτιμη βοήθεια και την καθοριστική συμβολή του σε όλα τα στάδια ολοκλήρωσης της εργασίας καθώς και για το εξαιρετικό κλίμα συνεργασίας που διαμόρφωσε. Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου και τους φίλους μου για την υποστήριξη που μου προσέφεραν καθ' όλη τη διάρκεια των σπουδών μου.

Αθήνα, Ιούλιος 2022

Έκτορας Καμβουσιώρας

ΕΝΤΟΠΙΣΜΟΣ ΕΠΙΚΙΝΔΥΝΗΣ ΣΥΜΠΕΡΙΦΟΡΑΣ ΟΔΗΓΟΥ ΜΕ ΔΕΔΟΜΕΝΑ ΕΥΡΕΙΑΣ ΚΛΙΜΑΚΑΣ ΑΠΟ ΕΞΥΠΝΑ ΣΥΣΤΗΜΑΤΑ ΚΑΤΑΓΡΑΦΗΣ ΚΑΙ ΤΕΧΝΙΚΕΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Έκτορας Καμβουσιώρας

Επιβλέπων: Γιώργος Γιαννής Καθηγητής Ε.Μ.Π

ΣΥΝΟΨΗ

Στόχος της παρούσας διπλωματικής εργασίας είναι ο εντοπισμός επικίνδυνης συμπεριφοράς οδηγού με δεδομένα ευρείας κλίμακας από έξυπνα συστήματα καταγραφής και τεχνικές μηχανικής μάθησης. Η συλλογή των στοιχείων έγινε από μία μεγάλη βάση δεδομένων που δημιουργήθηκε μέσω ενός πειράματος προσομοίωσης που έγινε σε οδηγούς. Στη συνέχεια χωρίστηκαν τρεις κατηγορίες οδήγησης: η φυσιολογική οδήγηση, η επικίνδυνη οδήγηση και η οδήγηση λίγο πριν το ατύχημα, χρησιμοποιώντας σαν κύρια μεταβλητή την μέγιστη ταχύτητα και ελέγχοντας αν οι οδηγοί ξεπερνούν το όριο ταχύτητας που υπάρχει μέσω αυτής. Επιπλέον η πλειονότητα των μελετών είχε πρόβλημα ανισορροπίας του δείγματος όσον αφορά τις διαφορετικές ταξινομήσεις, με τα δείγματα των επικίνδυνων οδηγικών καταστάσεων να είναι πολύ μικρότερα από τα δείγματα αυτά των ασφαλών οδηγικών συνθηκών. Γι' αυτό χρησιμοποιείται η μέθοδος επαναδειγματοληψίας SMOTE για την επίλυση της ανισορροπίας των δεδομένων στα επίπεδα ασφαλείας καθώς και για τη διασφάλιση της αμεροληψίας των μοντέλων. Για την ανάλυση των δεδομένων αναπτύχθηκαν μοντέλα Ridge Classifier, Support-vector machine, random forests και XgBoost. Σύμφωνα με τα αποτελέσματα τους τα μοντέλα random forests και XgBoost παρουσίασαν τα πιο αξιόπιστα αποτελέσματα στην ικανότητα πρόβλεψής με 95% ακρίβεια των τριών κατηγοριών οδηγών με χαμηλότερη πιθανότητα λάθους πρόβλεψης, συγκριτικά με τα Ridge Classifier και Support-vector machine. Στη συνέχεια για την καλύτερη κατανόηση αυτών των μοντέλων βρέθηκαν οι τιμές Shapley όπου μας έδειξαν της ποιο σημαντικές μεταβλητές που επηρεάζουν το κάθε μοντέλο. Τέλος, γίνονται προτάσεις για αξιοποίηση των αποτελεσμάτων, καθώς και για περαιτέρω έρευνα του αντικειμένου.

Λέξεις κλειδιά: εντοπισμός επικίνδυνης συμπεριφοράς, μέθοδος παλινδρόμησης κορυφογραμμής, μηχανής διανυσματικής υποστήριξης, μοντέλο τυχαίων δασών, μοντέλο λογισμικού ανοιχτού κώδικα, φυσιολογική οδήγηση, επικίνδυνη οδήγηση, οδήγηση λίγο πριν το ατύχημα.

DETECTION OF DANGEROUS DRIVER BEHAVIOR WITH WIDE-SCALE DATA FROM SMART RECORDING SYSTEMS AND MACHINE LEARNING TECHNIQUES

Hector Kamvoussioras

Supervisor: George Yannis, Professor N.T.U.A

ABSTRACT

The aim of this thesis is to identify dangerous driver behavior using large-scale data from intelligent sensor systems and machine learning techniques. The data was collected from a large database created through a simulation experiment conducted within the European H2020 project. Three categories of driving were extracted from the data; normal driving, dangerous driving and avoidable accident. The three categories were extracted using maximum speed as the concerned variable and checking whether drivers exceeded the speed limit through it. In addition, the majority of studies had a sample imbalance problem, with the samples of dangerous driving conditions being much smaller than those of safe driving conditions. Therefore, the SMOTE resampling method was used to resolve the imbalance of the data in the safety levels as well as to ensure the impartiality of the models. Ridge classifier, support-vector machine, random forests and XgBoost models were developed for data analysis. According to their results, the random forests and XgBoost models showed the most reliable results in the prediction ability with 95% accuracy of the three driver categories with lower probability of prediction error, compared to Ridge Classifier and Support-vector machine. Then to better understand these models, Shapley values were found where they showed us the most important variables affecting each model. Finally, suggestions are made to utilize the results and to further research the subject.

Keyword's: detection of dangerous behavior, Ridge Classifier, Support-vector machine, Random forest, XgBoost normal driving, dangerous driving, Avoidable Accident.

ΠΕΡΙΛΗΨΗ

Στόχος της παρούσας διπλωματικής εργασίας είναι ο εντοπισμός επικίνδυνης συμπεριφοράς οδηγού με δεδομένα ευρείας κλίμακας από έξυπνα συστήματα καταγραφής και τεχνικές μηχανικής μάθησης. Για το λόγο αυτό αναπτύχθηκαν μοντέλα στατιστικής ανάλυσης και διερευνήθηκε η ικανότητά τους να προβλέπουν 3 κατηγόριες οδηγών.

Τα δεδομένα προέκυψαν από το ερευνητικό έργο i-DREAMS, στο οποίο συμμετείχαν 36 οδηγοί 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. Από τα δεδομένα αυτά δημιουργήθηκε ένας πίνακας όπου περιέχει στοιχεία για την κατάσταση του οδηγού και την κατάσταση του αυτοκίνητου κατά την διάρκεια της οδήγησης. Οι πίνακες αυτοί χρησιμοποιήθηκαν στην στατιστική ανάλυση των δεδομένων και, χρησιμοποιώντας ως εξαρτημένη μεταβλητή το Speed_max, ξεχωρίστηκαν τρεις κατηγορίες:

- Φυσιολογική Οδήγηση (class: 0):
  Μέγιστη Ταχύτητα <= 0,8* Τρέχον όριο ταχύτητας
- Επικίνδυνη Οδήγηση (class: 1):
  0,8* Τρέχον όριο ταχύτητας <= Μέγιστη Ταχύτητα <= Τρέχον όριο ταχύτητας
- Οδήγηση Αποφεύγοντας Ατύχημα (class: 2):
  Μέγιστη Ταχύτητα => Τρέχον όριο ταχύτητας

Αρχικά για την ανάλυση των δεδομένων και για την εύρεση των σημαντικότερων μεταβλητών χρησιμοποιήθηκαν κατάλληλες μέθοδοί. Στη συνέχεια διαχωρίστηκαν οι μεταβλητές με την μεγαλύτερη συσχέτιση και χωρίστηκαν σε δύο ομάδες. Η (A) στην οποία περιέχονται μόνο οι μεταβλητές με την μεγαλύτερη συσχέτιση και η (B) στην οποία εμπεριέχονται επιπλέον και μεταβλητές οι οποίες θεωρούνται σημαντικές αλλά όχι στον ίδιο βαθμό με τις άλλες.

(A) [TTC_mean, Headway_std, Speed_std, ME_ForwardCollisionWarning_mean]

(B) [TTC_mean, Headway_median, HandsOnEvent_mean, FatigueEvent_median, ME_LaneDepartureWarningActive_mean, Speed_mean_1, Distance_travelled_sum]

Επισημαίνεται ότι η μεταβλητή Speed_max δεν λήφθηκε υπόψη στο πρώτο μέρος των αναλύσεων, καθώς θα αναπτύσσονταν προβλήματα μεροληψίας των μοντέλων ταξινόμησης.

Στη συνέχεια εφαρμόζοντας την SMOTE τεχνική (μία συνθετική τεχνική υπερδειγματοληψίας μειοψηφίας), η οποία μέσω της βιβλιογραφικής ανασκόπησης έδειξε τα καλύτερα αποτελέσματα, επιλύθηκε το πρόβλημα άνισης κατανομής των δεδομένων εκπαίδευσης στις διαφορετικές κλάσεις.

Αναπτύχθηκαν τέσσερεις αλγόριθμοι μηχανικής εκμάθησης με σκοπό την ταξινόμηση των οδηγών σε μία από τις τρεις κατηγορίες. Τα ονόματα και οι συμβολισμοί των τεσσάρων αλγορίθμων παρατίθενται στον πίνακα που ακολουθεί.

Όνομα μοντέλου (ελληνικά)	**Όνομα μοντέλου (αγγλικά)**	**Συμβολισμός μοντέλου**
μέθοδος παλινδρόμησης κορυφογραμμής	RidgeClassifier	RID
μηχανής διανυσματικής υποστήριξης	SupportVectorMachines	SVM
μοντέλο τυχαίων δασών	RandomForestClassifier	RF
μοντέλο ακραίας ενίσχυση κλίσης	XgBoost	XG

Ελέγχθηκαν και αξιολογήθηκαν τα αποτελέσματα των καλύτερων μοντέλων για την ομάδα (B) όπως φαίνεται στους παρακάτω πίνακες.

Για την ομάδα (B)

Πίνακας : Σύνοψη μοντέλου RandomForestClassifier και XgBoost

RandomForestClassifier	Ορθότητα	Ανάκληση	F1-score	false alarm rate	G-means
Μέσος όρος	92%	90%	91%	7%	95%
Σταθμισμένος Μέσος όρος	94%	94%	94%	5%	95%
XgBoost					
Μέσος όρος	91%	93%	92%	7%	95%
Σταθμισμένος Μέσος όρος	95%	95%	95%	5%	96%

Πίνακας ποσοστά ακρίβειας μοντέλων

	Ακρίβεια
RandomForestClassifier	95%
XgBoost	95%

Τέλος, για να έχουμε μια επισκόπηση των χαρακτηριστικών που είναι πιο σημαντικά για ένα μοντέλο, μπορούμε να χρησιμοποιήσουμε τις τιμές SHAP για κάθε χαρακτηριστικό, για κάθε δείγμα. Ειδικότερα προέκυψε για την ομάδα (B) ότι η πιο σημαντικές μεταβλητές που επηρεάζουν το μοντέλο RandomForestClassifier είναι η ταχύτητα, η απόσταση από το μπροστά αμάξι, ο χρόνος πρόσκρουσης και η ένδειξη

ότι τα χέρια του οδηγού βρίσκονται στο τιμόνι. Ενώ για το XgBoost σημαντικότερα είναι η ταχύτητα, ο χρόνος πρόσκρουσης, η απόσταση που διένυσε, η απόσταση από το μπροστά αμάξι και η ένδειξη ότι τα χέρια του οδηγού βρίσκονται στο τιμόνι. Συμπερασματικά παρατηρούμε πως και στα δύο μοντέλα η ταχύτητα είναι η ποιο σημαντική μεταβλητή αλλά στην συνέχεια βλέπουμε πως οι υπόλοιπες επηρεάζουν με διαφορετικό βαθμό το κάθε μοντέλο.

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

- Η μέθοδος Random Forest και η μέθοδος XgBoost από την ομάδα (B) σημείωσαν τις υψηλότερες επιδόσεις στην πλειοψηφία των μετρικών αξιολόγησης τους, με ακρίβεια 95%.
- Σημαντικό να αναφερθεί είναι το ότι βρέθηκε recall score 90% για το μοντέλο Random Forest και 94% για το μοντέλο XgBoost, με αποτέλεσμα τον καλύτερο εντοπισμό των τριών κατηγοριών συμπεριφοράς οδήγησης.
- Οι πιο σημαντικές μεταβλητές που ωθούν τον οδηγό στην πιο επικίνδυνη οδήγηση είναι η ταχύτητα, ο χρόνος απόστασης από το επόμενο όχημα και το χρονικό διάστημα που έχει ο οδηγός τα χέρια του στο τιμόνι.
- 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 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).
- Η μέθοδοι Ridge Classifier και Support Vector Machines δεν παρουσίασαν τόσο ικανοποιητικά αποτελέσματα όσο τα υπόλοιπα μοντέλα
- Και στα τέσσερα μοντέλα η ομάδα με τον μεγαλύτερο αριθμό μεταβλητών παρουσίασε πιο ικανοποιητικά αποτελέσματα.
- Τέλος από την εκπόνηση της συγκεκριμένης Διπλωματικής Εργασίας, προκύπτει ότι τα δεδομένα που συλλέγονται από τα έξυπνα συστήματα και περεταίρω έρευνες περιέχουν ιδιαίτερα σημαντικές πληροφορίες οι οποίες, μετά από κατάλληλη επεξεργασία και ανάπτυξη μαθηματικών μοντέλων, μπορούν να χρησιμεύσουν στην εξαγωγή χρήσιμων συμπερασμάτων για τις κρίσιμες παραμέτρους που επηρεάζουν την συμπεριφορά του οδηγού κατά τη διάρκεια οδήγησης αλλά και για τη γενικότερη κυκλοφοριακή συμπεριφορά των οδηγών

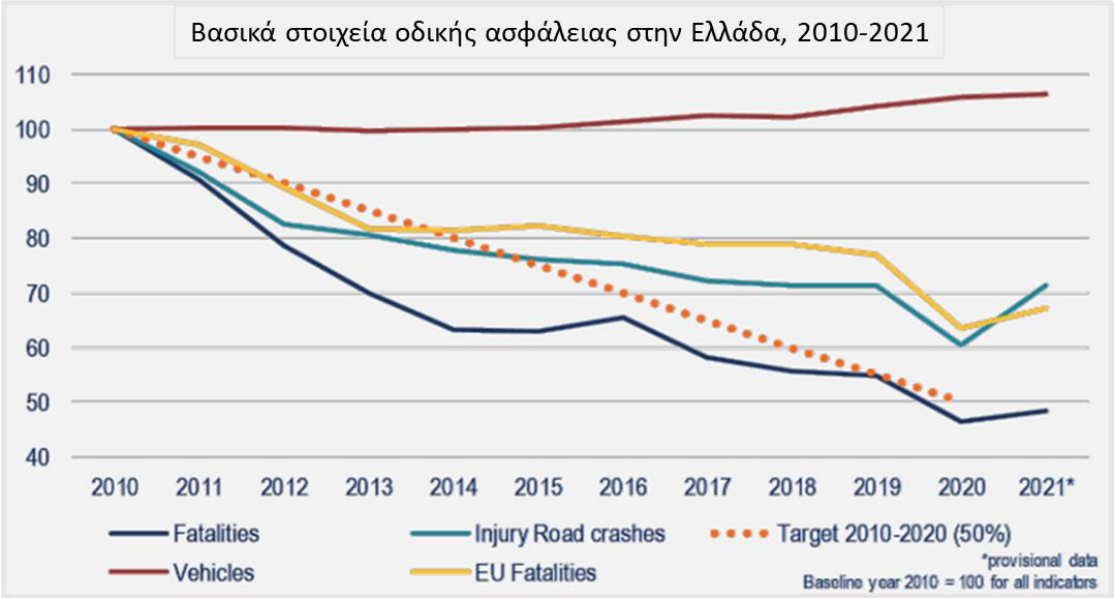
ΠΕΡΙΕΧΟΜΕΝΑ

1. ΕΙΣΑΓΩΓΗ

1.1 Γενική ανασκόπηση

Όντας σύνθετη και πολυπαραγοντική, η οδική ασφάλεια αποτελεί διαχρονικό μέλημα του κράτους, της κοινωνίας, της οικονομίας και πολλών άλλων τομέων. Τα οδικά ατυχήματα αποτελούν σήμερα την κύρια αιτία θνησιμότητας.

Η Ευρώπη έχει καταβάλει τεράστιες προσπάθειες τα τελευταία χρόνια για να μειώσει τον αριθμό των θανάτων από τροχαία ατυχήματα κατά 43% μεταξύ 2001 και 2010 και κατά 21% μεταξύ 2010 και 2010. Η Ελλάδα, ειδικότερα, κατάφερε να μειώσει τα τροχαία ατυχήματα κατά 51% μεταξύ 2010 και 2020, ποσοστό που αποτελεί τη μεγαλύτερη μείωση μεταξύ των κρατών μελών της ΕΕ. (γράφημα 1.1). (Kallidoni et al 2021)

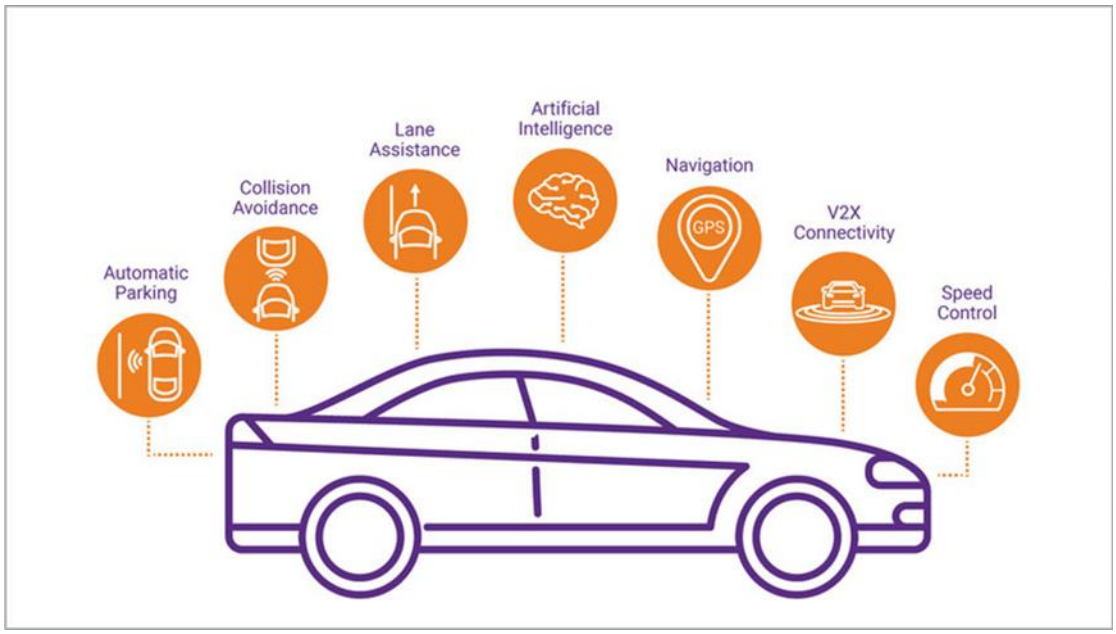


Γράφημα 1.1: Βασικά στοιχεία οδικής ασφάλειας στην Ελλάδα, 2010-2021
Πηγή: NTUA Road Safety Observatory (2022)

Η Ευρωπαϊκή Ένωση, καθώς και ο Παγκόσμιος Οργανισμός Υγείας σε συνεργασία με τα Ηνωμένα Έθνη έχουν θέσει στόχους για τη μείωση των θανάτων από τροχαία ατυχήματα στην ΕΕ κατά 50%. Για την επίτευξη του συγκεκριμένου στόχου, δίνεται ιδιαίτερη βαρύτητα στη συμβολή των νέων τεχνολογιών της αυτοκινητοβιομηχανίας και του αυτοματισμού στις μεταφορές, με στόχο την ενίσχυση της οδικής ασφάλειας. (Kallidoni et al 2021)

Μία τέτοια νέα τεχνολογία είναι η ADAS (advanced driver-assistance systems δηλαδή "Προηγμένα συστήματα υποβοήθησης οδηγού"), ένας όρος που καλύπτει μια ποικιλία τεχνολογιών της αυτοκινητοβιομηχανίας που μπορεί να προειδοποιεί για κινδύνους και ακόμη και να αυτοματοποιεί ορισμένες λειτουργίες οδήγησης, όπως

φαίνονται στην εικόνα 1.2. Τέτοιες πράξεις μπορούν να βελτιώσουν την ασφάλεια, ενώ παράλληλα καθιστούν την οδήγηση απλούστερη σε ορισμένες περιπτώσεις.



Εικόνα 1.2:Έξυπνα σύστημα και μηχανισμοί για την οδήγηση
Πηγή: Fred Meier (2021)

Είναι πιο πιθανό να αναγνωρίσουμε το αυξημένο όφελος αυτού του συστήματος. Ορισμένα από τα χαρακτηριστικά, όπως ο προσαρμοστικός έλεγχος ταχύτητας, καθιστούν τα καθήκοντα οδήγησης ευκολότερα και λιγότερο κουραστικά, αλλά ο κύριος στόχος αυτών των τεχνολογιών είναι η βελτίωση της ασφάλειας μέσω της μείωσης των λαθών του οδηγού- άλλωστε, σύμφωνα με την Εθνική Διοίκηση Οδικής Ασφάλειας, το ανθρώπινο λάθος ευθύνεται για το 94% των σοβαρών ατυχημάτων. Η τεχνολογία ADAS προσπαθεί να αποτρέψει ή, σε ορισμένες περιπτώσεις, να μειώσει τη σοβαρότητα των συγκρούσεων που δεν μπορούν να αποφευχθούν. Αισθητήρες όπως κάμερες, ραντάρ και τεχνολογίες υπερήχων χρησιμοποιούνται για τη συλλογή δεδομένων. Ορισμένα συστήματα χρησιμοποιούν επιπλέον δεδομένα GPS ή χαρτογράφησης για να "βλέπουν" το δρόμο και την επιθυμητή διαδρομή. Το ADAS μπορεί στη συνέχεια να προειδοποιεί τον οδηγό για πιθανούς κινδύνους και να επεμβαίνει αυτόματα για την αποφυγή συγκρούσεων.(Meier et al 2022)

Επιπλέον έχει αναπτυχθεί και η έρευνα με πειράματα οδήγησης σε φυσιολογικές συνθήκες, όπου τα ίδια τα αυτοκίνητα των συμμετεχόντων είναι εφοδιασμένα με εξοπλισμό που καταγράφει συνεχώς διάφορα στοιχεία της οδηγικής τους συμπεριφοράς με τρόπο δυσδιάκριτο και χωρίς την παρουσία επόπτη δοκιμής για μεγάλο χρονικό διάστημα. Περιλαμβάνονται πτυχές της κίνησης του οχήματος, της συμπεριφοράς του οδηγού και του άμεσου περιβάλλοντος. Οι σταθερές κάμερες που

βρίσκονται σε χώρο μπορούν να χρησιμοποιηθούν για τη διεξαγωγή φυσιοκρατικών παρατηρήσεων πεζών και ποδηλατών.

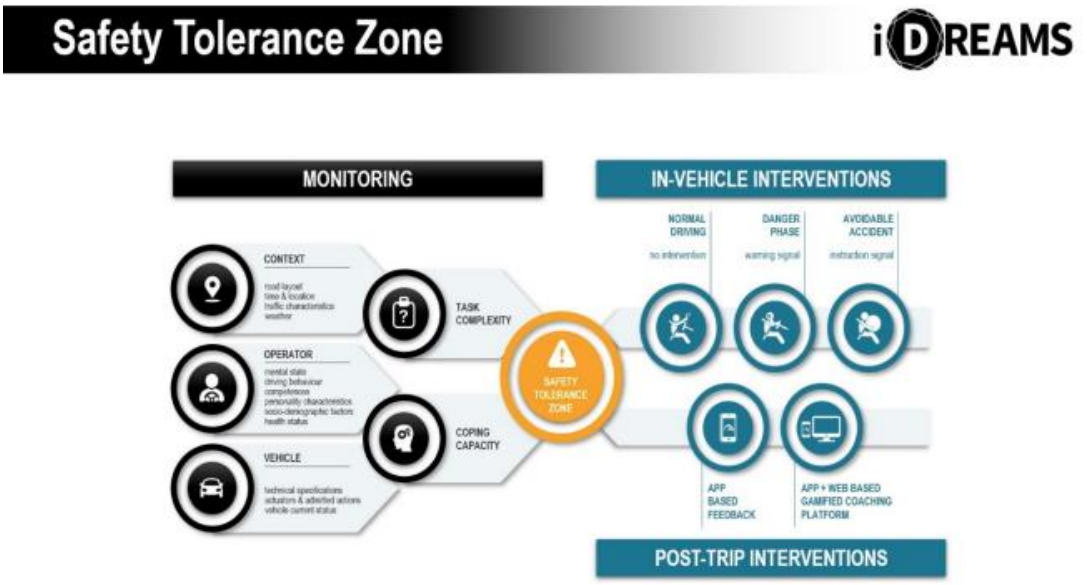
Οι παρατηρήσεις από πειράματα οδήγησης σε φυσιολογικές συνθήκες παρέχουν δεδομένα που είναι δύσκολο, αν όχι αδύνατο, να συγκεντρωθούν μέσω των παραδοσιακών ερευνητικών προσεγγίσεων. Για παράδειγμα, η ανάλυση δεδομένων ατυχημάτων και οι διεξοδικές έρευνες ατυχημάτων δεν μπορούν να αποκαλύψουν πολλά για τις δυσκολίες συμπεριφοράς πριν από ένα ατύχημα ή παραλίγο ατύχημα. Επειδή οι συμμετέχοντες στη δοκιμή έχουν συνήθως επίγνωση των πειραματικών ρυθμίσεων, οι παρατηρήσεις με τη χρήση οχημάτων με όργανα ή προσομοιώσεων δεν τους ενθαρρύνουν να ενεργήσουν με φυσιολογικό (naturalistic) τρόπο.(Winkelbauer et al 2022).

Η Ευρωπαϊκή Επιτροπή υποστηρίζει το ερευνητικό έργο i-DREAMS (2022) (https://idreamsproject.eu/) στο πλαίσιο του προγράμματος-πλαισίου για την έρευνα στον τομέα των μεταφορών "Horizons 2020". Στόχος της έρευνας αυτής είναι ο εντοπισμός επικίνδυνης συμπεριφοράς οδηγών, περιλαμβάνοντας διάφορες κατηγορίες οδήγησης. Θα είναι δυνατόν να ανιχνεύεται το επίπεδο στο οποίο βρίσκεται κάθε οδηγός και να σχεδιάζονται παρεμβάσεις για να αποτρέπεται η παρέκκλιση από την ασφαλή οδήγηση με τη χρήση ενός έξυπνου συστήματος παρακολούθησης της οδήγησης και των περιβαλλοντικών παραγόντων. Οι παρεμβάσεις θα πραγματοποιηθούν σε δύο στάδια. Το πρώτο μέρος λαμβάνει χώρα σε πραγματικό χρόνο, δηλαδή ενώ ο οδηγός οδηγεί, με στόχο να επιτρέψει στον οδηγό να λάβει αμέσως τα απαραίτητα μέτρα, και το δεύτερο στάδιο λαμβάνει χώρα αργότερα, με στόχο να βελτιώσει τις γνώσεις του οδηγού και συνεπώς τη συμπεριφορά του.

Η "Συμπεριφορά οδήγησης"χωρίζεται σε 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)

Τα τελευταία χρόνια, υπάρχει μεγάλο ενδιαφέρον για την ανάλυση της συμπεριφοράς των οδηγών με τη χρήση μηχανικών αλγορίθμων και αλγορίθμων μάθησης (Peppesetal., 2021). Επιπλέον, οι Michelaraki et al. (2021b) διαπίστωσαν ότι η χρήση ευφυών συστημάτων παρακολούθησης της συμπεριφοράς του οδηγού με στόχο παρεμβάσεις σε πραγματικό χρόνο είναι ιδιαίτερα επιτυχής στην ελαχιστοποίηση των ατυχημάτων. Η αναγκαιότητα κατασκευής τέτοιων συστημάτων με στόχο την αύξηση της οδικής ασφάλειας καθιστά αναγκαίο τον προσδιορισμό της επίδρασης διαφόρων μεταβλητών κινδύνου 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

Στόχος της παρούσας διπλωματικής εργασίας είναι ο εντοπισμός επικίνδυνης συμπεριφοράς οδηγού με δεδομένα ευρείας κλίμακας από έξυπνα συστήματα καταγραφής και τεχνικές μηχανικής μάθησης.

Καθώς η ταχύτητα έχει σημαντικό αντίκτυπο στην πιθανότητα ατυχήματος και στη σοβαρότητα του ατυχήματος, είναι ίσως ο πιο κρίσιμος παράγοντας κινδύνου. Έχει προσδιοριστεί ότι κάθε χιλιόμετρο αυξημένης μέσης ταχύτητας έχει ως αποτέλεσμα την αύξηση κατά 3% της πιθανότητας εμπλοκής σε ατύχημα. Η υπερβολική (πάνω

από το σχετικό όριο) και η ακατάλληλη (κάτω από το ισχύον όριο αλλά μη αποδεκτή για τις συνθήκες) ταχύτητα συμβάλλουν αμφότερες στα οδικά ατυχήματα και ευθύνονται για σημαντικό αριθμό θανάτων. Ο έλεγχος της ταχύτητας των οχημάτων μπορεί να συμβάλλει στην πρόληψη των ατυχημάτων και στην ελαχιστοποίηση του αριθμού των ανθρώπων που σκοτώνονται σε αυτά. Έτσι, όταν συμβαίνουν ατυχήματα, αυτά έχουν μικρότερες επιπτώσεις, με αποτέλεσμα λιγότερους σοβαρούς τραυματισμούς. Για αυτό επιλέχθηκε η μέγιστη ταχύτητα σαν εξαρτημένη μεταβλητή ώστε να προσδιοριστεί το επίπεδο οδικής ασφάλειας.

Οπότε, για να επιτευχθεί ο στόχος μας, δημιουργήσαμε μοντέλα κατηγοριοποίησης προκειμένου να εκτιμηθεί σε ποιο σημείο κάθε οδηγός εμπίπτει στις κατηγορίες που θέσαμε για την "Συμπεριφορά οδήγησης". Άρα, με τη χρήση των οδηγικών χαρακτηριστικών κάθε οδηγού καθώς και του σχετικού περιβάλλοντος οδήγησης ως εισροών, μπορέσαμε να κατατάξουμε τους οδηγούς σε μία από αυτές. Η κατηγοριοποίηση με μηχανική μάθηση είναι μια κρίσιμη τεχνική για την αναγνώριση της οδηγικής συμπεριφοράς και, κατά συνέπεια, για τη βελτίωση της οδικής ασφάλειας.

Στη συνέχεια εξετάστηκαν με κατάλληλους τρόπους τα αποτελέσματα και η ευστοχία των προβλέψεων και εξερευνήθηκε το καλύτερο σενάριο για τον καλύτερο εντοπισμό αυτών των κατηγοριών.

Η παρούσα έρευνα θα έχει διττή συμβολή, επιχειρώντας να συνεισφέρει και να ενισχύσει την έρευνα στον τομέα της ανάλυσης της οδήγησης και μέσω των αποτελεσμάτων αυτών μπορούν να προκύψουν προτάσεις για τη βελτίωση της οδικής ασφάλειας, καθώς και για περαιτέρω έρευνα του εν λόγω αντικειμένου.

1.3 Μεθοδολογία

Παρακάτω περιγράφεται η μεθοδολογία που ακολουθήθηκε για την επίτευξη του στόχου της διπλωματικής εργασίας.

Πρώτο βήμα αποτελεί ο καθορισμός του θέματος της μελέτης και του στόχου της. Έπειτα απαιτείται αναζήτηση συναφών ερευνών και μεθοδολογιών ανάλυσης στη διεθνή βιβλιογραφία και προσδιορισμός ζητημάτων που απαιτούν ανάλυση και έρευνα, που οδηγούν στην οριστικοποίηση του στόχου και στον τρόπο ανάλυσης των δεδομένων

Στη 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, τη βιβλιοθήκη ανάλυσης δεδομένων pandas την βιβλιοθήκη Επεξηγήσεις πρόσθετων SHapley (SHapley Additive Explanations) και την βιβλιοθήκη XgBoost.

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. Στη 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 προτάσεις

που μπορούν 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

Στόχος της βιβλιογραφικής ανασκόπησης είναι ο καθορισμός του αντικειμένου της διπλωματικής εργασίας και η εύρεση της κατάλληλης μεθοδολογίας που θα ακολουθηθεί για την εκπόνησή της. Αποτελεί βάση πάνω στην οποία δομείται η εργασία, καθώς παρουσιάζονται συναφείς έρευνες και μεθοδολογίες, ώστε να προκύψουν βασικά συμπεράσματα σύμφωνα με τα οποία θα επιλεχθεί η ερευνητική διαδρομή της παρούσας διπλωματικής εργασίας. Συγκεκριμένα, η δημοσιευμένη έρευνα στη διεθνή βιβλιογραφία που επικεντρώνεται στην ανάλυση, τον εντοπισμό και την οδηγική συμπεριφορά, καθώς και στην πρόβλεψη ατυχημάτων σε πραγματικές συνθήκες.

Από την σύγκριση και κριτική των ερευνών αυτών θα προκύψει ο στόχος της παρούσας μελέτης, καθώς και οι κατάλληλες μέθοδοι για την επίτευξή του.

2.2 Συναφείς έρευνες και μεθοδολογίες

Ο αριθμός των θανατηφόρων τροχαίων ατυχημάτων παγκοσμίως αυξάνεται παράλληλα με την αύξηση της οδικής κυκλοφορίας και την αποδοχή των ιδιωτικών αυτοκινήτων. Σύμφωνα με την Πράξη World Health Organization (WHO's) Road Road Injury Act, 1,35 εκατομμύρια άνθρωποι πεθαίνουν σε τροχαία ατυχήματα παγκοσμίως κάθε χρόνο, κοστίζοντας στα περισσότερα έθνη το 3% του GDP τους. (Farrag et al. 2020)

Για αυτό έχουν γίνει εκτενώς έρευνες και χρήση ειδικότερα σήμερα μοντέλων προσομοίωσης σε όλων τον κόσμο για την αξιολόγηση της απόδοσης διαφόρων κυκλοφοριακών εγκαταστάσεων και στρατηγικών διαχείρισης για αποδοτικά και βιώσιμα συστήματα μεταφορών. Ένας από τους βασικούς παράγοντες για τη διασφάλιση της αξιοπιστίας των μοντέλων όσον αφορά την αντανάκλαση των τοπικών συνθηκών είναι η βαθμονόμηση και η επικύρωση τους.

Επίσης είναι εξίσου σημαντικό να μπορούμε να αναγνωρίσουμε την οδική συμπεριφορά του οδηγού και να προβλέψουμε σε πραγματικό χρόνο την πιθανότητα ατυχήματος. Σε αυτή την προσπάθεια συνεισφέρουν σημαντικά η ανάπτυξη νέων τεχνολογιών όπως το "Προηγμένα συστήματα υποβοήθησης οδηγού" ή ADAS και τα πειράματα οδήγησης σε φυσιολογικές συνθήκες.

Η κατανόηση της επίδρασης των διαφόρων χαρακτηριστικών στην ανασφαλή οδηγική συμπεριφορά αποτελεί αντικείμενο πολλαπλών δημοσιευμένων ερευνών για την επίτευξη των προαναφερθέντων στόχων και την πρόοδο του επαγγέλματος.

Με την εξέταση της επιρροής, μπορούν να αναπτυχθούν σχετικά μοντέλα για τον εντοπισμό επικίνδυνων οδηγικών συμπεριφορών, τα οποία θα αυξήσουν την αποτελεσματικότητα των συστημάτων υποβοήθησης του οδηγού. Επιπλέον θα δώσουμε ιδιαίτερη σημασία στην κατανόηση αυτών τον μοντέλων και στην σημαντικότητα των μεταβλητών που χρησιμοποιούνται για να μπορέσουμε να βρούμε το τι οδηγεί τον οδηγό σε ποιο επικίνδυνη οδήγηση. Με αποτέλεσμα την ώθησε σε ποιο ασφαλή οδήγηση μέσο ποιο συγκεκριμένων οδηγιών και αλλαγών στους δρόμους και στην συμπεριφοράς του οδηγού. (Yu et al. 2021)

2.3 Αλγόριθμοι ταξινόμησης οδηγικής συμπεριφοράς

Τις τελευταίες δεκαετίες, τα οχήματα είναι εξοπλισμένα με πληθώρα αισθητήρων που μπορούν να παρέχουν χρήσιμες μετρήσεις και διαγνωστικά στοιχεία τόσο για την κατάσταση του οχήματος όσο και για τη συμπεριφορά του οδηγού. Επιπλέον, η ραγδαία αύξηση των αναγκών μεταφοράς ανθρώπων και αγαθών σε συνδυασμό με την εξέλιξη των τεχνολογιών πληροφορικής και επικοινωνιών (ΤΠΕ) ωθούν τον τομέα των μεταφορών προς μια νέα πιο έξυπνη και αποτελεσματική εποχή.(Peppes, Nikolaos, et al. 2021)

Η προσέγγιση που παρουσιάζεται στην μελέτη των Martinez et al, (2021) περιγράφει μια ολιστική ολοκληρωμένη πλατφόρμα η οποία συνδυάζει γνωστούς αλγορίθμους μηχανικής και βαθιάς μάθησης μαζί με εργαλεία που βασίζονται σε ανοιχτό κώδικα, προκειμένου να συλλέγει, να αποθηκεύει, να επεξεργάζεται, να αναλύει και να συσχετίζει διάφορες ροές δεδομένων που προέρχονται από οχήματα. Ειδικότερα, τα δεδομένα που ρέουν από διαφορετικά οχήματα επεξεργάζονται και αναλύονται με τη χρήση τεχνικών ομαδοποίησης προκειμένου να ταξινομηθεί η συμπεριφορά του οδηγού ως φιλική προς το περιβάλλον ή όχι, ενώ ακολουθεί συγκριτική ανάλυση των αλγορίθμων μηχανικής και βαθιάς μάθησης με επίβλεψη στο δεδομένο σύνολο δεδομένων με ετικέτες.

Η έρευνα των Tie-Qiao Tang και Zhi-Yan Yi (Tang & Yi 2017) προτείνει ένα μοντέλο παρακολούθησης αυτοκινήτων για τη διερεύνηση των επιπτώσεων του φωτεινού σηματοδότη στην οδηγική συμπεριφορά, την κατανάλωση καυσίμων και τις εκπομπές κατά τη διάρκεια ολόκληρης της διαδικασίας που κάθε όχημα διατρέχει τη διασταύρωση. Ειδικότερα, το προτεινόμενο μοντέλο έχει εξετάσει ρητά τις συμπεριφορές σε μια διασταύρωση με συσκευή αντίστροφης μέτρησης που παρέχει στιγμιαία πληροφόρηση στους οδηγούς. Το προτεινόμενο μοντέλο δοκιμάζεται με αριθμητική ανάλυση και τα αποτελέσματα δείχνουν ότι το μοντέλο μπορεί να βελτιώσει τη λειτουργική αποδοτικότητα και την ασφάλεια της κυκλοφορίας κοντά στη διασταύρωση, καθώς και να μειώσει τη μέση κατανάλωση καυσίμων των

οχημάτων. Η ανάλυση ευαισθησίας δείχνει ότι η αρχική χρονική απόσταση των οχημάτων στην αφετηρία της οδού μπορεί να έχει σημαντικές επιδράσεις στη χωρητικότητα ροής και στη συνολική κατανάλωση καυσίμου.

Στην έρευνα των Lu Yue και Xinsha Fu (Yue et al, 2021) σχεδιάζεται μια πραγματική εργασία οδήγησης για την εξαγωγή δεδομένων και προτείνει ένα σύστημα παρακολούθησης του οδηγικού στρες του οδηγού μοντέλο με βάση την οδηγική συμπεριφορά, το περιβάλλον οδήγησης και την εξοικείωση με τη διαδρομή. Η οδηγική συμπεριφορά είναι περιγράφεται από την ταχύτητα και την επιτάχυνση του οχήματος και το περιβάλλον οδήγησης ποσοτικοποιείται από ένα διευρυμένο μοντέλο υπολειμματικών δικτύων (DRN) που διαιρεί την εικόνα βίντεο από την πλήρη περιοχή σε υποπεριοχές. σύμφωνα με την κατανομή της προσοχής του οδηγού. Με βάση τα ψυχολογικά δεδομένα και το στρες του οδηγού απογραφής (DSI), η μελέτη χρησιμοποίησε μια τρισδιάστατη ανάλυση συστάδων K-means για να αποκτήσει τη μέθοδο αξιολόγησης των του οδηγικού στρες και κατασκεύασε ένα μοντέλο ενίσχυσης ακραίας κλίσης (XGBoost) για την παρακολούθηση του οδηγικού στρες. Οι συγκρίσεις των επιδόσεων με άλλα μοντέλα δείχνουν ότι το μοντέλο XGBoost υπερτερεί σημαντικά σε σχέση με τους άλλους τρεις κύριους αλγορίθμους μηχανικής μάθησης και υπερβαίνει τα περισσότερα παραδοσιακά μοντέλα χωρίς χρήση ψυχολογικών δεδομένων.

Στην έρευνα των Meiring, et al, (2015) διερευνώνται οι διάφορες λύσεις ανάλυσης του τρόπου οδήγησης. Πραγματοποιείται διεξοδική έρευνα για τον εντοπισμό των σχετικών αλγορίθμων μηχανικής μάθησης και τεχνητής νοημοσύνης που χρησιμοποιούνται στα τρέχοντα συστήματα ανάλυσης της συμπεριφοράς του οδηγού και του τρόπου οδήγησης. Συνεπώς, η παρούσα ανασκόπηση χρησιμεύει ως πλούτος πληροφοριών και θα ενημερώσει τον ειδικό και τον σπουδαστή σχετικά με την τρέχουσα κατάσταση της τεχνολογίας στα συστήματα ανάλυσης του τρόπου οδήγησης, την εφαρμογή αυτών των συστημάτων και τους υποκείμενους αλγορίθμους τεχνητής νοημοσύνης που εφαρμόζονται σε αυτές τις εφαρμογές. Σκοπός της έρευνας είναι η αξιολόγηση των δυνατοτήτων για μοναδική αναγνώριση του οδηγού με τη χρήση των προσεγγίσεων που έχουν εντοπιστεί σε άλλες μελέτες συμπεριφοράς του οδηγού.

Στην έρευνα των Soni & Lanka, (2021) εξετάζονται τα οδικά ατυχήματα που συμβαίνουν στα σήματα τροχαίας, παρόλο που οι περιπολίες και η ευαισθητοποίηση της τροχαίας είναι αυξημένες. Ατυχήματα συνέβησαν σε διασταυρώσεις σηματοδοτών αποτελούν σημαντικό ποσοστό των συνολικών αναφερόμενων οδικών ατυχημάτων. Συνήθως, όταν οι οδηγοί πλησιάζουν τα σήματα τροχαίας, κατά την έναρξη του κίτρινου χρώματος, οι οδηγοί εισέρχονται σε μια ζώνη διλήμματος, όπου βρίσκονται σε σύγχυση. εκτιμώντας τις δυνατότητές τους να διασχίσουν τη διασταύρωση ή να σταματήσουν. Έτσι, οποιαδήποτε λανθασμένη απόφαση μπορεί

να οδηγήσει σε σύγκρουση. Στο αποφύγουν τη σύγκρουση σε ορθή γωνία, οι οδηγοί εφαρμόζουν σκληρό φρένο για να σταματήσουν πριν από τα σήματα του φαναριού. Αυτό όμως μπορεί να οδηγήσει σε σύγκρουση σε οπισθοπορεία, όταν ο επόμενος οδηγός συναντήσει την απόφαση του πρώτου να σταματήσει απότομα. Η κατάσταση αυτή γίνεται πολύπλευρη όταν το σήμα είναι ετερογενής και περιέχει διάφορους τύπους οχημάτων. Έτσι, ο κύριος στόχος της παρούσας μελέτης είναι να εκτιμηθεί η απόδοση με τη χρήση τεχνικές μηχανικής μάθησης και εφαρμόζονται για να επικυρωθεί η ταξινόμηση της οδηγικής συμπεριφοράς όσον αφορά την ασφαλή/μη ασφαλή στάση σε σηματοδοτημένες διασταυρώσεις κατά την έναρξη του κίτρινου σήματος.

Η έρευνα της Marchegiani (Marchegiani, 2018), αξιοποιεί τη βιβλιογραφία σε εργασίες ελέγχου ταυτότητας (π.χ. ομιλητής αναγνώρισης) και παρουσιάζει ένα πλαίσιο για την ταυτοποίηση οδηγών το οποίο χρησιμοποιεί μηχανή διανυσμάτων υποστήριξης (SVM) και καθολική Universal Background Model. Το πλαίσιό της λειτουργεί σε σήματα του πεντάλ γκαζιού και του πεντάλ φρένων και, ως εκ τούτου, ενισχύει άλλες τεχνολογίες, όπως μικρόφωνα ή κάμερες, εάν υπάρχουν. Επιπλέον, το πλαίσιό μας είναι συμβατό με οχήματα που περιορίζονται σε παραδοσιακές μεθόδους ανίχνευσης. Στον πίνακα 2.1 παρατίθενται οι αλγόριθμοι ταξινόμησης με την υψηλότερη επίδοση από τις έρευνες που παρουσιάστηκαν.

Πίνακας 2.1: Αποτελεσματικότεροι αλγόριθμοι ταξινόμησης ανά έρευνα ανάλυσης οδηγικής συμπεριφοράς

Έρευνα	Σκοπός αλγορίθμων ταξινόμηση	Αλγόριθμοι ταξινόμησης με το υψηλότερο ποσοστό ορθών προβλέψεων
Francisco J. Martinez (9 July 2021)	Συμπεριφορά του οδηγού ως φιλική προς το περιβάλλον ή όχι με βάση 3 κατηγορίες	Τυχαία δάση (RF): ποσοστό ορθών προβλέψεων 95%
Tie-Qiao Tang και Zhi-Yan Yi (20 September 2017)	Αναγνώριση της κατάστασης του οδηγού με βάση 3 κατηγορίες	Τυχαία δάση (RF): ποσοστό ορθών προβλέψεων 82%
Lu Yue και Xinsha Fu (29 January 2021)	Ενα σύστημα παρακολούθησης του οδηγικού στρες με βάση 3 κατηγορίες	Ποντέλο ενίσχυσης ακραίας κλίσης (XGBoost): ποσοστό ορθών προβλέψεων 95%
Marthinus Meiring, Gys Albertus, and Hermanus Carel Myburgh(4 December 2015)	Αναγνώριση της κατάστασης του οδηγού με βάση 4 κατηγορίες	Μοντέλο παλινδρόμησης κορυφογραμμής (RidgeClassifier): ποσοστό ορθών προβλέψεων 68%
Karri Soni και Lanka (12 April 2021)	Αναγνώριση της κατάστασης του οδηγού με βάση 3 κατηγορίες	Μηχανές διανυσμάτων υποστήριξης (SVM): ποσοστό ορθών προβλέψεων 96%
Letizia Marchegiani (10 December 2018)	Ταξινόμηση επικίνδυνης οδηγικής κατάστασης σε 4 επίπεδα	Μηχανές διανυσμάτων υποστήριξης (SVM): ποσοστό ορθών προβλέψεων 94%

Ο αλγόριθμος "Random" Forest" έχει τις καλύτερες επιδόσεις, όπως προκύπτει από το μεγαλύτερο μέρος των πειραμάτων που εξετάστηκαν. Επιπλέον, ο αλγόριθμος "Support Vector Machine" και ο αλγόριθμος "XGBoost ", παρουσιάζουν επίσης καλά αποτελέσματα.

2.4 Ανάλυση οδηγικής συμπεριφοράς

Η ανάλυση της οδηγικής συμπεριφοράς αποτελεί έναν από τους μεγαλύτερους παράγοντες για την ασφαλή οδήγηση. Για να μπορέσει να μελετηθεί και να παρθούν αποφάσεις για την ασφάλεια του οδηγού δειερευνήθηκαν οι μελέτες που έχουν γίνει, σε μακροσκοπική ανάλυση δεδομένων ατυχήματος, σε μελέτες βασισμένες σε προσομοιώσεις οδήγησης, σε νατουραλιστικές μελέτες οδήγησης (NDS) και σε Προηγμένα συστήματα υποβοήθησης οδηγού (ADAS).

Η έρευνα του Toledo Tomer (Tomer, 2007) εξετάζει τα μοντέλα οδηγικής συμπεριφοράς που αποτυπώνουν τις τακτικές αποφάσεις ελιγμών των οδηγών σε διαφορετικές συνθήκες κυκλοφορίας. Τα μοντέλα αυτά είναι απαραίτητα για τα

μικροσκοπικά συστήματα προσομοίωσης της κυκλοφορίας. Εξετάζεται η τρέχουσα κατάσταση στους κύριους τομείς της έρευνας της οδηγικής συμπεριφοράς: επιτάχυνση, αλλαγή λωρίδας και αποδοχή διακένου. Συνολικά, ο κύριος περιορισμός των σημερινών μοντέλων είναι ότι σε πολλές περιπτώσεις δεν αποτυπώνουν επαρκώς την πολυπλοκότητα των οδηγών: δεν αποτυπώνουν τις αλληλεξαρτήσεις μεταξύ των αποφάσεων που λαμβάνονται από τους ίδιους οδηγούς με την πάροδο του χρόνου και μεταξύ των διαστάσεων της απόφασης- αναπαριστούν στιγμιαία λήψη αποφάσεων, γεγονός που δεν αποτυπώνει τις δυνατότητες σχεδιασμού και πρόβλεψης των οδηγών- και αποτυπώνουν μόνο μυωπικές εκτιμήσεις που δεν λαμβάνουν υπόψη τους εκτεταμένους στόχους και εκτιμήσεις της οδήγησης. Σε πολλές περιπτώσεις, αυτό οφείλεται στην περιορισμένη διαθεσιμότητα λεπτομερών δεδομένων τροχιάς, τα οποία απαιτούνται για την εκτίμηση. Ως εκ τούτου, η διαθεσιμότητα δεδομένων αποτελεί σημαντικό εμπόδιο στην πρόοδο της μοντελοποίησης της οδηγικής συμπεριφοράς.

Στην έρευνα των Shangguan et al. (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 του Utkarsh Agrawal (Agrawal, 2019) παρουσιάζει ένα σύστημα ταξινόμησης για τους οδηγούς βαρέων φορτηγών οχημάτων (HGV), χρησιμοποιώντας ένα βασικό σύνολο στερεοτύπων προτύπων οδήγησης που αποκαλύφθηκαν από περιστατικά οδήγησης κατά τη διάρκεια τριών ετών, δηλαδή το 2014, το 2015 και το 2016. Για να επιτευχθεί αυτό, τα στερεότυπα οδήγησης καθορίζονται με τη χρήση ενός πλαισίου ταξινόμησης συνόλου 2 σταδίων, ακολουθούμενου από έναν αλγόριθμο επισήμανσης προφίλ για τον καθορισμό του συνόλου των στερεοτύπων οδήγησης. Τα πολύ παρόμοια στερεότυπα συγχωνεύονται αργότερα για να σχηματίσουν τα βασικά στερεότυπα οδήγησης για τους οδηγούς φορτηγών αυτοκινήτων του Ηνωμένου Βασιλείου. Μετά τον καθορισμό των βασικών στερεοτύπων οδήγησης σε αυτά τα τρία έτη, ένας ταξινομητής δέντρων απόφασης μαθαίνει τους κανόνες ταξινόμησης για τον προσδιορισμό των στερεοτύπων οδήγησης για τους οδηγούς HGV σε ένα νέο σύνολο δεδομένων.

Η έρευνα του Sigve Oltedal (Oltedal, 2006) διερευνά τις επιδράσεις των χαρακτηριστικών προσωπικότητας και του φύλου στην επικίνδυνη οδηγική

συμπεριφορά και την εμπλοκή σε ατυχήματα. Συμμετείχε δείγμα Νορβηγών εφήβων σε δύο νομούς της Νορβηγίας (n=1356). Το άγχος συσχετίστηκε σημαντικά με την αναζήτηση ενθουσιασμού και την επικίνδυνη οδηγική συμπεριφορά, ενώ η αναζήτηση ενθουσιασμού συσχετίστηκε σημαντικά με την επικίνδυνη οδηγική συμπεριφορά και τις συγκρούσεις. Μέσω ανάλυσης παλινδρόμησης, τα χαρακτηριστικά της προσωπικότητας και το φύλο βρέθηκαν να εξηγούν το 37,3% της διακύμανσης της επικίνδυνης οδηγικής συμπεριφοράς. Ωστόσο, οι σχέσεις δεν ήταν πολύ ισχυρές και τα χαρακτηριστικά της προσωπικότητας εξηγούσαν μόνο ένα μέτριο μέρος της διακύμανσης.

Στη έρευνα του Verschuur William L. G., και Karel Hurts. (William L. G et al, 2008) εξετάστηκε ένα δείγμα 743 Ολλανδών οδηγών ερωτήθηκε σχετικά με τα λάθη και τις παραβάσεις που σχετίζονται με την οδήγηση, καθώς και σχετικά με τη συχνότητα εμπλοκής τους σε ατυχήματα τα τελευταία 3 χρόνια. Επιπλέον, μετρήθηκαν μέσω αυτοαναφοράς οι ακόλουθες επικίνδυνες συμπεριφορές και χαρακτηριστικά που σχετίζονται με την οδήγηση: στρατηγικές αποφάσεις που λαμβάνονται σχετικά με την οδήγηση πριν από την έναρξη ενός ταξιδιού, στάσεις που σχετίζονται με τη διάπραξη παραβάσεων, ψυχολογικές πρόδρομες καταστάσεις μη ασφαλούς οδήγησης (όπως η κούραση ή το άγχος κατά την οδήγηση) και σωματικές πρόδρομες καταστάσεις μη ασφαλούς οδήγησης (σωματικά ή ψυχολογικά μειονεκτήματα κατά την οδήγηση). Τα αποτελέσματα δείχνουν ότι αρκετές κλίμακες έχουν αποδεκτούς συντελεστές αξιοπιστίας, αν και αρκετές άλλες χρειάζονται βελτίωση. Στον πίνακα 2.2 παρουσιάζονται οι περιορισμοί, οι ελλείψεις και οι προτάσεις για περαιτέρω των παραπάνω μελετών.

Πίνακας 2.2: Ελλείψεις/Προτάσεις για μελλοντική διερεύνηση των ερευνών που παρουσιάστηκαν

Έρευνα	Ελλείψεις	Προτάσεις για μελλοντική διερεύνηση
Toledo Tomer (23 February 2007)	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
Shangguan et al. (23 April 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 πρόβλεψη.
Utkarsh Agrawal (28 November 2019)	Στην έρευνα χρησιμοποιούνται λίγα στοιχεία για τον προσδιορισμό της κατηγορίας του οδηγού χωρίς να περιλαμβάνει την εμπειρία και τον φόρτο εργασίας και άλλους παράγοντες που μπορεί να επηρεάζουν τον οδηγό.	Προτείνεται η περεταίρω έρευνα για την χρήση περισσότερων δεδομένων όπως το πόσο σκληρές στροφές παίρνει ο οδηγός και ο ψυχικός φόρτος εργασίας του οδηγού. Επίσης προτείνεται η ανάπτυξη ενός συστήματος ταξινόμησης βασισμένου στον συνδυασμό πολλαπλών μεθόδων.
Sigve Oltedal (August 2006)	Τα αποτελέσματα δεν προσφέρουν σαφή απάντηση και μπορεί να μην σχετίζονται και με την συμπεριφορά του οδηγού	Προτείνεται να γίνει διερεύνηση περισσότερων μεταβλητών που να μπορέσουν να σχετίσουν καλύτερη τη σχέση της προσωπικότητας του οδηγού όπως ο χρόνος ο οποίος αφιερώνεται στην οδήγηση .
Verschuur, William L. G., and Karel Hurts. (25 March 2008)	Η μεταβλητές που χρησιμοποιήθηκαν δεν έχουν καλή συσχέτιση μεταξύ τους και πολλά μοντέλα δεν προσφέρουν καλή ακρίβεια στα αποτελέσματα τους.	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.

2.5 Πρόβλημα ανισορροπίας δεδομένων σε κάθε τάξη

Συχνά συναντάμε το ζήτημα της ανισορροπίας των δεδομένων όσον αφορά τον τρόπο κατανομής τους στις διάφορες κατηγορίες σε πραγματικές περιπτώσεις. Ειδικότερα, σε σύγκριση με την ασφαλή οδηγική συμπεριφορά και τη μη πρόκληση ατυχήματος, αντίστοιχα, η επικίνδυνη οδηγική συμπεριφορά και η πιθανότητα ατυχήματος είναι σπάνια γεγονότα στη σχετική έρευνα. Η κύρια κλάση είναι αυτή με τον μεγαλύτερο όγκο δεδομένων, ενώ η μειοψηφική κλάση είναι αυτή με τον μικρότερο όγκο δεδομένων. Η αναλογία των γεγονότων που σχετίζονται με το ατύχημα προς τα γεγονότα που δεν σχετίζονται με το ατύχημα σε θέματα ανάλυσης ατυχημάτων σε πραγματικό χρόνο κυμαίνεται από 1:5.

Για την αντιμετώπιση αυτού του προβλήματος έγινε έρευνα από τον Cristian Padurariu (Padurariu, 2019), έχει δοθεί ένας αριθμός λύσεων που έχουν αναπτυχθεί μέχρι σήμερα. Μπορούν να ομαδοποιηθούν σε δύο διακριτές κατηγορίες: η μία κατηγορία αντιστοιχεί σε μεθόδους που λειτουργούν στο σύνολο δεδομένων σε ένα στάδιο προεπεξεργασίας που προηγείται της ταξινόμησης (resampling techniques), ενώ μια δεύτερη κατηγορία τροποποιεί τις αλγόριθμο ταξινόμησης προκειμένου να δοθεί μεγαλύτερη έμφαση στην κατηγορία της μειονότητας (Synthetic Minority Oversampling Technique, SMOTE). Οι μέθοδοι της πρώτης κατηγορίας είναι γνωστές ως μέθοδοι επαναδειγματοληψίας. Στο πλαίσιο της δεύτερης κατηγορίας η μάθηση με ευαισθησία στο κόστος είναι ένα παράδειγμα που δίνει έμφαση στην λανθασμένη ταξινόμηση των περιπτώσεων από την κατηγορία της μειονότητας κατά τη διάρκεια της διαδικασίας εκπαίδευσης, ενώ είναι ελάχιστα παρεμβατική για τον ταξινομητή με την έννοια ότι ο αλγόριθμος δεν απαιτεί σημαντικές αλλαγές.

Στην έρευνα των Zhu, Shengxue, et al.(Zhu, Shengxue, et al, 2022) εξετάζονται διάφορες τεχνικές SMOTE με boost διαδικασία, Τυχαία Υπερδειγματοληψία (Random Oversampling) και SMOTE-Adaboost. 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.

2.6 Δυσκολία κατανόησης μοντέλου

Πολλές φορές ένα υψηλά εκπαιδευμένο μοντέλο μπορεί να παρουσιάζει καλή ακρίβεια και αποτελέσματα αλλά να είναι δύσκολη η κατανόηση του. Δηλαδή είναι εξίσου σημαντικό να κατανοήσουμε το πόσο συνεισφέρει και επηρεάζει κάθε μεταβλητή το μοντέλο. Αυτό μπορεί να επιτευχθεί με τη επεξηγήσεις πρόσθετων SHapley (SHapley Additive Explanations)(slundberg 2022). Με αυτόν τον τρόπο θα μπορέσουμε να καταλάβουμε και να ξεχωρίσουμε τη θεωρείτε ποιο σημαντικό στα δεδομένα μας και να πάρουμε κατάλληλες αποφάσεις για την καλύτερη ασφάλεια του οδηγού στον δρόμο.

Στην έρευνα του Amir Bahador Parsa (Parsa, 2019) έγινε προσπάθεια να ανιχνευτούν τροχαία ατυχήματα με την χρήση ενός σύνολού δεδομένων πραγματικού χρόνου που αποτελείται από χαρακτηριστικά κυκλοφορίας, δικτύου, δημογραφικά, χρήσης γης και καιρικών συνθηκών. Με την χρήση του συνοπτικό διάγραμμα SHAP που διατάσσει τα χαρακτηριστικά με βάση τη σημασία τους για την ανίχνευση ατυχημάτων βλέπουμε ότι τα σχετιζόμενα με την κυκλοφορία χαρακτηριστικά είναι τα πιο σημαντικά χαρακτηριστικά στο μοντέλο. Με αποτέλεσμα να μπορούν να εντοπιστούν οι περιοχές που θεωρούνται επικίνδυνες προς την ασφάλεια του οδηγού.

Σε έρευνες διαφορετικού αντικειμένου του Sujith Mangalathu (Mangalathu, 2020) χρησιμοποιεί εκτεταμένες πειραματικές βάσεις δεδομένων για να προτείνει μοντέλα μηχανικής μάθησης τυχαίου δάσους για προβλέψεις τρόπων αστοχίας υποστυλωμάτων και διατμητικών τοίχων οπλισμένου σκυροδέματος, χρησιμοποιεί την πρόσφατα αναπτυχθείσα προσέγγιση SHapley Additive exPlanations (slundberg 2022) για την κατάταξη των μεταβλητών εισόδου για τον προσδιορισμό των τρόπων αστοχίας και εξηγεί γιατί το μοντέλο μηχανικής μάθησης προβλέπει έναν συγκεκριμένο τρόπο αστοχίας για ένα δεδομένο δείγμα ή πείραμα.

2.7 Σύνοψη

Λαμβάνοντας υπόψη την βιβλιογραφική ανασκόπηση που πραγματοποιήθηκε, στην παρούσα διπλωματική εργασία θα γίνει προσπάθεια να προσφερθεί επιπλέον γνώση στον τομέα των Ευφυών Μεταφορικών Συστημάτων (ITS) και βάση στον εντοπισμό της οδηγικής συμπεριφοράς του οδηγού.

Συγκεκριμένα θα μελετηθεί η επίδραση των διάφορων οδηγικών χαρακτηριστικών στην αναγνώριση των διαφορετικών επιπέδων της 'οδηγικής συμπεριφοράς' και θα εξεταστεί η μεταξύ τους σχέση. Επιπλέον μέσο την χρήση έξυπνων συστημάτων και μηχανικής μάθησης όπως αναφέρθηκε στης παραπάνω έρευνες θα γίνει η ταξινόμηση μέσο μοντέλων για τα τρία επίπεδα της 'οδηγικής συμπεριφοράς'. Για την αλγόριθμοι για την ταξινόμηση επιλέχθηκαν:

- Ο αλγόριθμος παλινδρόμησης κορυφογραμμής (RidgeClassifier)
- Ο αλγόριθμος ΄Μηχανών Διανυσμάτων Υποστήριξης' (Support Vector Machines)
- Ο αλγόριθμος ΄Τυχαία Δάση' (Random Forests)
- Ο αλγόριθμος ενίσχυσης ακραίας κλίσης (XGBoost)

Τέλος για την καλύτερη κατανόηση αυτόν τον μοντέλων θα γίνει όπως είδαμε σε παρόμοιες έρευνες η αξιολόγησή τους και μέσω του SHapley (SHapley Additive Explanations) και θα βρεθεί η σημαντικότητα και την συνεισφορά των μεταβλητών που επιλέχθηκαν.

3. ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

3.1 Εισαγωγή

Το μέρος αυτό εξηγεί το θεωρητικό πλαίσιο βάσει του οποίου διεξήχθη η μελέτη, καθώς και την επεξεργασία και την ανάλυση των δεδομένων. Αρχικά, γίνονται προσεγγίσεις επεξεργασίας δεδομένων και επαναδειγματοληψίας, καθώς κάθε τάξη έχει άνισο αριθμό δειγμάτων. Στη συνέχεια, παρουσιάζονται τα μοντέλα μηχανικής μάθησης που δημιουργήθηκαν για την κατηγοριοποίηση της οδηγικής συμπεριφοράς στα τρία επίπεδα της "Συμπεριφοράσ οδήγησης" και χωρίζονται σε ομάδες (A) και (B) με διαφορετικές μεταβλητές και εξετάζονται αυτές. Τέλος, τονίζεται η σημασία της αξιολόγησης των μετρικών, των κριτηρίων αποδοχής των μοντέλων και του στατιστικού ελέγχου των αποτελεσμάτων.

3.2 Σημαντικότητα ανεξάρτητων μεταβλητών

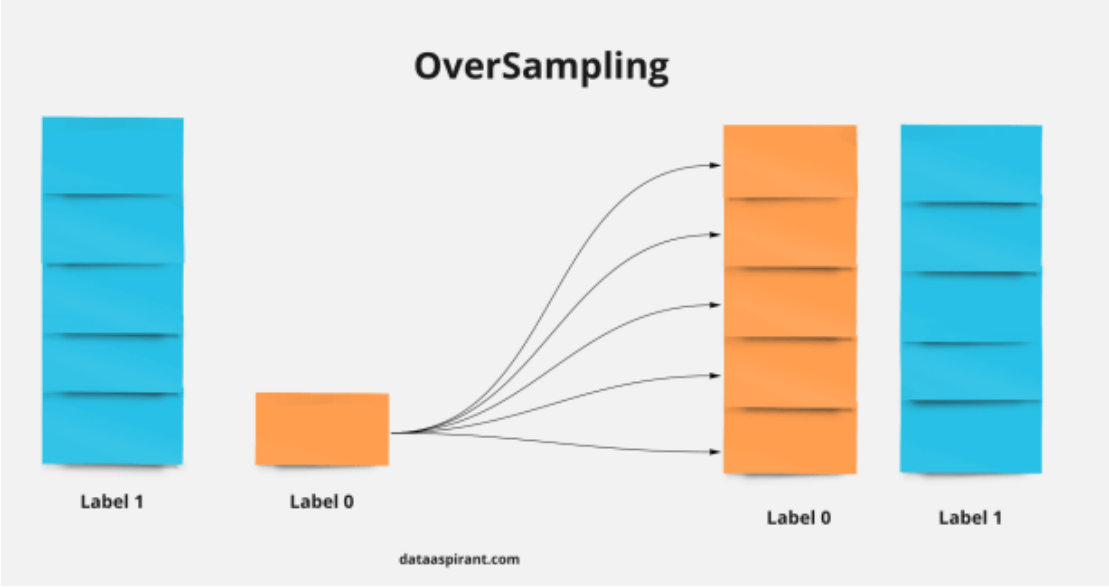
Για να βρούμε τα χαρακτηριστικά που λαμβάνονται από τους συντελεστές υπολογίζουμε την αύξηση του σφάλματος πρόβλεψης του μοντέλου μετά την αλλαγή ενός χαρακτηριστικού για να προσδιορίσουμε τη σημασία του. Επειδή το μοντέλο βασίζεται στο χαρακτηριστικό για την πρόβλεψη, ένα χαρακτηριστικό είναι "σημαντικό" εάν η μετατόπιση των τιμών του αυξάνει το σφάλμα του μοντέλου. Αντίθετα, ένα χαρακτηριστικό είναι λιγότερο σημαντικό εάν η αναδιάταξη των τιμών του δεν έχει καμία επίδραση στο σφάλμα του μοντέλου, επειδή το χαρακτηριστικό αγνοήθηκε για την πρόβλεψη σε αυτή την περίπτωση. Ο Breiman (2001) παρουσίασε την αξιολόγηση της σημαντικότητας των χαρακτηριστικών μετάθεσης για τυχαία δάση. Οι Fisher, Rudin και Dominici (2018) πρότειναν μια παραλλαγή της σημαντικότητας του χαρακτηριστικού που δεν επηρεάζει το μοντέλο, η οποία ονομάστηκε εξάρτηση από το μοντέλο και βασίζεται σε αυτή την έννοια. Προσέφεραν επίσης πιο σύνθετες έννοιες σχετικά με τη συνάφεια των χαρακτηριστικών, όπως μια (ειδική για το μοντέλο) εκδοχή που λαμβάνει υπόψη ότι πολλά μοντέλα πρόβλεψης μπορεί να προβλέπουν καλά τα δεδομένα.

3.3 Μέθοδοι επαναδειγματοληψίας για προβλήματα ανισορροπίας ταξινόμησης

Ένα πρόβλημα ταξινόμησης με ανισορροπία είναι ένα παράδειγμα προβλήματος ταξινόμησης όπου η κατανομή των παραδειγμάτων στις γνωστές κλάσεις είναι μεροληπτική ή λοξή. Η κατανομή μπορεί να ποικίλλει από μια ελαφρά μεροληψία έως μια σοβαρή ανισορροπία, όπου υπάρχει ένα παράδειγμα στην κατηγορία της μειονότητας για εκατοντάδες, χιλιάδες ή εκατομμύρια παραδείγματα στην κατηγορία ή στις κατηγορίες της πλειοψηφίας.

Οι ανισόρροπες ταξινομήσεις αποτελούν πρόκληση για την προγνωστική μοντελοποίηση, καθώς οι περισσότεροι αλγόριθμοι μηχανικής μάθησης που χρησιμοποιούνται για την ταξινόμηση σχεδιάστηκαν με βάση την υπόθεση ίσου αριθμού παραδειγμάτων για κάθε κλάση. Αυτό έχει ως αποτέλεσμα μοντέλα που έχουν φτωχή προβλεπτική απόδοση, ειδικά για την κλάση της μειονότητας. Αυτό είναι ένα πρόβλημα επειδή συνήθως η κλάση της μειονότητας είναι πιο σημαντική και επομένως το πρόβλημα είναι πιο ευαίσθητο στα σφάλματα ταξινόμησης για την κλάση της μειονότητας από ότι για την κλάση της πλειοψηφίας.(Brownlee, Jason et al 2019)

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.

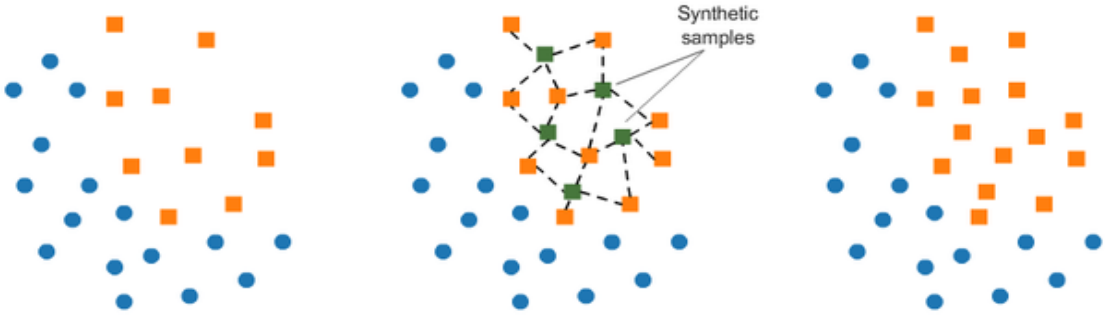


Γράφημα 3.1: Επαναδειγματοληψία δεδομένων που ανήκουν στην κλάση μειοψηφίας

Πηγή: dataasprirant (2022)

3.3.1 Τεχνική Συνθετικής Μειονοτικής Υπερδειγματοληψίας (SMOTE)

Η ταξινόμηση με ανισορροπία συνεπάγεται τη δημιουργία μοντέλων πρόβλεψης για σύνολα δεδομένων με σημαντική ανισορροπία κλάσεων. Όταν εργάζεστε με μη ισορροπημένα σύνολα δεδομένων, η δυσκολία είναι ότι οι περισσότερες προσεγγίσεις μηχανικής μάθησης θα παραβλέψουν τη μειοψηφική κλάση, με αποτέλεσμα την κακή απόδοση, παρά το γεγονός ότι η απόδοση στη μειοψηφική κλάση είναι συχνά η πιο σημαντική. Η υπερδειγματοληψία της κλάσης μειονότητας είναι ένας τρόπος αντιμετώπισης των μη ισορροπημένων συνόλων δεδομένων. Η αντιγραφή παραδειγμάτων στην κλάση της μειονότητας είναι ο ευκολότερος τρόπος, αλλά αυτά τα παραδείγματα δεν παρέχουν καμία νέα πληροφορία στο μοντέλο. Αντίθετα, νέα παραδείγματα μπορούν να δημιουργηθούν με τη σύνθεση παλαιών. Η τεχνική SyntheticMinorityOversamplingTechnique, ή εν συντομία SMOTE, είναι ένα είδος επαύξησης δεδομένων για τον πληθυσμό της μειονότητας. 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 (Brownlee, Jason. Et al 2020).



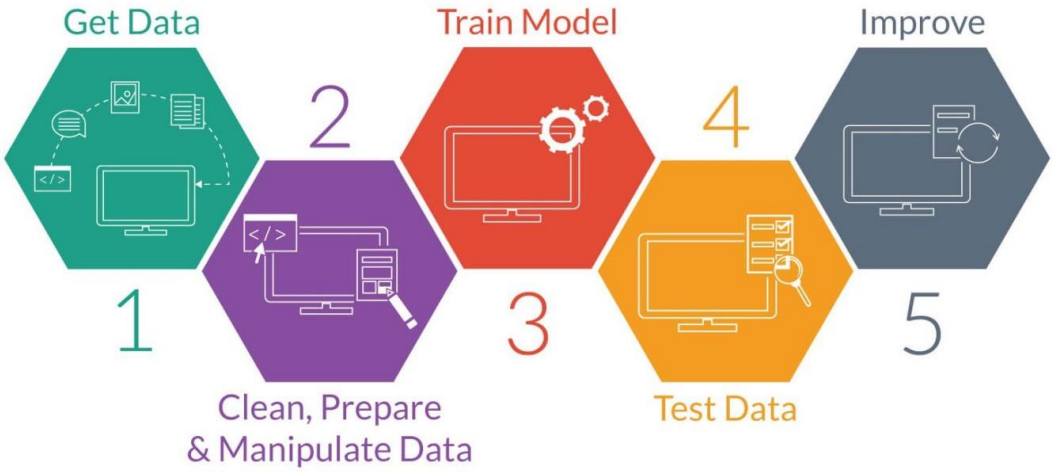
Γράφημα 3.2: SMOTE: Synthetic Minority Oversampling Technique
Πηγή: machinelearningmastery (2022)

3.4 Αλγόριθμοι ταξινόμησης (Classificationalgorithms)

Ο αλγόριθμος ταξινόμησης είναι μια τεχνική επιβλεπόμενης μάθησης που χρησιμοποιείται για τον προσδιορισμό της κατηγορίας των νέων παρατηρήσεων με βάση τα δεδομένα εκπαίδευσης. Στην Ταξινόμηση, ένα πρόγραμμα μαθαίνει από το υπάρχον σύνολο δεδομένων ή τις παρατηρήσεις και στη συνέχεια ταξινομεί τις νέες παρατηρήσεις σε έναν αριθμό κλάσεων ή ομάδων. Όπως, Ναι ή Όχι, 0 ή 1, Spam ή όχι Spam, γάτα ή σκύλος, κ.λπ. Οι κλάσεις μπορούν να αποκαλούνται ως στόχοι/ετικέτες ή κατηγορίες.(www.javatpoint.com 2022)

Στην παρούσα διπλωματική εργασία αναπτύχθηκαν 4 αλγόριθμοι μηχανικής εκμάθησης με σκοπό την ταξινόμηση δεδομένων πολλαπλών κλάσεων (multiclassclassification).

Γενικά, υπάρχουν μερικές βασικές φάσεις που εμπλέκονται στην κατασκευή αλγορίθμων μηχανικής μάθησης. Αρχικά, τα δεδομένα χωρίζονται σε δύο κατηγορίες: σύνολο δεδομένων εκπαίδευσης (trainingdataset) και σύνολο δεδομένων εξέτασης (testingdataset). Τα δεδομένα εκπαίδευσης χρησιμοποιούνται για την εκπαίδευση του αλγορίθμου και τα δεδομένα εξέτασης χρησιμοποιούνται για την αξιολόγηση του μοντέλου. Η αποτελεσματικότητα των μοντέλων κρίνεται βάση ορισμένων σημαντικών μετρικών αξιολόγησης.



Εικόνα 3.3 Διαδικασία Μηχανικής Εκμάθησης
Πηγή: javatpoint (2022)

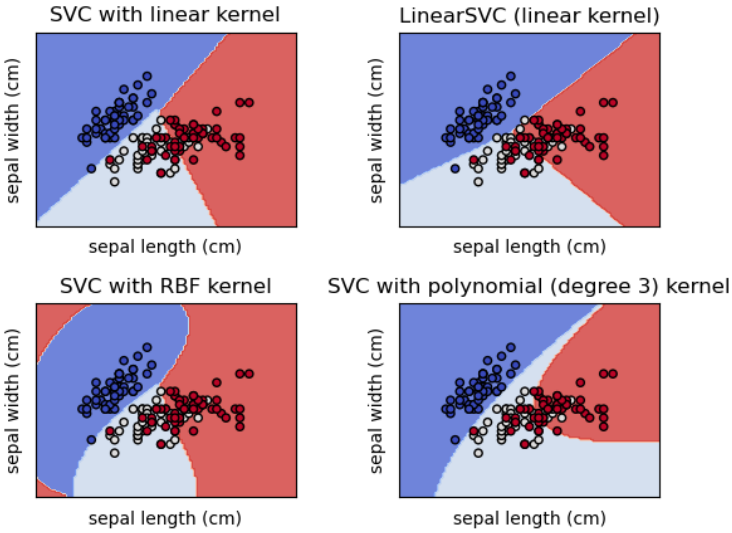
3.4.1 Μέθοδος παλινδρόμησης κορυφογραμμής (RidgeClassifier)

Ο RidgeClassifier είναι μια παραλλαγή ταξινομητή του παλινδρομητή Ridge. Αυτός ο ταξινομητής μετατρέπει τους δυαδικούς στόχους σε -1, 1 πριν αντιμετωπίσει το πρόβλημα ως πρόβλημα παλινδρόμησης και βελτιστοποιήσει τον ίδιο στόχο όπως και πριν. Η προβαλλόμενη κλάση καθορίζεται από το πρόσημο πρόβλεψης του παλινδρομιτή. Το πρόβλημα θεωρείται ως παλινδρόμηση πολλαπλών εξόδων για ταξινόμηση πολλαπλών κλάσεων και η προβλεπόμενη κλάση αντιστοιχεί στην έξοδο με τη μεγαλύτερη τιμή.

Μπορεί να φαίνεται αμφισβητήσιμη η χρήση μιας (penalized) απώλειας ελαχίστων τετραγώνων για την προσαρμογή ενός μοντέλου ταξινόμησης αντί των πιο παραδοσιακών λογιστικών ή αρθρωτών απωλειών. Ωστόσο, στην πράξη, όλα αυτά τα μοντέλα μπορούν να οδηγήσουν σε παρόμοια αποτελέσματα διασταυρούμενης επικύρωσης όσον αφορά την ακρίβεια ή την ακρίβεια/ανάκληση, ενώ η ποινικοποιημένη απώλεια ελαχίστων τετραγώνων που χρησιμοποιείται από τον RidgeClassifier επιτρέπει μια πολύ διαφορετική επιλογή των αριθμητικών επιλυτών με διακριτά προφίλ υπολογιστικής απόδοσης.

3.4.2 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. 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. Με την χρήση της μεθόδου των πυρήνων (kernel method) ο ταξινομητής διανυσμάτων υποστήριξης μπορεί να διαχειριστεί δεδομένα μη γραμμικά διαχωρίσιμα. (Lanka 2022)



Γράφημα 3.4: Αλγόριθμος SVC με την χρήση διαφορετικών μεθόδων πυρήνων Πηγή (Lanka 2022)

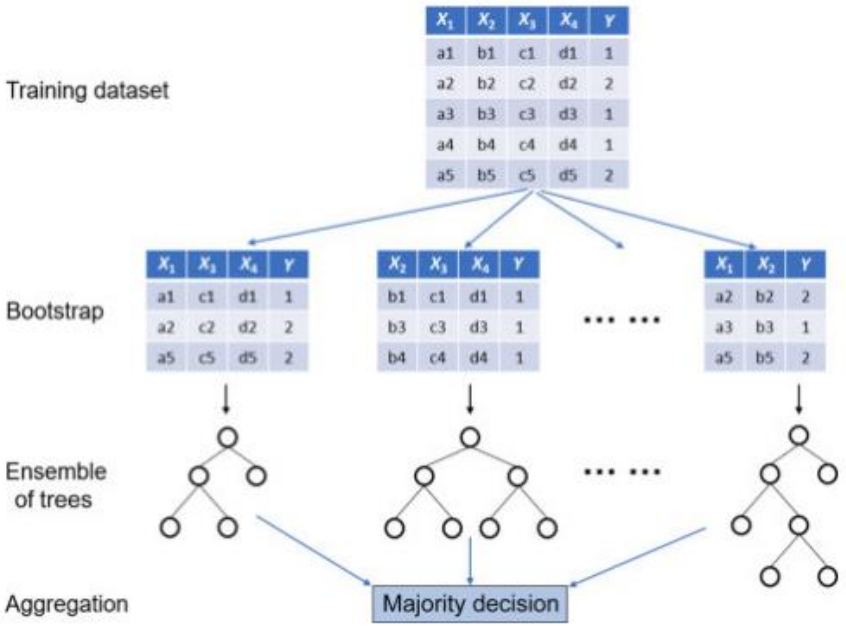
3.4.3 Τυχαία δάση (Random Forests)

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. Η διαδικασία που ακολουθεί το δένδρο απόφασης είναι η εξής:

- Αρχικοποίηση του κόμβου με το σύνολο των δεδομένων
- Διάσπαση του κόμβου με βάση κάποιο κριτήριο διαχωρισμού σε κάποιο από τα γνωρίσματα.

- Επανάληψη του βήματος 2 έως ότου ικανοποιηθεί το κριτήριο τερματισμού και τα δεδομένα έχουν ταξινομηθεί με βάση τα γνωρίσματα τους μέσω ενός συστήματος αποφάσεων

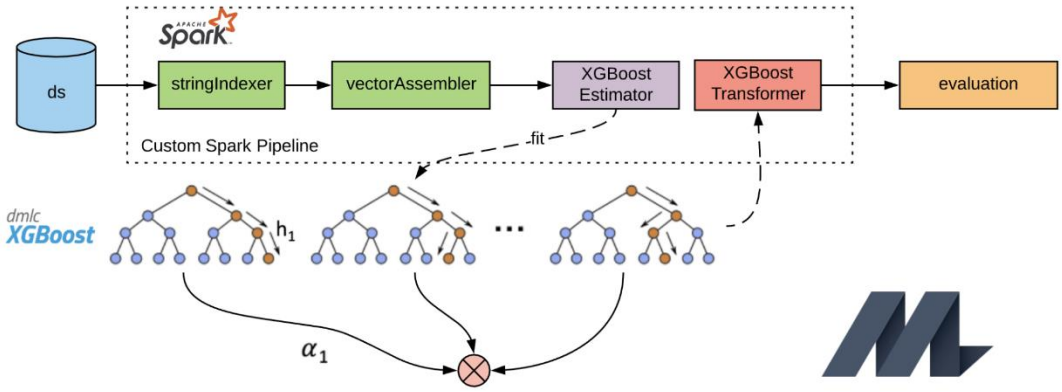
Ο δείκτης gini (gini index) και η εντροπία (entropy) αποτελούν τα κριτήρια υπολογισμού του κέρδους πληροφορίας. Οι αλγόριθμοι δένδρων απόφασης αξιοποιούν το κέρδος πληροφορίας για τον βέλτιστο αριθμό διαχωρισμών του κάθε κόμβου. Το μοντέλο των τυχαίων δασών αποτελεί συνδυαστική μέθοδο που εκπαιδεύει παράλληλα πολλαπλά δένδρα απόφασης αξιοποιώντας την τεχνική bagging δηλαδή των συνδυασμό bootstrapping και aggregation. 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. (Misra et al 2020)



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

3.4.4 XgBoost

Η XGBoost (Extreme Gradient Boosting) είναι μια βελτιστοποιημένη κατανεμημένη βιβλιοθήκη ενίσχυσης βαθμίδας που έχει σχεδιαστεί για να είναι εξαιρετικά αποδοτική, ευέλικτη και φορητή. Υλοποιεί αλγορίθμους μηχανικής μάθησης υπό το πλαίσιο Gradient Boosting. Το XGBoost παρέχει μια παράλληλη δενδρική ενίσχυση (επίσης γνωστή ως GBDT, GBM) που επιλύει πολλά προβλήματα της επιστήμης των δεδομένων με γρήγορο και ακριβή τρόπο. Ο ίδιος κώδικας τρέχει σε σημαντικά κατανεμημένα περιβάλλοντα (Hadoop, SGE, MPI) και μπορεί να επιλύσει προβλήματα πέρα από δισεκατομμύρια παραδείγματα.(Fu et al 2021)



Γράφημα 3.5: Διαδικασία ταξινόμησης XgBoost
Πηγή: Fu et al (2021)

3.5 Μετρικές αξιολόγησης για ταξινόμηση (Evaluation metrics for classification)

3.5.1 Μήτρα σύγχυσης (Confusion matrix)

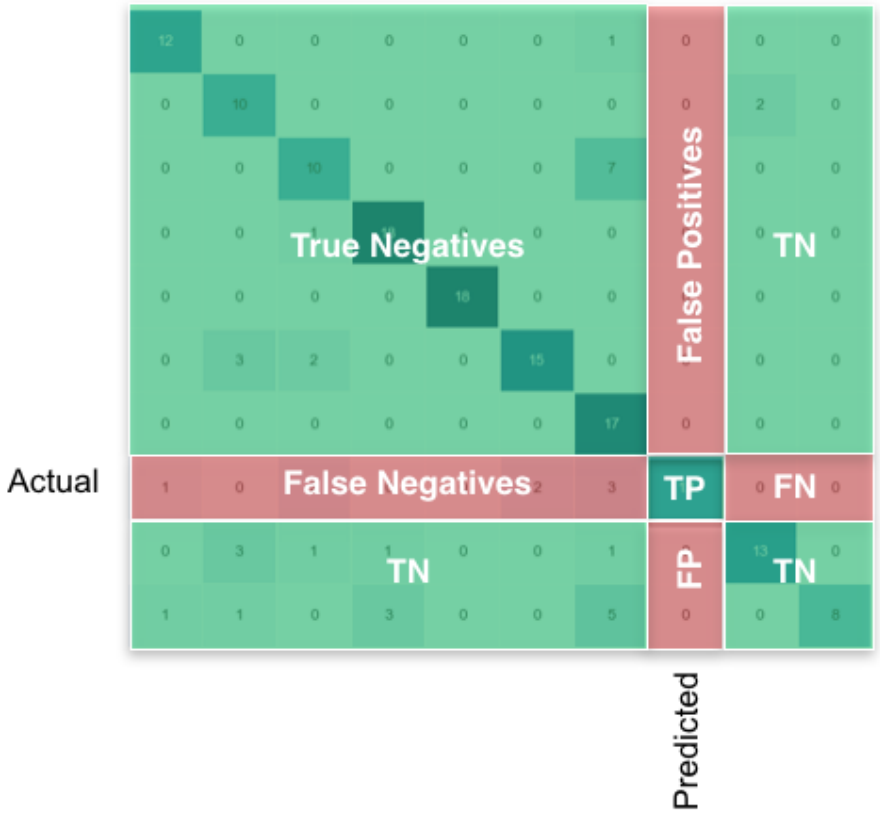
Το Confusion Matrix χρησιμοποιείται για τη γνώση της απόδοσης μιας ταξινόμησης μηχανικής μάθησης και παρουσιάζεται ως ένας πίνακας NxN, όπου N είναι ο αριθμός των κλάσεων ή των εξόδων. Ο πίνακας αυτός δίνει μια σύγκριση μεταξύ πραγματικών και προβλεπόμενων τιμών. Μπορεί να αποτελείτε από 2 κλάσεις, όπου παίρνουμε μήτρα σύγχυσης 2 x 2 ή από 3 τάξης όπου παίρνουμε μήτρα σύγχυσης 3 X 3. Στην παρούσα διπλωματική εργασία υπάρχουν 3 κλάσεις όπου κατατάσσουμε τους οδηγούς ως (0) φυσιολογική οδήγηση (Normal), (1) επικίνδυνη οδήγηση (Dangerous) και (2) οδήγηση λίγο πριν το ατύχημα (Unn_Acc). Έτσι, με αυτό τον τρόπο, μπορούμε να αξιολογήσουμε τα μοντέλα που τρέχουμε.

Η βάση δεδομένων μας που αποτελείται από 3 κλάσεις κατατάσσεται σε 4 κατηγορίες:

1. Αληθώς Θετικά (TruePositives – TP): Το πλήθος των στιγμιοτύπων της βάσης (+), ύπαρξη συμβάντος, που κατηγοριοποιήθηκαν ως (+) από τον ταξινομητή.
2. Αληθώς Αρνητικά (TrueNegative – TN): Το πλήθος των στιγμιοτύπων που ανήκουν στην κλάση (-), μη ύπαρξη συμβάντος, και ο ταξινομητής κατηγοριοποίησε ως (-).

3. Ψευδώς Θετικά (FalsePositive – FP): Είναι το πλήθος των παραδειγμάτων της κλάσης (-), μη ύπαρξη συμβάντος, που εσφαλμένα ο ταξινομητής κατηγοριοποίησε ως (+), ύπαρξη συμβάντος.
4. Ψευδώς Αρνητικά (FalseNegative – FN): Είναι το πλήθος των παραδειγμάτων της κλάσης (+), ύπαρξη συμβάντος, που εσφαλμένα κατηγοριοποιήθηκαν από τον ταξινομητή ως (-), μη συμβάντος.

Η μήτρα σύγχυσης έχει την μορφή που φαίνεται στον πίνακα:



Πίνακα 3.6: Confusion matrix for multiclass classification
Πηγή: Bharathi (2021)

3.5.2 Ακρίβεια (Accuracy,predict_score)

Δίνει τη συνολική ακρίβεια του μοντέλου δηλαδή το κλάσμα των συνολικών δειγμάτων που ταξινομούνται σωστά από τον ταξινομητή.

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Εναλλακτικά της ορθότητας μπορεί να χρησιμοποιηθεί το ποσοστό εσφαλμένης ταξινόμησης (MisclassificationRate) που υπολογίζει ποιο μέρος των προβλέψεων είναι λανθασμένο.

MisclassificationRate = 1 – accuracy.

3.5.3 Ορθότητα (Precision)

Υπολογίζει ποιο μέρος των προβλέψεων ως θετική τάξη είναι πραγματικά θετικό.

(Khanna, Mohit et al 2021)

$$precision = \frac{TP}{TP+FP}$$

3.5.4 Ανάκληση ή Ευαισθησία (Recallor Sensitivity)

Υπολογίζει ποιο κλάσμα όλων των θετικών δειγμάτων προβλέπονται σωστά ως θετικό από τον ταξινομητή. (Khanna, Mohit et al 2021)

$$sensitivity\ ή\ recall = \frac{TP}{TP+FN}$$

3.5.5 Εξειδικευτικότητα (Specificity)

Υπολογίζει ποιο κλάσμα όλων των αρνητικών δειγμάτων προβλέπονται σωστά ως αρνητικά από τον ταξινομητή. (Khanna, Mohit et al 2021)

$$specificity = \frac{TN}{TN+FP}$$

3.5.6 Μέτρο F (F-measure)

Συνδυάζει την ακρίβεια και την ανάκληση σε ένα μόνο μέτρο, μαθηματικά είναι το αρμονικό μέσο ακρίβειας και ανάκλησης (Khanna, Mohit et al 2021)

$$F\text{-}measure = \frac{2*precision*recall}{precision+recall}$$

3.5.7 Μέτρο G (G-means)

Υπολογίζει τον γεωμετρικό μέσο όρο των στοιχείων του πίνακα κατά μήκος του καθορισμένου άξονα του πίνακα. (MeteoInfo 2022)

$$G\text{-}means = \sqrt[n]{x1 * x2 * x3..}$$

3.5.8 Δείκτης λάθος συναγερμού (False alarm rate)

Υπολογίζεται ως ο αριθμός των λανθασμένων θετικών προβλέψεων δια του συνολικού αριθμού των αρνητικών. (Sklearn.Metrics.F1_score , 2022)

$$\text{False alarm rate} = \frac{FP}{TN + FP}$$

3.5.9 Μακροοικονομικός μέσος όρος (MacroAverage)

Χρησιμοποιείτε όταν όλες οι κατηγορίες πρέπει να αντιμετωπίζονται ισότιμα για να αξιολογήσετε τη συνολική απόδοση του ταξινομητή σε σχέση με τις πιο συχνές ετικέτες κλάσεων. (Micro-Average & Macro-Average Scoring Metrics – Python, 2020)

$$F1_{class1} + F1_{class2} + \cdots + F1_{classN}$$

3.5.10 Σταθμισμένος μέσος όρος (WeightedAverage)

Είναι ένας υπολογισμός που λαμβάνει υπόψη τη σχετική τιμή των ακεραίων σε μια συλλογή δεδομένων. Κάθε τιμή στο σύνολο δεδομένων κλιμακώνεται με ένα προκαθορισμένο βάρος πριν ολοκληρωθεί ο τελικός υπολογισμός κατά τον υπολογισμό ενός σταθμισμένου μέσου όρου. (Micro-Average & Macro-Average Scoring Metrics – Python, 2020)

$$F1_{class1} * W_1 + F1_{class2} * W_2 + \cdots + F1_{classN} * W_N$$

3.5.11 Χαρακτηριστική Καμπύλη Λειτουργίας Δέκτη (Receiver Operating Characteristic Curve - ROC Curve)

Η καμπύλη ROC είναι η γραφική παράσταση του αληθινού θετικού ποσοστού έναντι του ψευδούς θετικού ποσοστού για ένα δυαδικό σύστημα ταξινομητή καθώς το όριο της διάκρισής του είναι ποικίλο.

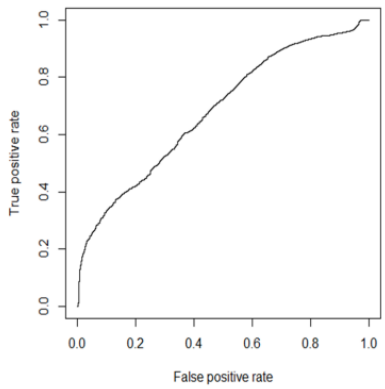
Ο κατακόρυφος άξονας της καμπύλης απεικονίζει το TP (=sensitivity) ενώ ο οριζόντιος το FP ποσοστό (1 - specificity).

Η ορθότητα (accuracy) υπολογίζεται ως το εμβαδόν κάτω από την καμπύλη ROC (Area Under the Curve – AUC)

AUC = 1: Πρόκειται για ιδανικό μοντέλο

AUC = 0.5: Πρόκειται για τυχαία πρόβλεψη

Ένα μοντέλο είναι αυστηρά καλύτερο αν έχει μεγαλύτερη περιοχή κάτω από την καμπύλη συνεπώς οι καλοί ταξινομητές βρίσκονται κοντά στην αριστερή πάνω γωνία του διαγράμματος. (Receiver Operating Characteristic (ROC), 2022)



Εικόνα 5.1: Καμπύλη ROC

3.5.12 Επεξηγήσεις πρόσθετων SHapley (SHapley Additive Explanations)

Το SHAP (SHapleyAdditiveExplanations) είναι μια θεωρητική προσέγγιση που εξηγεί τα αποτελέσματα οποιουδήποτε μοντέλου μηχανικής μάθησης. Συνδέει τη βέλτιστη κατανομή πιστώσεων με τοπικές εξηγήσεις χρησιμοποιώντας τις κλασικές τιμές Shapley από τη θεωρία και τις σχετικές επεκτάσεις τους. Ουσιαστικά, η τιμή Shapley είναι η μέση αναμενόμενη οριακή συνεισφορά μιας μεταβλητής αφού ληφθούν υπόψη όλοι οι πιθανοί συνδυασμοί. Η αξία Shapley βοηθά στον καθορισμό της ανταμοιβής για όλες τις μεταβλητές, όταν κάθε μία μπορεί να έχει συνεισφέρει περισσότερο ή λιγότερο από τις άλλες.(slundberg 2022)

4. ΣΥΛΛΟΓΗ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑ ΣΤΟΙΧΕΙΩΝ

4.1 Εισαγωγή

Ο στόχος του ερευνητικού έργου iDREAMS, όπως αναφέρθηκε σε προηγούμενες ενότητες, είναι ο εντοπισμός επικίνδυνης οδικής συμπεριφοράς και οι διαδικασίες και οι πράξεις με της οποίες μπορεί να γίνει η ελαχιστοποίηση της μη ασφαλούς οδηγικής συμπεριφοράς. Επιπλέων η μελέτη συγκεκριμένων δεδομένων σχετικά με την οδηγική συμπεριφορά και το οδικό περιβάλλον αποτελεί κρίσιμο βήμα για την επίτευξη των παραπάνω στόχων.

Σε αυτό το κεφάλαιο παρουσιάζεται η συλλογή και ο τρόπος επεξεργασίας των στοιχείων που διαμόρφωσαν την τελική βάση δεδομένων σύμφωνα με την οποία έγινε η ανάλυση της παρούσας έρευνας.

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.

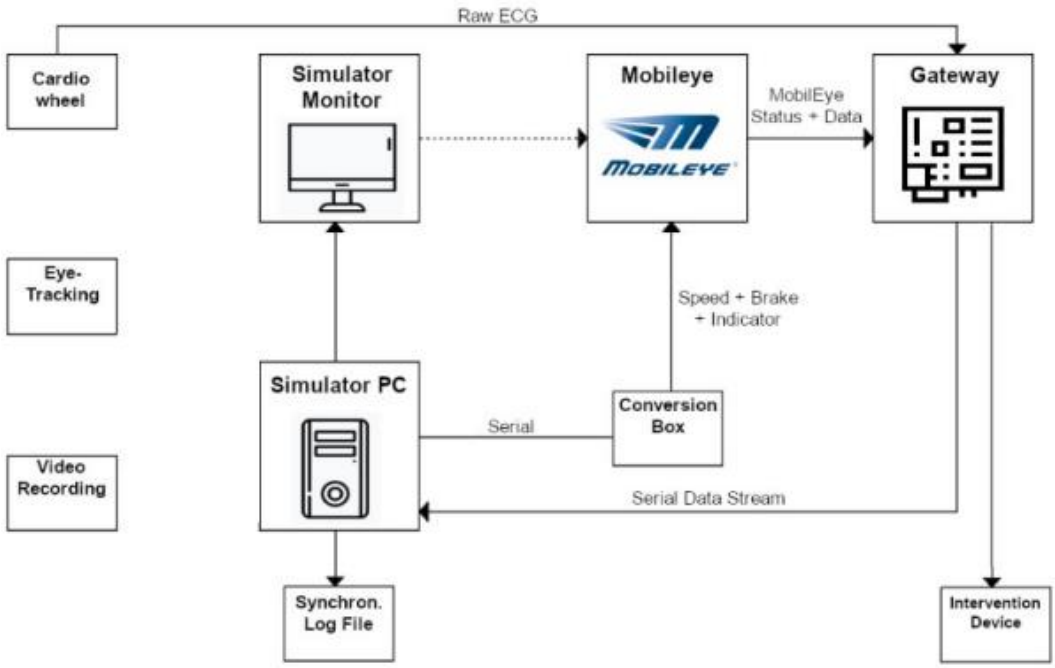
4.2.2 Προσομοιωτής οδήγησης

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 135ο.

Εικόνα 4.1: Προσομοιωτής οδήγησης

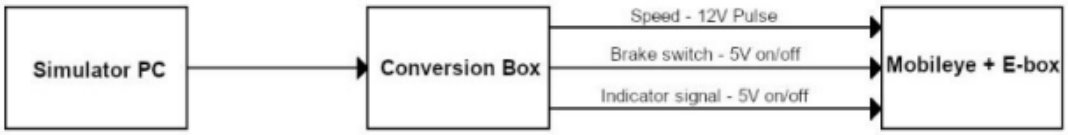
4.2.3 Αρχιτεκτονική προσομοιωτή οδήγησης

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



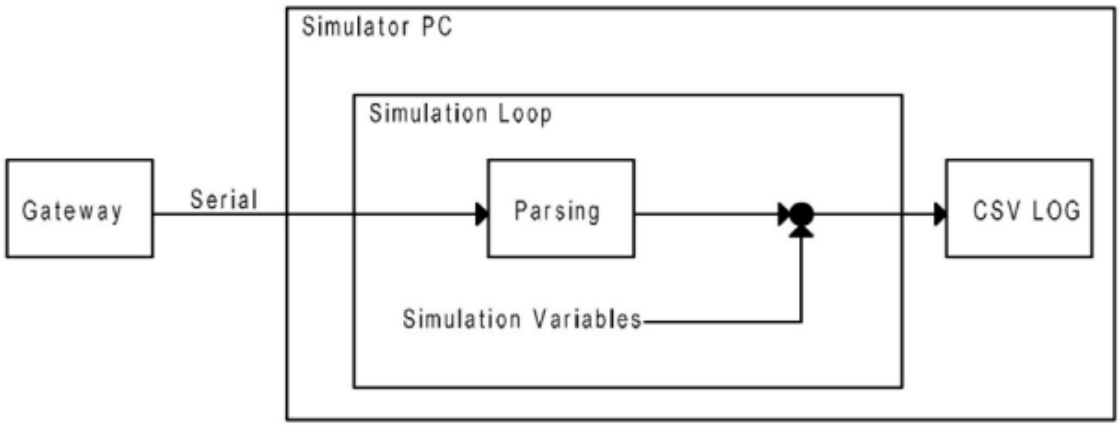
Εικόνα 4.2: Περιγραφή αρχιτεκτονικής προσομοιωτή

Σαν τα πραγματικά οχήματα, δεδομένα όπως η ταχύτητα, η θέση πέδησης και ο δείκτης χρήσης συλλέγονται από το Mobileye και είναι απαραίτητα 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 30 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 Ταχύτητα**
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
Σενάριο	**Οδικό Τμήμα**	**Αριθμός Λωρίδων**	**Όρια Ταχύτητα**
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

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 (correlation)

Η συσχέτιση μεταξύ των ανεξάρτητων μεταβλητών πρέπει να διερευνηθεί προκειμένου να δημιουργηθούν μοντέλα ταξινόμησης και παλινδρόμησης. Οι τιμές των συντελεστών συσχέτισης Pearson κυμαίνονται στο εύρος [-1,1] και η σύνδεση μεταξύ των ανεξάρτητων μεταβλητών είναι γραμμική.

Ακολουθούν τα χαρακτηριστικά των μεταβλητών:

- Ελάχιστη συσχέτιση για 0.00 ≤ |r| ≤ 0.30
- Μέτρια συσχέτιση για 0.31 ≤ |r| ≤ 0.70
- Υψηλή συσχέτιση για 0.71 ≤ |r| ≤ 1.00

Ως αποτέλεσμα, καθιερώθηκαν κατάλληλες τεχνικές υπολογισμού και απεικόνισης της συσχέτισης των μεταβλητών με τη χρήση του ίδιου αναλυτικού πακέτου στο προγραμματιστικό περιβάλλον python. Η συσχέτιση μεταξύ των διαφόρων παραγόντων απεικονίζεται στον τριγωνικό θερμικό χάρτη που ακολουθεί. Ένα θερμό χρώμα υποδηλώνει θετική συσχέτιση, ενώ ένα ψυχρό χρώμα υποδηλώνει αρνητική συσχέτιση.(Brownlee, Jason et al 2018) *Γράφημα 4.1: Τριγωνικός χάρτης συσχέτισης μεταβλητών*

	TTC_mean	TTC_std	Headway_median	Headway_mean	Headway_std	Speed_std	HandsOnEvent_median	HandsOnEvent_mean	FatigueEvent_median	ME_ForwardCollisionWarning_mean	ME_LaneDepartureWarningActive_mean	TTC_min	TTC_max	Headway_min	Headway_max	Speed_min	Speed_max	Speed_mean_1	Distance_travelled_sum	BSAV_SpeedLimitMS_max	BSAV_SpeedLimitKPH_max
TTC_mean																					
TTC_std	0.7663																				
Headway_median	0.0176	-0.0044																			
Headway_mean	0.0413	-0.0172	0.6106																		
Headway_std	0.0448	-0.0217	0.3254	0.8881																	
Speed_std	-0.0709	-0.0371	-0.0193	0.1120	0.1897																
HandsOnEvent_median	-0.0200	0.0007	-0.0284	-0.0619	-0.0612	0.1291															
HandsOnEvent_mean	-0.0216	0.0046	-0.0178	-0.0693	-0.0771	0.1363	0.9750														
FatigueEvent_median	-0.0164	-0.0310	-0.0219	-0.0669	-0.0782	0.0822	0.2700	0.2588													
ME_ForwardCollisionWarning_mean	0.0021	0.0097	-0.0017	-0.0047	-0.0054	-0.0068	-0.0165	-0.0197	0.0205												
ME_LaneDepartureWarningActive_mean	0.0626	0.0402	-0.0089	-0.0326	-0.0403	-0.0887	-0.0315	-0.0371	-0.0114	-0.0065											
TTC_min	0.3084	-0.0028	0.0827	0.2357	0.2712	-0.1657	-0.0895	-0.0927	-0.0627	-0.0205	0.0375										
TTC_max	0.7472	0.9961	-0.0039	-0.0159	-0.0202	-0.0437	-0.0027	0.0015	-0.0325	0.0093	0.0412	0.0052									
Headway_min	0.1253	0.0227	0.3209	0.5707	0.5046	-0.0706	-0.0439	-0.0494	-0.0670	-0.0106	-0.0024	0.3876	0.0251								
Headway_max	0.0437	-0.0213	0.2491	0.7858	0.9682	0.1991	-0.0565	-0.0744	-0.0750	-0.0052	-0.0397	0.2648	-0.0198	0.4588							
Speed_min	0.1048	0.0982	-0.1431	-0.4428	-0.5275	-0.6765	-0.0485	-0.0392	-0.0445	0.0298	0.1245	-0.0795	0.0989	-0.2467	-0.5179						
Speed_max	0.1047	0.1242	-0.1771	-0.3797	-0.3936	-0.3903	0.0006	-0.0047	-0.0672	0.0328	0.1432	-0.1519	0.1234	-0.2726	-0.3710	0.8099					
Speed_mean_1	0.0967	0.1172	-0.1822	-0.4728	-0.5274	-0.5592	-0.0187	-0.0154	-0.0492	0.0325	0.1326	-0.1266	0.1174	-0.2840	-0.5098	0.9475	0.9335				
Distance_travelled_sum	-0.0819	-0.0104	-0.0827	-0.2516	-0.2955	0.0341	0.0229	0.0223	0.1779	-0.0029	0.0268	-0.2253	-0.0126	-0.2265	-0.2840	0.1608	0.0641	0.1215			
BSAV_SpeedLimitMS_max	0.1139	0.0612	-0.0300	-0.0053	0.0125	-0.4861	-0.0721	-0.0924	-0.0919	0.0413	0.1241	0.0877	0.0660	0.0056	0.0164	0.5563	0.6892	0.6245	-0.0533		
BSAV_SpeedLimitKPH_max	0.1139	0.0612	-0.0300	-0.0053	0.0125	-0.4861	-0.0721	-0.0924	-0.0919	0.0413	0.1241	0.0877	0.0660	0.0056	0.0164	0.5563	0.6892	0.6245	-0.0533	1.0000	

Από το γράφημα 4.1 προκύπτουν τα εξής συμπεράσματα:

- Υπάρχει ισχυρή συσχέτιση μεταξύ πολλαπλών περιγραφικών στατιστικών για την ίδια μεταβλητή. Δεδομένου ότι η προαναφερθείσα ισχυρή συσχέτιση αφορά τη σύνδεση μεταξύ διαφορετικών μορφών του ίδιου πράγματος, είναι λογικό.
- Υπάρχει σημαντική συσχέτιση μεταξύ των μεταβλητών της ταχύτητας (Speed) και των περιορισμών ταχύτητας (BSAV SpeedLimit). Όταν αυξάνεται ο περιορισμός ταχύτητας, αυξάνεται και η ταχύτητα του οδηγού.

4.6 Σύνοψη

Συμπερασματικά, τα δεδομένα συλλέχθηκαν μέσω ενός πειράματος προσομοίωσης οδήγησης και θα ανακτηθούν δεδομένα που συνδέονται με τα χαρακτηριστικά της οδηγικής συμπεριφοράς. Ακολούθησε η κατάλληλη επεξεργασία των δεδομένων και η κατάρτιση περιγραφικών στατιστικών στοιχείων, προκειμένου να έχουν καλύτερη εικόνα της κατάστασής τους. Τέλος, διερευνήθηκε η συσχέτιση μεταξύ των μεταβλητών, η οποία αποτελεί προϋπόθεση για τις προκαταρκτικές διαδικασίες των αναλύσεων που θα ακολουθήσουν.

5. ΕΦΑΡΜΟΓΗ ΜΕΘΟΔΟΛΟΓΙΑΣ - ΑΠΟΤΕΛΕΣΜΑΤΑ

5.1 Εισαγωγή

Στο κεφάλαιο αυτό παρουσιάζεται η εφαρμογή της μεθοδολογίας που ακολουθήθηκε στην παρούσα έρευνα καθώς και τα αποτελέσματά της. Όπως αναφέρθηκε και στα παραπάνω κεφάλαια στόχος της εργασίας είναι ο προσδιορισμός της "Συμπεριφοράς οδήγησης".

Αρχικά για να μπορέσουμε να διερευνήσουμε την επιρροής των διαφορετικών παραγόντων της οδηγικής συμπεριφοράς, σύμφωνα με τις μεθοδολογίες προγενέστερων ερευνών θα αναπτυχθούν κατάλληλοι αλγόριθμοι μηχανικής εκμάθησης για την ταξινόμηση τους. Επιπλέον χωρίστηκαν τα δεδομένα σε δύο ομάδες (A) και (B) και ξεχωρίστηκαν οι ποιο σημαντικές μεταβλητές. Ειδικότερα θα αξιολογηθεί η σημαντικότητα των μεταβλητών στην ταξινόμηση καθώς και τα αποτελέσματα του κάθε μοντέλου.

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 (NumPy library for the Python programming 2022)
- Ανάλυση και χειρισμός δεδομένων: Pandas (Pandas - Python Data Analysis Library 2022)
- Χειρισμός ανομοιογένειας δεδομένων: Imbalanced Learn (Padurariu et al 2019)
- Γραφική απεικόνιση: Matplotlib, Seaborn (Matplotlib: Visualization with Python 2022)
- Μηχανική εκμάθηση: Scikit-Learn, Xgboost (Fu et al 2021)

5.2 Εντοπισμός επιπέδου ΄Επικίνδυνης Συμπεριφοράς Οδηγού’

Η πρώτη μέθοδος επικεντρώνεται στην εκτίμηση του αντίκτυπου κάθε πτυχής στην ανίχνευση της μη ασφαλούς οδηγικής συμπεριφοράς του οδηγού. Η εξέταση διαφόρων μεταβλητών κινδύνου με βάση την κατασκευή τεσσάρων αλγορίθμων κατηγοριοποίησης αποτελεί μέρος της τεχνικής που χρησιμοποιείται για την επίτευξη αυτού του στόχου. Τα ουσιώδη κριτήρια θα αξιολογηθούν με βάση τη συνολική απόδοση των μοντέλων ταξινόμησης.

5.2.1 Καθορισμός επιπέδων ασφαλείας

Απαιτείται η κατηγοριοποίηση των δεδομένων οδήγησης σε ένα από τα τρία επίπεδα της "Οδηγικής Συμπεριφοράς" πριν από την κατασκευή των αλγορίθμων ταξινόμησης και την έρευνα της επίδρασης των παραγόντων στην ανασφαλή οδήγηση. Με την βάση στη βιβλιογραφική ανάλυση, αξιολογήθηκαν διάφορες στρατηγικές ομαδοποίησης που έχουν χρησιμοποιηθεί σε προηγούμενες έρευνες και έγιναν περεταίρω δοκιμές πάνω σε αυτές. Συγκεκριμένα έγιναν δοκιμές αλλάζοντας το όριο ταχύτητα και συγκρίνοντας το με την μέγιστη ταχύτητα κάθε οδηγού. Τελικά η κατανομή των δειγμάτων στις τρεις κατηγορίες έγινε χρησιμοποιώντας τα καλύτερα αποτελέσματα από της δοκιμές.

- ‘Φυσιολογική Οδήγηση’ (class: 0): Μέγιστη Ταχύτητα <= 0,8* Τρέχον όριο ταχύτητας
- ‘Επικίνδυνη Οδήγηση’ (class: 1):
  0,8* Τρέχον όριο ταχύτητας <= Μέγιστη Ταχύτητα <= Τρέχον όριο ταχύτητας
- ‘Οδήγηση Αποφεύγοντας Ατύχημα’ (class: 2): Μέγιστη Ταχύτητα => Τρέχον όριο ταχύτητας

5.2.2 Επιλογή χαρακτηριστικών (Feature selection)

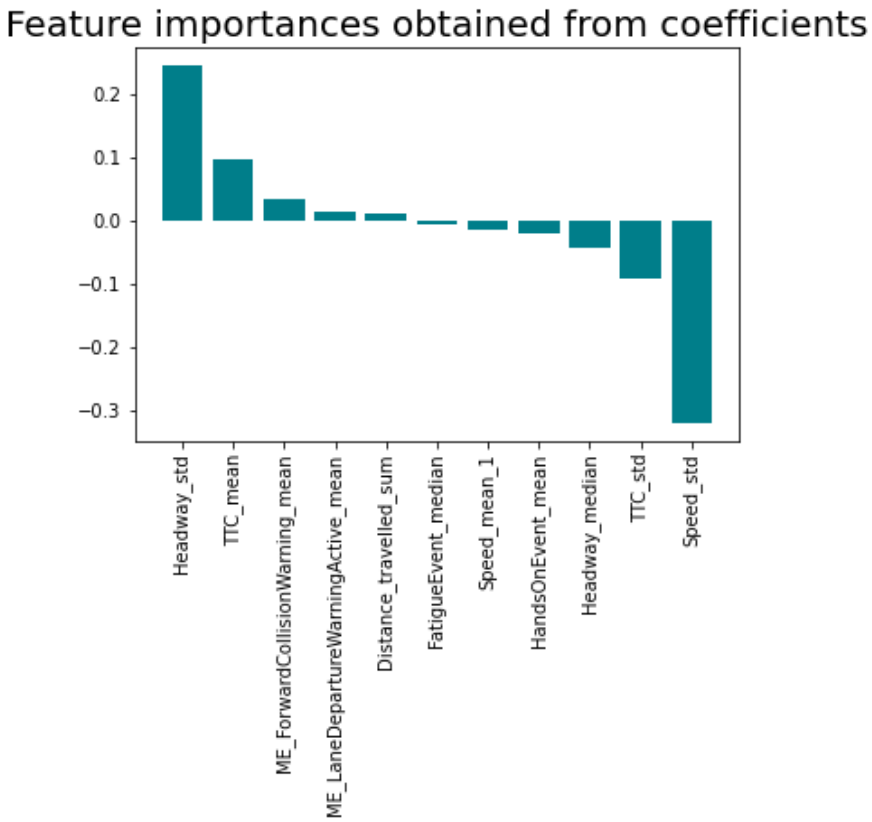
Η διαδικασία επιλογής χαρακτηριστικών αποτελεί σημαντικό μέρος της τεχνικής. Στόχος της μεθόδου είναι η μείωση του αριθμού των μεταβλητών εισόδου με παράλληλη μείωση του υπολογιστικού κόστους του μοντέλου και βελτίωση της προβλεπτικής του απόδοσης. Τα χαρακτηριστικά πρέπει να επιλέγονται με βάση τη συσχέτιση των μεταβλητών καθώς και την επίδραση κάθε μεταβλητής στη διαδικασία κατηγοριοποίησης. Η μέθοδος αυτή αποτελεί ένα πρώτο βήμα για τη μείωση του αριθμού των μεταβλητών εισόδου και τη βελτίωση των μοντέλων. Διερευνήθηκαν διάφορα σύνολα παραγόντων με τη συγχώνευσή τους ανάλογα με τη συσχέτιση και την επίδρασή 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

	TTC_mean	TTC_std	Headway_median	Headway_std	Speed_std	HandsOnEvent_mean	FatigueEvent_median	ME_ForwardCollisionWarning_mean	ME_LaneDepartureWarningActive_mean	Speed_mean_1	Distance_travelled_sum
TTC_mean	1.0000	-0.0233	0.3262	0.2261	0.0202	-0.0266	0.0545	-0.0042	0.0341	-0.0545	-0.1224
TTC_std	-0.0233	1.0000	-0.2853	-0.0865	0.2391	0.1324	0.0492	0.0339	-0.0144	0.1048	0.1257
Headway_media	0.3262	-0.2853	1.0000	0.5991	0.0086	-0.0824	0.0419	-0.0288	0.0525	-0.3011	-0.1111
Headway_std	0.2261	-0.0865	0.5991	1.0000	0.1820	-0.0291	-0.0121	-0.0275	0.0546	-0.2338	-0.0988
Speed_std	0.0202	0.2391	0.0086	0.1820	1.0000	0.2985	0.3568	-0.0036	-0.0866	-0.5686	0.0350
HandsOnEvent_mean	-0.0266	0.1324	-0.0824	-0.0291	0.2985	1.0000	0.2775	-0.0297	-0.0797	-0.1655	0.0656
FatigueEvent_median	0.0545	0.0492	0.0419	-0.0121	0.3568	0.2775	1.0000	-0.0106	-0.0837	-0.4289	0.1482
ME_ForwardCollisionWarning_me	-0.0042	0.0339	-0.0288	-0.0275	-0.0036	-0.0297	-0.0106	1.0000	-0.0070	0.0407	-0.0049
ME_LaneDepartureWarningActive_mean	0.0341	-0.0144	0.0525	0.0546	-0.0866	-0.0797	-0.0837	-0.0070	1.0000	0.1391	0.0229
Speed_mean_1	-0.0545	0.1048	-0.3011	-0.2338	-0.5686	-0.1655	-0.4289	0.0407	0.1391	1.0000	0.0394
Distance_travelled_sum	-0.1224	0.1257	-0.1111	-0.0988	0.0350	0.0656	0.1482	-0.0049	0.0229	0.0394	1.0000

Γράφημα 5.1: Συσχέτιση μεταβλητών προς εξέταση

Για τον εντοπισμό της σημαντικότητας των μεταβλητών στην ταξινόμηση χρησιμοποιήθηκε η τεχνική σημαντικότητας χαρακτηριστικών που λαμβάνονται από τους συντελεστές (Feature importances obtained from coefficients) όπως φαίνεται στο γράφημα 5.2.



Γράφημα 5.2: Σημαντικότητα μεταβλητών σύμφωνα με την μέθοδο

Στη συνέχεια παρατηρώντας της μεταβλητές που προέκυψαν χωρίστηκαν σε δύο ομάδες την

(A): TTC_mean, Headway_std, Speed_std, ME_ForwardCollisionWarning_mean

(B):TTC_mean,Headway_median,HandsOnEvent_mean,FatigueEvent_median,ME_LaneDepartureWarningActive_mean,Speed_mean_1,Distance_travelled_sum

Όπου στην (A) επιλέχθηκαν μόνο οι ποιος σημαντικές μεταβλητές που έχουν την μεγαλύτερη επιρροή για τον εντοπισμό της επικίνδυνης συμπεριφοράς οδήγησης και η (B) όπου περιλαμβάνει περισσότερες μεταβλητές για την δοκιμή και προσπάθεια παραγωγής καλύτερων αποτελεσμάτων.

5.2.3 Προετοιμασία δεδομένων

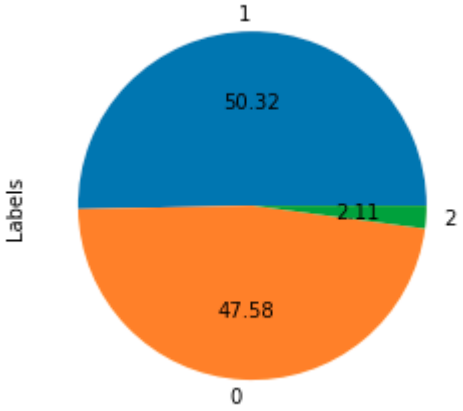
Τα δεδομένα από της δύο ομάδες (A) και (B) διαχωρίστηκαν στο σύνολο δεδομένων εκπαίδευσης (training dataset) και στο σύνολο δεδομένων δοκιμής (testing dataset), με το 70% των δεδομένων να αποτελούν το σύνολο δεδομένων εκπαίδευσης και το 30% του συνόλου δεδομένων δοκιμής, αντίστοιχα. Η λειτουργία της μηχανής, σύμφωνα με την περιγραφή στο κεφάλαιο 3 έχει ως εξής, τα δεδομένα εκπαίδευσης χρησιμοποιούνται για τη δημιουργία μοντέλων μηχανικής μάθησης για να διδαχθεί το μοντέλο να αναγνωρίζει το ποσό της ασφάλειας αναπτύσσοντας μοτίβα αναγνώρισης με βάση ορισμένα χαρακτηριστικά. Για τη διαδικασία αξιολόγησης των λειτουργιών του μοντέλου τα δεδομένα που συλλέγονται κατά τη διάρκεια της εξέτασης και τα κατατάσσει σε μία από τις τρεις βαθμίδες, επιτρέποντας τη σύγκρισή τους με την πραγματικότητα επίπεδα ασφάλειας.

5.2.4 Αντιμετώπιση άνισης κατανομής δεδομένων στις κλάσεις

Σύμφωνα με τη βιβλιογραφική ανάλυση, η πλειονότητα των μελετών είχε πρόβλημα ανισορροπίας του δείγματος όσον αφορά τις διαφορετικές ταξινομήσεις, με τα δείγματα των επικίνδυνων οδηγικών καταστάσεων να είναι πολύ μικρότερα από τα δείγματα των ασφαλών οδηγικών συνθηκών. Επιπλέον, όπως συζητήθηκε στο κεφάλαιο 3, τα μοντέλα ταξινόμησης υποθέτουν ότι τα δεδομένα κατανέμονται ομοιόμορφα μεταξύ των κλάσεων, καθιστώντας τα ευάλωτα σε λάθη ταξινόμησης όταν τα δεδομένα κατανέμονται άνισα. Η ανομοιόμορφη κατανομή που αναφέρεται στον πίνακα 5.3 είναι το αποτέλεσμα του προσδιορισμού των επιπέδων ασφαλείας και της κατηγοριοποίησης των δεδομένων στις διάφορες κατηγορίες.

Πίνακας 5.3: Κατανομή δειγμάτων στα διαφορετικά επίπεδα ασφαλείας

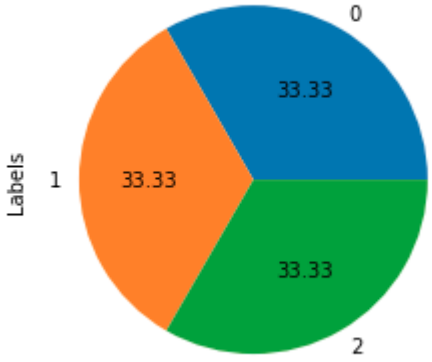
'Συμπεριφορά οδήγησης'	Αριθμός δειγμάτων	Ποσοστό δειγμάτων
Class 0 (Normal)	1580	47,58%
Class 1 (Dangerous)	1671	50,32%
Class 2 (Avoidable Accident)	70	2,11%



Γράφημα 5.4: Κατανομή δεδομένων εκπαίδευσης στα διαφορετικά επίπεδα ασφαλείας πριν την διαδικασία

Επιπλέον όταν χωριστούν σε δεδομένων εκπαίδευσης (training dataset) η τάξη ('Normal') και η τάξη ('Dangerous') μπορεί να οδηγήσουν σε σφάλματα πρόβλεψης της τάξης ('Avoidable Accident'). Ειδικότερα σε μοντέλα που έχουν σχέση με το επίπεδο ασφάλειας είναι πολύ σημαντικό να γίνεται η πρόβλεψη με σωστό τρόπο.

Για αυτό σύμφωνα με τη βιβλιογραφική μελέτη, χρησιμοποιούνται διάφορες στρατηγικές επαναδειγματοληψίας για την επίλυση της ανισορροπίας των δεδομένων στα επίπεδα ασφαλείας και για τη διασφάλιση της αμεροληψίας των μοντέλων. Συγκεκριμένα, η μέθοδος SMOTE και η τεχνική τυχαίας υπερδειγματοληψίας όπως φαίνεται στον πίνακα 5.4.



Γράφημα 5.5: Κατανομή δεδομένων εκπαίδευσης στα διαφορετικά επίπεδα ασφαλείας μετά την διαδικασίας

Η ανάπτυξη κάθε προσέγγισης επαναδειγματοληψίας έγινε ταυτόχρονα με την κατασκευή των μοντέλων ταξινόμησης των μοντέλων. Στη συνέχεια θα αναλυθούν τα μοντέλα ταξινόμησης που αναπτύχθηκαν καθώς και οι επιδόσεις που σημείωσαν.

5.2.5 Ανάπτυξη μοντέλων ταξινόμησης

Ορισμένοι αλγόριθμοι ταξινόμησης κατασκευάστηκαν για τον προσδιορισμό του επιπέδου "Συμπεριφορά Οδήγησης" που τοποθετείται σε κάθε χρονικό πλαίσιο 30 δευτερολέπτων του οδηγού, όπως περιγράφεται λεπτομερώς στις προηγούμενες ενότητες. Η βιβλιογραφική έρευνα οδήγησε στην επιλογή τεσσάρων μοντέλων. Η ορολογία και ο συμβολισμός των μοντέλων εξηγούνται στον πίνακα 5.6.

Πίνακας 5.6: Ονοματολογία και συμβολισμός μοντέλων ταξινόμησης

Όνομα μοντέλου (ελληνικά)	Όνομα μοντέλου (αγγλικά)	Συμβολισμός μοντέλου
μέθοδος παλινδρόμησης κορυφογραμμής	RidgeClassifier	RID
μηχανής διανυσματικής υποστήριξης	SupportVectorMachines	SVM
μοντέλο τυχαίων δασών	RandomForestClassifier	RF
μοντέλο ακραίας ενίσχυση κλίσης	XgBoost	XG

Για την καλύτερη επεξεργασία τον δεδομένων στον ΄RidgeClassifier’ τα δεδομένα επεξεργάστηκαν με τον ‘StandardScaler’ οπου είναι μια σημαντική τεχνική που εκτελείται κυρίως ως βήμα προ επεξεργασίας πριν από πολλά μοντέλα μηχανικής μάθησης, προκειμένου να τυποποιηθεί το εύρος της λειτουργικότητας του συνόλου δεδομένων εισόδου. Παρόλα αυτά τα αποτελέσματα δεν ήταν ικανοποιητικά.

Ειδικότερα η εκπαίδευση όλων τον μοντέλων έγινε με τα ‘Smote’ δεδομένα για να μπορέσουν με καλύτερο τρόπο να προβλέψουν της τρεις κατηγορίες που θέσαμε. Στη συνέχεια μετά την εκπαίδευση τους προβλέψαμε τα αρχικά δεδομένα.

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

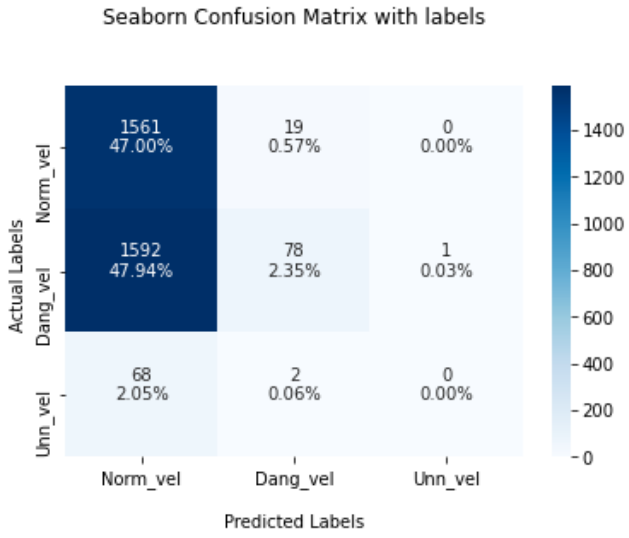
Παρακάτω παρατίθενται για της ομάδες (A) και (B) που έχουμε θέση οι μήτρες σύγχυσης για την γραφική αναπαράσταση της επίδοσης κάθε αλγορίθμου. Επίσης παρουσιάζονται οι μετρικές αξιολόγησης που προέκυψαν μετά την εξέταση του κάθε μοντέλου στο τέλος και τα αποτελέσματα συγκρίθηκαν μεταξύ τους.

Ομάδα (A)

1. Μέθοδος παλινδρόμησης κορυφογραμμής (RID)

Όπως φαίνεται στο γράφημα 5.7 ο αλγόριθμός RID έχει την δυνατότητα να προβλέψει το επίπεδο ‘Normal’ με μέτρια ποσότητα πρόβλεψης και τα επίπεδα ‘Avoidable Accident’ και ‘Dangerous’ με πολύ χαμηλή ευστοχία. Επομένως συνολικά

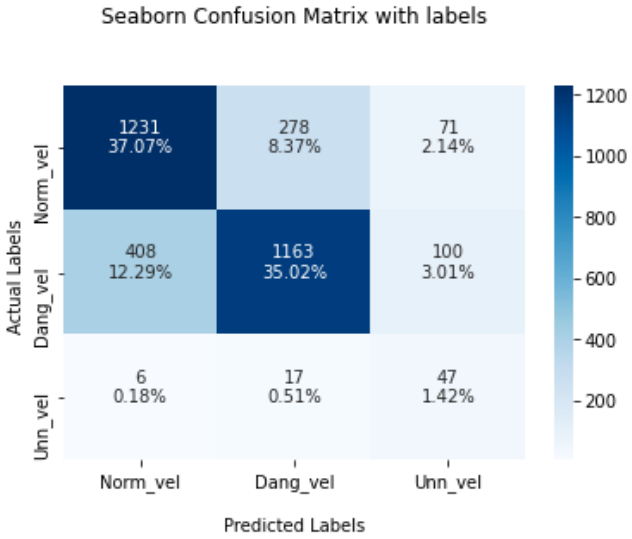
θεωρείται ένα μοντέλο με πολύ χαμηλή ικανότητά αναγνώρισης επικίνδυνων συμπεριφορών.



Γράφημα 5.7: Μήτρα σύγχυσης αλγόριθμου RID

2. Μηχανής διανυσματικής υποστήριξης (SVM)

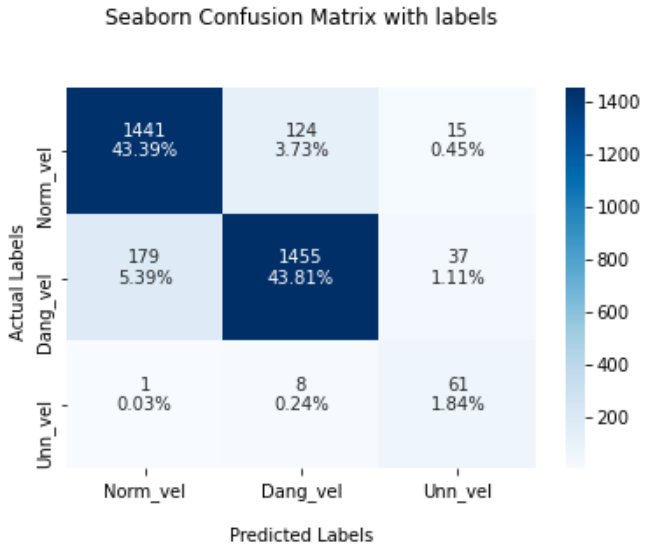
Ο αλγόριθμος (SVM) έχει την δυνατότητα να προβλέψει τα τρία επίπεδα με 73.5% ευστοχία που θεωρείτε ικανοποιητικό αποτέλεσμα. Ειδικότερα τα λανθασμένα αποτελέσματα στο επίπεδο ‘Dangerous’ είναι ιδιαίτερα χαμηλό με τιμή 8.88% όπως **φαίνεται** στο γράφημα 5.8.



Γράφημα 5.8: Μήτρα σύγχυσης αλγόριθμου SVM

3. Μοντέλο τυχαίων δασών (RF)

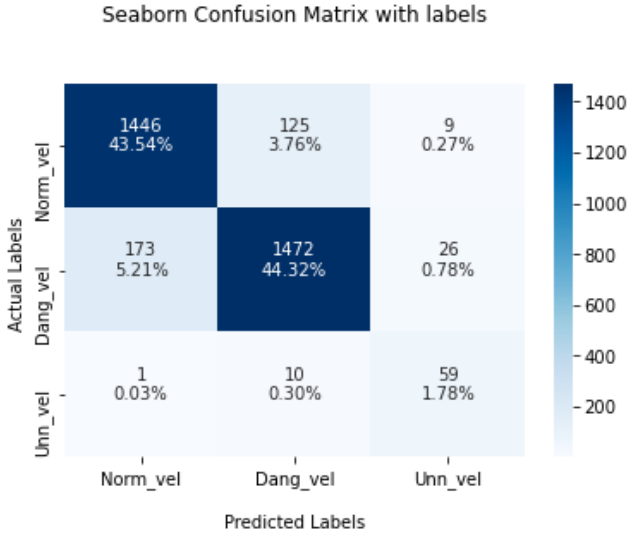
Ο αλγόριθμος (RF) παρουσιάζει εξαιρετικά αποτελέσματα στην πρόβλεψη και τον τριών επίπεδόν με ευστοχία 90% και πολύ χαμηλά ποσοστά λάθους όπως φαίνεται στο γράφημα 5.9.



Γράφημα 5.9: Μήτρα σύγχυσης αλγόριθμου RF

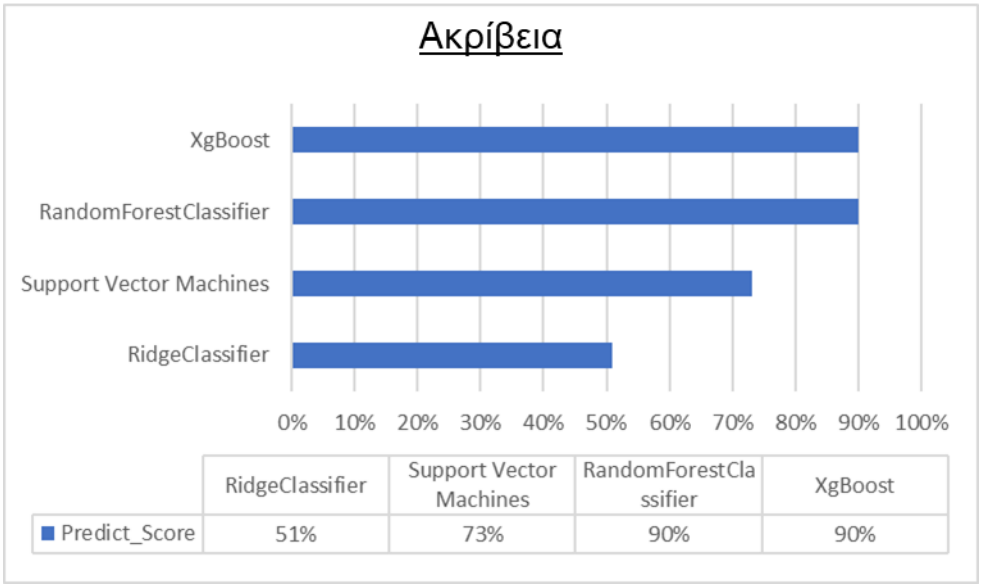
4. Μοντέλο λογισμικού ανοιχτού κώδικα (XG)

Το μοντέλο αυτό παρουσιάζει εξίσου εξαιρετικά αποτελέσματα προβλέψεις όπως το Μοντέλο τυχαίων δασών (RF) με 90% ευστοχία όπως φαίνεται στον πίνακα 5.10



Γράφημα 5.10: Μήτρα σύγχυσης αλγόριθμου XG

Συγκρίνοντας της ευστοχίας της ομάδας (Α) όπως φαίνεται στον πίνακα 5.11 βλέπουμε πως τα μοντέλα τυχαίων δασών (RF) και λογισμικού ανοιχτού κώδικα (XG) έχουν τα καλύτερα και παρόμοια αποτελέσματα

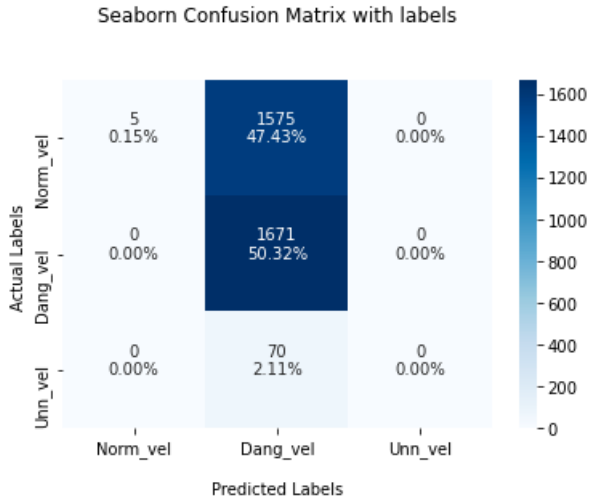


Πίνακας 5.11 ποσοστά ακρίβειας μοντέλων

Ομάδα (B)

1. Μέθοδος παλινδρόμησης κορυφογραμμής (RID)

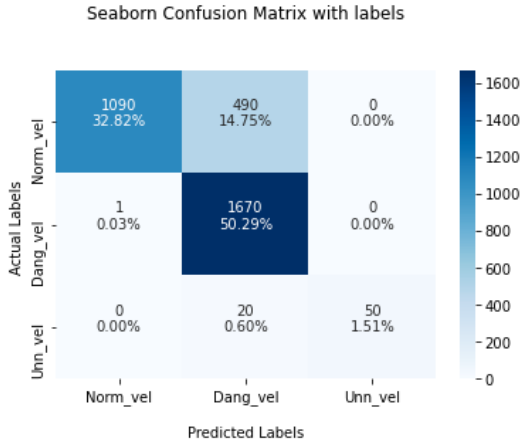
Το μοντέλο (RID) για την ομάδα (B) έχει παρόμοια αποτελέσματα με την ομάδα (A) ευστοχίας με χαμηλή απόδοση όπως φαίνεται στον πίνακα 5.12. Άρα βλέπουμε πως γενικότερα δεν μπορεί έχει την δυνατότητα να ικανοποίησή την πρόβλεψη των τριών επιπέδων.



Γράφημα 5.12: Μήτρα σύγχυσης αλγόριθμου RID

2. Μηχανής διανυσματικής υποστήριξης (SVM)

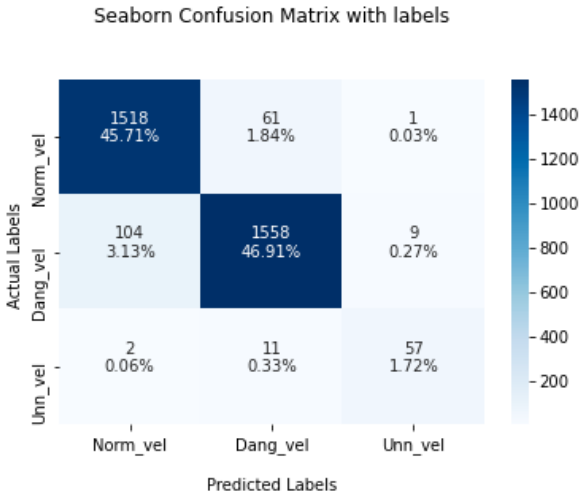
Το μοντέλο Μηχανής διανυσματικής υποστήριξης (SVM) προσφέρει καλύτερα αποτελέσματα στην ομάδα (B) με ευστοχία 85% όπως φαίνεται στον πίνακα 5.13 από την ομάδα (A) άλλα δεν μπορεί να προβλέψει το επίπεδο 'Avoidable Accident'.



Γράφημα 5.13: Μήτρα σύγχυσης αλγόριθμου SVM

3. Μοντέλο τυχαίων δασών (RF)

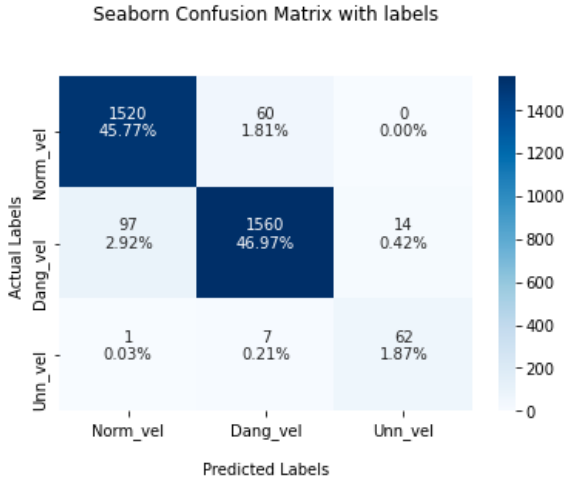
Το μοντέλο αυτό στην ομάδα (B) μπορεί και προσφέρει ακόμα καλύτερα αποτελέσματα πρόβλεψής με ευστοχία 95% από την ομάδα (A) όπως φαίνεται στον πίνακα 5.14.



Γράφημα 5.14: Μήτρα σύγχυσης αλγόριθμου RF

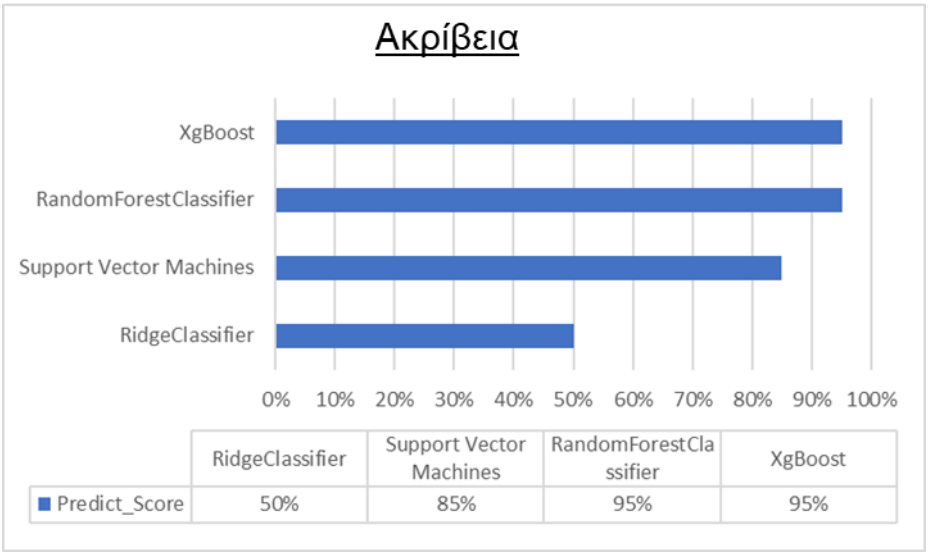
4. Μοντέλο λογισμικού ανοιχτού κώδικα (XG)

Ο αλγόριθμος (XG) έχει παρόμοια αποτελέσματα με το μοντέλο τυχαίων δασών (RF) και καλύτερα από την ομάδα (A) όπως φαίνεται στον πίνακα 5.15.



Γράφημα 5.15: Μήτρα σύγχυσης αλγόριθμου XG

Συνοπτικά βλέπουμε στον πίνακα 5.16 τα αποτελέσματα ευστοχίας όλων τον μοντέλων στην ομάδα (B) και μπορούμε να πούμε πως τα μοντέλα τυχαίων δασών (RF) και λογισμικού ανοιχτού κώδικα (XG) έχουν την καλύτερη δυνατότητα πρόβλεψής. Επίσης από της δύο ομάδες (A) και (B) παρατηρούμε πως Παρόλου που στην ομάδα (A) έχουμε μόνο της ποιο σημαντικές μεταβλητές η ομάδα (B) παράγει καλύτερα αποτελέσματα με την χρήση περισσότερων. Άρα είναι σημαντικό να αναφερθεί πως με την χρήση περισσότερων στοιχείων μπορούμε να προσφέρουμε ακόμα καλύτερα αποτελέσματα.

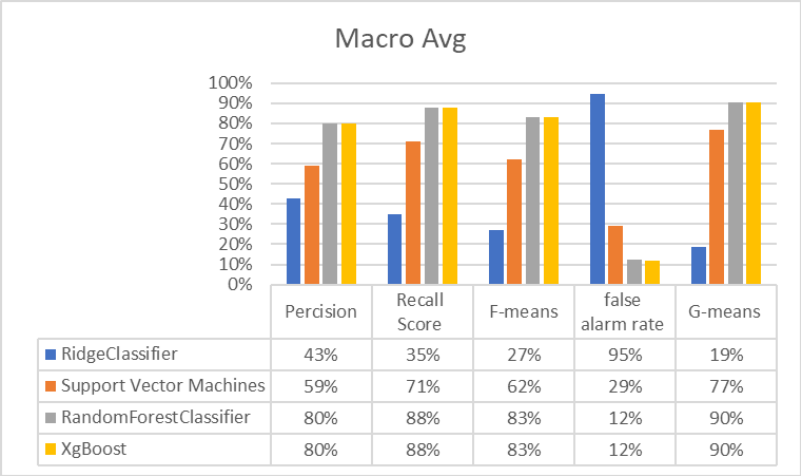


Πίνακας 5.16 ποσοστά ακρίβειας μοντέλων

5.2.6 Σύγκριση μετρικών αξιολόγησης των μοντέλων

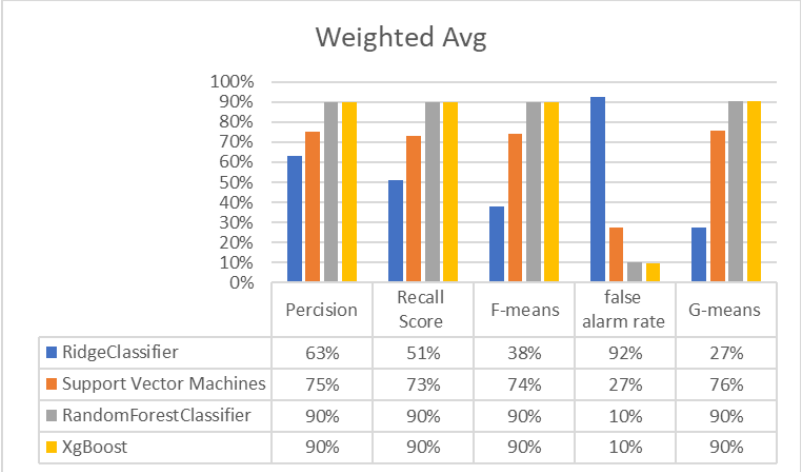
Στη συνέχεια αξιολογήσαμε τα μοντέλα και τον δύο ομάδων με τον τρόπο που αναφέραμε στο κεφάλαιο 3.5 και τα αποτελέσματα φαίνονται παρακάτω στον πίνακα 5.17, 5.18, 5.19, 5.20. Οι διαφορετικές τεχνικές επεξεργασίας των δεδομένων καθώς και η βελτιστοποίηση των παραμέτρων των αλγορίθμων είχαν ως στόχο την βελτίωση της προγνωστικής ικανότητας των μοντέλων.

Ομάδα (A)



	Percision	Recall Score	F-means	false alarm rate	G-means
RidgeClassifier	43%	35%	27%	95%	19%
Support Vector Machines	59%	71%	62%	29%	77%
RandomForestClassifier	80%	88%	83%	12%	90%
XgBoost	80%	88%	83%	12%	90%

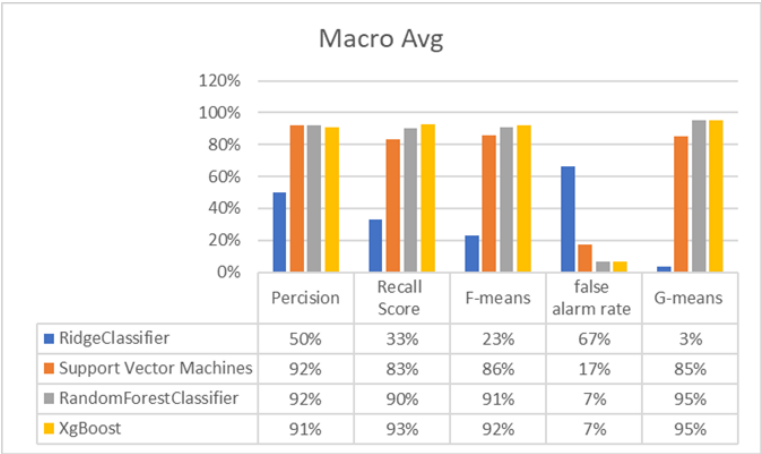
Γράφημα 5.17: Επίδοση των μοντέλων ταξινόμησης σύμφωνα με ορισμένες μετρικές αξιολόγησης



	Percision	Recall Score	F-means	false alarm rate	G-means
RidgeClassifier	63%	51%	38%	92%	27%
Support Vector Machines	75%	73%	74%	27%	76%
RandomForestClassifier	90%	90%	90%	10%	90%
XgBoost	90%	90%	90%	10%	90%

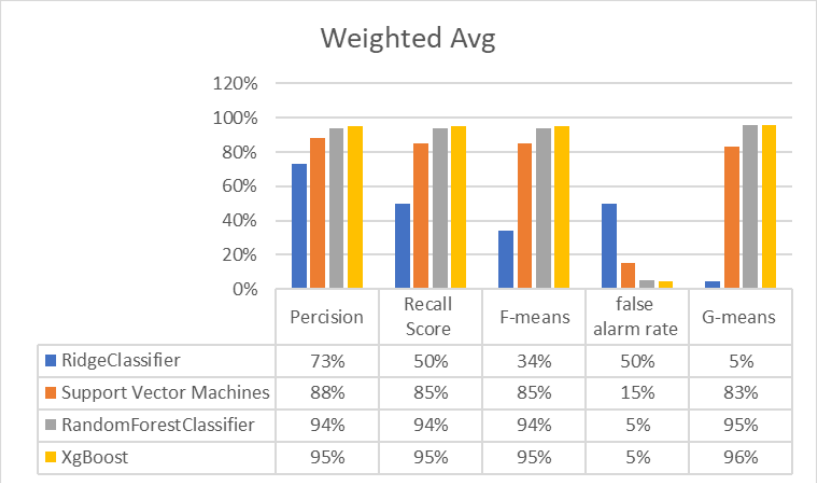
Γράφημα 5.18: Επίδοση των μοντέλων ταξινόμησης σύμφωνα με ορισμένες μετρικές αξιολόγησης

Ομάδα (B)



	Percision	Recall Score	F-means	false alarm rate	G-means
RidgeClassifier	50%	33%	23%	67%	3%
Support Vector Machines	92%	83%	86%	17%	85%
RandomForestClassifier	92%	90%	91%	7%	95%
XgBoost	91%	93%	92%	7%	95%

Γράφημα 5.19: Επίδοση των μοντέλων ταξινόμησης σύμφωνα με ορισμένες μετρικές αξιολόγησης



	Percision	Recall Score	F-means	false alarm rate	G-means
RidgeClassifier	73%	50%	34%	50%	5%
Support Vector Machines	88%	85%	85%	15%	83%
RandomForestClassifier	94%	94%	94%	5%	95%
XgBoost	95%	95%	95%	5%	96%

Γράφημα 5.20: Επίδοση των μοντέλων ταξινόμησης σύμφωνα με ορισμένες μετρικές αξιολόγησης

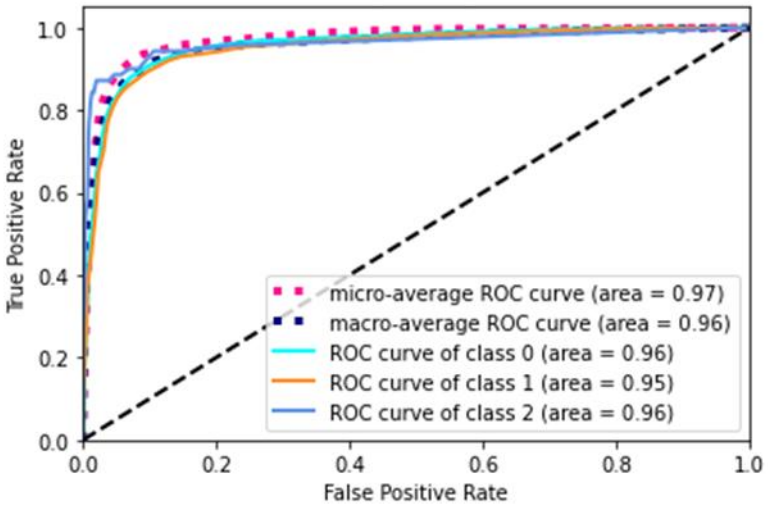
Για κάθε ένα από τα αποτελέσματα βρέθηκε ο μέσος όρος και ο μέσος όρος με βάρος στις κατηγορίες με τα περισσότερα στοιχεία έτσι ώστε να φανεί αν κάποιο από τα επίπεδα συνεισφέρει περισσότερο στα αποτελέσματα και σε τι βαθμό.

Σύμφωνα με τα αποτελέσματα που βλέπουμε παραπάνω οι αλγόριθμοι σημειώνουν υψηλά αποτελέσματα ανάκλησης (recall), ακρίβεια (precision), Μέτρο F (F-measure) και Μέτρο G (G-means) στης δύο ομάδες (A),(B) εκτός από τον αλγόριθμός RID και ειδικότερα όταν λαμβάνουμε υπόψη το βάρος που έχει κάθε επίπεδο βλέπουμε πως προσφέρουν ακόμα καλύτερη ακρίβεια. Ειδικότερα ο Random Forest Classifier και ο XgBoost παρουσίασαν την καλύτερη αξιολόγηση και την καλύτερη ακρίβεια.

Στη συνέχεια για τα μοντέλα Random Forest Classifier και ο XgBoost εξετάστηκε η καμπύλη ROC δηλαδή η γραφική παράσταση του αληθινού θετικού ποσοστού έναντι του ψευδούς θετικού ποσοστού για τον ταξινομητή.

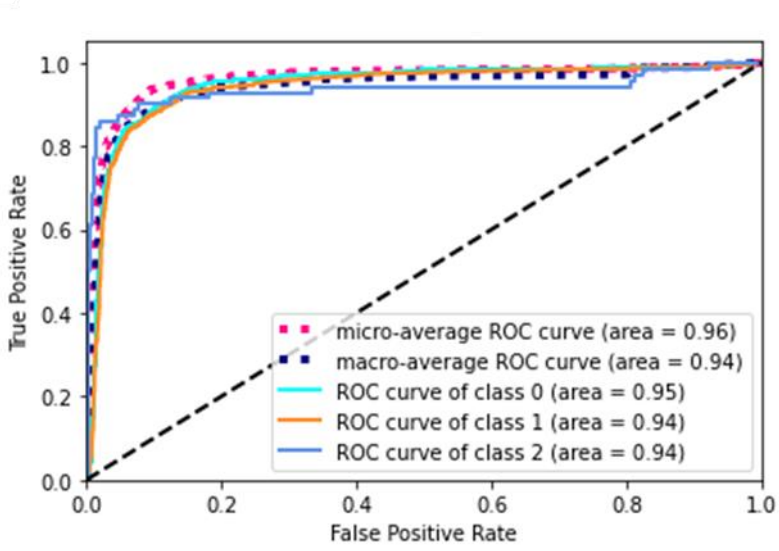
Ομάδα (Α)

Random Forest Classifier



Εικόνα 5.21: Καμπύλη ROC για RF

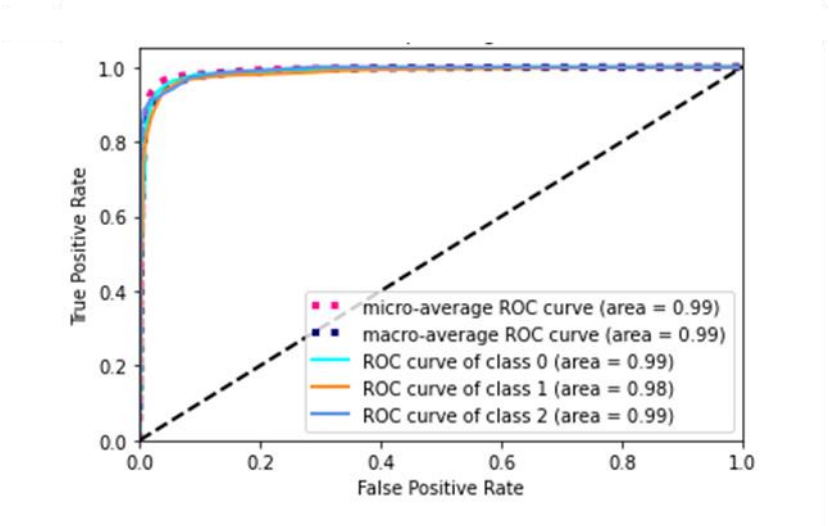
XgBoost



Εικόνα 5.22: Καμπύλη ROC για XG

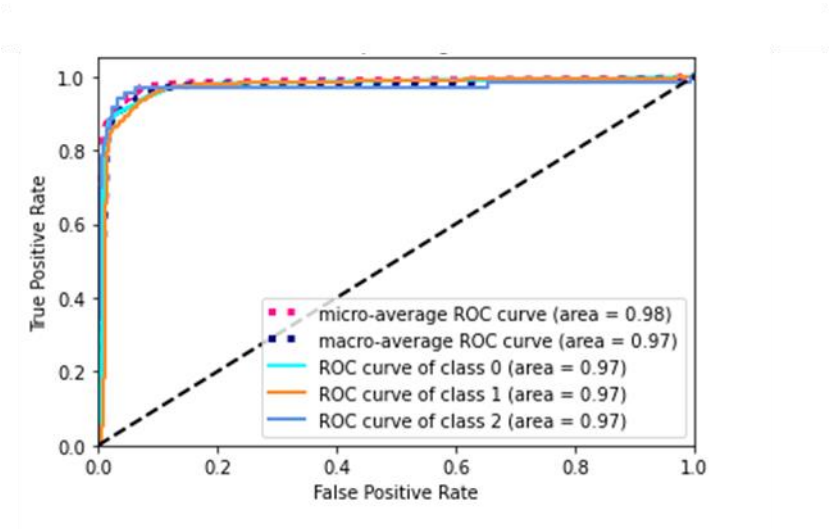
Ομάδα (Β)

Random Forest Classifier



Εικόνα 5.23: Καμπύλη ROC για RF

XgBoost



Εικόνα 5.24: Καμπύλη ROC για XG

Και στις δύο ομάδες (Α) και (Β) παρουσίασε πολύ ικανοποιητικά αποτελέσματα και για της τρεις κατηγορίες συμπεριφοράς οδήγησης έχοντας την δυνατότητα να της ξεχωρίσει με ακρίβεια. Συμπερασματικά είδαμε πως στην ομάδα (Β) στην οποία συμπεριλάβαμε περισσότερες μεταβλητές παρουσιάστηκαν καλύτερα αποτελέσματα στα μοντέλα που τρέξαμε. Οπότε καταλαβαίνουμε πως για να μπορούμε να προβλέψουμε με όσο την δυνατό καλύτερη ακρίβεια είναι αναγκαίο να έχουμε μεγάλη γκάμα στοιχείων. Ειδικότερα τα μοντέλα Random Forest Classifier και

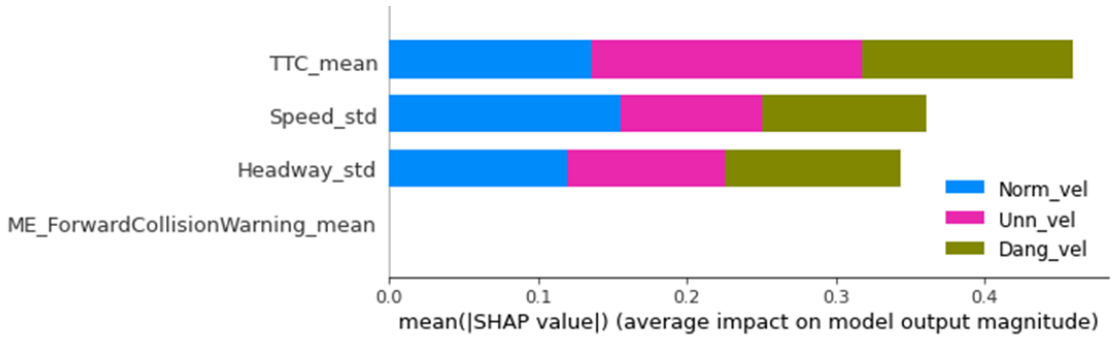
ο XgBoost παρουσίασαν την καλύτερη δυνατότητα εντοπισμού της επικίνδυνης οδήγησης με πολύ παρόμοια αποτελέσματα.

5.3 Εξήγηση λειτουργείας μοντέλων μηχανικής μάθησης

Τέλος θέλοντας να κατανοήσουμε τα μοντέλα μηχανικής μάθησης καλύτερα εξετάσαμε της τιμές SHAP (SHapley Additive exPlanations). Όπου το SHAP χρησιμοποιείται για να εξηγήσει πώς κάθε χαρακτηριστικό επηρεάζει το μοντέλο και επιτρέπει την τοπική και σφαιρική ανάλυση για το σύνολο δεδομένων και το συγκεκριμένο πρόβλημα. Τα μοντέλα Random Forest Classifier και ο XgBoost είναι αυτά με την καλύτερη ακρίβεια για τα οποία προέκυψαν τα διαγράμματα στον πιν για της ομάδες (A) και (B). Το διάγραμμά παρουσιάζει το πόσο συνεισφέρει η κάθε μεταβλητή το μοντέλο και σε τι βαθμό η κάθε κατηγορία στην οποία ανήκει.

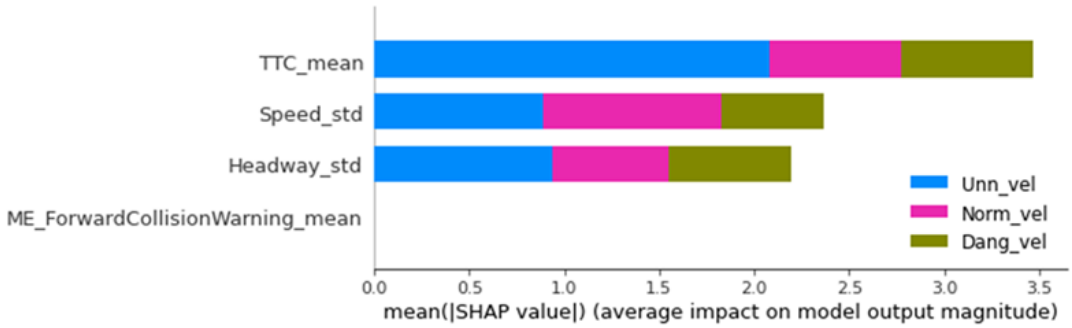
Ομάδα (A)

Random Forest Classifier



Εικόνα 1.25 Διάγραμμα τιμών Shaply για RF

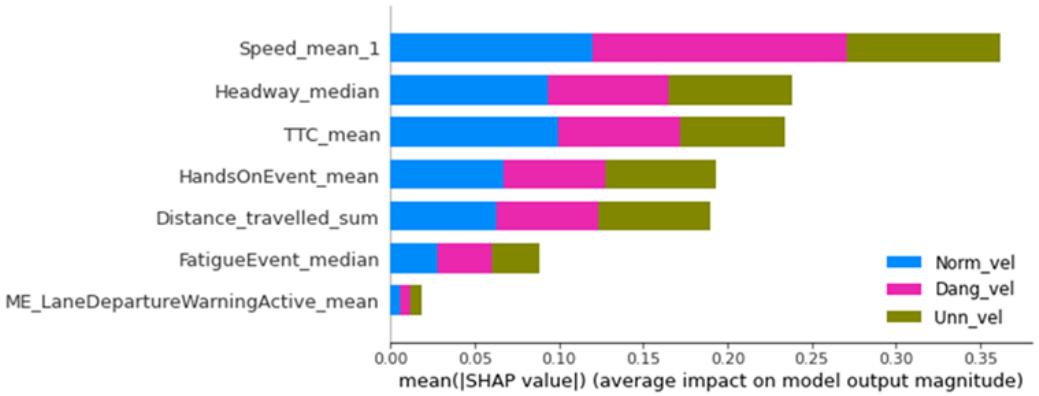
XgBoost



Εικόνα 2.26 Διάγραμμα τιμών Shaply για XG

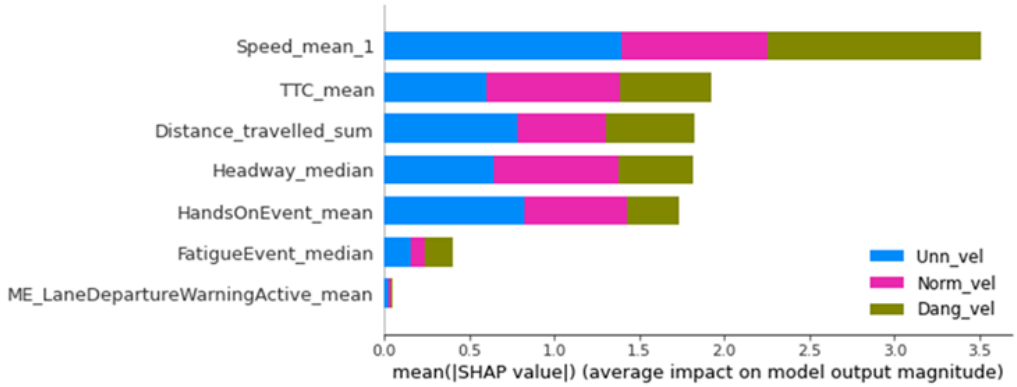
Ομάδα (B)

Random Forest Classifier



Εικόνα 3.27 Διάγραμμα τιμών Shaply για RF

XgBoost



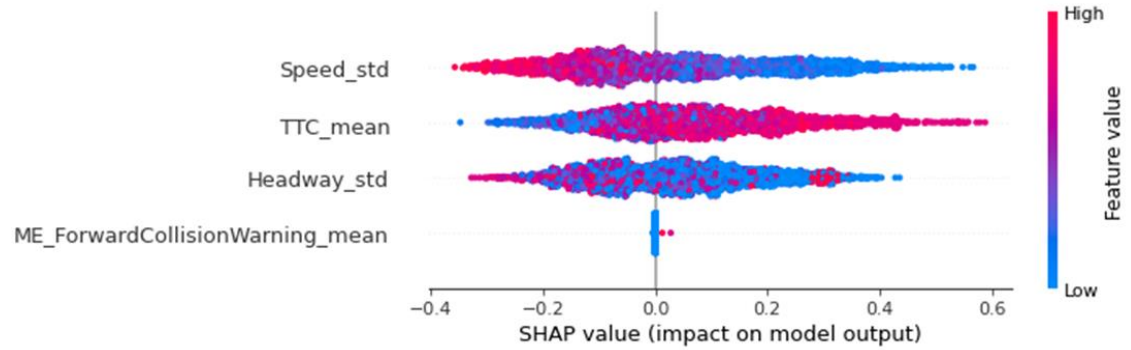
Εικόνα 4.28 Διάγραμμα τιμών Shaply για XG

Στη συνέχεια παρουσιάζονται τα διαγράμματα στα οποία φαίνονται τα χαρακτηριστικά που ωθούν την πρόβλεψη υψηλότερα εμφανίζονται με κόκκινο χρώμα, ενώ εκείνα που ωθούν την πρόβλεψη χαμηλότερα με μπλε χρώμα. Για την Ομάδα (A) και (B)

Ομάδα (Α)

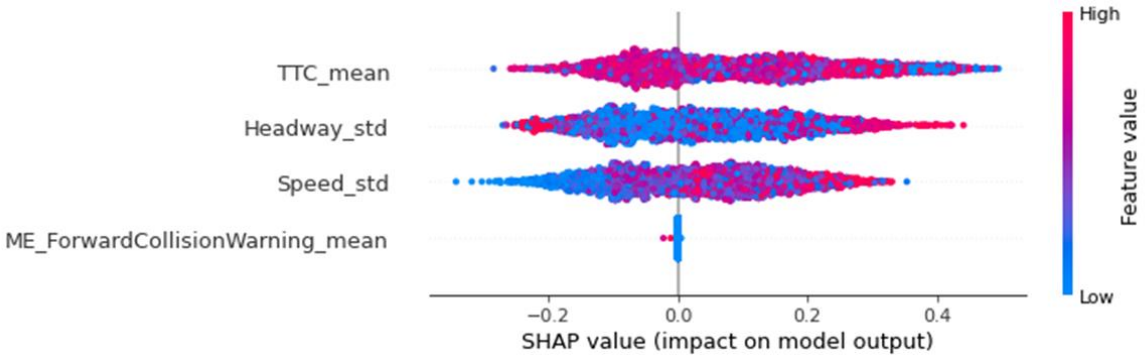
Random Forest Classifier

Κατηγορία: Normal (0)



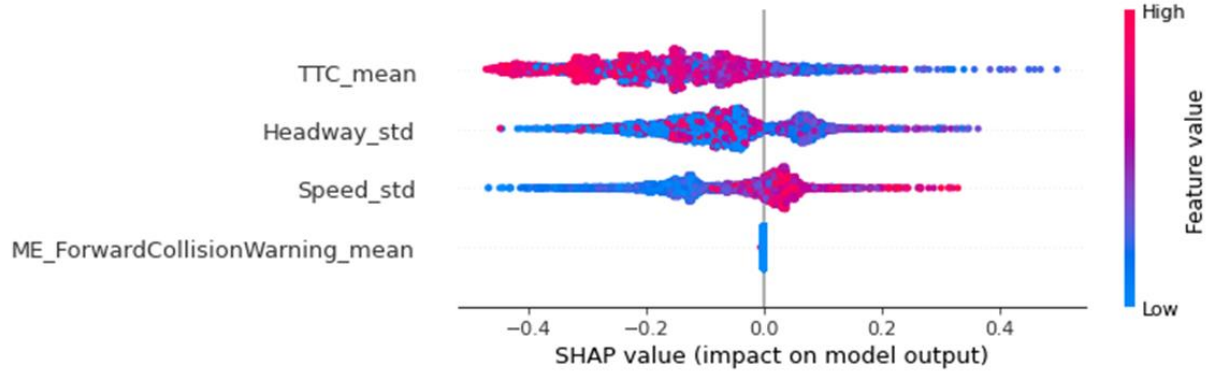
Εικόνα 5.29 Διάγραμμα τιμών Shaply για Κατηγορία: Normal (0)

Κατηγορία: Dangerous (1)



Εικόνα 6.30 Διάγραμμα τιμών Shaply για Κατηγορία: Dangerous (1)

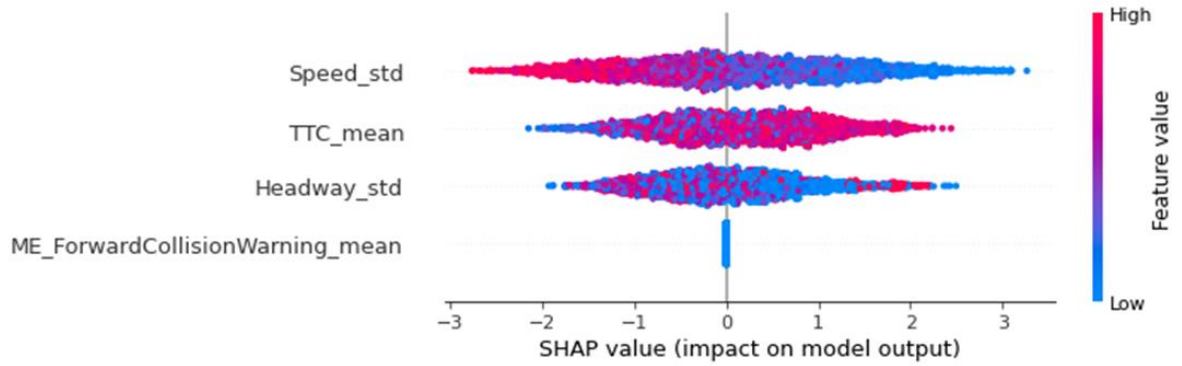
Κατηγορία: Unn_Acc (2)



Εικόνα 7.31 Διάγραμμα τιμών Shaply για Κατηγορία: Unn_Acc (2)

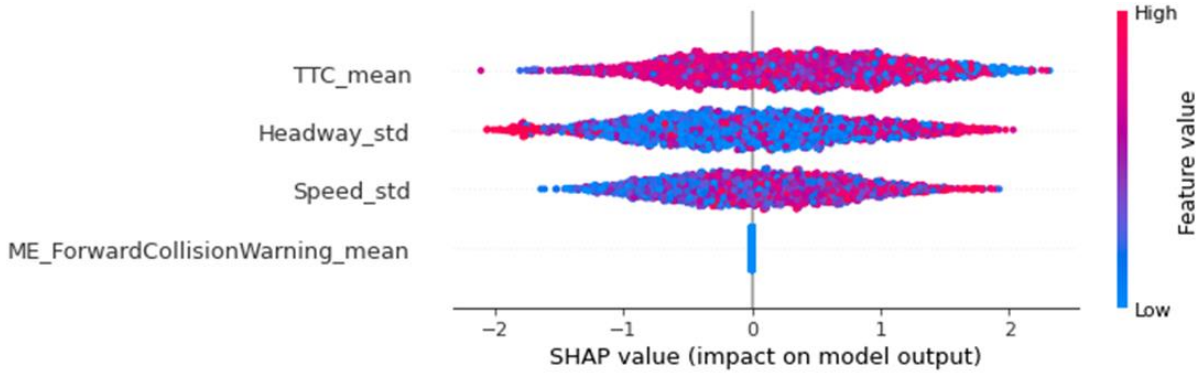
XgBoost

Κατηγορία: Normal (0)



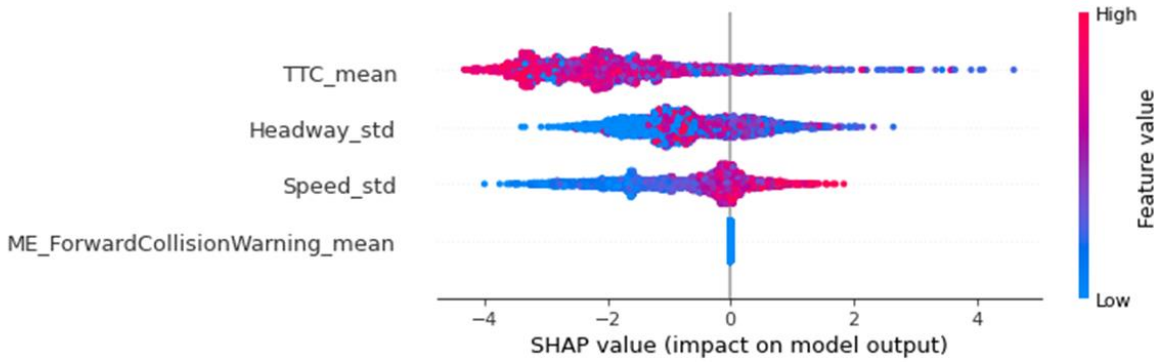
Εικόνα 8.32 Διάγραμμα τιμών Shaply για Κατηγορία: Normal (0)

Κατηγορία: Dangerous (1)



Εικόνα 9.33 Διάγραμμα τιμών Shaply για Κατηγορία: Dangerous (1)

Κατηγορία: Unn_Acc (2)



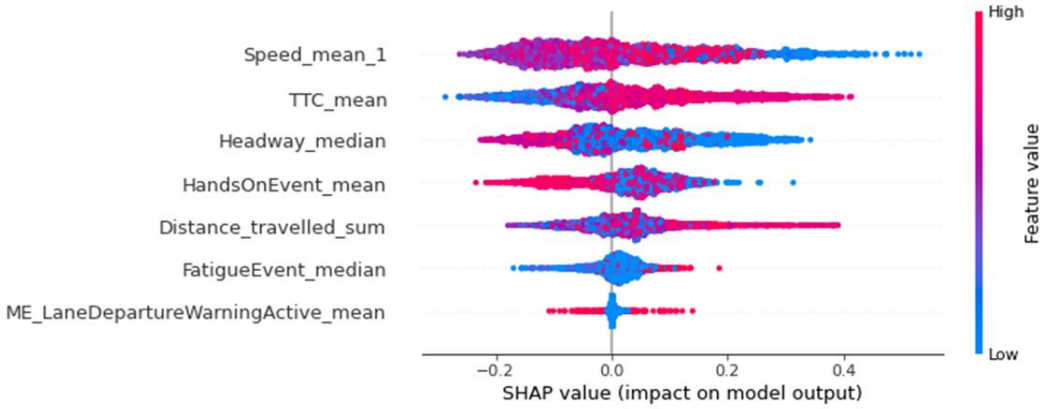
Εικόνα 10.34 Διάγραμμα τιμών Shaply για Κατηγορία: Unn_Acc (2)

Για την ομάδα (A) παρατηρήθηκε πως το TTC_mean όσο ποιο μικρές τιμές έχει τόσο περισσότερο ωθεί στην αλλαγή κατηγορίας και ειδικότερα στην Dangerous (1) και Unn_Acc (2) είναι η κύρια μεταβλητή που επηρεάζει το μοντέλο και για τα δύο μοντέλα. Το headway_std βλέπουμε πως στην κατηγορία Normal (0) όσο μικρότερη τιμή έχει ωθεί στην ποιο επικίνδυνη οδήγηση και ότι στις κατηγορίες Dangerous (1) και Unn_Acc (2) θεωρείτε ποιο σημαντική για το μοντέλο. Τέλος το Speed_std επηρεάζει την κατηγορία Normal (0) με πολύ μεγαλύτερο βαθμό από της υπόλοιπες και το ME_ForwarCollisionWarning_mean μα δίνει τιμές σε κλίμακα η οποία δεν επηρεάζει τα μοντέλα σε βαθμό που μπορούμε να παρατηρήσουμε.

Ομάδα (B)

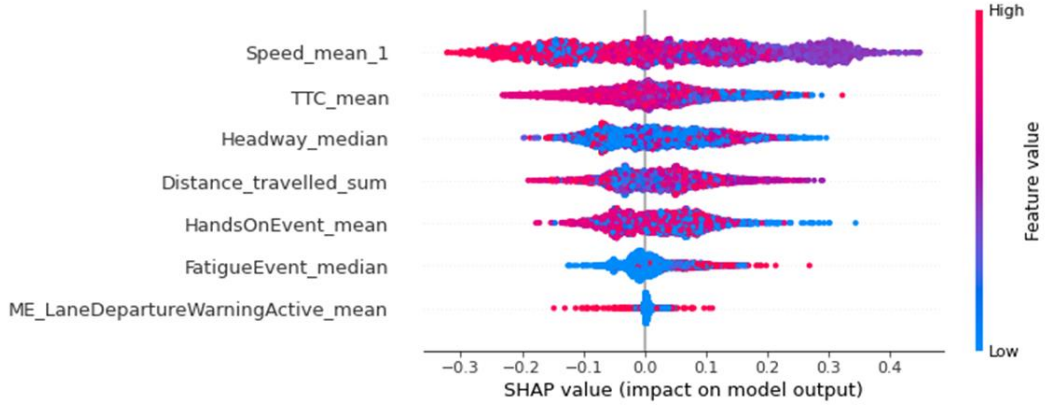
<u>Random Forest Classifier</u>

Κατηγορία: Normal (0)



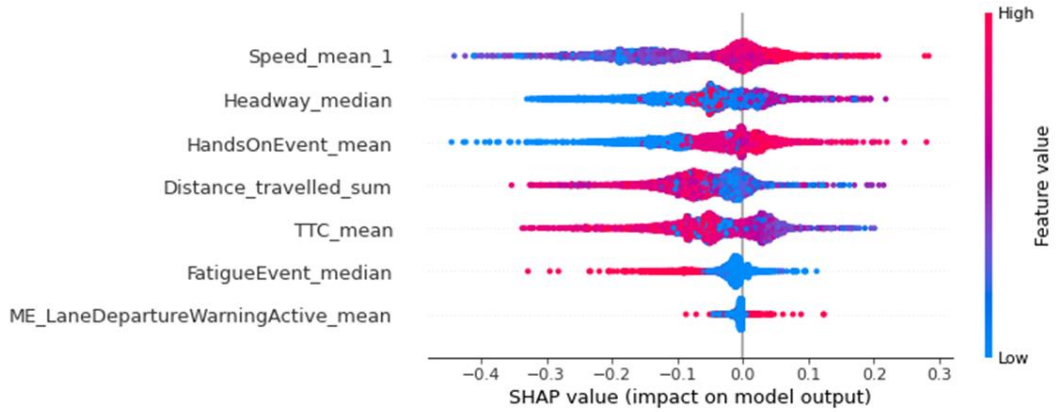
Εικόνα 11.35 Διάγραμμα τιμών Shaply για Κατηγορία: Normal (0)

Κατηγορία: Dangerous (1)



Εικόνα 12 Εικόνα 13.36 Διάγραμμα τιμών Shaply για Κατηγορία: Dangerous (1)

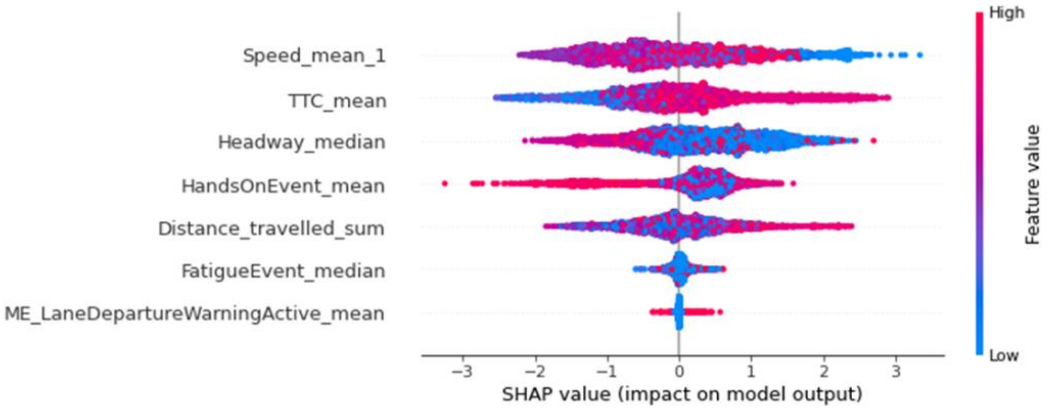
Κατηγορία: Unn_Acc (2)



Εικόνα 14.37 Διάγραμμα τιμών Shaply για Κατηγορία: Unn_Acc (2)

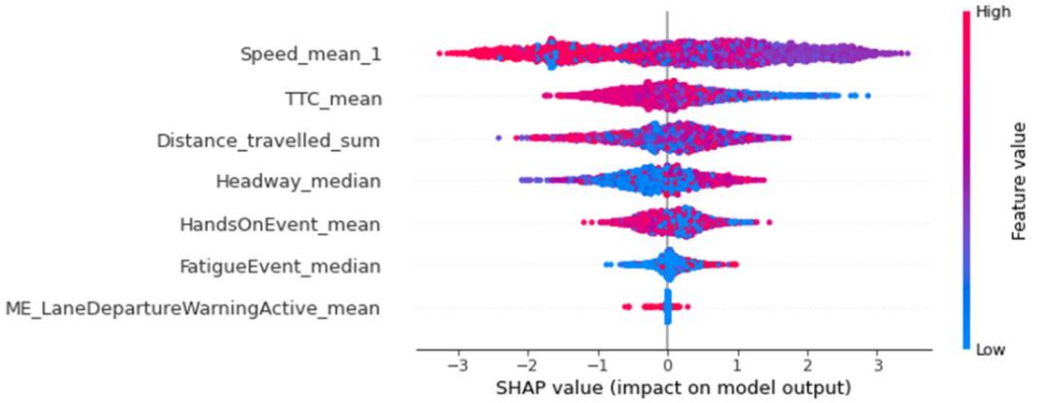
<u>XgBoost</u>

Κατηγορία: Normal (0)



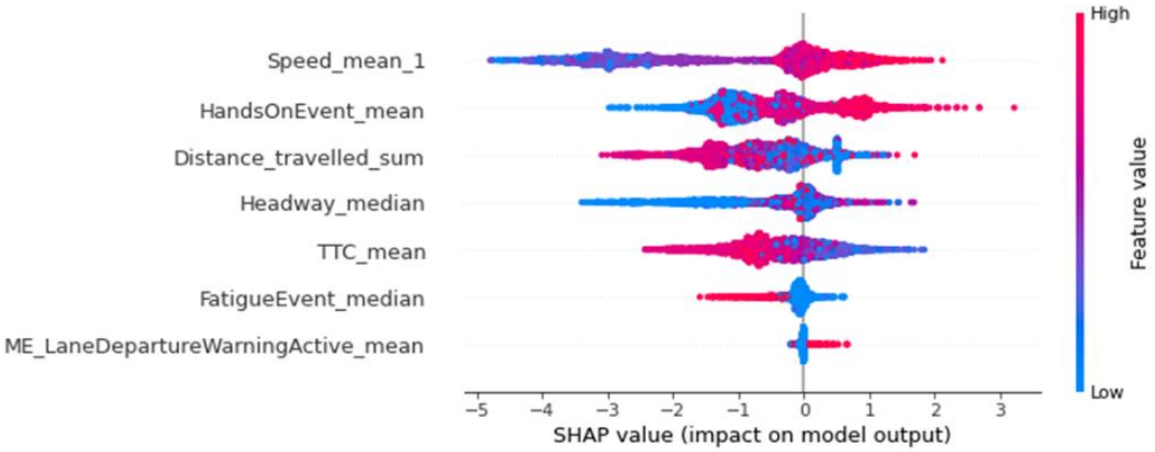
Εικόνα 15.37 Διάγραμμα τιμών Shaply για Κατηγορία: Normal (0)

Κατηγορία: Dangerous (1)



Εικόνα 16.38 Διάγραμμα τιμών Shaply για Κατηγορία: Dangerous (1)

Κατηγορία: Unn_Acc (2)



Εικόνα 17.39 Διάγραμμα τιμών Shaply για Κατηγορία: Unn_Acc (2)

Για την ομάδα (B) παρατηρούμε ότι το Speed_mean_1 είναι η ποιο σημαντική μεταβλητή για τα μοντέλα και όσο μεγαλύτερη τιμή έχει τόσο μεγαλύτερο ρόλο έχει στην ωθήσει σε ποιο επικίνδυνη οδήγηση. Η μεταβλητές HandsOnEvent_mean, TTC_Mean και Headway_meadian όταν έχουν μικρές τιμές επίσης ωθούν στην αλλαγή κατηγορίας ενώ η Distance_travelled_sum παροτρύνει στην αλλαγή με όσο μεγαλύτερη απόσταση έχει διανύσει ο οδηγός. Τέλος βλέπουμε πως οι μεταβλητές FatigueEvent_meadian ai ME_LaneDEpartureWarningActive_mean Παρόλου που δεν συνεισφέρουν σε μεγάλο βαθμό στην αλλαγή της οδηγικής συμπεριφοράς παρατηρούμε το ότι όσο ποιο ξεκούραστος είναι ο οδηγός τόσο λιγότερο επικίνδυνα οδηγεί.

Προκειμένου να διευρυνθεί η γνώση των στοιχείων που επηρεάζουν τον εντοπισμό της μη ασφαλούς οδήγησης, είναι απαραίτητο να αναλυθούν διεξοδικά τα ευρήματα και τα αποτελέσματα που προέκυψαν. Για να δημιουργηθεί το υπόβαθρο για περαιτέρω μελέτη, είναι σημαντικό να τονιστούν τόσο οι κύριες συνεισφορές της έρευνας όσο και οι θεμελιώδεις αδυναμίες.

5.4 Σύνοψη

Στο κεφάλαιο αυτό εφαρμόστηκε ένα σύστημα με το οποίο μπορούμε να εντοπίσουμε την **οδηγική** συμπεριφορά ενός οδηγού και να την κατατάξουμε σε τρεις κατηγορίες. Εξετάστηκαν οι παράγοντες οι οποίοι επηρεάζουν την επικίνδυνη οδήγηση και αναπτύχθηκαν τέσσερεις αλγόριθμοι ταξινόμησης και εξετάστηκε η επίδραση των διαφορετικών μεταβλητών στο σύνολο αυτών μέσω ορισμένων τεχνικών επιλογής και επεξεργασίας στοιχείων. Δημιουργήθηκαν δύο ομάδες με μεταβλητές (A): TTC_mean, Headway_std,, Speed_std, ME_ForwardCollisionWarning_mean και η (B): TTC_mean, Headway_median, HandsOnEvent_mean,FatigueEvent_median,ME_LaneDepartureWarningActive_mean,Speed_mean_1,Distance_travelled_sum οι οποίες επιδρούν άμεσα στην αναγνώρισής της οδηγηκής συμπεριφοράς.

Τα μοντέλα ταξινόμησης στο σύνολο τους είχαν ικανοποιητικά αποτελέσματα για την αναγνώριση του επιπέδου της ΄Συμπεριφοράς Οδήγησης του Οδηγού’ 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 ‘Λογισμικού ανοιχτού κώδικα ’ (Xgboost) είχαν τα καλύτερα αποτελέσματα για το σύνολο των επιπέδων ασφάλειας.

Επίσης μέσω της μελέτης και της κατανόησης με την χρήση του Shap τον μοντέλων με τα καλύτερα αποτελέσματα μπορέσαμε να δούμε το πόσο επηρεάζει κάθε μεταβλητή το κάθε μοντέλο και σε τη βαθμό. Ειδικότερα είδαμε πως στην ομάδα (A) η ταχύτητα και ο χρόνος προσκρούσεις είναι οι ποιο σημαντικές μεταβλητές που και η μεταβολές τους ωθούν τον οδηγώ σε ποιο επικίνδυνη οδήγηση ενώ στην ομάδα (B) παρατηρήσαμε και άλλες μεταβλητές οι οποίες επηρεάζουν τον οδηγώ σε εξίσου σημαντικό βαθμό.

Προκειμένου να διευρυνθεί η γνώση των στοιχείων που επηρεάζουν τον εντοπισμό της μη ασφαλούς οδήγησης, είναι απαραίτητο να αναλυθούν διεξοδικά τα ευρήματα και τα αποτελέσματα που προέκυψαν. Για να δημιουργηθεί το υπόβαθρο για περαιτέρω μελέτη, είναι σημαντικό να τονιστούν τόσο οι κύριες συνεισφορές της έρευνας όσο και οι θεμελιώδεις αδυναμίες.

6. ΣΥΜΠΕΡΑΣΜΑΤΑ

6.1 Σύνοψη Αποτελεσμάτων

Στόχος της παρούσας διπλωματικής εργασίας είναι ο εντοπισμός επικίνδυνης συμπεριφοράς οδηγού με δεδομένα ευρείας κλίμακας από έξυπνα συστήματα καταγραφής και τεχνικές μηχανικής μάθησης. Για το λόγο αυτό αναπτύχθηκαν μοντέλα στατιστικής ανάλυσης και διερευνήθηκε η ικανότητά τους να προβλέπουν 3 κατηγόριες οδηγών.

Τα δεδομένα προέκυψαν από το ερευνητικό έργο i-DREAMS, στο οποίο συμμετείχαν 36 οδηγοί 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. Από τα δεδομένα αυτά δημιουργήθηκε ένας πίνακας όπου περιέχει στοιχεία για την κατάσταση του οδηγού και την κατάσταση του αυτοκίνητου κατά την διάρκεια της οδήγησης. Οι πίνακες αυτοί χρησιμοποιήθηκαν στην στατιστική ανάλυση των δεδομένων και, χρησιμοποιώντας ως εξαρτημένη μεταβλητή το Speed_max, ξεχωρίστηκαν τρεις κατηγορίες:

- Φυσιολογική Οδήγηση (class: 0):
  Μέγιστη Ταχύτητα <= 0,8* Τρέχον όριο ταχύτητας
- Επικίνδυνη Οδήγηση (class: 1):
  0,8* Τρέχον όριο ταχύτητας <= Μέγιστη Ταχύτητα <= Τρέχον όριο ταχύτητας
- Οδήγηση Αποφεύγοντας Ατύχημα (class: 2):
  Μέγιστη Ταχύτητα > =Τρέχον όριο ταχύτητας

Αρχικά για την ανάλυση των δεδομένων και για την εύρεση των σημαντικότερων μεταβλητών χρησιμοποιήθηκαν κατάλληλες μέθοδοί. Στη συνέχεια διαχωρίστηκαν οι μεταβλητές με την μεγαλύτερη συσχέτιση και χωρίστηκαν σε δύο ομάδες. Η (Α) στην οποία περιέχονται μόνο οι μεταβλητές με την μεγαλύτερη συσχέτιση και η (Β) στην οποία εμπεριέχονται επιπλέον και μεταβλητές οι οποίες θεωρούνται σημαντικές αλλά όχι στον ίδιο βαθμό με τις άλλες.

(A) [TTC_mean, Headway_std, Speed_std, ME_ForwardCollisionWarning_mean]

(B) [TTC_mean, Headway_median, HandsOnEvent_mean, FatigueEvent_median, ME_LaneDepartureWarningActive_mean, Speed_mean_1, Distance_travelled_sum]

Επισημαίνεται ότι η μεταβλητή Speed_max δεν λήφθηκε υπόψη στο πρώτο μέρος των αναλύσεων, καθώς θα αναπτύσσονταν προβλήματα μεροληψίας των μοντέλων ταξινόμησης.

Στη συνέχεια εφαρμόζοντας την SMOTE τεχνική (μία συνθετική τεχνική υπερδειγματοληψίας μειοψηφίας), η οποία μέσω της βιβλιογραφικής ανασκόπησης έδειξε τα καλύτερα αποτελέσματα, επιλύθηκε το πρόβλημα άνισης κατανομής των δεδομένων εκπαίδευσης στις διαφορετικές κλάσεις.

Αναπτύχθηκαν τέσσερεις αλγόριθμοι μηχανικής εκμάθησης με σκοπό την ταξινόμηση των οδηγών σε μία από τις τρεις κατηγορίες. Τα ονόματα και οι συμβολισμοί των τεσσάρων αλγορίθμων παρατίθενται στον πίνακα που ακολουθεί.

Όνομα μοντέλου (ελληνικά)	**Όνομα μοντέλου (αγγλικά)**	**Συμβολισμός μοντέλου**
μέθοδος παλινδρόμησης κορυφογραμμής	RidgeClassifier	RID
μηχανής διανυσματικής υποστήριξης	SupportVectorMachines	SVM
μοντέλο τυχαίων δασών	RandomForestClassifier	RF
μοντέλο λογισμικού ανοιχτού κώδικα	XgBoost	XG

Ελέγχθηκαν και αξιολογήθηκαν τα αποτελέσματα των καλύτερων μοντέλων για την ομάδα (B) όπως φαίνεται στους παρακάτω πίνακες.

Ομάδα (B)

Πίνακας : Σύνοψη μοντέλου RandomForestClassifier και XgBoost

Random Forest Classifier	Ορθότητα	Ανάκληση	F1-score	false alarm rate	G-means
Μέσος όρος	92%	90%	91%	7%	95%
Σταθμισμένος Μέσος όρος	94%	94%	94%	5%	95%
XgBoost					
Μέσος όρος	91%	93%	92%	7%	95%
Σταθμισμένος Μέσος όρος	95%	95%	95%	5%	96%

Πίνακας ποσοστά ακρίβειας μοντέλων

	Ακρίβεια
RandomForestClassifier	95%
XgBoost	95%

Τέλος, για να έχουμε μια επισκόπηση των χαρακτηριστικών που είναι πιο σημαντικά για ένα μοντέλο, μπορούμε να χρησιμοποιήσουμε τις τιμές SHAP για κάθε χαρακτηριστικό, για κάθε δείγμα. Ειδικότερα προέκυψε για την ομάδα (B) ότι η πιο σημαντικές μεταβλητές που επηρεάζουν το μοντέλο RandomForestClassifier είναι η ταχύτητα, η απόσταση από το μπροστά αμάξι, ο χρόνος πρόσκρουσης και η ένδειξη

ότι τα χέρια του οδηγού βρίσκονται στο τιμόνι. Ενώ για το XgBoost σημαντικότερα είναι η ταχύτητα, ο χρόνος πρόσκρουσης, η απόσταση που διένυσε, η απόσταση από το μπροστά αμάξι και η ένδειξη ότι τα χέρια του οδηγού βρίσκονται στο τιμόνι. Συμπερασματικά παρατηρούμε πως και στα δύο μοντέλα η ταχύτητα είναι η ποιο σημαντική μεταβλητή αλλά στην συνέχεια βλέπουμε πως οι υπόλοιπες επηρεάζουν με διαφορετικό βαθμό το κάθε μοντέλο.

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.

- Η μέθοδος Random Forest και η μέθοδος XgBoost από την ομάδα (B) σημείωσαν τις υψηλότερες επιδόσεις στην πλειοψηφία των μετρικών αξιολόγησης τους, με ακρίβεια 95%.
- Σημαντικό να αναφερθεί είναι το ότι βρέθηκε recall score 90% για το μοντέλο Random Forest και 94% για το μοντέλο XgBoost, με αποτέλεσμα τον καλύτερο εντοπισμό των τριών κατηγοριών συμπεριφοράς οδήγησης.
- Οι πιο σημαντικές μεταβλητές που ωθούν τον οδηγό στην πιο επικίνδυνη οδήγηση είναι η ταχύτητα, ο χρόνος απόστασης από το επόμενο όχημα και το χρονικό διάστημα που έχει ο οδηγός τα χέρια του στο τιμόνι.
- 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 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).
- Η μέθοδοι Ridge Classifier και Support Vector Machines δεν παρουσίασαν τόσο ικανοποιητικά αποτελέσματα όσο τα υπόλοιπα μοντέλα
- Και στα τέσσερα μοντέλα η ομάδα με τον μεγαλύτερο αριθμό μεταβλητών παρουσίασε πιο ικανοποιητικά αποτελέσματα.
- Τέλος από την εκπόνηση της συγκεκριμένης Διπλωματικής Εργασίας, προκύπτει ότι τα δεδομένα που συλλέγονται από τα έξυπνα συστήματα και περεταίρω έρευνες περιέχουν ιδιαίτερα σημαντικές πληροφορίες οι οποίες, μετά από κατάλληλη επεξεργασία και ανάπτυξη μαθηματικών μοντέλων, μπορούν να χρησιμεύσουν στην εξαγωγή χρήσιμων συμπερασμάτων για τις κρίσιμες παραμέτρους που επηρεάζουν την συμπεριφορά του οδηγού κατά

τη διάρκεια οδήγησης αλλά και για τη γενικότερη κυκλοφοριακή συμπεριφορά των οδηγών.

6.3 Προτάσεις για αξιοποίηση των αποτελεσμάτων

Καταβάλλεται προσπάθεια να παρουσιαστούν ορισμένες προτάσεις για τη χρήση των ευρημάτων, οι οποίες μπορούν να συμβάλουν στη βελτίωση της κατανόησης των επιπτώσεων διαφόρων παραγόντων στην οδική ασφάλεια και στην προώθηση της έρευνας για τα ευφυή συστήματα μεταφορών (ITS). Οι προτάσεις αυτές βασίζονται στις διαπιστώσεις και τα συμπεράσματα που προέκυψαν κατά την εκπόνηση της παρούσας μελέτης.

- Χρήση μοντέλων κατηγοριοποίησης για τον προσδιορισμό του επιπέδου ασφάλειας ενός οδηγού υπό πραγματικές συνθήκες οδήγησης. Από τις επιδόσεις των τεσσάρων αλγορίθμων ταξινόμησης προκύπτει ότι μπορούν να αποδώσουν καλά αποτελέσματα και θα μπορούσαν να εφαρμοστούν σε περαιτέρω έρευνα σχετικά με την οδηγική συμπεριφορά.
- Μια βαθύτερη εξέταση των σημαντικότερων μεταβλητών που επηρεάζουν τον εντοπισμό της επικίνδυνης οδηγικής συμπεριφοράς Με τον τρόπο αυτό, θα υποστηρίξουμε τις προσπάθειες της επιστημονικής κοινότητας και της αυτοκινητοβιομηχανίας για την ενίσχυση των εξελιγμένων συστημάτων υποβοήθησης του οδηγού.
- Η δημιουργία ενός κατάλληλου συστήματος για τον προσδιορισμό της "Οδηγηκής συμπεριφοράς" όπου βρίσκεται ο οδηγός στο όχημα σε πραγματικό χρόνο. Ως αποτέλεσμα, ο οδηγός θα είναι σε θέση να παρακολουθεί το δρόμο και να αναλαμβάνει δράση σε περίπτωση παραβίασης της ζώνης ασφαλείας.
- Την ανάπτυξη μιας εφαρμογής για έξυπνα τηλέφωνα που θα συλλέγει τις πληροφορίες και θα προβλέπει πόσο χρόνο θα περάσει ο οδηγός σε κάθε επίπεδο από της τρεις κατηγορίες που έχουμε θέση για την οδηγηκή συμπεριφορά. Με αυτόν τον τρόπο, ο οδηγός θα μπορεί να κάνει τις απαιτούμενες προσαρμογές στον τρόπο οδήγησής του μετά τη λήξη της.
- Εξοπλισμός των οχημάτων με συστήματα που αξιοποιώντας τον αλγόριθμο θα μπορούν να προβλέψουν την ύπαρξη συμβάντος και θα εμφανίζουν προειδοποιητικό μήνυμα στον οδηγό να εκτελέσει κατάλληλες ενέργειες για την αποφυγή του.

6.4 Προτάσεις για περαιτέρω έρευνα

Οι σύγχρονες τεχνικές επεξεργασίας και ανάλυσης χρησιμοποιούνται όλο και περισσότερο στον τομέα της οδικής ασφάλειας. Οι ερευνητές ενδιαφέρονται πολύ για τη μελέτη της οδηγικής συμπεριφοράς με τη χρήση τεχνικών μηχανικής μάθησης. Στην έρευνα που αξιολογήθηκε αναδύθηκαν πολλές δυσκολίες. Οι ερευνητές συνέστησαν πρόσθετα στοιχεία και τεχνικές για την αντιμετώπισή τους. Μελετώντας διάφορες προσεγγίσεις μηχανικής μάθησης, η παρούσα μελέτη είχε ως στόχο να καλύψει το κενό που προέκυψε από τη βιβλιογραφική ανασκόπηση και να αποτελέσει τη βάση για περαιτέρω συγκριτικές αξιολογήσεις. Ωστόσο, κατά τη διάρκεια της κατασκευής της προσέγγισης και της αξιολόγησης των αποτελεσμάτων ανακαλύφθηκαν αρκετές αδυναμίες που θα πρέπει να ληφθούν υπόψη σε μελλοντική μελέτη. Οι συστάσεις για πρόσθετη μελέτη που παρατίθενται κατωτέρω θα μπορούσαν να μας βοηθήσουν να αποκτήσουμε βαθύτερη και εμπεριστατωμένη γνώση των προβλημάτων που επισημάνθηκαν.

- 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.
- 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.
- Εξέταση επιπλέον τύπων οδού με οδηγούς για την και σύγκριση των αποτελεσμάτων.
- 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.
- Ανάλυση μοντέλων βαθιάς μάθησης. Η βαθιά μάθηση συνεπάγεται τη χρήση αλγορίθμων μηχανικής μάθησης με περίπλοκη δομή που διαμορφώνεται σύμφωνα με τον ανθρώπινο εγκέφαλο. Η βαθιά μάθηση εξαλείφει την ανάγκη για χειροκίνητη ανίχνευση χαρακτηριστικών στα δεδομένα. Για την εύρεση σχετικών μοτίβων στα παραδείγματα εισόδου, αντί να βασίζεται σε οποιαδήποτε διαδικασία εκπαίδευσης. Η διαδικασία επισπεύδεται με αυτόν τον τρόπο, γεγονός που παράγει καλύτερα αποτελέσματα.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1]Kallidoni, Marianthi. “National Road Safety Strategic Plan 2021-2030.” *NRSO*, www.nrso.ntua.gr, https://www.nrso.ntua.gr/national-road-safety-strategic-plan-2021-2030/. Accessed 5 June 2022.

[2]Meier, Fred. “What Does ADAS Mean? | News | Cars.Com.” *Cars.Com*, www.cars.com, 3 June 2022, https://www.cars.com/articles/what-does-adas-mean-442753/.

[3]Winkelbauer, Martin, et al. “Naturalistic Driving | SpringerLink.” *Naturalistic Driving|SpringerLink*,link.springer.com, https://link.springer.com/chapter/10.1007/978-3-642-15503-1_15. Accessed 5 June 2022.

[4]Farrag, Siham G., et al. “A Microsimulation-Based Analysis for Driving Behaviour Modelling on a Congested Expressway - Journal of Ambient Intelligence and Humanized Computing.” *SpringerLink*, link.springer.com, 20 May 2020, https://link.springer.com/article/10.1007/s12652-020-02098-5.

[5]“Deep Learning-Based Traffic Safety Solution for a Mixture of Autonomous and Manual Vehicles in a 5G-Enabled Intelligent Transportation System.” *Deep Learning-Based Traffic Safety Solution for a Mixture of Autonomous and Manual Vehicles in a 5G-Enabled Intelligent Transportation System*, ieeexplore.ieee.org, https://ieeexplore.ieee.org/abstract/document/9303409. Accessed 29 June 2022.

[6]Peppes, Nikolaos, et al. “Sensors | Free Full-Text | Driving Behaviour Analysis Using Machine and Deep Learning Methods for Continuous Streams of Vehicular Data | HTML.” *MDPI*, www.mdpi.com, 9 July 2021, https://www.mdpi.com/1424-8220/21/14/4704/htm.

[7]Tang, Tie-Qiao, and Zhi-Yan Yi. *Modelling the Driving Behaviour at a Signalised Intersection with the Information of Remaining Green Time*. 20 Sept. 2017, https://ietresearch.onlinelibrary.wiley.com/doi/full/10.1049/iet-its.2017.0191.

[8]Lu, Yue, and Xinsha Fu. “XGBoost Algorithm-Based Monitoring Model for Urban Driving Stress: Combining Driving Behaviour, Driving Environment, and Route Familiarity.” *XGBoost Algorithm-Based Monitoring Model for Urban Driving Stress: Combining Driving Behaviour, Driving Environment, and Route Familiarity*, ieeexplore.ieee.org,29 Jan. 2021, https://ieeexplore.ieee.org/abstract/document/9340309.

[9]Marthinus Meiring, Gys Albertus, and Hermanus Carel Myburgh. “Sensors | Free Full-Text | A Review of Intelligent Driving Style Analysis Systems and Related Artificial

Intelligence Algorithms | HTML." *MDPI*, www.mdpi.com, 4 Dec. 2015, https://www.mdpi.com/1424-8220/15/12/29822/htm.

[10]Karri, Soni Lanka, et al. "Classification and Prediction of Driving Behaviour at a Traffic Intersection Using SVM and KNN - SN Computer Science." *SpringerLink*, link.springer.com, 12 Apr. 2021, https://link.springer.com/article/10.1007/s42979-021-00588-7.

[11]Toledo, Tomer. "Browse Journals by Subject." *Driving Behaviour: Models and Challenges*, www.tandfonline.com, 23 Feb. 2007, https://www.tandfonline.com/doi/full/10.1080/01441640600823940?scroll=top&needAccess=true.

[12]Shangguan, Qiangqiang. "An Integrated Methodology for Real-Time Driving Risk Status Prediction Using Naturalistic Driving Data - ScienceDirect." *An Integrated Methodology for Real-Time Driving Risk Status Prediction Using Naturalistic Driving Data - ScienceDirect*, www.sciencedirect.com, 23 Apr. 2021, https://www.sciencedirect.com/science/article/abs/pii/S0001457521001536?via%3Dihub.

[13]Agrawal, Utkarsh. "Towards Real-Time Heavy Goods Vehicle Driving Behaviour Classification in the United Kingdom." *Towards Real-Time Heavy Goods Vehicle Driving Behaviour Classification in the United Kingdom*, ieeexplore.ieee.org, 28 Nov. 2019, https://ieeexplore.ieee.org/abstract/document/8917446.

[14]Oltedal, Sigve. "The Effects of Personality and Gender on Risky Driving Behaviour and Accident Involvement - ScienceDirect." *The Effects of Personality and Gender on Risky Driving Behaviour and Accident Involvement - ScienceDirect*, www.sciencedirect.com, 31 Jan. 2006, https://www.sciencedirect.com/science/article/abs/pii/S0925753505001864.

[15]Verschuur, William L. G., and Karel Hurts. "Modeling Safe and Unsafe Driving Behaviour - ScienceDirect." *Modeling Safe and Unsafe Driving Behaviour - ScienceDirect*, www.sciencedirect.com, 25 Mar. 2008, https://www.sciencedirect.com/science/article/abs/pii/S0001457507001509.

[16]Padurariu, Cristian. "Dealing with Data Imbalance in Text Classification - ScienceDirect." *Dealing with Data Imbalance in Text Classification - ScienceDirect*, www.sciencedirect.com, 14 Oct. 2019, https://www.sciencedirect.com/science/article/pii/S1877050919314152.

[17]Zhu, Shengxue, et al. "Electronics | Free Full-Text | An Optimized Algorithm for Dangerous Driving Behavior Identification Based on Unbalanced Data." *MDPI*, www.mdpi.com, 13 May 2022, https://www.mdpi.com/2079-9292/11/10/1557.

[18]Parsa, Amir Bahador. “Toward Safer Highways, Application of XGBoost and SHAP for Real-Time Accident Detection and Feature Analysis - ScienceDirect.” *Toward Safer Highways, Application of XGBoost and SHAP for Real-Time Accident Detection and Feature Analysis - ScienceDirect*, www.sciencedirect.com, 20 Dec. 2019, https://www.sciencedirect.com/science/article/abs/pii/S0001457519311790.

[19]Mangalathu, Sujith, et al. “Failure Mode and Effects Analysis of RC Members Based on Machine-Learning-Based SHapley Additive exPlanations (SHAP) Approach - ScienceDirect.” *Failure Mode and Effects Analysis of RC Members Based on Machine-Learning-Based SHapley Additive exPlanations (SHAP) Approach - ScienceDirect*, www.sciencedirect.com, 12 June 2020, https://www.sciencedirect.com/science/article/abs/pii/S0141029620307513.

[20]Molnar, Christoph. “8.5 Permutation Feature Importance | Interpretable Machine Learning.” *8.5 Permutation Feature Importance | Interpretable Machine Learning*, christophm.github.io, https://christophm.github.io/interpretable-ml-book/feature-importance.html. Accessed 5 June 2022.

[21]Brownlee, Jason. “A Gentle Introduction to Imbalanced Classification - Machine Learning Mastery.” *Machine Learning Mastery*, machinelearningmastery.com, 22 Dec. 2019, https://machinelearningmastery.com/what-is-imbalanced-classification/.

[22]Brownlee, Jason. “SMOTE for Imbalanced Classification with Python - Machine Learning Mastery.” *Machine Learning Mastery*, machinelearningmastery.com, 16 Jan. 2020, https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/.

[23]“Classification Algorithm in Machine Learning - Javatpoint.” *Www.Javatpoint.Com*, www.javatpoint.com, https://www.javatpoint.com/classification-algorithm-in-machine-learning. Accessed 6 June 2022.

[24]“Support Vector Machine (SVM) Algorithm - Javatpoint.” *Www.Javatpoint.Com*, www.javatpoint.com, https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm. Accessed 7 June 2022.

[25]“What Is XGBoost?” *NVIDIA Data Science Glossary*, www.nvidia.com, https://www.nvidia.com/en-us/glossary/data-science/xgboost/. Accessed 7 June 2022.

[26]Bharathi. “Confusion Matrix for Multi-Class Classification - Analytics Vidhya.” Analytics Vidhya, www.analyticsvidhya.com, 24 June 2021, https://www.analyticsvidhya.com/blog/2021/06/confusion-matrix-for-multi-class-classification/.

[27]"3.3. Metrics and Scoring: Quantifying the Quality of Predictions." *Scikit-Learn*, scikit-learn.org, https://scikit-learn.org/stable/modules/model_evaluation.html. Accessed 7 June 2022.

[28]"Sklearn.Metrics.Precision_score." *Scikit-Learn*, scikit-learn.org, 1 Jan. 2000, https://scikitlearn.org/stable/modules/generated/sklearn.metrics.precision_score.html.

[29]Khanna, Mohit. "Sensitivity, Specificity and Accuracy - Decoding the Relationship." *Analytics Vidhya*, www.analyticsvidhya.com, 22 June 2021, https://www.analyticsvidhya.com/blog/2021/06/classification-problem-relation-between-sensitivity-specificity-and-accuracy/.

[30]"Sklearn.Metrics.F1_score." *Scikit-Learn*, scikit-learn.org, 1 Jan. 2000, https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html.

[31]"G-Means — MeteoInfo 3.3 Documentation." *G-Means — MeteoInfo 3.3 Documentation*,meteothink.org, http://meteothink.org/examples/miml/cluster/gmeans.htm l. Accessed 7 June 2022.

[32]"Micro-Average & Macro-Average Scoring Metrics - Python - Data Analytics." *Data Analytics*, vitalflux.com, 4 Sept. 2020, https://vitalflux.com/micro-average-macro-average-scoring-metrics-multi-class-classification-python/.

[33]"Receiver Operating Characteristic (ROC)." *Scikit-Learn*, scikit-learn.org, https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html. Accessed 7 June 2022.

[34]slundberg, editor. "GitHub - Slundberg/Shap: A Game Theoretic Approach to Explain the Output of Any Machine Learning Model." *GitHub*, github.com, 24 May 2022, https://github.com/slundberg/shap.

[35]Brownlee, Jason. "How to Calculate Correlation Between Variables in Python - Machine Learning Mastery." *Machine Learning Mastery*, machinelearningmastery.com, 26 Apr. 2018, https://machinelearningmastery.com/how-to-use-correlation-to-understand-the-relationship-between-variables/.

[36]Kharwal, Aman. "StandardScaler in Machine Learning." *Data Science | Machine Learning | Python | C++ | Coding | Programming | JavaScript*, thecleverprogrammer.com, 22 Sept. 2020, https://thecleverprogrammer.com/2020/09/22/standardscaler-in-machine-learning/.

[37]"I-DREAMS Project > A Smart Driver and Road Environment Assessment and Monitoring System." *I-DREAMS Project*, idreamsproject.eu, 30 June 2001, https://idreamsproject.eu/wp/.

[38] 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

[39] 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

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

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

[43] Matplotlib: Visualization with Python [WWW Document], 2022. URL https://matplotlib.org/